UNIVERSIDAD NACIONAL DE LOJA

TT-CIS

Área de la Energía, las Industrias y los Recursos Naturales No Renovables

CARRERA DE INGENIERÍA EN SISTEMAS

"Determinación de perfiles profesionales mediante técnicas de minería de datos"

"Trabajo de Titulación previo a la Obtención del título de Ingeniero en Sistemas"

Autora:

María José Rodríguez Ojeda.

Director:

Ing. Edwin René Guamán Quinche, Mg. Sc.

LOJA-ECUADOR

2014

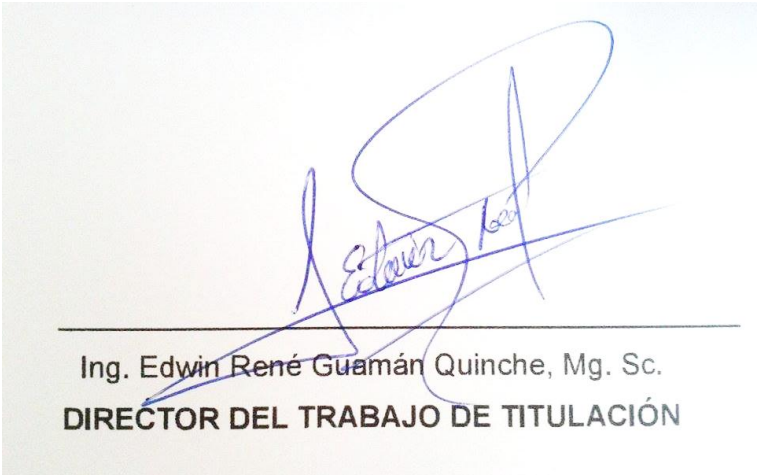
CERTIFICACIÓN DEL DIRECTOR

Ing. Edwin René Guamán Quinche, Mg. Sc.
DOCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS

CERTIFICA:

Que la Srta. María José Rodríguez Ojeda ha trabajado bajo mi tutoría el presente trabajo de titulación, previo a la obtención del título de Ingeniero en Sistemas, cuyo tema versa sobre "DETERMINACIÓN DE PERFILES PROFESIONALES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS", el mismo que ha sido dirigido, orientado y discutido bajo mi asesoramiento y cumple con la reglamentación pertinente, así como lo programado en el plan del proyecto; razones por las cuales reúne la suficiente validez técnica y práctica, por consiguiente autorizo su certificación para su posterior presentación y sustentación.

Loja; julio de 2014



AUTORÍA

Yo María José Rodríguez Ojeda declaro ser autora del presente trabajo de titulación y eximo expresamente a la Universidad de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales por el contenido del mismo.

Adicionalmente acepto y autorizo a la Universidad Nacional de Loja, la publicación de mi trabajo de titulación en el Repositorio Institucional – Biblioteca Virtual.

Autor: María José Rodríguez Ojeda

Firma: 

Cédula: 1105030256

Fecha: 31 de octubre de 2014

CARTA DE AUTORIZACIÓN DE TESIS POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.

Yo MARÍA JOSÉ RODRÍGUEZ OJEDA, declaro ser autora de la tesis titulada: "DETERMINACIÓN DE PERFILES PROFESIONALES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS", como requisito para optar al grado de INGENIERO EN SISTEMAS; autorizo al Sistema Bibliotecario de la Universidad Nacional de Loja para que con fines académicos, muestre al mundo la producción intelectual de la Universidad, a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional:

Los usuarios pueden consultar el contenido de este trabajo en el RDI, en las redes de información del país y del exterior, con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia de las tesis que realice el tercero.

Para constancia de esta autorización, en la ciudad de Loja, 31 días del mes de octubre del dos mil catorce.

Firma: 

Autor: María José Rodríguez Ojeda
Cédula: 1105030256
Dirección: San José Bajo (Ramón Burneo y Pedro Victor Falconí 0209)
Correo Electrónico: mjrodriguezo@unl.edu.ec
Teléfono: 2573982 Celular: 0991129568

DATOS COMPLEMENTARIOS

Director de Tesis: Ing. Edwin René Guamán Quinche, Mg. Sc.

Tribunal de Grado: Ing. Luis Alberto Jácome Galarza, Mg. Sc.
Ing. Henry Patricio Paz Arias, Mg. Sc.
Ing. Gabriela del Cisne Viñan Rueda, Mg. Sc.

DEDICATORIA

Dedico el presente trabajo de titulación a Dios sobre todas las cosas, por ser mi razón de vivir, mi primer pensamiento en los momentos más difíciles, la luz por la que tengo la fuerza para seguir en el camino, a pesar de todo lo que se pueda presentar por más difícil que sea.

A mis padres Robert Rodríguez y Enid Ojeda quienes me han dado la vida, me han cuidado, me han dado todo lo que tengo y me han hecho todo lo que soy. A mis hermanas; Mariutzi y María Fernanda quienes aseguro se alegran por mis momentos de felicidad y me han brindado su apoyo incondicional.

A mis amigos quienes considero mis hermanos los cuales han sido parte fundamental en mi vida; por su apoyo en los momentos de tristeza y gozo en los momentos de felicidad; amigos que perduraron durante todo este tiempo haciéndome saber que son verdaderos.

María José Rodríguez Ojeda

AGRADECIMIENTO

Agradezco en primero lugar a Dios creador de todas las cosas, por darme la vida y habitar en mi corazón, siendo la fuerza que me ha impulsado a seguir adelante. A mis padres Robert Rodríguez y Enid Ojeda por haber sido unos padres responsables en mi educación, por darme todo lo necesario para cumplir con mis metas, por nunca hacerme faltar techo, comida, por haber cuidado de mí y así brindarme la oportunidad de tener un mejor futuro, a mis hermanas Mariutzi y María Fernanda por estar pendientes durante todo el desarrollo del presente trabajo de titulación; incentivándome en todo momento.

A la Universidad Nacional de Loja por brindarme la oportunidad de formarme como profesional; a los docentes que me han contribuido de forma positiva en mi educación; a mi director Ing. René Guamán, y al Ing. Henry Paz por guiarme y ser un apoyo fundamental durante el desarrollo del presente trabajo de titulación.

Y finalmente, de manera especial agradezco a mis amigos quienes han sido clave para la culminación del presente trabajo de titulación, debido a sus palabras de aliento en todo momento, haciéndome saber que creen en mí.

María José Rodríguez Ojeda

a. Título

“DETERMINACIÓN DE PERFILES PROFESIONALES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS”.

b. Resumen

Actualmente en el sector educativo, podemos evidenciar algunas temáticas de vital importancia como es la realidad que enfrentan los profesionales, que año a año se titulan y salen de las universidades con grandes deseos de encontrar el empleo apropiado en donde puedan desarrollarse de forma plena en el ejercicio de su profesión. El problema principal que conlleva el desconocimiento del perfil profesional en un egresado y graduado es que pierde el horizonte del mundo laboral, estando expuesto a desarrollar actividades opuestas a sus verdaderas aspiraciones, desperdiciando así los conocimientos, capacidades, habilidades, intereses desarrollados a lo largo de su carrera profesional y reflejadas en su perfil profesional.

En base a lo descrito, el objetivo principal del trabajo de titulación se fundamenta en la determinación de perfiles profesionales. Para ello se ha realizado el estudio de casos de éxito [1-3], recopilación de técnicas y herramientas de minería de datos, seleccionadas de manera minuciosa y con base en fuentes bibliográficas confiables y la investigación permanente, con la finalidad de crear el escenario adecuado para solucionar el problema planteado y lograr la meta de minería, haciendo uso de la metodología CRISP-DM cumpliendo con todas las fases de la minería de manera exitosa.

Las fuentes de datos utilizadas tienen un enfoque cualitativo obtenido de un test aplicado a la población de interés, respecto de las capacidades, intereses y habilidades de inclinación hacia uno de los perfiles planteados. Y el enfoque cuantitativo que engloban los records académicos, los mismos que fueron obtenidos del Sistema de Gestión Académica (SGA) a través de su Web Services, además se realizó la recopilación de datos históricos para completar estas fuentes de datos, obtenidos de los libros físicos que reposan en la Universidad Nacional de Loja, datos y variables que fueron integradas y tomadas para la construcción de dos estructuras, con el fin de realizar un sin número de pruebas y encontrar el modelo predictivo que posteriormente fue validado mediante su aplicación en un contexto real, con el fin de obtener los mejores resultados que den valor y justifiquen la importancia de la realización del presente trabajo de titulación.

Summary

Currently field of education, we can prove some issues of vital importance as it is the reality faced by practitioners, het are entitled and out of universities with a great desire to find suitable employment where they can be fully develop in the exercise of their profession. The main problem involved the lack of professional profile on graduate is losing the horizon of the working world, I being exposed to develop opposite to their true aspirations activities, wasting knowledge, skills, abilities, interests developed along professional and reflected in his professional career profile.

Based on the described, the main objective of this entitled assignment is based in determining professional profiles. This has made the study of success stories [1-3], collection techniques and data mining tools, selected through way from reliable literature sources and continuous research, in order to create the right stage to solve the problem posed and achieve the goal of mining using the CRISP- DM methodology complying with all phases of mining successfully.

The data sources used have a qualitative approach obtained from a test in the population of interest, of the capabilities, interests and abilities guided toward one of the profiles. And the quantitative approach including academic records , they were obtained from the Academic Management System (EMS) through its Web Services , plus the historical data collection was performed to complete these data sources obtained from physical books lie up at the National University of Loja , data and variables that were integrated and taken to the construction of two structures , in order to perform a number of tests to find the predictive model that subsequently was then validated by applying it in a real context , in order to get the best results that add value and justify the importance achievement of this work.

Índice de Contenidos

Índice General

Índice de Figuras

Índice de Tablas

c. Introducción

En todas las instituciones de nivel superior se busca proporcionar una formación académica de excelencia, a través de docentes capacitados y una malla curricular de un alto nivel comprometida con la obtención de profesionales competentes y preparados para enfrentar los retos y aprovechar las oportunidades del mundo laboral, sin embargo la realidad es que existe un desconocimiento y falta de concienciación por parte de la mayoría de estudiantes, egresados y profesionales sobre cuáles son sus capacidades, intereses, habilidades y conocimientos reflejados en su perfil profesional; realidad que los limita en un futuro a las posibilidades de éxito en el proceso de inserción en el mundo laboral.

En vista del panorama presentado y el estudio de casos de éxito [1-3], la minería de datos, surge como una alternativa de solución en base la aplicación de las técnicas adecuadas que permitan determinar el perfil profesional de cada estudiante, el cual servirá como pauta en la toma de decisiones a nivel académico y profesional. Debido que al conocer su perfil profesional podrán postular y hacer una relación con cierta ocupación y/o vacante, con el fin de mejorar o asegurar sus probabilidades de éxito.

Posteriormente al desarrollo del proceso de minería de datos en base al análisis de fuentes bibliográficas [4-7], se ha obtenido el modelo para la determinación de los perfiles profesionales y se ha realizado un análisis comparativo de los perfiles profesionales obtenidos versus los empleos en los que han incurrido un grupo de la población encuestada, logrando obtener información concluyente y de gran valor que demuestra la importancia de la realización del presente trabajo de titulación.

Además, se elaboró un artículo científico con los resultados obtenidos, para su posterior difusión a la Comunidad Científica y como inicio para nuevos campos de aplicación.

Se debe destacar que la realización del presente trabajo de titulación ha permitido poner en práctica los conocimientos adquiridos a lo largo de la vida académica, siendo a la vez dichos conocimientos útiles en el ámbito profesional.

La Universidad Nacional de Loja y el Área de la Energía, las Industrias y los Recursos

Naturales no Renovables, poseen lineamientos establecidos que rigen la estructura del trabajo de titulación, el cual tiene el siguiente orden: RESUMEN que esla descripción síntetizada de lo que involucra todo el trabajo de titulación; ÍNDICE que describe los temas cada parte de la memoria y su ubicación, así como el índice de tablas y figuras; INTRODUCCIÓN que describe de manera global el ámbito del trabajo de titulación; REVISIÓN LITERARIA que comprende el contenido de la temática haciendo constar revisión bibliográfica de fuentes confiables; METODOLOGÍA donde se describen los recursos materiales, métodos científicos y técnicas de recolección de información empleados; así como la metodología para el cumplimiento de las fases del proyecto; RESULTADOS que contiene el desarrollo de la metodología empleada en el trabajo de titulación, haciendo constar todos los modelos como (descripción, diagramas, figuras, códigos, tablas, entre otros). DISCUSIÓN donde se discute los resultados obtenidos y tiene que ser respaldada con información científica relacionada con el trabajo; CONCLUSIONES siendo las descripciones concretas, experiencias de acuerdo a los objetivos específicos planteados; RECOMENDACIONES engloba los trabajos futuros, los consejos de acuerdo a la experiencia obtenida. Finamente el trabajo de titulación culmina con las FUENTES BIBLIOGRAFÍAS Y ANEXOS.

d. Revisión de Literatura

1. CAPÍTULO I. RECOPILACIÓN DE CASOS DE ÉXITO EN ARTÍCULOS CIENTÍFICOS DE FUENTES ACADÉMICAS, REVISTAS, PONENCIAS, RESPECTO A LA APLICACIÓN DE MINERÍA DE DATOS EN EL ÁMBITO EDUCATIVO.

Algunos modelos basados en el ámbito educativo [4-7] han demostrado mediante la aplicación de técnicas de minería de datos, que se puede obtener valiosa información que ayuda a la toma de decisiones**.** Estas técnicas han contribuido en la educación a detectar factores que influyen en la deserción y abandono de los estudiantes de su vida académica [7], como apoyo a los sistemas de enseñanza a distancia basados en web o sistemas de e-learning [4], contribuyendo a la mejora continua de cursos e-learning [5], en el análisis del rendimiento académico de los estudiantes [6]. Es por estas y un sin número de aplicaciones más que la utilización de estas técnicas es de suma importancia.

Se han recopilado algunos de los casos de éxito mencionados que están específicamente orientados al ámbito educativo, que demuestran el aporte que ha tenido la minería de datos en la educación, y que contribuyen de manera positiva con valiosa información para el desarrollo del presente trabajo de titulación después de un análisis a profundidad.

1.1. Caso de Éxito 1: Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web.

1.1.1. Introducción

Los sistemas basados en Web son cada vez más, es la tecnología más utilizada para la educación a distancia, debido a la facilidad de utilización y disponibilidad de las herramientas para navegar por la Web y la facilidad del desarrollo y mantenimiento de los recursos Web. El desarrollo actual de los sistemas de enseñanza basada en Web o sistemas de e-learning, se ha incrementado exponencialmente en los últimos

años y esto también ha motivado la aplicación de técnicas de minería de datos o descubrimiento de conocimiento como herramientas para poder mejorar el aprendizaje en los sistemas de e- learning [4].

Estas herramientas inteligentes utilizan técnicas de extracción de conocimiento o minería de datos para descubrir información útil para poder mejorar el sistema. Los métodos de descubrimiento de información utilizados en e-learning tienen como objetivo guiar a los estudiantes durante su aprendizaje para maximizarlo [4].

Las principales aplicaciones de las técnicas de minería de datos en educación, son como sistemas de personalización, sistemas recomendadores, sistemas de modificación, sistemas de detección de irregularidades, etc. debido a sus capacidades para: el descubrimiento de patrones de navegación regurales e irregurales, realización de clasificaciones de alumnos y de los contenidos, construcción adaptativa de planes de enseñanza, descubrimiento de relaciones entre actividades, diagnóstico incremental de los estudiantes, etc [4].

1.1.2. Puntos de vista de la aplicación de técnicas de minería de datos en educación.

- **Orientado hacia los autores:**

Con el objetivo de ayudar a los profesores y/o autores de los sistemas de e-learning para que puedan mejorar el funcionamiento o rendimiento de estos sistemas a partir de la información de utilización de los alumnos. Sus principales aplicaciones son: obtener una mayor realimentación de la enseñanza, conocer más sobre como los estudiantes aprenden en el web, evaluar a los estudiantes por sus patrones de navegación, reestructurar los contenidos del sitio web para personalizar los cursos, clasificar a los estudiantes en grupos, etc [4].

- **Orientado hacia los estudiantes:**

Con el objetivo de ayudar o realizar recomendaciones a los alumnos durante su interacción con el sistema de e-learning para poder mejorar su aprendizaje. Sus

principales aplicaciones son: sugerir buenas experiencias de aprendizaje a los estudiantes, adaptación del curso según el progreso del aprendiz, ayudar a los estudiantes dando sugerencias y atajos, recomendar caminos más cortos y personalizados, etc [4].

1.1.3. Educación basada en web

El desarrollo de las nuevas tecnologías de la educación y la comunicación han hecho posible la utilización de Internet y más concretamente la WWW (World Wide Web) en la educación a distancia, dando lugar a la denonimada Educación basada en Web o e-learning [8].

Cada uno los centros de enseñanza que existen en la actualidad utiliza un sistema o plataforma de enseñanza basado en web que puede ser: o bien un sistema propio desarrollado específicamente por ellos mismos, o bien uno de los múltiples sistemas comercial existentes como: Web-CT, Virtual-U, TopClass, etc. o de libre distribución como: ATutor, ILIAS, Moodle, etc. Estos sistemas proporcionan servicios útiles para la enseñanza a distancia como son herramientas para la comunicación sincronía y asíncrona, herramientas para la gestión de materiales de aprendizaje y herramientas para la gestión, seguimiento y evaluación de los estudiantes [4].

Para volver los sistemas anteriormente mencionados más que una red de páginas web estáticas se han desarrollado los Sistemas Hipermedia Adaptativos Basados en Web [9] que son un nuevo tipo de sistemas educativos que provienen de la evolución de los Sistemas Tutores Inteligentes (STI) y de los Sistemas Hipermedia Adaptativos (SHA), y que comparten con ellos características tales como: aumento de la interacción con los usuarios y adaptación de los contenidos a las necesidades de estos. Para ello, construyen un modelo del alumno y lo utilizan durante la interacción con dicho usuario para adaptarse a sus necesidades. Algunos ejemplos de Sistemas Hipermedia Adaptativos basados en Web para educación [9] son: Interbook, DCG, ELM-ART, CALAT, AHA!, etc.

Por último indicar también la existencia y el incremento en la utilización de múltiples estándares de e-learning [28]: IMS, ADL SCORM, AICC, IEEE LTSC, etc. que además de permitir la interoperatibilidad entre distintos sistemas, permiten la

reutilización de contenidos educativos, y también facilitan la incorporación de diferentes técnicas adaptativas [4].

1.1.4. Minería de datos web

La minería de datos es un área multidisciplinar donde convergen diferentes paradigmas de computación como son la construcción de árboles de decisión, la inducción de reglas, las redes neuronales artificiales, el aprendizaje basado en instancias, aprendizaje bayesiano, programación lógica, algoritmos estadísticos, etc. Las principales tareas y métodos de la minería de datos son: clasificación, agrupamiento, estimación, modelado de dependencias, visualización y descubrimiento de reglas [4].

Un caso particular de la minería de datos es la minería de Web o web mining [11], que como el propio nombre indica consiste en la aplicación de técnicas de minería de datos para extraer conocimiento a partir de datos de la Web. Se pueden distinguir tres tipos de minería de Web:

- **Minería de contenidos web.** Es el proceso de extraer información a partir de los contenidos de los documentos Web.
- **Minería de estructura web.** Es el proceso de descubrir información a partir de la estructura de la Web.
- **Minería de utilización web.** Es el proceso de descubrir información a partir de los datos de utilización de la Web.

De estos tres tipos de minería Web, el que más se ha utilizado para el descubrimiento de información en los sistemas de enseñanza basada en web es la minería de utilización Web o web usage mining [4].

1.1.5. Técnicas de minería web más utilizadas en sistemas de e-learning

Las técnicas más utilizadas en la minería de datos aplicada a los sistemas de e-learning son: clasificación y agrupamiento, descubrimiento de reglas de asociación, y análisis de secuencias. A continuación, se van a detallar los principales trabajos de investigación agrupados dentro de estos tres tipos de técnicas, aunque algunos de los investigadores no sólo utilizan una única técnica sino varias [4].

1.2. Caso de Éxito 2: Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning.

1.2.1. Introducción

En los últimos años hemos asistido a un gran incremento de los sistemas de educación on-line o sistemas de e-learning. Cada vez son más los centros de enseñanza públicos o privados que ponen a disposición de sus alumnos sistemas de gestión del aprendizaje (Learning Management Systems, LMS) basados en la web. El campo de aplicación de la minería de datos en educación, en particular orientado a los profesores para la mejora de sus cursos, plantea una serie de desafíos a resolver [5].

Por una parte, existe una amplia variedad de cursos e-learning sobre los que se puede aplicar minería de datos, pero los resultados obtenidos con un tipo de curso no necesariamente son válidos o aplicables a otro. La amplia gama de resultados que podría obtenerse dependiendo del tipo de curso, provoca que la búsqueda de patrones generales repetibles que puedan aplicarse a cualquier tipo de curso sea una tarea bastante difícil. Por otra parte, la aplicación de técnicas de minería de datos sobre un curso, de manera concreta y con parámetros específicos de filtrado, podría provocar un problema de descubrimiento de reglas de asociación en bases de datos pequeñas, donde la información de partida es insuficiente para construir un modelo que permita inferir comportamientos futuros [5].

En este caso de éxito se propone un sistema recomendador colaborativo que permite a profesores y expertos en educación intercambiar experiencias entre sí sobre cómo aprenden sus alumnos, de forma que este conocimiento les permita mejorar sus propios cursos on-line. Se presenta una arquitectura del sistema y el algoritmo de minería diseñado. Además se describen, respectivamente, la implementación del algoritmo y las pruebas realizadas para evaluar la efectividad del sistema [5].

1.2.2. Antecedentes

A continuación se van a describir los principales antecedentes en las distintas áreas de investigación relacionadas con el caso de éxito presentado:

1.2.2.1. Algoritmos de Minería de datos

Las técnicas de minería de datos más utilizadas entre los sistemas de educación online es el descubrimiento de reglas de asociación. Una regla de asociación [12] del tipo X ⇒ Y, expresa una fuerte correlación entre ítems (atributo-valor) de una base de datos. Se define el soporte de una regla como la probabilidad de que un registro satisfaga tanto a su antecedente como a su consecuente. El problema del descubrimiento de reglas de asociación consiste en encontrar todas las asociaciones que satisfagan ciertos requisitos de soporte y confianza mínimos, los cuales suelen expresarse mediante parámetros que define el usuario. El primer algoritmo que resolvió este problema fue A priori [5].

Una mejora del Apriori es el algoritmo denominado Apriori Predictivo [13], cuya principal ventaja es que el usuario no tiene que especificar los valores umbrales de soporte y confianza mínimos. El algoritmo intenta encontrar las N mejores reglas de asociación, donde N es un número fijo, buscando un balance adecuado entre el soporte y la confianza de forma que se maximice la probabilidad de hacer una predicción correcta sobre el conjunto de datos. Utilizando el método bayesiano, se define y calcula un parámetro llamado exactitud predictiva que nos dice el grado de exactitud de la regla encontrada [5].

1.2.2.2. Medidas de interés de las reglas descubiertas

Aunque la versión predictiva del algoritmo Apriori antes mencionado, representa una ventaja sobre la versión original, el algoritmo no asegura que las reglas obtenidas sean las más interesantes para detectar problemas en el curso e-learning. Por esta razón, es necesario llevar a cabo una evaluación del conocimiento extraído. Existe un sistema denominado IAS (Interestingness Analysis System) que compara las reglas descubiertas con el conocimiento que tiene el usuario del dominio de interés [5]. Sea U el conjunto de las especificaciones del usuario y A el conjunto de las reglas

descubiertas, la técnica propuesta clasifica y ordena las reglas dentro de los siguientes cuatro tipos de grupos:

a) Reglas conformes, tanto la condición como el consecuente (conform)
b) Reglas con el consecuente inesperado (unexpConseq)
c) Reglas con la condición inesperada (unexpCond)
d) Reglas con ambas, la condición y el consecuente inesperado (bsUnexp)

En general, los conjuntos de items frecuentes son beneficiosos para descubrir reglas de asociación en grandes bases de datos. Las bases de datos en educación son relativamente pequeñas (dependen de la cantidad de alumnos por clase), si las comparamos con otros campos de la minería de datos. Por tanto, es imprescindible aprender cómo aprovechar la experiencia, el sentido común y los modelos de otras personas que, previamente, hayan trabajado con bases de datos de características similares. Esto nos acerca al campo de la minería de datos distribuida y filtrado colaborativo [5].

1.2.2.3. Minería de datos distribuida y filtrado colaborativo

La minería de datos distribuida (MDD) asume que los datos están distribuidos en dos o más sitios y estos sitios cooperan para obtener resultados globales sin revelar los datos de cada sitio o revelando partes de éstos. Trabajos previos [14,15] han propuesto algoritmos MDD que agrupan los datos en subconjuntos. También se han propuesto algoritmos de minería paralelos [16] para trabajar con conjunto de datos grandes, dividiéndolos y distribuyéndolos entre los distintos procesos de una máquina virtual. Uno de los métodos más intuitivos para encontrar reglas de asociación de manera distribuida se conoce como partición horizontal de datos [17], donde el proceso de minería se aplica localmente y los resultados obtenidos en cada sitio se combinan finalmente para obtener reglas que se cumplen en la mayoría de las bases de datos locales.

Si a estas herramientas de MDD se añaden métodos pro- activos que utilizan herramientas para soportar trabajo colaborativo, estamos ante un desarrollo multidisciplinar que normalmente involucra expertos en diferentes áreas de conocimiento: ingenieros del conocimiento que modelan el conocimiento,

desarrolladores de bases de conocimiento que construyen, organizan, anotan y mantienen estas bases de datos y expertos que validan elementos de conocimiento antes de su inserción en un repositorio de contenidos. Las opiniones que dan estos expertos y los propios usuarios acerca de un problema a través del voto explícito o implícito constituyen la clave de los sistemas recomendadores colaborativos, que intentan sugerir las mejores soluciones basada en las experiencias del conjunto [5].

1.2.2.4. Sistemas recomendadores

Un campo de aplicación de los RS, que es muy reciente y está actualmente en auge es el e-learning [18,19] donde utilizando distintas técnicas de recomendación se brinda apoyo al alumno en su actividad de aprendizaje on-line o un camino de navegación óptimo basado en sus preferencias, dando recomendaciones de conocimientos, en base al histórico de navegación de otros alumnos de características similares [5].

Las técnicas de recomendación [20] poseen varias clasificaciones basándose en las fuentes de datos sobre las cuales se hacen las recomendaciones y el uso que se le da a estos datos. La aproximación de filtrado colaborativo (CFS: Collaborative Filtering System), también llamado filtrado social, depende de una base de datos de productos, así como datos demográficos y otras evaluaciones de un posible consumidor de algunos productos aún no experimentados. Esta técnica es quizás la más familiar, la más implementada y la más madura de las técnicas de recomendación [21].

1.2.3. Sistema Recomendador para la mejora contínua de Cursos E-Learning.

CIECoF (Continuous improvement of e-learning courses framework), es un sistema recomendador colaborativo aplicado a educación, cuya principal finalidad es ayudar a los profesores a mejorar sus cursos de e-learning de forma continua. El sistema utiliza técnicas de minería de datos distribuida, presentando al usuario las relaciones interesantes descubiertas a partir de su propia información y las descubiertas por otros usuarios con perfiles similares, que han obtenido dichas relaciones trabajando con sus

propias bases de datos. Mediante un procedimiento de valoración subjetiva, los usuarios evalúan el interés de las relaciones obtenidas [5].

De este modo, la base de conocimientos se reforzará con aquellas experiencias que por su peso satisfacen las necesidades de muchos usuarios, lo cual implica recomendaciones cada vez más efectivas. El sistema de minería de datos distribuida está basado en una arquitectura cliente-servidor con N clientes que aplican el mismo algoritmo de minería de reglas de asociación de manera local sobre los datos de utilización de un curso online por sus alumnos [5].

Los resultados de este algoritmo, se muestran al profesor en un formato comprensible de tuplas del tipo regla-problema-recomendación, para ayudarle a corregir los problemas detectados. Estos resultados puede compartirlos con otros profesores de perfil similar [5]. Se puede observar a detalle cada elemento de la arquitectura propuesta (ver figura 1).

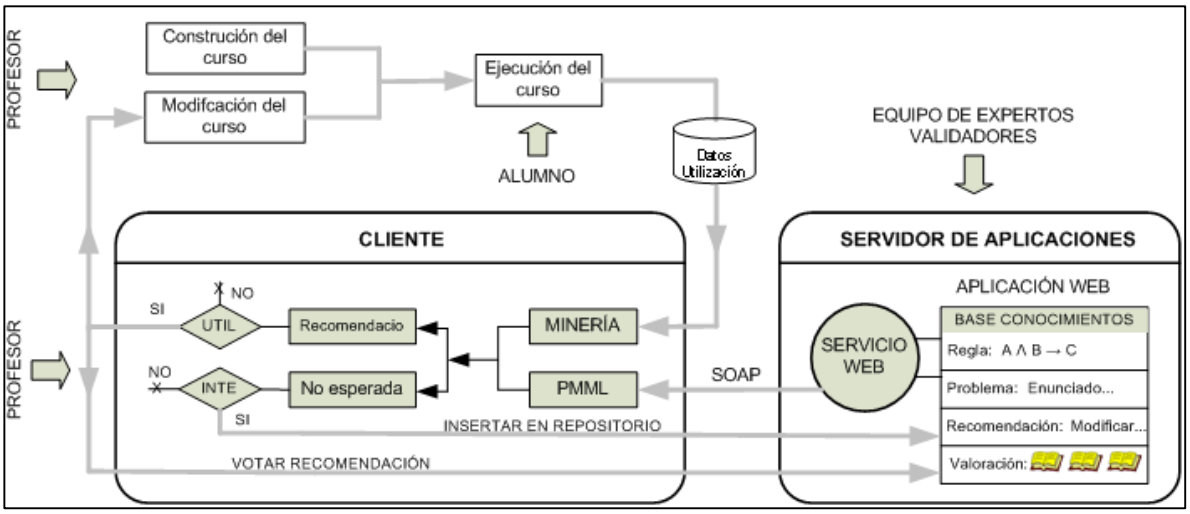


Figura 1. Arquitectura del Sistema CIECoF [5].

1.2.4. Diseño del algoritmo

Se ha diseñado e implementado un algoritmo para minería de reglas de asociación aplicado a educación, el cual se basa en los siguientes algoritmos: 1) Apriori Predictivo para el descubrimiento de reglas de asociación sin parámetros; 2) IAS para el análisis subjetivo y clasificación de las reglas inesperadas a través

de su comparación con una base de conocimientos sobre el dominio, previamente definida [5].

El algoritmo implementado incluye la nueva medida de interés basada en pesos propuesta anteriormente. El algoritmo implementado es especialmente útil para su uso en sistemas recomendadores colaborativos, donde se puede aprovechar la sinergia que ofrece la red, para producir recomendaciones cada vez más útiles y exactas. El algoritmo propuesto es interactivo e iterativo. En cada iteración el usuario ejecuta el algoritmo de minería para encontrar las reglas que servirán de base a las recomendaciones, pudiendo ejecutarlo tantas veces como desee [5].

El algoritmo utilizado en el sistema propuesto. El paso 1) se inicializa la variable Num al número de reglas N que desea encontrar el usuario; en 2) comienza un bucle cuyas instrucciones se ejecutarán mientras el usuario no decida parar. El paso 3) es el sub-algoritmo al que llamaremos Minería, que describiremos en la siguiente sección, y que devuelve el conjunto de recomendaciones (Rec) y reglas inesperadas (Rne) encontradas. De 4) a 6) el usuario vota si le ha sido útil o no la recomendación y de 7) a 11) evalúa las reglas no esperadas para determinar si son interesantes las cuales podrían añadirse previa validación por los expertos en la base de conocimientos KB (ver figura 2).

```
Entrada: El perfil de usuario: materia, nivel, estudios,
         Número de reglas N
 1)  Num = N;
 2)  while (usuario no desee parar) do
 3)     Rec, Rne = Algoritmo_Minería (Num);
 4)     for each i-rule in Rec do
 5)        UsuarioVotaRecom(Reci);
 6)     end
 7)     for each i-rule in Rne do
 8)        if ( Interesante(Rne)) ) then
 9)           Añadir_a_BaseConocimiento (Rne)
10)        end if
11)     end
12)     Num += N;
13)  end while
14)  end all
```

Figura 2. Algoritmo principal [5].

1.2.5. Descripción de la información descubierta

Los resultados que a continuación se exponen corresponden a pruebas realizadas con 150 alumnos, que ejecutaron el curso denominado “Hoja de Cálculo”. A continuación se van a describir un par de reglas descubiertas de tipo esperadas o sea que coinciden con la base de conocimientos. Indicar que también se descubrieron muchas reglas que no brindaban información alguna de utilidad para nuestros propósitos, como por ejemplo, aquellas que incluían en el antecedente y consecuente atributos de tiempo y que relacionaban ítems de conceptos que no estaban relacionados [5].

1) Si (e_time [25] = ALTO) entonces (e_score[25] = BAJO), exact = 0.85

Esta regla significa que, si el tiempo empleado en el ejercicio es alto, entonces la nota del ejercicio es baja. Se descubrió que existía un problema en ese ejercicio del curso hipermedia adaptativo, que pertenecía al tema “Uso de la aplicación”, la lección “Primeros pasos con el procesador de texto” y concepto “Renombrar y guardar un documento”, que era un escenario de INDESAHC de tipo video interactivo donde el alumno debe simular utilizando el ratón los pasos necesarios para completar una actividad. En este caso particular, se comprobó que el enunciado de la pregunta era ambiguo y podía interpretarse de varias maneras, con lo cual se corrigió. Otras reglas de formato similar se encontraron pero relacionadas con preguntas de tipo test o de relación de columnas [5].

2) Si (u_forum_read [2] = BAJO) Y (u_forum_post [2] = BAJO) entonces (u_final_score [1] = ALTO), exact = 0.75

Esta regla significa que, si los mensajes enviados y leídos del foro 2 que pertenece al tema 1 son bajos, entonces la nota del tema es alta. La regla descubre que ese foro del tema 1 no es necesario o que hay problemas con el tutor. Este tipo de regla descubierta cuestiona la necesidad de un foro a determinados niveles de la jerarquía del dominio, de hecho en nuestro caso se optó por eliminar el foro [5].

1.2.6. Resultados del caso de éxito 2

Los resultados desde el punto de vista del profesor al aplicar el sistema de manera consecutiva sobre los datos de utilización de tres grupos de alumnos, estos resultados están representados en la tabla I, donde la columna “Nuevas” se refiere a las recomendaciones iniciales que da el sistema a problemas detectados en el curso y que el profesor ha considerado útiles y aplicables; la columna “Rep” se refiere a aquellas recomendaciones iniciales que a pesar del profesor haberlas aplicado, se vuelven a repetir las mismas tuplas en ejecuciones consecutivas del curso. La Tabla III muestra los resultados desde el punto de vista del alumno. La columna “NRep” se refiere las tuplas que no se repiten, se muestran además las notas medias finales y desviaciones estándar de cada grupo y se calculan los valores de p-value comparando el grupo 1 con el grupo 2.

TABLA I

RESULTADOS DESDE EL PUNTO DE VISTA DEL PROFESOR

Grupo	Nuevas	Repetidas	Total	EfectRecom (%)
1	21	0	21	0
2	5	6	11	72,7

El estudio también arroja los resultados desde el punto de vista del alumno (ver tabla II).

TABLA II

RESULTADOS DESDE EL PUNTO DE VISTA DEL ALUMNO

Grupo	No Rep.	Nota	p-value 1-2
1	0	6,55 +0,30	< 0,0001
2	15	6,95 +0,56	

Del análisis de los datos de las Tablas I y II se pueden extraer varias conclusiones:

- Tal y como se suponía en nuestra hipótesis inicial, el porcentaje de efectividad se acerca al 100 % en la medida que el curso se ejecuta más veces. Se detectó que los problemas que se repiten se debieron a cambios en el diseño del curso, que tenían una alta componente subjetiva, por ejemplo el cambio de nivel de dificultad de una lección, o de la duración estimada para un tema [5].

- En cada grupo se han detectado nuevos problemas y por tanto nuevas recomendaciones asociadas para resolverlos, que no habían sido detectadas con anterioridad, la causa de esto podría estar en que, a pesar de los intentos por igualar la composición de cada grupo, estamos trabajando con personas con características muy subjetivas como el intelecto, habilidades, etc [5].

- Además de aumentar el porcentaje de efectividad, vemos que el total de recomendaciones asociadas a problemas encontrados disminuye, lo cual es un índice también de que el curso va mejorando continuamente [5].

- Comparando las notas de ambos grupos se observa una sensible mejoría, lo cual también indica la efectividad del sistema propuesto [5].

1.3. Caso de Éxito 3: Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.

1.3.1. Introducción

Desde hace algún tiempo se le está dedicando una creciente atención al rendimiento académico universitario. Algunos factores podrían, en gran medida, explicar el éxito o fracaso de un estudiante, como sus características socioeconómicas, edad, estudios previos, entorno al inicio de sus estudios, actividad, o no, laboral durante los estudios, características organizativas y docentes de los centros, planes de estudios, métodos evaluativos, etc [6].

Es por estas razones que se ha realizado un análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia (UPV) a lo largo de tres cursos, aunque también se ha trabajado con las titulaciones de Ingeniería Técnica en Informática de Gestión y de Ingeniería Informática [6].

Este análisis relaciona el rendimiento con las características socioeconómicas y académicas de los alumnos, que se obtienen en el momento de su matrícula, y que se

recogen en la base de datos de la universidad. Hemos definido un indicador del rendimiento para cada alumno, teniendo en cuenta las calificaciones obtenidas y las convocatorias utilizadas [6].

Para el estudio se ha utilizado técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, etc… Esto proporciona una herramienta importante para la acción tutorial, que puede apoyarse en las predicciones de los modelos que se obtienen para encauzar sus recomendaciones y encuadrar las expectativas y el esfuerzo necesario para cada alumno, lógicamente dentro de la cautela habitual a la hora de tratar modelos inferidos a partir de datos [6].

1.3.2. Objetivo del estudio y población

El estudio, a nivel global, pretende aplicar técnicas de minería de datos para analizar la influencia de los parámetros (socioeconómicos, características personales, nota de entrada…) más relevantes sobre el rendimiento académico de un alumno de primer curso en las titulaciones de informática de la UPV, de forma que nos permita predecir este rendimiento disponiendo únicamente de la información aportada por el alumno en el momento de su matrícula [6].

Estas titulaciones corresponden a Ingeniería Informática (II) impartida por la Facultad de Informática (FI), Ingeniería Técnica en Informática de Gestión (ITIG) e Ingeniería Técnica en Informática de Sistemas (ITIS), estas últimas impartidas por la Escuela Técnica Superior de Informática Aplicada (ETSIAp). En este trabajo se presentan únicamente resultados correspondientes a ITIS [6].

La adopción de políticas encaminadas a corregir situaciones de fracaso académico a partir de las conclusiones de este trabajo no es un objetivo del estudio. Estas políticas deben ser acometidas bien por los centros, bien por los estamentos responsables [6].

La población objeto de nuestro estudio está constituida por todos los alumnos de nuevo ingreso en cualquiera de las tres titulaciones de informática de la UPV antes

mencionadas. Con el fin de trabajar con una muestra representativa de la población, se ha considerado a los alumnos de nuevo ingreso matriculados en primero de alguno de los títulos de informática durante los cursos 01-02, 02-03 y 03-04, esto es, desde el último cambio del plan de estudios. Así, el estudio se ha realizado sobre 569 alumnos de II, 646 alumnos de ITIG y 572 alumnos de ITIS [6].

1.3.3. Obtención de la vista minable

Con el fin de crear un almacén de datos y un entorno que facilitara la obtención de datos para realizar el estudio, se decidió integrar los mismos en Oracle. De esta forma se ha podido utilizar como herramienta OLAP el Oracle Discoverer. Con ella se han extraído las la colección de individuos (vista minable), con todas sus características (atributos), que tiene como finalidad poder aplicar el proceso de la minería de datos sobre ella para poder extraer conocimiento útil [6].

Así, se ha creado una vista minable por titulación. Cada una de ellas contiene las notas y datos personales de los alumnos de la muestra seleccionada, de entre todos estos atributos disponibles, se han seleccionado aquellos que se consideraron, a priori, que podrían tener mayor influencia en el rendimiento académico, filtrando el resto (ver tabla III).

TABLA III

ATRIBUTOS SELECCIONADOS PARA LA VISTA MINABLE

Atributo	Descripción
Ocupacio P	Ocupación del padre
Ocupacio M	Ocupación de la madre
Ocupacio A	Ocupación del alumno
Ing Nota	Nota con la que el alumno aprueba estudios de acceso
Ing Est	Estudios con los que accede a la titulación

Seguidamente, se procedió a la agrupación de valores de algunos atributos por su elevado número de alternativas, con el fin de reducirlas y hacer más fácilmente interpretables los resultados obtenidos (ver tabla IV).

TABLA IV
AGRUPACIÓN DE VALORES DE LOS ATRIBUTOS

Atributo	Descripción
D_Altr Estud	Otros estudios universitarios del alumno al ingresar en la titulación
D_Estudis P	estudios del padre
D_Estudis M	estudios de la madre
Dpaises	Derivado del país de nacimiento del alumno, agrupando por zonas geográficas
Residencia Alumno	Derivado de la provincia y el código postal donde reside el alumno durante el curso
Residencia Familia Alumno	Derivado de la provincia y el código postal donde reside la familia del alumno durante el curso.
Edad Ingreso	Atributo derivado calculado como la diferencia entre el año de ingreso del alumno y año de nacimiento.

Finalmente, se especificó el tipo de cada atributo como nominal (o categórico) o numérico, siendo todos nominales excepto la nota de acceso a los estudios y la edad del alumno. Además de todos estos datos personales del alumno, para completar la vista minable se incorporó una columna con el Rendimiento de cada alumno, R, mediante un cálculo realizado [6].

1.3.4. Generación de los modelos con las técnicas de MD seleccionadas.

Entre las técnicas de minería de datos existentes, se ha utilizado dos de ellas para generar los modelos predictivos del rendimiento: los árboles de decisión y la regresión multivariante [6].

Para la generación de los modelos se ha utilizado la herramienta SPSS Clementine v.9.0. En concreto, de los árboles de decisión que incorpora el Clementine, se ha

utilizado para regresión el árbol C&R, que es un tipo de algoritmo de aprendizaje de árboles que se basa en el algoritmo CART de Leo Breiman et al [22].
Este árbol realiza particiones binarias con el objetivo que la media de cada rama sea diferente y, por tanto, discrimine con la suficiente precisión en un número de particiones razonable como para poder asignar a cada hoja un valor cercano a la media de los elementos que caen en ella. Asimismo, también hemos aplicado el método regression del Clementine, que implementa una regresión lineal [6].

En el modelo Árbol C&R obtenido para la vista minable de ITIS, cada línea corresponde a un nodo del árbol, y contiene el nombre del atributo usado en ese nodo así como su/s valor/es entre corchetes, si es un atributo nominal, o bien una expresión de la forma d V ó >V, si es un atributo numérico, siendo V un valor comprendido entre los valores mínimo y máximo de ese atributo en los ejemplos en ese nodo; además, si se trata de una hoja, se añade el símbolo "=>" tras el que se indica el valor predicho (el rendimiento); asimismo, el número que aparece entre paréntesis indica el número de instancias del conjunto de entrenamiento en ese nodo del árbol (ver figura 3).

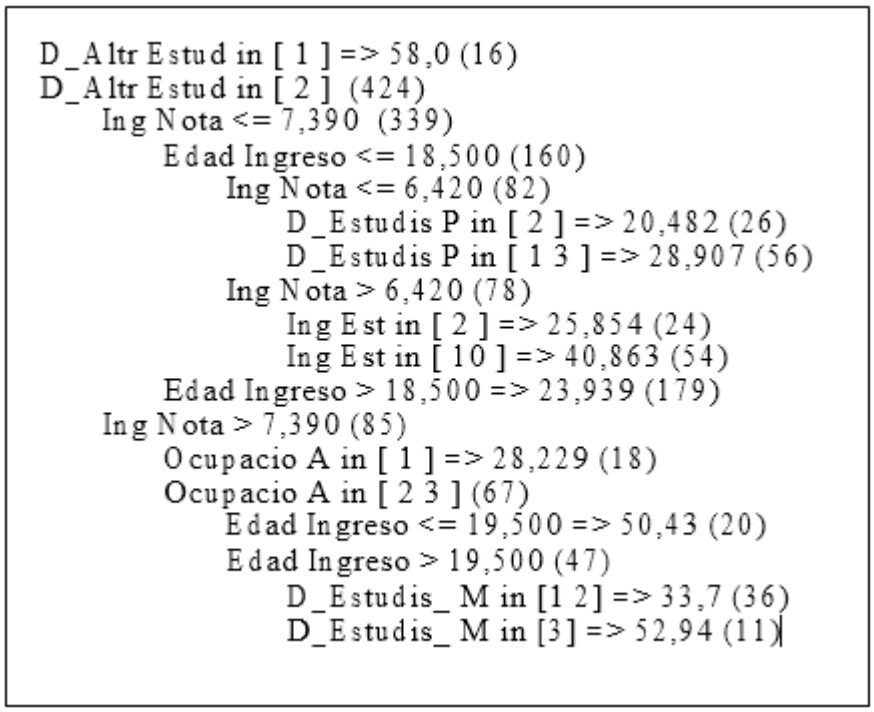


Figura 3. Árbol C&R para la titulación de ITIS [6].

Mientras que el modelo de regresión lineal multivariante generado para los alumnos de la titulación de ITIS, debe interpretarse de manera que determinados valores de ciertos atributos, los cuales hacen que el rendimiento de un alumno varíe en un determinado valor, dado por el coeficiente de cada atributo del modelo, y cuyo signo indica si el rendimiento aumenta debido a esa característica, o disminuye. Además, también aparece un término independiente de cualquier atributo (ver figura 4).

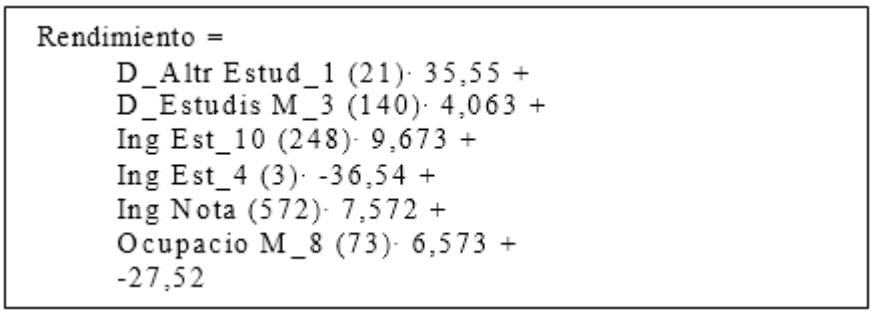


Figura 4. Modelo de regresión para la titulación ITIS [6].

1.3.5. Interpretación de los Resultados

En los modelos obtenidos podemos destacar que en el Árbol C&R un atributo puede utilizarse a diferentes niveles del árbol y además repetidamente, hemos calculado para cada atributo el número de ejemplos para los cuales dicho atributo se utiliza. Con esto tenemos que un atributo es relevante si se utiliza para un número importante de casos [6].

La tabla V muestra este número absoluto, que llamamos I (por importancia), el cual lo dividimos entre el número total de ejemplos utilizados para el entrenamiento, obteniendo un valor denominado IR (importancia relativa). Finalmente, sumamos todas las IR de cada atributo y normalizamos, obteniendo un valor IR2 tal que todos los IR2 para todos los atributos sumen 1 y nos muestre un valor de importancia que permita comparar más fácilmente entre diferentes árboles [6].

TABLA V

ANÁLISIS DE LOS ATRIBUTOS EN EL ÁRBOL DE DECISIÓN PARA ITIS

Atributo	I	IR	IR2
D_Altr Estud	440	1	0,343
Ing Nota	584	1,327	0,455
Edad Ingreso	406	0,922	0,316
D_Estudis P	82	0,186	0,063
D_Estudis M	47	0,106	0,036
Ocupacio A	85	0,193	0,066
Ing Est	78	0,177	0,060
TOTAL	1282	2,913	1

En base a los datos que presenta la tabla V se han dado los siguientes resultados del análisis [6]:

- Los mejores rendimientos se obtienen para el valor 1 del atributo D_Altr Estud, es decir, los alumnos que ya poseen estudios universitarios, aunque esta condición sólo la cumple un porcentaje relativamente pequeño de alumnos.
- El atributo Ing Nota (nota de ingreso) afecta positivamente.
- El atributo Edad Ingreso afecta negativamente (cuanto mayor es, peor rendimiento).
- El atributo D_Estudis P afecta positivamente para los valores 1 y 3 del padre sin estudios o estudios superiores, y negativamente para el valor 2 del padre con estudios equivalentes a bachillerato.
- El atributo Ocupacio A afecta positivamente para los valores 2 y 3 correspondientes a los alumnos con una ocupación inferior a 15 horas o que no realiza trabajo renumerado, y negativamente para el valor 1 de los alumnos con una ocupación mayor o igual a 15 horas semanales.
- El atributo Ing Est afecta positivamente para el valor 10 que corresponden a los alumnos que acceden desde bachillerato LOGSE con PAU, y negativamente para el valor 2 que corresponden a alumnos que acceden con prueba de acceso pero no provienen de bachillerato LOGSE

1.4. Caso de Éxito 4: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos.

1.4.1. Introducción

El caso de éxito descrito a continuación propone la utilización de técnicas de minería de datos para detectar, cuáles son los factores que más influyen para que los estudiantes de enseñanza media o secundaria fracasen, es decir, suspendan o abandonen. Además se propone utilizar diferentes técnicas de minería de datos debido a que es un problema complejo, los datos suelen presentar una alta dimensionalidad (hay muchos factores que pueden influir) y suelen estar muy desbalanceados (la mayoría de los alumnos suelen aprobar y sólo una minoría suele fracasar). El objetivo final es detectar lo antes posible a los estudiantes que presenten esos factores para poder ofrecerles algún tipo de atención o ayuda para tratar de evitar y/o disminuir el fracaso escolar [7]. Para realizar este estudio se ha hecho uso de un método que permite de manera organizada llegar a realizar la minería de datos con éxito (ver figura 5).

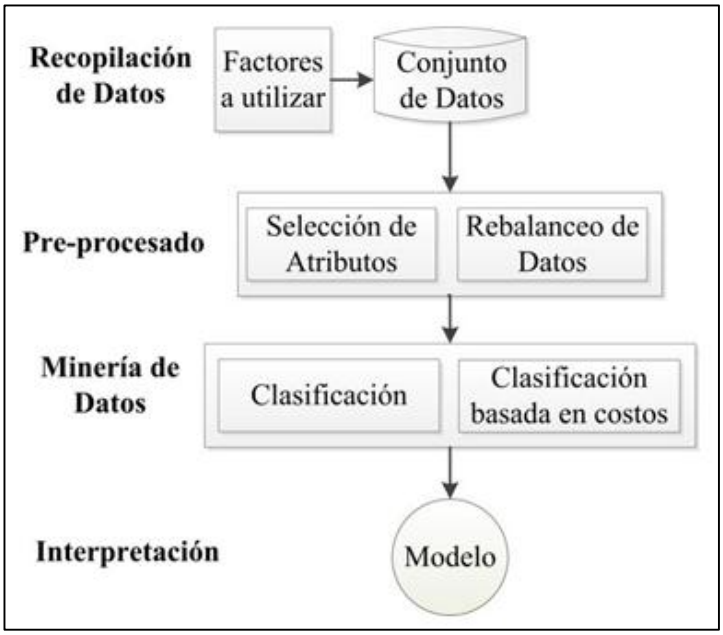


Figura 5. Método utilizado para predicción del fracaso escolar [7].

1.4.2. Recopilación de datos

En el caso de estudio concreto, los datos que se ha utilizado son de estudiantes del Programa II de la Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México. Todos los estudiantes que han participado en este estudio eran de nuevo ingreso en el curso académico 2009-2010 en el nivel medio-superior de educación Mexicana [7]. Toda la información se recopiló de tres fuentes diferentes:

a) De una encuesta que se les aplicó a todos los alumnos a mitad del curso, con la finalidad de obtener información para detectar factores importantes que pueden incidir en su rendimiento escolar.
b) Del Centro Nacional de Evaluación CENEVAL. Organismo que, entre otras actividades, realiza exámenes de ingreso o admisión en muchas instituciones de educación media y superior. En el examen, también se les hace un estudio socioeconómico, de éste se extrajo parte de la información.
c) Del Departamento de Servicios Escolares del Programa II, donde se recogen todas las notas obtenidas por los estudiantes.

1.4.3. Pre-procesado de datos

Se ha realizado algunas tareas de pre- procesado como: integración, la limpieza, la transformación y la discretización de datos, que ha permitido transformar los datos originales a una forma más adecuada para ser usada por el algoritmo en particular [7].

Después de realizar estas tareas se dispone de un primer fichero de datos con 77 atributos/variables sobre 670 alumnos. Este fichero de datos fue particionado (10 particiones) para poder hacer una validación cruzada en las pruebas de clasificación. Una partición es la división aleatoria del fichero original de datos en otros dos, uno para la etapa de entrenamiento o training y el otro para la etapa de prueba o test [7].

Para seleccionar los mejores atributos se revisaron los resultados obtenidos por los 10 algoritmos de selección y se contabilizaron los que han sido seleccionados por varios de ellos. Al hacer la selección de los atributos con mayor frecuencia, se ha pasado de tener los 77 atributos originales a solamente los 15 mejores (ver tabla VI). Nuevamente,

este fichero de datos se partió en 10 ficheros de entrenamiento y 10 ficheros de prueba [7].

TABLA VI

ATRIBUTOS SELECCIONADOS DE ACUERDO A LOS MÉTODOS APLICADOS.

Atributos	Frecuencia
Humanidades 1	10
Ingles 1	10
Ciencias Sociales 1	9
Matemáticas 1	9
Taller de Lectura y Redacción 1	9
Física 1	9
Computación 1	9
Nivel de motivación	5
Promedio de secuencia	3
Fumas	2
Calificación de EXANI I	2
Edad	2
Discapacidad Física	2
Semestre y grupo	2

1.4.4. Minería de datos y experimentación

Se ha realizado varios experimentos con el objetivo de obtener la máxima exactitud de clasificación. En un primer experimento hemos ejecutado 10 algoritmos de clasificación utilizando todos los 77 atributos con los que se cuenta, es decir de toda la información disponible. En un segundo experimento, utilizamos sólo los 15 mejores atributos o variables. En un tercer experimento, hemos repetido las ejecuciones pero utilizando los ficheros de datos re- balanceados. En un último experimento hemos consideramos diferentes costos de clasificación [7].

En el primer experimento, se han ejecutado los 10 algoritmos utilizando toda la información disponible, es decir, los ficheros de datos con los 77 atributos de los 670 alumnos. Hemos realizado una validación cruzada con 10 particiones. En este tipo de validación cruzada, se realiza el entrenamiento y el testeo diez veces con las diferentes

particiones. Los resultados obtenidos (la media de las 10 ejecuciones) con los ficheros de prueba/test de la aplicación de los algoritmos de clasificación contienen los porcentajes de exactitud obtenidos para la exactitud total y para los Aprobados son altos, no así para los que suspendieron y la media geométrica [7]. Concretamente, los algoritmos que obtienen los valores máximos son: JRip y ADTree (ver tabla VII).

TABLA VII

VALIDACIÓN CRUZADA UTILIZANDO LOS 77 ATRIBUTOS DISPONIBLES

Algoritmo	%Aciertos =Aprobó	%Aciertos Suspendió	%Exactitud =Total	Media Geométrica
JRip	97.7	**78.3**	96.0	**87.5**
NNge	98.5	73.3	96.3	85.0
OneR	98.9	41.7	93.7	64.2
Prism	99.5	25.0	93.1	49.9
Ridor	96.9	65.0	93.7	79.2
ADTree	**99.7**	76.7	**97.6**	87.4
J48	97.4	53.3	93.4	72.1
Random Tree	95.7	48.3	91.5	68.0
REPTree	98.0	56.7	94.3	74.5
SimpleCart	97.7	65.0	94.8	79.7

En el segundo experimento, se han utilizado los ficheros con los mejores 15 atributos, que consiste en ejecutar nuevamente los 10 algoritmos de clasificación para poder comprobar cómo ha afectado la selección de atributos en la predicción. Los resultados de la validación cruzada (la media de las 10 ejecuciones) de los algoritmos de clasificación utilizando solamente los 15 mejores atributos. Con estos atributos los algoritmos han mejorado el porcentaje de exactitud al utilizar sólo los mejores atributos. Aunque hay algunos algoritmos que empeoran un poco, en general la tendencia es de mejora. De hecho, se obtienen unos valores máximos mejores a los obtenidos con todos los atributos. Nuevamente los algoritmos que obtienen estos valores máximos son el JRip y ADTree (ver tabla VIII).

TABLA VIII
VALIDACIÓN CRUZADA CON LOS 15 MEJORES ATRIBUTOS PRIMERA PARTE

Algoritmo	%Aciertos =Aprobó	%Aciertos Suspendió	%Exactitud =Total	Media Geométrica
JRip	97.0	**81.7**	95.7	**89.0**
NNge	98.0	76.7	96.1	86.7
OneR	98.9	41.7	93.7	64.2
Prism	99.2	44.2	94.7	66.2
Ridor	95.6	68.3	93.1	80.8

TABLA IX
VALIDACIÓN CRUZADA CON LOS 15 MEJORES ATRIBUTOS SEGUNDA PARTE

Algoritmo	%Aciertos =Aprobó	%Aciertos Suspendió	%Exactitud =Total	Media Geométrica
ADTree	**99.2**	78.3	**97.3**	88.1
J48	97.7	55.5	93.9	73.6
Random Tree	98.0	63.3	94.9	78.8
REPTree	97.9	60.0	94.5	76.6
SimpleCart	98.0	65.0	95.1	79.8

1.4.5. Interpretación de Resultados

Los resultados que se observa en las tablas VII, VIII y IX observamos que la mayoría de los algoritmos en el segundo experimento han aumentado su exactitud en predicción, obteniendo nuevos valores máximos en casi todas las medidas excepto en el porcentaje de exactitud total. En este caso, los algoritmos que han obtenido los mejores resultados han sido el algoritmo Prism, OneR y nuevamente el algoritmo ADTree [7].

2. CAPÍTULO II. RECOPILACIÓN DE TÉCNICAS DE MINERÍA DE DATOS.

La minería de datos es el proceso de descubrimiento del conocimiento que se encuentra escondido en los datos almacenados en distintas fuentes. Es multidisciplinaria ya que combina ramas como la estadística, el análisis de datos, la inducción de reglas, etc. Este proceso se lo realiza de manera automatizada por lo que facilita el análisis de grandes bases de datos con características complejas [4, 23].

Durante el proceso de la minería de datos se realiza la aplicación de técnicas que permiten detectar fácilmente patrones en los datos. Estas técnicas persiguen el descubrimiento automático del conocimiento contenido en la información almacenada

de modo ordenado en grandes y complejas bases de datos [23, 26]. En la tabla X [28], se detalla las áreas de aplicación de las técnicas de minería de datos con el fin de resolver diversos problemas, corroborando la importancia de su estudio y aplicación:

TABLA X

ÁREAS DE APLICACIÓN DE LA MINERÍA DE DATOS.

Área de Aplicación	**Tipo de Problema**
Comercio/ Marketing	• Identificar patrones de compra de los clientes. • Buscar asociaciones entre clientes y características demográficas. • Predecir respuesta a campaña de mailing. • Análisis de la canasta de compra.
Banca	• Detectar patrones de uso fraudulento de tarjetas de crédito. • Identificar clientes leales. • Predecir clientes con probabilidad de cambiar su afiliación. • Determinar gasto de tarjeta de crédito por grupos. • Encontrar correlaciones entre indicadores y financieros. • Identificar reglas de mercado de valores a partir de datos históricos.
Seguros y Salud Privada	• Análisis de procedimientos médicos solicitados conjuntamente. • Predecir que clientes compran nuevas pólizas. • Identificar patrones de comportamiento clientes con riesgo. • Identificar comportamiento fraudulento.
Transportes	• Determinar la planificación de la distribución entre tiendas. • Analizar patrones de carga.
Medicina	Identificación de terapias médicas satisfactorias para diferentes enfermedades. • Asociación de síntomas y clasificación diferencial de patologías. • Estudio de factores (genéticos, precedentes, hábitos alimenticios, etc) de riesgo/salud en distintas patologías. • Segmentación de pacientes para una atención más inteligente según su grupo. • Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones. • Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.
Productos Industriales	• Extracción de modelos sobre comportamiento de compuestos.

	• Detección de piezas con trabas. • Predicción de fallos. • Modelos de calidad. • Estimación de composiciones óptimas en mezclas. • Extracción de modelos de coste. • Extracción de modelos de producción.

Estas técnicas se aplican utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos [23, 26], las cuales se clasifican en dos grandes categorías:

2.1. Técnicas Supervisadas o predictivas

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe comprobarse después del proceso de minería de datos antes de aceptarlo como válido [27]. El objetivo es conocer el comportamiento de la variable a predecir, para ello se debe contar con un conjunto de variables predictoras, que permitan predecir el valor de un atributo de un conjunto de datos, conocidos otros atributos. Una vez entrenado el modelo, sirve para realizar la predicción de datos cuyo valor es desconocida [24, 25].

2.2. Técnicas no supervisadas o descriptivas.

Estas técnicas ayudan a descubrir patrones y tendencias en los datos actuales, que no poseen variable a predecir o variable dependiente ya conocida. Los registros son agrupados por similaridad. El objetivo es descubrir el conocimiento para tomar acciones y obtener un beneficio. En otras palabras en las técnicas descriptivas no se asigna ningún papel determinado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos [24], [25, 27].

En general las técnicas de minería de datos se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones [23]. En la tabla X [24, 25] se ha detallado algunas de las diferencias de cada categoría:

TABLA XI

COMPARACIÓN DE LAS CATEGORÍAS DE LAS TÉCNICAS DE MINERÍA DE DATOS

Supervisadas o Predictivas	No supervisadas o descriptivas
Análisis de patrones secuenciales • Detección de secuencias de variables de tiempo.	Segmentos de datos • Agrupación no supervisada de variables. • Categorización automática de las variables.
Análisis de similitud temporales • Identificación de características comunes entre las variables.	Clasificación • Asignación de nuevos variables a segmentos predefinidos.
Predicción • Asignación de probabilidad de que ocurra un suceso. • Estimación de la demanda y el rendimiento por cliente.	Análisis de asociaciones • Análisis de características cruzadas entre variables. • Correlación de hábitos de consumo en base a su ocurrencia.

Como observamos en la tabla XI, se ha detallado algunas maneras de aplicar las técnicas de minería de datos, a continuación se las describe de forma más clara:

a). Descripción de clases: proporciona una clasificación concisa y resumida de un conjunto de datos los mismos que distingue unos de otros. La clasificación de los datos se conoce como caracterización, y la distinción entre ellos como comparación o discriminación [30].

b). Asociación: es el descubrimiento de relaciones de asociación o correlación en un conjunto de datos. También conocida como análisis de cesta de compra. Las asociaciones se expresan como condiciones atributo- valor y deben estar presentes varias veces en los datos [30, 31].

c). Clasificación: analiza un conjunto de datos de entrenamiento cuya clasificación de clase se conoce y construye un modelo de objetos para cada clase. Dicho modelo puede representarse con árboles de decisión o con reglas de clasificación, que muestran las características de los datos [30].

Se asigna una categoría a cada caso. Cada caso tiene un conjunto de atributos uno de ellos es el atributo clase. Se busca un modelo que describa el atributo clase como una función de los atributos de salida [31].

d). Predicción: Es una función que se aplica en la minería para predecir la distribución de valores de ciertos atributos en un conjunto de objetos o valores posibles de datos faltantes [30].

e). Agrupamiento o Clustering: Se utiliza directamente similaridad entre los datos en entrada, es decir se identifica clusters o grupos en el conjunto de datos, donde un cluster es una colección de datos “similares”. Los clusters deben ser de buena calidad es decir ser escalables a grandes bases de datos, esto es lo que intenta la minería de datos, la similitud determinada para formar los clusters puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos [29, 30].

Para cada categoría de las técnicas de minería de datos existen diferentes tipos de algoritmos, dependiendo del tipo de problema se pone en práctica ciertos algoritmos; debido a que no existe un único algoritmo que arroje la solución, dependiendo de algunos factores, como los datos, la meta de minería, las variables a predecir, se obtienen distintas aproximaciones. En la Tabla XII se observa una clasificación de los algoritmos más utilizados, que se describirán a continuación:

TABLA XII

ALGORITMOS DE LAS TÉCNICAS DE MINERÍA DE DATOS

Algoritmos Supervisados	Algoritmos no supervisados
Árboles de decisión	Agrupamiento(Clustering)
Redes Neuronales	K-Means
Regresión lineal	Reglas de Asociación
Máquinas de Soporte Vectorial	
Reglas de Inducción	

2.3. Algoritmos supervisados

2.3.1. Arboles de decisión

Algoritmo utilizado como herramienta para la clasificación, muy similar al conocido algoritmo C4.5. Consiste en la construcción de un árbol del que se pueden extraer reglas, dependiendo del problema. Cada operador puede requerir algunas entradas y entregar algunas salidas, para ello se debe realizar validaciones detectando elementos anómalos comparando si encajan o no con las reglas surgidas del árbol. Para predecir

el valor de un atributo de un árbol con precisión, se ve las correlaciones entre las variables predictoras y la variable a predecir [24, 33].

Ejemplo: En un banco se está evaluando si otorgarle el crédito a un cliente o no. Para ello se consideran atributos de los clientes para ver su nivel de riesgo, y según esto entregarle el crédito [32].

Figura 6. Árbol de decisión para evaluar riesgo de un cliente [32].

Como podemos observar en la figura 6. El árbol de decisión es planteado con una serie de condiciones, con el fin de predecir si se debe o no entregar el crédito al cliente respecto de la relación entre las distintas variables. El árbol de acuerdo a su estructura arroja algunos caminos de solución respecto al problema planteado [32].

Entre los algoritmos de árboles de decisión, se ha mencionado los siguientes:

2.3.1.1. ID3

Es uno de los algoritmos de árboles de decisión más populares, introducido por Quinlan en 1986. En el mismo el criterio escogido para seleccionar la variable más informativa está basado en el concepto de cantidad de información mutua entre dicha variable y la variable clase. La terminología usada en este contexto para denominar a la cantidad de información mutua es la de ganancia en información o en inglés information gain [34].

Esto es debido a que $I(X_i, C) = H(C) - H(C|X_i)$ y lo que viene a representar dicha cantidad de información mutua entre X_i y C es la reducción en incertidumbre en C debida al conocimiento del valor de la variable X_i. Matemáticamente se demuestra que este criterio de selección de variables utilizado por el algoritmo ID3 no es justo ya que favorece la elección de variables con mayor número de valores [34].

Además el algoritmo ID3 efectúa una selección de variables previa {denominada preprunning en este contexto {consistente en efectuar un test de independencia entre cada variable predictora X_i y la variable clase C, de tal manera que para la inducción del árbol de clasificación tan solo se van a considerar aquellas variables predictoras para las que se rechaza el test de hipótesis de independencia [34].

2.3.1.2. CHAID.

El modelo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos o numér icos o categóricos. CHAID es considerado un algoritmo que examina con precisión todas las divisiones posibles. Constituyen un conjunto de reglas que se pueden aplicar a un nuevo sin clasificar conjunto de datos para predecir cuáles registros tendrán un resultado determinado, al segmentar utilizando pruebas de chi cuadrado para crear divisiones en múltiples direcciones [34].

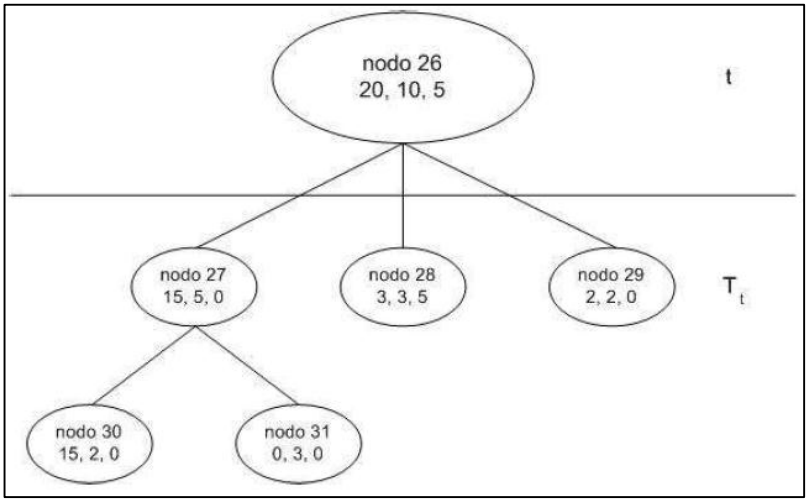


Figura 7. Proceso pos-poda, algoritmo C4.5 [34].

2.3.2. Redes Neuronales o inducción Neuronal

Las redes neuronales se han diseñado en cierto modo para simular el funcionamiento del cerebro humano y su capacidad de aprender. En otras palabras para realizar el modelo de minería de datos, como lo dice su nombre se basa en el estudio de las redes neuronales del cerebro humano respecto del comportamiento y su función, principalmente del sistema nervioso [24, 35].

El aprendizaje se da con el entrenamiento. El objetivo de este modelo es que exista un balance entre la habilidad para responder correctamente en relación a la entrada de patrones es decir usado para el entrenamiento y la habilidad de dar una respuesta buena que sea razonable para la entrada que es similar. Estos algoritmos son muy sofisticados para la detección de patrones y aprendizaje en la construcción de modelos predictivos a partir de una gran base de datos históricos [24, 35]. Para ello utiliza dos estructuras básicas:

- El nodo = simula a las neuronas humanas
- El vínculo=simula la conexión entre las neuronas

Las redes neuronales para predecir se basan en los nodos de entrada, los cuales envían información que luego es multiplicada por los vínculos y luego se une esta información en un nodo final que se le llama el nodo de salida. Una función podría ser aplicada a esta información que nos dará la predicción [24, 35].

Las redes neuronales se pueden utilizar tanto para problemas de clasificación (cuando se utilizan variables categóricas), como para problemas de regresión (cuando se cuenta con variables continuas). La estructura de una red neuronal está compuesta por un set de nodos interconectados, transmitiendo “señales” a través de sus conexiones, y cada una de esas conexiones o enlaces tiene un peso asociado [36].

La red neuronal funciona “aprendiendo de ejemplos”, por lo que necesita un conjunto de datos iniciales, presentados en una capa de entrada, donde cada nodo de esa capa corresponde a una variable predictora. Estos nodos iniciales se conectan con otros nodos en una capa oculta, donde se les aplica una “función de activación” utilizando los pesos que tienen asociadas las conexiones entre los nodos. La Figura 8 muestra un ejemplo de red, con nodos interconectados y pesos asociados [36].

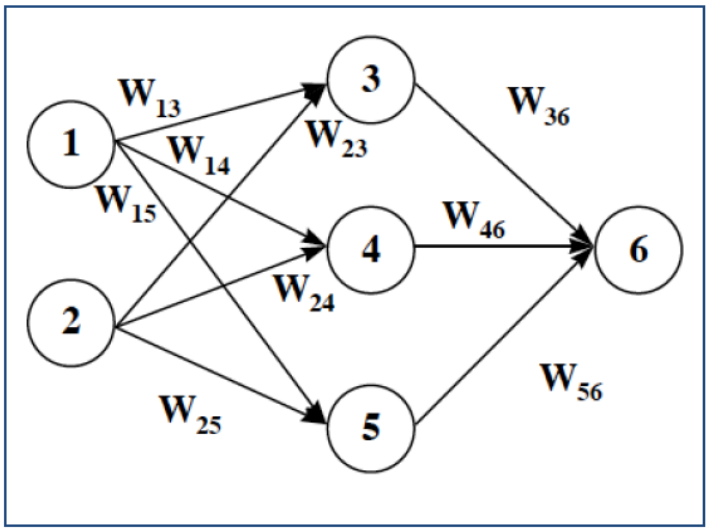


Figura 8. Neuronal con pesos asociados a cada nodo [37].

Los resultados obtenidos mediante la función de activación se traspasan de un nodo a otro hacia una capa de salida, que consiste en una o más variables de respuesta del modelo. En la Figura 8, el valor que se le entrega al nodo 6 corresponde a una composición de los valores ponderados obtenidos desde el nodo 1 y 2 [36], es decir:

$$\boldsymbol{W_{14} * ValorNodo_1 + W_{24} * ValorNodo_2}$$

Los pesos en la red son parámetros desconocidos del modelo, y se estiman a través del aprendizaje y métodos de entrenamiento [36].

2.3.3. Regresión Lineal

El modelo lineal ha sido ampliamente utilizado en el aprendizaje estadístico durante los últimos 30 años y continúa siendo una de las herramientas más importantes. El análisis de regresión es una herramienta estadística para evaluar la relación de una o más variables independientes $\{x_1, x_2 \ldots\ldots, x_k\}$ con una variable dependiente Y [38]. La regresión lineal puede ser apropiada para:

- Caracterizar la relación de variables dependientes e independientes para determinar la extensión, dirección y fuerza de la asociación entre ellas.
- Buscar la fórmula o ecuanción cuantitativa para describir la variable dependiente como función independiente de las variables $\{x_1, x_2 \ldots\ldots, x_k\}$

- Determinar las variables independientes que son importantes y cuáles no. Una variable es importante se ayuda a describir o predecir la variable dependiente.
- Obtener una estimación válida y precisa de uno o más coeficientes de correlación [38].

2.3.4. Máquinas de Soporte Vectorial (SVM)

El objetivo de este método es encontrar la mejor función para clasificar un conjunto de datos, encontrando los hiperplanos que mejor dividan la muestra, maximizando el grado de separación entre las clases generadas [36].

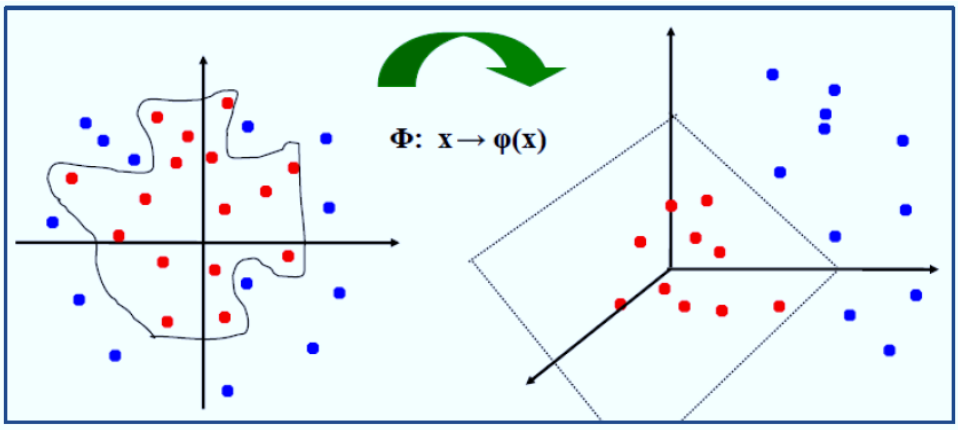


Figura 9. Transformación del espacio dimensional los datos [36].

Como podemos observar en la figura 9 para aplicar esta técnica se transforma los datos de entrada, desde un espacio de baja dimensión hacia uno dimensionalmente. Esto se realiza a partir de la elección de una función de kernel (polinomial), buscando entre los datos los parámetros del modelo a través de programación cuadrática. La medida utilizada para encontrar la mejor función es maximizar el margen o la distancia entre los objetos de dos clases [36].

2.3.5. Reglas de Inducción

Se trata de los algoritmos que arrojan como resultados un sin número de reglas respecto al análisis de los datos objeto de estudio de la minería de datos, se realiza el proceso buscando patrones, características similares entre los mismo y sus relaciones. Los algoritmos de inducción de reglas generalizan el conjunto de ejemplos de entrenamiento en forma de reglas que pueden ser evaluadas directamente para clasificar nuevas instancias [7].

A la vez las reglas generadas permiten la interpretación de conocimiento asociado a las reglas del negocio u objetivos propuestos. Estas reglas tienen la ventaja que son fáciles de entender. En la fase de crecimiento, para cada regla se añaden con glotonería condiciones a la regla hasta que la regla es perfecta (es decir, 100% de precisión). El procedimiento trata todos los valores posibles de cada atributo y selecciona el estado con la mayor ganancia de información [7].

2.3.5.1. JRip

El algoritmo JRip es que genera reglas basándose únicamente en la ganancia de información, está basado en el algoritmo de poda incremental de error reducido (RIPPER Repeated Incremental Pruning to Produce Error Reduction) que fue propuesto por William W. Este es un algoritmo que genera un listado de reglas conjuntivas y luego evaluarlas en orden para encontrar la primera regla que se cumple sobre el ejemplo a clasificar [7], [39], [40].

La generación de reglas en la minería de datos ha sido abordada en muchos estudios entre los cuales se tiene el algoritmo JRip el cual permite la inducción de reglas a partir de un conjunto de datos de gran tamaño y con estructuras complejas. Una vez encontrada dicha regla y que sea la más eficiente para el conjunto de datos, es asignado con una etiqueta de valor de salida final [39, 40, 48].

2.3.5.2. Ridor

Con sus siglas en inglés Ripple Down Rule, es el algoritmo que genera primero una regla por defecto (predeterminada) y luego toma las excepciones para la regla predeterminada con la mínima tasa de error. Entonces genera la mejor excepción para cada excepción iterando hasta lograr disminuir el error. Luego genera una expansión similar a un árbol de excepciones. La excepción es un conjunto de reglas que predice clases. Este algoritmo es usado para generar dichas excepciones [41, 42].

2.3.5.3. PART

El algoritmo PART de aprendizaje de reglas basado en árboles de decisión parciales (Frank y Witten, 1998) representa un enfoque alternativo híbrido para la inducción de listas de decisión, híbrido porque combina la estrategia divide-and-conquer de

aprendizaje de árboles de decisión con la estrategia separate-and-conquer de aprendizaje de reglas. Adopta la estrategia separate-and-conquer en el sentido de que construye una regla, elimina las instancias que ésta cubre y continúa creando reglas recursivamente para las instancias que permanecen hasta que no quede ninguna. Sin embargo, difiere del enfoque estándar en el modo en que se crea cada regla. En esencia, para crear una regla, se construye un árbol de decisión podado a partir del conjunto activo de instancias, la hoja de éste con mayor cobertura se convierte en una regla, y se desecha el árbol [43].

2.3.5.4. NNge

El algoritmo Nnge es conocido como el vecino más cercado; es un método que se originó en las estadísticas. Se consideró la primera regla de producción creada por Fix y Hodges (1951), quien realizó un análisis inicial de los propiedades de los elementos k-más cercanos al sistema, y estableció la consistencia del método como k varía de uno a infinito. También evaluado numéricamente la desempeño de vecino más cercano-k para muestras pequeñas, con los supuestos de estadísticas de distribución normal (Fix y Hodges, 1952) y ha sido ampliamente utilizada en el campo del reconocimiento de patrones desde 1963 [44].

- **Ejemplo del Algoritmo NNge:**

Supongamos que se tiene un objeto X que se quiere clasificar, y se tiene además un conjunto de datos de aprendizaje o entrenamiento (K), el cual consiste de una serie de objetos donde cada uno contiene su etiqueta correspondiente. El algoritmo lo que hará para clasificar al objeto X, es tomar del conjunto datos K el, estos K elementos son los K elementos del conjunto de entrenamiento que más se parecen a X. Se analizará cual etiqueta de los K elementos es la que se presenta con más frecuencia, y esta etiqueta será la que se le pondrá a X [44, 45].

Este algoritmo funciona siguiendo un modelo basado en la memoria. La memoria guarda un conjunto de objetos ó instancias que forman parte del entrenamiento. Para cada uno de estos objetos, se sabe cuál es su salida, esto es, los objetos están etiquetados. Cada ejemplo contiene un conjunto de valores independientes que producen un conjunto de salida dependiente de ellos. Dado un conjunto nuevo de valores independientes se

busca estimar mediante los k vecinos más cercanos, la salida de este. Esto se logra encontrando k ejemplos de entrenamiento, que en distancia estén más cercanos al objeto a clasificar, de allí el nombre k vecinos más cercanos [44, 45].

Los resultados que arroja el algoritmo son los siguientes:

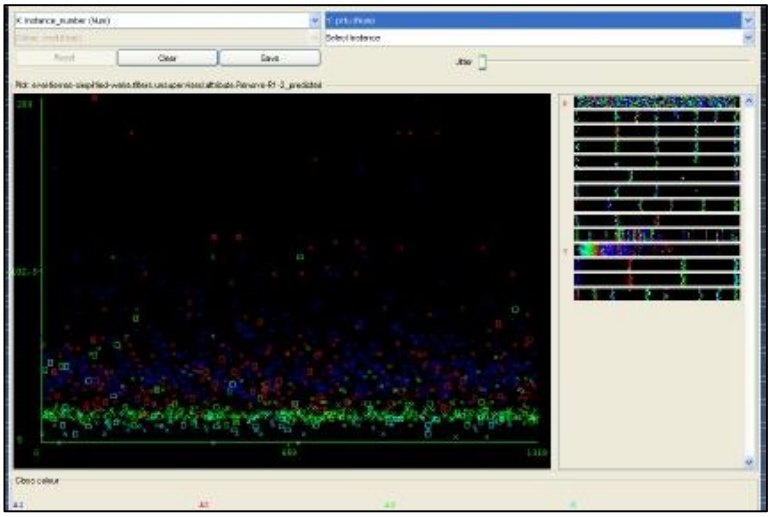


Figura 10. Gráfica del resultado del algoritmo NNge [45].

En la figura 10 podemos observar la gráfica y comparar los resultados obtenidos por el simulador en el cual las X indican las parcelas de manera correcta con respecto a la variable profundidad efectiva, es decir los datos ingresados son los más aptos para que en la parcela se cultive de la manera más prospera y los rectángulos indican lo contrario, parcelas no aptas [45].

2.3.5.5. Tabla de decisión (Decision Table)

El algoritmo permite modificar el número de búsquedas de correspondencias de los datos con las reglas una vez que ya no se produce mejore con la opción –S (valor por defecto 5); hicimos varias pruebas y vimos que el error cuadrático medio no variaba al variar el valor por defecto. También es posible variar el número de reglas (opción -x), para encontrar el menor error cuadrático medio. Sabemos que si un dato no encuentra correspondencia con ninguna regla, se asigna a la clase mayoritaria; el algoritmo esté más próxima. Incluyendo esta opción tampoco vimos mejore sobre el error [47, 48].

2.3.5.6. DTNB

Este algoritmo está basado en los algoritmos Decision Table y Naive Bayes. En cada momento de la búsqueda, el algoritmo evalúa la ventaja de dividir los atributos en dos subconjuntos disjuntos: uno para naive Bayes y otro para decisión table. Los atributos seleccionados son modelados en cada caso por Naive Bayes y por Decision Table respectivamente. En cada paso, el algoritmo considera también eliminación de un atributo completamente del modelo mientras que todos los atributos son modelados por la tabla de Decisión inicial [49].

2.4. Algoritmos no supervisados

2.4.1. Agrupamiento o Clustering

Se utiliza directamente similaridad entre los datos en entrada, es decir se identifica clusters o grupos en el conjunto de datos, donde un cluster es una colección de datos "similares". Los clusters deben ser de buena calidad es decir ser escalables a grandes bases de datos, esto es lo que intenta la minería de datos, la similitud determinada para formar los clusters puede medirse mediante funciones de distancia, especificadas por los usuarios o por expertos [29, 30].

En otras palabras mediante este algoritmo se agrupan registros. Usualmente es utilizado para darle al usuario final una visión más general de lo que ocurre en la base de datos. Después de haber trabajado algún tiempo con estas clasificaciones, se podrán empezar a hacer predicciones de lo que pasará con uno de los grupos ante un estímulo determinado. Esta técnica ayuda a encontrar los registros que se destacan del resto. Si se hace un cluster con ciertas características, quienes pertenezcan al grupo generalmente cumplirán con otras características [35].

- **Análisis de Clusters**

El análisis de clusters, tiene como objetivo principal segmentar un conjunto de datos en grupos o clusters, con objetos de características o atributos similares, pero al mismo tiempo entre los grupos los objetos sean diferentes. Los métodos de clustering intentan agrupar los objetos basándose en una "medida de similitud" [29].

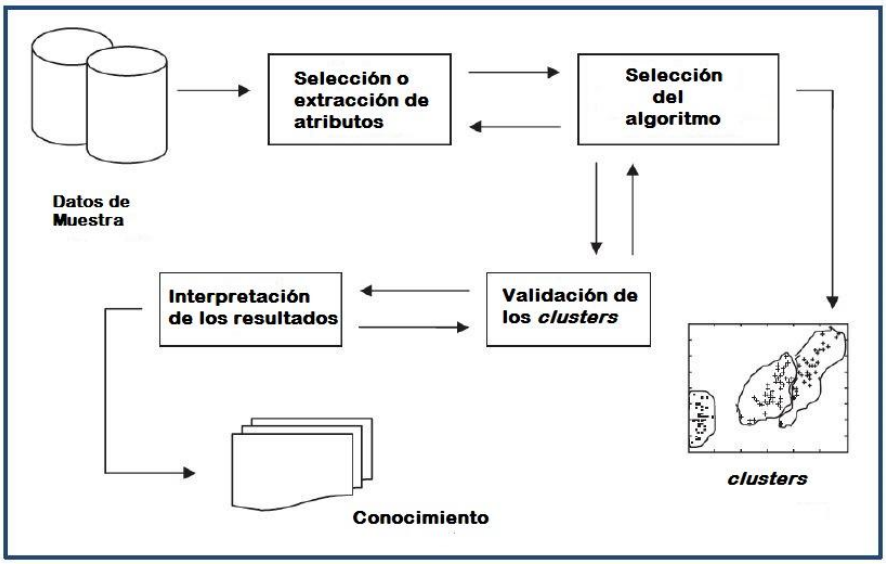


Figura 11. Etapas del análisis de clusters [36].

La figura 11 muestra las etapas del análisis de clusters descritas a continuación:

1. **Selección o extracción de atributos:** Se escogen los atributos principales con los que se quiere hacer clustering.
2. **Selección del algoritmo:** Elección del criterio de similitud adecuado.
3. **Validación de los clusters:** Se obtiene cantidad y composición de clusters distintos, dependiendo del algoritmo y los parámetros utilizados.
4. **Interpretación de los resultados:** Los resultados deben dar puntos de vista que tengan sentido para los usuarios finales a los que ha estado destinado el análisis.

2.4.1.2. K-Means

El algoritmo de las K – medias es un algoritmo altamente utilizado, de partición iterativo por naturaleza. Básicamente este algoritmo busca formar clusters o grupos de un conjunto de datos no etiquetados (sin previo agrupamiento) los cuales serán representados por K objetos llamados centroides [24, 38].

El algoritmo recibe como entrada el número de centros deseados y K-means mueve iterativamente los centros hasta minimizar la varianza dentro de cada grupo/partición. K-means opera sobre elementos que están representados por puntos en un espacio verctorial d – dimensional: es decir, clasificará un conjunto de vectores D con $d -$

$dimensiones$ donde $D = \{x_i | i = 1 \dots N\}$ y $x_i \in R^d$ denota al $i - ésimo$ punto o elemento [35, 38].

Para iniciar K-means debe determinar un centro inicial para cada grupo en que se vaya a dividir D. El algoritmo itera básicamente sobre dos pasos:

- **Asignación:** Cada punto es asignado al centro más cercano. Al final de la asignación se tienen una nueva partición de los datos.
- **Relocalización de medias (means):** se recalcula el centro de cada grupo, con la media aritmética de todos los puntos asignados a él.

Como se puede deducir con la explicación el algoritmo utiliza la noción de centroide. Cada uno de estos centroide es el valor medio de los objetos que pertenecen a dicho grupo o cluster. Este proceso termina cuando no hay cambio en las asignaciones y por tanto tampoco en ninguno de los valores de los centros, donde K es un valor dependiendo de las características que compartan los datos [35, 38].

A continuación tenemos la aplicación del algoritmo K-means en la herramienta rapidminer. El conjunto de datos conrresponde a la base de datos iris integrada dentro de los ejemplos que presenta la herramienta, base de datos que corresponde a 4 especies de flores y cuatro medidas respecto de dichas flores. Con los operadores aplicamos el algoritmo K-means sobre los datos (ver figura 12). Obteniendo los siguientes resultados:

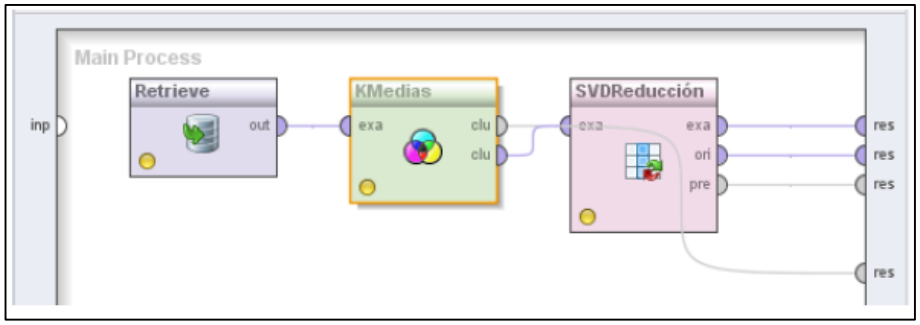


Figura 12. Proceso del algoritmo K-means [38].

Observamos que luego de ejecutar el algoritmo arroja el resultado de tres clusters que corresponden a los diferentes grupos de ítems (ver figura 13).

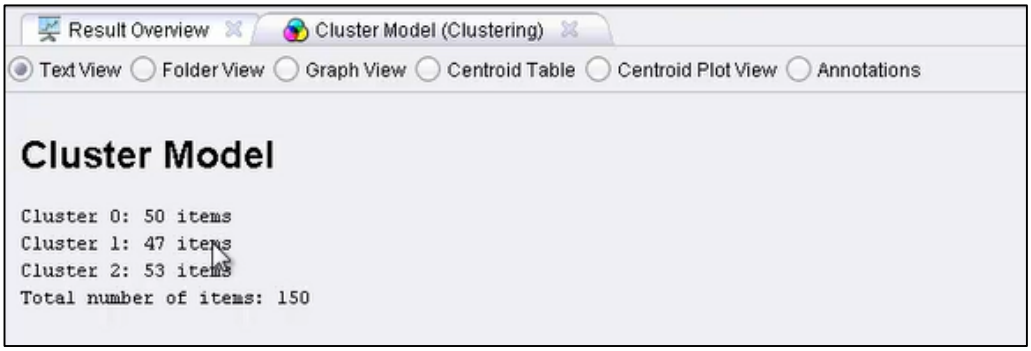
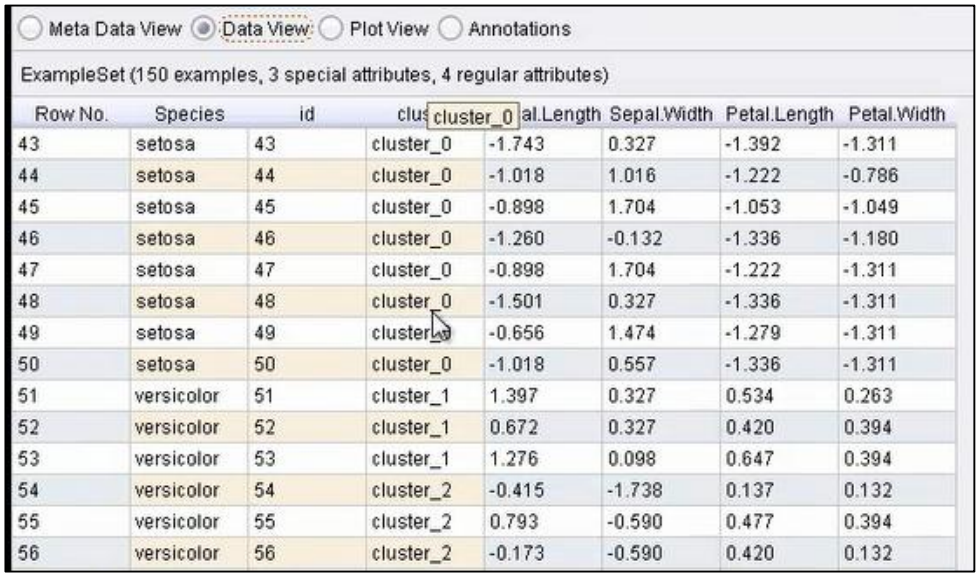


Figura 13. Resultado del algoritmo K-means [38].

Observamos que el algoritmo agrupa los ítems en tres clusters de acuerdo a las características en comunes de cada uno. Cluster es la variable que determina a que grupo pertenece cada una de las 150 observaciones que se tiene.



Row No.	Species	id	cluster_0	al.Length	Sepal.Width	Petal.Length	Petal.Width
43	setosa	43	cluster_0	-1.743	0.327	-1.392	-1.311
44	setosa	44	cluster_0	-1.018	1.016	-1.222	-0.786
45	setosa	45	cluster_0	-0.898	1.704	-1.053	-1.049
46	setosa	46	cluster_0	-1.260	-0.132	-1.336	-1.180
47	setosa	47	cluster_0	-0.898	1.704	-1.222	-1.311
48	setosa	48	cluster_0	-1.501	0.327	-1.336	-1.311
49	setosa	49	cluster_0	-0.656	1.474	-1.279	-1.311
50	setosa	50	cluster_0	-1.018	0.557	-1.336	-1.311
51	versicolor	51	cluster_1	1.397	0.327	0.534	0.263
52	versicolor	52	cluster_1	0.672	0.327	0.420	0.394
53	versicolor	53	cluster_1	1.276	0.098	0.647	0.394
54	versicolor	54	cluster_2	-0.415	-1.738	0.137	0.132
55	versicolor	55	cluster_2	0.793	-0.590	0.477	0.394
56	versicolor	56	cluster_2	-0.173	-0.590	0.420	0.132

Figura 14. Vista de los datos después de ejecución de K-means [38].

La gráfica de coordenadas paralelas representa en cada uno de los ejes las cuatro variables, donde el color permite distinguir los distintos clusters. Por ejemplo el clusters azul se caracteriza por tener el petal length mas bajo y el sepal width más alto, y as;i sucesivamente (ver figura 15).

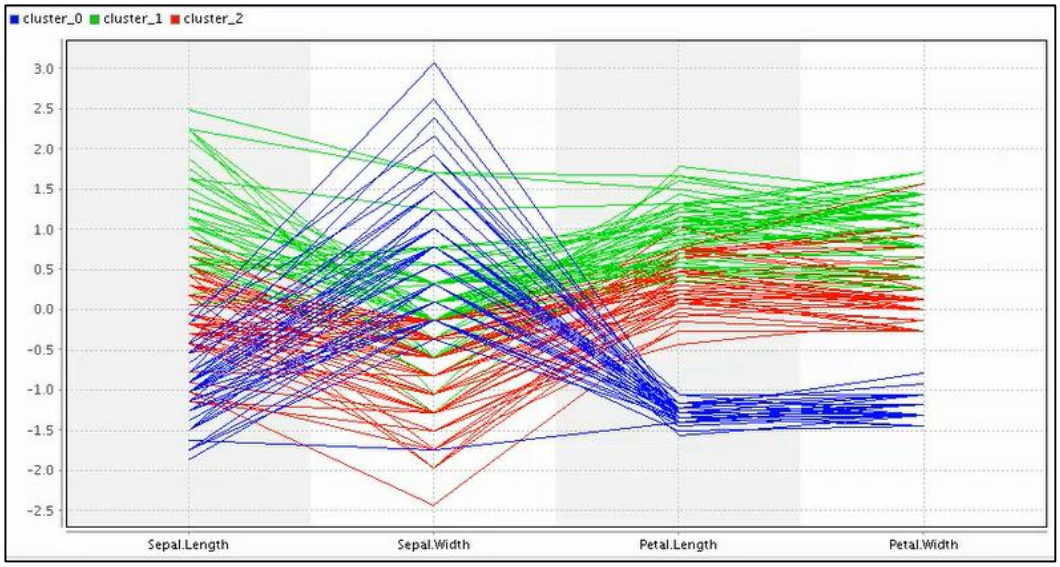


Figura 15. Gráfica de las coordenadas paralelas K-means [38].

2.4.2. Reglas asociación

Las reglas se asociación son parecidas a las reglas de clasificación. Tienen un lado izquierdo con condicionales que debe cumplir, y un lado derecho con las consecuencias de cumplir estas condiciones. Utilizan el proceso de cobertura que corresponde al número de instancias predichas correctamente. A pesar de ello, en el lado derecho de las reglas, puede aparecer cualquier para lo pares atributo-valor. Se ha encontrado ese tipo de reglas al considerar cada posible combinación de pares atributo-valor del lado derecho. De esta forma el paso próximo es podarlas usando cobertura y precisión respecto a la proporción de número de instancias a las cuales aplica la regla [35].

En las reglas de asociación, la cobertura también se la conoce como soporte y la precisión como confianza. Hablando probabilísticamente, el soporte y la confianza son determinados en base a las siguientes fórmulas:

$$soporte\ (A \Longrightarrow B) = P(A \cup B)$$

$$confianza(A \Longrightarrow B) = P(B|A) = \frac{soporte(A \cup B)}{soporte(A)}$$

La fórmula de soporte está dada por la unión de dos probabilidades, es decir el soporte de la regla $(A \rightarrow B)$ es la equivalente a probabilidad que se cumplan simultáneamente

A y B. Mientras que la fórmula de confianza está dada en términos de probabilidades condicionales, esto se lo ha interpretado como la probabilidad que curra B, dado que ocurra A, y su equivalente en términos de soporte [35].

Es muy importante tener en cuenta que en realidad la importancia está únicamente en reglas que tienen mucho soporte, por lo que se busca, pares atributo-valor o ítem que cubran una gran cantidad de instancias o ítem-sets. Para mayor entendimiento del algoritmo se lo ha explicado con un ejemplo práctico [35], [38]. En la tabla XIII tenemos los datos que contienen los ítems o atributos Ambiente, Temperatura, Humedad, viento, Clase.

TABLA XIII

DATOS DE EJEMPLO – REGLAS DE ASOCIACIÓN.

Ambiente	Temp.	Humedad	Viento	Clase
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
Lluvia	media	alta	no	P
lluvia	baja	Normal	no	P
lluvia	baja	Normal	si	N
nublado	baja	Normal	si	P
soleado	media	Alta	no	N
soleado	baja	normal	no	P
Lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvia	media	alta	si	N

Con los datos de la tabla, el itemset:

humedad = normal, viento = no, clase = P

Puede producir las siguientes posibles reglas:

If humedad=normal and viento=no Then clase=P 4/4
If humedad=normal and clase=P Then viento=no 4/6
If viento=no and clase=P Then humedad=normal 4/6
If humedad=normal Then viento=no and clase=P 4/8
If viento=no Then clase=P and humedad=normal 4/8
If clase=P Then viento=no and humedad=normal 4/9
If true Then humedad=normal and viento=no and clase=P 4/12

Si se piensa en 100% de éxito, entonces sólo la primera regla cumple. De hecho existen 58 reglas considerando la tabla completa que cubren al menos dos ejemplos con un 100% de exactitud (accuracy). El proceso es el siguiente:

1. Genera todos los ítems sets con un elemento. Usa estos para generar los de dos elementos, y así sucesivamente. Donde se toman todos los posibles pares que cumplen con las medidas mínimas de soporte. Esto permite ir eliminando posibles combinaciones ya que no todas se tienen que considerar.

2. Genera las reglas revisando que cumplan con el criterio mínimo de confianza. Donde se ha tomado en cuenta otro ejemplo con los datos de la Tabla XIV que contiene listas de compras de productos:

TABLA XIV

DATOS PRODUCTOS POR COMPRA

id1	**P1, p2, p5**
id2	P2, p4
id3	P2, p3
id4	p1, p2, p3
id5	p1, p3
id6	p2, p3
id7	p1, p3
id8	p1, p2, p3, p5
id9	p1, p2, p3

Los datos de la tabla XIV han sido representados gráficamente en la figura 16:

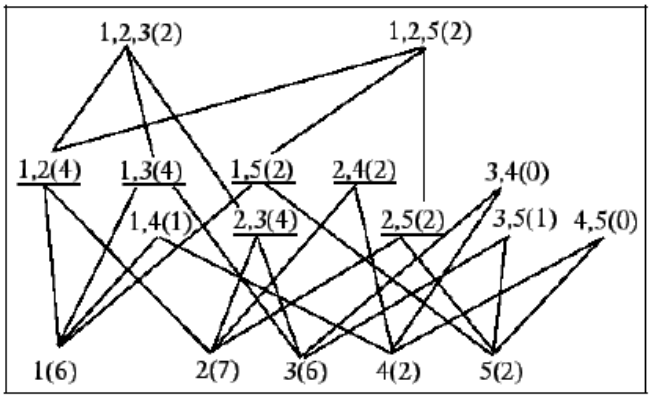


Figura 16. Generación de ítems-sets [35].

Se muestra la generación de candidatos por niveles. El primer número indica el producto y el número entre paréntesis las veces que ocurre.

Una vez que se tienen los conjuntos de ítems, generar las reglas es relativamente sencillo.

- Para cada conjunto l de ítems, genera todos sus subconjuntos.
- Para cada subconjunto , genera una regla:

$s \Rightarrow (l - s)$ si:

$$\frac{soporte(l)}{soporte(s)} \geq nivel\_confianza$$

Todas las reglas satisfacen los niveles mínimos de soporte.

Se han hecho algunas mejoras al algoritmo básico de reglas de asociación (apriori) para hacerlo más eficiente:

- Usar tablas hash para reducir el tamaño de los candidatos de los itemsets.
- Eliminar transacciones (elementos en la base de datos) que no contribuyan en súper conjuntos a considerar.
- Dividir las transacciones en particiones disjuntas, evaluar itemsets locales y luego, en base a sus resultados, estimar los globales.

- Hacer aproximaciones con muestreos en la lista de productos, para no tener que leer todos los datos.

2.4.3. Criterios de comparación de algoritmos

Se deben tener establecidos criterios o medidas que decidan que un algoritmo es de calidad, es decir que es apto para la resolución del problema como [35]:

- **Velocidad de Ejecución.-** Aquí se mide el tiempo que el algoritmo tarda para obtener el modelo, y arrojar los resultados, a partir de los datos con los que trabaja.

- **Escabilidad.-** es la habilidad de construir un modelo eficiente con grandes cantidades de datos, respecto al dominio con el que se trabaja.

- **Robustez.-** En caso de que se den datos con ruidos o con valores faltantes, para que el modelo sea robusto debe ser capaz de hacer predicciones correctas con estos datos.

- **Precisión para clasificación de datos.-** Esta es una medida para determinar si los datos son clasificados correctamente por el algoritmo de acuerdo a las reglas creadas.

- **Precisión para predecir datos futuros.-** Se realizan pruebas para determinar la precisión con la que trabaja el algoritmo clasificando registros con el valor que se desea predecir.

3. CAPÍTULO III. HERRAMIENTAS PARA EL PROCESO DE MINERÍA DE DATOS.

Existen diversas herramientas para el tratamiento automatizado de los datos, que permiten aplicar técnicas de minería de datos y obtener conocimiento de manera más rápida en el procesamiento de grandes cantidades de datos. Estas herramientas pueden ser comerciales o libres, para realizar un estudio y comparación entre ellas se han tomado en cuenta las herramientas de uso libre más utilizadas como:

3.1. Orange

Es una suite de software para minería de base de datos y aprendizaje automático basado en componentes que cuenta con un fácil y potente, rápido y versátil front-end de programación visual para el análisis exploratorio de datos y visualización, y librerias para Python y secuencias de comando. Contiene un completo juego de componentes para preprocesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo, y técnicas de exploración. Además cuenta con las siguientes características [51]:

- Basado en componentes que cuenta con un fácil y potente, rápido y versátil front-end de programación visual para el análisis exploratorio de datos y visualización, y librerias para Python y secuencias de comando.
- Contiene un completo juego de componentes para preprocesamiento de datos.
- Los componentes de Orange solo pueden ser manipulados desde programas desarrollados en Python.

3.2. Weka

Escrito en Java, Weka es una conocida suite de software para el aprendizaje y la máquina que soporta varias tareas de minería de datos típicos, especialmente los datos del proceso previo, el agrupamiento, clasificación, regresión, visualización y selección de características. WEKA proporciona acceso a bases de datos SQL utilizando Java Database Connectivity y puede procesar el resultado devuelto por una consulta de base de datos. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o una relación, donde se etiqueta cada punto de datos por un número fijo de atributos. Por ello se mencionan las características más relevantes [52, 53].

- Weka proporciona interfaces para la comunicación con el usuario.
- Proporciona una consola para poder introducir mandatos. Nos permite realizar tareas complejas.
- Nos permite ubicar patrones de comportamiento de la información a procesar de tal manera que es de gran ayuda en la toma de decisiones.

- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado.
- Es complicada de manejar ya que es necesario un conocimiento completo de la aplicación.
- Existe poca documentación sobre el uso de Weka dirigida al usuario.
- Un área importante que actualmente no cubren los algoritmos incluidos en Weka es el modelado de secuencia

3.3. KNIME

Es de uso fácil y comprensible, y de fuente abierta de integración de datos, procesamiento, análisis, y la plataforma de exploración. Se ofrece a los usuarios la capacidad de crear de forma visual los flujos de datos o tuberías, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas. KNIME está escrito en Java y está basado en Eclipse y hace uso de su método de extensión para apoyar plugins proporcionando así una funcionalidad adicional. A continuación se detallan las características principales de esta herramienta [54].

- Es de uso fácil y comprensible, y de fuente abierta de integración de datos, procesamiento, análisis, y la plataforma de exploración.
- Basado en Eclipse y hace uso de su método de extensión para apoyar plugins proporcionando así una funcionalidad adicional.
- A través de plugins, los usuarios pueden añadir módulos de texto, imagen, y el procesamiento de series de tiempo.
- Integración de varios proyectos de código abierto, tales como el lenguaje de programación de R, WEKA, el Kit de desarrollo de la Química, y LIBSVM.

3.4. JHepWork

Es un framework para análisis de datos libre y de código abierto que fue creado como un intento de hacer un entorno de análisis de datos usando paquetes de código abierto

con una interfaz de usuario comprensible y para crear una herramienta competitiva a los programas comerciales. Contiene bibliotecas científicas numéricas implementadas en Java para funciones matemáticas, números aleatorios, y otros algoritmos de minería de datos. jHepWork se basa en Jython un lenguaje de programación de alto nivel, pero codificación en Java también puede ser usada para llamar librerías jHepWork numéricas y gráficas. Entre otras de sus características tenemos [55]:

- Diseñado para los científicos, ingenieros y estudiantes.
- Es una herramienta competitiva a los programas comerciales.
- Especialmente para las ploteos científicos interactivos en 2D y 3D.
- Contiene bibliotecas científicas numéricas implementadas en Java para funciones matemáticas, números aleatorios, y otros algoritmos de minería de datos.

3.5. RapidMiner

Está disponible como una herramienta independiente para el análisis de datos y como un motor de datos, minería de datos que se pueden integrar en sus propios productos. Es un ambiente para aprendizaje automático y minería de datos de los experimentos que se utiliza para la investigación y en el mundo real los datos de tareas de minería de ambos. Combina el aprendizaje de los regímenes y los evaluadores de atributos del medio ambiente de Weka aprendizaje y presenta otras características como [56]:

- Cuenta con un módulo para integración con Weka y R.
- Multiplataforma
- Puedo usarse a través de diversas maneras, como de un GUI, en línea de comandos, por lotes, desde otros programas a través de llamadas a sus bibliotecas.
- Se usa en investigación educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales.
- RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización.

3.6. Análisis de las Herramientas más utilizadas

Se han realizado un sin número de investigaciones y un análisis de las herramientas de minería de datos desde diferentes puntos de vista para determinar cuál sería la mejor elección. Uno de los criterios ha sido de acuerdo a las que más se usan en la actualidad, así como las más fáciles de usar. A través de encuestas realizadas por diferentes autores se ha visto la evolución de las herramientas a lo largo de los años. En una encuesta realizada en el año 2005, se ve el dominio de la herramienta Rapidminer, tomando en cuenta algunas de las herramientas que han sido de nuestro objeto de estudio [57].

Figura 17. Estadísticas del uso de herramientas de minería de datos [57].

En la figura 17 podemos observar que en el año 2010 y 2011 Rapidminer es la herramienta más usada. Esta herramienta implementa un módulo compatible con R y Weka que le siguen en la estadística en utilización. Concluyendo que Rapidminer por la necesidad es una herramienta muy útil, fácil de usar y un apoyo a la toma de decisiones [57].

Así mismo en una encuesta realizada por KDnuggets, un periódico de minería de datos, RapidMiner ocupó el segundo lugar en herramientas de analítica y de minería de datos utilizadas para proyectos reales en el 2009 y fue el primero en el año 2010 [57].

e. Materiales y Métodos

Para el desarrollo del presente trabajo de titulación ha sido necesario un análisis de la información obtenida de las distintas fuentes de datos, realizando procesos como el filtrado, preparación, explotación de los datos para la identificación de las variables influyentes y generación del modelo final para la determinación de perfiles profesionales con la aplicación de técnicas de minería de datos. Es por ello que basándose en un estudio comparativo de las metodologías SEMMA, Catalyst o P3TQ y CRISP-DM [108], se ha determinado que CRISP-DM es la más utilizada actualmente para proyectos de minería de datos y al contar con un número de fases, sub-fases y actividades en mayor cantidad que las otras metologías permite de manera organizada documentar detalladamente el desarrollo del proceso de minería de datos en todas sus etapas permitiendo su culminación con éxito, es por ello que ha sido escogida como la metodología para el desarrollo del presente trabajo.

La metodología CRISP-DM propiamente enfocada a minería de datos ha permitido cumplir con los objetivos establecidos y realizar la validación del modelo de minería de datos, y así determinar la calidad de los datos sobre los que se ha trabajado para realizar todo el proceso. Los proyectos realizados con esta metodología cumplen un ciclo de vida (ver Figura 18) lo cual desemboca en resultados de confianza.

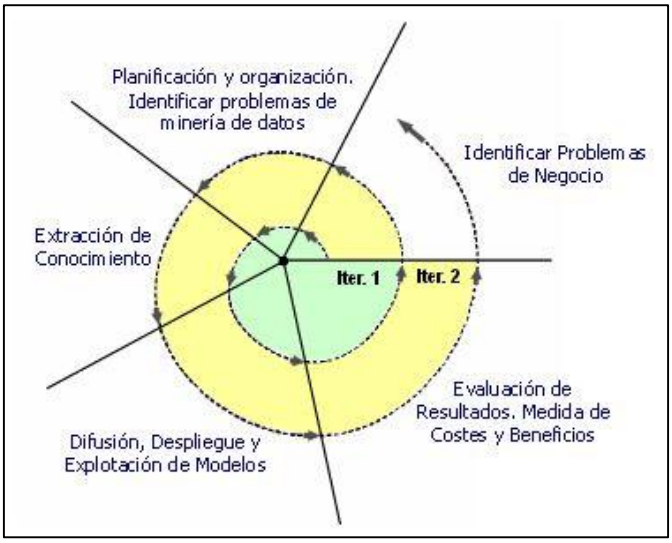


Figura 18. Ciclo de vida de un proyecto con CRISP-DM [58].

La metodología de la minería de datos CRISP-DM se describe en términos de un modelo jerárquico de proceso, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción, de lo general a lo específico [59], como podemos observar en la figura 19:

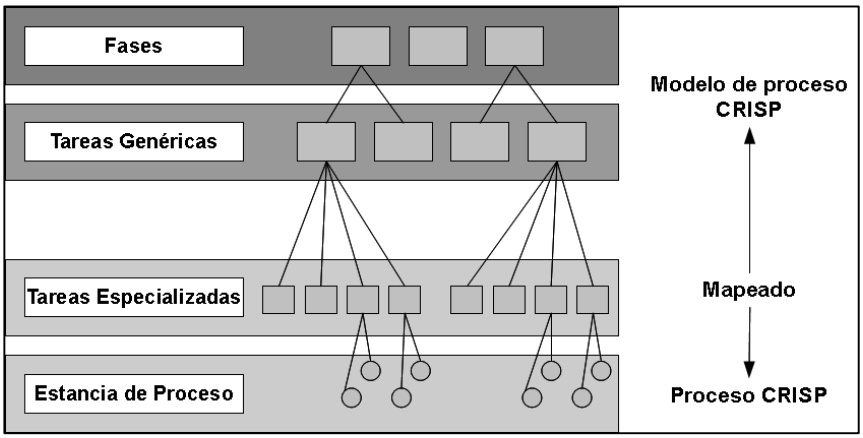


Figura 19. Los cuatro niveles de la Metodología CRISP-DM [59].

En el nivel superior, el proceso de minería de datos se organiza en una serie de Fases, cada fase consta de varias tareas genéricas de segundo nivel. El segundo nivel se llama Genérico o Tareas Genéricas, ya que está destinado a ser lo suficientemente general para abarcar todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sean posible. Completas significa abarcar tanto todo el proceso de minería de datos y todas las posibles aplicaciones de minería de datos. Estable significa que el modelo debe ser válido para las evoluciones todavía no previstas, como las nuevas técnicas de modelado [59].

El tercer nivel, el nivel de Tareas Especializadas, es dónde se describe cómo deberían llevarse a cabo acciones en las tareas genéricas de ciertas situaciones específicas. Por ejemplo: en el segundo nivel podría haber una tarea genérica llamada “Limpieza de datos”, entonces en el tercer nivel se diferencia cómo ésta tarea puede ser disímil en diferentes situaciones, como la “limpieza de los valores numéricos” vs. “valores categóricos” o si el tipo de problema es la “agrupación o el modelado predictivo”. La descripción de las fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y, a menudo será necesario dar marcha atrás en varias ocasiones a las tareas anteriores y repetir ciertas acciones. El modelo de proceso no intenta capturar todas las rutas posibles a través del proceso de

minería de datos, ya que esto requeriría un modelo de proceso demasiado complejo [59].

El cuarto nivel, la Estancia de Proceso, es un registro de las acciones, decisiones y resultados de una participación real de la minería de datos. Una instancia de proceso está organizada de acuerdo a las tareas definidas en los niveles superiores, pero representa lo que realmente ocurrió en un trabajo específico asignado, en lugar de lo que sucede en general [59].

Esta metodología consta de varias fases, como modelo de referencia para el desarrollo de proyectos de minería de datos, ayudando a la organización y mejora de las actividades, estas fases están conformadas a su vez por varias sub-fases y las mismas por actividades que son resumidas a continuación (ver Tabla XV) [58]:

TABLA XV

FASES DEL MODELO DE REFERENCIA CRISP-DM

<table>
<tr><th colspan="5">Comprensión del Negocio</th></tr>
<tr><th>Determinar los objetivos de negocio.</th><th>Evaluar la situación.</th><th>Determinar el objetivo de MD.</th><th colspan="2">Desarrollar el plan de proyecto</th></tr>
<tr><td>Antecedentes
Objetivos del negocio
Criterio de Éxito</td><td>Inventario de requerimientos de recursos, hipótesis y limitaciones.
Riesgos y Contingencias.
Terminología
Costes y beneficios.</td><td>Objetivos de minería de datos.
Criterio de éxito de MD.</td><td colspan="2">Plan del proyecto
Evaluación inicial de herramientas y técnicas.</td></tr>
<tr><th colspan="5">Comprensión de los datos</th></tr>
<tr><th>Obtener los datos iniciales</th><th>Describir los datos</th><th>Exploración de datos</th><th colspan="2">Verificación de la calidad de los datos</th></tr>
<tr><td>Reporte de la obtención de los datos iniciales.</td><td>Reporte de la descripción de los datos.</td><td>Reporte de la exploración de datos.</td><td colspan="2">Reporte de la calidad de los datos.</td></tr>
<tr><th colspan="5">Preparación de los datos</th></tr>
<tr><th>Selección de datos</th><th>Limpieza de datos</th><th>Construcción de datos</th><th>Integración de datos</th><th>Formateo de datos</th></tr>
<tr><td>Razones de inclusión / exclusión.</td><td>Reporte de limpieza de datos.</td><td>Atributos derivados.
Registros generados.</td><td>Datos combinados.</td><td>Datos formateados.</td></tr>
<tr><th colspan="5">Modelamiento</th></tr>
<tr><th>Selección de la técnica de modelado</th><th>Generar el diseño prueba</th><th>Construcción del modelo</th><th colspan="2">Evaluación del modelo</th></tr>
<tr><td>Técnica de modelado.</td><td>Diseño de prueba.</td><td>Parámetros del modelo.</td><td colspan="2">Evaluación del modelo.</td></tr>
</table>

Modelamiento. Hipótesis.		Descripción del modelo.	Revisión de la configuración de los parámetros del modelo.
Evaluación			
Evaluar los resultados		**Revisar el proceso**	**Establecimiento de los siguientes pasos**
Evaluación de los resultados de minería de datos, modelos aprobados.		Revisión del proceso.	Lista de posibles acciones, decisión.

Las fases de la metodología aplicadas en el desarrollo del presente trabajo de titulación han sido detalladas a continuación:

- **Fase uno:** Comprensión del negocio.- En esta tarea se describe los antecedentes o contexto inicial, los objetivos del negocio y los criterios de éxito, en otras palabras las necesidades del cliente, así como los factores que pueden influenciar en el resultado final, con el fin de ahorrar esfuerzo innecesario.

- **Fase dos:** Comprensión de los datos.- Esta es la fase de la metodología donde se pretende reducir la información a la únicamente necesaria para realizar la minería de datos, a su vez relacionarse directamente con la información para su mayor comprensión. Al realizar la comprensión se pasará a la preparación de los datos con las bases necesarias para realizar esta nueva fase con éxito.

- **Fase tres:** Preparación de los datos.- Esta fase permite construir el conjunto de datos final, debido a que engloba todas las actividades esenciales para llegar a este objetivo. La preparación de los datos requiere de esfuerzo y muchas veces de repetir algunas de las tareas. Se realizan desde tareas generales como selección de datos, con la selección de tablas, registros, y atributos a los, transformación de datos, cambios de formato, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos, etc. Datos a los que se ha aplicado técnicas de minería de datos específicas, detalladas en las posteriores fases.

- **Fase cuatro:** Modelamiento.- En esta fase se describen las diferentes técnicas de modelado elegidas y los parámetros aplicados en cada una de ellas. Previamente se aplicaron actividades de preparación de datos para generar atributos derivados a los cuales se aplicaron las técnicas. Una vez aplicadas estas técnicas se evaluó y comparo los resultados obtenidos de acuerdo al contexto del problema. En esta fase

se ha utilizado la herramienta RapidMiner luego de una previa selección para realizar el proceso de minería.

- **Fase cinco:** Evaluación.- Corresponde a la fase del análisis de los resultados obtenidos en la minería, se detalla los resultados más positivos, tomando en cuenta el porcentaje de clasificación en el conjunto de entrenamiento, los resultados del método de validación cruzada, la lógica de las reglas obtenidas, etc, respecto del rendimiento de los algoritmos que han arrojado los mejores resultados para la determinación de perfiles profesionales.

f. Resultados

A continuación se describen los resultados de cada uno de los objetivos planteados para el desarrollo del presente trabajo de titulación, para lo que se ha utilizado un sin número de herramientas, materiales, métodos, etc. Elementos que han sido de vital importancia para el logro del trabajo antes mencionado.

HIPOTESIS

Las técnicas de minería de datos permiten determinar los perfiles profesionales de estudiantes egresados y graduados de la carrera de ingeniería en sistemas, tomando como patrones el registro de sus records académicos, habilidades, capacidades e intereses propios de cada individuo.

A continuación se detalla de manera esquematizada las fases y actividades que engloban los objetivos del presente trabajo de titulación:

1. ETAPA UNO: Investigar las características y variables más influyentes de las fuentes de datos a utilizar.

1.1. Investigar sobre casos de éxito acerca de la aplicación de Minería de Datos, y específicamente respecto al ámbito educativo.

La minería de datos en la educación ofrece numerosas ventajas en comparación con los paradigmas más tradicionales de investigación relativa a la educación, como experimentos de laboratorio, estudios sociológicos o investigación de diseño. En particular, la creación de repositorios públicos de datos educacionales ha creado una base que hace posible aplicar la minería de datos en este campo. En particular los datos de estos repositorios son totalmente válidos y cada vez más fácilmente accesibles para comenzar una investigación. Estos puntos permiten a los investigadores ahorrar mucho tiempo en tareas como la búsqueda de individuos, organización de los estudios, recopilación de datos, ya que estos se encuentran directamente accesibles.

La Minería de datos ha sido muy utilizada en el ámbito educativo, para obtener valiosa información sobre el comportamiento de los estudiantes, información que ayuda a la toma de decisiones**.** Las técnicas de minería de datos aplicadas han dado apoyo a los sistemas de cursos e-learning y un sin número de sistemas enfocados en la educación. Las técnicas han permitido obtener modelos para detectar los factores que influyen en la deserción y abandono de los estudiantes en su vida académica, en el análisis del rendimiento académico de los estudiantes, etc.

Se han recopilado algunos de los casos de éxito orientados al ámbito educativo, que demuestran el impacto de la minería de datos en la educación. Estos casos de éxito se encuentran detallados en el apartado revisión de literatura, CAPÍTULO I. CASOS DE ÉXITO MD.

- **Caso de Éxito 1: Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web.**

El estudio realizado ha demostrado el éxito de la minería de datos dentro de áreas como: minería de utilización web, realizando clasificación y agrupamiento, descubrimiento de reglas de asociación o secuencias de patrones, minería de datos como minería de texto, detección de irregularidades y razonamiento basado en casos. Las aplicaciones de la minería de datos se las puede dividir en dos grupos dependiendo del sistema de enseñanza. Tomando en cuenta ello se tiene los entornos de enseñanza basados en web tradicionales como los sistemas de cursos e-learning y los sistemas Hipermedia Adaptativos así como los Sistemas tutores Inteligentes.

Finalmente el estudio arroja las líneas de investigación donde actualmente se trabaja con la minería de datos o áreas de aplicación a futuro enfocadas a la educación basadas en web donde se dará facilidad de utilización de los algoritmos de minería de datos. El desarrollo de herramientas más fáciles e intuitivas de utilizar con la integración de algoritmos de minería de datos dentro de las propias herramientas, dando mantenimiento permanente a los cursos, para la mejora continua de los sistemas. Desarrollo de algoritmos para realización de sugerencias sobre rutas, actividades en los entornos hipermedia adaptativos y sistemas tutores inteligentes basados en web, etc.

- **Caso de Éxito 2: Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning.**

En la investigación realizada se ha presentado un sistema recomendador colaborativo que utiliza minería de datos distribuida para la continua mejora de cursos de e-learning. Para el sistemas se ha diseñado e implementado un nuevo algoritmo de minería de reglas de asociación interactivo e iterativo que utiliza una nueva medida de evaluación de las reglas descubiertas basada en pesos y que tiene en cuenta la opinión de los expertos y de los propios profesores para producir recomendaciones cada vez más efectivas, para permitir que profesores de perfil similar, compartan los resultados de sus investigaciones como resultado de aplicar minería de manera local sobre sus propios cursos.

Los resultados finales de las pruebas realizadas demostraron que los problemas detectados se reducirían en consecutivas ejecuciones del curso y por otra, que las notas finales de los alumnos mejorarían en la medida que el profesor iba corrigiendo los problemas. Existen aún trabajos futuros por hacer sobre el sistema, debido a que la aplicación de técnicas de minería de datos no es estática de acuerdo a las nuevas variables, nuevos datos; se pueden obtener nuevos modelos que permitan tomar decisiones que aporten significativamente en el mejoramiento de la educación de manera continua.

- **Caso de Éxito 3: Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.**

El presente trabajo se trata del análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia (UPV) y con titulaciones de Ingeniería Técnica en Informática de Gestión y de Ingeniería Informática. Se ha hecho uso de técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, económico, etc. Este caso de éxito ha proporcionado una herramienta importante para la acción tutorial, que puede apoyarse

en las predicciones de los modelos que se obtienen para encauzar sus recomendaciones y encuadrar las expectativas y el esfuerzo necesario para cada alumno.

Las técnicas de minería de datos proporcionan una herramienta que permite determinar qué características de los alumnos de nuevo ingreso son más relevantes permitiendo estimar su rendimiento académico el primer año. En el caso estudiado, factores como los estudios previos del alumno y la nota de ingreso en la titulación aparecen de manera repetida como claramente correlacionados con el rendimiento académico el primer año. También aparecen factores que podrían influir en el rendimiento, como las ocupaciones y estudios de los padres, o la edad de ingreso del alumno, aunque estos dependen de la técnica utilizada. En cambio, el país de procedencia o el lugar de residencia (del alumno o de su familia) no aparecen en ningún caso.

Con este estudio se busca la mejora de la calidad del proceso educativo. En este sentido, es interesante analizar el rendimiento académico de los estudiantes para tomar medidas oportunas tanto de forma individual como global.

- **Caso de Éxito 4: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos**

Este proyecto está orientado al ámbito Educativo con la aplicación de técnicas de Minería de Datos para detectar los factores que influyen para que los estudiantes de enseñanza media o secundaria fracasen. El objetivo final es detectar lo antes posible a los estudiantes que presenten esos factores para poder ofrecerles algún tipo de atención o ayuda para tratar de evitar y/o disminuir el fracaso escolar. Se han hecho uso de tres fuentes de datos: una encuesta en papel pasada al formato electrónico, los datos del CENEVAL y los datos del departamento escolar.

Para el proceso de selección se aplicó 10 algoritmos de selección de mejores atributos sobre el fichero de datos. De esta manera partiendo de 77 atributos originales se ha reducido a los 15 mejores, seleccionados por la frecuencia de aparición de acuerdo a los resultados arrojado por algoritmos. Para la aplicación de la minería de datos se partió en 10 ficheros de entrenamiento y 10 ficheros de prueba.

Se realizaron cuatro experimentos. En un primer experimento se ha ejecutado 10 algoritmos de clasificación utilizando los 77 atributos iniciales. En un segundo experimento, se ha utilizado los 15 mejores atributos escogidos por los algoritmos de selección. Un tercer experimento, donde se ha repetido las ejecuciones pero utilizando los ficheros de datos re- balanceados y finalmente se ha considerado diferentes costos de clasificación, todo ello con el objetivo de obtener la máxima exactitud de clasificación que determinen los resultados más óptimos.

- **Análisis Final de los Casos expuestos**

Las técnicas y algoritmos de minería de datos usados en los casos de éxito que han sido objeto de análisis se resumen en la tabla XVI presentada a continuación:

TABLA XVI

RESUMEN DE LAS TÉCNICAS DE MD APLICADAS EN LOS CASOS DE ÉXITO

Caso de Éxito	Tipo de MD Aplicada	Técnicas de MD
Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web	web mining	Clasificación y agrupamiento, descubrimiento de reglas de asociación y secuencias de patrones
Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning	Minería de datos distribuida y filtrado colaborativo	Algoritmo Apriori predictivo para el descubrimiento de las reglas de asociación
Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.	Minería de datos educativa	Árbol de decisión Regresión Lineal Clustering
Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos	Minería de datos educativa	Algoritmos de inducción de reglas de clasificación: JRip, NNge, OneR, Prism y Ridor

1.2. Analizar qué características sirven para determinar los perfiles profesionales de los egresados y titulados de la Carrera de Ingeniería en Sistemas.

Partiendo de que la minería de datos se la realiza a través de la creación de modelos y que estos a su vez se los desarrolla partiendo del reconocimiento de patrones, los mismos que pueden ser llamados características. Se debe realizar un análisis de cuáles son las características que sirven para lograr el objetivo que se intenta alcanzar con la aplicación de la minería de datos. En el caso del presente trabajo de titulación donde se busca determinar los perfiles profesionales existen un sin número de características importantes como:

Nombres de los estudiantes, nombre del módulo, las notas en cada materia, las notas en cada módulo, las horas planificadas de cada materia, las horas dictadas en cada materia, las horas asistidas en cada materia, los aspectos más relevantes del perfil de egreso de la carrera de Ingeniería en Sistemas.

Tomando en cuenta que los perfiles profesionales son las actividades de las distintas ramas de empleos por los que puede optar un profesional de la carrera de Ingeniería en sistemas. La determinación de los perfiles profesionales, se la realizará en base a estas características que son la representación del registro que presenta cada estudiante durante su vida universitaria.

Para la carrera de Ingeniería en Sistemas se realiza un análisis de la malla curricular base, que consta con diez módulos formados por distintas unidades, que permiten la formación del ingeniero en sistemas (ver tabla XVII):

TABLA XVII

MALLA CURRICULAR CIS 1990-2013

Mód	Nombre	Unidades
1	LAS PROBLEMATICAS GLOBALES DE LA REALIDAD SOCIAL Y LAS ALTERNATIVAS DE NIVELACION EN EL CAMPO INGENIERIL.	Química, proceso Investigativo, matemáticas, física, geometría plana.
2	BASES CIENTÍFICO TÉCNICAS DE LA FORMACIÓN PROFESIONAL EN SISTEMAS	Matemáticas, física, algebra lineal, geometría analítica, proceso investigativo.

	INFORMÁTICOS Y COMPUTACIONALES.	
3	FUNDAMENTACIÓN CIENTÍFICA DE LA INGENIERÍA. EN SISTEMAS.	Matemáticas, matemáticas discretas, fundamentos básicos de computación, cálculo diferencial, física 1, algebra lineal, investigación.
4	FORMACIÓN BÁSICA DEL PROGRAMADOR	Cálculo integral, metodología de la programación, contabilidad general, física II, proceso investigativo, estadística, programación básica.
5	FORMACIÓN DEL PROGRAMADOR PARA LA CONSTRUCCIÓN DE APLICACIONES ESPECÍFICAS.	Programación avanzada, estructura de datos orientada a objetos, estadística inferencial, contabilidad de costos, electrónica básica, diseño y gestión de base de datos, proceso investigativo.
6	FORMACIÓN DE ANALISTA DE SISTEMAS.	Economía, administración de empresas, arquitectura de computadores, proceso investigativo, lenguaje ensamblador, diseño digital, análisis y diseño de sistemas.
7	DESARROLLO DE SISTEMAS INFORMÁTICOS.	Proceso investigativo, proyectos informáticos I, teoría de telecomunicaciones, derecho informático, diseño de sistemas, ecuaciones diferenciales, sistemas operativos.
8	GESTIÓN DE REDES Y CENTROS DE CÓMPUTO.	Proyectos informáticos II, análisis numérico, administración de centros de cómputo, auditoría informática investigación de operaciones, gestión de redes, proceso investigativo.
9	DESARROLLO DE COMPONENTES Y MODELOS PARA SOFTWARE BASE.	Sistemas de información I, lenguajes formales, modelamiento matemático, ingeniería del software, compiladores, sistemas de información II, proceso investigativo.
10	SISTEMAS INTELIGENTES Y AUTOMATIZADOS.	Inteligencia artificial, control automatizado asistido por computadores, anteproyectos de tesis, sistemas expertos, simulación, ética profesional, proceso investigativo.

Sin embargo en la malla curricular se ha dado variaciones en las unidades de ciertos módulos con el transcurso de los diferentes periodos académicos desde los inicios de la carrera (1999) hasta el año 2013, variaciones que se aprecia en las siguientes tablas:

TABLA XVIII

VARIACIONES DE LAS UNIDADES DEL MÓDULO CUATRO

Periodo Académico	Unidades
Marzo 2005 - Julio 2005	○ Calculo Integral ○ Metodología de la Programación ○ Estructura De Datos I ○ Física II ○ Estadística ○ Programación Básica

Marzo 2006 - Julio 2006	○ Calculo Integral ○ Metodología de la Programación ○ Contabilidad General Electromagnetismo ○ Estadística ○ Programación Básica

TABLA XIX

VARIACIONES DE LAS UNIDADES DEL MÓDULO CINCO

Periodo Académico	Unidades
Septiembre 2005 - Febrero 2006	○ Programación II ○ Estructura de datos II ○ Ecuaciones Diferenciales ○ Contabilidad General ○ Estadística II ○ Teoría De Circuitos
Septiembre 2008 - Febrero 2009	○ Programación Avanzada ○ Estructura de datos Orientada a Objetos ○ Estadística Inferencial ○ Contabilidad de Costos ○ Teoría de los Circuitos ○ Ecuaciones Diferenciales

TABLA XX

VARIACIONES DE LAS UNIDADES DEL MÓDULO SEIS

Periodo Académico	Unidades
Marzo 2004 - Julio 2004	○ Diseño y Gestión de Base de Datos ○ Redes I ○ Contabilidad De Costos ○ Electrónica Básica ○ Diseño Digital ○ Análisis Y Diseño De Sistemas

TABLA XXI

VARIACIONES DE LAS UNIDADES DEL MÓDULO SIETE

Periodo Académico	Unidades
Septiembre 2006 - Febrero 2007	○ Redes II ○ Teoría de Telecomunicaciones ○ Ingeniería del Software ○ Análisis y Diseño II ○ Investigación de Operaciones

	○ Teoría de Autómatas
Septiembre 2008 - Febrero 2009	○ Redes II ○ Teoría de Telecomunicaciones ○ Ingeniería del Software ○ Análisis y Diseño II ○ Inteligencia Artificial ○ Teoría de Autómatas

TABLA XXII

VARIACIONES DE LAS UNIDADES DEL MÓDULO OCHO

Periodo Académico	Unidades
Marzo 2007 - Julio 2007	○ Sistemas Operativos ○ Arquitectura De Computadores ○ Inteligencia Artificial ○ Lenguaje Ensamblador ○ Microprocesadores

TABLA XXIII

VARIACIONES DE LAS UNIDADES DEL MÓDULO NUEVE

Periodo Académico	Unidades
Septiembre 2007 - Febrero 2008	○ Sistemas Expertos ○ Sistemas de Información I ○ Mantenimiento de Computadores ○ Ingeniería del Software II ○ Control automatizado asistido por computadores ○ Administración de Centros de Computo

TABLA XXIV

VARIACIONES DE LAS UNIDADES DEL MÓDULO DIEZ

Periodo Académico	Unidades
Marzo 2008 - Julio 2008	○ Sistemas de Información II ○ Legislación Laboral ○ Presupuestos e Inversiones ○ Proyectos I ○ Auditoria Informática ○ Cad

Cabe mencionar que la malla curricular de la carrera de ingeniería en sistemas desde sus inicios (1999) hasta el periodo septiembre 2008 – febrero 2009 estaba estructurada

por once módulos, detallando a continuación (ver tabla XXV) las unidades que conforman el módulo once:

TABLA XXV

VARIACIONES DE LAS UNIDADES DEL MÓDULO ONCE

Periodo Académico	Unidades
Septiembre 2008 - Febrero 2009	○ Aplicaciones Web ○ Anteproyectos ○ Proyectos Informáticos ○ Administración Bases de datos Sql Server ○ Aplicaciones Mysql y Uml

1.3. Recoger y realizar un análisis de las fuentes de datos que se va a necesitar para realizar la minería de datos.

Las fuentes de datos necesarias para la realización del presente trabajo de titulación provienen de la Universidad de Loja, como son los servidores que contienen datos de los estudiantes como sus records académicos e información personal.

Otra fuente importante ha sido el Web Service, herramienta que permite la consulta y obtención de datos personales, académicos y estadísticos del Sistema de Gestión Académica (creado en el año 2008 para administrar la información académica de los estudiantes Universitarios).

Los servicios web como fuente de información, son de gran utilidad ya que permiten la comunicación entre aplicaciones o componentes de aplicaciones de manera estándar a través de protocolos comunes y de forma independiente al lenguaje de programación, plataforma de implantación, formato de presentación o sistema operativo, debido a que es un contenedor que encapsula funciones específicas y hace que estas funciones puedan ser utilizadas en otros servidores.

Existe una gran cantidad de información que es explotada gracias al Web services para fines institucionales; siendo éstos considerados como una revolución informática de la nueva generación de aplicaciones que trabajan colaborativamente, manteniendo la validez e integridad de la información. La propuesta de utilizar esta tecnología, es tener datos centralizados, siempre y cuando exista mayor flexibilidad a la hora de ser utilizados.

Para acceder a ésta herramienta se contó con la ayuda del departamento de Informática y Telecomunicaciones, en la persona del Ingeniero Milton Palacios, quien después de los trámites pertinentes y la firma respectiva del acuerdo de confidencialidad (ver anexo 1), ha proporcionado un usuario y contraseña (ver anexo 2), para el acceso (ver figura 20) y posterior obtención de los datos necesarios para el desarrollo del presente trabajo.

Figura 20. Pantalla principal del Servidor Web del SGA de la UNL.

Los servicios han sido agrupados en distintas categorías de acuerdo a la información que retornan (ver figura 21) y contienen métodos y parámetros de consultas ya predeterminados de acuerdo a sus funciones para la explotación de datos con mayor rapidez (ver Figura 22):

Figura 21. Página Principal de Usuario Autorizado.

Figura 22. Lista de Categorías de Los Servicios Web.

Las categorías son Académica, Institucional, Personal, Validación y Estadística, de las cuales para el presente proyecto se ha hecho uso de 3 de ellas (ver tabla XXVI):

TABLA XXVI

CATEGORÍAS DEL WEB SERVICE AGRUPADAS DE ACUERDO A LOS SERVICIOS

CATEGORÍA	DESCRIPCIÓN
Académica	En esta categoría se encontrarán los métodos o servicios relacionados a la información académica. Como datos de: • Estudiantes • Docentes
Institucional	En esta categoría se encontrarán los métodos o servicios relacionados con la información institucional. Como datos de: áreas, carreras, módulos, paralelos.
Personal	En esta categoría se encontrarán los métodos o servicios relacionados con información personal de: datos de docentes, datos de estudiantes.

Cada una de estas categorías cuentan con métodos que facilitan la obtención de los datos de manera directa del sistema de gestión académica (ver tabla XXVII):

TABLA XXVII

MÉTODOS UTILIZADOS DE LAS CATEGORÍAS DEL WEB SERVICE

CATEGORÍA	MÉTODOS
Académica	sgaws_periodos_lectivos sgaws_ofertas_academicas sgaws_fechas_matriculasoa sgaws_estadoestudiantes_paralelo sgaws_plan_estudio sgaws_carga_horaria_docente sgaws_notas_estudiante sgaws_carreras_estudiante sgaws_reporte_matricula
Institucional	sgaws_lista_areas sgaws_datos_area sgaws_carreras_area sgaws_datos_carreras sgaws_modulos_carrera sgaws_paralelos_carrera
Personal	sgaws_datos_docente sgaws_datos_estudiante sgaws_datos_usuario

Cada uno de estos métodos devuelven datos específicos (ver tabla XXVIII), lo que facilita la obtención de la información requerida de manera organizada:

TABLA XXVIII

DESCRIPCIÓN DE LOS MÉTODOS DE LAS CATEGORÍAS

Cat.	MÉTODOS	DESCRIPCION
Académica	sgaws_periodos_lectivos	Devuelve una lista de periodos lectivos.
	sgaws_ofertas_academicas	Devuelve una lista de ofertas académicas correspondientes a un periodo lectivo específico.
	sgaws_fechas_matriculasoa	Devuelve una lista de fechas de matrículas de sus diferentes tipos respecto de una oferta académica.
	sgaws_estadoestudiantes_paralelo	Devuelve una lista de estudiantes respecto del paralelo de acuerdo a su estado de matrícula (aprobada, reprobada).
	sgaws_plan_estudio	Devuelve el plan de estudio de una oferta académica o paralelo.
	sgaws_carga_horaria_docente	Devuelve la carga horaria de cada docente en las diferentes ofertas académicas.

	sgaws_notas_estudiante	Devuelve las notas de los estudiantes respecto de su paralelo, módulo, carrera, etc.
	sgaws_carreras_estudiante	Devuelve la lista de matrículas de un estudiante en caso de estar en diferentes carreras.
	sgaws_reporte_matricula	Devuelve un reporte completo de la matrícula de un estudiante, como notas, porcentaje de asistencia, etc.
Institucional	sgaws_lista_areas	Devuelve una lista de las áreas de la institución.
	sgaws_datos_area	Devuelve las siglas del área buscada.
	sgaws_carreras_area	Devuelve las carreras que conforman cada área.
	sgaws_datos_carreras	Devuelve un detalle de cada carrera respecto de una oferta académica.
	sgaws_modulos_carrera	Devuelve los datos de todos los módulos de una carrera específica.
	sgaws_paralelos_carrera	Devuelve los paralelos de cierta oferta académica de una determinada carrera.
Personal	sgaws_datos_docente	Devuelve los datos de uno o varios docentes respecto de su cédula.
	sgaws_datos_estudiante	Devuelve los datos personales del estudiante o los estudiantes respecto de la carrera ingresada.
	sgaws_datos_usuario	Devuelve los datos de cualquier usuario según su tipo estudiante o docente respecto de la cédula ingresada.

1.3.1. Análisis de los datos recopilados

Con la fuente de datos Web Service, se obtuvieron los récords académicos e información personal de los egresados de la carrera de ingeniería en sistemas que estaban registrados en el Sistema de Gestión Académica desde su creación en el año 2008 hasta el 2013. A medida que se realizó el análisis de la información se advirtió que era necesario recopilar los datos históricos anteriores al año de creación del SGA para completar la información académica de todos los egresados registrados a partir del año 2008.

Para ello ha sido necesario recurrir a los libros que contienen los registros académicos de los estudiantes de la Carrera de Ingeniería en Sistemas (ver figura 23), indagando de forma individual el proceso académico reflejado en la malla curricular con sus diferentes

variaciones a través de los periodos académicos, realizando un seguimiento minucioso del desarrollo académico con criterio; descartando los módulos en los que el estudiante haya tenido un estado de reprobado o retirado; dando así la veracidad del caso en la información de los módulos en estado como aprobado y así completar la Base de Datos de 260 estudiantes que consiguieron culminar sus estudios universitarios y egresaron desde el año 2008.

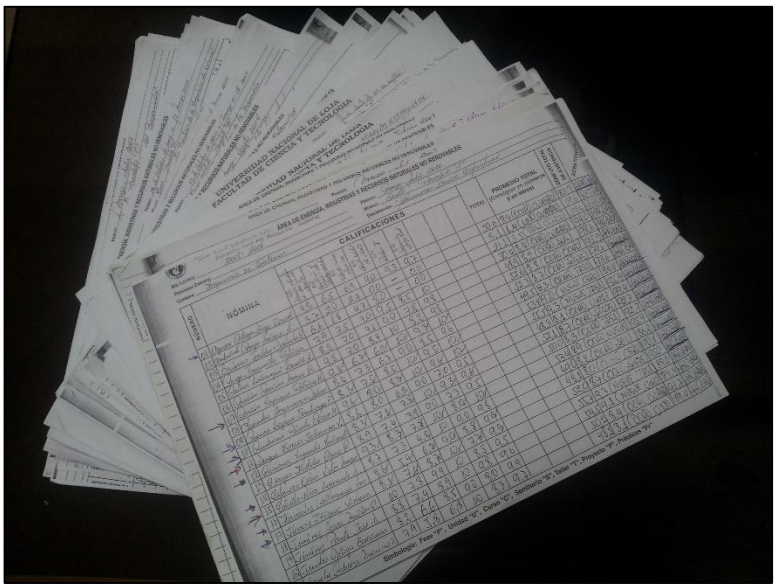
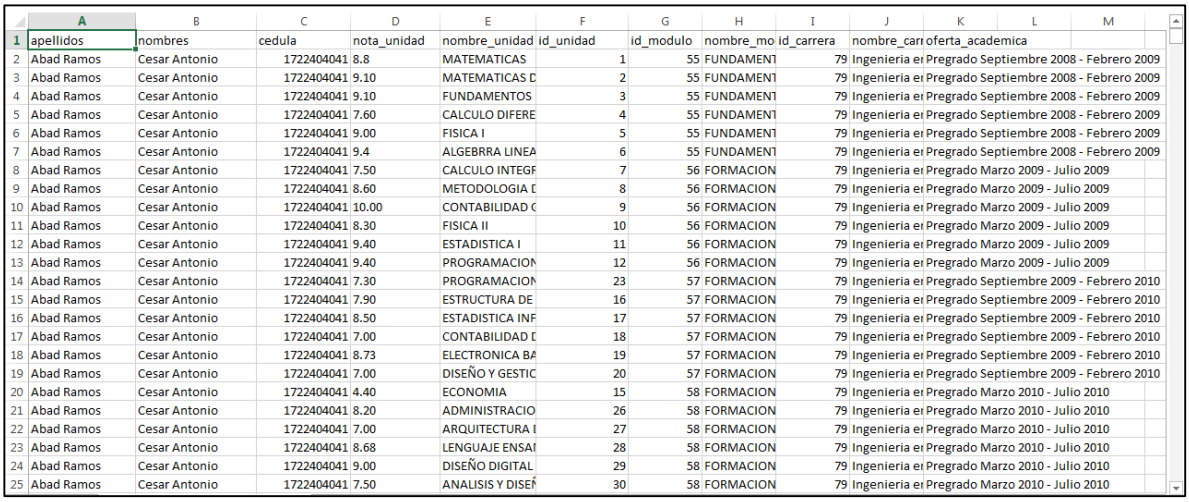


Figura 23. Libros con el registro de los datos históricos.

Ésta tarea permitió tener un conocimiento a profundidad de algunos factores que más adelante serán determinantes con la aplicación de Minería de Datos para contrastar y analizar los resultados en cuanto a los perfiles profesionales determinados. La recopilación de los datos se realizó a partir de copias sacadas de los libros que se encuentran en poder de la secretaría del Área de la Energía, las industrias y los recursos naturales no renovables; con la autorización pertinente (ver anexo 3), siendo trasladados a un archivo en formato Excel (ver Figura 24), para su posterior migración a la base de datos ya establecida de los datos obtenidos del SGA a través del web service, con ello los registros académicos de los egresados desde el año 2008 han sido completados, para su posterior consumo en la minería de datos.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	apellidos	nombres	cedula	nota_unidad	nombre_unidad	id_unidad	id_modulo	nombre_mo	id_carrera	nombre_car	oferta_academica		
2	Abad Ramos	Cesar Antonio	1722404041	8.8	MATEMATICAS	1	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
3	Abad Ramos	Cesar Antonio	1722404041	9.10	MATEMATICAS D	2	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
4	Abad Ramos	Cesar Antonio	1722404041	9.10	FUNDAMENTOS	3	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
5	Abad Ramos	Cesar Antonio	1722404041	7.60	CALCULO DIFERE	4	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
6	Abad Ramos	Cesar Antonio	1722404041	9.00	FISICA I	5	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
7	Abad Ramos	Cesar Antonio	1722404041	9.4	ALGEBRRA LINEA	6	55	FUNDAMENT	79	Ingenieria e	Pregrado Septiembre 2008 - Febrero 2009		
8	Abad Ramos	Cesar Antonio	1722404041	7.50	CALCULO INTEGF	7	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
9	Abad Ramos	Cesar Antonio	1722404041	8.60	METODOLOGIA D	8	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
10	Abad Ramos	Cesar Antonio	1722404041	10.00	CONTABILIDAD (	9	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
11	Abad Ramos	Cesar Antonio	1722404041	8.30	FISICA II	10	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
12	Abad Ramos	Cesar Antonio	1722404041	9.40	ESTADISTICA I	11	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
13	Abad Ramos	Cesar Antonio	1722404041	9.40	PROGRAMACION	12	56	FORMACION	79	Ingenieria e	Pregrado Marzo 2009 - Julio 2009		
14	Abad Ramos	Cesar Antonio	1722404041	7.30	PROGRAMACION	23	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
15	Abad Ramos	Cesar Antonio	1722404041	7.90	ESTRUCTURA DE	16	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
16	Abad Ramos	Cesar Antonio	1722404041	8.50	ESTADISTICA INF	17	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
17	Abad Ramos	Cesar Antonio	1722404041	7.00	CONTABILIDAD D	18	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
18	Abad Ramos	Cesar Antonio	1722404041	8.73	ELECTRONICA BA	19	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
19	Abad Ramos	Cesar Antonio	1722404041	7.00	DISEÑO Y GESTIC	20	57	FORMACION	79	Ingenieria e	Pregrado Septiembre 2009 - Febrero 2010		
20	Abad Ramos	Cesar Antonio	1722404041	4.40	ECONOMIA	15	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		
21	Abad Ramos	Cesar Antonio	1722404041	8.20	ADMINISTRACIO	26	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		
22	Abad Ramos	Cesar Antonio	1722404041	7.00	ARQUITECTURA I	27	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		
23	Abad Ramos	Cesar Antonio	1722404041	8.68	LENGUAJE ENSAI	28	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		
24	Abad Ramos	Cesar Antonio	1722404041	9.00	DISEÑO DIGITAL	29	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		
25	Abad Ramos	Cesar Antonio	1722404041	7.50	ANALISIS Y DISEÑ	30	58	FORMACION	79	Ingenieria e	Pregrado Marzo 2010 - Julio 2010		

Figura 24. Registro de datos históricos en formato Excel.

En base a un análisis de los datos recopilados del SGA de la UNL a partir del año 2008 y los datos anteriores a este año que corresponden a datos históricos, se observó que la información presenta algunos detalles importantes. Los datos corresponden a la categoría académica de los egresados de la carrera de ingeniería en sistemas a partir del año 2008 hasta el año 2013; distribuidos en los siguientes periodos académicos formados por las diferentes ofertas académicas (ver tabla XXIX):

TABLA XXIX

PERIODOS ACADÉMICOS Y OFERTAS ACADÉMICAS DE LOS DATOS RECOPILADOS

Nro.	Periodo Académico	Oferta Académica
1	2003 – 2004	Septiembre 2003 – Febrero 2004
		Marzo 2004 – Julio 2004
2	2004 – 2005	Septiembre 2004 – Febrero 2005
		Marzo 2005 – Julio 2005
3	2005 – 2006	Septiembre 2005 – Febrero 2006
		Marzo 2006 – Julio 2006
4	2006 – 2007	Septiembre 2006 – Febrero 2007
		Marzo 2007 – Julio 2007
5	2007 – 2008	Septiembre 2007 – Febrero 2008
		Marzo 2008 – Julio 2008
6	2008 – 2009	Septiembre 2008 – Febrero 2009
		Marzo 2009 – Julio 2009

7	2009 – 2010	Septiembre 2009 – Febrero 2010
		Marzo 2010 – Julio 2010
8	2010 – 2011	Septiembre 2010 – Febrero 2011
		Marzo 2011 – Julio 2011
9	2011 – 2012	Septiembre 2011 – Febrero 2012
		Marzo 2012 – Julio 2012
10	2012 – 2013	Septiembre 2012 – Febrero 2013
		Marzo 2013 – Julio 2013

Los egresados de la carrera de ingeniería en sistemas en estos diez períodos académicos han culminado sus estudios sin dificultades, otros han interrumpido sus estudios por reprobar uno o varios módulos en el transcurso de su formación y el resto se ha retirado en algún momento de su preparación por motivos desconocidos (ver figura 25).

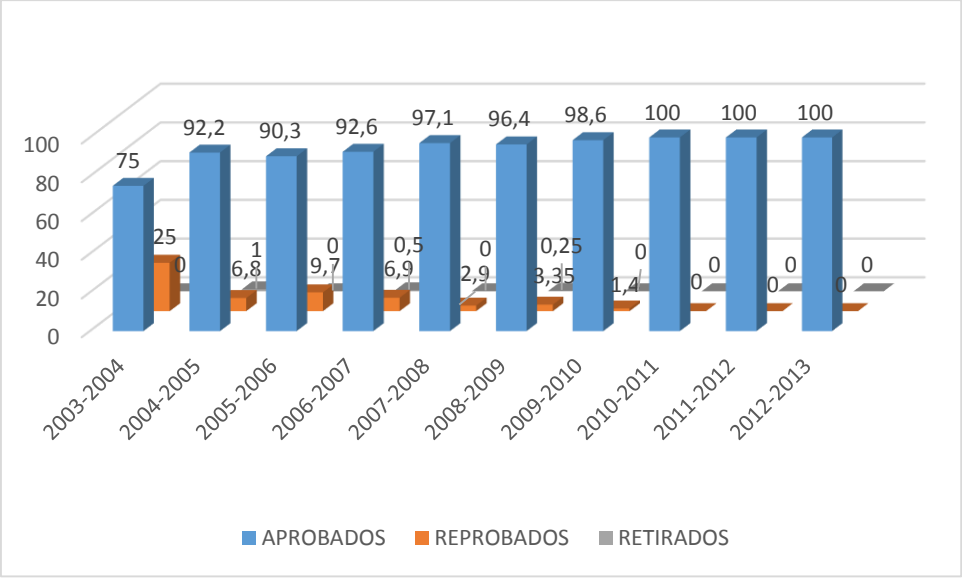


Figura 25. Seguimiento de estudios académicos.

De las estadísticas obtenidas a partir de los datos recopilados y representadas en la Figura 25, las representaremos al mismo tiempo en porcentajes para su mayor comprensión (ver figura 26), donde podemos observar que el 94% de los egresados ha culminado sus estudios sin dificultades, el 6% ha interrumpido sus estudios por haber reprobado uno o alguno de los módulos cursados y prácticamente el 0% se ha retirado

por distintos motivos en algún punto de sus estudios antes de llegar al punto de culminación.

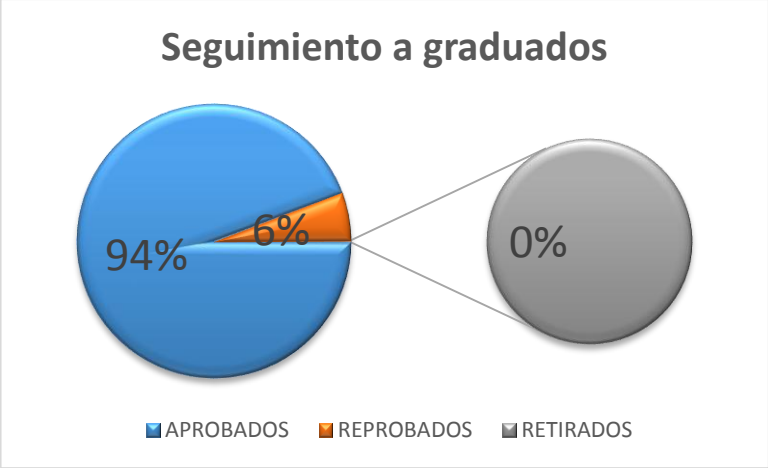


Figura 26. Diagrama de pasteles del seguimiento de estudios académicos.

Cabe recalcar que los datos recopilados corresponden a los estudiantes egresados y graduados de la carrera de ingeniería en sistemas del año 2003 al 2013, en base a estos datos hacemos una exploración observando sus más importantes características:

- **Estudiantes Egresados/graduados:**

El número de graduados varía al número de egresados y va cambiando continuamente (ver Tabla XXX). En el estado actual y de acuerdo al estudio realizado el 55% de los estudiantes son graduados, mientras que el 45% restante corresponde a los egresados (ver figura 27), observando un déficit de graduados, pero a su vez conservando el pensamiento que un egresado ha sido considerado como un profesional, por el hecho de haber culminado sus estudios universitarios dentro de las aulas.

TABLA XXX

EGRESADOS/GRADUADOS

Estudiantes	Nro.
Egresado	117
Graduado	143
TOTAL	260

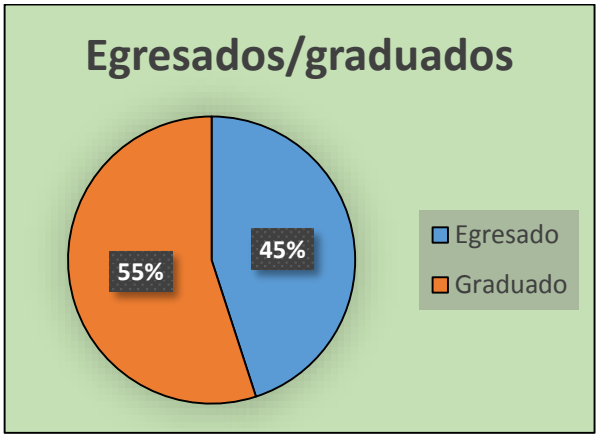


Figura 27. Distribución de egresados y graduados.

- **Género:**

Del total de estudiantes egresados y graduados de la carrera de ingeniería en sistemas, el 53% corresponde al género masculino, mientras que el 47% al género femenino (ver figura 28), por lo que observamos que el género predominante es el masculino con una mínima diferencia.

TABLA XXXI

GENERO EGRESADOS Y GRADUADOS

Genero	Nro.
Masculino	138
Femenino	122
TOTAL	260

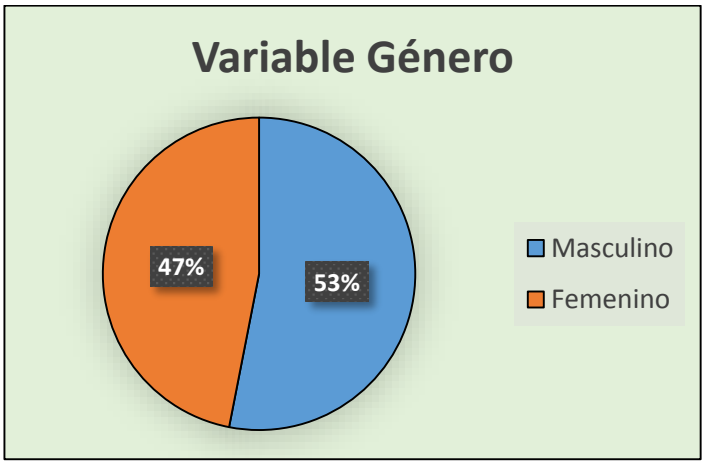


Figura 28. Distribución de egresados y graduados respecto del Género.

- **Recorrido Académico:**

En base a los periodos académicos se ha hecho un análisis del número de estudiantes egresados y graduados han aprobado, reprobado o se han retirado antes de culminar con éxito sus estudios universitarios (ver figura 29). Donde podemos evidenciar que de esta población son muy pocos los estudiantes que han reprobado al menos un módulo y aún menos lo que se han retirado, concluyendo que en su mayoría han culminado sus estudios sin sufrir ningún tropiezo.

TABLA XXXII

RECORRIDO ACADÉMICO

Periodo Académico	Aprobados	Reprobados	Retirados
2003 – 2004	24	8	0
2004 – 2005	81	6	1
2005 – 2006	214	23	0
2006 – 2007	200	15	1
2007 – 2008	371	11	0
2008 – 2009	375	14	1
2009 – 2010	365	6	0
2010 – 2011	350	0	0
2011 – 2012	147	0	0
2012 – 2013	51	0	0
TOTAL	2178	83	3

Figura 29. Gráfica del Recorrido Académico de los Egresados y graduados.

- **Matriculados/Egresados en cada Periodo Académico:**

Realizando una exploración profunda de los datos, en base a los registros físicos y a la información contenida en el Sistema de Gestión Académica, se ha podido determinar los datos estadísticos de acuerdo a los periodos académicos, cuántos de los estudiantes matriculados han egresado o se han graduado.

TABLA XXXIII

MATRICULADOS/EGRESADOS EN CADA PERIODO ACADÉMICO

Periodo Académico	Matriculados	Estudiantes	Egresados/Graduados
2003 – 2004	146	86	60
2004 – 2005	151	103	48
2005 – 2006	115	72	43
2006 – 2007	122	75	47
2007 – 2008	153	110	43
2008 – 2009	270	251	19
TOTAL	957	697	260

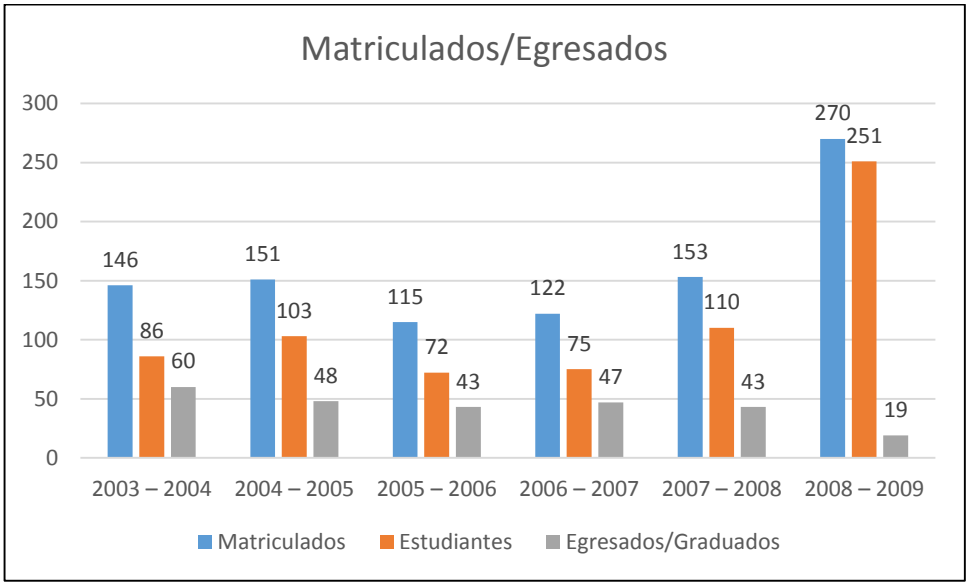


Figura 30. Descripción gráfica de los matriculados egresados o graduados.

- **Matriculados/Egresados (2003-2008):**

Realizando un cálculo final de los estudiantes que se han matriculado y han culminado sus estudios con éxito, podemos observar que un 27% corresponde a los egresados o graduados mientras que un 73% se han matriculado pero han reprobado o se han retirado en el camino (ver figura 31), observando que de los matriculados son extremadamente una cantidad menor que ha terminado sus estudios con éxito.

TABLA XXXIV

MATRICULADOS/EGRESADOS 2003 AL 2008

Categoría	Nro.
Estudiantes	697
Egresados/Graduados	260
TOTAL MATRICULADOS	957

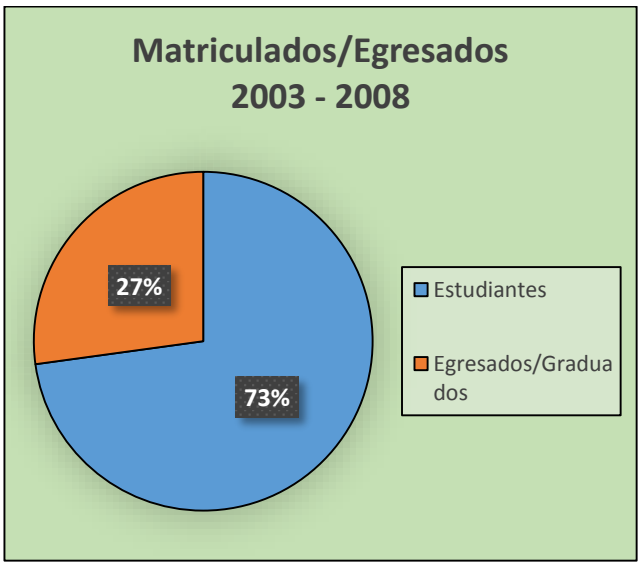


Figura 31. Matriculados/Egresados 2003-2008.

- **Proyecto de Seguimiento a graduados**

De la información recopilada a través del web services pertenece a los egresados de la carrera de ingeniería en sistemas, desde la creación del SGA en la UNL. Entre la masa de los egresados se encuentran los graduados que han culminado y obtenido su título académico por el desarrollo y disertación de sus proyectos de fin de carrera. El proyecto de seguimiento a graduados como su nombre lo dice intenta mantener la información actualizada de los profesionales de la carrera, información muy relevante para la validación de los resultados obtenidos a través de la minería de datos. Se ha recopilado del proyecto de seguimiento a graduados, información en archivos digitales proporcionados por el ing. René Guamán Quinche encargado del proyecto de seguimiento a graduados. Estos datos nos revelan estadísticas respecto a la cantidad de estudiantes egresados que han obtenido su título. Además estos datos han servido para contactarlos y recopilar nuevos datos importantes.

- **Perfil de egreso de la carrera de ingeniería en sistemas**

El Ingeniero en Sistemas, está capacitado para desenvolverse profesionalmente con pensamiento crítico, reflexivo, analítico y estratégico, pero sin dejar de lado la formación humanista con sensibilidad a las necesidades de su entorno social en un ambiente competitivo, interdisciplinario y globalizado, con competencia para generar soluciones técnicas a problemas reales, aplicando sus conocimientos en las áreas del

Hardware, Software, Análisis de Sistemas e Ingeniería del Software, Telecomunicaciones y Organización.

Además de su formación académica, posee actitudes como la disciplina, responsabilidad y comportamiento ético en su ejercicio profesional, así como la capacidad humana de solidaridad y compromiso social. Al incorporarse a la actividad dentro de su campo profesional, estará en condiciones para desarrollar el siguiente perfil profesional [61]:

● Poseer conocimientos en las Ciencias Básicas, Sociales y humanísticas.

● Desenvolverse con solvencia técnica y actitudes suficientes para planificar, organizar, ejecutar, controlar y evaluar las actividades inherentes al campo de la Informática y la Computación, con una clara visión de la realidad, vinculando la teoría con la práctica y con interés permanente por la innovación y la investigación en los diversos campos de su profesión.

● Capacidad para identificar, definir y analizar problemas de procesamiento de datos y generación de sistemas de información así como para interactuar interdisciplinariamente en la implementación de soluciones técnicas y económicamente ventajosas para resolver problemas relacionados a su campo profesional.

● Trabajar y asesorar en el uso de herramientas y técnicas en el análisis, diseño, gestión y evaluación de Soluciones Informáticas incluyendo el hardware, software, redes y telecomunicaciones que sirvan de manera eficaz y eficiente para resolver las necesidades operacionales y de gestión de la organización.

● Dirigir, trabajar y/o asesorar actividades de planificación, ejecución, control y evaluación de: Infraestructura Tecnológica, Seguridad de Sistemas de Información, Diseño e implementación de Redes de Datos, Software de aplicación y Sistemas de Telecomunicación.

● Desempeñarse con solvencia en lo técnico-administrativo y económico-financiero dentro del campo inherente a la Informática y computación.

● Realizar auditorías técnicas de Hardware, Software, Redes y Telecomunicaciones.

● Elaborar, gestionar y evaluar proyectos de investigación e inversión de sistemas Informáticos y Computación.

● Proceder con conocimiento de causa en actividades relacionadas con legislación laboral, contratación pública y propiedad intelectual, dentro del campo y ámbito de la Informática y computación.

● Generar y presentar soluciones eficientes, eficaces e innovadoras que ayuden a la toma de decisiones en la empresa u organización.

● Mantener una actitud autodidacta e investigativa frente a las nuevas tecnologías de la informática y las telecomunicaciones.

El perfil de egreso del ingeniero en sistemas ha sido utilizado para realizar una encuesta que determine el perfil profesional de cada estudiante para realizar posteriormente la predicción de los perfiles en el proceso de minería de datos. A continuación se describe esta encuesta:

1.3.2. Obtención del perfil profesional aplicando un test de habilidades, capacidades e intereses, útil para el proceso de minería de datos.

El test aplicado ha sido construido en base framework de desarrollo web de código abierto django, y el lenguaje de programación python, tomando como referencia el Sistema de Evaluación del Desempeño Docente de la Institución que está desarrollado en base al mismo framework y que ha sido publicado en la nube como código abierto [103].

El objetivo este test desarrollado (ver anexo 4) es con el fin de obtener el perfil profesional de los egresados y graduados de la carrera de ingeniería en sistemas. Los perfiles planteados se los ha obtenido realizando una consulta bibliográfica de los perfiles con las respectivas características, habilidades, capacidades que puede tener un estudiante al egresar de la carrera de ingeniería en constancia en el documento del

rediseño de la carrera de ingeniería en sistemas de la UNL [60] y en la documentación de algunas universidades del país y del mundo [61-68]. A su vez se realizó una entrevista al ing. del Banco de Loja quien ha pasado por varios departamentos relacionados con la carrera de ingeniería en sistemas, por lo cual conoce las características, habilidades y capacidades que debería cumplir una persona para tener cierto perfil profesional (ver anexo 5). Finalmente para realizar un filtro de estos perfiles se ha analizado las unidades de la malla curricular base y sus cambios a través de los diferentes periodos ya detallada anteriormente, observando si de acuerdo a la estructura de las unidades se podría enfocar a un determinado perfil. Por lo tanto los perfiles profesionales escogidos son 8 nombrados a continuación:

- Analista de Sistemas de Información
- Arquitecto y Diseñador de Software.
- Desarrollador de software
- Administrador de Sistemas de Bases de Datos
- Auditor Informático
- Administrador de Centros de computo
- Administrador de Redes computacionales.
- Especialista en mantenimiento hardware y software.

De cada perfil o especialidad se ha determinado 5 características propias y únicas, que formarán los ítems de selección del test planteado. A continuación se da un concepto de cada perfil y se nombra las características correspondientes:

1.3.2.1. Analista de Sistemas de Información

El analista es un solucionador de problemas, aborda como un reto el análisis de problemas y disfruta al diseñar soluciones factibles. Cuenta con la capacidad de afrontar sistemáticamente cualquier situación mediante la correcta aplicación de herramientas, técnicas y su experiencia. Debe ser un comunicador con capacidad para relacionarse con los demás durante extensos periodos [69-71]. A continuación se presenta en la tabla XXXV las características de este perfil:

TABLA XXXV

CARACTERÍSTICAS DEL PERFIL ANALÍSTA DE SISTEMAS

Característica	Capacidad	Habilidad	Interés
Enfrentar los problemas de sistemas de información de manera específica, enfocándose en las necesidades del negocio.		X	
Determinación de los requerimientos de información, y análisis de las necesidades del Sistema.	X		
Piensa que tiene la facilidad para integrar cosas e interactuar con las personas y por ello promover un cambio que involucre el uso de los sistemas de información.		X	
Se considera una persona auto disciplinado y auto motivado, con la capacidad de administrar y coordinar los sistemas de información, incluyendo a otras personas.	X		
Piensa que tiene la habilidad de trabajar y asesorar con el uso de herramientas y técnicas en el análisis eficiente para resolver las necesidades operacionales y de gestión de la organización.	X		

1.3.2.2. Arquitecto y Diseñador de Software

Un arquitecto debe entender el propósito del software, debe ser capaz de ver todos sus usos, y los problemas que conllevan diseñarlo para que este sea funcional. Define la configuración de los componentes de las aplicaciones de acuerdo a la estructura del problema planteado, los requerimientos funcionales, los no-funcionales y las necesidades de negocios de la organización [72, 73]. A continuación se presenta la tabla con las características de este perfil (ver tabla XXXVI).

TABLA XXXVI

CARACTERÍSTICAS DEL PERFIL ARQUITECTO Y DISEÑADOR DE SOFTWARE

Característica	Capacidad	Habilidad	Interés
Se preocupa por entender el propósito del software, ver todos sus usos, y los problemas que conllevan diseñarlo para que este sea funcional.		X	
Investiga nuevas tecnologías y comprende Frameworks arquitectónicos y las mejores prácticas.	X		
Tiene la capacidad de definir y documentar la solución, asegurándose que esté acorde con el sistema deseado y además sea la correcta para su soporte y evaluación.	X		
Sólidos conocimientos de arquitectura de software y aplicaciones N-Capas así como en bases de datos relacionales.	X		
Se le facilita seguir los requerimientos del software especificados, buscar algo nuevo e innovador, dar los lineamientos del software en base a las soluciones del analista.		X	

1.3.2.3. Desarrollador de software

Es la persona capacitada para realizar programas o componentes de sistemas de computación, interpretar especificaciones de diseño, documentar productos realizados, verificar los componentes programados, buscar causas de mal funcionamiento y corregir los programas o adaptarlos a cambios en las especificaciones [74, 75]. A continuación se presenta la tabla con las características de este perfil (ver tabla XXXVII).

TABLA XXXVII

CARACTERÍSTICAS DEL PERFIL DESARROLLADOR DE SOFTWARE

Característica	Capacidad	Habilidad	Interés
Tiene la capacidad de interpretar especificaciones de diseño o requisitos de las asignaciones a programar.	X		
Comprende con facilidad la definición o instanciación de clases, escritura de algoritmos, estructuración de datos necesarios, o la incorporación eventual de componentes obtenidos de otros programas.	X		
Facilidad para documentar productos realizados, verificar los componentes programados, buscar causas de mal funcionamiento y corregir los programas o adaptarlos a cambios en las especificaciones.		X	
Facilidad para trabajar en equipo y bajo presión, tomando en cuenta el uso eficiente de recursos y el ambiente de desarrollo, aportando con propuestas de cambios tendientes a mejorar la calidad, mantenibilidad y eficiencia del software.		X	
Capacidad de resolver un problema de manera automatizada, evaluar la funcionalidad del software a través de entornos de pruebas, depuración de código, etc. Observar a detalle lo que va realizando.	X		

1.3.2.4. Administrador de Sistemas de Bases de Datos

Es el individuo que está a cargo del rendimiento de la base de datos, de la retención y la seguridad. Ellos asisten a los procesos de desarrollo necesarios para un rendimiento óptimo de la base de datos. Deben identificar las señales de posibles fallas en el sistema y otros desastres que pudieran generar la pérdida información. Si evitar una falla o un desastre está fuera de su alcance, tienen que estar preparados para recuperarlos. También tienen que mantener los registros de usuarios y contraseñas para asegurar la seguridad de los datos en la base de datos [76]. A continuación se presenta la tabla con las características de este perfil (ver tabla XXXVIII).

TABLA XXXVIII

CARACTERÍSTICAS DEL PERFIL ADMINISTRADOR DE SISTEMAS DE BASE DE DATOS

Característica	Capacidad	Habilidad	Interés
Está en la capacidad de diseñar, probar e implementar bases de datos, tomando en cuenta la escalabilidad de las bases de datos, independencia lógica y física de los datos, redundancia mínima, integridad de los datos, respaldo y recuperación.	X		
Le sería fácil entrenar a otros en la organización para que tengan acceso y usen de la información localizada en la base de datos de manera correcta.		X	
Le llama la atención el control masivo de datos, la integración del desarrollo para hacer funcional las bases de datos, el uso de complejos grupos de datos y modelos.			X
Posee habilidades para mantener, modificar y actualizar los paquetes de software de bases de datos y desarrollar pruebas de aceptación en nuevos sistemas y software.		X	
Capacidad para identificar, definir y analizar problemas de procesamiento de datos, salvaguardando las bases de datos mediante el análisis, control y evaluación.	X		

1.3.2.5. Auditor Informático

Es la persona encargada de la verificación y certificación de la informática dentro de las organizaciones. Debe estar en la capacidad de recoger, agrupar y evaluar evidencias para determinar si un sistema de información salvaguarda el activo empresarial, mantiene la integridad de los datos, lleva a cabo eficazmente los fines de la organización y utiliza eficientemente los recursos [77-79]. A continuación se presenta la tabla con las características de este perfil (ver tabla XXXIX).

TABLA XXXIX

CARACTERÍSTICAS DEL PERFIL AUDITOR INFORMÁTICO

Característica	Capacidad	Habilidad	Interés
Capacidad de gestión de seguridad de sistemas y planes de contingencia, gestión de problemas y cambios en entornos informáticos, gestión del departamento de sistemas y operaciones y planificación informática.	X		
Conocimiento y práctica de las normas estándares para la auditoría interna.	X		
Habilidad para el manejo de herramientas de control y verificación de la seguridad, herramientas de monitoreo de actividades y simuladores o generadores de datos, en base a la aplicación de técnicas de evaluación de riesgos, muestreo, etc.		X	
Generar y presentar soluciones eficientes, eficaces e innovadoras que ayuden a la toma de decisiones en la empresa u organización, en base a auditorías técnicas de		X	

Hardware, Software, Redes y Telecomunicaciones que realice.			
Es curioso, le gusta estar vigilante a que los procesos y las personas cumplen con las políticas de la organización, estando a la expectativa que se realicen cambios sin control.			X

1.3.2.6. Administrador de Centros de cómputo

Un administrador de centros de cómputo es aquella persona dentro de la empresa que soluciona problemas, mide recursos, planea su aplicación, desarrolla estrategias, efectúa diagnósticos de situaciones exclusivos de la organización a la que pertenece. El administrador sumado a sus conocimientos académicos debe tener ciertas características de personalidad, de conocimiento tecnológico de administración [80, 81]. A continuación se presenta la tabla con las características de este perfil (ver tabla XL).

TABLA XL

CARACTERÍSTICAS DEL PERFIL ADMINISTRADOR DE CENTROS DE CÓMPUTO

Característica	Capacidad	Habilidad	Interés
Administrar los servicios de la red y la comunicación electrónica en un espacio físico determinado, estando en constante actualización de los avances tecnológicos y proveedores de equipamiento que existen.		X	
Es un negociador, ya que posee una fuerte visión para los negocios. Supervisa y dirige a la vez.		X	
Capacidad para interactuar interdisciplinariamente en la implementación de soluciones técnicas y económicamente ventajosas para resolver problemas relacionados a su campo profesional.	X		
Dirigir, trabajar y/o asesorar actividades de planificación, ejecución, control y evaluación de la infraestructura Tecnológica en el diseño e implementación de un centro de cómputo.		X	
Interés por el liderazgo, la gestión, la planificación y el trabajo en equipo respecto a la parte operativa de la organización, es decir en el ámbito de infraestructura tecnológico.			X

1.3.2.7. Administrador de Redes computacionales

Es la persona con perfil de un técnico de nivel superior del área de las tecnologías, con un dominio adecuado de los conocimientos y técnicas capaz de instalar, configurar y administrar redes de datos y voz de mediana envergadura, permitiéndole utilizar de manera efectiva las herramientas de comunicación entre departamento o instituciones. Asimismo, cumple labores de soporte de infraestructura de redes aplicando técnicas de

cableado estructurado y de certificación de redes, y da soporte a los sistemas computacionales de los usuarios de una organización [82, 83]. A continuación se presenta la tabla con las características de este perfil (ver tabla XLI).

TABLA XLI
CARACTERÍSTICAS DEL PERFIL ADMINISTRADOR DE REDES COMPUTACIONALES

Característica	Capacidad	Habilidad	Interés
Conocimientos y aplicación de técnicas para instalar, configurar y administrar redes de datos y voz de mediana envergadura, permitiendo utilizar de manera efectiva las herramientas de comunicación.	X		
Ofrecer servicios como: soporte en dispositivos de red, sistemas computacionales, monitoreo, administración y seguridad de redes.	X		
Se desempeña como un agente acelerador de las transformaciones en materias informáticas orientadas a redes computacionales, comprometiéndose con su propio aprendizaje, a través de la formación continua.		X	
Diseño, gestión y evaluación de Soluciones Informáticas respecto a redes y telecomunicaciones que sirvan de manera eficaz.	X		
Capacidad de análisis de la infraestructura necesaria para la seguridad de redes, cableado, comunicaciones, controlando la parte funcional, tomando en cuenta estándares y la escalabilidad de la Red en caso de cambios a futuro.	X		

1.3.2.8. Especialista en mantenimiento hardware y software

Un especialista en mantenimiento de hardware y software es una persona con la capacidad de modificar un producto después de la entrega para corregir errores, mejorar el rendimiento y otros atributos. El mantenimiento implica la reparación tanto física como de software de su computador en el momento que se presenta el problema [84]. A continuación se presenta la tabla con las características de este perfil (ver tabla XLII).

TABLA XLII
CARACTERÍSTICAS DEL PERFIL ESPECIALISTA EN MANTENIMIENTO HARDWARE Y SOFTWARE

Característica	Capacidad	Habilidad	Interés
Capacidad de modificar un producto después de la entrega para corregir errores, mejorar el rendimiento, u otros atributos, abarcando el ensamblado y la reparación tanto física y lógica de su computador en el momento que se presenta el problema.	X		

Capacidad de brindar soporte técnico, con destrezas para instalar, configurar, operar y dar mantenimiento, a nivel de sistemas operativos, paquetes de software, electrónica y principalmente en hardware.	X		
Capacidad de realizar un seguimiento y control continuo de los recursos informáticos existentes en la empresa para identificar posibles averías y proceder a su reparación.	X		
Saber y aplicar las técnicas específicas para la localización de las causas que provocan disfunciones en el rendimiento de los equipos y sistemas informáticos, y detectar posibilidades de mejora y actuaciones para evitar dichas disfunciones.		X	
Interés por las cosas prácticas, la toma de decisiones rápidas para la solución de problemas a nivel de hardware y sistemas operativos de los equipos, que permitan el correcto funcionamiento de los sistemas que maneja la organización.			X

El test fue implantado en el servidor de la Universidad Nacional de Loja (http://192.188.49.12:8080/), después de los trámites, autorización respectiva (ver anexo 6) y un sin número de configuraciones necesarias (ver anexo 7). La difusión del test se la enfocó al 100 % de la población teniendo una respuesta positiva del 80% de los egresados y graduados de la carrera de ingeniería en sistemas de La UNL (2003-2013), concluyendo que la difusión ha tenido una excelente acogida.

El propósito de aplicar este test fue el de obtener el perfil profesional de los estudiantes en base a sus habilidades, capacidades e intereses, siendo una de las fuentes de datos más importante para realizar posteriormente el proceso de minería de datos.

2. ETAPA DOS: Comparar y seleccionar la técnica de minería de datos y herramientas de acuerdo al ambiente de estudio.

2.1. Recopilación de información, evaluación y selección de las herramientas disponibles para realizar el proceso de minería de datos.

- **Herramientas de gestión de Bases de Datos.**

Es muy alta la oferta de herramientas de gestión de bases de datos desarrolladas con una licencia de software libre o pagado hoy en día; la selección de la herramienta adecuada depende de las necesidades, el presupuesto y otros factores para que una se diferencie de otra. En el presente trabajo de titulación ha sido necesario el uso de una herramienta de gestión de base de datos para la preparación, limpieza y generación

de estructuras de los datos, para esta importante selección se tomó en cuenta las herramientas de licencia libre y se analizó sus características más fuertes como las interfaces en las que trabaja, los sistemas operativos para los cuales son compatibles, los formatos de exportación y exportación que ofrecen, etc (ver tabla XLIII [85-89]).

TABLA XLIII

CARACTERÍSTICAS DE HERRAMIENTAS DE GESTIÓN DE BASES DE DATOS

Nombre	Formatos de Exportación /Importación	Bases de datos	Licencia	Sistema operativo		
				Windows	Linux	MacOS
SQLyog Enterprise	SQL, CSV,HTML, XML	MYSQL	Trial	X	X	
PhpMyAdmin	SQL, CSV, TXT, XLS,XML	MYSQL	Free	X	X	X
MySQL Workbench	SQL	MYSQL	GPL	X	X	X
DatAdmin	SQL, ZIP, CSV, XML, DBF, TXT, XLS, HTML	Oracle, MySQL, Postgre SQL, Interbase, Firebird, MSSQL y SQLite	Personal	X	X	

Haciendo un análisis estadístico de las características de las herramientas recopiladas anteriormente en la tabla XLIII, podemos observar algunas diferencias importantes. La tabla XLIV representa el número de formatos con los que se puede exportar o importar en las herramientas.

TABLA XLIV

NÚMERO DE FORMATOS EXPORTACIÓN/IMPORTACIÓN DE CADA HERRAMIENTA

Herramienta	Formatos Exportación/Importación
SQLyog	4
PhpMyAdmin	5
Workbench	1
DatAdmin	8

La figura 32 representa en un diagrama de pastel los porcentajes de formatos de exportación e importación de cada herramienta, en donde observamos que la herramienta DatAdmin es la predominante en esta característica con un 44%, a la que le sigue PhpMyAdmin con un 28%, SQLyog con un 22% y finalmente con un porcentaje sumamente bajo workbench con un 6%.

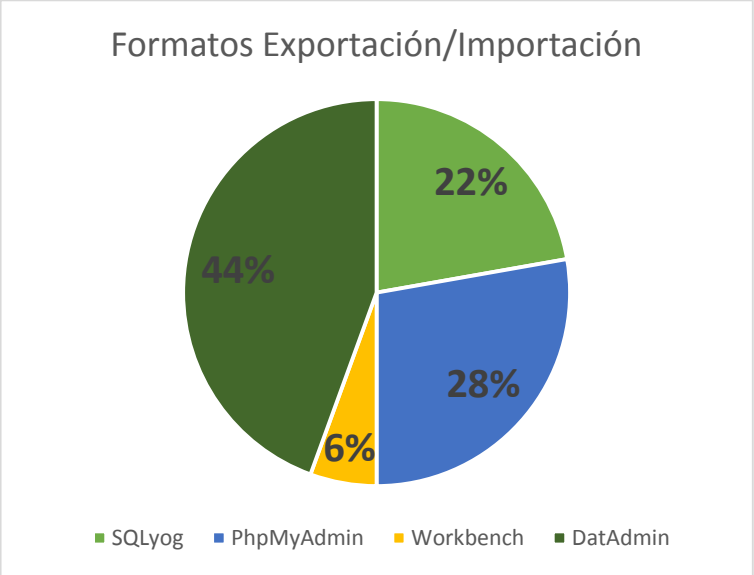


Figura 32. Formatos Exportación/Importación de las herramientas para GBD.

Otra de las características analizadas para la correcta selección de la herramienta de gestión de base de datos corresponde al número de gestores de base de datos que la herramienta está en capacidad de dar soporte (ver tabla XLV).

TABLA XLV

NÚMERO DE SGBD QUE DA SOPORTE CADA HERRAMIENTA

Herramienta	SGBD
SQLyog	1
PhpMyAdmin	1
Workbench	1
DatAdmin	7

La figura 33 muestra de manera gráfica la representación del número de Sistemas Gestores de Base de Datos que cada herramienta brinda soporte, donde observamos que las herramientas SQLyog, PhpMyAdmin y Workbench están empatadas con el valor de uno, debido a que dan soporte únicamente al SGBD MySQL, mientras que la

herramienta DatAdmin claramente tiene una mayor ventaja, por su capacidad de dar soporte a siete SGBD que corresponden a Oracle, Postgre SQL, Interbase, Firebird, MSSQL, SQLite y MySQL como las otras herramientas. Por lo que llegamos a la conclusión que en esta característica claramente DatAdmin está en el primer lugar.

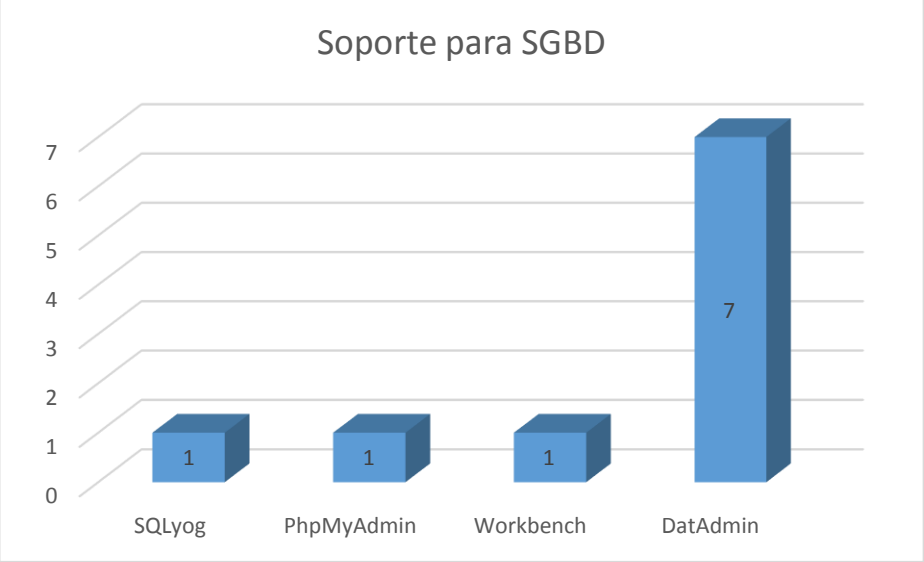


Figura 33. Soporte para SGBD de cada herramienta.

Finalmente la característica analizada entre las herramientas para la gestión de base de datos corresponde a la compatibilidad con los sistemas operativos más utilizados en la actualidad, que son Windows, Linux y MacOS (ver tabla XLVI).

TABLA XLVI

SISTEMAS OPERATIVOS COMPATIBLES CON CADA HERRAMIENTA

Herramienta	# Sistemas Operativos compatibles
SQLyog	2
PhpMyAdmin	3
Workbench	3
DatAdmin	2

La figura 34 nos muestra a través de un gráfico de barras el número de sistemas operativos para los cuales las herramientas son compatibles. Donde podemos observar que existe un empate entre SQLyog y DatAdmin con 2 sistemas operativos y PhpMyadmin vs workbench con 3, siendo los predominantes en esta característica.

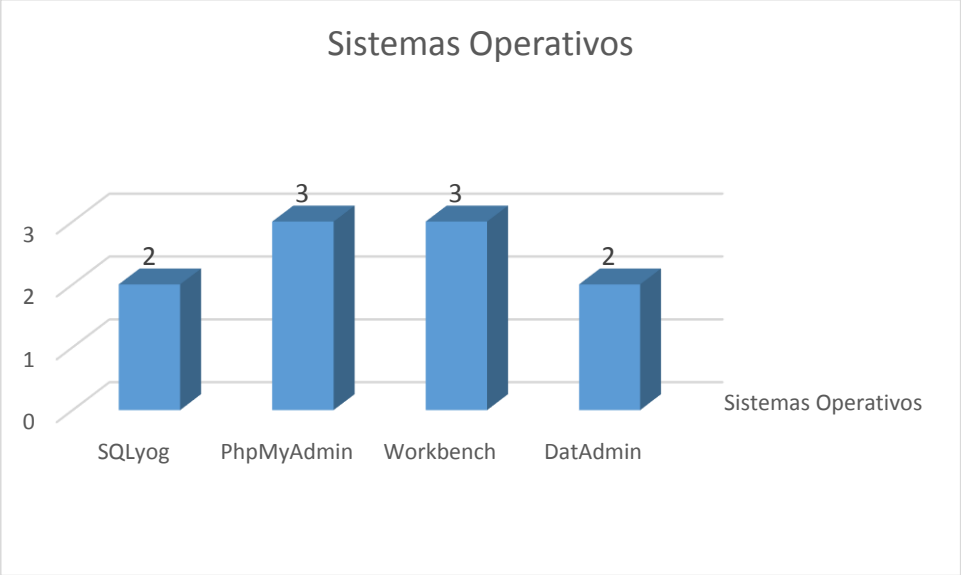


Figura 34. Compatibilidad con Sistemas Operativos de cada herramienta.

En conclusión se han tomado en cuenta tres características principales para la selección de la herramienta de administración o gestión de bases de datos las cuales son: capacidad de exportación e importación de archivos de base de datos en distintos tipos de formatos, soporte a sistemas de gestión de base de datos y compatibilidad con los sistemas operativos más utilizados en la actualidad Windows, Linux y macOS. La herramienta que ha predominado con una gran ventaja en dos de tres de las características analizadas ha sido DatAdmin siendo la herramienta seleccionada para el desarrollo del presente trabajo de titulación.

- **Herramientas enfocadas al proceso de Minería de datos**

Existen diferentes herramientas que dan apoyo al proceso de minería de datos, toda la información recopilada de las principales herramientas gratuitas para aplicar el proceso de minería de datos se encuentra en la revisión literaria, CAPÍTULO III. HERRAMIENTAS PARA EL PROCESO DE MINERÍA DE DATOS.

La información recopilada respecto de las herramientas para la minería de datos ha sido respecto a su uso, sus características y a los objetivos que se pretende alcanzar. En la tabla XLVII [90-92] se detalla estos y otros aspectos importantes de manera comparativa permitiendo diferenciar las herramientas y determinar la más adecuada para el análisis de minería de datos.

TABLA XLVII

CARACTERÍSTICAS DE LAS HERRAMIENTAS ENFOCADAS AL PROCESO DE MD

SOFTWARE	Modo de manipulación				Integración con Herra-mientas	Licencia	Limitación
	IGU	Batch (Lotes)	Línea de comandos	Creando Aplicación			
ORANGE	X		X		X	Open Source	Los componentes son manipulados solamente desde programas desarrollados en Python.
WEKA	X		X	X	X	GNU	Problemas de manejo de memoria. Es muy lento en grandes conjuntos de datos.
KNIME	X				X	GPLv3	Weka proporciona un mayor número de componentes que esta herramienta.
RAPIDMINER	X	X	X	X	X	GNU GPL	Algunos algoritmos derivados de la integración con weka no funcionan bien en su totalidad, ejm: NNge.

Realizando un análisis de la tabla XLVII, observamos que las herramientas que predominan en cuanto a la cantidad de formas de manipulación de las funciones son la herramienta Rapidminer seguida del Software Weka. En cuando a la integración con herramientas todas ofrecen esta posibilidad, mientras que prestando atención a la limitación principal de cada herramienta todas presentan una desventaja desde diferentes perspectivas.

- **Comportamiento de las herramientas y evaluación con datos de Prueba.**

La comparación de las características herramientas en base a su manipulación (ver anexo 8) y la evaluación de las mismas con datos de prueba (ver anexo 9), son otros de los criterios tomados en cuenta en la selección de la herramienta adecuada para el proceso de minería de datos. Las pruebas han sido efectuadas con la base de datos denominada Golf. Esta base de datos contiene variables como el estado del clima, temperatura, humedad, presencia o ausencia de viento, que de acuerdo a sus valores reflejan en las condiciones que debe existir para decidir salir o no a jugar el Golf.

Para esta evaluación se han tomado en cuenta tres de las herramientas analizadas, RapidMiner y weka que tienen la ventaja en comparación con las otras herramientas en cuanto a la cantidad de modos de manipular los datos, y la herramienta KNIME cuya limitación es de menor importancia que las demás herramientas analizadas. Este proceso se ha realizado con el fin de observar en un entorno práctico el comportamiento de cada herramienta y hacer una comparación entre sus características, esta evaluación ha sido detallada en la tabla XLVIII.

TABLA XLVIII

COMPARACION Y EVALUACION DE LAS CARACTERISTICAS DE LAS HERRAMIENTAS CON DATOS DE PRUEBA

Herramienta	Manipulación GUI	Presentación de los datos	Visualización de Resultados	Manejo y presentación de procesos.	Valoración Final [1-10]
KNIME	Interfaz poco amigable al usuario, confusión entre los elementos, desorganización.	Los datos se los puede visualizar a través de una sencilla tabla en una pantalla externa, donde no presenta muchos detalles.	Grafos y detalladamente de acuerdo al algoritmo aplicado.	Interacción entre nodos de manera gráfica.	6
WEKA	Interfaz simple, componentes mezclados en diferentes pestañas.	Permite observar los datos inicialmente cargados, en diagramas de barras, representando cada una de las variables, para una mejor interpretación.	Matrices, Texto plano, grafos.	Cuenta tres entornos (Explorer, Experimenter, Simple CLI) para clasificar los datos, y además la interacción entre componentes en el entorno KnowledgeFlow.	8
RAPIDMINER	Presenta una interfaz amigable, intuitiva para el usuario, con paneles muy bien organizados.	Los datos se presentan en múltiples formas: - En una tabla detalladamente, con opciones de visualización de acuerdo a las características de los datos. - A través de gráficas. - Tabla de detalles de Tipo de datos, estadísticas, rango, etc.	Matrices, texto plano, grafos con opción a múltiples configuraciones y un sin número formas de visualización dependiendo de los operadores y algoritmos que se apliquen.	Procesos compuestos de operadores que interactúan entre sí, estos operadores se arrastran a la zona de trabajo, con opción de modificar sus parámetros de manera individual. La descripción de los procesos se la puede observar en formato XML.	9.5

Finalmente en base a los aspectos analizados de cada una de las herramientas, y de acuerdo a la comparación de las mismas en base a sus diferentes características y comportamiento, se ha llegado a la conclusión de que las herramientas más predominantes son RapidMiner y weka. A su vez se ha tomado en cuenta que la herramienta weka se puede asociar a la herramienta Rapidminer a través de uno de sus módulos. Por lo tanto la herramienta seleccionada para el desarrollo del proceso de minería de datos en el presente trabajo de titulación es Rapidminer por ser considerada como la más adecuada.

Como acotación a este proceso de selección podemos mencionar que la cualidad más importante en la herramienta escogida es su sencillez, su aplicación, reduciendo además los costos de implantación en un equipo de desarrollo debido a ser libre, sus módulos de integración con otras herramientas que permiten realizar también procesos de minería de datos como lo son R y Weka, su capacidad para solventar con éxito problemas de minería de datos, las cuatro formas distintas de utilización que ofrece: a través de una interfaz gráfica, en línea de comandos, creación de una aplicación y en batch (lotes). Es por esto y todos los criterios tomados en cuenta anteriormente que se justifica como conveniente su elección.

2.2. Hacer un análisis comparativo de las técnicas de minería de datos que se acoplen al problema de investigación planteado.

Las técnicas de minería de datos representan la aplicación automatizada de algoritmos de minería de datos permiten la detección de fácil de patrones en los datos. Se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o de descubrimiento del conocimiento. Estas técnicas se encuentran en una constante evolución al mezclar algunos campos como: inteligencia artificial, sistemas expertos, bases de datos, estadística, reconocimiento de patrones, lógica, etc.

Los resultados varían de acuerdo a las técnicas aplicadas (ver tabla XVI), no se puede asegurar que una técnica tiene ventajas sobre otra, o que exista una única técnica de solución al problema dado, debido a que los resultados que arroje dependen de los datos analizados, el objetivo que se pretende alcanzar, de los parámetros de cada proceso

aplicado, etc. Toda la información recopilada de estas técnicas de minería de datos se encuentra en la revisión literaria, CAPÍTULO II. TÉCNICAS DE MINERÍA DE DATOS.

3. ETAPA TRES: Diseñar el modelo de minería de datos en base a las técnicas seleccionadas.

3.1. Primera Fase: Determinar objetivos del negocio, evaluar la situación actual y determinar el objetivo de la minería.

En esta primera fase se realiza un análisis de la situación inicial del problema, con el fin de comprender los objetivos del negocio, y desarrollar el plan del proyecto. Se analiza a detalle los factores alrededor del problema como recursos, hipótesis, limitaciones, riesgos y contingencia y demás características.

3.1.1. Tarea Uno: Comprensión del Negocio

En esta tarea se describe los antecedentes o contexto inicial, los objetivos del negocio y los criterios de éxito, en otras palabras las necesidades del cliente, así como los factores que pueden influenciar en el resultado final, con el fin de ahorrar esfuerzo innecesario.

3.1.1.1. Actividad 1: Antecedentes

La carrera de Ingeniería en Sistemas, perteneciente a el Área de la Energía las Industrias y los Recursos Naturales No Renovables, que a su vez es parte de la Universidad Nacional de Loja, tiene como misión “Ofertar estudios de grado y formación continua, en el campo de la informática, sistemas computacionales y redes de datos, a través de la investigación científica para vincularse con la sociedad” [60], es por ello que es de suma importancia formar profesionales capaces de enfrentarse a la competitividad que existe en el ámbito laboral.

Si nos adentramos más en el sector educativo, podemos evidenciar algunas temáticas de vital importancia como es la realidad que enfrentan los profesionales, que año a año se

titulan y salen de las universidades con grandes deseos de encontrar el empleo apropiado en donde puedan desarrollarse de forma plena en el ejercicio de su profesión. Es así desde diciembre 2009, el Servicio Público de Empleo opera a través de la Red Socio Empleo Ecuador[1], un proyecto emblemático del Ministerio de Relaciones Laborales[2]. Constituye una red de oficinas nacional que presta servicios de intermediación laboral y capacitación especializada a través de un soporte informático que facilita los procesos de reclutamiento y selección de personal. Para que éste medio pueda resultar útil es necesario que el aspirante sea capaz de conocer y hacer que todas sus cualidades personales, nivel de formación, experiencia y habilidades, sean de acuerdo con el perfil profesional específico de cada estudiante, con ello al postular al empleo relacionado con su perfil tenga mayores probabilidades de éxito.

Es por ello que una solución óptima a esta problemática sería la determinación del perfil profesional de cada estudiante, una de las formas más idóneas para lograr esta meta es con la aplicación de técnicas de minería de datos, un proceso que ha sido aplicado en algunas Instituciones Educativas [93], donde aplicaron dichas técnicas para crear un modelo en la prevención de fraudes, gestión y monitoreo de un negocio, el proyecto fue innovador constituyéndose como un gran aporte investigativo para la Universidad.

Estas técnicas han contribuido por ejemplo a detectar factores que influyen en la disertación y abandono de los estudiantes en su vida académica, de ahí su importancia [7, 94]. Demostrando que en base a la aplicación de técnicas de minería de datos, se puede obtener valiosa información que ayuda a la toma de decisiones.

3.1.1.2. Actividad 2: Objetivos del negocio

- Identificar los perfiles profesionales enfocados en la carrera de ingeniería en sistemas, a través de la formación de los estudiantes.
- Identificar los factores que determinan el perfil profesional de los estudiantes.
- Conocer los perfiles profesionales a los cuales se orienten un grupo de estudiantes.

[1] Página Oficial Red Socio Empleo Ecuador: http://www.socioempleo.gob.ec. [Fecha: 20-febrero-2014].
[2] Página Oficial Ministerio de Relaciones Laborales: http://www.relacioneslaborales.gob.ec/. [Fecha: 20-febrero-2014].

3.1.1.3. Actividad 3: Criterios de éxito (factores)

- Lograr identificar el perfil profesional en al menos un 60% de los egresados de la carrera de ingeniería en Sistemas.
- Determinar el perfil profesional de los egresados en dos aspectos: cualitativo y cuantitativo.
- Determinar el perfil profesional de los egresados de acuerdo a lo cuantitativo tomando en cuenta los records académicos.
- Determinar el perfil profesional de los egresados de acuerdo a lo cualitativo aplicando un test para descubrir las capacidades y habilidades de los egresados.
- Comparar los resultados en el aspecto cuantitativo y cualitativo.
- Comprobar los resultados obtenidos tomando al menos un 10% de egresados y graduados con el fin de averiguar si los empleos se ajustan a los perfiles profesionales obtenidos.

3.1.2. Tarea Dos: Evaluación de la Situación

Esta tarea abarca el inventario de requerimientos de recursos, hipótesis y limitaciones, riesgos y contingencias, terminología y costes y beneficios, es decir todo el escenario del PFC enfocado en los objetivos y metas del mismo.

3.1.2.1. Actividad 1: Inventario de requerimientos de recursos

Para la elaboración del presente trabajo de titulación se ha especificado los recursos necesarios en cuanto al talento humano, hardware, software, y fuentes de datos necesarios para el desarrollo y culminación exitosa del proyecto.

- **Talento Humano:** En el desarrollo del presente trabajo de titulación se contó con la intervención del talento humano descrito a continuación:

  - Postulante: Egda. María José Rodríguez Ojeda.
  - Asesores de Tesis: Ing. Henry Paz Arias.

- Director de Tesis: Ing. René Guamán Quinche.

- **Recursos Hardware:** Para el desarrollo exitoso del proyecto se utilizó las herramientas a continuación descritas:

  - Portátil Compaq 515
  - Portátil HP Pavilion dv4
  - Impresora canon mp230
  - Flash memory Kingston de 8GB
  - Disco Duro Toshiba de 1TB

- **Recursos Software:** Existen diversas herramientas software que permiten realizar los procesos de manera automatizada, ahorrando tiempo y permitiendo obtener resultados confiables, además herramientas libres que ayudan a optimizar los costos de desarrollo del proyecto, herramientas tales como las que se describen a continuación:

  - Sistema Operativo: Windows 7 Home Premium
  - Gestor de Referencias Bibliográficas: Zotero
  - Web Service SGA UNL.
  - Sistema Gestor de Base de Datos: MySQL
  - Frontal de Base de Datos: DatAdmin
  - Herramienta Google para desarrollo de cronogramas: Gantter Project
  - Paquete de Ofimática: Microsoft Office 2013
  - Software de minería de datos: Rapidminer Studio
  - TexMaker: Desarrollo del Artículo Científico
  - Prezi: Desarrollo de presentaciones

- **Fuentes de Datos**

En cuanto al origen de los datos para la aplicación de técnicas de minería de datos se utilizará bases de datos relacionales e información arrojada por encuestas realizadas a los mismos estudiantes que han culminado sus estudios universitarios. Datos que se

encuentran almacenados en Bases de Datos gestionadas por la Unidad de Telecomunicaciones e Información a través de un web services. El perfil profesional, la malla curricular de CIS y los datos obtenidos del Proyecto de Seguimiento a graduados. Las fuentes de datos son descritas a continuación:

a). Web Service del Sistema de Gestión Académico de la UNL

La Institución cuenta con el departamento de Informática y Telecomunicaciones que gestiona el Web Service del Sistema de Gestión Académico, que se trata del servidor que contiene datos de los estudiantes académicos, institucionales, personales y estadísticos. Siendo desarrollado para proporcionar métodos que faciliten la obtención de estos datos solo a personas autorizadas que se autentiquen en el sistema a través de un usuario y contraseña.

Los servicios han sido agrupados en distintas categorías de acuerdo a la información que retornan y contienen métodos ya predeterminados de acuerdo a sus funciones que ayudan a realizar la explotación de los datos con mayor rapidez entre las que tenemos la Académica, Institucional, Personal, Validación y Estadística, de las cuales para el presente proyecto se ha utilizado las tres primeras, descritas a continuación:

- **Académica:** En esta categoría se encontrarán los métodos o servicios relacionados a la información académica, como datos de estudiantes y docentes.

- **Institucional:** En esta categoría se encontrarán los métodos o servicios relacionados con la información institucional. Como datos de: áreas, carreras, módulos, paralelos.

- **Personal:** En esta categoría se encontrarán los métodos o servicios relacionados con información personal de: datos de docentes, datos de estudiantes.

b). Área de la Energía, las Industrias y los Recursos Naturales no Renovables. Carrera de Ingeniería en Sistemas Informáticos y Computacionales.

A medida que se realizó el análisis de la información se advirtió que era necesario recopilar los datos académicos-históricos de los egresados y graduados CIS, anteriores al año de creación del SGA, para completar la información académica de todos los egresados/graduados registrados a partir del año 2008.

Para ello ha sido necesario recurrir a los libros que contienen los registros académicos de los estudiantes de la Carrera de Ingeniería en Sistemas, que se encuentran en poder de la secretaría del AEIRNNR, indagando los registros del proceso académico de los estudiantes de forma individual, reflejado en la malla curricular con sus diferentes variaciones a través de los periodos académicos, filtrando los egresados y realizando un seguimiento minucioso del desarrollo académico con criterio; descartando los módulos en los que el estudiante egresado haya tenido un estado de reprobado o retirado; dando así la veracidad del caso en la información de los estudiantes en su estado como aprobado y así completar la Base de Datos de los que consiguieron culminar sus estudios universitarios y egresaron desde el año 2008.

c). Perfil de Egreso del estudiante de Ingeniería en Sistemas

En general todo egresado de la carrera de ingeniería en sistemas al culminar sus estudios profesionales debe estar en la capacidad de cumplir con el perfil profesional especificado en la estructuración de la carrera [60], sin embardo cada estudiante se destaca más en ciertos aspectos respecto a su pensamiento crítico, reflexivo, analítico y estratégico, siendo capaz de generar soluciones técnicas a problemas reales. Aplicando sus conocimientos obtenidos a lo largo de sus estudios en las diferentes áreas del Hardware, Software, Análisis de Sistemas e Ingeniería del Software, Telecomunicaciones y Organización, donde tienen mayores potencialidades. Los datos del perfil profesional de cada estudiante se los ha recopilado mediante el Test aplicado.

Este test ha sido desarrollado en el framework django; con el objetivo de tener almacenados los datos directamente en una base de datos en tiempo real. El test está conformado por dos preguntas que contienen 20 items respectivamente. Estas preguntas han sido planteadas en base a búsqueda bibliográfica de las 5 características que son propias de cada una de las 8 especialidades escogidas. Al test solo pueden acceder los usuarios registrados, mediante su cedula como usuario y contraseña. El objetivo es almacenar estas características y determinar su perfil profesional utilizando técnicas de minería de datos.

d). Proyecto de Seguimiento a graduados

De la información recopilada a través del web services pertenece a los egresados de la carrera de ingeniería en sistemas, desde la creación del SGA en la UNL. Entre la masa de los egresados se encuentran los graduados que han culminado y obtenido su título académico por el desarrollo y disertación de sus proyectos de fin de carrera. El proyecto de seguimiento a graduados como su nombre lo dice intenta mantener la información actualizada de los profesionales de la carrera, información muy relevante para contactarlos y pedir su colaboración en el test aplicado.

3.1.2.2. Actividad 2: Hipótesis y limitaciones

En cuanto a las hipótesis las que se han determinado son las siguientes:

- Retraso en el desarrollo del proyecto en la fase de limpieza de los datos.
- Retraso al completar los datos obtenidos del SGA de los egresados en base a los registros que se encuentran en los libros físicos.
- Resultados obtenidos corresponden a una certidumbre que supera el 80%.
- En base a una muestra los perfiles profesionales determinados de los estudiantes concuerdan con los empleos en los que se desempeñan actualmente los profesionales.
- Los resultados obtenidos al final del proyecto ayuden a la toma de decisiones en varios aspectos.
- El aprendizaje de las herramientas a usar en el desarrollo del proyecto no retrasen la obtención de los resultados.

- La obtención de los datos, mediante la disfusión y respuesta del test planteado retrasen el desarrollo normal del proyecto.
- Complicaciones para conseguir de manera rápida alojamiento para el test desarrollado en el framework django.

En cuanto a las limitaciones se han determinado:

- El volumen de los datos que se encuentran en el web service de los egresados de la carrera de ingeniería a sistema corresponde a registros desde el 2008 hasta la actualidad, debido a que en esa fecha se creó el Sistema de Gestión Académica, y no existen los datos de los egresados desde la creación de la carrera, es decir los datos de los años anteriores son datos históricos.

- El proyecto de seguimiento a graduados no contiene todos los registros de los graduados de la carrera de ingeniería en sistemas debido a la dificultad para recopilar los datos, por la desvinculación de los graduados de la institución, por razones como: encontrarse en el exterior, muertes, no ejercer la profesión, etc.

- Son escasas las empresas que brindan alojamiento para aplicaciones hechas en python. El framework django trabaja en base a este lenguaje de programación, por lo que ha existido dificultad en encontrar alojamiento. Sin embargo se ha encontrado apoyo en la unidad de tele comunicaciones e información de la Universidad nacional de Loja, donde se ha brindado una ip publica, para su publicación, debido que el proceso de evaluación docente está hecho bajo la misma plataforma. Por lo que ha sido necesario administrar el servidor y realizar las configuraciones necesarias, después de haber contado con los permisos respectivos.

- La difusión del test se la ha realizado por distintos medios como: la red social Facebook mediante mensajes directos y una publicación en el grupo de graduados de la carrera de ingeniería en sistemas, las redes profesionales linkedin y viadeo, página de la carrera a través de una publicación, envió de correos masivos obtenidos del webservice en la

categoría de datos personales, dirigido a los egresados y graduados. A pesar de ello la respuesta no ha sido total.

3.1.2.3. Actividad 3: Riesgos y contingencias

En el desarrollo del presenta trabajo de titulación se han suscitado algunos riesgos, para los cuales se ha planteado algunas estrategias de contingencia, para culminar los objetivos planteados (ver tabla XLIX).

TABLA XLIX

RIESGOS/CONTINGENCIAS DEL TRABAJO DE TITULACIÓN

Riesgo	Contingencia
Contratiempos en la obtención de los datos debido a la gestión de los permisos respectivos.	Realizar las peticiones correspondientes de los datos a las autoridades en un periodo de tiempo ligado al cronograma de actividades.
Tiempo del proyecto dure más que lo planificado.	Desarrollo de un cronograma donde se tome en cuenta un tiempo prudente para cada una de las fases.
Perdida de los datos obtenidos de las fuentes de datos.	Realizar respaldos en diferentes dispositivos, subirlos en un espacio en la nube de forma periódica con cada cambio.
Problemas con el presupuesto en los requisitos software y talento humano.	Utilizar herramientas de uso libre, solicitar asesoría por parte de los docentes de la Institución.
Resultados menos dramáticos que los esperados.	Obtener varios modelos en el marco de la resolución del problema haciendo varias pruebas hasta obtener los resultados deseados.
Retraso en recopilar los datos históricos anteriores al año de creación del SGA, para completar la información.	Emplear estrategias de búsqueda para el filtrado de los datos usando horas diarias extras a lo planificado.
Dificultad para encontrar alojamiento del test desarrollado en el framework django en base a la plataforma python.	Realizar las gestiones pertinentes en la Unidad de Telecomunicaciones e Información, para adquirir alojamiento en el servidor de la UNL.
Retraso en recopilar los datos del test desarrollado en el framework django.	Utilizar distintos medios de difusión para crear un impacto que genere una buena respuesta por parte de los egresados y graduados de la carrera de ingeniería en sistemas.

3.1.2.4. Actividad 4: Terminología (Glosario)

- **Del negocio**

- **UNL:** Proviene de las siglas Universidad Nacional de Loja
- **AEIRNNR:** Área de Energía, las Industrias y los Recursos Naturales no Renovables
- **CIS:** Carrera de ingeniería en Sistemas
- **UTI:** Unidad de Telecomunicaciones e Información.
- **Perfil Profesional:** conjunto de capacidades y competencias que identifican la formación de una persona para asumir en condiciones óptimas las responsabilidades propias del desarrollo de funciones y tareas de una determinada profesión [95].

- **De la minería de datos**

- **KDD:** Proviene de las siglas descubrimiento de información en bases de datos, en inglés knowledge discovery in databases.
- **Minería de datos:** Es la ciencia de extracción de información útil de grandes conjuntos de datos o de bases de datos [64].
- **Fuentes de datos:** Representan los datos, los orígenes, la recolección de datos válidos y confiables, para la resolución de problemas, de ello depende que el modelo sea confiable.
- **Web Service:** Es una tecnología que utiliza un conjunto de protocolos y estándares que sirven para intercambiar datos entre aplicaciones.
- **Modelo:** Es una representación del mundo real, es la solución del problema. Un modelo es la forma de tratar la información.
- **Clustering:** Se utiliza directamente similaridad entre los datos en entrada, es decir se identifica clusters o grupos en el conjunto de datos, donde un cluster es una colección de datos “similares”. Los clusters deben ser de buena calidad es decir ser escalables a grandes bases de datos, esto es lo que intenta la minería de datos, la similitud determinada para formar los clusters [29, 30].
- **Algoritmos Supervisados o predictivos:** Para conocer el comportamiento de la variable a predecir, se debe contar con un conjunto de variables predictoras, que

permitan predecir el valor de un atributo de un conjunto de datos, conocidos otros atributos. Una vez entrenado el modelo, sirve para realizar la predicción de datos cuyo valor es desconocida [24, 25, 27].

- **Algoritmos no supervisados:** Estos algoritmos ayudan a descubrir patrones y tendencias en los datos actuales, que no poseen variable a predecir. Los registros son agrupados por similaridad. El objetivo es descubrir el conocimiento para tomar acciones y obtener un beneficio [24, 25, 27].

3.1.2.5. Actividad 5: Análisis de Costo/Beneficio

La actividad engloba los costes detallados del proyecto, los mismos que serán agrupados de acuerdo a su finalidad para su entendimiento.

- **Costes de Personal**

Estos costes hacen referencia al tiempo que el talento humano invirtió en el desarrollo del presente trabajo de titulación. Tomando en cuenta el tiempo del director del proyecto en un 20% aproximadamente respecto del tiempo empleado por el postulante. El precio asignado por hora de tutoría es un estimado del 25% del costo total en horas (ver tabla L). Justificándose en base a la revisión de fuentes bibliografías [96] y plasmado a continuación:

TABLA L

COSTO POR HORA DEL TALENTO HUMANO

Rol	Costo por Hora ($)
Postulante	$ 5,00
Asesor del Trabajo de Titulación	$ 6,00
Director del Trabajo de Titulación	$ 7,00

Posteriormente se determinó las distintas actividades a realizar en el proyecto junto con su duración correspondiente (ver tabla LI).

TABLA LI

ACTIVIDADES DEL PROYECTO Y SU DURACIÓN

Nro.	Actividad	Duración (horas)
1	Revisión Bibliográfica respecto de la minería de datos	96
2	Análisis y recopilación de las fuentes de datos.	240
3	Examinar las técnicas y herramientas de minería de datos.	48
4	Análisis y limpieza de los datos recopilados.	480
5	Buscar los patrones que puedan generarse por el procesamiento de los datos.	48
6	Plasmar el o los modelos de acuerdo a los datos y la técnica seleccionada.	160
7	Generar modelos para la determinación de los perfiles profesional.	168
8	Validar el modelo generado de acuerdo a los objetivos planteados.	240
9	Evaluar los resultados obtenidos y compararlos con los esperados.	160
TOTAL:		1640

Las actividad requieren de uno o más del personal involucrado en el proyecto, debido a ello detallamos el porcentaje de participación de cada uno de los integrantes en las actividades (ver tabla LII).

TABLA LII

PORCENTAJE DE PARTICIPACIÓN DEL PERSONAL EN CADA ACTIVIDAD

Act.	Personal	% de Implicación	Horas
1	Postulante	60 %	57,6
	Asesor del Proyecto	0 %	0
	Director del Proyecto	40 %	38,4
2	Postulante	70 %	168,0
	Asesor del Proyecto	0 %	0
	Director del Proyecto	30 %	72,0
3	Postulante	60 %	28,8
	Asesor del Proyecto	10 %	4,8
	Director del Proyecto	30 %	14,4
4	Postulante	70 %	336
	Asesor del Proyecto	10 %	48
	Director del Proyecto	20 %	96
5	Postulante	80 %	38,4
	Asesor del Proyecto	10 %	4,8

	Director del Proyecto	10 %	4,8
6	Postulante	70 %	112
	Asesor del Proyecto	20 %	32
	Director del Proyecto	10 %	16
7	Postulante	70 %	117,6
	Asesor del Proyecto	20 %	33,6
	Director del Proyecto	10 %	16,8
8	Postulante	70 %	168
	Asesor del Proyecto	20 %	48
	Director del PFC	10 %	24
9	Postulante	80 %	128
	Asesor del PFC	10 %	16
	Director del PFC	10 %	16

A continuación, se realiza el coste del talento humano del proyecto por horas invertidas, tomando en cuenta la tabla 8 anterior, (ver tabla LIII).

TABLA LIII

COSTE FINAL DEL PROYECTO DE ACUERDO A LAS HORAS INVERTIDAS

Integrante	Costo/Hora ($)	Horas Invertidas	Coste Total ($)
Postulante	$ 5.00	1154.4	$ 5772.00
Asesor del Proyecto	$ 6.00	187.2	$ 1127.20
Director del Proyecto	$ 7.00	298.4	$ 2088.80
TOTAL:			$ 8988.00

- **Costos de Hardware**

El hardware juega un papel importante en el desarrollo del presente proyecto, los costes deben ser analizados de acuerdo al precio de cada equipo, el número de equipos y el total. A continuación se detallan estos costes, para los cuales se tomó en cuenta su vida útil de 5 años y el cálculo de su depreciación (ver tabla LIV).

TABLA LIV

COSTES DE LOS RECURSOS HARDWARE

Descripción	Costo	T. De Vida (5 Años)	Total
Portátil Compaq 515	$ 805.00	161	$ 107.33
Portátil HP Pavilion dv4	$ 1070.71	214.14	$ 142.76
Impresora	$ 50.00	10	$ 6.67
Flash memory 8GB	$ 15.00	2	$ 30.00
TOTAL			**$ 286.76**

- **Costes de Software**

La minería de datos pretende analizar grandes volúmenes de datos de forma automatizada, por lo tanto el software toma un papel muy importante en el desarrollo del presente proyecto. A continuación se detallan los costes asociados al mismo, necesario para el cumplimiento de los objetivos del presente proyecto (ver tabla LV).

TABLA LV

COSTES DE LOS RECURSOS SOFTWARE

Software	Descripción	Cant	V. Unit.	V. Total
Gantter Project.	Desarrollo de cronogramas	1	$ 00.00	$ 00.00
Mysql	Gestor de Base de Datos	1	$ 00.00	$ 00.00
DatAdmin	Administrador de Base de datos	1	$ 00.00	$ 00.00
Rapidminer	Software de minería de datos	1	$ 0.00	$ 0.00
TexMaker	Software de documentación	2	$ 0.00	$ 0.00
django	Desarrollo de aplicaciones web.	1	$ 0.00	$ 0.00
python	Lenguaje de programación	1	$ 0.00	$ 0.00
TOTAL:				$ 0.00

Las herramientas detalladas en la tabla LV, son de uso libre, debido a ello no constituyeron ningún coste en el proyecto.

- **Materiales y Servicios**

Dentro de los costes asociados al proyecto se encuentran los materiales de oficina que han sido necesarios para los trámites que conlleva el trabajo de titulación, la presentación de las distintas fases y la entrega de la memoria final (ver tabla LVI).

TABLA LVI

COSTES DE LOS MATERIALES DE OFICINA

Material	Cantidad	V. Unit.	Total
Resmas de Hoja	2	$ 4.00	$ 8.00
Tinta para Impresora	4	$ 7.00	$ 28.00
Empastado	6	$ 5.00	$ 30.00
TOTAL			**$ 66.00**

Por otra parte los costes respecto a los servicios representan un 6% del total del costo de desarrollo del presente trabajo de titulación (ver tabla LVII).

TABLA LVII

COSTE DE LOS SERVICIOS.

Categoría	Descripción	Total
Agua y luz eléctrica	Servicios de 10 meses aproximadamente	$ 50.00
Internet	100 horas de internet de telefonía fija	$ 80.00
Llamadas a Celular	60 min aproximadamente	$ 15.00
Transporte en bus y taxi	Aproximadamente 120 veces en bus y 20 en taxi	$ 105.00
Publicación de Artículo científico	Publicación en revista indexada.	$ 200.00
Cursos de Capacitación	Seminarios, talleres.	$ 120.00
TOTAL		**$ 570.00**

Como podemos observar en la tabla LVII se han especificado servicios de comunicación, servicios básicos, servicios de transporte, gastos de publicación y costes en capacitación. Sumando un total de $ 570.00 en estos servicios y $ 66.00 en costes de materiales de oficina, dando como resultado $ 636.00 de costes de materiales y servicios detallados en esta sección.

- **Coste Total**

Se ha realizado el cálculo de costes total que implica el desarrollo del presente trabajo de titulación, incluyendo los costos de personal, hardware, software, materiales y servicios detallados anteriormente, así como los valores de imprevistos correspondientes al diez por ciento del total del valor del costo total calculado (ver tabla LVIII).

TABLA LVIII

DETALLE DEL COSTO FINAL DEL TRABAJO DE TITULACIÓN

Recursos	Coste Total ($)
Costes de Personal	$ 8915.20
Costes de Hardware	$ 286.76
Costes de Software	$ 0.00
Costes de Materiales y Servicios.	$ 636.00
Subtotal:	**$ 9837.96**
Imprevistos 10%	$ 983.79
TOTAL:	**$ 10821.75**

3.1.2.6. Actividad 6: Cronograma del Proyecto

El cronograma del PFC ha sido elaborado en la herramienta Gantter Project de google a través del correo institucional, plasmado en un diagrama de Gantt donde se pueden visualizar las distintas fases del proyecto con sus respectivas actividades que han sido de ayuda para cumplir con los objetivos del proyecto de manera ordenada con el uso de los recursos antes detallados en cada fase planificada (ver Figura 35).

		Nombre	Duración	Inicio	Fin
1		**FASE 1: Investigar las características y variables más influyentes de los datos almacenados, para resolver el problema.**	19d?	**01/11/2013**	**27/11/2013**
2		Realizar una investigación bibliográfica respecto de la minería de Datos.	6d?	01/11/2013	08/11/2013
3		Análisis de las características que servirán para determinar los empleos apropiados en los profesionales.	4d?	11/11/2013	14/11/2013
4		Recopilación y análisis de las fuentes de datos que se va a necesitar para realizar la minería de datos.	4d?	15/11/2013	20/11/2013
5		Documentación de la memoria final.	4d?	20/11/2013	25/11/2013
6		Tutoria de la Fase	1d?	26/11/2013	26/11/2013
7		Fin FASE 1	1d?	27/11/2013	27/11/2013
8		**FASE 2: Comparar y seleccionar la técnica de minería de datos más adecuada.**	27d?	**28/11/2013**	**03/01/2014**
9		Hacer un análisis comparativo de las técnicas de minería de datos ya utilizadas de acuerdo a sus ventajas y desventajas.	3d?	28/11/2013	02/12/2013
10		Seleccionar la o las herramientas disponibles para realizar el proceso de minería de datos.	6d?	03/12/2013	10/12/2013
11		Realizar la transformación de los datos al formato requerido para el posterior procesamiento de minería.	12d?	11/12/2013	26/12/2013
12		Documentación de la memoria final.	4d?	27/12/2013	01/01/2014
13		Tutoria de la Fase	1d?	02/01/2014	02/01/2014
14		Fin FASE 2	1d?	03/01/2014	03/01/2014
15		**FASE 3: Elaboración y diseño del modelo para la toma de decisiones en la selección de los empleos apropiados para los profesionales.**	75d?	**06/01/2014**	**18/04/2014**
16		Proceso de Minería de Datos.	25d?	06/01/2014	07/02/2014
17		Crear la estructura de minería de datos tomando en cuenta las variables influyentes en el problema.	20d?	10/02/2014	07/03/2014
18		Buscar los patrones que puedan generarse por el procesamiento de los datos	9d?	10/03/2014	20/03/2014
19		Plasmar el modelo planteado de acuerdo a los datos de prueba y a la técnica seleccionada.	15d?	21/03/2014	10/04/2014
20		Documentación de la memoria final.	4d?	11/04/2014	16/04/2014
21		Tutoria de la Fase	1d?	17/04/2014	17/04/2014
22		Fin FASE 3	1d?	18/04/2014	18/04/2014
23		**FASE 4: Interpretar y evaluar el modelo de minería de datos aplicado en un contexto real, culminación del PFC.**	39d?	**21/04/2014**	**12/06/2014**
24		Validar los patrones encontrados para garantizar el cumplimiento de los objetivos planteados.	12d?	21/04/2014	06/05/2014
25		Comparar los resultados obtenidos con los propósitos esperados.	8d?	07/05/2014	16/05/2014
26		Documentación Final de la memoria.	4d?	19/05/2014	22/05/2014
27		Redacción de un artículo científico con el estándar IEEE.	7d?	23/05/2014	02/06/2014
28		Difución de los resultados al campo científico.	6d?	03/06/2014	10/06/2014
29		Culminación del PFC	1d?	11/06/2014	11/06/2014
30		Fin FASE 4	1d?	12/06/2014	12/06/2014

Figura 35. Cronograma del Trabajo de Titulación.

3.1.3. Tarea Tres: Determinación de metas de la minería de datos

La minería de datos es el proceso de exploración, análisis, extracción y refinamiento de grandes volúmenes de información de manera automatizada, con el fin de descubrir conocimiento, es decir información que ayude a la toma de decisiones. El conocimiento es descubierto con la aplicación de técnicas de minería de datos que permiten revelar modelos en los datos tomando en cuenta patrones esenciales que ayuden a cumplir con la meta de la minería de datos [58, 97].

La meta depende del proyecto que se esté realizando, por ello en el presente proyecto al aplicar el proceso de minería de datos se busca que a partir de un conjunto de datos se descubran uno o varios modelos que determinen los perfiles profesionales mediante la aplicación de técnicas de minería de datos. Los datos se los ha recopilado enfocados de dos formas: datos cualitativos y cuantitativos. Los datos cuantitativos engloban los records académicos de estudiantes egresados, mientras que los datos cualitativos correspondes a los obtenidos de un test que se ha desarrollado para obtener los intereses, las capacidades, habilidades e interés de cada uno.

Ha sido necesario para el cumplimiento de esta meta el análisis de las fuentes de datos disponibles en la Universidad Nacional de Loja, así como la recopilación y filtrados de los datos obtenidos de dichas fuentes. Parte esencial ha sido a su vez la selección de las herramientas que permitirán realizar la minería de datos en menor tiempo, con mayor calidad, etc. Dichas herramientas deberán permitir aplicar las técnicas de minería más adecuada de acuerdo al objetivo que se desea alcanzar.

A su vez para alcanzar la meta o propósito se ha planteado algunas fases como: Investigar las características y variables más influyentes de las fuentes de datos a utilizar, comparar y seleccionar la técnica de minería de datos de acuerdo al ambiente de estudio, diseñar el modelo de minería de datos en base a las técnicas seleccionadas, interpretar y evaluar el modelo de minería de datos aplicado en un contexto real.

Cabe recalcar que las herramientas escogidas son de uso libre y permiten que los datos recopilados y filtrados sean tratados de manera rápida y realizar la minería de datos con calidad en base a patrones de datos válidos y procesables.

3.1.4. Tarea Cuatro: Elaboración del plan de Proyecto

El plan del Proyecto se lo realizó en base a diferentes aspectos, con el objetivo de llegar a cumplir con la meta de la minería de datos. Para ello se ha desglosado el desarrollo del presente proyecto en diferentes actividades especificando el tiempo utilizado, el porcentaje de implicación de los involucrados, las herramientas manipuladas y los resultados obtenidos en cada actividad (ver tabla LIX). En esta tabla se ha utilizado la nomenclatura PT, AT y DT; donde:

PT = Postulante del Trabajo de Titulación
AT= Asesor del Trabajo de Titulación
DT= Director del Trabajo de Titulación

TABLA LIX

PLAN DEL PROYECTO

Actividad	% Implicación			Tiempo (Horas)	Herramienta	Resultado
	PT	AT	DT			
Revisión Bibliográfica respecto de la minería de datos.	60	30	10	96	Gestor Bibliográfico Zotero	Estado del Arte respecto a la minería de datos.
Análisis y recopilación de las fuentes de datos.	70	10	20	48	SGBD Mysql, Frontal DatAdmin, Web Service del SGA, Excel, django, python.	Acceso a los libros con los datos históricos de los egresados, acceso a los datos del SGA a través del web service.
Examinar las técnicas y herramientas de minería de datos.	60	10	30	48	Gestor Blibliográfico Zotero, Kmine, Rapidminer, Weka	Análisis de casos de éxitos, prácticas con los datos a través de herramientas de MD.
Análisis y limpieza de los datos recopilados.	70	0	30	480	SGBD Mysql, Frontal DatAdmin, Excel.	Datos del SGA combinados con los datos históricos de los libros.
Buscar los patrones que puedan generarse por el procesamiento de los datos.	90	0	10	160	Rapidminer, Excel	Análisis de los datos ya filtrados, identificando factores que arrojen el perfil profesional de los egresados.
Plasmar el o los modelos de acuerdo a los datos y la técnica seleccionada.	80	0	20	160	Rapidminer, Excel	Modelo adecuado a la técnica escogida, arrojando ya resultados de los perfiles profesionales.
Generar modelos para la determinación de los perfiles profesional.	90	0	10	168	Rapidminer, Excel	Modelos completos probados datos reales.
Validar el modelo generado de acuerdo a los objetivos planteados.	80	0	20	240	Rapidminer, TexMaker	Comparación del modelo final respecto de los resultados esperados.
Evaluar los resultados obtenidos y compararlos con los esperados.	80	0	20	240	Rapidminer, TexMaker	Analizar resultados con un contexto real.

3.2. Segunda Fase: Comprensión de los Datos. Recopilación, Exploración y verificación de los Datos obtenidos.

Esta es la fase de la metodología donde se pretende reducir la información a la únicamente necesaria para realizar la minería de datos, a su vez relacionarse directamente con la información para su mayor comprensión. Al realizar la comprensión se pasará a la preparación de los datos con las bases necesarias para realizar esta nueva fase con éxito.

3.2.1. Tarea Uno: Obtener los datos iniciales

- **Reporte de la obtención de los datos iniciales**

El conjunto de datos obtenidos provienen de las fuentes de datos procedentes de la Universidad Nacional de Loja, debido a que el presente trabajo de titulación es de carácter académico. A continuación se detalla brevemente los datos obtenidos:

- Datos de los egresados de la carrera de ingeniería en sistemas respecto a las categorías académica y personal provenientes del Sistema de Gestión Académica de la institución creado en el 2008. Estos datos se obtuvieron a través del Web Service para su posterior explotación, en base a distintos métodos ya predefinidos para la consulta de los mismos; y almacenados en una base de datos con el Gestor de Base de datos Mysql y administrados en el Frontal de Base de datos libre DatAdmin.
- Registros de los records académicos de los estudiantes egresados de la carrera de ingeniería en sistema, que se encuentran en los Libros físicos que están en poder de la secretaría del Área de la Energía, las Industrias y los Recursos Naturales no Renovables de la institución.
- Datos personales de los graduados de la carrera de ingeniería en sistemas del Sistema de seguimiento a Graduados de la carrera.
- Test de habilidades, capacidades e intereses desarrollado en el framework django bajo el lenguaje de programación python.

En el transcurso de la Recolección inicial de los datos se presentaron algunos inconvenientes, por lo cual se aplicaron algunas soluciones para continuar con el desarrollo normal del presente proyecto (ver tabla LX).

TABLA LX

RECOLECCIÓN INICIAL DE LOS DATOS, INCONVENIENTES/SOLUCIONES

Inconvenientes	Soluciones
Dificultad al obtener los datos directamente del Sistema de Gestión Académico.	Utilización del Web Service para la consulta de los datos específicos del SGA, a través de los servicios agrupados en distintas categorías y sus métodos predeterminados de acuerdo a sus funciones.
Problemas de administración de la base de datos debido a la herramienta en uso.	Consulta bibliográfica de las distintas herramientas de administración de base de datos y selección de la más adecuada.
Datos incompletos, debido a que el SGA se creó en el año 2008.	Los registros anteriores al año 2008 en que se creó el SGA, corresponden a datos históricos. Por lo que se ha tenido que recurrir a los registros de los libros físicos, filtrar los datos y pasarlos manualmente.
Retraso en recopilar los datos históricos, para completar los datos.	Emplear estrategias de búsqueda para el filtrado de los datos usando horas diarias extras a lo planificado.
Perdida de los datos obtenidos de las fuentes de datos.	Realizar respaldos en diferentes dispositivos, subirlos en un espacio en la nube de forma periódica con cada cambio.
Dificultad para encontrar alojamiento del test desarrollado en el framework django.	Realizar las gestiones pertinentes en la Unidad de Telecomunicaciones e Información, para adquirir alojamiento de la aplicación en el servidor de la UNL, considerando que el proceso de evaluación docente se también está desarrollado bajo django.
Retraso en recopilar los datos del test aplicado.	Se ha utilizado distintos medios de difusión para crear un impacto y generar respuesta en la contestación del test.

Al observar la tabla LX, podemos constatar los inconvenientes que se ha sufrido en esta etapa. De todos estos inconvenientes se puede decir que el de mayor grado de complicación ha sido el de los datos incompletos, debido que a medida que se realizó el análisis de la información se advirtió que era necesario recopilar los datos históricos anteriores al año de creación del SGA, para completar la información académica de todos los egresados registrados a partir del año 2008.

Para ello ha sido necesario recurrir a los libros que contienen los registros académicos de los estudiantes de la Carrera de Ingeniería en Sistemas, indagando de forma individual el proceso académico reflejado en la malla curricular con sus diferentes variaciones a través de los periodos académicos, realizando un seguimiento minucioso del desarrollo académico con criterio; descartando los módulos en los que el estudiante haya tenido un estado de reprobado o retirado; dando así la veracidad del caso en la información de los estudiantes en su estado como aprobado y así completar la Base de Datos de los que consiguieron culminar sus estudios universitarios y egresaron desde el 2008.

Ésta tarea de solución ha permitido tener un conocimiento a profundidad de algunos factores que más adelante serán determinantes con la aplicación de Minería de Datos para contrastar y analizar los resultados en cuanto a los perfiles profesionales determinados. La recopilación de los datos se realizó a partir de copias sacadas de los libros que se encuentran en poder de la secretaría del Área de la Energía, las industrias y los recursos naturales no renovables; con la autorización pertinente, siendo trasladados a un archivo en formato Excel, para su posterior migración a la base de datos ya establecida de los datos obtenidos del SGA a través del web service, con ello los registros académicos de los egresados desde el año 2008 han sido completados, para su posterior consumo en la minería de datos. En la solución a este inconveniente se ha ocupado gran parte del tiempo planificado para el desarrollo del presente proyecto, permitiendo continuar con la recolección inicial de datos.

Los datos que se ha visto necesario analizar para determinar el perfil profesional a más del aspecto cuantitativo respecto al record académico, son respecto a la parte cualitativa de cada egresado o graduado de la carrera de ingeniería en sistemas que corresponde a sus

habilidades, capacidades e intereses; de esta forma se ha obtenido un perfil más real, y a su vez se ha podido realizar una comparación de los resultados cuantitativos y cualitativos. Para la obtención de estos datos se ha desarrollado un test en el framework para aplicaciones web django que trabaja en base al lenguaje de programación python. La dificultad no ha sido el desarrollo sino más bien encontrar alojamiento para la aplicación, debido a que no existen muchas empresas que den soporte para aplicaciones hechas en python. La solución se ha dado al realizar los trámites pertinentes para que el alojamiento sea en el servidor de la institución, debido que el proceso de evaluaciones docentes también está desarrollado con django. Después de los trámites respectivos se ha realizado la administración del servidor para correr la aplicación y con esfuerzo se ha logrado tenerla corriendo en la web a través de la ip pública: 192.188.49.12:8080.

Teniendo el test ya en producción se ha realizado la difusión del mismo para terminar con la recolección de los datos, ha sido un proceso duro para llegar a despertar el interés por parte de los evaluados, pero utilizando distintos medios se ha logrado terminar la recolección inicial de los datos con éxito.

A continuación se presenta un reporte de Recolección Inicial de Datos:

- **Datos del Sistema de Gestión Académica**

- Tipo: Base de Datos Mysql
- Locación: Servidor de la institución
- Método de acceso: Web Service de la Institución
- Problemas encontrados: Retraso para obtener los datos hasta familiarizarse con la herramienta de consulta.
- Fortalezas: Métodos ya incorporados en el Web service de acuerdo a los servicios para la consulta de los datos.

- **Datos Históricos de los records académicos**

- Tipo: Plantillas Excel 2010

- Locación. Libros Físicos de la secretaría del AEIRNNR
- Métodos de Acceso: Copias de los Libros Físicos y registro manual.
- Problemas encontrados: Desorden en los datos, dificultad para encontrar los datos necesarios con rapidez.
- Fortalezas: Datos completados de los records académicos de los egresados de la carrera de ingeniería en sistemas que constaban en el SGA a partir del año 2008.

- **Datos personales de los graduados**

- Tipo: Plantillas Excel 2010
- Locación: Sistema de Seguimiento a Graduados
- Método de acceso: Descarga directa de los registros en el Sistema
- Problemas encontrados: Datos incompletos debido a la dificultad de la recolección de los datos personales.
- Fortalezas: Datos de suma importancia debido a su utilidad para la validación de los resultados de la minería de datos.

- **Test de habilidades, capacidades e intereses**

- Tipo: aplicación web hecha en base al framework django.
- Locación: alojamiento en el servidor de la institución.
- Método de acceso: descarga de las respuestas del test de la base de datos de la aplicación realizada en mysql.
- Problemas encontrados: dificultad para la recolección de los datos.
- Fortalezas: determinación del perfil profesional más realista.

3.2.2. Tarea Dos: Describir los datos

- **Reporte con la descripción de los datos**

En esta tarea se pretende examinar las propiedades de los datos adquiridos, de manera superficial. Datos que fueron recolectados de distintas fuentes de datos y que al unirlos se consiguió tener en una base de datos, de forma que nos ofrezca ventajas como:

- Explotar la información de forma ordenada y separada.
- Tener independencia de los Datos
- Reducción de la Redundancia
- Mejor disponibilidad de los datos.
- Mayor eficiencia en la recogida de los datos.
- Mayor coherencia en los resultados
- Mayor valor informativo.
- Una Base de datos que contienen los datos estrictamente necesarios para el proceso de minería de manera ordenada.
- Acceso más rápido y sencillo
- Mayor flexibilidad para adaptarse a trabajos futuros.
- Reducción del espacio de almacenamiento

Para familiarizarse de manera más clara con los datos adquiridos se procede a examinar algunas de sus propiedades, respecto de las fuentes de datos de las que han sido adquiridos:

Para acceder a los datos del Sistema de Gestión Académica de la UNL, se hizo uso de la herramienta Web service que permite la consulta y obtención de los datos contenidos en este sistema. El Web service ofrece un sin número de servicios que son explotados en base a algunos métodos de consulta ya incorporados en la herramienta distribuidos en 5 categorías básicas de acuerdo a los datos que arrojan (ver tabla LXI).

TABLA LXI

CATEGORÍAS DEL WEB SERVICE DE ACUERDO A LOS DATOS DE RETORNO

Categoría	Métodos
Académica	En esta categoría se encontrarán los métodos o servicios relacionados a la información académica, como datos de estudiantes y docentes.
Institucional	En esta categoría se encontrarán los métodos o servicios relacionados con la información institucional, como datos de áreas, carreras, módulos y paralelos.
Personal	En esta categoría se encontrarán los métodos o servicios relacionados con información personal de datos de docentes y datos de estudiantes.
Validación	En esta categoría se encontrarán los métodos o servicios relacionados con la validación de docentes y estudiantes.
Estadística	En esta categoría se encontrarán los métodos o servicios relacionados con información estadística, como número de estudiantes matriculados, número de estudiantes aprobados y número de estudiantes reprobados.

De las 5 categorías que se observan en la tabla LXI, para el presente proyecto se ha hecho uso de tres de ellas: Académica, Institucional y Personal. Cada categoría contiene métodos que de consultas ya predeterminados de acuerdo a sus funciones para la explotación de datos con mayor rapidez, los cuales se detallan a continuación respecto de la categoría a la que corresponden:

- **Categoría Académica**

En esta categoría se encontrarán los métodos o servicios relacionados a la información académica, referentes a datos de estudiantes y docentes. Pero específicamente de esta categoría se obtuvo las notas de los egresados, utilizando los siguientes métodos (ver tabla LXII).

TABLA LXII

ESPECIFICACIONES DE LOS MÉTODOS UTILIZADOS EN LA CATEGORÍA ACADÉMICA

Método	Parámetros	Salida
sga_periodos_lectivos()		Devuelve una lista de periodos lectivos.
sgaws_notas_estudiante()	(cedula, idCarrera, idOferta)	Devuelve las notas de los estudiantes respecto de su cedula, idCarrera, id oferta, asociados a un paralelo, módulo, carrera, etc.
sgaws_ofertas_academicas()	(id_periodo)	Devuelve una lista de ofertas académicas correspondientes a un periodo lectivo específico.
sgaws_carreras_estudiante()		Devuelve las carreras con su id correspondiente.

- **Categoría Institucional**

En esta categoría se encontrarán los métodos o servicios relacionados con la información institucional, referentes específicamente a las distintas áreas, carreras, módulos y paralelos. De la cual se obtuvo datos de los estudiantes respecto al Área de la Energía, las Industrias y los Recursos Naturales no Renovables, así como la carrera específicamente de ingeniería en sistemas y todos los módulos registrados en el SGA, con el uso ciertos métodos definidos (ver tabla LXIII).

TABLA LXIII

ESPECIFICACIONES DE LOS MÉTODOS UTILIZADOS EN LA CATEGORÍA INSTITUCIONAL

Método	Parámetros	Salida
sgaws_carreras_area	(sigla_carrera)	Devuelve las carreras que conforman cada área.
sgaws_datos_carrera()	(id_oferta)	Devuelve un detalle de cada carrera respecto de una oferta académica.
sgaws_lista_areas()		Devuelve una lista de las áreas de la institución.
sgaws_modulos_carrera()	(oferta_id, carrera_id)	Devuelve los datos de todos los módulos de una carrera específica.
sgaws_paralelos_carrera()	(oferta_id, carrera_id)	Devuelve los paralelos de cierta oferta académica de una determinada carrera.

- **Categoría Personal**

En esta categoría se encontrarán los métodos o servicios relacionados con información personal tanto de docentes como datos de estudiantes. Pero específicamente se obtuvo datos personales de los estudiantes del Área de la Energía, las Industrias y los Recursos Naturales no Renovables, de la carrera de ingeniería en sistemas de la UNL. Datos como nombres, apellidos y cédula. Haciendo uso de un método estratégico en esta categoría, orientado al estudiante (ver tabla LXIV).

TABLA LXIV

ESPECIFICACIONES DE LOS MÉTODOS UTILIZADOS EN LA CATEGORÍA PERSONAL

Método	Parámetros	Salida
sgawd_datos_estudiante()	(idCarrera)	Devuelve los datos personales del estudiante o los estudiantes respecto de la carrera ingresada.

Es así que para obtener los datos requeridos del Sistema de Gestión Académica se lo realiza, con la clave de usuario y contraseña otorgados para el accedo al webservice, donde nos autentificamos a través del navegador en la url: [http://ws.unl.edu.ec/]; en esta dirección se encuentran los datos que se obtienen haciendo un llamado a los métodos ya

incorporados en los servicios que ofrece el web service. El llamado a estos métodos ha sido realizado a través de una app creada en php con lo cual se lee los datos que están en formato Json en el web services y se los almacena en una base de datos para su posterior explotación.

El formato Json (*JavaScript Object Notation - Notación de Objetos de JavaScript*), es un formato ligero para el intercambio de datos. Leerlo y escribirlo es simple para humanos, mientras que para las máquinas es simple interpretarlo y generarlo. En texto en formato JSON, se presentan de la forma:

- Un objeto es un conjunto desordenado de pares nombre/valor. Un objeto comienza con { (llave de apertura) y termine con } (llave de cierre). Cada nombre es seguido por: (dos puntos) y los pares nombre/valor están separados por, (coma).

A partir de estos datos en este formato se realiza una limpieza en cuanto a eliminación de caracteres innecesarios como corchetes, llaves y comillas, para posteriormente almacenar la información en la base de datos.

Específicamente con el Web Service, se obtuvieron los récords académicos e información personal de los egresados de la carrera de ingeniería en sistemas que estaban registrados en el Sistema de Gestión Académica desde su creación en el año 2008 hasta el 2013. Con un aproximado de 9000 registros en notas ha sido la base de datos obtenida y administrada con el Gestor de Base de datos Mysql y manipulada con el frontal libre de administración de BD DatAdmin.

La base de datos generada, tiene una estructura que contiene alrededor de 15 tablas, detalladas a continuación:

- *area:* Contiene la información relacionada a las áreas de la Universidad Nacional de Loja. La estructura de la misma está formada por los atributos: nombre_area, secretario y director.

- *carrera:* Contiene los datos de las carreras de las distintas áreas de la UNL, la cual se relaciona con la tabla modalidad y titulación de la base de datos. La estructura de la misma está formada por los atributos: nombre, especialidad y costo.
- *estudiante:* Almacena la información personal de los estudiantes. La estructura de la tabla estudiante está formada por los atributos: nombres, apellidos, cedula, fecha de nacimiento, teléfono, celular, dirección, etc.
- *estudiante_paralelo:* contiene los datos personales de los estudiantes asociados a un paralelo en particular.
- *genero:* contiene los datos respecto del género sexual de la persona, la tabla está estructurada por un atributo denominado nombre, que puede contener dos únicos valores: masculino y femenino.
- *modalidad:* contiene las 3 modalidades de estudio que ofrece la Universidad Nacional de Loja (distancia, presencial, semipresencial).
- *modulo:* contiene el registro de los módulos de las distintas carreras que oferta la Universidad Nacional de Loja.
- *modulo_oferta_academica:* contiene la relación de las ofertas académicas con sus respectivos módulos.
- *nota_unidad:* contiene el registro de las notas de los estudiantes en las distintas unidades de la malla curricular de las ofertas académicas de las distintas carreras de la Universidad Nacional de Loja.
- *oferta_academica:* contiene los datos de las ofertas académicas a partir del año 2008 de todas las carreras de la UNL, las mismas que tienen un tiempo de duración de 6 meses habitualmente.
- *oferta_carrera:* Corresponde a la relación de las ofertas académicas respecto de las carreras, lo que permite hacer referencia a los datos asociados.
- *paralelo:* contiene el registro de los paralelos asociados a los módulos y carrera correspondiente. Con una estructura de atributos como: sección, número y nombre.
- *periodo_academico:* Contiene los datos asociados a los periodos académicos a partir del año 2008, con una estructura formada por los atributos id y fecha_periodo.
- *titulación:* contiene la información de los títulos obtenidos por los estudiantes de las diferentes carreras que oferta la UNL.

- *unidad:* Contiene el registro de las unidades que conforman la malla curricular respecto de los distintos módulos ofertados.

Cada una de las tablas tiene una estructura diferente de acuerdo a los datos que almacenan y la funcionalidad que tienen dentro de la base de datos generada, a continuación su descripción (ver tabla LXV a LXXIX).

TABLA LXV

ESTRUCTURA DE LA TABLA AREA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
sigla	Siglas de del área.	varchar(50)	8	PK
nombre_area	Nombre del área.	varchar(250)		-
secretario	Nombre del Secretario de área.	varchar(250)		-
director	Nombre del Director de área.	varchar(250)	6	-

TABLA LXVI

ESTRUCTURA DE LA TABLA CARRERA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la carrera	integer(20)	142	PK
nombre	Nombre de la carrera	varchar(250)		-
especialidad	Nombre de la especialidad	varchar(250)		-
modalidad_id	Id asociado a la tabla *modalidad*	varchar(250)	3	FK
titulación_id	Id asociado a la tabla *titulacion*	varchar(250)	5	FK
area_id	Id asociado a la tabla *area*	varchar(250)	8	FK
costo	Valor del costo de la carrera	double(12,2)	500	-

TABLA LXVII

ESTRUCTURA DE LA TABLA ESTUDIANTE

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del estudiante	Integer(20)	3438	-
numeroIdentificacion	Numero de cedula	varchar(250)		PK
nombres	Nombres del estudiante	varchar(250)		-
apellidos	Apellidos del estudiante	varchar(250)		-
fecha_nacimiento	Fecha de nacimiento del estudiante	date	1 a 3438	-
telefono	Número de teléfono del estudiante	varchar(250)		-
celular	Numero de celular del estudiante	varchar(250)		-
direccion	Dirección del domicilio del estudiante	varchar(250)		-
pais	Nacionalidad del estudiante	varchar(250)		-
provincia	Provincia a la que pertenece el estudiante	varchar(250)		-
email	Correo electrónico de estudiante	varchar(250)		-
genero_id_fk	Id asociado a la tabla *genero*	varchar(250)	3438	FK
es_egresado	Referencia al estado de egresado o no del estudiante	Integer(20)	3438 (1 y 0)	-

TABLA LXVIII

ESTRUCTURA DE LA TABLA ESTUDIANTE_PARALELO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
estudiante_id	Id asociado a la tabla *estudiante*	varchar(250)	11418	FK
paralelo_id	Id asociado a la tabla *paralelo*	integer(20)		FK

TABLA LXIX

ESTRUCTURA DE LA TABLA GÉNERO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
nombre	Descripción del género (femenino, masculino)	varchar(250)	2	PK

TABLA LXX

ESTRUCTURA DE LA TABLA MODALIDAD

Atributo	Descripción	Tipo de dato	# de Registros	Keys
modalidad	Descripción del tipo de estudios (distancia, presencial, semipresencial).	varchar(250)	3	PK

TABLA LXXI

ESTRUCTURA DE LA TABLA MODULO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del modulo	integer(20)	468	PK
nombre	Nombre que diferencia al modulo	varchar(250)		-

TABLA LXXII

ESTRUCTURA DE LA TABLA MODULO_OFERTA_ACADEMICA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la tabla *modulo_oferta_academica*	varchar(250)	468	PK
modulo_id_fk	Id asociado a la tabla *modulo*	integer(20)		FK
oferta_carr_id	Id asociado a la tabla *carrera*	integer(20)		FK

TABLA LXXIII

ESTRUCTURA DE LA TABLA NOTA_UNIDAD

Atributo	Descripción	Tipo de dato	# de Registros	Keys
unidad_id	Identificador de la nota	integer(20)	48487	PK
nota	Valor de la nota de la unidad	double(12,2)		-
estudiante_id_fk	Id asociado a la tabla *estudiante*	varchar(250)		FK
oferta_carrera_id_fk	Id asociado a la tabla oferta_*carrera*	integer(20)		FK

TABLA LXXIV

ESTRUCTURA DE LA TABLA OFERTA_ACADEMICA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
Id	Identificador de la nota	integer(20)	49	PK
Nombre	Valor de la nota de la unidad	varchar(250)		-
fecha_inicio	Id asociado a la tabla *estudiante*	date		FK
fecha_fin	Id asociado a la tabla oferta_*carrera*	date		FK

TABLA LXXV

ESTRUCTURA DE LA TABLA OFERTA_CARRERA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la oferta académica	integer(20)	1283	PK
oferta_id_fk	Id asociado a la tabla *oferta_academica*	integer(20)		FK
carrera_id_fk	Id asociado a la tabla *carrera*	integer(20)		FK

TABLA LXXVI

ESTRUCTURA DE LA TABLA PARALELO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del paralelo	integer(20)	12695	PK
seccion	Nombre de la carrera	varchar(250)		-

numero	Número del paralelo de acuerdo a los tenga cada modulo	varchar(250)		-
nombre	Nombre del paralelo	varchar(250)		-
modulo_id_fk	Id asociado a la tabla *modulo*	integer(20)		FK
oferta_carrera_id_fk	Id asociado a la tabla *oferta_carrera*	integer(20)		FK

TABLA LXXVII

ESTRUCTURA DE LA TABLA PERIODO_ACADEMICO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del periodo académico.	integer(20)	11	PK
fecha_periodo	Fecha del periodo académico en la forma (año de inicio - año fin)	varchar(250)		-

TABLA LXXVIII

ESTRUCTURA DE LA TABLA TITULACION

Atributo	Descripción	Tipo de dato	# de Registros	Keys
titulacion	Descripción del título obtenido por el estudiante.	varchar(250)	5	PK

TABLA LXXIX

ESTRUCTURA DE LA TABLA UNIDAD

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador y nombre de la unidad	varchar(250)	567	PK
modulo_id_fk	Id asociado al módulo al que pertenece la unidad	integer(20)		FK

Tomando en cuenta la información que contienen las tablas antes mencionadas, se ha realizado un análisis determinando cuales de ellas y que datos de los almacenados son los necesarios para el proceso de minería, por lo cual se han escogido las siguientes tablas con los datos estrictamente necesarios:

- *periodo_academico:* contiene el registro de los periodos académicos desde el año 2008, donde cada periodo académico tiene un tiempo de duración de un año, cuya estructura de esta tabla consta de los atributos id y fecha_periodo (ver tabla LXXX).

TABLA LXXX

PERIODOS ACADÉMICOS QUE ALMACENA LA TABLA PERIODO_ACADEMICO.

id	fecha_periodo
1	2008 – 2009
2	2009 – 2010
3	2010 – 2011
4	2011 – 2012
5	2012 - 2013

- *oferta_academica:* contiene el registro de las ofertas que se han generado desde el año 2008, donde cada oferta tiene un tiempo de duración de 6 meses generalmente, cuya estructura de esta tabla consta de los atributos id, nombre, fecha-inicio, fecha_fin (ver tabla LXXXI).

TABLA LXXXI

OFERTAS ACADÉMICAS QUE ALMACENA LA TABLA OFERTA_ACADEMICA

id	nombre	fecha_inicio	fecha_fin
1	Septiembre 2008 – Febrero 2009	Septiembre 2008	Febrero 2009
2	Marzo 2009 – Julio 2009	Marzo 2009	Julio 2009
3	Septiembre 2009 – Febrero 2010	Septiembre 2009	Febrero 2010
4	Marzo 2010 – Julio 2010	Marzo 2010	Julio 2010
5	Septiembre 2010 – Febrero 2011	Septiembre 2010	Febrero 2011
6	Marzo 2011 – Julio 2011	Marzo 2011	Julio 2011
7	Septiembre 2011 – Febrero 2012	Septiembre 2011	Febrero 2012
8	Marzo 2012 – Julio 2012	Marzo 2012	Julio 2012
9	Septiembre 2012 – Febrero 2013	Septiembre 2012	Febrero 2013
10	Marzo 2013 – Julio 2013	Marzo 2013	Julio 2013

Cabe mencionar que existen 5 periodos académicos asociados a sus ofertas académicas correspondientes, donde a cada periodo académico le corresponden dos ofertas académicas respectivamente (ver tabla LXXXII).

TABLA LXXXII

PERIODOS ACADÉMICOS Y OFERTAS ACADÉMICAS CORRESPONDIENTES

Nro.	Periodo Académico	Oferta Académica
1	2008 – 2009	Septiembre 2008 – Febrero 2009
		Marzo 2009 – Julio 2009
2	2009 – 2010	Septiembre 2009 – Febrero 2010
		Marzo 2010 – Julio 2010
3	2010 – 2011	Septiembre 2010 – Febrero 2011
		Marzo 2011 – Julio 2011
4	2011 – 2012	Septiembre 2011 – Febrero 2012
		Marzo 2012 – Julio 2012
5	2012 – 2013	Septiembre 2012 – Febrero 2013
		Marzo 2013 – Julio 2013

- *modulo:* contiene el registro de los módulos de las distintas carreras que ofrece la Universidad Nacional de Loja. La estructura de la tabla está formada por los atributos id y nombre (ver tabla LXXXIII).

TABLA LXXXIII

MÓDULOS QUE ALMACENA LA TABLA MODULO EN LA BASE DE DATOS

id	nombre
1	FUNDAMENTACION CIENTIFICA DE LA INGENIERIA EN SISTEMAS
2	FUNDAMENTOS CIENTIFICO - TECNICOS DE LA INGENIERIA DEL SOFTWARE
3	FORMACION DEL PROGRAMADOR PARA LA CONSTRUCCION DE APLICACIONES ESPECÍFICAS.
4	FORMACION DE ANALISTA DE SISTEMAS
5	DESARROLLO DE SISTEMAS INFORMATICOS
6	GESTION DE REDES Y CENTROS DE COMPUTO
7	DESARROLLO DE COMPONENTES Y MODELOS PARA SOFTWARE BASE
8	SISTEMAS INTELIGENTES Y AUTOMATIZADOS
9	DESARROLLO DE PROYECTOS

La tabla LXXXIII como podemos observar contiene 9 módulos correspondientes a los módulos que van desde el modulo tres hasta el módulo once; ordenados del id = 1 al id = 9, relacionados a los estudiantes egresados de la carrera de ingeniería en sistemas, registrados en el SGA desde el año 2008.

- *unidad:* contiene el registro de las unidades, que corresponde a la malla curricular de la carrera de ingeniería en sistemas con todas sus variaciones; a continuación se presenta una tabla de las unidades existentes relacionándolas a cada módulo respectivamente (ver tabla LXXXIV).

TABLA LXXXIV

UNIDADES QUE CONTIENE LA TABLA UNIDAD RESPECTO A LOS DATOS RECOLECTADOS

Nombre	Modulo
MATEMATICAS, MATEMATICAS DISCRETAS, FUNDAMENTOS BASICOS DE COMPUTACION, CALCULO DIFERENCIAL, FISICA I, ALGEBRRA LINEAL.	Módulo 3
CALCULO INTEGRAL, METODOLOGIA DE LA PROGRAMACION, CONTABILIDAD GENERAL, FISICA II, ESTADISTICA, PROGRAMACION BASICA, ELECTROMAGNETISMO, ESTRUCTURA DE DATOS I	Módulo 4
PROGRAMACION AVANZADA, ESTRUCTURA DE DATOS ORIENTADA A OBJETOS, ESTADISTICA INFERENCIAL, CONTABILIDAD DE COSTOS, ELECTRONICA BASICA, DISEÑO Y GESTION DE BASE DE DATOS, TEORIA DE LOS CIRCUITOS, ECUACIONES DIFERENCIALES, PROGRAMACION II, ESTRUCTURA DE DATOS II, CONTABILIDAD GENERAL, ESTADISTICA II	Módulo 5
ECONOMIA, ADMINISTRACION DE EMPRESAS, ARQUITECTURA DE COMPUTADORES, LENGUAJE ENSAMBLADOR, DISEÑO DIGITAL, ANALISIS Y DISEÑO DE SISTEMAS, DISEÑO Y GESTION DE BASE DE DATOS, ELECTRONICA BASICA, CONTABILIDAD DE COSTOS, REDES I	Módulo 6
PROYECTOS INFORMATICOS I, TEORIA DE TELECOMUNICACIONES, DERECHO INFORMATICO, DISEÑO DE SISTEMAS, ECUACIONES DIFERENCIALES, SISTEMAS OPERATIVOS, REDES II, INGENIERIA DEL SOFTWARE, ANALISIS Y DISEÑO DE SISTEMAS II, INVESTIGACION DE OPERACIONES, TEORIA DE AUTOMATAS, INTELIGENCIA ARTIFICIAL	Módulo 7

PROYECTOS INFORMATICOS II, ANALISIS NUMERICO, ADMINISTRACION DE CENTROS DE COMPUTO, AUDITORIA INFORMATICA, INVESTIGACION DE OPERACIONES, GESTION DE REDES, SISTEMAS OPERATIVOS, ARQUITECTURA DE COMPUTADORES, INTELIGENCIA ARTIFICIAL, LENGUAJE ENSAMBLADOR, MICROPROCESADORES	Módulo 8
SISTEMAS DE INFORMACION I, LENGUAJES FORMALES, MODELAMIENTO MATEMATICO, INGENIERIA DEL SOFTWARE, COMPILADORES, SISTEMAS DE INFORMACION II, SISTEMAS EXPERTOS, MANTENIMIENTO DE COMPUTADORES, CONTROL AUTOMATIZADO ASISTIDO POR COMPUTADORES, ADMINISTRACION DE CENTROS DE COMPUTO, SISTEMAS DE INFORMACION I	Módulo 9
INTELIGENCIA ARTIFICIAL, CONTROL AUTOMATIZADO ASISTIDO POR COMPUTADORES, ANTEPROYECTOS DE TESIS, SISTEMAS EXPERTOS, SIMULACION, ETICA PROFESIONAL, SISTEMAS DE INFORMACION II, LEGISLACION LABORAL, PRESUPUESTOS E INVERSIONES, PROYECTOS I, AUDITORIA INFORMATICA, CAD	Módulo 10
APLICACIONES WEB, ANTEPROYECTOS, PROYECTOS INFORMATICOS, ADMINISTRACION BASES DE DATOS SQL SERVER, APLICACIONES MYSQL Y UML, PROGRAMACIONES .NET	Módulo 11

- *estudiante:* corresponde a la tabla de los datos personales del estudiante, donde su estructura contiene los atributos: nombres, apellidos, cedula, fecha de nacimiento, teléfono, celular, dirección, etc. De estos atributos se ha tomado nombres, apellidos y cedula respecto a los datos almacenados.

- *nota_unidad:* contiene el registro de las notas de cada estudiante por cada unidad dictada. Por ello se ha considerado tomar los 9089 registros en notas que corresponden a los estudiantes egresados de la carrera de Ingeniería en Sistemas objeto del proceso de minería de datos.

Para completar está información académica recolectada de todos los egresados de la carrera de Ingeniería en Sistemas registrados en el SGA a partir del año 2008, ha sido necesario recopilar los datos históricos anteriores a este año, donde examinamos algunas

propiedades, detalles, características de los datos que se han aumentado a esta base de datos de forma manual.

- Los libros físicos pertenecientes a la carrera de ingeniería en sistemas, contienen aproximadamente 20000 registros en notas de los estudiantes a partir del año 1999 en el que se creó la carrera. A partir de ello se ha realizado un proceso de filtrado de las notas, considerando puntualmente las de los egresados registrados en SGA, con el fin de completar su información académica con las notas de ciertos módulos que no constan en este sistema. A partir del proceso de búsqueda y digitación manual de los las notas de todas las unidades de los módulos respectivos se obtuvieron un total de 3500 datos, que fueron trasladados directamente a la base de datos que contiene 9089 registros en notas, completando un total de 12476 datos para realizar el proceso de minería aproximadamente.

- Los datos registrados en los libros físicos se encontraban desorganizados, además existía inconsistencia en los mismos, debido a duplicidad en las notas en los distintos módulos, las variaciones de la malla curricular en los diferentes periodos académicos y los estudiantes registrados en sus diferentes estados de reprobado, aprobado y retirado en módulos específicos.

Se ha diseñado el modelo de la base de datos final, estructurada con las tablas estrictamente necesarias para el proceso de minería de datos (ver Figura 36):

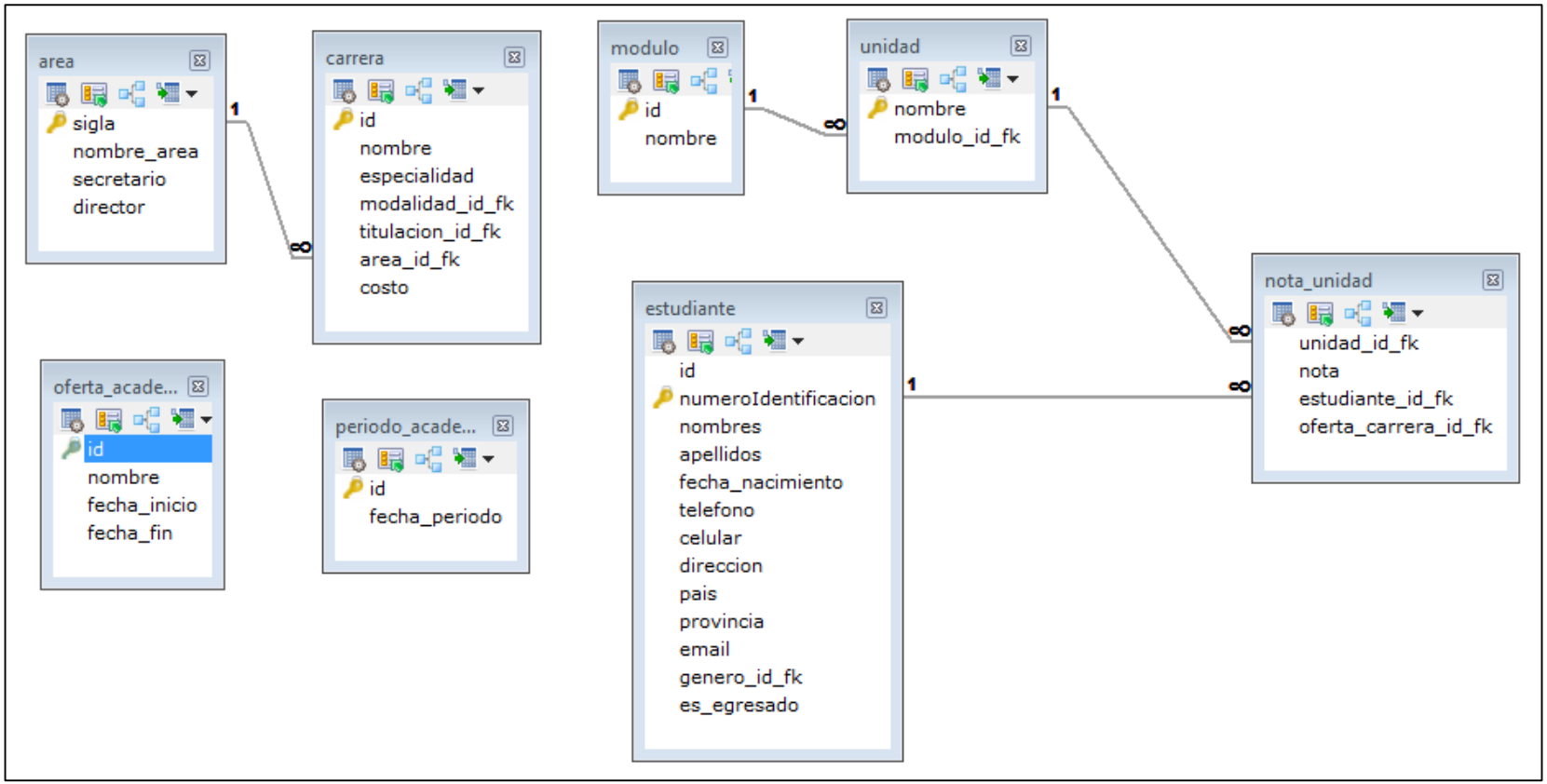


Figura 36. Modelo de la Base de Datos generada

Los datos recolectados del test de habilidades, capacidades e intereses desarrollado en el framework django bajo el lenguaje de programación python, se almacenan en una base de datos propia estructurada durante el desarrollo de la aplicación, esta base de datos generada, contiene las siguientes tablas:

- *app_caracteristica:* contiene los ítems o características de respuesta correspondientes a cada perfil.
- *app_caracteristica_perfil:* contiene la relación respecto de las características y el perfil de pertenencia.
- *app_catalogo_caracteristica:* contiene los catálogos o categorías de las características (habilidad, capacidad, interés).
- *app_contestacion:* contiene los datos de los test contestados: las preguntas, las respuestas es decir las características,
- *app_cuestionario:* contiene los cuestionarios creados
- *app_estudiante:* contiene los estudiantes a los que va dirigido el test
- *app_estudiante_temp:* contiene los temporalmente los estudiantes hasta configurarlos como usuarios de la aplicación.
- *app_item_pregunta:* contiene las características y su relación con las preguntas.
- *app_perfil: contiene los perfiles creados.*
- *app_periodo_actual:* contiene el periodo en que el o los test están activos.
- *app_periodo_test:* contiene los registros de los periodos de los test, su fecha_incio y su fecha_fin.
- *app_pregunta:* contiene las preguntas del cuestionario.
- *app_seccion:* contiene las secciones de preguntas que tiene cada cuestionario
- *app_test:* contiene el registro de los test creados.
- *app_tipo_pregunta:* contiene los tipos de las preguntas del cuestionario (ejm: contestación múltiple).
- *app_usuario:* contiene los estudiantes que tiene el rol de usuarios.
- *auth_group:* contiene el registro de los grupos de usuarios (ejm: egresados).
- *auth_group_permissions:* contiene el registro de los grupos de la aplicación con sus respectivos permisos.
- *auth_permission:* contiene el registro *de los* permisos de autentificación

- *auth_user:* contiene los datos de la autentificación del usuario; como: contraseña encriptado, nombre de usuario, nombres apellidos, fecha en que se autentifico, si es usuario administrador, etc.
- *auth_user_groups:* contiene la relación del usuario con el grupo correspondiente.
- *auth_user_user_permissions:* contiene la relación del usuario y los permisos correspondientes.
- *django_admin_log:* contiene el registro de la autentificación de los usuarios, con los objetos manipulados y otras especificaciones.
- *django_content_type:* contiene los registros de las autentificaciones, la historia, los objetos añadidos o modificados a través de la interfaz de administración.
- *django_session:* contiene los datos de las sesiones anónimas, es decir se almacena y recupera datos arbitrarios en función de cada visitante del sitio.

Cada una de las tablas, tiene una estructura diferente de acuerdo a la funcionalidad que tienen dentro de la base de datos de la aplicación django, a continuación su descripción (ver tabla LXXXV a CIX).

TABLA LXXXV

ESTRUCTURA DE LA TABLA APP_CARACTERISTICA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador la característica	int(11)	40	PK
nombre	Nombre de la característica	varchar(250)		-
catalogo_caracteristica_fk_id	Identificador del catálogo de la característica	int(11)		FK

TABLA LXXXVI

ESTRUCTURA DE LA TABLA APP_CARACTERISTICA_PERFIL

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la correspondencia del perfil con sus características	int(11)	40	PK
perfil_id_fk_id	Identificador perfil	Int(11)		FK
caracteristica_id_fk_id	Identificado de la característica	int(11)		FK

TABLA LXXXVII

ESTRUCTURA DE LA TABLA APP_CATALOGO_CARACTERISTICA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de los catálogos de las características	int(11)	3	PK
nombre	Nombre del catálogo	varchar(250)		-

TABLA LXXXVIII

ESTRUCTURA DE LA TABLA APP_CONTESTACION

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador contestaciones del test	int(11)	3622	PK
pregunta_id	Identificador pregunta correspondiente (1 o 2)	Int(11)		-
respuesta_id	Identificador de la característica escogida	int(11)		-
test_id	Identificador del número de test contestado	int(11)		-

TABLA LXXXIX

ESTRUCTURA DE LA TABLA APP_CUESTIONARIO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del cuestionario	int(11)	1	PK
titulo	Título principal del cuestionario creado	varchar(250)		-
nombre	Nombre que describe al cuestionario	varchar(250)		-
periodo_test_id	Identificador del periodo del cuestionario creado	int(11)		-

TABLA XC

ESTRUCTURA DE LA TABLA APP_ESTUDIANTE

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del estudiante	int(11)	260	PK
apellidos	Apellidos del estudiante	varchar(250)		-
nombres	Nombres del estudiante	varchar(250)		-
numero_identificacion	Cedula del estudiante	int(11)		-
fecha_nacimiento	Fecha de nacimiento del estudiante	date		-
email	Dirección de correo del estudiante	varchar(250)	249	-
genero	Género del estudiante (masculino/femenino)	varchar(250)	260	-
estado_civil	Situación civil del estudiante	varchar(250)		-
pais	País donde nació el estudiante	varchar(250)		-
provincia	Provincia correspondiente al país	varchar(250)	256	-
ciudad	Ciudad respecto de la provincia	varchar(250)	254	-
direccion	Lugar de domicilio del estudiante	varchar(250)	255	-
telefono	Teléfono convencional	varchar(250)	199	-
celular	Teléfono celular	varchar(250)	250	-

TABLA XCI

ESTRUCTURA DE LA TABLA APP_ESTUDIANTE_TEMP

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del estudiante	int(11)	260	PK
apellidos	Apellidos del estudiante	varchar(250)		-
nombres	Nombres del estudiante	varchar(250)		-
cedula	Número de identificación del estudiante	varchar(250)		-

TABLA XCII

ESTRUCTURA DE LA TABLA APP_ITEM_PREGUNTA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del número de ítems formador de acuerdo a las características y la pregunta	int(11)	40	PK
caracteristica_id	Identificador de la característica y su relación cierta pregunta	int(11)		-
texto	Texto de la pregunta	varchar(250)		-
pregunta_id	Identificador de la pregunta	int(11)		-

TABLA XCIII

ESTRUCTURA DE LA TABLA APP_PERFIL

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del perfil	int(11)	8	PK
nombre	Nombre del perfil	varchar(250)		-
sigla	Sigla que describe el perfil	varchar(250)		-
descripcion	Concepto del perfil	lomgtext		-

TABLA XCIV

ESTRUCTURA DE LA TABLA APP_PERIODO_ACTUAL

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de los periodos actuales de los test	int(11)	1	PK
periodo_test_id	Identificador que relaciona al periodo de contestación con el test en especifico	Int(11)		-

TABLA XCV

ESTRUCTURA DE LA TABLA APP_PERIODO_TEST

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del periodo	int(11)	1	PK
nombre	Nombre respecto al test	varchar(250)		-
descripcion	Descripción como la finalidad del test	longtext		-
fecha_inicio	Fecha de habilitación del test	datetime		-
fecha_fin	Fecha de culminación y cierre automático del test	datetime		

TABLA XCVI

ESTRUCTURA DE LA TABLA APP_PREGUNTA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la pregunta	int(11)	2	PK
texto	Descripción de la pregunta	varchar(250)		-
orden	Numero de pregunta	int(11)		-
tipo_pregunta_id_fk_id	Identificador de de la pregunta respecto del test	int(11)		FK
seccion_id	Identificador de la sección a la que pertenece	int(11)		-

TABLA XCVII

ESTRUCTURA DE LA TABLA APP_SECCION

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la sección	int(11)	2	PK
titulo	Título de la sección	varchar(250)		-
descripcion	Descripción de la sección	varchar(250)		-
cuestionario_id	Identificador del cuestionario al que pertenece la sección	int(11)		-

TABLA XCVIII

ESTRUCTURA DE LA TABLA APP_TEST

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del test	int(11)	208	PK
nombre	Nombre del test	varchar(250)		-
fecha	Fecha de contestación del test	date		-
hora_inicio	Hora en que se inició la contestación de test.	time		-
hora_fin	Hora de finalización de la contestación del test.	time		-
estudiante_id	Identificador del estudiante de acuerdo al test contestado	int(11)		-
cuestionario_id	Identificador del cuestionario en caso de haber más de uno	int(11)		-

TABLA XCIX

ESTRUCTURA DE LA TABLA APP_TIPO_PREGUNTA

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la pregunta	int(11)	1	PK
nombre	Nombre del tipo de pregunta	varchar(250)		-
descripcion	Descripción de la pregunta	varchar(250)		-

TABLA C

ESTRUCTURA DE LA TABLA APP_USUARIO

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del estudiante con el rol de usuario	int(11)	260	PK
estudiante_id	Identificador del estudiante	int(11)		-
user_id	Identificador en relación al registro de permisos del usuario	int(11)		-

TABLA CI

ESTRUCTURA DE LA TABLA AUTH_GROUP

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del grupo	int(11)	1	PK
name	Nombre del grupo de usuarios (ejmp: egresados)	varchar(80)		-

TABLA CII

ESTRUCTURA DE LA TABLA AUTH_GROUP_PERMISSIONS

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la tabla	int(11)	NULL	PK
group_id	Identificador del grupo al que se relaciona	int(11)		-
permission_id	Identificador de los permisos	int(11)		-

TABLA CIII

ESTRUCTURA DE LA TABLA AUTH_PERMISSION

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del permiso de autentificación	int(11)	63	PK
Name	Nombre del permiso de autentificación	varchar(50)		-
content_type_id	Identificador del tipo de contenido	int(11)		-
codename	Nombre del código de permiso (ejm: delete_user)	varchar(100)		-

TABLA CIV

ESTRUCTURA DE LA TABLA AUTH_USER

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la autentificación del usuario	int(11)	261	PK
password	Contraseña encriptado del usuario	varchar(128)		-
last_login	Fecha en que se autentifico	tinyint(1)		-
is_superuser	Describe si el usuario tiene permisos de super usuario (1 y 0)	varchar(30)		-
username	Nombre que identifica al usuario para la autentificación	varchar(30)		-
first_name	Nombres del usuario	varchar(30)		-
last_name	Apellidos del usuario	varchar(30)		-
email	Correo electrónico del usuario	varchar(75)		-
is_staff	Indica si este usuario puede tener acceso al sitio de administración (1/0)	tinyint(1)		-
is_active	Indica si el usuario está en estado activo (1/0)	tinyint(1)		-
date_joined	Fecha y hora de creación de la cuenta	datetime		-

TABLA CV

ESTRUCTURA DE LA TABLA AUTH_USER_GROUPS

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la tabla de la relación de los usuarios y el grupo	int(11)	260	PK
user_id	Identificador del usuario	int(11)		-
group_id	Identificador del grupo al que pertenece el usuario	int(11)		-

TABLA CVI

ESTRUCTURA DE LA TABLA AUTH_USER_USER_PERMISSIONS

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador de la relación de los usuarios con los respectivos permisos	int(11)	NULL	PK
user_id	Identificador del usuario	int(11)		-
permission_id	Identificador del tipo de permiso	int(11)		-

TABLA CVII

ESTRUCTURA DE LA TABLA django_admin_log

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del registro de autentificación de los usuarios	int(11)	185	PK
action_time	Tiempo en que se ha realizado la acción	datetime		-
user_id	Identificador del usuario correspondiente	int(11)		-
content_type_id	Identificador del tipo de contenido o acción realizada	int(11)		-
object_id	Identificador del objeto que ha sido manipulado	longtext		-
object_repr	Descripción o nombre del objeto manipulado	varchar(200)		-
action_flag	Número de la acción realizada	smallint(5)		-
change_message	Mensaje del cambio realizado	Longtext		-

TABLA CVIII

ESTRUCTURA DE LA TABLA DJANGO_CONTENT_TYPE

Atributo	Descripción	Tipo de dato	# de Registros	Keys
id	Identificador del historial del administrados	int(11)	21	PK
name	Nombre de la acción realizada	varchar(100)		-

app_label	Nombre de la parte de la aplicación afectada	varchar(100)		-
model	Nombre de la tabla del modelo afectada	varchar(100)		-

TABLA CIX

ESTRUCTURA DE LA TABLA DJANGO_SESSION

Atributo	Descripción	Tipo de dato	# de Registros	Keys
session_key	Identificador de las sesiones anónimas almacenadas	varchar(40)	108	PK
session_data	Código encriptado de la sesión y acción realizada	longtext		-
expire_date	Fecha y hora en que termino la sesión	datetime		-

Se ha diseñado el modelo de la base de datos final de la aplicación del test desarrollado en la herramienta django, que muestra la estructurada con todas las tablas que ayudan a la correcta funcionalidad del test (ver figura 37):

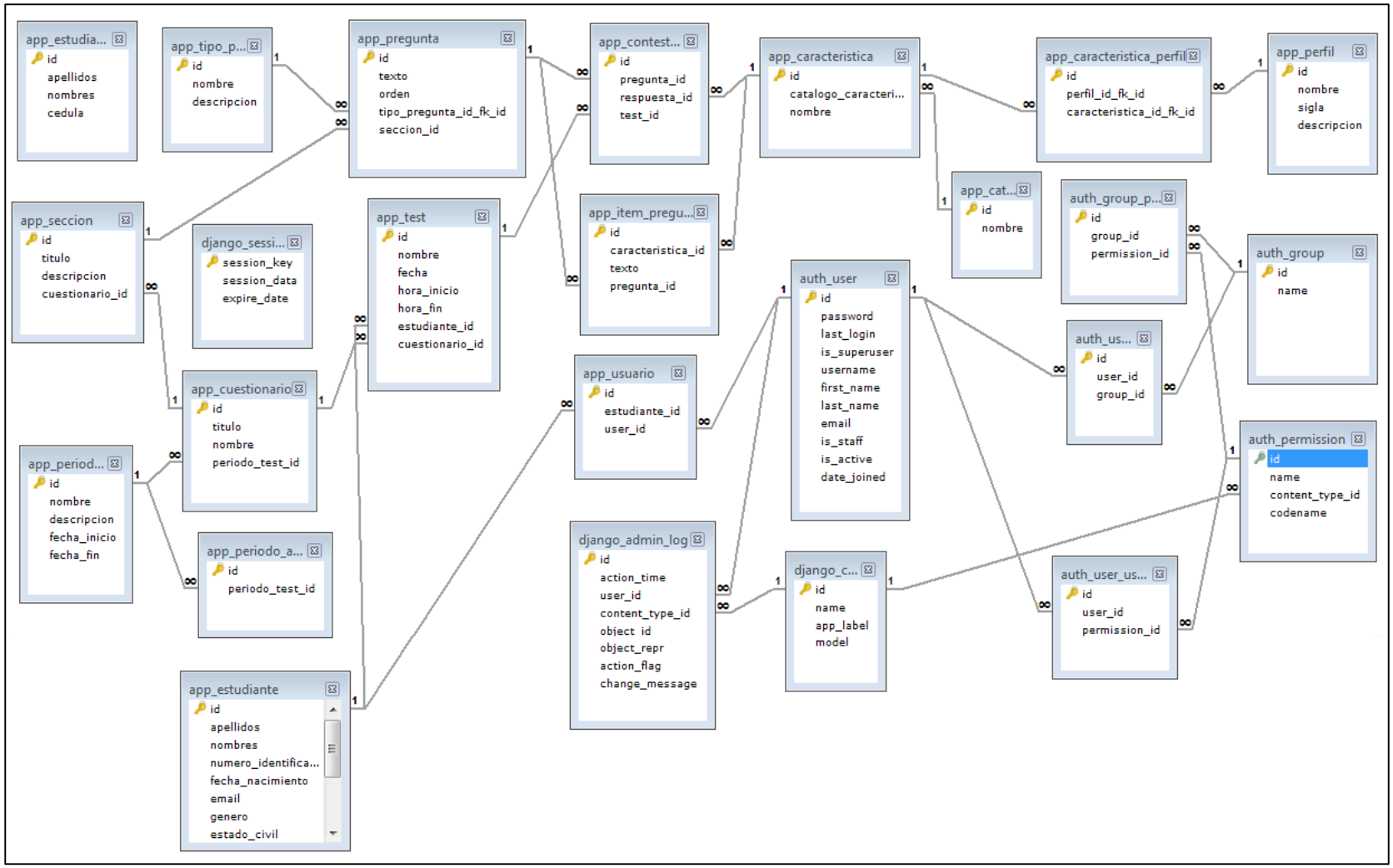


Figura 37. Modelo de la base de datos (Test django).

3.2.3. Tarea Tres: Explorar los datos

En la siguiente tarea se realiza la exploración de datos con el objetivo de relacionarse a profundidad con los mismos. Los datos son descritos para su mejor entendimiento a través de gráficos especificando sus características.

3.2.3.1. Actividad 1: Reporte con la exploración de los datos

Como se ha mencionado en el transcurso del desarrollo del presente trabajo de titulación, se ha realizado un test enfocado a los egresados y graduados con objeto de recabar datos de suma importancia. Para llamar la atención a la población encuestada se ha hecho un gran esfuerzo en la difusión a través de redes sociales profesionales e informales, grupos de graduados, una publicación en la página de la carrera, etc. Lo que ha dado como resultado de test contestados en un 80% (ver figura 38), llegando a la conclusión que la difusión ha sido todo un éxito.

TABLA CX

RESULTADOS DE LA DIFUSIÓN DEL TEST PERFIL PROFESIONAL

Test Perfil	Nro.
Contestados	208
No contestados	52
TOTAL	260

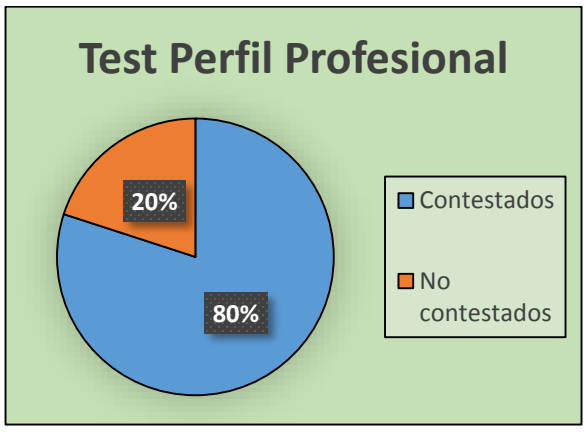


Figura 38. Resultados Difusión del Test Perfil Profesional.

La finalidad del Test aplicado es que en base a las características escogidas por cada participante que corresponden a las capacidades, habilidades e intereses propios de cada uno de los ocho perfiles planteados, se determina el perfil profesional predominante, para posteriormente tomarlo como nuestra variable dependiente dentro del proceso de minería de datos. A continuación se muestra los resultados del test aplicado, por cada perfil profesional (ver tabla CXI).

TABLA CXI

RESULTADOS DEL TEST PERFIL PROFESIONAL

Nomenclatura	Perfil Profesional	Cantidad
AS	Analista de Sistemas de Información	27
ADS	Arquitecto y Diseñador de Software.	30
DS	Desarrollador de software	25
DBA	Administrador de Sistemas de Bases de Datos	30
AI	Auditor Informático	18
ACC	Administrador de Centros de computo	19
AR	Administrador de Redes computacionales.	38
MHS	Especialista en mantenimiento hardware y software.	21
TOTAL		208

En la figura 39 se observa de manera gráfica los resultados del test aplicado, donde el perfil profesional predominante entre los egresados y graduados encuestados es el perfil de 'Administrador de Redes computacionales' con un total de 38 perfiles de este tipo, seguido por los perfiles 'Administrador de Sistemas de Bases de Datos' y Arquitecto y Diseñador de Software empatados con la cantidad de 30, seguidos de 'Analista de Sistemas de Información' con 27, 'Desarrollador de software' con 25, 'Especialista en mantenimiento hardware y software' con 21, 'Administrador de Centros de cómputo' con 19 y finalmente 'Auditor Informático' con una totalidad de 18.

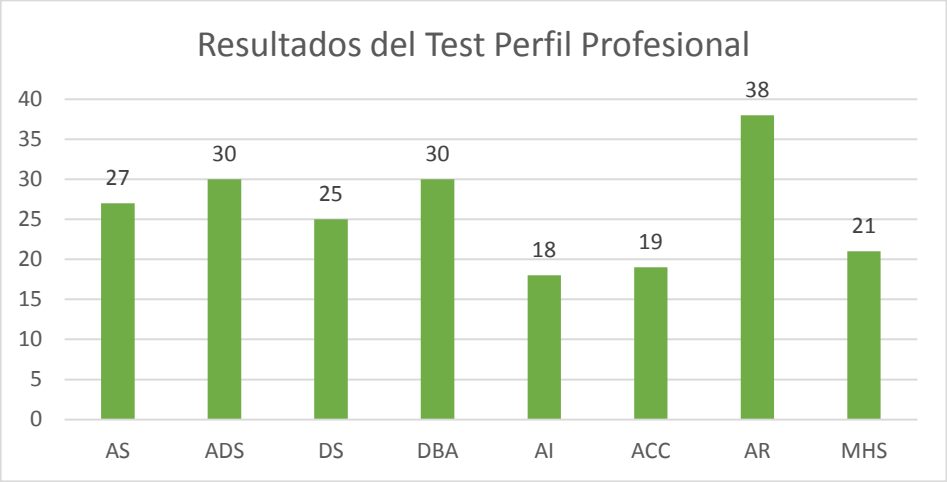


Figura 39. Diagrama de Barras de los Resultados del Test Perfil Profesional.

En la figura 40 se han representado los mismos resultados pero de manera porcentual, donde podemos observar que no existe una gran diferencia entre los resultados de un perfil profesional con otro, debido a que el porcentaje más elevado es el 18% correspondiente al perfil 'Administrador de Redes computacionales''.

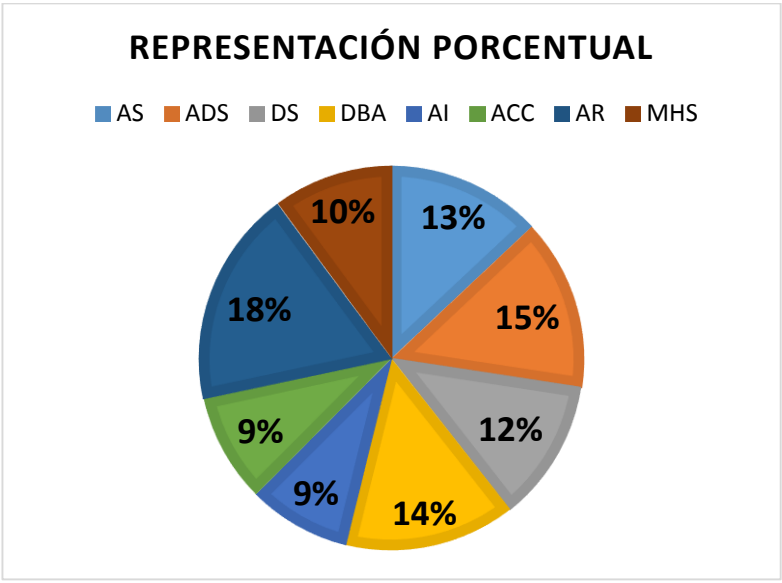


Figura 40. Gráfica de la representación porcentual de los resultados.

Tomando en cuenta la población que ha dado contestación al test aplicado se han determinado algunas estadísticas importantes. Al realizar el análisis de estos datos se ha

determinado cuales de los estudiantes matriculados en los periodos académicos respectivos han culminado sus estudios con éxito (ver tabla CXII).

TABLA CXII

ALUMNOS MATRICULADOS QUE CULMINARON SUS ESTUDIOS

PERIODOS ACADEMICOS	MATRICULADOS	EGRESADOS
2003-2004	129	22
2004-2005	151	48
2005-2006	115	47
2006-2007	49	33
2007-2008	153	43
2008-2009	280	15
TOTAL		208

Para tener una visión más clara del índice de egresados respecto de los matriculados en cada periodo académico se ha representado los datos estadísticos en un diagrama de barras que muestra los resultados más claramente (ver figura 41).

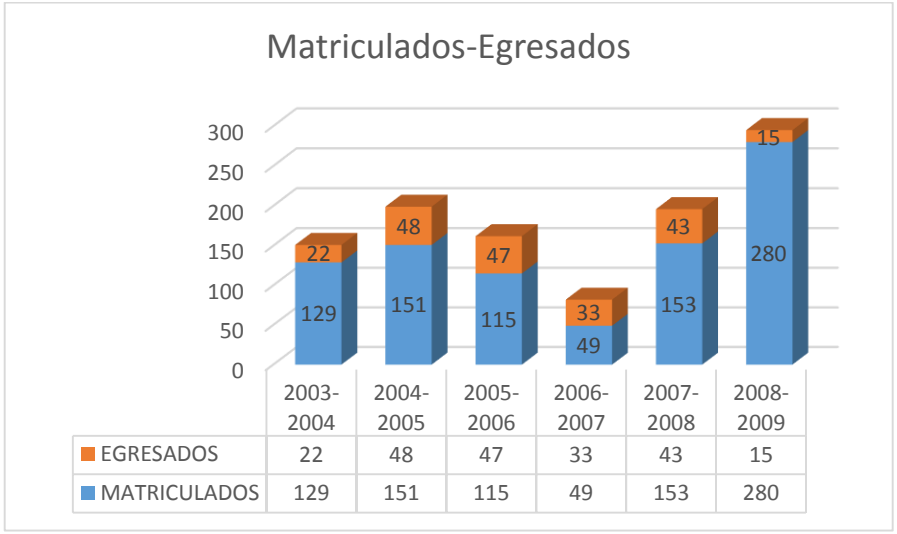


Figura 41. Matriculados que han egresado en los diferentes periodos académicos.

En la figura 41 se puede observar que durante todos los periodos académicos el índice de egresados respecto a los matriculados es una cantidad extremadamente pequeña. Para mayor claridad vamos a analizar los resultados en cada periodo académico:

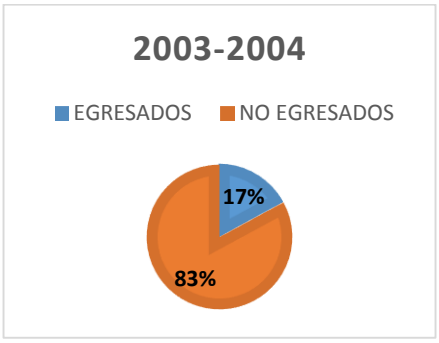


Figura 42. Egresados del Periodo académico 2003-2004.

La figura 42 muestra que el 17% de los matriculados en el periodo académico 2003-2004 culminaron sus estudios académicos, mientras el 83% restante pertenece a los estudiantes reprobados o retirados en el transcurso del desarrollo de sus estudios, dejando el registro de sus records académicos incompletos.

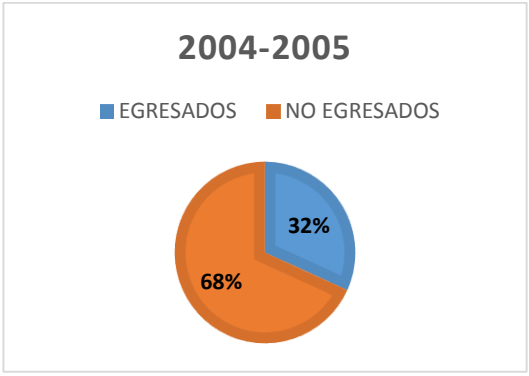


Figura 43. Egresados del Periodo académico 2004-2005.

En la figura 43 podemos observar que el 32% de los matriculados en el periodo académico 2004-2005 corresponden a los estudiantes que han egresado con o sin dificultades en el transcurso de sus estudios, mientras el 68% restante pertenece a los estudiantes que luego del proceso de matriculación sufrieron algún percance que ha sido el impedimento para la culminación de los estudios superiores en la carrera de ingeniería en sistemas.

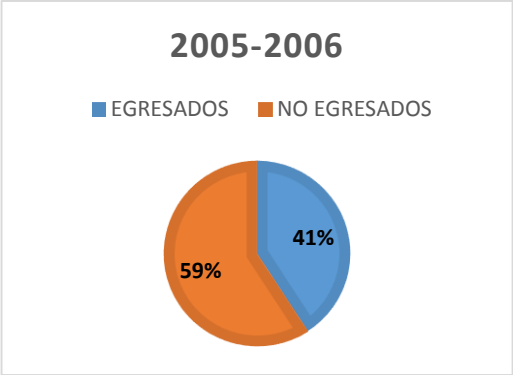


Figura 44. Egresados del Periodo académico 2005-2006.

La figura 44 corresponde al diagrama de pastel que representa la totalidad de matriculados en el periodo académico 2005-2006, de los cuales el 41% ha egresado en la carrera de ingeniería en sistemas, mientras que el 59% restante no ha terminado sus estudios con éxito.

Figura 45. Egresados del Periodo académico 2006-2007.

La figura 45 corresponde al análisis del periodo académico 2006-2007 donde del total de matriculados el 32% completaron sus estudios académicos sin embargo el 68% han tenido dificultades en el camino, dejando sus estudios incompletos.

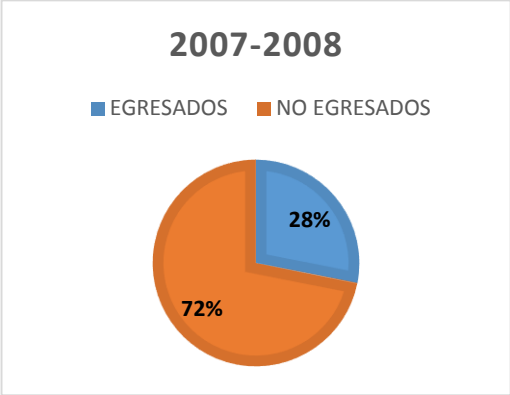


Figura 46. Egresados del Periodo académico 2007-2008.

La figura 46 representa el conjunto de los estudiantes matriculados en el periodo 2007-2008, de los cuales el 28% de los matriculados en ese periodo ha egresado con éxito y el 72% restante no termino los estudios superiores.

Figura 47. Egresados del Periodo académico 2008-2009.

La figura 47 representa los estudiantes matriculados en el periodo académico 2008-2009 en un diagrama de pastel. En el gráfico observamos que el 5%, mientras que el 95% restante han perdido o se han retirado en el transcurso de sus estudios.

De todos estos datos estadísticos se ha realizado el cálculo final del porcentaje de egresados existentes de los estudiantes matriculados en los años 2003 al 2008, debido a

que en el año 2013 son los últimos egresados que han sido objeto de análisis en el presente trabajo de titulación. A continuación la figura 46 representa estos datos estadísticos:

Figura 48. Totalidad de Egresados respecto de matriculados periodo 2003 al 2008.

Como podemos observar en la figura 48 se observa el conjunto de estudiantes matriculados entre los años 2003 al 2008, que corresponden a 930 estudiantes representando el 100%. De los cuales el 22% son egresados, mientras que el 78% no han culminado sus estudios, pudiendo concluir que son muy pocos los estudiantes que terminan sus estudios con éxito.

3.2.4. Tarea Cuatro: Verificar la calidad de los datos

Esta tarea corresponde a la verificación de la calidad o confiabilidad de la información obtenida a través de diferentes métodos, esta verificación es esencial para el desarrollo de la minería de datos ya que aumenta la calidad de los datos. No se trata de un grupo de pruebas aplicadas a los datos recolectados sino de verificaciones realizadas antes de empezar la recolección de datos y monitoreadas durante el proceso de desarrollo del presente trabajo de titulación.

Reporte de la calidad de los datos

Se han efectuado las verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y la distribución de los valores nulos y para encontrar valores fuera de rango, para evitar que estos pueden constituirse en ruido en los

datos y que intervengan en el proceso de minería de manera negativa. La idea en esta actividad ha sido asegurar la completitud y corrección de los datos.

En el transcurso de la recopilación de los datos se han presentado los siguientes problemas:

- Los datos de los egresados y graduados de la carrera de ingeniería en sistemas que se encuentran en el Sistema de Gestión Académica de la Institución, están incompletos debido a que este sistema ha sido puesto en marcha en el año 2008, mientras la carrera tiene origen desde el año 1999, por lo tanto esos datos no fueron actualizados y son considerados como datos históricos, cuyo registro se encuentra simplemente en libros físicos en poder del AEIRNNR. Posible solución: recolección de los datos históricos manualmente.
- Los datos personales de la población objeto de estudio no están completos y actualizados, lo que causa inconvenientes como retraso al realizar el proceso de difusión y respuesta del test aplicado para la recolección de la variable perfil profesional. Posible solución: apoyo en la información personal de los graduados proveniente del Sistema de Seguimiento a graduados de la CIS.
- Las mallas curriculares han sufrido un sin número de cambios con el transcurso del tiempo de la carrera de ingeniería en sistemas, incluso en la actualidad está formada por diez módulos, mientras que al inicio de la carrera estaba conformada por once módulos. Así mismo las unidades han variado con el paso de los periodos académicos sufriendo algunos cambios a pesar de pertenecer a los mismos módulos. Posible solución: tratamiento de los datos tomando en cuenta propiamente las unidades de manera apartadas a la relación con cada módulo, debido a que existen inconsistencias.
- La recolección de los datos desde los libros físicos ha sido una tarea muy extensa debido a la desorganización, cambios en la malla curricular de los registros de los egresados y graduados de cis. Posible solución: emplear diferentes métodos de búsqueda en los libros físicos, por ejemplo tomando en cuenta los periodos académicos de cada estudiante.
- La limpieza de los datos ocasiono una fuerte inversión de tiempo debido a que se ha realizado la integración de los datos recolectados a través del Web Service del SGA, con los datos históricos de los libros físicos de la cis, debido a incoherencias en los

datos, como la existencia de módulos con estado de reprobados y retirados por parte de los estudiantes, etc. Posible solución: invertir una mayor cantidad de horas diarias de trabajo, para evitar retrasos en el trabajo de titulación.

3.3. Tercera Fase: Preparación de los Datos (Selección, limpieza e integración de los datos).

Esta fase permite construir el conjunto de datos final, debido a que engloba todas las actividades esenciales para llegar a este objetivo. La preparación de los datos requiere de esfuerzo y muchas veces de repetir algunas de las tareas. Se realizan desde tareas generales como selección de datos, con la selección de tablas, registros, y atributos a los, transformación de datos, cambios de formato, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos, etc. Datos a los que se ha aplicado técnicas de minería de datos específicas, detalladas en las posteriores fases.

3.3.1. Tarea Uno: Selección de los Datos

Como se ha descrito anteriormente el estudio está enfocado en los egresados y graduados de la carrera de ingeniería en sistemas (2003-2013), formando un total de 260 estudiantes. De los cuales finalmente se cuenta con 208 estudiantes egresados y graduados en la base de datos manejada, que corresponden al 80% los cuales respondieron al test aplicado para determinar su perfil profesional. De los 208, se tomará un 72% que corresponden a 150, con el objeto de realizar el proceso de minería y determinar las reglas que expresen en relación con las unidades como se llega a un determinado perfil. El restante 28 % que corresponde a 58 estudiantes se tomará para realizar la validación de las reglas obtenidas del proceso.

Cabe recalcar que a medida que se realizó el análisis de la información se advirtió que era necesario recopilar los datos históricos anteriores al año de creación del SGA, para completar la información académica de todos los egresados o graduados registrados a partir del año 2008. Para ello ha sido necesario recurrir a los libros que contienen los registros académicos de los estudiantes de la Carrera de Ingeniería en Sistemas,

indagando de forma individual el proceso académico reflejado en la malla curricular con sus diferentes variaciones a través de los periodos académicos, realizando un seguimiento minucioso del desarrollo académico con criterio; descartando los módulos en los que el estudiante haya tenido un estado de reprobado o retirado; dando así la veracidad del caso en la información de los estudiantes en su estado como aprobado y así completar la Base de Datos de los que consiguieron culminar sus estudios universitarios y egresaron a partir del año 2008.

3.3.2. Tarea Dos: Limpieza de los datos

Durante este proceso se manipularon dos bases de datos. La principal que corresponde a la base de datos del sistema de Gestión Académica, de donde se seleccionaron los datos de los egresados y graduados de la carrera de ingeniería en sistemas, específicamente sus datos personales y el record académico de cada estudiante. Por lo que se ha creído conveniente descartar algunas de las tablas que maneja esa base de datos, que serían innecesaria respecto de los datos que vamos a utilizar. Las tablas que se ha eliminado son estudiante_paralelo, genero, modalidad, modulo, modulo_oferta_academica, oferta_carrera, paralelo, titulación, de esta manera solo quedaron las tablas necesarias.

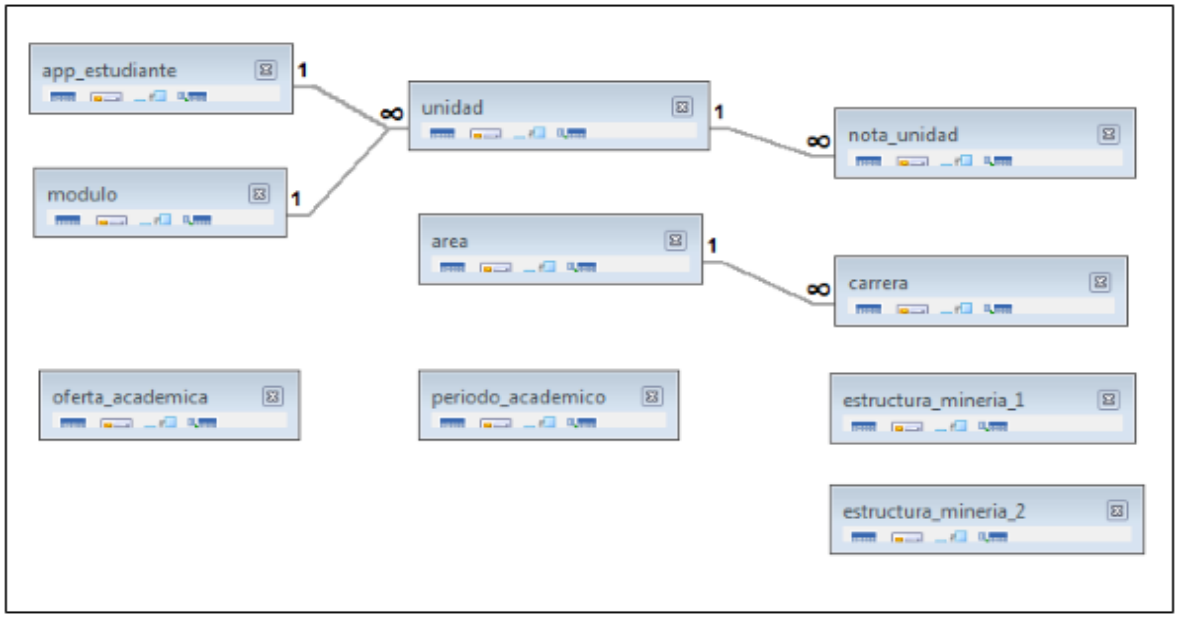


Figura 49. Estructura de la base de datos depurada.

Para obtener el perfil profesional, luego de que se ha realizado la difusión y obtención de los resultados del test; se han utilizado algunas de las tablas de la base de datos, descritas a continuación:

- *app_caracteristica:* contiene los ítems o características de respuesta correspondientes a cada perfil.
- *app_caracteristica_perfil:* contiene la relación respecto de las características y el perfil de pertenencia.
- *app_contestacion:* contiene los datos de los test contestados: las preguntas, las respuestas es decir las características,
- *app_estudiante:* contiene los estudiantes a los que va dirigido el test
- *app_item_pregunta:* contiene las características y su relación con las preguntas.
- *app_perfil: contiene los perfiles creados.*
- *app_pregunta:* contiene las preguntas del cuestionario.
- *app_test:* contiene el registro de los test creados.

Estas tablas se han utilizado para a través de funciones y métodos con consultas sql obtener el perfil profesional de los estudiantes en base a los resultados del test aplicado, a continuación se muestra el modelo entidad – relacional de las tablas detalladas anteriormente, para observar de manera gráfica su interacción.

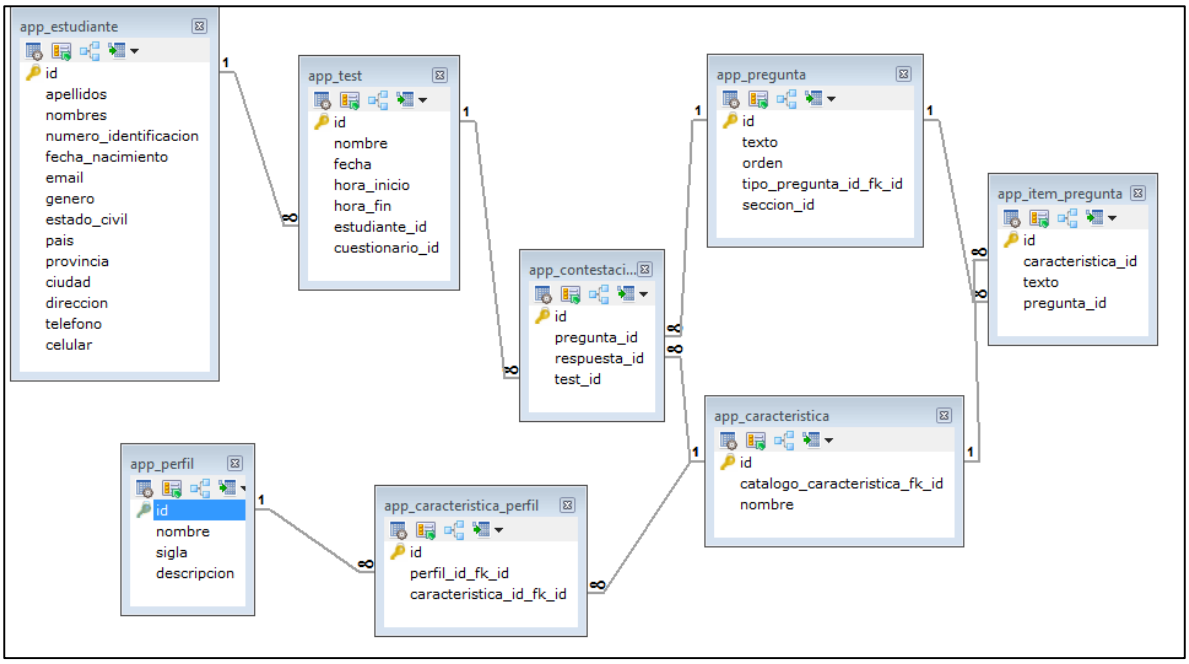


Figura 50. Tablas de la bd_test_django para determinar el perfil profesional.

3.3.3. Tarea Tres: Construcción de Datos

En esta tarea se detalla la o las estructuras de minería de datos diseñadas, considerando las fuentes de datos recopiladas y los atributos que influyen para llegar a la determinación de perfiles profesionales con el uso de técnicas de minería de datos. Esta construcción se la realiza con el fin de mejorar la calidad de los datos y aplicar algunos criterios de mejora [107]. Es por ello que en base a la meta de minería de datos y para realizar varias pruebas en búsqueda de los mejores resultados se han diseñado dos estructuras:

- **Estructura_uno: Datos no agrupados**

Esta estructura está conformada por 67 variables; 66 variables correspondientes al conjunto total de unidades que existen entre todos los datos; una gran cantidad de los registros contienen los atributos de las unidades con valores nulos debido al cambio en las mallas curriculares antes mencionado y finalmente la estructura contiene la variable dependiente

perfil_profesional obtenida del test aplicado clave para el proceso de predicción dentro de la minería de datos (ver Tabla CXIII).

TABLA CXIII

ESTRUCTURA_UNO DE MINERÍA DE DATOS PARA DETERMINAR EL PERFIL PROFESIONAL

Atributo	Tipo de datos	Tipo de contenido	Valores
Matemática	double	continuo	Regular Bueno Excelente
matematicas_discretas	double	continuo	Regular Bueno Excelente
fundamentos_basicos_computacion	double	continuo	Regular Bueno Excelente
calculo_diferencial	double	continuo	Regular Bueno Excelente
fisica_I	double	continuo	Regular Bueno Excelente
algebra_lineal	double	continuo	Regular Bueno Excelente
calculo_integral	double	continuo	Regular Bueno Excelente
metodologia_programacion	double	continuo	Regular Bueno Excelente
contabilidad_general	double	continuo	Regular Bueno Excelente
fisica_II	double	continuo	Regular Bueno Excelente
estadistica_I	double	continuo	Regular Bueno Excelente
programacion_I	double	continuo	Regular Bueno Excelente
Electromagnetismo	double	continuo	Regular Bueno Excelente
estructura_datos_I	double	continuo	Regular

			Bueno Excelente
Economía	double	continuo	Regular Bueno Excelente
estructura_datos_oo	double	continuo	Regular Bueno Excelente
estadistica_inferencial	double	continuo	Regular Bueno Excelente
contabilidad_costos	double	continuo	Regular Bueno Excelente
electronica_basica	double	continuo	Regular Bueno Excelente
diseno_gestion_bd	double	continuo	Regular Bueno Excelente
teoria_circuitos	double	continuo	Regular Bueno Excelente
ecuaciones_diferenciales	double	continuo	Regular Bueno Excelente
programacion_II	double	continuo	Regular Bueno Excelente
estructura_datos_II	double	continuo	Regular Bueno Excelente
estadistica_II	double	continuo	Regular Bueno Excelente
administracion_empresas	double	continuo	Regular Bueno Excelente
arquitectura_computadores	double	continuo	Regular Bueno Excelente
lenguaje_ensamblador	double	continuo	Regular Bueno Excelente
diseno_digital	double	continuo	Regular Bueno Excelente
analisis_diseno_sistemas_I	double	continuo	Regular Bueno Excelente
ingenieria_software_I	double	continuo	Regular Bueno

			Excelente
redes_I	double	continuo	Regular Bueno Excelente
proyectos_informaticos_I	double	continuo	Regular Bueno Excelente
teoria_telecomunicaciones	double	continuo	Regular Bueno Excelente
derecho_informatico	double	continuo	Regular Bueno Excelente
sistemas_informacion	double	continuo	Regular Bueno Excelente
analisis_diseno_sistemas_II	double	continuo	Regular Bueno Excelente
ingenieria_software_II	double	continuo	Regular Bueno Excelente
sistemas_operativos	double	continuo	Regular Bueno Excelente
redes_II	double	continuo	Regular Bueno Excelente
investigacion_operaciones	double	continuo	Regular Bueno Excelente
teoria_automatas	double	continuo	Regular Bueno Excelente
inteligencia_artificial	double	continuo	Regular Bueno Excelente
proyectos_informaticos_II	double	continuo	Regular Bueno Excelente
analisis_numerico	double	continuo	Regular Bueno Excelente
administracion_cc	double	continuo	Regular Bueno Excelente
auditoria_informatica	double	continuo	Regular Bueno Excelente
gestion_redes	double	continuo	Regular Bueno Excelente

sistemas_informacion_I	double	continuo	Regular Bueno Excelente
Microprocesadores	double	continuo	Regular Bueno Excelente
lenguajes_formales	double	continuo	Regular Bueno Excelente
modelamiento_matematico	double	continuo	Regular Bueno Excelente
Compiladores	double	continuo	Regular Bueno Excelente
sistemas_informacion_II	double	continuo	Regular Bueno Excelente
sistemas_expertos	double	continuo	Regular Bueno Excelente
mantenimiento_computadores	double	continuo	Regular Bueno Excelente
control_automatizado_asistido_c	double	continuo	Regular Bueno Excelente
anteproyectos_tesis	double	continuo	Regular Bueno Excelente
Simulación	double	continuo	Regular Bueno Excelente
etica_profesional	double	continuo	Regular Bueno Excelente
legislacion_laboral	double	continuo	Regular Bueno Excelente
presupuestos_invesiones	double	continuo	Regular Bueno Excelente
diseno_asistido_computadores	double	continuo	Regular Bueno Excelente
aplicaciones_web	double	continuo	Regular Bueno Excelente
administracion_bd_mysql_uml	double	continuo	Regular Bueno Excelente
programacion_net	double	continuo	Regular

			Bueno Excelente
perfil_profesional	varchar	discreto	AS ADS DS DBA AI ACC AR MHS

A continuación se describen los 66 atributos correspondientes a las unidades y que han sido mencionados en la Tabla CXIII, algunos de estos atributos se obtuvieron al realizar funciones con los datos de la misma base de datos, en base al análisis de lo que se pretende alcanzar:

- ✓ ***matemática:*** representa el valor de la nota de la unidad matemática.
- ✓ ***matematicas_discretas:*** representa el valor de la nota de la unidad matemáticas discretas.
- ✓ ***fundamentos_basicos_computacion:*** representa el valor de la nota de la unidad fundamentos básicos de computación.
- ✓ ***calculo_diferencial:*** representa el valor de la nota de la unidad calculo diferenicial.
- ✓ ***fisica_I:*** representa el valor de la nota de la unidad física uno.
- ✓ ***algebra_lineal:*** representa el valor de la nota de la unidad algebra lineal.
- ✓ ***calculo_integral:*** este atributo representa el valor de la nota de la unidad cálculo integral.
- ✓ ***metodologia_programacion*** este atributo representa el valor de la nota de la unidad metodología de la programación.
- ✓ ***contabilidad_general:*** este atributo representa el valor de la nota de la unidad contabilidad general.
- ✓ ***fisica_II:*** este atributo representa el valor de la nota de la unidad física dos.
- ✓ ***estadistica_I:*** este atributo representa el valor de la nota de la unidad estadística uno.
- ✓ ***programacion_I:*** este atributo representa el valor de la nota de la unidad programación básica o programación uno.
- ✓ ***electromagnetismo:*** este atributo representa el valor de la nota de la unidad electromagnetismo.

- ✓ ***estructura_datos_I:*** representa el valor de la nota de la unidad estructura de datos uno.
- ✓ ***economía:*** representa el valor de la nota de la unidad economía.
- ✓ ***estructura_datos_oo:*** representa el valor de la nota de la unidad estructura de datos orientada a objetos.
- ✓ ***estadistica_inferencial:*** representa el valor de la nota de la unidad estadística inferencial.
- ✓ ***contabilidad_costos:*** representa el valor de la nota de la unidad contabilidad de costos.
- ✓ ***electronica_basica:*** representa el valor de la nota de la unidad electrónica básica.
- ✓ ***diseno_gestion_bd:*** representa el valor de la nota de la unidad diseño y gestión de base de datos.
- ✓ ***teoria_circuitos:*** representa el valor de la nota de la unidad teoría de circuitos.
- ✓ ***ecuaciones_diferenciales:*** representa el valor de la nota de la unidad ecuaciones diferenciales.
- ✓ ***programacion_II:*** representa el valor de la nota de la unidad programación avanzada o programación dos.
- ✓ ***estructura_datos_II:*** representa el valor de la nota de la unidad estructura de datos dos.
- ✓ ***estadistica_II:*** representa el valor de la nota de la unidad estadística dos.
- ✓ ***administracion_empresas:*** representa el valor de la nota de la unidad administración de empresas.
- ✓ ***arquitectura_computadores:*** representa el valor de la nota de la unidad arquitectura de computadores.
- ✓ ***lenguaje_ensamblador:*** representa el valor de la nota de la unidad lenguaje ensamblador.
- ✓ ***diseno_digital:*** representa el valor de la nota de la unidad diseño digital.
- ✓ ***analisis_diseno_sistemas_I:*** representa el valor de la nota de la unidad análisis y diseño de sistemas uno.
- ✓ ***ingenieria_software_I:*** representa el valor de la nota de la unidad ingeniería del software uno.
- ✓ ***redes_I:*** atributo que representa el valor de la nota de la unidad redes uno.

- ✓ ***proyectos_informaticos_I:*** atributo que representa el valor de la nota de la unidad proyectos informáticos uno.
- ✓ ***teoria_telecomunicaciones:*** atributo que representa el valor de la nota de la unidad teoría de telecomunicaciones.
- ✓ ***derecho_informatico:*** atributo que representa el valor de la nota de la unidad derecho informático.
- ✓ ***sistemas_informacion:*** atributo que representa el valor de la nota de la unidad sistemas de información.
- ✓ ***analisis_diseno_sistemas_II:*** atributo que representa el valor de la nota de la unidad análisis y diseño de sistemas dos.
- ✓ ***ingenieria_software_II:*** atributo que representa el valor de la nota de la unidad ingeniería del software dos.
- ✓ ***sistemas_operativos;*** atributo que representa el valor de la nota de la unidad sistemas operativos.
- ✓ ***redes_II:*** atributo que representa el valor de la nota de la unidad redes dos.
- ✓ ***investigacion_operaciones:*** atributo que representa el valor de la nota de la unidad investigación de operaciones.
- ✓ ***teoria_automatas:*** atributo que representa el valor de la nota de la unidad teoría de autómatas.
- ✓ ***inteligencia_artificial:*** atributo que representa el valor de la nota de la unidad inteligencia artificial.
- ✓ ***proyectos_informaticos_II:*** atributo que representa el valor de la nota de la unidad proyectos informáticos dos.
- ✓ ***analisis_numerico:*** representa el valor de la nota de la unidad análisis numérico.
- ✓ ***administracion_cc:*** atributo que representa el valor de la nota de la unidad administración de centros de cómputo.
- ✓ ***auditoria_informatica:*** atributo que representa el valor de la nota de la unidad auditoría informática.
- ✓ ***gestion_redes:*** atributo que representa el valor de la nota de la unidad gestión de redes.
- ✓ ***sistemas_informacion_I:*** atributo que representa el valor de la nota de la unidad sistemas de información uno.

- ✓ ***Microprocesadores:*** atributo que representa el valor de la nota de la unidad microprocesadores.
- ✓ ***lenguajes_formales:*** atributo que representa el valor de la nota de la unidad lenguajes formales.
- ✓ ***modelamiento_matematico:*** atributo que representa el valor de la nota de la unidad modelamiento matemático.
- ✓ ***Compiladores:*** atributo que representa el valor de la nota de la unidad compiladores.
- ✓ ***sistemas_informacion_II:*** atributo que representa el valor de la nota de la unidad sistemas de información dos.
- ✓ ***sistemas_expertos:*** atributo que representa el valor de la nota de la unidad sistemas expertos.
- ✓ ***mantenimiento_computadores:*** atributo que representa el valor de la nota de la unidad mantenimiento de computadores.
- ✓ ***control_automatizado_asistido_c:*** atributo que representa el valor de la nota de la unidad control automatizado asistido por computadores.
- ✓ ***anteproyectos_tesis:*** atributo que representa el valor de la nota de la unidad anteproyectos de tesis.
- ✓ ***Simulación:*** atributo que representa el valor de la nota de la unidad simulación.
- ✓ ***etica_profesional:*** atributo que representa el valor de la nota de la unidad ética profesional.
- ✓ ***legislacion_laboral:*** atributo que representa el valor de la nota de la unidad legislación laboral.
- ✓ ***presupuestos_invesiones:*** atributo que representa el valor de la nota de la unidad presupuestos e inversiones.
- ✓ ***diseno_asistido_computadores:*** atributo que representa el valor de la nota de la unidad diseño asistido por computadores.
- ✓ ***aplicaciones_web:*** atributo que representa el valor de la nota de la unidad aplicaciones web.
- ✓ ***administracion_bd_mysql_uml:*** atributo que representa el valor de la nota de la unidad administración de base de datos mysql y uml.
- ✓ ***programacion_net:*** atributo que representa el valor de la nota de la unidad programación en .net.

- **Estructura_dos: Datos agrupados**

La estructura_dos está formada por 28 variables; 18 variables correspondientes a los grupos generados en base de 47 unidades de las 66 totales; con la finalidad de eliminar la gran cantidad de valores nulos existentes y por su relación entre sí, 9 atributos de unidades que no han sido alteradas manteniéndose de la estructura_uno y la variable dependiente denominada perfil_profesional obtenida por cada estudiante de manera personal en base al test aplicado (ver Tabla CXIV).

TABLA CXIV

ESTRUCTURA_DOS DE MINERÍA DE DATOS PARA DETERMINAR EL PERFIL PROFESIONAL

Atributo	Tipo de datos	Tipo de contenido	Valores
Matemática	double	continuo	Regular Bueno Excelente
Física	double	continuo	Regular Bueno Excelente
Calculo	double	continuo	Regular Bueno Excelente
Programación	double	continuo	Regular Bueno Excelente
estructura_datos	double	continuo	Regular Bueno Excelente
Estadística	double	continuo	Regular Bueno Excelente
presupuestos_contabilidad	double	continuo	Regular Bueno Excelente
Redes	double	continuo	Regular Bueno Excelente
proyectos_informaticos	double	continuo	Regular Bueno Excelente
sistemas_informacion	double	continuo	Regular Bueno Excelente

analisis_diseno_sistemas	double	continuo	Regular Bueno Excelente
ingenieria_software	double	continuo	Regular Bueno Excelente
arquitectura_computadores	double	continuo	Regular Bueno Excelente
electronica_telecomunicaciones	double	continuo	Regular Bueno Excelente
base_datos	double	continuo	Regular Bueno Excelente
lenguaje_ensamblador	double	continuo	Regular Bueno Excelente
Derecho	double	continuo	Regular Bueno Excelente
teoria_automatas	double	continuo	Regular Bueno Excelente
inteligencia_artificial	double	continuo	Regular Bueno Excelente
administracion_centros_computo	double	continuo	Regular Bueno Excelente
auditoria_informatica	double	continuo	Regular Bueno Excelente
lenguajes_formales	double	continuo	Regular Bueno Excelente
compiladores	double	continuo	Regular Bueno Excelente
sistemas_expertos	double	continuo	Regular Bueno Excelente
anteproyectos_tesis	double	continuo	Regular Bueno Excelente
simulación	double	continuo	Regular Bueno Excelente
etica_profesional	double	continuo	Regular Bueno Excelente
perfil_profesional	varchar	discreto	AS

			ADS DS DBA AI ACC AR MHS

A continuación se ha mencionado los 27 atributos, 18 de ellos corresponden a los grupos de las unidades afines determinados para la estructura_dos en base a las 47 de las 66 unidades originales y los 9 restantes que no han sido alterados siendo comunes con la estructura_uno.

Cada unidad ha sido identificada con su id específico de la base de datos, a su vez se ha tomado en cuenta que para obtener estos atributos se han realizado métodos directos para manipular la estructura de la base de datos, en el frontal DatAdmin escogido como herramienta para la gestión de la base de datos:

✓ ***matematica:*** cálculo realizado del promedio de las unidades afines a matemáticas.

TABLA CXV

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO MATEMATICA

id	nombre_unidad
1	MATEMATICAS
2	MATEMATICAS DISCRETAS
6	ALGEBRA LINEAL
40	INVESTIGACION DE OPERACIONES
44	ANALISIS NUMERICO
51	MODELAMIENTO MATEMATICO
56	CONTROL AUTOMATIZADO ASISTIDO POR COMPUTADORES

✓ ***física:*** cálculo realizado del promedio de las unidades afines a física.

TABLA CXVI

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO FISICA

id	nombre_unidad
5	FISICA I
10	FISICA II

- ✓ ***calculo:*** cálculo realizado del promedio de las unidades afines a cálculo.

TABLA CXVII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO CALCULO

id	nombre_unidad
4	CALCULO DIFERENCIAL
7	CALCULO INTEGRAL
22	ECUACIONES DIFERENCIALES

- ✓ ***programación:*** cálculo realizado del promedio de las unidades afines a programación.

TABLA CXVIII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO PROGRAMACION

id	nombre_unidad
12	PROGRAMACION I
23	PROGRAMACION II
66	PROGRAMACION .NET
8	METODOLOGIA DE LA PROGRAMACION
63	APLICACIONES WEB

- ✓ ***estructura_datos:*** cálculo realizado del promedio de las unidades afines a estructura de datos.

TABLA CXIX

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO ESTRUCTURA_DATOS

id	nombre_unidad
14	ESTRUCTURA DE DATOS I
16	ESTRUCTURA DE DATOS ORIENTADA A OBJETOS
24	ESTRUCTURA DE DATOS II

- ✓ ***estadística:*** cálculo realizado del promedio de las unidades afines a estadística.

TABLA CXX

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO ESTADISTICA

id	nombre_unidad
11	ESTADISTICA I
17	ESTADISTICA INFERENCIAL
25	ESTADISTICA II

- ✓ ***presupuestos_contabilidad:*** cálculo realizado del promedio de las unidades afines a contabilidad, economía y administración.

TABLA CXXI

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO PRESUPUESTOS_CONTABILIDAD

id	nombre_unidad
9	CONTABILIDAD GENERAL
18	CONTABILIDAD DE COSTOS
26	ADMINISTRACION DE EMPRESAS
61	PRESUPUESTOS E INVERSIONES
15	ECONOMIA

✓ ***redes:*** cálculo realizado del promedio de las unidades afines a redes.

TABLA CXXII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO REDES

id	nombre_unidad
32	REDES I
39	REDES II
47	GESTION DE REDES

✓ ***proyectos_informaticos:*** cálculo realizado del promedio de las unidades afines a proyectos informáticos.

TABLA CXXIII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO PROYECTOS_INFORMATICOS

id	nombre_unidad
33	PROYECTOS INFORMATICOS I
43	PROYECTOS INFORMATICOS II

✓ ***sistemas_informacion:*** cálculo realizado del promedio de las unidades afines a sistemas de información.

TABLA CXXIV

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO SISTEMAS_INFORMACION

id	nombre_unidad
48	SISTEMAS DE INFORMACION I
53	SISTEMAS DE INFORMACION II

✓ ***analisis_diseno_sistemas:*** cálculo realizado del promedio de las unidades afines con la materia análisis y diseño de sistemas.

TABLA CXXV

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO

ANALISIS_DISENO_SISTEMAS

id	nombre_unidad
30	ANALISIS Y DISEÑO DE SISTEMAS I
36	ANALISIS Y DISEÑO DE SISTEMAS II
62	CAD

- ✓ ***ingenieria_software:*** cálculo realizado del promedio de las unidades afines a ingeniería de software.

TABLA CXXVI

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO

INGENIERIA_SOFTWARE

id	nombre_unidad
31	INGENIERIA DEL SOFTWARE I
37	INGENIERIA DEL SOFTWARE II

- ✓ ***arquitectura_computadores:*** cálculo realizado del promedio de las unidades afines a mantenimiento, arquitectura de computadores y sistemas operativos.

TABLA CXXVII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO

ARQUITECTURA_COMPUTADORES

id	nombre_unidad
27	ARQUITECTURA DE COMPUTADORES
3	FUNDAMENTOS BASICOS DE COMPUTACION
38	SISTEMAS OPERATIVOS
49	MICROPROCESADORES
55	MANTENIMIENTO DE COMPUTADORES

- ✓ ***electronica_telecomunicaciones:*** cálculo realizado del promedio de las unidades afines a electrónica.

TABLA CXXVIII

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO ELECTRONICA_TELECOMUNICACIONES

Id	nombre_unidad
19	ELECTRONICA BASICA
21	TEORIA DE LOS CIRCUITOS
29	DISEÑO DIGITAL
34	TEORIA DE TELECOMUNICACIONES
13	ELECTROMAGNETISMO

- ✓ ***base_datos:*** cálculo realizado del promedio de las unidades afines a bases de datos.

TABLA CXXIX

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO BASE_DATOS

id	nombre_unidad
20	DISEÑO Y GESTION DE BASE DE DATOS
64	ADMINISTRACION BASES DE DATOS SQL SERVER
65	APLICACIONES MYSQL Y UML

- ✓ ***derecho_informatica:*** cálculo realizado del promedio de las unidades afines derecho y legislación.

TABLA CXXX

UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO DERECHO_INFORMATICA

id	nombre_unidad
35	DERECHO INFORMATICO
60	LEGISLACION LABORAL

- ✓ ***automatas_ lenguajes_formales:*** cálculo realizado del promedio de las unidades afines a teoría de autómatas y lenguajes formales.

TABLA CXXXI
UNIDADES AGRUPADAS PARA DETERMINAR EL ATRIBUTO AUTOMATAS_LENGUAJES_FORMALES

id	nombre_unidad
41	TEORÍA DE AUTÓMATAS
50	LENGUAJES FORMALES

A continuación el listado de unidades que no han sido agrupadas por contener en mínimas cantidades valores nulos y por no tener una relación fuerte con el resto de unidades para realizar una agrupación. Estos atributos son comunes con la estructura_uno, debido a que no han sufrido alteración alguna:

- ✓ ***lenguaje_ensamblador***
- ✓ ***inteligencia_artificial***
- ✓ ***administracion_centros_computo***
- ✓ ***auditoria_informatica***
- ✓ ***compiladores***
- ✓ ***sistemas_expertos***
- ✓ ***anteproyectos_tesis***
- ✓ ***simulacion***
- ✓ ***etica_profesional***

Existe un atributo en común que comparten estas dos estructuras y el más importante que corresponde a la variable tomada como dependiente denominada perfil_profesional, descrita a continuación:

- ✓ ***perfil_profesional:*** Es el atributo que representa el perfil profesional del egresado o graduado, obtenido a partir del test aplicado. El perfil profesional puede tener los valores especificados a continuación en la tabla CXXXII:

TABLA CXXXII

VALORES DEL ATRIBUTO PERFIL_PROFESIONAL

id	sigla	nombre_especialidad
1	AS	Analista de Sistemas de Información
2	ADS	Arquitecto y Diseñador de Software.
3	DS	Desarrollador de software
4	DBA	Administrador de Sistemas de Bases de Datos
5	AI	Auditor Informático
6	ACC	Administrador de Centros de computo
7	AR	Administrador de Redes computacionales.
8	MHS	Especialista en mantenimiento hardware y software.

Cabe mencionar que las notas de las unidades tienen un valor comprendido entre 0 y 10, para realizar el proceso de minería se ha visto importante realizar la discretización de estos valores descritos en la Tabla CXXXIII:

TABLA CXXXIII

DISCRETIZACIÓN DE LAS NOTAS DE CADA UNIDAD

Nro.	Nomenclatura	Rango
1	Regular	Menor a 7.5
2	Bueno	Entre 7.5 a 8.5
3	Excelente	Mayor a 8.5

3.3.4. Tarea Cuatro: Integración de datos

Esta etapa involucra la creación de nuevas estructuras a partir de los datos seleccionados, por lo que cabe mencionar que con las bases de datos manipuladas durante el proceso de las fases anteriores, se ha realizado una integración obteniendo la base de datos final denominada bd_mineria_cis (ver figura 51) y las dos estructuras de la minería de datos manejadas: la estructura_mineria_1 (ver figura 52) de los datos no agrupados y la estructura_mineria_dos (ver figura 53) de los datos agrupados que se han detallado en la tarea anterior. Las estructuras de minería de datos corresponden a la construcción propiamente de los datos, han sido creadas a través de funciones que contienen métodos generados aplicando la lógica de sentencias sql, dichas estructuras que contiene las variables necesarias para realizar el proceso de minería de datos y así poder continuar con la etapa posterior de manera exitosa.

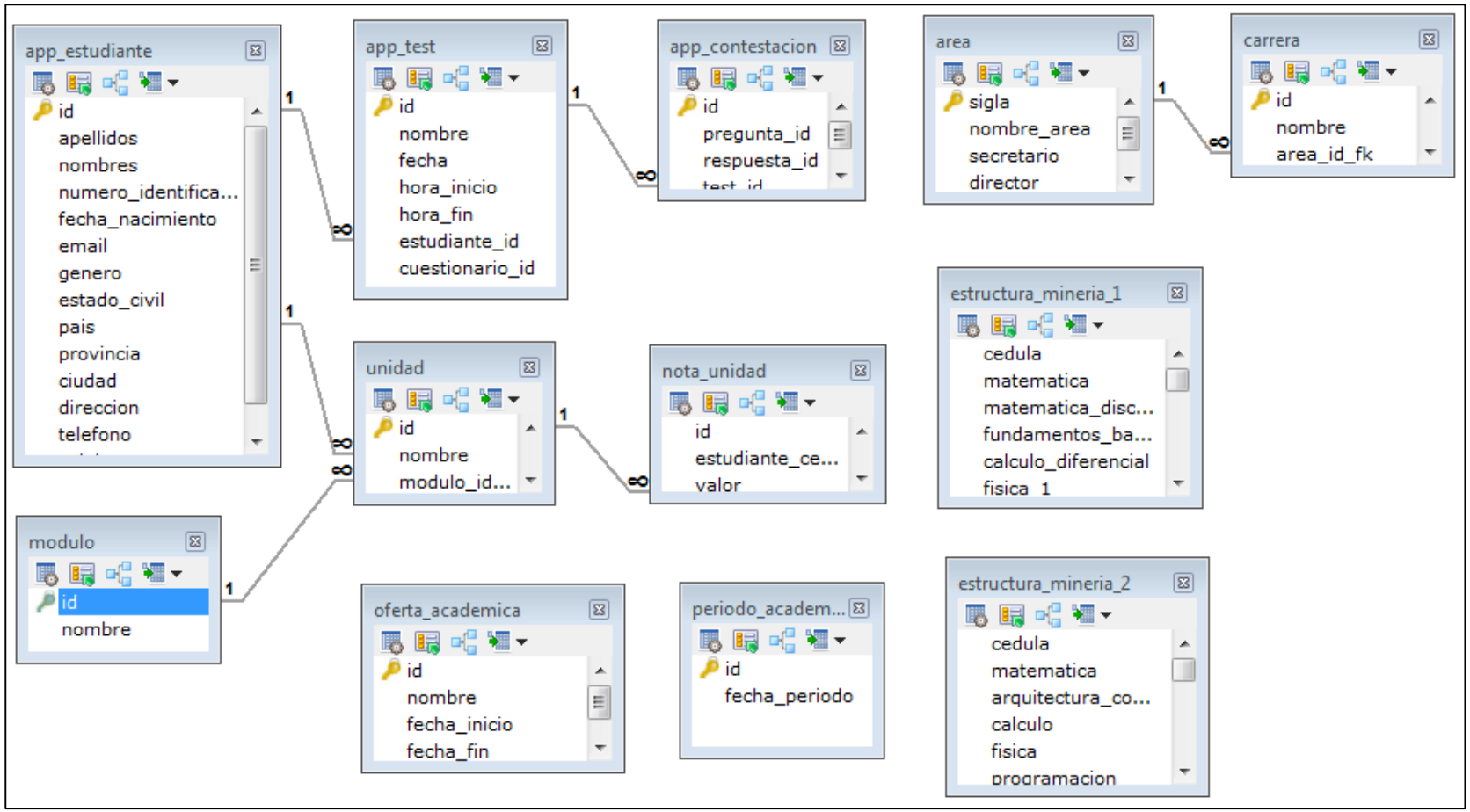
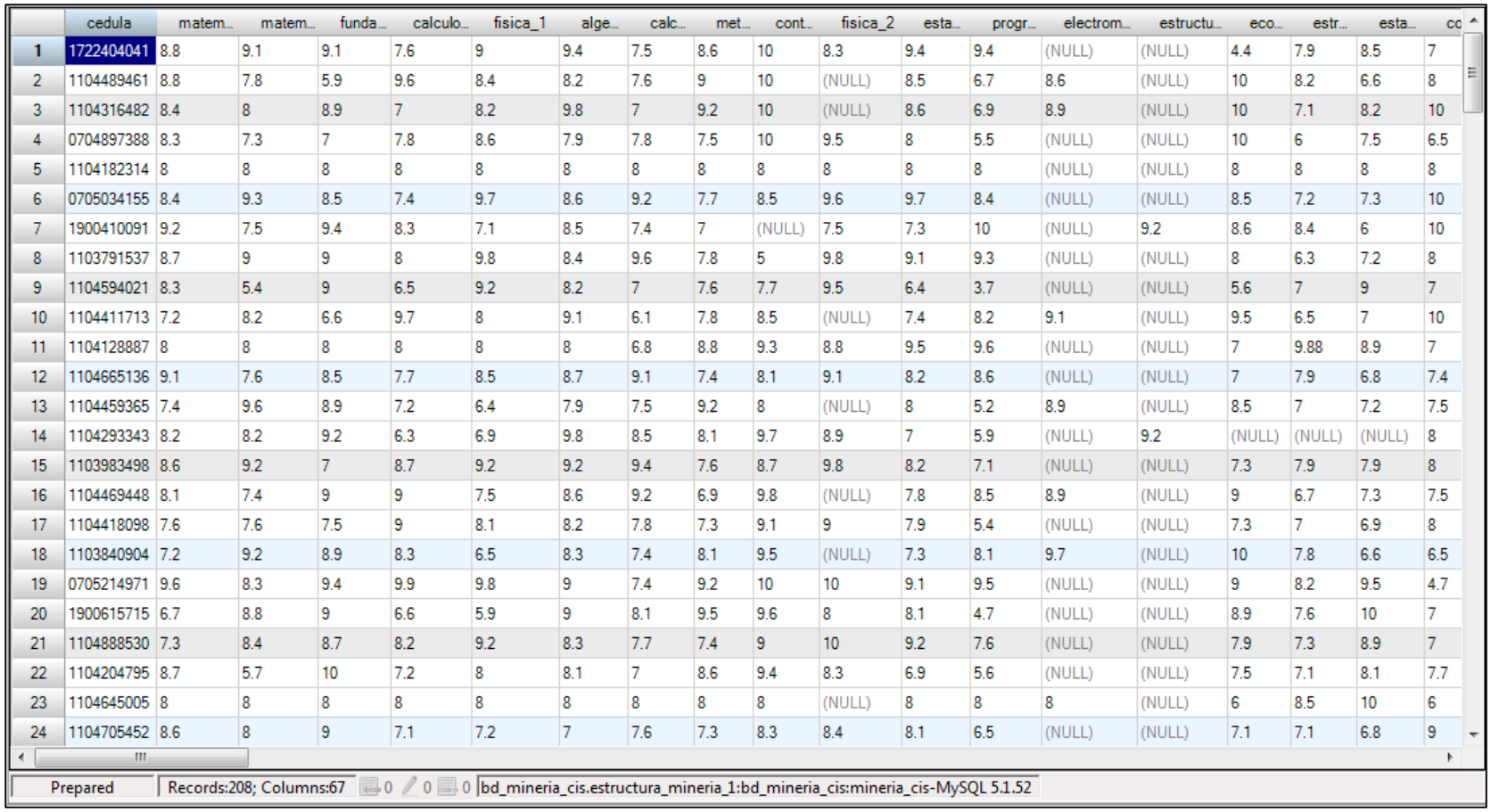
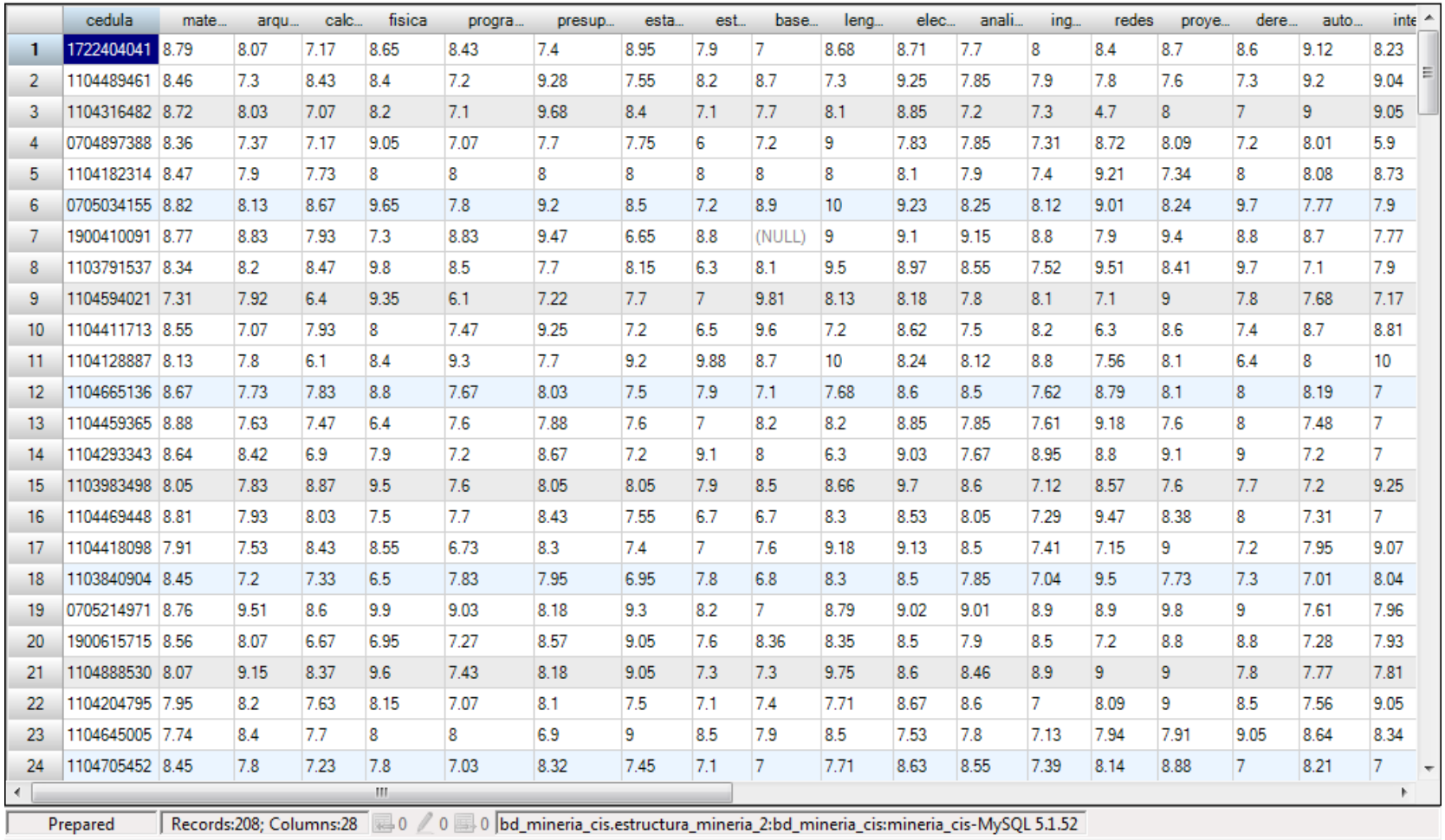


Figura 51. Diseño final de la Base de datos.



	cedula	matem...	matem...	funda...	calculo...	fisica_1	alge...	calc...	met...	cont...	fisica_2	esta...	progr...	electrom...	estructu...	eco...	estr...	esta...	cc
1	1722404041	8.8	9.1	9.1	7.6	9	9.4	7.5	8.6	10	8.3	9.4	9.4	(NULL)	(NULL)	4.4	7.9	8.5	7
2	1104489461	8.8	7.8	5.9	9.6	8.4	8.2	7.6	9	10	(NULL)	8.5	6.7	8.6	(NULL)	10	8.2	6.6	8
3	1104316482	8.4	8	8.9	7	8.2	9.8	7	9.2	10	(NULL)	8.6	6.9	8.9	(NULL)	10	7.1	8.2	10
4	0704897388	8.3	7.3	7	7.8	8.6	7.9	7.8	7.5	10	9.5	8	5.5	(NULL)	(NULL)	10	6	7.5	6.5
5	1104182314	8	8	8	8	8	8	8	8	8	8	8	8	(NULL)	(NULL)	8	8	8	8
6	0705034155	8.4	9.3	8.5	7.4	9.7	8.6	9.2	7.7	8.5	9.6	9.7	8.4	(NULL)	(NULL)	8.5	7.2	7.3	10
7	1900410091	9.2	7.5	9.4	8.3	7.1	8.5	7.4	7	(NULL)	7.5	7.3	10	(NULL)	9.2	8.6	8.4	6	10
8	1103791537	8.7	9	9	8	9.8	8.4	9.6	7.8	5	9.8	9.1	9.3	(NULL)	(NULL)	8	6.3	7.2	8
9	1104594021	8.3	5.4	9	6.5	9.2	8.2	7	7.6	7.7	9.5	6.4	3.7	(NULL)	(NULL)	5.6	7	9	7
10	1104411713	7.2	8.2	6.6	9.7	8	9.1	6.1	7.8	8.5	(NULL)	7.4	8.2	9.1	(NULL)	9.5	6.5	7	10
11	1104128887	8	8	8	8	8	8	6.8	8.8	9.3	8.8	9.5	9.6	(NULL)	(NULL)	7	9.88	8.9	7
12	1104665136	9.1	7.6	8.5	7.7	8.5	8.7	9.1	7.4	8.1	9.1	8.2	8.6	(NULL)	(NULL)	7	7.9	6.8	7.4
13	1104459365	7.4	9.6	8.9	7.2	6.4	7.9	7.5	9.2	8	(NULL)	8	5.2	8.9	(NULL)	8.5	7	7.2	7.5
14	1104293343	8.2	8.2	9.2	6.3	6.9	9.8	8.5	8.1	9.7	8.9	7	5.9	(NULL)	9.2	(NULL)	(NULL)	(NULL)	8
15	1103983498	8.6	9.2	7	8.7	9.2	9.2	9.4	7.6	8.7	9.8	8.2	7.1	(NULL)	(NULL)	7.3	7.9	7.9	8
16	1104469448	8.1	7.4	9	9	7.5	8.6	9.2	6.9	9.8	(NULL)	7.8	8.5	8.9	(NULL)	9	6.7	7.3	7.5
17	1104418098	7.6	7.6	7.5	9	8.1	8.2	7.8	7.3	9.1	9	7.9	5.4	(NULL)	(NULL)	7.3	7	6.9	8
18	1103840904	7.2	9.2	8.9	8.3	6.5	8.3	7.4	8.1	9.5	(NULL)	7.3	8.1	9.7	(NULL)	10	7.8	6.6	6.5
19	0705214971	9.6	8.3	9.4	9.9	9.8	9	7.4	9.2	10	10	9.1	9.5	(NULL)	(NULL)	9	8.2	9.5	4.7
20	1900615715	6.7	8.8	9	6.6	5.9	9	8.1	9.5	9.6	8	8.1	4.7	(NULL)	(NULL)	8.9	7.6	10	7
21	1104888530	7.3	8.4	8.7	8.2	9.2	8.3	7.7	7.4	9	10	9.2	7.6	(NULL)	(NULL)	7.9	7.3	8.9	7
22	1104204795	8.7	5.7	10	7.2	8	8.1	7	8.6	9.4	8.3	6.9	5.6	(NULL)	(NULL)	7.5	7.1	8.1	7.7
23	1104645005	8	8	8	8	8	8	8	8	8	(NULL)	8	8	8	(NULL)	6	8.5	10	6
24	1104705452	8.6	8	9	7.1	7.2	7	7.6	7.3	8.3	8.4	8.1	6.5	(NULL)	(NULL)	7.1	7.1	6.8	9

Figura 52. Estructura de minería de datos no agrupados.



	cedula	mate...	arqu...	calc...	fisica	progra...	presup...	esta...	est...	base...	leng...	elec...	anali...	ing...	redes	proye...	dere...	auto...	inte
1	1722404041	8.79	8.07	7.17	8.65	8.43	7.4	8.95	7.9	7	8.68	8.71	7.7	8	8.4	8.7	8.6	9.12	8.23
2	1104489461	8.46	7.3	8.43	8.4	7.2	9.28	7.55	8.2	8.7	7.3	9.25	7.85	7.9	7.8	7.6	7.3	9.2	9.04
3	1104316482	8.72	8.03	7.07	8.2	7.1	9.68	8.4	7.1	7.7	8.1	8.85	7.2	7.3	4.7	8	7	9	9.05
4	0704897388	8.36	7.37	7.17	9.05	7.07	7.7	7.75	6	7.2	9	7.83	7.85	7.31	8.72	8.09	7.2	8.01	5.9
5	1104182314	8.47	7.9	7.73	8	8	8	8	8	8	8	8.1	7.9	7.4	9.21	7.34	8	8.08	8.73
6	0705034155	8.82	8.13	8.67	9.65	7.8	9.2	8.5	7.2	8.9	10	9.23	8.25	8.12	9.01	8.24	9.7	7.77	7.9
7	1900410091	8.77	8.83	7.93	7.3	8.83	9.47	6.65	8.8	(NULL)	9	9.1	9.15	8.8	7.9	9.4	8.8	8.7	7.77
8	1103791537	8.34	8.2	8.47	9.8	8.5	7.7	8.15	6.3	8.1	9.5	8.97	8.55	7.52	9.51	8.41	9.7	7.1	7.9
9	1104594021	7.31	7.92	6.4	9.35	6.1	7.22	7.7	7	9.81	8.13	8.18	7.8	8.1	7.1	9	7.8	7.68	7.17
10	1104411713	8.55	7.07	7.93	8	7.47	9.25	7.2	6.5	9.6	7.2	8.62	7.5	8.2	6.3	8.6	7.4	8.7	8.81
11	1104128887	8.13	7.8	6.1	8.4	9.3	7.7	9.2	9.88	8.7	10	8.24	8.12	8.8	7.56	8.1	6.4	8	10
12	1104665136	8.67	7.73	7.83	8.8	7.67	8.03	7.5	7.9	7.1	7.68	8.6	8.5	7.62	8.79	8.1	8	8.19	7
13	1104459365	8.88	7.63	7.47	6.4	7.6	7.88	7.6	7	8.2	8.2	8.85	7.85	7.61	9.18	7.6	8	7.48	7
14	1104293343	8.64	8.42	6.9	7.9	7.2	8.67	7.2	9.1	8	6.3	9.03	7.67	8.95	8.8	9.1	9	7.2	7
15	1103983498	8.05	7.83	8.87	9.5	7.6	8.05	8.05	7.9	8.5	8.66	9.7	8.6	7.12	8.57	7.6	7.7	7.2	9.25
16	1104469448	8.81	7.93	8.03	7.5	7.7	8.43	7.55	6.7	6.7	8.3	8.53	8.05	7.29	9.47	8.38	8	7.31	7
17	1104418098	7.91	7.53	8.43	8.55	6.73	8.3	7.4	7	7.6	9.18	9.13	8.5	7.41	7.15	9	7.2	7.95	9.07
18	1103840904	8.45	7.2	7.33	6.5	7.83	7.95	6.95	7.8	6.8	8.3	8.5	7.85	7.04	9.5	7.73	7.3	7.01	8.04
19	0705214971	8.76	9.51	8.6	9.9	9.03	8.18	9.3	8.2	7	8.79	9.02	9.01	8.9	8.9	9.8	9	7.61	7.96
20	1900615715	8.56	8.07	6.67	6.95	7.27	8.57	9.05	7.6	8.36	8.35	8.5	7.9	8.5	7.2	8.8	8.8	7.28	7.93
21	1104888530	8.07	9.15	8.37	9.6	7.43	8.18	9.05	7.3	7.3	9.75	8.6	8.46	8.9	9	9	7.8	7.77	7.81
22	1104204795	7.95	8.2	7.63	8.15	7.07	8.1	7.5	7.1	7.4	7.71	8.67	8.6	7	8.09	9	8.5	7.56	9.05
23	1104645005	7.74	8.4	7.7	8	8	6.9	9	8.5	7.9	8.5	7.53	7.8	7.13	7.94	7.91	9.05	8.64	8.34
24	1104705452	8.45	7.8	7.23	7.8	7.03	8.32	7.45	7.1	7	7.71	8.63	8.55	7.39	8.14	8.88	7	8.21	7

Figura 53. Estructura de minería de datos agrupados.

3.4. Cuarta Fase: Selección de técnicas y generación de pruebas

En esta fase se detallan las técnicas de modelado elegidas y los parámetros óptimos para cada algoritmo. Una vez aplicadas las técnicas, se evaluó y comparo los resultados obtenidos. Para realizar las pruebas con cada algoritmo fue necesario utilizar la herramienta RapidMiner, previamente seleccionada, la cual cuenta con todos los componentes y características necesarias para culminar esta fase.

3.4.1. Tarea Uno: Selección de técnicas de modelado

En esta etapa se eligen los algoritmos que serán aplicados parcialmente o en su totalidad, al conjunto de datos para llevar a cabo las pruebas. En la minería de datos existen ocho familias de clasificadores, pero los más utilizados son cuatro: los bayesianos, los de agrupación, las reglas y los árboles de decisión.

En base a los casos de estudio analizados y sus técnicas aplicadas se determinó que la técnica idónea para este trabajo es la Clasificación mediante árboles de decisión, además se tomara en cuenta las técnicas de reglas basadas en inducción, estos algoritmos se encuentra detallados en el apartado CAPÍTULO II. TÉCNICAS DE MINERÍA DE DATOS.

A continuación se muestra una descripción de los algoritmos que se aplicaran en el presente Trabajo de Titulación.

3.4.1.1. Algoritmos de Clasificación basados en árboles de decisión

Este tipo de algoritmos son robustos a datos con ruido, la función aprendida es representada como un árbol, permiten obtener de forma visual las reglas de clasificación bajo las cuales operan los datos del experimento, una de sus principales ventajas es la facilidad de interpretación, debido a la representación de los resultados en forma de árbol, donde los nodos están interconectados mostrando los diferentes caminos [98, 99].

Los algoritmos más utilizados para la clasificación son los algoritmos de inducción. Además Existen varias perspectivas para los algoritmos de inducción, pero en este caso se trabajará con los que generan árboles de decisión. Los Árboles de decisión establecen un conjunto de condiciones organizadas jerárquicamente, de tal modo que la decisión final, se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [98, 99].

Los árboles de decisión trabajan mejor en tareas de clasificación. Puesto que clasificar es determinar entre varias clases, la clase a la que pertenece una entidad; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema. Además los árboles de decisión manejan muy bien datos no numéricos. Estos árboles de decisiones generan modelos para examinar los datos, realizar predicciones o describir situaciones. La calidad de un árbol se basa en la precisión de la clasificación y del tamaño del árbol [98, 99].

Para realizar los experimentos con la aplicación de algoritmos de árboles de clasificación se han tomado en cuenta los algoritmos ID3 y CHAID en la herramienta RapidMiner. La utilización de estos algoritmos es con la finalidad de generar modelos para la determinación de perfiles profesionales de egresados y graduados de la carrera de ingeniería en sistemas, La descripción de estos algoritmos se encuentra en el apartado CAPÍTULO II. TÉCNICAS DE MINERÍA DE DATOS.

3.4.1.2. Algoritmos Basados en Reglas de Inducción

Este tipos de algoritmos arrojan como resultados un sin número de reglas respecto al análisis de los datos objeto de estudio de la minería de datos, el proceso interno que realiza es la búsqueda de patrones, relaciones y características similares entre los datos. Estas reglas generadas permiten la interpretación de conocimiento asociado a las reglas de los objetivos propuestos [100, 101].

Los algoritmos de inducción de reglas tratan el conjunto de entrenamiento en forma de reglas que pueden son evaluadas directamente para clasificar nuevas instancias. Estas reglas tienen la ventaja que son fáciles de entender. Estos algoritmos realizan un proceso constante de análisis para cada regla hasta que la regla es perfecta de acuerdo a ciertas condiciones, con una precisión del 100%. Las reglas con seleccionada de acuerdo al criterio de la mayor ganancia de información [100, 101].

Para la aplicación de los algoritmos de reglas de inducción se han escogido los algoritmos JRip, Part, Ridor, Decisión Table, DTNB y NNge, detallados en el apartado CAPÍTULO II. TÉCNICAS DE MINERÍA DE DATOS.

3.4.2. Tarea Dos: Diseño de pruebas

El diseño de pruebas contiene las especificaciones de los experimentos realizados con las dos estructuras de minería construidas, enfocadas a datos no agrupados y datos agrupados. En el transcurso del desarrollo de esta tarea se especificara los aspectos que se va a evaluar de cada prueba realizada, evaluación de los resultados obtenidos para darles valor a los modelos generados con varios conjuntos de datos a lo largo del presente trabajo. A continuación, se enumeran y describen de forma general los experimentos realizados.

- **Algoritmos de clasificación en base a los árboles de decisión: ID3 y CHAID**

Con estos algoritmos se realizaron dos pruebas, la primera consiste en seleccionar un subconjunto de ejemplos denominado conjunto de entrenamiento que corresponde al 72% del total de ejemplos; la segunda prueba consiste en aplicar un método denominado validación cruzada mediante cinco subconjuntos, posteriormente se presentan los resultados obtenidos a través de la matriz de confusión, también se evalúa el número de hojas obtenidas por cada algoritmo y el número de líneas creadas para generar el árbol, también se evaluar que atributo selecciono cada algoritmo como raíz para generar los árboles, hay que mencionar que se aplicó parámetros óptimos para obtener resultados precisos.

- **Algoritmos de Inducción de Reglas JRip, PART, Ridor, Decisión Table, DTNB y NNge**

El diseño de pruebas realizadas con estos algoritmos se refiere a los experimentos realizados con las dos estructuras de minería construidas, enfocadas a datos no agrupados y datos agrupados. En las pruebas con el conjunto de entrenamiento se ha tomado un 72% de los datos mientras que el 28% restante será utilizado para la evaluación de los modelos, a su vez se detalla la validación cruzada con cinco subconjuntos, estos resultados serán evaluados a través de una matriz de confusión, y también se evaluara el número de reglas generadas en cada algoritmo y las reglas más relevantes.

3.4.3. Tarea Tres: Construcción de modelos

En esta actividad se describe los resultados obtenidos con los modelos seleccionados y se describen de forma general los experimentos realizados con los parámetros y configuraciones realizadas en la herramienta.

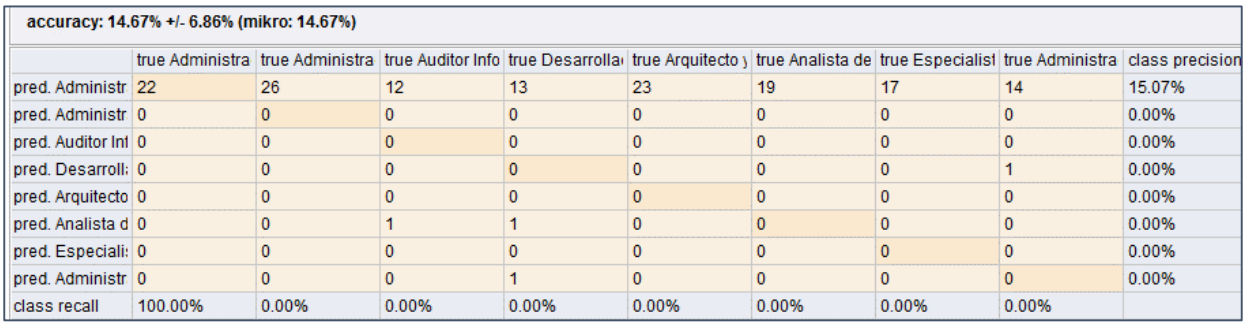
- **ID3**

Para el algoritmo ID3 se fijaron los siguientes parámetros: en criterio se seleccionó *Accuracy,* aquí se especifica el criterio de selección de atributos y de divisiones numéricas, el parámetro *minimal size for split=4* es el número mínimo de divisiones que se pueden dar por cada nodo, el parámetro *minimal leaf size=2* es el tamaño mínimo de cada hoja, el parámetro *minimal gain=0.25* se fijó como la ganancia mínima que debe lograrse con el fin de producirse una división, para la validación cruzada se fijó el parámetro *number of validations=5* que es el número de subconjuntos que se generan para evaluar el algoritmo. Los resultados arrojados por el algoritmo ID3 son los siguientes (ver tabla CXXXIV):

TABLA CXXXIV
RENDIMIENTO DEL ALGORITMO ID3.

ID3	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	28.67%	71.33%	0.167	0.000	0.00%	0.000	-
Validación Cruzada	14.67%	85.33%	0.002	1.000	100%	1.000	-

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 54).

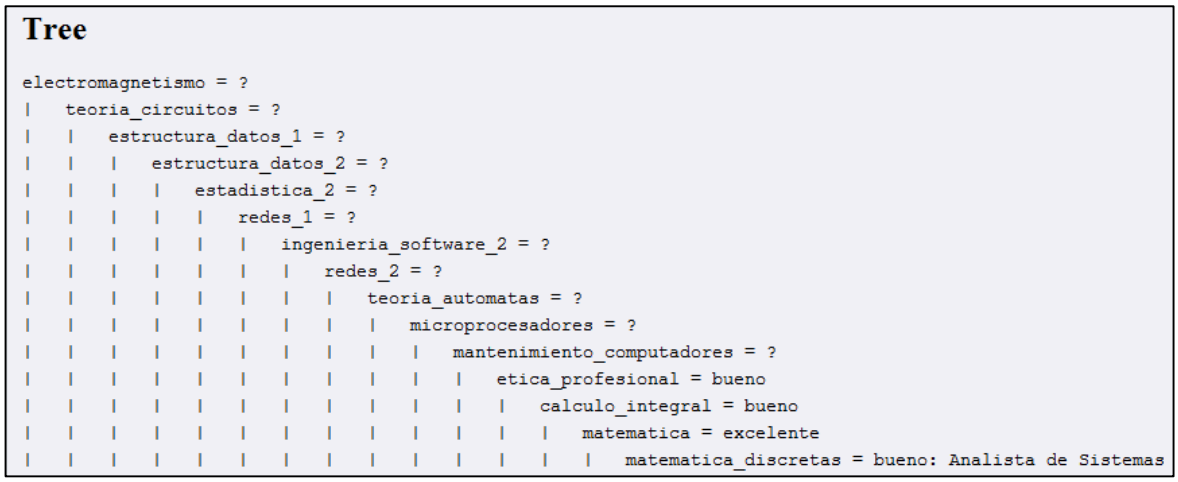


accuracy: 14.67% +/- 6.86% (mikro: 14.67%)

	true Administra	true Administra	true Auditor Info	true Desarrolla	true Arquitecto	true Analista de	true Especialist	true Administra	class precision
pred. Administr	22	26	12	13	23	19	17	14	15.07%
pred. Administr	0	0	0	0	0	0	0	0	0.00%
pred. Auditor Int	0	0	0	0	0	0	0	0	0.00%
pred. Desarroll:	0	0	0	0	0	0	0	1	0.00%
pred. Arquitecto	0	0	0	0	0	0	0	0	0.00%
pred. Analista d	0	0	1	1	0	0	0	0	0.00%
pred. Especiali:	0	0	0	0	0	0	0	0	0.00%
pred. Administr	0	0	0	1	0	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	

Figura 54. Matriz de confusión obtenida con el algoritmo ID3.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, se puede observar porcentaje de 0% para la mayoría de clases y solo el 15% para la primera clase denominada *Administrador de Sistemas de Bases de Datos.* El modelo generado por este algoritmo fue un árbol de decisión de grandes dimensiones, en modo texto ocupa aproximadamente 350 líneas, a continuación se describe un fragmento de las condiciones generadas para generar el árbol (ver figura 55).



Tree

```
electromagnetismo = ?
|   teoria_circuitos = ?
|   |   estructura_datos_1 = ?
|   |   |   estructura_datos_2 = ?
|   |   |   |   estadistica_2 = ?
|   |   |   |   |   redes_1 = ?
|   |   |   |   |   |   ingenieria_software_2 = ?
|   |   |   |   |   |   |   redes_2 = ?
|   |   |   |   |   |   |   |   teoria_automatas = ?
|   |   |   |   |   |   |   |   |   microprocesadores = ?
|   |   |   |   |   |   |   |   |   |   mantenimiento_computadores = ?
|   |   |   |   |   |   |   |   |   |   |   etica_profesional = bueno
|   |   |   |   |   |   |   |   |   |   |   |   calculo_integral = bueno
|   |   |   |   |   |   |   |   |   |   |   |   |   matematica = excelente
|   |   |   |   |   |   |   |   |   |   |   |   |   |   matematica_discretas = bueno: Analista de Sistemas
```

Figura 55. Fragmento del árbol generado por el algoritmo ID3.

Como se puede observar este modelo comprende muchas condiciones, el algoritmo eligió la unidad electromagnetismo como raíz para generar en este caso el árbol de decisión. Si bien se generaron demasiadas líneas para construir el árbol, estas permiten generar reglas o condiciones para describir cada perfil profesional.

- **CHAID**

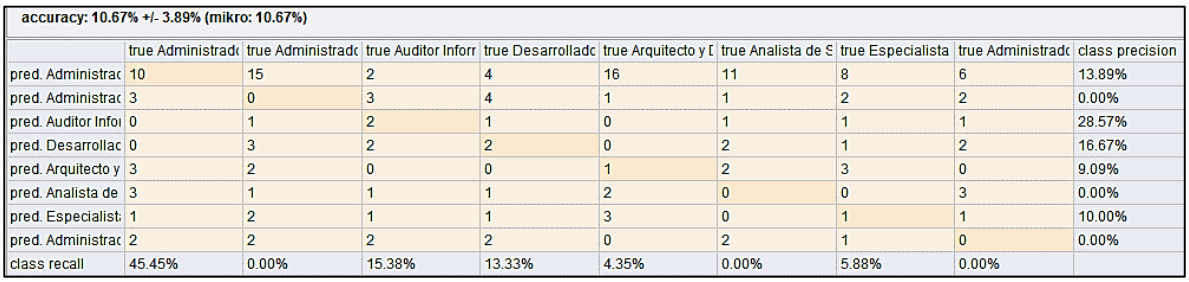
Para el algoritmo CHAID se fijaron los siguientes parámetros: el parámetro minimal size for split=4 es el número mínimo de divisiones que se pueden dar por cada nodo, el parámetro minimal leaf size=2 es el tamaño mínimo de cada hoja, el parámetro minimal gain=0.1 se fijó como la ganancia mínima que debe lograrse con el fin de producirse una división, se fijó el parámetro maximal deph=10 es el nivel máximo de profundidad al momento de crecer del árbol, también se fijó el parámetro confidence=0.5 que es el nivel de confianza utilizado para el cálculo del error pesimista de la poda, el parámetro number of prepruning=3 es el número de nodos alternativos probados cuando la técnica de la poda evitaría una división, los se estableció la opción true para el parámetros no prepruning que sirve para aplicar las reglas de poda se luego de cada iteración y se estableció false para el parámetro prenuning que permite aplicar las reglas de poda después de generar el árbol. Para la validación cruzada se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo CHAID son las siguientes (ver tabla CXXXV):

TABLA CXXXV
RENDIMIENTO DEL ALGORITMO CHAID.

CHAID	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	98.00%	2.00%	0.977	0.022	2.22%	0.105	0.046
Validación Cruzada	10.67%	89.33%	-0.033	0.894	89.43%	0.944	-

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 56).



accuracy: 10.67% +/- 3.89% (mikro: 10.67%)

	true Administrad	true Administrad	true Auditor Inforr	true Desarrollad	true Arquitecto y [	true Analista de S	true Especialista	true Administrad	class precision
pred. Administrac	10	15	2	4	16	11	8	6	13.89%
pred. Administrac	3	0	3	4	1	1	2	2	0.00%
pred. Auditor Infor	0	1	2	1	0	1	1	1	28.57%
pred. Desarrollac	0	3	2	2	0	2	1	2	16.67%
pred. Arquitecto y	3	2	0	0	1	2	3	0	9.09%
pred. Analista de	3	1	1	1	2	0	0	3	0.00%
pred. Especialist	1	2	1	1	3	0	1	1	10.00%
pred. Administrac	2	2	2	2	0	2	1	0	0.00%
class recall	45.45%	0.00%	15.38%	13.33%	4.35%	0.00%	5.88%	0.00%	

Figura 56. Matriz de confusión obtenida con el algoritmo CHAID.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, se puede observar porcentaje de 0% para las clases *Administrador de Redes computacionales, Analista de Sistemas de Información y Administrador de Centros de Cómputo,* en cambio para el resto de clases se observa un porcentaje entre 9% y 28%. El modelo generado por este algoritmo fue un árbol de decisión también de grandes dimensiones, en modo texto ocupa aproximadamente 300 líneas, a continuación se describe un fragmento de las condiciones generadas para generar el árbol (ver figura 57).

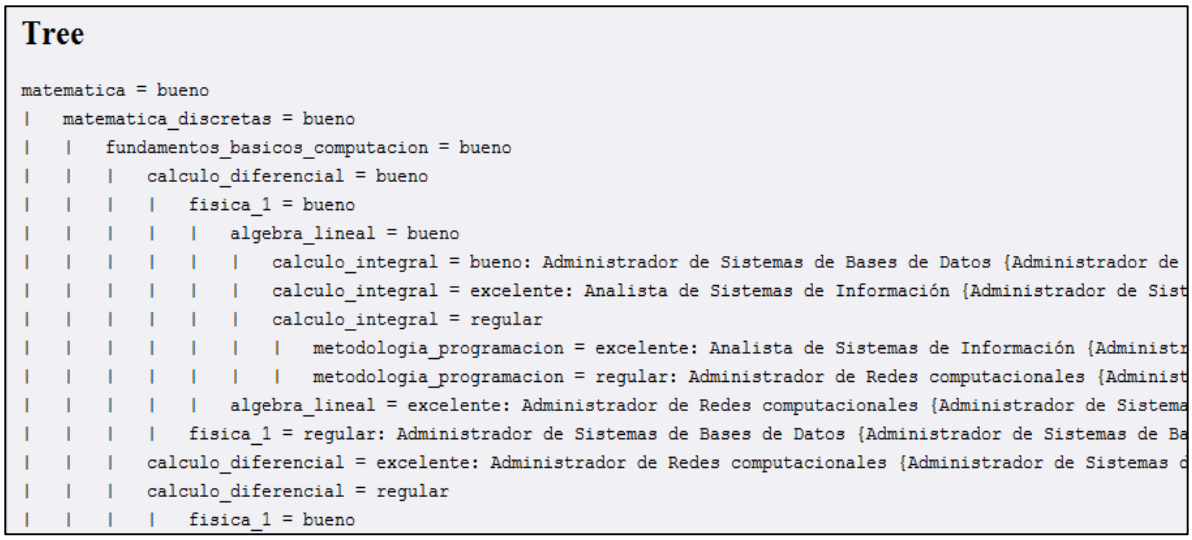


Figura 57. Fragmento del árbol generado por el algoritmo CHAID.

Como se puede observar este modelo comprende muchas condiciones, el algoritmo eligió la unidad matemática como raíz para generar en este caso el árbol de decisión. Si bien se generaron demasiadas líneas para construir el árbol, estas permiten generar reglas o condiciones para describir cada perfil profesional.

- **Decision Table**

Para el algoritmo Decisión Table se fijaron los siguientes parámetros: el parámetro S= weka.attributeSelection.BestFirst -D 1 -N 5, se lo fijo por defecto ya que define al método de búsqueda, el parámetro X=1.0, define el número de validaciones cruzadas, el parámetro E=null es el criterio de evaluación por defecto es accuracy, el parámetro I=true se define el uso de vecinos cercanos en lugar de tabla global. Para la validación cruzada como en todas las pruebas que se realizó se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo. Los resultados arrojados por el algoritmo Decisión Table son las siguientes (ver tabla CXXXVI).

TABLA CXXXVI
RENDIMIENTO DEL ALGORITMO DECISION TABLE.

DECISION TABLE	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	33.33%	66.67%	0.222	0.825	82.47%	0.827	0.364
Validación Cruzada	10.00%	90.00%	-0.034	0.891	89.08%	0.892	0.410

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 58).

accuracy: 10.00% +/- 5.58% (mikro: 10.00%)

	true Administrad	true Administrad	true Auditor Infor	true Desarrollad	true Arquitecto y	true Analista de	true Especialista	true Administrad	class precision
pred. Administra	3	4	4	1	6	9	3	5	8.57%
pred. Administra	5	7	5	7	11	1	5	5	15.22%
pred. Auditor Infc	1	2	2	1	1	0	1	1	22.22%
pred. Desarrolla	4	4	1	0	0	1	1	1	0.00%
pred. Arquitecto y	7	7	1	4	3	6	5	3	8.33%
pred. Analista de	2	1	0	0	1	0	2	0	0.00%
pred. Especialis	0	0	0	1	0	0	0	0	0.00%
pred. Administra	0	1	0	1	1	2	0	0	0.00%
class recall	13.64%	26.92%	15.38%	0.00%	13.04%	0.00%	0.00%	0.00%	

Figura 58. Matriz de confusión obtenida con el algoritmo Decision Table.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, se puede observar un porcentaje de 0% para las clases *Desarrollador de software, Analista de Sistemas de Información, Especialista en mantenimiento hardware & software y Administrador de Centros de Cómputo,* en cambio para el resto de clases se observa un porcentaje entre 8% y 22%. La tabla de Decisión generada por este algoritmo consta de 29 reglas, a continuación se describe un fragmento (ver figura 59).

Rules:

matematica_discretas	fisica_1	teoria_circuitos
excelente	excelente	excelente
bueno	bueno	excelente
regular	bueno	excelente
excelente	bueno	excelente
bueno	regular	excelente
excelente	regular	excelente
excelente	excelente	excelente
excelente	excelente	excelente
bueno	excelente	excelente
regular	regular	excelente
excelente	bueno	excelente
regular	excelente	?
bueno	excelente	?
excelente	excelente	?
excelente	bueno	excelente

Figura 59. Fragmento de la tabla de decisión generada por el algoritmo Decision Table.

Como se puede observar este modelo comprende reglas comprensibles pero se basa en 6 del total de atributos o unidades. Si bien se generaron reglas comprensibles es necesario tomar en cuenta la mayoría de los atributos para describir cada perfil profesional.

- **DTNB**

Para el algoritmo DTNB se fijaron los siguientes parámetros: el parámetro X=1.0, se fijó este valor como el número validaciones cruzadas, el parámetro E=false, define el criterio de evaluación como defecto se establece en accuracy, el parámetro I=false establece el uso del vecino más cercano en lugar de la mayoría en la tabla global, el parámetro R=true sirve para visualizar las reglas en la tabla de decisión. Para la validación cruzada como en todas las pruebas que se realizó se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo DTNB son las siguientes (ver tabla CXXVII):

TABLA CXXXVII
RENDIMIENTO DEL ALGORITMO DTNB.

DTNB	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	35.33%	64.67%	0.255	0.800	80.01%	0.810	0.356
Validación Cruzada	10.00%	90.00%	-0.014	0.881	88.14%	0.883	0.407

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 60).

accuracy: 10.00% +/- 3.65% (mikro: 10.00%)

	true Administrac	true Administrac	true Auditor Info	true Desarrollac	true Arquitecto y	true Analista de	true Especialist	true Administrac	class precision
pred. Administra	1	5	5	4	7	2	3	1	3.57%
pred. Administra	7	7	4	1	4	5	3	7	18.42%
pred. Auditor Inf	1	1	0	0	0	1	3	2	0.00%
pred. Desarrolla	1	1	0	0	1	1	1	0	0.00%
pred. Arquitecto	5	7	3	6	4	5	3	4	10.81%
pred. Analista d	3	1	0	0	4	1	2	0	9.09%
pred. Especialis	2	1	1	1	0	2	1	0	12.50%
pred. Administra	2	3	0	3	3	2	1	1	6.67%
class recall	4.55%	26.92%	0.00%	0.00%	17.39%	5.26%	5.88%	6.67%	

Figura 60. Matriz de confusión obtenida con el algoritmo DTNB.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidas como los perfiles profesionales, se puede observar un porcentaje de 0% para las clases *Auditor Informático, Desarrollador de software,* en cambio para el resto de clases se observa un porcentaje entre 3% y 18%. La tabla de Decisión generada por este algoritmo consta de 138 reglas, a continuación se describe un fragmento (ver figura 61).

Rules:

matematica	matematica_discretas	fundamentos_basicos_co
bueno	excelente	regular
bueno	regular	regular
regular	regular	regular
excelente	excelente	excelente
bueno	regular	excelente
excelente	excelente	bueno
excelente	bueno	bueno
excelente	excelente	excelente
excelente	bueno	excelente
bueno	excelente	excelente
excelente	excelente	excelente
excelente	bueno	bueno
excelente	excelente	bueno
excelente	bueno	excelente
excelente	excelente	excelente
excelente	excelente	excelente
bueno	regular	excelente
excelente	excelente	bueno
excelente	excelente	regular
bueno	regular	excelente
excelente	regular	regular

Figura 61. Fragmento de tabla de decisión generado por el algoritmo DTNB.

Como se puede observar este modelo comprende reglas comprensibles pero se basa en 7 del total de atributos o unidades. Si bien se generaron reglas comprensibles, es necesario tomar en cuenta la mayoría de las unidades como atributos para describir cada perfil profesional.

- **JRIP**

Para el algoritmo JRip se fijaron los siguientes parámetros: el parámetro F=3 es el número mínimo de divisiones que se pueden dar por cada nodo, el parámetro N=2 es el tamaño mínimo de cada hoja, el parámetro O=2 el número de ejecuciones para lograr una para optimización, D=false indica el modo de depuración, S=10 indica la semilla para generar aleatoriedad, E=false si no marca la tasa de error, P=true para indicar se utiliza la poda. Para la validación cruzada se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo JRip son las siguientes (ver tabla CXXXVIII):

TABLA CXXXVIII
RENDIMIENTO DEL ALGORITMO JRIP.

JRIP	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	94.00%	6.00%	0.931	0.101	10.10%	0.254	0.112
Validación Cruzada	12.00%	88.00%	-0.001	0.879	87.90%	0.931	0.429

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 62).

accuracy: 12.00% +/- 3.40% (mikro: 12.00%)

	true Administra	true Administra	true Auditor Info	true Desarrollac	true Arquitecto y	true Analista de	true Especialist	true Administra	class precision
pred. Administr	2	3	3	3	3	3	2	1	10.00%
pred. Administr	9	4	3	3	8	0	3	6	11.11%
pred. Auditor Inf	2	0	1	0	0	2	2	1	12.50%
pred. Desarrolla	3	1	0	1	2	2	3	1	7.69%
pred. Arquitecto	3	3	2	3	2	2	0	2	11.76%
pred. Analista d	1	5	0	3	2	5	3	1	25.00%
pred. Especialis	2	3	4	2	1	3	1	1	5.88%
pred. Administr	0	7	0	0	5	2	3	2	10.53%
class recall	9.09%	15.38%	7.69%	6.67%	8.70%	26.32%	5.88%	13.33%	

Figura 62. Matriz de confusión obtenida con el algoritmo JRIP.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, no existen clases con un porcentaje de clasificación de 0%, se observa un porcentaje de clasificación entre 5% y 25%. Las reglas generadas por este algoritmo son 38, a continuación se describe un fragmento (ver figura 63).

W-JRip

```
JRIP rules:
===========

(fundamentos_basicos_computacion = regular) and (auditoria_informatica = excelente) and (fisica_2 = bueno) => perf
(diseno_digital = regular) and (administracion_cc = regular) and (matematica_discretas = excelente) => perfil_prof
(estadistica_1 = bueno) and (contabilidad_costos = regular) and (analisis_diseno_sistemas_1 = regular) => perfil_p
(estadistica_1 = bueno) and (matematica_discretas = regular) and (estadistica_inferencial = excelente) => perfil_p
(aplicaciones_web = regular) and (estadistica_2 = bueno) => perfil_profesional=Administrador de Centros de Cómputo
(auditoria_informatica = bueno) and (sistemas_informacion_2 = excelente) and (fisica_1 = bueno) => perfil_profesio
(teoria_telecomunicaciones = regular) and (fisica_2 = bueno) and (matematica = bueno) => perfil_profesional=Admini
(ingenieria_software_1 = regular) and (inteligencia_artificial = regular) and (diseno_gestion_bd = bueno) => perfi
(analisis_diseno_sistemas_1 = excelente) and (contabilidad_costos = excelente) and (algebra_lineal = excelente) =>
(estadistica_1 = excelente) and (calculo_integral = excelente) and (inteligencia_artificial = excelente) and (ecor
(programacion_1 = regular) and (administracion_empresas = regular) and (fundamentos_basicos_computacion = bueno) =
(calculo_integral = bueno) and (economia = bueno) and (calculo_diferencial = regular) and (proyecto_informaticos_1
(programacion_2 = bueno) and (compiladores = regular) => perfil_profesional=Desarrollador de software (2.0/0.0)
(calculo_diferencial = regular) and (fisica_2 = bueno) and (etica_profesional = excelente) => perfil_profesional=E
(calculo_integral = excelente) and (electronica_basica = bueno) and (proyecto_informaticos_1 = bueno) => perfil_pr
(sistemas_informacion_1 = excelente) and (calculo_diferencial = regular) and (administracion_cc = regular) => perf
(metodologia_programacion = bueno) and (sistemas_informacion_1 = excelente) and (fundamentos_basicos_computacion =
(modelamiento_matematico = regular) and (compiladores = bueno) and (fundamentos_basicos_computacion = regular) =>
(metodologia_programacion = excelente) and (fundamentos_basicos_computacion = bueno) and (matematica_discretas = b
```

Figura 63. Fragmento de reglas generadas por el algoritmo JRIP.

Como se puede observar este modelo genero menos reglas que los algoritmos anteriores las cuales son extensas, pero generadas en base a todas las unidades o atributos seleccionados inicialmente para describir cada perfil profesional.

- **PART**

Para el algoritmo PART se fijaron los siguientes parámetros: el parámetro C=0.6 el umbral de confianza para la poda, el parámetro M=2 en este es el número mínimo de objetos por hoja, el parámetro R=false es el uso reducido de poda error, el parámetro N=default establece el número de pliegues para la reducción de poda error, a veces se usa como poda, el parámetro B=true se usa para divisiones binarias, el parámetro U=true genera una lista de decisión sin podar, el parámetro Q=10 es la semilla de datos aleatorios. Para la validación cruzada como en todas las pruebas que se realizó se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo PART son las siguientes (ver tabla CXXXIX):

TABLA CXXXIX
RENDIMIENTO DEL ALGORITMO PART.

PART	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	81.33%	18.67%	0.785	0.376	37.61%	0.456	0.200
Validación Cruzada	15.33%	84.67%	0.037	0.843	84.32%	0.891	0.410

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 64).

accuracy: 15.33% +/- 3.40% (mikro: 15.33%)

	true Administrador	true Administrador	true Auditor Inform	true Desarrollador	true Arquitecto y D	true Analista de Si	true Especialista e	true Administrador	class precision
pred. Administrad	1	5	3	3	5	3	0	1	4.76%
pred. Administrad	5	2	2	3	0	5	1	4	9.09%
pred. Auditor Infor	1	5	1	0	4	1	1	2	6.67%
pred. Desarrollad	5	4	3	3	6	3	2	0	11.54%
pred. Arquitecto y [	2	1	0	0	2	2	0	5	16.67%
pred. Analista de S	1	6	1	4	3	5	3	1	20.83%
pred. Especialista	6	0	2	2	2	0	7	0	36.84%
pred. Administrad	1	3	1	0	1	0	3	2	18.18%
class recall	4.55%	7.69%	7.69%	20.00%	8.70%	26.32%	41.18%	13.33%	

Figura 64. Matriz de confusión obtenida con el algoritmo PART.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, no existen clases con un porcentaje de clasificación de 0%, se observa un porcentaje de clasificación entre 4% y 36%. Las reglas generadas por este algoritmo son 50, a continuación se describe un fragmento (ver figura 65).

W-PART

```
PART decision list
------------------

aplicaciones_web = bueno AND
estructura_datos_2 != excelente AND
programacion_2 = excelente AND
ingenieria_software_1 != bueno AND
matematica_discretas = bueno: Administrador de Redes computacionales (3.46/0.59)

sistemas_expertos = regular AND
inteligencia_artificial != excelente AND
fisica_1 != excelente: Administrador de Sistemas de Bases de Datos (3.0)

sistemas_expertos = regular AND
contabilidad_general = bueno: Administrador de Centros de Cómputo (3.0/1.0)
```

Figura 65. Fragmento de reglas generadas por el algoritmo PART.

Como se puede observar este modelo generó más reglas que el algoritmo anterior, estas reglas son cortas pero descriptivas, esto no ocurrió con el algoritmo anterior, se puede observar que las generadas son base a todas las unidades o atributos seleccionados inicialmente para describir cada perfil profesional.

- **RIDOR**

Para el algoritmo Ridor se fijaron los siguientes parámetros: el parámetro F=2.0, se lo fijo este valor como el número de pliegues, para realizar la poda, el parámetro S=4.0, define el número de barajaduras para seleccionar al azar los datos con el fin de obtener una mejor regla, el parámetro A=false, establece el uso de bandera o también si se utiliza la tasa de error de todos los datos para seleccionar la clase predeterminada en cada paso, el parámetro M=false establece el uso de bandera o también la clase mayoritaria como clase predeterminada en cada paso en lugar de elegir la clase predeterminada basada en la tasa de error, el parámetro N=2.0 establece los pesos mínimos de las instancias en una división.

Los resultados arrojados por el algoritmo RIDOR son las siguientes (ver tabla CXL):

TABLA CXL
RENDIMIENTO DEL ALGORITMO RIDOR.

RIDOR	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	51.33%	48.67%	0.437	0.487	48.67%	0.698	0.307
Validación Cruzada	10.67%	89.33%	-0.025	0.893	89.33%	0.944	0.436

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 66).

accuracy: 10.67% +/- 7.12% (mikro: 10.67%)

	true Administra	true Administra	true Auditor Info	true Desarrolla	true Arquitecto y	true Analista de	true Especialist	true Administra	class precision
pred. Administr	2	8	0	2	7	4	3	2	7.14%
pred. Administr	2	4	3	0	2	2	2	3	22.22%
pred. Auditor Inf	0	3	0	1	6	1	5	2	0.00%
pred. Desarroll	4	1	3	0	0	2	1	1	0.00%
pred. Arquitecto	5	4	4	2	3	2	1	1	13.64%
pred. Analista d	2	2	0	3	4	3	2	2	16.67%
pred. Especiali	4	1	2	3	1	2	3	3	15.79%
pred. Administr	3	3	1	4	0	3	0	1	6.67%
class recall	9.09%	15.38%	0.00%	0.00%	13.04%	15.79%	17.65%	6.67%	

Figura 66. Matriz de confusión obtenida con el algoritmo RIDOR.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, se puede observar a las clases *Auditor Informático y Desarrollador de software*, con un porcentaje de clasificación de 0%, mientras que las demás clases presentan un porcentaje de clasificación entre 6% y 22%. Las reglas generadas por este algoritmo son 6, a continuación se describe un fragmento (ver figura 67).

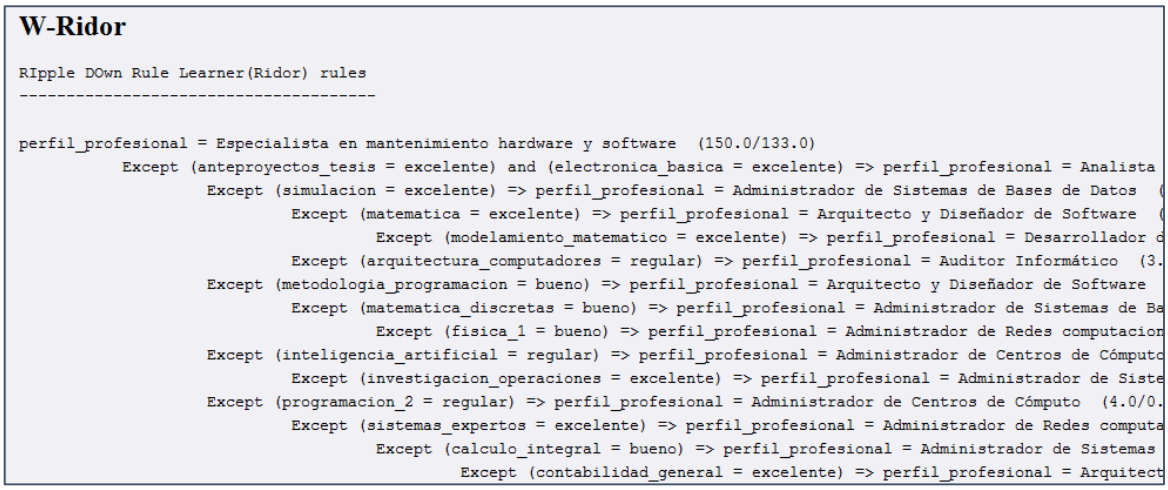


Figura 67. Fragmento de reglas generadas por el algoritmo RIDOR.

Como se puede observar este modelo generó más reglas que el algoritmo anterior, estas reglas, pero estas son extensas lo cual dificulta su comprensión, se puede observar además que las generadas son base a todas las unidades o atributos seleccionados inicialmente para describir cada perfil profesional.

- **NNGE**

Haciendo uso de la librería de weka.jar, descargada de la Página Oficial Weka, han desarrollado una aplicación en base al lenguaje de programación java [102], que permite hacer el llamado directo de algunos de los algoritmos de clasificación de la herramienta weka. Entre los algoritmos que incluye esta aplicación tenemos el algoritmo NNge, el cual no se ha podido probar en la herramienta RapidMiner a pesar de incluir el módulo de weka, debido a que no soporta este algoritmo y no genera la matriz de confusión. Por lo tanto haciendo uso de esta aplicación se han realizado las pruebas en el formato .arff.

Para este algoritmo se fijaron los siguientes parámetros: el parámetro *G=5.0*, se fijó este valor como el número de intentos de generalización, el parámetro *I=5.0*, define el número de carpeta para el cálculo de la información mutua.

Para la validación cruzada como en todas las pruebas que se realizó, se fijó el parámetro *number of validations=5* que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo NNge son las siguientes (ver tabla CXLI):

TABLA CXLI
RENDIMIENTO DEL ALGORITMO NNGE.

NNGE	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	100%	0.0%	1.000	0	0	0	0
Validación Cruzada	10.67%	89.33%	-0.039	0.223	102.7%	0.472	1.433

Luego de aplicar la validación cruzada con el algoritmo NNge, estas es la matriz de confusión que se generó (ver figura 68).

```
=== Confusion Matrix ===

a b c d e f g h   <-- classified as
4 6 3 1 2 3 2 1 | a = Administrador de Sistemas de Bases de Datos
4 7 4 1 3 5 1 1 | b = Administrador de Redes computacionales
3 7 1 0 0 1 0 1 | c = Auditor Inform□tico
3 7 1 1 0 3 0 0 | d = Desarrollador de software
2 9 0 2 1 5 2 2 | e = Arquitecto y Dise□ador de Software
6 4 1 1 4 1 0 2 | f = Analista de Sistemas de Informaci□n
6 4 0 2 0 1 0 4 | g = Especialista en mantenimiento hardware y software
5 3 0 2 2 1 1 1 | h = Administrador de Centros de C□mputo
```

Figura 68. Matriz de confusión obtenida por el algoritmo NNge.

La matriz de confusión describe el porcentaje de precisión para cada perfil profesional, en donde se observar que presentan bajas probabilidades. Las reglas generadas por el algoritmo son 67, a continuación se muestra un fragmento de las mismas (ver figura 69).

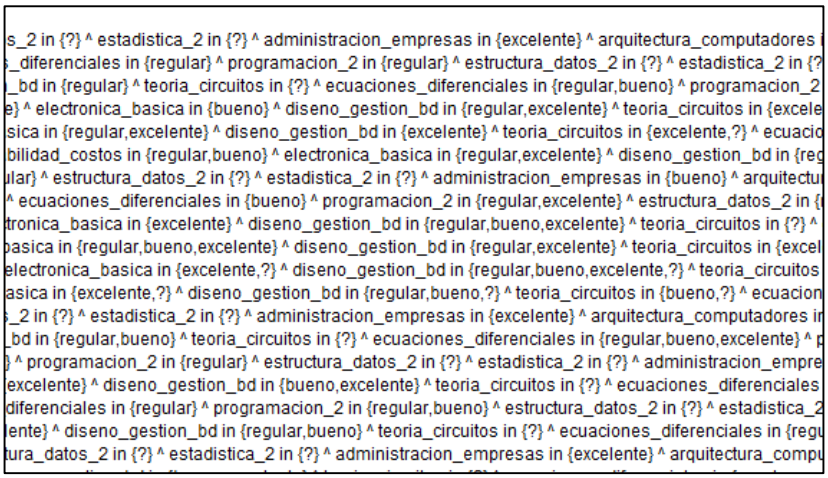


Figura 69. Fragmento de las reglas obtenidas por el algoritmo NNge.

El algoritmo género un gran número de reglas, se puede observar que las reglas construidas son en base a todos los atributos de la estructura, cabe mencionar que algunas de las reglas generadas por el algoritmo tomo en cuenta valores nulos.

3.4.4. Tarea Cuatro: Evaluación General de Modelos

La tarea se enfoca en realizar una evaluación global de los algoritmos aplicados, en otras palabras corresponde a la tarea del análisis de los resultados obtenidos en la minería, se detalla los resultados más positivos, tomando en cuenta el porcentaje de clasificación en el conjunto de entrenamiento, el proceso de validación, la lógica de las reglas obtenidas, etc, respecto del rendimiento de los algoritmos que han arrojado los mejores resultados para la determinación de perfiles profesionales.

Para establecer el mejor algoritmo, se eligieron las siguientes medidas de error: índice de Kappa, error absoluto, error relativo, error cuadrático medio y error cuadrático relativo, las cuales son el resultado de los algoritmos aplicados (ver tabla CXLII).

TABLA CXLII

COMPARACIÓN DEL RENDIMIENTO DE ALGORITMOS CON DATOS NO AGRUPADOS

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)	Índice de Kappa	Error Absoluto	Error Relativo	Error Cuadrático Medio	Error Cuadrático Relativo
CHAID	Conjunto de Entrenamiento	98.00%	2.00%	0.977	0.022	2.22%	0.105	0.046
	Validación Cruzada	10.67%	89.33%	-0.033	0.894	89.43%	0.944	-
DECISION TABLE	Conjunto de Entrenamiento	33.33%	66.67%	0.222	0.825	82.47%	0.827	0.364
	Validación Cruzada	10.00%	90.00%	-0.034	0.891	89.08%	0.892	0.410
DTNB	Conjunto de Entrenamiento	35.33%	64.67%	0.255	0.800	80.01%	0.810	0.356
	Validación Cruzada	10.00%	90.00%	-0.014	0.881	88.14%	0.883	0.407
ID3	Conjunto de Entrenamiento	28.67%	71.33%	0.167	0.000	0.00%	0.000	-
	Validación Cruzada	14.67%	85.33%	0.002	1.000	100.00%	1.000	-
JRIP	Conjunto de Entrenamiento	94.00%	6.00%	0.931	0.101	10.10%	0.254	0.112
	Validación Cruzada	12.00%	88.00%	-0.001	0.879	87.90%	0.931	0.429
PART	Conjunto de Entrenamiento	81.33%	18.67%	0.785	0.376	37.61%	0.456	0.200
	Validación Cruzada	15.33%	84.67%	0.037	0.843	84.32%	0.891	0.410
RIDOR	Conjunto de Entrenamiento	51.33%	48.67%	0.437	0.487	48.67%	0.698	0.307
	Validación Cruzada	10.67%	89.33%	-0.025	0.893	89.33%	0.944	0.436
NNGE	Conjunto de Entrenamiento	100%	0.0%	1.000	0.000	0.0%	0.000	0.000
	Validación Cruzada	10.67%	89.33%	-0.039	0.223	102.76%	0.472	1.433

- **Resultados de los algoritmos con datos no agrupados.**

Como podemos observar en la tabla CXLII, en las pruebas realizadas con el conjunto de entrenamiento los algoritmos que presentan mejor rendimiento son: CHAID, PART, JRip y NNge; siendo el porcentaje de clasificación superior a las realizadas mediante validación cruzada, esto es lógico puesto que los mismos datos sirvieron de base para la generación de reglas, es por esto que se produce una sobreestimación de los resultados; con ello se ha logrado una aproximación de los algoritmos con mejor rendimiento (ver figura 70).

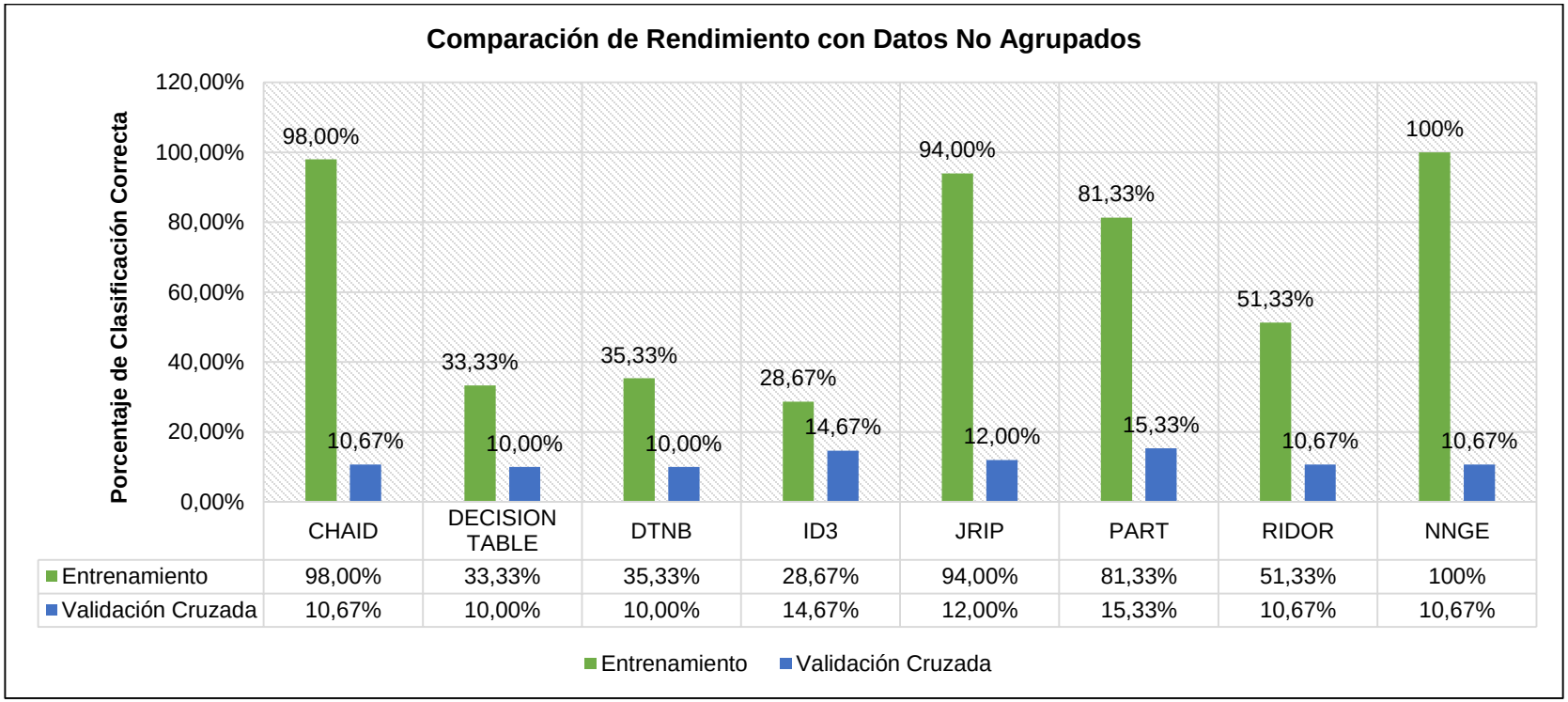


Figura 70. Comparación Rendimiento en pruebas de Entrenamiento y Validación Cruzada con datos no agrupados.

Los algoritmos con mejor rendimiento son: CHAID, JRip, PART, Ridor y NNge, estos presentan un rendimiento similar en pruebas de Entrenamiento. De estos algoritmos NNge se ubica en el primer lugar sin duda aparente debido a que presenta un rendimiento del 100%, sin embargo a pesar de ser un algoritmo muy robusto, es sensible al generar reglas con datos que contengan valores nulos, por ello las reglas generadas por este algoritmo se construyeron tomando en cuenta valores nulos o perdidos representados por el signo *"?",* perdiendo el sentido lógico y dificultando la tarea de análisis [104], es por ello que ha sido descartado como uno de los mejores.

En las pruebas de Validación cruzada todos los algoritmos presentaron bajos porcentajes de clasificación debido a la poca cantidad de datos [105], por tal motivo los algoritmos no superan el 15% de clasificación, sin embargo esto no ha significado que los modelos obtenidos por los algoritmos que han presentado un buen rendimiento sean erróneos.

- **Medidas de error de los algoritmos con mejor rendimiento con datos no agrupados.**

La primera medida de error es el índice de Kappa, el algoritmo que presenta el mejor índice es: PART con 0.037, en otra medida de error denominada Error Absoluto tiene como mejor resultado también el obtenido por el algoritmo NNGE con 0.223, en la medida de error denominada Error Relativo igualmente tiene como mejor resultado el obtenido también por el algoritmo PART con un 84.32%, en la medida de error denominada Error Cuadrático Medio el mejor resultado fue el obtenido por el algoritmo NNGE con una medida de 0.472, mientras que en la última medida de error denominada Error Cuadrático Relativo el algoritmos que tiene el mejor resultado es DTNB con una medida de 0.407.

A continuación se describen los porcentajes de clasificación para cada clase o perfil profesional (ver figura 71).

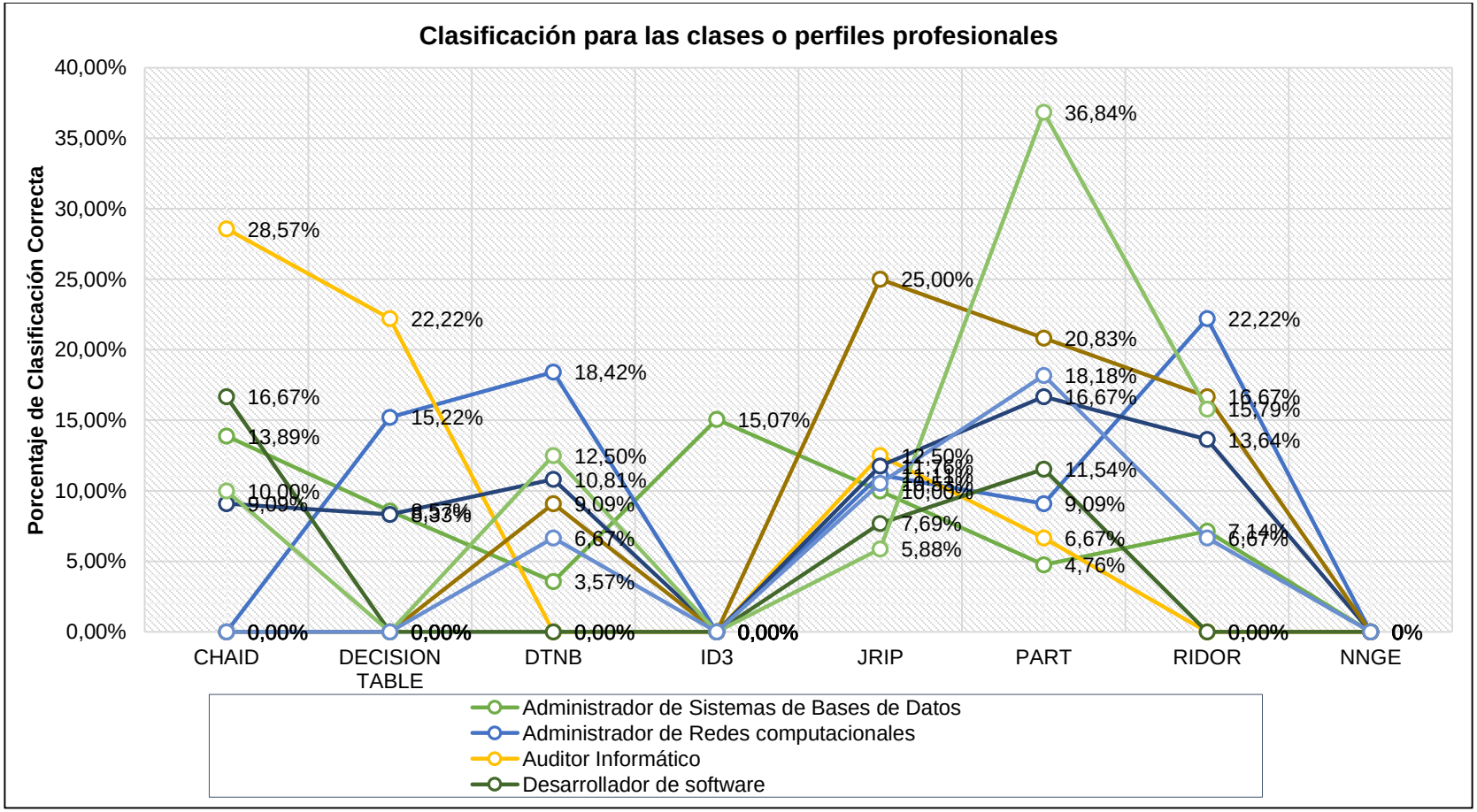


Figura 71. Clasificación para las clases perfiles profesionales.

En la figura 71 se puede observar que los algoritmos que mejor porcentaje de clasificación obtuvieron para las clases o perfiles profesionales: PART, Ridor y JRip.

Con el propósito de incrementar el porcentaje de clasificación en las pruebas de validación cruzada, se agrupo algunas unidades consiguiendo los siguientes resultados (ver tabla CXLIII).

TABLA CXLIII

COMPARACIÓN DEL RENDIMIENTO DE ALGORITMOS CON DATOS AGRUPADOS

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)	Índice de Kappa	Error Absoluto	Error Relativo	Error Cuadrático Medio	Error Cuadrático Relativo
CHAID	Conjunto de Entrenamiento	96.67%	3.33%	0.962	0.036	3.56%	0.133	0.059
	Validación Cruzada	14.00%	86.00%	0.004	0.866	86.56%	0.928	-
DECISION TABLE	Conjunto de Entrenamiento	27.33%	72.67%	0.139	0.842	84.23%	0.845	0.372
	Validación Cruzada	8.67%	91.33%	-0.022	0.881	88.13%	0.883	0.407
DTNB	Conjunto de Entrenamiento	24.67%	75.33%	0.153	0.850	85.02%	0.852	0.375
	Validación Cruzada	9.33%	90.67%	-0.025	0.874	87.45%	0.876	0.403
ID3	Conjunto de Entrenamiento	87.33%	12.67%	0.854	0.000	0.00%	0.000	-
	Validación Cruzada	14.00%	86.00%	0.009	0.872	87.18%	0.931	-
JRIP	Conjunto de Entrenamiento	94.67%	5.33%	0.938	0.086	8.62%	0.220	0.097
	Validación Cruzada	14.00%	86.00%	0.025	0.869	86.91%	0.920	0.424
PART	Conjunto de Entrenamiento	82.67%	17.33%	0.800	0.197	19.70%	0.317	0.139
	Validación Cruzada	8.00%	92.00%	-0.045	0.915	91.48%	0.945	0.436
RIDOR	Conjunto de Entrenamiento	42.67%	57.33%	0.345	0.573	57.33%	0.757	0.333
	Validación Cruzada	12.67%	87.33%	-0.002	0.873	87.33%	0.934	0.428
NNGE	Conjunto de Entrenamiento	100.0%	0.0%	1.000	0.000	0.0%	0.000	0.000
	Validación Cruzada	11.33%	88.67%	-0.026	0.221	102.0%	0.470	1.428

- **Resultados de los algoritmos con datos agrupados.**

En pruebas de entrenamiento se puede observar que los algoritmos que presentan mejor rendimiento son: CHAID, ID3, PART, JRip y NNge; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento. De estos algoritmos NNge se ubica en el primer lugar sin duda aparente debido a que presenta un rendimiento del 100%, sin embargo a pesar de ser un algoritmo muy robusto, es sensible al generar reglas con datos que contengan valores nulos, por ello las reglas generadas por este algoritmo se construyeron tomando en cuenta valores nulos o perdidos representados por el signo *"?",* perdiendo el sentido lógico y dificultando la tarea de análisis [104], es por ello que ha sido descartado como uno de los mejores.

En pruebas de entrenamiento el porcentaje de clasificación es superior a las realizadas mediante validación cruzada, esto es lógico puesto que los mismos datos sirvieron de base para la generación de reglas, es por esto que se produce una sobreestimación de los resultados, sin embargo luego de aplicar la validación cruzada se obtuvo porcentajes bajos debido a la poca cantidad de datos para realizar el proceso de minería [105] (ver figura 72).

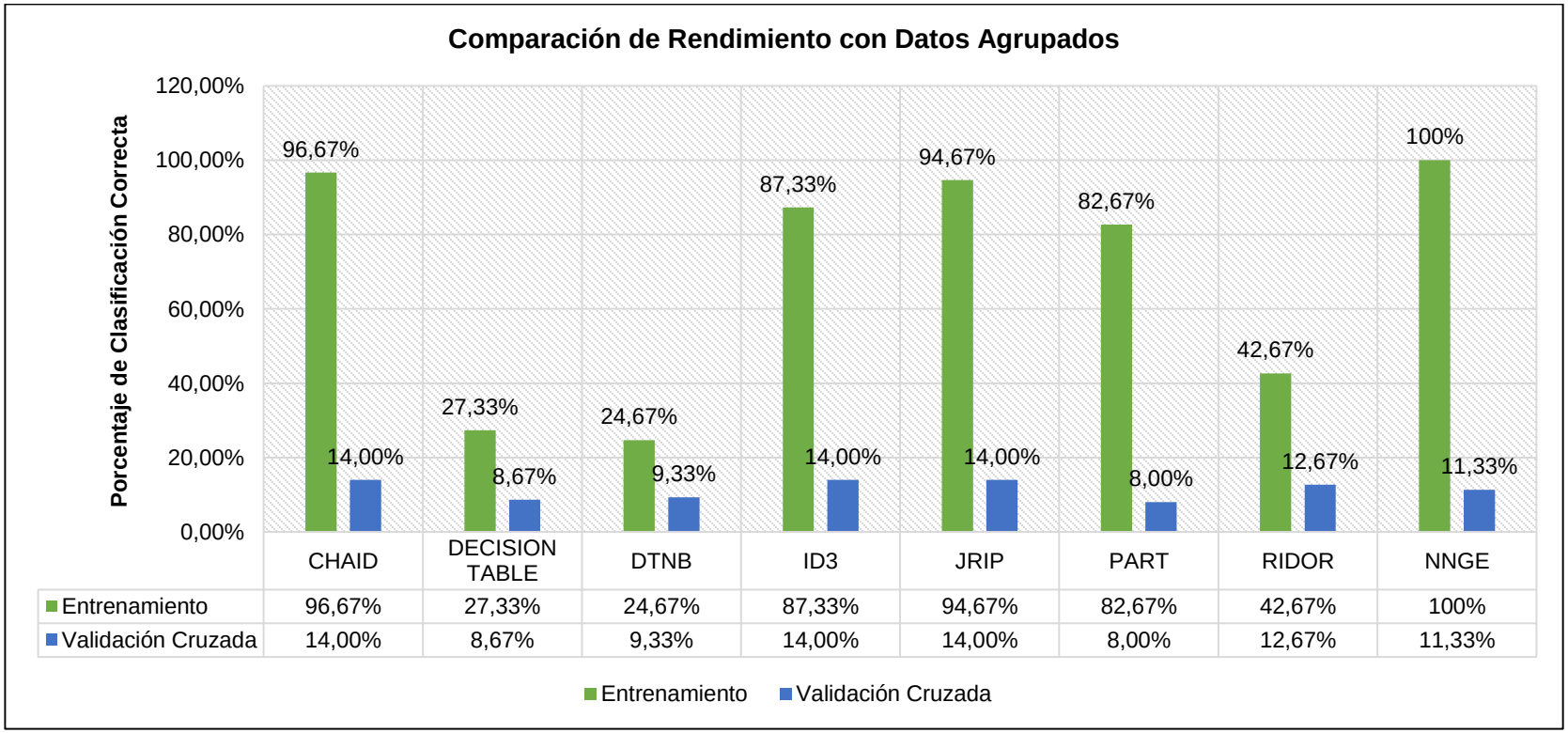


Figura 72. Comparación Rendimiento en pruebas de Entrenamiento y Validación Cruzada con datos agrupados.

En las pruebas de Validación Cruzada todos los algoritmos muestran porcentajes bajos en la clasificación debido a la poca cantidad de datos [105] y a la presencia de outliers o valores atípicos que le restan calidad a los datos [106], de la misma manera ocurrió en las pruebas anteriores sin agrupar unidades, en esta prueba disminuyo en 1% el mayor porcentaje de clasificación con el 14%, sin embargo esto no significó que los resultados obtenidos fueron descartados de inmediato.

- **Medidas de error de los algoritmos con mejor rendimiento con datos agrupados.**

La primera medida de error es el índice de Kappa, el algoritmo que presenta el mejor índice es: JRIP con 0.025, en otra medida de error denominada Error Absoluto tiene como mejor resultado también el obtenido por el algoritmo NNGE con 0.221, en la medida de error denominada Error Relativo igualmente tiene como mejor resultado el obtenido también por el algoritmo CHAID con un 86.56%, en la medida de error denominada Error Cuadrático Medio el mejor resultado fue el obtenido por el algoritmo NNGE con una medida de 0.470, mientras que en la última medida de error denominada Error Cuadrático Relativo el algoritmos que tiene el mejor resultado es DTNB con una medida de 0.403.

A continuación se compara el rendimiento de los algoritmos en pruebas de entrenamiento y validación cruzada con datos agrupados y no agrupados (ver figura 73).

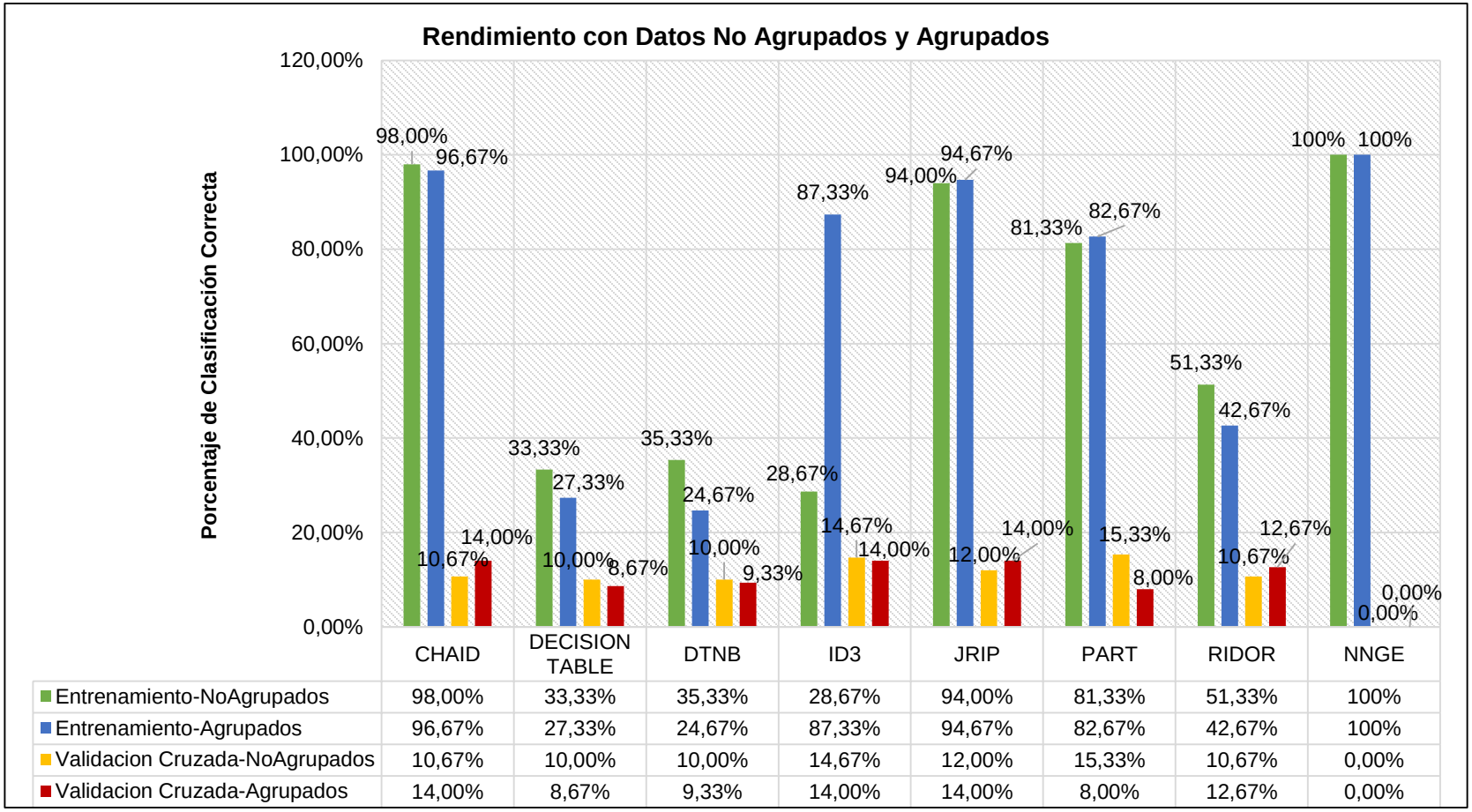


Figura 73. Comparación de rendimiento en pruebas de entrenamiento y validación cruzada con datos agrupados y no agrupados.

- **Comparación general de la evaluación de modelos con datos agrupados y no agrupados**

Como se puede observar en la figura anterior al agrupar las unidades los porcentajes de clasificación varían muy poco entre las dos estructuras, el único algoritmo que logro tener un considerable aumento en pruebas de entrenamiento es el ID3, este algoritmo obtuvo con datos no agrupados un porcentaje de 28% y luego de agrupar los datos aumento hasta 87% en clasificación.

Por lo tanto se ha notado que realizar la agrupación en los datos no es tan esencial, debido a que no existe un cambio significativo en los resultados, por lo que es mucho mejor trabajar con los datos originales sin realizar una alteración.

Hasta el momento hemos analizado que PART está entre uno de los dos mejores seguido de CHAID, debido al análisis de las medidas de error, sin embargo estas medidas provienen de la validación cruzada que es baja debido a los pocos datos [105], por lo tanto se ha considerado muy importante que el algoritmo JRip ha presentado un porcentaje de clasificación del 94% superior al del algoritmo PART que le corresponde el 81.33%, a su vez el Algoritmo JRip ha producido reglas más entendibles que PART.

Por lo tanto se ha llegado a la considerar al algoritmo JRip como uno de los mejores junto con CHAID. Es por ello y en base a estos resultados obtenidos se decidió tomar en cuenta los dos algoritmos: CHAID y JRip de la estructura de los datos no agrupados, ya que presentan los mejores resultados tanto en rendimiento, clasificación de la clase perfil profesional y las medidas de error [106], no se tomó en cuenta el algoritmo NNge ya que las reglas generadas por este toman en cuenta los valores nulos por lo tanto son difíciles de interpretar y de utilizar [104].

Cabe mencionar que los procesos para generar los algoritmos descritos anteriormente se encuentran detallados en el anexo 10 del presente trabajo de titulación.

4. ETAPA CUATRO: Interpretar y evaluar el modelo de minería de datos y aplicación en un contexto real

4.1. Quinta Fase: Evaluación de resultados obtenidos

En esta fase se realiza el análisis de los resultados obtenidos en la minería, mediante una evaluación de los modelos, detallando los resultados más positivos, para ello se ha tomado en cuenta el 28% restante de los datos, con la finalidad de observar las coincidencias entre los perfiles profesionales, y corroborar la validez de los algoritmos elegidos como lo más óptimos, y determinar finalmente el mejor.

4.1.1. Evaluación de algoritmos CHAID y JRip

En este apartado se realizó un análisis del rendimiento con los mejores algoritmos en este caso CHAID y JRip, los datos utilizados para realizar esta son 28% restante de los recopilados inicialmente es decir se trabajara con notas de 58 estudiantes egresados y graduados de la Carrera de Ingeniería en Sistemas.

4.1.1.1. Evaluación de Algoritmo JRip

A continuación se muestra los resultados obtenidos con este algoritmo, clasificando el 91.38% de las instancias (ver figura 74):

accuracy: 91.38%

	true Administrador	true Administrador	true Auditor Inform	true Desarrollador	true Arquitecto y D	true Analista de Si	true Especialista e	true Administrador	class precision
pred. Administrad	13	0	0	0	0	0	0	0	100.00%
pred. Administrad	0	9	2	0	0	0	1	2	64.29%
pred. Auditor Infor	0	0	4	0	0	0	0	0	100.00%
pred. Desarrollad	0	0	0	4	0	0	0	0	100.00%
pred. Arquitecto y [	0	0	0	0	5	0	0	0	100.00%
pred. Analista de S	0	0	0	0	0	6	0	0	100.00%
pred. Especialista	0	0	0	0	0	0	7	0	100.00%
pred. Administrad	0	0	0	0	0	0	0	5	100.00%
class recall	100.00%	100.00%	66.67%	100.00%	100.00%	100.00%	87.50%	71.43%	

Figura 74. Matriz de confusión obtenida con el algoritmo JRip con datos de prueba.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, en donde se puede observar que solamente el perfil profesional *Administrador de Redes computacionales* tiene el 64.29% de precisión mientras que para los restantes siete perfiles hay el 100% de precisión al clasificar las instancias.

A continuación se describe un fragmento de los valores de predicción realizada a los estudiantes (ver figura 75).

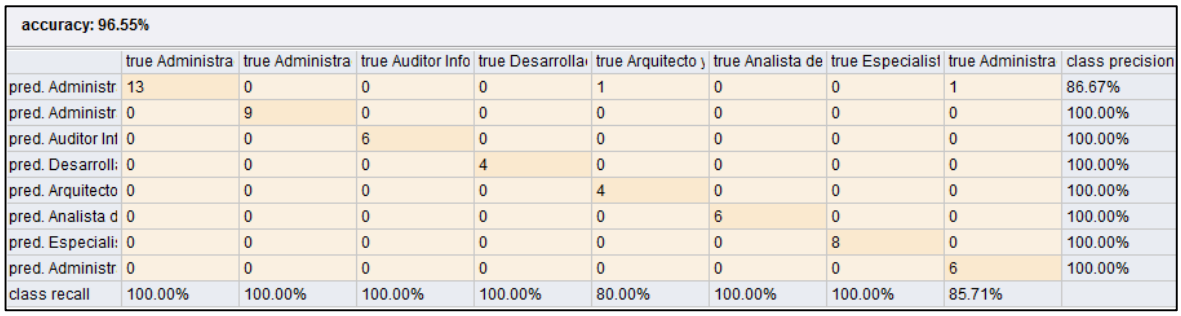
cedula	perfil_profesional	confid...	confiden...	confide...	confid...	confi...	confide...	confi...	confide...	prediction(perfil_profesional)
0705034155	Administrador de Redes computacionales	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104665136	Administrador de Redes computacionales	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104459365	Auditor Informático	0	0	1	0	0	0	0	0	Auditor Informático
1104469448	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1104418098	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1103840904	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1900615715	Especialista en mantenimiento hardware y software	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
1104600398	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1104665110	Especialista en mantenimiento hardware y software	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
1104377435	Especialista en mantenimiento hardware y software	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
1104307895	Administrador de Redes computacionales	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104742703	Auditor Informático	0	0	1	0	0	0	0	0	Auditor Informático
1104058423	Analista de Sistemas de Información	0	0	0	0	0	1	0	0	Analista de Sistemas de Información
1104580939	Administrador de Redes computacionales	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104012792	Especialista en mantenimiento hardware y software	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
1104401961	Especialista en mantenimiento hardware y software	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
1104426901	Auditor Informático	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104107196	Administrador de Redes computacionales	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1103803233	Administrador de Centros de Cómputo	0	0	0	0	0	0	0	1	Administrador de Centros de Cómputo
1103740054	Desarrollador de software	0	0	0	1	0	0	0	0	Desarrollador de software
1103378125	Auditor Informático	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1104335417	Administrador de Centros de Cómputo	0	0	0	0	0	0	0	1	Administrador de Centros de Cómputo
1104241284	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1105139453	Administrador de Sistemas de Bases de Datos	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos

Figura 75. Predicción de Perfil Profesional por el algoritmo JRip.

La figura anterior muestra el grado de predicción para cada perfil profesional y define para cada estudiante al mayor de los valores distribuidos.

4.1.1.2. Evaluación de Algoritmo CHAID

A continuación se muestra los resultados obtenidos con este algoritmo, clasificando el 96.55% de las instancias (ver figura 76):



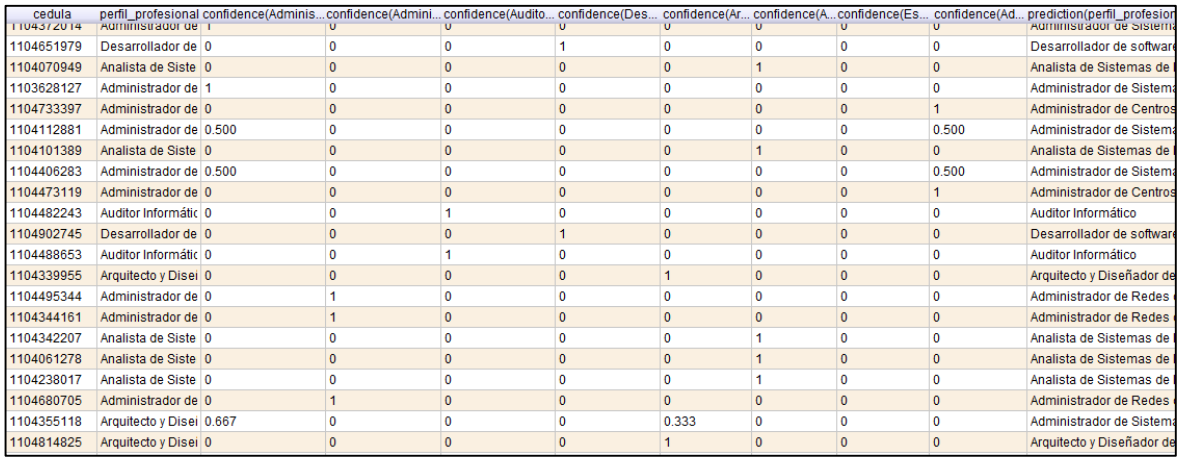
accuracy: 96.55%

	true Administra	true Administra	true Auditor Info	true Desarrolla	true Arquitecto y	true Analista de	true Especialist	true Administra	class precision
pred. Administr	13	0	0	0	1	0	0	1	86.67%
pred. Administr	0	9	0	0	0	0	0	0	100.00%
pred. Auditor Int	0	0	6	0	0	0	0	0	100.00%
pred. Desarroll:	0	0	0	4	0	0	0	0	100.00%
pred. Arquitecto	0	0	0	0	4	0	0	0	100.00%
pred. Analista d	0	0	0	0	0	6	0	0	100.00%
pred. Especiali:	0	0	0	0	0	0	8	0	100.00%
pred. Administr	0	0	0	0	0	0	0	6	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	80.00%	100.00%	100.00%	85.71%	

Figura 76. Matriz de confusión obtenida con el algoritmo CHAID con datos de prueba.

La matriz de confusión generada describe el porcentaje de precisión para las clases definidos como los perfiles profesionales, en donde se puede observar que el siete de los ocho perfiles profesionales se clasifico correctamente el 100%, mientras que el perfil profesional *Administrador de Sistemas de Bases de Datos,* solo obtuvo el 86.67% de instancias clasificadas.

A continuación se describe un fragmento de los valores de predicción realizada a los estudiantes (ver figura 77)



cedula	perfil_profesional	confidence(Adminis...	confidence(Admini...	confidence(Audito...	confidence(Des...	confidence(Ar...	confidence(A...	confidence(Es...	confidence(Ad...	prediction(perfil_profesion
1104372014	Administrador de	1	0	0	0	0	0	0	0	Administrador de Sistema
1104651979	Desarrollador de	0	0	0	1	0	0	0	0	Desarrollador de software
1104070949	Analista de Siste	0	0	0	0	0	1	0	0	Analista de Sistemas de I
1103628127	Administrador de	1	0	0	0	0	0	0	0	Administrador de Sistema
1104733397	Administrador de	0	0	0	0	0	0	0	1	Administrador de Centros
1104112881	Administrador de	0.500	0	0	0	0	0	0	0.500	Administrador de Sistema
1104101389	Analista de Siste	0	0	0	0	0	1	0	0	Analista de Sistemas de I
1104406283	Administrador de	0.500	0	0	0	0	0	0	0.500	Administrador de Sistema
1104473119	Administrador de	0	0	0	0	0	0	0	1	Administrador de Centros
1104482243	Auditor Informátic	0	0	1	0	0	0	0	0	Auditor Informático
1104902745	Desarrollador de	0	0	0	1	0	0	0	0	Desarrollador de software
1104488653	Auditor Informátic	0	0	1	0	0	0	0	0	Auditor Informático
1104339955	Arquitecto y Diseı	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de
1104495344	Administrador de	0	1	0	0	0	0	0	0	Administrador de Redes
1104344161	Administrador de	0	1	0	0	0	0	0	0	Administrador de Redes
1104342207	Analista de Siste	0	0	0	0	0	1	0	0	Analista de Sistemas de I
1104061278	Analista de Siste	0	0	0	0	0	1	0	0	Analista de Sistemas de I
1104238017	Analista de Siste	0	0	0	0	0	1	0	0	Analista de Sistemas de I
1104680705	Administrador de	0	1	0	0	0	0	0	0	Administrador de Redes
1104355118	Arquitecto y Diseı	0.667	0	0	0	0.333	0	0	0	Administrador de Sistema
1104814825	Arquitecto y Diseı	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de

Figura 77. Predicción de Perfil Profesional por el algoritmo CHAID.

Como se puede observar este algoritmo define claramente el perfil profesional de cada estudiante, es decir existen pocos estudiantes que se dividen en proporciones para cada perfil profesional.

- **Comparación de los resultados de la Evaluación de los Algoritmos CHAID y JRip.**

La evaluación realizada del rendimiento de los modelos generados por los mejores algoritmos CHAID y JRip, han presentado algunas variaciones, las cuales podemos observar y realizar una comparación en la tabla CXLIV.

TABLA CXLIV

RESULTADOS DE LA EVALUACIÓN DE LOS MODELOS GENERADOS CON CHAID Y JRip.

PERFIL PROFESIONAL	Instancias bien clasificadas (%)	
	JRip	CHAID
Analista de Sistemas de Información	100	100
Arquitecto y Diseñador de Software.	100	100
Desarrollador de software	100	100
Administrador de Sistemas de Bases de Datos	100	86.67
Auditor Informático	100	100
Administrador de Centros de computo	100	100
Administrador de Redes computacionales.	64.29	100
Especialista en mantenimiento hardware y software.	100	100

A continuación se describe el rendimiento de los dos algoritmos aplicados de manera gráfica para una mejor interpretación (ver figura 78).

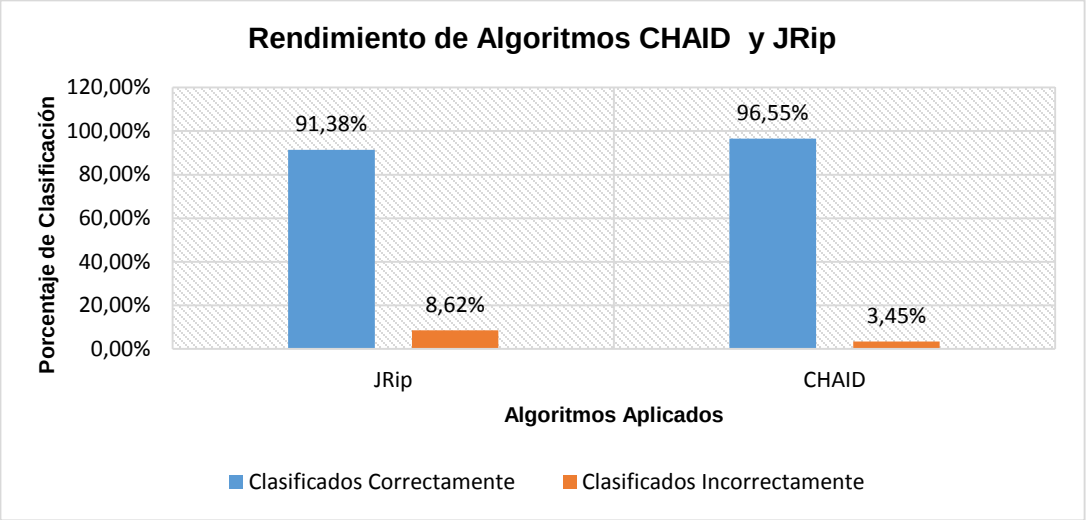


Figura 78. Rendimiento de algoritmos CHAID y JRip en predicción de perfiles profesionales.

Como se puede observar en la figura anterior el algoritmo que mejor rendimiento presentó ha sido CHAID logrando clasificar el 96.55% de las instancias mientras que el algoritmo JRip clasificó el 91.38%; en base a estos resultados obtenidos se ha seleccionado al algoritmo CHAID como el óptimo para predecir los perfiles profesionales de estudiantes de la carrera Ingeniería en Sistemas.

4.1.2. Aplicación de los modelos de minería de datos en un contexto real.

Evaluando los modelos y asegurando su calidad, se ha realizado la aplicación de los mismos en nuevos datos que corresponden a los últimos egresados de la carrera de ingeniería en sistemas año 2014, para probar su rendimiento en un contexto real y tomar la decisión final. Los modelos utilizados provienen de los algoritmos CHAID y JRip, considerando hasta el momento que CHAID ha resultado ser el más óptimo seguido de JRip.

4.1.2.1. Aplicación del Algoritmo JRip para la evaluación final.

A continuación se describe un fragmento de los valores de predicción realizada a los estudiantes del último año de la carrera Ingeniería en Sistemas (ver figura 79).

cedula	perfil_profe...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	prediction(perfil_profesional)
1104121346	?	0	0	0	0	0	1	0	0	Analista de Sistemas de Información
1105035941	?	0	0	0	1	0	0	0	0	Desarrollador de software
172095834	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
110378149	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1105027054	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
110502887	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
110500384	?	0	0	0	1	0	0	0	0	Desarrollador de software
1104999147	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1105137143	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1103720122	?	0	0	0	0	0	1	0	0	Analista de Sistemas de Información
110495270	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
110357084	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales
1103535991	?	0	0	1	0	0	0	0	0	Auditor Informático
110479368	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
1104598113	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
110487188	?	0	0	0	0	0	0	0	1	Administrador de Centros de Cómputo
1104009871	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
110461596	?	0	0.743	0.057	0.029	0	0.029	0.057	0.086	Administrador de Redes computacionales

Figura 79. Predicción de Perfil Profesional por el algoritmo JRip de los últimos egresados CIS año 2014.

La figura 79 muestra el grado de predicción para cada perfil profesional y define el grado de aproximación para cada estudiante, se puede observar que el algoritmo genera la predicción de la mayoría de estudiantes con el valor de 1 para el perfil profesional que es más cercano. A continuación se describe la cantidad de perfiles profesionales identificados (ver figura 80).

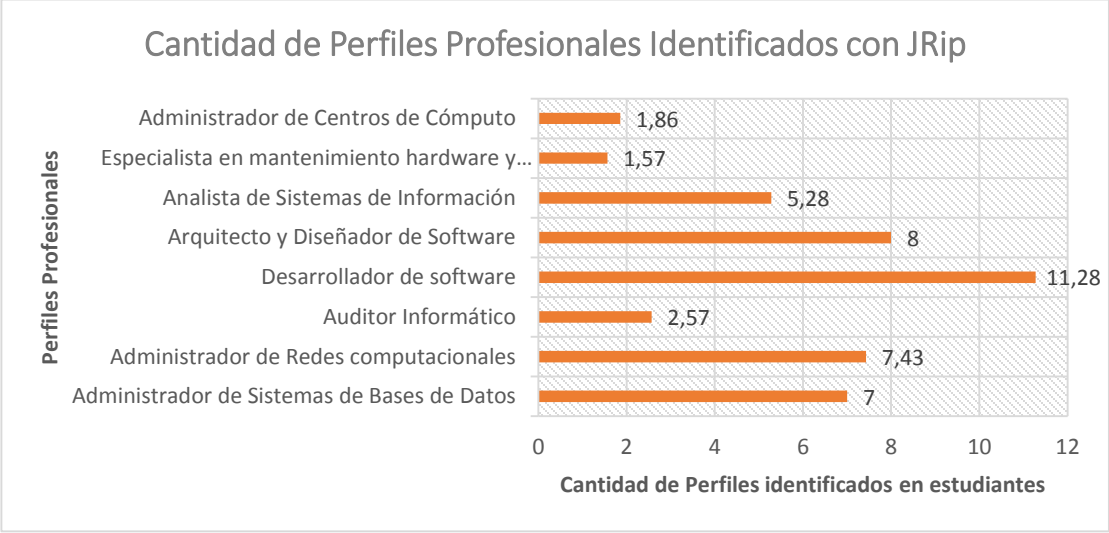


Figura 80. Cantidad de Perfiles Profesionales identificados de los últimos egresados CIS año 2014.

La figura 80 muestra que no se han dado instancias no clasificadas y que existen un gran número de desarrolladores de software (11), mientras que los otros perfiles profesionales están entre valores de 1 y 8.

4.1.2.2. Aplicación del Algoritmo CHAID para la evaluación final.

A continuación se describe un fragmento de los valores de predicción realizada a los estudiantes del último año de la carrera Ingeniería en Sistemas (ver figura 81).

ExampleSet (45 examples, 11 special attributes, 65 regular attributes)

Row No.	cedula	perfil_profesional	confid...	confiden...	confide...	confide...	co...	co...	co...	confide...	prediction(perfil_profesional)
1	1104121346	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
2	1105035941	?	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
3	172095834[illegible]	?	?	?	?	?	?	?	?	?	Administrador de Sistemas de Bases de Datos
4	1103781496	?	0	1	0	0	0	0	0	0	Administrador de Redes computacionales
5	1105027054	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
6	1105028870	?	0	0	1	0	0	0	0	0	Auditor Informático
7	1105003840	?	0	0.500	0	0.500	0	0	0	0	Desarrollador de software
8	1104999147	?	0	0	0	1	0	0	0	0	Desarrollador de software
9	1105137143	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
10	1103720122	?	?	?	?	?	?	?	?	?	Administrador de Sistemas de Bases de Datos
11	1104952708	?	?	?	?	?	?	?	?	?	Administrador de Sistemas de Bases de Datos
12	1103570840	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
13	1103535991	?	0	0	0	0	0	0	1	0	Especialista en mantenimiento hardware y software
14	1104793680	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
15	1104598113	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
16	1104871882	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
17	1104009871	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
18	1104615966	?	?	?	?	?	?	?	?	?	Administrador de Sistemas de Bases de Datos
19	1105032161	?	1	0	0	0	0	0	0	0	Administrador de Sistemas de Bases de Datos
20	0105220347	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
21	1104969033	?	0	0	1	0	0	0	0	0	Auditor Informático
22	1105154502	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
23	0705379287	?	0	0	0	0	1	0	0	0	Arquitecto y Diseñador de Software
24	1105582827	?	?	?	?	?	?	?	?	?	Administrador de Sistemas de Bases de Datos

Figura 81. Predicción de Perfil Profesional por el algoritmo CHAID de los últimos egresados CIS año 2014.

La figura 81 muestra el grado de predicción para cada perfil profesional y define el grado de aproximación para cada estudiante, se puede observar que el algoritmo generó la predicción de la mayoría de estudiantes con grados de aproximación siendo el de mayor valor el perfil profesional de cada estudiante.

A continuación se describe la cantidad de perfiles profesionales identificados con el algoritmo CHAID (ver figura 82).

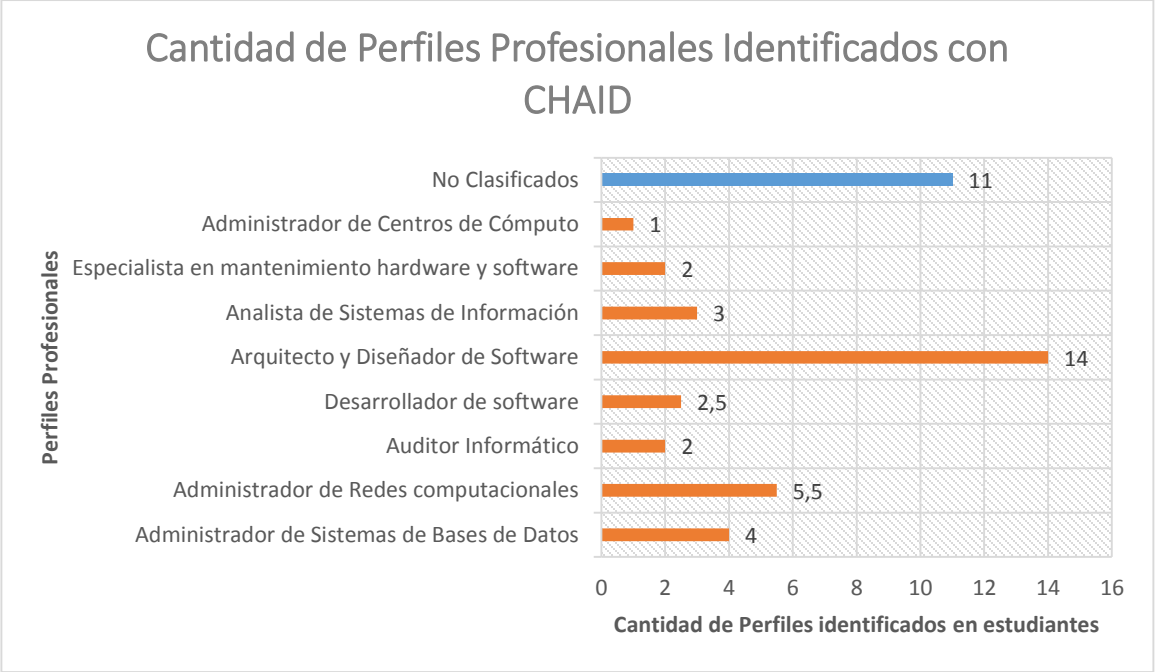


Figura 82. Cantidad de Perfiles Profesionales identificados de los últimos egresados CIS año 2014.

La figura muestra que no se han clasificado 11 estudiantes, además existen un gran número de arquitectos y diseñadores de software (14), mientras que los otros perfiles profesionales están entre 1 y 5.

Finalmente se ha comparado el porcentaje global de determinación de perfiles profesionales mediante los modelos CHAID y JRip para los últimos egresados de la carrera de ingeniería en sistemas (ver figura 83).

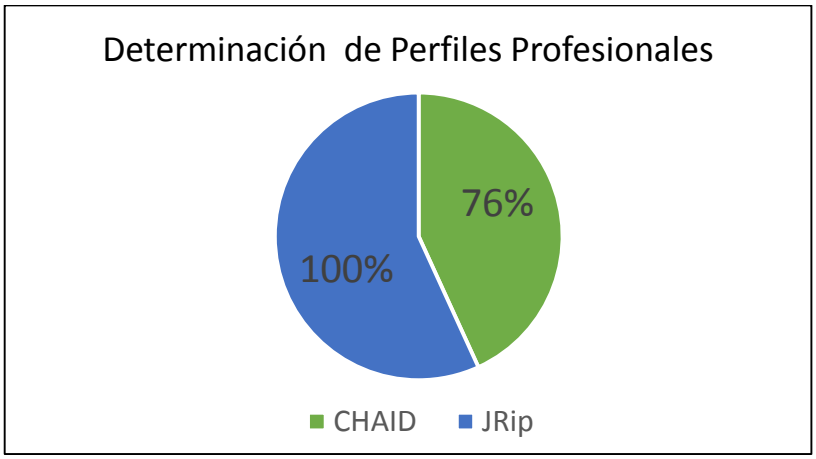


Figura 83. Porcentaje de Predicción final de los modelos CHAID y JRip.

Al observar los resultados de la figura 83 se demuestra que ha sido fundamental la aplicación de los modelos generados en un contexto real, debido a que JRip realiza la determinación de los perfiles profesionales en un 100%, mientras que CHAID lo hace en un 76%, cambiando la perspectiva y validando el modelo generado por el algoritmo JRip perteneciente a las técnicas de reglas de inducción como el óptimo para la determinación de perfiles profesionales en la carrera de ingeniería en sistemas de la UNL.

4.1.2.3. Perfiles Profesionales que se ajustan a los empleos desempeñados

De acuerdo a una encuesta aplicada en la herramienta online SurveyMonkey, de la población de egresados y graduados 100 respondieron la encuesta de los cuales el 50% aseguraron que están trabajando y colocaron el empleo en el que se están desempeñando actualmente, que corresponden al 24% de los egresados y graduados tomados para la determinación de los perfiles profesionales mediantes las técnicas de minería de datos (ver figuras 84).

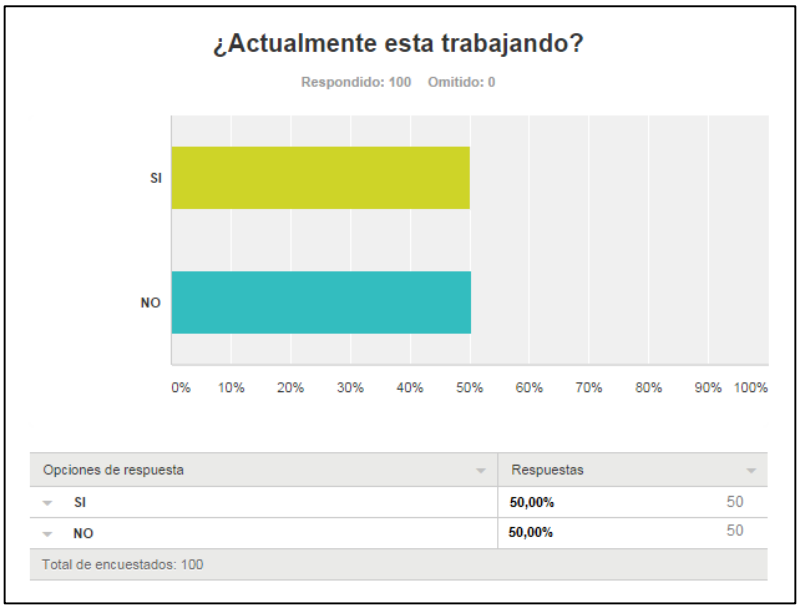


Figura 84. Gráfica de los resultados de la pregunta uno de la encuesta en la herramienta.

La figura 84 muestra que el 50% de los encuestados señalan que están trabajando, mientras que el otro 50% no lo está. La población que está trabajando tiene acceso a la siguiente pregunta con el fin de que especifiquen el empleo en el que se están desempeñando para realizar una comparación con los perfiles determinados (ver figura 85).

Figura 85. Respuestas del cargo de trabajo que desempeñan los egresados y graduados.

A continuación en la tabla CXLV se han recopilado los empleos especificados por cada participante y se han organizado de acuerdo al perfil profesional, es decir ubicando los empleos que se acercan y los que no se acercan o relacionan con cada uno de los perfiles profesionales determinados.

TABLA CXLV

ESPECIFICACIÓN DE LOS EMPLEOS POR CADA PERFIL PROFESIONAL

No.	Perfiles Profesionales	Se acercan al perfil	No se acercan al perfil
1	Analista de Sistemas de Información	Analista de Sistemas de Información	Secretaria, Técnica Informática, venta de equipos informáticos, Técnico en sistemas
2	Arquitecto y Diseñador de Software.	Diseñador de Software.	Analista desarrolladora de software, propietario de hospital computer (servicio técnico de tecnología), Desarrollador de software, Coordinadora del Patronato de Centinela del Cóndor
3	Desarrollador de software	Programador Junior, desarrollador de software	Asistente de Tecnología, técnico en computadores, venta de equipos informáticos, ayudante en ventas.
4	Administrador de Sistemas de Bases de Datos	oficial de seguridad de la información	Analista diseñador de pruebas de SW, secretaria, cajera, venta de equipos, mantenimiento de software
5	Auditor Informático	Auditor informático	Cajera, desarrollador de software
6	Administrador de Centros de computo	Administrando un centro de cómputo	Cajera, Soporte Técnico de Pc's, secretaria, mantenimiento
7	Administrador de Redes computacionales.	Analista Zonal TIC. Ministerio de Educación, mantenimiento de redes, Administrador de Red y Técnico	Desarrollador de aplicativos y sistemas informáticos para el Ejército Ecuatoriano, Técnico de soporte, secretaria, Venta y Reparación de Equipos Informáticos, programador de sistemas, Técnico en mantenimiento de computadoras, técnico, cajera en un supermercado,
8	Especialista en mantenimiento hardware y software.	Mantenimiento de Recursos informáticos	Programador de aplicaciones web, Desarrollo de Software, programador junior, Analista Zonal de TIC CZ7

En base a estos datos recopilados se ha realizado una comparación cuantitativa de los perfiles profesionales determinados a través del test de django con los encontrados con los empleos desempeñados recopilados de la encuesta aplicada (ver tabla CXLVI).

TABLA CXLVI
EMPLEOS QUE SE AJUSTAN A LOS PERFILES PROFESIONALES

sigla	Perfiles Profesionales	Cantidad de perfiles	Se ajustan al perfil	No se ajustan al perfil
AS	Analista de Sistemas de Información	8	3	5
ADS	Arquitecto y Diseñador de Software.	6	2	4
DS	Desarrollador de software	7	2	5
DBA	Administrador de Sistemas de Bases de Datos	6	1	5
AI	Auditor Informático	3	1	2
ACC	Administrador de Centros de computo	5	1	4
AR	Administrador de Redes computacionales.	10	3	7
MHS	Especialista en mantenimiento hardware y software.	5	1	4
TOTAL		50	14	36

Como podemos observar en la tabla CXLV se encuentra detallada la cantidad de egresados y graduados que existen por cada perfil y el número de empleos que se ajustan o acercan al perfil determinado, para tener una mejor perspectiva de estos resultados podemos observarlos representados en la figura 86.

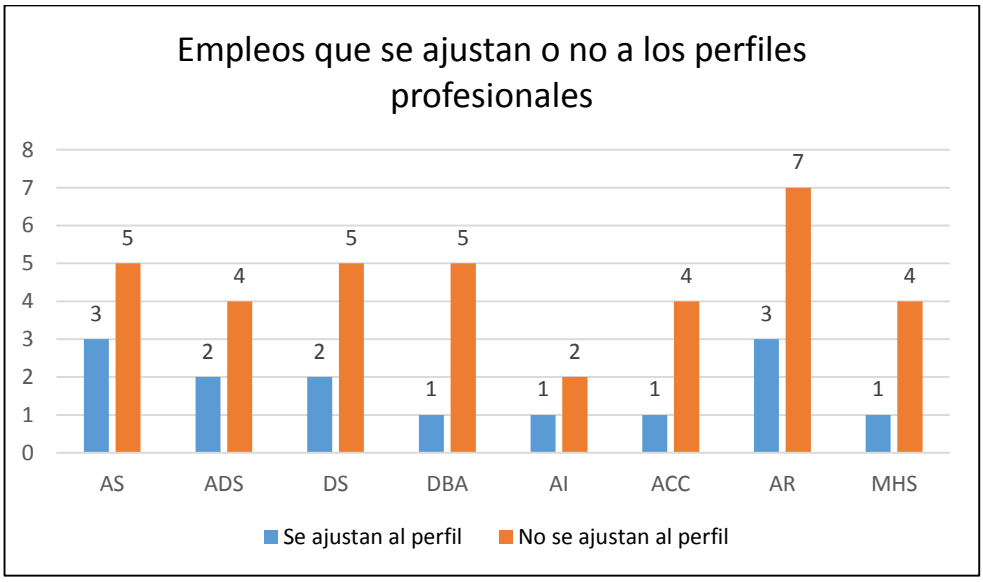


Figura 86. Cantidad de empleos que se ajustan por cada uno de los ocho perfiles.

Como podemos observar en la figura 86, para todos los perfiles profesionales es mayor la cantidad de los empleos que no se ajustan a cada perfil que los que se ajustan a los mismos, ahora realizamos un análisis de manera general respecto del porcentaje de empleos que se ajustan los perfiles profesionales (ver figura 87).

Figura 87. Porcentaje de los empleos que se ajustan a los perfiles profesionales de forma general.

La figura 87 nos muestra que el 72% de los empleos no se ajustan al perfil profesional, mientras que el 28% se acercan al perfil determinado, lo que nos dice que. Con esto se ha podido observar (ver la tabla CXLV) que los empleos que no se ajustan al perfil profesional determinado corresponden a otro de los perfiles planteados o a su vez se salen totalmente del contexto de las ramas de la carrera de ingeniería en sistemas.

Por lo que podemos decir que existen muy pocos egresados o graduados que están ejerciendo su profesión de acuerdo a su perfil profesional, por ello se nota lo necesario que es el conocer y hacer conocer todas sus cualidades personales, nivel de formación, experiencia, habilidades, capacidades e intereses que lo caracterizan y lo hacen diferente de los demás reflejadas en este perfil profesional, lo que implicaría que al tener conocimiento del mismo podrán postular y hacer una relación con cierta ocupación y/o vacante, con el fin de mejorar o asegurar sus probabilidades de éxito.

g. Discusión

- Investigar las características y variables más influyentes de las fuentes de datos a utilizar.

Para el desarrollo del presente trabajo de titulación se ha realizado el análisis de casos de éxito respecto a la aplicación de minería de datos en la resolución de problemas que se presentan en el ámbito educativo, lo que ha permitido conocer las técnicas de minería de datos existentes, observando sus potencialidades, esto con el fin de seleccionar las más adecuadas que se acoplen al problema, lo que ha conllevado a la obtención de resultados importantes y transcendentales.

Además se analizó las fuentes de datos que se van a utilizar para la determinación de los perfiles profesionales tomando como población los egresados y graduados de la carrera de ingeniería en sistemas, basándonos en el principio que el perfil profesional se ve reflejado en los conocimientos, habilidades, intereses y capacidades obtenidas durante la formación académica por parte del estudiante, por lo que se ha visto necesario tomar en cuenta datos cualitativos y cuantitativos.

Los datos cualitativos obtenidos de un test aplicado a la población con el fin de obtener el perfil profesional y realizar la predicción obteniendo un modelo para aplicarlo en nuevos datos futuros. La parte cuantitativa fue obtenida de los records académicos almacenados en del Sistema de Gestión Académica a través de la herramienta Web Service después de los permisos respectivos (ver anexo 2), la misma que fue completada manualmente con los datos históricos de los libros físicos que reposan en la secretaría del Área de la energía, las Industrias y los Recursos Naturales no Renovables (ver anexo 3).

- Comparar y seleccionar la técnica de minería de datos de acuerdo al ambiente de estudio.

Se ha realizado una recopilación bibliográfica de las diferentes herramientas de licencia libre para gestión de base de datos como de minería de datos, haciendo una comparación de las características de estas herramientas observando cual nos presenta mayores beneficios y se acopla a las necesidades del problema a resolver para la selección adecuada. Se manipularon las herramientas de minería de datos evaluando su rendimiento con datos de prueba y evidenciando la que se destaca en características, llegando a determinar DatAdmin como herramienta gestora de la base de datos como apoyo en la preparación, limpieza y generación de estructuras de los datos y RapidMiner para el proceso de minería propiamente.

Las técnicas fueron seleccionadas en base al análisis de fuentes bibliográficas y la meta de minería de datos que se pretende alcanzar, determinando como las adecuadas las técnicas de clasificación en base a árboles de decisión y las técnicas basadas en reglas de inducción.

- Diseñar el modelo de minería de datos en base a las técnicas seleccionadas.

La meta de minería de datos busca a partir de un conjunto de datos descubrir uno o varios modelos que determinen los perfiles profesionales mediante la aplicación de técnicas de data mining. Las fuentes las conforman datos cualitativos que corresponden al perfil profesional identificado del test aplicado el cual tuvo una acogida considerable del 80% de los egresados y graduados, en conjunto con los records académicos de los mismos participante, con lo que se realizó la construcción de dos estructuras de minería para datos no agrupados y agrupados con el fin de hacer varias pruebas y obtener los mejores resultados.

En esta fase se detallan las técnicas de modelado elegidas y los parámetros óptimos para cada algoritmo. Los algoritmos han sido distribuidos en los dos grupos de técnicas seleccionadas ID3 y CHAID que son en base a árboles de decisión ubicados en las técnicas de clasificación y los algoritmos JRip, PART, Ridor, Decisión Table, DTNB y NNge que pertenecen a las técnicas de reglas de inducción.

Con estos algoritmos se realizaron dos pruebas, la primera consistió en seleccionar un subconjunto de ejemplos denominado conjunto de entrenamiento que corresponde al 72% del total de ejemplos; la segunda prueba consiste en aplicar un método denominado validación cruzada mediante cinco subconjuntos, aplicadas para cada uno de las estructuras de minería construidas, de donde se obtuvo los siguientes resultados:

Al agrupar las unidades los porcentajes de clasificación varían muy poco entre las dos estructuras, el único algoritmo que logro tener un considerable aumento en pruebas de entrenamiento es el ID3, este algoritmo obtuvo con datos no agrupados un porcentaje de 28% y luego de agrupar los datos aumento hasta 87% en clasificación. Por lo tanto se ha notado que realizar la agrupación en los datos no es tan esencial, debido a que no existe un cambio significativo en los resultados, por lo que es mucho mejor trabajar con los datos sin su alteración.

En las pruebas de Validación Cruzada todos los algoritmos muestran porcentajes bajos en la clasificación debido a la poca cantidad de datos [105] y a la presencia de outliers o valores atípicos que le restan calidad a los datos [106], sin embargo esto no significó que los resultados obtenidos fueron descartados.

Después del análisis del porcentaje de clasificación, la matriz de confusión, la lógica de las reglas generadas, las medidas de error arrojadas, existe un algoritmo que muestra un rendimiento en clasificación del 100%, correspondiente a NNge, sin embargo no se tomó en cuenta este algoritmo ya que las reglas generadas por este toman en cuenta los valores nulos por lo tanto son difíciles de interpretar y de utilizar [104]. Finalmente los mejores algoritmos de generación de los modelos han resultado CHAID que está basado en árboles de decisión y JRip algoritmo de reglas de inducción.

- Interpretar y evaluar el modelo de minería de datos aplicado en un contexto real.

En esta etapa se han realizado tres tareas que han sumado sido de gran importancia en el desarrollo del trabajo de titulación, las cuales han son:

- ✓ Análisis de los resultados obtenidos en la minería, mediante una evaluación de los modelos, recalcando los resultados más positivos, para ello se ha tomado en cuenta el

28% restante de los datos, con la finalidad de observar las coincidencias entre los perfiles profesionales, y verificar la validez de los algoritmos elegidos CHAID y JRip elegidos hasta este punto como lo más óptimos. De donde ha resultado que el algoritmo que mejor rendimiento presentó fue CHAID logrando clasificar el 96.55% de las instancias mientras que el algoritmo JRip clasificó el 91.38%; la clasificación se refiere a las coincidencias entre los perfiles iniciales con la predicción generada por estos modelos, por ello el algoritmo CHAID continúa siendo el más óptimo para predecir los perfiles profesionales de estudiantes de la carrera Ingeniería en Sistemas.

- ✓ Una vez evaluado los modelos y asegurando su calidad, se ha realizado la aplicación de los mismo en nuevos datos que corresponden a los últimos egresados de la carrera de ingeniería en sistemas año 2014, para probar su rendimiento en un contexto real y tomar la decisión final. Los modelos utilizados provienen de los algoritmos CHAID y JRip, considerando hasta el momento que CHAID ha resultado ser el más óptimo seguido de JRip. De esta aplicación CHAID realizó una predicción del 76% mientras que el modelo generado por el algoritmo JRip es el de mejor rendimiento con una predicción del 100%, el cual pertenece a las técnicas de reglas de inducción. Debido a esta aplicación se ha verificado que es correcto tomar el modelo generado por el algoritmo JRip para la determinación de perfiles profesionales en la carrera de ingeniería en sistemas de la UNL. A su vez de la aplicación de los modelos podemos decir que el perfil profesional que más se ha destacado entre los últimos egresados de la carrera de ingeniería en sistemas es el de desarrollador de software.

- ✓ De acuerdo a una encuesta aplicada en la herramienta online SurveyMonkey, de la población de egresados y graduados 100 respondieron la encuesta de los cuales el 50% aseguraron que están trabajando y colocaron el empleo en el que se están desempeñando actualmente, que corresponden al 24% de los egresados y graduados tomados para la determinación de los perfiles profesionales mediantes las técnicas de minería de datos, de lo cual resultó que el 72% de los empleos no se ajustan al perfil profesional debido a que corresponden a otro de los perfiles planteados o a su vez se salen totalmente del contexto de especialidades de la carrera de ingeniería en sistemas. Mientras que el 28% se acercan al perfil determinado, porque podemos decir que existen

muy pocos egresados o graduados que están ejerciendo su profesión de acuerdo a su perfil profesional, por ello se nota lo necesario que es el conocer y hacer conocer todas sus cualidades personales, nivel de formación, experiencia, habilidades, capacidades e intereses que lo caracterizan y lo hacen diferente de los demás reflejadas en este perfil profesional, lo que implicaría que al tener conocimiento del mismo podrán postular y hacer una relación con cierta ocupación y/o vacante, con el fin de mejorar o asegurar sus probabilidades de éxito.

h. Conclusiones

Una vez finalizadas las etapas del presente trabajo de titulación, cumplimiento con los objetivos; se ha llegado a las siguientes conclusiones:

- La minería de datos hoy en día se ha convertido en una herramienta de vital importancia en el tratamiento, análisis y obtención de resultados del procesamiento de grandes cantidades de datos; para guiar la toma de decisiones tanto en las instituciones educativas como en todo tipo de organizaciones que así lo requieran.

- La tarea más costosa a lo largo del proyecto, tanto en tiempo como esfuerzo, fue la recolección y armado de la Base de Datos, puesto que se obtuvieron de una fuente digital así como de una física; lo que conllevó al análisis, clasificación y limpieza de la información para luego agruparlos en una sola Base.

- Posterior al desarrollo del presente proyecto se puede concluir la importancia de determinar el perfil profesional de los egresados y graduados con los factores determinantes como: el récord académico que muestra el desempeño del estudiante a lo largo de la carrera y la parte cualitativa de cada individuo, ésta última se determinó mediante la aplicación de una encuesta que permitió posteriormente contrastar sus conocimientos con sus habilidades, intereses y capacidades que los hacen únicos y candidatos potenciales y competentes a diferentes Áreas y temáticas dentro del mundo laboral.

- En presente trabajo de titulación desarrollado se evidenció que el perfil profesional que más predomina en los últimos egresados del año 2014 es desarrollador de software cuyos conocimientos obtenidos en las aulas universitarias serán puestos en práctica en el desarrollo de su vida profesional, ésta información es muy útil para los futuros cambios que se realicen a nivel de la malla curricular y perfil de carrera.

- La metodología que se utilizó para la Minería de datos en el presente proyecto fue CRISP-DM, la misma que ayudó a organizar mediante fases, sub fases y tareas que apoyaron a la documentación del proyecto a más de ser una guía para el desarrollo durante todo el proceso, permitiendo su culminación con éxito.

- Para el proceso de minería de datos se escogió los algoritmos ID3 y CHAID que pertenecen a las técnicas de clasificación basadas en árboles de decisión y los algoritmos JRip, PART, Ridor, Decisión Table, DTNB y NNge, pertenecientes al grupo de técnicas de reglas de inducción. Ya en el desarrollo y generación de modelos, los mejores algoritmos fueron CHAID y JRip los cuáles se hicieron con el 72% de los datos y con el 28% restante se hizo la evaluación de los mismos para verificar su validez, donde CHAID resultó el más óptimo al clasificar el 96.55% de las instancias; mientras que JRip clasificó el 91,38%. Posterior a ello se realizó la aplicación de éstos algoritmos en un contexto real para validar y realizar sí la elección final, en donde JRip tuvo el mejor rendimiento en la predicción con el 100%; mientras que CHAID realizó la predicción del 76%, llegando a la conclusión que JRip es el modelo que se debe aplicar para la obtención de los perfiles profesionales.

i. Recomendaciones

- De acuerdo a la experiencia adquirida en el desarrollo del presente proyecto, se recomienda actualizar, alimentar y mejorar el Sistema de Gestión Académica, puesto que se evidenció la ausencia de datos históricos de los récords académicos de los estudiantes, desde la creación de la carrera en 1999 hasta el año 2008 en el que se implementó el SGA.

- Al realizar el análisis y procesamiento de la información para la minería de datos se advirtió la importancia de recomendar que las mallas curriculares se actualicen con materias en cada periodo académico, es decir, sean mejoradas de acuerdo a los avances que representa seguir una carrera que va de la mano con el progreso tecnológico.

- Es importante en la fase de análisis de información, aplicar varias técnicas para la aplicación de minería de datos y en base a la comparación de resultados confirmar cuál de ellas resulta ser la más apropiada para obtener el resultado esperado.

- Para la generación de modelos de forma correcta y ordenada es importante que la técnica de minería de datos sea apropiada, es por ello que se recomienda la utilización de la metodología de CRISP - DM ya que se enfoca a proyectos de minería conformada por varias fases que permiten el desarrollo de forma ordenada del proyecto, hacia la consecución de los objetivos establecidos.

- Para la generación de los modelos para analizar los perfiles profesionales, es recomendable aplicar la técnica de reglas de inducción; puesto que se ha comprobado durante el desarrollo del presente trabajo que ésta técnica es con la que se obtiene mejores resultados en cuanto a las predicciones que se realizan en un contexto real.

- Se recomienda que se integre el modelo de minería de datos obtenido para la determinación de perfiles profesionales en el Sistema de Gestión Académica para que

el estudiante al culminar su carrera profesional conozca cuál es su perfil profesional, el mismo que le servirá para la toma de decisiones tanto en su vida profesional como laboral.

- Se recomienda mantener el programa de seguimiento a graduados en una constante y permanente actualización, con el fin de mantener ésta información para futuros estudios y aportes en beneficio de la universidad y de los profesionales egresados de la carrera de ingeniería de sistemas.

j. Bibliografía

Referencias Bibliográficas

[1] COBO O. Angel, ROCHA B. Rocío. *Selección de atributos predictivos del rendimiento académico de estudiantes en un modelo de B-Learning.* [En línea]: http://edutec.rediris.es/Revelec2/Revelec37/pdf/Edutec-e_n37_Cobo_Rocha_Alvarez.pdf. [Acceso: 21-Junio-2013].

[2] ECKERT Karina, SUÉNAGA Roberto. *Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD.* [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/27103/Documento_completo.pdf?sequence=1. [Acceso: 21-Junio-2013].

[3] ÁLVARO J. Galindo, ÁLVAREZ G. Hugo. *Minería de Datos en la Educación*. [En línea]: http://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf. [Acceso: 21-Junio-2013].

[4] ROMERO M. Cristóbal, VENTURA S. Sebastián, HERVÁS M. Cesar. *Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web*. [En línea]: http://www.lsi.us.es/redmidas/CEDI/papers/189.pdf. [Acceso: 6-Noviembre-2013].

[5] GARCÍA S. Enrique, ROMERO M. Cristóbal, VENTURA S. Sebastián, CASTRO Carlos. *Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning.* [En línea]: http://rita.det.uvigo.es/200805/uploads/IEEE-RITA.2008.V3.N1.A3.pdf. [Acceso: 8-Noviembre-2013].

[6] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica. *Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos*. [En línea]:

http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf. [Acceso: 10-Noviembre-2013].

[7] MÁRQUEZ V. Carlos, ROMERO M. Cristóbal, VENTURA S. Sebastián. *Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos.* [En línea]: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf. [Acceso: 12 Noviembre-2013].

[8] Horton, W. *Designing Web-Based Training.* John Wiley&Sons. 2000. [Acceso: 13-Noviembre-2013].

[9] Brusilovsky, P. *Adaptive and Intelligent Technologies for Web-based Education. Special Issue on Intelligent Systems and Teleteaching, Künstliche Intelligenz,* 4, 19-25. 1999. [Acceso: 13-Noviembre-2013].

[10] Perkowitz, M.; Etzioni, O. *Adaptive web sites: Automatically synthesizing web pages. National Conference on Artificial Intelligence.* WI. 1998. [Acceso: 13-Noviembre-2013].

[11] Scime, A. *Web Mining: Applications and Techniques. Idea Group.* 2004. [Acceso: 13-Noviembre-2013].

[12] Agrawal, R., et al, *"Fast discovery of association rules", In Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI Press*, 1996, pp. 307-328. [Acceso: 13-Noviembre-2013].

[13] Tobias S., *"Finding Association Rules That Trade Support Optimally against Confidence", Lecture Notes in Computer Science*, Vol. 2168, 2001, 424. [Acceso: 13-Noviembre-2013].

[14] A. Savasere, E. Omiecinski, and S. B. Navathe, *"An efficient algorithm for mining association rules in large databases," in Proceedings of 21st International*

Conference on Very Large Data Bases. VLDB, Sept. 11-15 1995, pp. 432–444. [En línea]: http://www.vldb.org/dblp/db/conf/vldb/. [Acceso: 14-Noviembre-2013].

[15] P. Scheuermann. *"Distributed web log mining using maximal large itemsets," Knowledge and Information Systems,* vol. 3, no. 4, Nov. 2001, pp. 389–404. [Acceso: 14-Noviembre-2013].

[16] D. B. Skillicorn and Y. Wang, *"Parallel and sequential algorithms for data mining using inductive logic," Knowledge and Information Systems,* vol. 3, no. 4, pp. 405–421, Nov. 2001. [Acceso: 14-Noviembre-2013].

[17] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu, *"Efficient mining of association rules in distributed databases," IEEE Transactions on Knowledge and Data Engineering,* vol. 8, no. 6, pp. 911–922, Dec. 1996. [Acceso: 14-Noviembre-2013].

[18] Rosta F., Brusilovsky, P., *"Social navigation support in a course recommendation system", Adaptive Hypermedia and Adaptive Web- Based Systems:* 4th International Conference, AH 2006, 2006, pp. 91-100. [Acceso: 14-Noviembre-2013].

[19] Tang T., McCalla, G., *"Smart Recommendation for an Evolving E- Learning System: Architecture and Experiment".* International Journal on E-Learning, 4(1), 2005, pp. 105-129. [Acceso: 14-Noviembre-2013].

[20] Terveen, L. and Hill, W., *"Beyond Recomendador Systems: Helping People Help Each Other".* In J. M. Carroll (Ed.) Human-Computer Interaction in the New Millennium, Addison-Wesley. ACM Press, New York, ch 22, 2001, pp. 487-509. [Acceso: 14-Noviembre-2013].

[21] Burke, R., *"Semantic ratings and heuristic similarity for collaborative filtering", In Proceedings of the Seventeenth National Conference on Artificial Intelligence,* Austin, Texas, July 30th–August 3rd, 2000. [Acceso: 14-Noviembre-2013].

[22] Breiman, L., Friedman, J., Losen, R., Stone, *Classification and Regression Trees. Wadsworth and Brooks,* 1984 (new edition 1993). [Acceso: 14-Noviembre-2013].

[23] MORENO G. María N, QUINTALES Miguel, GARCÍA P. Francisco J, POLO M. José M. *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software.* [En línea]: http://ceur-ws.org/Vol-84/paper4.pdf. [Acceso: 18-Noviembre-2013].

[24] SILVA Maximiliano. *Minería de Datos y descubrimiento del conocimiento.* [En línea]: http: //exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_de_Datos_y_KDD .pdf. [Acceso: 22-Noviembre-2013].

[25] FACENA Unne. *Teleprocesos y sistemas distribuidos.* [En línea]: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/SDataMining.pdf. [Acceso: 22-Noviembre-2013].

[26] LÓPEZ, C. P. (2007). *Minería de datos: técnicas y herramientas.* [En línea]: http://books.google.com.ec/books?id=wz-D_8uPFCEC. [Acceso: 26-Noviembre-2013].

[27] RUIZ S. Almudena, NISTAL G. Ismael. (2007. *Minería de Datos y Reconocimiento de Dígitos Manuscritos.* [En línea]: http://www.it.uc3m.es/~jvillena/irc/practicas/09-10/19mem.pdf. [Acceso: 29-Noviembre-2013].

[28] Instituto Superior Tecnológico de Teziutlán. *Minería de Datos. Área de oportunidades.* [En línea]:http://www.itsteziutlan.edu.mx/site2010/pdfs/2012/11/articulo_mineria_datos.pdf. [Acceso: 02-Diciembre-2013].

[29] FEBLES R. Juan P. *KDD y MD*. [En línea]: http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASH018e.dir/doc.pdf. [Acceso: 10-Diciembre-2013].

[30] GONZÁLEZ B. Jesús A. *Minería de Datos*. [En línea]: http://ccc.inaoep.mx/~jagonzalez/AI/Sesion13_Data_Mining.pdf. [Acceso: 10-Diciembre-2013].

[31] SOTO Antonio. *Minería de Datos*. [En línea]: http://www.slideshare.net/antoniosql/minera-de-datos. [Acceso: 10-Diciembre-2013].

[32] MARTÍNEZ Á. Clemente A. *Aplicación de Técnicas de Minería de Datos para mejorar el Proceso de Control de Gestión en Entel*. Universidad de Chile. [En línea]: http://www.tesis.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1. [Acceso: 15-Diciembre-2013].

[33] Leonardo M. Tito, Felipe Mullicundo, Bernabeu Ricardo Dario. *RapidMiner: Tutorial online + Operadores.* [En línea]: http://www.dataprix.com/files/RapidMiner_Tutorial_online_Operadores.pdf. [Acceso: 15-Diciembre-2013].

[34] LARRAÑAGA Pedro, INZA Iñaki, MOUJAHID Abdelmalik. *Árboles de clasificación.* Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad del País Vasco - Euskal Herriko Unibertsitatea. [Acceso: 18-Diciembre-2013].

[35] GALVIS Mario, MARTÍNEZ Fabricio. *Confrontación de dos técnicas de minería de datos aplicadas a un dominio específico.* [En línea]: http://www.javeriana.edu.co/biblos/tesis/ingenieria/Tesis184.pdf. [Acceso: 21-Agosto-2013]. [Acceso: 20-Diciembre-2013].

[36] MARTÍNEZ A. CLEMENTE A. *Aplicación de Técnicas de Minería de datos para mejorar el proceso de control de gestión en Entel.* [En línea]: www.tesis.uchile.cl/handle/2250/112065. [Acceso: 20-Diciembre-2013].

[37] TWO CROWS CORPORATION. 1999. *Introduction to Data Mining and Knowledge Discovery. 3° ed*. Two Crows Corporation. [Acceso: 20-Diciembre-2013].

[38] COTONIETO J. León. *Estudio del desempeño de modelos de minado de datos implementados con SQL, funciones definidas por el usuario y funciones nativas.* [En línea]: http://148.204.210.201/tesis/1323991798111TesisJavierLeon.pdf. [Acceso: 21-Diciembre-2013].

[39] COHEN W. W. Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference on Machine Learning*; 1995; 1995. [Acceso: 22-Diciembre-2013].

[40] RIVEST, R. *Learning Decision Lists*. Machine Learning, 1987, 2(3), 229-246. [Acceso: 22-Diciembre-2013].

[41] Brian R. Gaines, Paul Compton (1995). *Induction of Ripple-Down Rules Applied to Modeling Large Databases*. J. Intell. Inf. Syst.. 5(3):211-228. [Acceso: 22-Diciembre-2013].

[42] VELANDIA O. Ronald A, HERNÁNDEZ S. Fredy L. *Evaluación de algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión.* [En línea]: http://tangara.uis.edu.co/biblioweb/tesis/2010/134742.pdf. [Acceso: 22-Diciembre-2013].

[43] MARTÍNEZ D. Zuleyka, MENÉNDEZ F. José, DEL POZO G. Eva M. *La inteligencia artificial como una alternativa viable en la predicción de la insolvencia de Empresa de Seguros.* [En línea]: www.ieaf.es/new/.../142_cf536507626439897feb4fa379184c02.html. [Acceso: 22-Diciembre-2013].

[44] Martin B. Instance-Based learning: Nearest Neighbor With Generalization [Master Thesis]. Hamilton, New Zealand: University of Waikato; 1995. [Acceso: 22-Diciembre-2013].

[45] LAZO María J, CARAGUAY Nina, CASTAÑEDA Claudia. *Inteligencia Artificial Avanzada*. [En línea]: http://es.slideshare.net/majitol/aplicacin-de-aprendizaje-automtico-en-minera-de-datos. [Acceso: 22-Diciembre-2013].

[46] *Inteligencia en Redes de Comunicaciones*. [En línea]: http://www.it.uc3m.es/jvillena/irc/practicas/03-04/21.mem.pdf. [Acceso: 22-Diciembre-2013].

[47] KOHAVI, Ron. *The power of decision tables*. En Machine Learning: ECML-95. Springer Berlin Heidelberg, 1995. p. 174-189. [Acceso: 22-Diciembre-2013].

[48] AGUILAR C. Nora M. *Aplicación de métodos de aprendizaje automático para la desambicuación del PP Attachmente en español.* [En línea]: http://161.116.36.206/~publicacions/research_reports/TIM_GRIAL_REPORT3.pdf. [Acceso: 22-Diciembre-2013].

[49] HALL, Mark; FRANK, Eibe. *Combining Naive Bayes and Decision Tables*. En FLAIRS Conference. 2008. p. 318-319. [Acceso: 22-Diciembre-2013].

[50] VALLEJOS Sofia J. *Minería de Datos.* [En línea]: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf. [Acceso: 22-Diciembre-2013].

[51] GONZÁLEZ S. Gustavo, DELFÍN A. Sonia, DE LA ROSA Josep L. *Preprocesamiento de bases de datos masivas y multi-dimensionales en minería de uso web para modelar usuarios: comparación de herramientas y técnicas con un caso de estudio*. [En línea]: http://www.lsi.us.es/redmidas/CEDI/papers/822.pdf. [Acceso: 3-Enero-2014].

[52] DRUCKER Peter. Revista Aprende Avanza contigo. *Seis herramientas gratuitas para análisis de datos.* [En línea]: http://www.revistamprende.com/gestion/item/139. [Acceso: 3-Enero-2014].

[53] Weka. *Página Oficial de documentación de Weka.* [En línea]: http://www.hakank.org/weka/. [Acceso: 3-Enero-2014].

[54] Knime. *Página de información de Knime*. [En línea]: http://tech.knime.org/knime-labs. [Acceso: 3-Enero-2014].

[55] CHELANOV S. *jHepWork a java-based analysis framework*. [En línea]: http://www.desy.de/dvsem/WS0708/chekanov_talk.pdf. [Acceso: 3-Enero-2014].

[56] GARCÍA G. Francisco J. *Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)* [En línea]: http://masteres.ugr.es/moea/pages/tfm-1213/tfm_garciagonzalezfrancisco_1/! [Acceso: 3-Enero-2014].

[57] RAMOS M. Andres A. *Análisis de la deserción de alumnos de una Universidad, mediante árboles de clasificación.* [En línea]: http://opac.ucv.cl/pucv_txt/pucv/Txt-9500/UCF9984_01.pdf. [Acceso: 3-Enero-2014].

[58] HERNÁNDEZ O. José. *Minería de Datos. Otros Aspectos.* [En línea]: http://users.dsic.upv.es/~jorallo/master/dm5.pdf. [Acceso: 5-Enero-2014].

[59] GALLARDO A. José A. *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* [En línea]: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf. [Acceso: 5-Enero-2014].

[60] Universidad Nacional de Loja. Reglamento de Rediseño de la carrera de Ingeniería en Sistemas. *Perfil de Egreso del Ingeniero en Sistemas.* [En línea]: https://drive.google.com/file/d/0By4B2OXR-vftTHpZV3F0a2NiNWs/edit?usp=sharing. [Acceso: 02-Noviembre-2013].

[61] Página Web de la Universidad Técnica del Norte. Ecuador. *Carrera Sistemas. Campo Ocupacional, CIS.* [En línea]: http://www.utn.edu.ec/fica/carreras/sistemas/?page_id=153. [Acceso: 08-Enero-2014].

[62] Página Web oficial de la Universidad Peruana de Ciencias e Informática. *Ingeniería de sistemas e Informática.* [En línea]: http://www.upci.edu.pe/facultades.php?ac=ci_ing&op=ing. [Acceso: 08-Enero-2014].

[63] Página web oficial de la Pontificia Universidad Católica del Ecuador. *Facultad de Ingeniería. Ingeniero en Sistemas.* [En línea]: http://www.puce.edu.ec/portal/content/Ingenier%C3%ADa%20en%20Sistemas/292;jsessionid=248B1730885DCEC54DFC19EC224DCF7F.node0?link=oln30.redirect. [Acceso: 09-Enero-2014].

[64] Página web oficial de la Facultad de Sistemas, México. *Ingeniero en Sistemas Computacionales.* [En línea]: http://www.sistemas.uadec.mx/index.php/carreras/isc. [Acceso: 09-Enero-2014].

[65] Página web oficial Universidad del Valle. Cede central Cochabamba - Bolivia. *Ingeniería de Sistemas Informáticos.* [En línea]: http://www.univalle.edu/index.php/facultades/informatica/sistemas. [Acceso: 09-Enero-2014].

[66] Página web oficial de la ESPE. Universidad de las Fuerzas Armadas. *Ingeniería de Sistemas e Informática.* [En línea]: http://www.espe.edu.ec/portal/portal/main.do?sectionCode=107. [Acceso: 09-Enero-2014].

[67] Instituto Tecnológico Superior de Irapuato. *Ingeniería en Sistemas Computacionales.* Carretera Irapuato - Silao Km. 12.5, C.P. 36821 Irapuato, Guanajuato, México. [En línea]: http://www.itesi.edu.mx/Oferta%20Educativa/Nivel%20Superior/IngSistemas.html. [Acceso: 09-Enero-2014].

[68] Universidad del Valle. Sede Central Cochabamba, Bolivia. *Ingeniería de Sistemas Informáticos.* [En línea]: http://www.univalle.edu/index.php/facultades/informatica/sistemas. [Acceso: 09-Enero-2014].

[69] KENDALL Kenneth E. KENDALL Julie E. *Análisis y diseño de sistemas*. [En línea]: http://books.google.com.ec/books?id=5rZA0FggusC&printsec=frontcover&hl=es#v=onepage&q&f=false. [Acceso: 10-Enero-2014].

[70] DOMINGUEZ C. Luis A. *Análisis de Sistemas de Información*. [En línea]: http://www.aliatuniversidades.com.mx/bibliotecasdigitales/pdf/sistemas/Analisis_de_sistemas_de_informacion.pdf. [Acceso: 10-Enero-2014].

[71] CÁCERES Ana M. *Análisis y Diseño de Sistemas.* Universidad Don Bosco. [En línea]: http://rcruz0423.galeon.com/docs/clase1.pdf. [Acceso: 10-Enero-2014].

[72] BARRAZA A. Fernando. *Modelado y Diseño de Arquitectura de Software.* [En línea]: http://cic.puj.edu.co/wiki/lib/exe/fetch.php?media=materias:s2_conceptosdemodelado.pdf. [Acceso: 10-Enero-2014].

[73] Instituto Politécnico Nacional. *Ingeniería en Sistemas Computacionales.* [En línea]: http://www.ipn.mx/educacionsuperior/Paginas/Ing-Sist-Compu.aspx. [Acceso: 10-Enero-2014].

[74] MINISTERIO de EDUCACIÓN CIENCIA Y TECNOLOGÍA. PRESIDENCIA DE LA NACIÓN. *Perfil Profesional Técnico en Programación V.6.* [En línea]: http://www.cicomra.org.ar/cicomra2/archivos/notas/Perfil%20Prof%20Tecnico%20Programador%20V6.pdf. [Acceso: 11-Enero-2014].

[75] Universidad de Colima. Facultad de telemática. *Ingeniero en Software.* Av. Universidad #333, Colima, Col. [Enlínea]: http://www.ucol.mx/acerca/coordinaciones/cgd/pregrado/PE/is.pdf. [Acceso: 12-Enero-2014].

[76] GÓMEZ S. Jorge. *Funciones del DBA (Administrador de Base de Datos) bajo la Plataforma Oracle.* [En línea]:

http://cdigital.uv.mx/bitstream/123456789/28520/1/Gomez%20Sanchez.pdf. [Acceso: 12-Enero-2014].

[77] Artículo de la Universidad Autónoma del Estado de Hidalgo. Área académica: Sistemas Computacionales. *Auditoría Informática.* [En línea]: http://www.uaeh.edu.mx/docencia/P_Presentaciones/tlahuelilpan/sistemas/auditoria_informatica/auditoria_informatica.pdf. [Acceso: 14-Enero-2014].

[78] PIATTINI V. Mario G, PESO N. Emilio. *Auditoría Informática*. Un enfoque práctico. Editor: Alfaomega. ISBN: 9701507312, 9789701507315. [Acceso: 14-Enero-2014].

[79] Echenique G. J.A. (2001). *Auditoría en Informática*. 2da. Ed. Mc Graw- Hill. [Acceso: 14-Enero-2014].

[80] RECALDE Ch. Tanya. *Administración de Centros de Cómputo*. [En línea]: http://www.ecotec.edu.ec/documentacion%5Cinvestigaciones%5Cdocentes_y_directivos%5Carticulos/5817_TRECALDE_00227.pdf. [Acceso: 14-Enero-2014].

[81] PALOM I. Francisco J*. La tecnología punta del mando.* Colección “”Produćtica”. Marcombo, S.A. 1989. Boxareu Editores. [Acceso: 14-Enero-2014].

[82] Escuela Ingeniería y Gestión. *Técnico en administración de redes Computacionales.* 600 4600 477 ipp.cl. [En línea]: http://www.ipp.cl/cl/descargas/tecnico-administracion-redes-computacionales.pdf. [Acceso: 15-Enero-2014].

[83] Universia Chile Estudios. *Administración de Redes Computacionales Duoc UC.* [En línea]: http://estudios.universia.net/chile/estudio/duoc-administracion-redes-computacionales. [Acceso: 15-Enero-2014].

[84] BERNAL A. Eduardo. *Mantenimiento de Hardware y Software*. [En línea]: http://tic.ucuenca.edu.ec/index.php/mantenimiento-de-hardwareysoftware-unid. [Acceso: 15-Enero-2014].

[85] ACEVEDO F. Eduardo. *Herramienta de gestión de base de datos SQLyog.* [En línea]: http://www.eduardoaf.com/blog-miscelanea/miscelanea-aplicaciones/herramienta-gestion-base-de-datos-sqlyog/. [Acceso: 20-Enero-2014].

[86] Página Oficial PhpMyAdmin*. PhpMyAdmin Bringing MySQL to the web*. [En línea]: http://www.phpmyadmin.net/home_page/index.php. [Acceso: 20-Enero-2014].

[87] ÁLVAREZ Miguel A. *PhpMyAdmin.* [En línea]: http://www.desarrolloweb.com/articulos/844.php. [Acceso: 22-Enero-2014].

[88] PIREÑO Rafael. MySQL *Workbench, editor visual de bases de datos MySQL*. [En línea] http://www.gizmos.es/programas-y-aplicaciones/mysql-workbench-editor-visual-de-bases-de-datos-mysql.html. [Acceso: 24-Enero-2014].

[89] JenaSoft. Database Management. *Administrador de datos eficiente herramienta de gestión de datos DatAdmin*. [En línea]: http://www.datadmin.com/features/data-admin-reference-browser. [Acceso: 24-Enero-2014].

[90] ASTORGA Nathalia, SALINAS Maruxa. *Weka para minería de datos*, 2013. [En línea]: http://prezi.com/_gli7zt6vv0t/weka-para-mineria-de-datos-2013/. [Acceso: 26-Enero-2014].

[91] CORSO Cynthia L, GIBELLINI Fabián. Facultad Regional Córdoba/Universidad Tecnológica Nacional. Argentina. *Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios.* [En línea]: http://41jaiio.sadio.org.ar/sites/default/files/16_JSL_2012.pdf. [Acceso: 26-Enero-2014].

[92] CUBERO Juan C, BERZAL Fernando. Departamento de Ciencias de la computación, Universidad de Granada. *Guión de prácticas de minería de datos, herramientas de minería de datos, KNIME.* [En línea]: http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf. [Acceso: 26-Enero-2014].

[93] ARRUELAS A. Sergio A. *Sistema Para La Prevención De Fraudes, Gestión Y Monitoreo Del Negocio.* [En línea]: http://www.dspace.uce.edu.ec/bitstream/25000/143/1/T-UCE-0011-8.pdf. [Acceso: 29-Enero-2014].

[94] FIGUEROA Mauricio. *Minería de Datos Aplicada a Credit Scoring.* [En línea]: http://repositorio.usfq.edu.ec/bitstream/23000/547/1/82579.pdf. [Acceso: 29-Enero-2014].

[95] Página Web Oficial de la Universidad Generalitat Valenciana. Conselleria d'Economia, Indústria, Turisme i Ocupació. Ciutat Administrativa 9 d'octubre. Torre 2. c/ Castán Tobeñas, 77 - 46018 Valencia. *Perfil Profesional.* [En línea]: http://www.recursoseees.uji.es/fichas/fc12.pdf. [Acceso: 05-Febrero-2014].

[96] Página Web Oficial Stad Center Ecuador. Ingeniería en Estadística. *Tutorías Universitarias.* [En línea]: http://www.stadcenterecuador.com/component/content/article/28-portada/42-tutorias-universitarias.html. [Acceso: 05-Febrero-2014].

[97] MOLINA L. José M, GARCIA H. Jesús. *Técnicas de Análisis de Datos.* [En línea]: tesis-algoritmo-c45.googlecode.com/files/131469066-apuntesAD.pdf. [Acceso: 06-Febrero-2014].

[98] ALUJA Tómas. *La minería de datos entre la estadística y la inteligencia artificial.* [En línea]: http://upcommons.upc.edu/revistes/bitstream/2099/4162/4/article.pdf. [Acceso: 20-Abril-2014].

[99] VIZCAINO G. Paula A. *Aplicación de técnicas de inducción de Árboles de Decisión a problemas de clasificación mediante el uso de weka.* [En línea]: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf. [Acceso: 20-Abril-2014].

[100] SERVENTE Magdalena. *Algoritmos TDIDT aplicados a la minería de datos inteligente.* [En línea]: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf. [Acceso: 21-Abril-2014].

[101] HERNÁNDEZ O. José. *Práctica 2 de minería de datos, profundizando en el Clementine.* [En línea]: http://users.dsic.upv.es/~jorallo/cursoDWDM/kdd-lab2.pdf. [Acceso: 21-Abril-2014].

[102] PAZ A. Henry P. *Publicaciones Henry. Weka with Data Mining done in java.* [En línea]: http://publicacioneshenry.wordpress.com/. [Acceso: 08-Febrero-2014].

[103] LABANDA Milton. *Sitema de Evaluación del Desempeño Docente*. [En línea]: https://github.com/miltonlab/sedd. [Acceso: 5-febrero-2014].

[104] KAISER Jiří. *Dealing with Missing Values in Data*. [En línea]: http://www.si-journal.org/index.php/JSI/article/viewFile/178/134. [Acceso: 12-abril-2014].

[105] ANDREW Y. Ng. *Preventing "Overfitting" of Cross-Validation Data*. [En línea]: http://ai.stanford.edu/~ang/papers/cv-final.pdf. [Acceso: 08-mayo-2014].

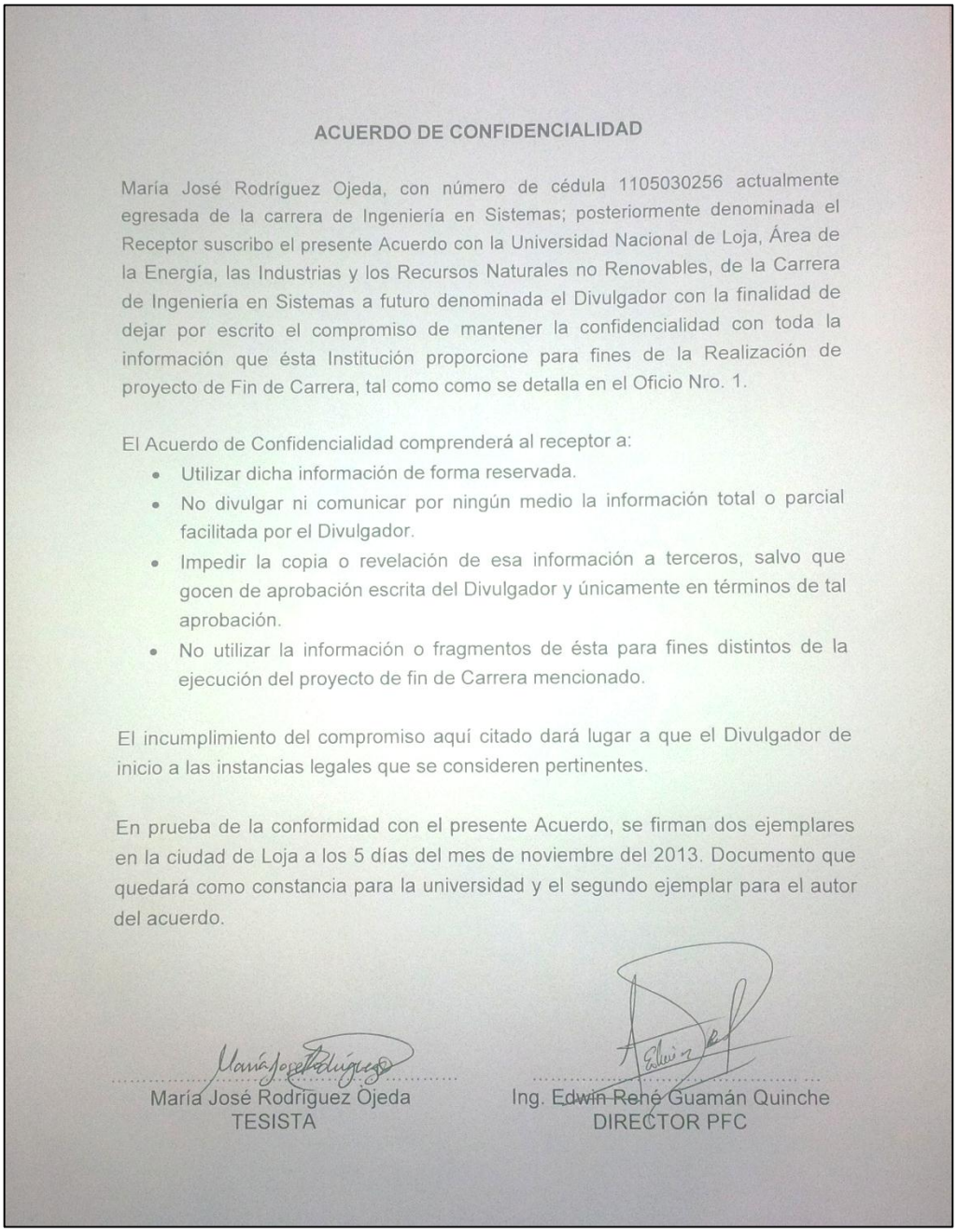
[106] HAN Jiawei, KAMBER Micheline. *Data Mining, Southeast Asia EditionL Concepts and Techniques*. Morgan Kaufmann, pp. 2006. [Acceso: 08-mayo-2014].

[107] DAPOZO Gladys, PORCEL Eduardo, LÓPEZ María V, BOGADO Verónica. *Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios.* [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/20453/Documento_completo.pdf?sequence=1. [Acceso: 08-junio-2014].

[108] MOINE Juan M. HAEDO Ana S. GORDILLO Silvia. Estudio Comparativo de metodologías para minería de datos. [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1. [Acceso: 05-enero-2014].

k. Anexos

Anexo 1: Acuerdo de confidencialidad para el acceso a la herramienta Web Service.



ACUERDO DE CONFIDENCIALIDAD

María José Rodríguez Ojeda, con número de cédula 1105030256 actualmente egresada de la carrera de Ingeniería en Sistemas; posteriormente denominada el Receptor suscribo el presente Acuerdo con la Universidad Nacional de Loja, Área de la Energía, las Industrias y los Recursos Naturales no Renovables, de la Carrera de Ingeniería en Sistemas a futuro denominada el Divulgador con la finalidad de dejar por escrito el compromiso de mantener la confidencialidad con toda la información que ésta Institución proporcione para fines de la Realización de proyecto de Fin de Carrera, tal como como se detalla en el Oficio Nro. 1.

El Acuerdo de Confidencialidad comprenderá al receptor a:

- Utilizar dicha información de forma reservada.
- No divulgar ni comunicar por ningún medio la información total o parcial facilitada por el Divulgador.
- Impedir la copia o revelación de esa información a terceros, salvo que gocen de aprobación escrita del Divulgador y únicamente en términos de tal aprobación.
- No utilizar la información o fragmentos de ésta para fines distintos de la ejecución del proyecto de fin de Carrera mencionado.

El incumplimiento del compromiso aquí citado dará lugar a que el Divulgador de inicio a las instancias legales que se consideren pertinentes.

En prueba de la conformidad con el presente Acuerdo, se firman dos ejemplares en la ciudad de Loja a los 5 días del mes de noviembre del 2013. Documento que quedará como constancia para la universidad y el segundo ejemplar para el autor del acuerdo.

María José Rodriguez Ojeda
TESISTA

Ing. Edwin René Guamán Quinche
DIRECTOR PFC

Figura 1. Documento del acuerdo de confidencialidad firmado.

Anexo 2: Permiso de acceso a la herramienta Web Service.

UNIVERSIDAD NACIONAL DE LOJA
UNIDAD DE TELECOMUNICACIONES E NFORMACIÓN

Of. N.620-UTI-UNL
Loja, 6 de Noviembre 2013

Sra.
María José Rodríguez Ojeda.
EGRESADA DE LA CARRERA DE INGENIERIA EN SISTEMAS

De mi Consideración

Por medio del presente me es grato saludarle y expresarle éxitos en las delicadas funciones, y a la vez hacerle llegar la clave y usuario con respecto al oficio emitido por su persona
Usuario.-
Clave.- [redacted]

Particular que pongo a su conocimiento para los fines pertinentes.

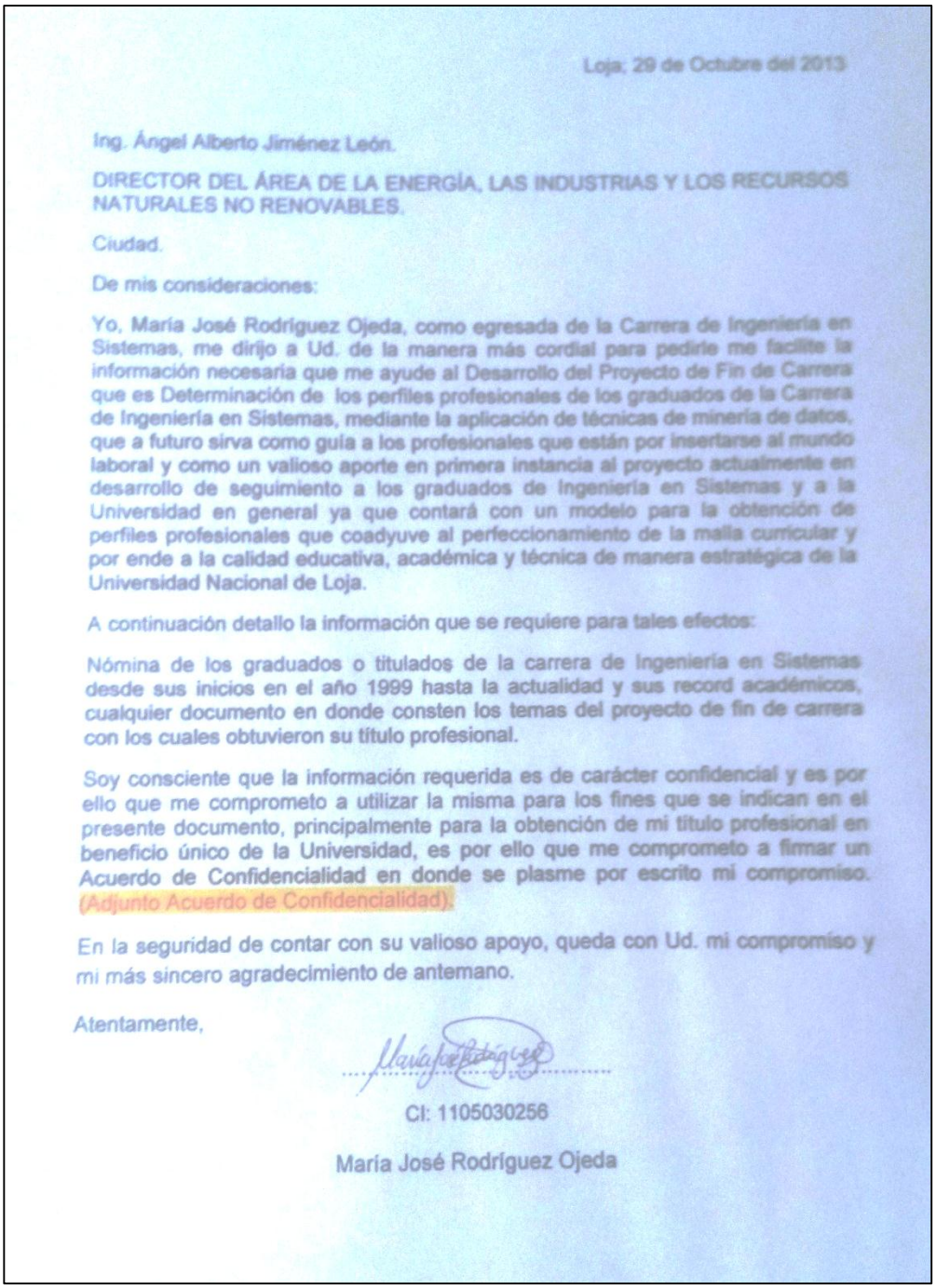
Atentamente,

Ing. Milton Palacios
DIRECTOR DE LA UNIDAD
DE TELECOMUNICACIONES E INFORMACIÓN

CIUDAD UNIVERSITARIA "GUILLERMO FALCONÍ ESPINOSA" La Argelia, Loja-Ecuador
PBX: (593 07 2) 547252 Casilla: Letra "S" E-mail: telecomunicaciones@unl.edu.ec Sitio Web: www.unl.edu.ec

Figura 1. Oficio positivo de respuesta al acceso del Web Service.

Anexo 3: Autorización de acceso a los libros físicos de la carrera de ingeniería en sistemas.



Loja; 29 de Octubre del 2013

Ing. Ángel Alberto Jiménez León.

DIRECTOR DEL ÁREA DE LA ENERGÍA, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES.

Ciudad.

De mis consideraciones:

Yo, María José Rodríguez Ojeda, como egresada de la Carrera de Ingeniería en Sistemas, me dirijo a Ud. de la manera más cordial para pedirle me facilite la información necesaria que me ayude al Desarrollo del Proyecto de Fin de Carrera que es Determinación de los perfiles profesionales de los graduados de la Carrera de Ingeniería en Sistemas, mediante la aplicación de técnicas de minería de datos, que a futuro sirva como guía a los profesionales que están por insertarse al mundo laboral y como un valioso aporte en primera instancia al proyecto actualmente en desarrollo de seguimiento a los graduados de Ingeniería en Sistemas y a la Universidad en general ya que contará con un modelo para la obtención de perfiles profesionales que coadyuve al perfeccionamiento de la malla curricular y por ende a la calidad educativa, académica y técnica de manera estratégica de la Universidad Nacional de Loja.

A continuación detallo la información que se requiere para tales efectos:

Nómina de los graduados o titulados de la carrera de Ingeniería en Sistemas desde sus inicios en el año 1999 hasta la actualidad y sus record académicos, cualquier documento en donde consten los temas del proyecto de fin de carrera con los cuales obtuvieron su título profesional.

Soy consciente que la información requerida es de carácter confidencial y es por ello que me comprometo a utilizar la misma para los fines que se indican en el presente documento, principalmente para la obtención de mi título profesional en beneficio único de la Universidad, es por ello que me comprometo a firmar un Acuerdo de Confidencialidad en donde se plasme por escrito mi compromiso. (Adjunto Acuerdo de Confidencialidad).

En la seguridad de contar con su valioso apoyo, queda con Ud. mi compromiso y mi más sincero agradecimiento de antemano.

Atentamente,

CI: 1105030256

María José Rodríguez Ojeda

Figura 1. Oficio de la autorización al acceso a los libros físicos CIS.

Anexo 4: Funcionalidad del Test Perfil Profesional.

El test ha sido desarrollado con el uso del framework django, que trabaja en base al lenguaje de programación python, para hacerlo automatizado y a través de acceso remoto, de esta forma facilitar la recolección de los datos con el uso de las tecnologías.

Las aplicaciones django permiten trabajar con una base de datos propia que se acopla a la estructura de la aplicación mientras se la va desarrollando. Este test también permite crear un sin número de test que se podrán usar con fines académicos.

A continuación se describe la interfaz de usuario, correspondiente al grupo de usuarios egresados, del test Perfil Profesional:

La figura 1 muestra la primera pantalla que corresponde a la autentificación del usuario, en este caso con su número de identificación como usuario y contraseña:

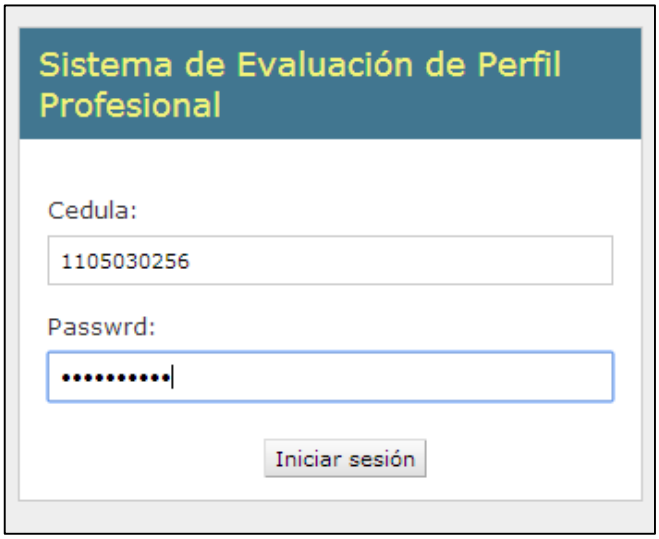


Figura 1. Pantalla de ingreso al test.

Una vez autentificado correctamente el usuario se carga su sesión y muestra la lista del o los test disponibles, en este caso el test para determinar el perfil profesional - CIS:

Figura 2. Pantalla de presentación del test disponible.

Al seleccionar el test correspondiente se carga el test con las preguntas correspondientes, y muestra en la cabecera los datos del usuario autentificado:

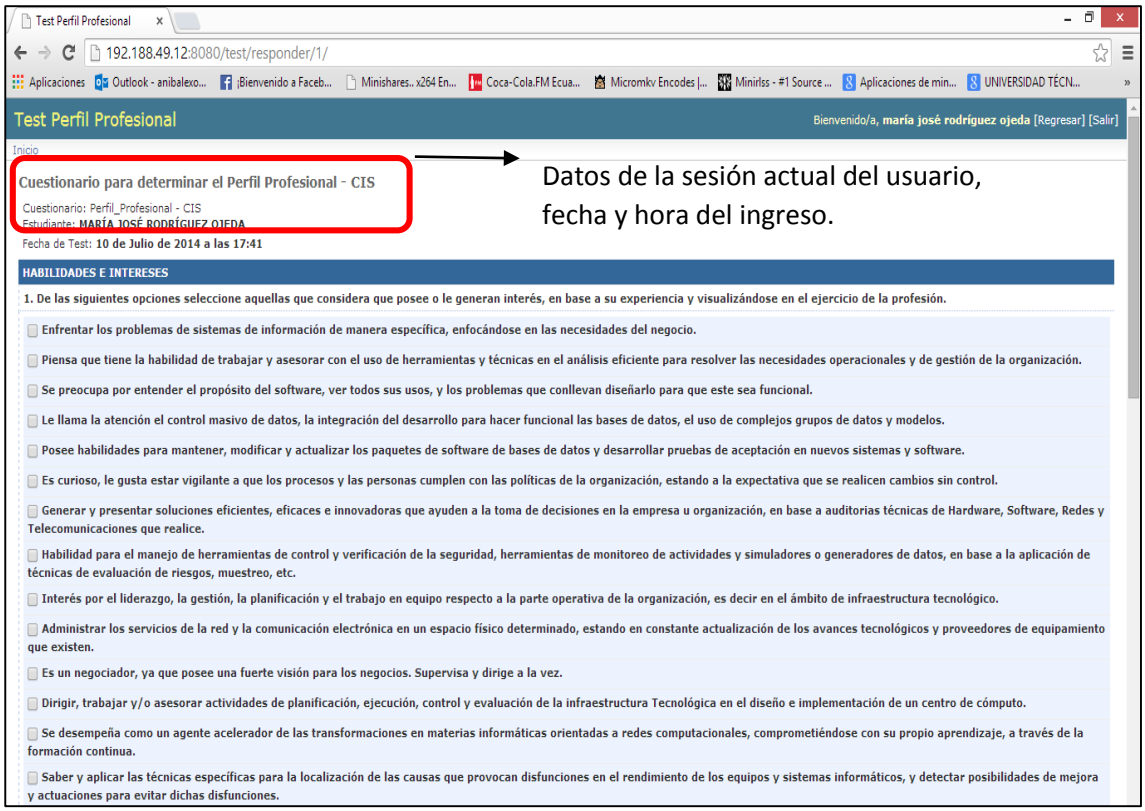


Figura 3. Pantalla del test con las preguntas planteadas.

Una vez realizada la selección de los ítems se procede a grabar y salir de la sesión:

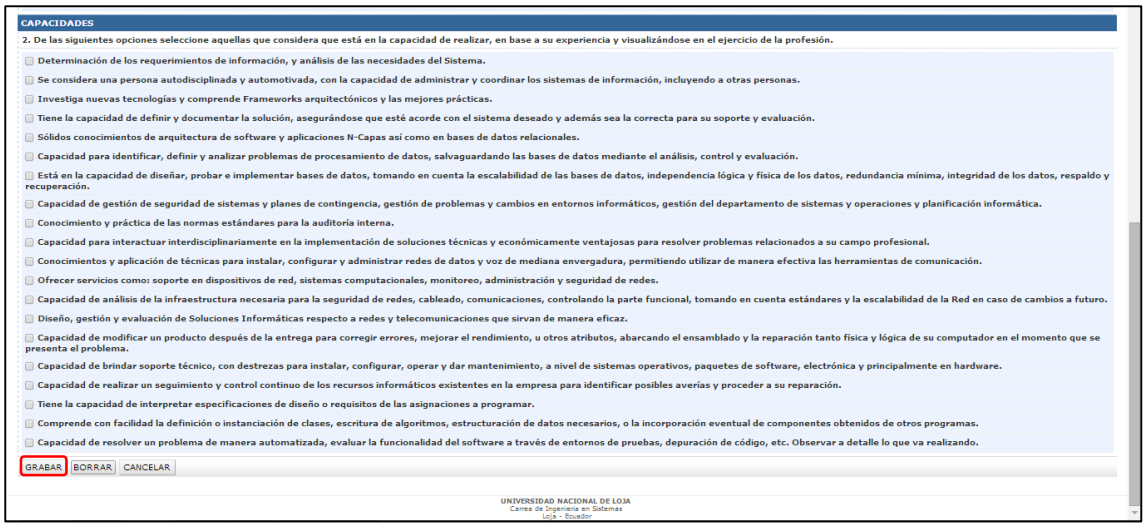


Figura 4. Finalización del test contestado.

Existe también la parte de administración de la aplicación, donde se puede realizar un sin número de acciones que así mismo se actualiza en la base de datos automáticamente; acciones como crear, modificar, eliminar test, administración de usuarios, administración de test contestados, verificaciones de seguridad, etc.

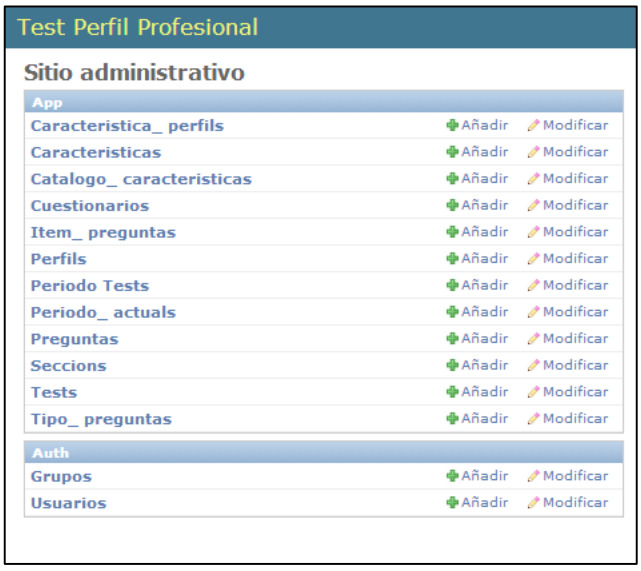


Figura 5. Pantalla de Administración del Test.

En la opción test podemos observar los test que han sido contestados, el número, a que usuario le pertenece, la fecha y hora que ingreso a la sesión y la fecha y hora de finalización del test.

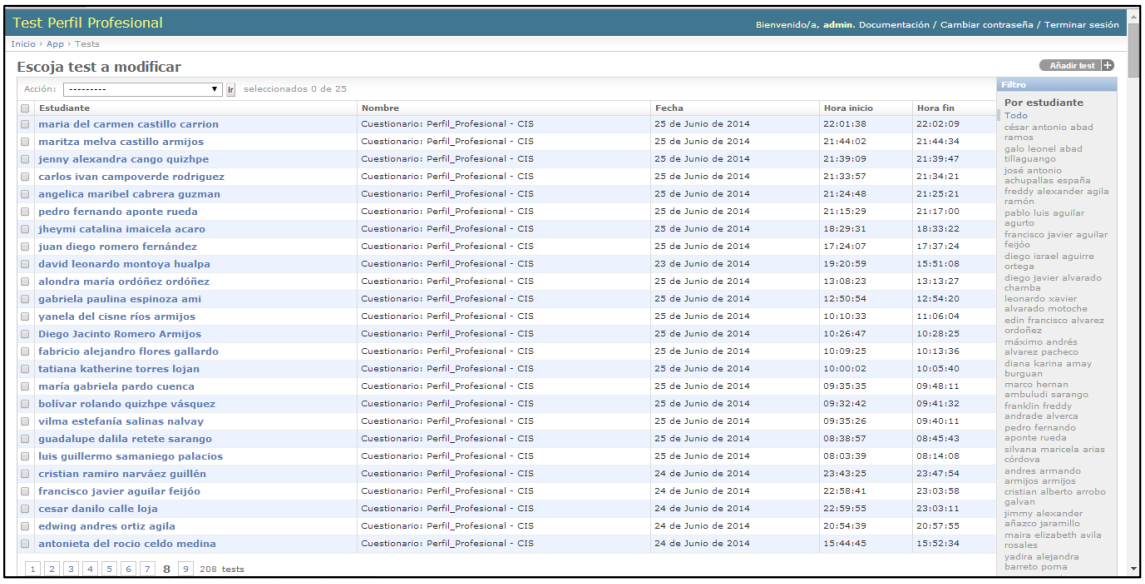
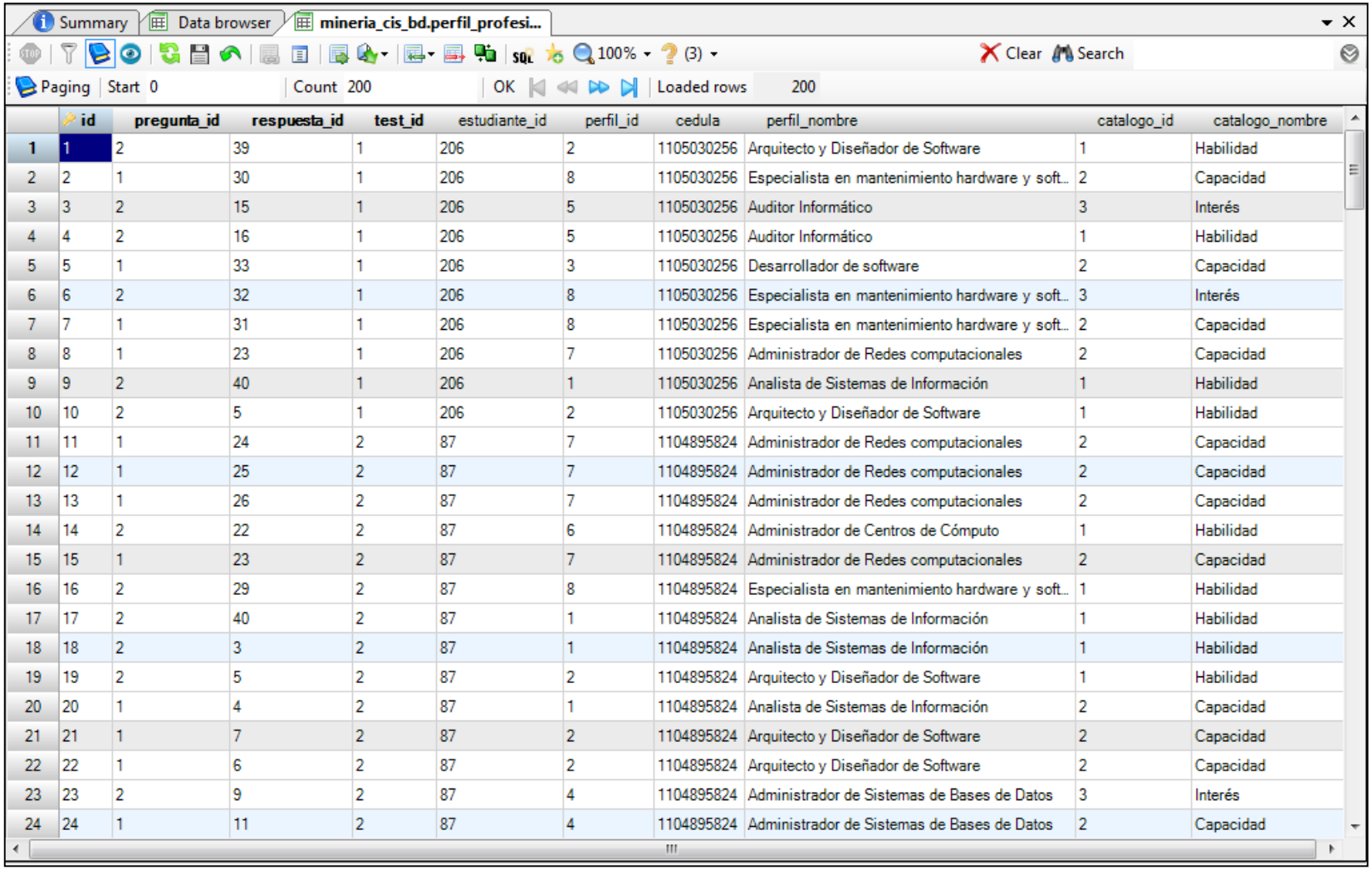


Figura 6. Pantalla de visualización de contestaciones.

A continuación observamos en la figura 7 la estructura de tabla de las contestaciones, que ha sido obtenida a través de funciones mysql en la herramienta de gestión de bases de datos DatAdmin, para la obtención del perfil profesional de cada egresado o graduado de acuerdo al test planteado y realizar la predicción a través del proceso de minería de datos:



Summary | Data browser | mineria_cis_bd.perfil_profesi...

Paging | Start 0 | Count 200 | OK | Loaded rows 200

	id	pregunta_id	respuesta_id	test_id	estudiante_id	perfil_id	cedula	perfil_nombre	catalogo_id	catalogo_nombre
1	1	2	39	1	206	2	1105030256	Arquitecto y Diseñador de Software	1	Habilidad
2	2	1	30	1	206	8	1105030256	Especialista en mantenimiento hardware y soft...	2	Capacidad
3	3	2	15	1	206	5	1105030256	Auditor Informático	3	Interés
4	4	2	16	1	206	5	1105030256	Auditor Informático	1	Habilidad
5	5	1	33	1	206	3	1105030256	Desarrollador de software	2	Capacidad
6	6	2	32	1	206	8	1105030256	Especialista en mantenimiento hardware y soft...	3	Interés
7	7	1	31	1	206	8	1105030256	Especialista en mantenimiento hardware y soft...	2	Capacidad
8	8	1	23	1	206	7	1105030256	Administrador de Redes computacionales	2	Capacidad
9	9	2	40	1	206	1	1105030256	Analista de Sistemas de Información	1	Habilidad
10	10	2	5	1	206	2	1105030256	Arquitecto y Diseñador de Software	1	Habilidad
11	11	1	24	2	87	7	1104895824	Administrador de Redes computacionales	2	Capacidad
12	12	1	25	2	87	7	1104895824	Administrador de Redes computacionales	2	Capacidad
13	13	1	26	2	87	7	1104895824	Administrador de Redes computacionales	2	Capacidad
14	14	2	22	2	87	6	1104895824	Administrador de Centros de Cómputo	1	Habilidad
15	15	1	23	2	87	7	1104895824	Administrador de Redes computacionales	2	Capacidad
16	16	2	29	2	87	8	1104895824	Especialista en mantenimiento hardware y soft...	1	Habilidad
17	17	2	40	2	87	1	1104895824	Analista de Sistemas de Información	1	Habilidad
18	18	2	3	2	87	1	1104895824	Analista de Sistemas de Información	1	Habilidad
19	19	2	5	2	87	2	1104895824	Arquitecto y Diseñador de Software	1	Habilidad
20	20	1	4	2	87	1	1104895824	Analista de Sistemas de Información	2	Capacidad
21	21	1	7	2	87	2	1104895824	Arquitecto y Diseñador de Software	2	Capacidad
22	22	1	6	2	87	2	1104895824	Arquitecto y Diseñador de Software	2	Capacidad
23	23	2	9	2	87	4	1104895824	Administrador de Sistemas de Bases de Datos	3	Interés
24	24	1	11	2	87	4	1104895824	Administrador de Sistemas de Bases de Datos	2	Capacidad

Figura 7. Contestaciones del test desarrollado en django.

Anexo 5: Entrevista respecto de las habilidades, capacidades e intereses de cada Perfil Profesional.

Loja; 10 de marzo de 2014

Ingeniero
Gabriel Requelme
TÉCNICO EN INFRAESTRUCTURA Y COMUNICACIONES DEL BANCO DE LOJA.

CERTIFICA:

Que, María José Rodríguez Ojeda portadora de la cédula de identidad número 1105030256, se dirigió a mí con el objetivo de que le conceda una entrevista para la elaboración de la tesis titulada "Determinación de perfiles profesionales mediante técnicas de minería de datos", la entrevista se basó en dialogar sobre las características de los perfiles profesionales relacionados con la carrera de Ingeniería en Sistemas y pude facilitar información debido a que he tenido experiencia en algunos cargos de diferentes departamentos y conozco las habilidades, capacidades e intereses que debe cumplir el personal para orientarse a cada uno de los perfiles profesionales.

Certifico esto en honor a la verdad y autorizo a la interesada hacer uso del presente en lo crea conveniente.

Atentamente,

Ing. Gabriel Requelme
TÉCNICO EN INFRAESTRUCTURA Y COMUNICACIONES DEL BANCO DE LOJA.

Figura 1. Documento que certifica la veracidad de la entrevista.

Anexo 6: Autorización para el alojamiento de la aplicación django con el Test Perfil Profesional

UNIVERSIDAD NACIONAL DE LOJA
UNIDAD DE TELECOMUNICACIONES E NFORMACIÓN

Of. N.380-UTI-UNL
Loja, 16 de Junio de 2014

Señor
María José Rodríguez
Ciudad.

De mi Consideración:

Por medio de la presente tengo a bien dirigirme a su persona, para hacerle llegar la Clave y usuario, en el oficio emitido por su persona

Clave:
Usuario:
IP.
PUERTO.-

Además se debe indicar, que se debe de enviar la configuración para desplegar aplicaciones en el servidor..

Particular que pongo a su conocimiento para los fines pertinentes.

Atentamente,

Ing. Milton Palacios Morocho
DIRECTOR DE LA UNIDAD
DE TECOMUNICACIONES E INFORMACIÓN

CIUDAD UNIVERSITARIA "GUILLERMO FALCONÍ ESPINOSA" La Argelia, Loja-Ecuador
PBX: (593 07 2) 547252 Casilla: Letra "S" E-mail: telecomunicaciones@unl.edu.ec Sitio Web: www.unl.edu.ec

Figura 1. Autorización del alojamiento del test en el servidor de la Institución.

Anexo 7: Proceso del alojamiento del Test desarrollado en la herramienta django.

Para habilitar el test en la web ha sido necesario gestionar alojamiento que solvente los requerimientos de la aplicación como el soporte de python 2.7, sistema gestor de base de datos, etc. Para ello se ha recurrido a realizar las distintas gestiones en la Unidad de Telecomunicaciones e información de la institución, debido a que el sistema de evaluación docente también está realizado en django. Una vez obtenidos los permisos ha sido necesario realizar algunas configuraciones en el servidor:

1. Se ha utilizado el repositorio de datos dropbox, para cargar e instalar en el servidor los documentos o herramientas que se requieran:

Figura 1. Descarga de acceso directo a dropbox como aplicación de escritorio.

Se ha realizado la conexión al servidor a través de la herramienta de acceso remoto putty:

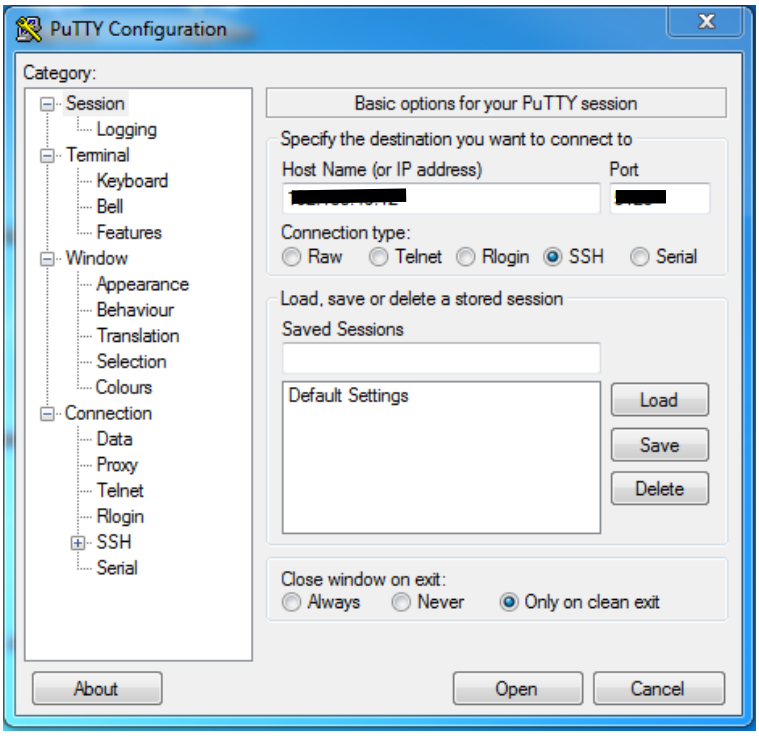


Figura 2. Accediendo al servidor desde la aplicación Putty.

Una vez que nos autentificamos en el servidor con el nombre de usuario y contraseña otorgados:

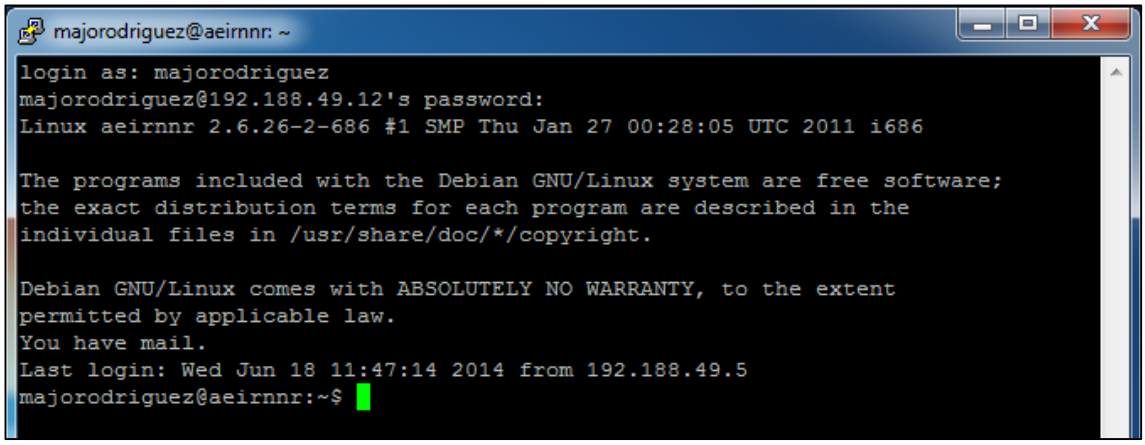


Figura 3. Acceso correcto al servidor.

Copiamos en el repositorio de dropbox el proyecto django y procedemos a ejecutar el comando para cargarlo en el espacio asignado para trabajar en el servidor:

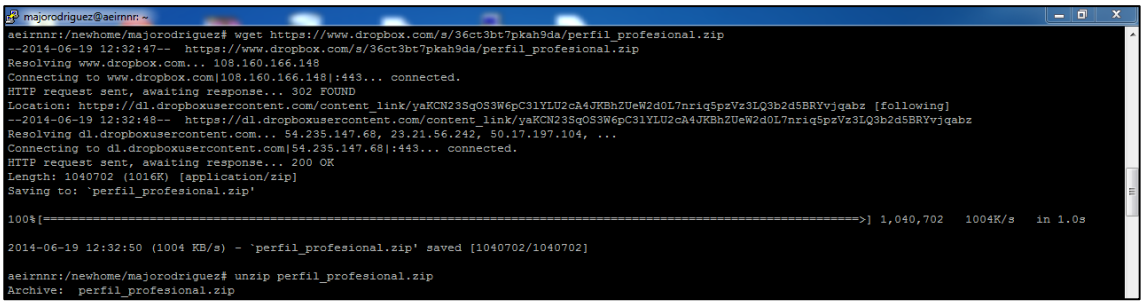


Figura 4. Cargando proyecto django desde drobpox.

Una vez cargador el archivo llamado perfil_profesional.zip, procedemos a descomprimirlo con el comando unzip que llama a la herramienta zip que se encuentra disponible en el servidor. Cargamos la base de datos de la aplicación django llamada perfil_profesional_db, a través de phpmyadmin. Posteriormente al levantar la aplicación django, se ha tomado en cuenta que el servidor cuenta con django y python, requerimientos esenciales para la ejecución:

```
aeirnnr:/newhome/majorodriguez/perfil_profesional# python manage.py runserver 0.0.0.0:8080
Validating models...

0 errors found
Django version 1.3.1, using settings 'perfil_profesional.settings'
Development server is running at http://0.0.0.0:8080/
Quit the server with CONTROL-C.
Traceback (most recent call last):
  File "/usr/lib/python2.6/dist-packages/django/core/servers/basehttp.py", line 283, in run
    self.result = application(self.environ, self.start_response)
  File "/usr/lib/python2.6/dist-packages/django/contrib/staticfiles/handlers.py", line 68, in __call__
    return self.application(environ, start_response)
  File "/usr/lib/python2.6/dist-packages/django/core/handlers/wsgi.py", line 250, in __call__
    self.load_middleware()
  File "/usr/lib/python2.6/dist-packages/django/core/handlers/base.py", line 47, in load_middleware
    raise exceptions.ImproperlyConfigured('Error importing middleware %s: "%s"' % (mw_module, e))
ImproperlyConfigured: Error importing middleware django.middleware.clickjacking: "No module named clickjacking"
[19/Jun/2014 12:47:45] "GET / HTTP/1.1" 500 784
Traceback (most recent call last):
  File "/usr/lib/python2.6/dist-packages/django/core/servers/basehttp.py", line 283, in run
    self.result = application(self.environ, self.start_response)
  File "/usr/lib/python2.6/dist-packages/django/contrib/staticfiles/handlers.py", line 68, in __call__
    return self.application(environ, start_response)
  File "/usr/lib/python2.6/dist-packages/django/core/handlers/wsgi.py", line 250, in __call__
    self.load_middleware()
  File "/usr/lib/python2.6/dist-packages/django/core/handlers/base.py", line 47, in load_middleware
    raise exceptions.ImproperlyConfigured('Error importing middleware %s: "%s"' % (mw_module, e))
ImproperlyConfigured: Error importing middleware django.middleware.clickjacking: "No module named clickjacking"
^Caeirnnr:/newhome/majorodriguez/perfil_profesional#
```

Figura 5. Proyecto django cargado.

Al llamar al enlace para la puesta en marcha de la aplicación se presenta el siguiente error:

Figura 6. Error de respuesta presentado desde el navegador.

Visualizamos que en el momento de realizar la carga internamente se ha dado un error aparentemente por la ausencia del paquete clickjacking:

```
aeirnnr:/newhome/majorodriguez/perfil_profesional# pip install clickjacking
Downloading/unpacking clickjacking
  Could not find any downloads that satisfy the requirement clickjacking
No distributions at all found for clickjacking
Storing complete log in /root/.pip/pip.log
aeirnnr:/newhome/majorodriguez/perfil_profesional# pip install clickjacking.py
Downloading/unpacking clickjacking.py
  Could not find any downloads that satisfy the requirement clickjacking.py
No distributions at all found for clickjacking.py
Storing complete log in /root/.pip/pip.log
aeirnnr:/newhome/majorodriguez/perfil_profesional#
```

Figura 7. Error paquete clickjacking.

Después de un análisis profundo se ha llegado a concluir que el error radica en que la versión actual de django es la 1.3 en el servidor, mientras que la aplicación está desarrollada con la versión 1.6. Para solcuionar el problema se hará uso de virtualenv de python que permite crear entornos virtuales con diferentes versiones en este caso del framework django, para ello procedemos a la instalación:

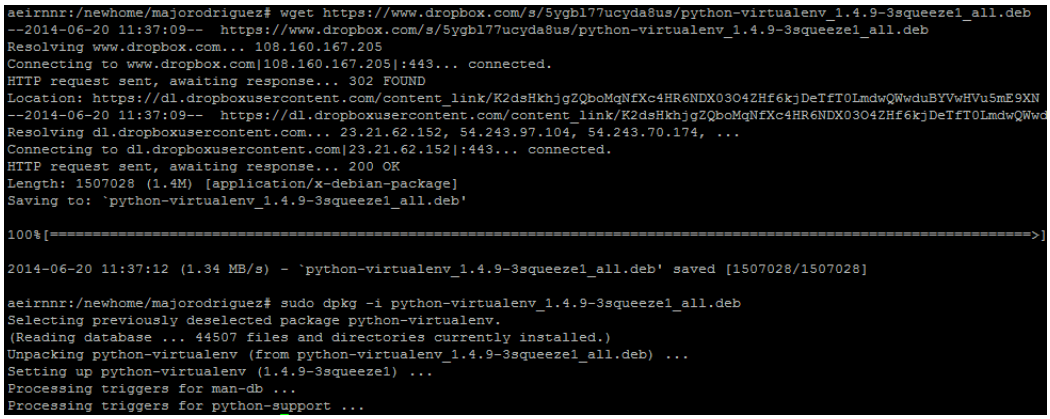


Figura 8. Instalación de virtualenv.

Creamos un Python virtualenv, con el nombre mi_proyecto:

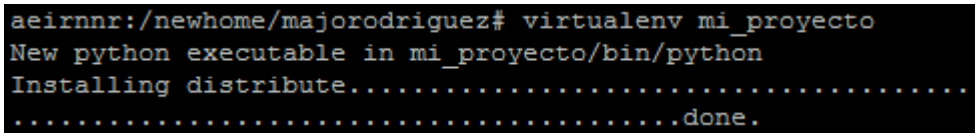


Figura 9. Creación de mi_proyecto con python virtualenv.

Lo que crea el directorio **mi_proyecto/** con la siguiente estructura.

mi_proyecto/

bin/

include/

lib/

En el directorio bin/ se encuentran los ejecutables necesarios para interactuar con el virtualenv. En include/ se encuentran algunos archivos de cabecera de C (archivos .h) necesarios para compilar algunas librerías de Python. Y finalmente en lib/ se encuentra una copia de Python así como un directorio llamado site-packages/ en el cual se aloja el código fuente de los paquetes Python instalados en el virtualenv. A continuación se activa virtualenv: Se procesa el archivo **bin/activate** que se encuentra en la carpeta que se ha creado al ejecutar la utilidad virtualenv, y el prompt de la terminal indica que el virtualenv ha sido activado con éxito:

```
aeirnnr:/newhome/majorodriguez# cd mi_proyecto
aeirnnr:/newhome/majorodriguez/mi_proyecto# source bin/activate
(mi_proyecto)aeirnnr:/newhome/majorodriguez/mi_proyecto#
```

Figura 10. Activación de virtualenv con el comando source bin/actívate.

Después de activarlo, se instala la versión django que se requiere, en este caso django 1.6.2, usando el ejecutable pip que viene por defecto en cada virtualenv creado, y realizando la descarga de django desde dropbox:

```
(mi_proyecto)aeirnnr:/newhome/majorodriguez/mi_proyecto# wget https://www.dropbox.com/s/rw19ofdw0gbseyq/Django.zip
```

Figura 11.Instalando django 1.6.2. del repositorio dropbox.

Descomprimimos el archivo:

```
(mi_proyecto)aeirnnr:/newhome/majorodriguez/mi_proyecto# unzip Django.zip
```

Figura 12. Descomprimiendo software django descargado.

Instalar django:

```
(mi_proyecto)aeirnnr:/newhome/majorodriguez/mi_proyecto/Django# python setup.py install
```

Figura 13. Instalación de la versión nueva de django.

Levanto proyecto django:

```
(mi_proyecto)aeirnnr:/newhome/majorodriguez/perfil_profesional# python manage.py runserver 0.0.0.0:8000
Validating models...

0 errors found
June 20, 2014 - 12:47:51
Django version 1.6.2, using settings 'perfil_profesional.settings'
Starting development server at http://0.0.0.0:8000/
Quit the server with CONTROL-C.
/newhome/majorodriguez/perfil_profesional/app/views.py:17: DeprecationWarning: django.utils.simplejson is deprecated; use json instead.
  from django.utils import simplejson
[20/Jun/2014 12:48:04] "GET / HTTP/1.1" 200 2065
```

Figura 14. Proceso de levantamiento del proyecto django desde el servidor.

Para mantener el proceso de la aplicación en django en ejecución permanentemente se ha ejecutado un proceso escondido conocido como demonio. Para ello se ha creado un script de ejecución:

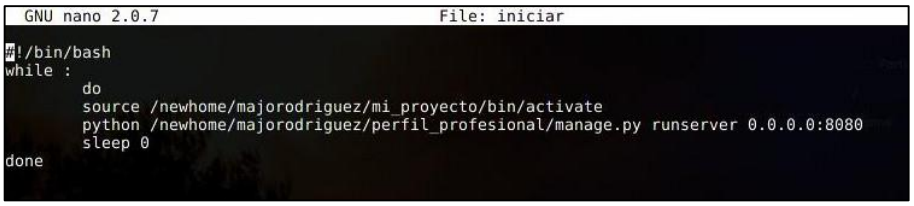


Figura 15. Script para ejecución para el levantamiento de la aplicación django.

Corremos el script:

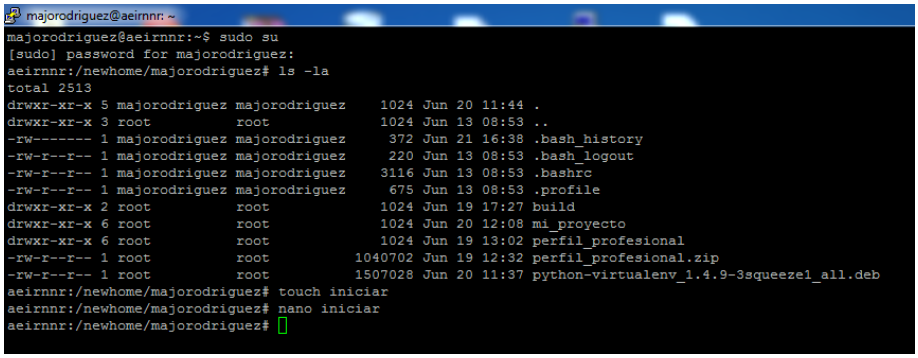


Figura 16. Script de ejecución ejecutado.

Finalmente probamos que la aplicación has sido levantada correctamente:

Figura 17. Aplicación test perfil profesional en django corriendo normalmente.

Anexo 8: Comparación de las característica de Herramientas como apoyo para el proceso de Minería de Datos.

En este apartado de características de las herramientas KNIME, Weka y RapidMiner, realizando la evaluación de las características relevantes como manipulación de la interfaz gráfica, manejo de procesos de minería y visualización de resultados.

- **KNIME**

Esta herramienta presenta en su interfaz distintos nodos agrupados en fichas (ver figura 1), como por ejemplo:

- ✓ Entrada de datos [IO > Read].
- ✓ Salida de datos [IO > Write].
- ✓ Preprocesamiento [Data Manipulation], para filtrar, discretizar, normalizar, filtrar, seleccionar variables…
- ✓ Minería de datos [Mining], para construir modelos (reglas de asociación, clustering, clasificación, MDS, PCA…).
- ✓ Salida de resultados [Data Views] para mostrar resultados en pantalla (textual o gráfica).

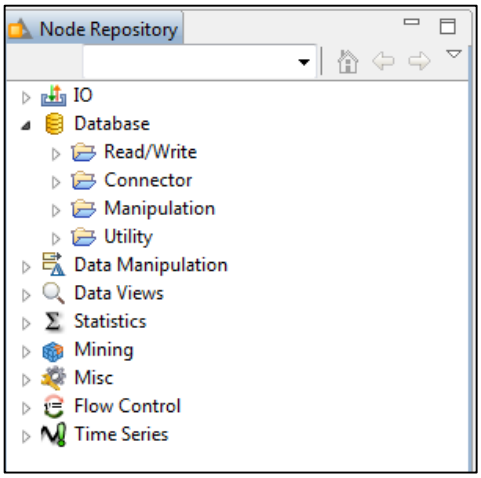


Figura 1. Repositorio de Nodos o procesos de KNIME.

Para crear un flujo de ejecución, las salidas de unos nodos se utilizan como entradas de otros. Por ejemplo (ver figura 2), un flujo básico podría ser de la forma:

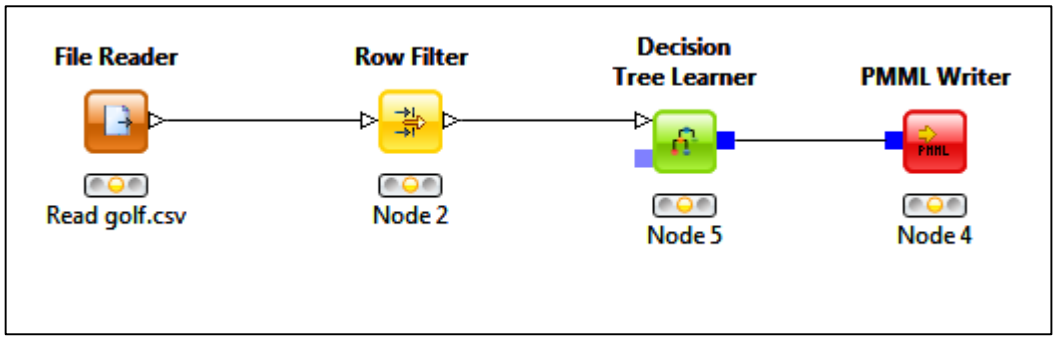


Figura 2. Ejemplo de un flujo de procesos básico en KNIME.

La visualización de resultados, depende del tipo de algoritmo aplicado, cada algoritmo tiene su forma de visualizar los datos, al darle click derecho – view en el nodo del algoritmo. Por ejemplo (ver figura 3) al utilizar el algoritmo Decision Tree (árbol de decisión):

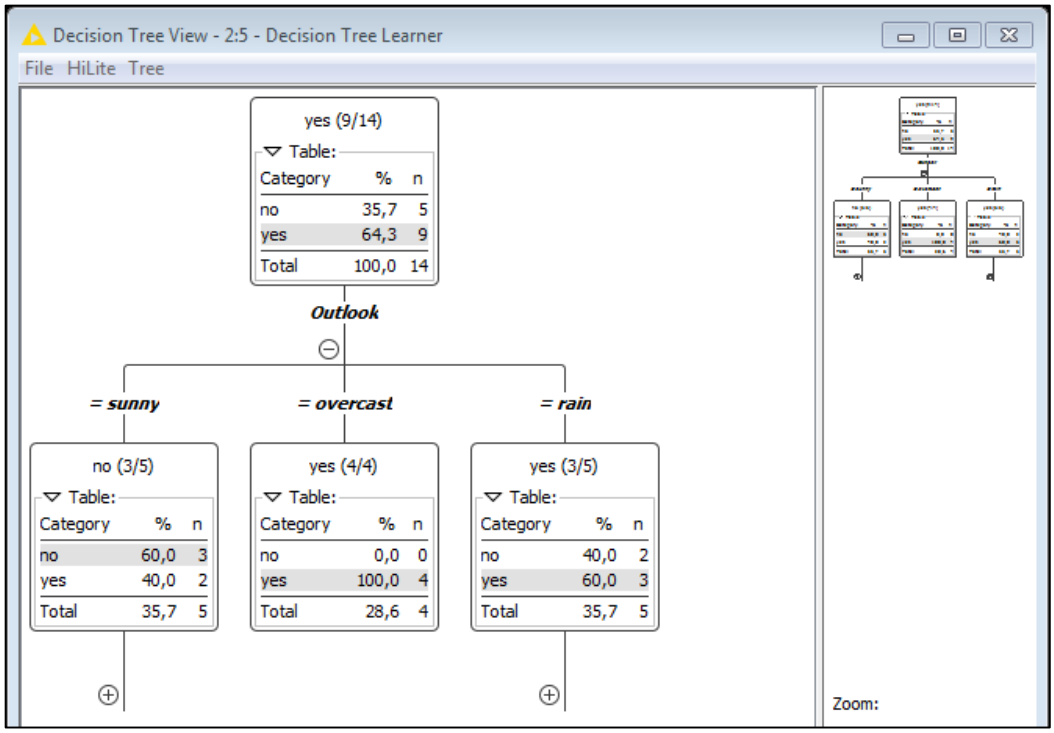


Figura 3. Resultados utilizando Árboles de decisiones.

- **Weka**

La interfaz principal de weka nos presenta una sencilla pantalla donde tenemos las opciones de escoger entre aplicaciones como explorer, experimenter, knowledgeFlow y Simple CLI para comenzar a trabajar con la herramienta (ver figura 4).

Figura 4. Pantalla inicial de la herramienta weka.

La forma más común de iniciar el trabajo es dentro de la opción explorer, donde se presenta una interfaz con múltiples opciones, al cargar los datos la herramienta los presenta de forma gráfica en un diagrama de barras donde se muestra todos los atributos de la estructura cargada para realizar el proceso de minería (ver figura 5).

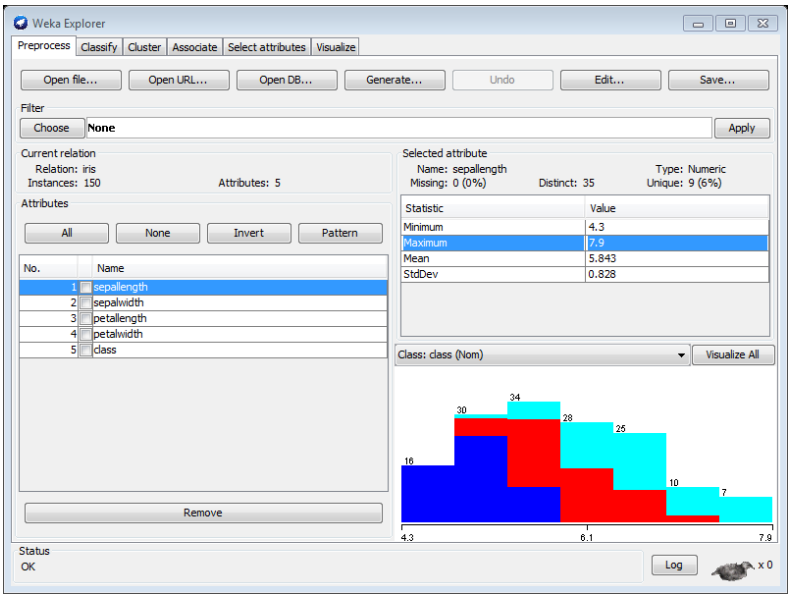


Figura 5. Visualización de los datos cargados.

Tenemos un sin número de algoritmos de aplicación, la herramienta presenta los resultados directamente en forma de texto, detallando algunas características de los mismos (ver figura 6):

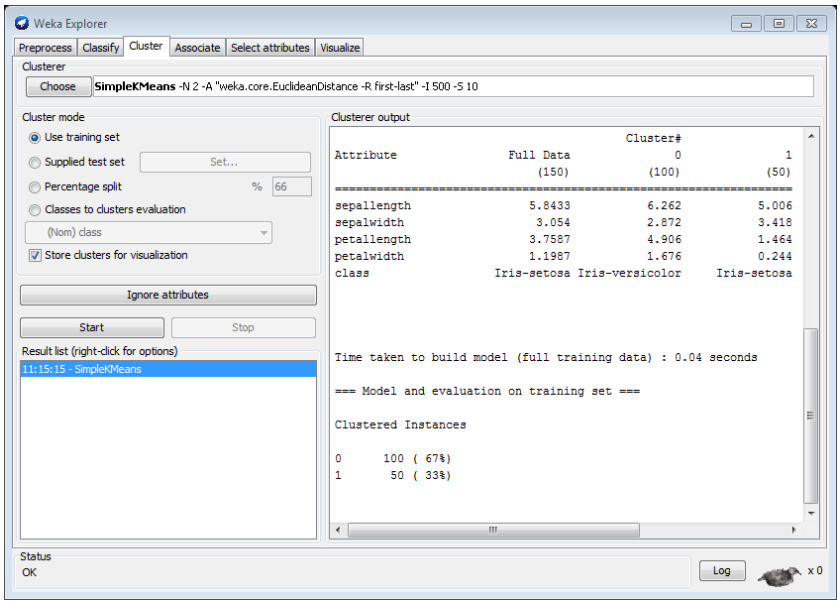


Figura 6. Pantalla de Resultados en formato texto.

Weka a su vez nos permite realizar la visualización de los resultados a su vez de manera gráfica para una mayor comprensión (ver figura 7).

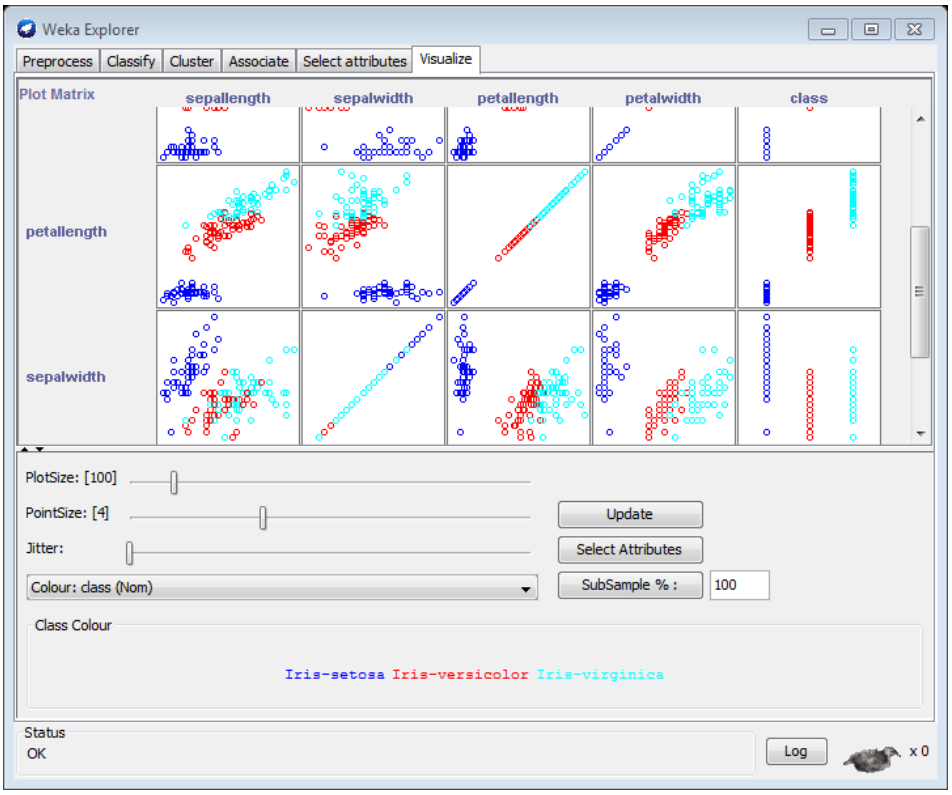


Figura 7. Representación gráfica de los resultados.

- **RapidMiner**

La herramienta RapidMiner, muestra un sin número de fuertes características, está estructurada por una interfaz gráfica muy organizada e intuitiva, formada por múltiples paneles (ver figura 8).

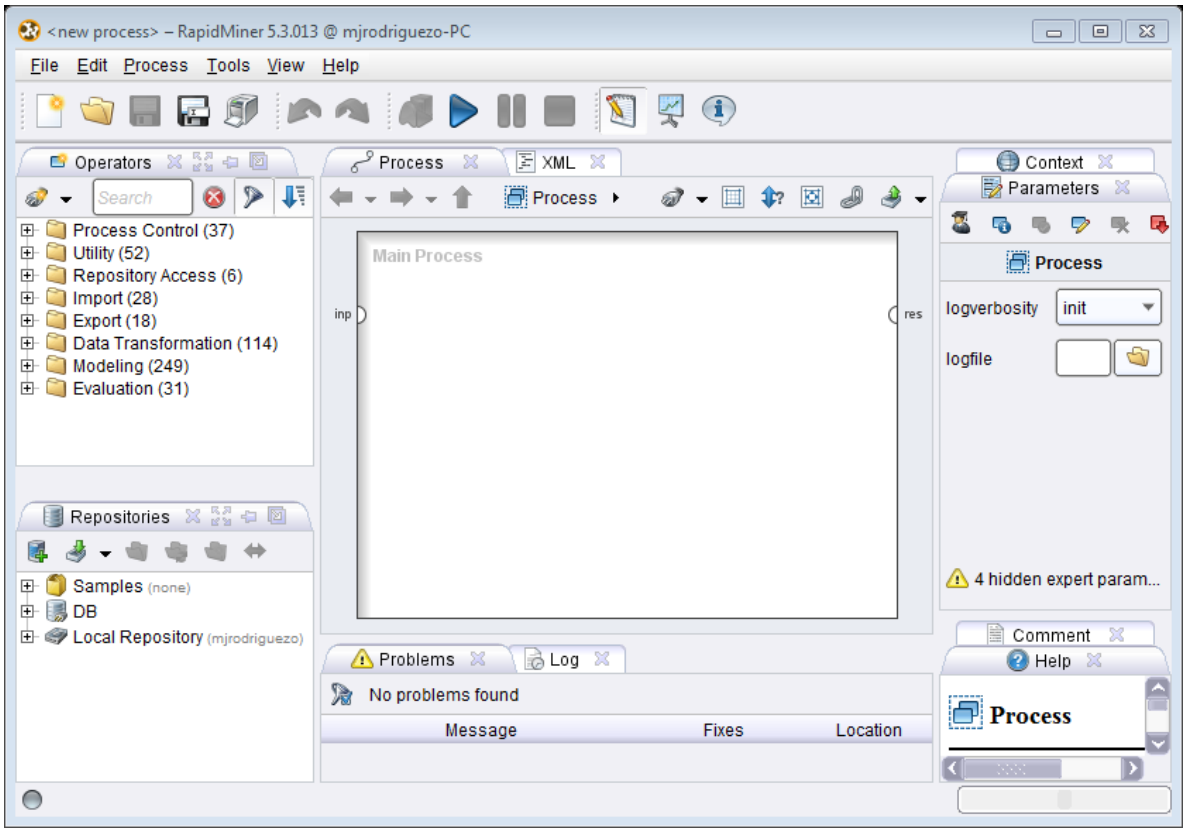


Figura 8. Interfaz gráfica general de RapidMiner.

El panel operators, como la palabra lo dice contiene los más de 500 operadores que ofrece esta potente herramienta para realizar el proceso de minería (ver figura 9).

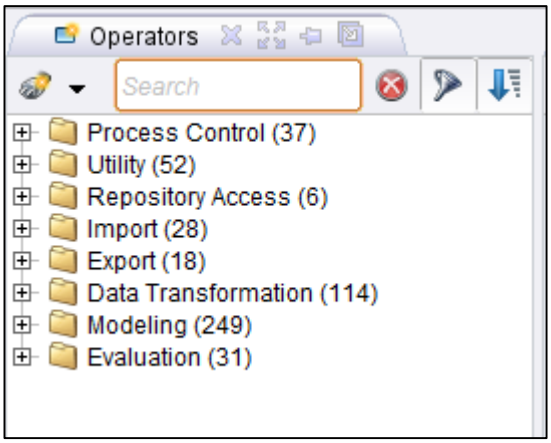


Figura 9. Panel operadores de Rapidminer.

El panel Repositories, como su nombre lo dice es el almacén de lo realizado en la herramienta, contiene las copias de los archivos como procesos y datos, organizados a través de un árbol jerárquico, desde aquí se realiza la carga de los datos examples ya almacenados en la herramienta, las conexiones a las bases de datos, la carga de tablas de manera directa, la llamada a archivos en distintos formatos como: csv, Excel, arff, etc (ver figura 10).

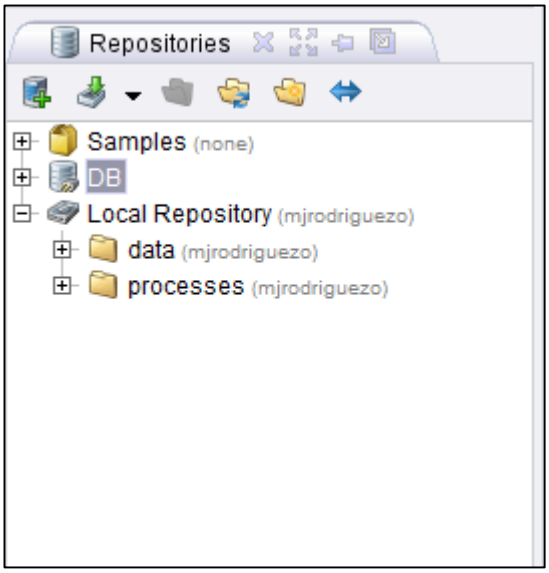


Figura 10. Panel Repositorios RapidMiner.

La herramienta además está formada por el Panel de procesos permite de manera visual, ir armando nuestra estructura de minería, jugando con los distintos operadores, con el fin de obtener los mejores resultados (ver figura 11). El Panel de parámetros (ver figura 12) que permite cambiar los parámetros de cada proceso y el panel de ayuda (ver figura 13). Todos estos paneles son los principales de la herramienta.

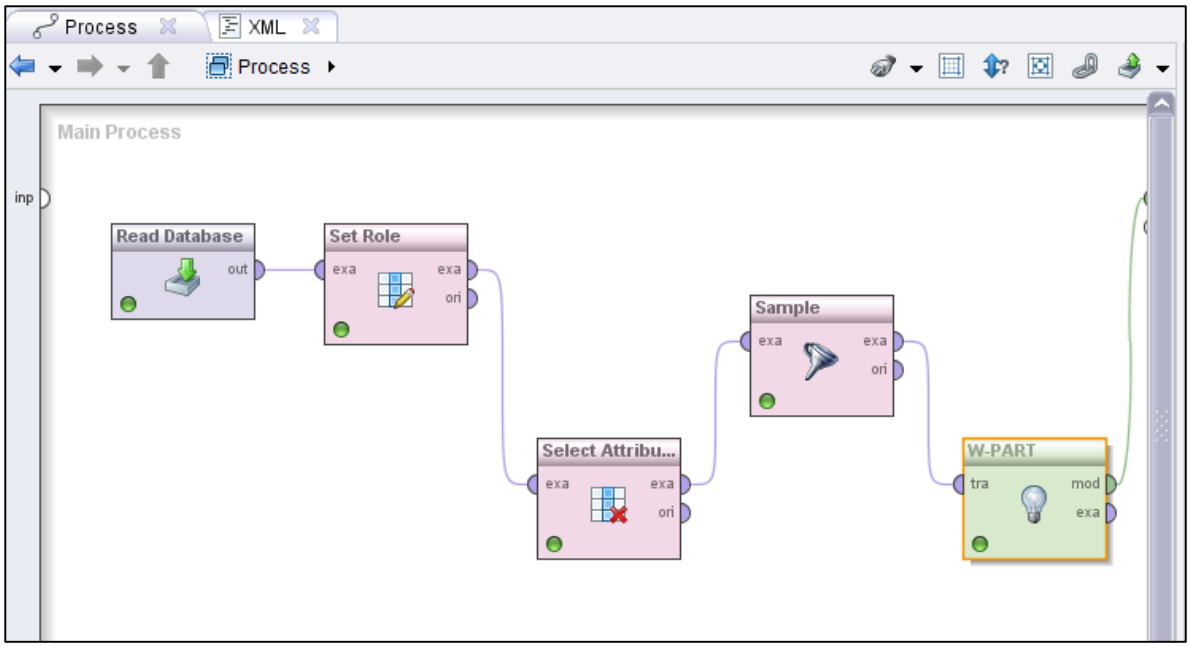


Figura 11. Panel gráfico de la interacción de procesos.

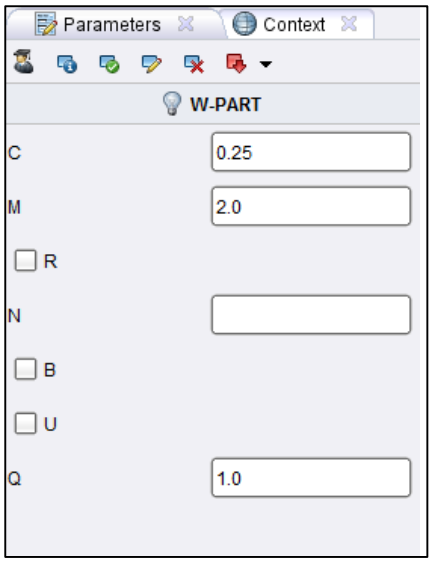


Figura 12. Panel de Parámetros

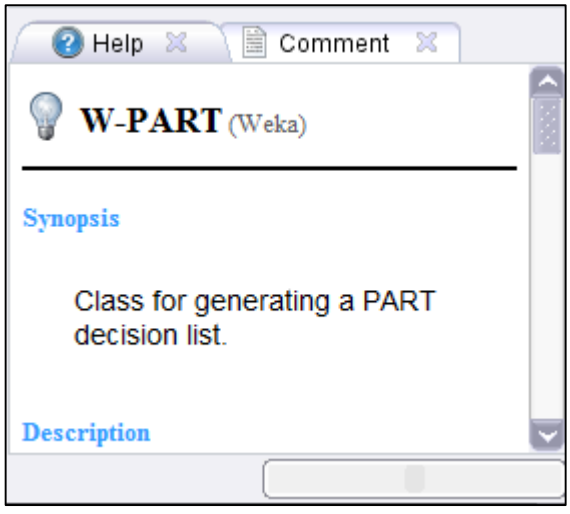


Figura 13. Panel de Ayuda

RapidMiner permite visualización los resultados de forma detallada y gráficamente es muy amigable al usuario; pero los tipos de gráficos o presentación de los resultados depende de cada algoritmo aplicado (ver figuras 14, 15 y 16).

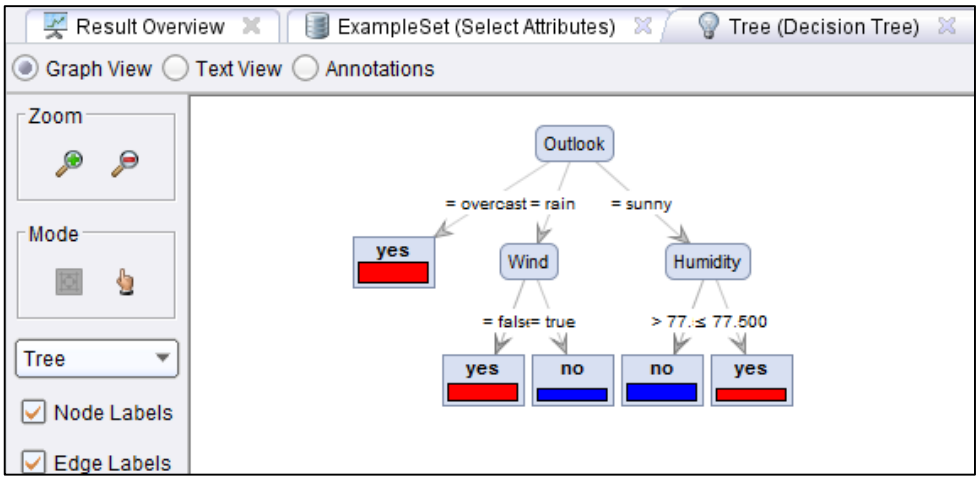


Figura 14. Gráfica del algoritmo árbol de decisión.

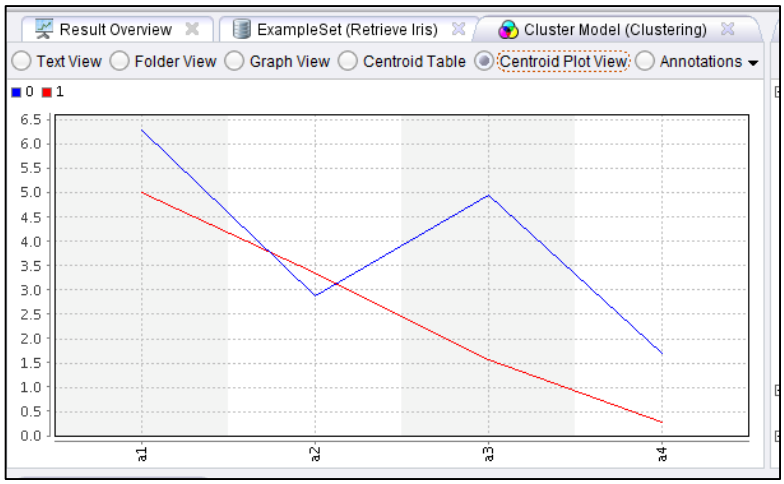


Figura 15. Gráfica de los resultados con algoritmo K-means.

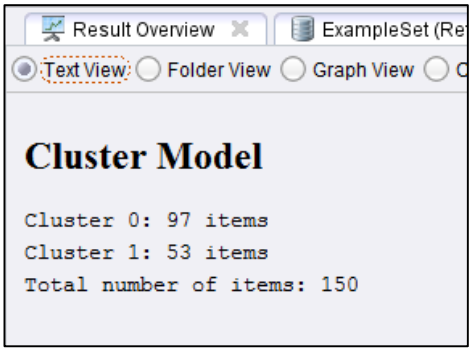


Figura 16. Resultados con el Algortimo K-means.

Anexo 9: Evaluación las herramientas de minería con datos de Prueba.

Para realizar una comparación entre las diferentes herramientas, se han ejecutado algunas pruebas de minería de datos con la base de datos denominada Golf. Esta base de datos contiene 5 variables cada una con sus correspondientes valores:

- **Outlook** (clima)**:** sunny, overcast, rain
- **Temperature** (temperatura)**:** numerical
- **Humidity** (humedad)**:** numerical
- **Wind** (viento)**:** true, false
- **Play** (jugar)**:** yes, no

La estructura de la base de datos descrita, representa las condiciones como clima, estado de la temperatura, humedad, existencia o no de viento; que deben cumplirse para tomar la decisión de salir a jugar o no al golf (ver tabla I).

TABLA I

BASE DE DATOS GOLF

Outlook	Temperature	Humidity	Wind	Play
sunny	85.0	85.0	false	no
sunny	80.0	90.0	true	no
overcast	83.0	78.0	false	yes
rain	70.0	96.0	false	yes
rain	68.0	80.0	false	yes
rain	65.0	70.0	true	no
overcast	64.0	65.0	true	yes
sunny	72.0	95.0	false	no
sunny	69.0	70.0	false	yes
rain	75.0	80.0	false	yes
sunny	75.0	70.0	true	yes
overcast	72.0	90.0	true	yes
overcast	81.0	75.0	false	yes
rain	71.0	80.0	true	no

El proceso de minería de datos se lo ha realizado con la utilización del algoritmo árbol de decisión. Las pruebas realizadas con los datos especificados en la tabla I se han con el fin de comparar tres aspectos importantes en cada herramienta que son: la visualización e interacción de los operadores del proceso de minería de datos para resolver el problema planteado, lo amigable de la interfaz gráfica de usuario respecto del manejo de cada herramienta así como la visualización de los resultados y facilidad de interpretación de los mismos en cada herramienta (ver figuras 1 - 14). Estos aspectos han sido analizados con el fin de realizar la adecuada selección de la herramienta, a continuación se han detallado las pruebas realizadas:

- **KNIME**

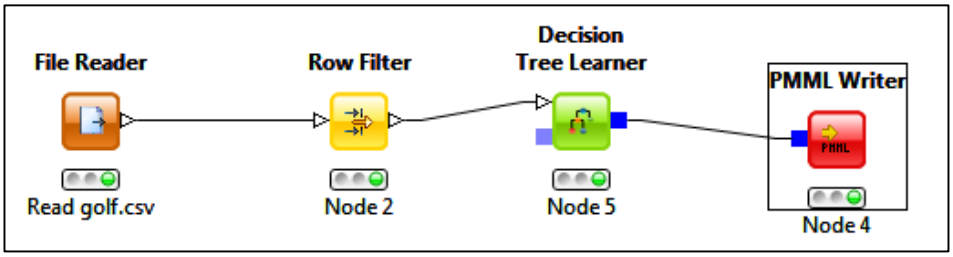
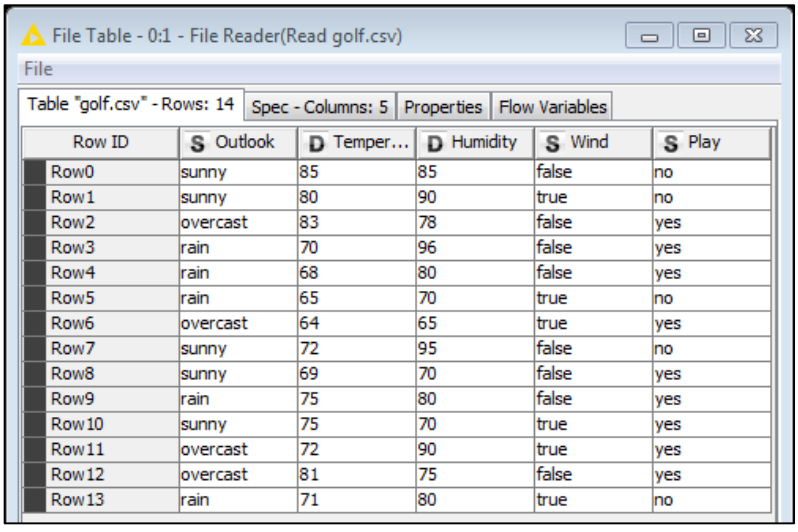


Figura 1. Operadores del Proceso de Minería en KNIME.



Row ID	Outlook	Temper...	Humidity	Wind	Play
Row0	sunny	85	85	false	no
Row1	sunny	80	90	true	no
Row2	overcast	83	78	false	yes
Row3	rain	70	96	false	yes
Row4	rain	68	80	false	yes
Row5	rain	65	70	true	no
Row6	overcast	64	65	true	yes
Row7	sunny	72	95	false	no
Row8	sunny	69	70	false	yes
Row9	rain	75	80	false	yes
Row10	sunny	75	70	true	yes
Row11	overcast	72	90	true	yes
Row12	overcast	81	75	false	yes
Row13	rain	71	80	true	no

Figura 2. Forma de visualizar los datos iniciales en KNIME

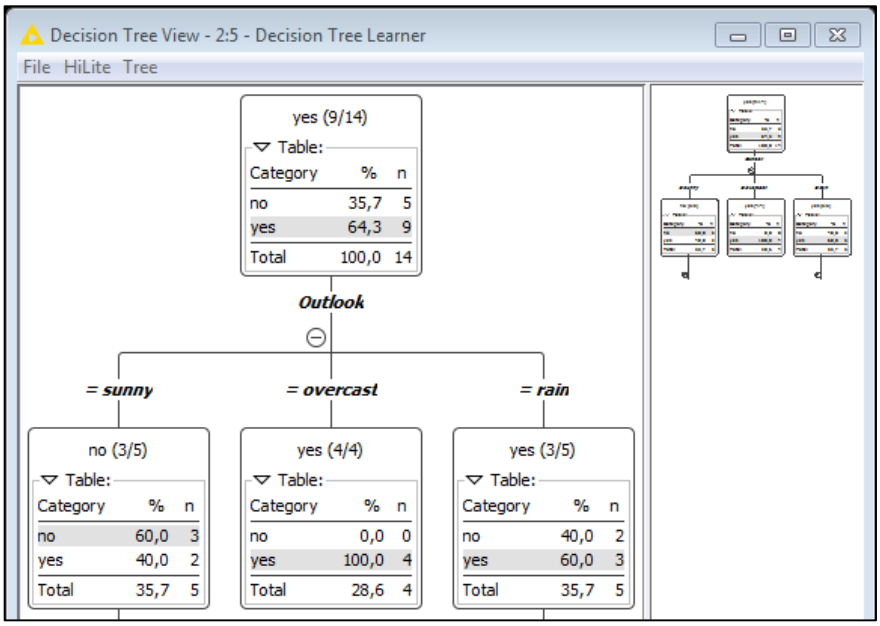


Figura 3. Visualización gráfica de los resultados en KNIME.

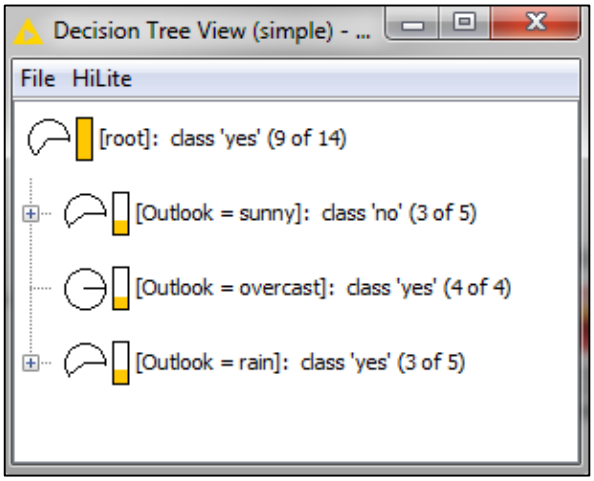


Figura 4. Visualización de detalles los resultados en KNIME.

- **Weka**

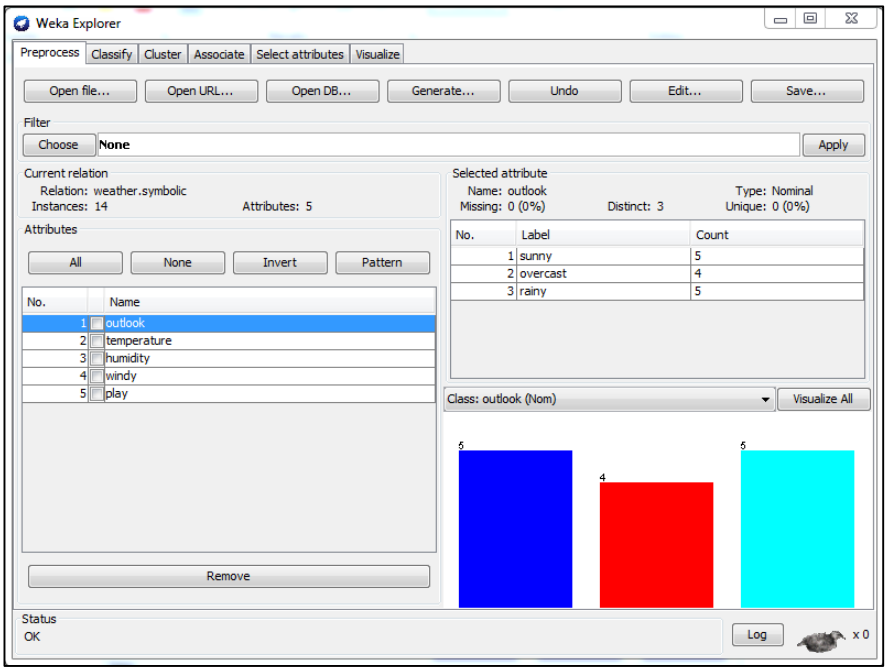


Figura 5. Visualización inicial de los datos en Weka.

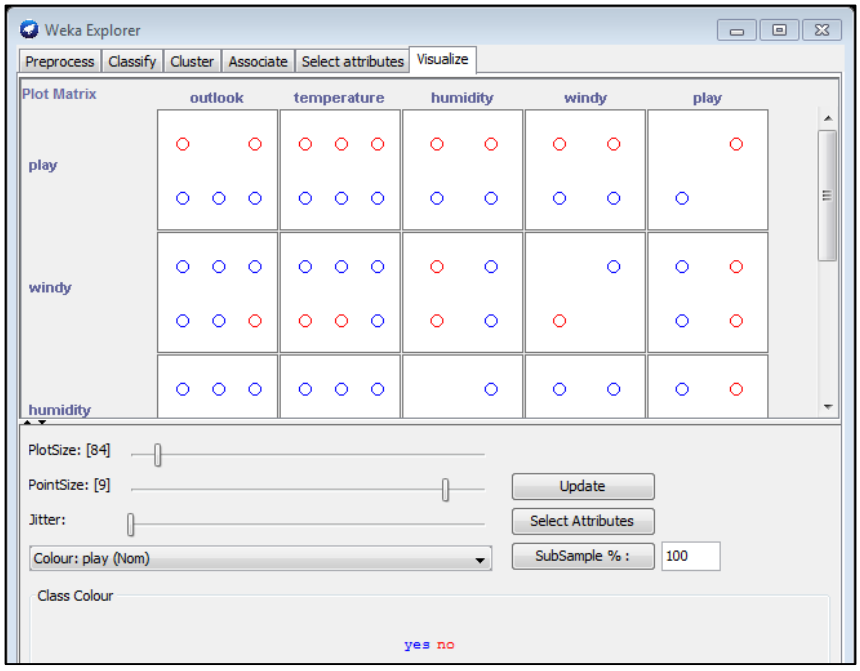


Figura 6. Visualización gráfica de los resultados en Weka.

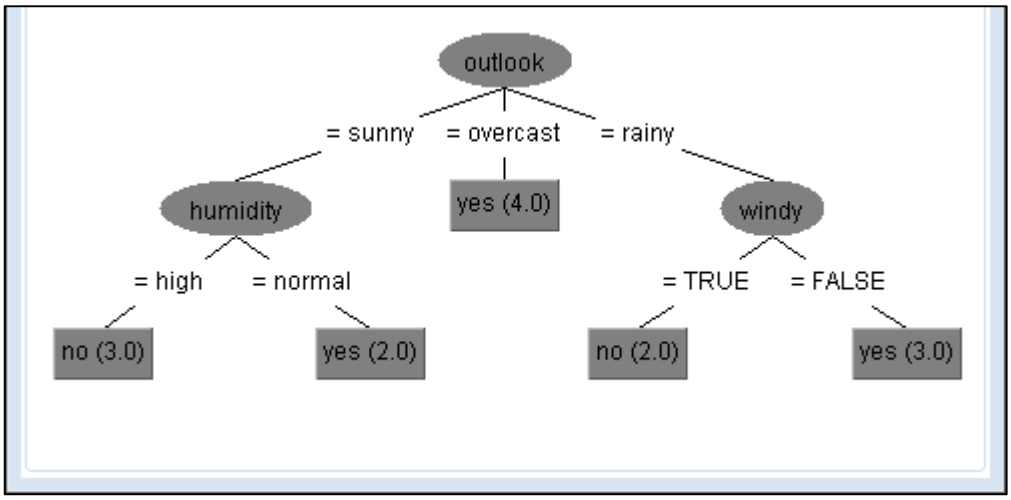


Figura 7. Visualización gráfica de los resultados en Weka, árbol de decisión.

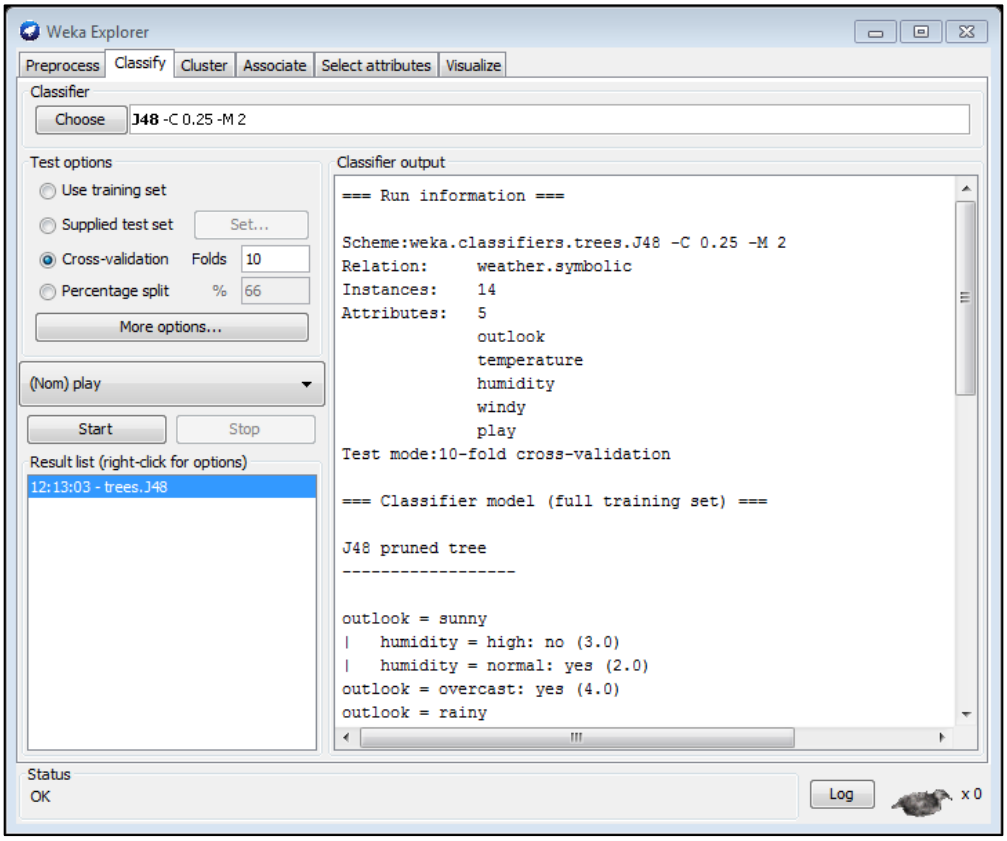


Figura 8. Visualización de detalles los resultados en Weka.

- **RapidMiner**

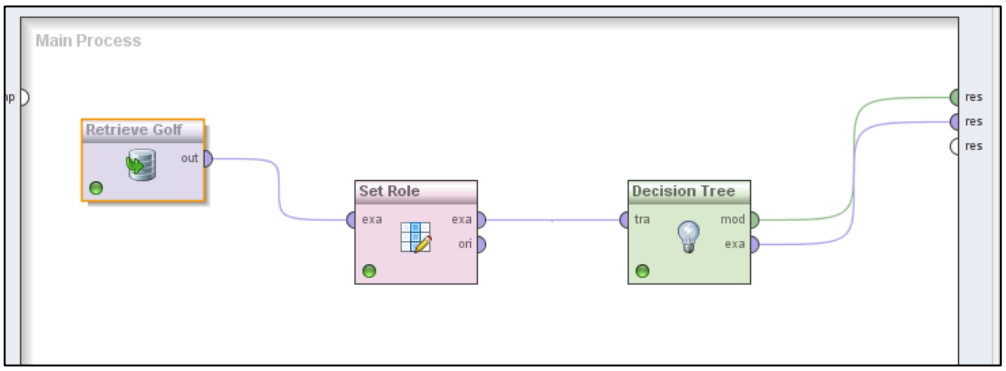


Figura 9. Operadores del Proceso de Minería en RapidMiner.

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

Figura 10. Visualización en tabla de los datos iniciales.

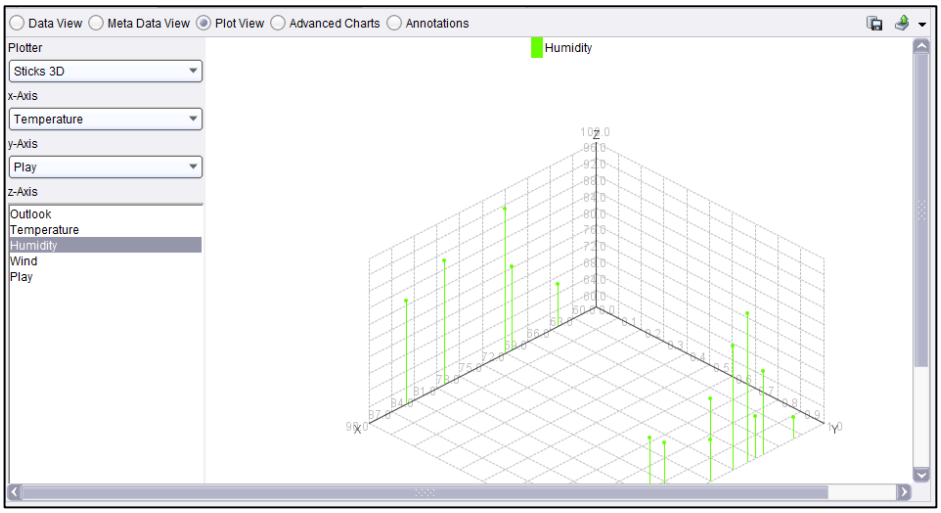


Figura 11. Visualización gráfica de cada atributo de los datos iniciales.

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	Play	nominal	mode = yes (9), least = no (5)	no (5), yes (9)	0
regular	Outlook	nominal	mode = sunny (5), least = overcast (4)	rain (5), overcast (4), s	0
regular	Temperature	integer	avg = 73.571 +/- 6.572	[64.000 ; 85.000]	0
regular	Humidity	integer	avg = 80.286 +/- 9.840	[65.000 ; 96.000]	0
regular	Wind	nominal	mode = false (8), least = true (6)	true (6), false (8)	0

Figura 12. Datos estadísticos de los datos Iniciales generados por RapidMiner.

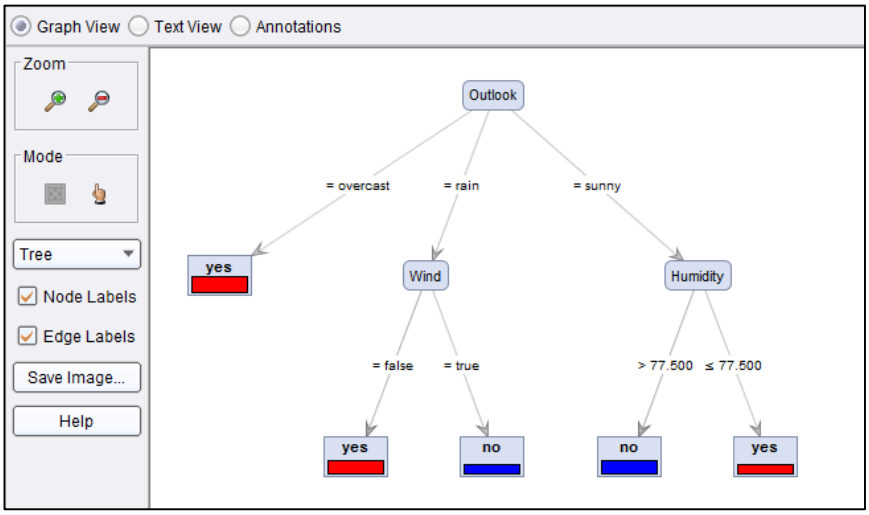


Figura 13. Visualización gráfica de los resultados en RapidMiner.

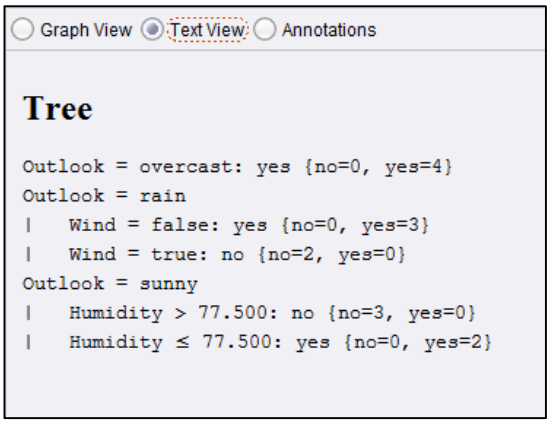


Figura 14. Visualización de detalles los resultados en RapidMiner.

Anexo 10: Procesos y operadores de Minería de Datos en Rapid Miner.

Rapidminer es una potente herramienta que trabaja con un sin número de operadores que juntos conforman todo un proceso, estos operadores se los combina para que interactúan entre sí y lograr el correcto funcionamiento del algoritmo aplicado. En este apartado se detalla todos los procesos de minería construidos: con el 72% para el conjunto de entrenamiento, la validación cruzada de cada algoritmo, los procesos respecto al 28% restante para la evaluación de los mejores algoritmos y finalmente la aplicación de los modelos obtenidos por los mejores algoritmos con los egresados actuales del año 2014 de la carrera de ingeniería en sistemas; detallando los operadores que se ha utilizado, así como su funcionalidad dentro de cada proceso.

✓ **Pruebas de entrenamiento y validación cruzada con datos no agrupados.**

A continuación se describen los procesos formados para los algoritmos CHAID, Decision__Table, DTNB, ID3, Jrip, NNge, Part, Ridor para los datos no agrupados.

- **Clasificación mediante CHAID**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 1).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Numerical to Polynominal: Este operados convierte los tipos de datos numericos a polinomiales para mejorar la calidad de los datos y acoplarse al algoritmo empleado; se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

CHAID: Este operador se encuentra dentro del grupo “algoritmos de clasificación en base a la aplicación de árboles de decisión” y se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

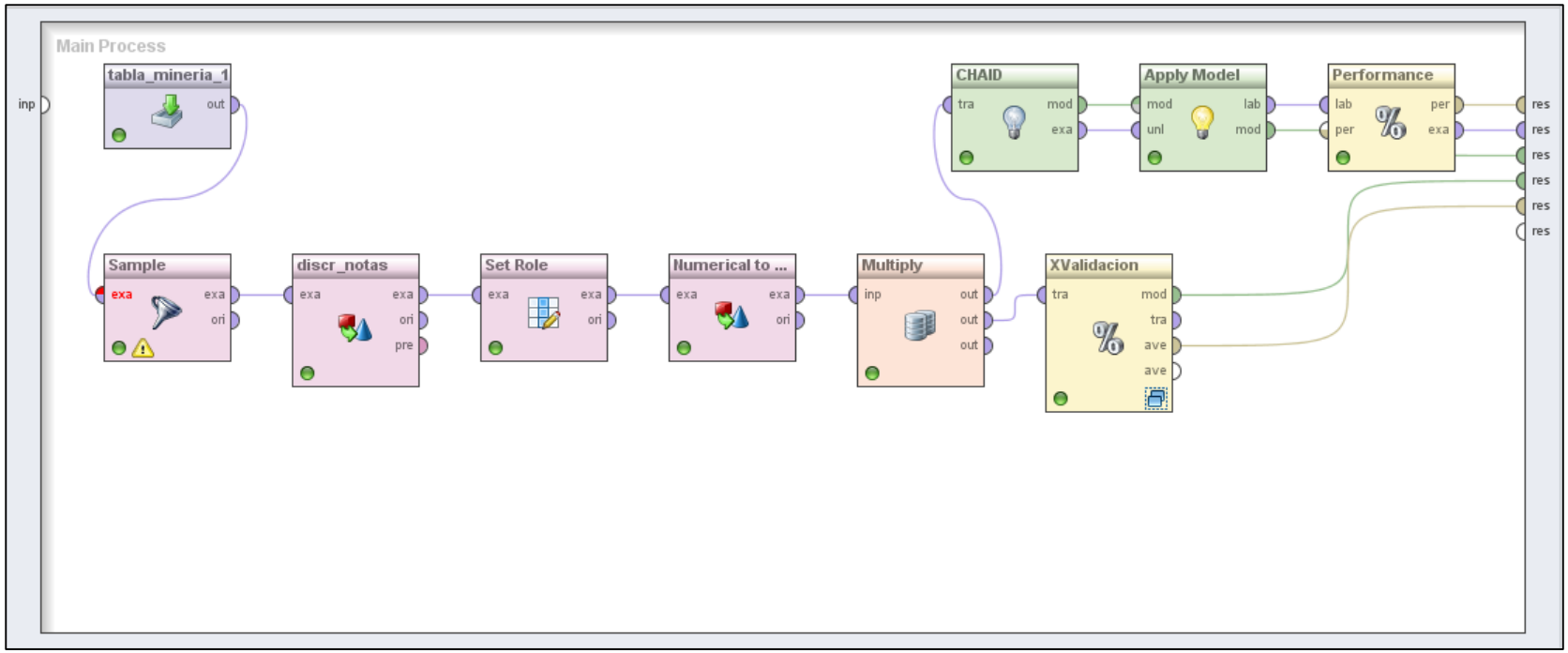


Figura 1. Conjunto de operadores que conforman el proceso para el algoritmo CHAID.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 2).

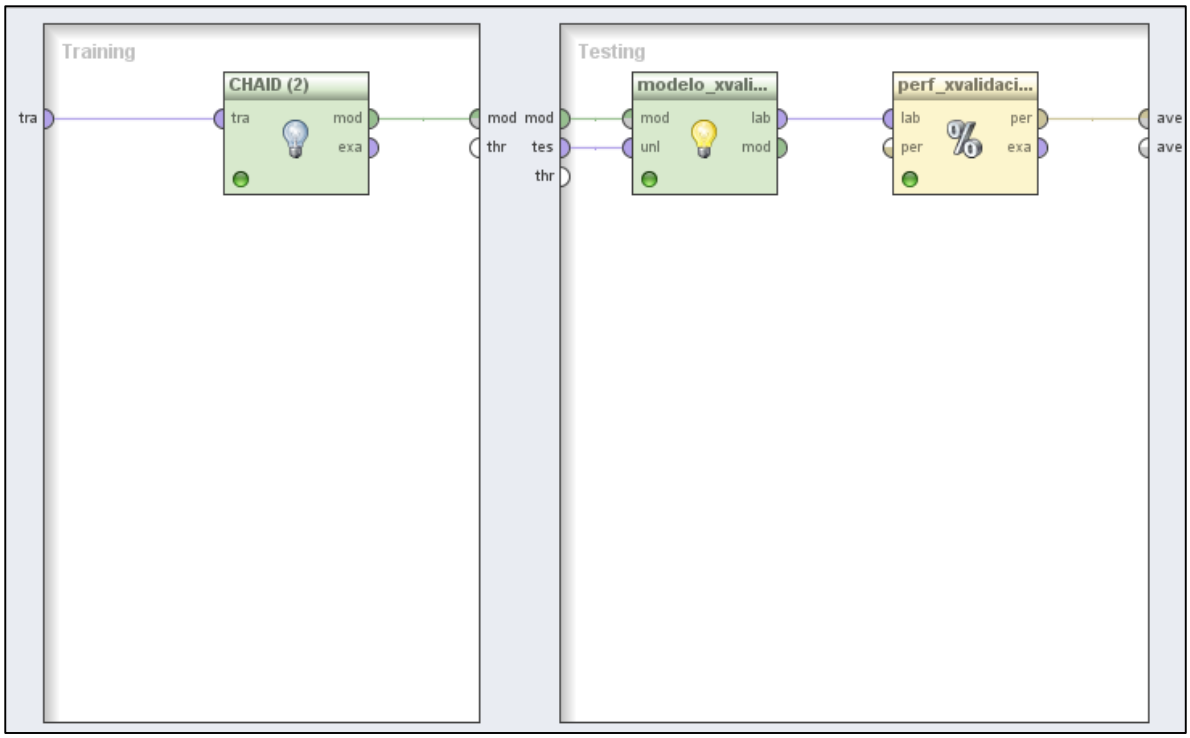


Figura 2. Procesos de Validación Cruzada para el algoritmo CHAID.

- **Clasificación mediante Decision__Table**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 3).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Numerical to Polynominal: Este operados convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse al algoritmo empleado; se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-DecisionTable: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Decision Table, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

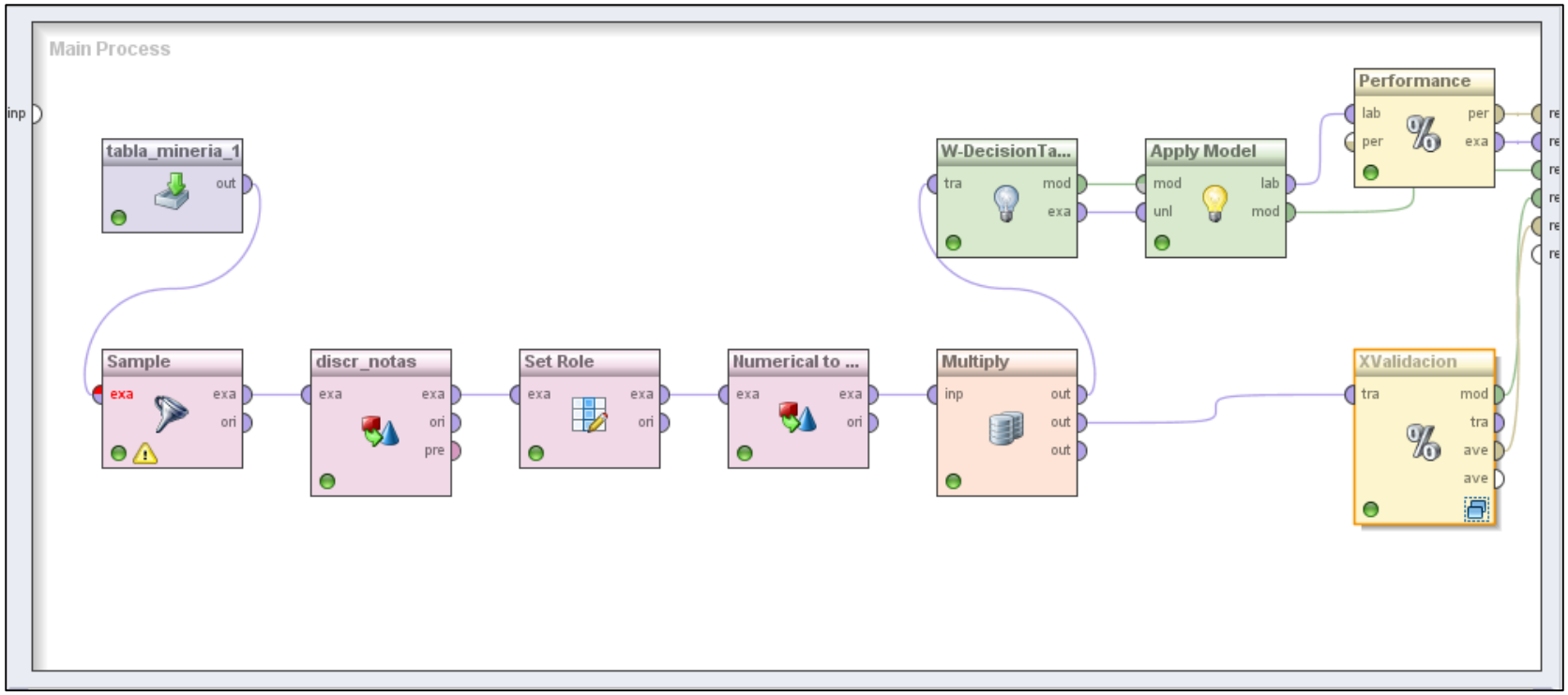


Figura 3. Conjunto de operadores que conforman el proceso para el algoritmo Decision__Table.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 4).

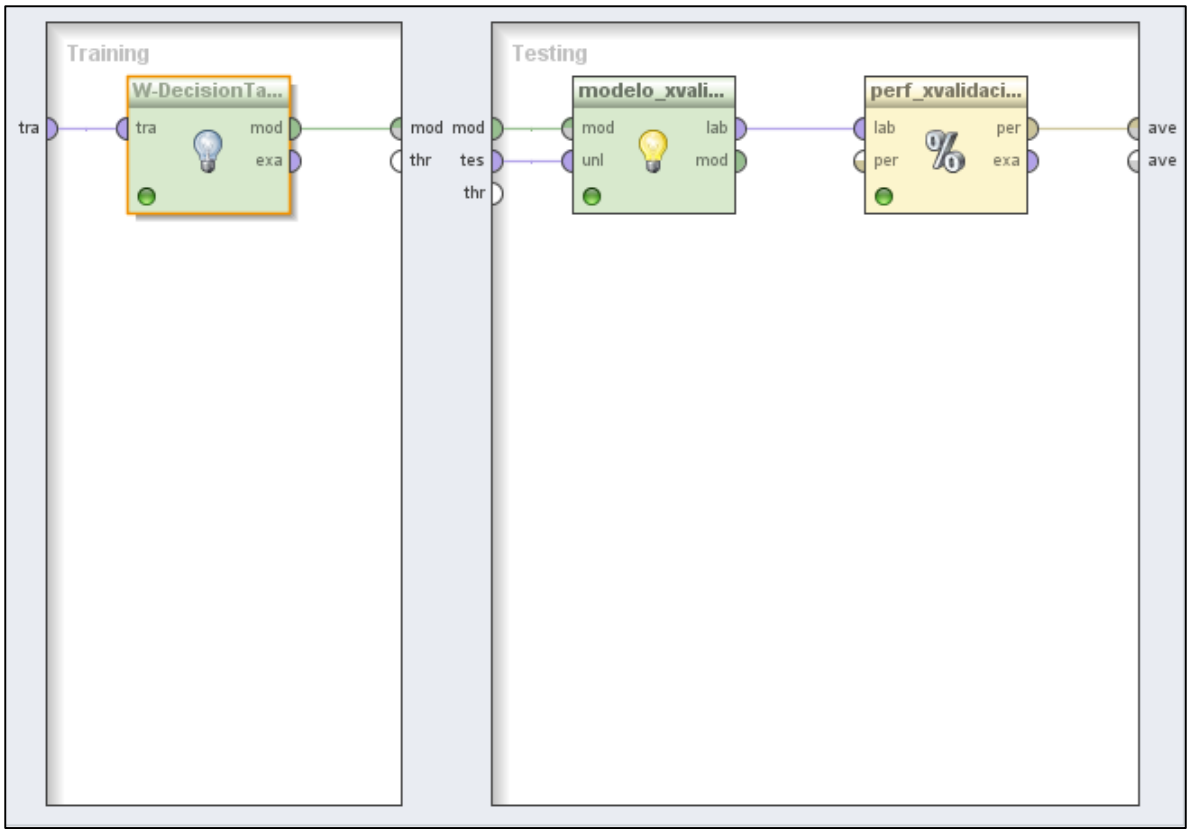


Figura 4. Procesos de Validación Cruzada para el algoritmo Decision__Table.

- **Clasificación mediante DTNB**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 5).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validacion cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-DTNB: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado DTNB, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

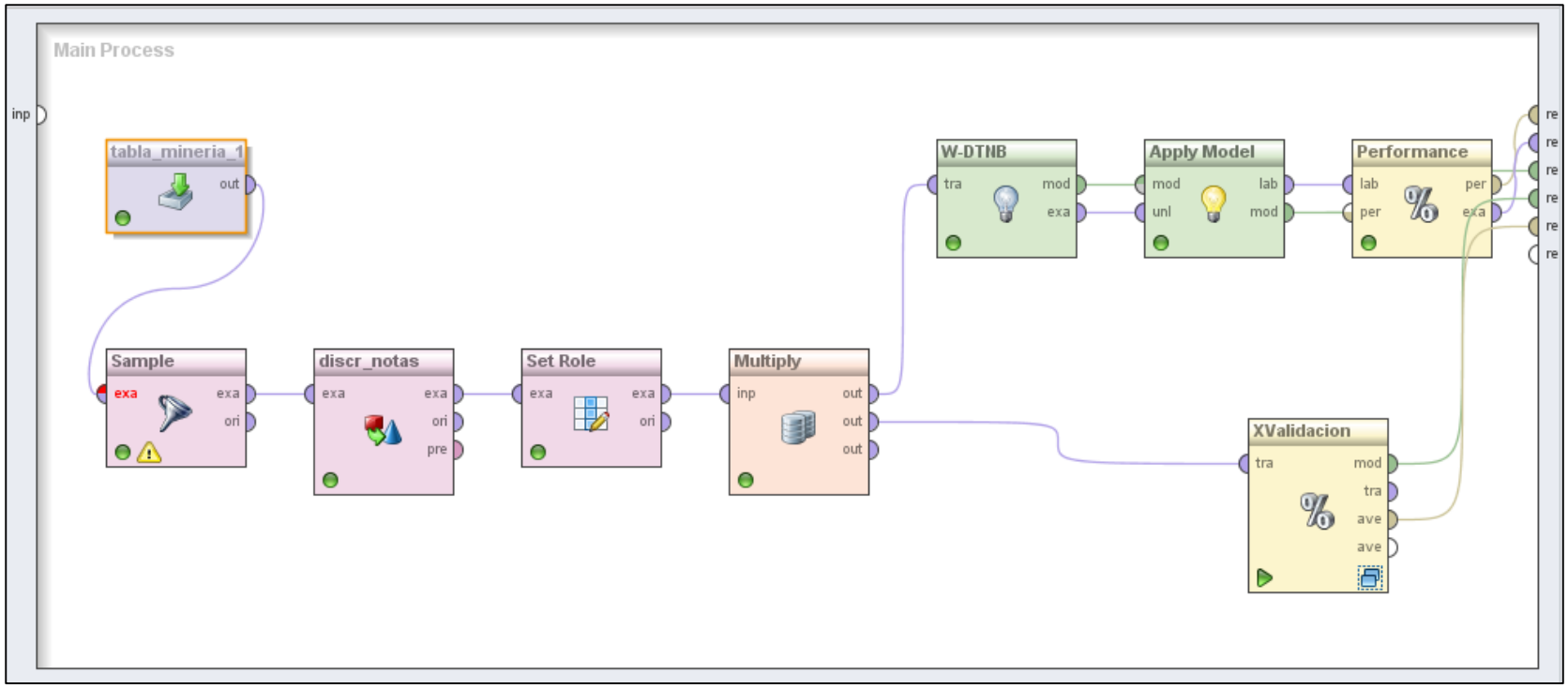


Figura 5. Conjunto de operadores que conforman el proceso para el algoritmo DTNB.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 6).

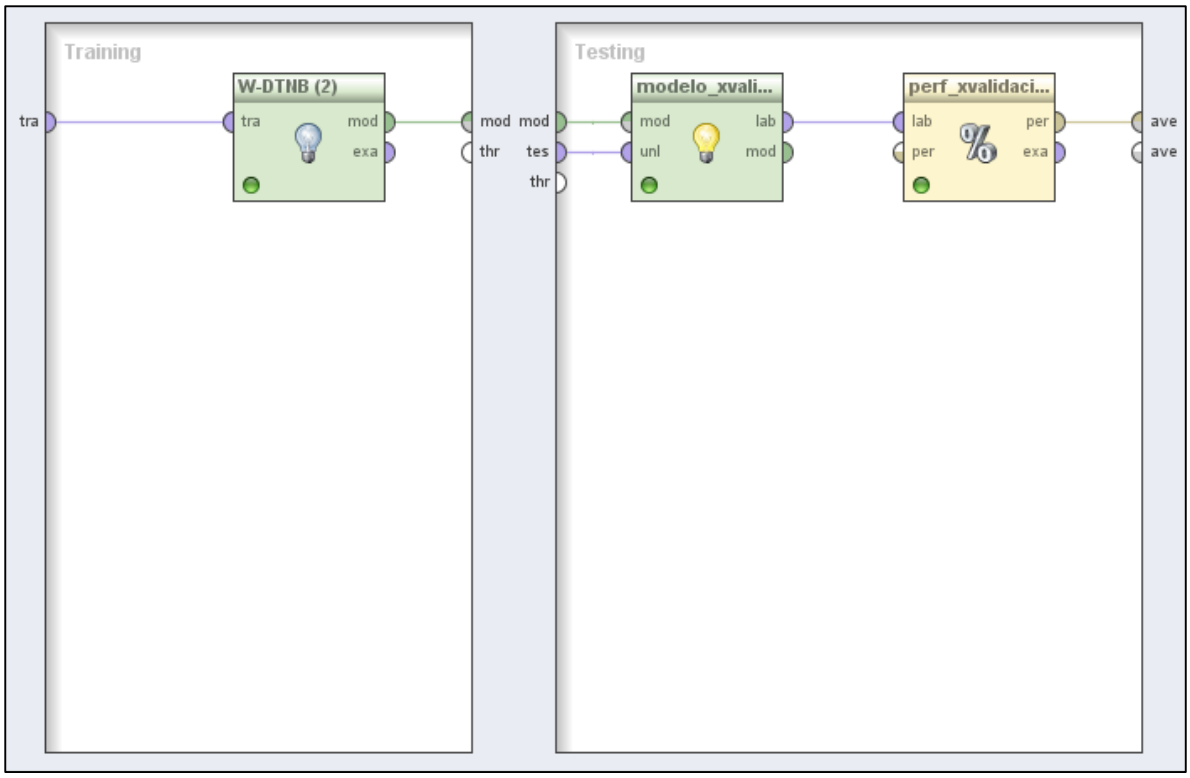


Figura 6. Procesos de Validación Cruzada para el algoritmo DTNB.

- **Clasificación mediante ID3**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 7).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Numerical to Polynominal: Este operados convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse al algoritmo empleado; se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

ID3: Este operador se encuentra dentro del grupo “algoritmos de clasificación en base a la aplicación de árboles de decisión” y se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

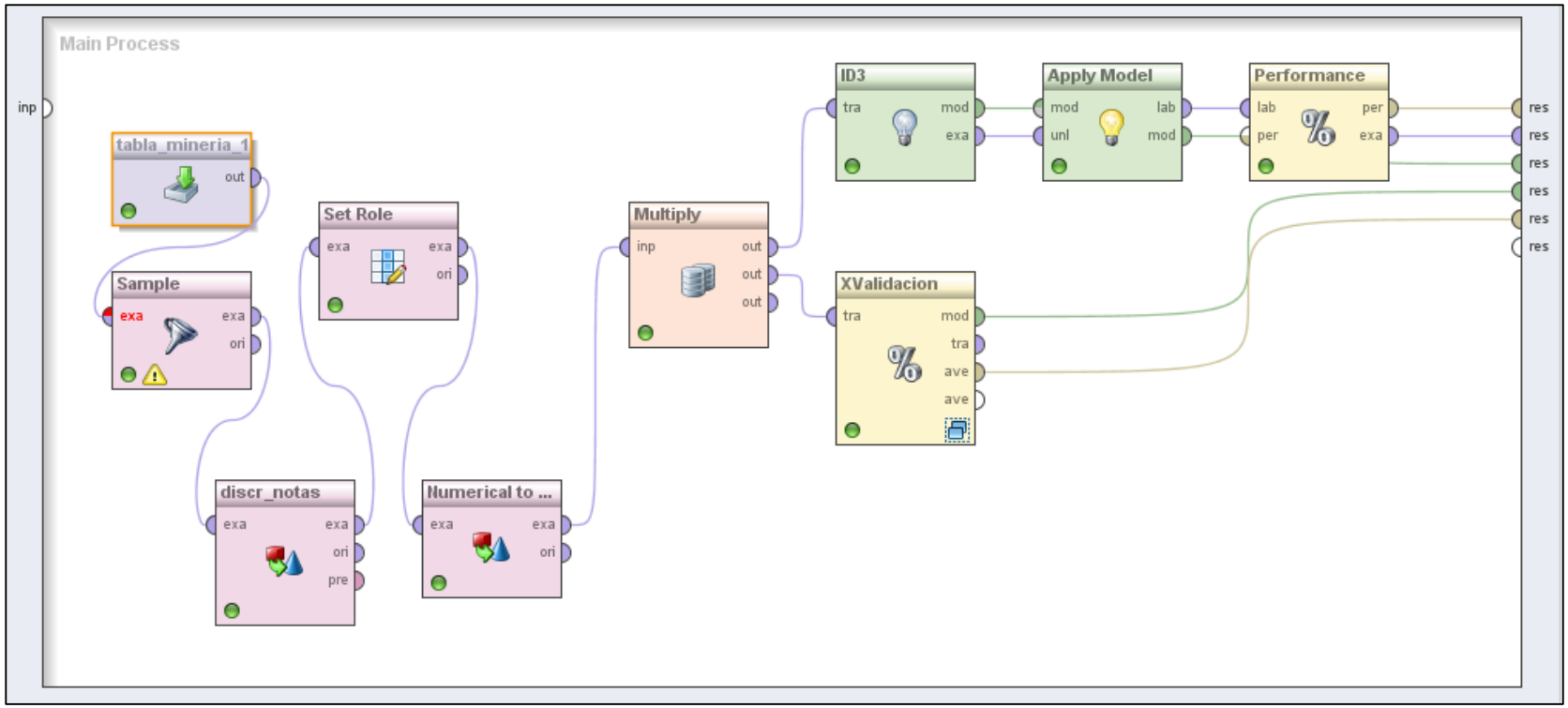


Figura 7. Conjunto de operadores que conforman el proceso para el algoritmo ID3.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 8).

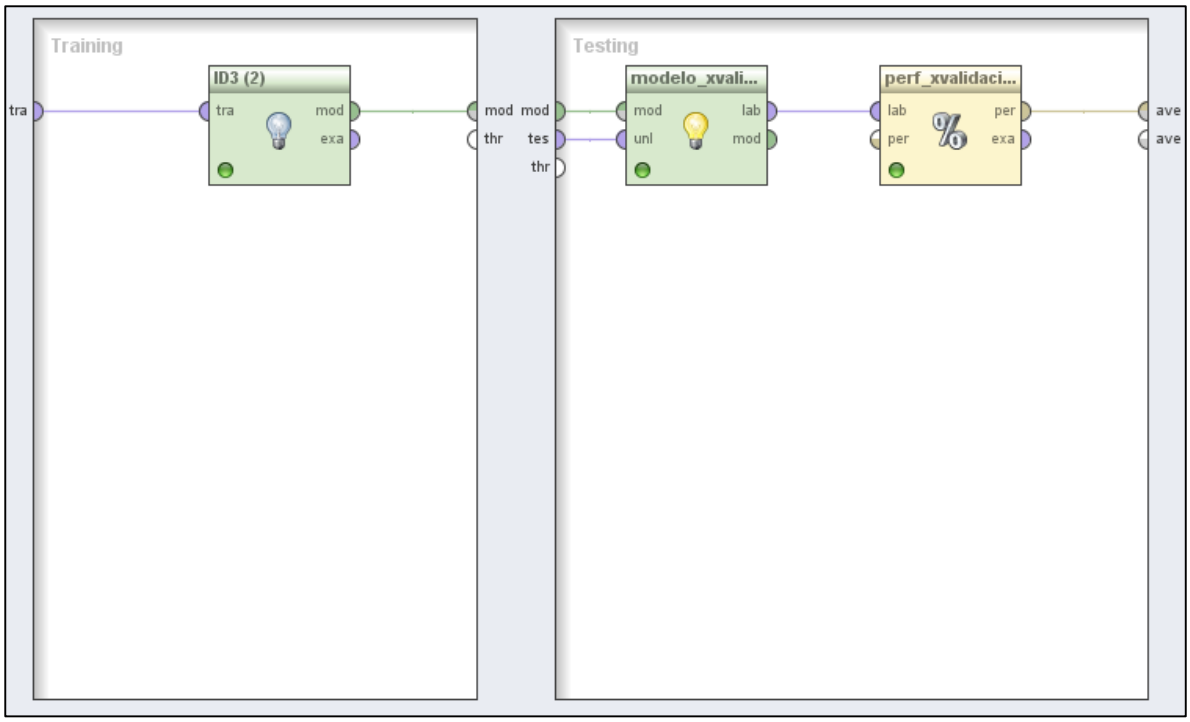


Figura 8. Procesos de Validadción Cruzada para el algoritmo ID3.

- **Clasificación mediante Jrip**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 9).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-JRip: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado JRip, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

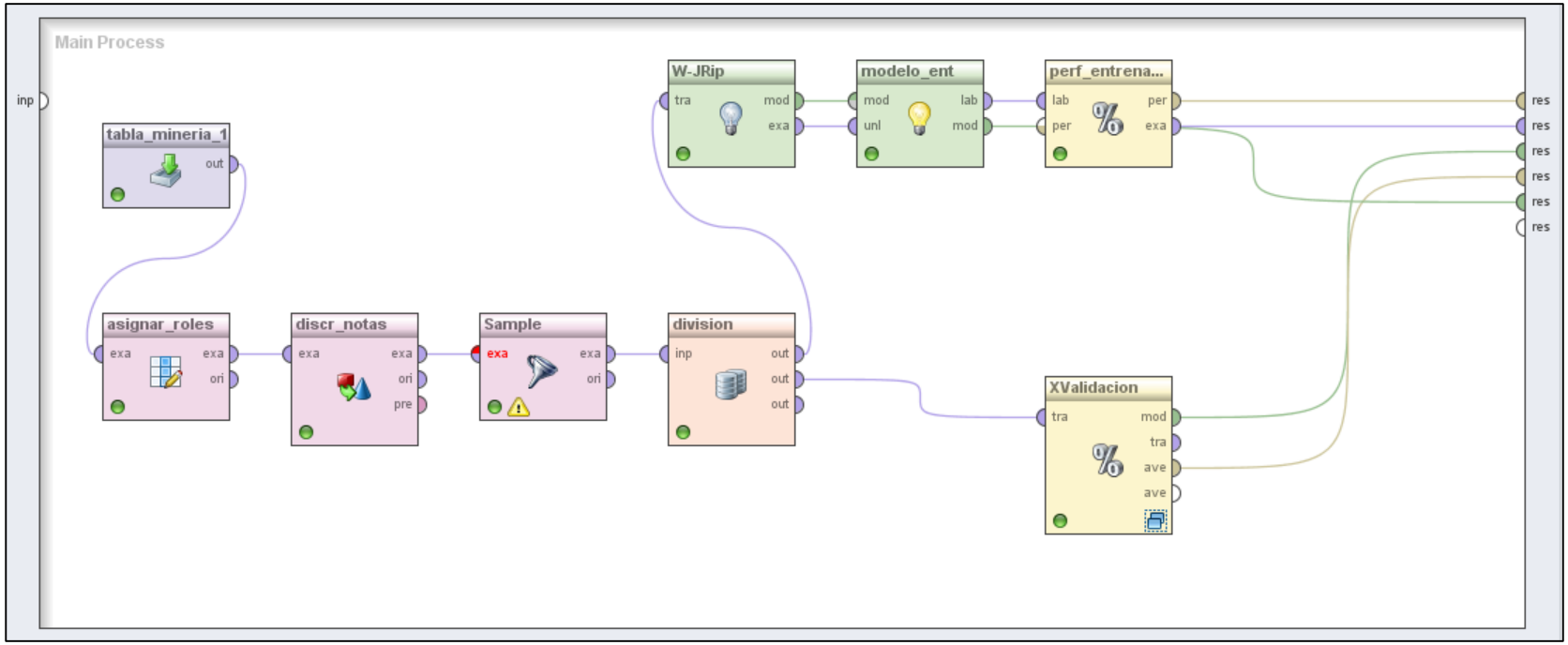


Figura 9. Conjunto de operadores que conforman el proceso para el algoritmo Jrip.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 10).

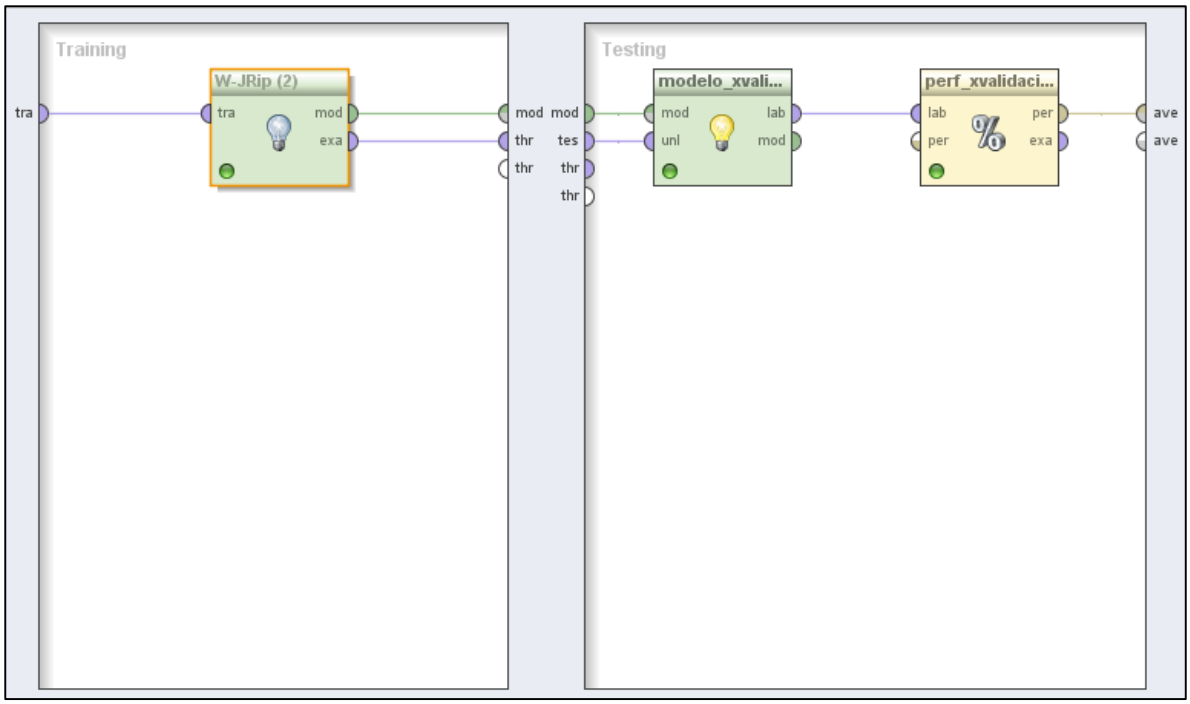


Figura 10. Procesos de Validación Cruzada para el algoritmo Jrip.

- **Clasificación mediante NNge**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 11).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-NNge: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado NNge, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

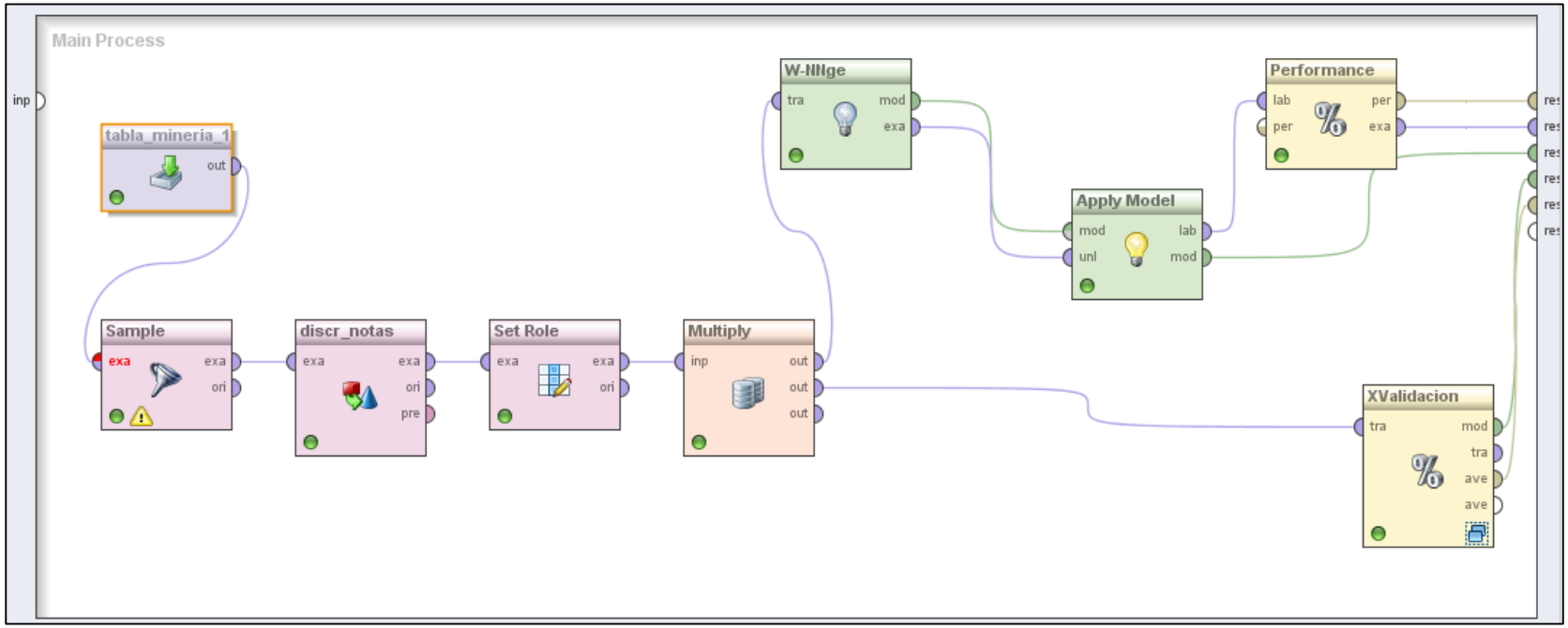


Figura 11. Conjunto de operadores que conforman el proceso para el algoritmo NNge.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 12).

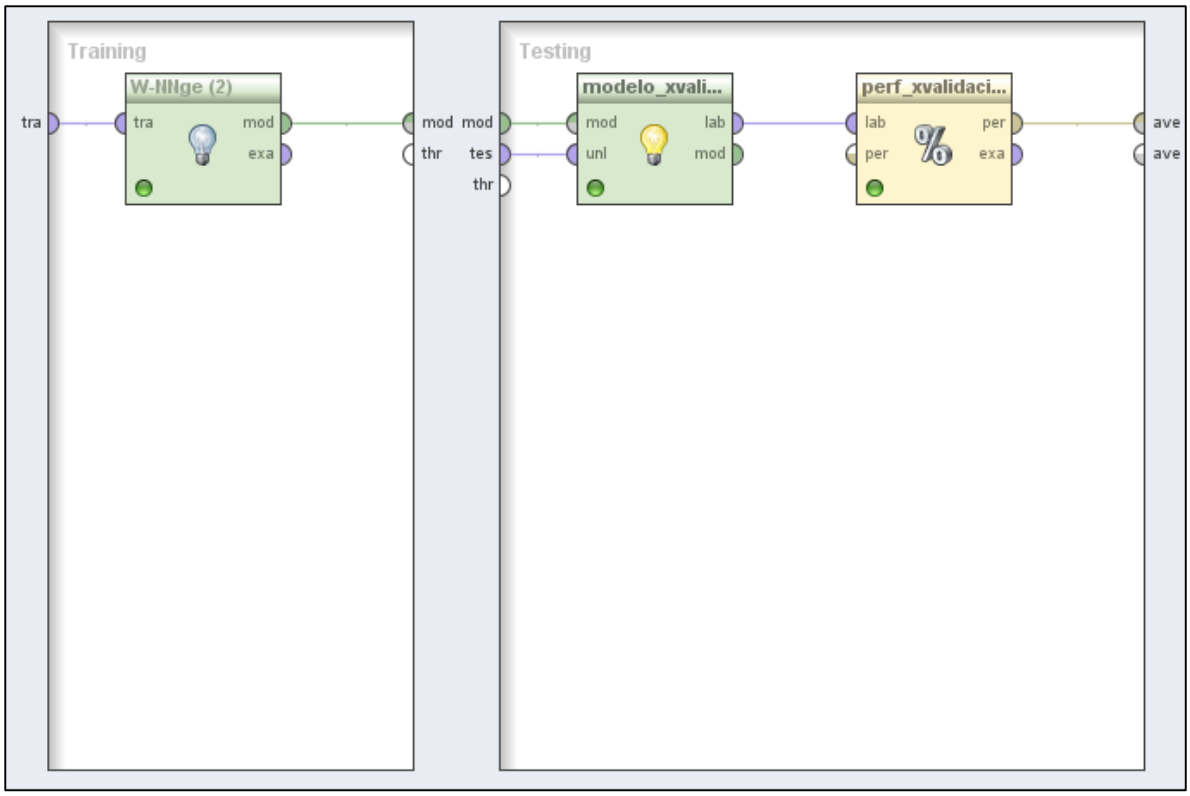


Figura 12. Procesos de Validación Cruzada para el algoritmo NNge.

- **Clasificación mediante Part**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 13).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-PART: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Part, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

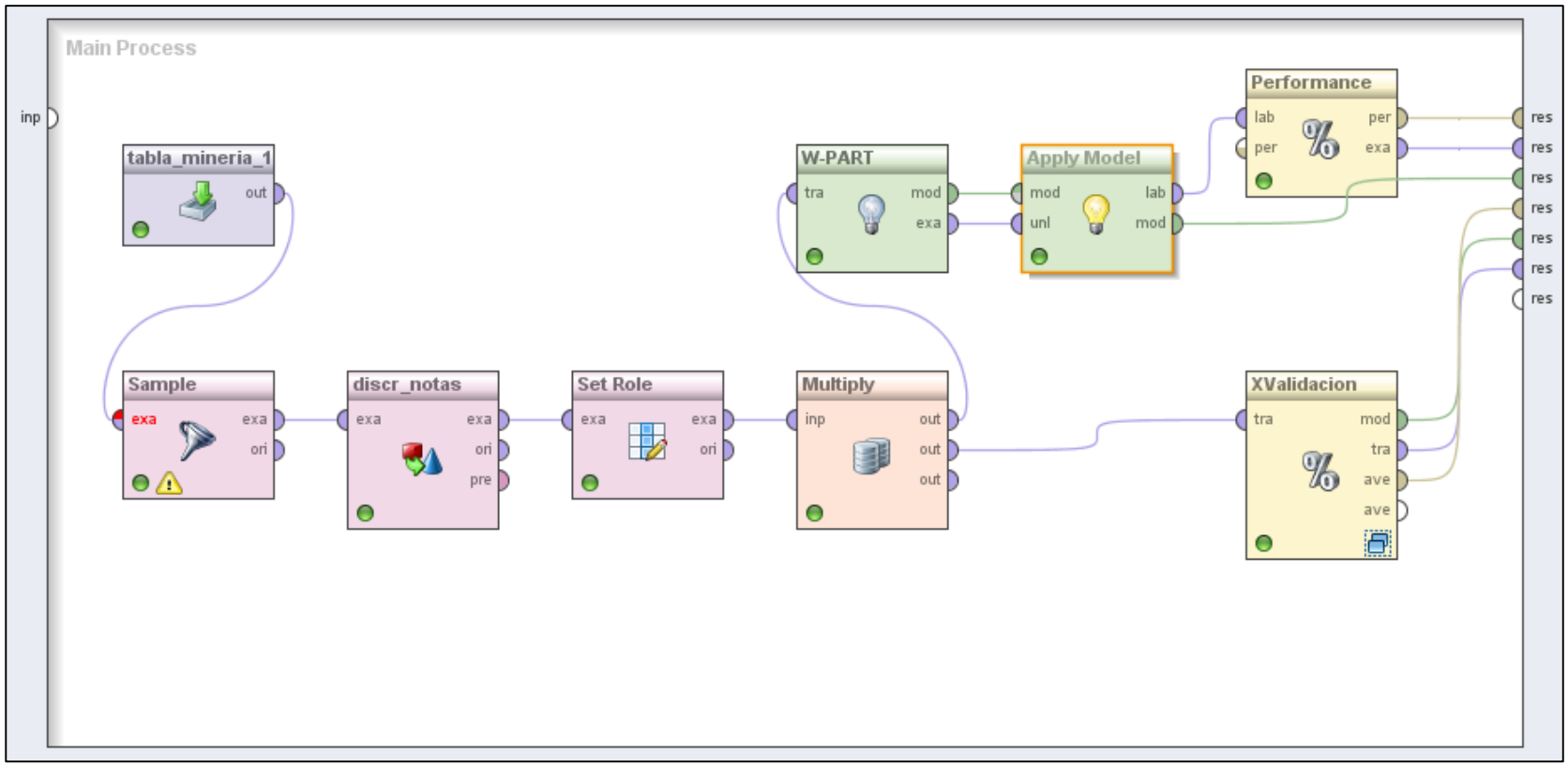


Figura 13. Conjunto de operadores que conforman el proceso para el algoritmo Part.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 14).

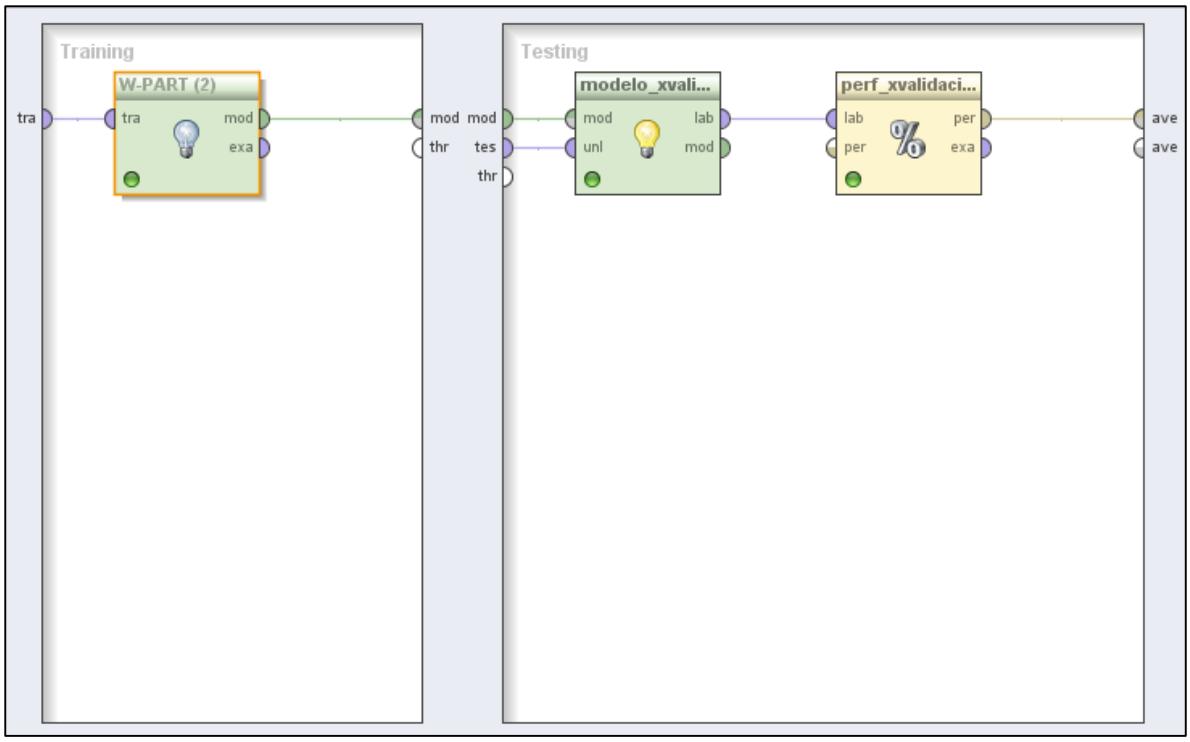


Figura 14. Procesos de Validación Cruzada para el algoritmo Part.

- **Clasificación mediante Ridor**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 15).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados; se lo ha utilizado a su vez en el subproceso de validación cruzada.
Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Set Role: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Multiply: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-Ridor: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Ridor, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Apply Model: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

Performance: Este proceso genera la matriz de confusión del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

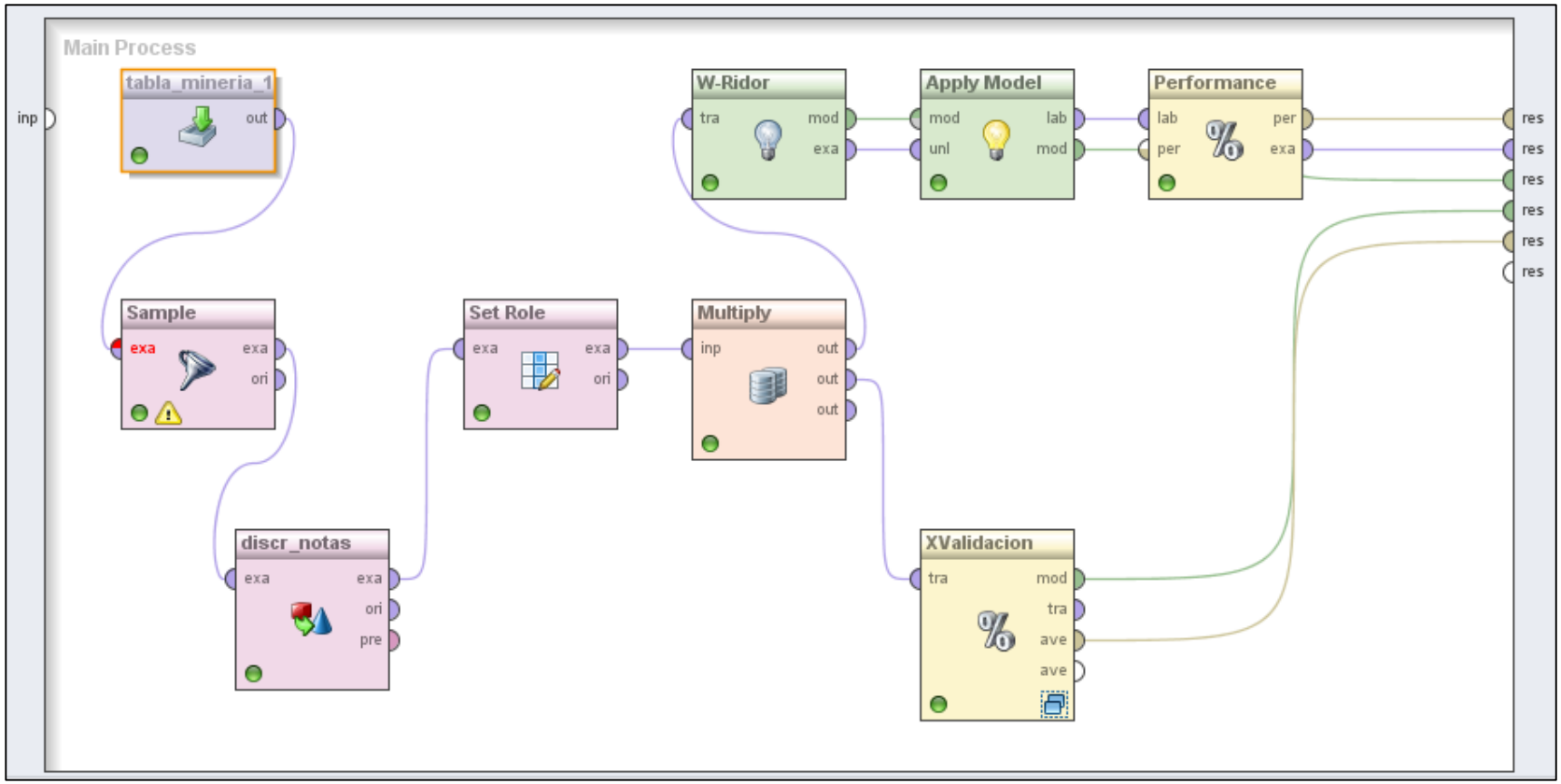


Figura 15. Conjunto de operadores que conforman el proceso para el algoritmo Ridor.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 16).

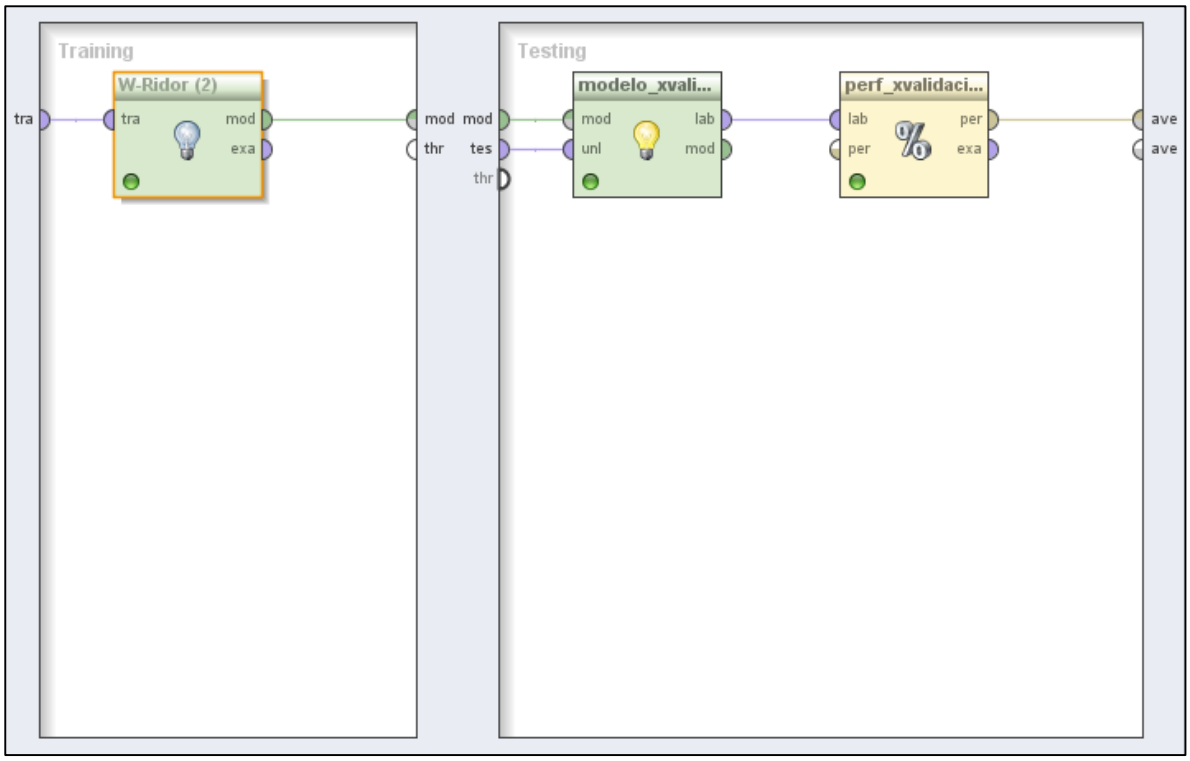


Figura 16. Procesos de Validación Cruzada para el algoritmo Ridor.

✓ **Pruebas de entrenamiento y validación cruzada con dattos no Agrupados.**

A continuación se describen los procesos formados para los algoritmos CHAID, Decision__Table, DTNB, ID3, Jrip, NNge, Part, Ridor para los datos agrupados.

- **Clasificación mediante CHAID**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 17).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

CHAID: Este operador se encuentra dentro del grupo “algoritmos de clasificación en base a la aplicación de árboles de decisión” y se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

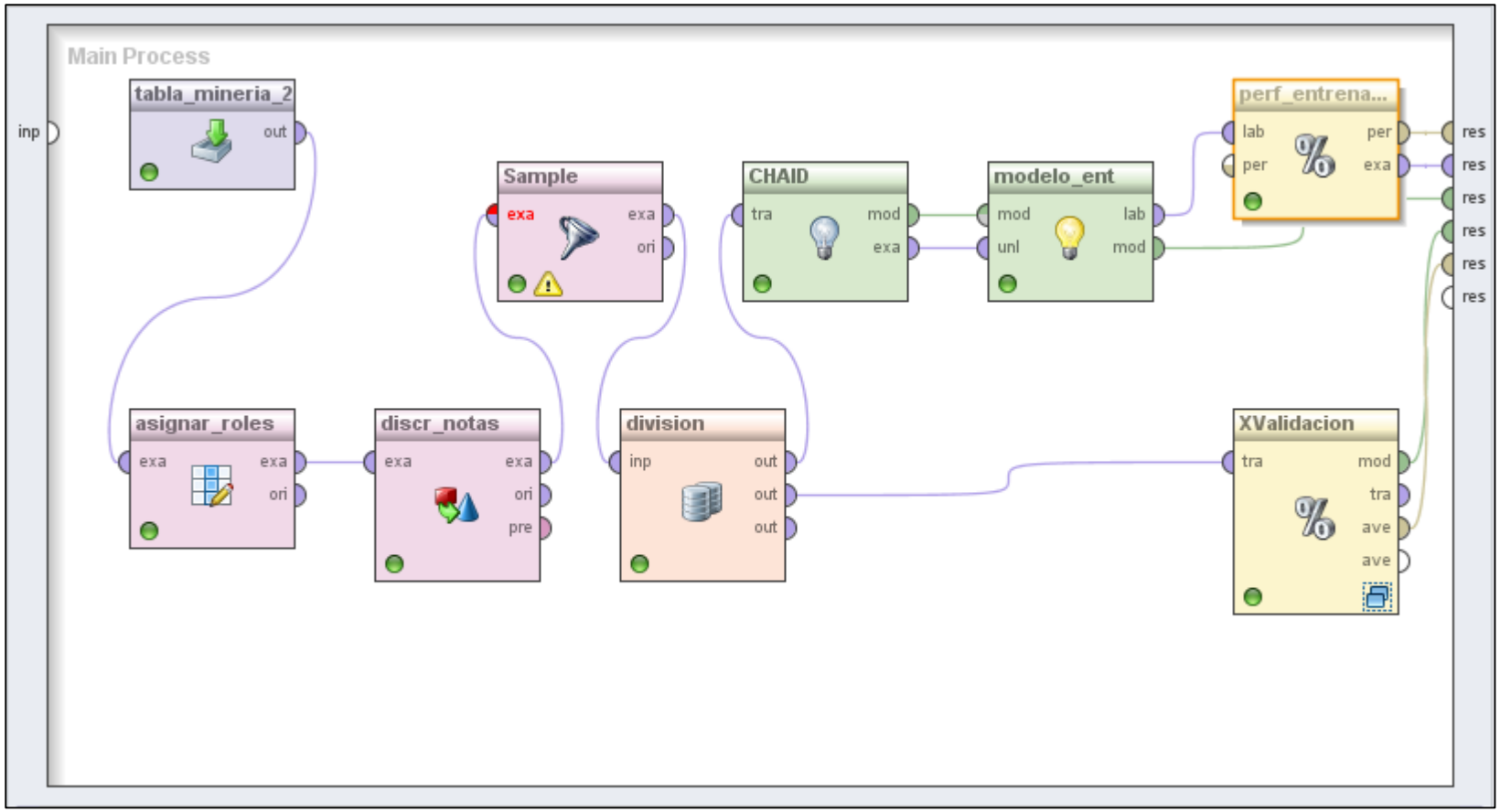


Figura 17. Conjunto de operadores que conforman el proceso para el algoritmo CHAID.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 18).

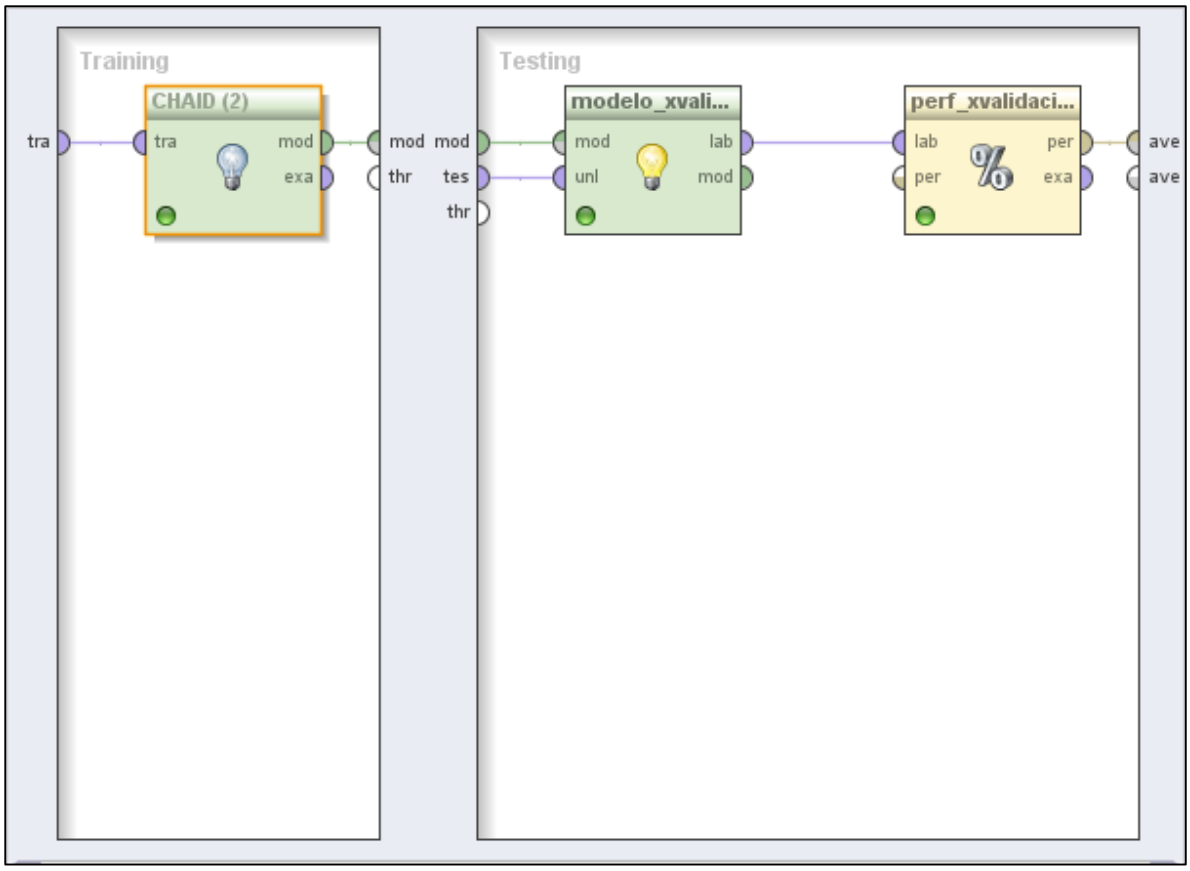


Figura 18. Procesos de Validación Cruzada para el algoritmo CHAID.

- **Clasificación mediante Decision__Table**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 19).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-DecisionTable: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Decision Table, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

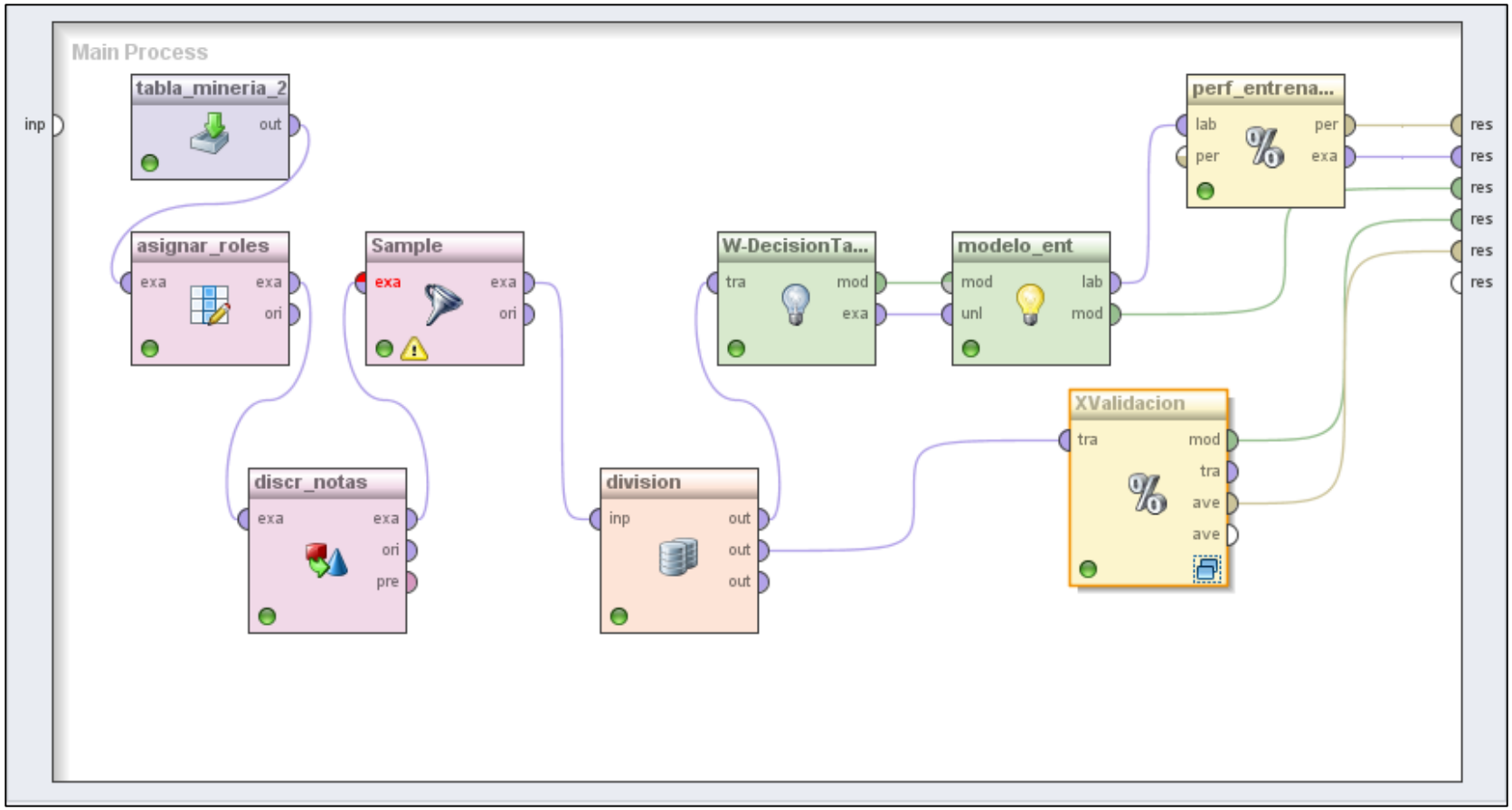


Figura 19. Conjunto de operadores que conforman el proceso para el algoritmo Decision__Table.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 20).

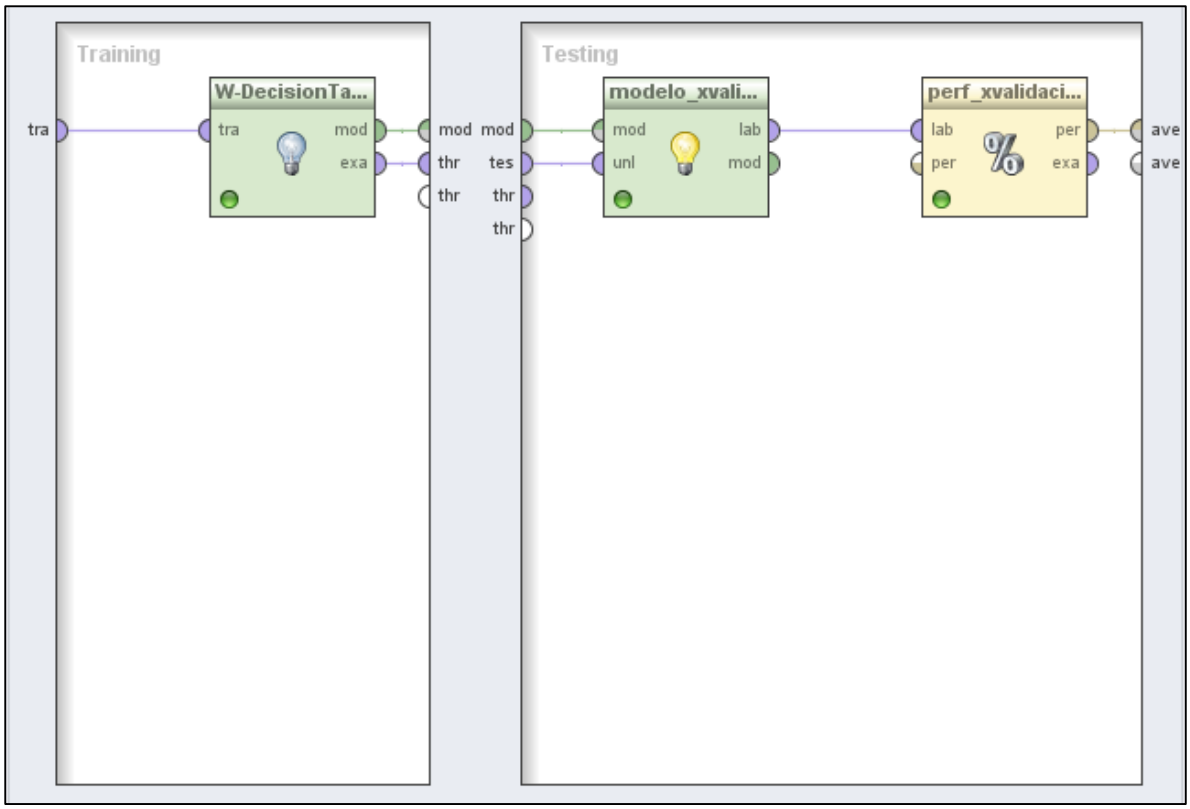


Figura 20. Procesos de Validación Cruzada para el algoritmo Decision__Table.

- **Clasificación mediante DTNB**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 21).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-DTNB: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado DTNB, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

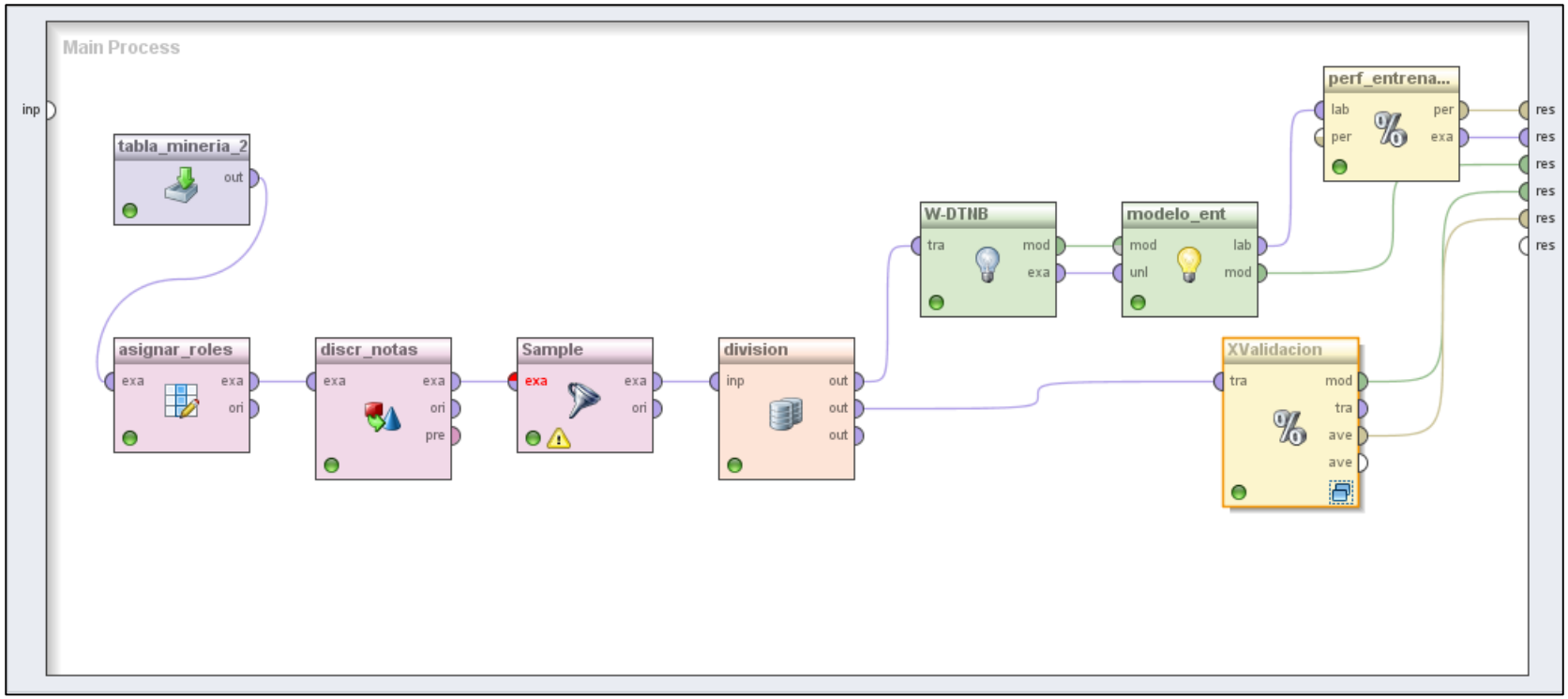


Figura 21. Conjunto de operadores que conforman el proceso para el algoritmo DTNB.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 22).

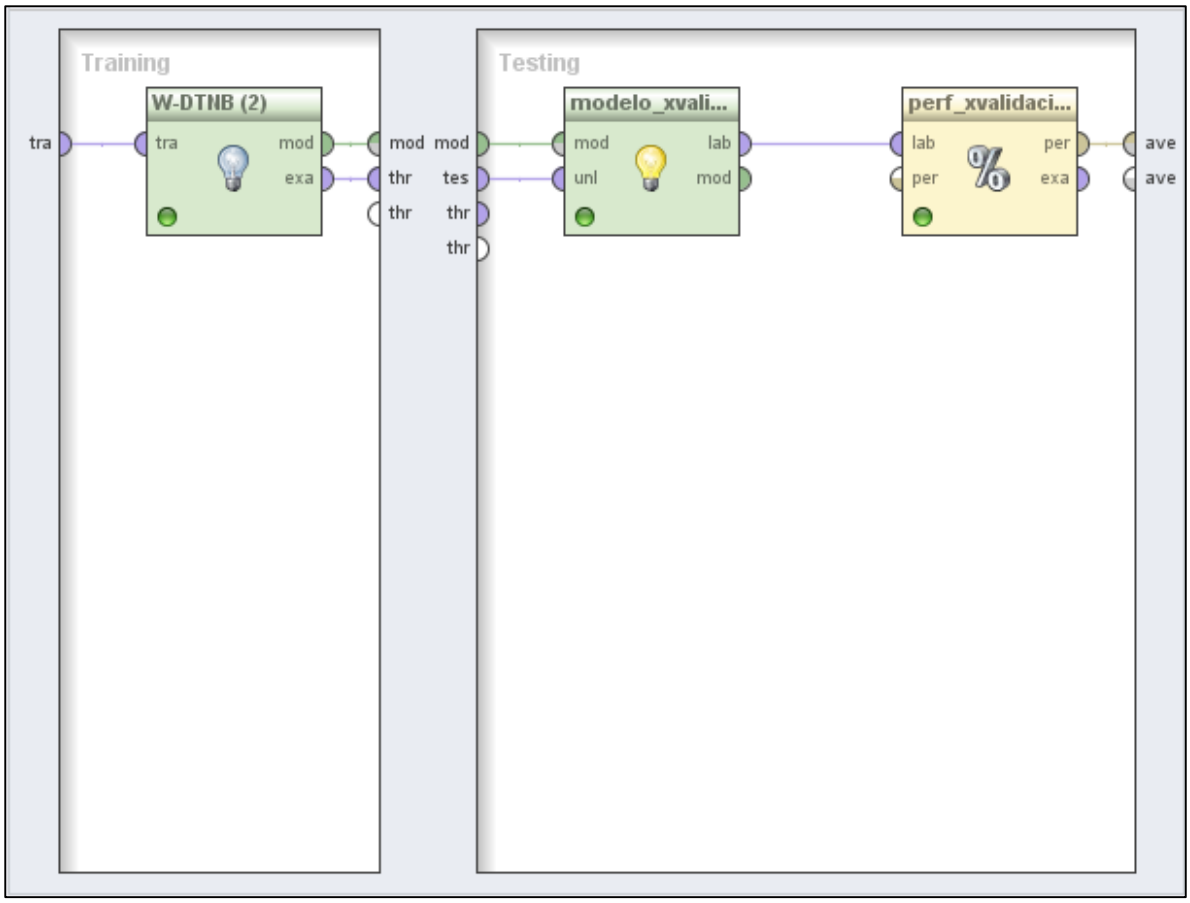


Figura 22. Procesos de Validación Cruzada para el algoritmo DTNB.

- **Clasificación mediante ID3**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 23).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

ID3: Este operador se encuentra dentro del grupo “algoritmos de clasificación en base a la aplicación de árboles de decisión” y se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

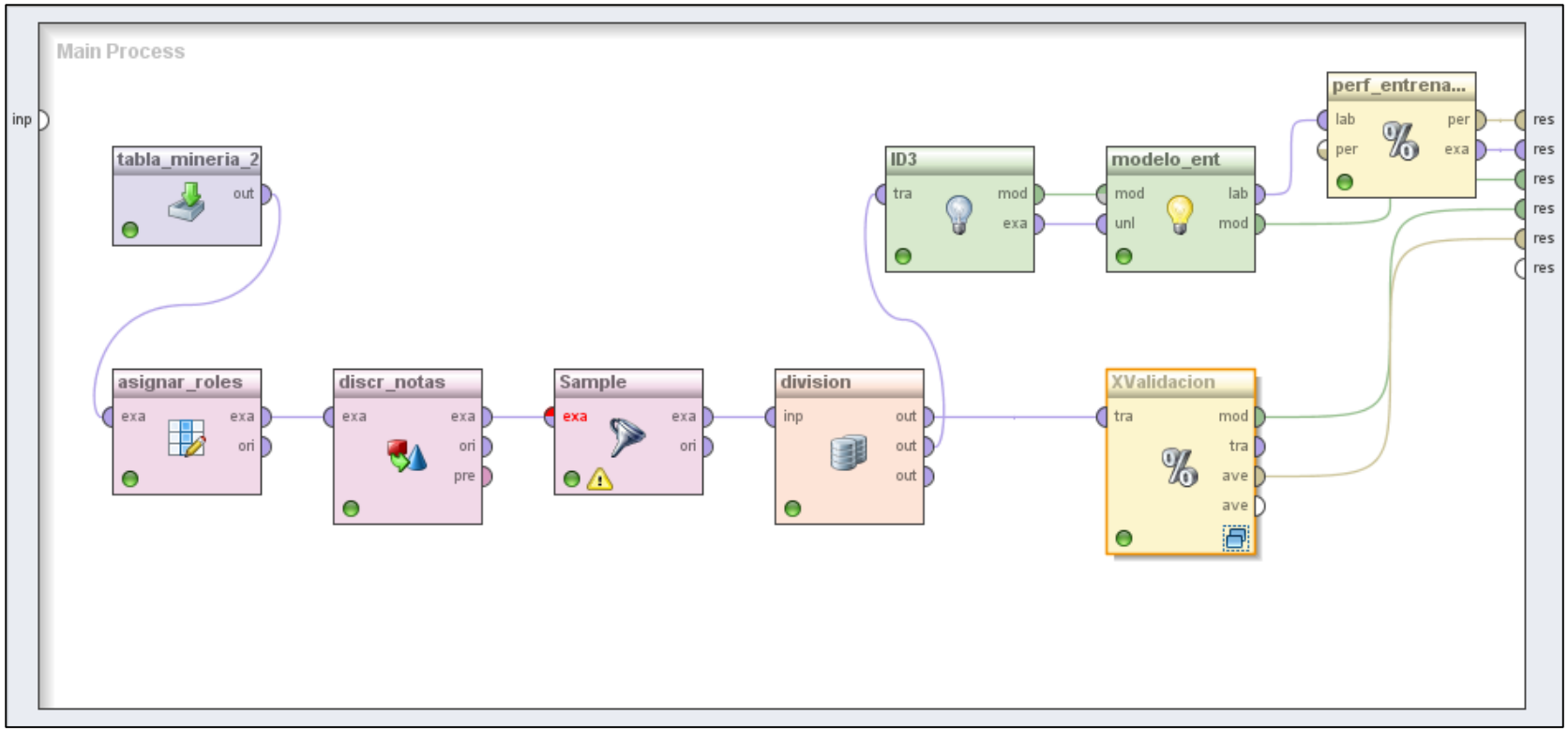


Figura 23. Conjunto de operadores que conforman el proceso para el algoritmo ID3.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 24).

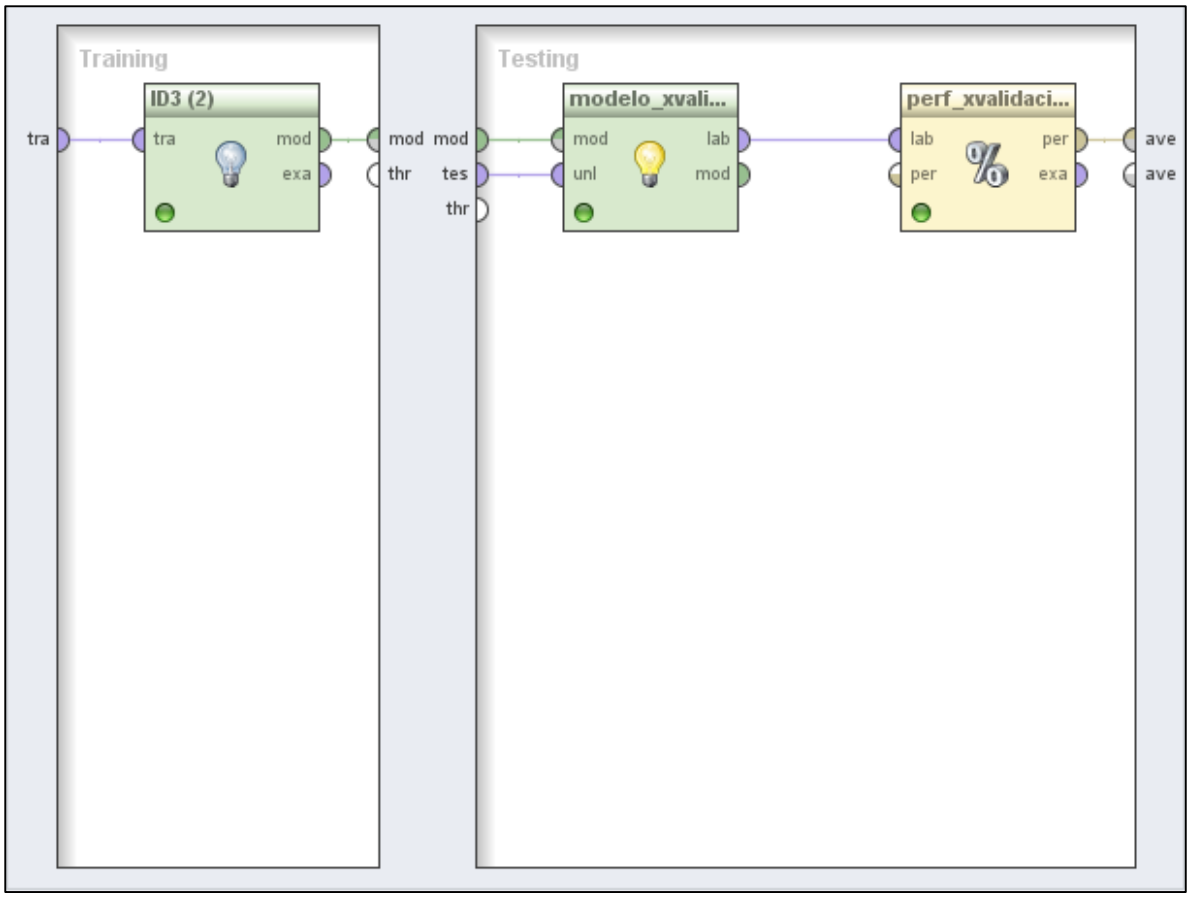


Figura 24. Procesos de Validación Cruzada para el algoritmo ID3.

- **Clasificación mediante Jrip**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 25).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-JRip: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado JRip, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

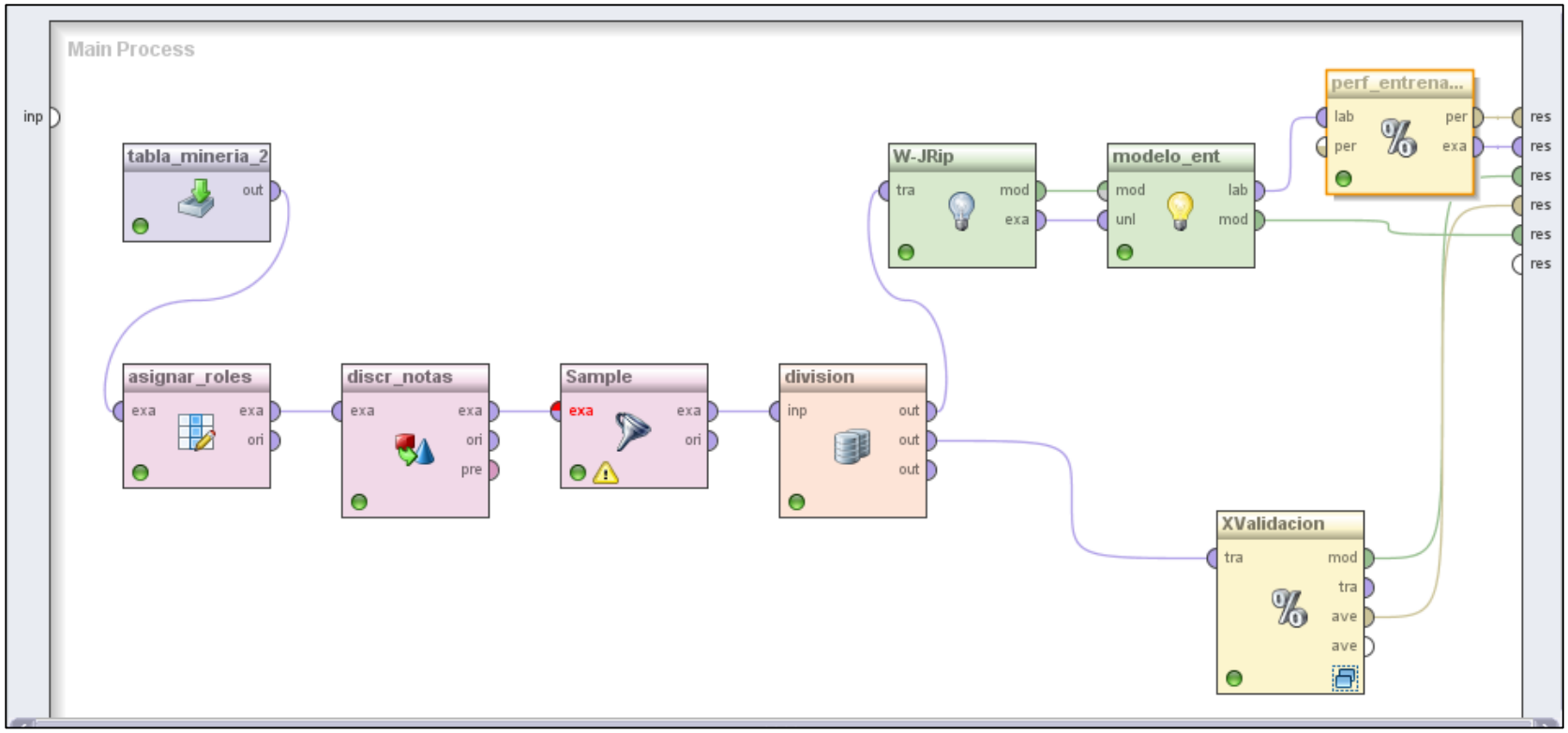


Figura 25. Conjunto de operadores que conforman el proceso para el algoritmo Jrip.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 26).

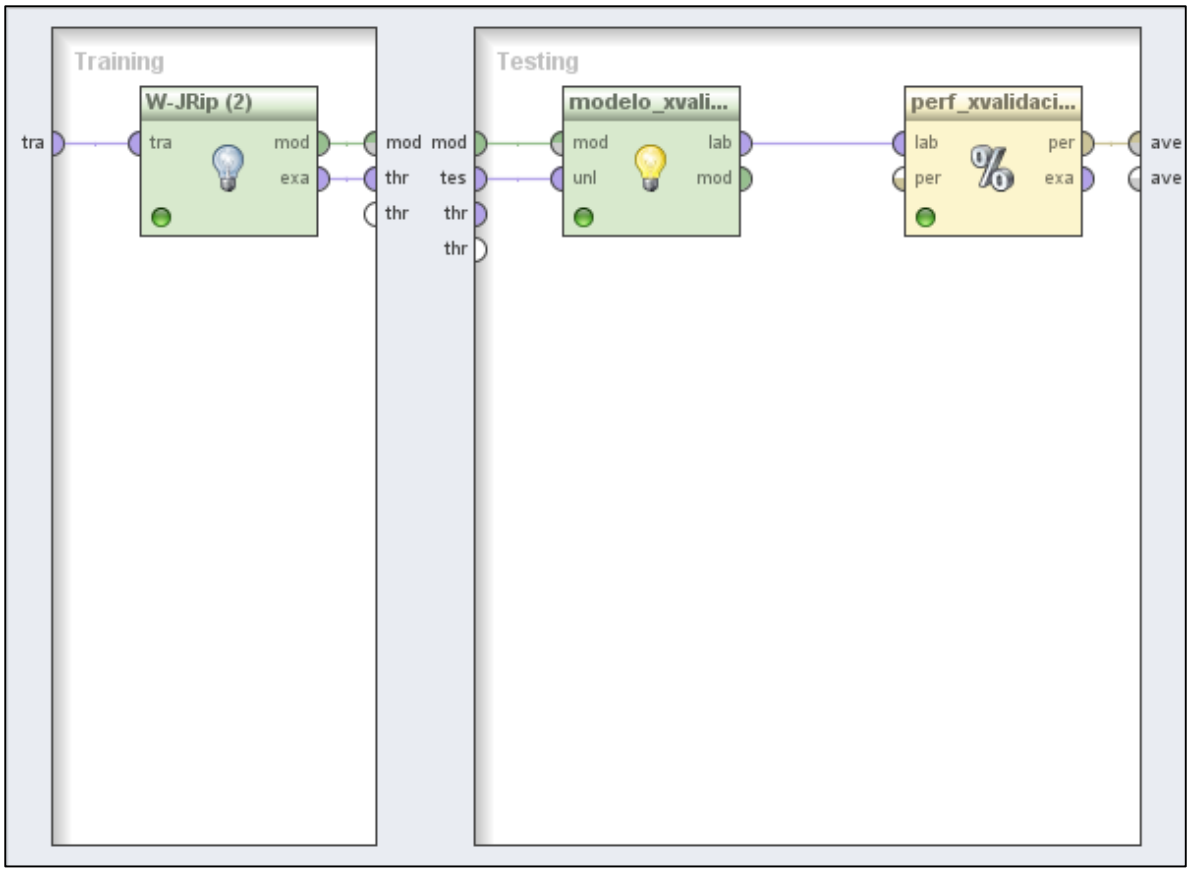


Figura 26. Procesos de Validación Cruzada para el algoritmo Jrip.

- **Clasificación mediante NNge**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 27).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-NNge: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado NNge, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidacion: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

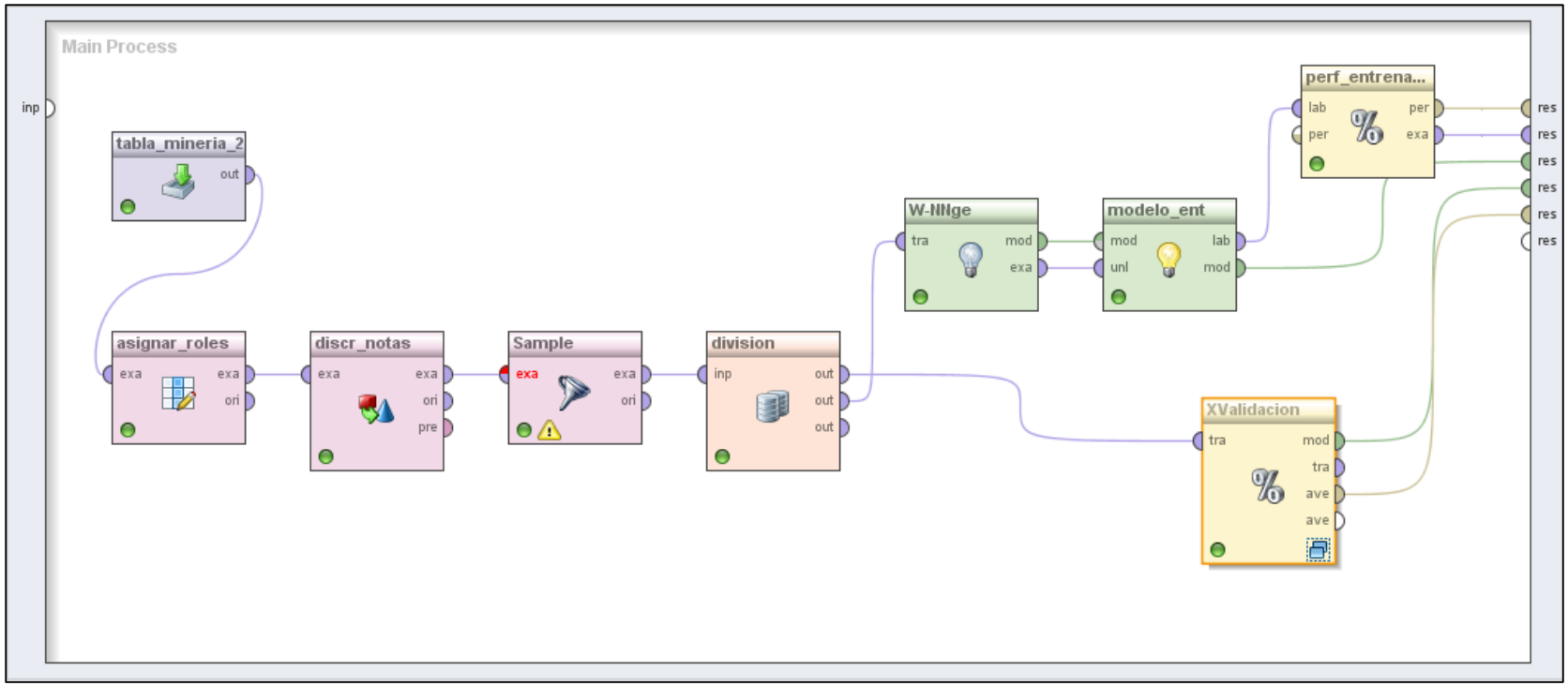


Figura 27. Conjunto de operadores que conforman el proceso para el algoritmo NNge.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 28).

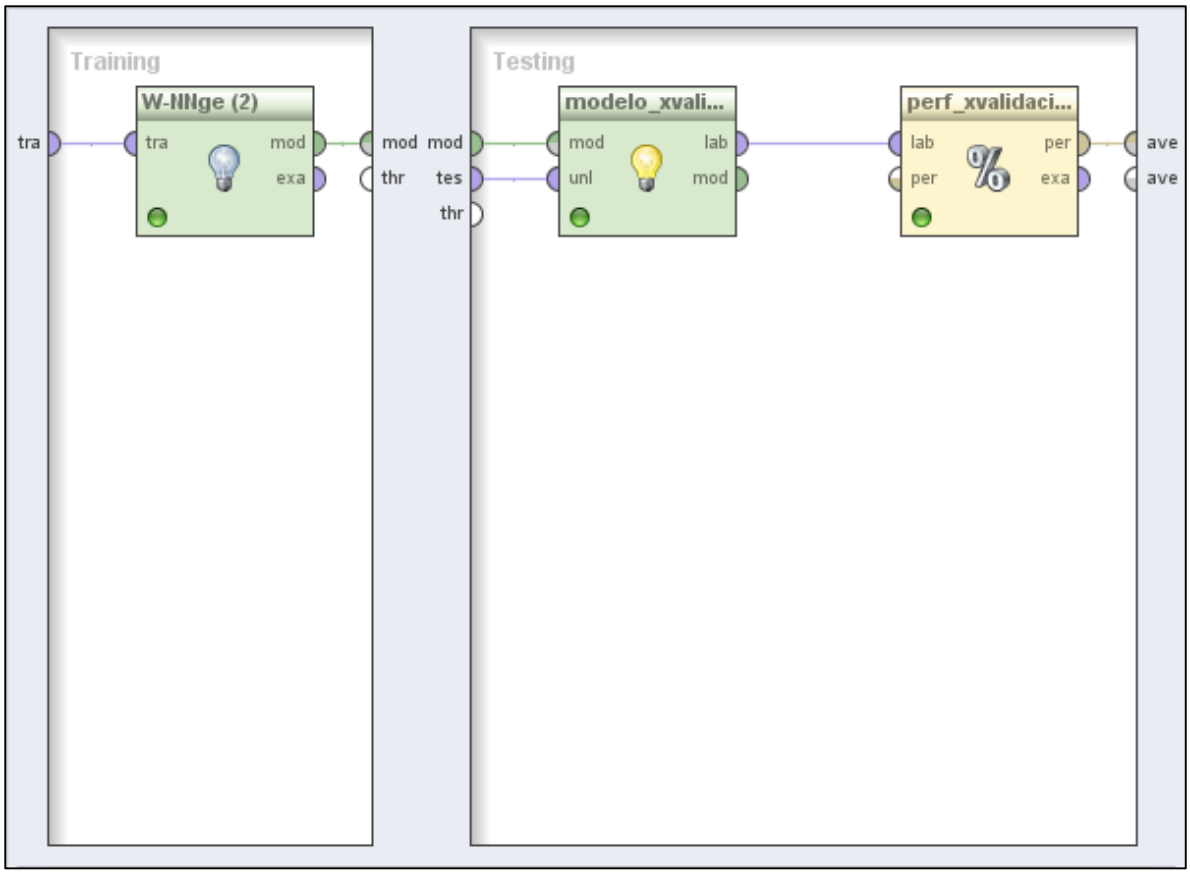


Figura 28. Procesos de Validadción Cruzada para el algoritmo NNge.

- **Clasificación mediante Part**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 29).

tabla_mineria_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validacion cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-PART: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Part, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidación: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

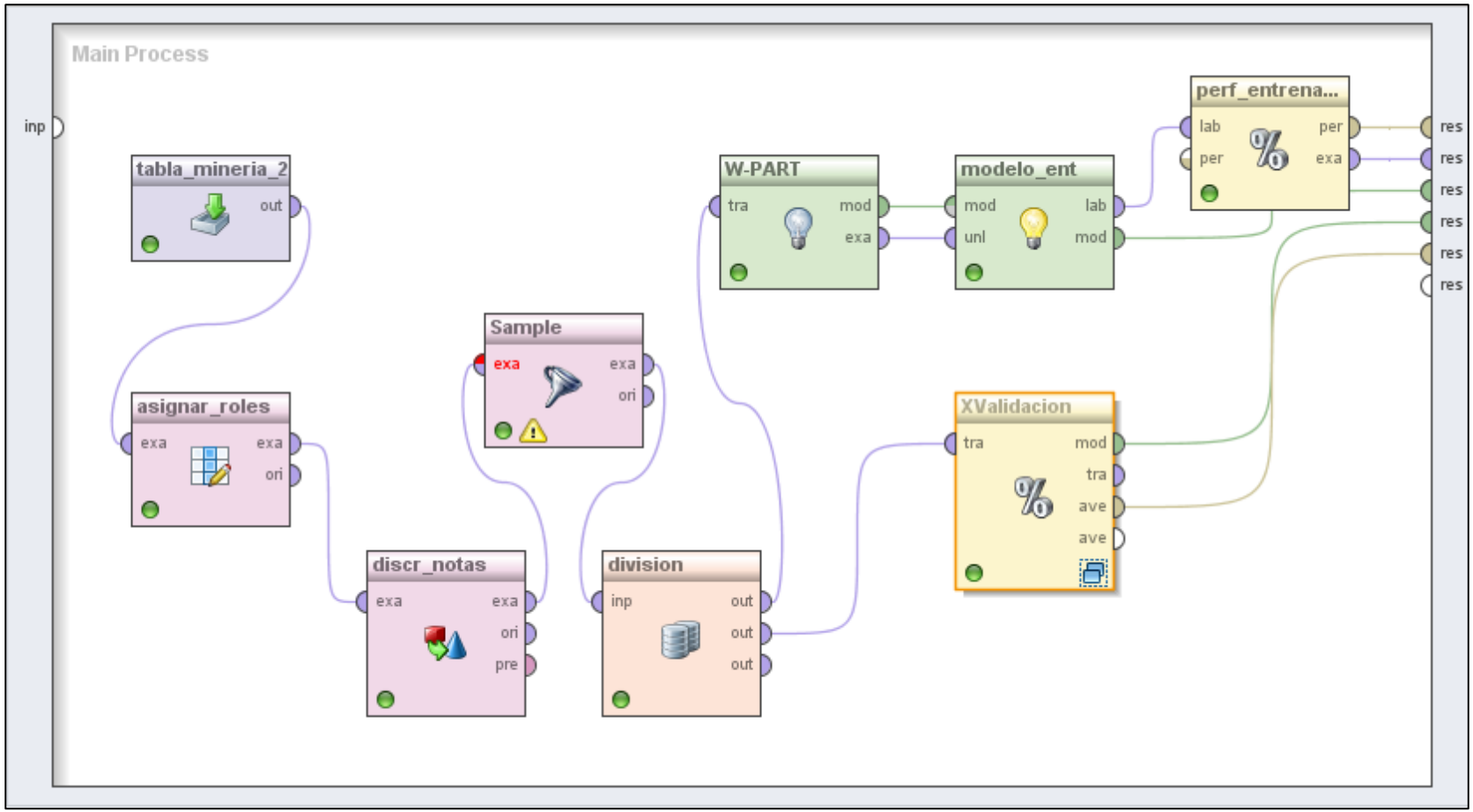


Figura 29. Conjunto de operadores que conforman el proceso para el algoritmo Part.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 30).

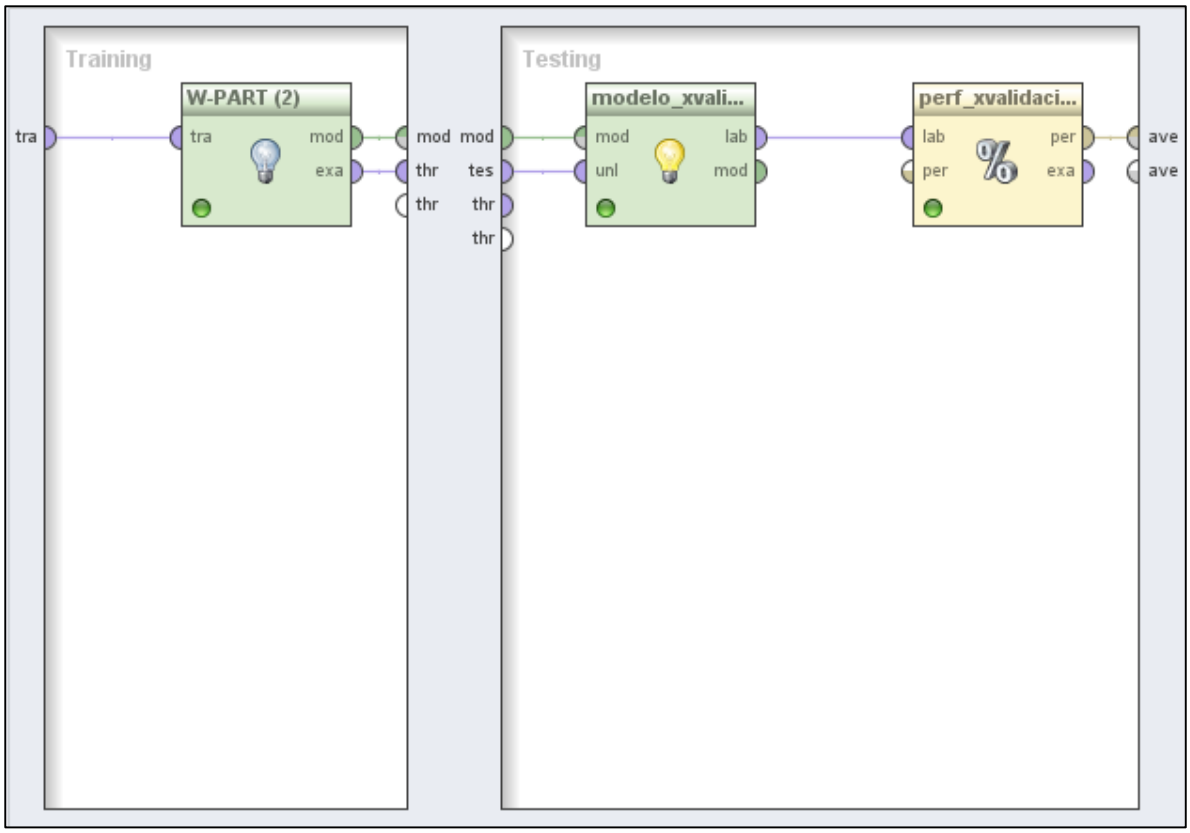


Figura 30. Procesos de Validación Cruzada para el algoritmo Part.

- **Clasificación mediante Ridor**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 31).

table_miner_2: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería dos de los datos agrupados. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

asignar_roles: Este operador se ha utilizado para asignar los roles necesarios, “label” a la variable dependiente perfil profesional y el atributo “cedula” como ID. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

Sample: Corresponde al 72% de los datos que se han utilizado para el conjunto de entrenamiento en el proceso de minería; se lo ha utilizado a su vez en el subproceso de validación cruzada.

division: Este operador divide la secuencia de procesos; para encontrar el modelo en base al algoritmo utilizado y para realizar la validación cruzada.

W-Ridor: Este operador pertenece a uno de los algoritmos de inducción de reglas denominado Ridor, que internamente determina el modelo. Se lo ha utilizado a su vez en el subproceso de validación cruzada.

modelo_ent: Este operador se utiliza para generar el modelo del algoritmo aplicado y realizar su evaluación y ha sido utilizado también dentro del subproceso de validación cruzada.

perf_entrenamiento: Este proceso genera la matriz de confusion del algoritmo aplicado, presentando los resultados a detalle de valores como el “error absoluto”. “error relativo”, “error de clasificación” etc; y ha sido utilizado también en el subproceso de validación cruzada.

XValidación: Este operador realiza la evaluación del modelo utilizando el método de validación cruzada.

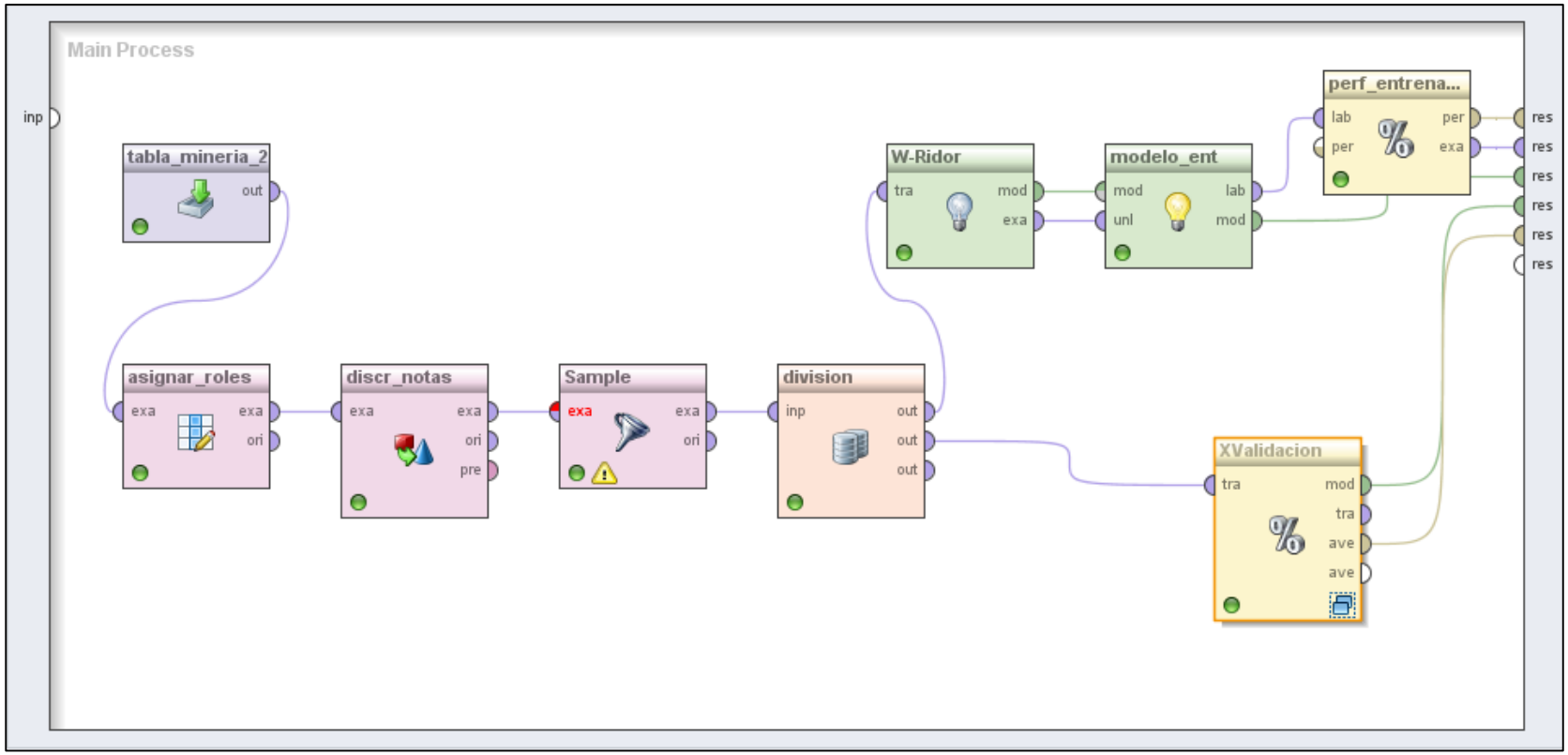


Figura 31. Conjunto de operadores que conforman el proceso para el algoritmo Ridor.

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 32).

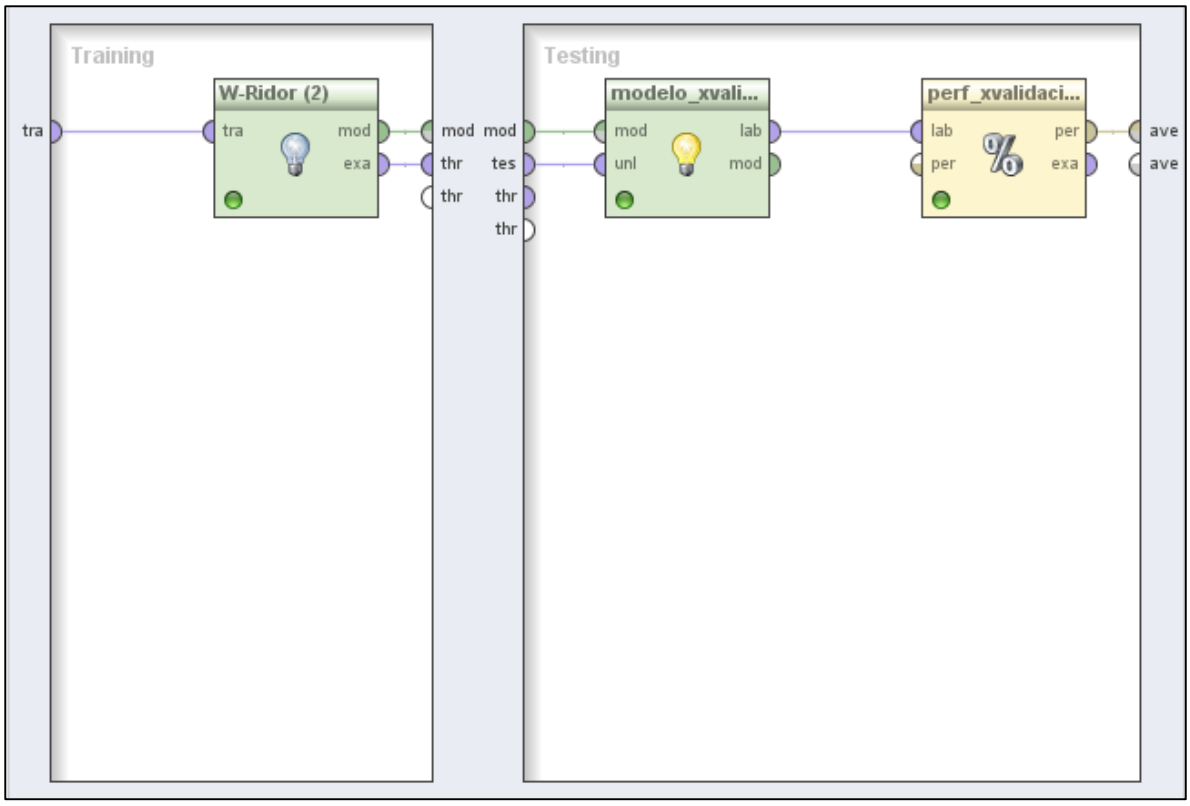


Figura 32. Procesos de Validación Cruzada para el algoritmo Ridor

- ✓ **Procesos de la evaluación de los modelos generados por los mejores algoritmos CHAID Y JRip.**

En este apartado se ha detallado los procesos de la evaluación de los modelos con los mejores algoritmos en este caso CHAID y JRip, los operadores utilizados para probar los modelos generados tomando el 28% de los datos restantes recopilados inicialmente de estudiantes egresados y graduados de la Carrera de Ingeniería en Sistemas.

- **Evaluación del modelo CHAID.**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 33).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados.
Sample: Corresponde al 28% de los datos que se han utilizado para la evaluación del modelo, que es el porcentaje restante del conjunto de entrenamiento tomado.
discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente.
Set Role: Este operador se ha utilizado para asignar los roles necesarios, "label" a la variable dependiente perfil profesional y el atributo "cedula" como ID.
Numerical to Polynominal: Este operador convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse a la evaluación del modelo realizado.
Rd Model CHAID: este operador carga el modelo generado por el algoritmo CHAID que ha sido almacenado previamente.
Apply Model: Este operador se utiliza para generar el modelo de la evaluación del algoritmo CHAID.
Performance: Este proceso genera la matriz de confusion de la evaluación del modelo, presentando los resultados a detalle como la clasificación o coincidencia de la predicción para cada perfil profesional.

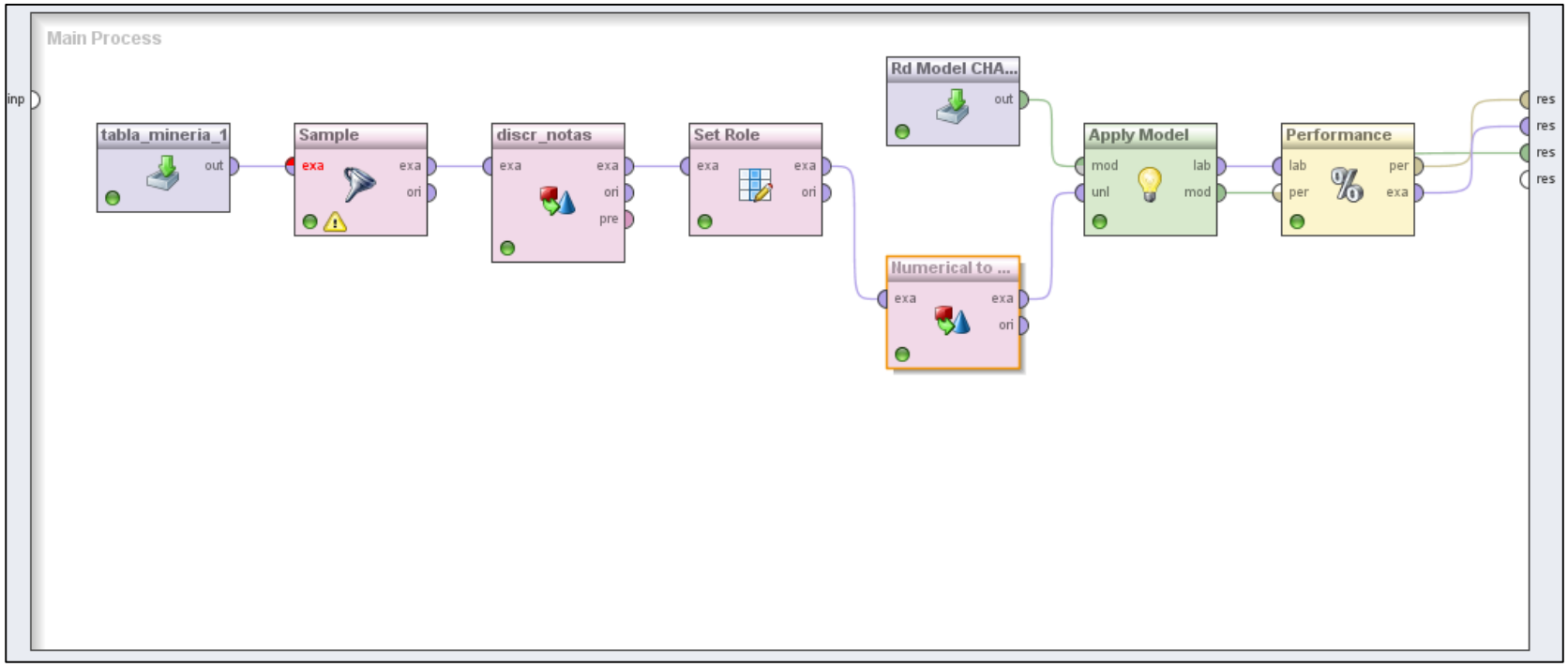


Figura 33. Conjunto de operadores que forman el proceso de evaluación del modelo del algoritmo CHAID.

- **Evaluación del modelo JRip.**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 34).

tabla_mineria_1: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería uno de los datos no agrupados.
Sample: Corresponde al 28% de los datos que se han utilizado para la evaluación del modelo, que es el porcentaje restante del conjunto de entrenamiento tomado.
discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente.
Set Role: Este operador se ha utilizado para asignar los roles necesarios, "label" a la variable dependiente perfil profesional y el atributo "cedula" como ID. S
Numerical to Polynominal: Este operador convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse a la evaluación del modelo.
Rd Model JRip: este operador carga el modelo generado por el algoritmo JRip que ha sido almacenado previamente.
Apply Model: Este operador se utiliza para generar el modelo de la evaluación del algoritmo JRip.
Performance: Este proceso genera la matriz de confusion de la evaluación del modelo, presentando los resultados a detalle como la clasificación o coincidencia de la predicción para cada perfil profesional.

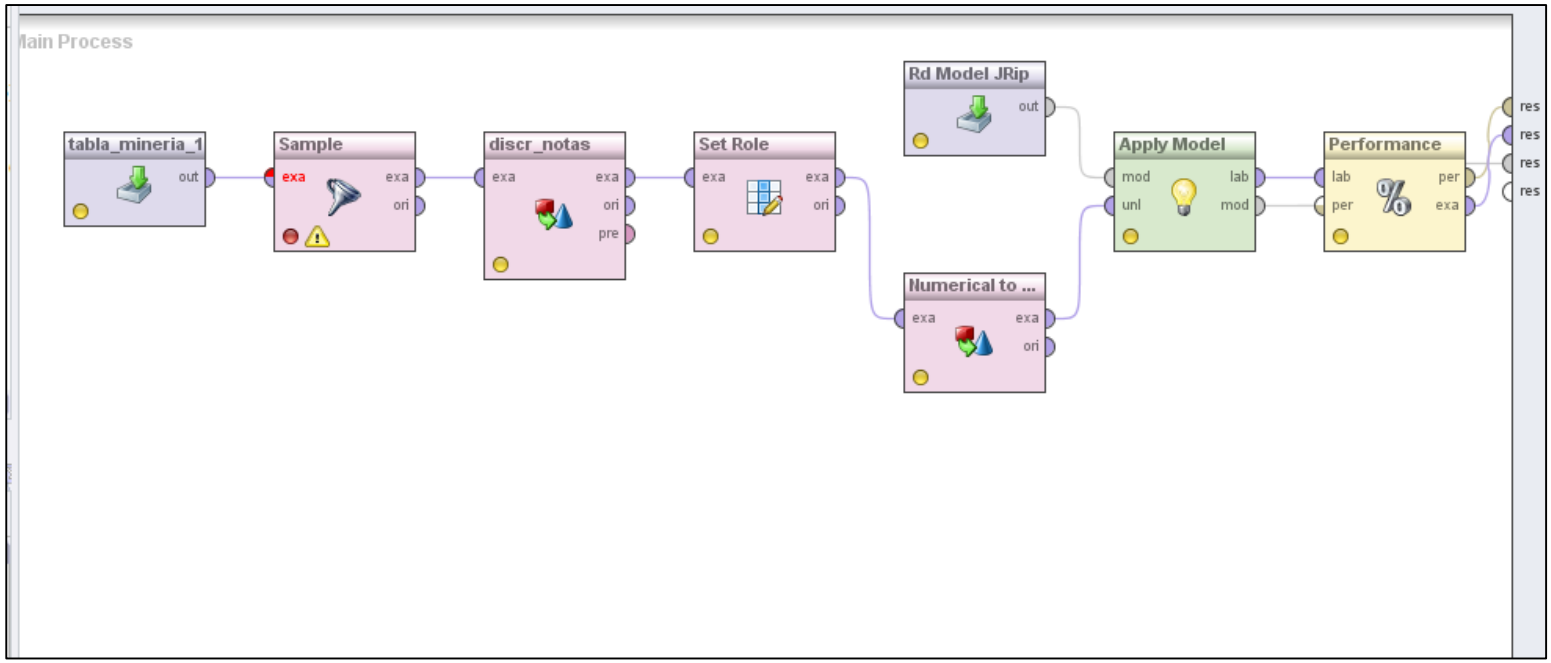


Figura 34. Conjunto de operadores que forman el proceso de evaluación del modelo del algoritmo JRip.

- ✓ **Procesos de la aplicación de los modelos de minería de datos generados en un contexto real.**

En este apartado se realizó la descripción de los procesos que contienen los operadores de la aplicación de los modelos obtenidos de los mejores algoritmos resultantes CHAID Y JRip, con datos de los egresados del año 2014 de la carrera ingeniería en sistemas:

- **Procesos de la evaluación de algoritmos CHAID, últimos egresados CIS.**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 35).

tabla_egr_2014: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería de los datos de los últimos egresados de la carrera.
Select Attributes: este operador sirve para seleccionar los atributos de la estructura con los que vamos a trabajar para aplicar el modelo, en este caso se ha dejado a parte el id, y se han seleccionado las variables unidad.
discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente.
Set Role: Este operador se ha utilizado para asignar los roles necesarios, "label" a la variable dependiente perfil profesional y el atributo "cedula" como ID.
Numerical to Polynominal: Este operador convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse a la aplicación del modelo realizado.
Rd Model CHAID: este operador carga el modelo generado por el algoritmo CHAID que ha sido almacenado previamente.
Apply Model: Este operador se utiliza para generar el modelo de la evaluación del algoritmo CHAID con los nuevos datos.

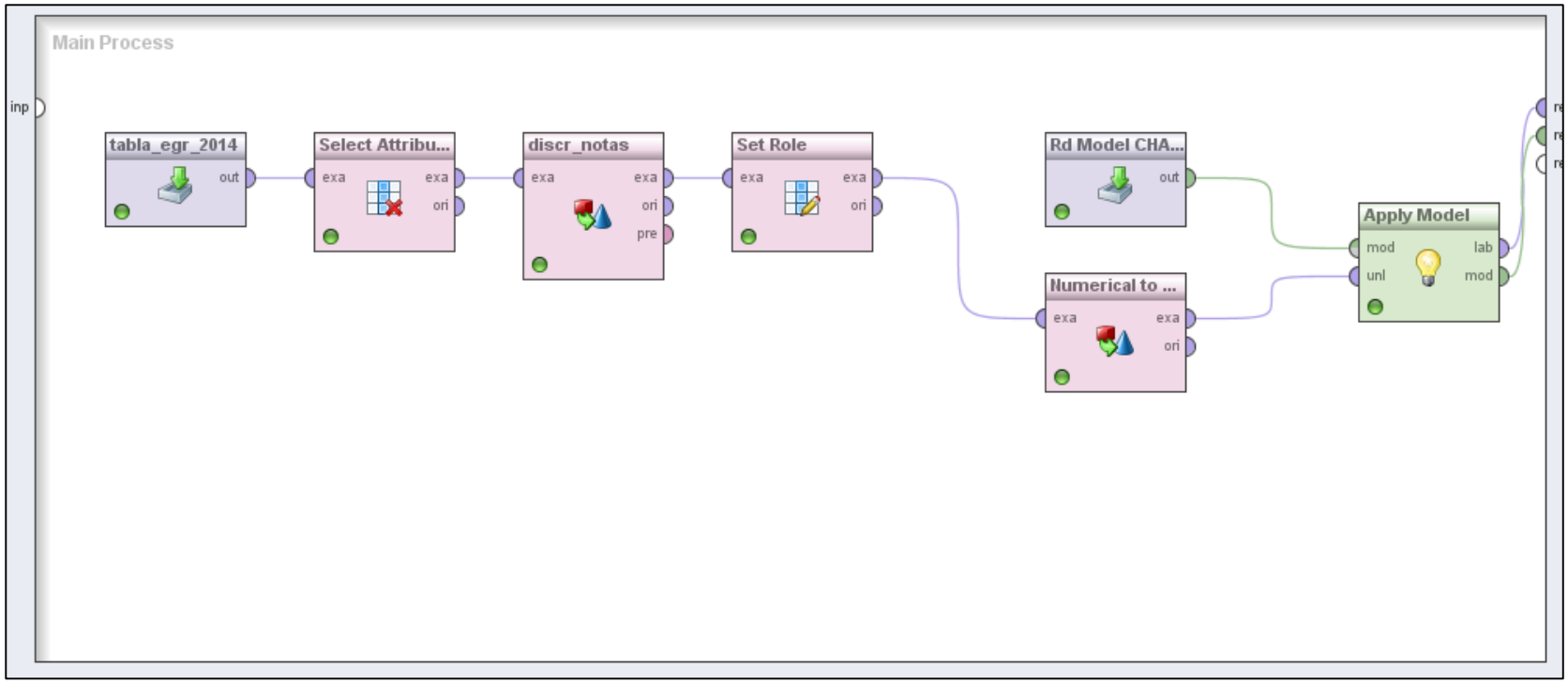


Figura 35. Procesos de la aplicación del modelo generado por el algoritmo CHAID en un contexto real.

- **Procesos de la evaluación del algoritmo JRip, últimos egresados 2014.**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 36).

tabla_egr_2014: Corresponde al operador de conexión a la base de datos que contiene la estructura de minería de los datos de los últimos egresados de la carrera.
Select Attributes: este operador sirve para seleccionar los atributos de la estructura con los que vamos a trabajar para aplicar el modelo, en este caso se ha dejado a parte el id, y se han seleccionado las variables unidad.
discr_notas: Este operador realiza el proceso de discretización de las notas de cada unidad en tres rangos: regular, bueno y excelente.
Set Role: Este operador se ha utilizado para asignar los roles necesarios, "label" a la variable dependiente perfil profesional y el atributo "cedula" como ID.
Numerical to Polynominal: Este operador convierte los tipos de datos numéricos a polinomiales para mejorar la calidad de los datos y acoplarse a la aplicación del modelo realizado.
Rd Model JRip: este operador carga el modelo generado por el algoritmo JRip que ha sido almacenado previamente.
Apply Model: Este operador se utiliza para generar el modelo de la evaluación del algoritmo JRip con los nuevos datos.

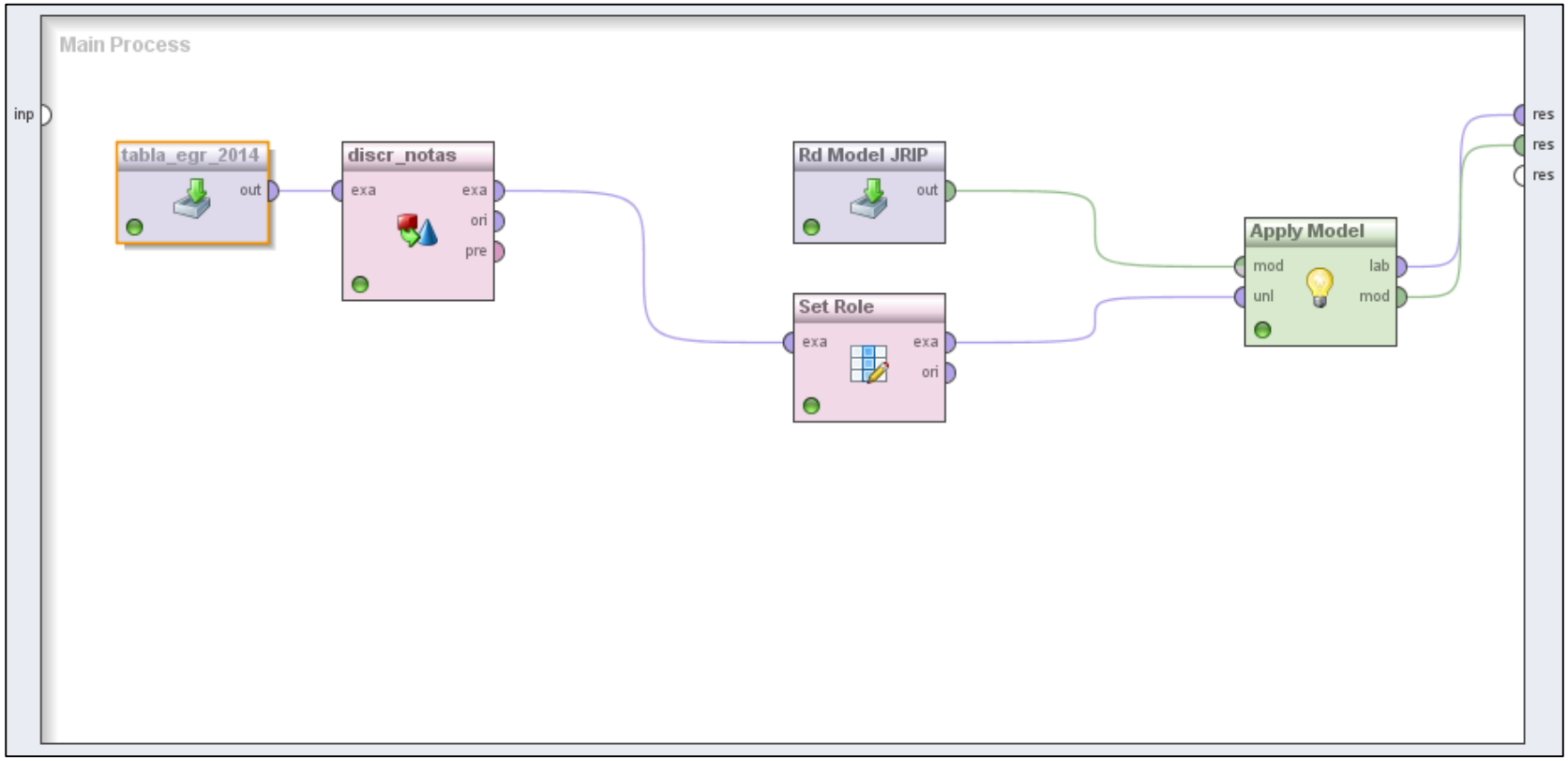


Figura 36. Procesos de la aplicación del modelo generado por el algoritmo JRip en un contexto real.

- **Algoritmo NNge:**

Como podemos observar el algoritmo ha sido probado en la herramienta Rapidminer, sin embargo al ejecutar el proceso no se crea la matriz de confusión, por lo tanto se ha hecho uso de una aplicación en base de la librería de weka.jar, bajo la plataforma java mencionada en el trabajo de titulación. Las pruebas con las estructuras de minería de datos en base al formato .arff; las cuales se especifican a continuación:

- **Algoritmo NNge para datos no agrupados en herramienta en base a java.**

Se carga directamente la estructura en el formato .aff y se ejecuta el algoritmo NNge para el conjunto de entrenamiento (ver figura 37).

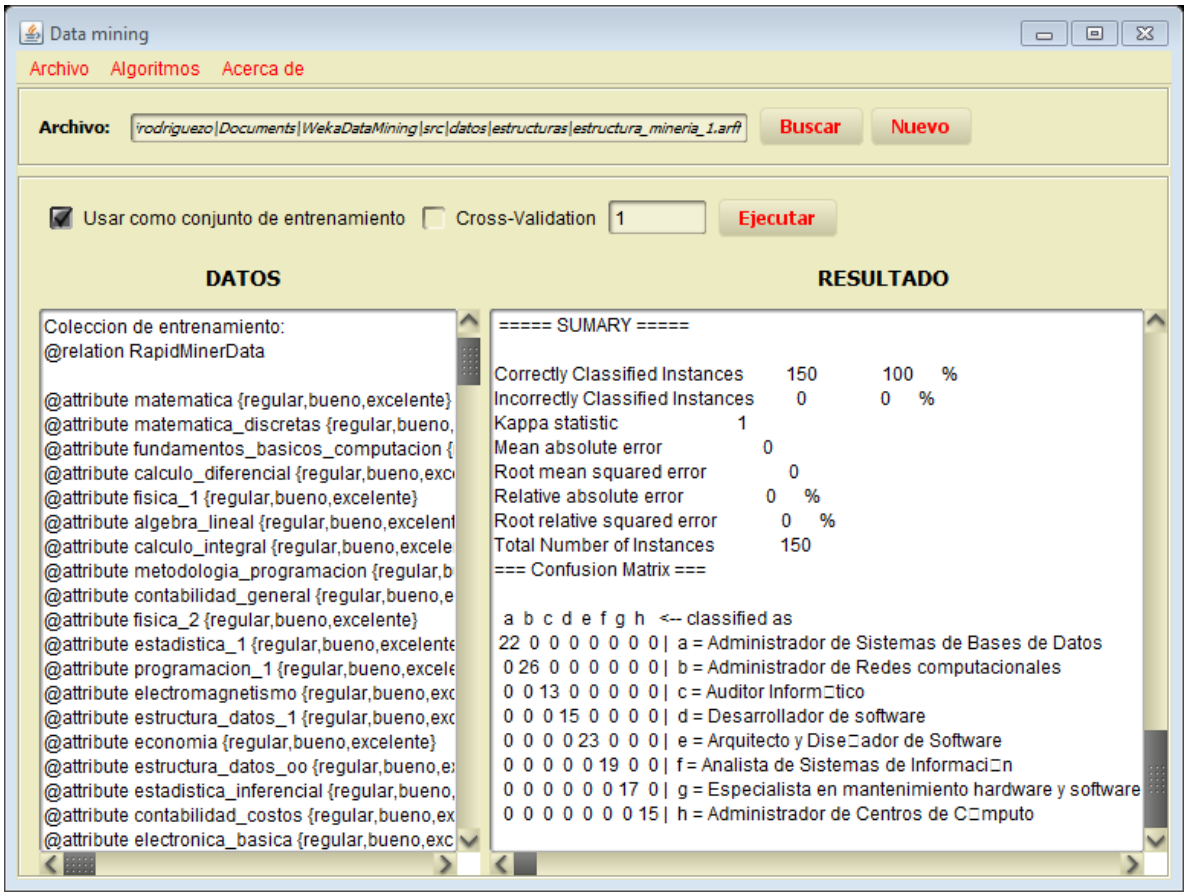


Figura 37. Prueba con el conjunto de entrenamiento para datos no agrupados.

La ejecución del algoritmo genera reglas que contienen valores nulos o perdidos representados por le signo '?' (ver figura 38).

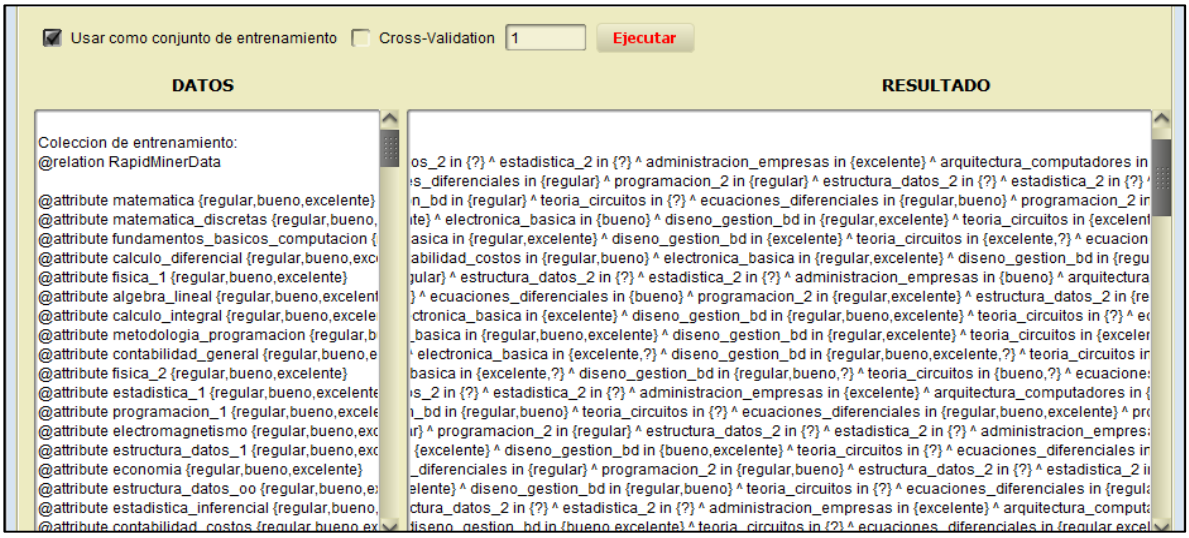


Figura 38. Reglas generadas de la ejecución con datos no agrupados.

Con los mismos datos se realiza la evaluación con el método de validación cruzada tomando un valor de 5 (ver figura 39).

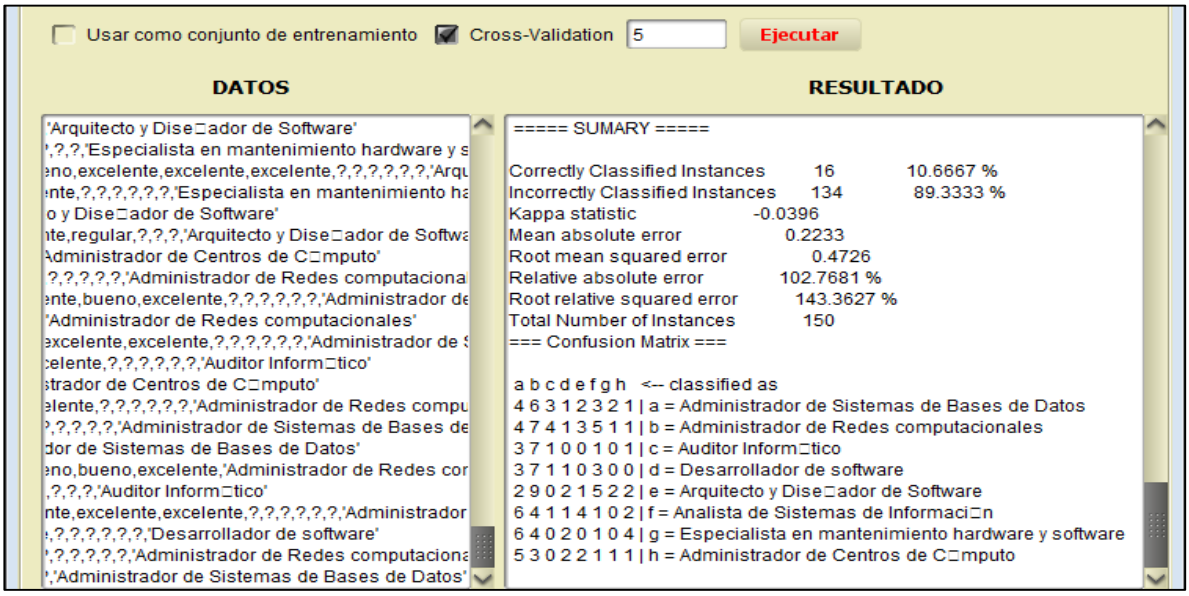


Figura 39. Resultados de validación cruzada con NNge.

- **Algoritmo NNge para datos no agrupados en herramienta en base a java.**

Se carga directamente la estructura en el formato .aff y se ejecuta el algoritmo NNge para el conjunto de entrenamiento (ver figura 40).

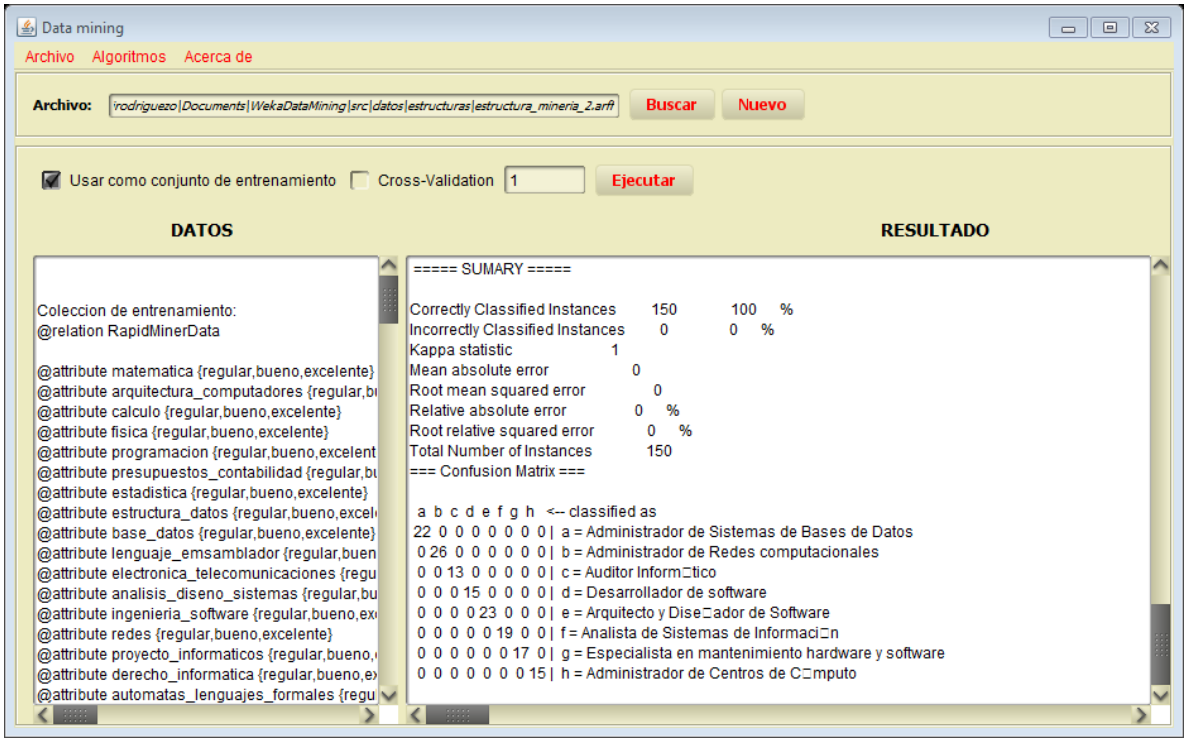


Figura 40. Prueba con el conjunto de entrenamiento para datos no agrupados.

La ejecución del algoritmo genera reglas que contienen valores nulos o perdidos representados por le signo '?' (ver figura 41).

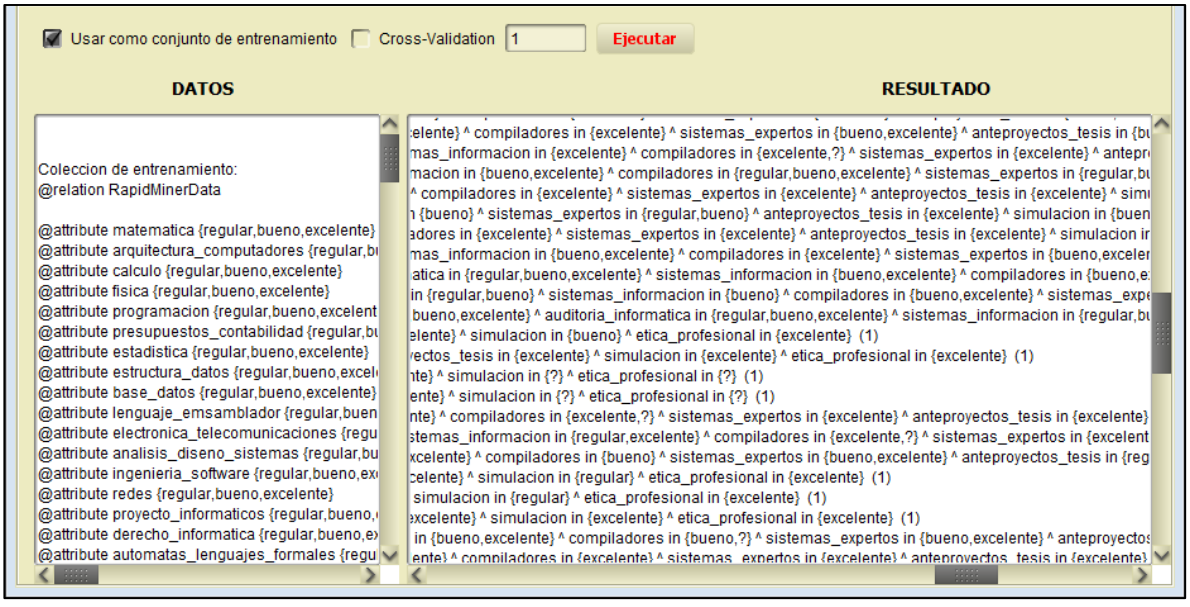


Figura 41. Reglas generadas de la ejecución con datos no agrupados.

Con los mismos datos se realiza la evaluación con el método de validación cruzada tomando un valor de 5 (ver figura 42).

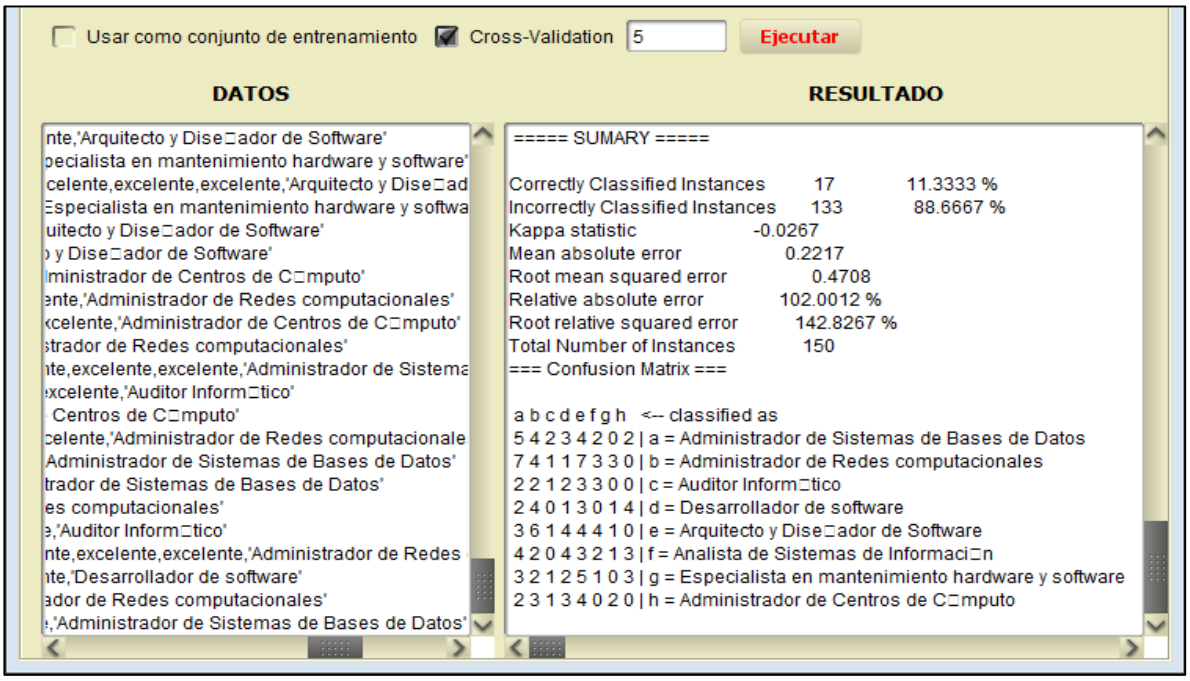


Figura 42. Resultados de validación cruzada con NNge.

Anexo 11: Artículo Científico

Determination of professional profiles using data mining techniques.

María J. Rodríguez*, coauthors: Henry P. Paz; René E. Guamán
Universidad Nacional de Loja, Carrera de Ingeniería en Sistemas,
Loja, Ecuador, e-mail:{mjrodriguezo, hpaz, rguaman}@unl.edu.ec

***Abstract*— This paper presents the determination of professional profiles using data mining techniques focused on the education sector. This has made the study of success stories, gathering tools and techniques of data mining, selected based on reliable literature sources and continuous research, using the CRISP-DM methodology in the development of mining process obtaining data models that have been evaluated and applied in a real context in order to select the most optimal.**

***Keywords*— Data mining, professional profile, models, confusion matrix, cross-validation, error measures, data sources, outliers.**

I. INTRODUCCION

La Universidad Nacional de Loja se ha caracterizado por impulsar la investigación, hacia la búsqueda de alternativas de solución a las problemáticas más preocupantes de hoy en día y de esta manera aportar al progreso académico-profesional. En el sector educativo, se ha podido evidenciar la realidad que enfrentan los egresados y profesionales, que año a año egresan y se titulan con el deseo de desarrollarse de forma plena en el ejercicio de su profesión.

Se puede evidenciar que las aptitudes adquiridas por un estudiante a lo largo de su formación académica en el aspecto cualitativo como: habilidades, capacidades, intereses y cuantitativamente el conocimiento reflejado en su record académico son algunos de los factores que determinan su perfil profesional. En vista del panorama presentado, la minería de datos surge como una alternativa de solución respaldándonos en el éxito de su aplicación en el ámbito educativo [1-3]. Es por ello que mediante la aplicación de técnicas adecuadas se buscará determinar el perfil profesional de cada estudiante, el cual servirá como pauta en la toma de decisiones a nivel académico y profesional.

Para el desarrollo del presente trabajo se ha realizado el estudio de casos de éxito [4-7], recopilación de técnicas y herramientas de minería de datos [8-10], con la finalidad de crear el escenario adecuado para solucionar el problema planteado y lograr la meta de minería. Las técnicas de minería de datos seleccionadas en base al análisis realizado son técnicas de clasificación en base a los árboles de decisión y reglas de inducción. Para el proceso de minería se ha construido dos estructuras de minería de datos, para datos agrupados y no agrupados con el fin de realizar un sin número de pruebas y encontrar los mejores resultados.

La organización del trabajo es la siguiente: en la Sección II. METODOLOGÍA, se explica la metodología utilizada para el desarrollo con todas sus fases. La Sección III. DESARROLLO, muestra un caso de estudio realizado con la herramienta RapidMiner y aplicación de técnicas de minería de datos, detallando el proceso en base a la metodología, hasta obtener los resultados. La Sección IV CONCLUSIONES. Finalmente se pueden encontrar las REFERENCIAS BIBLIOGRÁFICAS.

II. METODOLOGÍA

La metodología utilizada como modelo de referencia se denomina CRISP-DM, la cual está propiamente enfocada a proyectos de minería de datos. Está formada por varias fases, que han permitido de manera organizada detallar todo el proceso de minería de datos y cumplir con los objetivos establecidos (ver figura 1).

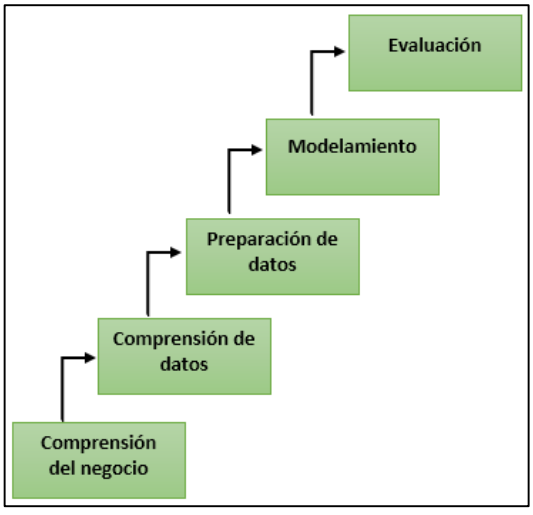


Figura 1. Ciclo de vida de un proyecto con CRISP-DM [13].

A continuación se ha descrito cada una de las fases, con la finalidad de comprender la estructura de esta metodología [13]:

- Fase uno: Comprensión del negocio.- En esta tarea se describe los antecedentes o contexto inicial, los objetivos del negocio y los criterios de éxito.
- Fase dos: Comprensión de los datos.- Esta es la fase de la metodología donde se pretende mantener la información únicamente necesaria para realizar la minería de datos y

familiarizarse con la información.

- Fase tres: Preparación de los datos.- Esta fase permite construir el conjunto de datos final, realizando tareas de selección de datos, selección de tablas, registros, y atributos a los, transformación de datos, cambios de formato, limpieza de datos, generación de variables adicionales, etc.
- Fase cuatro: Modelamiento.- En esta fase se describen las diferentes técnicas de modelado elegidas y se realiza su aplicación obteniendo los modelos y generando los resultados para su posterior evaluación.
- Fase cinco: Evaluación.- Corresponde a la fase del análisis de los resultados obtenidos en la minería, y validación de los resultados en un contexto real.

III. DESARROLLO

A. Fase Uno: Comprensión del Negocio

La minería de datos es el proceso de exploración, análisis, extracción y refinamiento de grandes volúmenes de información de manera automatizada, con el fin de descubrir conocimiento, es decir información que ayude a la toma de decisiones [11, 12].

La meta depende del proyecto que se esté realizando, por ello en el presente proyecto al aplicar el proceso de minería de datos se busca que a partir de un conjunto de datos se descubran uno o varios modelos que determinen los perfiles profesionales mediante la aplicación de técnicas de minería de datos.

Para la elaboración del presente trabajo se ha especificado los recursos necesarios en cuanto al talento humano, hardware, software, y fuentes de datos necesarios para el desarrollo y culminación exitosa del mismo.

- Objetivos del negocio

  - Identificar los perfiles profesionales enfocados en la carrera de ingeniería en sistemas, a través de la formación de los estudiantes.
  - Identificar los factores que determinan el perfil profesional de los estudiantes.
  - Conocer los perfiles profesionales a los cuales se orienten un grupo de estudiantes.

B. Fase Dos: Comprensión de los datos.

Se ha realizado un análisis de las variables más influyentes en el problema con la finalidad de determinar las fuentes de datos a utilizar; dichas variables se las ha enfocado de dos formas: cualitativas y cuantitativas. Las cualitativas corresponden a los obtenidos de un test aplicado a la población objeto de estudio y las cuantitativas que engloban los records académicos de estudiantes egresados y graduados, estas fuentes de datos utilizadas son detalladas a continuación:

1. Datos del Sistema de Gestión Académica

Datos de los egresados de la carrera de ingeniería en sistemas respecto a las categorías académica y personal provenientes del Sistema de Gestión Académica de la institución creado en el 2008. Estos datos se obtuvieron a través del Web Service para su posterior explotación.

2. Datos Históricos de los records académicos

Registros de los records académicos de los estudiantes egresados de la carrera de ingeniería en sistemas, que se encuentran en los libros físicos que están en poder de la secretaría del Área de la Energía, las Industrias y los Recursos Naturales no Renovables de la institución. Estos datos se han recopilado desde el año 2003, con el fin de completar la información académica de los egresados y graduados respecto de las notas de ciertos módulos que no constan en el Sistema de Gestión Académica.

3. Test de habilidades, capacidades e intereses

- El test ha sido desarrollado con el uso de la herramienta django, en base a los intereses, capacidades y habilidades de 8 perfiles planteados.

- Explorar los datos

Como se ha mencionado en el transcurso del desarrollo del presente trabajo, se ha realizado un test enfocado a los egresados y graduados con objeto de recabar una nueva variable de suma importancia, obteniendo un 80% de acogida por parte de la población llegando a la conclusión que la difusión ha tenido éxito (ver figura 2).

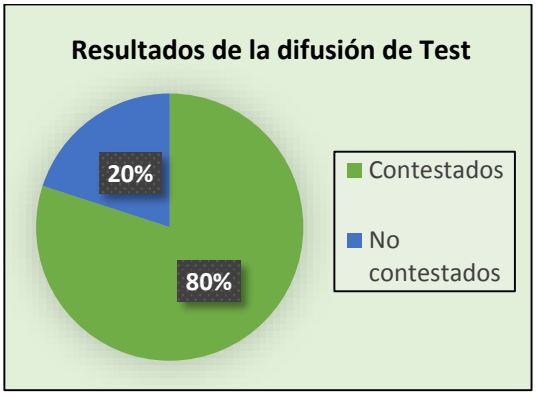


Figura 2. Resultados Difusión del Test Perfil Profesional.

El objetivo del test desarrollado es obtener el perfil profesional de los egresados y graduados, estos perfiles se los ha seleccionado realizando una consulta bibliográfica de las respectivas características, habilidades, capacidades de cada perfil que puede tener un estudiante al egresar de la carrera de ingeniería en sistemas, en base al documento del rediseño de la esta carrera de la Universidad Nacional de Loja [14] y en la documentación de algunas universidades del país y del mundo [15-22].

Finalmente para realizar un filtro de estos perfiles se ha analizado las unidades de la malla curricular base y sus cambios a través de los diferentes periodos académicos. Por lo tanto los perfiles profesionales escogidos son 8 nombrados a continuación:

TABLA I
PERFILES PROFESIONALES

Sigla	Perfil Profesional
AS	Analista de Sistemas de Información
ADS	Arquitecto y Diseñador de Software.
DS	Desarrollador de software
DBA	Administrador de Sistemas de Bases de Datos
AI	Auditor Informático
ACC	Administrador de Centros de computo
AR	Administrador de Redes computacionales.
MHS	Especialista en mantenimiento hardware y software.

En la figura 3 se observa de manera gráfica los resultados del test aplicado, donde el perfil profesional predominante es el perfil de 'Administrador de Redes computacionales' con el 18% lo cual no se diferencia demasiado del resto de perfiles.

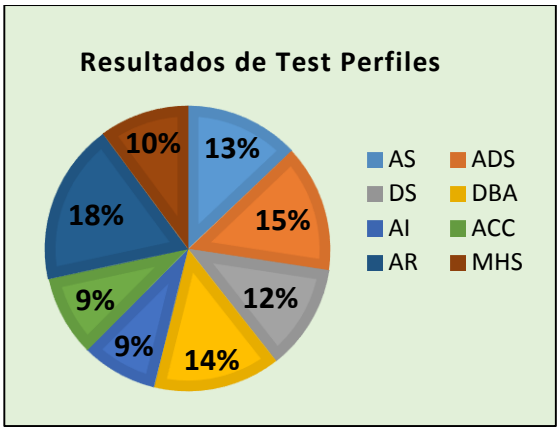


Figura 3. Resultados del Test Perfil Profesional.

El perfil profesional posteriormente será tomado como nuestra variable dependiente dentro del proceso de minería de datos para realizar la predicción.

C. Fase tres: Preparación de los datos. (Selección, limpieza e integración de los datos).

Como se ha descrito anteriormente el estudio está enfocado en los egresados y graduados de la carrera de ingeniería en sistemas (2003-2013). De los cuales finalmente se cuenta con el 80% que respondieron al test aplicado para determinar su perfil profesional. De este porcentaje final, se tomará un 72% con el objeto de realizar el proceso de minería y determinar las reglas que expresen en relación con las unidades como se llega a un determinado perfil. El 28% restante se tomará para realizar la validación de las reglas obtenidas del proceso.

- Construcción de Datos

En base a la meta de minería de datos y para realizar varias pruebas en búsqueda de los mejores resultados se han diseñado dos estructuras:

1. Estructura_uno de datos no agrupados: Esta estructura está conformada por 67 variables; 66 variables correspondientes al conjunto total de unidades que existen entre todos los datos; una gran cantidad de los registros contienen los atributos de las unidades con valores nulos debido al cambio en las mallas curriculares antes mencionado y finalmente la estructura contiene la variable dependiente perfil_profesional obtenida del test aplicado clave para el proceso de predicción dentro de la minería de datos (ver Tabla II).

TABLA II
TABLA DE VARIABLES UTILIZADAS Y FUENTE DE PROVIDENCIA DE ESTRUCTURA UNO

Fuente	Variable
SGA y Registros de Libros Físicos	matemática, matematicas_discretas, fundamentos_basicos_computacion, calculo_diferencial, fisica_I, algebra_lineal, calculo_integral, metodologia_programacion, contabilidad_general, fisica_II, estadistica_I, programacion_I, electromagnetismo, estructura_datos_I, economía, estructura_datos_oo, estadistica_inferencial, contabilidad_costos, electronica_basica, diseno_gestion_bd, teoria_circuitos, ecuaciones_diferenciales, programacion_II, estructura_datos_II, estadistica_II, administracion_empresas, arquitectura_computadores, lenguaje_ensamblador, diseno_digital, analisis_diseno_sistemas_I, ingenieria_software_I, redes_I, proyectos_informaticos_I, teoria_telecomunicaciones, derecho_informatico, sistemas_informacion, analisis_diseno_sistemas_II, ingenieria_software_II, sistemas_operativos, redes_II, investigacion_operaciones, teoria_automatas, inteligencia_artificial, proyectos_informaticos_II, analisis_numerico, administracion_cc, auditoria_informatica, gestion_redes, sistemas_informacion_I, microprocesadores, lenguajes_formales, modelamiento_matematico, compiladores, sistemas_informacion_II, sistemas_expertos, mantenimiento_computadores, control_automatizado_asistido_c, anteproyectos_tesis, simulación, etica_profesional, legislacion_laboral, presupuestos_invesiones, diseno_asistido_computadores, aplicaciones_web, administracion_bd_mysql_uml, programacion_net.
Test Perfil Profesional	perfil_profesional

2. Estructura_dos de datos agrupados: La estructura_dos está formada por 28 variables; 18 variables correspondientes a los grupos generados en base de 47 unidades de las 66 totales; con la finalidad de eliminar la gran cantidad de valores nulos existentes y por su relación entre sí, 9 atributos de unidades que no han sido alteradas manteniéndose de la estructura_uno y la variable dependiente denominada perfil_profesional obtenida por cada estudiante de manera personal en base al test aplicado (ver tabla III).

TABLA III
TABLA DE VARIABLES UTILIZADAS Y FUENTE DE PROVIDENCIA DE ESTRUCTURA DOS

Fuente	Variable
SGA y Registros de Libros Físicos	matemática, física, calculo, programación, estructura_datos, estadística, presupuestos_contabilidad, redes, proyectos_informaticos, sistemas_informacion, analisis_diseno_sistemas, ingenieria_software, arquitectura_computadores, electronica_telecomunicaciones, base_datos, lenguaje_ensamblador, derecho, teoria_automatas, inteligencia_artificial, administracion_centros_computo, auditoria_informatica, lenguajes_formales, compiladores, sistemas_expertos, anteproyectos_tesis, simulación, etica_profesional
Test Perfil Profesional	perfil_profesional

Cabe mencionar que las notas de las unidades tienen un valor comprendido entre 0 y 10, en base a ello para el proceso de minería se ha visto importante realizar la discretización de estos valores en las dos estructuras. (ver tabla IV):

TABLA IV
DISCRETIZACIÓN DE LAS NOTAS DE CADA UNIDAD

Nro.	Nomenclatura	Rango
1	Regular	Menor a 7.5
2	Bueno	Entre 7.5 a 8.5
3	Excelente	Mayor a 8.5

D. Fase Cuatro: Selección de técnicas y generación de pruebas.

En esta fase se ha utilizado la herramienta RapidMiner luego de una previa selección para realizar el proceso de minería y la aplicación de dos grupos de técnicas: Clasificación y las técnicas de reglas basadas en inducción, seleccionadas en base a los casos de estudio analizados. Dentro de cada grupo de técnicas se ha utilizado diferentes algoritmos descritos a continuación:

- Técnicas de Clasificación

En este tipo de técnicas los algoritmos son robustos a los datos con ruido, la función aprendida es representada como un árbol, permitiendo obtener a su vez de forma visual las reglas de clasificación bajo las cuales operan los datos del experimento [23, 24]. La aplicación de estas técnicas han sido basadas en los algoritmos de árboles de decisión, específicamente 2 algoritmos: ID3 y CHAID.

- Técnicas Basadas en Reglas de Inducción

Este tipos de algoritmos arrojan como resultados un sin número de reglas respecto al análisis de los datos, el proceso interno que realiza es la búsqueda de patrones, relaciones y características similares entre los datos. Estas reglas tienen la ventaja que son fáciles de entender [25, 26]. Para la aplicación de estas técnicas se han escogido 6 algoritmos: JRip, Part, Ridor, Decisión Table, DTNB y NNge.

✓ **Diseño de pruebas**

En las pruebas con el conjunto de entrenamiento se ha tomado un 72% de los datos mientras que el 28% restante será utilizado para la evaluación de los modelos. A su vez se realizará la evaluación de los modelos con el método de validación cruzada. Los resultados obtenidos al aplicar los distintos algoritmos han sido detallados en las siguientes tablas; tanto para la estructura de datos no agrupados (ver tabla V), como para la estructura de datos agrupados (ver tabla VI).

TABLA V
COMPARACIÓN DEL RENDIMIENTO DE ALGORITMOS CON DATOS NO AGRUPADOS

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)
CHAID	Conjunto de Entrenamiento	98.00%	2.00%
	Validación Cruzada	10.67%	89.33%
DECISION TABLE	Conjunto de Entrenamiento	33.33%	66.67%
	Validación Cruzada	10.00%	90.00%
DTNB	Conjunto de Entrenamiento	35.33%	64.67%
	Validación Cruzada	10.00%	90.00%
ID3	Conjunto de Entrenamiento	28.67%	71.33%
	Validación Cruzada	14.67%	85.33%
JRIP	Conjunto de Entrenamiento	94.00%	6.00%
	Validación Cruzada	12.00%	88.00%
PART	Conjunto de Entrenamiento	81.33%	18.67%
	Validación Cruzada	15.33%	84.67%
RIDOR	Conjunto de Entrenamiento	51.33%	48.67%
	Validación Cruzada	10.67%	89.33%
NNGE	Conjunto de Entrenamiento	100%	0.0%
	Validación Cruzada	10.67%	89.33%

La tabla V nos muestra que en pruebas de entrenamiento con datos no agrupados los algoritmos que presentan mejor rendimiento son: CHAID, PART, JRip y NNge; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento.

TABLA VI
COMPARACIÓN DEL RENDIMIENTO DE ALGORITMOS CON DATOS AGRUPADOS

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)
CHAID	Conjunto de Entrenamiento	96.67%	3.33%
	Validación Cruzada	14.00%	86.00%
DECISION TABLE	Conjunto de Entrenamiento	27.33%	72.67%
	Validación Cruzada	8.67%	91.33%
DTNB	Conjunto de Entrenamiento	24.67%	75.33%
	Validación Cruzada	9.33%	90.67%
ID3	Conjunto de Entrenamiento	87.33%	12.67%
	Validación Cruzada	14.00%	86.00%
JRIP	Conjunto de Entrenamiento	94.67%	5.33%
	Validación Cruzada	14.00%	86.00%
PART	Conjunto de Entrenamiento	82.67%	17.33%
	Validación Cruzada	8.00%	92.00%
RIDOR	Conjunto de Entrenamiento	42.67%	57.33%
	Validación Cruzada	12.67%	87.33%
NNGE	Conjunto de Entrenamiento	100.0%	0.0%
	Validación Cruzada	11.33%	88.67%

La tabla VI nos muestra que en pruebas de entrenamiento con datos agrupados los algoritmos que presentan mejor rendimiento son: CHAID, ID3, PART, JRip y NNge; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento.

- Comparación general de la evaluación de modelos con datos agrupados y no agrupados:

En las pruebas de Validación Cruzada todos los algoritmos muestran porcentajes bajos en la clasificación debido a la poca cantidad de datos [28] y a la presencia de outliers o valores atípicos que le restan calidad a los datos [29].

Podemos observar que al agrupar las unidades los porcentajes de clasificación varían muy poco entre las dos estructuras, el único algoritmo que logro tener un considerable aumento en pruebas de entrenamiento es ID3, por lo tanto se ha notado que realizar la agrupación en los datos no es tan esencial, por ello se ha seleccionado los mejores algoritmos en base a los resultados con datos no agrupados (ver figura 4), descartando los modelos de los datos agrupados.

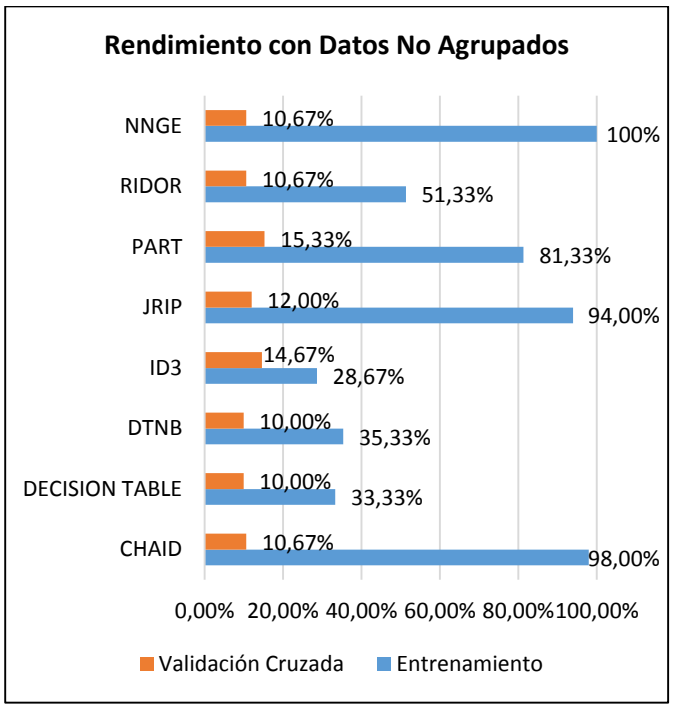


Figura 4. Comparación del rendimiento en pruebas de entrenamiento y validación cruzada con datos no agrupados.

En la figura 4, se observa el rendimiento de los algoritmos, donde aparentemente NNge es el más óptimo con un porcentaje del 100%, el cual ha sido aplicado en una herramienta diferente a RapidMiner, debido a que esta no soporta el algoritmo. Esta nueva herramienta hace uso de la librería de weka.jar, y ha sido desarrollada en base al lenguaje de programación java [30]. Sin embargo a pesar del perfecto porcentaje de clasificación siendo un algoritmo robusto no se lo tomó en cuenta debido que no maneja correctamente la presencia de valores nulos, reflejado en que las reglas generadas por el algoritmo, debido a que toma en cuenta los valores nulos o perdidos [27], por lo tanto son difíciles de interpretar y utilizar.

Continuando con el análisis CHAID presenta el 98% de instancias bien clasificadas seguido de JRip con el 94%, siendo los mejores en esta característica así como en el análisis del rendimiento, lógica de reglas y medidas de error por lo que han sido seleccionados como los mejores hasta el momento.

E. Fase Cinco: Evaluación

En este apartado se realizó un análisis del rendimiento con los mejores algoritmos en este caso CHAID y JRip, los datos utilizados para realizar estas pruebas corresponden al 28% restante de los recopilados inicialmente. Donde la evaluación realizada del rendimiento de los modelos, han presentado algunas variaciones del porcentaje de clasificación para cada perfil profesional, las cuales las observamos en la tabla VII.

TABLA VII
RESULTADOS DE LA EVALUACIÓN DE LOS MODELOS GENERADOS CON CHAID Y JRip

PERFIL PROFESIONAL	Instancias bien clasificadas (%)	
	JRip	CHAID
Analista de Sistemas de Información	100	100
Arquitecto y Diseñador de Software.	100	100
Desarrollador de software	100	100
Administrador de Sistemas de Bases de Datos	100	86.67
Auditor Informático	100	100
Administrador de Centros de computo	100	100
Administrador de Redes computacionales.	64.29	100
Especialista en mantenimiento hardware y software.	100	100

Los resultados de clasificación mostrados en la tabla VII corresponden a los valores de cada matriz de confusión generada, donde se describe el porcentaje de precisión para las clases definidas como perfiles profesionales, en donde se puede observar que varían muy poco, ya que cada uno rinde el 100% en 7 perfiles, pero JRip presenta el 64.29% para el perfil Administrador de Redes computacionales y CHAID lo supera con el 86.67% en el perfil Administrador de Sistemas de Bases de Datos. A continuación realizaremos una evaluación global del porcentaje de clasificación (ver figura 5).

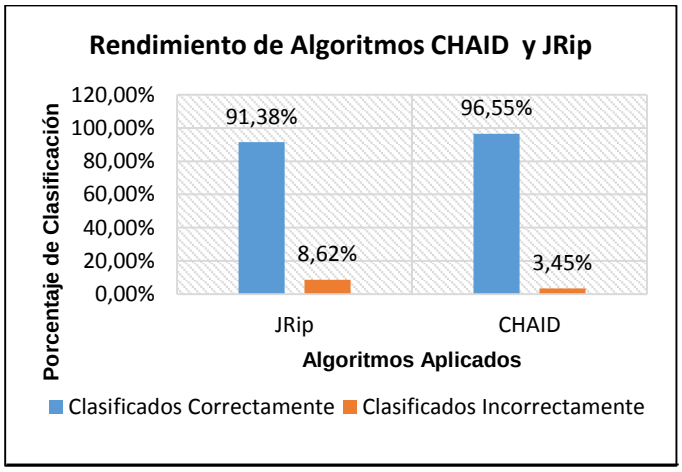


Figura 5. Porcentaje de clasificación de CHAID y JRip.

Como se puede observar en la figura 5 el algoritmo que mejor rendimiento presenta es CHAID logrando clasificar el 96.55% de las instancias mientras que el algoritmo JRip clasificó el 91.38%, siendo una leve diferencia, sin embargo en base a ello se ha seleccionado al algoritmo CHAID como el más óptimo para la predicción.

- Aplicación de los modelos de minería de datos en un contexto real.

Finalmente se ha realizado la validación de los modelos generados al probarlos en un contexto real para seleccionar el más óptimo. Los datos tomados corresponden a los egresados de la carrera de ingeniería en sistemas año 2014. De esta validación se ha obtenido el porcentaje global de determinación de los perfiles profesionales mediante los modelos CHAID y JRip (ver figura 6).

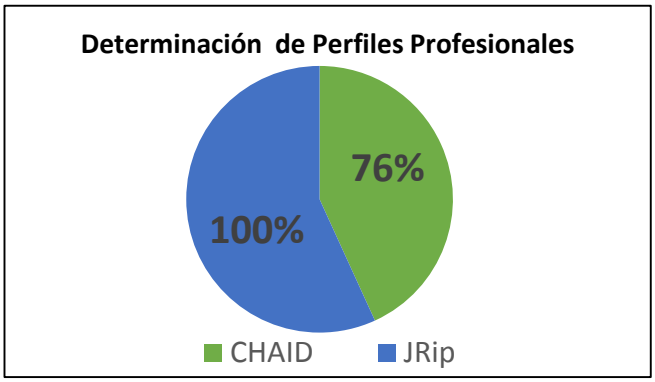


Figura 6. Porcentaje de predicción final de los modelos CHAID y JRip.

Al observar los resultados de la figura 6 se demuestra que ha sido una sido fundamental la aplicación de los modelos en un contexto real, debido a que JRip realiza la determinación de los perfiles profesionales en un 100%, mientras que CHAID lo hace en un 76%, cambiando la perspectiva y validando el modelo generado por el algoritmo JRip perteneciente a las técnicas de reglas de inducción como el óptimo para la determinación de perfiles profesionales en la carrera de ingeniería en sistemas de la Universidad Nacional de Loja. A continuación analizaremos el porcentaje de los perfiles profesionales determinados por el modelo final (ver figura 7).

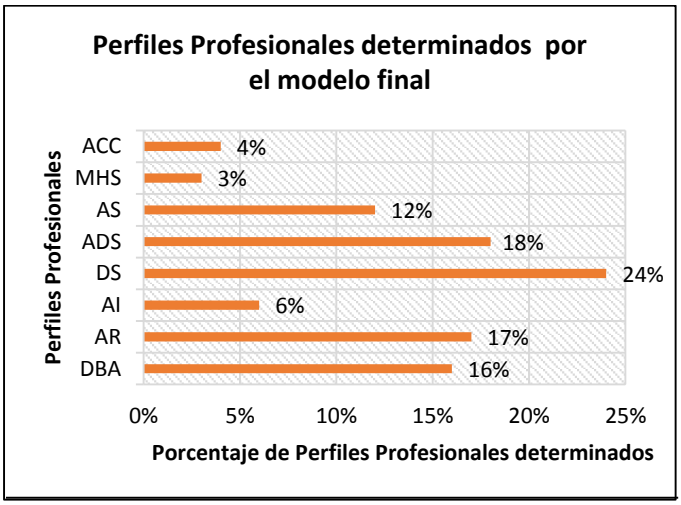


Figura 7. Determinación de los perfiles profesionales mediante el modelo JRip.

En la figura 7 se puede observar que el perfil profesional que más se destaca entre la población es el perfil 'Desarrollador de software', seguido de 'Arquitecto y Diseñador de software', mientras que el perfil con menor porcentaje corresponde a 'Especialista en mantenimiento Hardware y Software', observando que la mayor parte de los

egresados de la carrera de ingeniería en sistemas del año 2014 son desarrolladores.

IV. CONCLUSIONES

La minería de datos hoy en día se ha convertido en una herramienta de vital importancia en el tratamiento, análisis y obtención de resultados del procesamiento de grandes cantidades de datos; para guiar la toma de decisiones tanto en las instituciones educativas como en todo tipo de organizaciones que así lo requieran.

La tarea más costosa a lo largo del proyecto, tanto en tiempo como esfuerzo, fue la recolección y armado de la Base de Datos, puesto que se obtuvieron de una fuente digital así como de una física; lo que conllevó al análisis, clasificación y limpieza de la información para luego agruparlos en una sola Base.

Posterior al desarrollo del presente proyecto se puede concluir la importancia de determinar el perfil profesional de los egresados y graduados con los factores determinantes como: el récord académico que muestra el desempeño del estudiante a lo largo de la carrera y la parte cualitativa de cada individuo, ésta última se determinó mediante la aplicación de una encuesta que permitió posteriormente contrastar sus conocimientos con sus habilidades, intereses y capacidades que los hacen únicos y candidatos potenciales y competentes a diferentes Áreas y temáticas dentro del mundo laboral.

En el trabajo desarrollado se evidenció que el perfil profesional que más predomina en los últimos egresados del año 2014 es desarrollador de software cuyos conocimientos obtenidos en las aulas universitarias serán puestos en práctica en el desarrollo de su vida profesional, ésta información es muy útil para los futuros cambios que se realicen a nivel de la malla curricular y perfil de carrera.

La metodología que se utilizó para la Minería de datos en el presente trabajo fue CRISP-DM, la misma que ayudó a organizar mediante fases, sub fases y tareas que apoyaron a la documentación del proyecto a más de ser una guía para el desarrollo durante todo el proceso, permitiendo su culminación con éxito.

Para el proceso de minería de datos se escogió los algoritmos ID3 y CHAID que pertenecen a las técnicas de clasificación basadas en árboles de decisión y los algoritmos JRip, PART, Ridor, Decisión Table, DTNB y NNge, pertenecientes al grupo de técnicas de reglas de inducción. Ya en el desarrollo y generación de modelos, los mejores algoritmos fueron CHAID y JRip los cuáles se hicieron con el 72% de los datos y con el 28% restante se hizo la evaluación de los mismos para verificar su validez, donde CHAID resultó el más óptimo al clasificar el 96.55% de las instancias; mientras que JRip clasificó el 91,38%. Posterior a ello se realizó la aplicación de éstos algoritmos en un contexto real para validar y realizar sí la elección final, en donde JRip tuvo el mejor rendimiento en la predicción con el 100%; mientras que CHAID realizó la predicción del 76%, llegando a la conclusión que JRip es el modelo que se debe aplicar para la obtención de los perfiles profesionales.

REFERENCIAS BIBLIOGRÁFICAS

[1] COBO O. Angel, ROCHA B. Rocío. *Selección de atributos predictivos del rendimiento académico de estudiantes en un modelo de B-Learning*. [En línea]: http://edutec.rediris.es/Revelec2/Revelec37/pdf/Edutec-e_n37_Cobo_Rocha_Alvarez.pdf. [Acceso: 21-Junio-2013].

[2] ECKERT Karina, SUÉNAGA Roberto. *Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD.* [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/27103/Documento_completo.pdf?sequence=1. [Acceso: 21-Junio-2013].

[3] ÁLVARO J. Galindo, ÁLVAREZ G. Hugo. *Minería de Datos en la Educación*. [En línea]: http://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf. [Acceso: 21-Junio-2013].

[4] ROMERO M. Cristóbal, VENTURA S. Sebastián, HERVÁS M. Cesar. Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. [En línea]: http://www.lsi.us.es/redmidas/CEDI/papers/189.pdf. [Acceso: 6-Noviembre-2013].

[5] GARCÍA S. Enrique, ROMERO M. Cristóbal, VENTURA S. Sebastián, CASTRO Carlos. Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning. [En línea]: http://rita.det.uvigo.es/200805/uploads/IEEE-RITA.2008.V3.N1.A3.pdf. [Acceso: 8-Noviembre-2013].

[6] R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica. Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. [En línea]: http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf. [Acceso: 10-Noviembre-2013].

[7] MÁRQUEZ V. Carlos, ROMERO M. Cristóbal, VENTURA S. Sebastián. Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. [En línea]: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf. [Acceso: 12 Noviembre-2013].

[8] ASTORGA Nathalia, SALINAS Maruxa. *Weka para minería de datos*, 2013. [En línea]: http://prezi.com/_gli7zt6vv0t/weka-para-mineria-de-datos-2013/. [Acceso: 26-Enero-2014].

[9] CORSO Cynthia L, GIBELLINI Fabián. Facultad Regional Córdoba/Universidad Tecnológica Nacional. Argentina. Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios. [En línea]: http://41jaiio.sadio.org.ar/sites/default/files/16_JSL_2012.pdf. [Acceso: 26-Enero-2014].

[10] CUBERO Juan C, BERZAL Fernando. Departamento de Ciencias de la computación, Universidad de Granada. *Guión de prácticas de minería de datos, herramientas de minería de datos, KNIME.* [En línea]: http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf. [Acceso: 26-Enero-2014].

[11] HERNÁNDEZ O. José. *Minería de Datos. Otros Aspectos*. [En línea]: http://users.dsic.upv.es/~jorallo/master/dm5.pdf. [Acceso: 5-Enero-2014].

[12] MOLINA L. José M, GARCIA H. Jesús. *Técnicas de Análisis de Datos*. [En línea]: tesis-algoritmo-c45.googlecode.com/files/131469066-apuntesAD.pdf. [Acceso: 06-Febrero-2014].

[13] GALLARDO A. José A. *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* [En línea]: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf. [Acceso: 5-Enero-2014].

[14] Universidad Nacional de Loja. Reglamento de Rediseño de la carrera de Ingeniería en Sistemas. *Perfil de Egreso del Ingeniero en Sistemas.* [En línea]: https://drive.google.com/file/d/0By4B2OXR-vftTHpZV3F0a2NiNWs/edit?usp=sharing. [Acceso: 02-Noviembre-2013].

[15] Página Web de la Universidad Técnica del Norte. Ecuador. Carrera Sistemas. *Campo Ocupacional, CIS.* [En línea]: http://www.utn.edu.ec/fica/carreras/sistemas/?page_id=153. [Acceso: 08-Enero-2014].

[16] Página Web oficial de la Universidad Peruana de Ciencias e Informática. *Ingeniería de sistemas e Informática*. [En línea]: http://www.upci.edu.pe/facultades.php?ac=ci_ing&op=ing. [Acceso: 08-Enero-2014].

[17] Página web oficial de la Pontificia Universidad Católica del Ecuador. *Facultad de Ingeniería. Ingeniero en Sistemas.* [En línea]:

http://www.puce.edu.ec/portal/content/Ingenier%C3%ADa%20en%20Sistemas/292;jsessionid=248B1730885DCEC54DFC19EC224DCF7F.node0?link=oln30.redirect. [Acceso: 09-Enero-2014].
[18] Página web oficial de la Facultad de Sistemas, México. *Ingeniero en Sistemas Computacionales.* [En línea]: http://www.sistemas.uadec.mx/index.php/carreras/isc. [Acceso: 09-Enero-2014].
[19] Página web oficial Universidad del Valle. Cede central Cochabamba - Bolivia. *Ingeniería de Sistemas Informáticos.* [En línea]: http://www.univalle.edu/index.php/facultades/informatica/sistemas. [Acceso: 09-Enero-2014].
[20] Página web oficial de la ESPE. Universidad de las Fuerzas Armadas. *Ingeniería de Sistemas e Informática.* [En línea]: http://www.espe.edu.ec/portal/portal/main.do?sectionCode=107. [Acceso: 09-Enero-2014].
[21] Instituto Tecnológico Superior de Irapuato. *Ingeniería en Sistemas Computacionales.* Carretera Irapuato - Silao Km. 12.5, C.P. 36821 Irapuato, Guanajuato, México. [En línea]: http://www.itesi.edu.mx/Oferta%20Educativa/Nivel%20Superior/Ing Sistemas.html. [Acceso: 09-Enero-2014].
[22] Universidad del Valle. Sede Central Cochabamba, Bolivia. *Ingeniería de Sistemas Informáticos.* [En línea]: http://www.univalle.edu/index.php/facultades/informatica/sistemas. [Acceso: 09-Enero-2014].
[23] ALUJA Tómas. *La minería de datos entre la estadística y la inteligencia artificial.* [En línea]: http://upcommons.upc.edu/revistes/bitstream/2099/4162/4/article.pdf. [Acceso: 20-Abril-2014].
[24] VIZCAINO G. Paula A. *Aplicación de técnicas de inducción de Árboles de Decisión a problemas de clasificación mediante el uso de weka (Waikato Environment For Knowledge Analysis).* [En línea]: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf. [Acceso: 20-Abril-2014].
[25] SERVENTE Magdalena. *Algoritmos TDIDT aplicados a la minería de datos inteligente.* [En línea]: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf. [Acceso: 21-Abril-2014].
[26] HERNÁNDEZ O. José. *Práctica 2 de minería de datos, profundizando en el Clementine.* [En línea]: http://users.dsic.upv.es/~jorallo/cursoDWDM/kdd-lab2.pdf. [Acceso: 21-Abril-2014].
[27] KAISER Jiří. *Dealing with Missing Values in Data*. [En línea]: http://www.si-journal.org/index.php/JSI/article/viewFile/178/134. [Acceso: 12-abril-2014].
[28] ANDREW Y. Ng. *Preventing "Overfitting" of Cross-Validation Data*. [En línea]: http://ai.stanford.edu/~ang/papers/cv-final.pdf. [Acceso: 08-mayo-2014].
[29] DAPOZO Gladys, PORCEL Eduardo, LÓPEZ María V, BOGADO Verónica. *Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios*. En línea]:http://sedici.unlp.edu.ar/bitstream/handle/10915/20453/Documento_completo.pdf?sequence=1. [Acceso: 08-junio-2014].
[30] PAZ A. Henry P. Publicaciones Henry. *Weka with Data Mining done in java*. [En línea]: http://publicacioneshenry.wordpress.com/. [Acceso: 08-Febrero-2014].

Egda. María José Rodríguez He studied at the Engineering in Computer Systems at the National University of Loja - Loja, Ecuador, in 2013 Researcher of techniques and data mining tools to solve problems in education. Passionate about software development with free tools.

Ing. Henry Patricio Paz Arias He received an engineering degree in Systems at the National University of Loja - Ecuador in 2010, with a degree of Master in Computer Science in 2012 at the Universidad Autonoma del Estado de Hidalgo - Mexico specializing in Intelligent Computing. His current research interest is the development and testing of artificial intelligence techniques to solve problems.

Ing. Edwin René Guamán Quinche Graduate Engineer in IT and computer systems at the National University of Loja, Loja - Ecuador. Currently a professor at the Engineering Systems. Researcher of ubiquitous computing and the development of free software applications.

Anexo 12. RESUMEN EJECUTIVO

Introducción:

La minería de datos, surge como una alternativa de solución a un sin número de problemas, es por ello que en base la aplicación de las técnicas adecuadas se pretende determinar el perfil profesional los egresados y graduados de la carrera de ingeniería es sistemas que han ingresados en los años 2008 al 2013, de la Universidad Nacional de Loja.

Las fuentes de datos utilizadas tienen un enfoque cualitativo obtenido de un test aplicado a la población de interés, respecto de las capacidades, intereses y habilidades de inclinación hacia uno de los perfiles planteados. Y el enfoque cuantitativo que engloban los records académicos, los mismos que fueron obtenidos del Sistema de Gestión Académica (SGA) a través de su Web Services, además se realizó la recopilación de datos históricos para completar estas fuentes de datos, obtenidos de los libros físicos que reposan en la Universidad Nacional de Loja, datos y variables que fueron integradas y tomadas para la construcción de dos estructuras, con el fin de realizar un sin número de pruebas y encontrar el modelo predictivo que posteriormente fue validado mediante su aplicación en un contexto real, con el fin de obtener los mejores resultados que den valor y justifiquen la importancia de la realización del presente trabajo de titulación.

Con el proceso de minería de datos se ha obtenido el modelo para la determinación de los perfiles profesionales y se ha realizado un análisis comparativo de los perfiles profesionales obtenidos versus los empleos en los que han incurrido un grupo de la población encuestada, logrando obtener información concluyente y de gran valor que demuestra la importancia de la realización del presente trabajo de titulación.

En base a lo descrito, el objetivo principal del trabajo de titulación se fundamenta en la determinación de perfiles profesionales. Para ello se ha realizado el estudio de casos de éxito [1-3], recopilación de técnicas y herramientas de minería de datos, seleccionadas de manera minuciosa y con base en fuentes bibliográficas confiables y la investigación permanente, con la finalidad de crear el escenario adecuado para solucionar el problema planteado y lograr la meta de minería, haciendo uso de la

metodología CRISP-DM cumpliendo con todas las fases de la minería de manera exitosa.

Problemática:

Actualmente en el sector educativo, podemos evidenciar algunas temáticas de vital importancia como es la realidad que enfrentan los profesionales, que año a año se titulan y salen de las universidades con grandes deseos de encontrar el empleo apropiado en donde puedan desarrollarse de forma plena en el ejercicio de su profesión. El problema principal que conlleva el desconocimiento del perfil profesional en un egresado y graduado es que pierde el horizonte del mundo laboral, estando expuesto a desarrollar actividades opuestas a sus verdaderas aspiraciones, desperdiciando así los conocimientos, capacidades, habilidades, intereses desarrollados a lo largo de su carrera profesional y reflejadas en su perfil profesional.

Para justificar esta perspectiva se ha realizado una encuesta en la herramienta online SurveyMonkey, de la población de egresados y graduados objeto de estudio se aplicó la encuesta al 48%, de los cuales el 50% aseguraron que están trabajando y colocaron el empleo en el que se están desempeñando actualmente, que corresponden al 24% de los egresados y graduados tomados para la determinación de los perfiles profesionales mediantes las técnicas de minería de datos.

Analizando las respuestas y realizando la comparación de los perfiles profesionales obtenidos con el proceso de minería de datos y el empleo actualmente ejercido por la población se ha podido evidenciar que para todos los perfiles profesionales es mayor la cantidad de los empleos que no se ajustan a cada perfil que los que se ajustan a los mismos (ver figura 1).

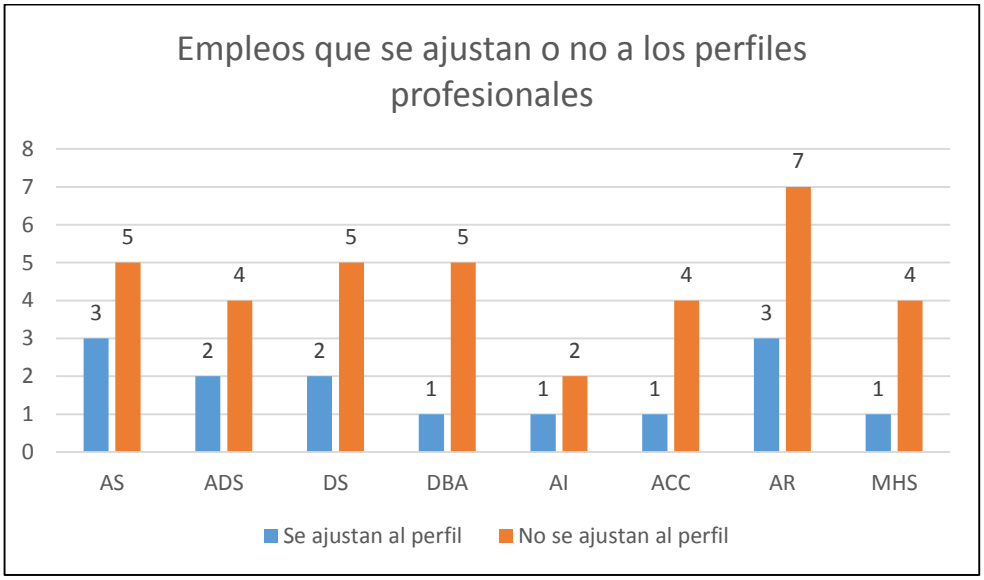


Figura 1. Cantidad de empleos que se ajustan por cada uno de los ocho perfiles.

Ahora realizamos un análisis de manera general respecto del porcentaje de empleos que se ajustan los perfiles profesionales (ver figura 2).

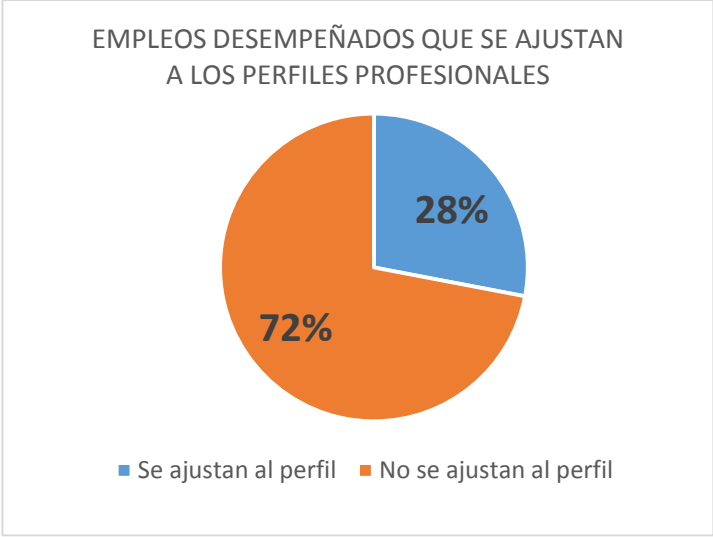


Figura 2. Porcentaje de los empleos que se ajustan a los perfiles profesionales de forma general.

La figura 2 nos muestra que el 72% de los empleos no se ajustan al perfil profesional, mientras que el 28% se acercan al perfil determinado, lo que nos dice que. Con esto se ha podido observar (ver la tabla I) que los empleos que no se ajustan al perfil profesional

determinado corresponden a otro de los perfiles planteados o a su vez se salen totalmente del contexto de las ramas de la carrera de ingeniería en sistemas.

TABLA I

ESPECIFICACIÓN DE LOS EMPLEOS POR CADA PERFIL PROFESIONAL

No.	Perfiles Profesionales	Se acercan al perfil	No se acercan al perfil
1	Analista de Sistemas de Información	Analista de Sistemas de Información	Secretaria, Técnica Informática, venta de equipos informáticos, Técnico en sistemas
2	Arquitecto y Diseñador de Software.	Diseñador de Software.	Analista desarrolladora de software, propietario de hospital computer (servicio técnico de tecnología), Desarrollador de software, Coordinadora del Patronato de Centinela del Cóndor
3	Desarrollador de software	Programador Junior, desarrollador de software	Asistente de Tecnología, técnico en computadores, venta de equipos informáticos, ayudante en ventas.
4	Administrador de Sistemas de Bases de Datos	oficial de seguridad de la información	Analista diseñador de pruebas de SW, secretaria, cajera, venta de equipos, mantenimiento de software
5	Auditor Informático	Auditor informático	Cajera, desarrollador de software
6	Administrador de Centros de computo	Administrando un centro de cómputo	Cajera, Soporte Técnico de Pc's, secretaria, mantenimiento
7	Administrador de Redes computacionales.	Analista Zonal TIC. Ministerio de Educación, mantenimiento de redes, Administrador de Red y Técnico	Desarrollador de aplicativos y sistemas informáticos para el Ejército Ecuatoriano, Técnico de soporte, secretaria, Venta y Reparación de Equipos Informáticos, programador de sistemas, Técnico en mantenimiento de computadoras, técnico, cajera en un supermercado,
8	Especialista en mantenimiento hardware y software.	Mantenimiento de Recursos informáticos	Programador de aplicaciones web, Desarrollo de Software, programador junior, Analista Zonal de TIC CZ7

Por lo que podemos decir que existen muy pocos egresados o graduados que están ejerciendo su profesión de acuerdo a su perfil profesional, por ello se nota lo necesario que es el conocer y hacer conocer todas sus cualidades personales, nivel de formación, experiencia, habilidades, capacidades e intereses que lo caracterizan y lo hacen diferente de los demás reflejadas en este perfil profesional, lo que implicaría que al tener

conocimiento del mismo podrán postular y hacer una relación con cierta ocupación y/o vacante, con el fin de mejorar o asegurar sus probabilidades de éxito.

Objetivos del Negocio:

- Identificar los perfiles profesionales enfocados en la carrera de ingeniería en sistemas, a través de la formación de los estudiantes.
- Identificar los factores que determinan el perfil profesional de los estudiantes.
- Conocer los perfiles profesionales a los cuales se orienten un grupo de estudiantes.

Resultados y Aporte:

Del desarrollo del presente trabajo se han obtenido algunos resultados, de acuerdo a los perfiles profesionales determinados mediante las técnicas de minería de datos de los egresados y graduados de la carrera de ingeniería en sistemas del Área de Energía las Industrias y los Recursos Naturales No Renovables de la Institución.

Los principales resultados encontrados al culminar el presente trabajo son:

- **Respecto a los modelos generados con el proceso de minería de datos**

A través de la herramienta RapidMiner seleccionada, en el proceso de minería se ha realizado la aplicación de dos grupos de técnicas: Clasificación y las técnicas de reglas basadas en inducción. Dentro de cada grupo de técnicas se ha utilizado diferentes algoritmos. Para las técnicas de Clasificación basadas en los algoritmos de árboles de decisión, se ha aplicado específicamente 2 algoritmos: ID3 y CHAID. Mientras que en las basadas en Reglas de Inducción se han escogido 6 algoritmos: JRip, Part, Ridor, Decisión Table, DTNB y NNge.

En las pruebas con el conjunto de entrenamiento se ha tomado un 72% de los datos mientras que el 28% ha sido utilizado para la evaluación de los modelos, para las dos estructuras de minería diseñadas de datos no agrupados y para la estructura de datos agrupados. Donde se ha podido observar que al agrupar las unidades los porcentajes de clasificación varían muy poco entre las dos estructuras, el único algoritmo que logro tener un considerable aumento en pruebas de entrenamiento es el ID3, por lo tanto se

ha notado que realizar la agrupación en los datos no es tan esencial, por ello se ha realizado la selección de los mejores algoritmos en base a datos no agrupados (ver figura 3), descartando los modelos de los datos agrupados.

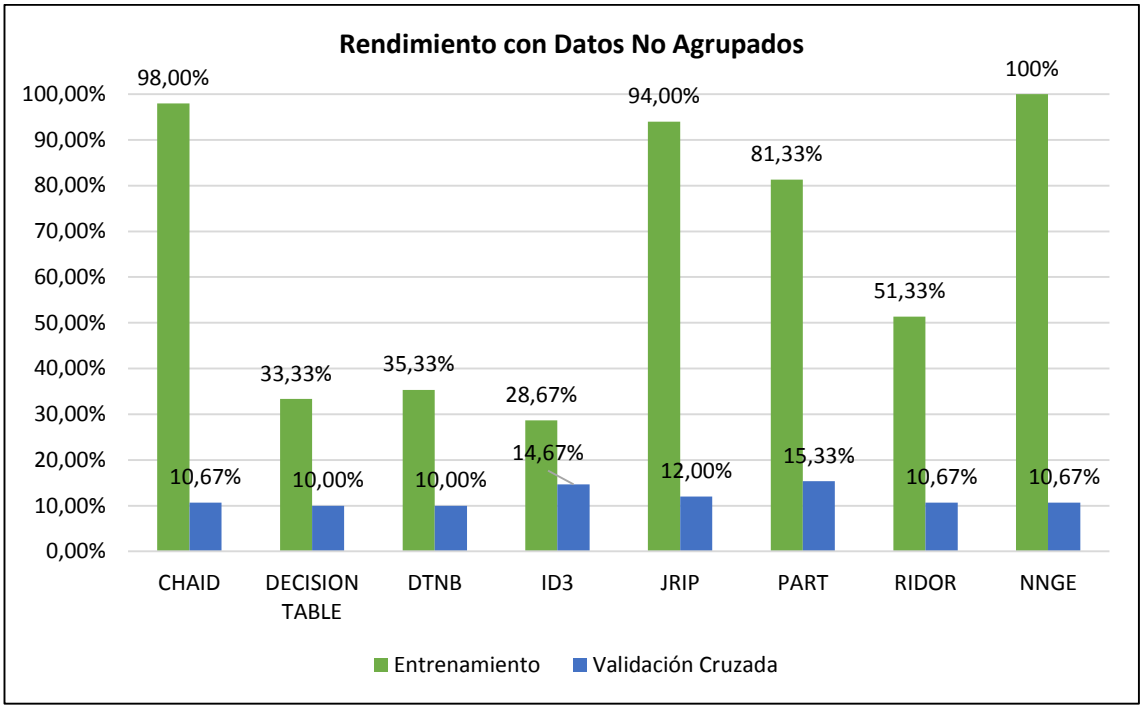


Figura 3. Comparación Rendimiento en pruebas de Entrenamiento y Validación Cruzada con datos no agrupados.

En la figura 3, se observa el rendimiento de los algoritmos, donde aparentemente NNge es el más óptimo con un porcentaje de rendimiento del 100%, el cual ha sido aplicado en una herramienta diferente a RapidMiner, debido a que es el único algoritmo que no soporta, esta nueva herramienta hace uso de la librería de weka.jar, y ha sido desarrollada en base al lenguaje de programación java, herramienta que ha sido mencionada en el transcurso del desarrollo del presente trabajo. Sin embargo a pesar del perfecto porcentaje de clasificación siendo un algoritmo robusto no se lo tomó en cuenta debido que no maneja correctamente la presencia de valores nulos, reflejado en que las reglas generadas por el algoritmo toman en cuenta los valores nulos o perdidos, por lo tanto son difíciles de interpretar y de utilizar.

Continuando con el análisis CHAID presenta el 98% de instancias bien clasificadas seguido de JRip con el 94%, siendo los mejores en esta característica así como en el

análisis del rendimiento, lógica de reglas y medidas de error por lo que han sido seleccionados como los mejores hasta el momento.

- **Resultados de la evaluación de mejores modelos obtenidos**

Se ha realizado el análisis de los resultados obtenidos en la minería, mediante una evaluación de los modelos, para ello se ha tomado en cuenta el 28% restante de los datos, con la finalidad de observar las coincidencias entre los perfiles profesionales, y corroborar la validez de los algoritmos CHAID y JRip elegidos como lo más óptimos hasta el momento y así determinar finalmente el mejor (ver figura 4).

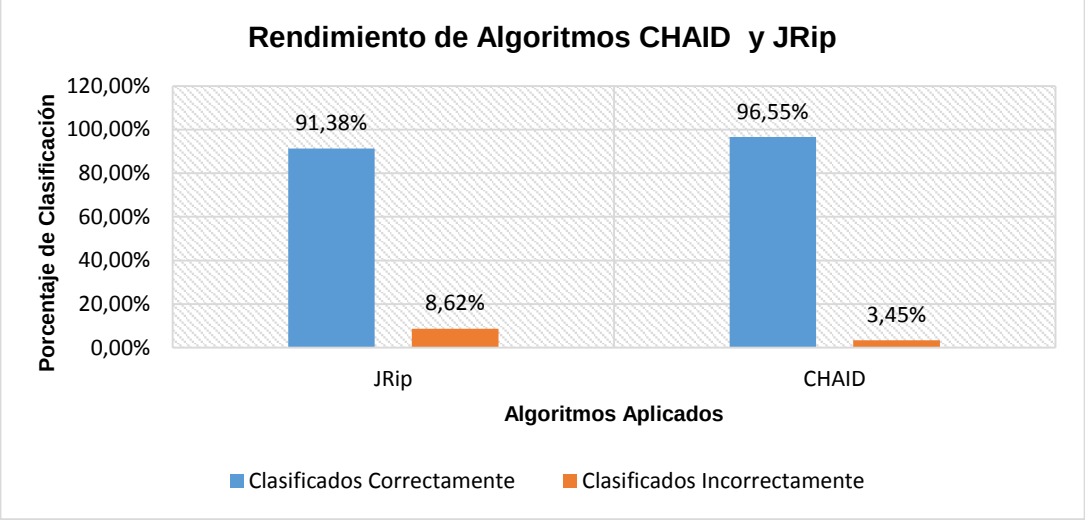


Figura 4. Rendimiento de algoritmos CHAID y JRip en predicción de perfiles profesionales.

Como se puede observar en la figura 4 el algoritmo que mejor rendimiento presenta es CHAID logrando clasificar el 96.55% de las instancias mientras que el algoritmo JRip clasificó el 91.38%, siendo una leve diferencia, sin embargo en base a ello se ha seleccionado al algoritmo CHAID como el más óptimo para la predicción de los perfiles profesionales enfocados a los egresados y graduados de la carrera de ingeniería en sistemas.

- **Resultados de la aplicación de los modelos de minería de datos en un contexto real.**

Evaluando los modelos y asegurando su calidad, se ha realizado la aplicación de los mismos en nuevos datos que corresponden a los últimos egresados de la carrera de ingeniería en sistemas año 2014, para probar su rendimiento en un contexto real y tomar la decisión final. Los modelos utilizados provienen de los algoritmos CHAID y JRip, considerando hasta el momento que CHAID ha resultado ser el más óptimo seguido de JRip.

Finalmente se ha comparado el porcentaje global de determinación de perfiles profesionales mediante los modelos CHAID y JRip para los últimos egresados de la carrera de ingeniería en sistemas (ver figura 5).

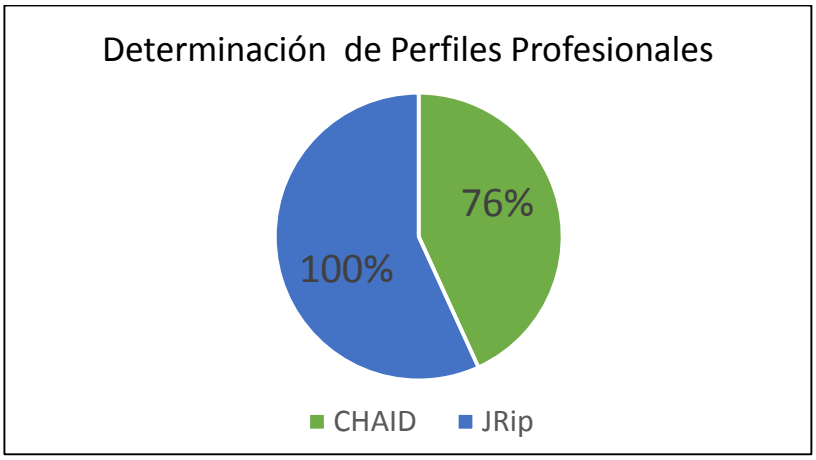


Figura 5. Porcentaje de Predicción final de los modelos CHAID y JRip.

Al observar los resultados de la figura 5 se demuestra que ha sido fundamental la aplicación de los modelos generados en un contexto real, debido a que JRip realiza la determinación de los perfiles profesionales en un 100%, mientras que CHAID lo hace en un 76%, cambiando la perspectiva y validando el modelo generado por el algoritmo JRip perteneciente a las técnicas de reglas de inducción como el óptimo para la determinación de perfiles profesionales en la carrera de ingeniería en sistemas de la UNL.

Observemos la simbología de los perfiles profesionales planteados durante el desarrollo del presente trabajo (ver tabla II).

TABLA II

SIMBOLOGÍA DE LOS PERFILES PROFESIONALES

Sigla	Perfil Profesional
AS	Analista de Sistemas de Información
ADS	Arquitecto y Diseñador de Software.
DS	Desarrollador de software
DBA	Administrador de Sistemas de Bases de Datos
AI	Auditor Informático
ACC	Administrador de Centros de computo
AR	Administrador de Redes computacionales.
MHS	Especialista en mantenimiento hardware y software.

Finalmente tenemos el porcentaje de los perfiles profesionales determinados por el modelo JRip, escogido finalmente como el más óptimo al ser validado en un contexto real (ver figura 6).

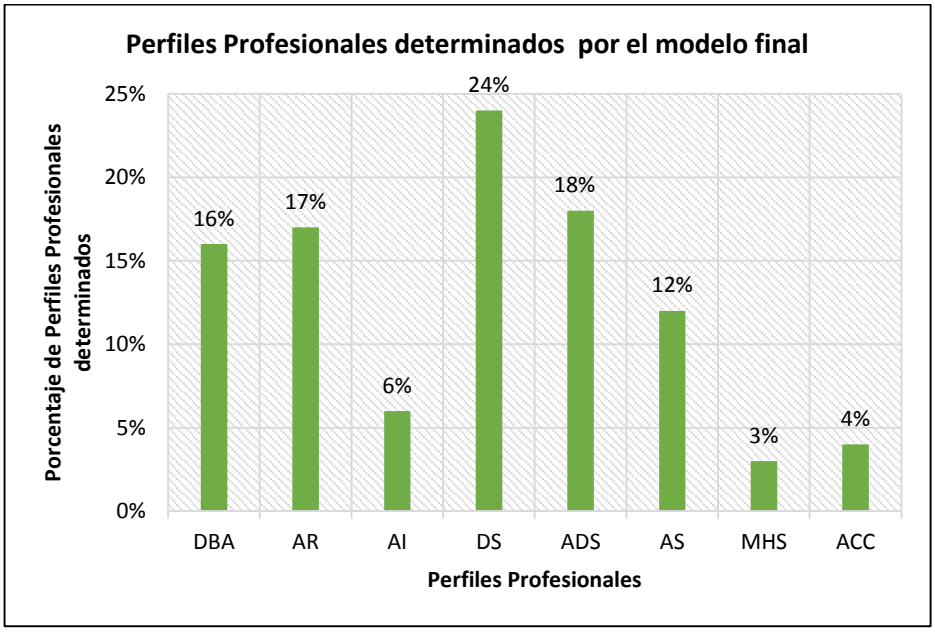


Figura 6. Determinación de perfiles profesionales mediante el modelo JRip.

En la figura 6 se puede observar que el perfil profesional que más se destaca entre la población es el perfil ‘Desarrollador de software’, seguido de ‘Arquitecto y Diseñador de

software', mientras que el perfil con menor porcentaje corresponde a 'Especialista en mantenimiento Hardware y Software', observando que la mayor parte de los egresados de la carrera de ingeniería en sistemas del año 2014 son desarrolladores.

Con los resultados descritos anteriormente a continuación se detallan las estrategias de solución, recomendaciones y trabajos futuros:

- Determinación y difusión del perfil profesional a través del Sistema de Gestión Académica de los estudiantes de la carrera de ingeniería en sistemas al finalizar sus estudios académicos, tomando el modelo generado por el algoritmo JRip determinado como el más óptimo.
- De acuerdo al perfil profesional determinado de cada estudiante actualizar los datos del Sistema de seguimiento a graduados de la carrera de ingeniería en sistemas y brindarles información de maestrías, cursos, empleos relacionados a su perfil, que les brinden mejores oportunidades en la vida profesional.
- Tomar como referencia el proceso que se ha realizado en el presente trabajo para encontrar los modelos de determinación de perfiles profesionales en el resto de carreras del Área de la Energía, las Industrias y los Recursos Naturales no Renovables, así como en el resto de áreas de la institución, con el fin de dar apoyo a todos los egresados y graduados que año a año salen de la Universidad Nacional de Loja.

Podemos mencionar que al realizar el análisis y procesamiento de la información para la minería de datos se advirtió la importancia de recomendar que las mallas curriculares se actualicen con materias en cada periodo académico, es decir, sean mejoradas de acuerdo a los avances que representa seguir una carrera que va de la mano con el progreso tecnológico.

Se ha visto la importancia que en la fase de análisis de información, aplicar varias técnicas para la aplicación de minería de datos y en base a la comparación de resultados confirmar cuál de ellas resulta ser la más apropiada para obtener el resultado esperado.

Se recomienda que para la generación de modelos futuros para a la determinación de perfiles profesionales enfocados a otras carreras de la Institución, se aplique la técnicas de reglas de inducción; puesto que se ha comprobado durante el desarrollo del presente

trabajo que ésta técnica es con la que se obtiene mejores resultados en cuanto a las predicciones que se realizan en un contexto real.

Finalmente se puede mencionar que para el desarrollo del presente trabajo y alcanzar los objetivos del mismo, se utilizaron herramientas de Software libre como: DatAdmin para la gestión de la base de datos, manipulación directa de los datos recopilados y creación de las vistas minables. RapidMiner para el proceso de minería de datos con el apoyo de sus sin número de operadores para la aplicación de los algoritmos.

Y para la generación de modelos de forma correcta y ordenada es importante que la técnica de minería de datos sea apropiada, se ha utilizado la metodología de CRISP – DM, por lo que se la recomienda ya que se enfoca a proyectos de minería conformada por varias fases que permiten el desarrollo de forma ordenada del proyecto, hacia la consecución de los objetivos establecidos.

Anexo 13: Certificado de Traducción del Trabajo de Titulación

MARÍA NATACHA RUÍZ SALCEDO

LICENCIADA EN LENGUA EXTRANJERA ESPECIALIDAD INGLÉS DEL COLEGIO DE BACHILLERATO FISCOMISIONAL "DANIEL ÁLVAREZ BURNEO".

CERTIFICA:

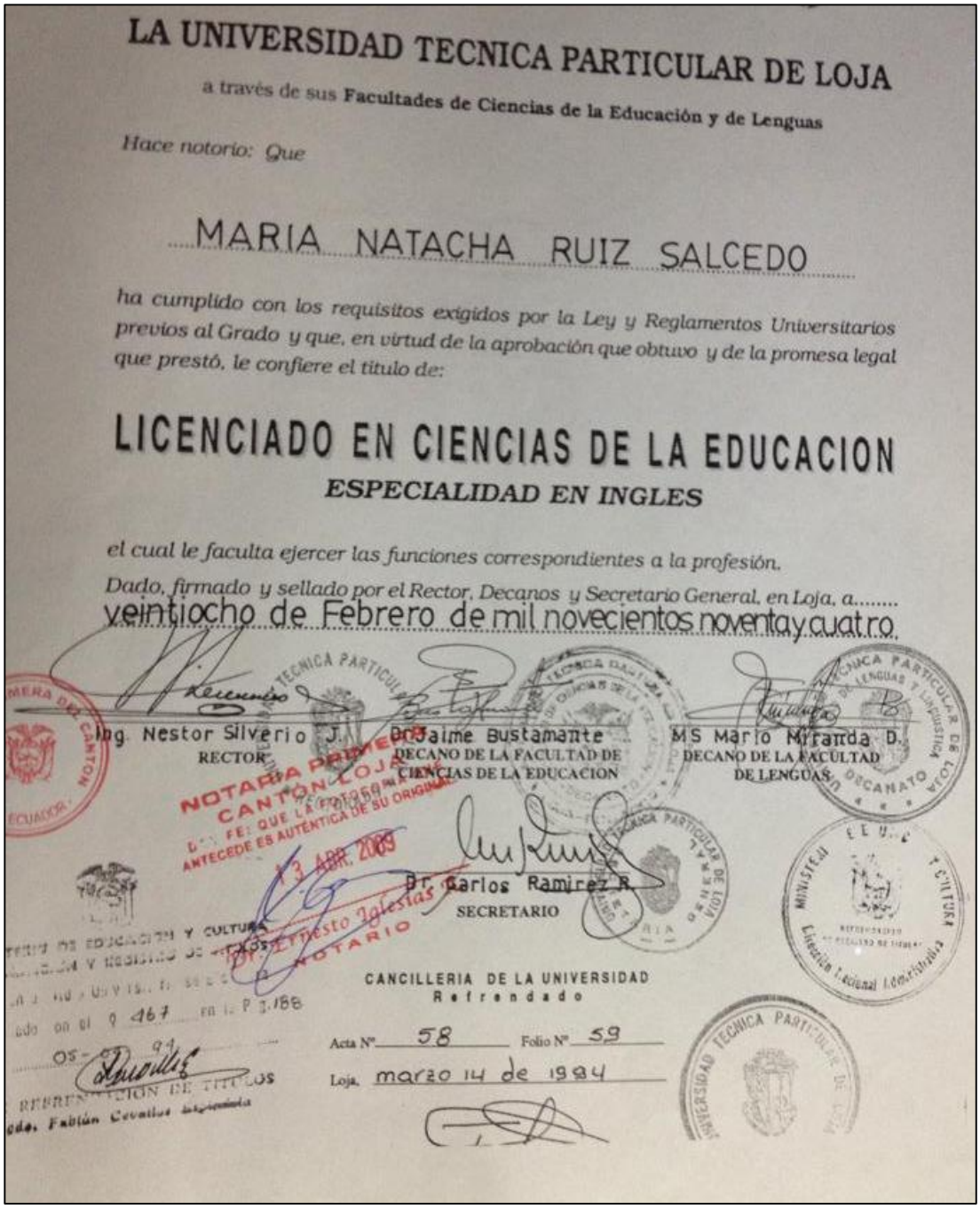
Que el resumen del Trabajo de Titulación denominado "DETERMINACIÓN DE PERFILES PROFESIONALES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS", realizado por la Srta. Egresada MARÍA JOSÉ RODRÍGUEZ OJEDA, previa a la obtención del título de INGENIERO EN SISTEMAS, es una traducción correcta y verdadera del idioma español a inglés, con lo mejor de mis conocimientos y entendimiento.

Por lo cual autorizo su presentación, sustentación y defensa.

Loja, 30 de octubre de 2014.

MARÍA NATACHA RUÍZ SALCEDO
LICENCIADA EN LENGUA EXTRANJERA

Figura 1. Certificado de Traducción



LA UNIVERSIDAD TECNICA PARTICULAR DE LOJA

a través de sus **Facultades de Ciencias de la Educación y de Lenguas**

Hace notorio: Que

MARIA NATACHA RUIZ SALCEDO

ha cumplido con los requisitos exigidos por la Ley y Reglamentos Universitarios previos al Grado y que, en virtud de la aprobación que obtuvo y de la promesa legal que prestó, le confiere el título de:

LICENCIADO EN CIENCIAS DE LA EDUCACION

ESPECIALIDAD EN INGLES

el cual le faculta ejercer las funciones correspondientes a la profesión.

Dado, firmado y sellado por el Rector, Decanos y Secretario General, en Loja, a.....

veintiocho de Febrero de mil novecientos noventa y cuatro

Ing. Nestor Silverio J. — RECTOR

Dr. Jaime Bustamante — DECANO DE LA FACULTAD DE CIENCIAS DE LA EDUCACION

MS Mario Miranda D. — DECANO DE LA FACULTAD DE LENGUAS

Dr. Carlos Ramirez R. — SECRETARIO

CANCILLERIA DE LA UNIVERSIDAD
Refrendado

Acta Nº 58 Folio Nº 59

Loja, marzo 14 de 1994

NOTARIA PRIMERA CANTON LOJA — DOY FE: QUE LA FOTOCOPIA QUE ANTECEDE ES AUTÉNTICA DE SU ORIGINAL — 13 ABR. 2009 — Dr. Ernesto Iglesias — NOTARIO

Figura 2. Título de la persona que certifica la traducción del trabajo de titulación