UNIVERSIDAD
NACIONAL
DE LOJA

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

CARRERA DE INGENIERÍA EN SISTEMAS

"Identificación de Factores en la Reprobación y Deserción mediante técnicas de Minería de Datos en el Área de la Energía de la Universidad Nacional de Loja"

"Tesis previa a la Obtención del título de Ingeniero en Sistemas"

Autor:

- González Pineda, Anibal Israel

Director:

- Ing. Pablo Fernando, Ordoñez Ordoñez, Mg. Sc.

LOJA-ECUADOR
2014

CERTIFICACIÓN DEL DIRECTOR

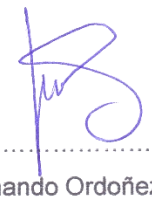
Ing. Pablo Fernando Ordoñez Ordoñez, Mg. Sc.

DOCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS

CERTIFICA:

Que el egresado Anibal Israel González Pineda autor del presente trabajo de titulación, cuyo tema versa sobre **"IDENTIFICACIÓN DE FACTORES EN LA REPROBACIÓN Y DESERCIÓN MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN EL ÁREA DE LA ENERGÍA DE LA UNIVERSIDAD NACIONAL DE LOJA",** ha sido dirigido, orientado y discutido bajo mi asesoramiento y reúne a satisfacción los requisitos exigidos en una investigación de este nivel por lo cual autorizó su presentación y sustentación.

Loja, Junio de 2014



.....

Ing. Pablo Fernando Ordoñez Ordoñez, Mg. Sc.

DIRECTOR DEL TRABAJO DE TITULACIÓN

AUTORÍA

Yo Anibal Israel González Pineda declaro ser autor del presente trabajo de tesis y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales por el contenido de la misma.

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

Autor: Anibal Israel González Pineda

Firma:.....

Cédula: 1104895824

Fecha: 17 de Octubre de 2014

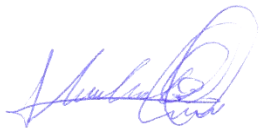
CARTA DE AUTORIZACIÓN DE TESIS POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.

Yo **ANIBAL ISRAEL GONZÁLEZ PINEDA**, declaro ser autor de la tesis titulada: **"IDENTIFICACIÓN DE FACTORES EN LA REPROBACIÓN Y DESERCIÓN MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN EL ÁREA DE LA ENERGÍA DE LA UNIVERSIDAD NACIONAL DE LOJA"**, como requisito para optar el 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 la tesis que realice un tercero.

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



Firma:.....

Autor: Anibal Israel González Pineda

Cédula: 1104895824

Dirección: Loja (Dolores Cacuango y Farabundo Martí)

Correo Electrónico: aigonzalezp@unl.edu.ec

Teléfono: 2326268 **Celular:** 0980572522

DATOS COMPLEMENTARIOS

Director de Tesis: Ing. Pablo Fernando Ordoñez Ordoñez, Mg. Sc.

Tribunal de Grado: Ing. Luis Roberto Jácome Galarza, Mg. Sc.

Ing. Henry Patricio Paz Arias, Mg. Sc.

Ing. Franco Hernán Salcedo López, Mg. Adm.

AGRADECIMIENTO

Agradezco a DIOS todopoderoso por brindarme la oportunidad de obtener este triunfo personal, por concederme salud, sabiduría y entendimiento para lograr esta meta.

A la ilustre Universidad Nacional de Loja, al Área de la Energía, las Industrias y los Recursos Naturales No Renovables, a la Carrera de Ingeniería en Sistemas, por darme la oportunidad de formarme como profesional; al coordinador y distinguidos catedráticos, por guiarme y compartir sus sabias enseñanzas en el transcurso de mi formación académica.

De manera personal agradezco al Ing. Pablo Fernando Ordoñez Ordoñez, quien con sus valiosos conocimientos y dedicación supo guiarme en el desarrollo del presente trabajo de Titulación.

Anibal Israel González Pineda.

DEDICATORIA

A dios sobre todo y a mis padres José y María que con sus consejos y apoyo constante me han ayudado a formarme como persona y académicamente. Gracias por todos los esfuerzos que han tenido que realizar para que pudiera hacer lo que me gusta en todo momento. Gracias por guiarme por el camino correcto y corregirme. Gracias en definitiva por todo.

A mis queridos hermanos Danny y Lizbeth, que de alguna manera u otra celebran mi éxito.

Dios les pague a todos y todas aquellas personas que han contribuido conmigo.

Anibal Israel González Pineda.

CESIÓN DE DERECHOS

Anibal Israel González Pineda autor principal del presente trabajo de titulación, autoriza a la Universidad Nacional de Loja, al Área de la Energía, las Industrias y los Recursos Naturales No Renovables y por ende a la Carrera de Ingeniería en Sistemas hacer uso del mismo en lo que estime sea conveniente.

a. Título

“IDENTIFICACIÓN DE FACTORES EN LA REPROBACIÓN Y DESERCIÓN MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN EL ÁREA DE LA ENERGÍA DE LA UNIVERSIDAD NACIONAL DE LOJA”

b. Resumen

Actualmente las Instituciones de Educación Superior enfrentan un problema, este es la deserción universitaria y surge desde que el estudiante ingresa al colegio y luego pasa por el Sistema Nacional de Nivelación y Admisión (SNNA), y posteriormente cursa una carrera profesional. Para ello se han empleado una gran cantidad de esfuerzos con el fin de que los alumnos trabajen directamente o aumenten su compromiso académico. Para contrarrestar este problema se evidencia una cantidad de actividades como: talleres, tareas extra clase entre otros, sin lograr mejorar esta situación, es decir estos esfuerzos no han sido suficientes y aún el problema persiste en las instituciones educativas de nivel superior.

En base a lo descrito anteriormente, el objetivo principal del presente Trabajo de Titulación es la identificación de factores que inciden en la deserción y reprobación universitaria. Para realizar todo este proceso, se aplicaron de técnicas de minería de datos que permitirán extraer conclusiones e información relevante sobre la deserción y reprobación. La selección de las técnicas de minería de datos se justifica por fuentes bibliográficas, puesto que son esenciales para el caso de estudio. En cuanto a los métodos aplicados para lograr el desarrollo eficiente del presente trabajo se ha utilizado la metodología CRISP-DM, la cual da la posibilidad de llevar un trabajo ordenado e iterativo.

Los datos utilizados para llevar a cabo la identificación de los factores, se obtuvo de las bases de datos del Sistema de Gestión Académica (S.G.A.) a través de su Web Services, además se recopiló datos del Área de Bienestar Universitario, todo esto para luego integrar en una sola fuente de datos, para su posterior procesamiento.

Finalmente luego de identificar los factores de deserción y reprobación, se generó un modelo predictivo de deserción con el propósito de validar los resultados obtenidos, el cual se evaluó con datos de estudiantes que cursan actualmente las carreras del Área de Energía de las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja.

Summary

Currently the Higher Education Institutions face a complex problem, this problem is the college dropout and emerges from the student starts school and then passes through the National System of Equalization and Admission (SNNA), and then studying for a career. This has been used a lot of efforts in order to have students work directly or enhance their academic commitment. For countering this problem a number of activities have been evidenced such as workshops, extra class, tasks, among others. Without achieving improve the situation, it means, these efforts have not been sufficient and the problem still persists in educational higher-level institutions

Based on writing above, the main objective of this work is to identify the factors that influence on university dropout and fail. To achieve this process were applied to data mining techniques that will allow drawing conclusions and relevant information on dropout and failure rates. The selection of data mining techniques is justified by literature sources; they are essentially for this study. By other hand the methods used to achieve, the efficient development of this work has used the CRISP-DM methodology, which gives the possibility to bring an orderly and iterative work.

The data used to carry out the identification of the factors was obtained from the databases of SISTEMA DE GESTIÓN ACADÉMICA (SGA) through its Web services also details University Welfare Area was collected, and then integrate this on a single source of data for later processing.

Finally, after identifying factors dropout and failure, a predictive model of attrition in order to validate the results was obtained, which was evaluated with data from students currently sign on of Energy Industries and Natural Resources Non-Renewable carriers of Universidad Nacional de Loja.

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 preparados o una infraestructura adecuada; sin embargo la deserción y reprobación universitaria persiste como problema que se desea reducir [1-6]. En ese aspecto es necesaria la identificación de factores de deserción y reprobación en los estudiantes universitarios, cuando están cursando los primeros años de su carrera profesional, siendo imprescindible para emprender acciones oportunas y poder mitigar este índice; y no menos significativo, predecir su deserción en cualquier momento [7-10].

En base a lo descrito, el presente Trabajo de Titulación se enfoca en identificar los factores de deserción y reprobación, utilizando información previamente proporcionada por la institución y luego mediante la aplicación de técnicas de minería de datos descubrir patrones ocultos en estos datos.

Para el desarrollo del presente Trabajo de Titulación, se inició con la recopilación y estudio de las diferentes técnicas de minería de datos, en el cual se evidencio que las técnicas de árboles de decisión y técnicas de inducción de reglas son las que mejores resultados ofrecen; posteriormente luego de estudiar las técnicas de minería de datos, se procedió a evaluar los modelos generados por varios algoritmos para identificar los factores de deserción y reprobación, además de un modelo predictivo de deserción. Cabe mencionar que la minería de datos aplicada a la educación ha dado como resultado de gran apoyo a predecir cualquier tipo de factor o característica de un caso, fenómeno o situación en la educación de nivel superior de un estudiante [11,15].

Posteriormente se evaluó los modelos generados por los algoritmos con mejor rendimiento enfocados a deserción, con estudiantes que cursan actualmente sus carreras en el Área de Energía las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja.

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

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 muestra una síntesis de lo que involucra todo el Trabajo de Titulación; ÍNDICE que describe los temas tratados, su ubicación, así como el índice de tablas y figuras; INTRODUCCIÓN que engloba una descripción general de lo relevante que es el trabajo y un abstracto del proceso desarrollado para la obtención de resultados; REVISIÓN LITERARIA comprende las temáticas útiles para el desarrollo del Trabajo de Titulación; METODOLOGÍA comprende los materiales, métodos y técnicas que fueron empleados; RESULTADOS se centra en las actividades que fueron realizadas en el trascurso de todo el trabajo; DISCUSIÓN involucra una descripción en el que deben constar los objetivos y el proceso que se realizó en cada uno de ellos para el cumplimiento de los mismos, además se detalla la valoración técnica, económica, ambiental del presente trabajo; CONCLUSIONES describe las ideas que se generó tras la culminación del trabajo; RECOMENDACIONES engloba los trabajo futuros. Finamente el trabajo de titulación culmina con sus respectivas FUENTES BIBLIOGRAFÍA Y ANEXOS.

d. Revisión de Literatura

1. CAPÍTULO I: RECOPILACIÓN DE CASOS DE ÉXITO EN FUENTES ACADÉMICAS, REVISTAS, PONENCIAS, ARTÍCULOS CIENTÍFICOS, SOBRE LA APLICACIÓN DE MINERÍA DE DATOS EN LA DESERCIÓN.

En el campo educacional, las técnicas de minería de datos han sido aplicadas con la finalidad de entender el comportamiento de los estudiantes [4,12], para recomendar actividades [5], para ofrecer nuevas experiencias de aprendizaje [6] o con el objetivo de mejorar la efectividad del curso, para promover el trabajo en grupo [7] o incluso predecir el rendimiento de los alumnos [8, 9, 12].

Se han recopilado algunos casos de éxito orientados a la deserción universitaria, los cuales generan conclusiones y recomendaciones que podrán ser tomados en cuenta al realizar el presente Trabajo de Titulación [8-10, 11].

1.1 CASO DE ÉXITO 1: Aplicando Minería de Datos al Marketing Educativo

Este caso de estudio trata sobre la importancia de la inteligencia de negocios y en especial, la minería de datos en el sector educativo, aplicado esencialmente en los registros del estudiante, desde que ingresa en la universidad y las posibles causas que originan la deserción en cada periodo académico.

1.1.1 Introducción

La inteligencia de negocios, y en especial la técnica de minería de datos, ocupan un lugar relevante en las ventajas empresariales y la toma de decisión en el sector privado. Muchas de las técnicas de minería de datos son utilizadas en el mundo corporativo.

Sin embargo, son transferibles a la educación superior, dado que dentro de las mismas se almacenan grandes cantidades de datos que enmarcan el expediente de los estudiantes.

Con dicha información las Universidades podrían ayudar a conocer los perfiles del estudiante que actualmente estudia en la universidad, así como el perfil del estudiante desertor. Sobre este último tema, actualmente no solo las instituciones de educación superior privadas y públicas, sino el mismo Ministerio de Educación del país, enfrentan el problema que muchos de los estudiantes que terminan su bachillerato e ingresan en la educación superior, por diversos factores se retiran, sin llegar a culminar su ciclo de estudio o lo culminan en instituciones diferentes a las que empezaron, mientras otros sí la terminan en la misma institución, pero con un título diferente del que iniciaron.

1.1.2 Minería de datos y grupo de semillero de investigación Perceptron

A comienzos del año 2010, se conformó un grupo de investigación en la Escuela de Marketing y Publicidad, específicamente en el programa de Marketing y Negocios Internacionales llamado, Perceptron. Dicho grupo de investigación debió en primera instancia, capacitar a estudiantes (semilleros de investigación) y docentes en técnicas de Minería de datos aplicadas al mercadeo [16].

Posteriormente, como primer objetivo de investigación, se centró en el marketing educativo, haciendo gestión de conocimiento en las bases de datos existentes en la universidad, para caracterizar el perfil de los estudiantes que ingresan y desertan en los programas de Publicidad Internacional y Marketing & Negocios Internacionales.

Debido a que la investigación partía de extraer conocimiento, que en algunos casos la dirección de la escuela suponía, se inició aplicando un método no supervisado; es decir que no se tiene variable objetivo, para primero tratar de comprender su base datos en busca de descubrir patrones y tendencias; con ellos se usó la técnica de Agrupamiento bajo un método No-Jerárquico con el algoritmo K-means.

1.1.3 Caracterización del perfil del estudiante de la escuela y del perfil del estudiante desertor.

La escuela de Marketing y Publicidad consta de dos programas académicos. El primero es Marketing y Negocios Internacionales, que comenzó sus labores en el segundo semestre del año 2001; y el segundo programa es Publicidad Internacional, que comenzó labores el primer semestre de 2006. Ambos programas tienen en común un

alto porcentaje de mujeres inscritas en la escuela. Según los datos entregados por el sistema de información de la Universidad Sergio Arboleda SINFA [47], aproximadamente el 57% de los estudiantes es de género femenino, y el 43 % de género masculino. Específicamente el programa de Marketing y el programa de Publicidad cuentan con un 56% y 63 % respectivamente de estudiantes de género femenino.

La información de los estudiantes fue recopilada por medio de un censo en el momento en que los estudiantes se inscribieron y matricularon. Estos datos incluyen información demográfica y socioeconómica de la familia del estudiante. Posteriormente se retroalimenta dicha información con las notas y el promedio acumulado.

Para ello, se transformó la base de datos, incluyendo datos categóricos a numéricos, codificando cada uno de los metadatos. Se tuvieron en cuenta las siguientes variables:

- DNI: documento de identificación
- Semestre
- Ciudad de domicilio
- Ciudad de domicilio del acudiente
- Departamento de domicilio del acudiente
- Sexo o género
- Edad
- Estado civil
- País de nacimiento
- Estrato
- Medio por el cual se enteró del programa y de la universidad
- Idioma

1.1.4 Razones de deserción

En la investigación las bases de datos tuvieron que reprocesarse, debido a que se incluyen en el expediente estudiantes graduados, estudiantes que aplazaron semestre, estudiantes que están cursando y estudiantes que no han vuelto a matricularse. Al depurar las bases de datos, posteriormente fueron normalizadas y se les aplicó el algoritmo de K-Means (ver figura 1) [17].

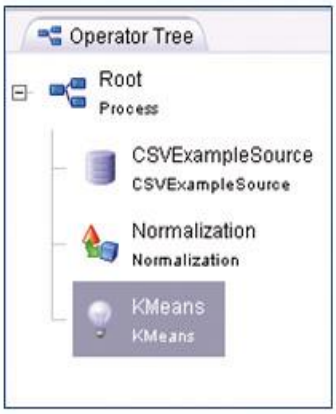


Figura 1: Procedimiento del modelo usando Rapid Miner.

1.1.5 Caracterización del perfil estudiante de la escuela

Aplicando la técnica de agrupamiento en la base de datos, se encontró que existen unos factores relevantes que tipifican el perfil del estudiante del programa de Marketing y Negocios Internacionales como son, el nivel educativo y profesión de los padres, estado civil, origen de la inscripción, descripción de colegios en cuanto al calendario, la jornada y la integración de géneros [18]. En base la base de datos, se hizo variar el número de K que definiría el número de clúster óptimo que caracterizará el perfil del estudiante.

Como resultado final, se obtuvo para el programa de Marketing y Negocios Internacionales una elección de K=3, que representó 3 tipos de clústers que para el caso de la investigación, constituyeron perfiles significativos (ver figura 2).

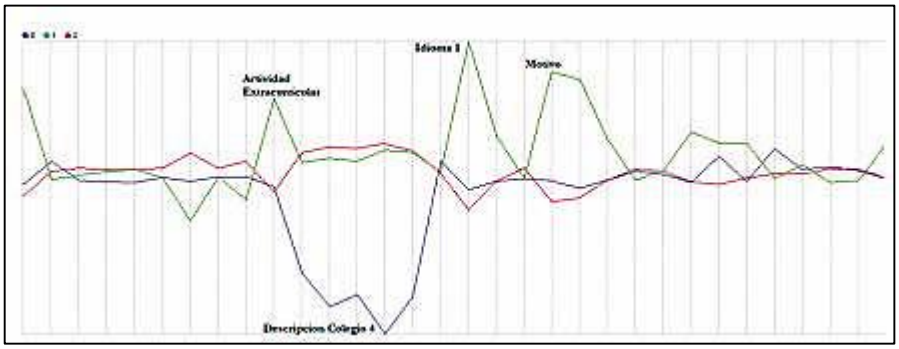


Figura 2: Gráfica del Clúster de perfiles de los estudiantes de Marketing y Negocios Internacionales.

1.1.6 Caracterización del perfil desertor

Aplicando la técnica de agrupamiento en la base de datos, se encontró que existen unos factores relevantes dentro del proceso de deserción de los estudiantes de Publicidad Internacional y de los estudiantes de Marketing y Negocios Internacionales, como son la edad, el sexo, el medio por el cual se enteraron y la ciudad donde se encuentran sus acudientes.

Las principales razones por las cuales los estudiantes desertan del programa, según la base de datos complementada por tele mercadeo son: bajo rendimiento académico, motivos económicos, personales, cambio de carrera, motivos salud, embarazo, viaje, motivos laborales, traslado de ciudad, motivos académicos, traslado de país, cambio de universidad, calamidad doméstica, cambio de sede a Santa Marta, motivos seguridad, matriculó pero nunca asistió, motivos disciplinarios y otros estudios.

Para ello, se puso a varia el número de K que definiría el número de clúster óptimo que caracterizará el perfil del estudiante desertor. En cada uno de las bases de datos de estudiantes desertores se puso a variar el valor de k = 2 hasta k = 7.Como resultado final, se obtuvo para el programa de Marketing y Negocios Internacionales una elección de K=3, con tres clúster (ver figura 3) [19,20].

Clúster 0: Se caracteriza por ser un grupo únicamente de mujeres, principalmente estudiantes desertores de los semestres I, II y III, en el que sus edades más representativas son 19, 20, 21 y 23 años; la gran mayoría con sus acudientes vive en la ciudad de Bogotá; ingresaron en los períodos 2006-02, 2007-02 y 2008-02 y las principales razones de retiro del semestre son bajo rendimiento académico y problemas económicos.

Clúster 1: Este grupo se caracteriza por ser únicamente de hombres, desertores de los semestres I y II; su edad más representativa es 19 años; la gran mayoría con sus acudientes viviendo en la ciudad de Bogotá, que ingresaron principalmente en los períodos 2007-02 y 2008-02. Las principales razones para retirar el semestre son bajo rendimiento académico, problemas económicos y por el requisito del idioma inglés.

Clúster 2: Este clúster está representado por mujeres y hombres, en un porcentaje equivalente; con alumnos que han desertado principalmente en los semestres V y VI, su edad más representativa es de 26 años, la gran mayoría tiene acudientes que viven en la ciudad de Bogotá. Son alumnos que ingresaron principalmente en el período 2003-01 y cuya razón de deserción es el bajo rendimiento académico.

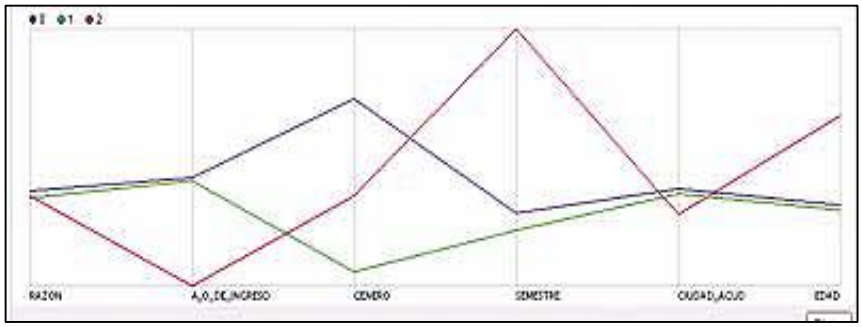


Figura 3: Gráfica del Clúster de deserción de los estudiantes de Marketing y Negocios Internacionales.

1.2 CASO DE ÉXITO 2: Aplicación de Técnicas de Minería de Datos para la Evaluación del Rendimiento Académico y la Deserción Estudiantil.

El presente caso de estudio presenta los resultados de la evaluación del rendimiento académico y de la deserción estudiantil en el Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM). La investigación se efectuó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008.

1.2.1 Introducción

Debido a la gran cantidad de información generada por las distintas áreas de cualquier institución resulta imprescindible la utilización de las Tecnologías de la Información y la Comunicación (TIC) para que la información pueda ser almacenada, transformada, analizada y visualizada. Algunas de las TIC utilizadas en este proyecto como parte del proceso de descubrimiento del conocimiento en bases de datos (DCBD) fueron: base de datos, análisis estadístico unidimensional y multidimensional y aprendizaje automático. Esta investigación forma parte del conjunto de acciones que fueron planificadas en el marco de las acreditaciones de las carreras de Ingeniería Electrónica e Ingeniería Industrial.

Las carreras que se dictan en la UNLAM están distribuidas en 4 Departamentos (Unidades Académicas) y tomando los datos del año 2008 se encuentran matriculados aproximadamente 35000 estudiantes. En el DIIT se dictan las carreras de Ingeniería Informática, Ingeniería Electrónica e Ingeniería Industrial cuyas matrículas son 4480, 919 y 613 respectivamente.

En la Tabla I se pueden observar los resultados de un primer análisis cuantitativo del rendimiento de los estudiantes del DIIT durante el año 2008.

TABLA I:
DESCRIPCIÓN RENDIMIENTO DE LOS ESTUDIANTES POR CARRERA.

Asignaturas Aprobadas	**Cantidad de alumnos Informática**	**Cantidad de alumnos Electrónica**	**Cantidad de alumnos Industrial**
0	467	199	70
1	784	106	79
2	1182	186	147
3	799	130	119
4	542	160	64
5	392	56	56
Más de 5	314	82	78
Total	4480	919	613

El objetivo de este trabajo es presentar un estudio que utilizando el proceso DCDB permita, a través de clasificadores, identificar:

- El rendimiento académico de los alumnos.
- Los patrones determinantes de la deserción estudiantil.

Durante las distintas etapas de este proceso se utilizaron los datos de los alumnos desde el año 2003 hasta el año 2008. Las herramientas de software utilizadas fueron:

- El motor de base de datos MS SQL Server para realizar la recopilación, integración y almacenamiento de los datos.
- El programa estadístico SPPS para realizar la depuración, selección y transformación de los datos.
- El programa Weka para obtener los clasificadores aplicando técnicas de minería de datos.

1.2.2 Resultados del proceso de descubrimiento del conocimiento en base de datos.

A continuación de describen los resultados obtenidos en el presente trabajo, por cada fase del proceso de descubrimiento.

1.2.2.1 Fase de Recopilación e Integración

El resultado de esta fase fue la generación de un almacén de datos conformado por 7 tablas cuyas descripciones se pueden ver en la Tabla II. Para la generación del almacén de datos se tomaron e integraron datos de la base de datos de alumnos de la UNLaM, de la base de datos de encuestas del DIIT y de la bases de datos de colegios de educación secundaria del Ministerio de Educación.

TABLA II:
DESCRIPCIÓN POR TABLAS.

Tablas	Descripción
Alumnos	Datos del Estudiante
Carreras	Datos de las carreras del DIIT.
Planes Estudio	Datos de los planes de estudio, vigentes y no vigentes, de las carreras.
Materias	Datos de las materias de los planes de estudio.
Exámenes	Datos de las notas, por carrera, plan de estudio y materia, de los estudiantes.
Censos	Datos de los censos realizados a los estudiantes.
Secundarios	Datos de los colegios de educación secundaria.

En el área de integración se transformaron los siguientes atributos:

- fecha de nacimiento: se redefinió este campo de tipo carácter con una longitud de 8 a tipo fecha con longitud determinada por la configuración del motor de base de datos.
- año de ingreso: se redefinió este campo de tipo carácter con una longitud de 2 a tipo numérico con una longitud de 4 sin decimales.
- fecha de examen: se redefinió este campo de tipo carácter con una longitud de 8 a tipo fecha con longitud determinada por la configuración del motor de base de datos.

1.2.2.2 Fase de Limpieza, Selección y Transformación

La calidad de los patrones que se obtienen con la minería de datos es directamente proporcional a la calidad de los datos utilizados. Esta fase es la responsable de obtener datos de alta calidad. Para lograr este objetivo se buscó detectar valores anómalos (outliers) y datos faltantes, se realizó una selección de los atributos relevantes y se construyeron nuevos atributos a partir de los existentes.

Los atributos generados, para cada estudiante, fueron:

- *edad:* este atributo se generó a partir de la fecha de nacimiento.
- *indice_materias:* este atributo se generó tomado el resultado de la división de la cantidad de materias aprobadas por la cantidad de años entre la fecha actual o fecha de abandono y la fecha de ingreso. La cantidad de materias aprobadas se obtuvo de la cantidad de instancias en la tabla Exámenes con un valor en el atributo Nota igual o mayor que 4. En la Tabla III se puede ver la discretización de este atributo.
- *reprobadas:* este atributo se generó a partir de la cantidad de instancias en la tabla Exámenes con un valor en el atributo Nota menor que 4.
- *promedio:* este atributo es el cálculo del promedio del alumno.

TABLA III:
DISCRETIZACIÓN DEL ATRIBUTO INDICE_MATERIAS.

indice_materias	**Valor**
Menor a 2	1 – Malo
Mayor a 1,99 y menor a 3	2 – Regular
Mayor a 2,99 y menor a 4,5	3 – Bueno
Mayor a 4,49 y menor a 5,5	4 – Muy bueno
Mayor a 5,49	5 – Excelente

1.2.2.3 Fase de Minería de Datos.

Dentro del proceso de DCDB está fase es la encargada de producir nuevo conocimiento [21]. En este trabajo se decidió utilizar:

- La clasificación como tipo de área de minería.
- El árbol de decisión como tipo de modelo.
- El J48 (implementación en Weka del algoritmo C4.5) [22] y el FT [23] como algoritmos de minería.

En la Tabla IV se pueden ver los atributos del archivo elaborado para la fase de minería de datos. Este archivo contiene 9545 instancias que representan a los alumnos inactivos, activos y reincorporados. Para entrenar los modelos se utilizó un archivo con 2865 instancias (30% del original), que fueron seleccionadas en forma aleatoria.

TABLA IV:
ATRIBUTOS DEL ARCHIVO UTILIZADO EN LA FASE DE MINERÍA DE DATOS.

Nombre	Descripción	Tipo
sexo	1 – Masculino 2 – Femenino	Nominal
edad	Edad	Numérico
estado_civil	1 – Casada/o 2 – Divorciada/o 3 – Soltera/o 4 – Separada/o 5 – Viuda/o	Nominal
carrera	201 – Ing. en Informática 202 – Ing. Electrónica 203 – Ing. Industrial	Nominal
estado	1 – inactivo 2 – activo 3 – reincorporado	Nominal
indice_materias	1 – malo 2 – regular 3 – bueno 4 – muy bueno 5 – excelente	Nominal
promedio	Promedio del alumno	Numérico
reprobadas	Cantidad de materias no aprobadas	Numérico
trabajo	1 – No trabaja 2 – Trabaja	Nominal
horas	Total de horas trabajadas diariamente	Numérico
horario	1 – Mañana 2 – Tarde 3 – Noche	Nominal
gestion_escuela	1 – Estatal 2 – Privada	Nominal
tipo_escuela	1 – Bachiller 2 – Comercial 3 – Polimodal 4 – Técnica	Nominal
estudio_padre	1 – Sin Estudios 2 – Estudios primarios 3 – Estudios secundarios	Nominal

	4 – Estudios superiores	
estudio_madre	1 – Sin Estudios 2 – Estudios primarios 3 – Estudios secundarios 4 – Estudios superiores	Nominal

Se eligieron como clases los siguientes atributos:

- *indice_materias:* para encontrar los patrones determinantes del rendimiento académico.
- *estado:* para encontrar los patrones determinantes de la deserción estudiantil.

1.2.2.3.1 Rendimiento Académico

El mejor resultado fue obtenido por el algoritmo FT que alcanzó un 78,07% de instancias clasificadas correctamente, mientras que el algoritmo J48 clasificó en forma correcta un 72,53% de las instancias. En las siguiente tabla (ver tabla V) se puede observar la matriz de confusión generada por el algoritmo FT y en la siguiente (ver tabla VI) la generada por el algoritmo J48.

TABLA V:
MATRIZ DE CONFUSIÓN GENERADA POR EL ALGORITMO FT.

a	**b**	**c**	**d**	**e**	
3197	437	276	41	5	a=1-Malo
184	2093	151	27	6	b=2-Regular
79	380	1357	34	13	c=3-Bueno
199	69	26	552	28	d=4-Muy Bueno
16	54	23	45	253	e=5-Excelente

TABLA VI:
MATRIZ DE CONFUSIÓN GENERADA POR EL ALGORITMO J48.

a	**b**	**c**	**d**	**e**	
3588	276	28	41	23	a=1-Malo
387	1794	201	46	33	b=2-Regular
460	253	1081	52	17	c=3-Bueno
169	254	138	276	37	d=4-Muy Bueno
32	60	69	46	184	e=5-Excelente

1.2.2.3.2 Deserción Estudiantil

Al igual que en la clasificación anterior el mejor resultado fue obtenido por el algoritmo FT que alcanzó un 77,86% de instancias clasificadas correctamente contra el 72,78%

logrado por el algoritmo J48. En las siguientes tablas (ver tablas VII y VIII) se pueden ver las matrices de confusión generadas por los algoritmos FT y J48 respectivamente.

TABLA VII:
MATRIZ DE CONFUSIÓN GENERADA EN DESERCIÓN
POR EL ALGORITMO FT

Inactivo	Activo	Reincorporado	
2344	716	183	Inactivo
390	3799	135	Activo
367	322	1289	Reincorporado

TABLA VIII:
MATRIZ DE CONFUSIÓN GENERADA EN DESERCIÓN
POR EL ALGORITMO J48

Inactivo	Activo	Reincorporado	
2435	694	114	Inactivo
687	3546	91	Activo
389	623	966	Reincorporado

1.2.2.4 Fase de Evaluación e Interpretación

En un contexto ideal los patrones descubiertos por la fase de minería de datos deben reunir 3 cualidades: ser precisos, comprensibles e interesantes [21]. En este trabajo nos interesó mejorar principalmente la comprensibilidad.

Para efectuar la evaluación de los modelos se tomó como medida el porcentaje de aciertos al clasificar una instancia en su respectiva clase. Por cada algoritmo se realizaron 30 iteraciones y en la siguiente tabla (ver tabla IX) se pueden ver los mejores porcentajes de aciertos.

TABLA IX:
PORCENTAJE DE ACIERTOS DE LOS ALGORITMOS
DE CLASIFICACIÓN

	FT	J48
Rendimiento Académico	78.07%	72,53%
Deserción estudiantil	77,86%	72,78%

En la Tabla 9 se puede observar que el algoritmo FT tuvo un mejor desempeño que el algoritmo J48. Pero si se analizan las matrices de confusión (Tablas 5, 6,7 y 8) se puede ver que para detectar un rendimiento académico malo y alumnos inactivos el algoritmo J48 supera al FT (ver tabla X).

TABLA X:
PORCENTAJE DE ACIERTOS DEL RENDIMIENTO ACADÉMICO MALO Y DE LOS ALUMNOS INACTIVOS

	FT	J48
Rendimiento Académico	80.81%	90,70%
Alumnos inactivos	72,28%	75,08%

- Que el algoritmo J48 generó un árbol de decisión muy grande y por lo tanto poco comprensible y difícil de interpretar.
- Que el árbol generado por el algoritmo FT no permite explicar el rendimiento académico y las causas de la deserción estudiantil.

1.2.2.5 Fase de Difusión y Uso

Durante el curso del primer cuatrimestre se espera trabajar con un archivo que contenga los datos del año 2009 para poder seguir evaluando el poder predictivo de los modelos.

1.3 CASO DE ÉXITO 3: Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos.

En el presente caso de estudio se describen los resultados de la investigación realizada en la Universidad de Nariño, del país Colombia, cuyo propósito fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de minería de datos, con información almacenada en las bases de datos durante los últimos 15 años.

1.3.1 Introducción

Debido al avance de la tecnología en los sistemas computacionales, se hace indispensable y necesaria la utilización de tecnologías informáticas que contribuyan a resolver ciertos problemas que sin la utilización de ellas, se haría prácticamente imposible el tratamiento de los mismos, brindando soluciones eficientes y sustentadas en la realidad para aplicarlas en el contexto en el que se encuentran. Una de estas tecnologías es la minería de datos, en la que se fundamentó todo el proceso investigativo de este proyecto.

La Universidad de Nariño es una institución pública de educación superior cuya área de influencia es el suroccidente de Colombia, cuya sede principal se encuentra en la ciudad de Pasto, capital del departamento de Nariño. En ella se encuentra la mayoría de estudiantes universitarios de la región. Los estudiantes de educación secundaria aspiran obtener un cupo en ésta, por su calidad educativa, y prestigio de sus egresados. Desafortunadamente, en algunos casos, cuando el estudiante se matricula a un determinado programa, su rendimiento no es el esperado, generando índices de deserción altos y bajo rendimiento académico. Por lo tanto se genera un interrogante acerca de cuáles son las causas que motivan la deserción y/o el bajo rendimiento y que perfiles tienen este tipo de estudiantes.

1.3.2 Etapa de Selección

El objetivo de esta etapa es obtener las fuentes de datos internas y externas que sirven de base para el proceso de Minería de Datos. Como fuente interna, se seleccionó la base de datos histórica de los estudiantes de la Universidad de Nariño, compuesta por información personal y académica de 46173 estudiantes. Como fuente externa, se seleccionó la información de los colegios de educación secundaria del país, que se obtuvo con el Ministerio de Educación Nacional de Colombia.

Estas fuentes de datos se integraron en la base de datos UDENARDB, construida con el sistema gestor de base de datos PostgreSQL, UDENARDB la componen siete tablas, cuyas descripciones se pueden ver adelante (ver tabla XI).

1.3.3 Etapa de Pre-procesamiento de Datos

El objetivo de esta etapa es obtener datos limpios, i.e. datos sin valores nulos o anómalos que permitan obtener patrones de calidad. Por medio de consultas ad-hoc sobre la base de datos UDENARDB, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de las tablas y se dejaron únicamente los atributos más relevantes para la investigación y aquellos que no contenían valores nulos. Como resultado de esta etapa, quedaron únicamente datos de 20329 estudiantes, para su posterior análisis. De la tabla alumnos se seleccionaron 19 atributos, de la tabla carreras 4, de la tabla facultades 2, de la tabla materias 3, de la tabla notas 8, de la tabla liquidación 11 y de la tabla colegios 3 atributos. Los atributos seleccionados de las

diferentes tablas en su gran mayoría no contenían valores nulos ni anómalos (*outliers*), pero en aquellos casos que se presentaban, estos fueron reemplazados utilizando técnicas estadísticas tales como la media y la moda o derivando sus valores a través de otros como por ejemplo la edad de ingreso del estudiante conocida la fecha de ingreso y la fecha de nacimiento.

TABLA XI:
DESCRIPCIÓN DE TABLAS DE LA BASE DE DATOS UDENAR.

Tablas	No. Atributos	Descripción
ALUMNOS	69	Se encuentran todos los datos personales del estudiante.
CARRERAS	10	Se encuentra información de todas las carreras existentes en la Universidad de Nariño
FACULTADES	4	Contiene información de las facultades de la Universidad de Nariño.
MATERIAS	4	Se encuentran toda la información de las materias existentes en el plan académico de cada carrera.
NOTAS	8	Contiene información de las notas por materia de cada estudiante.
LIQUIDACIÓN	27	Se encuentra toda la información financiera del estudiante
COLEGIOS	7	Contiene información de los colegios del país

1.3.4 Etapa de Transformación de Datos

En la etapa de transformación, se buscan características útiles para representar los datos dependiendo de la meta del proceso de minería de datos. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos.

1.3.5 Etapa de Minería de Datos

La etapa de minería de datos es la más característica del proceso DCBD [24]. El objetivo de esta etapa es la búsqueda y descubrimiento de patrones insospechados y de interés utilizando áreas de descubrimiento tales como clasificación, clustering [27], patrones secuenciales [26] y asociación [25] entre otras. Para el descubrimiento de patrones de deserción estudiantil y bajo rendimiento académico se utilizaron las áreas de Clasificación y Asociación.

1.3.5.1 Reglas de Clasificación

Para predecir los perfiles de bajo rendimiento académico, el conjunto de datos *udenar.arff* se clasificó escogiendo como clase el atributo *clasepromedio*. Este atributo indica el rendimiento académico del estudiante basado en el promedio acumulado de las notas hasta el semestre cursado.

Entre las reglas de clasificación más representativas están:

- Si el estrato socioeconómico es 2, el ponderado de exámenes de estado ICFES está entre 50 y 70, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Humanas, entonces su rendimiento es Bajo. El 68% con estas características se clasifican de esta manera.
- Si la edad de ingreso es menor o igual a 18 años, el estrato socioeconómico es 2, género masculino, el ponderado ICFES está entre 50 y 70, vive con la familia, es del Sur de Nariño, está en primer semestre, está en la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 67% con estas características se clasifican de esta manera.
- Si la edad de ingreso es menor o igual a 18 años, proviene de un colegio privado, el calendario del colegio es septiembre a junio, género femenino, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 70% con estas características se clasifican de esta manera.

Para predecir los perfiles de deserción estudiantil se escogió como clase el atributo Clase_al. Este atributo indica si el estudiante no se ha retirado, ha reingresado o se retiró definitivamente de la Universidad.

Entre las reglas de clasificación más representativas están:

- Más del 50% de los estudiantes retirados que pertenecen a la facultad de ingeniería, reingresan.
- Los estudiantes retirados que pertenecen a las facultades de Ciencias Naturales y Matemáticas y Ciencias Humanas no reingresan.

1.3.5.2 Reglas de Asociación

Entre las reglas de Asociación más representativas, que permiten identificar relaciones no explícitas entre los atributos del conjunto de datos *udenar.arff* que involucran bajo rendimiento y deserción están:

- El 95% de los estudiantes que tienen promedio bajo está en primer semestre. El 10% de todos los estudiantes son de primer semestre y tienen promedio bajo.
- El 84% de los estudiantes retirados son de estrato socioeconómico 2 y provienen de municipios del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, son de estrato 2 y provienen del Sur de Nariño.
- El 89 % de los estudiantes retirados son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y proceden del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y provienen del Sur de Nariño.
- El 88% de los estudiantes retirados, tienen una edad de ingreso menor que 18 años provienen del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, tienen una edad de ingreso menor que 18 años y son del Sur de Nariño.
- El 86% de estudiantes retirados terminaron su bachillerato en colegios públicos, son de primer semestre y provienen del Sur de Nariño. El 2.5% de todos los estudiantes se han retirado, terminaron su bachillerato en colegios públicos, son de primer semestre y provienen del Sur de Nariño.

1.3.6 Etapa de Interpretación y Evaluación de Resultados

De acuerdo a los resultados obtenidos, la mayoría de los estudiantes de primer semestre, provenientes de la zona sur del departamento de Nariño, de estratos socioeconómicos bajos y matriculados en algún programa de la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas, presentan un bajo rendimiento académico. Este perfil es similar al perfil de la mayoría de estudiantes que se retiran. Por otra parte la mayoría de estudiantes que se retiran de estas dos facultades no reingresan, lo que no sucede en la facultad de Ingeniería, donde casi la mayoría de estudiantes retirados reingresan.

1.4 CASO DE ÉXITO 4: Aplicación de Técnicas de Minería de Datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL.

El presente caso de estudio está orientado en la identificación de patrones de comportamiento relacionados con acciones de los estudiantes que utilizan el entorno virtual de aprendizaje (EVA) de la Universidad Técnica Particular de Loja.

1.4.1 Introducción

En el campo de Tecnologías para la Educación, la adaptación de sistemas educativos ofrece una forma avanzada en un ambiente de aprendizaje que intenta satisfacer las necesidades de los diferentes estudiantes, así como sistemas de construcción de un modelo de conocimiento del estudiante, las metas y preferencias del mismo. Aquí es donde juega un papel fundamental la aplicación de las técnicas de minería contribuyendo al descubrimiento de patrones de comportamiento de los estudiantes durante su interacción con el EVA de la UTPL.

Algunos entornos de aprendizaje virtual (EVA) son adecuados para la investigación del comportamiento de los alumnos en el aprendizaje. Uno de los más populares de éstos es Moodle ampliamente utilizado para la presentación de materiales de aprendizaje, así como para los debates entre los alumnos. Esta herramienta permite a un profesor no sólo informar materiales de aprendizaje de forma flexible, sino que también permite proporcionar la posibilidad de que los alumnos participen en las discusiones comunes, chats sincrónicos, crear sus blogs, archivos de revisión de vídeo de las conferencias, utilizar el correo electrónico, etc. Moodle es también una poderosa herramienta para el seguimiento de las acciones de los estudiantes y la interpretación de éstos resultados.

1.4.2 Limpieza y Transformación de Datos

La recopilación de datos debe ir acompañada de una limpieza e integración de los mismos, para que estos datos estén en condiciones para su análisis. Los benéficos del análisis y de la extracción de conocimiento a partir de datos dependen, en gran medida, de la calidad de los datos recopilados. Además, generalmente, debido a las características propias de las técnicas de minería de datos, es necesario realizar una transformación de los datos para obtener una “materia prima” que sea adecuada para

el propósito concreto y las técnicas que se quieren emplear. En definitiva, el éxito de un proceso de minería de datos depende, no solo de tener todos los datos necesarios (una buena recopilación), sino de que estén íntegros, completos y consistentes.

1.4.3 Minería de Datos

En la minería de datos, es necesario tener los datos, ya generados como atributos, además de que en esta fase se evaluarán técnicas y algoritmos de aprendizaje automático para elegir cuál será el más apropiado para los datos que se ha seleccionado. El proceso de la minería será efectuado a través de la herramienta, de minería de datos WEKA.

Dentro de este proyecto de Trabajo de Titulación se mencionó que los atributos a identificar del modelo de estudiante atenderán a un modelo predictivo de las instancias, y para este tipo de modelo se optará por utilizar las áreas que son clasificación y análisis de secuencias.

Terminado esto se procedió a hacer uso de los algoritmos que posee WEKA y se guardó los modelos generados por cada experiencia de algoritmo utilizada ya, que luego las matrices de confusión también serían sujetas a análisis. La técnica de evaluación utilizada fue la de cross-validation (validación cruzada).

1.4.4 Resultados de las Técnicas de Minería de Datos

En la clasificación, se experimentó con los algoritmos J48, REPTree, BayesNet, NaiveBayes y JRip con previo un análisis de los resultados que presentó cada uno de éstos, se seleccionó el algoritmo que ofreció los mejores resultados en cuanto al número de clasificaciones correctas para ser sujeto a análisis, resultando ser la mejor opción el REPTree, en el caso de informática manifiesta que un 97.6804% presenta un nivel de participación en el curso escaso (E) y el 2.3196% está conformado por los estudiantes que presentan un nivel moderado (M) y permanente (P), para el nivel de utilización de las herramientas denota un 97.4227% en el nivel escaso y en el 2.5773% están situados los que presentan un nivel moderado (M) y permanente (P), para el caso de abogacía el algoritmo idóneo según sus resultados fue el J48, aquí en el 99.4962% intervienen los que presentan un nivel de participación en el curso escaso (E), mientras que en el

0.5038% están ubicados los que tienen un nivel moderado (M) y permanente (P), todo esto refleja un nivel bajo como el dominante en la participación o interés en el curso en los estudiantes de las dos carreras.

La técnicas de clustering fueron utilizadas para el indicador del nivel de utilización de las herramientas, se experimentó con el algoritmo SimpleKMeans eligiendo a dos grupos de población el primero lógica de la programación (informática), presentando una mayor incidencia en el nivel permanente y moderado en las herramientas (recursos, áreas, mensajería, twitter y cuestionario), el segundo ética y derechos humanos (abogacía), manifestando tener mayor incidencia solamente en la herramienta foros con un nivel moderado, en las demás herramientas presentan un nivel de utilización escaso.

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

Las técnicas de Minería de Datos (una etapa dentro del proceso completo de KDD) pretenden obtener patrones o modelos a partir de datos recopilados. Y es aquí donde el usuario realiza una evaluación subjetiva y decide si los modelos obtenidos son útiles.

Las técnicas de Minería de Datos se clasifican (ver tabla XII) en dos grandes categorías: las no supervisadas o descriptivas y las supervisadas o predictivas [28,29].

TABLA XII:
TÉCNICAS DE MINERÍA DE DATOS.

Técnicas No Supervisadas o Descriptivas	**Técnicas Supervisadas o Predictivas**
Reglas de Asociación	Árboles de Decisión
Clustering (Agrupamiento)	Redes Neuronales
	Máquinas de Soporte Vectorial
	Clasificadores Bayesianos
	Reglas de Inducción

Cualquiera sea el problema que se quiere resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado manejando aproximaciones distintas. El número de técnicas de minería de datos es muy grande y estas se pueden ampliar en el futuro, a continuación se muestra una breve reseña de las técnicas.

2.1 Técnicas No Supervisadas o Descriptivas

Las técnicas descriptivas son aquellas en las cuales no se asigna ningún papel predeterminado a las variables. No está considerada la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos, los modelos se crean automáticamente partiendo del reconocimiento de patrones [29, 30]. A continuación se describen algunas de las técnicas descriptivas.

2.1.1 Reglas de Asociación

Esta técnica de minería de datos también denominada reglas de asociación de cesta de la compra, porque esta técnica se utilizó inicialmente para descubrir la relación entre productos vendidos en una tienda, trabaja principalmente con atributos nominales,

además esta es muy utilizada en análisis de textos, búsquedas de patrones en páginas Web, etc. [29].

En las reglas de asociación se trabaja con dos medidas de calidad, la cobertura (support) y la confianza (confidence). La cobertura o soporte es el número de instancias que la regla predice correctamente. La confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar [30]. A continuación un ejemplo de lo que se considera una regla de asociación.

Si yogurt griego Danoni y cervezas El Murciélago ENTONCES pañales Dodit
Soporte=10%, Confianza=70%.

Esta regla nos muestra que había un 10% de casos donde aparecía una compra de yogurt griego Danoni y de cerveza El Murciélago, y en 70% de ellas se había comprado además pañales Dodit [30].

2.1.2 Clustering

Es una técnica también denominada agrupamiento, permite la identificación de grupos donde los elementos tienen gran similitud entre sí y muchas diferencias con los de otros grupos. Así por ejemplo se puede segmentar un conjunto de clientes, el conjunto de valores e índices financieros, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas. [29,31].

La agrupación o clustering está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma separada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado. [29,31].

La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar data mining sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización [29,31].

Hay varios algoritmos de clustering. A continuación se muestra el más conocido de estos.

2.1.2.1 Clustering Numérico (K-Medias)

Uno de los algoritmos más utilizados para hacer clustering es el *K-Medias (k-means),* que se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos clústeres se van a crear, este es el parámetro k, para lo cual se seleccionan *k* elementos aleatoriamente, que representaran el centro o media de cada clúster. A continuación cada una de las instancias, ejemplos, es asignada al centro del clúster más cercano de acuerdo con la distancia euclídea que le separa de él. Para cada uno de los clústers así construidos se calcula el centroide de todas sus instancias.

Estos centroides son tomados como los nuevos centros de sus respectivos clústers. Finalmente se repite el proceso completo con los nuevos centros de sus respectivos clústers. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clústers, ya que los puntos centrales de los clústers se han estabilizado y permanecerán invariables después de cada iteración. El algoritmo de k-medias es el siguiente (ver figura 4) [29].

1. Elegir k ejemplos que actúan como semillas (*k* número de clusters).
2. Para cada ejemplo, añadir ejemplo a la clase más similar.
3. Calcular el centroide de cada clase, que pasan a ser las nuevas semillas
4. Si no se llega a un criterio de convergencia (por ejemplo, dos iteraciones no cambian las clasificaciones de los ejemplos), volver a 2.

Figura 4: Pseudocódigo de algoritmo k-medias.

Para obtener centroides, se calcula la media o la moda según se trate de atributos numéricos o simbólicos. A continuación, se muestra un ejemplo de clustering con el algoritmo k-medias.

En este caso se parte de un total de nueve ejemplos o instancias, luego se configura el algoritmo para obtener tres clústers, y se inicializan aleatoriamente los centroides de los clústers a un ejemplo determinado. Una vez inicializados los datos, se comienza el bucle del algoritmo.

Cada uno de los ejemplos se representa con un tono de color diferente que indica la pertenencia del ejemplo a un clúster determinado, mientras que los centroides siguen

mostrándose como círculos de mayor tamaño y sin relleno. Por último el proceso de clustering finaliza en el paso 3 (ver figura 5) [29].

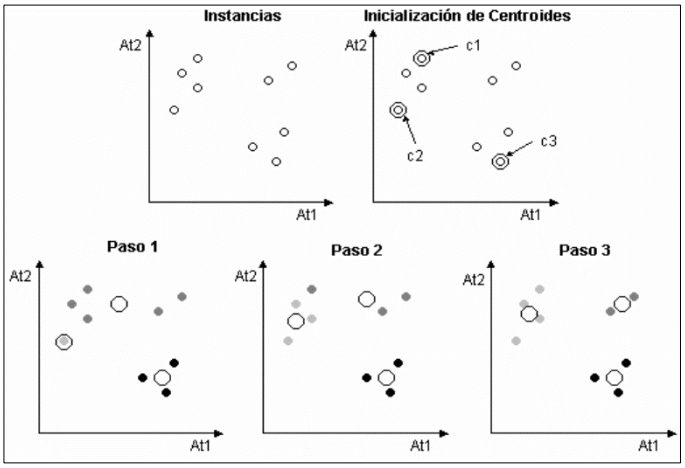


Figura 5: Ejemplo de clustering con k-medias.

2.2 Técnicas Supervisadas o Predictivas

Las técnicas predictivas son aquellas en las cuales se especifica el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como valido [29,30].

2.2.1 Árboles de Decisión

Esta es una de las técnicas más utilizadas de Minería de Datos porque tienen una representación sencilla de problemas con un número finito de clases. Además son modelos comprensibles y proposicionales, es decir algoritmos que aprenden modelos sobre una única tabla de datos y que no establecen relaciones entre más de una fila de la tabla a la vez ni sobre más de un atributo a la vez. Las condiciones expresan sobre el valor de un atributo [29,31].

Un árbol de decisión puede interpretarse como una serie de reglas compactadas para su representación en forma de árbol para una mejor forma visual [29,31].

Se observa en la siguiente figura (ver figura 6) como a partir del valor de la variable X8, si el valor es menor de 3.2 se continuara la toma de decisiones por la rama izquierda y si es mayor o igual se continuará por la rama de la derecha. A partir de aquí cada rama

tiene una variable separadora con un valor de separación, y así sucesivamente formando un árbol [29,31].

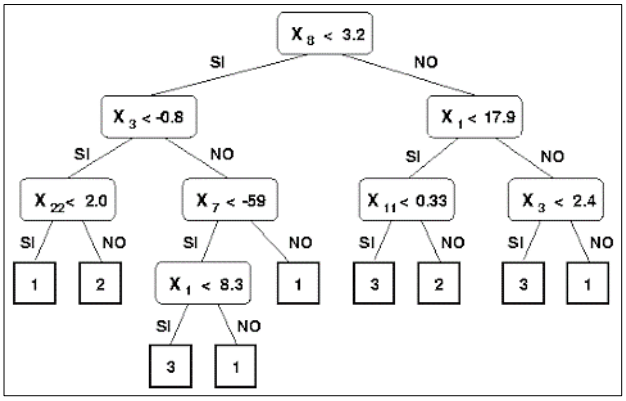


Figura 6: Ejemplo de árbol de decisión.

Algunos tipos de árboles de decisión se describen a continuación:

- **ID3**

Algoritmo desarrollado por J. Ross Quinlan en la década de los 80, este algoritmo construye árboles de decisión a partir de un conjunto de ejemplos, estos ejemplos están constituidos por un conjunto de atributos y un clasificador o clase. Los dominios de los atributos y de las clases deben ser discretos [29,31].

El algoritmo ID3 genera descripciones que clasifican a cada uno de las entidades del conjunto. El nivel de precisión en la clasificación comúnmente es alto. Sin embargo, tiene algunas desventajas, puesto que los atributos y clases deben ser discretos y no pueden ser continuos, además si se cuenta con conocimientos de dominio o conocimientos previos, el sistema no hace uso de ellos.

Algunas de las veces, los árboles generados son demasiado densos, los cuales son difíciles de interpretar y en caso de que esto suceda pueden ser transformados en reglas de decisión para mejorar su comprensión.

- **C4.5**

El algoritmo C4.5 es una variación del anterior ID3 ya que excluye muchas de sus limitaciones como: permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos *a>b* y otra para *a<=b*. Además,

los árboles son menos densos porque cada hoja no cubre una clase en particular sino una distribución de clases, lo cual los hace menos profundos y densos.

El C4.5 genera un árbol de decisión partiendo de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero *(depth-first)* [68,69]. Antes de realizar la partición de datos, el algoritmo toma en cuenta todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información.

Para cada atributo discreto, se toma en cuenta una prueba con *n* resultados, siendo *n* el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

Algunas de las mejoras que ofrece este algoritmo son [74]:

- ✓ Evitar un sobreajuste de los datos.
- ✓ Determinar la profundidad al momento de crecer un árbol de decisión.
- ✓ Reducir errores en la poda.
- ✓ Condicionar la Post-Poda.
- ✓ Manejar atributos continuos.
- ✓ Escoger un rango de medida apropiado.
- ✓ Manejo de datos de entrenamiento con valores faltantes.
- ✓ Manejar atributos con diferentes valores.
- ✓ Mejorar la eficiencia computacional

- **CHAID**

El algoritmo *CHAID* fue diseñada a partir de una técnica conocida como Detección Automática de Interacciones *(AID: Automatic Interaction Detection)*. De este varia su nombre, *Detección Automática de Interacciones con Prueba*, *CHAID (Chi Squared Automatic Interaction Detection)* [75]. Es utilizada como una técnica de explotación de datos como un modelo estadístico y forma parte, de los modelos llamados árboles de decisión y se aplica en la clasificación y predicción, con el fin de pronosticar el objetivo según una serie de criterios o variables independientes.

El modelo utiliza el algoritmo *CHAID* para crear grupos de registros que presenten la misma probabilidad de resultado, basándose en los valores de las variables independientes. El algoritmo parte del nodo raíz y se bifurca en nodos descendientes hasta llegar a los nodos hoja, donde finaliza la ramificación.

Posteriormente se aplica la prueba *(chi squared test)* para determinar si se debe continuar con la ramificación y, de ser positivo, qué variables independientes se deben usar. Esta prueba se realiza mediante la tabulación cruzada entre el resultado y cada una de las variables independientes. El resultado es un "valor-p", este valor representa la probabilidad de que la hipótesis nula sea correcta, luego los "valores-p" para cada tabulación cruzada de todas las variables independientes se clasifican, y si el mejor es decir, el valor más pequeño se encuentra bajo un umbral determinado, se realiza una ramificación del nodo raíz en esa ubicación. La bifurcación termina cuando el mejor "valor-p" ya no se encuentra bajo el umbral determinado y los nodos hoja del árbol son aquellos que no han sufrido ramificaciones [76].

Entre las ventajas de CHAID la más relevante es la construcción sencilla del modelo, que permite manejar variables independientes tanto discretas como continuas.

2.2.2 Redes Neuronales

Las redes neuronales es una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro. Esta técnica constituye una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales y es que estas son capaces de detectar y aprender complejos patrones y características dentro de los datos [30,31].

Las redes de neuronas constituyen una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender complejos patrones y características dentro de los datos. Esta técnica parte de un conjunto de datos de entrada específico y el objetivo es lograr que la red aprenda de manera automática las propiedades del sistema [30,31].

Para ello se debe emplear ejemplos positivos y negativos y se debe dividir los datos en tres grupos que no deben ser semejantes. El más grande es el conjunto de

entrenamiento *(Training Set)* que nos sirve para calcular los valores de activación de las neuronas. Para comprobar que estos valores calculados son correctos se dispone del conjunto de Validación *(Validation Set)*. Por último se utiliza un conjunto de Prueba *(Test Set)* para dar los parámetros de fiabilidad de la red neuronal [30,31].

Las redes neuronales se construyen estructurando en una serie de niveles o capas (entrada, procesamiento u oculta y salida) compuestas por nodos o conocidas como neuronas, como se muestran en la siguiente figura (ver figura 7) [30,31].

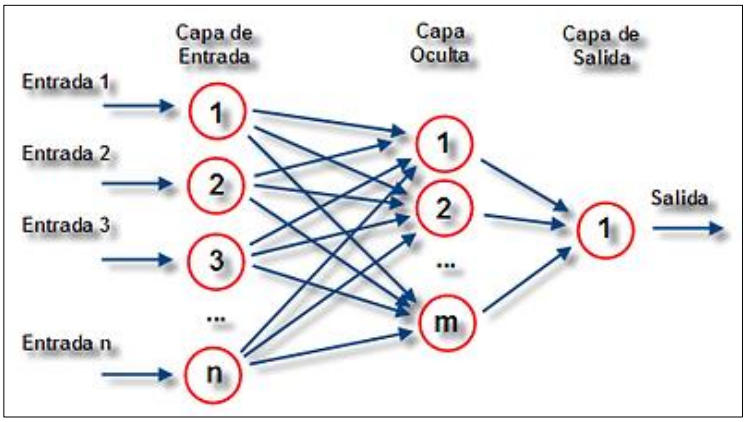


Figura 7: Ejemplo de Red Neuronal.

En la figura anterior se observa que las capas en una red neuronal artificial pueden clasificarse en tres tipos:

- Capa de entrada: constituida por aquellas neuronas que introducen los patrones de entrada en la red. En esas neuronas no se produce procesamiento.
- Capa oculta: formada por aquellas neuronas cuyas entradas provienen de capas anteriores y las salidas pasan a neuronas de capas posteriores.
- Capa de salida: neuronas cuyos valores de salida corresponden con las salidas de toda la red.

Sin embargo las redes de neuronas tienen el inconveniente de la dificultad de acceder y comprender los modelos que generan y presentan dificultades para extraer reglas de tales modelos. Otra característica es que son capaces de trabajar con datos incompletos e, incluso, contradictorios lo que, dependiendo del problema, puede resultar una ventaja o un inconveniente [30,31]. Las ventajas de las redes neuronales son:

- Aprendizaje: tienen la habilidad de aprender mediante una etapa de aprendizaje consistente en proporcionar a la RNA datos como entrada, a la vez que se le indica cual es la salida esperada.
- Auto organización: crea su propia representación de la información en su interior, sin necesidad de una programación explicita. Por ello, junto a la programación evolutiva se le llama computación flexible.
- Tolerancia a fallos: Almacena la información de forma redundante, esta puede seguir respondiendo aceptablemente aún si se daña parcialmente.
- Flexibilidad: puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada. Por ejemplo, si la información de entrada es la imagen de un objeto, la respuesta correspondiente es acertada incluso si la imagen tiene parámetros de luz ligeramente distintos o el objeto cambia ligeramente de posición.
- Tiempo real: es paralela, si se implementa con computadoras o en dispositivos electrónicos que utilicen dicha paralelización se pueden obtener respuestas en tiempo real, de la misma manera que el cerebro es capaz de procesar cantidades ingentes de información en paralelo sin esfuerzo aparente.

2.2.3 Máquinas de soporte Vectorial

Esta técnica también denominada *(SVM por Support Vector Machines)*, es una reciente técnica de clasificación que ha tomado mucha atención en años recientes. Se basa en un clasificador lineal muy sencillo *(el que maximiza la distancia de los tres ejemplos, que son los vectores soporte)*, precedido de una transformación de espacio *(a través de la función núcleo)* para darle potencia expresiva [30,32].

El clasificador lineal que se usa, sencillamente obtiene la línea en más dimensiones, el hiperplano, que divide limpiamente las dos clases y además, donde los tres ejemplos más próximos a la frontera están lo más distantes posibles. Son eficientes incluso para cientos de dimensiones, pues el separador lineal solo tiene que mirar unos pocos puntos *(vectores soporte)* y puede descartar muchos que estarán lejos de la frontera [30,32].

Si los datos no son separables linealmente se aplica una función núcleo *(de alemán "kernel")* que suele aumentar el número de dimensiones de tal manera que los datos sean separables.

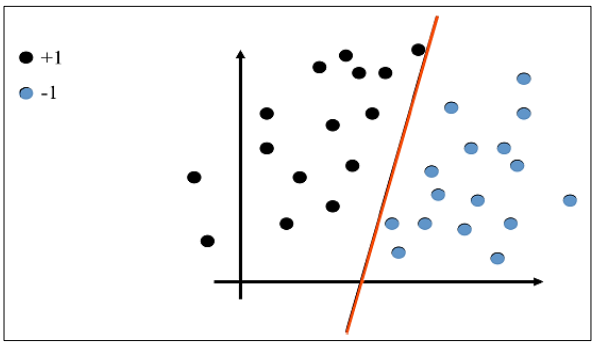


Figura 8: Ejemplo de hiperplano en máquinas de soporte vectorial.

En la figura anterior (ver figura 8) se observa como los ejemplos positivos y negativos son separados por un plano [30,32].

2.2.4 Clasificadores Bayesianos

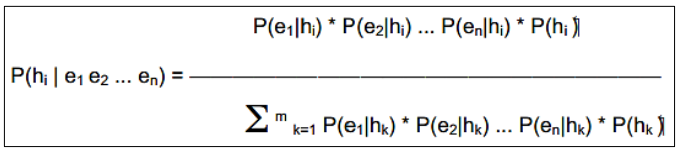
Los clasificadores Bayesianos son clasificadores estadísticos, cuya tarea es predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular. La clasificación Bayesiana tiene sus cimientos en el teorema de Bayes, y se ha comprobado que los clasificadores Bayesianos poseen una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos [29,30].

Diferentes estudios comparando los algoritmos de clasificación han determinado que un clasificador Bayesiano sencillo conocido como el clasificador *"Naive Bayes"* es comparable en rendimiento a un árbol de decisión y a clasificadores de redes de neuronas [29,30].

A continuación se explica el más conocido de los clasificadores bayesianos el clasificador *Naive Bayes*.

- **Clasificador Naive Bayes**

La simplificación propuesta por Naive Bayes es que las evidencias son independientes unas de otras, es decir, no están condicionadas entre sí. Además, supone que las hipótesis son disjuntas, es decir que son excluyentes entre si y que cubren todas las posibles hipótesis. El teorema de Bayes simplificado en Naive Bayes según la *hi* aplicado a un ejemplo con n evidencias y k hipótesis se escribe así (ver figura 9) [30,32].



$$P(h_i \mid e_1\, e_2 \ldots e_n) = \frac{P(e_1|h_i) * P(e_2|h_i) \ldots P(e_n|h_i) * P(h_i)}{\sum_{k=1}^{m} P(e_1|h_k) * P(e_2|h_k) \ldots P(e_n|h_k) * P(h_k)}$$

Figura 9: Teorema de Bayes.

Naive Bayes tiene como dificultad que si hay escasos datos, es posible que alguna probabilidad sea 0, porque no ha aparecido un determinado valor de un atributo para una cierta clase [30,32].

Esto se mejoró con la implementación del algoritmo *Naive Bayes m-estimado.* Se calcula un m-estimado de la probabilidad *(con una formula)* y el atributo considerado toma un cierto valor, m es una constante denominada “tamaño equivalente de muestra” [30].

2.2.5 Reglas de Inducción

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.

Estas reglas pueden ser representadas de muchas maneras, incluyendo árboles de decisión y reglas modulares. Los métodos de inducción de reglas evalúan los atributos del conjunto de entrenamiento y deciden cuales usar para discriminar entre las diferentes clases.

- **JRip**

Basado en el algoritmo de poda incremental de error reducido (RIPPER Repeated Incremental Pruning to Produce Error Reduction) [33], 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. Una vez encontrada dicha regla y que sea la más eficiente para ese ejemplo, es asignado con una etiqueta de valor de salida [34,36].

- **PART**

Utiliza la estrategia divide y vencerás en la cual se construye una regla, se remueven las instancias cubiertas por ella, y continúa creando reglas recursivamente para las instancias restantes hasta que no quede ninguna. Para elaborar una regla se construye un árbol de decisión podado para el conjunto de instancias en cuestión, se toma la hoja que cubra el mayor número de instancias y se transforma en regla, y se descarta el árbol [35].

- **Ridor**

Este algoritmo genera primero una regla por defecto y luego toma las excepciones para la regla por defecto con la misma 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 [79].

- **Decision Table**

Este algoritmo construye una tabla a partir de los datos de entrenamiento por un subconjunto llamado “esquema” de sus atributos y una selección de las instancias para entrenamiento. Para clasificar una instancia el algoritmo busca en la tabla todos los ejemplos que coinciden, teniendo en cuenta sólo los atributos que forman el esquema. Si no se encuentra un ejemplo que coincida, el algoritmo devuelve la clase más cercana de la tabla; en otro caso, devuelve la clase mayoritaria del conjunto de ejemplos que coincidieron [80].

- **DTNB**

Este algoritmo es un hibrido basado en el algoritmo Decisión 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 la tabla de decisiones, el otro para Naive Bayes. En cada paso los atributos seleccionados son modelados por Naive Bayes y el resto por el Decision Table, y todos los atributos son modelados por la tabla de decisión inicial.

En cada paso, el algoritmo considera también eliminación de un atributo completamente del modelo [81].

- **NNge**

Este algoritmo es un híbrido entre los algoritmos basados en instancias y los de inducción de reglas. Aprende de forma incremental, primero clasificando y luego generalizando cada nuevo ejemplo. La generalización consiste en agrupar la nueva instancia con el ejemplar de la misma clase más próximo. Si el ejemplar más próximo es un ejemplo aislado se crea un hiperrectángulo que contiene a ambos. De lo contrario, si el ejemplar más próximo era un hiperrectángulo, este crece para abarcar el nuevo ejemplo. Los hiperrectángulos se describen mediante reglas “si - entonces”. Para determinar el vecino más cercano se aplica la función de distancia Euclideana modificada capacitada para manejar hiperrectángulos [82].

3. CAPÍTULO III: HERRAMIENTAS ÚTILES EN MINERÍA DE DATOS.

Existen algunas herramientas diseñadas para extraer conocimientos desde bases de datos que contienen grandes cantidades de información [37,38]. La minería de datos es una técnica compuesta por fases la cual integra varias áreas y no se la debe confundir con un gran software [39].

En el desarrollo de un proyecto de este tipo se utilizan diferentes aplicaciones de software en cada fase, estas pueden ser de estadística [40], visualización de datos [41] y principalmente de inteligencia artificial [42].

En la actualidad existen aplicaciones o herramientas comerciales de minería de datos muy completas que permiten extraer conocimientos desde bases de datos y a su vez contienen un sin número de utilidades que facilitan el desarrollo de un proyecto. Sin embargo, en casi todo desarrollo de proyecto suelen complementarse con otras herramientas.

Algunas de las herramientas más populares son [43]: Orange, RapidMiner, Knime, Weka y R.

3.1 Orange

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 librerías para Python y cuenta con bibliotecas de secuencias de comando [44].

Este software está enfocado en las personas que no tienen mucha experiencia en minería de datos [47]. Utiliza la metodología de SEMMA de SAS para organizar sus procedimientos [45,46]. Una desventaja de esta herramienta es el limitado acceso a los datos, ya que no cuenta con muchas herramientas para esto [47].

Características:

- Diseñar procesos de análisis de datos a través de programación visual [47].

- Ofrece diferentes visualizaciones, de diagramas de dispersión, gráficos de barras, los árboles, a dendrogramas, redes y mapas de calor [47].
- Se puede interactuar sin problemas a través del esquema de análisis de datos [47].
- También están disponibles complementos especializados, como Bioorange para la bioinformática [47].
- Además de desarrollar componentes propios y luego integrar al resto de Orange [47].
- Orange es de código abierto con la comunidad activa [47].
- Se ejecuta en Windows, Mac OS X, y la variedad de sistemas operativos Linux [47].

3.2 RapidMiner

RapidMiner, antiguamente llamado YALE es un entorno para aprendizaje computacional y minería de datos que se utiliza para áreas de minería de datos tanto en investigación como en el mundo real [48,49].

Permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML y trabajan con la interfaz gráfica de usuario de RapidMiner. Esta herramienta contiene más de 500 operadores para todos los principales procedimientos de máquina de aprendizaje, y además combina esquemas de aprendizaje y evaluadores de atributos de la herramienta Weka [49].

Se puede utilizar como una herramienta stand-alone para el análisis de datos y como motor para minería de datos que a su vez puede integrarse en productos propios [50].

Características:

- RapidMiner es un sistema prototipado para el descubrimiento del conocimiento y Data Mining [51].
- Es un software de tipo Open-Source con licencia GNU GPL, basado en java [51].
- Trabaja bajo las plataformas Windows y Linux [78]. Posee alrededor de 500 operadores que pueden ser combinados [49, 51].
- Usa el lenguaje de scripting XML para describir los operadores y su configuración [51].

- La característica más importante es la capacidad de jerarquizar cadenas del operador y de construir complejos árboles de operadores [51].
- El lenguaje de encriptación permite automáticamente una gran cantidad de experimentos [51].
- Posee una interfaz gráfica, línea comando, y API de Java para usar RapidMiner desde tus propios programas [51].
- Una gran cantidad de extensiones (plugins) [51].
- Las aplicaciones incluyen: Text Mining, Multimedia Mining, entre otras [51].

3.3 Knime

KNIME (Konstanz Information Miner) es un entorno gratuito de código abierto y fácil uso para el proceso y ejecución de técnicas de minería de datos [52].

Esta herramienta ofrece a los usuarios la posibilidad de crear de forma visual flujos o tuberías de datos, ejecutar parcialmente o todos los pasos de análisis, y posteriormente examinar los resultados [53]. KNIME está escrito en Java y hace uso de sus métodos de extensión para soportar plugins [52].

A través de plugins, existe la posibilidad de que los usuarios pueden agregar módulos de texto, imagen, procesamiento de series de tiempo y la integración de diversos proyectos de código abierto, por ejemplo el lenguaje de programación R, y la herramienta WEKA [52,54].

Características:

- KNIME está desarrollado en la plataforma Eclipse y programado en java.
- Creación de modelos estadísticos y de minería de datos, así como árboles de decisión, regresiones [53].
- Está concebido como una herramienta gráfica y dispone de una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva [53].
- Validación de modelos [53].
- Scoring o aplicación de dichos modelos sobre conjuntos nuevos de datos [53].

- Al ser de código abierto la herramienta permite su extensión mediante la creación de nuevos nodos que implementan algoritmos a la medida del usuario [53].
- La interfaz gráfica de usuario permite el montaje fácil y rápido de nodos para procesamiento de datos (ETL: extracción, transformación, carga), para el análisis de datos, modelado y visualización [53].

3.4 Weka

Es un conjunto de librerías java para la extracción de conocimientos desde bases de datos [55]. Es una herramienta desarrollada bajo licencia GPL lo cual ha permitido que sea una de las más utilizadas en los últimos años en el área de la Minería de datos [56,57].

Una de las propiedades más interesantes de este software, es su facilidad para añadir extensiones, modificar métodos, entre otros [55, 57].

Características:

- Está disponible bajo la licencia pública general de GNU [57,58].
- Es muy portable porque está completamente desarrollado en lenguaje Java y puede ejecutarse en cualquier plataforma [57, 58].
- Contiene una amplia colección de técnicas para el pre procesamiento de datos y modelado [57].
- Facilidad de uso por un principiante gracias a su interfaz gráfica de usuario [56, 58].
- Diversas fuentes de datos (ASCII, JDBC) [57].
- Interfaz visual basada en procesos / flujos de datos (rutas) [58].
- Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas (T-test) [58].

3.5 R

R es un software gratuito (GNU), para el análisis estadístico y gráfico el cual proporciona una amplia gama de técnicas y que se desarrolla a través de paquetes compartidos por los usuarios y para los usuarios [59].

Proporciona una amplia variedad de estadística (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, etc.) y las técnicas gráficas, por lo que es altamente extensible [59,60].

Características:

- Almacenamiento y manipulación efectiva de datos [60].
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices [60].
- Amplia, coherente e integrada colección de herramientas para análisis de datos [60].
- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora [61].
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R) [62].

e. Materiales y Métodos

La identificación de los factores de deserción y reprobación se lo realizó con la utilización de la herramienta Rapid Miner y el administrador de bases de datos DatAdmin, para obtener la información necesaria en el presente Trabajo de Titulación se utilizó el Web Services del Sistema de Gestión Académica (S.G.A) de la Universidad Nacional de Loja, el cual contiene datos: académicos, personales de estudiantes como docentes y además se recopiló datos del Área de Bienestar Universitario.

En este apartado se realizó un análisis de las características más relevantes de las metodologías de Minería de Datos y con una tabla comparativa elegir la más apropiada para el presente proyecto.

- **Metodología CRISP-DM**

La metodología CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, actualmente es la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining [63].

El proceso se estructura en seis fases o etapas (ver figura 10), a continuación se muestra la estructura.

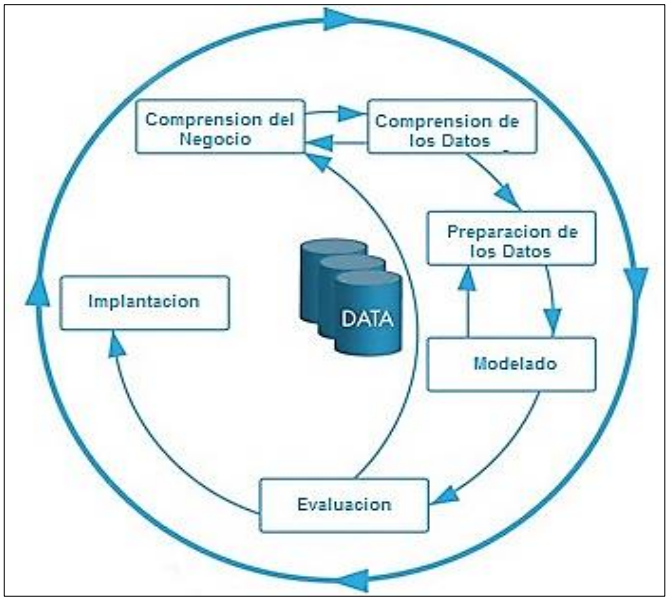


Figura 10: Estructura de procesos en CRISP-DM.

La primera etapa es *Comprensión del Negocio*: esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego este conocimiento de los datos es convertido en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos [64,65].

Comprensión de los Datos: esta fase empieza con la colección de datos inicial y continúa con las actividades que le permiten familiarizarse primero con los datos, estudiar la calidad de datos, descubrir los primeros conocimientos en los datos o descubrir subconjuntos interesantes para luego formar hipótesis en cuanto a la información oculta [64,65].

Preparación de los Datos: esta fase envuelve todas las actividades necesarias para construir el conjunto de datos final, es decir los datos que se manipularan en las herramientas de modelado a partir de los iniciales [65], estas actividades probablemente van a ser ejecutadas muchas veces y no en un orden establecido.

En la fase de *Modelado* distintas técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos, además pueden existir varias técnicas para el mismo tipo de problema de minería de datos, algunas técnicas tienen requerimientos específicos sobre la forma de datos es por ello que volver a la fase de preparación de datos es a menudo necesario.

La fase de *Evaluación* consiste en la construcción de un modelo (o modelos) con alta calidad de una perspectiva de análisis de datos, antes del empezar el despliegue final del modelo es importante evaluar detalladamente el mismo y revisar los pasos ejecutados para crearlo, para comparar el modelo obtenido con los objetivos de negocio [64,65].

En la fase de *Implantación* la creación del modelo no es generalmente el final del proyecto, incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado de manera comprensible para el usuario [64,65].

En la metodología CRISP DM la sucesión de fases no es necesariamente rígida, es decir cada fase es estructurada en varias áreas generales de segundo nivel. Las áreas generales se proyectan a áreas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas [65].

- **Metodología de Berry y Linoff**

Esta metodología propuesta por Michael Berry y Gordon Linoff para el desarrollo de proyecto de minería de datos, se basa en etapas que son: Traducir el problema de negocio a un problema de minería de datos; seleccionar los datos adecuados; conocer los datos; crear un model set; solucionar problemas con los datos; transformar datos; construir modelos; evaluar modelos, desplegar modelos y evaluar resultados [66,67]. Estas etapas se pueden agrupar en las siguientes fases (ver figura 11).

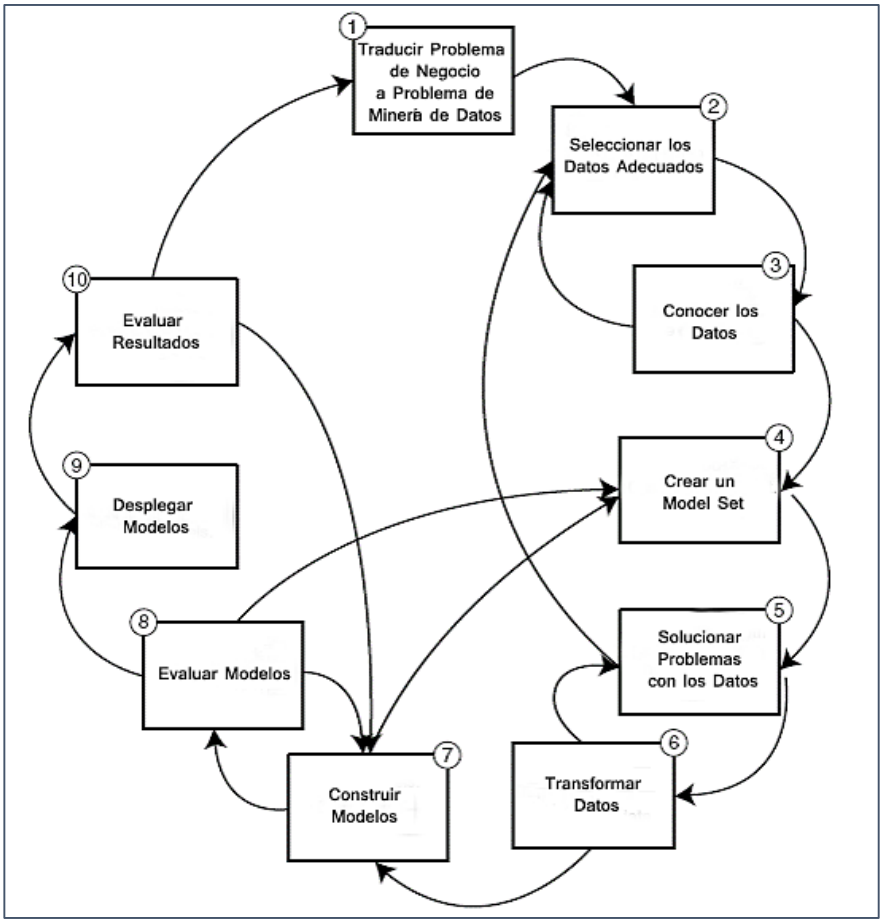


Figura 11: Etapas de la Metodología de Berry y Linoff.

Filtrado de Datos: al obtener los datos es necesario recurrir a varias fuentes, en muchos casos el formato de los datos no es idóneo, y la mayoría de las veces no se puede aplicar algún algoritmo de minería. Por lo tanto es necesario realizar un pre procesamiento y filtrado los datos, ya sea con la eliminación de valores incorrectos, datos no válidos, datos desconocidos [66,67].

Selección de variables: esta fase se realiza con el objetivo de reducir la cantidad de datos, seleccionando las variables más relevantes para el problema, sin sacrificar la calidad del modelo obtenido. Los métodos para la seleccionar las características son dos: el primero basado en la elección de los mejores atributos del problema; el segundo basado en buscar variables independientes mediante pruebas de sensibilidad, algoritmos de distancia o algoritmos heurísticos [66,67].

Extracción de Conocimiento: El objetivo de esta fase es obtener un modelo de conocimiento, que represente patrones de comportamiento basados en los valores de cada variable del problema o relaciones entre variables [66,67].

Interpretación y Evaluación: en esta fase se realiza la validación, se comprueba que las conclusiones son válidas y satisfactorias. En caso de obtener varios modelos mediante el uso de diferentes técnicas, se comparan los modelos buscando el que se ajuste mejor al problema. Si ninguno de los modelos logra los resultados esperados, se vuelve a realizar alguno de los procesos anteriores para obtener mejores modelos [66,67].

- **Metodología SEMMA**

La metodología SEMMA fue desarrollada por SAS Institute para descubrir patrones de negocio desconocidos [68]. El nombre refiere a las cinco fases básicas del proceso (ver figura 12).

Figura 12: Fases de la metodología SEMMA.

Muestreo: se busca extraer una parte de los datos lo suficientemente grande para contener información significativa, pero reducida para manipular fácilmente. Si los patrones frecuentes aparecen, estos se pueden diferenciar en una muestra representativa. Si un conjunto de datos es tan pequeño que no se puede extraer una muestra, puede ser descubierto por medio de métodos de síntesis [68,69].

Explorar: en esta fase se explora los datos buscando tendencias y anomalías inesperadas para lograr una comprensión general de los mismos. Esta fase ayuda a refinar el proceso denominado descubrimiento [68,69].

Modificar: en esta fase se modifican los datos por medio de la creación, selección y transformación de variables, para concentrarse en el proceso de selección del modelo. Basado en los descubrimientos en la fase de exploración, se puede dar la necesidad de manipular los datos para incluir información como la de agrupamiento de compradores y subgrupos significativos, o introducir nuevas variables [68,69].

Modelar: en esta fase se modelan los datos permitiendo que el software busque automáticamente una combinación de datos que prediga con cierta certeza un resultado deseado [68,69].

Muestreo o Evaluación: en esta fase se califican los datos mediante la evaluación de la utilidad y fiabilidad de los resultados del proceso de minería de datos. Un método común de evaluación es aplicar el modelo a una porción separada de resultados obtenidos durante el muestreo. Si el modelo es válido, debería funcionar para esta muestra, así como para la muestra utilizada en la construcción del modelo [68,69].

Principalmente SEMMA es una organización lógica para el manejo de una herramienta funcional de SAS llamada *Enterprise Manager* para el manejo de tareas de minería de datos. SEMMA pretende llevar de una manera fácil la tarea de la exploración estadística y la visualización de técnicas, seleccionando y transformando las variables predictivas más importantes, modelándolas para obtener resultados, y finalmente confirmar la precisión del modelo [69,70].

El ciclo de desarrollo de la metodología SEMMA se puede apreciar a continuación (ver figura 13).

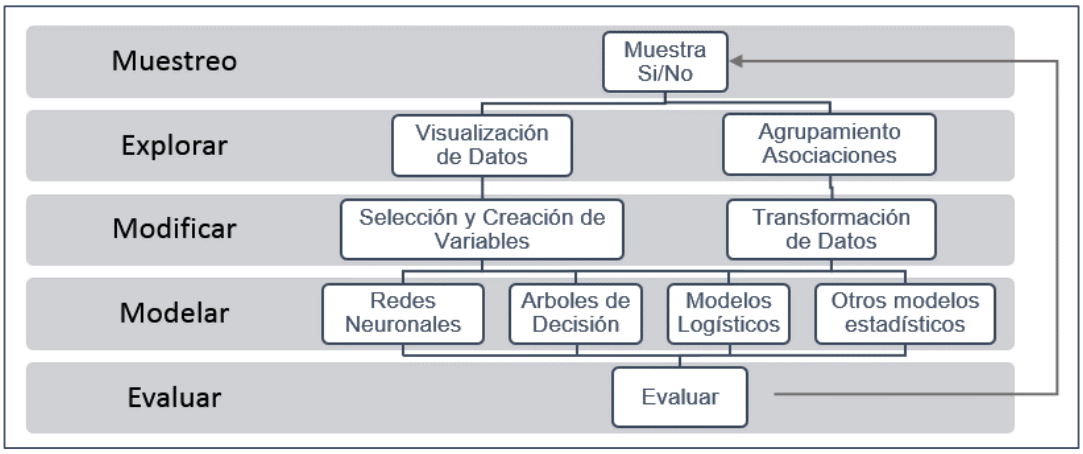


Figura 13: Ciclo de desarrollar en la metodología SEMMA.

- **Tabla Comparativa de Metodologías**

Para la ejecución ordenada de los proyectos de Minería de Datos, se han tomado en cuenta características o especificaciones sobre los procesos o guías de modelado, con el propósito de seleccionar la más idónea para el presente Trabajo de Titulación, a continuación se describen las características más relevantes de cada metodología (ver tabla XIII):

TABLA XIII:
CARACTERÍSTICAS DE METODOLOGÍAS PARA MINERÍA DE DATOS.

Características	CRISP-DM	BERRY Y LINOFF	SEMMA
Implica Retroalimentación y es Cíclica	Si	Si	Si
Aplicación en Herramientas	Libre y Gratuita [63-65].	Libre y Gratuita [66,67]	Ligada a productos SAS [69,70].
Fases o Etapas	Fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, implantación [63-65].	Etapas: traducir el problema de negocio a un problema de minería de datos, seleccionar los datos adecuados, conocer los datos, crear un model set, solucionar problemas con los datos, transformar datos, construir modelos, evaluar modelos, desplegar modelos y evaluar resultados [66,67].	Fases: Muestreo, Explorar, Modificar, Modelar, Evaluar [68,69].
Distribución	Libre [63]	Libre [66,67]	Distribución en clientes SAS [69].

Tipo	Metodología	Metodología	Secuencia Lógica
Orientación o Enfoque	Enfocada a resultados empresariales [70,63].	Enfocada a resultados de proceso y comerciales [66,67].	Enfocada a resultados de proceso [70].

Las metodologías más conocidas son CRISP-DM y SEMMA, porque estructuran el proyecto de Minería de Datos en fases o etapas. Al comparar las metodologías el primer aspecto en analizar es la perspectiva, donde las metodologías de Berry-Linoff y CRISP-DM, tienen en cuenta una perspectiva más amplia, ya que toman en cuenta los objetivos empresariales del proyecto mientras que la metodología SEMMA se centra en características de técnicas del desarrollo como muestreo de los datos.

Otra característica que diferencia a las metodologías mencionadas es la relación con herramientas comerciales. La metodología SEMMA por ejemplo está ligada con productos SAS en las cuales se encuentra implementada. La metodología de Berry y Linoff no depende de una herramienta para su utilización y la metodología CRISP-DM también es una metodología libre y gratuita que no depende de la herramienta que se utilice.

Las metodologías CRISP-DM y SEMMA son bastante sólidas, pero en la práctica la metodología SEMMA tiene un alcance de resultados más reducida que CRISP-DM, debido a que SEMMA trabaja perfectamente cuando se tiene un sistema SAS, el cual es muy popular en empresas grandes.

La metodología CRISP-DM se ajusta mejor a los parámetros de la Minería de Datos, y a los procesos que una empresa realiza al trabajar con sus datos. Además esta metodología es ampliamente utilizada en ambientes académicos y en la mayoría de proyectos Data Mining [63], según la plataforma líder de información de Minería de Datos y Descubrimiento de Conocimiento en los EE.UU *kdnuggets.com,* la siguiente figura (ver figura 14) representa el resultado obtenido en sucesivas encuestas efectuadas durante los últimos años, respecto del grado de utilización de las principales guías de desarrollo de proyectos de Data Mining [65,78].

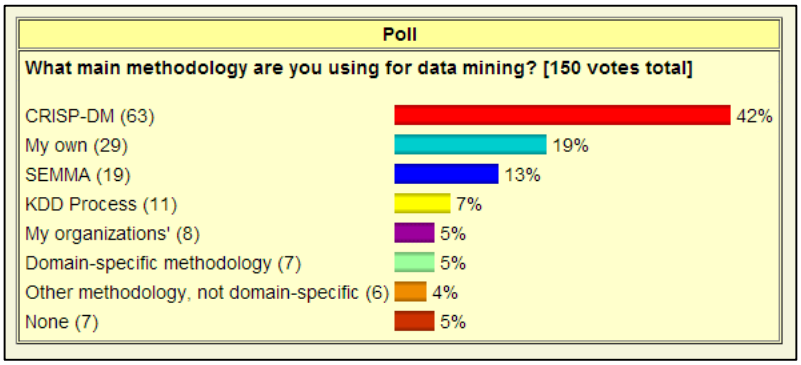


Figura 14: Encuesta de metodologías más utilizadas según kdnuggets.

Como conclusión y en base a lo antes mencionado, en la presente investigación se determinó utilizar la metodología CRISP-DM ya que permitió cumplir con los objetivos planteados, por su amplia utilización y principalmente por ser la más utilizada en ambientes académicos [63].

Esta metodología nos permitió llevar un desarrollo ordenado y sobre todo alcanzable de acuerdo a los parámetros propuestos en el presente Trabajo de Titulación.

f. Resultados

En el presente apartado se detalla los resultados obtenidos en cada etapa, los mismos que serán de importancia para verificar el cumplimiento de la hipótesis planteada.

HIPÓTESIS

La minería de datos permite identificar los factores de deserción y reprobación universitaria, en base a información personal, académica e institucional.

Por lo mencionado anteriormente se describe los aspectos abordados a lo largo de cada uno de los objetivos desarrollados.

1. ETAPA UNO: Análisis y Muestreo de los Datos existentes en las bases de datos de la Universidad Nacional de Loja para su procesamiento.

En la presente etapa se realizó un análisis de la información almacenada en las bases de datos del Web Services del Sistema de Gestión Académica y además información del Área Bienestar Universitario de la Universidad Nacional de Loja con el propósito examinar que datos serán necesarios para el desarrollo del presente Trabajo de Titulación.

1.1. Realizar entrevistas y solicitar autorización con el responsable de la Unidad de Telecomunicaciones e Información (UTI), en cuanto a manipulación de los datos y acceso a los mismos.

Se realizó reuniones con el director de la Unidad de Telecomunicaciones e Información sobre el desarrollo del presente Trabajo de Titulación por lo cual se obtuvo un certificado de acceso a los datos a través del Web Services del Sistema de Gestión Académica de la Universidad Nacional de Loja (ver anexo 1).

1.2. Análisis de la fuente de datos Web Services de la Universidad Nacional de Loja.

En esta actividad se realizó un análisis acerca de los datos existentes en la Universidad Nacional de Loja, a través del Web Services (ver figura 15) del Sistema de Gestión Académica de la Universidad Nacional de Loja, este proporciona una biblioteca de métodos que facilitan a obtención de datos personales académicos y estadísticos que se han generado durante la vigencia del sistema principal de la institución, la misma que es administrada por la Unidad de Telecomunicaciones.

Figura 15: Pagina del Web Services del Sistema de Gestión Académica de la Universidad Nacional de Loja.

El Sistema de Gestión Académico de la Universidad Nacional de Loja, posee una amplia cantidad de información, la cual puede ser aprovechada por aplicaciones con fines académicos, por esta razón la biblioteca de métodos o servicios *SGAWebServices* ofrece distintas funciones, utilizando la tecnología de Web Services.

En cuanto a los servicios estos están agrupados por categorías (ver tabla XIV) de acuerdo a la información que estos retornan.

TABLA XIV:
SERVICIOS AGRUPADOS EN CATEGORÍAS.

Categoría	Descripción
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.

Para obtener los datos de las categorías antes mencionadas se utilizan archivos *wsdl* este es un formato *xml* donde se describen los métodos del Web Services, es la interfaz pública de los mismos.

Estos indican la forma de comunicación, es decir, los requisitos del protocolo y los formatos de los mensajes necesarios para interactuar con los servicios listados en su catálogo. Posteriormente un programa cliente que se conecta a un servicio web puede leer el WSDL para determinar qué funciones están disponibles en el servidor. A continuación se muestran los métodos o servicios de cada categoría descrita anteriormente.

- **Categoría Académica**

La categoría Académica contiene los siguientes servicios (ver tabla XV) para obtener los datos de los estudiantes, docentes y de la Universidad Nacional de Loja.

TABLA XV:
SERVICIOS DE LA CATEGORÍA ACADÉMICA.

Método	Descripción
sgaws_periodos_lectivos:	Retorna una lista de períodos lectivos.
sgaws_ofertas_academicas:	Retorna una lista de ofertas académicas en un periodo lectivo.

sgaws_fechas_matriculaoa:	Retorna las fechas de matrículas ordinaria, extraordinaria y especial de una oferta académica.
sgaws_estadoestudiantes_paralelo:	Retorna la lista de estudiantes de un paralelo con el estado de matrícula (aprobada, reprobado, matriculada), de acuerdo a la oferta académica, módulo, carrera y paralelo.
sgaws_plan_estudio:	Retorna el plan de estudio asignado o un paralelo en una oferta académica.
sgaws_carga_horaria_docente:	Retorna la carga horaria asignada a un docente en una oferta académica.
sgaws_notas_estudiante:	Retorna las notas de un estudiante en cada unidad, la carrera, módulo, paralelo al que pertenece y estado de la matrícula en la oferta académica.
sgaws_carreras_estudiante:	Retorna los datos de un estudiante y las carreras en las que está matriculado.
sgaws_reporte_matricula:	Retorna los datos de un estudiante, oferta académica, nota y porcentaje de asistencia y estado de la matrícula del estudiante en esa oferta.

- **Categoría Institucional**

La categoría Institucional contiene los siguientes servicios (ver tabla XVI) los cuales permiten obtener los datos en cuanto a estructura de la Universidad Nacional de Loja.

TABLA XVI:
SERVICIOS DE LA CATEGORÍA INSTITUCIONAL.

Método	Descripción
sgaws_lista_areas	Retorna una lista de las áreas que conforman la universidad.
sgaws_datos_area	Retorna los datos de las siglas del área ingresada.
sgaws_carreras_area	Retorna las carreras que forman parte de un área.
sgaws_datos_carreras	Retorna una lista de carreras con sus datos en una oferta académica.
sgaws_modulos_carrera	Retorna los datos de todos los módulos de una carrera en una oferta académica.
sgaws_paralelos_carrera	Retorna todos los paralelos de una carrera en una oferta académica.

- **Categoría Personal**

La categoría Personal se encuentran los siguientes servicios (ver tabla XVII) los cuales permiten obtener los datos personales de estudiantes, docentes de la Universidad Nacional de Loja.

TABLA XVII:
SERVICIOS DE LA CATEGORÍA PERSONAL.

Método	Descripción
sgaws_datos_docente:	Retorna los datos o información del docente, que corresponden a la cédula ingresada.
sgaws_datos_estudiante:	Retorna los datos o información del estudiante, que corresponden a la cédula ingresada.
sgaws_datos_usuario:	Retorna los datos del usuario ingresado e indica el tipo de usuario estudiante / docente o ambos que corresponden a la cédula ingresada.

- **Categoría Validación**

La categoría Validación contiene servicios (ver tabla XVIII), los cuales permiten validar a docentes como a estudiantes de la Universidad Nacional de Loja.

TABLA XVIII:
SERVICIOS DE LA CATEGORÍA VALIDACIÓN.

Método	Descripción
sgaws_validar_docente:	Retorna Verdadero o Falso si el docente está en la base de datos del Sistema de Gestión Académica.
sgaws_validar_estudiante:	Retorna Verdadero o Falso si el estudiante está en la base de datos del Sistema de Gestión Académica.

- **Categoría Estadística**

La categoría Estadística contienen servicios (ver tabla XIX), que permiten obtener información estadística como el número de estudiantes matriculados, número de estudiantes aprobados y número de estudiantes reprobados en los diferentes periodos académicos que ofrece la Universidad Nacional de Loja.

TABLA XIX:
SERVICIOS DE LA CATEGORÍA ESTADÍSTICA.

Método	Descripción
sgaws_nmatriculados_oferta:	Devuelve el número de estudiantes matriculados en una oferta académica.
sgaws_nmatriculados_area:	Devuelve el número de estudiantes matriculados en un área.
sgaws_nmatriculados_carrera:	Devuelve el número de estudiantes matriculados en una carrera.
sgaws_nmatriculados_modulo:	Devuelve el número de estudiantes matriculados en un módulo
sgaws_nmatriculados_paralelo:	Devuelve el número de estudiantes matriculados en un paralelo.
sgaws_naprobados_oferta:	Devuelve el número de estudiantes aprobados en una oferta académica.
sgaws_naprobados_area:	Devuelve el número de estudiantes aprobados de un área en una oferta académica.
sgaws_naprobados_carrera:	Devuelve el número de estudiantes aprobados en una carrera en una oferta académica.
sgaws_naprobados_modulo:	Devuelve el número de estudiantes aprobados en un módulo en una oferta académica.
sgaws_naprobados_paralelo:	Devuelve el número de estudiantes aprobados en un paralelo en una oferta académica.
sgaws_nreprobados_oferta:	Devuelve el número de estudiantes reprobados en una oferta académica.
sgaws_nreprobados_area:	Devuelve el número de estudiantes reprobados en un área y en una oferta académica.
sgaws_nreprobados_carrera:	Devuelve el número de estudiantes reprobados en una carrera en una oferta académica.
sgaws_nreprobados_modulo:	Devuelve el número de estudiantes reprobados en un módulo en una oferta académica.
sgaws_nreprobados_paralelo:	Devuelve el número de estudiantes reprobados en un paralelo en una oferta académica.

En base a un análisis a los datos almacenados en el Sistema de Gestión Académico de la Universidad Nacional de Loja, se observó que la información almacenada hasta la fecha actual se encuentra en una base de datos centralizada, a continuación se describen algunos detalles de la misma.

En la categoría Académica se obtuvo información en cuanto a datos de estudiantes y docentes. La siguiente figura (ver tabla XX), muestra información acerca de 11 periodos académicos que han transcurrido desde el 2003 hasta el momento.

Cabe recalcar en cuanto a notas académicas y datos personales de estudiantes almacenados en el Sistema de Gestión Académico de la Universidad Nacional de Loja son desde el año 2008 hasta la actualidad ya que este fue el año en el que se implementó, los datos almacenados de años anteriores corresponden a datos históricos.

TABLA XX:
PERIODOS ACADÉMICOS DE LA UNIVERSIDAD NACIONAL DE LOJA.

Número de Periodo	Descripción
1	2003 - 2004
2	2004 - 2005
3	2005 - 2006
4	2006 - 2007
5	2007 - 2008
6	2008 - 2009
7	2009 - 2010
8	2010 - 2011
9	2011 - 2012
10	2012 - 2013
11	2013 - 2014

La Universidad Nacional de Loja ofrece una número de dos ofertas académicas por año y los datos almacenados que se han obtenido son desde el Marzo 2004 hasta la fecha con un total de 49 ofertas académicas (ver figura 16) (ver anexo 2).

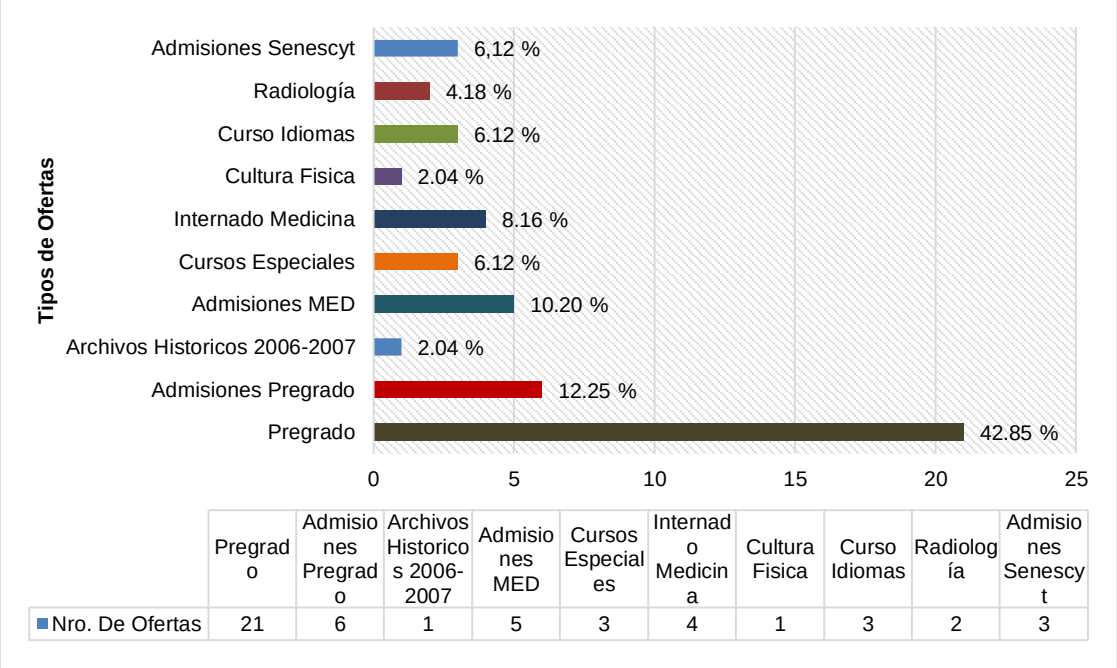


Figura 16: Ofertas académicas en la Universidad Nacional de Loja.

De las 49 ofertas académicas se identificaron 10 tipos almacenadas Sistema de Gestión Académico de la Universidad Nacional de Loja, las cuales son: tres cursos de admisiones Senescyt, dos cursos de Radiología, tres cursos de Idiomas, uno de cultura Física, cuatro de internado Medicina, tres de cursos Especiales, cinco de admisiones MED, seis de admisiones Pregrado, 21 cursos de Pregrado y también se observó una oferta con la descripción Archivos Históricos 2006-2007(ver anexo 2).

Con respecto a los estudiantes que han pasados por el Área de Energía las Industrias y los recursos Naturales No renovables de la Universidad Nacional de Loja se observa que existen en registros un número de 11418 (ver figura 17) (ver anexo 2).

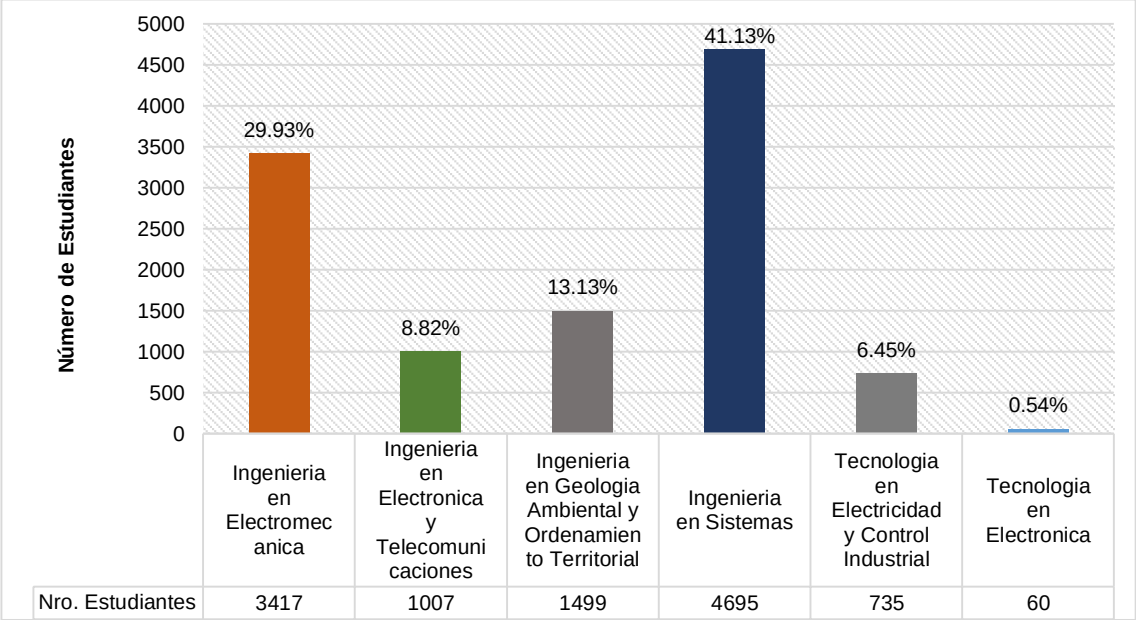


Figura 17: Estudiantes del Área de Energía de la Universidad Nacional de Loja

De los registros mencionados anteriormente, del Área de la Energía se observa que 3417 es decir el 29.93% corresponde a la Carrera Ingeniería en Electromecánica, con 1007 o el 8.82% corresponden a la Carrera Ingeniería en Electrónica y Telecomunicaciones, con 1499 o el 13.13% corresponden a la Carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial, con 735 o 6.45% corresponden a la Tecnología en Electricidad y Control Industrial, con 60 o 0.54% corresponden a la Tecnología en Electrónica y por último la mayor parte de estudiantes corresponden a la Carrera de Ingeniería en Sistemas con 41.13% o 4695 estudiantes. De los estudiantes antes mencionados se obtuvo las notas de materias que han cursado obteniendo un total aproximado de 48487 registros (ver anexo 2).

La Universidad Nacional de Loja se encuentra actualmente dividido con ocho áreas (ver tabla XXI), dicha información fue obtenida a través de los métodos de la categoría institucional.

TABLA XXI:
ÁREAS DE LA UNIVERSIDAD NACIONAL DE LOJA.

Área	Descripción
AARNR	Área Agropecuaria y de Recursos Naturales Renovables.
ACE	Cursos Especiales.
AEAC	Área de Educación, el Arte y la Comunicación.
AEIRNNR	Área de la Energía, Industrias y Recursos Naturales No Renovables.
AJSA	Área Jurídica, Social y Administrativa.
ASH	Área de la Salud Humana.
MED	Modalidad de Estudios a Distancia.
PREUNIVERSITARIO	Preuniversitario.

La categoría institucional contiene información acerca de 142 carreras que se han ofertado hasta el momento en la Universidad Nacional de Loja (ver figura 18, ver anexo 2).

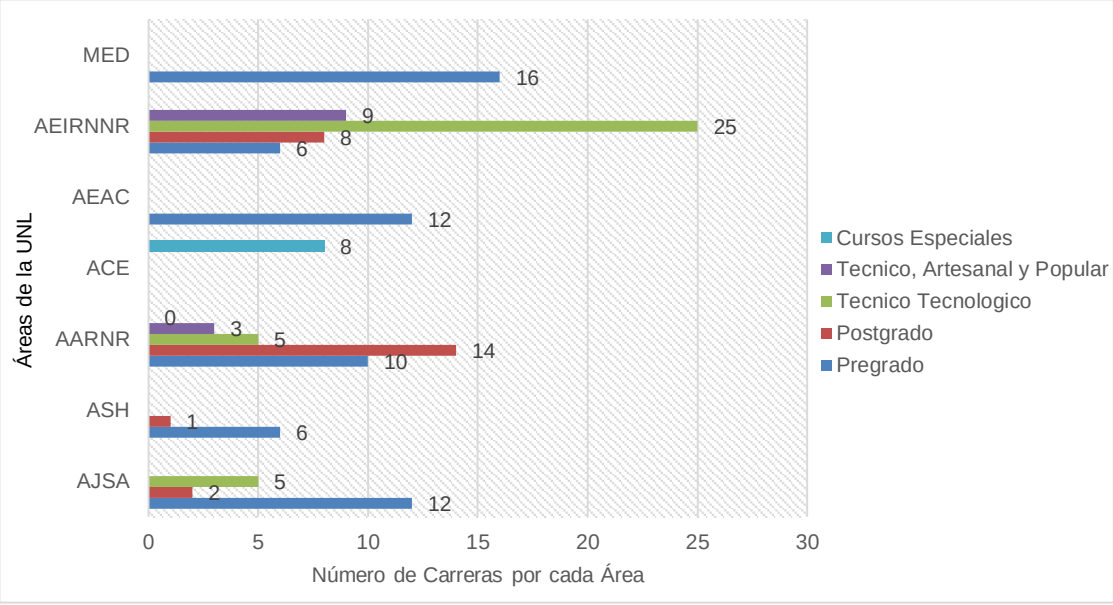


Figura 18: Carreras que se ofertan en la Universidad Nacional de Loja.

Las carreras que ofrecen cada una de las Áreas de la Universidad son: MED con 16 carreras de Pregrado que corresponde al 11.26%, el área AEIRNNR se ofrecen un total de 48 carreras de las cuales nueve son de Técnico, Artesanal y Popular, ocho de Postgrado, seis de Pregrado y 25 de Técnico Tecnológico y corresponden al 33.80% de todas las carreras, el área AEAC ofrece solamente 12 carreras de Pregrado que corresponde al 8.45% de todas las carreras, el área ACE corresponden a los cursos especiales que son un total de ocho y corresponden al 5.63%, el área AARNR ofrece un total de 32 carrera que corresponden al 22.54% de las cuales 14 son de Postgrado, 10 son de Pregrado, tres de Técnico, Artesanal y Popular y cinco de Técnico Tecnológico, el área ASH ofrece un total de siete carreras de las cuales, 6 son de Pregrado y una de Postgrado y corresponde al 4.93% de todas las carreras, el área AJSA ofrece un total de 19 carreras que corresponde al 13.38% de las cuales, dos son de Postgrado, 12 son de Pregrado y cinco de Técnico Tecnológico. En la siguiente figura (ver figura 16) se muestra la distribución de carreras en las diferentes modalidades de estudio (ver anexo 2).

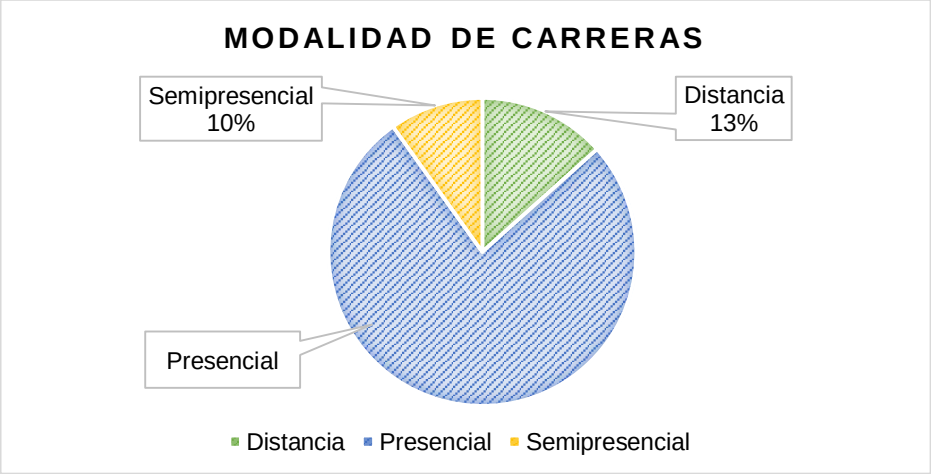


Figura 19: Modalidades de estudio en la Universidad Nacional de Loja.

De las 142 carreras mencionadas anteriormente con el 10% es decir 14 carreras son de modalidad semipresencial, con el 13% es decir 19 carreras son de modalidad a distancia y con el 77% es decir 109 carreras son de modalidad presencial (ver anexo 2).

Se obtuvo información acerca de los módulos que se ofertan actualmente para las carreras de la Universidad Nacional de Loja y se obtuvo un total de 468 módulos y la descripción de los mismos (ver anexo 2).

El número de paralelos que se han generado desde la implementación del sistema y que se ofertan en cada periodo hasta el momento son un aproximado de 12695 (ver anexo 2). Estos corresponden a todas las carreras de la Universidad Nacional de Loja.

La categoría Personal ofrece información personal acerca de los estudiantes y docentes que pertenecen a la Universidad Nacional de Loja. Los datos obtenidos en la (ver anexo 3) pertenecen a estudiantes de las carreras: Ingeniería Electromecánica, Ingeniería en Geología Ambiental y Ordenamiento Territorial, Ingeniería en Sistemas, Ingeniería en Electrónica y Telecomunicaciones que pertenecen al Área de la Energía las Industrias y los recursos Naturales no Renovables de la Universidad Nacional de Loja. Los datos obtenidos que se obtuvo son: nombres, apellidos, fecha de nacimiento, teléfono, ciudad, país, email, género y una variable para establecer si ya egreso, el número de estudiantes que cursan actualmente en el área antes mencionadas son de 3438.

En la categoría Validación se encuentran métodos que permiten verificar si un docente o estudiante está en los registros de la base de datos de la Universidad Nacional de Loja, por tal motivo no se realizó ninguna observación acerca de esta categoría.

Con respecto a la categoría Estadística se pudo obtener información acerca de estudiantes matriculados, aprobados y reprobados de todas las Áreas, Carreras y Módulos que se ofertan en la Universidad Nacional de Loja.

A continuación se describen los datos analizados de esta categoría:

En la siguiente figura (ver figura 20) se muestran los estudiantes que han aprobado y reprobado a lo largo de todos los periodos académicos desde Septiembre de 2009 hasta la actualidad. Los datos pertenecen al Área de Cursos Especiales (ver anexo 2).

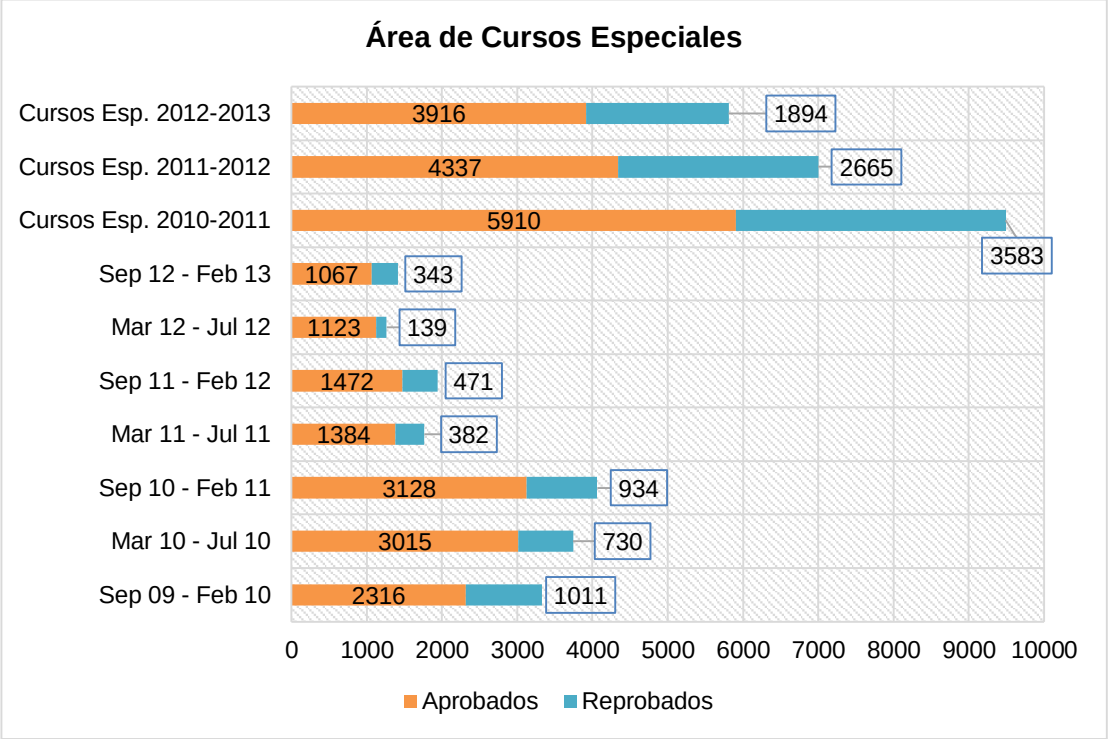


Figura 20: Estudiantes que aprobaron y reprobaron en el Área ACE.

En la figura anterior se puede observar que el periodo donde más estudiantes reprobaron es en el periodo de Cursos Especiales 2011 – 2012, con una cantidad notable de 2665 estudiantes reprobados que corresponde al 36.11% de estudiantes que se matricularon en ese periodo. Mientras que el periodo que reporta con menos reprobados es el periodo Marzo 2012– Julio 2012 con una cantidad de 139 estudiantes que corresponde al 10.84% de estudiantes matriculados en ese periodo.

En la siguiente figura (ver figura 21) muestra los estudiantes que han aprobado y reprobado en las carreras del Área Agropecuaria y de Recursos Naturales Renovables los datos existentes tienen origen desde el periodo Septiembre 2008 – Febrero 2009 (ver anexo 2)..

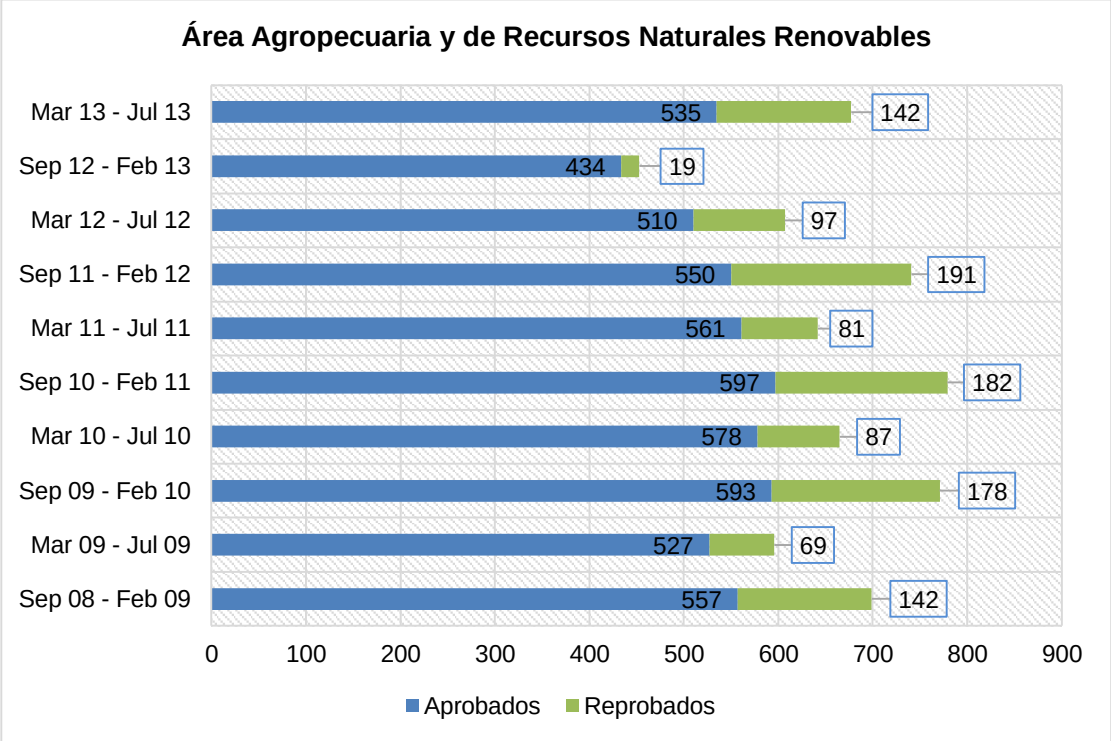


Figura 21: Estudiantes que aprobaron y reprobaron en el Área AARNR.

En los datos que se muestran en la figura anterior se puede notar que el periodo en el cual más estudiantes reprobaron es Septiembre 2011 – Febrero 2012 con un número de 191 estudiantes que corresponde al 24.70% de estudiantes matriculados en ese periodo, mientras que el periodo Septiembre 2012 – Febrero 2013 es el periodo con menos estudiantes reprobados con un número de 19 estudiantes que corresponde al 4.03% de estudiantes que se matricularon en ese periodo.

Los datos que se muestran en la siguiente figura (ver figura 22) corresponden a estudiantes que aprobaron y reprobaron en los diferentes periodos académicos desde el periodo septiembre 2008 – Febrero 2009 del Área de Educación, el Arte y la Comunicación de la Universidad Nacional de Loja (ver anexo 2).

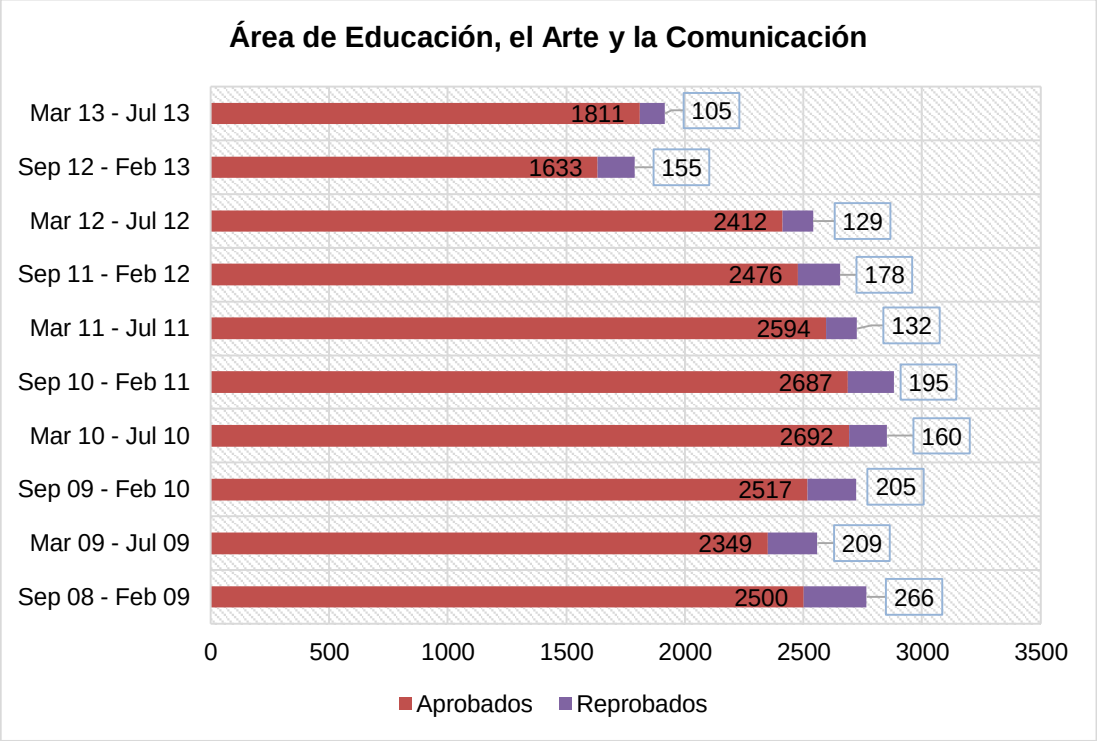


Figura 22: Estudiantes que aprobaron y reprobaron en el Área AEAC.

En la figura anterior el periodo donde más estudiantes reprobaron es el periodo Septiembre 2008 – Febrero 2009 con un número de 266 estudiantes que corresponde al 9.61% de estudiantes que se matricularon inicialmente en ese periodo, mientras que el periodo donde más se evidencio la reprobación de estudiantes es en el último periodo, Marzo 2011 – Julio 2011 con un número de 132 estudiantes que corresponde al 4.82% de estudiantes que se matricularon inicialmente en ese periodo.

En la siguiente figura (ver figura 23) se describe los estudiantes aprobados y reprobados en el Área de la Energía, Industrias y Recursos Naturales No Renovables.

Los datos analizados corresponden a estudiantes que cursaron los periodos desde Septiembre 2008 – Febrero 2009 hasta la actualidad (ver anexo 2).

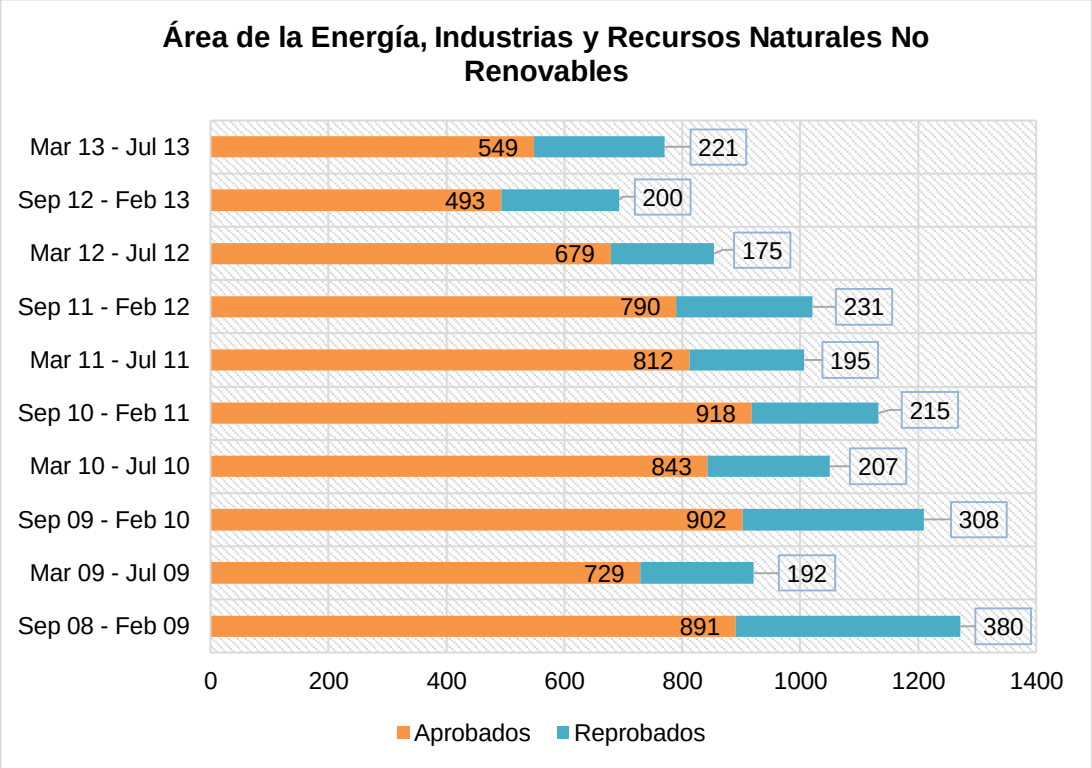


Figura 23: Estudiantes que aprobaron y reprobaron en el Área AEIRNNR.

En la figura anterior el período que registra más estudiantes reprobados es el periodo Septiembre 2008 – Febrero 2009 con un número de 380 estudiantes que corresponde al 29.89% de estudiantes que se matricularon al inicio del periodo académico, mientras que le periodo con menos estudiantes reprobados es el periodo Septiembre 2010 – Febrero 2011 con un número de 215 estudiantes que corresponde al 18.90% de estudiantes que se matricularon al inicio del periodo académico.

En la siguiente figura (ver figura 24) se describe información acerca de los estudiantes que han aprobado y reprobado en las diferentes carreras del Área Jurídica, Social y Administrativa, los datos analizados corresponden desde el periodo Septiembre 2008 – Febrero 2009 hasta la actualidad (ver anexo 2).

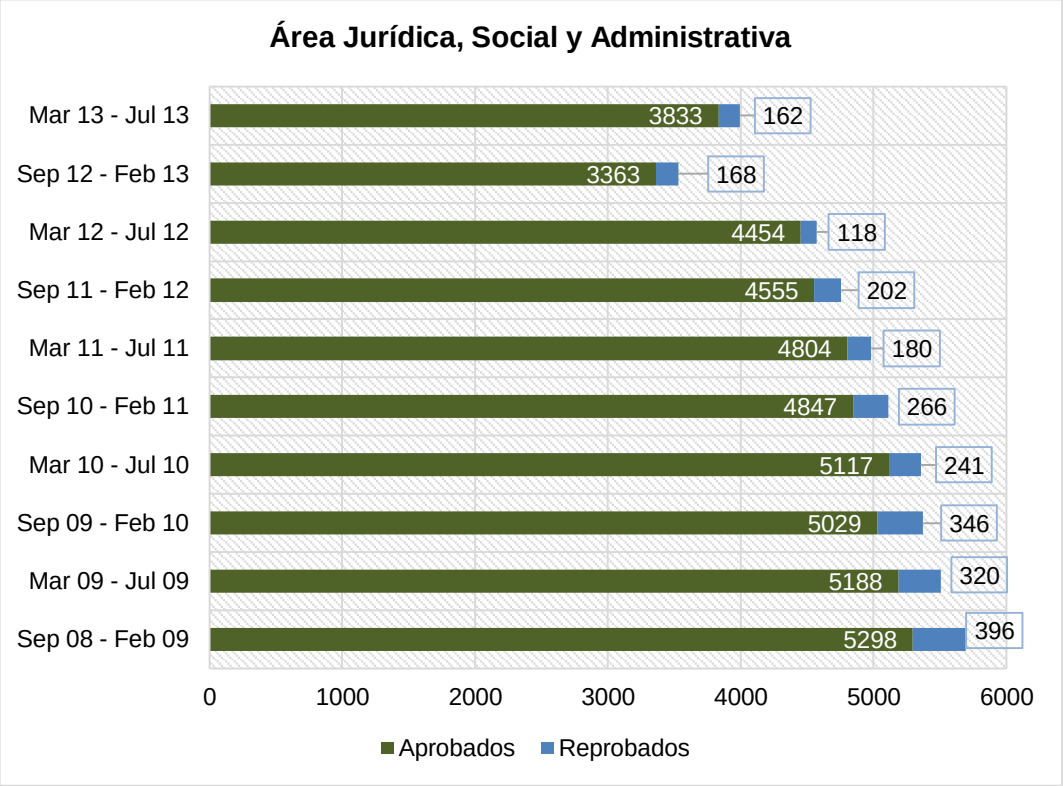


Figura 24: Estudiantes que aprobaron y reprobaron en el Área AJSA.

Los datos que se observan en la figura anterior muestran que el periodo con más estudiantes reprobados es el periodo Septiembre 2008 – Febrero 2009 con un número de 396 estudiantes, que corresponde al 6.94% de estudiantes matriculados inicialmente, mientras que el periodo con menos estudiantes reprobados es el periodo Marzo 2012 – Julio 2012 con un número de 118 estudiantes, que corresponde al 2.56% de estudiantes matriculados al inicio del periodo académico.

En la siguiente figura (ver figura 25) se muestra información acerca de los estudiantes que han aprobado y reprobado en las diferentes carreras del Área de la Salud Humana de la Universidad Nacional de Loja, los datos almacenados son desde el periodo Septiembre 2008 – Febrero 2009 hasta el momento.

Además de los periodos de pregrado que se cursan en esta área también se observó información acerca de periodos de Internado de los últimos tres años, por lo cual también se incluyó en la gráfica de análisis (ver anexo 2).

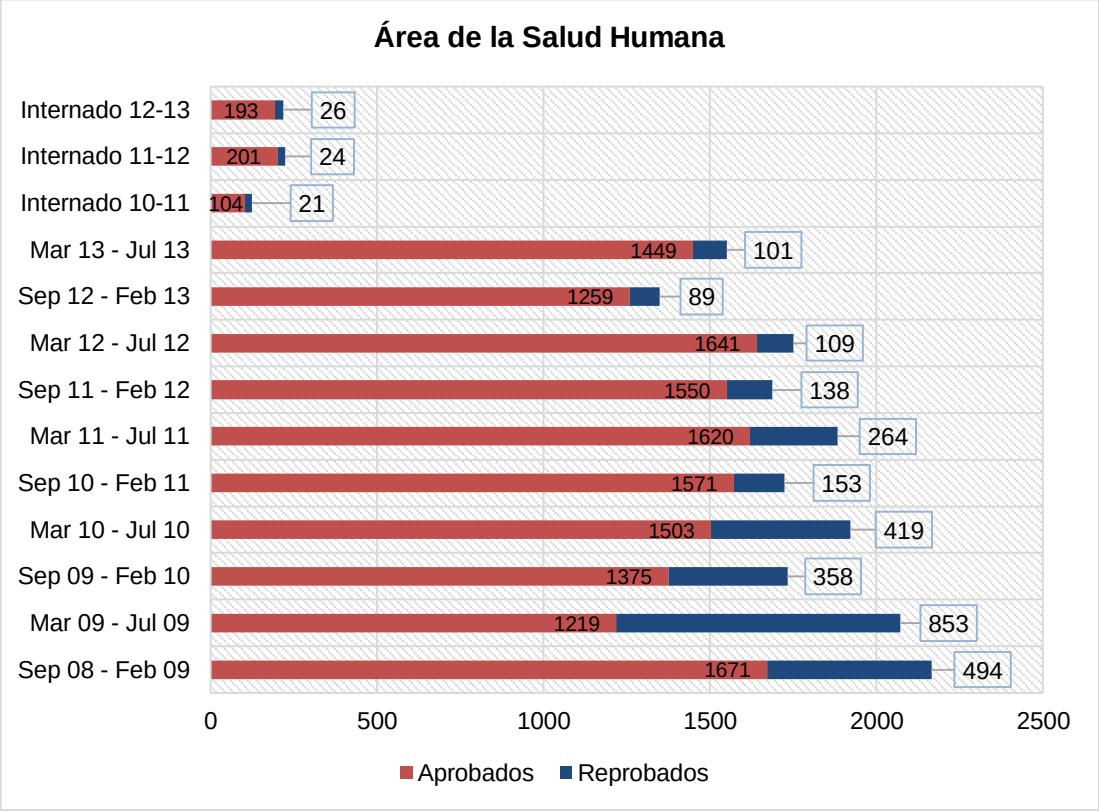


Figura 25: Estudiantes que aprobaron y reprobaron en el Área ASH.

En la figura anterior describe el número de estudiantes reprobados y aprobados en el Área de la Salud Humana en el cual se observa que existe un notable número de estudiantes reprobados de estudiantes en el periodo Marzo 2009 – Julio 2009, con un número de 853 que corresponden al 40.06% de estudiantes que se matricularon al inicio del periodo, mientras que el periodo Marzo 2012 –Julio 2012 se registró solo un número de 109 estudiantes reprobados, que corresponde al 6.13% de estudiantes matriculados inicialmente.

En la siguiente figura (ver figura 26) se muestra información acerca de los estudiantes aprobados y reprobados en las diferentes carreras del Área de Estudios a Distancia de la Universidad Nacional de Loja, los datos analizados corresponden desde el periodo Septiembre 2009 – Febrero 2010 hasta la actualidad (ver anexo 2).

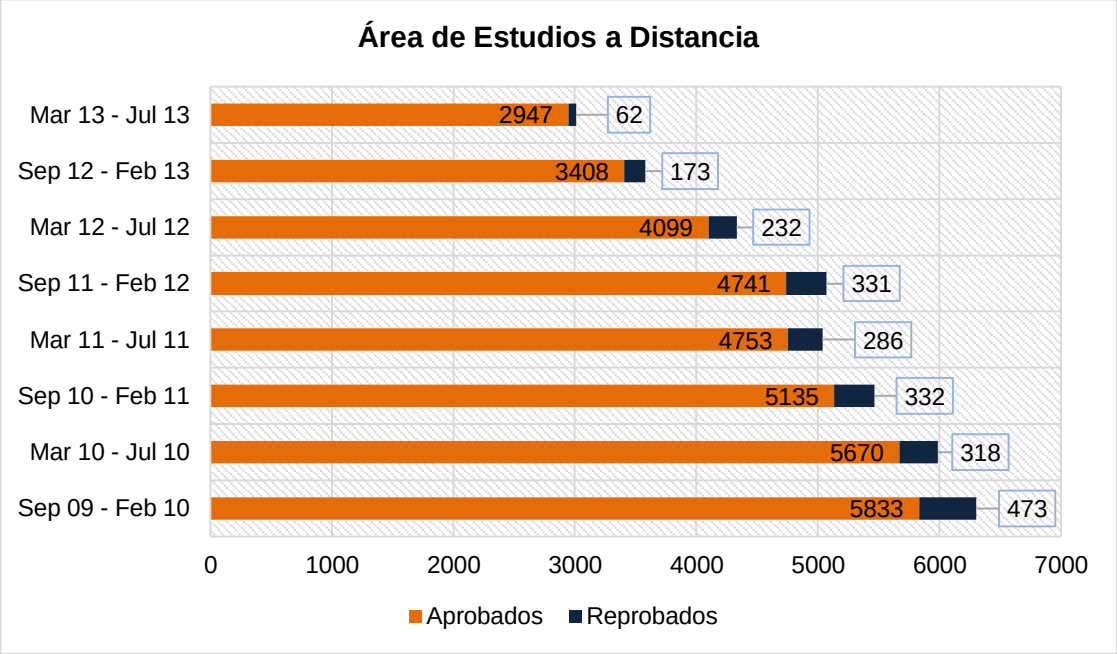


Figura 26: Estudiantes que aprobaron y reprobaron en el Área MED.

Los datos que se observan en la figura anterior, muestran que el periodo con más estudiantes reprobados es el periodo Septiembre 2009 – Febrero 2010 con un número de 473 estudiantes, que corresponde al 7.49% de estudiantes matriculados inicialmente, mientras que el periodo con menos estudiantes reprobados es el periodo Marzo 2013 – Julio 2013 con un número de 62 estudiantes, que corresponde al 2.03% de estudiantes matriculados al inicio del periodo académico.

En la siguiente tabla (ver tabla XXII) se muestra en promedio en porcentaje el promedio de reprobación en cada una de las Áreas de la Universidad Nacional de Loja (ver figura 27).

TABLA XXII:
ESTUDIANTES QUE REPRUEBAN EN CADA ÁREA.

Área	**Promedio**
Cursos Especiales	18.99%
Área Agropecuaria y de Recursos Naturales Renovables	16.76%
Área de Educación, el Arte y la Comunicación	6.81%
Área de la Energía, Industrias y Recursos Naturales No Renovables	23.32%
Área Jurídica, Social y Administrativa	4.79%
Área de la Salud Humana	14.94%
Modalidad de Estudios a Distancia	5.40%

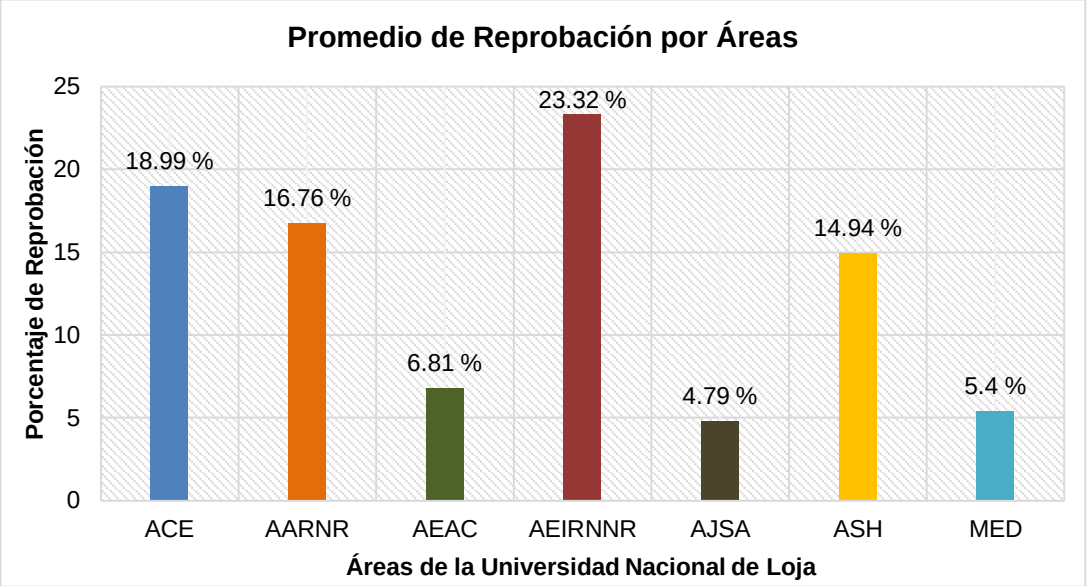


Figura 27: Porcentaje de estudiantes que reprueban en cada Área.

En el análisis de los datos acerca de los estudiantes que aprueban y reprueban en cada área de la Universidad Nacional de Loja se observó que el área que más estudiantes reprueban en promedio es el Área de la Energía, Industrias y los Recursos Naturales No Renovables, que en promedio por cada periodo académico reprueba el 23,32% de estudiantes, mientras que el Área Jurídica, Social y Administrativa en el área que menos estudiantes reprobados tiene con un 4,79%.

2. ETAPA DOS: Examinar Herramientas para exploración de Bases de Datos, Proceso de Minería De Datos, y revisión de técnicas que permitan resolver el problema planteado.

La presente etapa se realizara actividades relacionadas con el análisis y evaluación de herramientas para la gestión de bases de datos y de apoyo al proceso de minería de datos, además se realizara una revisión de las técnicas de minería de datos.

2.1. Selección de herramientas útiles para exploración de Bases de Datos

La siguiente figura (ver figura 28) muestra algunas herramientas para administrar bases de datos y sus características, que se consideran importantes para el presente Trabajo de Titulación, las mismas que se describen detalladamente en el anexo 3. Hay que

aclarar que estas son algunas y no las únicas herramientas para exploración y administración de bases de datos (ver anexo 3).

Producto	Licencia	Importar (ver anexo 1)			Exportar (ver anexo 1)											Asistente para crear / modificar datos
		SQL	ZIP	CSV	SQL	DBF	TXT	XLS	CSV	Database Estructure	WK	DIF	XML	HTML	Postgre Native SQL	
Database Workbench 4.4.1	Free Lite Edition	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓ [2]
DataAdmin 5.4.2.5 Personal	Personal	✓	✓	✗	✓	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓ [1]
My SQL Workbench 6.0.8	GPL	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
SQLyog Enterprise	Trial	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓ [3]

Figura 28: Características de Herramientas para explorar Bases de Datos.

Se analizó las herramientas que se muestran en la tabla y se seleccionó la herramienta DataAdmin 5.4.2.5 Personal, se decidió en base a la posibilidad de exportar o emitir reportes en diferentes formatos para su posterior uso (ver figura 29), también por la posibilidad de manipular datos directamente. Y por su licencia aunque estando limitada en algunas características cuenta con las suficientes para explorar y manipular los datos almacenados.

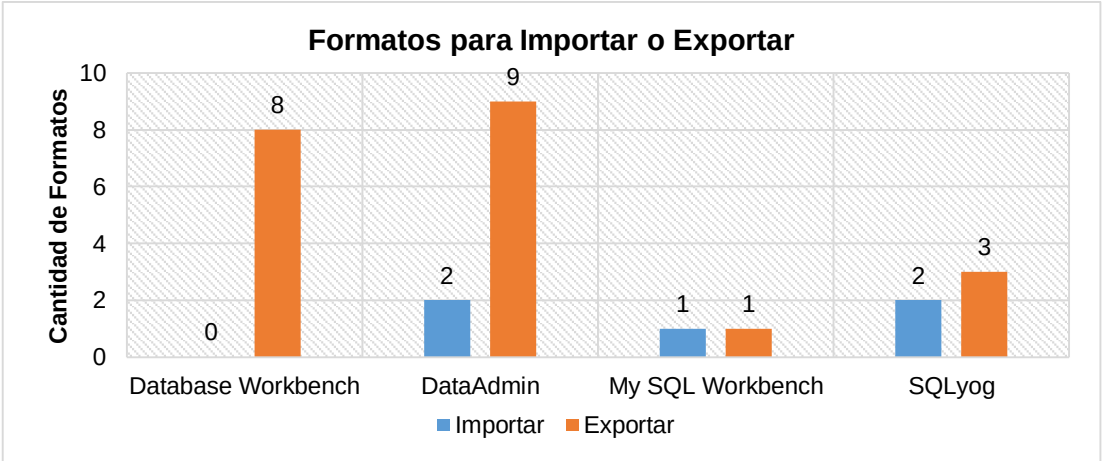


Figura 29: Formato de archivos por cada herramienta.

En la figura anterior se muestra la cantidad de formatos de archivos para exportar e importar de cada una de las herramientas para la exploración de bases de datos: la herramienta *Database Workbench* no tiene disponible ningún formato para importar por su tipo de licencia mientras que si posee un número de ocho formatos para exportar, la herramienta *My SQL Workbench* posee un formato disponible tanto para exportar y para importar, la herramienta *SQLyog* posee un número de dos formatos para importar y tres para exportar y por último la herramienta DatAdmin contiene el mayor número de formatos para exportar e importar, dos formatos para importar y nueve para exportar siendo esta la herramienta seleccionada para el desarrollo del presente trabajo.

2.2. Recopilación de casos de éxito en fuentes académicas, revistas, ponencias, artículos científicos, sobre la aplicación de minería de datos en la deserción.

En el campo educacional, las técnicas de minería de datos han sido usadas para entender el comportamiento de los estudiantes para recomendar actividades, ofrecer nuevas experiencias de aprendizaje o con el objetivo de mejorar la efectividad del curso, promover el trabajo en grupo o incluso predecir el rendimiento de los alumnos.

Se han recopilado algunos casos de éxito orientados a la deserción universitaria, los cuales generan conclusiones y recomendaciones que podrán ser tomados en cuenta al realizar el presente Trabajo de Titulación. Estos casos de estudio se encuentran detallados en la *Sección revisión literaria, Capitulo 1: Casos de Éxito.*

- **Caso de Éxito 1: Aplicando Minería de Datos al Marketing Educativo**

En el análisis se evidenció que el medio por el cual se enteraron los estudiantes sobre la carrera que cursan, en la gran mayoría de los casos fue a través de amigos y familiares, por lo cual se realizó un estudio para establecer si el tipo de información que recibió el estudiante, pudo ser mal interpretada o incorrecta y su influencia sobre la decisión de abandono.

Aunque en los análisis se reportó un grupo mínimo de estudiantes que desertaron en los programas de la Escuela de Marketing y Publicidad debido a problemas económicos, se debe tener en cuenta que ambos programas son los más costosos en su área de

conocimiento en Bogotá y que en general el perfil de estudiante que ingresa en la Universidad Sergio Arboleda no es sensible al costo de la matrícula.

- **Caso de Éxito 2: Aplicación de Técnicas de Minería de Datos para Evaluación del rendimiento Académico y la Deserción Estudiantil.**

El presente caso de éxito presenta resultados de la evaluación del rendimiento académico y de deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento (KDD), sobre los datos de alumnos del periodo 2003-2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un pre procesamiento de los datos y un software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

Al concluir el trabajo se pudo consolidar en el DIIT un grupo de investigación en las técnicas de Data Mining y además la implementación de un almacén de datos que permitirá tomar decisiones con menor incertidumbre. Además no se logró encontrar un clasificador del rendimiento académico y de la deserción estudiantil con un alto grado de precisión y compresibilidad, se adquirió experiencia en el uso de los programas SPSS y Weka que permitirá que el grupo avance en esta línea de investigación.

- **Caso de Éxito 3: Detección de Patrones de bajo rendimiento académico y Deserción Estudiantil con Técnicas de Minería De Datos.**

En el presente trabajo se muestran los resultados de la investigación realizada en la Universidad de Nariño (Colombia), el objetivo planteado fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios de DCBD del Departamento de Ingeniería.

En el desarrollo del presente Trabajo de Titulación, las fases de pre procesamiento y transformación de datos fueron las más costosas en tiempo, debido a la mala calidad de los datos de las bases de datos existentes.

En cuanto a los patrones obtenidos, la Universidad de Nariño deberá emprender actividades y proponer estrategias de seguimiento a estudiantes con estos perfiles con el fin de prevenir que caigan en bajo rendimiento y disminuir el grado de deserción que se presenta. Con este proyecto, se demostró que TariyKDD es una herramienta fiable, que puede ser utilizada en cualquier proyecto de Minería de Datos y su distribución es libre.

- **Caso de Éxitos 4: Aplicación de Técnicas de Minería de Datos para Identificar Patrones de Comportamientos Relacionados con las Acciones del Estudiante con el Eva de la UTPL.**

Para predecir el nivel de participación en el curso y el nivel de utilización de las herramientas, en informática y abogacía después de la aplicación de algoritmos de clasificación, el que presentó los mejores resultados fue el REPTree en informática y J48 en abogacía, y con fundamento en las matrices de confusión presentadas por éstos algoritmos se pudo apreciar notablemente que en ambas carreras existe mayor cantidad de estudiantes que presentan una escasa, participación en el curso, así como en la utilización de herramientas.

Se realizó algunas experimentaciones con clustering a los datos de los cursos: lógica de la programación (informática), ética y derechos humanos (abogacía), el algoritmo seleccionado fue SimpleKMeans para determinar grupos de estudiantes con comportamientos similares. En relación al indicador del nivel de utilización de las herramientas, se encontró que para lógica de la programación el uso de las herramientas foros, recursos, áreas, mensajería, twitter y cuestionario, es en un nivel permanente (alto) y moderado (medio), mientras que en ética y derechos humanos los estudiantes presentan un nivel bajo de utilización de las herramientas a excepción de los foros dónde se pudo observar que existe un nivel permanente (alto) de interacción y en algunos casos moderado.

Haciendo una comparación de los grupos generados en ambas materias, en los resultados se encontró que los estudiantes del curso de lógica de la programación presentan mayor interacción en la mayoría de las herramientas siendo estas: Foros, recursos, áreas, mensajería, twitter y cuestionario, mientras que los estudiantes de ética y derechos humanos revelan mayor interacción en una sola herramienta la cual es foros, por lo que se deduce que a lo mejor en este curso el profesor habilitó para el estudiante más actividades a desarrollar en esta herramienta, que en las demás, demostrando que la cantidad de utilización de las herramientas depende de las actividades que habilite el profesor en ellas.

En la búsqueda de los estilos de aprendizaje también se realizaron experimentaciones con clustering, se aplicó el algoritmo FarthestFirst, encontrando dos grupos: en el primer grupo se ubican los estudiantes que presentan mayor tendencia en el estilo visual-sensorial (V-SN), reflexivo-global (R-G), reflejando mayor complacencia para trabajar en sus áreas con la ayuda de gráficos, siendo muy observadores y creativos, aprendiendo mejor trabajando solos, en el segundo grupo están los estudiantes que tienen mayor incidencia en el estilo activo-secuencial (AC-SC), sensorial-auditivo (SN-AU), denotando que receptan mejor la información a través de debates con otras personas, exponiendo diferentes puntos de vista y tienen mayor gusto y predisposición para trabajar en grupo.

- **Evaluación Final**

Los puntos de interés sobre los cuales se analizó los casos de éxito descritos anteriormente son en base las técnicas o algoritmos que aplicaron y las herramientas que utilizaron para el desarrollo de los proyectos (ver tabla XXIII).

TABLA XXIII:
TÉCNICAS APLICADAS Y HERRAMIENTAS UTILIZADAS.

Casos de Éxito	Técnicas Aplicadas	Herramientas
Caso 1: "Aplicando minería de datos al marketing educativo"	Se aplicó técnicas de agrupamiento como clúster o algoritmo *Kmeans* a la bases de datos.	▪ Rapid Miner ▪ DatAdmin
Caso 2: "Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil"	En este trabajo se aplicó la clasificación como tipo de tarea de minería, el árbol de decisión como tipo de modelo y el J48 (implementación en Weka del algoritmo C4.5) y el FT como algoritmos de minería de datos.	▪ MS SQL Server ▪ SPSS ▪ Weka

Caso 3: "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos"	Para el descubrimiento de patrones de deserción estudiantil y bajo rendimiento académico se utilizaron las Áreas de Clasificación y Asociación. Para generar las reglas de clasificación su utilizó el algoritmo C4.5 y para las reglas de Asociación, el algoritmo EquipAsso, disponibles en la herramienta TariyKDD.	▪ TariyKDD
Caso 4: "Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL"	El trabajo se apoya en las técnicas de clasificación, como son los árboles de decisión, reglas de decisión y métodos bayesianos éstos se utilizan para establecer indicadores de participación del estudiante en el curso.	▪ Weka ▪ SQLyog

Al analizar cada uno de los casos se debe señalar que en la mayoría de los casos de éxito, para alcanzar los objetivos planteados se utilizó mayormente algoritmos de clasificación como árboles de decisión y técnicas basadas en reglas, entre las herramientas más utilizadas están Weka en dos casos y RapidMiner en uno.

2.3. Selección de Herramienta, como apoyo al proceso de Minería de Datos.

Existen algunas herramientas diseñadas para extraer conocimientos desde bases de datos que contienen grandes cantidades de información. La minería de datos es una técnica compuesta por fases la cual integra varias áreas y por tal motivo a esta no se la debe confundir con un gran software.

En la actualidad existen aplicaciones o herramientas comerciales de minería de datos muy completas que permiten extraer conocimiento desde bases de datos y estas a su vez contienen un sin número de utilidades que facilitan el desarrollo de un proyecto. Sin embargo, en casi todo desarrollo de proyecto suelen complementarse con otras herramientas. Las características de cada herramienta se encuentran descritas en el capítulo 3: *Herramientas útiles para el proceso de minería de datos.*

Luego de evaluar las herramientas para el proceso de Minería de Datos (ver anexo 4) se muestra a continuación la tabla con las características más importantes de las herramientas Orange, RapidMiner y Weka (ver tabla XXIV).

TABLA XXIV:
CARACTERÍSTICAS RELEVANTES DE HERRAMIENTAS DATA MINING.

Producto		orange	RAPID MINER	WEKA The University of Waikato
Licencia		Open Source	Gnu GPL	GNU
Metodología		Semma	Varias	Varias
Manipulación a través de:	GUI	✓	✓	✓
	Batch (Lotes)	X	✓	X
	Línea de comandos	✓	✓	✓
	Creando Aplicación	X	✓	✓
Integración con Herramientas		X	✓	✓

En base a las características analizadas en cada herramienta, con Orange se observó que está basado en varios componentes y cuenta con un cómodo, potente, rápido y versátil front-end de programación visual para el análisis exploratorio de datos y visualización. Emplea una completa variedad de componentes para pre procesamiento de datos, filtrado, modelado, evaluación del modelo, y técnicas de exploración importantes para el proceso de minería de datos.

La herramienta Rapid Miner es un ambiente para realizar experimentos en minería de datos además de aprendizaje automático, tanto en investigación como el mundo real. Se pudo observar que los experimentos pueden componerse de un gran número de operadores anidables, y cuentan con descripción individual que no es muy explícita como en otras herramientas como Weka y Orange, esta herramienta ofrece más de 500 operadores y evaluadores de atributos, que están incluidos en la herramienta Weka. Una de las características más atractivas de esta herramienta es que también puede ser utilizada como motor de minería de datos para integrarse en otros productos.

La herramienta Weka es una herramienta muy conocida y que cuenta con la mayor cantidad de documentación, al examinar la herramienta se observó que es una herramienta que se apoya más en el procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Las técnicas de minería de datos se basan en la hipótesis de que los datos están disponibles en un único

archivo plano o relación (.arff), además la herramienta proporciona la opción de acceso a bases de datos SQL utilizando conectividad de bases de datos Java y puede procesar el resultado devuelto como una consulta de base de datos.

Finalmente se determinó que Weka y RapidMiner eran las herramientas convenientes para el trabajo que se realizó, ya que ambas podían complementarse. Pero finalmente se ha seleccionado la herramienta RapidMiner por sus amplias y flexibles características que ofrece ya que están más acordes a nuestras necesidades para realizar actividades en el proceso de minería de datos, además las técnicas de procesamiento y evaluadores de atributos de la herramienta Weka se encuentran disponibles también en la herramienta RapidMiner

- **Evaluación con datos de Prueba**

Para la evaluación de las herramientas descritas anteriormente se utilizaron datos de prueba y se evaluaron sus funciones, los datos obtenidos para su evaluación corresponden a un Data Set denominada “Iris” que describen las características de 3 tipos de flores (ver anexo 5).

Para la evaluación de las herramientas se tomaron en cuenta características como: descripción y manejo de procesos, visualización de resultados, manipulación de datos, estos permitieron aclarar la funcionalidad de cada una de las herramientas. Además se eligió el algoritmo de clasificación K-Means, para obtener clusters con los datos de prueba y evaluar los resultados en cada herramienta.

A continuación (ver tabla XXV) se describen algunos criterios de evaluación que también se tomaron en cuenta la evaluar cada una de las herramientas de Minería de Datos (ver anexo 5).

TABLA XXV:
COMPARACIÓN DE CARACTERÍSTICAS DE HERRAMIENTAS DATA MINING.

Herramienta	Descripción y Manejo de Procesos	Visualización Resultados	Manipulación de Datos
Rapid Miner	Los procesos para realizar la clasificación de las instancias, se realiza a través de componentes que se arrastran a un espacio	Ofrece 6 maneras de visualizar los resultados: - Text View - Folder View - Graph View - Centroid Table	Los datos a importar se pueden hacer desde una base de datos o archivos con varias extensiones, hacia un repositorio

	de trabajo de la cual se obtiene una perspectiva eficiente. Además se puede obtener una descripción de los procesos en XML.	- Centroid Plot View - Annotations	de Rapid Miner por lo cual la herramienta no se rige a una sola extensión de archivo, sino que lo hace directamente sobre los datos.
Weka	Los procesos para realizar la clasificación de los datos se realiza a través de sus cuatro entornos *(Explorer, Experimenter, KnowledgeFlow, Simple CLI)*, estos se pueden realizar a través de componentes en el entorno *KnowledgeFlow.*	La visualización de resultados en el entorno *Explorer* se realiza a través de matrices y se compara cada uno de los atributos evaluados mientras que en el entorno *KnowledgeFlow* se realiza a través de: Matrices, Texto Plano y Grafos.	Los datos se pueden importar de tres maneras, mediante una dirección url, Conexión a Base de Datos y la opción que más trabajo toma realizar que es a través de un archivo con extensión *arff.*
Orange	Los procesos se gestionan por medio de componentes sobre un área de trabajo y estos a su vez se encuentran en categorías. Esto permite que se manipulen de forma eficaz también contiene una categoría donde se encuentran componentes que aún están en proceso de desarrollo.	La visualización de los datos se realiza a través de ciertos componentes los cuales son: - Distributions - Attribute Statistics - Scatter Plot - Linear Projection - Radviz - Polyviz - Parallel Cordinates - Survey Plot - Correspondence Analysis - Multi Correspondence Analysis - Mosaic Display - Sieve Diagram	Los datos fuente para aplicar algoritmos se manejan a través de un componente el cual se configura para para importar archivos con varias extensiones, entre las cuales están *txt, xml* y *arff*, la desventaja de esta herramienta es que para conectarse a una base de datos no es posible.

A continuación se describe el análisis que se realizó a cada herramienta antes mencionada y las características seleccionadas en la tabla anterior (ver figura 30), los valores que se establecieron para evaluar las herramientas fueron en un rango de 0-25, se lo realizó a criterio personal, en base a la experiencia al manejar cada una de las herramientas.

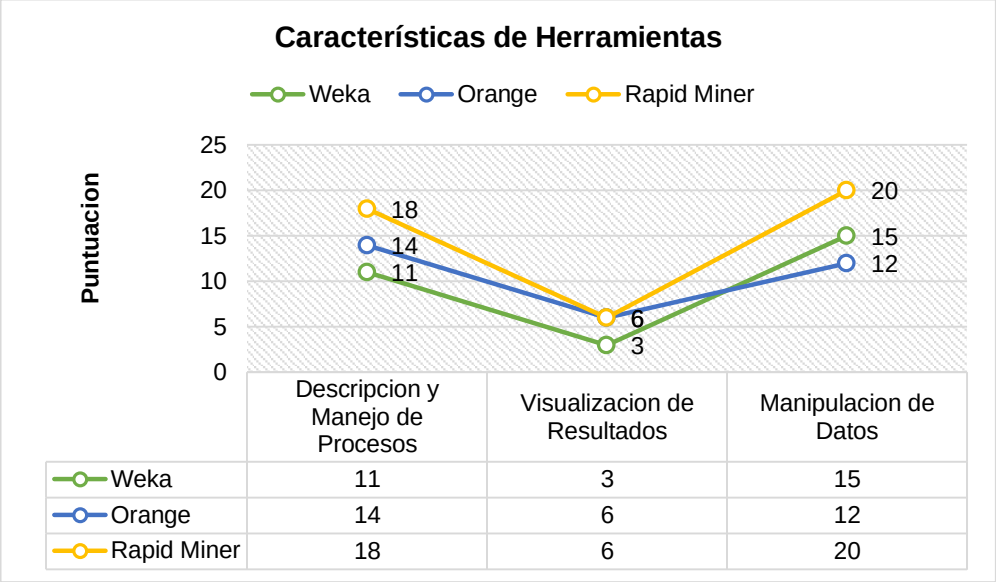


Figura 30: Evaluación de Características de Herramientas.

En la figura se observa que la herramienta Weka en cuanto a descripción y manejo de procesos tiene una puntuación de 11 esto porque sus procesos no contenían ninguna descripción o guía del manejo de estos, en visualización de resultados con una puntuación de tres este valor se estableció por el número de poder visualizar los resultados y en cuanto a manipulación de datos se estableció 15 por la posibilidad de manejar los datos desde fuentes distintas o archivos que se podían importar.

En cuanto a la herramienta Orange en cambio obtuvo una puntuación de 14 en descripción y manejo de procesos, en visualización de resultados obtuvo una puntuación de 6 y en manipulación de datos un valor de 12, la herramienta Rapid Miner es la que muestra mejores valore en cuanto a características evaluadas con un valor de 18 en descripción y manejo de procesos, en visualización de resultados se estableció un valor de 6 y un valor de 20 en manipulación de datos.

2.4. Recopilación de Información acerca de las Técnicas de Minería de Datos.

Las técnicas de Minería de Datos (una etapa dentro del proceso completo de KDD) pretenden obtener patrones o modelos a partir de datos recopilados. Y es aquí donde el usuario realiza una evaluación subjetiva y decide si los modelos obtenidos son útiles.

Existen diversas técnicas de minería de datos las cuales se describen en la tabla XII, y mencionan dos categorías: las no supervisadas o descriptivas y las supervisadas o predictivas. Además se debe recalcar que cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado manejando aproximaciones distintas. Estas técnicas se encuentran descritas en el *Capítulo 2: Recopilación de Técnicas de Minería de Datos* de la Sección Revisión Literaria.

3. ETAPA TRES: Generar Modelos para la Identificación de Factores y Patrones de Comportamiento.

En la presente etapa se aplican las técnicas seleccionadas, se generan las estructuras de minería de datos y se generan modelos, posteriormente se compara su rendimiento.

3.1. Realizar la Integración y Recopilación de datos Iniciales

En esta actividad se realizaran tareas de análisis en cuanto a los objetivos del Trabajo de Titulación, criterios de éxito, también se realiza la recopilación de los datos con los cuales se trabajó y luego una exploración de los mismos.

3.1.1. Primera Fase: Comprensión del Negocio

A continuación se describen los objetivos que se desean lograr con el desarrollo del presente Trabajo de Titulación. El mismo que se originó con la idea de aplicar técnicas de Minería de Datos para identificar los factores en la deserción y reprobación de estudiantes.

3.1.1.1. Tarea Uno: Determinar los objetivos del negocio

Luego de haber descrito cuales son los objetivos e idea general del presente Trabajo de Titulación, a continuación se detallan los criterios de éxito, desde un punto de vista del negocio, además el problema se afronta desde dos puntos de vista.

El Área de la Energía las Industrias y los Recursos Naturales No Renovables, pertenece a la Universidad Nacional de Loja. Esta es una institución de Educación Superior con

una fuerte presencia a nivel regional y nacional. Ofrece formación académica y profesional de calidad en el marco del SAMOT, con sólidas bases científicas y técnicas, pertinencia social y valores.

La institución es un referente para el desarrollo profesional y se encuentra en la Región Sur y del Ecuador además posee varias extensiones por todo el país. Sus estudiantes son hombres y mujeres pertenecientes a todas las etnias ecuatorianas y de todos los estratos sociales. La misión principal de la institución es ofrecer formación en los niveles: técnico y tecnológico superior; profesional o de tercer nivel; y, de postgrado o cuarto nivel; que realiza investigación científico-técnica sobre los problemas del entorno, con calidad, pertinencia y equidad, a fin de coadyuvar al desarrollo sustentable de la región y del país.

Objetivos del Negocio

- Conocer los factores que determinan que un grupo de sus estudiantes: reprueben o abandonen sus estudios.
- Identificar los patrones de comportamiento de estudiantes reprueban los estudios.

Factores Críticos del Éxito

- En este Trabajo de Titulación se consideró como factor de éxito la posibilidad de aplicar técnicas de minería de datos para predecir si un estudiante tiene probabilidades de abandonar los estudios, para ello se tiene en cuenta la utilización de datos personales, notas académicas.
- Uno de los criterios que se toma en cuenta en el desarrollo del presente Trabajo de Titulación es identificar al menos un factor que incida en la deserción y reprobación, es por ello que se utilizó datos personales de estudiantes que pertenecen al Área de Energía de la Universidad Nacional de Loja, estos datos obtenidos del Web Services se migraron a una base de datos relacional para su posterior selección, limpieza y transformación de los mismos.

3.1.1.2. Tarea Dos: Evaluación de la Situación

En esta tarea se realizó un análisis de todos los recursos disponibles para el desarrollo del Trabajo de Titulación entre estos están: los recursos humanos, hardware software y

otros. Igualmente se describen los requerimientos, restricciones y los riesgos con sus actividades de contingencia.

3.1.1.2.1. Recursos Disponibles

En el desarrollo del presente Trabajo de Titulación se tomaron en cuenta recursos humanos, hardware, software, datos y materiales, a continuación se describen los mismos.

- **Recursos Humanos**

En el presente Trabajo de Titulación el coste del personal se entiende por el número de horas en la realización de las actividades del Trabajo de Titulación. Los trabajadores que pertenecen al equipo de trabajo son los siguientes.

Asesor de Proyecto: Sus principales funciones fueron llevar a cabo la revisión del proyecto y la consolidación de las conclusiones.

Investigador: sus funciones fueron el manejo de la herramienta de minería de datos, en caso Rapid Miner, y de los diferentes algoritmos de minería de datos incluidos en la herramienta.

Administrador de Bases de Datos: sus funciones fueron la preparación de los datos y de programar las consultas requeridas por el analista para los distintos experimentos.

- **Recursos Hardware**

El desarrollo del Trabajo de Titulación no se necesitó de una gran inversión económica para adquirir equipos Hardware ya que el personal de trabajo no necesitó de varios equipos para trabajar en las actividades.

Ordenador Personal: fue necesario un ordenador portátil de un coste medio ya que para la elaboración del presente Trabajo de Titulación no fue necesaria la compra de un ordenador con características de último modelo.

Impresora: para la revisión de avances del Trabajo de Titulación fue necesaria la impresión en papel por lo cual se requirió de una impresora. Además no fue necesario que esta sea de un alto costo ya que no se necesitó de imágenes de alta calidad.

- **Recursos Software**

Por recursos Software se entiende las licencias de pago de las herramientas que se utilizaron para la realización del Trabajo de Titulación.

Rapid Miner: es una herramienta informática para el análisis y minería de datos puesto que permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación educación, capacitación, creación rápida de prototipos. Además que proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, pre procesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en Weka. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004.

DatAdmin: es un software intuitivo para la administración de las bases de datos que admite numerosas bases de datos, su diseño se centra principalmente en el uso fácil y gracias a la modularidad y la flexibilidad de la herramienta puede ser, desde el punto de vista de la funcionalidad, en comparación con los programas centrados en una única base de datos, además de ser libre solo para fines no comerciales.

TexMaker: esta herramienta está desarrollada como un editor de LaTeX de código abierto, fácil de usar, potente y completo. Porque logra integrar muchas herramientas necesarias para elaborar documentos con LaTeX. También incluye compatibilidad con Unicode, un corrector ortográfico, autocompletado, plegado de código y un visor de PDF integrado que admite synctex y el modo de vista continua.

- **Datos**

Para el desarrollo del presente Trabajo de Titulación se utilizaron datos provenientes del Sistema de Gestión Académica y el Área de Bienestar Universitario estos datos comprenden información de estudiantes y docentes de la institución y los servicios que han recibido.

SGA (Sistema de Gestión Académica): este es un Sistema de Gestión que posee la Universidad Nacional de Loja y es administrado por el departamento Unidad de Telecomunicaciones e Información el cual proporciono una cuenta de acceso a los mismos. El sistema contiene información tanto personal como académica de los estudiantes de la institución y datos en cuanto a planificación de los Docentes.

Área de Bienestar Universitario: esta área de la Universidad Nacional de Loja se encarga de contribuir al mejoramiento de la calidad de vida y al desarrollo integral de la comunidad universitaria, la información se encontró almacenada en papel, su procesamiento se describe en áreas posteriores.

- **Materiales**

El material necesario para el desarrollo del Trabajo de Titulación fue material de oficina como: carpetas plásticas, bolígrafos, cartuchos de impresora, etc. Dicho material fue necesario para la elaboración de los documentos, para revisiones y para la entrega de la memoria final del Trabajo de Titulación.

3.1.1.2.2. Riesgos y Contingencias

A continuación se realizara una identificación de los posibles riesgos y también se describen los planes de contingencia para cada uno de ellos, en caso de que ocurran se puede estar prevenido ante dichas situaciones y evitando impactos negativos en la planificación y el coste del Trabajo de Titulación.

Estos son los riesgos (ver tabla XXVI) que se identificaron en base a una observación personal y además se describen los planes de contingencia para reducir el impacto en caso de que suceda alguno.

TABLA XXVI:
RIESGOS Y CONTINGENCIAS DEL PROYECTO.

Riesgos	Contingencias
Perdida de los datos.	Realizar respaldo de los datos en la nube o en espacios físicos diferentes al de origen.
Dificultad de comunicación con el tutor o el asesor.	Solicitar reunión con el tutor o asesor y obtener datos personales para facilitar la comunicación.
Fallos en hardware.	Mantenimiento preventivo, inspección técnica y reparación o realizar nueva adquisición.
Datos no confiables e incompletos para las pruebas.	Realizar nuevamente la selección de atributos para evaluar o seleccionar nuevamente los algoritmos utilizados.

Fallos en el software	Realizar pruebas de funcionamiento antes de utilizar software en actividades del proyecto.
Mala estimación del tiempo para cada una de las actividades, es decir retrasos en alguna actividad.	Estimar detenidamente el esfuerzo necesario para las actividades y prevenir problemas y retrasos añadiendo un tiempo prudente para solucionarlos.

3.1.1.2.3. Terminología

Estos son los términos más relevantes en el proyecto, este glosario contiene términos propios del negocio así como aquellos específicos de la minería de datos.

Términos del Negocio

- **Web Services:** Un Web Services es un sistema diseñado para soportar la interoperabilidad y la interacción entre máquina-maquina a través de una red.
- **WSDL:** es un formato XML para describir servicios de red como un conjunto de puntos finales que operan en mensajes y estos a su vez se orientan a documentos, orientados a información o procedimiento. Las operaciones y los mensajes se describen de forma abstracta, y luego se enlaza a un protocolo de red y el mensaje de formato concreto para definir un punto final.

Términos de Minería de Datos

- **Base de Datos:** conjunto de datos organizados de modo tal que resulte fácil acceder a ellos, gestionarlos y actualizarlos.
- **Minería de Datos:** conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de pronosticar de forma automatizada tendencias y comportamientos.
- **Información:** Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada.
- **Conocimiento:** Entendimiento, inteligencia, razón natural.
- **Algoritmos de asociación:** es un procedimiento que obtiene un conjunto de valores que se repiten de un determinado tamaño, para combinarlos con reglas.

- **Algoritmos de clasificación:** es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio.
- **Árbol:** estructura de datos en la cual los registros son almacenados de manera jerárquica.
- **Regla:** conjunto de operaciones que deben llevarse a cabo para realizar una inferencia o deducción correcta.

3.1.1.2.4. Presupuesto

En esta sección se detalla los costes asociados con el Trabajo de Titulación. Los costes estarán agrupados por categorías para lograr una mayor comprensión de los mismos.

- **Costes de Personal**

Estos costes se refieren a los honorarios de todos los integrantes del equipo de desarrollo, durante el periodo de desarrollo del Trabajo de Titulación. Cada integrante está definido por un rol que desempeño en el Trabajo de Titulación y las horas de trabajo.

Se determinó el coste por hora de cada uno de los integrantes que participaron en el desarrollo del Trabajo de Titulación. A continuación, (ver tabla XXVII) se detallan los sueldos de integrante.

TABLA XXVII:
COSTO POR HORA DE INTEGRANTES DE PROYECTO.

Rol	Sueldo por Hora ($)
Asesor de Proyecto	$ 10,31
Investigador	$ 12,07
Administrador de Base de Datos.	$ 14,21

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

TABLA XXVIII:
RELACIÓN DE ACTIVIDADES DEL PROYECTO Y DURACIÓN DE LAS MISMAS.

Nro.	Actividad	Duración (horas)
1	Análisis de datos existentes en las Bases de Datos de la Universidad Nacional de Loja.	85

2	Examinar las técnicas y herramientas de minería de datos.	120
3	Generar modelos para la identificación de factores y patrones de comportamiento.	282
4	Evaluar modelo generado y análisis de resultados.	144
	Total:	631

Cada actividad no necesariamente requiere de un integrante para realizarla, por lo que se detalló el porcentaje de implicación en cada actividad (ver tabla XXIX).

TABLA XXIX:
ASIGNACIÓN DE ACTIVIDADES POR ROLES Y CÁLCULO DE HORAS DEDICADAS

Actividad	Profesional	% de Implicación	Horas
1	Asesor de Proyecto	10 %	7,5
	Investigador	60 %	51
	Administrador de Base de Datos.	30 %	26,5
2	Asesor de Proyecto	10 %	12
	Investigador	70 %	84
	Administrador de Base de Datos.	20 %	24
3	Asesor de Proyecto	10 %	28,2
	Investigador	65 %	183,3
	Administrador de Base de Datos.	25 %	70,5
4	Asesor de Proyecto	25 %	36
	Investigador	60 %	86,4
	Administrador de Base de Datos.	15 %	21,6

A continuación, se han de sumaron las horas de cada uno de los integrantes que intervinieron en el desarrollo del proyecto de esta forma poder calcular el coste del personal por hora invertidas (ver tabla XXX).

TABLA XXX:
RECOPILACIÓN DE HORAS Y COSTES POR ROL DEL PERSONAL.

Integrante	Sueldo por hora ($)	Horas invertidas	Coste Total ($)
Asesor de Proyecto	$ 10,31	83,7	$ 862,95
Investigador	$ 12,07	404,7	$ 4884,73
Administrador de Base de Datos.	$ 14,21	142,6	$ 2026.35
		Total:	$ 7774,03

- **Costes de Hardware**

A continuación se detallan los costes relacionados con el hardware necesario para el desarrollo del Trabajo de Titulación. También se incluyen los precios totales de cada equipo, la vida útil de los equipos informáticos es de tres años, por lo que se tomó en cuenta la depreciación de los mismos.

La depreciación de los equipos es proporcional al año de utilización por lo que se aplica una reducción del tercio de su precio por cada año de uso (ver tabla XXXI).

TABLA XXXI:
COSTES DEL HARDWARE.

Equipo	**Cantidad**	**Precio Unitario ($)**	**Precio Total ($)**
Ordenador Portátil	1	$ 940,00	$ 940,00
Impresora	1	$ 73,00	$ 73,00
Total			$ 1013,00

Como se describe anteriormente la duración del Trabajo de Titulación es de ocho meses y el tiempo de vida útil de equipos informáticos es de tres años el coste total incluyendo la depreciación es de $ 337,00.

Las características de los Equipos son los siguientes:

Ordenador Portátil: Hp Pavillion dv2610, Procesador AMD Turion (tm) 64 X2 TL-58 1.90 Ghz, 2.00 GB RAM, Sistema Operativo Windows 7 32 bits, Tarjeta gráfica Nvidia ® GeForce G310 de 512 MB, Disco duro SATA de 500 GB (7.200 rpm), Unidad óptica de DVD+/-RW a 16X (lectura y escritura de DVD/CD).

Impresora: Impresora multifunción de inyección de tinta en color Canon MP230. Multifunción impresora/copiadora/escáner, conectividad USB 2.0 de alta velocidad (conector tipo B), hasta 4.800 x 1.200 ppp.

- **Costes de Software**

A continuación se detallan los costes asociados al software necesario para el desarrollo del Trabajo de Titulación. Este recurso se diferencia del Hardware ya que se adquiere mediante licencias, es decir permisos para utilizar dicho software en un número determinado de ordenadores durante un periodo de tiempo. En algunos casos el tiempo suele ser ilimitado (ver tabla XXXII).

La licencia del Sistema Operativo tiene duración ilimitada y este se consideró como un coste incluido en el del ordenador personal.

TABLA XXXII:
COSTES DEL SOFTWARE.

Producto	Cantidad	Precio Unitario ($)	Coste Total ($)
Rapid Miner	1	$ 0.00	$ 0,00
TexMaker	1	$ 0.00	$ 0,00
DatAdmin	1	$ 0.00	$ 0,00
		Total:	$ 0,00

Tanto la herramienta Rapid Miner, TexMaker y DatAdmin son software libre por lo que no representaron ningún coste en el Trabajo de Titulación. Por lo tanto los costes de depreciación no se aplican a los mismos.

- **Materiales y Servicios**

En esta sección se detallan los costes relacionados con los materiales de oficina. Al igual que en la sección anterior (Costes de software) se incluyen los precios totales de cada material.

El material de oficina es necesario para el desarrollo del Trabajo de Titulación como carpetas plásticas, bolígrafos, tinta para la impresora, etc. Dicho material fue necesario para la elaboración de los documentos, para la gestión interna y para la entrega de la memoria final del Trabajo de Titulación (ver tabla XXXIII).

Los gastos en servicios básicos se calculan como el 10% sobre el total del resto de los costes. En este caso incluyeron principalmente los siguientes gastos:

Suministro eléctrico con la compañía EERSSA durante los ocho meses que duró el desarrollo del Trabajo de Titulación. Con esto se pudo hacer uso de las herramientas necesarias para llevar a cabo el desarrollo del mismo.

Contratación de una línea telefónica y a su vez una conexión a internet del tipo ADSL que garantizase una conexión estable a internet y así poder trabajar con todas sus ventajas. Transporte para reuniones, revisiones periódicas del desarrollo del Trabajo de Titulación.

Sumando los costes anteriores se obtuvo un resultado de 3297,9 $ por tanto, el coste total de los gastos en servicios fue de 329,79 $.

TABLA XXXIII:
COSTES DE MATERIALES.

Ítems	Descripción	Coste Total ($)
Materiales de Oficina	Papel, tinta, carpetas, etc.	$ 101,80
Servicios Básicos	Servicios de agua, luz e internet.	$ 329,79
Transporte	Recorridos en un aproximado de 60.	$ 26,00
Publicación de Resultados.	Publicación en revista indexada.	$ 260,00
Cursos de Capacitación	Seminarios, talleres.	$ 250,00
Total:		$ 967,59

- **Coste Total**

A continuación se resumen todos los gastos del Trabajo de Titulación, desglosados en las mismas secciones anteriores (ver tabla XXXIV).

En caso de imprevistos se calculan como el 10% del valor total del Trabajo de Titulación y el costo corresponde a $ 373,75.

TABLA XXXIV:
RESUMEN DE COSTES DEL PRESUPUESTO.

Recursos	Coste Total ($)
Costes de Personal	$ 7774,03
Costes de Hardware	$ 1013,00
Costes de Software	$ 0,00
Costes de Material de Oficina y Servicios.	$ 967,59
Subtotal:	$ 9754,62
Imprevistos 10%	$ 975,46
Total:	$ 10730,08

3.1.1.2.5. Cronograma del Proyecto

A continuación, se muestra el cronograma del Trabajo de Titulación, se elaboró un diagrama de Gantt en el que se muestran las distintas actividades que permitieron el desarrollo del mismo y los recursos empleados para la realización de cada una de ellas (ver figura 31).

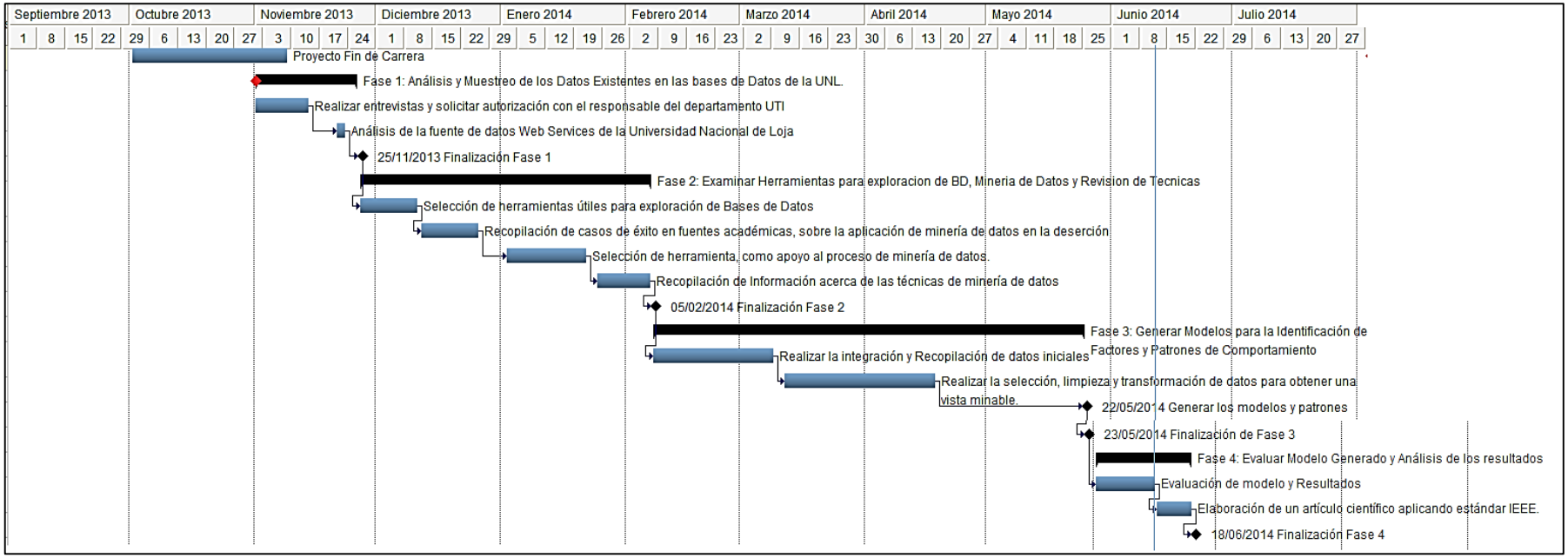


Figura 31: Cronograma de Proyecto.

3.1.1.3. Tarea Tres: Determinación de metas de la Minería de Datos

La minería de datos consiste en el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Para implementarlo se hace uso de algoritmos de clasificación, asociación de atributos y detección de patrones secuenciales para resolver problemas de agrupamiento automático y en esta sección se explican los objetivos a lograr con la aplicación de algoritmos de minería de datos en este Trabajo de Titulación.

Los proyectos de minería de datos generalmente tienen como metas una de las siguientes: Ahorro de dinero optimizando la eficacia de sus actividades y obtener ganancias económicas descubriendo nuevas fuentes de beneficios.

Sin embargo el presente Trabajo de Titulación está orientado a un ámbito claramente académico, de no haber sido así estaría más acorde con el segundo objetivo. La meta de la minería de datos en este Trabajo de Titulación es que a partir de un conjunto de datos, en este caso notas académicas, datos personales de estudiantes, periodos en los cuales los estudiantes reprueban y un conjunto de técnicas presentes en la herramienta Rapid Miner, poder identificar los factores por los cuales reprueban o abandonan los estudios universitarios.

Para poder alcanzar el objetivo descrito anteriormente se tiene como metas las siguientes: el análisis de los datos existentes en la institución a través de entrevistas y reuniones con encargados, la siguiente meta en alcanzar es la de examinar las herramientas que permitirán llevar de mejor manera los procesos de minería de datos, posterior a esto la meta a lograr es la generación de modelos que permitan identificar los factores que inciden en la reprobación y deserción universitaria.

Puesto que existe una gran cantidad de datos disponibles, es necesario aplicar técnicas de selección de atributos, para pulir los resultados eliminando atributos innecesarios o que aporten con muy poca información para la identificación de los factores de deserción y reprobación.

3.1.1.4. Tarea Cuatro: Plan de Proyecto

En esta sección se detalla el plan que se realizó para alcanzar los objetivos de la minería de datos y con ellos los del negocio. Para mejor comprensión se muestra la planificación (ver tabla XXXV) incluyendo el nombre de las actividades ejecutadas, con su duración, los recursos que requiere, herramientas, técnicas entradas y salidas.

TABLA XXXV:
IDENTIFICADOR DE INTEGRANTES.

Identificador	Integrante
AP	Asesor de Proyecto
AI	Analista/Investigador
AD	Administrador de Bases de Datos.

Los recursos empleados se establecen en porcentajes, es decir la suma entre todos los porcentajes de participación debe ser el 100%. A continuación, se detalla (ver tabla XXXVI) el plan de proyecto.

TABLA XXXVI:
PLAN DEL PROYECTO.

Actividad	Duración (horas)	Dependencias	Recursos %			Herramientas	Técnicas	Entradas	Salidas
			AP	AI	AD				
1. Análisis de datos existentes en las Bases de Datos de la UNL.	84		10	60	30	Excel, TexMaker, DatAdmin	CRISP-DM		Acceso a Datos de Web Services de la UNL, Análisis de Fuente de Datos.
2. Examinar las técnicas y herramientas de minería de datos.	120	1	10	70	20	TexMaker	CRISP-DM	Objetivos del Proyecto.	Herramienta para Explorar BD, Herramienta para proceso de Minería de Datos.
3.Generar modelos para la identificación de factores y patrones de comportamiento	120	2	10	65	25	Rapid Miner, TexMaker	Árboles, Clasificación, Reglas de Asociación y Selección de Atributos	Análisis de Fuente de Datos, Datos Limpios.	Modelos para identificar factores de deserción y reprobación.
4.Evaluar modelo generado y análisis de resultados	486	3	25	60	15	TexMaker		Modelo seleccionado	Informe con el conocimiento adquirido, Informe con las conclusiones del Proyecto, Memoria del Trabajo de Titulación.

3.1.2. Segunda Fase: Comprensión de los Datos

En este capítulo se presentan las diferentes actividades realizadas con el objetivo de familiarizarse con los datos, para posteriormente, comprender el problema y tener los conocimientos suficientes para emprender con seguridad la siguiente fase, que consiste en la preparación de los datos.

3.1.2.1. Tarea Uno: Recolección Inicial de los Datos

En esta tarea se decidió que datos se van a utilizar para realizar el proceso de minería de datos, y se concluyó que se trabajara con información personal del estudiante, registros académicos de los mismos y servicios recibidos por la institución. Y en cuanto al origen de esta información se consideró el seleccionar los datos de una sola fuente para no complicar demasiado la tarea de formar bases de datos, pues no es lo mismo trabajar con una única fuente de datos, que formar una base de datos a partir de distintas fuentes ya que esta segunda opción extendería la carga de trabajo al tener que unificar todo en una sola base de datos.

Tomando en cuenta las consideraciones anteriores se decidió trabajar con información almacenada en la base de datos del Sistema de Gestión Académica de la Universidad Nacional de Loja, la cual es administrada por el departamento Unidad de Telecomunicaciones e Información, la misma que proporciono un acceso a esta base de datos a través de una interfaz Web Services. Los datos con cuales se van a trabajar pertenecen a estudiantes del Área de la Energía de las Industrias y los Recursos Naturales No Renovables que cuenta con un número de 2821 estudiantes los cuales están distribuidos entre las carreras: Ingeniería en Electromecánica, Ingeniería en Electrónica y Telecomunicaciones, Ingeniería en Geología Ambiental y Ordenamiento Territorial y la de mayor número de estudiantes Ingeniería en Sistemas.

Los datos almacenados se encuentran distribuidos en categorías los cuales a su vez contienen métodos y que parámetros se requieren para obtener los datos, a continuación se describen las categorías que se encuentran en el Web Services.

- **Académica:** contiene información académica como datos de estudiantes y docentes.

- **Institucional:** contiene información institucional como datos de áreas, carreras, módulos y paralelos.
- **Personal:** contiene información personal de docentes y estudiantes.
- **Validación:** contiene métodos o servicios relacionados con la validación de docentes y estudiantes.
- **Estadística:** contiene información estadística como número de estudiantes matriculados, estudiantes aprobados y estudiantes reprobados.

Para obtener los datos se procedió a desarrollar una aplicación que permita consumir los datos del Web Services, y estos a su vez sean almacenados en una base de datos SQL. El explorador de Base de Datos seleccionado para la administración de la base de datos fue DatAdmin y este será usado a lo largo del Trabajo de Titulación.

En cuanto a datos del Área de Bienestar Universitario, estos se encuentran almacenados en archivos digitales y en papel como libros, informes mensuales, los cuales se procedieron a introducir en archivos digitales para su posterior tratamiento e integración con datos del sistema de gestión académica.

3.1.2.2. Tarea Dos: Descripción de los datos

En esta tarea se procede a describir los datos adquiridos en su formato original, como en la mayor parte de los casos estos datos tuvieron que ser tratados para poder formar con ellos una base de datos coherente y consistente, la cual permita trabajar en todo el desarrollo del Trabajo de Titulación. Como se mencionó anteriormente los datos fueron obtenidos de varias fuentes como el Área de Bienestar Universitario y el Sistema de Gestión Académica de la Universidad Nacional de Loja, los datos con los que se cuentan corresponden desde el año 2008, pero también se cuenta con datos históricos desde el año 2003, a continuación se muestran los datos obtenidos.

Para obtener los datos se emplearon algunos métodos y parámetros que permiten obtener los datos a continuación se describen los utilizados:

- **Categoría Académica:**

De esta categoría se obtuvo los datos de estudiantes como notas académicas para ello se usó los siguientes métodos y los parámetros requeridos (ver tabla XXXVII).

TABLA XXXVII:
MÉTODOS UTILIZADOS PARA OBTENER INFORMACIÓN DE CATEGORÍA ACADÉMICA.

Método	Parámetros
sga_periodos_lectivos()	
sgaws_egresados()	$fecha
sgaws_notas_estudiante()	$cedula, $idCarrera, $idOferta
sgaws_ofertas_academicas()	$id_periodo

- **Categoría Institucional**

De la presente categoría se obtuvo datos respecto a las áreas de la universidad, como carreras, lista de áreas, módulos y paralelos creados en cada periodo. A continuación se muestran los métodos utilizados de esta categoría para obtener dicha información (ver tabla XXXVIII).

TABLA XXXVIII:
MÉTODOS UTILIZADOS PARA OBTENER INFORMACIÓN DE CATEGORÍA INSTITUCIONAL.

Método	Parámetros
sgaws_carreras()	$sigla_carrera
sgaws_datos_carrera()	$id_oferta
sgaws_lista_areas()	
sgaws_modulos_carrera()	$oferta_id, $carrera_id
sgaws_paralelos_carrera()	$oferta_id, $carrera_id
sgaws_paralelos_carreras()	

- **Categoría Personal**

De esta categoría se obtuvo información personal de estudiantes del Área de Energía de la Universidad Nacional de Loja, datos como nombres, teléfono, lugar de origen, dirección, fecha de nacimiento, estado civil, etc. A continuación se muestran los métodos utilizados de esta categoría (ver tabla XXXIX).

TABLA XXXIX:
MÉTODOS UTILIZADOS PARA OBTENER INFORMACIÓN DE CATEGORÍA PERSONAL.

Método	Parámetros
sgawd_datos_estudiante()	$cedula

Los datos obtenidos por el servicio Web se visualizaron como texto plano en una página del navegador por lo que se tuvo que hacer una limpieza en cuanto a eliminar caracteres

innecesarios como corchetes, llaves y comillas, ya que posteriormente la información será almacenada en una base de datos.

Esta base de datos que se creó, se almacenó la información obtenida a través del Web Services, que previamente paso un tratamiento de limpieza, la base de datos contiene las siguientes tablas.

- *oferta_academica:* contiene información acerca de las ofertas que se han generado desde el año 2008 por lo general estas ofertas tienen una duración aproximada de seis meses.
- *oferta_carrera:* contiene información acerca de todas las ofertas que se han generado para cada una de las carreras de la Universidad Nacional de Loja.
- *modalidad:* contiene las tres modalidades de estudio que se ofrecen.
- *titulación:* contiene descripción de todas las titulaciones que se proporcionan a los estudiantes.
- *carrera:* contiene información acera de las carreras que ofrece la Universidad Nacional de Loja.
- *Área:* contiene información básica de los directivos de las áreas de la Universidad Nacional de Loja.
- *nota_unidad:* contiene notas académicas que se han obtenido en todas las materias cursadas por los estudiantes.
- *paralelo:* contiene información de todos los paralelos que se han generado en módulos, carreras y periodos.
- *modulo_oferta_carrera:* contiene información acerca de todos los módulos que se ofrecen en cada oferta académica.
- *modulo:* contiene el nombre de todos los módulos de las distintas carreras que ofrece la Universidad Nacional de Loja.
- *unidad:* contiene nombres de las materias, que se ofrecen y el modulo en el que se imparten.
- *estudiante:* contiene información personal del estudiante como: nombres, apellidos, fecha de nacimiento, teléfono, etc.
- *género:* contiene los géneros masculino y femenino.
- *estudiante_paralelo:* contiene información de estudiantes y los paralelos a los cuales pertenece.

- *periodo_academico:* contiene información de los periodos académicos hasta la presente fecha.
- *reprobado_oferta:* contiene el número de estudiantes reprobados en cada oferta academica.
- *aprobado_oferta:* contiene el número de estudiantes aprobados en cada oferta academica.
- *matriculado_oferta:* contiene el número de estudiantes matriculados en cada oferta academica.
- *matriculado_area:* contiene el número de estudiantes matriculados por cada área de la universidad nacional de Loja.
- *aprobado_area:* contiene el número de estudiantes aprobados por cada área de la universidad nacional de Loja.
- *reprobado_area:* contiene el número de estudiantes reprobados en cada área de la universidad nacional de Loja.
- *reprobado_carrera:* contiene el número de estudiantes reprobados en cada carrera.
- *matriculado_carrera:* contiene el número de estudiantes matriculados en cada carrera.
- *aprobado_carrera:* contiene el número de estudiantes aprobados en cada carrera.
- *aprobado_modulo:* contiene el número de estudiantes aprobados en cada módulo.
- *aprobado_paralelo:* contiene el número de estudiantes aprobados por cada paralelo.
- *matriculado_paralelo:* contiene el número de estudiantes matriculados en un paralelo.
- *reprobado_paralelo:* contiene el número de estudiantes reprobados en cada paralelo.
- *reporte_matricula:* contiene información acerca del rendimiento del estudiante en cada módulo que se ha matriculado.

A continuación se describe la estructura de cada tabla que se encuentra en la base de datos generada (ver tabla XL a LXVIII).

TABLA XL:
ESTRUCTURA DE TABLA OFERTA_ACADEMICA.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de oferta académica	integer	49 registros	PK
nombre	Nombre de la oferta académica	varchar	49 diferentes ofertas	
fecha_inicio	Fecha de inicio de oferta	datetime		
fecha_fin	Fecha de finalización de oferta	datetime		

TABLA XLI:
ESTRUCTURA DE TABLA OFERTA_CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de registro	integer	1283 registros	PK
oferta_id_fk	Identificador de oferta académica	integer	1283 registros	FK->oferta_academica
carrera_id_fk	Identificador de carrera	integer	1283 registros	FK->carrera

TABLA XLII:
ESTRUCTURA DE TABLA MODALIDAD.

Nombre	Descripción	Tipo	Rango	Keys
modalidad	Nombre de modalidad de estudio	varchar	3 tipos de modalidad	PK

TABLA XLIII:
ESTRUCTURA DE TABLA TITULACIÓN.

Nombre	Descripción	Tipo	Rango	Keys
titulacion	Nombre de la titulación	varchar	5 tipos de titulación	PK

TABLA XLIV:
ESTRUCTURA DE TABLA CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de la carrera	Integer	142 registros	PK
nombre	Nombre completo de la carrera	Varchar	1 a 142 nombres	
especialidad	Nombre completo de especialidad	Varchar	142 tipos de especialidad	

modalidad_id_fk	Modalidad de estudio	Varchar	3 tipos de modalidades	FK->modalidad
titulacion_id_fk	Nombre completo de titulación	Varchar	5 tipos de titulación	FK->titulacion
Área_id_fk	Siglas del área que pertenece	Varchar	8 tipos de áreas	FK->Área
costo	Valor de cursar carrera	float	0 a 500	

TABLA XLV:
ESTRUCTURA DE TABLA ÁREA

Nombre	Descripción	Tipo	Rango	Keys
sigla	Siglas de del área.	varchar	8 registros	PK
nombre_Área	Nombre completo del área.	varchar	8 áreas de la universidad	
secretario	Nombre del Secretario de área.	varchar	1 a 8 nombres	
director	Nombre del Director de área.	varchar	1 a 6 nombres	

TABLA XLVI:
ESTRUCTURA DE TABLA NOTA_UNIDAD.

Nombre	Descripción	Tipo	Rango	Keys
unidad_id_fk	Nombres de materias	varchar	48487 registros	FK->unidad
nota	Notas obtenidas en materias	float	0 a 10	
estudiante_id_fk	Número de cedula de estudiante	varchar	48487 registros	FK->estudiante
oferta_carrera_id_fk	Identificador de oferta académica	integer	48487 registros	FK->oferta_carrera

TABLA XLVII:
ESTRUCTURA DE TABLA PARALELO.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de paralelo	integer	12695 registros	PK
seccion	No contiene valores	varchar		
número	Número de paralelo	varchar		

nombre	Nombre de paralelo	varchar		
modulo_id_fk	Identificador de modulo	integer	12695 registros	
oferta_carrera_id_fk	Identificador de carrera	integer	12695 registros	FK->oferta_carrera

TABLA XLVIII:
ESTRUCTURA DE TABLA MODULO_OFERTA_CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de registro	integer	6094 registros	PK
modulo_id_fk	Identificador de modulo	integer	6094 registros	FK->modulo
oferta_carr_id_fk	Identificador de carrera	integer	6094 registros	FK->oferta_carrera

TABLA XLIX:
ESTRUCTURA DE TABLA MODULO.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de modulo	integer	468 registros	PK
nombre	Nombre completo de modulo	varchar	468 nombres de modulo	

TABLA L:
ESTRUCTURA DE TABLA UNIDAD.

Nombre	Descripción	Tipo	Rango	Keys
nombre	Nombre de cada materia	varchar	567 registros	PK
modulo_id_fk	Identificador de modulo	integer	567 registros	FK->modulo

TABLA LI:
ESTRUCTURA DE TABLA ESTUDIANTE.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de estudiante	integer	3438 registros	
numeroIdentificacion	Número de cedula de estudiante	varchar	3438 números de cedula	PK
nombres	Nombres de estudiante	varchar	3438 nombres	
apellidos	Apellidos de estudiante	varchar	3438 apellidos	
fecha_nacimiento	Fecha de nacimiento de estudiante	datetime		

telefono	Número de teléfono del estudiante	varchar	1 a 3438	
celular	Número de celular del estudiante	varchar	1 a 3438	
direccion	Dirección de domicilio de estudiante	varchar	1 a 3438	
pais	País de origen de estudiante	varchar	1 a 3438	
provincia	Provincia de origen de estudiante	varchar	1 a 3438	
email	Correo electrónico de estudiante	varchar	1 a 3438	
genero_id_fk	Género de estudiante	varchar	2 tipos de género	FK->género
es_egresado	Valor para describir si estudiante es egresado	integer	2 tipos de valores (1 y 0)	

TABLA LII:
ESTRUCTURA DE TABLA GÉNERO.

Nombre	Descripción	Tipo	Rango	Keys
nombre	Nombre de género	varchar	2 tipos de género	PK

TABLA LIII:
ESTRUCTURA DE TABLA ESTUDIANTE_PARALELO.

Nombre	Descripción	Tipo	Rango	Keys
estudiante_id_fk	Número de cedula de estudiante	varchar	11418 registros	FK->estudiante
paralelo_id_fk	Identificador de paralelo	integer	11418 registros	FK->paralelo

TABLA LIV:
ESTRUCTURA DE TABLA PERIODO_ACADEMICO.

Nombre	Descripción	Tipo	Rango	Keys
id	Identificador de periodo académico	integer	11 registros	PK
fecha_periodo	Año de periodo académico	varchar	11 registros	

TABLA LV:
ESTRUCTURA DE TABLA REPROBADO_OFERTA.

Nombre	Descripción	Tipo	Rango	Keys
numero_reprob	Número de estudiantes reprobados	integer	0 registros	
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LVI:
ESTRUCTURA DE TABLA APROBADO_OFERTA.

Nombre	Descripción	Tipo	Rango	Keys
numero_aprobado	Número de estudiantes aprobados	integer	0 registros	
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LVII:
ESTRUCTURA DE TABLA MATRICULADO_OFERTA.

Nombre	Descripción	Tipo	Rango	Keys
num_matric	Número de matricula	integer	0 registros	
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LVIII:
ESTRUCTURA DE TABLA MATRICULADO_ÁREA.

Nombre	Descripción	Tipo	Rango	Keys
num_matriculado	Número de matricula	integer	0 registros	
Área_id_fk	Identificador del Área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LIX:
ESTRUCTURA DE TABLA APROBADO_ÁREA.

Nombre	Descripción	Tipo	Rango	Keys
numero_aprob	Número de aprobados	integer	0 registros	
Área_id_fk	Identificador del Área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LX:
ESTRUCTURA DE TABLA REPROBADO_ÁREA

Nombre	Descripción	Tipo	Rango	Keys
numero_aprob	Número de aprobados	integer	0 registros	
Área_id_fk	Identificador del área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LXI:
ESTRUCTURA DE TABLA REPROBADO_CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
numero_reprob	Número de reprobados	integer	632 registros	
Área_id_fk	Identificador del Área	integer	632 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	632 registros	FK->oferta

TABLA LXII:
ESTRUCTURA DE TABLA MATRICULADO_CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
numero_matriculado	Número de matriculado	integer	0 registros	
Área_id_fk	Identificador del Área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LXIII:
ESTRUCTURA DE TABLA APROBADO_CARRERA.

Nombre	Descripción	Tipo	Rango	Keys
numero_aprob	Número de aprobados	integer	0 registros	
Área_id_fk	Identificador del Área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LXIV:
ESTRUCTURA DE TABLA APROBADO_MODULO.

Nombre	Descripción	Tipo	Rango	Keys
num_aprobados	Número de aprobados	integer	0 registros	
Área_id_fk	Identificador del Área	integer	0 registros	FK->Área
oferta_id_fk	Identificador de la oferta	integer	0 registros	FK->oferta

TABLA LXV:
ESTRUCTURA DE TABLA APROBADO_PARALELO.

Nombre	Descripción	Tipo	Rango	Keys
numero_aprob	Número de aprobados	integer	0 registros	
paralelo_id_fk	Identificador del paralelo	integer	0 registros	FK->paralelo

TABLA LXVI:
ESTRUCTURA DE TABLA MATRICULADO_PARALELO.

Nombre	Descripción	Tipo	Rango	Keys
num_matriculado	Número de matriculados	integer	0 registros	
paralelo_id_fk	Identificador del paralelo	integer	0 registros	FK->paralelo

TABLA LXVII:
ESTRUCTURA DE TABLA REPROBADO_PARALELO.

Nombre	Descripción	Tipo	Rango	Keys
num_reprob	Número de reprobados	integer	0 registros	
paralelo_id_fk	Identificador del paralelo	integer	0 registros	FK->paralelo

TABLA LXVIII:
ESTRUCTURA DE TABLA REPORTE_MATRICULA.

Nombre	Descripción	Tipo	Rango	Keys
Id	Identificador de reporte	integer	10637 registros	PK
nota_promedio	Nota de modulo	varchar	10637 registros	
asistencia_promedio	Promedio de asistencias en cada modulo	double	10637 registros	
estudiante_id_fk	Identificador de estudiante	varchar	10637 registros	FK->estudiante
oferta_id_fk	Identificador de oferta	integer	10637 registros	FK->oferta
estado	Estado de matricula	varchar	10637 registros	

La siguiente fuente de información es el Área de Bienestar Universitario que comprende cinco áreas de servicio las cuales son: Psicopedagógico, Becas, Salud, Defensa de los derechos estudiantiles, Programa para estudiantes en estado de gestación, Programa de atención para estudiantes con capacidades diferentes y el Infocentro Universitario. Los datos obtenidos del Área de Bienestar Universitario se almacenan en informes

mensuales que posteriormente se adhieren a un libro que describe las atenciones realizadas en un año, a continuación se describe los datos obtenidos:

Servicio de Salud: en este servicio se obtuvo acceso a los libros de cada año, de los cuales se obtuvieron los siguientes datos (ver tabla LXIX):

TABLA LXIX:
DATOS OBTENIDOS DEL ÁREA SALUD.

Atributo	Descripción
Cedula	Número de cedula de estudiante
Nombres	Nombres de estudiante
Apellidos	Apellidos de estudiante
Edad	Edad de atención del estudiante
Carrera de estudio	Carrera que cursa el estudiante
Módulo	Módulo de atención
Servicio	Servicio que solicito

Servicio de Becas: en esta área de servicio la información se gestionaba en archivos lo cual facilito la recopilación, los datos encontrados son los siguientes (ver tabla LXX):

TABLA LXX:
DATOS OBTENIDOS DEL ÁREA BECAS.

Atributo	Descripción
Cedula	Número de cedula de estudiante
Nombres	Nombres de estudiante
Apellidos	Apellidos de estudiante
Carrera de estudio	Carrera que cursa el estudiante
Duración de beca	El periodo de vigencia de la beca
Tipo de beca	Tipo de beca solicitada
Observaciones	Observaciones de la beca

Servicio de Psicopedagógico: en este servicio se procedió a extraer la información de los libros los cuales son (ver tabla LXXI):

TABLA LXXI:
DATOS OBTENIDOS DEL ÁREA PSICOPEDAGÓGICO.

Atributo	Descripción
Cedula	Número de cedula de estudiante
Nombres	Nombres de estudiante
Apellidos	Apellidos de estudiante
Edad	Edad de atención del estudiante
Carrera	Carrera que cursa el estudiante
Módulo	Módulo de atención

Servicio de Estudiantes en estado de gestación: de este servicio se pudo obtener los siguientes datos (ver tabla LXXII):

TABLA LXXII:
DATOS OBTENIDOS DEL PROGRAMA PARA ESTUDIANTES EN ESTADO DE GESTACIÓN.

Atributo	Descripción
Cedula	Número de cedula de estudiante
Nombres	Nombres de estudiante
Apellidos	Apellidos de estudiante
Carrera	Carrera que cursa el estudiante
Módulo	Módulo de gestación
Observaciones	Observaciones de seguimiento
Año	Periodo del estado de gestación

Además de los datos recopilados del Área de Bienestar Universitario y del Sistema de Gestión de Académica, existen datos que no disponibles a través del Web Services, por ese motivo se procedió a solicitar al departamento de UTI los siguientes datos (ver tabla LXXIII):

TABLA LXXIII:
DATOS OBTENIDOS DEL DEPARTAMENTO UTI.

Atributo	Descripción
Cedula	Número de cedula de estudiante
Cantón	Cantón de procedencia del estudiante
Dirección Actual	Dirección actual de residencia
Dirección de Procedencia	Dirección de origen de estudiante
Estado Civil	Estado civil de estudiante
Etnia	Etnia de estudiante
Fecha de nacimiento de madre	Fecha de nacimiento de la madre
Fecha de nacimiento de padre	Fecha de nacimiento del padre
Lugar de trabajo de madre	Lugar de trabajo de la madre
Lugar de trabajo de padre	Lugar de trabajo del padre
Nombres de madre	Nombres de la madre
Nombres de padre	Nombres del padre
Nombres de cónyuge	Nombres del cónyuge
Número de hijos	Número de hijos que tiene el estudiante
Tipo de sangre	Tipo de sangre del estudiante

A continuación se muestra el modelo de la base de datos generada para almacenar la información obtenida (ver figura 32).

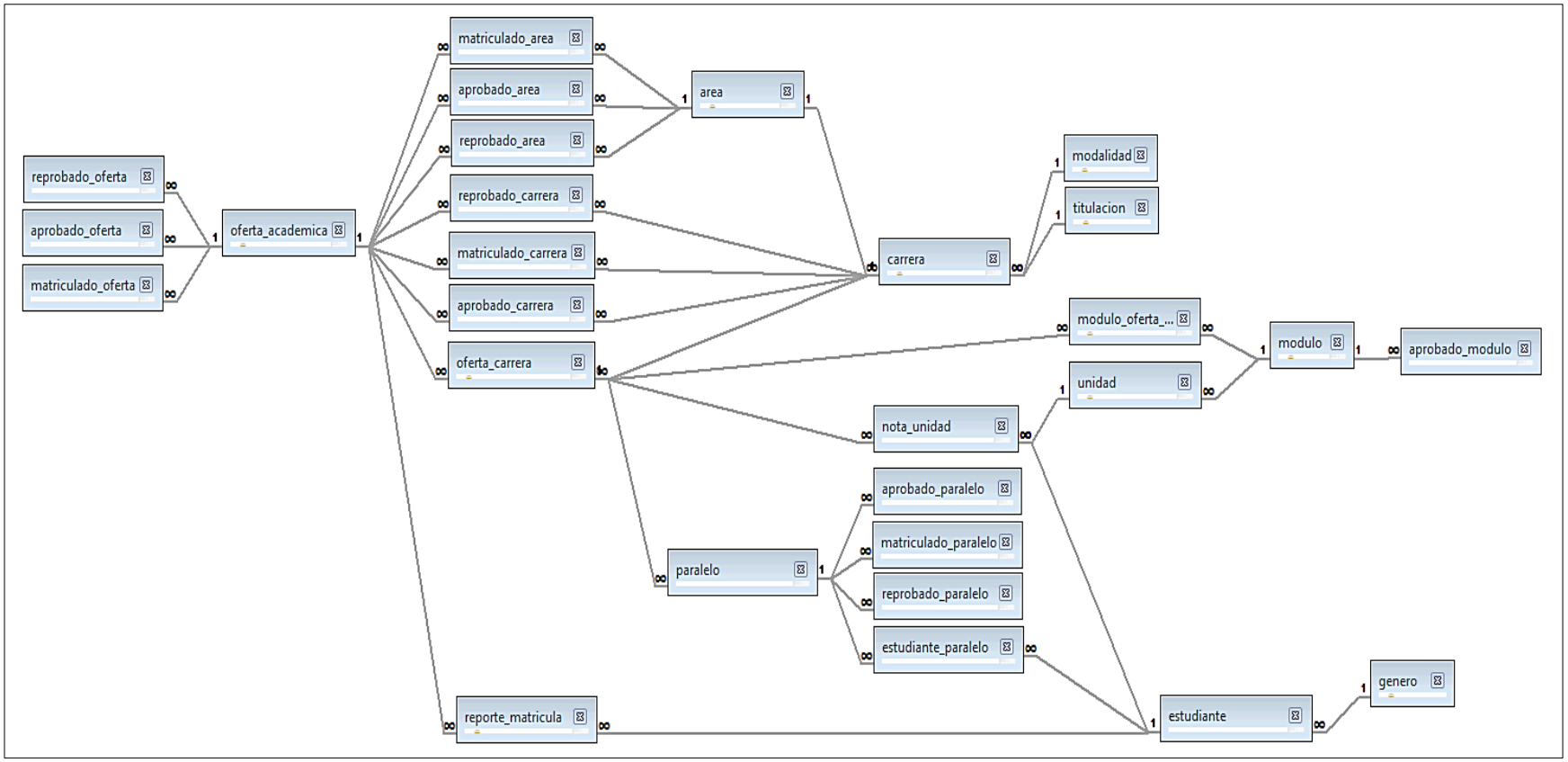


Figura 32: Modelo de base de datos generada.

3.1.2.3. Tarea Tres: Exploración de los datos

En esta tarea se lleva a cabo la exploración de los datos para familiarizarse con los mismos. Estos pueden describirse mediante gráficos para indicar las características de los datos, además en esta etapa no se realiza ningún tratamiento de los datos.

Con respecto a datos estadísticos, el Área de la Energía las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja, la carrera que ha recibido el mayor número de estudiantes es la de Ingeniería en Sistemas (ver tabla LXXIV, ver figura 33).

TABLA LXXIV:
NÚMERO DE ESTUDIANTES POR CARRERAS.

Carrera	Nro. Estudiantes
Ingeniería en Electromecánica	3417
Ingeniería en Electrónica y Telecomunicaciones	1007
Ingeniería en Geología Ambiental y Ordenamiento Territorial	1499
Ingeniería en Sistemas	4695
Tecnología en Electricidad y Control Industrial	735
Tecnología en Electrónica	60

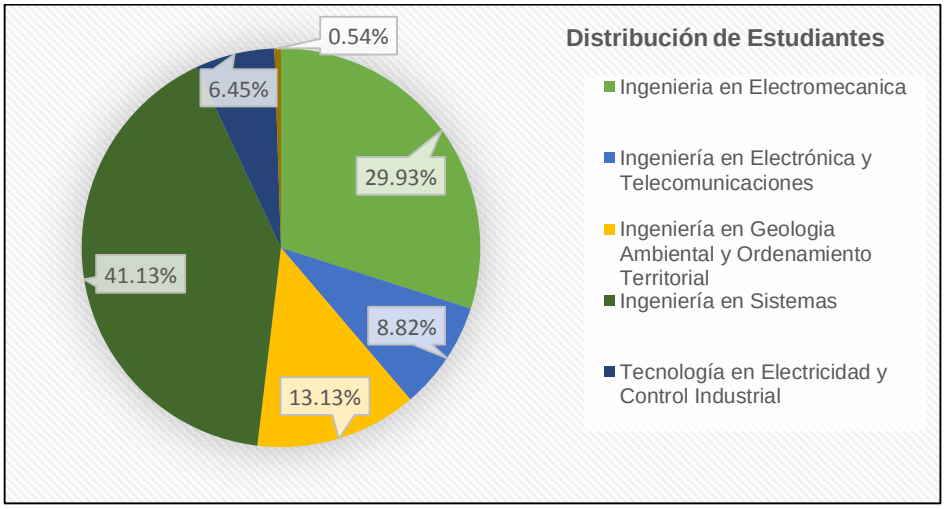


Figura 33: Distribución de estudiantes por Carreras.

De los estudiantes del Área de Energía, las Industrias y los Recursos Naturales No Renovables, a continuación se describe la distribución por género de cada uno de ellos (ver tabla LXXV, ver figura 34).

TABLA LXXV:
SEXO DE ESTUDIANTES.

Sexo	Nro.
Masculino	2199
Femenino	622
Total:	2821

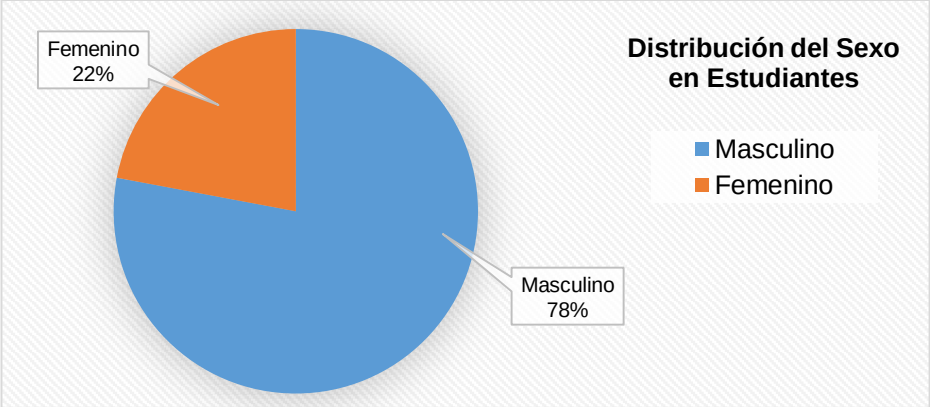


Figura 34: Distribución de sexo de estudiantes del Área AEIRNNR.

El número de estudiantes aprobados y reprobados se obtuvo a través de datos estadísticos los cuales describen la cantidad de reprobados y aprobados por cada periodo académico, el periodo que presenta más estudiantes reprobados es el periodo Septiembre 2008 – Febrero 2009 (ver tabla LXXVI, ver figura 35).

TABLA LXXVI:
PERÍODOS DEL ÁREA AEIRNNR.

Periodo	Aprobados	Reprobados
Septiembre 08 - Febrero 09	891	380
Marzo 09 - Julio 09	729	192
Septiembre 09 - Febrero 10	902	308
Marzo 10 - Julio 10	843	207
Septiembre 10 - Febrero 11	918	215
Marzo 11 - Julio 11	812	195
Septiembre 11 - Febrero 12	790	231
Marzo 12 - Julio 12	679	175
Septiembre 12 - Febrero 13	493	200
Marzo 13 - Julio 13	549	221

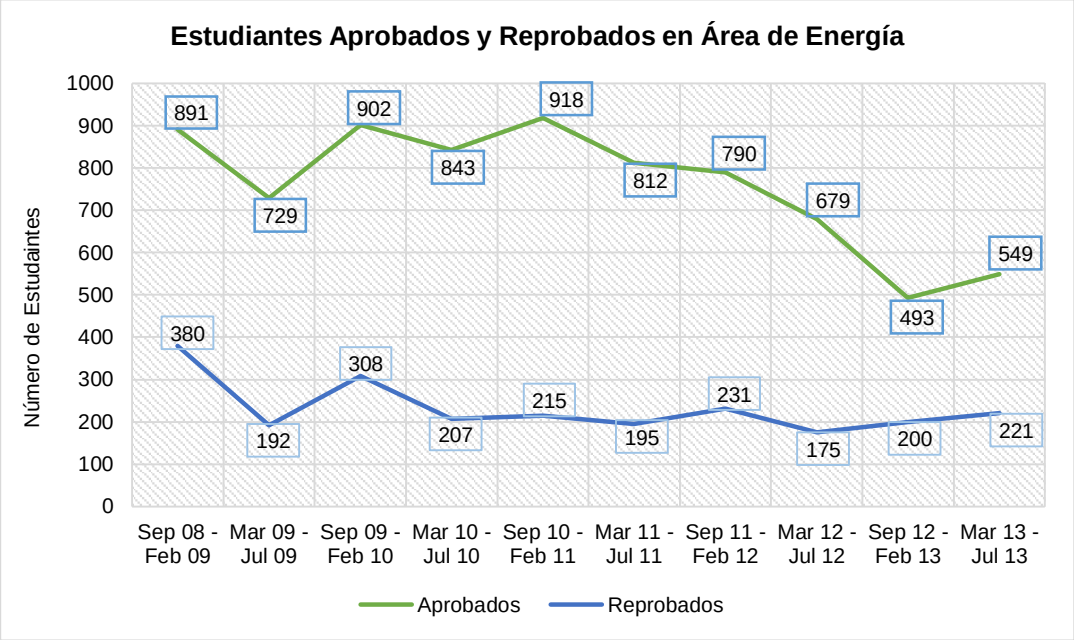


Figura 35: Distribución de estudiantes aprobados y reprobados en cada periodo.

Con respecto a los datos recopilados del Área de Bienestar Universitario a continuación se describe el número de estudiantes del Área de Energía, que recibieron cada uno de los servicios (ver tabla LXXVII, ver figura 36).

TABLA LXXVII:
DISTRIBUCIÓN DE SERVICIOS PARA ESTUDIANTES DEL ÁREA DE ENERGÍA.

Servicio	Número de Estudiantes
Salud	181
Psicopedagógico	65
Becas	140
Programa para estudiantes en estado de gestación	18

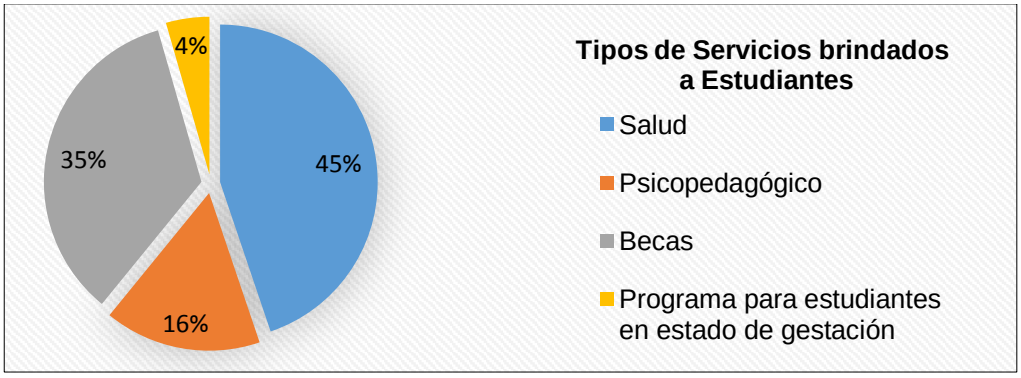


Figura 36: Distribución de servicios a estudiantes.

3.1.2.4. Tarea Cuatro: Verificación de la Calidad de los Datos

En esta tarea se examina la calidad de los datos con el objetivo de analizar si estos están completos, son correctos o tienen errores y en caso de tenerlos que tan frecuentes son. Durante la tarea anterior de exploración se pudo observó que existían problemas de calidad de los mismos y también se encontraron campos con datos nulos.

Algunos problemas encontrados fueron:

- Problema al instante de extraer la información del Web Services, puesto que el identificador de estudiantes, la columna *estudiante_id_fk*, de la tabla *estudiante_paralelo*, contiene números de cedula de aproximadamente 2700 estudiantes y se encontraron caracteres que describían el nombre de los mismos. Posible Solución: corregir estos errores obteniendo el número de cedula de estos estudiantes de la tabla estudiante y modificando manualmente los campos incorrectos.
- En el periodo académico septiembre 2012 - febrero 2013, que cursaron algunos estudiantes de la carrera Electromecánica se encuentran incompletos. Posible Solución: identificar el método del Web Services para obtener los datos correspondientes a ese periodo.
- Otro problema ocurrido en el desarrollo del Trabajo de Titulación fue la recopilación de datos almacenados en el Área de Bienestar Universitario, puesto que actualmente no cuenta con un sistema que permita gestionar dicha información, por lo cual se tuvo que obtener dicha información. Posible Solución: obtener la información almacenada y luego filtrarla para optimizar tiempo en el traslado de la misma y su integración con la base de datos.

Los errores fueron encontrados analizando los datos de la base de datos generada, estos hay que tenerlos en cuenta para que no derivasen en futuros problemas.

3.2. Realizar la Selección, Limpieza y Transformación de los datos para obtener una vista minable.

En esta actividad se realizaron tareas en cuanto a preparación de datos, seleccionar cuales serán útiles y la construcción de estructuras de minería para la aplicación de técnicas en la siguiente fase.

3.2.1. Tercera Fase: Preparación de los Datos

En esta fase se describen las actividades que se realizaron para la preparación del conjunto final de datos, la selección y limpieza de los datos, para luego de esto proceder con la construcción los datos a los cuales previamente se modifican. Es decir, es en esta etapa en donde se prepara todo lo relativo a la base de datos, para poder luego ingresarlas en la herramienta de modelado.

3.2.1.1. Tarea Uno: Selección de los Datos

En la etapa anterior se describieron los datos iniciales y las categorías en las que se encontraba la información almacenada, además se detallan los datos que fueron seleccionados y los que fueron excluidos.

Como se describió, en apartados anteriores, el dominio del problema estaba centrado en identificar factores de deserción y reprobación en estudiantes del Área de Energía las Industrias y los Recursos Naturales No Renovables, por lo tanto, el primer filtro que se debió aplicar fue el que eliminar los datos de estudiantes que no pertenecían a dicha área. Esto ocurrió al momento de extraer la información a través del Web Services de la Universidad Nacional de Loja, resultando con un número aproximado de 2821 estudiantes con sus registros.

Se procedió a eliminar también tablas con datos estadísticos obtenidos, estas contienen datos como, número de estudiantes aprobados y reprobados en la Universidad Nacional de Loja, las tablas descartadas son: reprobado_oferta, aprobado_oferta, matriculado_oferta, matriculado_area, aprobado_area, reprobado_area, reprobado_carrera, matriculado_carrera, aprobado_carrera, aprobado_modulo, aprobado_paralelo, matriculado_paralelo y reprobado_paralelo de esta forma quedo

reducido el número de tablas de la base de datos inicial con la cual se contaba, para luego en las áreas posteriores realizar correspondiente depuración de los datos.

A continuación se muestra (ver figura 37) la estructura de la base de datos después de eliminar las tablas con datos estadísticos.

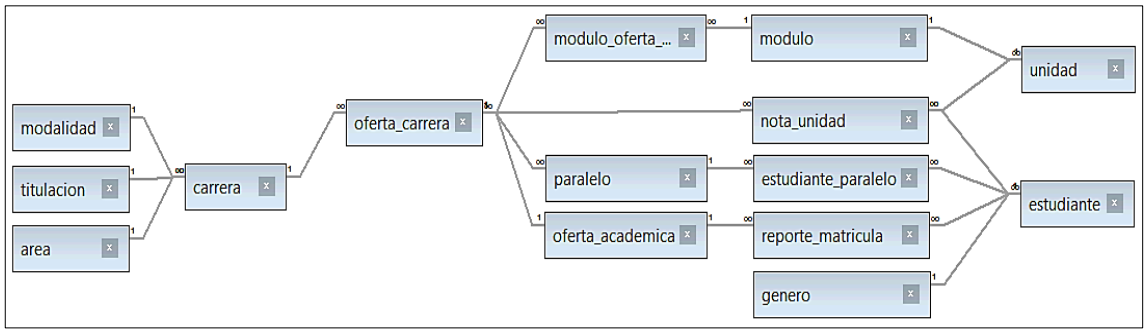


Figura 37: Estructura de la base de datos depurada.

3.2.1.2. Tarea Dos: Limpieza de los Datos

En esta tarea se describe como se mejoró la calidad de los datos seleccionados en la tarea anterior. Como se observó en la tarea verificación de la fase dos se encontraron problemas con los datos, es por ello que en esta tarea se les aplicó una solución.

El primer problema encontrado fueron datos incorrectos en la columna identificador de la tabla estudiante_paralelo, la cual debería contener los números de cedula de estudiantes, se identificó que algunos que tenían como identificador nombres o apellidos para solucionar este problema se procedió a buscar los números de cedula de estos estudiantes para remplazarlos manualmente.

Otro problema que se descubrió era que algunos estudiantes de la carrera Electromecánica no contenían información acerca del periodo Septiembre 2012 – Febrero 2013, para dar solución a este problema se volvió a acceder al Web Services, para luego proceder a bajar la información acerca del periodo que faltaba.

Además los estudiantes de la carrera Ingeniería en Electrónica y Telecomunicaciones se encontraban sin registro de los periodos: Marzo 2012 – Julio 2012, Septiembre 2012 – Febrero 2013, Marzo 2013 – Julio 2013, por lo cual aplico la solución al problema anterior recuperando dichos periodos ausentes.

Se eliminó la columna sección de la tabla paralelo ya que esta contenía datos nulos, por lo cual era innecesario su uso para el proceso de minería de datos, también se realizaron otras áreas de limpieza en la base de datos como descripción de periodos académicos y números de cedula de estudiantes, además se eliminó los registros creados al momento de realizar las pruebas del sistema cuando este se implementó y también se eliminó registros duplicados de la tabla estudiantes.

Culminando así con las áreas mencionadas anteriormente se puede decir que se han limpiado los datos, pero no de forma definitiva, puesto que como se ha explicado anteriormente este es un proceso iterativo, pero con lo realizado hasta el momento permitió continuar con la construcción de los datos.

3.2.1.3. Tarea Tres: Construcción de Datos

En esta tarea se realizaron más actividades de limpieza sobre los datos, para su utilización en la tarea posterior de integración para ello se crearon atributos derivados o se alteraron valores de atributos ya existentes.

Se ha preparado una estructura de minería de datos, considerando los siguientes atributos descritos en la siguiente tabla (ver tabla LXXVIII), con el objetivo de aplicar diferentes modelos de minería de datos, y evaluar el comportamiento de cada modelo, esto con el fin de identificar los factores de deserción universitaria.

TABLA LXXVIII:
ESTRUCTURA DE MINERÍA DE DATOS PARA IDENTIFICAR FACTORES DE DESERCIÓN.

Atributo	Tipo de Datos	Tipo de Contenido	Valores
cedula	integer	continuo	
edad_de_ingreso	integer	continuo	16 – 52
estado	varchar	discreto	▪ desertor ▪ egresado
modulos_reprobados	integer	continuo	0 – 10
periodo_reprobacion	varchar	discreto	▪ 0 ▪ 1-3 ▪ 4-7 ▪ 8-11
servicios	varchar	discreto	▪ 1 ▪ 2 ▪ 3 ▪ 4

sexo	varchar	discreto	▪ m ▪ f
distancia_origen	integer	discreto	▪ 1 ▪ 2 ▪ 3 ▪ 4
carrera	varchar	discreto	▪ IE ▪ IET ▪ IGAOT ▪ IS ▪ TECI ▪ TE
cambio_carrera	integer	discreto	▪ si ▪ no
promedio_asistencia	varchar	discreto	▪ bajo ▪ medio ▪ alto
promedio_notas	varchar	discreto	▪ malo ▪ regular ▪ bueno ▪ muy bueno ▪ excelente
tipo _beca	varchar	discreto	▪ a ▪ b ▪ c ▪ d ▪ ninguna
estado_gestacion	varchar	discreto	▪ si ▪ no
bienestar_servicios	varchar	discreto	▪ si ▪ no
estado_civil	varchar	discreto	▪ soltero ▪ casado ▪ divorciado ▪ union libre ▪ viudo
etnia	varchar	discreto	▪ indigena ▪ blanco ▪ mestizo ▪ montubio
hijos	varchar	discreto	▪ Si ▪ no
horario_estudio	varchar	discreto	▪ matutino ▪ vespertino

A continuación se describen los atributos mencionados en la tabla anterior algunos de estos atributos se derivaron del cálculo entre atributos de la misma base de datos.

- ***Cedula:*** representa el número de identificación de cada estudiante.
- ***Edad de ingreso:*** calculado de diferencia entre la fecha de nacimiento y la fecha de inicio del primer módulo que se matriculo.
- ***Estado:*** representa el estado actual del estudiante estos pueden ser egresado, desertor, estudiante.
- ***Número Módulos reprobados:*** este atributo es calculado con el número de módulos reprobados por cada estudiante en las carreras del Área de Energía.
- ***Periodo de Reprobación:*** representa el rango en el cual un estudiante reprobó un periodo académico (ver tabla LXXIX).

TABLA LXXIX:
DISCRETIZACIÓN DE ATRIBUTO PERIODO DE REPROBACIÓN.

Siglas	Descripción
0	El estudiante no reprobó ninguno de los módulos.
1-3	El estudiante reprobó entre los primeros módulos que van desde primero a tercero.
4-7	El estudiante reprobó entre los módulos intermedios que van desde cuarto a séptimo.
8-11	El estudiante reprobó entre los últimos módulos que van desde octavo a decimo.

- ***Servicios:*** representa los servicios de comunicación contratados por el estudiante y la representación para este atributo se describe a continuación en la siguiente tabla (ver tabla LXXX):

TABLA LXXX:
SERVICIOS CONTRATADOS POR EL ESTUDIANTE.

Siglas	Descripción
1	Fijo y Celular
2	Solo Fijo
3	Solo Celular
4	Ninguno de los servicios

- ***Sexo:*** este atributo representa el género de cada estudiante en donde el valor de “f” representa estudiantes femeninos y “m” a estudiantes masculinos.
- ***Distancia de origen:*** calculado a partir de la distancia entre el lugar de origen y la institución donde se encuentra realizando los estudios, a continuación se puede observar la representación de este atributo en la siguiente tabla (ver tabla LXXXI). A continuación se describe como se encuentra dividido el cantón Loja en las siguientes parroquias urbanas (San Sebastián, El Valle, Sagrario, Sucre) y las siguientes rurales (Yangana, Malacatos, Vilcabamba, Quinara, San Lucas,

Jimbilla, El Cisne, Santiago, Gualel, Taquil, Chuquiribamba, San Pedro de Vilcabamba, Chantaco) y los cantones de la provincia de Loja (Calvas, Catamayo, Célica, Chaguarpamba, Espíndola, Gonzanamá, Loja, Macara, Olmedo, Paltas, Pindal, Puyango, Quilanga, Saraguro, Sozoranga, Zapotillo) [72].

TABLA LXXXI:
DISCRETIZACIÓN DE ATRIBUTO DISTANCIA ORIGEN.

Siglas	Descripción
1	El origen es del sector urbano.
2	El origen es del sector rural.
3	El origen es de alguno cantón de la provincia de Loja.
4	El origen es de otra provincia.

- ***Carrera:*** representa la carrera en la cual se matriculo el estudiante al ingresar a la institución, la representación de este atributo se puede observar a continuación (ver tabla LXXXII):

TABLA LXXXII:
CARRERAS DEL ÁREA DE ENERGÍA.

Siglas	Descripción
IE	Ingeniería en Electromecánica.
IET	Ingeniería en Electrónica y Telecomunicaciones.
IGAOT	Ingeniería en Geología Ambiental y Ordenamiento Territorial.
IS	Ingeniería en Sistemas.
TECI	Tecnología en Electricidad y Control Industrial.
TE	Tecnología en Electrónica.

- ***Cambio de carrera:*** este atributo pretende describir si un estudiante durante el transcurro de sus estudios se cambió de carrera dentro del Área de Energía.
- ***Promedio de asistencia:*** este atributo representa el promedio de asistencias de cada estudiante a las clases impartidas en la institución (ver tabla LXXXIII).

TABLA LXXXIII:
DISCRETIZACIÓN DE ATRIBUTO PROMEDIO ASISTENCIAS.

Valor	Descripción
bajo	Promedio menor a 85.0 %
medio	Promedio entre 85.1% y 95.0%
alto	Promedio mayor a 95.0 %

- ***Promedio de notas:*** este atributo representa el promedio de notas de cada estudiante por cada periodo académico (ver tabla LXXXIV).

TABLA LXXXIV:
DISCRETIZACIÓN DE ATRIBUTO PROMEDIO NOTAS.

Siglas	Descripción
malo	Promedio menor a 7.4
regular	Promedio entre 7.5 y 8.4
bueno	Promedio entre 8.5 y 8.9
muy bueno	Promedio entre 9.0 y 9.5
excelente	Promedio mayor a 9.5

- ***Beca Tipo:*** este atributo describe si recibió o no algún tipo de beca por parte de del Área Bienestar Universitario (ver tabla LXXXV).

TABLA LXXXV:
TIPOS DE BECA QUE RECIBEN ESTUDIANTES.

Siglas	Descripción
a	Estudiantes recibieron beca tipo A
b	Estudiantes recibieron con beca tipo B
c	Estudiantes recibieron con beca tipo C
d	Estudiantes recibieron con beca tipo D
ninguna	Estudiante que no ha recibido ningún tipo de beca

- ***Estado Gestación:*** este atributo describe si una estudiante de sexo femenino estuvo en estado de gestación.
- ***Bienestar Servicios:*** este atributo describe si el estudiante recibió algunos de los servicios del Área Bienestar Universitario.
- ***Estado civil:*** este atributo describe el estado civil del estudiante.
- ***Etnia:*** este atributo describe la etnia de un estudiante (ver tabla LXXXVI).

TABLA LXXXVI:
ETNIA DEL ESTUDIANTE.

Siglas	Descripción
1	Indígena
2	Blanco
3	Mestizo
4	Montubio

A continuación se describe la estructura formada con el propósito de identificar los factores de reprobación universitaria (ver tabla LXXXVII).

TABLA LXXXVII:
ESTRUCTURA DE MINERÍA DE DATOS PARA IDENTIFICAR FACTORES DE REPROBACIÓN.

Atributo	Tipo de Datos	Tipo de Contenido	Valores
cedula	integer	continuo	
edad_de_ingreso	integer	continuo	16 – 52
servicios	varchar	discreto	▪ 1 ▪ 2 ▪ 3 ▪ 4
distancia_origen	varchar	discreto	▪ 1 ▪ 2 ▪ 3 ▪ 4
carrera	varchar	discreto	▪ IE ▪ IET ▪ IGAOT ▪ IS ▪ TECI ▪ TE
promedio_asistencia	varchar	discreto	▪ bajo ▪ medio ▪ alto
promedio_notas	varchar	discreto	▪ malo ▪ regular ▪ bueno ▪ muy bueno ▪ excelente
tipo _beca	varchar	discreto	▪ a ▪ b ▪ c ▪ d ▪ ninguna
bienestar_servicios	varchar	discreto	▪ si ▪ no
estado_civil	varchar	discreto	▪ soltero ▪ casado ▪ divorciado ▪ union libre ▪ viudo
hijos	varchar	discreto	▪ si ▪ no
padre_trabaja	varchar	discreto	▪ si ▪ no
madre_trabaja	varchar	discreto	▪ si ▪ no
horario_estudio	varchar	discreto	▪ matutino ▪ vespertino

reprobo	varchar	discreto	▪ si ▪ no

En la estructura presentada se removió los siguientes atributos *estado, modulos reprobados, periodo reprobación, sexo, estado gestación, etnia*, y se dio la inclusión de nuevos atributos denominado *padre trabaja*, *madre trabaja*, *reprobo* el cual define si un estudiante reprobó o no en sus estudios, a continuación se describen los nuevos atributos.

- ***Padre trabaja:*** este atributo describe si el padre de un estudiante se encuentra trabajando.
- ***Madre trabaja:*** este atributo describe si la madre de un estudiante se encuentra trabajando.
- ***reprobo:*** este atributo describe si un estudiante reprobó o no un periodo académico a lo largo de sus estudios hasta la actualidad.

3.2.1.4. Tarea Cuatro: Integración de datos

Una vez se inició esta tarea de integración se realizaron cambios pertenecientes a los áreas anteriores lo cual no supuso ningún inconveniente, pues ya se advirtió que el proceso de preparación de los datos seguía un ciclo iterativo.

La tarea de integración empezó con la base de datos que se ha venido trabajando hasta el momento. El primer paso fue crear tablas la primera corresponde a datos solicitados al departamento UTI y con respecto a los servicios que ofrece el Área de Bienestar Universitario las tablas agregadas son:

- Tabla bienestar_becas que contiene los siguientes atributos (*cedula, nombres, apellidos, carrera, duracion_beca, tipo_beca, observaciones*).
- Tabla bienestar_estudiantes_gestacion que contiene los siguientes atributos (cedula, nombres, apellidos, carrera, modulo_atencion, observaciones, año).
- Tabla bienestar_psicopedagogico que contiene los siguientes atributos (cedula_fk, nombres, apellidos, edad_atencion, carrera, modulo_atencion, trabajo_social).

- Tabla bienestar_salud que contiene los siguientes atributos (cedula_fk, nombres, apellidos, edad_atencion, carrera, modulo_atencion, laboratorio, enfermería, odontología, medico).
- Tabla estudiante_datos_personales que contienen los siguientes atributos (cedula, canton, dirección_actual, dirección_procedencia, estado_civil, etnia, fecha_nacimiento_madre, fecha_nacimiento_padre, lugar_trabajo_madre, lugar_trabajo_padre, nombres_conyuge, nombres_madre, nombres_padre, hijos, tipo_sangre).

Una vez creadas las tablas, se introdujeron los datos recopilados anteriormente del Área Bienestar Universitario, posteriormente para mejorar la calidad de estas se establecieron relaciones, una vez realizada la tarea anterior, se lo utilizo al conjunto inicial de datos para crear atributos derivados como: edad (calculado de la diferencia entre la fecha de nacimiento y la fecha actual), edad_ingreso (calculado de la diferencia entre la fecha de nacimiento y el periodo del primer módulo de ingreso). Para obtener estos atributos derivados, se procedió a analizar los atributos de otras tablas de la misma base de datos, estos atributos de detallan anteriormente (ver tabla LXXVIII y LXXXVII).

En este punto surgió un problema con estudiantes que han cursado periodos anteriores al 2008 puesto que el sistema empezó a trabajar a partir de este año, por lo cual se observó información académica solo de los últimos módulos. Por ejemplo, estudiantes que ingresaron a la universidad en el año 2006 solo se tiene información académica desde el cuarto modulo que corresponde al periodo Marzo 2008 – Julio 2008, en vista de este inconveniente para completar el campo edad_ingreso de la tabla antes mencionada se procedió a extraer los registros faltantes de los registros en papel que gestionan las secretarias de cada carrera del área de la energía. Una vez terminado la integración de los datos se muestra a continuación (ver figura 38) el diagrama de la base de datos final.

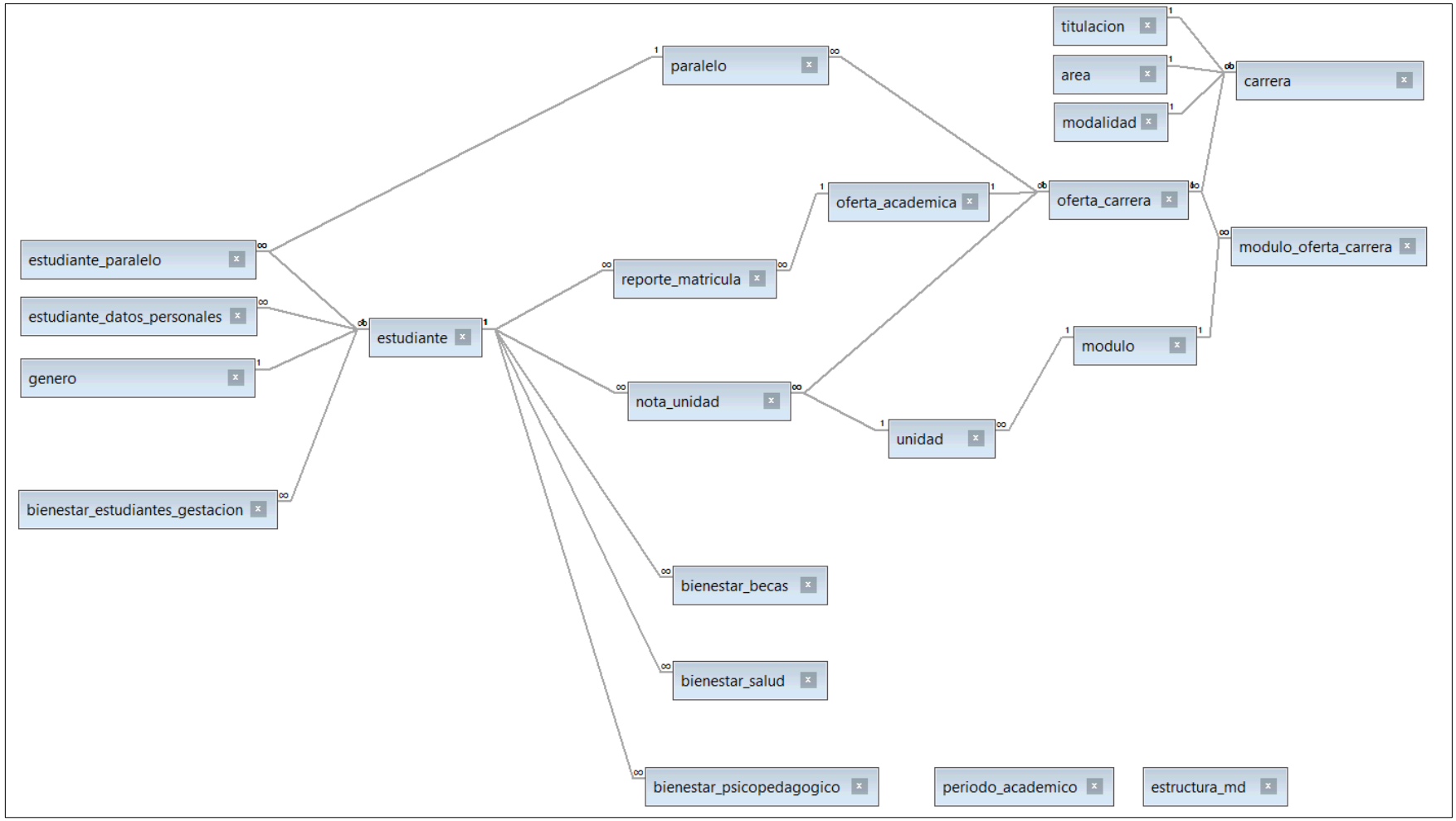


Figura 38: Diseño de la base de datos final.

3.3. Generar los Modelos y Patrones elegidos utilizando una herramienta o paquete de Minería de Datos.

En esta actividad se realizaron tareas en cuanto a generación de modelos y también se compara el rendimiento de los mismos en la herramienta RapidMiner.

3.3.1. Cuarta Fase: Modelado

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.

Para llevar a cabo esta actividad es necesario utilizar la herramienta Rapid Miner que fue previamente seleccionada, la cual cuenta con todos los componentes y características necesarias para llevar a cabo esta fase.

3.3.1.1. Tarea Uno: Selección de Técnica de Modelado

En esta etapa se seleccionara algunas de las técnicas de modelado que posteriormente fueron aplicados parcialmente o en su totalidad, al conjunto de datos para llevar a cabo los experimentos. 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. A continuación se muestra una descripción de los algoritmos que se aplicaran en el presente Trabajo de Titulación.

3.3.1.1.1. Algoritmos de Clasificación

Los algoritmos más utilizados para la clasificación son los algoritmos de inducción. Existen varios enfoques para los algoritmos de inducción, pero se trabajará con aquellos que generan árboles de decisión *(TDIT Top Down Induction Trees)* [29,31].

Los Árboles de decisión es una técnica de minería de datos, que establece una colección de condiciones organizadas jerárquicamente, de tal manera que la decisión final a elegir se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [73].

Los árboles de decisión se adecuan mejor a la clasificación. Puesto que clasificar es determinar de entre varias clases, a qué clase pertenece una entidad; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema. Los árboles de decisión manejan datos no numéricos muy bien [73,31].

Estos árboles de decisiones generan modelos que comúnmente son usados para examinar los datos e inducir las reglas para realizar predicciones. La calidad de un árbol depende de la precisión de la clasificación y del tamaño del árbol. En particular se hará hincapié en los algoritmos *C4.5, ID3 y CHAID.*

Se ha planificó la utilización de los árboles de decisión con el fin de generar modelos que permita establecer patrones de deserción y reprobación de estudiantes. Estos algoritmos se encuentran descritos en el *Capítulo II: Recopilación de Técnicas de Minería de Datos: Técnicas Supervisadas o Predictivas.*

3.3.1.1.2. Algoritmos Basados en Reglas de Inducción

Estos algoritmos representan el conjunto de entrenamiento en forma de reglas que posteriormente son evaluadas directamente para clasificar nuevas instancias. Y pueden ser representadas de diversas maneras, incluyendo árboles de decisión y reglas modulares. Los métodos de inducción de reglas evalúan los atributos del conjunto de entrenamiento y concluyen cuales se deben utilizar para separar entre las diferentes clases. Se planificó la utilización de algoritmos basados en reglas de inducción JRip, PART, Ridor, Decisión Table, DTNB y NNge los cuales se encuentran detallados en el

3.3.1.2. Tarea Dos: Generación de Diseño de Pruebas

En este apartado se describe los experimentos realizados, también cómo se evaluaron los resultados. Se realizaron a lo largo del Trabajo de Titulación distintos experimentos con distintos subconjuntos de datos, también se utilizó las estructuras generadas para identificar factores de deserción y reprobación. A continuación, se enumeran y describen de forma general los experimentos realizados.

- **Algoritmos de Árboles de Decisión: ID3, C4,5 y CHAID**

Con estos algoritmos se realizó una primera evaluación con un conjunto de entrenamiento que corresponde al 70% de la muestra total, y los cuales posteriormente se comparan con los resultados obtenidos mediante la validación cruzada con cinco subconjuntos y seguido de esto 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, cabe recalcar que se aplicó los mejores parámetros para obtener los mejores resultados.

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

Con estos algoritmos se realizó de igual manera que los anteriores pruebas con un conjunto de entrenamiento del 70% de la muestra total y luego estos comparar con los resultados aplicando la validación cruzada con cinco subconjuntos y luego estos resultados serán evaluados a través de una matriz de confusión, y también se evaluara el número de reglas generadas en cada algoritmo y también las reglas más relevantes.

3.3.1.3. Tarea Tres: Construcción de Modelo

En esta actividad se procede a explicar los resultados obtenidos de los modelos seleccionados. Además, se enumeran y describen de forma general los experimentos

realizados con los parámetros y configuraciones realizadas en la herramienta para el desarrollo de los modelos.

3.3.1.3.1. Construcción de Modelos para Deserción

En este apartado se realizara el diseño de modelos que aplicaran en la siguiente fase, estos están enfocados en encontrar los factores que inciden en la deserción, cabe mencionar que se formaron dos grupos: el primer grupo es de estudiantes y el segundo de egresados y desertores, con este segundo grupo se trabajara a lo largo de la construcción de los modelos y luego en la fase de evaluación se examinara el desempeño de los modelos con el grupo de estudiantes.

Previamente a esto se estableció roles para los siguientes atributos *numeroIdentificacion* se seleccionó como el rol *id* para cada uno de los registros y se definió al atributo *estado* como clase o *label* es decir el objetivo a predecir. Para llevar a cabo las estas pruebas primeramente se utilizara un conjunto instancias como entrenamiento y posteriormente con el total de instancias se aplicara la validación cruzada. Para las pruebas de entrenamiento se utilizó el 70% de la muestra total y para las pruebas de validación cruzada se utilizara el total de instancias, para ver los procesos formados para completar esta pruebas (ver anexo 6).

- **Modelado con algoritmo ID3**

El algoritmo ID3 se fijaron los siguientes parámetros: en criterio se seleccionó *ganancia de información,* aquí se especifica el criterio de selección de atributos y de divisiones numéricas, el parámetro *minimal size for split=3* 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=0.6* 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 LXXXVIII):

TABLA LXXXVIII:
RENDIMIENTO DE ALGORITMO ID3 EN DESERCIÓN.

ID3	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	99.51%	0.49%	0.990	0.006	0.55%	0.053	0.088
Validación Cruzada	93.48%	6.52%	0.863	0.060	5.97%	0.241	-

Luego de aplicar la validación cruzada con el algoritmo, estas es la matriz de confusión que se generó (ver figura 39).

accuracy: 93.48% +/- 0.80% (mikro: 93.48%)

	true desertor	true egresado	class precision
pred. desertor	1177	78	93.78%
pred. egresado	56	743	92.99%
class recall	95.46%	90.50%	

Figura 39: Matriz de confusión obtenida con el algoritmo ID3 en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre los estados desertor y egresado, se puede observar un porcentaje similar en ambos, la clase desertor con un 93.78% y la clase egresado con 92.99%. El modelo generado por este algoritmo fue un árbol de decisión de grandes dimensiones, en modo texto ocupa 724 líneas y 280 hojas, a continuación se describe un fragmento de las condiciones generadas para generar el árbol (ver figura 40).

Tree

```
periodo_reprobacion = 0: egresado {desertor=0, egresado=693}
periodo_reprobacion = 1-3
|   edad_ingreso = 17
|   |   modulos_reprobados = 1
|   |   |   distancia_origen = 1
|   |   |   |   promedio_asistencia = alto: desertor {desertor=7, egresado=0}
|   |   |   |   promedio_asistencia = bajo
|   |   |   |   |   promedio_notas = malo
|   |   |   |   |   |   servicios = 1
|   |   |   |   |   |   |   sexo = f: desertor {desertor=4, egresado=0}
|   |   |   |   |   |   |   sexo = m
|   |   |   |   |   |   |   |   carrera = IE
|   |   |   |   |   |   |   |   |   cambio_carrera = no: desertor {desertor=5, egresado=0}
|   |   |   |   |   |   |   |   |   cambio_carrera = si
|   |   |   |   |   |   |   |   |   |   tipo_beca = nignuna
|   |   |   |   |   |   |   |   |   |   |   bienestar_servicios = no
```

Figura 40: Fragmento del árbol generado por el algoritmo ID3 en deserción.

Como se puede observar este modelo comprende muchas condiciones para generar el árbol, también el algoritmo eligió el atributo periodo reprobación como raíz para generar en este caso el árbol de decisión. Si bien se generaron demasiadas líneas que permiten generar el árbol, las condiciones generadas, también permiten tener una mayor precisión, considerando todas las variables posibles para obtener el resultado.

- **Modelado con algoritmo C4.5**

El algoritmo C4.5 se fijaron los siguientes parámetros: en criterio se seleccionó *precisión*, aquí se especifica el criterio de selección de atributos y de divisiones numéricas, el parámetro minimal size for split=3 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=0.6 se fijó como la ganancia mínima que debe lograrse con el fin de producirse una división, se fijó el parámetro *maximal deph*=8 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 fijaron en *false* parámetros *no prepruning* que sirve para aplicar las reglas de poda se luego de cada iteración y *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 C4.5 son las siguientes (ver tabla LXXXIX):

TABLA LXXXIX:
RENDIMIENTO DE ALGORITMO C4.5 EN DESERCIÓN.

C4.5	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	93.18%	6.82%	0.857	0.115	11.46%	0.239	0.398
Validación Cruzada	90.85%	9.15%	0.808	0.130	12.96%	0.262	-

Luego de aplicar la validación cruzada con el algoritmo C4.5, estas es la matriz de confusión que se generó (ver figura 41):

accuracy: 90.85% +/- 1.75% (mikro: 90.85%)			
	true desertor	true egresado	class precision
pred. desertor	1158	113	91.11%
pred. egresado	75	708	90.42%
class recall	93.92%	86.24%	

Figura 41: Matriz de confusión obtenida por el algoritmo C4.5 en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre las clases desertor y egresado, en este caso se observa un porcentaje mayor pero no significativo de la clase desertor con un 91.11% sobre la clase egresado con 90.42%. El modelo generado por este algoritmo fue un árbol de decisión con 95 líneas de condiciones el cual consta de 49 hojas y una profundidad de 8 un fragmento de las condiciones se puede observar en la siguiente figura (ver figura 42).

Tree

```
carrera = IE
|   estado_civil = casado
|   |   promedio_notas = bueno: egresado {desertor=0, egresado=11}
|   |   promedio_notas = malo: desertor {desertor=19, egresado=8}
|   |   promedio_notas = muy_bueno: egresado {desertor=0, egresado=5}
|   |   promedio_notas = regular: egresado {desertor=2, egresado=11}
|   estado_civil = divorciado: desertor {desertor=5, egresado=0}
|   estado_civil = soltero
|   |   horario_estudio = vespertino
|   |   |   promedio_notas = bueno: egresado {desertor=7, egresado=87}
|   |   |   promedio_notas = excelente: egresado {desertor=2, egresado=3}
|   |   |   promedio_notas = malo: desertor {desertor=357, egresado=36}
|   |   |   promedio_notas = muy_bueno
|   |   |   |   servicios = 1
```

Figura 42: Fragmento del árbol generado por el algoritmo C4.5 en deserción.

Como se puede observar este modelo comprende una menor cantidad de condiciones que el anterior de ID3 para generar el árbol, en este caso el modelo eligió el atributo carrera como raíz para generar en este caso el árbol de decisión. Los resultados generados por este algoritmo muestran que no se generaron demasiadas reglas puesto que se ha establecido un minino de profundidad de 8 esto a su vez permite que el algoritmo sea más corto y presente reglas más compresibles que el algoritmo anterior.

- **Modelado con algoritmo 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=0.6 se fijó como la ganancia mínima que debe lograrse con el fin de producirse una división, se fijó el parámetro *maximal deph=8* 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 fijaron en *false* parámetros *no prepruning* que sirve para aplicar las reglas de poda se luego de cada iteración y 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 XC):

TABLA XC:
RENDIMIENTO DE ALGORITMO CHAID EN DESERCIÓN.

CHAID	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	95.82%	4.18%	0.912	0.072	7.21%	0.190	0.315
Validación Cruzada	93.91%	6.09%	0.871	0.084	8.35%	0.216	-

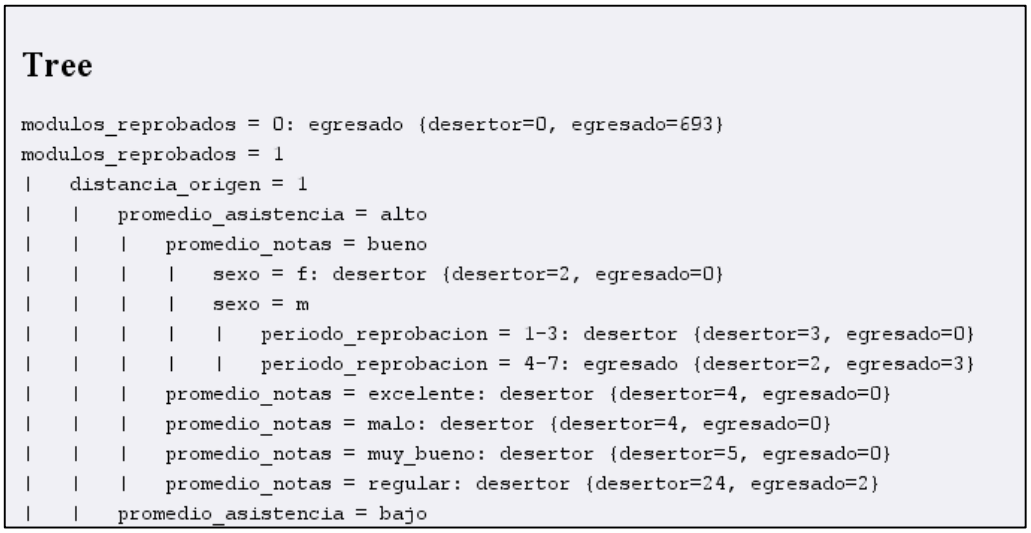
Luego de aplicar la validación cruzada con el algoritmo CHAID, estas es la matriz de confusión que se generó (ver figura 43):

accuracy: 93.91% +/- 1.68% (mikro: 93.91%)

	true desertor	true egresado	class precision
pred. desertor	1205	97	92.55%
pred. egresado	28	724	96.28%
class recall	97.73%	88.19%	

Figura 43: Matriz de confusión obtenida por el algoritmo CHAID en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre los clase desertor y egresado, en donde la clase egresado predomina con un 96.28% respecto a la clase desertor que posee un 92.55%, superando a las obtenidas por los algoritmos anteriores ID3 y C4.5. El modelo generado por este algoritmo fue un árbol de decisión con 103 líneas de condiciones el cual consta de 61 hojas y una profundidad de 8 un fragmento de las condiciones se puede observar en la siguiente figura (ver figura 44).



Tree

```
modulos_reprobados = 0: egresado {desertor=0, egresado=693}
modulos_reprobados = 1
|   distancia_origen = 1
|   |   promedio_asistencia = alto
|   |   |   promedio_notas = bueno
|   |   |   |   sexo = f: desertor {desertor=2, egresado=0}
|   |   |   |   sexo = m
|   |   |   |   |   periodo_reprobacion = 1-3: desertor {desertor=3, egresado=0}
|   |   |   |   |   periodo_reprobacion = 4-7: egresado {desertor=2, egresado=3}
|   |   |   promedio_notas = excelente: desertor {desertor=4, egresado=0}
|   |   |   promedio_notas = malo: desertor {desertor=4, egresado=0}
|   |   |   promedio_notas = muy_bueno: desertor {desertor=5, egresado=0}
|   |   |   promedio_notas = regular: desertor {desertor=24, egresado=2}
|   |   promedio_asistencia = bajo
```

Figura 44: Condiciones del árbol generado por el algoritmo CHAID en deserción.

Como se puede observar este modelo comprende una menor cantidad de condiciones que los dos modelos anteriores ID3 y C4.5 para generar el árbol, en este caso el modelo eligió el atributo *módulos reprobados* como raíz para generar en este caso el árbol de decisión. Como se puede observar los tres tipos de árboles seleccionaron diferentes atributos para generar los árboles.

Este último algoritmo de clasificación generó pocas líneas con lo cual se generó un árbol corto sin muchas condiciones, esto a su vez se debe como en el algoritmo anterior a que se definió el número máximo de profundidad del árbol pese a ello las condiciones generadas son bastante comprensibles.

- **Modelado con algoritmo JRip**

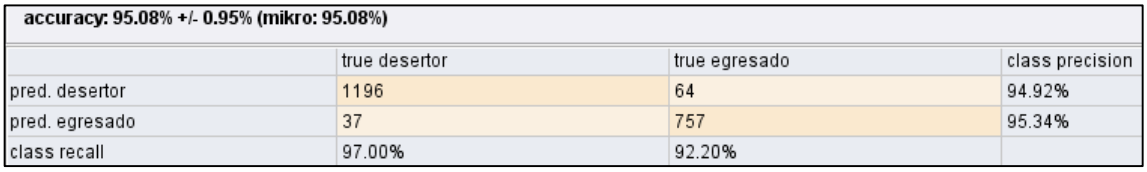
Para el algoritmo JRip se fijaron los siguientes parámetros: el parámetro *F=3* es el número mínimo de divisiones que se pueden dar por cada nodo, el parámetro *N=2* es el tamaño mínimo de cada hoja, el parámetro *O=2* el número de ejecuciones para lograr una para optimización, *D=false* indica el modo de depuración, *S=10* indica la semilla para generar aleatoriedad, *E=false* si no marca la tasa de error, *P=false* 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 XCI):

TABLA XCI:
RENDIMIENTO DE ALGORITMO JRIP EN DESERCIÓN.

JRip	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	95.48%	4.52%	0.904	0.084	8.40%	0.205	0.341
Validación Cruzada	95.08%	4.92%	0.897	0.072	7.19%	0.205	0.342

Luego de aplicar la validación cruzada con el algoritmo JRip, estas es la matriz de confusión que se generó (ver figura 45):

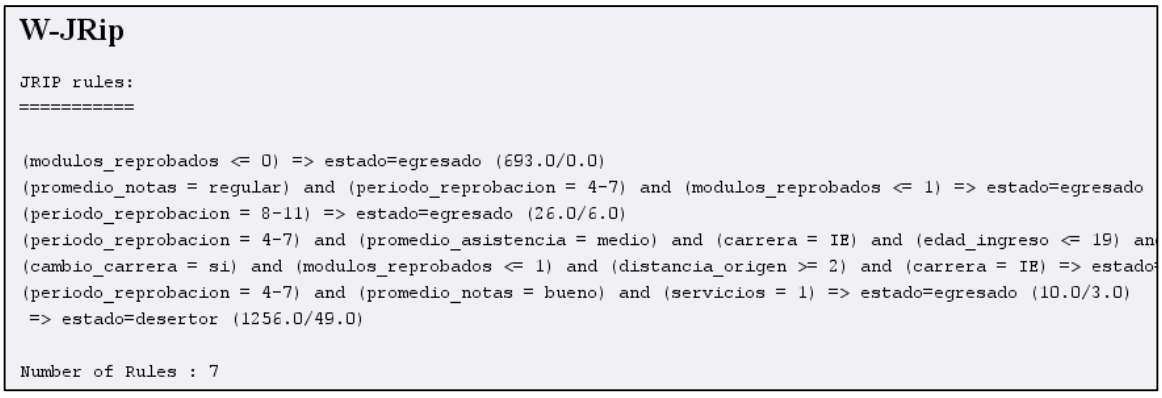


accuracy: 95.08% +/- 0.95% (mikro: 95.08%)

	true desertor	true egresado	class precision
pred. desertor	1196	64	94.92%
pred. egresado	37	757	95.34%
class recall	97.00%	92.20%	

Figura 45: Matriz de confusión obtenida por el algoritmo JRip en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre las clases desertor y egresado, se puede observar que este algoritmo obtuvo mejores resultados que los anteriores, sin embargo se observa una mejor clasificación para la clase egresado con 95.34%, sobre la clase desertor con 94.92%. Las reglas generadas por el modelo constan de 8 las cuales se muestran a continuación (ver figura 46):



W-JRip

```
JRIP rules:
===========

(modulos_reprobados <= 0) => estado=egresado (693.0/0.0)
(promedio_notas = regular) and (periodo_reprobacion = 4-7) and (modulos_reprobados <= 1) => estado=egresado
(periodo_reprobacion = 8-11) => estado=egresado (26.0/6.0)
(periodo_reprobacion = 4-7) and (promedio_asistencia = medio) and (carrera = IE) and (edad_ingreso <= 19) an
(cambio_carrera = si) and (modulos_reprobados <= 1) and (distancia_origen >= 2) and (carrera = IE) => estado=
(periodo_reprobacion = 4-7) and (promedio_notas = bueno) and (servicios = 1) => estado=egresado (10.0/3.0)
 => estado=desertor (1256.0/49.0)

Number of Rules : 7
```

Figura 46: Reglas generadas por el algoritmo JRip en deserción.

Como se puede observar este modelo tiene pocas reglas de decisión, pero más comprensibles que las generadas por los algoritmos ID3, C4.5 y CHAID. Si bien las reglas generadas clasifican un buen porcentaje de las instancias, estas solo describen condiciones para los estudiantes egresados, generando solo una regla para estudiantes desertores, como se muestra en la siguiente figura (ver figura 47).

```
=> estado=desertor (1256.0/49.0)
```

Figura 47: Regla generada para definir a estudiantes desertores.

- **Modelado con algoritmo PART**

Para el algoritmo PART se fijaron los siguientes parámetros: el parámetro *C=0.5* el umbral de confianza para la poda, el parámetro *M=2* en 2 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=false* se usa para divisiones binarias, el parámetro *U=false* 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 XCII):

TABLA XCII:
RENDIMIENTO DE ALGORITMO PART EN DESERCIÓN.

PART	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	97.15%	2.85%	0.940	0.043	4.33%	0.147	0.244
Validación Cruzada	94.64%	5.36%	0.888	0.064	6.38%	0.207	0.344

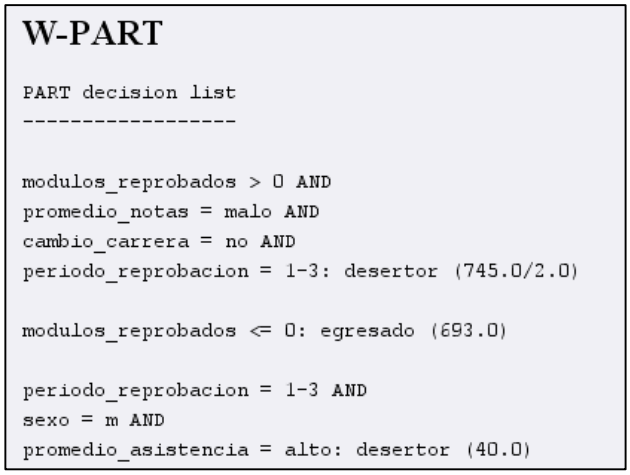
Luego de aplicar la validación cruzada con el algoritmo PART, esta es la matriz de confusión que se generó (ver figura 48).

accuracy: 94.64% +/- 0.71% (mikro: 94.64%)

	true desertor	true egresado	class precision
pred. desertor	1183	60	95.17%
pred. egresado	50	761	93.83%
class recall	95.94%	92.69%	

Figura 48: Matriz de confusión obtenida por el algoritmo PART en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre las clases desertor y egresado, en donde se observa una mejor clasificación para la clase desertor con un 95.17% superando a la clase egresado con 93.83%. Las reglas generadas por el modelo PART constan de 38 las cuales se muestran un fragmento a continuación (ver figura 49):



W-PART

```
PART decision list
------------------

modulos_reprobados > 0 AND
promedio_notas = malo AND
cambio_carrera = no AND
periodo_reprobacion = 1-3: desertor (745.0/2.0)

modulos_reprobados <= 0: egresado (693.0)

periodo_reprobacion = 1-3 AND
sexo = m AND
promedio_asistencia = alto: desertor (40.0)
```

Figura 49: Fragmento de Reglas generadas por el algoritmo PART en deserción.

Las reglas generadas por el algoritmo PART son muy comprensibles que el algoritmo JRip, pero superada en rendimiento por el algoritmo ID3, C4.5 y CHAID. En este caso el algoritmo generó más reglas que el algoritmo anterior, y la mayoría de las reglas generadas tienen como objetivo describir al estado desertor.

- **Modelado con algoritmo Ridor**

Para el algoritmo Ridor se fijaron los siguientes parámetros: el parámetro *F=4.0*, se lo fijo este valor como el número de pliegues, para realizar la poda, el parámetro *S=1.0*, define el número de barajaduras para seleccionar al azar los datos con el fin de obtener una mejor regla, el parámetro *A=true* 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.

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 Ridor son las siguientes (ver tabla XCIII):

TABLA XCIII:
RENDIMIENTO DE ALGORITMO RIDOR EN DESERCIÓN.

Ridor	*Instancias bien clasificadas (%)*	*Instancias mal clasificadas (%)*	*Índice de kappa*	*Error Absoluto*	*Error Relativo*	*Error Cuadrático Medio*	*Error Cuadrático Relativo*
Conjunto de Entrenamiento	95.20%	4.80%	0.900	0.048	4.80%	0.219	0.364
Validación Cruzada	94.30%	5.70%	0.879	0.057	5.70%	0.236	0.394

Luego de aplicar la validación cruzada con el algoritmo Ridor, esta es la matriz de confusión que se generó (ver figura 50).

accuracy: 94.30% +/- 1.78% (mikro: 94.30%)

	true desertor	true egresado	class precision
pred. desertor	1207	91	92.99%
pred. egresado	26	730	96.56%
class recall	97.89%	88.92%	

Figura 50: Matriz de confusión obtenida por el algoritmo Ridor en deserción.

La matriz de confusión generada describe el porcentaje de precisión sobre las clases desertor y egresado, en donde se observa una mejor clasificación para la clase egresado con un 96.56% superando a la clase desertor con 92.99%.

Las reglas generadas por el modelo Ridor constan de 14, a continuación se muestra un fragmento (ver figura 51).

W-Ridor

```
RIpple DOwn Rule Learner(Ridor) rules
--------------------------------------

estado = egresado  (2054.0/1233.0)
           Except (modulos_reprobados > 0.5) and (periodo_reprobacion = 1-3) and (camb
           Except (modulos_reprobados > 0.5) and (promedio_notas = malo) and (modulos_
           Except (modulos_reprobados > 0.5) and (promedio_notas = malo) and (periodo_
           Except (modulos_reprobados > 0.5) and (periodo_reprobacion = 1-3) and (prom
           Except (modulos_reprobados > 0.5) and (promedio_notas = malo) and (periodo_
           Except (modulos_reprobados > 0.5) and (modulos_reprobados > 1.5) and (edad_
           Except (modulos_reprobados > 0.5) and (periodo_reprobacion = 1-3) and (dist
           Except (modulos_reprobados > 0.5) and (promedio_notas = malo) => estado = d
```

Figura 51: Fragmento de reglas obtenidas por el algoritmo Ridor en deserción.

Las reglas generadas por el algoritmo Ridor son muy comprensibles y además muy extensas, en este caso el algoritmo generó menos reglas que los algoritmos anteriores, y la mayoría de las reglas generadas tienen como objetivo describir al estado desertor.

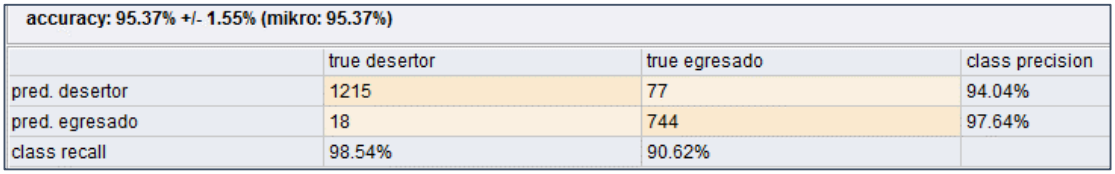
- **Modelado con algoritmo Decisión Table**

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=5.0*, define el número de validaciones cruzadas, el parámetro *E=null* es el criterio de evaluación por defecto es *accuracy*, el parámetro *I=false* 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 XCIV):

TABLA XCIV:
RENDIMIENTO EL ALGORITMO DECISION TABLE EN DESERCIÓN.

DECISION TABLE	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	96.10%	3.90%	0.918	0.071	7.08%	0.180	0.298
Validación Cruzada	95.37%	4.63%	0.902	0.082	8.24%	0.193	0.322

Luego de aplicar la validación cruzada con el algoritmo Decisión Table, estas es la matriz de confusión que se generó (ver figura 52):



accuracy: 95.37% +/- 1.55% (mikro: 95.37%)

	true desertor	true egresado	class precision
pred. desertor	1215	77	94.04%
pred. egresado	18	744	97.64%
class recall	98.54%	90.62%	

Figura 52: Matriz de confusión obtenida por el algoritmo Decisión Table en deserción.

La matriz de confusión describe el porcentaje de precisión sobre las clases desertor y egresado, en donde se observa una mayor clasificación correcta para la clase que presenta posibilidades de egresar con 97.64%, sobre estudiantes con posibilidades de abandonar los estudios con 94.04%. Las reglas generadas y descritas a través de la tabla son en total 52, de los cuales ha seleccionado ocho atributos de los 17 con los que se trabajó inicialmente, a continuación se muestra un fragmento de la tabla de decisiones generada (ver figura 53).

Rules:

promedio_notas	sexo	edad_ingreso	modulos_reprobados	periodo_reprobacion	distancia_origen
malo	m	'All'	'(1.5-inf)'	4-7	'All'
malo	m	'All'	'(1.5-inf)'	1-3	'All'
malo	m	'All'	'(1.5-inf)'	8-11	'All'
bueno	f	'All'	'(0.5-1.5]'	8-11	'All'
malo	m	'All'	'(1.5-inf)'	4-7	'All'
malo	f	'All'	'(0.5-1.5]'	8-11	'All'
regular	f	'All'	'(0.5-1.5]'	8-11	'All'
muy_bueno	f	'All'	'(0.5-1.5]'	8-11	'All'
bueno	m	'All'	'(0.5-1.5]'	8-11	'All'
malo	m	'All'	'(0.5-1.5]'	1-3	'All'
regular	m	'All'	'(0.5-1.5]'	8-11	'All'
malo	m	'All'	'(0.5-1.5]'	8-11	'All'
malo	m	'All'	'(1.5-inf)'	1-3	'All'
bueno	f	'All'	'(1.5-inf)'	4-7	'All'

Figura 53: Fragmento de la tabla de decisiones obtenida por el algoritmo Decisión Table en deserción.

En este caso el algoritmo género más reglas que el algoritmo JRip, pero no más que las generadas por los demás algoritmos, se puede observar que las reglas generadas y representadas por la tabla son comprensibles, y además solo ha seleccionado 8 atributos de los 17 presentes en la estructura enfocada en deserción.

- **Modelado con algoritmo 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=true* 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 XCV):

TABLA XCV:
RENDIMIENTO DE ALGORITMO DTNB EN DESERCIÓN.

DTNB	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	95.55%	4.45%	0.906	0.060	6.00%	0.185	0.308
Validación Cruzada	91.28%	8.72%	0.817	0.097	9.71%	0.273	0.456

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

accuracy: 91.28% +/- 1.33% (mikro: 91.29%)

	true desertor	true egresado	class precision
pred. desertor	1160	106	91.63%
pred. egresado	73	715	90.74%
class recall	94.08%	87.09%	

Figura 54: Matriz de confusión obtenida por el algoritmo DTNB en deserción.

La matriz de confusión describe el porcentaje de precisión sobre las clases desertor y egresado, en donde se observa una mayor clasificación para la clase que presenta posibilidades de abandonar los estudios con 91.63%, sobre estudiantes con posibilidades de egresar con 90.74%. Las reglas generadas por el algoritmo son 499, a continuación se muestra un fragmento de las mismas (ver figura 55).

Rules:

promedio_asistencia	promedio_notas	servicios	sexo	edad_ingreso	distancia_origen
alto	muy_bueno	1	f	'All'	'All'
medio	regular	4	f	'All'	'All'
medio	bueno	1	m	'All'	'All'
alto	bueno	3	f	'All'	'All'
medio	bueno	1	f	'All'	'All'
bajo	malo	1	f	'All'	'All'
medio	muy_bueno	3	m	'All'	'All'
alto	muy_bueno	3	m	'All'	'All'
bajo	malo	1	m	'All'	'All'
medio	bueno	3	m	'All'	'All'
medio	muy_bueno	1	f	'All'	'All'
alto	bueno	3	m	'All'	'All'
alto	bueno	1	f	'All'	'All'
bajo	regular	1	f	'All'	'All'
bajo	regular	3	f	'All'	'All'
medio	regular	1	f	'All'	'All'
bajo	malo	1	m	'All'	'All'
medio	bueno	1	m	'All'	'All'

Figura 55: Fragmento de la tabla de decisiones obtenida por el algoritmo DTNB en deserción.

El algoritmo género las reglas mediante una tabla de decisión, se puede observar que las reglas construidas son en base a 15 atributos de los 17 presentes en la estructura enfocada en deserción, por lo tanto las reglas generadas presentan la característica de ser bastante extensas.

- **Modelado con algoritmo NNge**

Para el algoritmo NNge 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 XCVI):

TABLA XCVI:
RENDIMIENTO DE ALGORITMO NNGE EN DESERCIÓN.

NNge	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	94.97%	5.03%	0.892	0.050	0.105	0.224	0.462
Validación Cruzada	93.82%	6.18%	0.871	0.062	0.129	0.249	0.508

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

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1173   60 |    a = desertor
   67  754 |    b = egresado
```

Figura 56: Matriz de confusión obtenida por el algoritmo NNge en deserción.

La matriz de confusión describe el porcentaje de precisión sobre las clases desertor y egresado, en donde se observa una mayor clasificación para la clase que presenta posibilidades de abandonar los estudios con un 95.13%, sobre estudiantes sin peligro de abandonar con 91.84%.

Las reglas generadas por el algoritmo son de 143, a continuación se muestra un fragmento de las mismas (ver figura 57).

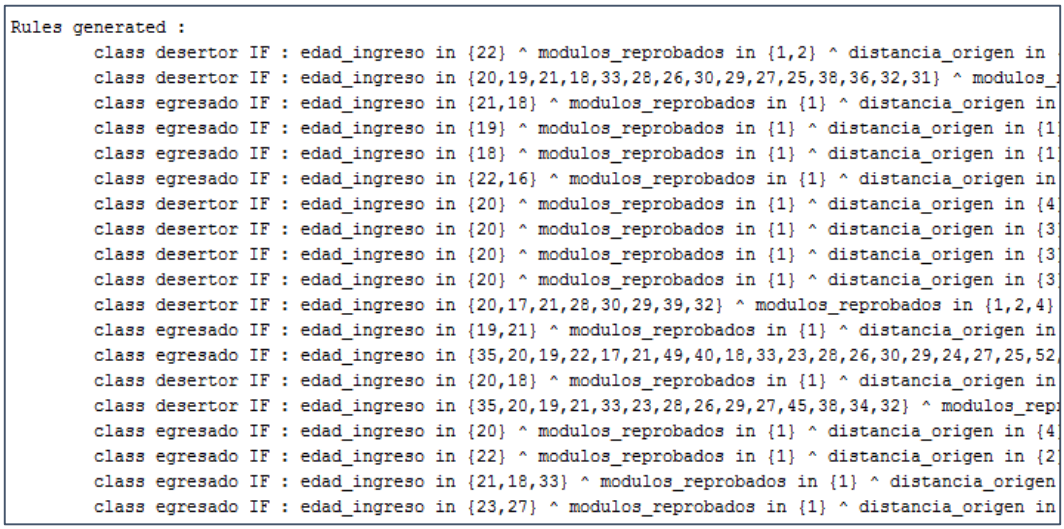


Figura 57: Fragmento de las reglas obtenidas por el algoritmo NNge en deserción.

El algoritmo género un gran número de reglas, se puede observar que las reglas construidas son en base a tres atributos de los 17 presentes en la estructura enfocada en deserción, además cabe mencionar que algunas de las reglas generadas por el algoritmo tomo en cuenta valores nulos.

3.3.1.3.2. Construcción de Modelos para Reprobación

En este apartado se utilizó la estructura enfocada a encontrar los factores que inciden en la reprobación. Se estableció roles para los siguientes atributos *numeroIdentificacion* se seleccionó como el rol de *id* para cada uno de los registros y se definió al atributo *reprobó* como clase es decir el objetivo a predecir.

Para llevar a cabo las estas pruebas se realizaron con el conjunto instancias totales y posteriormente aplicando la validación cruzada en los mismos. Para las pruebas de entrenamiento se utilizó el 70% de la muestra total, y para el proceso de validación cruzada mediante subconjuntos se utilizó en 100%, para ver los procesos formados para completar esta pruebas (ver anexo 6).

- **Modelado con algoritmo JRip**

Para el algoritmo JRip se fijaron los siguientes parámetros: el parámetro F=4 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

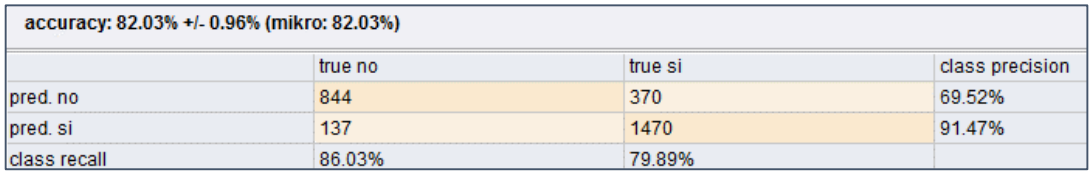
una para optimización, D=false indica el modo de depuración, S=10 indica la semilla para generar aleatoriedad, E=false si no marca la tasa de error, P=false para indicar se utiliza la poda. Para la validación cruzada se fijó el parámetro number of validations=5 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Los resultados arrojados por el algoritmo JRip son las siguientes (ver tabla XCVII):

TABLA XCVII:
RENDIMIENTO DE ALGORITMO JRIP EN REPROBACIÓN.

JRip	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	82.62%	17.38%	0.634	0.260	25.95%	0.360	1.028
Validación Cruzada	82.03%	17.97%	0.624	0.261	26.09%	0.362	1.043

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

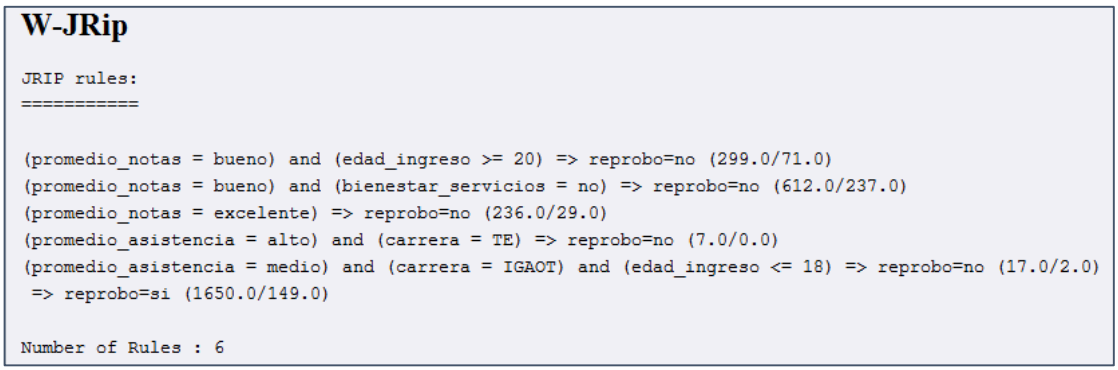


accuracy: 82.03% +/- 0.96% (mikro: 82.03%)

	true no	true si	class precision
pred. no	844	370	69.52%
pred. si	137	1470	91.47%
class recall	86.03%	79.89%	

Figura 58: Matriz de confusión obtenida por el algoritmo JRip en reprobación.

La matriz de confusión generada describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mejor clasificación para la clase que presenta posibilidades de reprobar con 91.47%, sobre estudiantes sin peligro de reprobar con 69.52%. A continuación se muestran las reglas generadas por algoritmo JRip que constan de seis (ver figura 59).



W-JRip

```
JRIP rules:
===========

(promedio_notas = bueno) and (edad_ingreso >= 20) => reprobo=no (299.0/71.0)
(promedio_notas = bueno) and (bienestar_servicios = no) => reprobo=no (612.0/237.0)
(promedio_notas = excelente) => reprobo=no (236.0/29.0)
(promedio_asistencia = alto) and (carrera = TE) => reprobo=no (7.0/0.0)
(promedio_asistencia = medio) and (carrera = IGAOT) and (edad_ingreso <= 18) => reprobo=no (17.0/2.0)
 => reprobo=si (1650.0/149.0)

Number of Rules : 6
```

Figura 59: Reglas generadas por el algoritmo JRip en reprobación.

Como se puede observar las reglas generadas, son comprensibles, sin embargo las reglas describen el comportamiento o situación en la que un estudiante no reprueba y solo generó una regla para describir a estudiantes que han reprobado.

- **Modelado con algoritmo 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 mínimo de objetos por hoja, el parámetro R=false es el uso reducido de poda error, el parámetro N=vacío establece el número de pliegues para la reducción de poda error, a veces se usa como poda, el parámetro B=false se usa para divisiones binarias, el parámetro U=false 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 XCVIII):

TABLA XCVIII:
RENDIMIENTO DE ALGORITMO PART EN REPROBACIÓN.

PART	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	86.93%	13.07%	0.713	0.194	19.39%	0.311	0.888
Validación Cruzada	80.36%	19.64%	0.572	0.246	24.62%	0.382	1.100

Luego de aplicar la validación cruzada con el algoritmo PART, estas es la matriz de confusión que se generó (ver figura 60):

accuracy: 80.36% +/- 0.88% (mikro: 80.36%)

	true no	true si	class precision
pred. no	730	303	70.67%
pred. si	251	1537	85.96%
class recall	74.41%	83.53%	

Figura 60: Matriz de confusión obtenida por el algoritmo PART en reprobación.

La matriz de confusión describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mejor clasificación para la clase que presenta posibilidades de reprobar con 85.96%, sobre estudiantes sin

peligro de reprobar con 70.67%. A continuación se muestra un fragmento de las reglas generadas por el algoritmo PART las cuales constan de 103 (ver figura 61).

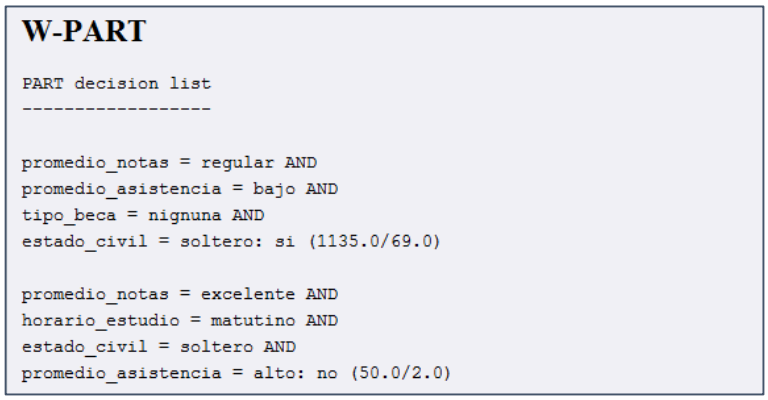


Figura 61: Reglas generadas por el algoritmo PART en reprobación.

En este caso el algoritmo género más reglas que el algoritmo JRip, además se puede observar describen el comportamiento o situación en la que un estudiante reprueba, esto no ocurrió con el algoritmo anterior.

- **Modelado con algoritmo Decisión 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=5.0*, 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.

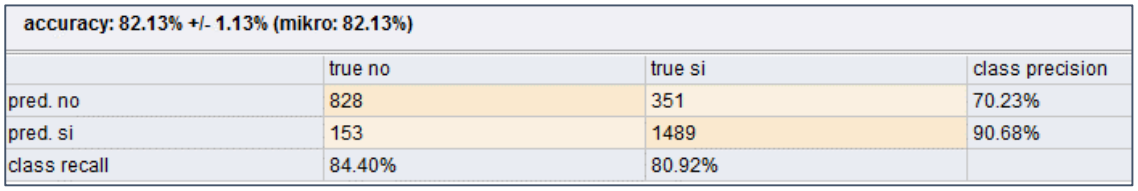
Los resultados arrojados por el algoritmo Decisión Table son las siguientes (ver tabla C):

TABLA XCIX:
RENDIMIENTO DE ALGORITMO DECISION TABLE EN REPROBACIÓN.

DECISION TABLE	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	82.17%	17.83%	0.623	0.267	26.73%	0.365	1.041

Validación Cruzada	82.13%	17.87%	0.624	0.264	26.35%	0.363	1.043

Luego de aplicar la validación cruzada con el algoritmo Decisión Table, estas es la matriz de confusión que se generó (ver figura 62):

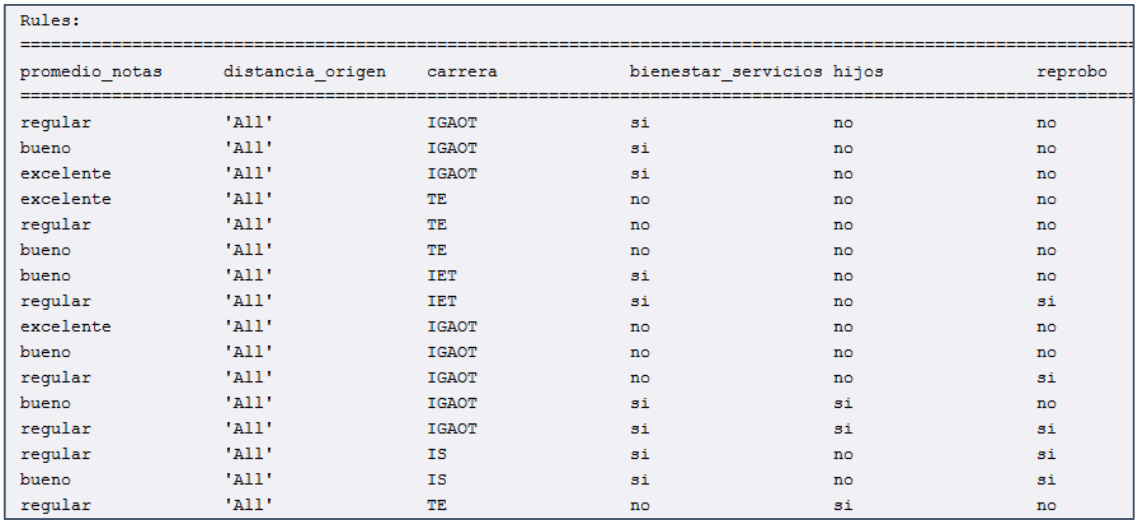


accuracy: 82.13% +/- 1.13% (mikro: 82.13%)

	true no	true si	class precision
pred. no	828	351	70.23%
pred. si	153	1489	90.68%
class recall	84.40%	80.92%	

Figura 62: Matriz de confusión obtenida por el algoritmo Decisión Table en reprobación.

La matriz de confusión describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mayor clasificación correcta para la clase que presenta posibilidades de reprobar con 90.68%, sobre estudiantes sin peligro de reprobar con 70.23%. Las reglas generadas por el algoritmo son en total 48, a continuación se muestra un fragmento de la tabla de decisiones generada (ver figura 63).



Rules:

promedio_notas	distancia_origen	carrera	bienestar_servicios	hijos	reprobo
regular	'All'	IGAOT	si	no	no
bueno	'All'	IGAOT	si	no	no
excelente	'All'	IGAOT	si	no	no
excelente	'All'	TE	no	no	no
regular	'All'	TE	no	no	no
bueno	'All'	TE	no	no	no
bueno	'All'	IET	si	no	no
regular	'All'	IET	si	no	si
excelente	'All'	IGAOT	no	no	no
bueno	'All'	IGAOT	no	no	no
regular	'All'	IGAOT	no	no	si
bueno	'All'	IGAOT	si	si	no
regular	'All'	IGAOT	si	si	si
regular	'All'	IS	si	no	si
bueno	'All'	IS	si	no	si
regular	'All'	TE	no	si	no

Figura 63: Fragmento de la tabla de decisiones obtenida por el algoritmo Decisión Table en reprobación.

En este caso el algoritmo género más reglas que el algoritmo JRip, pero no más que las generadas por el algoritmo PART, se puede observar que las reglas generadas y representadas por la tabla son comprensibles, y además solo ha seleccionado cinco atributos de los 13 presentes en la estructura enfocada en reprobación.

- **Modelado con algoritmo Ridor**

Para el algoritmo Ridor se fijaron los siguientes parámetros: el parámetro *F=5.0*, se lo fijo este valor como el número de pliegues, para realizar la poda, el parámetro *S=1.0*, define el número de barajaduras para seleccionar al azar los datos con el fin de obtener una mejor regla, el parámetro *A=true* 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.

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 Ridor son las siguientes (ver tabla C):

TABLA C:
RENDIMIENTO DE ALGORITMO RIDOR EN REPROBACIÓN.

RIDOR	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	82.57%	17.43%	0.636	0.174	17.43%	0.417	1.191
Validación Cruzada	80.50%	19.50%	0.574	0.195	19.50%	0.441	1.270

Luego de aplicar la validación cruzada con el algoritmo Ridor, estas es la matriz de confusión que se generó (ver figura 64).

accuracy: 80.50% +/- 0.96% (mikro: 80.50%)

	true no	true si	class precision
pred. no	726	295	71.11%
pred. si	255	1545	85.83%
class recall	74.01%	83.97%	

Figura 64: Matriz de confusión obtenida por el algoritmo Ridor en reprobación.

La matriz de confusión describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mayor clasificación para la clase que presenta posibilidades de reprobar con 85.83%, sobre estudiantes sin

peligro de reprobar con 71.11%. Las reglas generadas por el algoritmo son 36, a continuación se muestra un fragmento de estas reglas (ver figura 65).

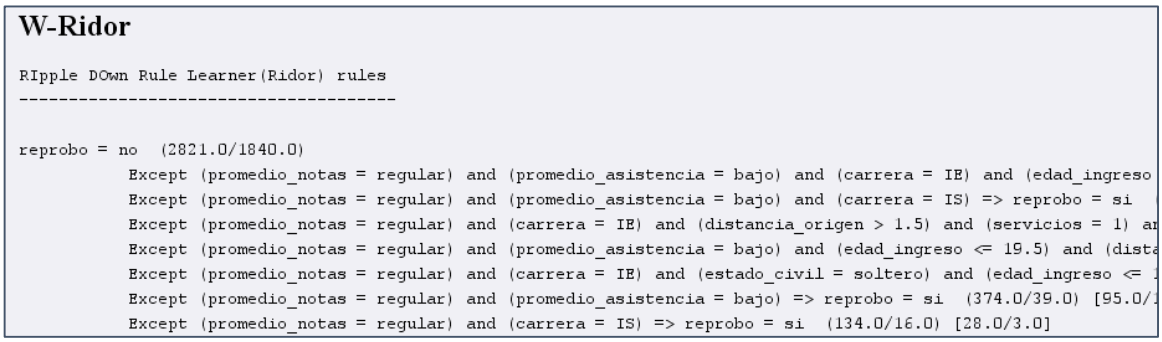


Figura 65: Fragmento de la tabla de decisiones obtenida por el algoritmo Ridor en reprobación.

En este caso el algoritmo género menos reglas que el algoritmo Decisión Table y Ridor pero más que las generadas por el algoritmo JRip, también se observó que las reglas generadas tienen la característica de ser demasiadas extensas.

- **Modelado con algoritmo 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 CI):

TABLA CI:
RENDIMIENTO DE ALGORITMO DTNB EN REPROBACIÓN.

DTNB	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	81.97%	18.03%	0.619	0.221	22.10%	0.372	1.062
Validación Cruzada	81.35%	18.65%	0.604	0.231	23.06%	0.373	1.072

Luego de aplicar la validación cruzada con el algoritmo DTNB, estas es la matriz de confusión que se generó (ver figura 66).

accuracy: 81.35% +/- 1.71% (mikro: 81.35%)			
	true no	true si	class precision
pred. no	798	343	69.94%
pred. si	183	1497	89.11%
class recall	81.35%	81.36%	

Figura 66: Matriz de confusión obtenida por el algoritmo DTNB en reprobación.

La matriz de confusión describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mayor clasificación para la clase que presenta posibilidades de reprobar con 89.11%, sobre estudiantes sin peligro de reprobar con 69.94%. Las reglas generadas por el algoritmo son de 27, a continuación se muestra un fragmento de las mismas (ver figura 67).

Rules:

carrera	bienestar_servicios	madre_trabaja	reprobo
IE	si	?	si
IET	si	?	si
IS	si	?	si
IGAOT	si	?	no
IE	no	?	si
TE	no	?	no
IS	no	?	si
IGAOT	no	?	no
IET	no	?	si
TECI	no	?	no
IGAOT	si	no	no
IS	si	no	no
IS	no	no	si
IGAOT	no	no	no
TECI	no	no	no

Figura 67: Fragmento de la tabla de decisiones obtenida por el algoritmo DTNB en reprobación.

El algoritmo género las reglas mediante una tabla de decisión, se puede observar que las reglas construidas son en base a tres atributos de los 13 presentes en la estructura enfocada en reprobación.

- **Modelado con algoritmo NNge**

Para el algoritmo NNge 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 CII):

TABLA CII:
RENDIMIENTO DE ALGORITMO NNGE EN REPROBACIÓN.

NNge	***Instancias bien clasificadas (%)***	***Instancias mal clasificadas (%)***	***Índice de kappa***	***Error Absoluto***	***Error Relativo***	***Error Cuadrático Medio***	***Error Cuadrático Relativo***
Conjunto de Entrenamiento	77.42%	22.57%	0.472	0.225	0.501	0.475	1.011
Validación Cruzada	75.96%	24.03%	0.472	0.240	0.529	0.490	1.029

Luego de aplicar la validación cruzada con el algoritmo NNge, estas es la matriz de confusión que se generó (ver figura 68).

```
   a    b   <-- classified as
 652  329 |    a = no
 349 1491 |    b = si
```

Figura 68: Matriz de confusión obtenida por el algoritmo NNge en reprobación.

La matriz de confusión describe el porcentaje de precisión sobre las clases estudiantes con posibilidades de reprobar y estudiantes que no, se observa una mayor clasificación para la clase que presenta posibilidades de reprobar con un 81.03%, sobre estudiantes sin peligro de reprobar con 66.46%.

Las reglas generadas por el algoritmo son de 714, a continuación se muestra un fragmento de las mismas (ver figura 69).

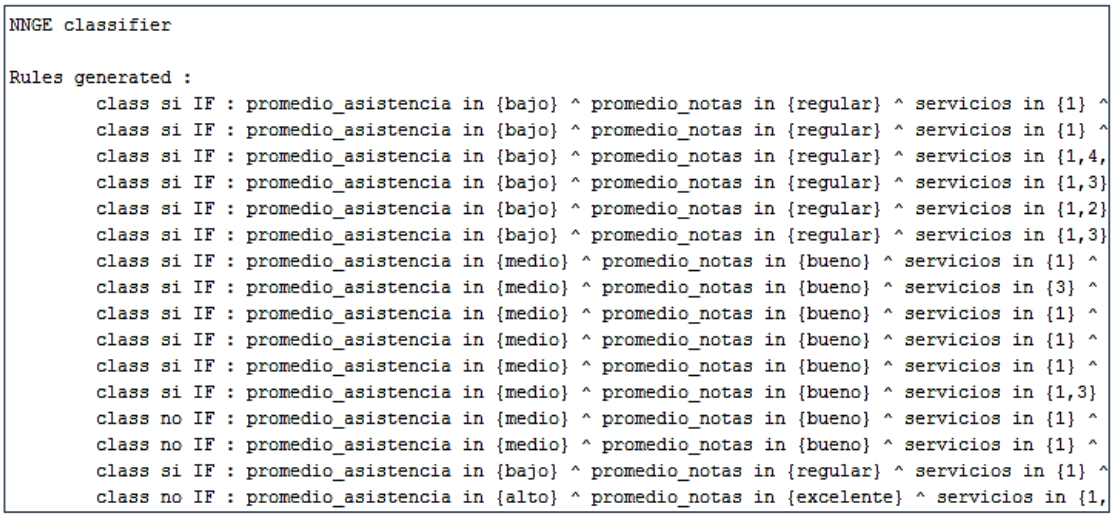


Figura 69: Fragmento de las reglas obtenidas por el algoritmo NNge en reprobación.

El algoritmo género un gran número de reglas, se puede observar que las reglas construidas se formaron en base a todos los 14 atributos presentes en la estructura enfocada en reprobación, además cabe mencionar que algunas de las reglas generadas presentan el carácter interrogación (?) representación de valores nulos.

3.3.1.4. Tarea Cuatro: Evaluación de Modelo

En esta actividad se realizaron tareas con el propósito de evaluar el rendimiento de cada uno de los algoritmos aplicados y las medidas de error obtenidas, con la finalidad de seleccionar el mejor algoritmo para identificar factores de deserción y reprobación.

3.3.1.4.1. Evaluación de Modelos de Deserción

En este punto se hace una evaluación global de los algoritmos aplicados a encontrar factores de deserción, esto con el fin de ir comparando los resultados obtenidos por los distintos algoritmos de clasificación y basados en reglas, sin embargo se debe mencionar que no hay regla que pueda indicar si un modelo es completamente bueno o confiable, en la siguiente tabla se muestran los resultados (ver tabla CIII).

En la siguiente tabla se muestran los parámetros utilizados para comparar el rendimiento de los algoritmos.

El índice Kappa, esta es una medida de concordancia entre las categorías pronosticadas por el clasificador y las categorías observadas, que tiene en cuenta las posibles concordancias debidas al azar. Dónde:

- Si el valor es 1: Concordancia perfecta.
- Si el valor es 0: Concordancia debida al azar.
- Si el valor es negativo: Concordancia menor que la que cabría esperar por azar.

Por lo tanto, en cada uno de los algoritmos aplicados se muestra que, tenemos un alto grado de concordancia superando el (0.800). Además se tomaron en cuenta las variables de error absoluto, error relativo, error cuadrático medio y error cuadrático relativo las cuales se describen a continuación:

Error Absoluto: este valor indica la distancia de acierto entre el valor estimado y el valor real.
Error Relativo: es el cociente entre el valor absoluto y el valor real expresado en porcentaje.

Error Cuadrático Medio: esta medida de error es la principal y más utilizada; a veces se toma la raíz cuadrada para darle la misma dimensión que el valor pronosticado. Muchas técnicas como la regresión lineal, utiliza el error cuadrático medio, ya que tiende a ser medida más fácil de manipular matemáticamente, sin embargo todas las medidas del rendimiento son fáciles de calcular, por lo que el error medio cuadrado no tiene ninguna ventaja particular [93].

Error Cuadrático Relativo: esta medida de error toma el error cuadrático total y la normaliza dividiendo por el error cuadrático total del predictor por defecto [93].

TABLA CIII:
COMPARACIÓN DE RENDIMIENTO DE ALGORITMOS PARA DESERCIÓN.

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)	Índice de Kappa	Error Absoluto	Error Relativo	Error Cuadrático Medio	Error Cuadrático Relativo
ID3	Conjunto de Entrenamiento	99.51%	0.49%	0.990	0.006	0.55%	0.053	0.088
	Validación Cruzada	93.48%	6.52%	0.863	0.060	5.97%	0.241	-
C4.5	Conjunto de Entrenamiento	93.18%	6.82%	0.857	0.115	11.46%	0.239	0.398
	Validación Cruzada	90.85%	9.15%	0.808	0.130	12.96%	0.262	-
CHAID	Conjunto de Entrenamiento	95.82%	4.18%	0.912	0.072	7.21%	0.190	0.315
	Validación Cruzada	93.91%	6.09%	0.871	0.084	8.35%	0.216	-
JRIP	Conjunto de Entrenamiento	95.48%	4.52%	0.904	0.084	8.40%	0.205	0.341
	Validación Cruzada	95.08%	4.92%	0.897	0.072	7.19%	0.205	0.342
PART	Conjunto de Entrenamiento	97.15%	2.85%	0.940	0.043	4.33%	0.147	0.244
	Validación Cruzada	94.64%	5.36%	0.888	0.064	6.38%	0.207	0.344
RIDOR	Conjunto de Entrenamiento	95.20%	4.80%	0.900	0.048	4.80%	0.219	0.364
	Validación Cruzada	94.30%	5.70%	0.879	0.057	5.70%	0.236	0.394
DECISION TABLE	Conjunto de Entrenamiento	96.10%	3.90%	0.918	0.071	7.08%	0.180	0.298
	Validación Cruzada	95.37%	4.63%	0.902	0.082	8.24%	0.193	0.322
DTNB	Conjunto de Entrenamiento	95.55%	4.45%	0.906	0.060	6.00%	0.185	0.308
	Validación Cruzada	91.28%	8.72%	0.817	0.097	9.71%	0.273	0.456
NNGE	Conjunto de Entrenamiento	94.97%	5.03%	0.892	0.050	0.105	0.224	0.462
	Validación Cruzada	93.82%	6.18%	0.871	0.062	0.129	0.249	0.508

En pruebas de entrenamiento se puede observar un porcentaje de datos clasificados sobre el 93%, entre los cuales se destaca el algoritmo ID3 y PART con un porcentaje superior al 97%, con estos resultados obtenidos ya se logró tener una aproximación de los algoritmos que mejor rendimiento obtienen.

En estas pruebas de entrenamiento se puede observar que existe un elevado porcentaje de clasificación esto es lógico puesto que los datos que sirvieron de base para la generación de las reglas, esto produce una sobreestimación de los resultados, sin embargo mediante la aplicación de pruebas de validación cruzada se obtuvieron resultados más reales y precisos, es por ello que algunos de estos algoritmos bajaron considerablemente su porcentaje de clasificación como es el caso del algoritmo *ID3* que obtuvo en la prueba de entrenamiento un porcentaje de 99.51% y luego de aplicar la validación cruzada bajo hasta 93.48%, como se observa en la siguiente figura (ver figura 70).

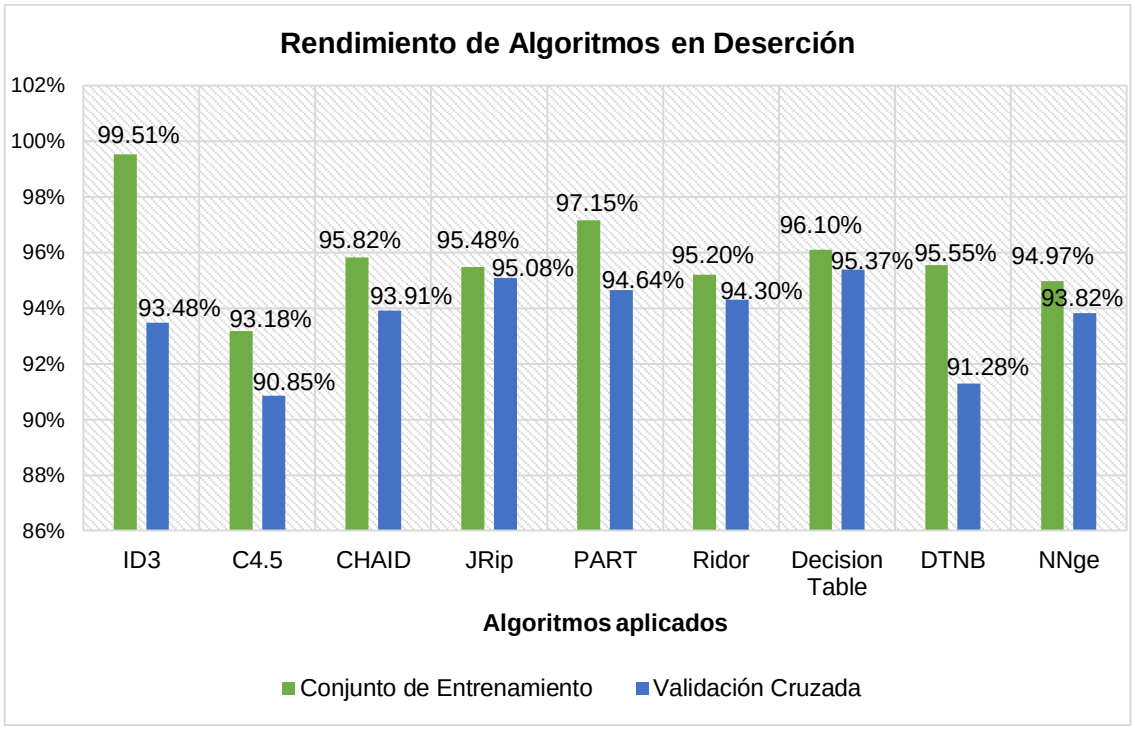


Figura 70: Rendimiento de algoritmos con pruebas de Entrenamiento y Validación Cruzada en Deserción.

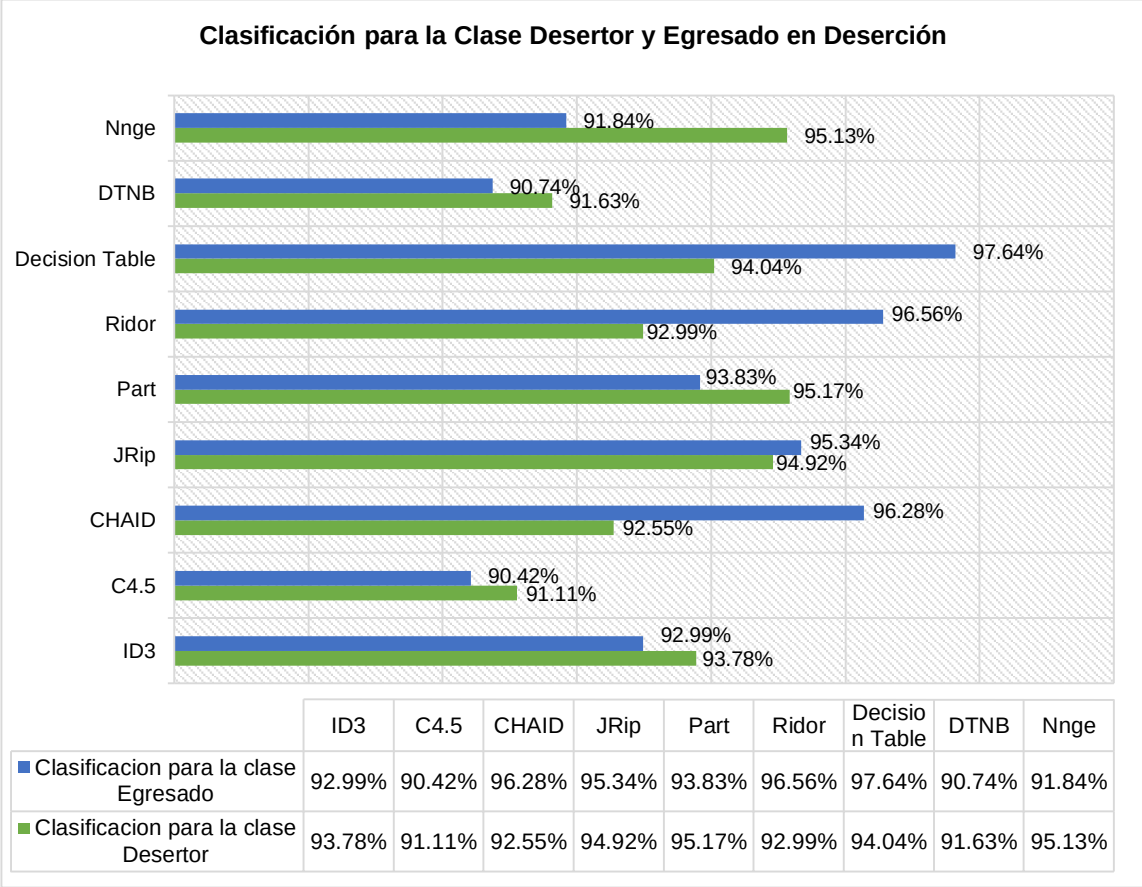
Entre los algoritmos que obtuvieron bajos resultados en las pruebas de validación cruzada están: C.4.5, DTNB, ID3, CHAID y NNge, si bien el rendimiento es bajo respecto a los demás algoritmos esto no significa que los resultados obtenidos fueron descartados de inmediato.

Mientras que en pruebas de validación cruzada los algoritmos que obtuvieron mejor rendimiento son: Decisión Table, JRip, PART y Ridor, con respecto a este último algoritmo las reglas generadas fueron muy pocas y dificultoso de interpretar, por lo tanto los porcentajes de clasificación se encuentran sobrestimados.

En cuanto a los algoritmos con mejor rendimiento a continuación se describen las medidas de error obtenidas.

La primera medida de error es el índice de Kappa y el algoritmo que presenta mejor índice es Decisión Table con 0.902, en otra medida de error denominada Error Absoluto el algoritmo con mejor resultado es Ridor con un 0.057, en la medida de error denominada Error Relativo el algoritmo que mejor se desempeña también es Ridor con un 5.70%, en la medida de error denominada Error Cuadrático Medio el algoritmo con mejor resultado es Decisión Table con 0.180 mientras que en la última medida de error denominada Error Cuadrático Relativo el algoritmo Decisión Table también obtuvo el mejor índice con 0.322.

A continuación se describen los porcentajes de clasificación para clase desertor y egresado (ver figura 71).



	ID3	C4.5	CHAID	JRip	Part	Ridor	Decision Table	DTNB	Nnge
Clasificacion para la clase Egresado	92.99%	90.42%	96.28%	95.34%	93.83%	96.56%	97.64%	90.74%	91.84%
Clasificacion para la clase Desertor	93.78%	91.11%	92.55%	94.92%	95.17%	92.99%	94.04%	91.63%	95.13%

Figura 71: Resultados de Clasificación para la clase Desertor y Egresado.

En la tabla anterior se puede observar que los algoritmos que mejor porcentaje de clasificación obtuvieron para la clase desertor son: PART, JRip, NNge y Decision Table. En base a los porcentajes de clasificación correcta de instancias, en las medidas de error examinadas, se ha establecido que los algoritmos que mejor se desempeñan son Decision Table, PART y Ridor.

Con lo mencionado anteriormente el algoritmo que presenta mejor índice de precisión sobre la clase desertor es el algoritmo PART con 95.17% sobre el algoritmo Decisión Table con 94.04% y el algoritmo Ridor con 92,99%, estos valores fueron obtenidos de la matriz de confusión generada por los algoritmos [89,90].

Por lo tanto en base a los resultados obtenidos se decidió tomar en cuenta los dos algoritmos: Decisión Table y PART, ya que presentan los mejores resultados tanto en rendimiento, clasificación de la clase desertor y las medidas de error [91], no se tomó en cuenta el algoritmo Ridor ya que las reglas generadas por este son muy pocas y redundantes en algunos casos por lo tanto son difíciles de interpretar y de utilizar [87,88].

3.3.1.4.2. Evaluación de Modelos de Reprobación

En este punto se hace una evaluación global de los algoritmos aplicados, enfocados a encontrar factores de reprobación en los estudiantes, se comparan los resultados obtenidos por los distintos algoritmos utilizados, sin embargo se debe mencionar que no hay regla que pueda indicar si un modelo es completamente bueno o confiable, en la siguiente tabla se muestran los resultados (ver tabla CIV).

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 se describen a continuación:

TABLA CIV: COMPARACIÓN DE RENDIMIENTO DE ALGORITMOS PARA REPROBACIÓN.

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)	Índice de Kappa	Error Absoluto	Error Relativo	Error Cuadrático Medio	Error Cuadrático Relativo
JRIP	Conjunto de Entrenamiento	82.62%	17.38%	0.634	0.260	25.95%	0.360	1.028
	Validación Cruzada	82.03%	17.97%	0.624	0.261	26.09%	0.362	1.043
PART	Conjunto de Entrenamiento	86.93%	13.07%	0.713	0.194	19.39%	0.311	0.888
	Validación Cruzada	80.36%	19.64%	0.572	0.246	24.62%	0.382	1.100
DECISION TABLE	Conjunto de Entrenamiento	82.17%	17.83%	0.623	0.267	26.73%	0.365	1.041
	Validación Cruzada	82.13%	17.87%	0.624	0.264	26.35%	0.363	1.043
RIDOR	Conjunto de Entrenamiento	82.57%	17.43%	0.636	0.174	17.43%	0.417	1.191
	Validación Cruzada	80.50%	19.50%	0.574	0.195	19.50%	0.441	1.270
DTNB	Conjunto de Entrenamiento	81.97%	18.03%	0.619	0.221	22.10%	0.372	1.062
	Validación Cruzada	81.35%	18.65%	0.604	0.231	23.06%	0.373	1.072
NNGE	Conjunto de Entrenamiento	77.42%	22.57%	0.472	0.225	0.501	0.475	1.011
	Validación Cruzada	75.96%	24.03%	0.472	0.240	0.529	0.490	1.029

En pruebas de entrenamiento se puede observar que los algoritmos que presentan mejor rendimiento son: PART, JRip y Ridor; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento.

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 más reales y precisos, por lo tanto una mejor clasificación y reglas bien construidas (ver figura 72).

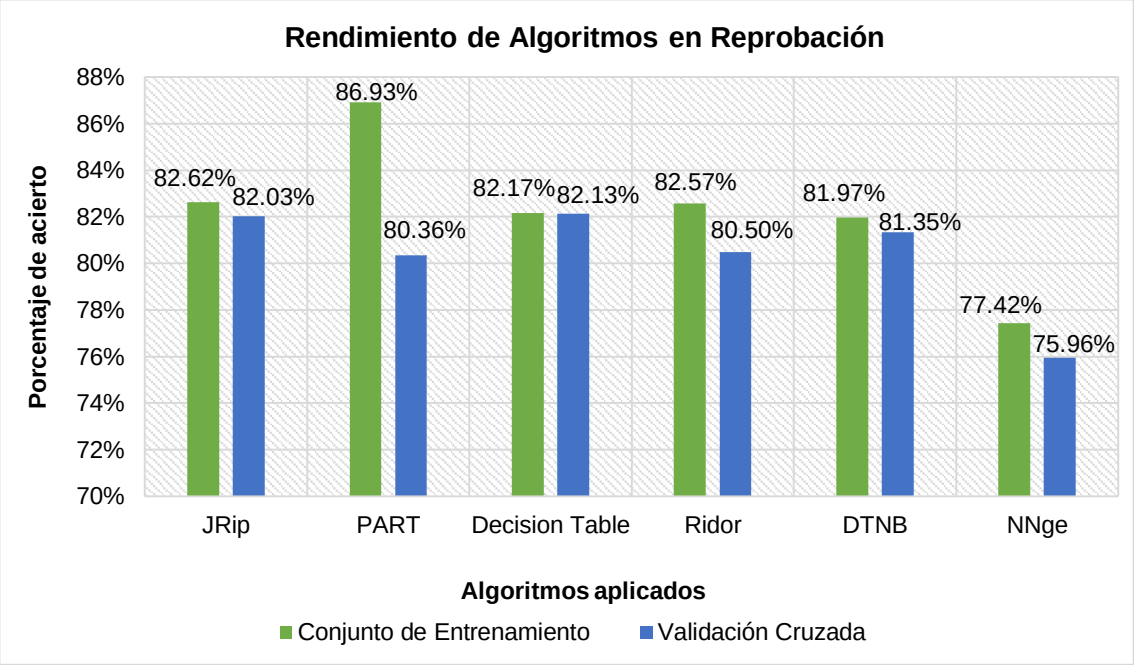


Figura 72: Resultados de clasificación correcta de algoritmos en Modelos de Reprobación.

Los algoritmos con mejor rendimiento son: Decisión Table, JRip, DTNB y Ridor, estos presentan un rendimiento similar en pruebas de Entrenamiento y Validación Cruzada, es decir las reglas obtenidas con el conjunto de entrenamiento no cambian mucho luego de aplicar la validación cruzada, sin embargo las reglas generadas por el algoritmo JRip son pocas y se enfocan en describir el comportamiento o situación de estudiantes que egresan dificultando la tarea de análisis de factores de reprobación [87,88].

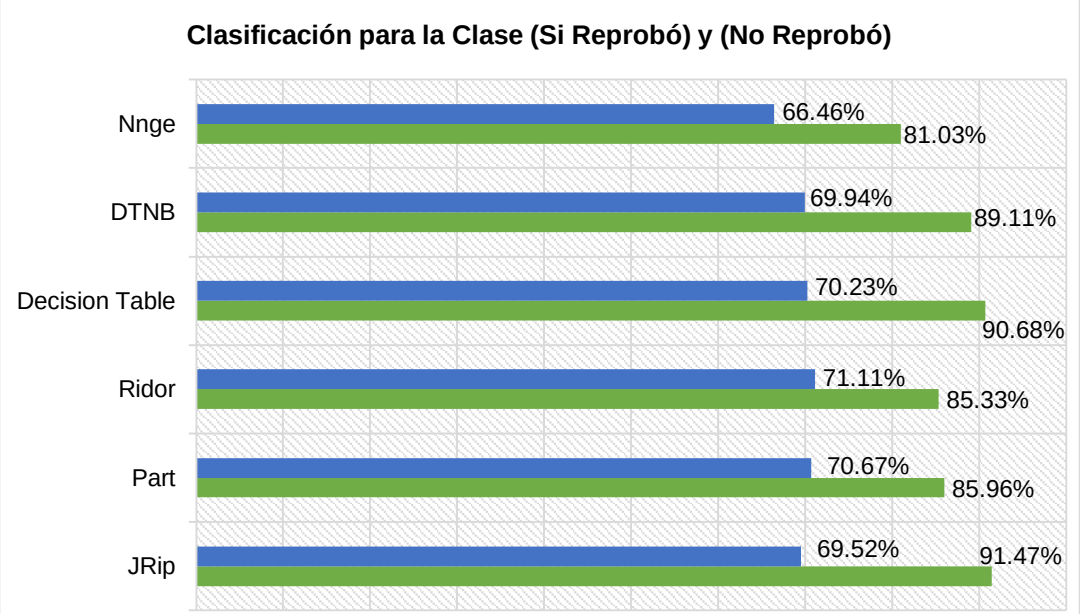
Entre los algoritmos que no obtuvieron los mejores resultados en las pruebas de validación cruzada están: PART y NNge, si bien el rendimiento es bajo respecto a los

demás algoritmos esto no significó que los resultados obtenidos fueron descartados de inmediato.

En cuanto a los algoritmos con mejor rendimiento a continuación se describen las medidas de error.

La primera medida de error es el índice de Kappa, los algoritmos que presentan el mejor índice son: Decisión Table y JRip ambos con 0.624, en otra medida de error denominada Error Absoluto tiene como mejor resultado el obtenido por el algoritmo Ridor con 0.195, en la medida de error denominada Error Relativo igualmente tiene como mejor resultado el obtenido por el algoritmo Ridor con un 19.50%, en la medida de error denominada Error Cuadrático Medio el mejor resultado fue el obtenido por el algoritmo DTNB con una medida de 0.362, mientras que en la última medida de error denominada Error Cuadrático Relativo dos al algoritmos obtuvieron los mejores resultados: Decisión Table y JRip con una medida de 1.043.

A continuación se describen los porcentajes de clasificación para clase que reprobó (Si) y la clase que no ha reprobado (No) (ver figura 73).



	JRip	Part	Ridor	Decision Table	DTNB	Nnge
Clasificacion para la clase (No)	69.52%	70.67%	71.11%	70.23%	69.94%	66.46%
Clasificacion para la clase (Si)	91.47%	85.96%	85.33%	90.68%	89.11%	81.03%

Figura 73: Resultados de clasificación para clase (Si) reprobó y (No) reprobó.

En la tabla anterior se puede observar los algoritmos que mejor porcentaje de clasificación obtuvieron para la clase que reprueba (Si) fueron: Decisión Table, DTNB, PART y Ridor. En base a los porcentajes de clasificación correcta de instancias, en las medidas de error examinadas, se ha establecido que los algoritmos que presentan mejor índice de precisión sobre la clase que reprueba (Si) es el algoritmo Decisión Table con 90.68% sobre el algoritmo DTNB con 89.11%, PART con 85.96% y el algoritmo Ridor con 85.96%, estos valores fueron obtenidos de la matriz de confusión generada por los algoritmos [89,90].

Por lo tanto en base a los resultados obtenidos se decidió tomar en cuenta los algoritmos: Decisión Table y Ridor, ya que presentan los mejores resultados tanto en rendimiento, clasificación de la clase que reprueba (Si) y las medidas de error [91], no se tomó en cuenta el algoritmo DTNB, porque la tabla de decisión generada construyó reglas poco coherentes y escasamente descriptivas [87,88], ya que formo reglas en base a tres atributos de los 13 seleccionados inicialmente.

Como se puede observar los algoritmos aplicados muestran una clasificación correcta de instancias entre 75% y 87%, debido a la existencia de valores nulos en algunos de los atributos del conjunto de datos [83-85], sin embargo existen sugerencias para solucionar este problema entre las cuales están: *a) Ignorar:* Algunos algoritmos son robustos a datos faltantes; *b) Eliminar/Reemplazar toda la columna:* Si hay muchos valores faltantes no nos servirá. En algunas ocasiones se puede “rehacer” a partir de otra o varias columnas dependientes, pero en este caso no se puede aplicar esta solución ya que no existen datos dependientes; *c) Eliminar la Fila:* Si hay muchas instancias con valores faltantes o nulos, nos quedamos sin ejemplos; *d) Reemplazar el Valor por la media, varianza o moda o bien predecirlo* [84].

Se analizaron y probaron cada una de las posibles soluciones, sin lograr mejorar los resultados obtenidos inicialmente por tal motivo, se aplicó la primera opción, de ignorar los valores faltantes y trabajar con los resultados obtenidos.

Los bajos porcentajes obtenidos en general por los algoritmos enfocados en reprobación demuestra una baja calidad de datos almacenados en las bases de datos de la Universidad Nacional de Loja [86], este problema no se presentaría si la fuente de datos utilizada sería un Data Warehouse, puesto que los datos pasarían por un proceso de

limpieza, selección de datos precisos, selección de datos consistentes y la eliminación de información redundante [92,94]. Todos los inconvenientes detallados anteriormente incurrieron en el rendimiento de los algoritmos enfocados en reprobación y por consiguiente que la clasificación sea inferior al 90%, afectando además los valores obtenidos por las medidas de error, escogidas para determinar el mejor algoritmo en la identificación de factores de reprobación.

4. ETAPA CUATRO: Evaluar Modelo Generado y Análisis de Resultados.

En esta etapa se realizaron actividades con el propósito de evaluar el rendimiento de los algoritmos con el mejor rendimiento con datos nuevos, además se hizo un análisis de los posibles factores de deserción y reprobación.

4.1. Evaluación de modelo y Resultados

En esta actividad se realizó un análisis de los algoritmos con mejor rendimiento para la tarea de predicción de estudiantes desertores.

4.1.1. Quinta Fase: Evaluación

En esta fase se muestra un resumen de los resultados más relevantes de la comparación de los modelos generados y de resultados. Además se evalúa el rendimiento de los algoritmos que ofrecen mejores resultados con estudiantes que cursan actualmente en las carreras del área de energía de la Universidad Nacional de Loja.

4.1.1.1. Tarea Uno: Evaluación de Resultados

En este apartado se realizó una evaluación del rendimiento de dos algoritmos Decisión Table y PART enfocados a predecir los estudiantes desertores, se analizó los factores que inciden tanto en deserción como en reprobación y se trabajó con datos de estudiantes que cursan actualmente en las carreras del Área de Energía, obteniendo los siguientes resultados, para ver los procesos formados para estas pruebas (ver anexo 7).

4.1.1.1.1. Resultados de Predicción con el modelo Decisión Table

En la siguiente figura (ver figura 74) se observa como el algoritmo clasificó cada instancia en base al modelo generado por el algoritmo Decisión Table, estableciendo valores de 1 para las instancias positivas en cada estado (desertor, egresado) a predecir y 0 para las instancias negativas, sin embargo existen instancias que el algoritmo no clasifica correctamente, debido a que el modelo solo maneja sus resultados de forma binomial ya

sea 1 o 0, es decir para valores intermedios no se establece un resultado, mostrando el signo de interrogación como valor.

ExampleSet (767 examples, 4 special attributes, 17 regular attributes)

Row No.	numeroIdentificacion	confidence(desertor)	confidence(egresado)	prediction(estado)
132	1104136880	0.963	0.037	desertor
133	1104141260	0.147	0.853	egresado
134	1104143035	0.688	0.312	desertor
135	1104148307	0.006	0.994	egresado
136	1104151285	0.854	0.146	desertor
137	1104166374	0.969	0.031	desertor
138	1104175367	0.006	0.994	egresado
139	1104178437	0.936	0.064	desertor
140	1104182348	0.026	0.974	egresado
141	1104187180	0.500	0.500	desertor
142	1104191158	0.147	0.853	egresado
143	1104194608	0.333	0.667	egresado
144	1104203672	0.966	0.034	desertor
145	1104208275	0.969	0.031	desertor
146	1104218910	0.969	0.031	desertor
147	1104220007	0.854	0.146	desertor

Figura 74: Resultados de predicción de deserción con el algoritmo Decisión Table.

En la siguiente figura (ver figura 75) se describen a los estudiantes pertenecen al grupo de posibles desertores y egresados, dando un total de 375 posibles desertores y 392 posibles egresados que corresponden al 48.89% y 51.11% respectivamente, dando así un total de 100% de instancias clasificadas.

ExampleSet (767 examples, 4 special attributes, 17 regular attributes)

Role	Name	Type	Range	Sum
id	numeroIdentificacion	nominal	0104646815 (1), 0105220347 (1), 0603951815	?
confidence_desertor	confidence(desertor)	real	[0.004 ; 0.990]	366.438
confidence_egresado	confidence(egresado)	real	[0.010 ; 0.996]	400.562
prediction	prediction(estado)	nominal	desertor (375), egresado (392)	?

Figura 75: Distribución de estudiantes clasificados por el modelo obtenido del algoritmo Decisión Table.

4.1.1.1.2. Resultados de Predicción con el modelo PART

En la siguiente figura (ver figura 76) se observa la clasificación del algoritmo para cada instancia en base al modelo generado estableciendo valores de 1 para las instancias positivas en cada estado a predecir (desertor, egresado) y 0 para las instancias negativas, y a los valores intermedios que el algoritmo no clasifico en su totalidad se describen en proporciones en donde más se aproxime a 1, pertenecerá a dicho estado.

ExampleSet (767 examples, 4 special attributes, 17 regular attributes)				
Row No.	numeroIdentificacion	confidence(desertor)	confidence(egresado)	prediction(estado)
1	0104646815	0.941	0.059	desertor
2	0105220347	0	1	egresado
3	0603951815	0.966	0.034	desertor
4	0703745349	1	0	desertor
5	0704300821	0.966	0.034	desertor
6	0704405596	0.997	0.003	desertor
7	0704406271	0	1	egresado
8	0704412923	0	1	egresado
9	0704418144	0	1	egresado
10	0704634427	0	1	egresado
11	0704655646	0.966	0.034	desertor
12	0704674704	0.966	0.034	desertor
13	0704813476	0.997	0.003	desertor
14	0705006369	0	1	egresado

Figura 76: Resultados de predicción de deserción con el algoritmo PART.

En la siguiente figura (ver figura 77) se describen los estudiantes que pertenecen al grupo de posibles desertores y egresados, dando un total de 369 posibles desertores y 398 posibles egresados que corresponden al 48.11% y 51.89% respectivamente, dando así un total de 100% de instancias correctamente clasificadas.

ExampleSet (767 examples, 4 special attributes, 17 regular attributes)				
Role	Name	Type	Range	Sum
id	numeroIdentificacion	nominal	0104646815 (1), 0105220347 (1), (	?
confidence_desertor	confidence(desertor)	real	[0.000 ; 1.000]	371.140
confidence_egresado	confidence(egresado)	real	[0.000 ; 1.000]	395.860
prediction	prediction(estado)	nominal	desertor (369), egresado (398)	?

Figura 77: Distribución de estudiantes clasificados por el modelo obtenido del algoritmo PART.

4.1.1.1.3. Comparación de Resultados de Predicción

Los resultados que se presentan a continuación reflejan el rendimiento de los modelos generados por cada algoritmo en donde se puede observar rendimientos similar en ambos algoritmos, puesto que se clasificó todas las instancias correctamente, es decir no hubo instancias sin clasificar.

El algoritmo Decisión Table, tomo en cuenta siete de los 17 atributos seleccionados para generar la tabla de decisión con las reglas de predicción, mientras que el algoritmo PART si tomo en cuenta todos los atributos y generar reglas, dejando una cantidad menor de registros sin clasificar (ver figura 78).

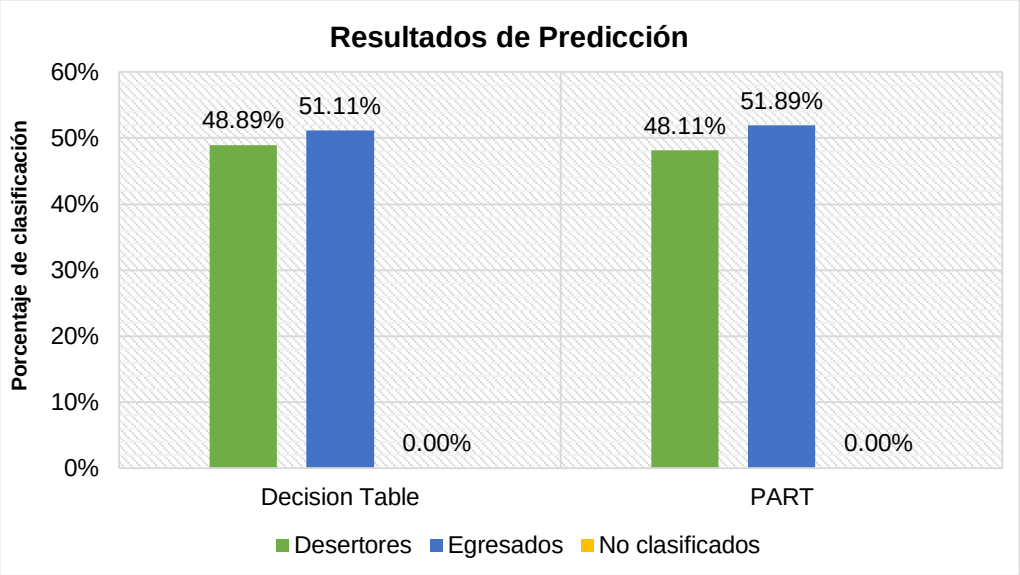


Figura 78: Comparación de resultados obtenidos por Decision Table y PART en predicción de Desertores.

El gráfico muestra que ambos modelos generados por el algoritmo Decisión Table y PART clasificaron todas las instancias y obtuvieron grados de predicción para estudiantes desertores y egresados.

Es por ello que el modelo generado por el algoritmo PART (ver anexo 8) ofrece los mejores resultados, puesto que presenta medidas de error bajos y genera resultados de predicción para todos los estudiantes y con ello determinar si un estudiante está en peligro de abandonar los estudios (ver anexo 9).

4.1.1.1.4. Análisis de Factores de Deserción

Como se estableció al comienzo del Trabajo de Titulación, se tiene como objetivo encontrar los factores de deserción y reprobación, es por ello que se ha planteado un conjunto de atributos asociados al problema de deserción universitaria [72,77] agrupados en factores individuales, académicos, institucionales como se describe en la siguiente tabla (ver tabla CV).

También se destaca la importancia de agregar atributos asociados al factor socioeconómico del estudiante para tener una mejor apreciación de la deserción universitaria.

TABLA CV:
ATRIBUTOS ASOCIADOS A DESERCIÓN.

Factores	Atributos
Individuales	Servicios, etnia, estado civil, hijos, sexo, distancia de origen, edad de ingreso, estado de gestación.
Académicos	Número de módulos reprobados, periodo de reprobación, promedio de notas, promedio de asistencia, cambio de carrera.
Institucionales	Tipo de beca, servicios de bienestar universitario, carrera, horario de estudio.

Sin embargo, a través de la realización de este trabajo de investigación con los datos de los estudiantes del área de energía de la Universidad Nacional de Loja, se puede apreciar que los sistemas de información actuales de esta institución, no están orientados a capturar las variables relevantes para la realización de estudios de minería de datos educativos asociados a la deserción o reprobación.

En base a la propuesta descrita anteriormente (ver tabla CVIII), a continuación se describen los resultados obtenidos, para determinar los factores de mayor incidencia en la deserción universitaria.

Se realizó una evaluación de los atributos de mayor peso al momento de generar el modelo con los algoritmos Decisión Table y PART.

Con el algoritmo Decisión Table se obtuvieron los siguientes resultados (ver figura 79).

attribute	weight
modulos_reprobados	1
periodo_reprobacion	0.639
promedio_notas	0.396
promedio_asistencia	0.251
tipo_beca	0.227
edad_ingreso	0.185
estado_gestancion	0.140
etnia	0.104
carrera	0.050
estado_civil	0.037
cambio_carrera	0.020
hijos	0.010
servicios	0.010
distancia_origen	0.007
bienestar_servicios	0.007
sexo	0.004
horario_estudio	0.001

Figura 79: Pesos de atributos con el modelo Decision Table.

A continuación se describe el peso de cada atributo y al factor asociado en este caso (ver tabla CVI).

TABLA CVI:
COMPARACIÓN DE PESOS CON EL ALGORITMO DECISION TABLE.

Factores	Atributos	Peso	Suma	% de aproximación
Individuales	servicios	0.010	0.497	6.2125% / 100%
	etnia	0.104		
	estado_civil	0.037		
	hijos	0.010		
	sexo	0.004		
	distancia_origen	0.007		
	edad_ingreso	0.185		
	estado_gestancion	0.140		
Académicos	modulos_reprobados	1	2.306	46.12% / 100%
	periodo_reprobacion	0.639		
	promedio_notas	0.396		
	promedio_asistencia	0.251		
	cambio_carrera	0.020		
Institucionales	tipo_beca	0.227	0.285	7.125% / 100%
	bienestar_servicios	0.007		
	Carrera	0.050		
	horario_estudio	0.001		

En base a los resultados obtenidos y descritos en la tabla anterior el factor que más incide es el académico y esto es lógico puesto que el atributo *modulos reprobados* describe perfectamente a los estudiantes desertores de los egresados, sin dejar de lado a los atributos *periodo reprobacion, promedio notas, promedio asistencias* y *cambio carrera,* además hay que señalar que los atributos asociados con el factor institucional se ubican por sobre los atributos asociados a los factor individuales.

Con el algoritmo PART también se realizó una evaluación de los atributos de mayor peso al momento de generar el modelo, a continuación se describen los resultados obtenidos, para determinar los factores de mayor incidencia en la deserción universitaria (ver figura 80).

attribute	weight
modulos_reprobados	1
periodo_reprobacion	0.639
promedio_notas	0.396
promedio_asistencia	0.251
tipo_beca	0.227
edad_ingreso	0.185
estado_gestancion	0.140
etnia	0.104
carrera	0.050
estado_civil	0.037
cambio_carrera	0.020
hijos	0.010
servicios	0.010
distancia_origen	0.007
bienestar_servicios	0.007
sexo	0.004
horario_estudio	0.001

Figura 80: Pesos de atributos en el modelo PART.

A continuación se describe el peso de cada atributo y al factor asociado en este caso (ver tabla CVII).

TABLA CVII:
COMPARACIÓN DE PESOS CON EL ALGORITMO PART.

Factores	Atributos	Peso	Suma	% de aproximación
Individuales	servicios	0.010	0.497	6.2125% / 100%
	etnia	0.104		
	estado_civil	0.037		
	hijos	0.010		
	sexo	0.004		
	distancia_origen	0.007		
	edad_ingreso	0.185		
	estado_gestancion	0.140		
Académicos	modulos_reprobados	1	2.306	45.94% / 100%
	periodo_reprobacion	0.639		
	promedio_notas	0.396		
	promedio_asistencia	0.251		
	cambio_carrera	0.020		
Institucionales	tipo_beca	0.227	0.285	7.125% / 100%
	bienestar_servicios	0.007		
	carrera	0.050		
	horario_estudio	0.001		

En base a los resultados obtenidos y descritos en la tabla anterior el factor que más incide es el académico y en parte es lógico puesto que el atributo *módulos reprobados* describe perfectamente a los estudiantes desertores de los egresados, sin dejar de lado a los

atributos *periodo reprobación, promedio notas, promedio asistencias* y *cambio carrera,* hay que mencionar que los atributos asociados con el factor institucional se ubica en el segundo lugar por sobre los individuales.

Luego de haber evaluado los atributos más relevantes con cada uno de los algoritmos de mejor rendimiento (Decisión Table y PART), se puede observar resultados similares en cuanto a los atributos más relevantes.

En base a lo descrito anteriormente se concluye que el factor que más incide en la deserción de estudiantes del Área de Energía las Industrias y los Recursos Naturales No Renovables es el Académico seguido del factor Institucional y por último el Individual.

4.1.1.1.5. Análisis de Factores de Reprobación

Como se estableció al comienzo del Trabajo de Titulación, se tiene como objetivo encontrar los factores de reprobación, es por ello que en esta tarea se realizó analizando el rendimiento y medidas de error de seis algoritmos basados en reglas: Decisión Table, JRip, Part, Ridor, DTNB y NNge y en base a las reglas generadas interpretar las situación o comportamiento en la que un estudiante reprueba.

En base a lo descrito anteriormente, se estableció utilizar las reglas generadas por el algoritmo Ridor, puesto que generó reglas más coherentes y utilizando todos los atributos seleccionados inicialmente, esto no ocurrió en el algoritmo Decisión Table que solamente generó reglas en base a cinco de los 13 atributos iniciales [87,88].

En cuanto a las medidas de error el algoritmo Decisión Table obtuvo mejor porcentaje de clasificación, mejor índice de Kappa y Error Cuadrático Relativo, mientras que el algoritmo Ridor obtuvo mejores resultados en las medidas de Error Absoluto y Error Relativo.

Por lo tanto se llega a la conclusión de utilizar las reglas generadas por el algoritmo Ridor para interpretar los factores o comportamiento en la reprobación de estudiantes, a continuación se describen las mejores reglas por orden de relevancia.

- **Regla 1:** Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo y pertenecen a la carrera de Ingeniería en Sistemas entonces si reprueba.
- **Regla 2:** Si el estudiante tiene un promedio de notas regular, pertenecen a la carrera de Ingeniería en Electromecánica, son origen rural de la ciudad de Loja, poseen los servicios básicos y son solteros entonces si reprueba.
- **Regla 3:** Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo, pertenecen a la carrera de Ingeniería en Electromecánica con edad de ingreso a la universidad mayor a 20 años entonces si reprueba.
- **Regla 4:** Si el estudiante tiene un promedio de notas regular, estudia por la tarde, tiene un promedio de asistencia medio, son de origen rural y no han recibido servicios de bienestar universitario entonces si reprueba.

4.1.1.2. Tarea Dos: Evaluación de los Resultados de la Minería con respecto a los Factores Críticos.

En esta tarea se realizó una evaluación de los resultados obtenidos en el Trabajo de Titulación, con los factores críticos de éxito que se establecieron al inicio del mismo.

4.1.1.2.1. Evaluación con respecto al primero de los Factores Críticos

En este Trabajo de Titulación también se consideró como factor de éxito la posibilidad de aplicar técnicas de minería de datos para predecir si un estudiante tiene probabilidades de abandonar los estudios, para ello se tiene en cuenta la utilización de datos personales así como notas académicas y el modulo y las materias en las cuales los estudiantes reprobaron.

Con respecto a este factor crítico se utilizaron los dos algoritmos que ofrecían los mejores resultados en rendimiento como es Decisión Table y PART.

- El modelo generado por el algoritmo basado en reglas Decisión Table, generó 62 reglas en base a cinco atributos de los 17 seleccionados inicialmente y lograron clasificar o establecer el grado de deserción de todos los 767 (100%) estudiantes resultando así 48.89% de posibles desertores y 51.11% de posibles egresados.

- El modelo generado por el algoritmo de inducción de reglas PART, generó 38 reglas en base a los 17 atributos seleccionados inicialmente, clasificó o estableció el grado de deserción de los 767 (100%) estudiantes resultando así 48.11% de posibles desertores y 51.89% de posibles egresados.

4.1.1.2.2. Evaluación con respecto al segundo de los Factores Críticos

Uno de los criterios que se toma en cuenta en el desarrollo del presente Trabajo de Titulación es identificar al menos un factor que incida en la deserción y reprobación, es por ello que se utilizó datos personales de estudiantes que pertenecen al Área de Energía de la Universidad Nacional de Loja, estos datos obtenidos del Web Services se migraron a una base de datos relacional para su posterior selección, limpieza y transformación de los mismos.

Con respecto a este factor crítico, después de haber evaluado los resultados con los modelos y reglas generados por los algoritmos Decisión Table y Ridor estos son los resultados obtenidos:

- Para la identificación de factores de deserción se aplicó una evaluación de los atributos más relevantes de los modelos y reglas generados, esto se obtuvo al momento que los datos fueron procesados por el modelo, los resultados arrojados por parte de los dos algoritmos son similares, de los cuales se obtuvo como factor predominante al Académico, y lo que genera más interés es el factor ubicado en segundo lugar, el Institucional y colocando al factor Individual en el último lugar.
- Para los factores de reprobación se realizó un análisis de las reglas más relevantes asociadas a la reprobación en el Área de Energía de la Universidad Nacional de Loja, estas permiten interpretar la situación en la que un estudiante reprueba, el algoritmo con mejor rendimiento fue PART se clasificó aproximadamente el 80% de los estudiantes.

4.2. Elaboración de un artículo científico aplicando estándar IEEE

Los resultados obtenidos en el presente trabajo fueron plasmados de formas diferentes, siendo la memoria final del Trabajo de Titulación el que contiene todos los procesos

realizados para el cumplimiento de los objetivos, además de un resumen ejecutivo en el cual se detallan los resultados obtenidos (ver anexo 10) y un artículo científico con el estándar IEEE (ver anexo 12).

g. Discusión

1. Desarrollo de la propuesta alternativa

La propuesta alternativa describe el proceso realizado para el cumplimiento de los objetivos.

- **Análisis y Muestreo de los datos existentes en las Bases de Datos de la Universidad Nacional de Loja para su Procesamiento.**

En el primer objetivo se realizó entrevistas con responsables del departamento UTI, entidad que gestiona la información, esto se realizó con la finalidad de obtener permisos de acceso a la información almacenada en las bases de datos del Sistema de Gestión Académica (S.G.A.) de la Universidad Nacional de Loja (ver anexo 1).

Posteriormente una vez concedidos los permisos necesarios se analizó los datos almacenados, en cuanto a estructura y calidad de los datos, también se analizó las categorías en las que se encuentran los datos y los métodos descritos para obtener los datos a través de un Web Services.

Todo lo descrito anteriormente, se encuentra detallado en la Etapa Uno de la sección Resultados.

- **Examinar Herramientas para Exploración de Bases de Datos, Proceso de Minería de Datos, y Revisión de Técnicas que permitan resolver el problema planteado.**

En este objetivo primeramente se realizó una recopilación de información y análisis de herramientas de administración de datos utilizada en la preparación de los datos, luego se seleccionó una herramienta la cual fue DatAdmin, puesto que ofrece varias posibilidades de exportar la información, además de una licencia libre para su utilización.

Posteriormente se realizó una búsqueda y recopilación de casos de estudio la aplicación de diferentes técnicas de minería de datos, útiles para la identificación de factores de deserción y reprobación, en base a los casos de estudio recopilados se realizó una

comparación entre los mismos. Se realizó una recopilación de información y evaluación de las herramientas de apoyo al proceso de minería de datos y como resultado de esto fue elegida la herramienta Rapid Miner por su extenso número de componentes a utilizar y por su licencia Open Source.

Luego se realizó una recopilación de información sobre las diversas técnicas de minería de datos y como resultado de esto se trabajó con árboles de decisión y reglas de inducción.

Todo lo descrito en este apartado se encuentra detallado en la Etapa Dos de la sección resultados.

- **Generar Modelos para la Identificación de Factores y Patrones de Comportamiento**

Para completar el presente objetivo se procedió a desarrollar cada una de las fases que definen la minería de datos, para ello se realizó una integración y recopilación de los datos que se van a utilizar en el presente Trabajo de Titulación. Posteriormente se realizó una selección, limpieza y transformación de los datos para luego generar dos estructuras una para deserción y otra para reprobación, luego de ello proceder a realizar la minería sobre los datos.

Luego de las actividades mencionadas, se generó modelos con los diferentes algoritmos seleccionados (ID3, C4.5, CHAID, Decision Table, DTNB, JRip, Ridor, NNge y Part) enfocados en identificar factores de deserción y los algoritmos (JRip, Decision Table, DTNB, Ridor, NNge y Part) enfocados en identificar factores de reprobación. Para generar los modelos se realizó una primera prueba con un conjunto de entrenamiento con el 70% de la muestra total de estudiantes y luego mediante la técnica de validación cruzada se comprobaron los modelos, esto se realizó con el total de los estudiantes.

Todo lo descrito en este apartado se encuentra detallado en la Etapa Tres de la sección resultados.

- **Evaluar Modelo Generado y Análisis de los Resultados**

Para completar el presente objetivo se realizó una evaluación de los algoritmos que mejores resultados ofrecieron enfocados en identificar los factores de deserción (Decisión Table y PART), luego se realizó una valoración de los pesos que reciben los atributos al momento de generar los modelos, y en base a dicha valoración, asociarlos con tres tipos de factores: individuales, académicos e institucionales dando como predominante al factor académico seguido del institucional y en último lugar al individual.

Posteriormente el modelo que se obtuvo los mejores resultados fue el generado por el algoritmo PART, este se aplicó con el fin de predecir si un estudiante que cursa actualmente las carreras del área de energía esta con posibilidades de desertar o abandonar los estudios.

Además de evaluar los factores de deserción también se aplicó los algoritmos JRip, Decision Table, DTNB, Ridor, NNge y PART, enfocados en identificar factores de reprobación, estos generaron reglas que permiten establecer situaciones o circunstancias en las cuales un estudiante reprueba. Siendo el algoritmo Ridor el que mejores resultados obtuvo. Para finalizar se desarrolló, un informe ejecutivo con todos los resultados obtenidos y un artículo científico siguiendo los estándares IEEE para su posterior uso, y aquel que lo considere necesario.

Todo lo descrito en este apartado se encuentra detallado en la Etapa Cuatro de la sección resultados.

Para culminar, en base a lo descrito se puede asegurar el cumplimiento total y exitoso de la hipótesis planteada, ya que la minería de datos permitió identificar los factores que inciden en la deserción y reprobación de los estudiantes del área de energía. Puesto que, los resultados obtenidos confirman en que los objetivos planteados fueron alcanzados y cumplidos en su totalidad, logrando en su conjunto la culminación exitosa del trabajo de titulación, concluyendo además que dichos resultados dejan abierta una importante línea de investigación.

2. Valoración Técnica Económica Ambiental

En el presente trabajo de titulación denominado “Identificación de Factores en la Reprobación y Deserción mediante técnicas de Minería de Datos en el Área de la Energía de la Universidad Nacional de Loja”, da como resultado la identificación de los factores que inciden mayormente en la deserción y reprobación, siendo un aporte viable para las institución de nivel superior, ya que sirve para la toma de decisiones en cuanto a estrategias para disminuir la deserción y mejorar la formación académica del estudiante.

La valoración económica del proyecto tiene su base en que el desarrollo del proyecto se ajusta a los intereses de la institución puesto que el software utilizado y los métodos empleados no implican costes, ni pagos, ni licencias permitiendo obtener ahorros significativos. Destacando además que el trabajo no tiene ningún impacto negativo en el ecosistema ya que no existe peligro alguno para el medio ambiente al momento de aplicar técnicas de minería de datos.

A continuación se detalla el talento humano, los bienes y servicios utilizados en el trabajo de titulación:

El talento humano que participó en el trabajo de Titulación, está conformado principalmente por el investigador quien fue el encargado de llevar acabo el desarrollo del presente trabajo, también como administrador de base de datos, el asesor del proyecto como director, quien fue guía para esquematizar el anteproyecto así como para el desarrollo y culminación del mismo. En la siguiente tabla (ver tabla CVIII) se detalla un estimado del tiempo y costo asignado al investigador, asesor y director, responsables de la culminación exitosa del Trabajo de Titulación.

TABLA CVIII:
COSTO POR HORA DE INTEGRANTES DE PROYECTO.

Integrante	Sueldo por hora ($)	Horas invertidas	Coste Total ($)
Asesor de Proyecto	$ 10,31	83,7	$ 862,95
Investigador	$ 12,07	404,7	$ 4884,73
Administrador de Base de Datos.	$ 14,21	142,6	$ 2026.35
		Total:	$ 7774,03

A continuación se, detalla los Recursos Hardware que fueron empleados, siendo una de ellas el computador portátil utilizado para la manipulación de los datos y manejo de la herramienta de minería de datos, así como para la redacción de los informes que detallan todo el proceso realizado (ver tabla C).

También se describen en la misma tabla, los recursos Software para el procesamiento de los datos y aplicación de los algoritmos de minería de datos, siendo la herramienta Rapid Miner la cual no posee costo alguno, así como el administrador de base de Datos DatAdmin y el procesador de Textos científicos TexMaker.

Los Recursos Materiales implicados en la elaboración del presente trabajo se detallan igual en la misma tabla CIX, siendo estos necesarios para la presentación de borradores e informes finales.

TABLA CIX:
HARDWARE, SOFTWARE, MATERIALES Y SERVICIOS.

HARDWARE			
Equipo	**Cantidad**	**Precio Unitario ($)**	**Precio Total ($)**
Ordenador Portátil	1	$ 940,00	$ 940,00
Impresora	1	$ 73,00	$ 73,00
Subtotal			$ 1.013,00
SOFTWARE			
Producto	**Cantidad**	**Precio Unitario ($)**	**Coste Total ($)**
Rapid Miner	1	$ 0.00	$ 0.00
TexMaker	1	$ 0.00	$ 0.00
DatAdmin	1	$ 0.00	$ 0.00
Subtotal:			$ 0.00
MATERIALES Y SERVICIOS			
Ítems	**Descripción**		**Coste Total ($)**
Materiales de Oficina	Papel, tinta, carpetas, etc.		$ 101,80
Servicios Básicos	Servicios de agua, luz e internet.		$ 329,79
Transporte	Recorridos en un aproximado de 60.		$ 26,00
Publicación de Resultados.	Publicación en revista indexada.		$ 260,00
Cursos de Capacitación	Seminarios, talleres.		$ 250,00
Subtotal:			$ 967,59
Total:			$1980.00

Finalmente se presenta la suma total del talento humano, Hardware, Software, Materiales y Servicios, utilizados en el trabajo de Titulación, siendo una aproximación del coste real (Ver tabla CX).

TABLA CX:
RESUMEN DE COSTES DEL PRESUPUESTO.

Recursos	**Coste Total ($)**
Costes de Personal	$ 7774,03
Costes de Hardware	$ 1013,00
Costes de Software	$ 0,00
Costes de Material de Oficina y Servicios.	$ 967,59
Subtotal:	$ 9754.62
Imprevistos 10%	$ 975.46
Total:	$ 10730.08

La tabla anterior detalla el costo total por cada recurso utilizado, además hay que mencionar que el incremento en el presupuesto, con respecto al propuesto inicialmente en el anteproyecto. Debido a que uno de los recursos hardware, concretamente el ordenador portátil sufrió daños irreparables, por tal motivo se hizo uso de nuevo ordenador portátil para continuar con el desarrollo del proyecto el cual se encuentra detallado en la tabla CIX.

En cuanto a tiempos estimados, cabe recalcar que se cumplieron con éxito cada uno de los puntos establecidos con anticipación.

h. Conclusiones

Al culminar el presente Trabajo de Titulación, se describe las conclusiones obtenidas:

- En el trabajo realizado se llegó a la conclusión de que el factor que incide mayormente en la deserción de los estudiantes es el académico, seguido por el institucional dejando por último al factor individual, mientras que en el ambiente de reprobación los estudiantes que reprueban presentan la siguiente condición: Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo y pertenecen a la carrera de Ingeniería en Sistemas entonces si reprueba.
- Se analizaron los datos que posee la institución y se pudo confirmar que en el área de energía los estudiantes reprueban en cada periodo académico un promedio de 23%.
- En el transcurso de la tarea de modelado se han obtenido excelentes resultados con los algoritmos basados en reglas como Ridor, Part y JRip.
- Se han examinado distintos algoritmos de clasificación y una de las conclusiones que se obtiene del análisis de los resultados obtenidos es que en este caso los algoritmos que generaban modelos más sencillos son los que presentan mejores resultados.
- Con el desarrollo del presente trabajo se ha aprendido a seguir la metodología CRISP-DM en un proyecto. Ya que permitió retomar y repetir fases anteriores, para generar nuevos modelos que permitan alcanzar los objetivos propuestos.
- Como conclusión final hay que recalcar que se alcanzaron todos los objetivos establecidos al inicio del Trabajo de Titulación. El primero de ellos fue el análisis de las fuentes de datos, en donde se analizó la calidad y estructura de los datos. También se consiguió elaborar distintos modelos y reglas con los que se pudo identificar factores de deserción y reprobación, además de un modelo predictivo de deserción con el cual se obtuvo excelentes resultados.

i. Recomendaciones

Una vez concluido el Trabajo de Titulación, se considera interesante proporcionar las siguientes recomendaciones:

- De acuerdo a la experiencia adquirida a través del Trabajo de Titulación con los datos de los estudiantes del Área de Energía de la Universidad Nacional de Loja, se puede apreciar que los sistemas de información actuales de la institución no están orientados a capturar variables relevantes para la realización de estudios de minería, por lo tanto se recomienda capturar datos socioeconómicos y de los niveles de formación académica hasta el momento, al momento de matricularse en la institución.
- Es necesario al momento de analizar la información, implementar varias técnicas de minería de datos, y en base a ello comparar los resultados, y confirmar cual técnica resulta ser la más eficaz para el problema que se desea resolver.
- Se recomienda invertir tiempo en el estudio en métodos y parámetros para obtener información y luego almacenarla en una base de datos ya que la tarea más costosa a lo largo del proyecto fue la recopilación y preparación de los datos puesto que se obtuvieron a través de un Web Services.
- Para poder generar modelos de forma correcta y ordenada es importante, usar una metodología para el desarrollo de proyectos de minería de datos, siendo CRISP-DM una de las más usadas en el ámbito académico e industrial, ya que propone las fases necesarias para generar uno o varios modelos de calidad.
- Al momento de generar modelos para analizar factores de deserción y reprobación es recomendable aplicar la técnica de reglas de inducción, puesto que al desarrollar el presente trabajo esta técnica es la que obtiene los mejores resultados.
- Se recomienda la construcción de un almacén de datos (Data Warehouse) para la Universidad Nacional de Loja, que permita obtener datos de calidad para proyectos de minería de datos que se realicen en un futuro.
- Además se encomienda implantar el modelo predictivo de deserción en una plataforma informática de la universidad.
- Se propone además, continuar con el análisis de factores de deserción, en todas las áreas y carreras de la Universidad Nacional de Loja, tomando en cuenta datos académicos, personales, institucionales y socioeconómicos.

j. Bibliografía

Referencias Bibliográficas

[1] L. L. PINZON CADENA, “Aplicando minería de datos al marketing educativo”, Enero-Junio 2011, vol. No. 1, no. Notas D Marketing, pp. 45–61.

[2] M. Y. SARANGO SEDAMANOS, “Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL”, Universidad Técnica Particular de Loja, Loja Ecuador, 2012.

[3] B. M. Latiesa Rodríguez, La deserción universitaria: desarrollo de la escolaridad en la enseñanza superior: éxitos y fracasos. 1992.

[4] E. Apaza and F. Huamán, “Factores determinantes que inciden en la deserción de los estudiantes universitarios”, Apuntes Universitarios, vol. 2, no. 1, pp. 77–86, 2012.

[5] A. V. D. López, “Estrategias para vencer la deserción universitaria”, Educación y educadores, vol. 7, 2004.

[6] O. M. F. González, M. M.-C. Beluzan, and R. M. Araneda, “Estrategias de aprendizaje y autoestima. Su relación con la permanencia y deserción universitaria,” Estudios pedagógicos, vol. 35, no. 1, pp. 27–45, 2009.

[7] M. Boado, “Una aproximación a la deserción estudiantil universitaria en Uruguay,” Universidad de la República, Montevideo, Uruguay, en cooperación con el Instituto Internacional para la Educación Superior en América Latina y el Caribe, pp. 10–24, 2005.

[8] G. J. Paramo and C. A. C. Maya, “Deserción estudiantil universitaria. Conceptualización,” Revista Universidad EAFIT, vol. 35, no. 114, pp. 65–78, 2012.

[9] M. C. G. AFONSO, P. R. A. Pérez, L. C. PÉREZ, and J. T. B. Benítez, “El abandono de los estudios universitarios: factores determinantes y medidas preventivas,” Revista española de pedagogía, vol. 65, no. 236, pp. 71–88, 2007.

[10] E. Castaño, S. Gallón, and J. Vásquez, “Análisis de los factores asociados a la deserción estudiantil en la educación superior: un estudio de caso,” Revista de Educación, no. 345, pp. 255–280, 2008.

[11] E. B. Durán and R. N. Costaguta, “Minería de datos para descubrir estilos de aprendizaje”, Revista Iberoamericana de Educación, vol. 42, no. 2, p. 6, 2007.

[12] E. G. Salcines, C. Romero, S. Ventura, and C. de-Castro-Lozano, “Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning.,” IEEE-RITA, vol. 3, no. 1, pp. 19–30, 2008.

[13] M. J. E. Pinazo and R. G. Pérez, “Minería de datos educativos en plataformas virtuales de aprendizaje musical”, Revista electrónica de LEEME, no. 27, pp. 1–16, 2011.

[14] J. H. Orallo, M. J. R. Quintana, and C. F. Ramírez, Introducción a la Minería de Datos. Pearson Prentice Hall, 2004.

[15] K. Gilbert, R. R. Sánchez, and J. C. R. Santos, “Mineria de datos: Conceptos y tendencias,” Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, vol. 10, no. 29, pp. 11–18, 2006.

[16] R. Nisbet, J. Elder IV, and G. Miner, Handbook of statistical analysis and data mining applications. Access Online vía Elsevier, 2009.

[17] E. Yolis, P. Britos, G. Perichisky, and R. García-Martínez, “Algoritmos Genéticos Aplicados a la Categorización Automática de Documentos,” Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI, vol. 2, no. 2, 2003.

[18] “Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior - ICETEX-Ministerio de Educación Nacional de Colombia” [En línea]. Disponible: http://www.mineducacion.gov.co/1621/article-85399.html. [Acceso: 13-Ago-2013].

[19] Forero, L., Narváez, S. Modelo de predicción de la deserción universitaria por medio de algoritmos de Data Mining, 2011

[20] Timarán, P.R., Millán, M.: EquipAsso: an Algorithm based on New Relational Algebraic Operators for Association Rules Discovery. In proceedings of the Fourth IASTED International Conference on Computational Intelligence. ACTA Press, Calgary, Alberta, Canadá, 2005.

[21] Hernandez, O. J., Ramirez, Q. M., Ferri, R. C.: “Introducción a la Minería de Datos”. Editorial Pearson Prentice Hall, Madrid, España ,2004.

[22] Witten, I. H., Frank, E: “Data Mining Practical Machine Learning Tools and Techniques”. Morgan Kaufmann Publishers, San Francisco, California, USA, 2005.

[23] Gama J.: “Functional Trees”. Machine Learning pp: 219-250, Springer Netherlands, 2004

[24] Hernandez, O.J., Ramirez, Q.M., Ferri, R.C.: Introducción a la Minería de Datos. Editorial Pearson Prentice Hall, Madrid, España ,2004.

[25] Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules, VLDB Conference, Santiago, Chile, 1994.

[26] Agrawal R., Srikant R.: Mining Sequential Patterns. In Proceedings of the 11th International Conference on Data Engineering, 1995.

[27] Berry, M., Linoff, G.: Data Mining Techniques for Marketing, Sales and Customer Support. Wiley Computer Publishing, 1997.

[28] C. P. LOPEZ, Minería de datos: técnicas y herramientas. Paraninfo Cengage Learning, 2007.

[29] J. M. MOLINA, J. GARCIA, Técnicas de Minería de Datos basadas en Aprendizaje Automático: Técnicas de Análisis de Datos, [En Línea], Disponible: http://santiagozapatakdd.files.wordpress.com/2011/03/curso-kdd-full-cap-3.pdf, [Acceso: 20-Dic-2013]

[30] P. M. ROMEU GUALLART, Minería de Datos Aplicada al Análisis del Tratamiento Informativo de la Drogadicción, [En Línea] Disponible: http://dspace.ceu.es/bitstream/10637/5020/1/Minería de datos aplicada aI análisis del tratamiento informativo de la drogadicción_Romeu Guallart, Pablo María.pdf, [Acceso: 20-Dic-2013]

[31] T. ALUJA BANET, La minería de datos, entre la estadística y la inteligencia artificial. Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, 2001, vol. 25, no 3, p. 479-498.

[32] G. A., BETANCOURT, Las máquinas de soporte vectorial (svms). Scientia et Technica, 2005, vol. 1, no 27.

[33] FÜRNKRANZ J, WIDMER G. Incremental reduced error pruning. Proceedings of the Eleventh International Conference of Machine Learning; 1994; New Brunswick, New Jersey: Morgan Kaufmann; 1994. p. 70-7.

[34] COHEN W. W. Fast Effective Rule Induction. Proceedings of the Twelfth International Conference on Machine Learning; 1995; 1995.

[35] FRANK E, WITTEN IH. Generating Accurate Rule Sets Without Global Optimization. In: Shavlik J, editor. Proceedings of the Fifteenth International Conference on Machine Learning; 1998; San Francisco, CA: Morgan Kaufmann Publishers; 1998.

[36] RIVEST, R. Learning Decision Lists. Machine Learning, 1987, 2(3), 229-246.

[37] V. Valcárcel Asencios, “Data Mining y el descubrimiento del conocimiento”, *Ind. Data*, vol. 7, no. 2, pp. 83–86, 2004.

[38] J. Alcalá-Fdez, M. J. del Jesus, J. M. Garrell, F. Herrera, C. Herbás, and L. Sánchez, “Proyecto KEEL: Desarrollo de una herramienta para el análisis e implementación de

algoritmos de extracción de conocimiento evolutivos," *Tendencias de la Mineria de Datos en Espana, Red Espanola de Mineria de Datos y Aprendizage*, pp. 413–424, 2004.

[39] C. A. González, G. de Sistemas Inteligentes, and J. J. R. Diez, "Introducción a la Minería de Datos y al Aprendizaje Automático".

[40] T. Aluja, "Los nuevos retos de la estadística, el Data Mining".

[41] J. C. Dürsteler, "Visualización de información", *Una visita guiada. Barcelona: Gestión*, 2000.

[42] T. A. Banet, "La minería de datos, entre la estadística y la inteligencia artificial", *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, vol. 25, no. 3, pp. 479–498, 2001.

[43] "Herramientas Software-Minería de datos." [En linea]. Disponible: http://www.uco.es/grupos/ayrna/es/enlaces/45-herramientas-software-datamining, [Acceso: 17-Ago-2013].

[44] "Características - Orange." [En Línea]. Disponible: http://orange.biolab.si/features/ [Acceso: 18-Ago-2013].

[45] Guido Sagasti, "Guía para hacer Data Mining y Framework de Negocios". [En Línea]. Disponible: http://Web.austral.edu.ar/descargas/facultad-ingenieria/20100921-IV-Jornada-DM&BI-SAS-Guido-Sagasti.pdf, [Acceso: 18-Ago-2013].

[46] "SAS | SEMMA." [En Línea]. Disponible: http: //www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html, [Acceso: 18-Ago-2013].

[47] "Software Libre para Minería de Datos - Data Mining Consulting", Ago-2013. [En Línea]. Disponible: http://www.dataminingperu.com/index.php?option=com_content&view=article&id=36&catid=12&Itemid=133, [Acceso: 18-Ago-2013].

[48] J. M. Cadenas, J. V. Carrillo, M. C. Garrido, and E. Munoz, "Nip1. 5: Una herramienta software para la generación de conjuntos de datos con imperfección para minería de datos," in Actas del XV Congreso Español Sobre Tecnologías y Lógica Fuzzy (ESTYLF-2010), pp. 411–416.

[49] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholzand T. Euler, YALE: Rapid Prototyping for Com-plex Data Mining Tasks, in Proc. 12th ACMSIGKDD Int. Conference on Knowledge Disco-very and Data Mining (KDD-06) (2006) 935–940.

[50] A. Gómez, “Inteligencia de negocios, una ventaja competitiva para las organizaciones”, Revista CIENCIA Y TECNOLOGÌA, vol. 8, no. 22, pp. 85–96, 2013.

[51] A. S. Román García, “Minería de datos en encuestas de profesores al fin de semestre de la Facultad de Ingeniería, UNAM”, 2012.

[52] J. C. Cubero and F. Berzal, “Sistemas Inteligentes de Gestión Guión de Prácticas de Minería de Datos Práctica 1 Herramientas de Minería de Datos Introd”, Ago-2013. [En Línea]. Disponible: http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf. [Acceso: 05-Sep-2013].

[53] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, KNIME: The Konstanz information miner. Springer, 2008.

[54] R. Mikut and M. Reischl, “Data mining tools,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 5, pp. 431–443, 2011.

[55] S. R. Garner, “Weka: The waikato environment for knowledge analysis,” in *Proceedings of the New Zealand computer science research students conference*, 1995, pp. 57–64.

[56] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, “Weka”, in *Data Mining and Knowledge Discovery Handbook*, Springer, 2005, pp. 1305–1314.

[57] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, “WEKA Manual for Version 3-7-8,” 2013.

[58] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, “Weka-a machine learning workbench for data mining,” in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, pp. 1269–1277.

[59] “The R Project for Statistical Computing.” [En Línea]. Disponible: http://www.r-project.org/. [Acceso: 05-Sep-2013].

[60] Introducción a R Notas sobre R: Un entorno de programación para Análisis de Datos y GráficosVersión 1.0.1 (2000-05-16) R Development Core Team, “Introducción a R”. [En Línea]. Disponible: http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf [Acceso: 05-Sep-2013].

[61] H. Carrión, “Internet en el Ecuador”, *Recuperado el*, vol. 4, 2011.

[62] M. F. ARTEAGA VÉLEZ and R. R. VINCES PACHECO, “Análisis del Sistema de Comercialización de los Servicios de Internet y Líneas Telefónicas Convencional

Domiciliarias de la Empresa Cnt en la ciudad de Portoviejo y su incidencia en el Crecimiento y Desarrollo Empresarial Durante el Periodo 2010-2011.," 2013.

[63] J. M. Moine, A. Haedo, and S. Gordillo, "Estudio comparativo de metodologías para minería de datos", in *XIII Workshop de Investigadores en Ciencias de la Computación*, 2011.

[64] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0", *CRISP-DM Consortium*, 2000.

[65] J. Gallardo, "Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM" [En Línea], Disponible: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf 2000.

[66] M. Berry, G. Linoff, "Data mining techniques for marketing, sales and customer relationship management", Wiley Publishing, Inc. Indianapolis, 2004, pp. 43 – 86.

[67] SANTOS, Rafael Castillo. Detección de Intrusos mediante Técnicas de Minería de Datos. Facultad de Ingeniería, 2006, p. 31.

[68] Sas Institute, Disponible en: http://www.sas.com/technologies/analytics/datamining/miner/semma.html; Fecha de Acceso: 14-02-2014.

[69] D. Olson, D.Delen, "Advanced Data Mining Techniques". Berlin: Springer - Verlag, 2008, pp 19.

[70] CAMARGO, Hernando; SILVA, Mario. Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP. Rev. Tecnol, vol. 9, no 1.

[71] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[72] Plan Participativo de Fortalecimiento de la Democracia y Desarrollo del Cantón Loja, [En Línea] <http://www.loja.gob.ec/files/docman/diagnostico.pdf.>, [Consulta: 12 de Abril del 2014].

[73] SERVENTE M. y MARTINEZ García R. Algoritmos TDIDT Aplicados a la Minería Inteligente. Revista del Instituto Tecnológico de Buenos Aires, 2002, no 26, p.39-57.

[74] ESPINO José A.; TIJERENA Javier E.; CEDANO Manuel; AMAYA Eleazar de la Fuente; PEREZ Juan J.; CARBALLO Aníbal C., Algortimo C4,5, Nuevo Loredo, Tamaulipas, 2005.

[75] HARTIGAN, J.A.; Clustering Algorithms, 1975, New York: John Wiley & Sons.

[76] VILLARREAL, Rafael Martínez. Modelo de referencia CHISP-DM para el desarrollo de proyectos de minería de datos y CHAID como técnica de análisis para obtener información en minería de datos. InterSedes, 2012, vol. 13, no 25.

[77] Universidad Nacional ICFES, "Estudio de la deserción estudiantil en la educación superior en Colombia", Bogotá, Documento sobre estado del arte 2002.

[78] Data Mining Methodology (Aug 2007), What main methodology are you using for data mining?, [En Línea] <http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.>, [Consulta: 11 de Abril del 2014].

[79] Brian R. Gaines, Paul Cmpton (1995). Induction of Ripple-Down Rules Applied to Modeling Large Databases. J. Intell. Syst.5 (3):211-228.

[80] KOHAVI, Ron. The power of decision tables. En Machine Learning: ECML-95. Springer Berlin Heidelberg, 1995. p. 174-189.

[81] HALL, Mark; FRANK, Eibe. Combining Naive Bayes and Decision Tables. En FLAIRS Conference. 2008. p. 318-319.

[82] Martin B. Instance-Based learning: Nearest Neighbor With Generalization [Master Thesis]. Hamilton, New Zealand: University of Waikato; 1995.

[83] Pérez López, C. "Técnicas de análisis multivariante de datos. Aplicaciones con SPSS". Editorial Pearson Prentice-Hall. Madrid. España. p. 21-22, 39-40, 48. 2004.

[84] DAPOZO, Gladys N., et al. Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios. En IX Workshop de Investigadores en Ciencias de la Computación. 2007.

[85] FORMIA, Sonia. Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. 2013. Tesis Doctoral. Facultad de Informática.

[86] Cortizo Pérez, J. C. "Preprocesado de datos". D. Sistemas Informáticos. Esc. Superior Politécnica. Universidad Europea de Madrid. AINetLab (AINetSolutions). http://www.ainetsolutions.com.

[87] BRITOS, Paola Verónica; GARCÍA MARTÍNEZ, Ramón. Propuesta de Procesos de Explotación de Información. En XV Congreso Argentino de Ciencias de la Computación. 2009.

[88] COHEN, William W.; SINGER, Yoram. A simple, fast, and effective rule learner. En Proceedings of the National Conference on Artificial Intelligence. JOHN WILEY & SONS LTD, 1999. p. 335-342.

[89] CARDENAS, Miguel M.; Precisón del Modelo Gráficas, estadística y minería de datos, Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, Madrid, Spain, 2013.

[90] ZAMORANO, Jordi P., "Técnicas cuantitativas para la extracción de términos en un corpus", Universidad autónoma de Madrid, 2006

[91] HAN, Jiawei; KAMBER, Micheline. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann, 2006.

[92] KANTARDZIC, Mehmed. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[93] WITTEN, Ian H.; FRANK, Eibe. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[94] ZORRILLA, Marta. Data warehouse y OLAP. Universidad de Cantabria, 2007.

k. Anexos

ANEXO 1: Autorización de Acceso a los datos por el departamento Unidad de Telecomunicaciones e Información.

UNIVERSIDAD NACIONAL DE LOJA
UNIDAD DE TELECOMUNICACIONES E NFORMACIÓN

Of. N.605 UTI-UNL
Loja, 29 de Octubre del 2013

Sr.
ANIBAL Israel Gonzales
Ciudad.-

De mi consideración:

Por medio del presente me es grato saludarle a su persona y a la vez comunicarle que se hace la entrega de la clave de permiso para el acceso a los datos a través de los Web Services del Sistema de Gestión Académica

Usuario.- [redacted]
Clave.- [redacted]

Particular que pongo a su conocimiento para los fines pertinentes.

Atentamente,

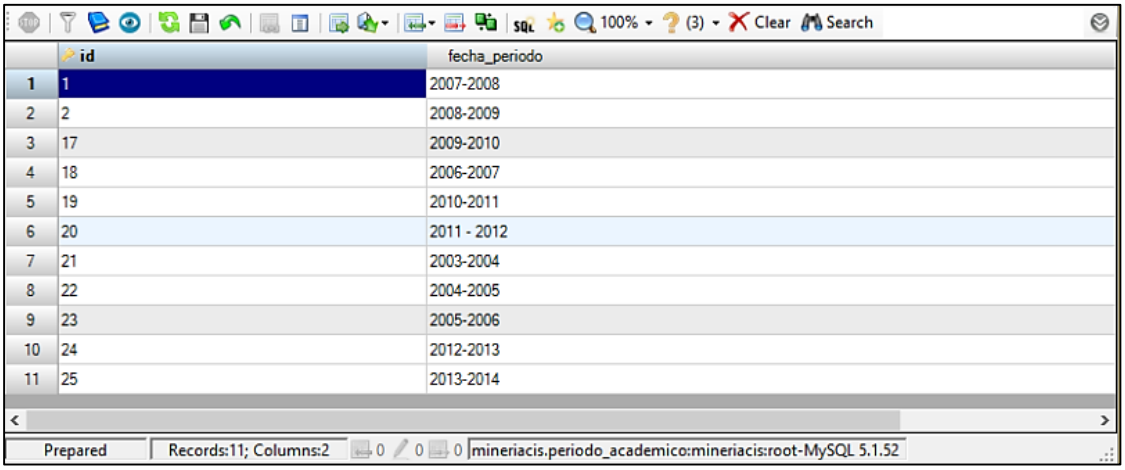
Ing. Milton Palacios
DIRECTOR DE LA UNIDAD
DE TELECOMUNICACIONES E INFORMACIÓN

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: Permiso de acceso al Web Services.

ANEXO 2: Datos del Web Services

Estos son los periodos académicos que se encuentran almacenados en el Sistema de Gestión Académico de la Universidad Nacional de Loja ha ofrecido hasta la actualidad (ver figura 1).

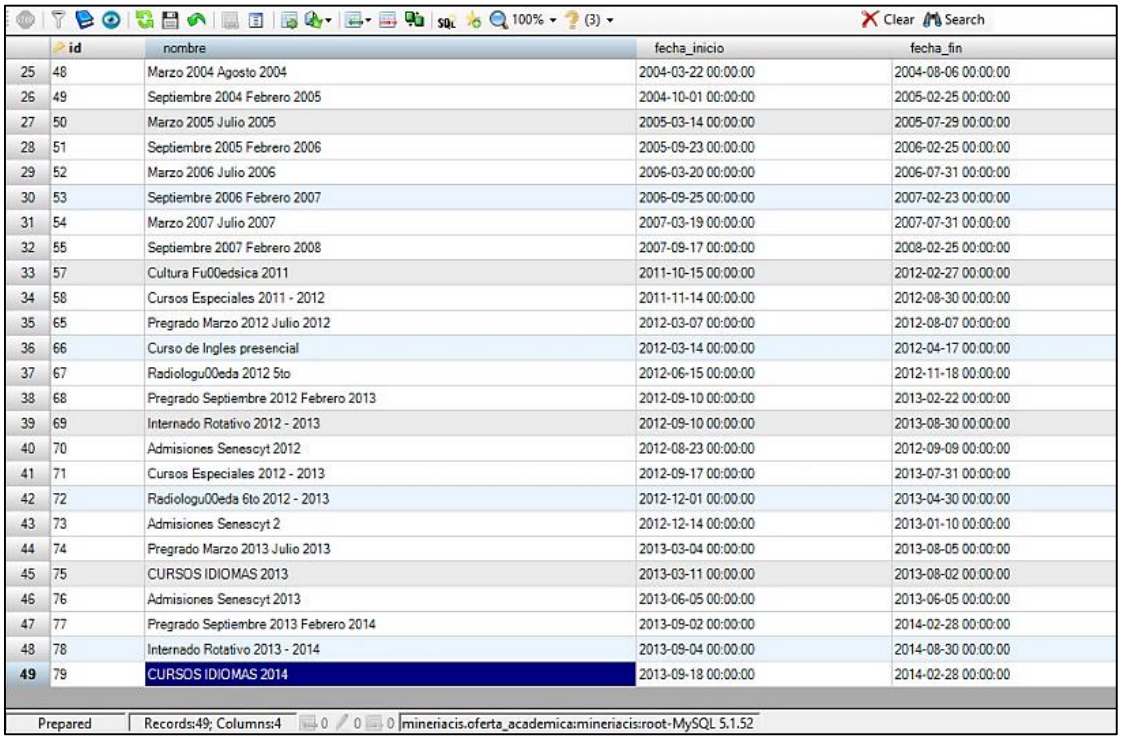


	id	fecha_periodo
1	1	2007-2008
2	2	2008-2009
3	17	2009-2010
4	18	2006-2007
5	19	2010-2011
6	20	2011 - 2012
7	21	2003-2004
8	22	2004-2005
9	23	2005-2006
10	24	2012-2013
11	25	2013-2014

Prepared | Records:11; Columns:2 | mineriacis.periodo_academico:mineriacis:root-MySQL 5.1.52

Figura 1: Periodos académicos de la Universidad Nacional de Loja

En la siguiente figura se muestran algunos de las ofertas académicas que la universidad ha ofrecido y que se encuentran almacenados hasta la actualidad (ver figura 2).



	id	nombre	fecha_inicio	fecha_fin
25	48	Marzo 2004 Agosto 2004	2004-03-22 00:00:00	2004-08-06 00:00:00
26	49	Septiembre 2004 Febrero 2005	2004-10-01 00:00:00	2005-02-25 00:00:00
27	50	Marzo 2005 Julio 2005	2005-03-14 00:00:00	2005-07-29 00:00:00
28	51	Septiembre 2005 Febrero 2006	2005-09-23 00:00:00	2006-02-25 00:00:00
29	52	Marzo 2006 Julio 2006	2006-03-20 00:00:00	2006-07-31 00:00:00
30	53	Septiembre 2006 Febrero 2007	2006-09-25 00:00:00	2007-02-23 00:00:00
31	54	Marzo 2007 Julio 2007	2007-03-19 00:00:00	2007-07-31 00:00:00
32	55	Septiembre 2007 Febrero 2008	2007-09-17 00:00:00	2008-02-25 00:00:00
33	57	Cultura Fu00edsica 2011	2011-10-15 00:00:00	2012-02-27 00:00:00
34	58	Cursos Especiales 2011 - 2012	2011-11-14 00:00:00	2012-08-30 00:00:00
35	65	Pregrado Marzo 2012 Julio 2012	2012-03-07 00:00:00	2012-08-07 00:00:00
36	66	Curso de Ingles presencial	2012-03-14 00:00:00	2012-04-17 00:00:00
37	67	Radiologu00eda 2012 5to	2012-06-15 00:00:00	2012-11-18 00:00:00
38	68	Pregrado Septiembre 2012 Febrero 2013	2012-09-10 00:00:00	2013-02-22 00:00:00
39	69	Internado Rotativo 2012 - 2013	2012-09-10 00:00:00	2013-08-30 00:00:00
40	70	Admisiones Senescyt 2012	2012-08-23 00:00:00	2012-09-09 00:00:00
41	71	Cursos Especiales 2012 - 2013	2012-09-17 00:00:00	2013-07-31 00:00:00
42	72	Radiologu00eda 6to 2012 - 2013	2012-12-01 00:00:00	2013-04-30 00:00:00
43	73	Admisiones Senescyt 2	2012-12-14 00:00:00	2013-01-10 00:00:00
44	74	Pregrado Marzo 2013 Julio 2013	2013-03-04 00:00:00	2013-08-05 00:00:00
45	75	CURSOS IDIOMAS 2013	2013-03-11 00:00:00	2013-08-02 00:00:00
46	76	Admisiones Senescyt 2013	2013-06-05 00:00:00	2013-06-05 00:00:00
47	77	Pregrado Septiembre 2013 Febrero 2014	2013-09-02 00:00:00	2014-02-28 00:00:00
48	78	Internado Rotativo 2013 - 2014	2013-09-04 00:00:00	2014-08-30 00:00:00
49	79	CURSOS IDIOMAS 2014	2013-09-18 00:00:00	2014-02-28 00:00:00

Prepared | Records:49; Columns:4 | mineriacis.oferta_academica:mineriacis:root-MySQL 5.1.52

Figura 2: Ofertas académicas en la Universidad Nacional de Loja.

Se obtuvieron los estudiantes del Área de la Energía las Industrias y los Recursos Naturales no Renovables con el id de los paralelos en los que se encontraron matriculados. Los datos corresponden a un número de 11418 paralelos creados cada periodo académico para al área antes mencionada (ver figura 3).

Paging | Start 0 | Count 200 | OK | Loaded rows 200

	estudiante_id_fk	paralelo_id_fk
163	alex mauricio cau00f1ar sarango	812
164	jairo rodrigo veintimilla ludeu00f1a	812
165	jose salomon yaguache maza	812
166	leonardo leoncio calderon camacho	812
167	luis gustavo luna jaramillo	812
168	carlos ivan mejia torres	812
169	julio david alvarado abad	812
170	edgar lenin merchu00e1n arce	812
171	roberth javier castillo chavez	812
172	jose alberto oviedo jaramillo	812
173	shubert alexis castro meza	812
174	josu00e9 luis patiu00f1o guachu00f3n	812
175	alex geovanny zaavedra ortiz	812
176	johnson benjamin reyes jaramillo	812
177	manuel geovanny ru00edos cuenca	812
178	lenin efren ordu00f3u00f1ez mora	812
179	josu00e9 daniel granda paccha	812
180	diego ernesto soto briceu00f1o	813
181	david alejandro romero zhingre	813
182	robinson stalin apolo aguilar	813
183	julio mauricio chalan jumbo	813
184	miguel eduardo diaz robles	813
185	cesar augusto flores aguirre	813
186	diego fernando piedra betancourt	813
187	diego vinicio puglla remache	813

Prepared | Records:11.418; Columns:2 | mineriacis.estudiante_paralelo:mineriacis:root-MySQL 5.1.52

Figura 3: Estudiantes del Área de Energía de la Universidad Nacional de Loja.

De los estudiantes antes mencionados también se obtuvo las notas obtenidas en las materias cursadas (ver figura 4).

Paging | Start 0 | Count 200 | OK | Loaded rows 200

	unidad_id_fk	nota	estudiante_id_fk		oferta_carrera_id_fk
46	Anu00e1lisis Matemu00e1tico II	7.8	1104725948	jimmy stalin	416
47	Taller Elu00e9ctrico	5.6	1104725948	jimmy stalin	416
48	Electru00f3nica	8.1	1104725948	jimmy stalin	416
49	Herramaientas CAD II	9.1	1104725948	jimmy stalin	416
50	Estu00e1tica	6.5	1104725948	jimmy stalin	416
51	Circuitos I	8.3	1104725948	jimmy stalin	416
52	ANALISIS MATEMATICO I	8.1	1104524143	byron geovanny	159
53	TALLER MECANICO	8	1104524143	byron geovanny	159
54	PROYECTO DE CIEN.BAS 2	7.4	1104524143	byron geovanny	159
55	HERRAMIENTAS CAD II	8.9	1104524143	byron geovanny	258
56	GEOMETRIA PLANA	7	1900391499	jenny maribel	326
57	QUIMICA	7.2	1900391499	jenny maribel	326
58	MATEMATICAS	7	1900391499	jenny maribel	326
59	FISICA	7	1900391499	jenny maribel	326
60	PROCESO INVESTIGATIVO	8.6	1900391499	jenny maribel	326
61	ANALISIS MATEMATICO I	0.3	1104572415	fausto armando	80
62	DIBUJO TECNICO	8.4	0705008373	diego vinicio	80

Figura 4: Notas de estudiantes del Área de Energía.

En la siguiente figura se puede observar que la Universidad Nacional de Loja se encuentra dividida en ocho Áreas (ver figura 5).

Data

	sigla	nombre_area	secretario	director
1	AARNR	Area Agropecuaria y de Recursos Naturales Renovables	Dr. Angel Hernu00e1n Torres Jumbo	Dr. u00c3u0081ngel Benigno Cabrera Achupallas
2	ACE	Cursos Especiales	Dr. Matu00edas Ramu00edrez Bravo, Mg Sc	, null, null
3	AEAC	Area de la Educaciu00f3n, el Arte y la Comunicaciu00f...	Dra. Victoria Raquel Torres Torres	Dra. Rocio Toral
4	AEIRNNR	Area de Energu00eda, Industria y Recursos Naturales...	Dr. Estuardo Napoleu00f3n Rodru00edguez Lu...	Dr. Jorge Reyes Jaramillo
5	AJSA	Area Juru00eddica, Social y Administrativa	Dra. Aida Leticia Carriu00f3n Vega	, null
6	ASH	Area de la Salud Humana	Dra. Ena Regina Pelaez Soria	Dr. Edgar Enrique Benitez Gonzu00c3u00a1lez
7	MED	Modalidad de Estudios a Distancia	Dr. Vicente Cristu00f3bal Analuisa Leu00f3n	Dr. Cu00c3u00a9sar Antonio Leu00c3u00b3n Ag...
8	PREUNIVERSITARI...	PREUNIVERSITARIO	Secretario Preuniversitario	Director Preuniversitario

Read only | Records:8; Columns:4 | mineriacis.area:mineriacis:root-mysql | Open for editing

Figura 5: Áreas de la Universidad Nacional de Loja

En la siguiente figura se muestran las carreras que se ofertan actualmente en la Universidad Nacional de Loja (ver figura 6).

	id	nombre	especialidad	modalidad_id_fk	titulacion_id_fk	area_id_fk	cc
118	248	Ninguna	TECNOLOGIA EN MECANICA INDUSTRIAL	presencial	TECNICO_TECNOLOGICO	AEIRNNR Ar...	0
119	249	Ninguna	TECNOLOGIA AGROPECUARIA	presencial	TECNICO_TECNOLOGICO	AEIRNNR Ar...	0
120	251	Ninguna	TECNOLOGIA EN TOPOGRAFIA	presencial	TECNICO_TECNOLOGICO	AEIRNNR Ar...	0
121	253	Ninguna	TECNOLOGIA AGRICOLA	presencial	TECNICO_TECNOLOGICO	AEIRNNR Ar...	0
122	254	Ninguna	Ingenieria en Acuacultura	presencial	PREGRADO	AARNR Area...	30
123	256	LIDERAZGO EDUCATIVO	PEDAGOGIA	presencial	PREGRADO	AEAC Area d...	30
124	258	Abogado	Derecho	presencial	POSTGRADO	AJSA Area J...	0
125	259	Ninguna	ESPECIALIZACION EN GESTION DE SISTEMAS DE RI...	presencial	POSTGRADO	AARNR Area...	0
126	261	Ninguna	ESPECIALIZACIoN EN AGRONEGOCIOS	presencial	POSTGRADO	AARNR Area...	0
127	264	Ninguna	PROFESIONALIZACION DOCENTE EN CIENCIAS DE L...	presencial	PREGRADO	AEAC Area d...	0
128	267	Ebanisteria, Tallado	Ebanisteria, Tallado	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
129	268	DIGITADOR, PROGRAMADOR EN COMPUTACI...	DIGITADOR, PROGRAMADOR EN COMPUTACION	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
130	269	Ebanisteria	Ebanisteria	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
131	270	Instalador electricista	TECNICAS ELECTRICAS	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
132	271	Mecanica Automotriz	Mecanica Automotriz	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
133	272	Tecnicas Electricas	Tecnicas Electricas	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
134	273	CONSTRUCCION CIVIL	TECNICAS CONSTRUCTIVAS	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
135	274	Ofimatica	Ofimatica	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0
136	275	mantenimiento preventivo y correctivo de equipo d...	mantenimiento preventivo y correctivo de equipo de audio...	presencial	TECNICA, ARTESANAL Y POP...	AEIRNNR Ar...	0

Figura 6: Carreras de la Universidad Nacional de Loja.

En la siguiente figura (ver figura 7) se muestran las modalidades que ofrece la Universidad Nacional de Loja para cada carrera.

	modalidad
1	distancia
2	presencial
3	semipresencial

Figura 7: Modalidades de estudio en la Universidad Nacional de Loja.

En la siguiente figura se observan los módulos que se han ofertado para cada una de las carreras de la Universidad Nacional de Loja hasta la actualidad (ver figura 8,9).

	id	nombre
445	666	9
446	671	modulo 1
447	672	modulo 2
448	673	modulo 1
449	674	modulo 2
450	675	modulo 1
451	676	modulo 2
452	677	Nivel 1
453	678	NIvel 2
454	679	Modulo UNO
455	681	MODULO DOS
456	682	Derecho (extension)
457	683	4
458	684	5
459	686	6
460	687	11

Figura 8: Módulos que se ofertan en la Universidad Nacional de Loja.

Paging | Start 12600 | Count 200 | OK | Loaded rows 95

	id	seccion	numero	nombre	modulo_id_fk	oferta_carrera_id_fk
12672	17867		1	DR.SARMIENTO 15-17 LU-M-V	598	1281
12673	17868		1	DRA.SARMIENTO17-19 MA-J-V	598	1281
12674	17869		1	MG.VIVANCO 17H-19 LU-MI-V	598	1281
12675	17870		1	MG.VIVANCO 19H-21 MA-J-V	598	1281
12676	17871		1	MG.GONZu00c1LEZ 17H-19LU-MI-V	598	1281
12677	17872		1	MG.GONZu00c1LEZ 19-21 MA-JU-V	598	1281
12678	17873		1	PROFESOR 1: 8-10 L-MI-V	598	1281
12679	17874		1	PROFESOR 1:7H-8 LU A V	598	1281
12680	17875		1	PROFESOR 1:12-13 LU A V	598	518
12681	17876		1	PROFESOR 1:12H-13 LU A V	598	1281
12682	17877		2	Ciclo 2 Paralelo 1	141	1272
12683	17878		7	EXTENSIu00d3N PARTICIPATIVA	29	1256
12684	17879		2	A	319	1263
12685	17880		2	B	319	1263
12686	17881		5	A	322	1263

Figura 9: Paralelos que se ofertan en la Universidad Nacional de Loja.

Se obtuvo información personal acerca de los 3438 estudiantes que pertenecen al Área de la Energía (ver figura 10).

Paging | Start 0 | Count 200 | OK | Loaded rows 200

	id	numeroIdentificacion	nombres	apellidos	fecha_nacimiento	telefono	celular	direccion
175	9683	0705493799	andres bruno	ochoa au00f1asco	1991-01-19 00:00:00	2909767	089214650	mariscal sucre y republica del ecuador
176	7208	0705497162	Jonathan Bladimir	Arrobo Ajila	1991-08-21 00:00:00	091229751	091229751	Huaquillas
177	9650	0705497576	eduardo luis	quezada romero	1988-08-27 00:00:00	072510933	0993290248	Esteban Godoy..(1ra Etapa))
178	3968	0705559508	leydi jacqueline	cevallos quezada	1991-10-09 00:00:00	072148545	085837329	Cdla. Brisas del Mar
179	2570	0705602316	Josu00e9 Andru00e9s	Moreno Moreno	1993-08-12 00:00:00	072909243	090044455	9 de octubre y portovelo
180	1270	0705620540	ruddy mauricio	conde lopez	1991-08-03 00:00:00	07 2964 232	080676064	Sucre y Colon
181	1955	0705620946	Lenin Alexander	Romero Betancourt	1993-01-08 00:00:00	2964225	0990221167	Balsalito
182	398	0705624062	juan eduardo	granda granda	(NULL)	(NULL)	(NULL)	(NULL)
183	2556	0705624278	Marlon Victor	Salvatierra Celi	1993-10-28 00:00:00	2953-387	087369471	Av.occidental
184	10415	0705634590	Juan Pablo	Romero Armijos	1995-04-29 00:00:00	0997577348	0997577348	Quinara 23-100 y Atahualpa
185	2244	0705639292	Richard de Jesus	Matamoros Benavides	1992-10-24 00:00:00	072958451	080240273	Loja
186	2607	0705642734	jose miguel	Sanchez Matamoros	(NULL)	(NULL)	(NULL)	(NULL)
187	10220	0705649796	Jenny Judith	Luzuriaga Jimenez	1993-02-07 00:00:00	0991834855	0991834855	Urbanizacion Vizcaya
188	10386	0705650448	Marjorie Maribel	Romero Serrano	(NULL)	(NULL)	(NULL)	(NULL)

Figura 10: Datos de estudiantes del AEIRNNR de la Universidad Nacional de Loja.

En la siguiente figura (ver figura 11) se muestra información almacenada acerca de estudiantes matriculados en la Universidad Nacional de Loja.

Data

Paging | Start 200 | Count 200 | OK | Loaded rows 200

	num_matriculado	area_id_fk	oferta_id_fk
380	4260	ACE Cursos Especiales	36 Pregrado Septiembre 2010 Febrero 2011
381	4265	PREUNIVERSITARIO PREUNIVERSITARIO	4 Preuniversitario Agosto 2008
382	4331	MED Modalidad de Estudios a Distancia	65 Pregrado Marzo 2012 Julio 2012
383	4598	AJSA Area Juru00eddica, Social y Administrativa	65 Pregrado Marzo 2012 Julio 2012
384	4786	AJSA Area Juru00eddica, Social y Administrativa	45 Pregrado Septiembre 2011 Febrero 2012
385	4994	AJSA Area Juru00eddica, Social y Administrativa	42 Pregrado Marzo 2011 - Julio 2011
386	5039	MED Modalidad de Estudios a Distancia	42 Pregrado Marzo 2011 - Julio 2011
387	5072	MED Modalidad de Estudios a Distancia	45 Pregrado Septiembre 2011 Febrero 2012
388	5154	AJSA Area Juru00eddica, Social y Administrativa	36 Pregrado Septiembre 2010 Febrero 2011
389	5384	AJSA Area Juru00eddica, Social y Administrativa	34 Pregrado Marzo 2010 - Julio 2010
390	5385	AJSA Area Juru00eddica, Social y Administrativa	28 Pregrado Septiembre 2009 - Febrero 2010
391	5467	MED Modalidad de Estudios a Distancia	36 Pregrado Septiembre 2010 Febrero 2011
392	5514	AJSA Area Juru00eddica, Social y Administrativa	26 Pregrado Marzo 2009 - Julio 2009

Figura 11: Estudiantes matriculados en cada área de la Universidad Nacional de Loja.

En la siguiente figura (ver figura 12) se muestra información almacenada acerca de estudiantes aprobados en varias Áreas en la Universidad Nacional de Loja.

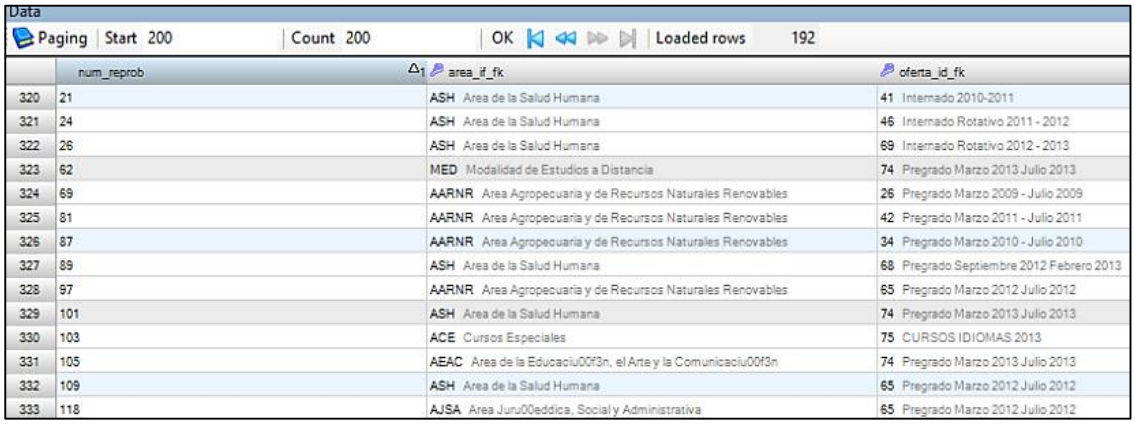
Data

Paging | Start 200 | Count 200 | OK | Loaded rows 192

	numero_aprob	area_id_fk	oferta_id_fk
357	1641	ASH Area de la Salud Humana	65 Pregrado Marzo 2012 Julio 2012
358	1671	ASH Area de la Salud Humana	25 Pregrado Septiembre 2008 - Febrero 2009
359	1811	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	74 Pregrado Marzo 2013 Julio 2013
360	2316	ACE Cursos Especiales	28 Pregrado Septiembre 2009 - Febrero 2010
361	2349	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	26 Pregrado Marzo 2009 - Julio 2009
362	2412	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	65 Pregrado Marzo 2012 Julio 2012
363	2476	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	45 Pregrado Septiembre 2011 Febrero 2012
364	2500	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	25 Pregrado Septiembre 2008 - Febrero 2009
365	2517	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	28 Pregrado Septiembre 2009 - Febrero 2010
366	2594	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	42 Pregrado Marzo 2011 - Julio 2011
367	2687	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	36 Pregrado Septiembre 2010 Febrero 2011
368	2692	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	34 Pregrado Marzo 2010 - Julio 2010
369	2947	MED Modalidad de Estudios a Distancia	74 Pregrado Marzo 2013 Julio 2013
370	3015	ACE Cursos Especiales	34 Pregrado Marzo 2010 - Julio 2010

Figura 12: Estudiantes aprobados en cada área de la Universidad Nacional de Loja.

En la siguiente figura (ver figura 13) se muestra información estadística acerca de estudiantes reprobados por cada área de la Universidad Nacional de Loja.



Data

Paging | Start 200 | Count 200 | OK | Loaded rows 192

	num_reprob	area_if_fk	oferta_id_fk
320	21	ASH Area de la Salud Humana	41 Internado 2010-2011
321	24	ASH Area de la Salud Humana	46 Internado Rotativo 2011 - 2012
322	26	ASH Area de la Salud Humana	69 Internado Rotativo 2012 - 2013
323	62	MED Modalidad de Estudios a Distancia	74 Pregrado Marzo 2013 Julio 2013
324	69	AARNR Area Agropecuaria y de Recursos Naturales Renovables	26 Pregrado Marzo 2009 - Julio 2009
325	81	AARNR Area Agropecuaria y de Recursos Naturales Renovables	42 Pregrado Marzo 2011 - Julio 2011
326	87	AARNR Area Agropecuaria y de Recursos Naturales Renovables	34 Pregrado Marzo 2010 - Julio 2010
327	89	ASH Area de la Salud Humana	68 Pregrado Septiembre 2012 Febrero 2013
328	97	AARNR Area Agropecuaria y de Recursos Naturales Renovables	65 Pregrado Marzo 2012 Julio 2012
329	101	ASH Area de la Salud Humana	74 Pregrado Marzo 2013 Julio 2013
330	103	ACE Cursos Especiales	75 CURSOS IDIOMAS 2013
331	105	AEAC Area de la Educaciu00f3n, el Arte y la Comunicaciu00f3n	74 Pregrado Marzo 2013 Julio 2013
332	109	ASH Area de la Salud Humana	65 Pregrado Marzo 2012 Julio 2012
333	118	AJSA Area Juru00eddica, Social y Administrativa	65 Pregrado Marzo 2012 Julio 2012

Figura 13: Estudiantes reprobados en cada área de la Universidad Nacional de Loja.

En las siguientes tablas (ver tabla I, II, III, IV, V, VI, VII) se describe información acerca de los estudiantes matriculados, aprobados y reprobados en cada área de la Universidad Nacional de Loja.

TABLA I:

ESTUDIANTES MATRICULADOS, APROBADOS Y REPROBADOS DEL ÁREA ACE.

ACE	Estudiantes			% de bajas
Periodo	Matriculados	Aprobados	Reprobados	
Septiembre 2009 – Febrero 2010	3396	2316	1011	29.77
Marzo 2010 – Julio 2010	3906	3015	730	18.68
Septiembre 2010 – Febrero 2011	4260	3128	934	21.92
Marzo 2011 – Julio 2011	1775	1384	382	21.52
Septiembre 2011 – Febrero 2012	1986	1472	471	23.71
Marzo 2012 – Julio 2012	1282	1123	139	10.84
Septiembre 2012 – Febrero 2013	1440	1067	343	23.81
Cursos Especiales 2010-2011	9999	5910	3583	35.86
Cursos Especiales 2011-2012	7380	4337	2665	36.11
Cursos Especiales 2012-2013	12959	3916	1894	14.61

TABLA II:

ESTUDIANTES MATRICULADOS, APROBADOS Y REPROBADOS DEL ÁREA AARNR.

AARNR	Estudiantes			% de bajas
Periodo	Matriculados	Aprobados	Reprobados	
Septiembre 2008 – Febrero 2009	707	557	142	20.08
Marzo 2009 – Julio 2009	600	527	69	11.5

Septiembre 2009 – Febrero 2010	774	593	178	22.99
Marzo 2010 – Julio 2010	679	578	87	12.81
Septiembre 2010 – Febrero 2011	804	597	182	22.63
Marzo 2011 – Julio 2011	651	561	81	12.44
Septiembre 2011 – Febrero 2012	773	550	191	24.70
Marzo 2012 – Julio 2012	625	510	97	15.52
Septiembre 2012 – Febrero 2013	471	434	19	4.03
Marzo 2013 – Julio 2013	679	535	142	20.91

TABLA III:
ESTUDIANTES MATRICULADOS, APROBADOS
Y REPROBADOS DEL ÁREA AEAC.

AEAC	Estudiantes			% de
Periodo	Matriculados	Aprobados	Reprobados	bajas
Septiembre 2008 – Febrero 2009	2767	2500	266	9.61
Marzo 2009 – Julio 2009	2563	2349	209	8.15
Septiembre 2009 – Febrero 2010	2735	2517	205	7.49
Marzo 2010 – Julio 2010	2874	2692	160	5.56
Septiembre 2010 – Febrero 2011	2901	2687	195	6.72
Marzo 2011 – Julio 2011	2734	2594	132	4.82
Septiembre 2011 – Febrero 2012	2673	2476	178	6.65
Marzo 2012 – Julio 2012	2557	2412	129	5.04
Septiembre 2012 – Febrero 2013	1791	1633	155	8.65
Marzo 2013 – Julio 2013	1917	1811	105	5.47

TABLA IV:
ESTUDIANTES MATRICULADOS, APROBADOS
Y REPROBADOS DEL ÁREA AEIRNNR.

AEIRNNR	Estudiantes			% de
Periodo	Matriculados	Aprobados	Reprobados	bajas
Septiembre 2008 – Febrero 2009	1271	891	380	29.89
Marzo 2009 – Julio 2009	921	729	192	20.84
Septiembre 2009 – Febrero 2010	1210	902	308	25.45

Marzo 2010 – Julio 2010	1052	843	207	19.67
Septiembre 2010 – Febrero 2011	1137	918	215	18.90
Marzo 2011 – Julio 2011	1013	812	195	19.24
Septiembre 2011 – Febrero 2012	1046	790	231	22.08
Marzo 2012 – Julio 2012	893	679	175	19.59
Septiembre 2012 – Febrero 2013	693	493	200	28.86
Marzo 2013 – Julio 2013	770	549	221	28.70

TABLA V:
ESTUDIANTES MATRICULADOS, APROBADOS
Y REPROBADOS DEL ÁREA AJSA.

AJSA	Estudiantes			% de bajas
Periodo	Matriculados	Aprobados	Reprobados	
Septiembre 2008 – Febrero 2009	5700	5298	396	6.94
Marzo 2009 – Julio 2009	5514	5188	320	5.80
Septiembre 2009 – Febrero 2010	5385	5029	346	6.42
Marzo 2010 – Julio 2010	5384	5117	241	4.47
Septiembre 2010 – Febrero 2011	5154	4847	266	5.16
Marzo 2011 – Julio 2011	4994	4804	180	3.60
Septiembre 2011 – Febrero 2012	4786	4555	202	4.22
Marzo 2012 – Julio 2012	4598	4454	118	2.56
Septiembre 2012 – Febrero 2013	3531	3363	168	4.75
Marzo 2013 – Julio 2013	3999	3833	162	4.05

TABLA VI:
ESTUDIANTES MATRICULADOS, APROBADOS
Y REPROBADOS DEL ÁREA ASH.

ASH	Estudiantes			% de bajas
Periodo	Matriculados	Aprobados	Reprobados	
Septiembre 2008 – Febrero 2009	2172	1671	494	22.74
Marzo 2009 – Julio 2009	2129	1219	853	40.06
Septiembre 2009 – Febrero 2010	1743	1375	358	20.53
Marzo 2010 – Julio 2010	1936	1503	419	21.64
Septiembre 2010 – Febrero 2011	1732	1571	153	8.83

Marzo 2011 – Julio 2011	1884	1620	264	14.01
Septiembre 2011 – Febrero 2012	1701	1550	138	8.11
Marzo 2012 – Julio 2012	1778	1641	109	6.13
Septiembre 2012 – Febrero 2013	1348	1259	89	6.60
Marzo 2013 – Julio 2013	1580	1449	101	6.39
Internado 2010-2011	126	104	21	16.66
Internado 2011-2012	225	201	24	10.66
Internado 2012-2013	219	193	26	11.87

TABLA VII:
ESTUDIANTES MATRICULADOS, APROBADOS
Y REPROBADOS DEL ÁREA MED.

MED	Estudiantes			% de bajas
Periodo	Matriculados	Aprobados	Reprobados	
Septiembre 2009 – Febrero 2010	6308	5833	473	7.49
Marzo 2010 – Julio 2010	5988	5670	318	5.31
Septiembre 2010 – Febrero 2011	5467	5135	332	6.07
Marzo 2011 – Julio 2011	5039	4753	286	5.67
Septiembre 2011 – Febrero 2012	5072	4741	331	6.52
Marzo 2012 – Julio 2012	4331	4099	232	5.35
Septiembre 2012 – Febrero 2013	3581	3408	173	4.83
Marzo 2013 – Julio 2013	3048	2947	62	2.03

ANEXO 3: Evaluación de características de Herramientas para Administración de Base de Datos.

Análisis de características de Database Workbench 4.4.1. En este apartado se realiza una evaluación de las características más importantes de la herramienta como tipos de formatos que se pueden exportar (ver figura 1).

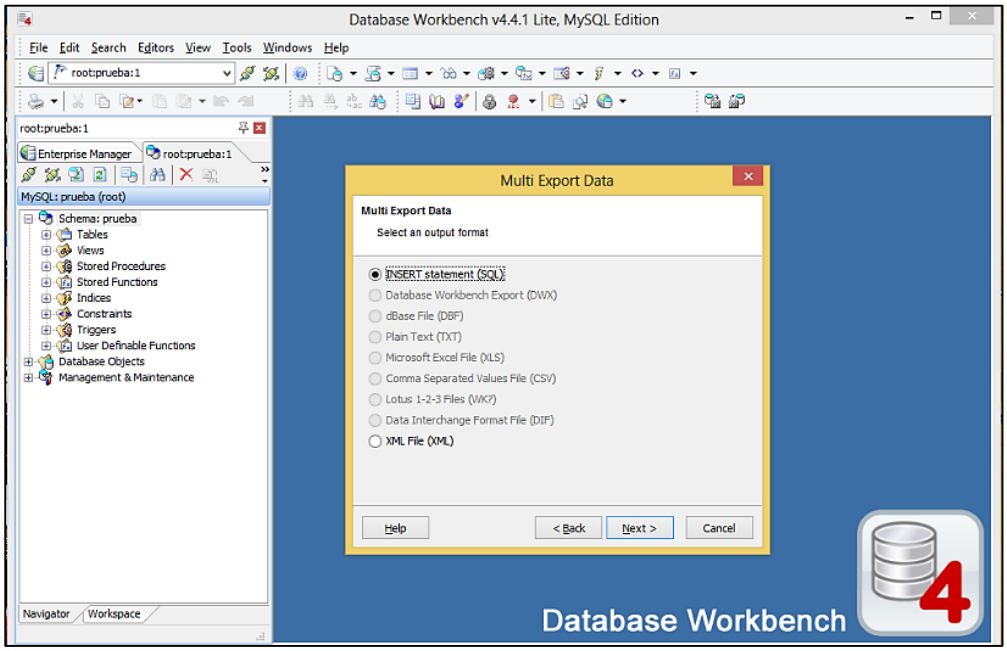


Figura 1: Exportar archivos desde Database Workbench.

A continuación se analiza la posibilidad de importar en esta herramienta (ver figura 2).

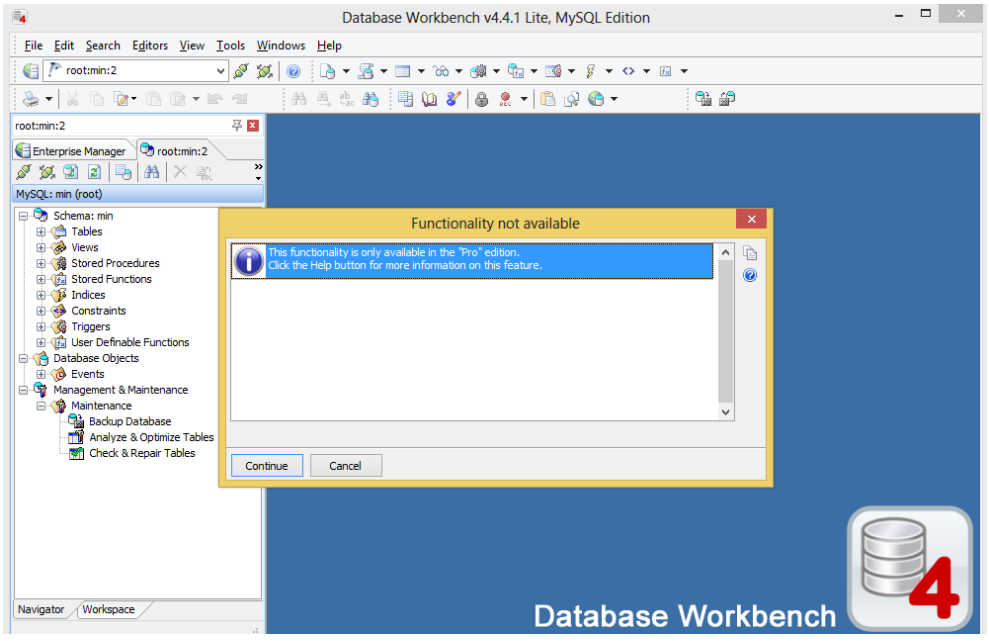


Figura 2: Importar archivos hacia Database Workbench.

Análisis de características de DataAdmin 5.4.2.5 Personal. En este apartado se realiza una evaluación de las características más importantes de la herramienta como tipos de formatos que se pueden exportar (ver figura 3).

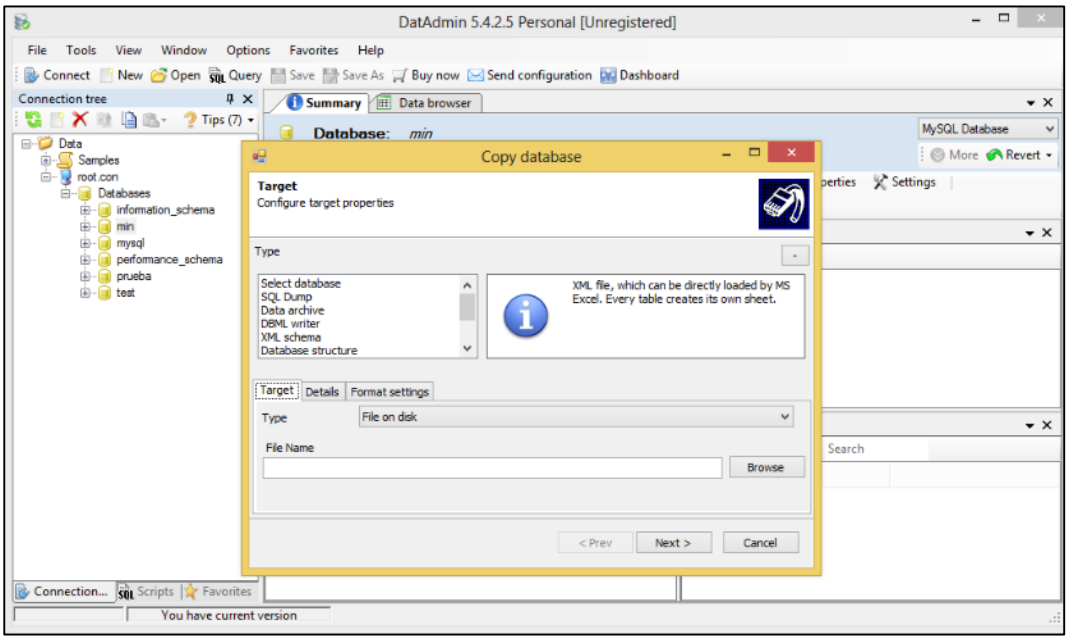


Figura 3: Exportar archivos desde DataAdmin.

A continuación se analiza la posibilidad de importar información hacia esta herramienta (ver figura 4).

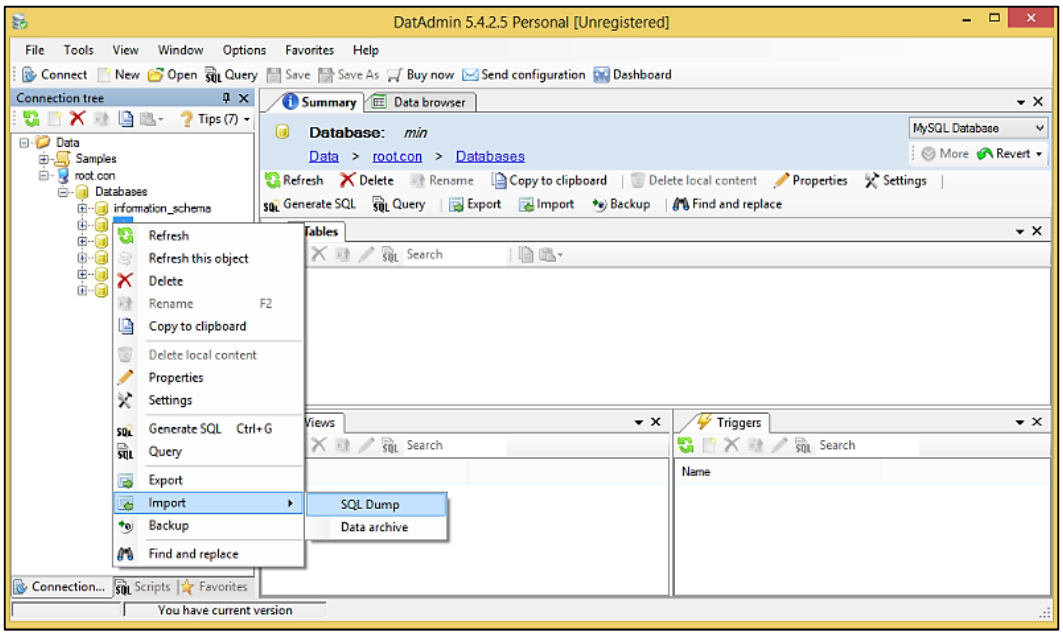


Figura 4: Importar archivos hacia DataAdmin.

Análisis de características de My SQL Workbench 6.0.8. En este apartado se realiza una evaluación de las características más importantes de la herramienta como tipos de formatos que se pueden exportar (ver figura 5).

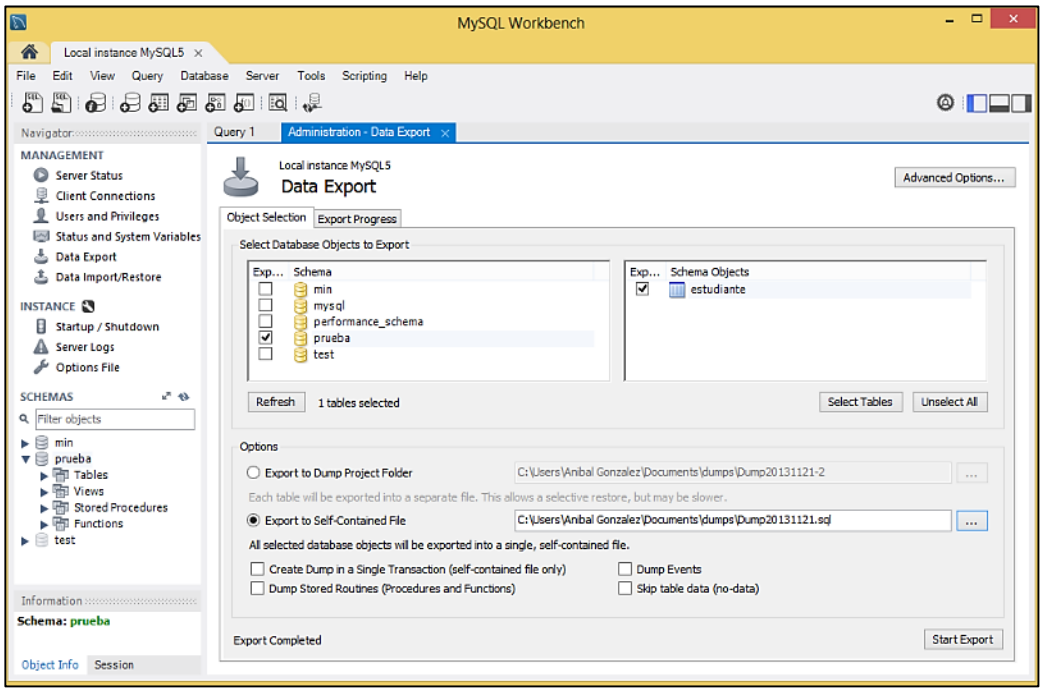


Figura 5: Exportar archivos desde My SQL Workbench.

A continuación se analiza las formas de importar información hacia esta herramienta (ver figura 6).

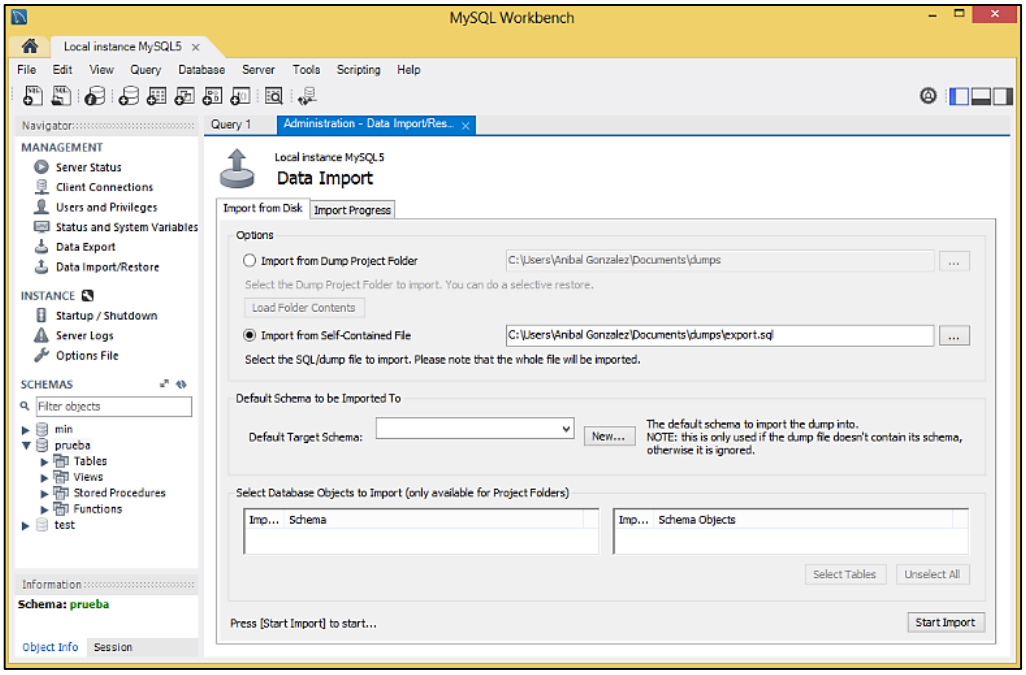


Figura 6: Importar archivos hacia My SQL Workbench.

Análisis de características de SQLyog Enterprise. En este apartado se realiza una evaluación de las características más importantes de la herramienta como tipos de formatos que se pueden exportar (ver figura 7).

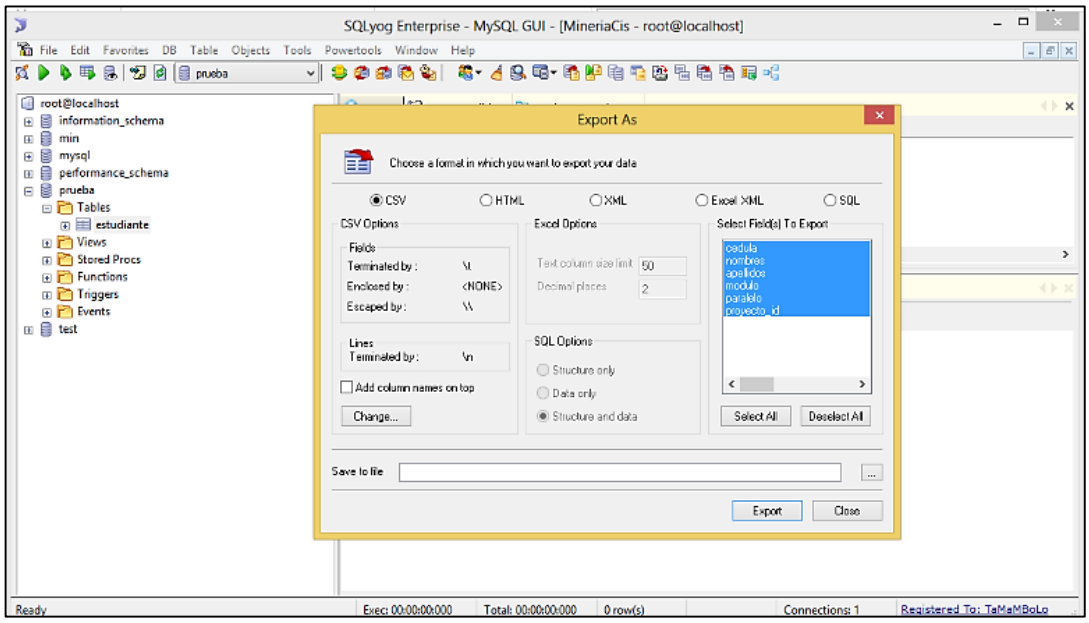


Figura 7: Exportar archivos desde SQLyog Enterprise.

A continuación se analiza las formas de importar información hacia esta herramienta (ver figura 8, 9).

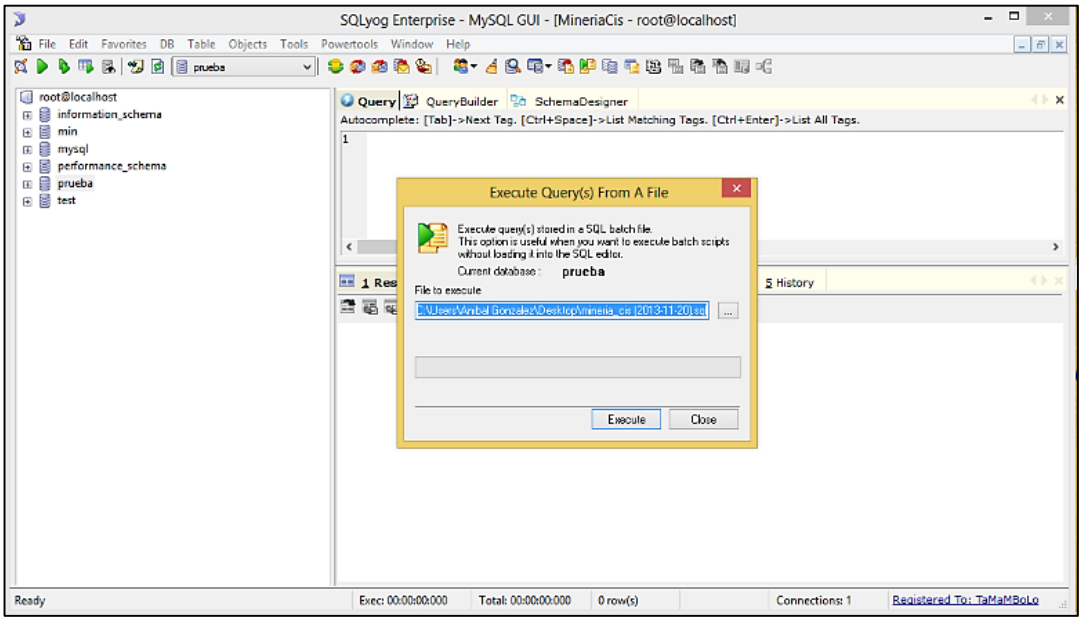


Figura 8: Importar archivos .sql hacia SQLyog Enterprise.

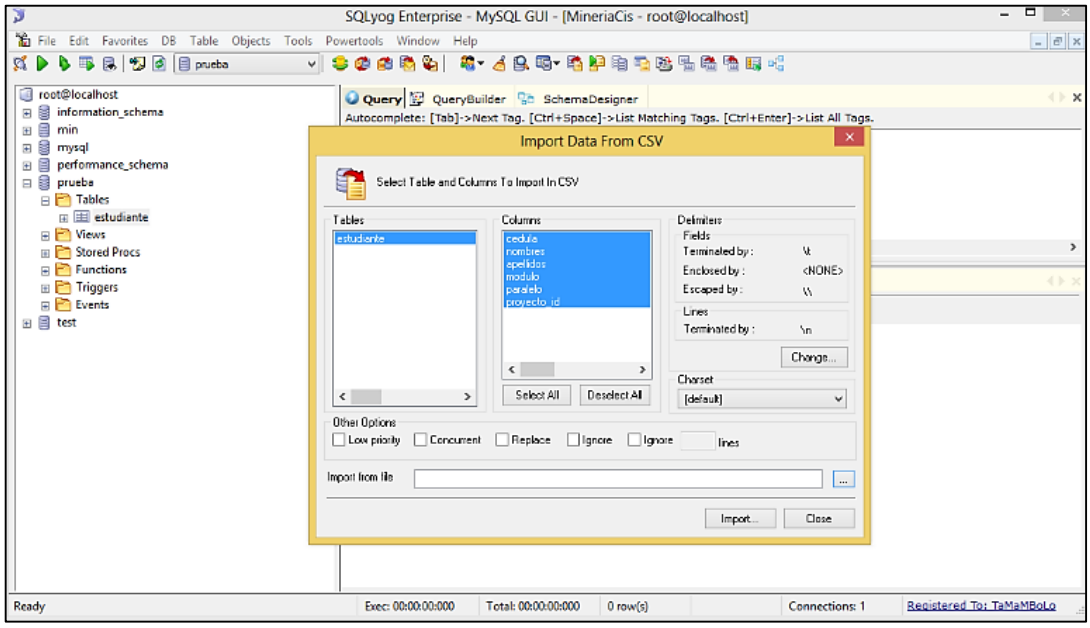


Figura 9: Importar archivos .csv hacia SQLyog Enterprise.

ANEXO 4: Evaluación de característica de Herramientas como apoyo para el proceso de Minería de Datos.

En este apartado de características de la herramienta Orange, se realizó una evaluación de las características más importantes como manejo de procesos de minería y visualización de resultados (ver figura 1, 2, 3, 4).

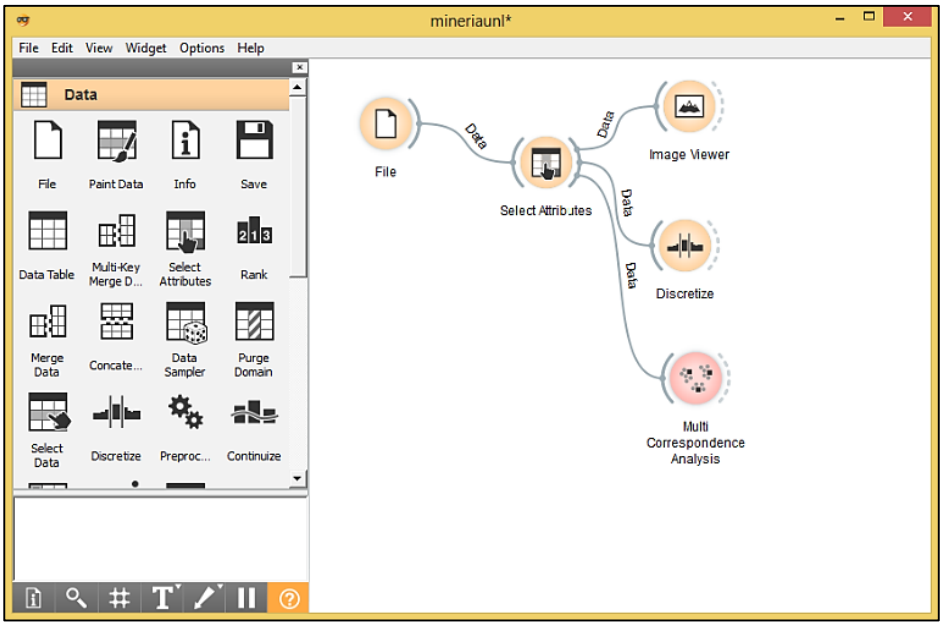


Figura 1: Procesos visuales en Orange.

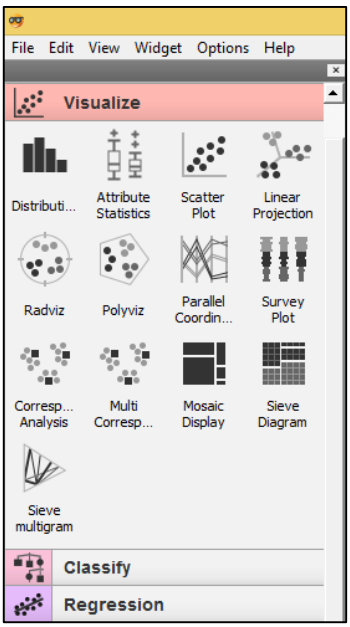


Figura 2: Visualizadores de datos en Orange.

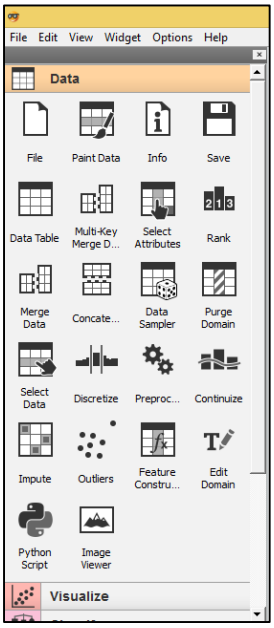


Figura 3: Análisis de datos

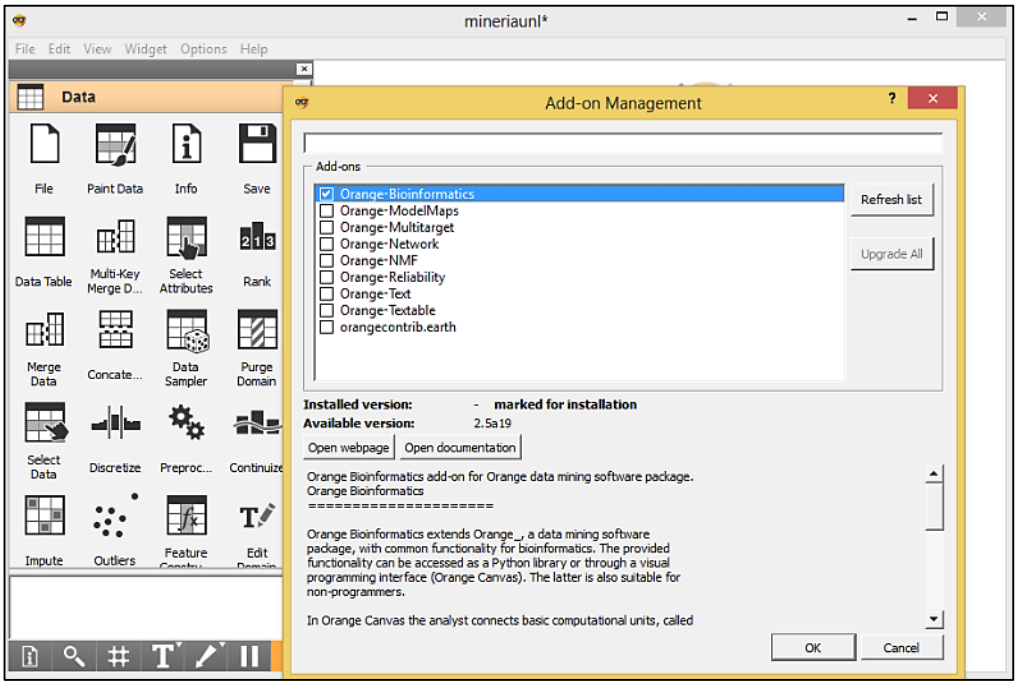


Figura 4: Añadir componentes en Orange.

En este apartado de características de la herramienta RapidMiner, se realizó una evaluación de las características más importantes como manejo de procesos de minería y visualización de resultados (ver figura 5, 6, 7).

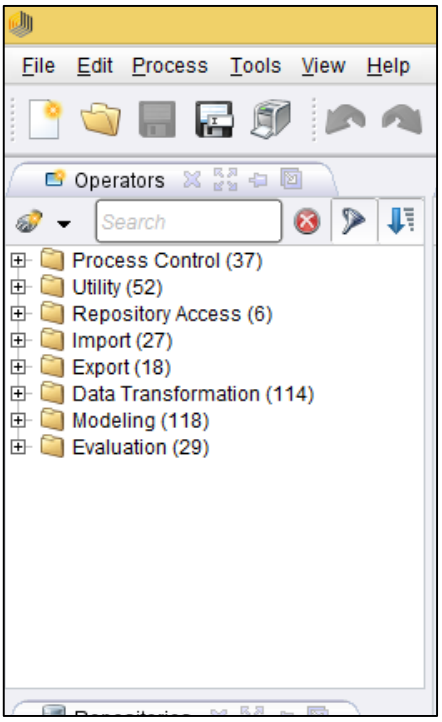


Figura 5: Operadores de Rapid Miner.

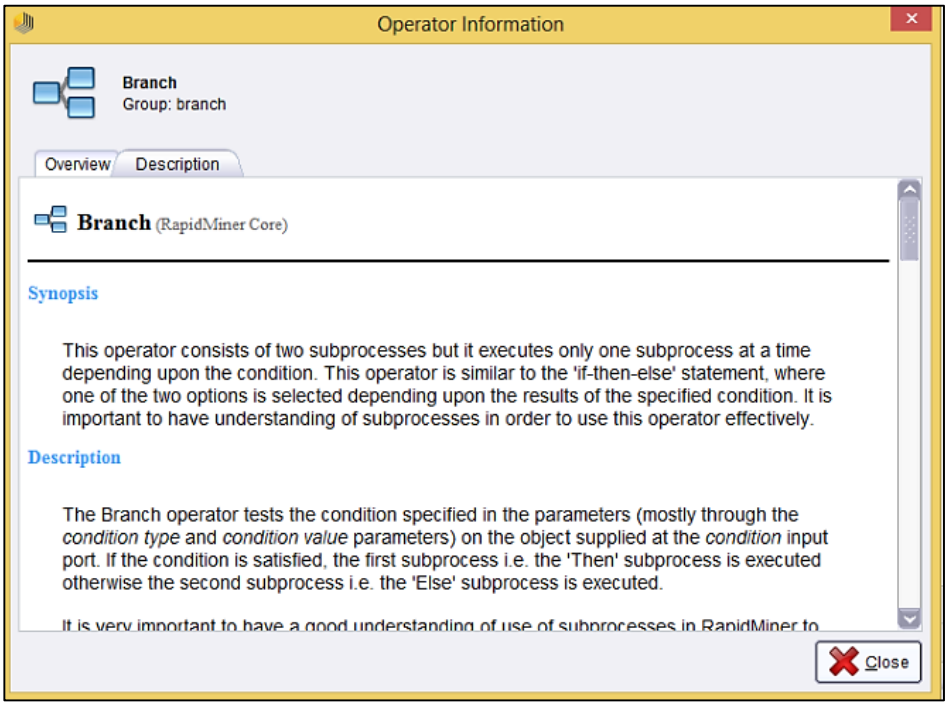


Figura 6: Descripción de Operadores de Rapid Miner.

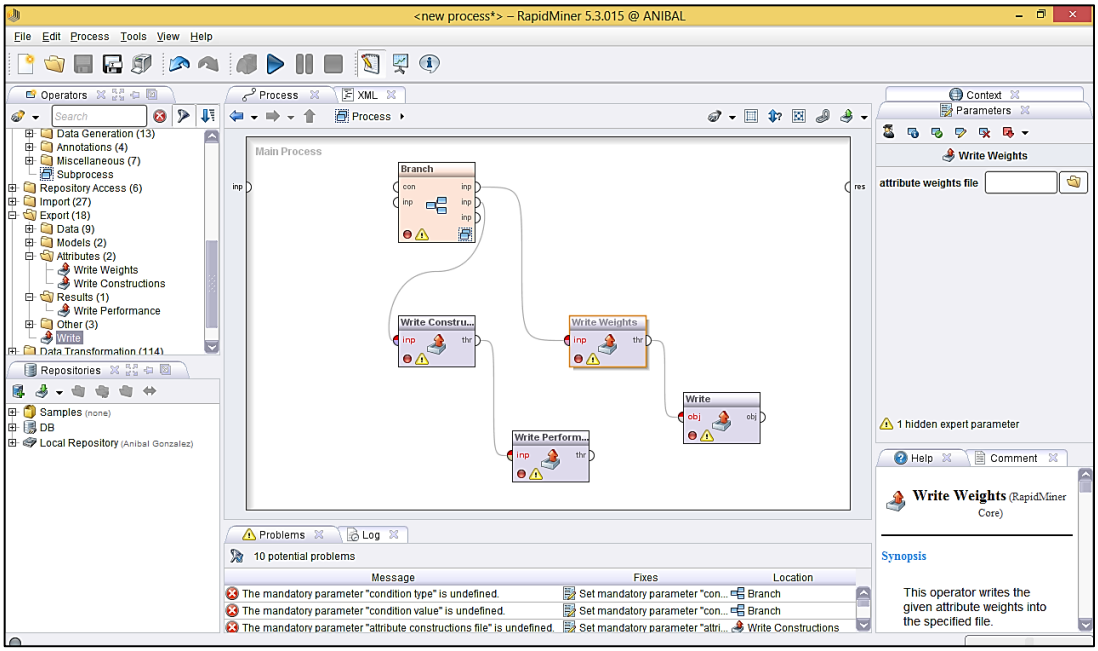


Figura 7: Jerarquización de Operadores de Rapid Miner.

En este apartado de características de la herramienta Weka, se realizó una evaluación de las características más importantes como manejo de procesos de minería y visualización de resultados (ver figura 8, 9, 10).

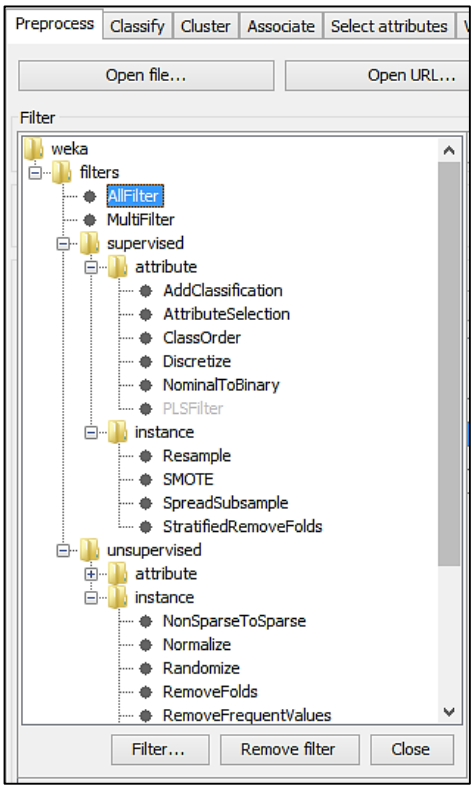


Figura 8: Filtros de Pre procesamiento

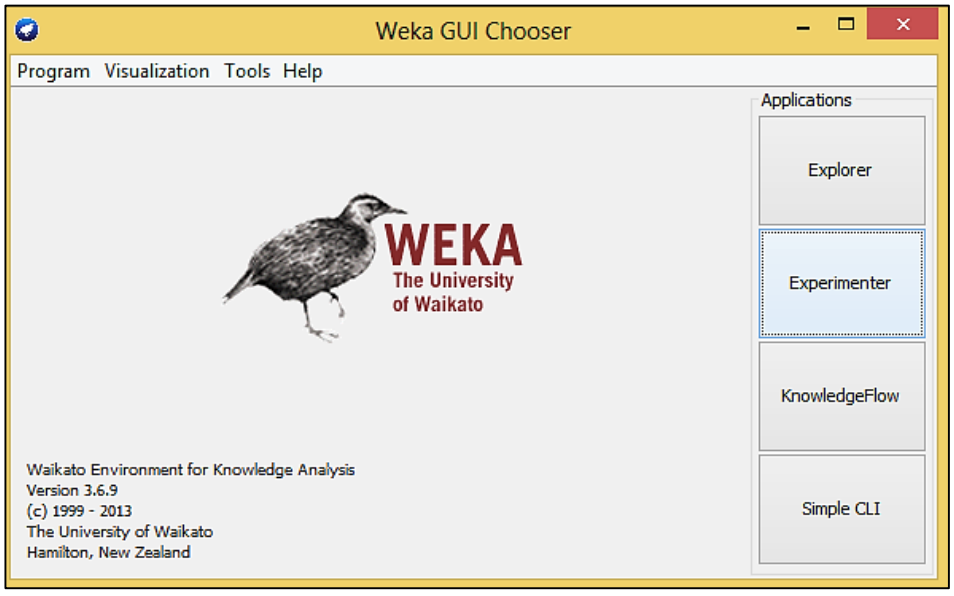


Figura 9: Interfaz Gráfica de Weka.

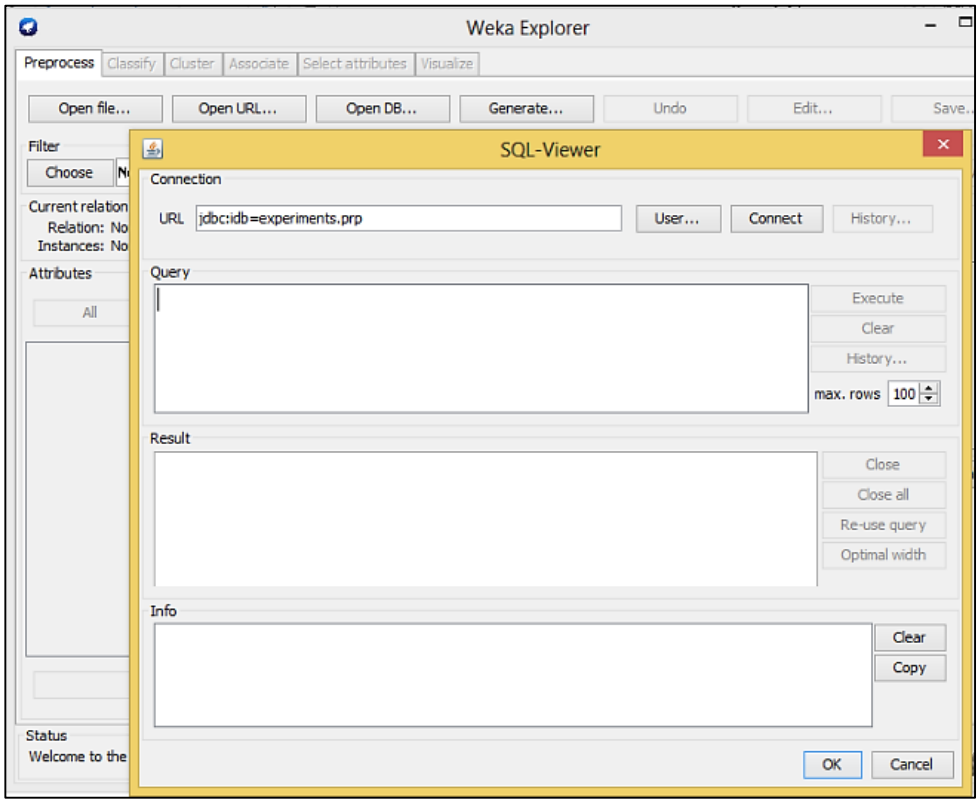


Figura 10: Conexión JDBC en Weka.

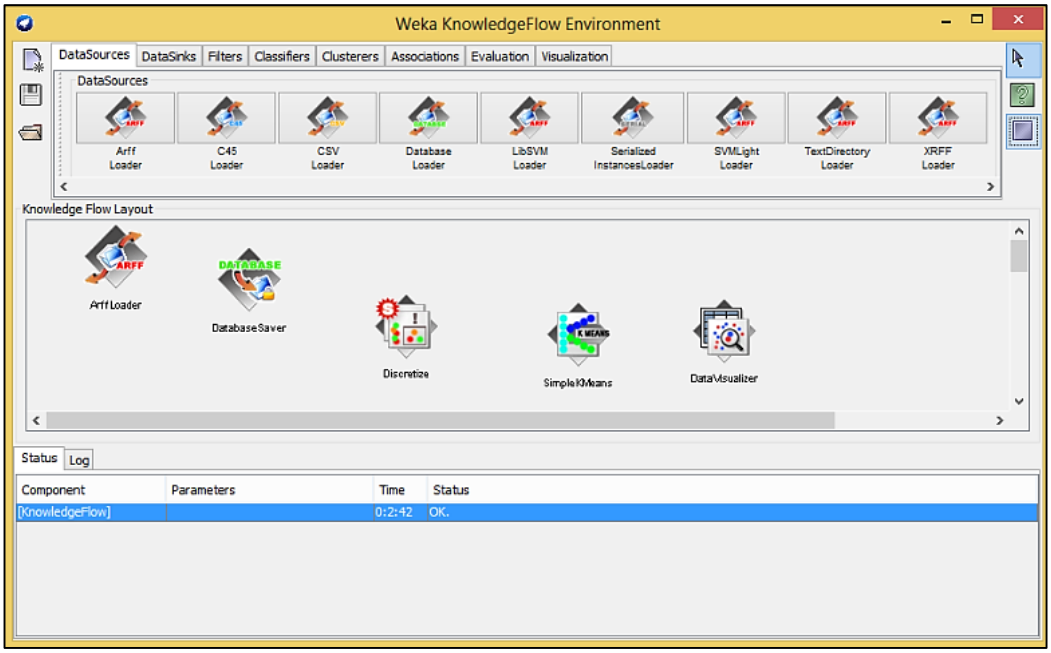


Figura 11: Escenario flujos de datos en Weka.

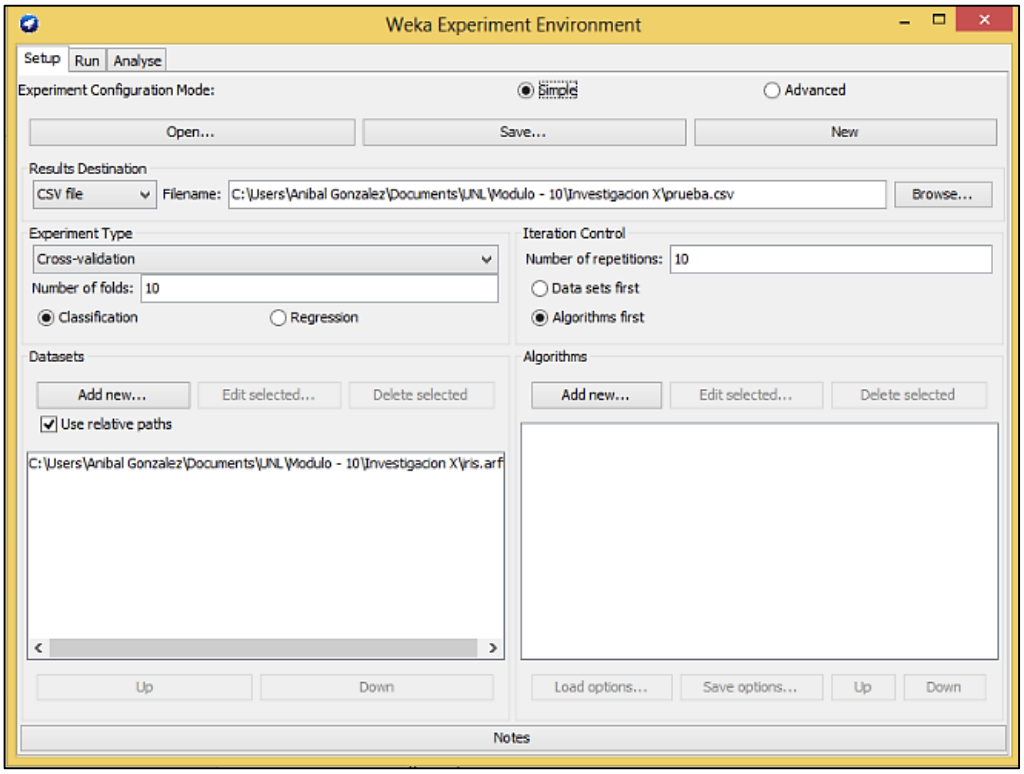


Figura 12: Escenario Experimentes en Weka.

ANEXO 5: Análisis de Herramientas apoyo para el proceso de Minería de Datos con datos de Prueba.

Se realizó pruebas con las herramientas de minería de datos para ello se tomó como muestra un dataset libre de usar la misma que se encuentra en el siguiente repositorio (ver figura 1): http://mlr.cs.umass.edu/ml/datasets/Iris, en el cual se describen las características de 3 tipos de flores.

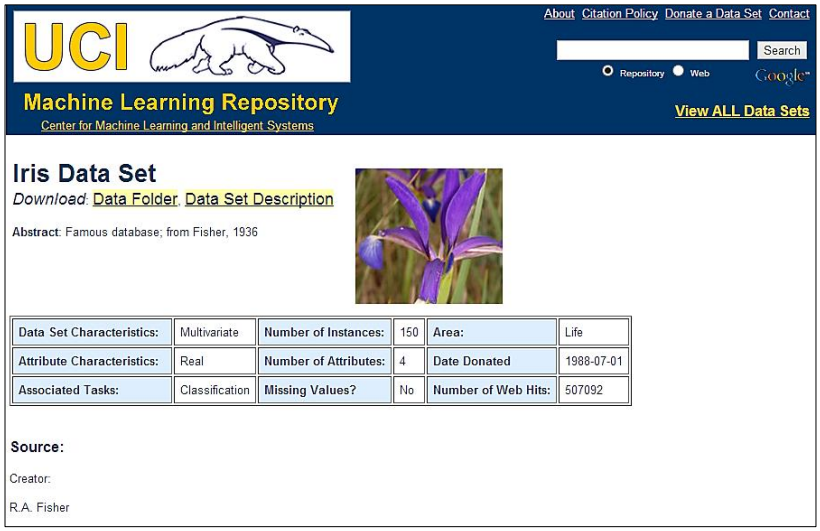


Figura 1: Repositorio de Base de datos para pruebas.

El conjunto de datos consta de 3 clases de 50 casos cada uno, donde cada clase se refiere a un tipo de planta de iris. Una clase es linealmente diferente de las otras 2, estos últimos no son linealmente diferentes entre sí. Los atributos que se encuentran en este conjunto de datos son los siguientes (ver tabla VIII):

TABLA VIII:
DESCRIPCIÓN DE ATRIBUTOS DE LOS DATOS DE PRUEBA.

Nro.	Atributo	Medida	Descripción
1	sepal length	Cm	Longitud del sépalo
2	sepal width	Cm	Ancho de sépalo
3	petal length	Cm	Longitud de pétalo
4	petal width	Cm	Ancho de pétalo
5	class: -- Iris Setosa -- Iris Versicolour -- Iris Virginica		Puede ser una clase de flor

Resultados en Rapid Miner

A continuación (figura 2, 3, 4, 5, 6) se muestra los resultados obtenidos al evaluar las herramientas con los datos.

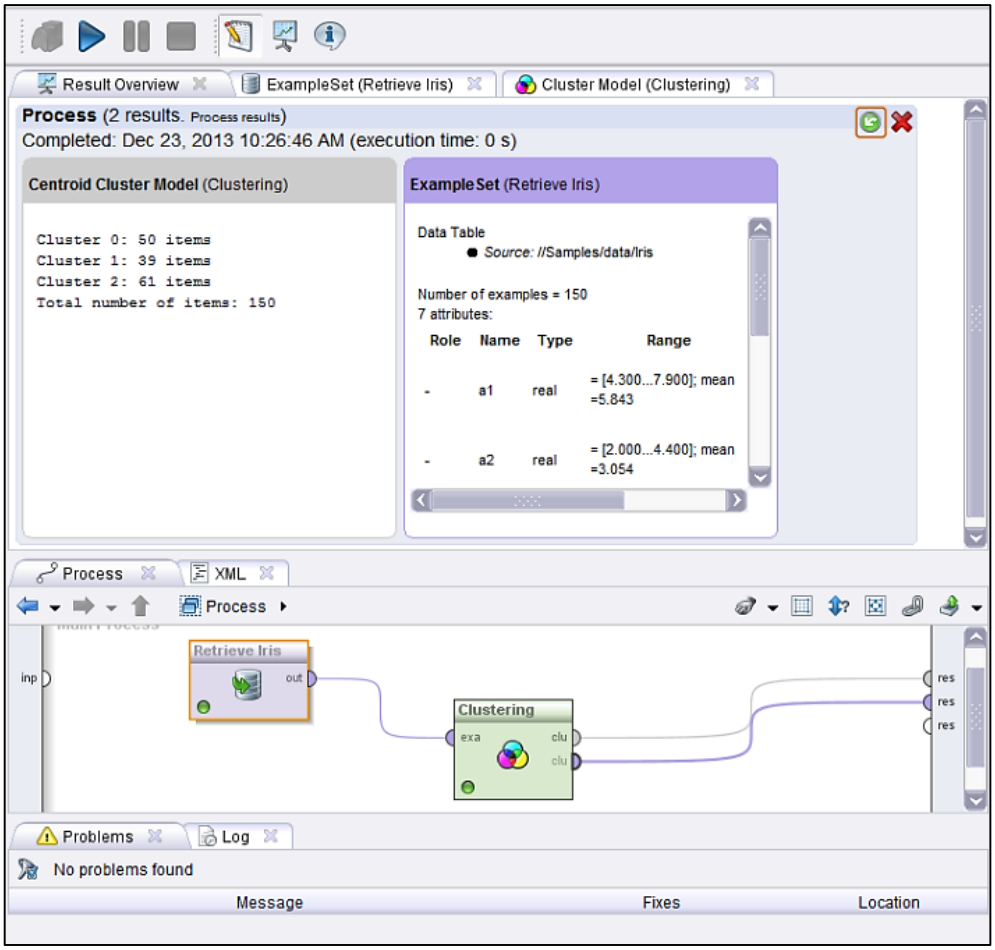


Figura 2: Resultados de pruebas en Rapid Miner.

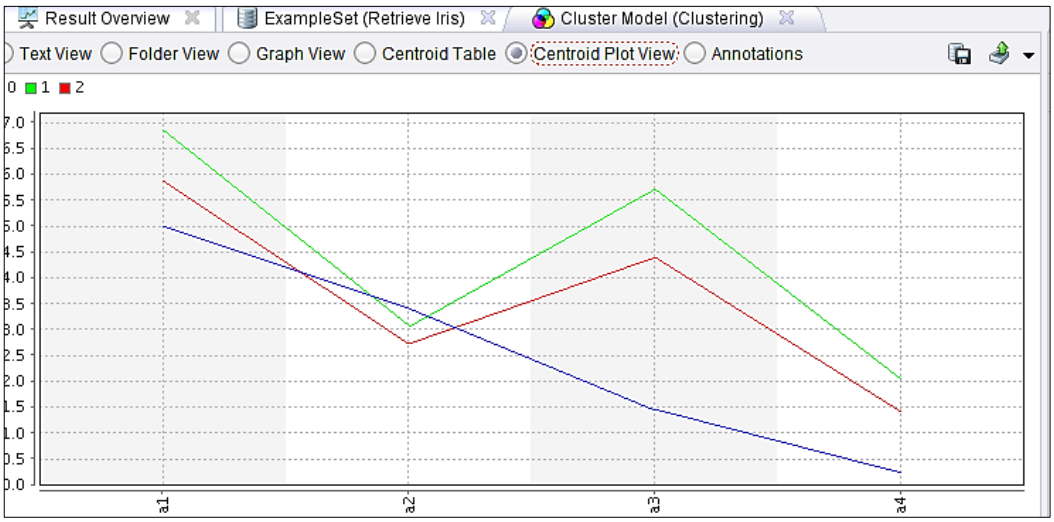


Figura 3: Grafica de resultados de pruebas en Rapid Miner

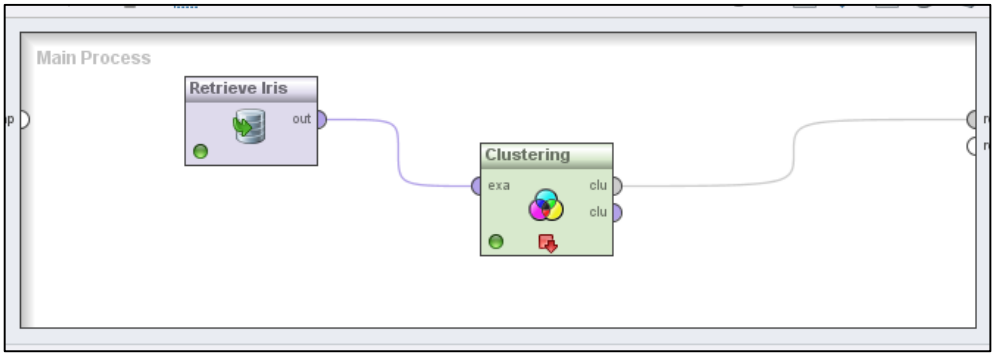


Figura 4: Estructura de Procesos en Rapid Miner.

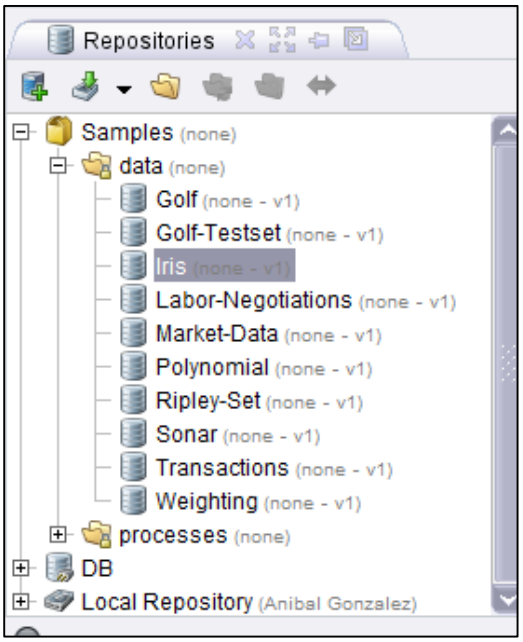


Figura 5: Repositorio de datos en Rapid Miner.

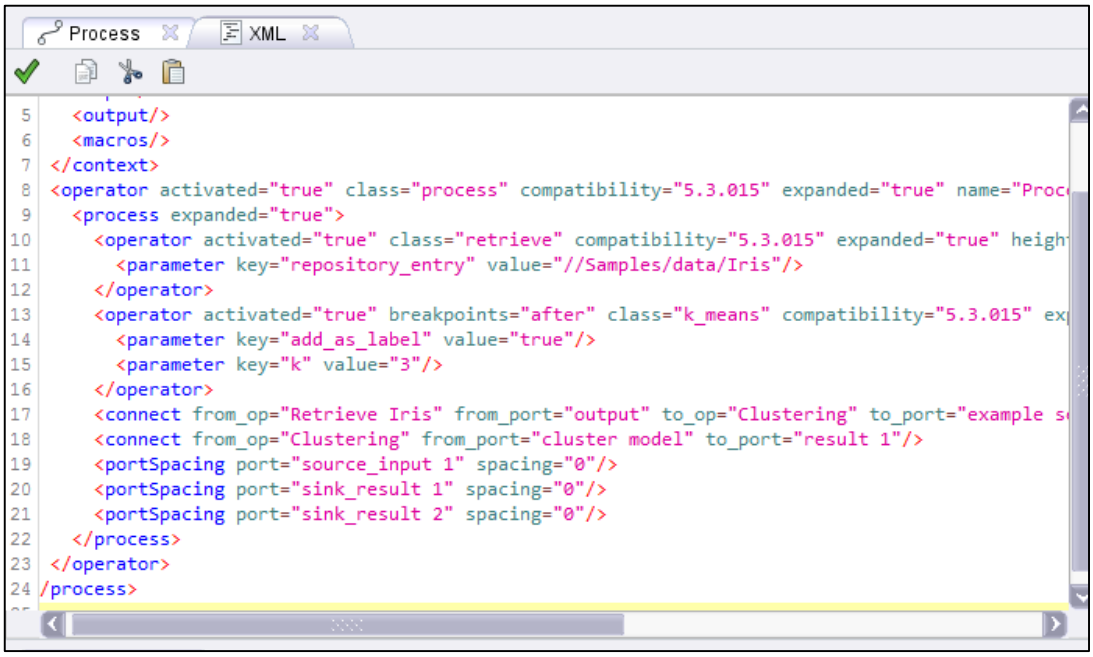


Figura 6: Estructura de los procesos en XML.

Resultados en Weka

En la siguiente figura (ver figura 7 ,8) se muestra los resultados obtenidos a partir de la aplicación del algoritmo de clasificación K-Means.

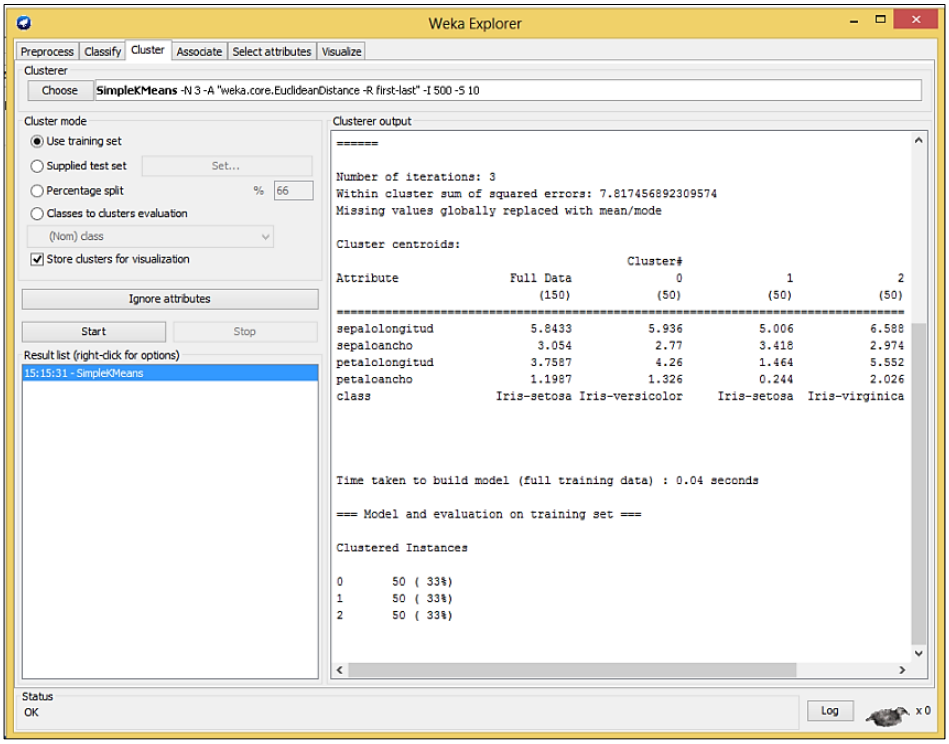


Figura 7: Resultados de aplicar algoritmo K-Means.

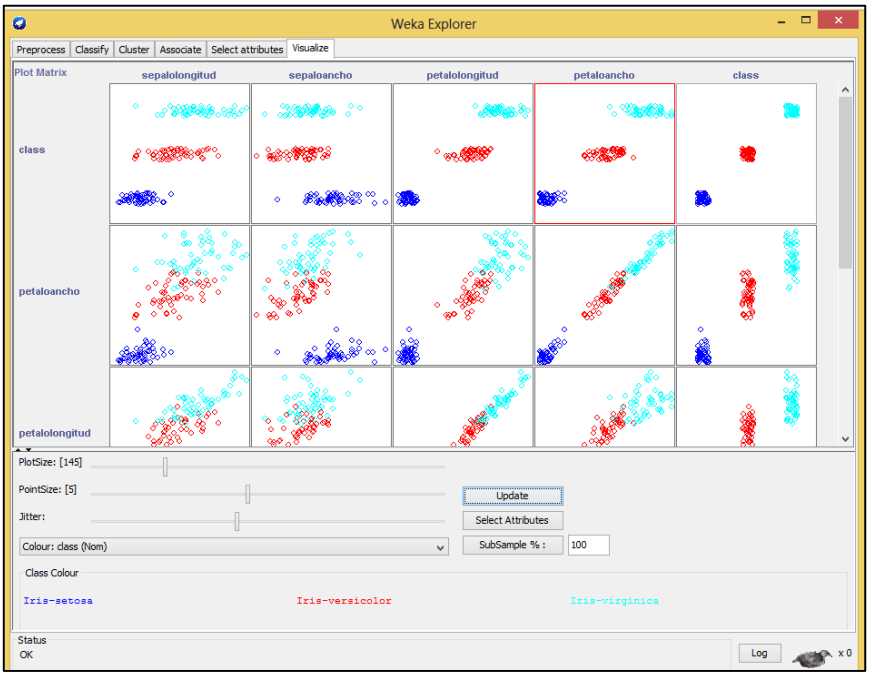


Figura 8: Visualización de Resultados.

Resultados en Orange

A continuación se muestran los resultados obtenidos a partir de pruebas realizadas en la Herramienta Orange (ver figura 9, 10, 11).

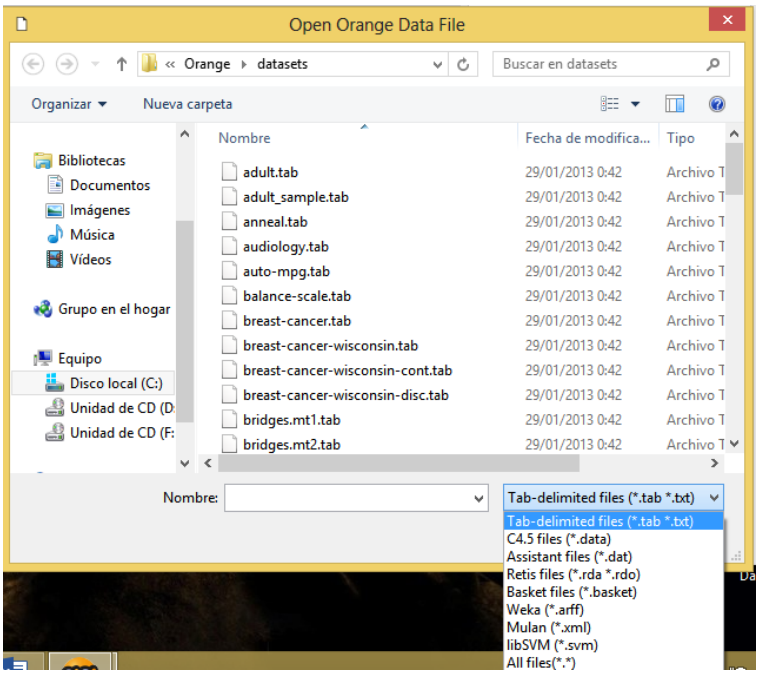


Figura 9: Formatos para importar datos a Orange

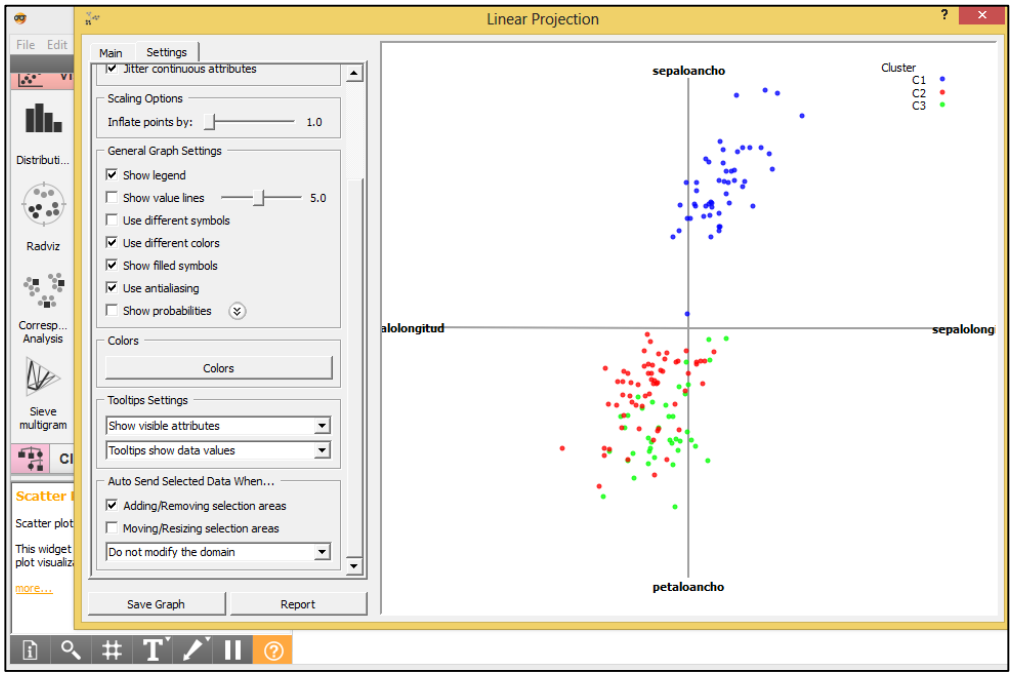


Figura 10: Gráfico de clúster obtenidos en Orange.

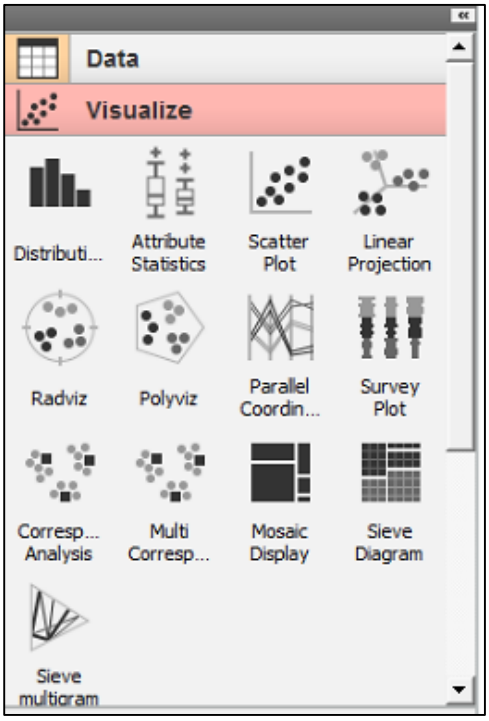


Figura 11: Componentes de visualización de datos en Orange.

ANEXO 6: Procesos de Minería de Datos formados en RapidMiner.

En RapidMiner los proyectos se trabajan mediante una definición de procesos, una definición de proceso es un conjunto de operadores o módulos encadenados en una secuencia.

Una definición de procesos es un archivo que se almacena en un repositorio. Un repositorio, además de almacenar archivos de definición de procesos, puede almacenar datos pre-cargados en archivos de lectura rápida.

En este apartado se detalla cómo se formaron estos procesos, que operadores se necesitó para completar la estructura de análisis y los parámetros que se utilizaron para cada algoritmo seleccionado.

DESERCIÓN

A continuación se describen los procesos formados para los algoritmos ID3, C4.5, CHAID, JRip, PART, Decision Table, DTNB y Ridor enfocados al análisis de factores de deserción.

- **Clasificación mediante ID3**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 2):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.

union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.

filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.

Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.

Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.

asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.

Numerical to Polynominal: este operador se utilizó con el fin de convertir los tipos de datos numéricos en datos polinomiales con el fin de mejorar la calidad de los mismos en la estructura de minería.

división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada, también se utilizó el mismo componente dentro del subproceso de validación cruzada.

muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.

ID3: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada con el nombre de modeloxval.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada con el nombre de evaluación_xval.

Write model: este operador cumple la función de exportar el modelo generado en un archivo externo el cual se puede utilizar en un proceso posterior.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada.

ranking de atributos: este atributo cumple la función de evaluar los pesos de los atributos que pasaron por el modelo generado por el algoritmo ID3 (ver figura 1).

A continuación se muestra el subproceso de validación cruzada formado con el fin de validar el modelo generado (ver figura 1).

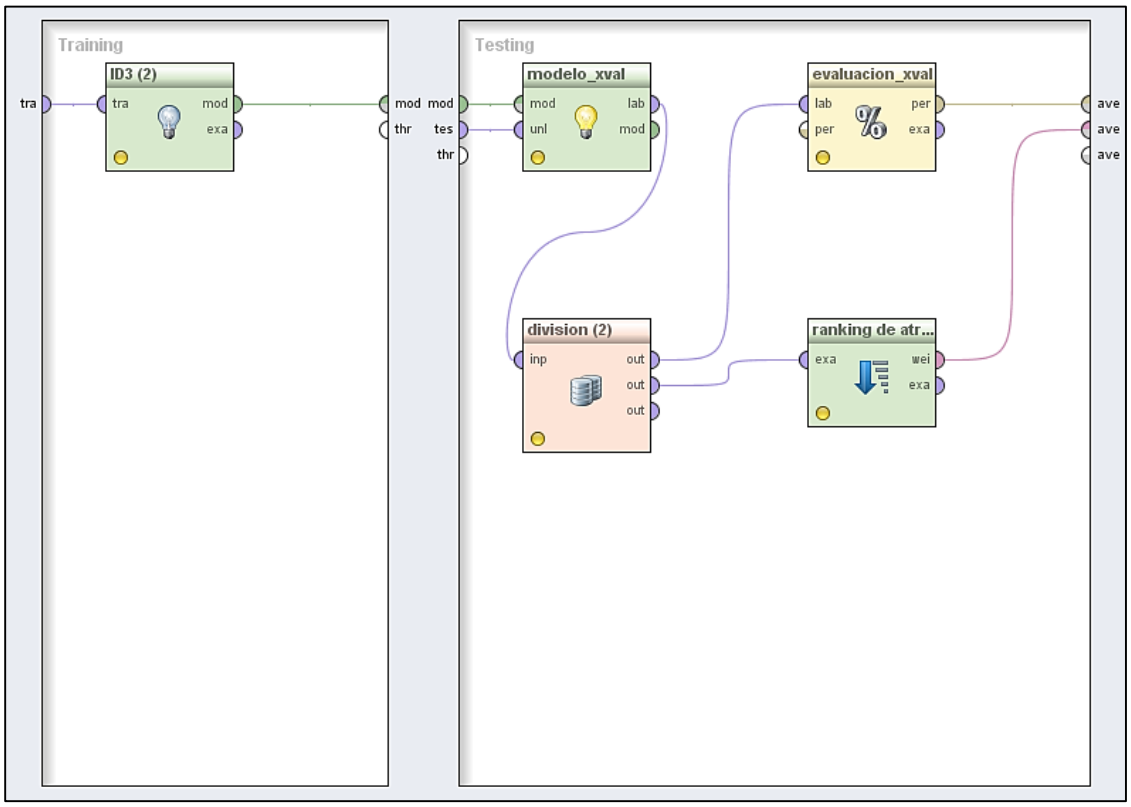


Figura 1: Proceso de Validación Cruzada para el algoritmo ID3.

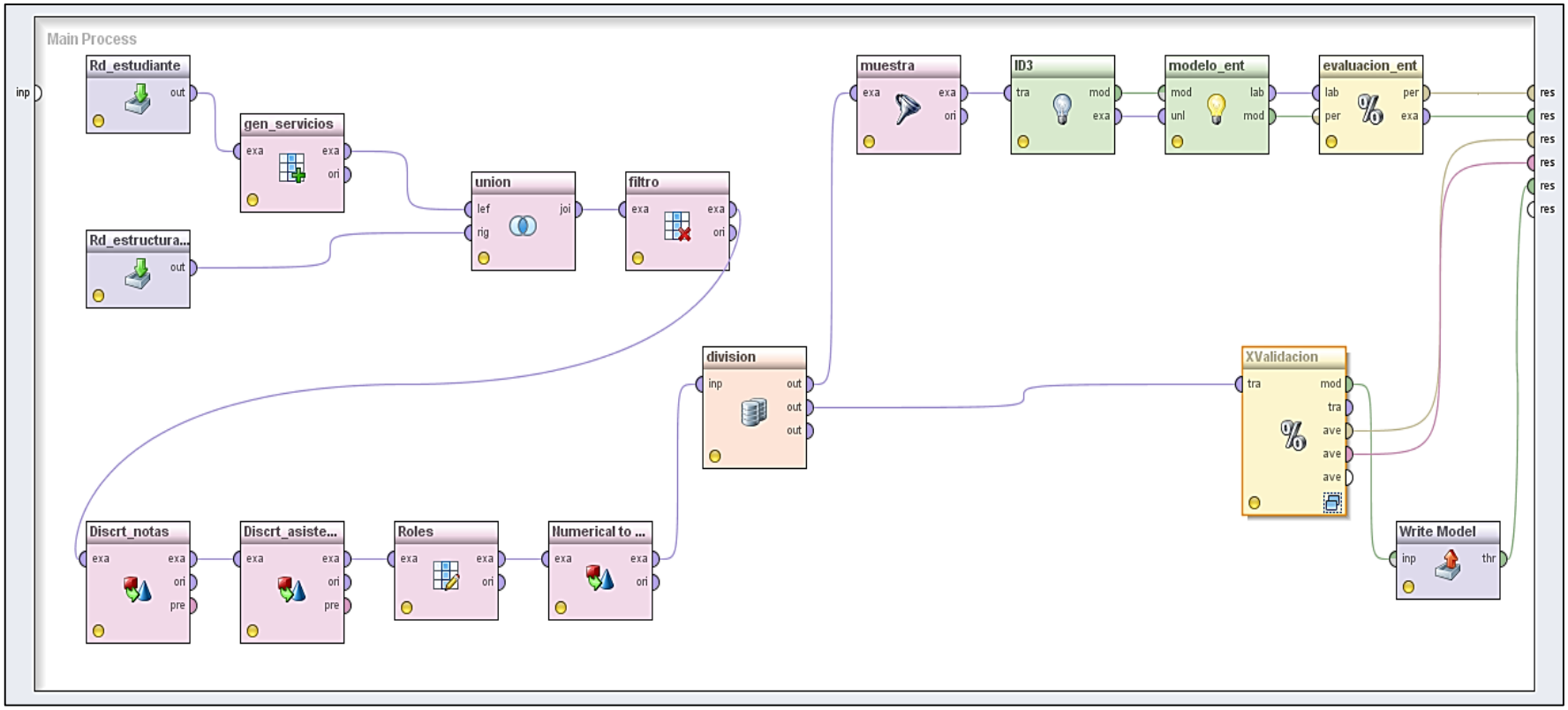


Figura 2: Conjunto de operadores que conforman el proceso para el algoritmo ID3.

- **Clasificación mediante C4.5**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 4):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
Decisión Tree: este operador contiene el algoritmo C4.5 para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.
modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 3).

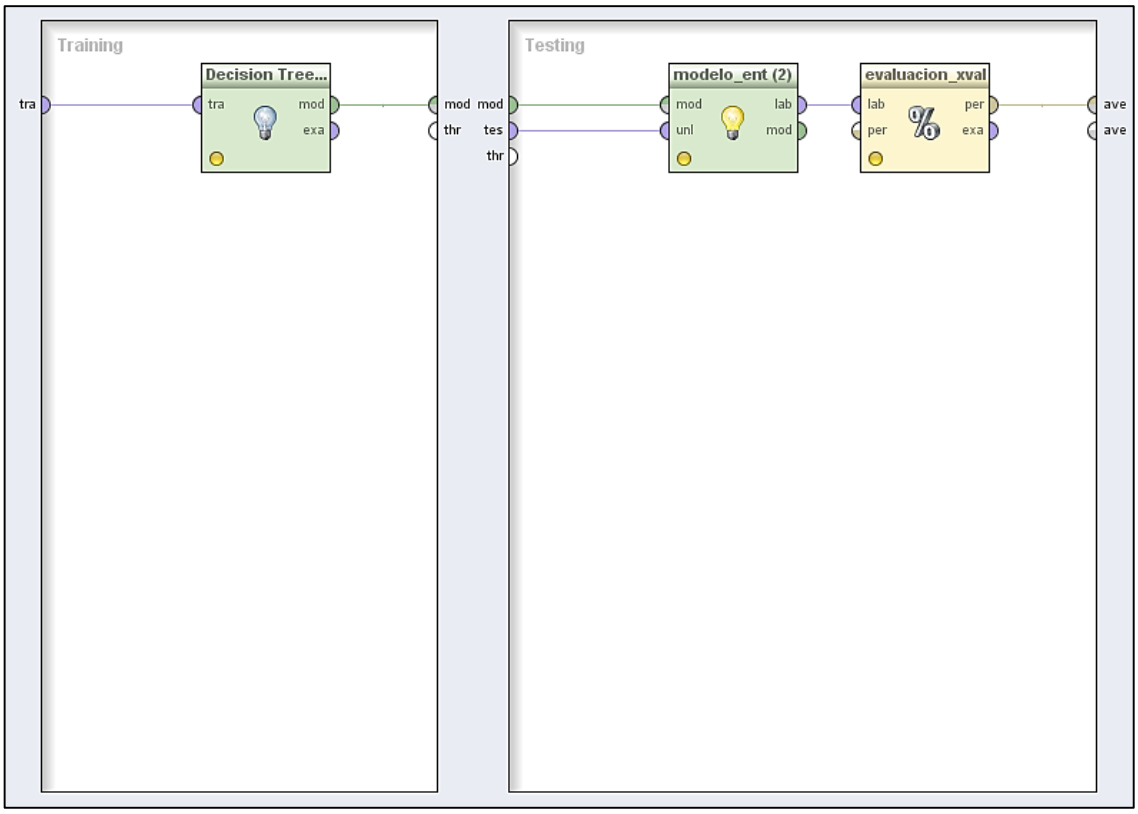


Figura 3: Proceso de Validación Cruzada para el algoritmo C4.5

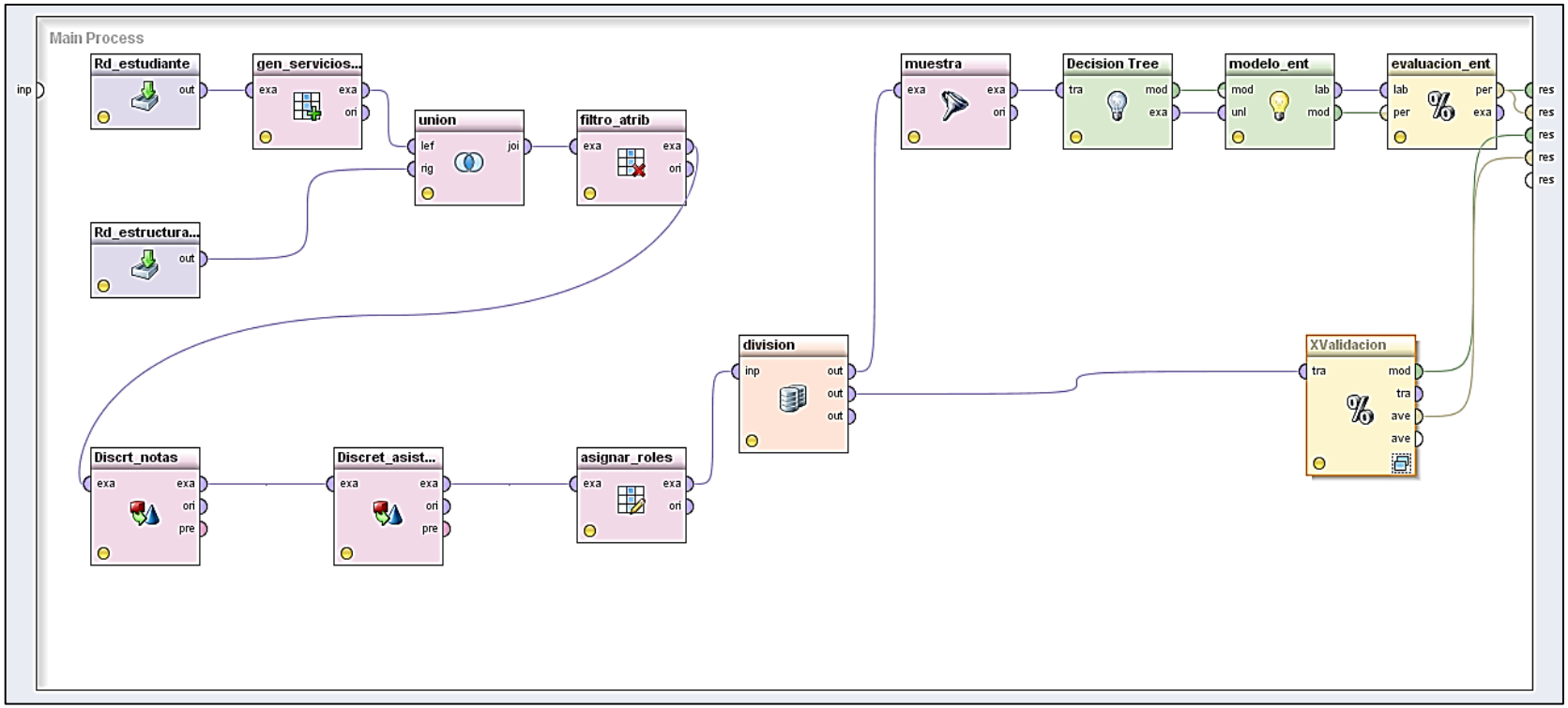


Figura 4: Conjunto de operadores que conforman el proceso para el algoritmo C4.5

- **Clasificación mediante CHAID**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 6):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
Numerical to Polynominal: este operador se utilizó con el fin de convertir los tipos de datos numéricos en datos polinomiales con el fin de mejorar la calidad de los mismos en la estructura de minería.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
CHAID: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion..

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 5).

Figura 5: Proceso de Validación Cruzada para el algoritmo CHAID.

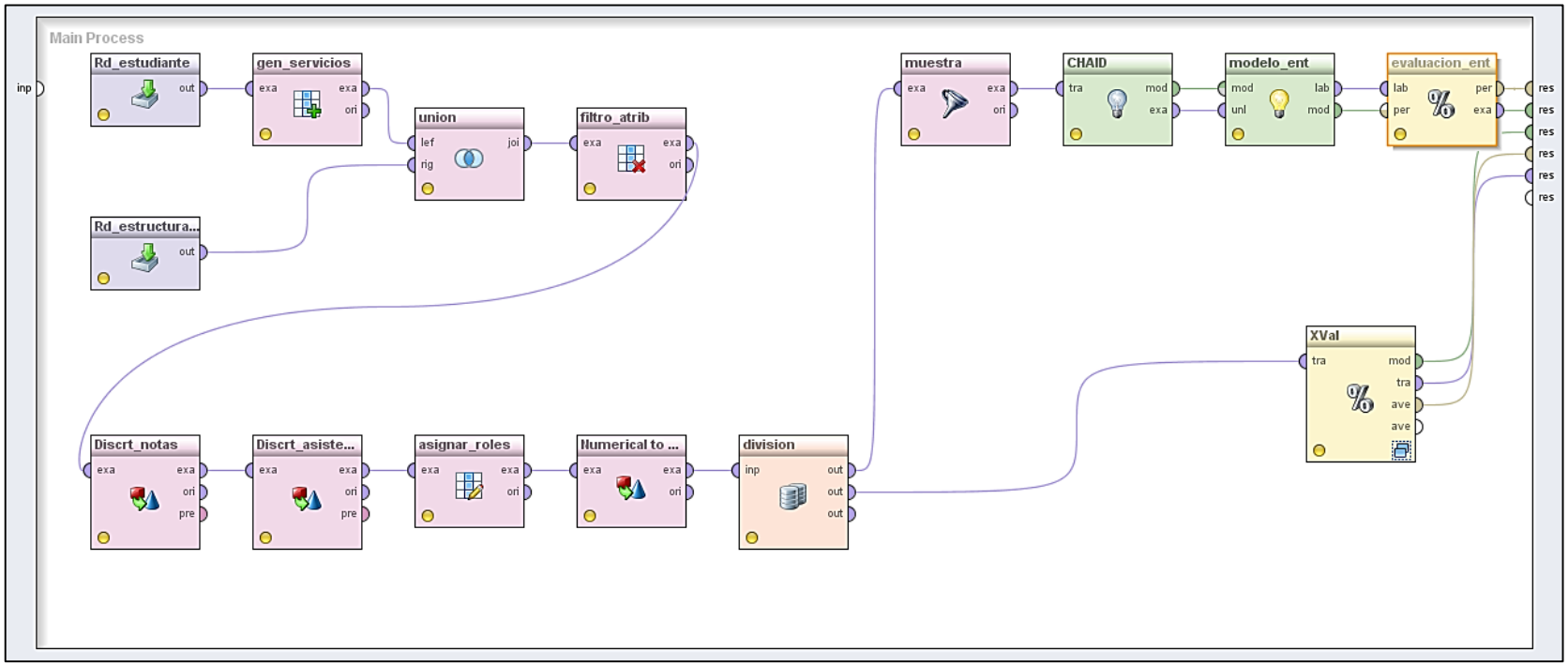


Figura 6: Conjunto de operadores que conforman el proceso para el algoritmo CHAID.

- **Clasificación mediante JRip**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 8):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos de la tabla estudiante.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-JRip: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 7).

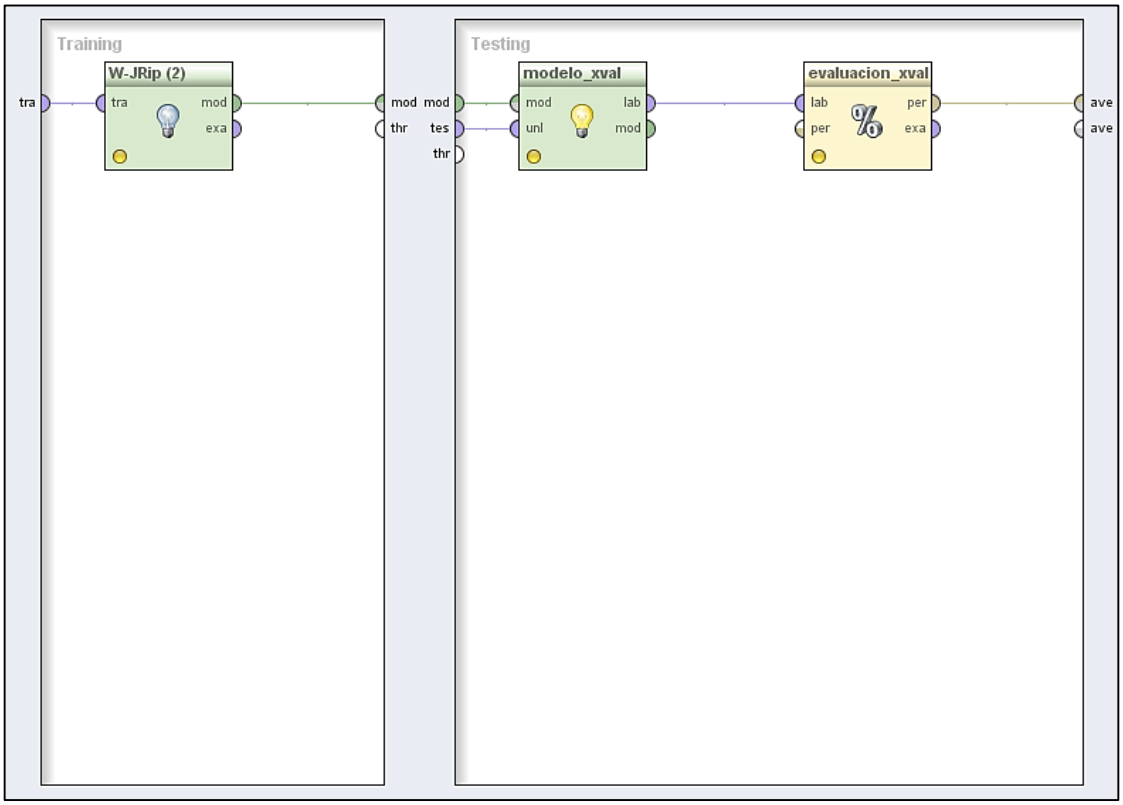


Figura 7: Proceso de Validación Cruzada para el algoritmo JRip.

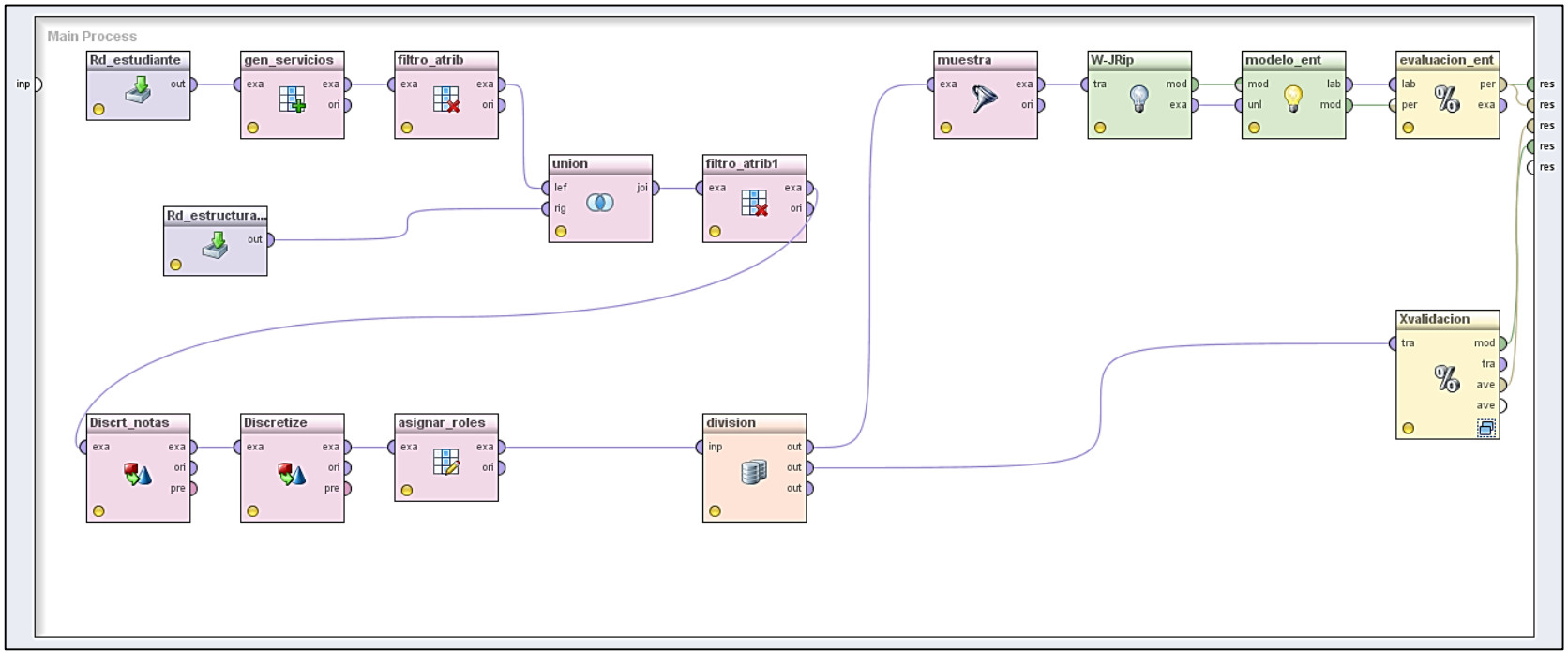


Figura 8: Conjunto de operadores que conforman el proceso para el algoritmo JRip.

- **Clasificación mediante PART**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 10):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos de la tabla estudiante.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada, el mismo componente se utilizó dentro del subproceso de validación cruzada denominado XValidacion.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-PART: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 9).

Write model: este operador cumple la función de exportar el modelo generado en un archivo externo el cual se puede utilizar en un proceso posterior.

ranking de atributos: este atributo cumple la función de evaluar los pesos de los atributos que pasaron por el modelo generado por el algoritmo PART (ver figura 9).

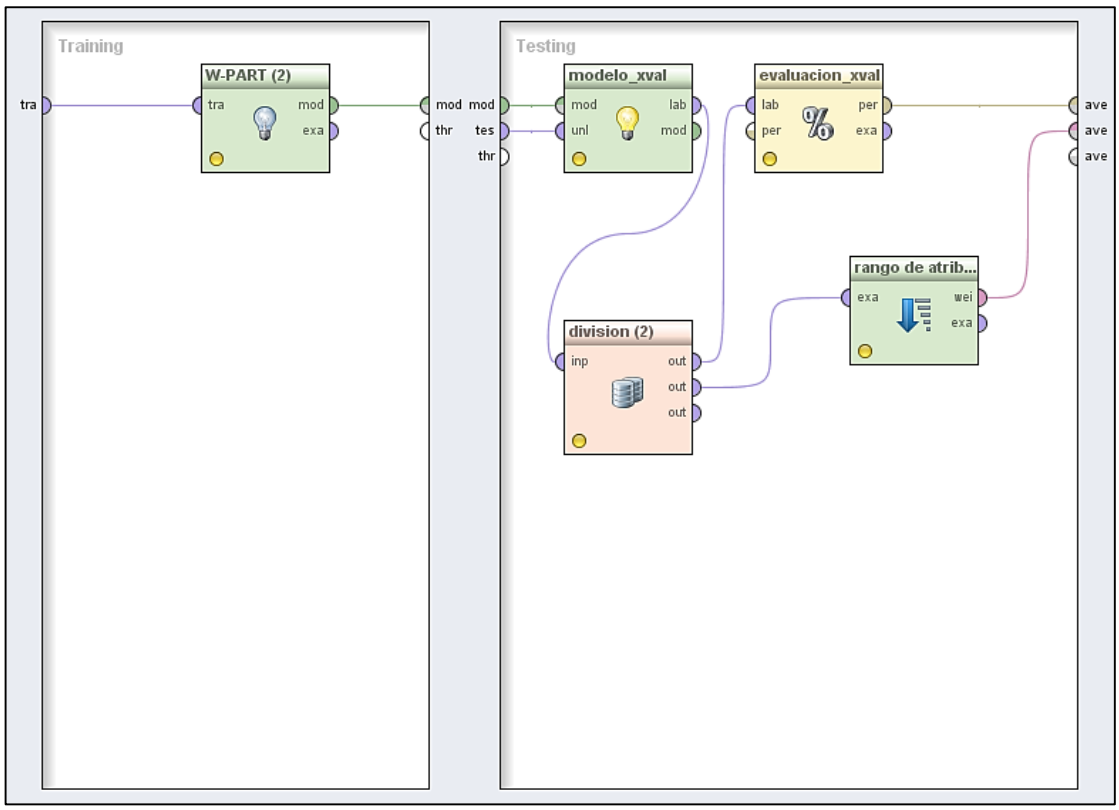


Figura 9: Proceso de Validación Cruzada para el algoritmo PART.

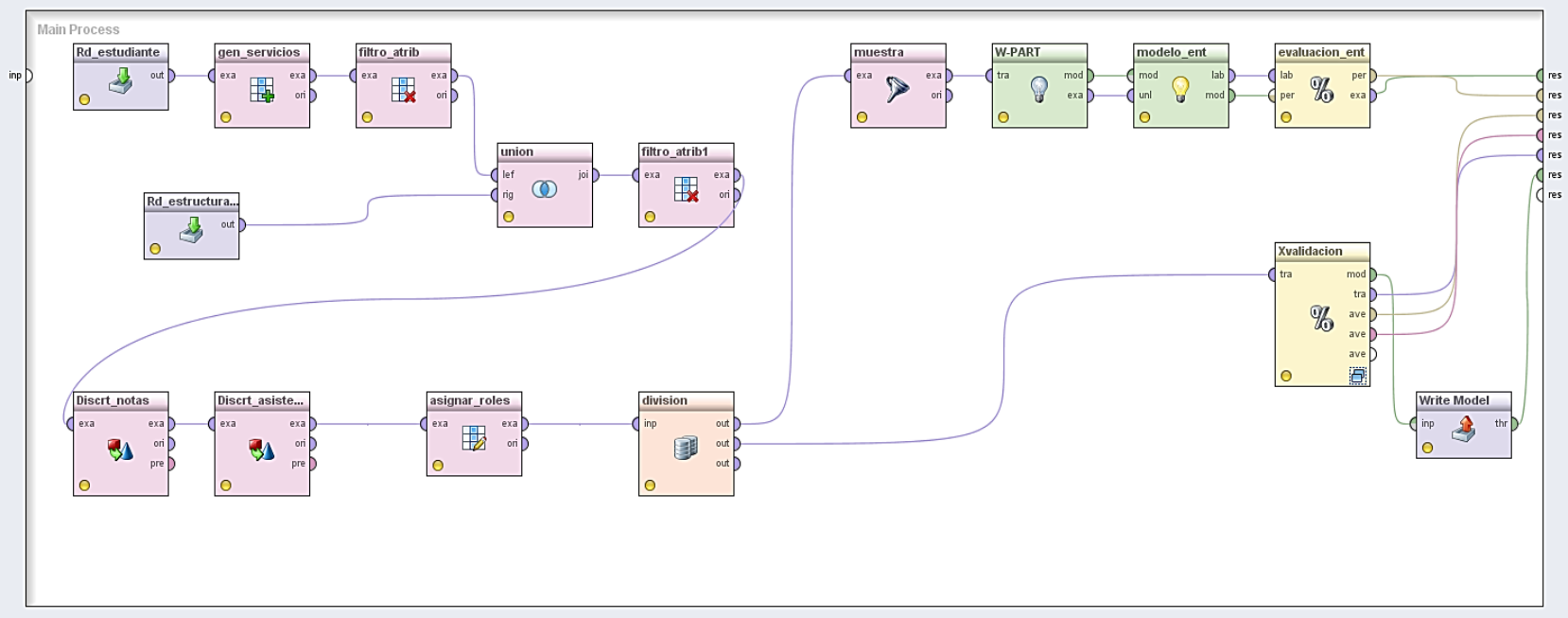


Figura 10: Conjunto de operadores que conforman el proceso para el algoritmo PART.

- **Clasificación mediante DECISION TABLE**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 12):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos de la tabla estudiante.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada, el mismo componente se utilizó dentro del subproceso de validación cruzada denominado XValidacion.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-DecisionTable: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 9).

Write model: este operador cumple la función de exportar el modelo generado en un archivo externo el cual se puede utilizar en un proceso posterior.

ranking de atributos: este atributo cumple la función de evaluar los pesos de los atributos que pasaron por el modelo generado por el algoritmo PART (ver figura 11).

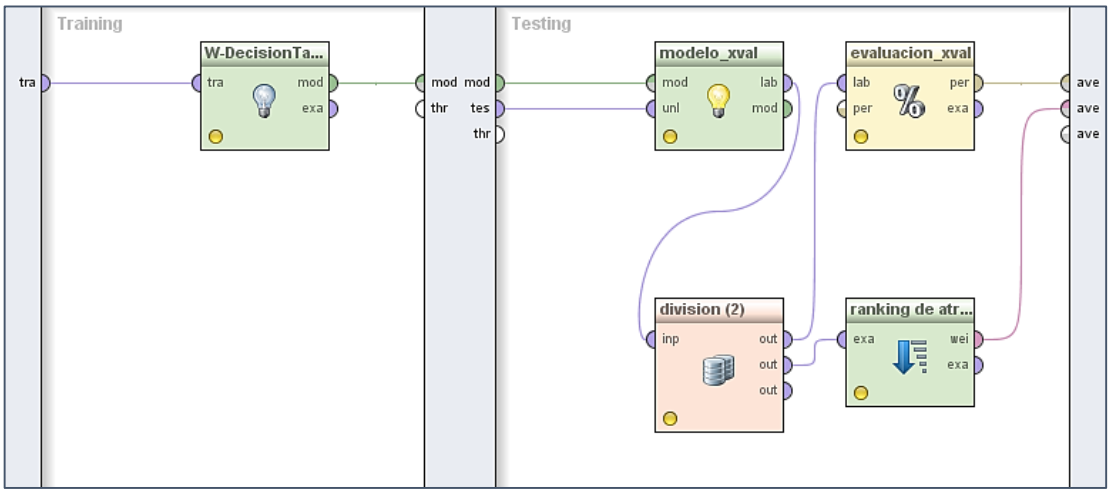


Figura 11: Proceso de Validación Cruzada para el algoritmo Decision Table.

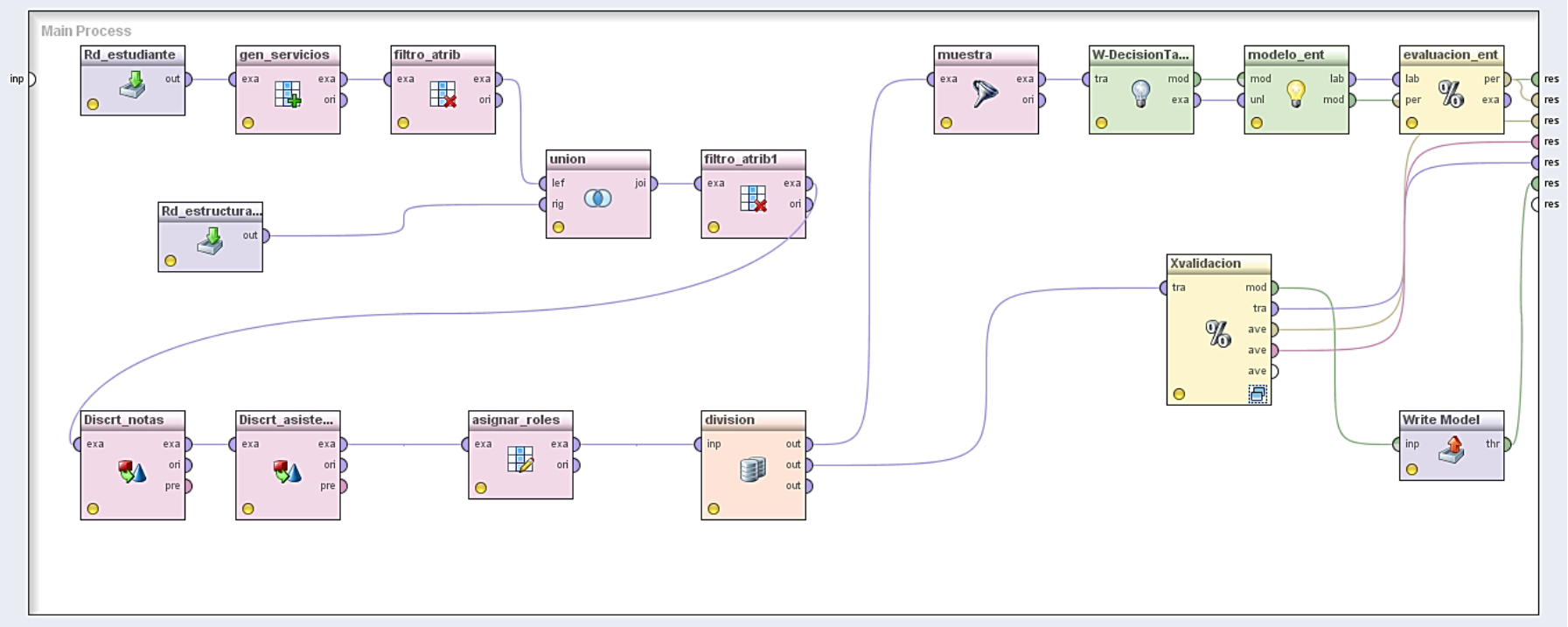


Figura 12: Conjunto de operadores que conforman el proceso para el algoritmo Decision Table.

- **Clasificación mediante DTNB**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 14):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos de la tabla estudiante.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada, el mismo componente se utilizó dentro del subproceso de validación cruzada denominado XValidacion.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-DTNB: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 9).

ranking de atributos: este atributo cumple la función de evaluar los pesos de los atributos que pasaron por el modelo generado por el algoritmo PART (ver figura 13).

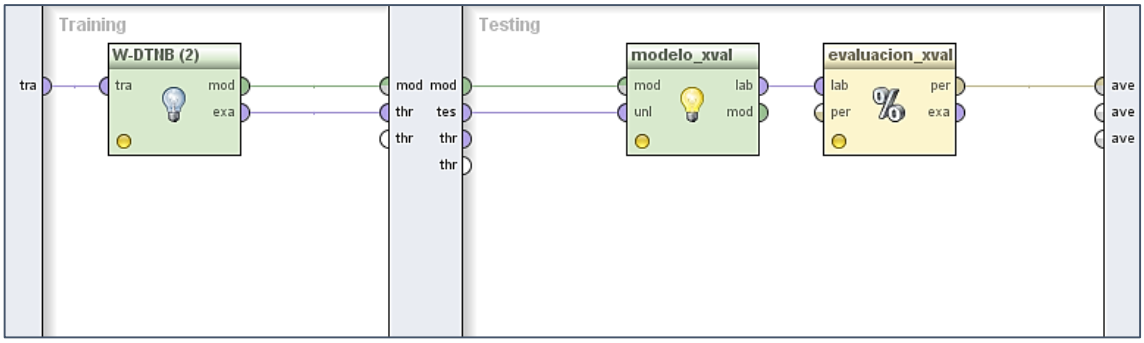


Figura 13: Proceso de Validación Cruzada para el algoritmo DTNB.

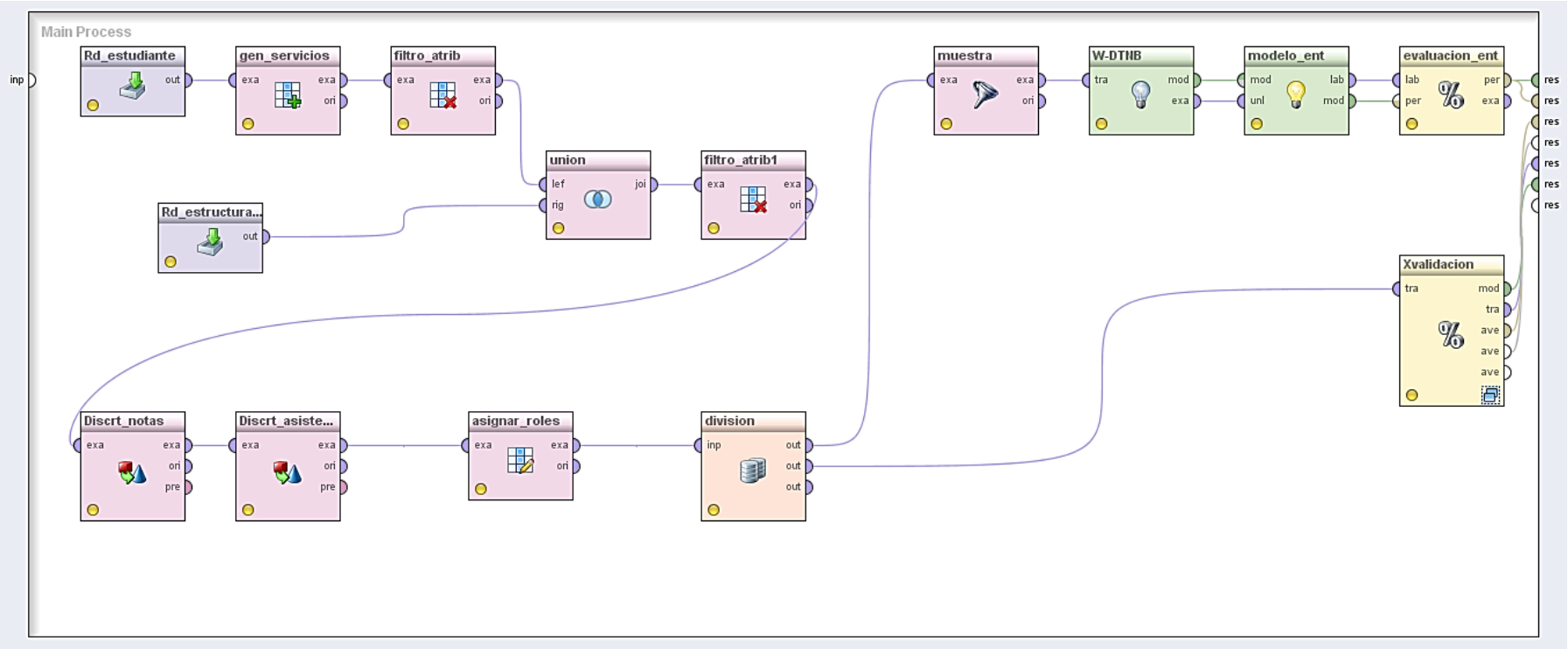


Figura 14: Conjunto de operadores que conforman el proceso para el algoritmo DTNB.

- **Clasificación mediante Ridor**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 16):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md: este operador se utilizó para leer los registros de la tabla estructura_md de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos de la tabla estudiante.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo estado se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada, el mismo componente se utilizó dentro del subproceso de validación cruzada denominado XValidacion.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-Ridor: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 9).

ranking de atributos: este atributo cumple la función de evaluar los pesos de los atributos que pasaron por el modelo generado por el algoritmo PART (ver figura 15).

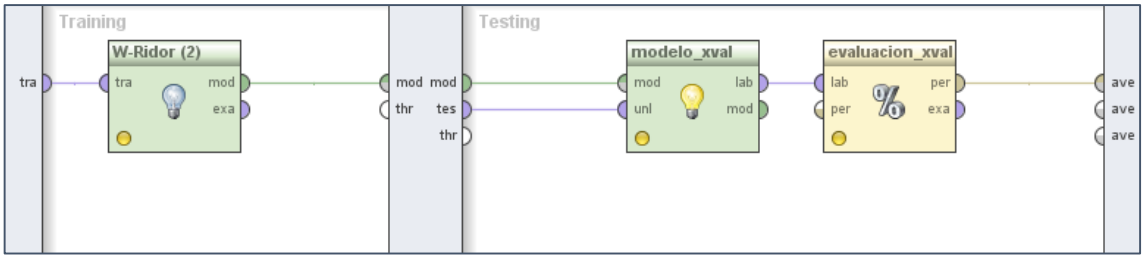


Figura 15: Proceso de Validación Cruzada para el algoritmo Ridor.

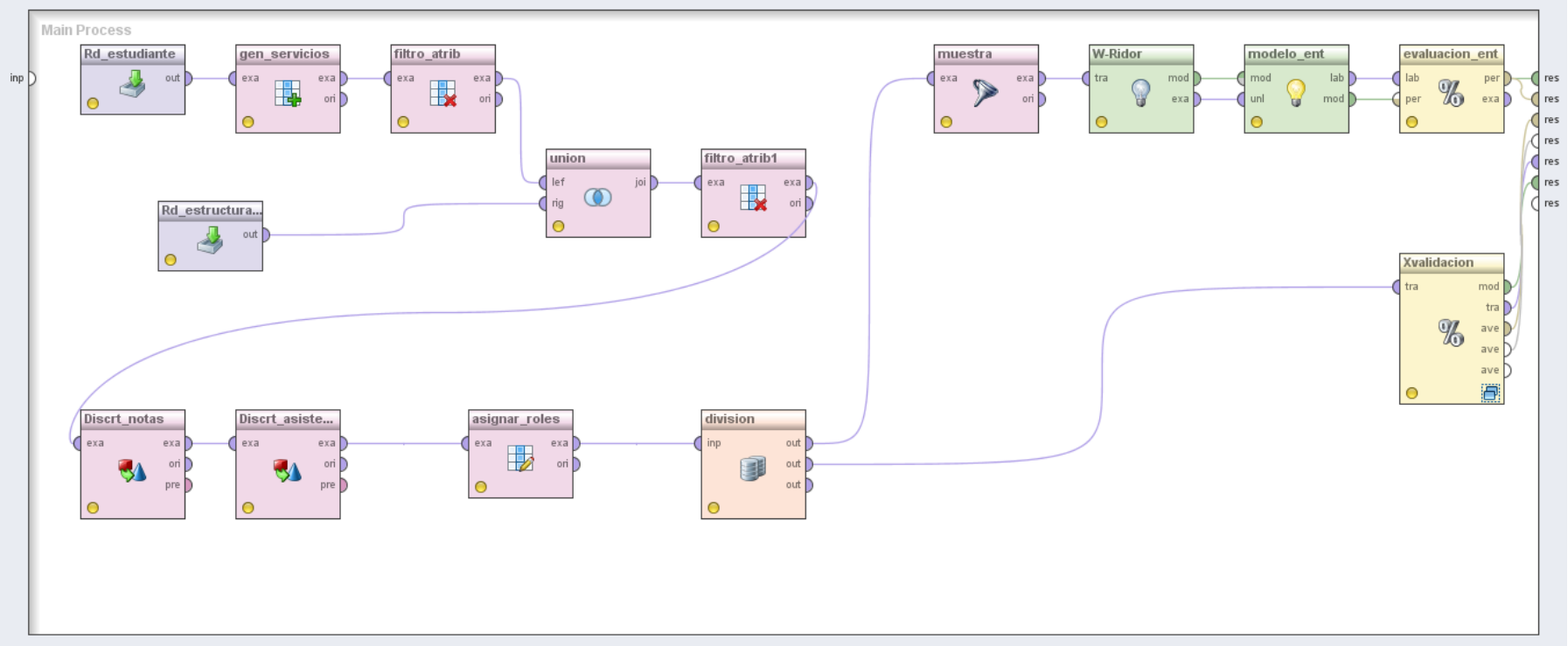


Figura 16: Conjunto de operadores que conforman el proceso para el algoritmo Ridor.

REPROBACIÓN

A continuación se describen los procesos formados para los algoritmos JRip, PART, Ridor, Decision Table y DTNB, enfocados al análisis de factores de reprobación.

- **Clasificación mediante JRip**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 18):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md_reprob: este operador se utilizó para leer los registros de la tabla estructura_md_reprob de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo reprobó se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.

W-JRip: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 17).

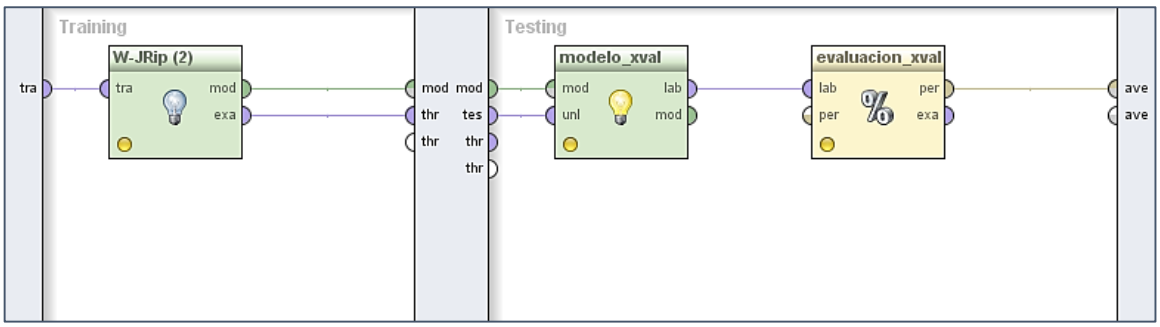


Figura 17: Proceso de Validación Cruzada para el algoritmo JRip.

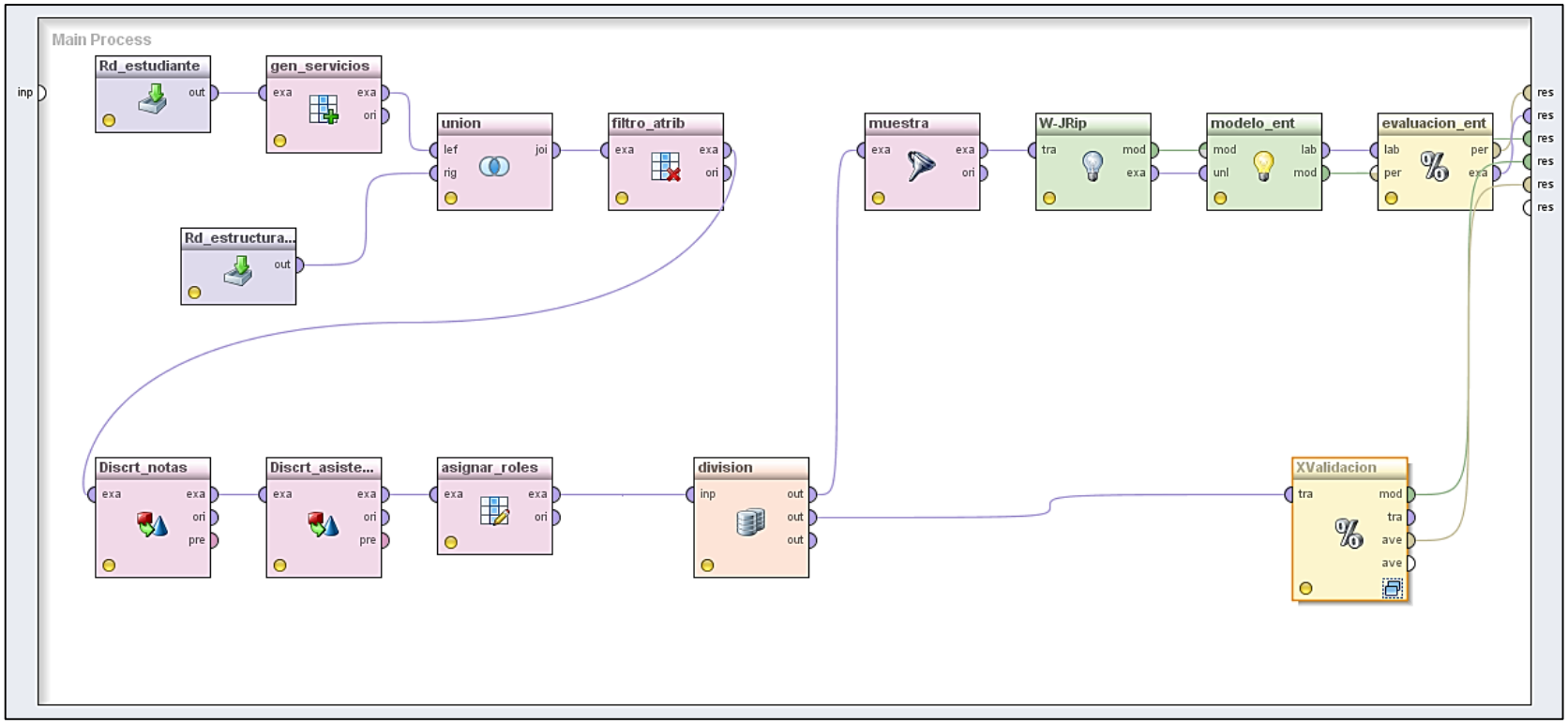


Figura 18: Conjunto de operadores que conforman el proceso para el algoritmo JRip.

- **Clasificación mediante PART**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 20):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md_reprob: este operador se utilizó para leer los registros de la tabla estructura_md_reprob de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo reprobo se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-PART: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.
modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion..

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 19).

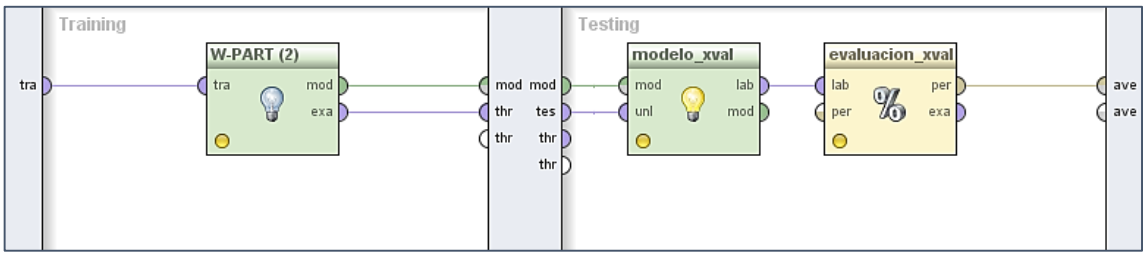


Figura 19: Proceso de Validación Cruzada para el algoritmo PART.

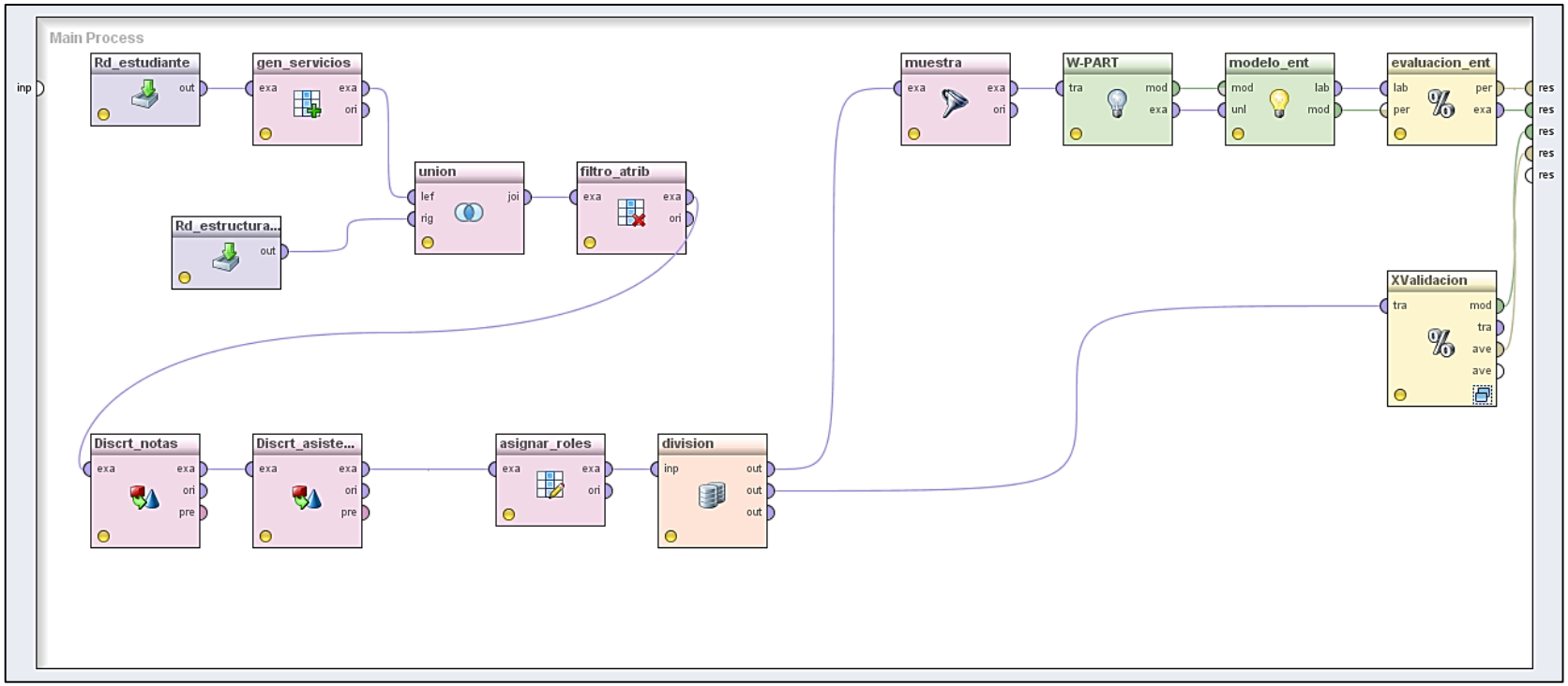


Figura 20: Conjunto de operadores que conforman el proceso para el algoritmo PART.

- **Clasificación mediante Decision Table**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 22):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md_reprob: este operador se utilizó para leer los registros de la tabla estructura_md_reprob de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib1: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo reprobo se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-DecisionTable: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.
modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion..

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 21).

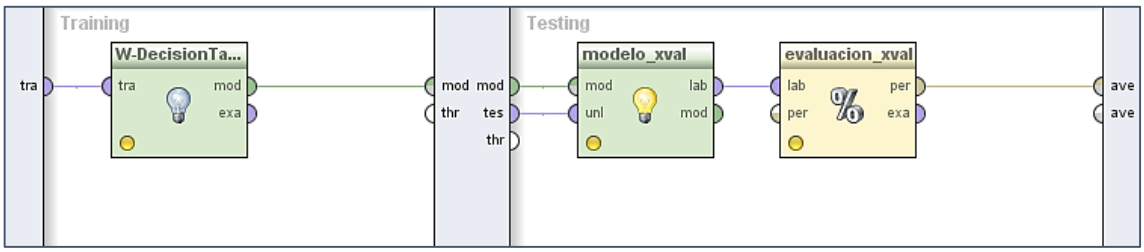


Figura 21: Proceso de Validación Cruzada para el algoritmo Decision Table.

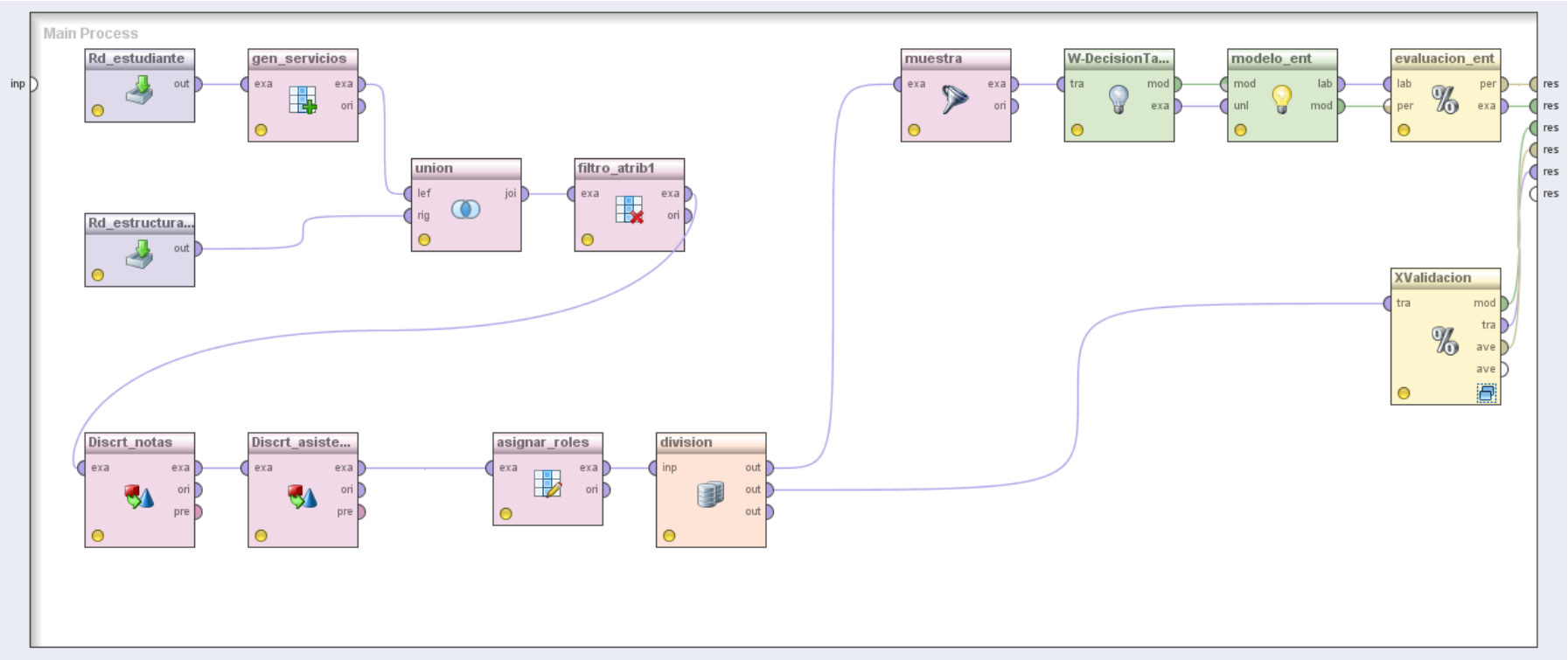


Figura 22: Conjunto de operadores que conforman el proceso para el algoritmo Decision Table.

- **Clasificación mediante DTNB**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 24):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md_reprob: este operador se utilizó para leer los registros de la tabla estructura_md_reprob de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo reprobo se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-DTNB: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.
modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 23).

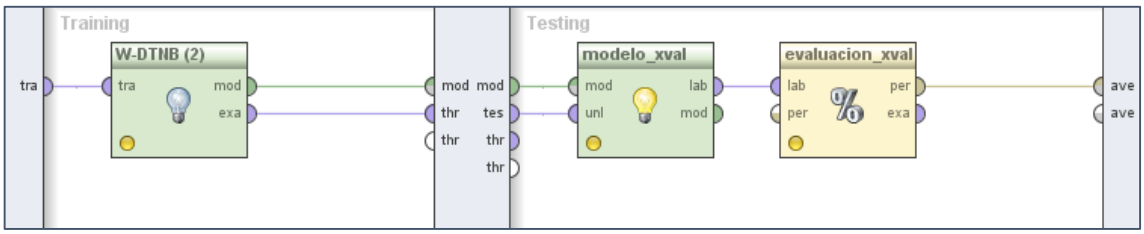


Figura 23: Proceso de Validación Cruzada para el algoritmo DTNB.

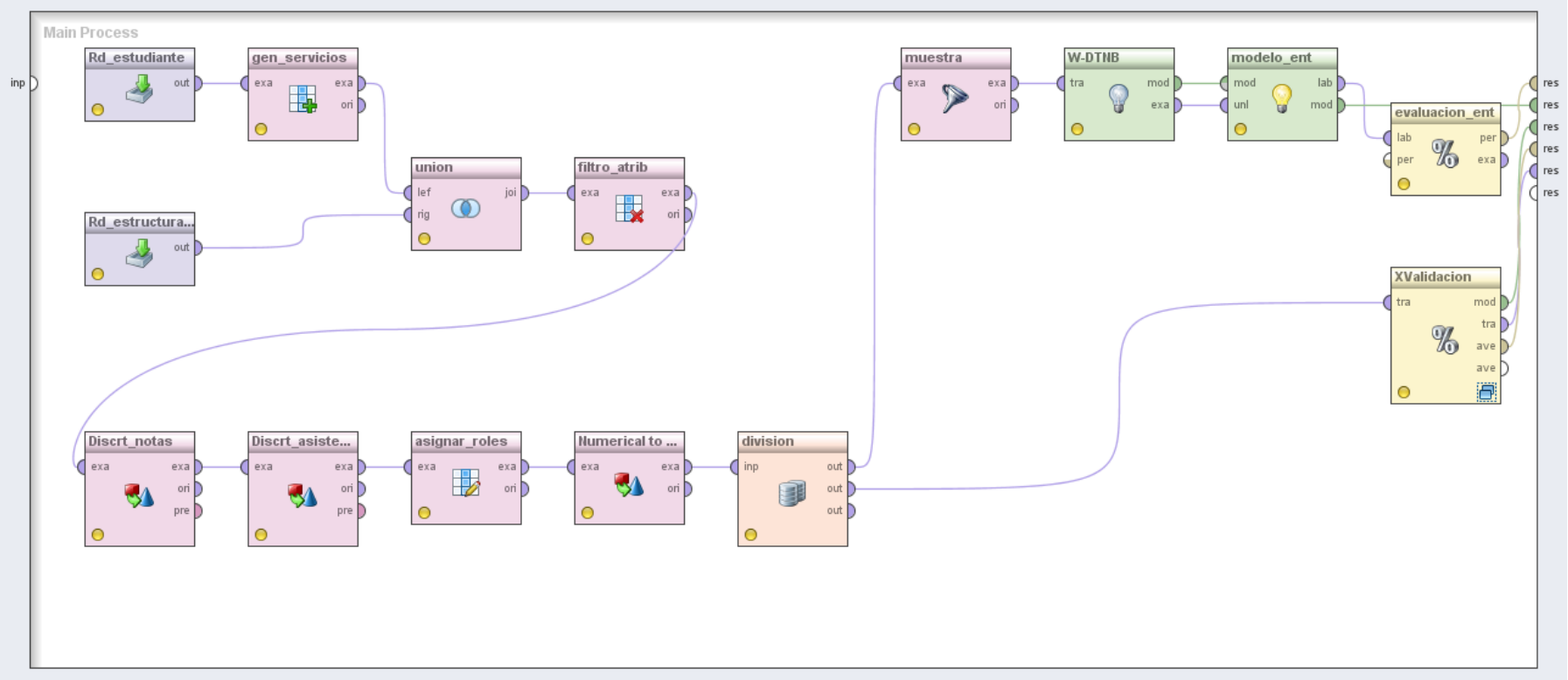


Figura 24: Conjunto de operadores que conforman el proceso para el algoritmo DTNB.

- **Clasificación mediante Ridor**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 26):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md_reprob: este operador se utilizó para leer los registros de la tabla estructura_md_reprob de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, los roles asignados son: para el atributo reprobo se le asignó el rol label y para el atributo numeroIdentificacion se le asignó el rol de id.
división: este operador cumple la función de crear copias idénticas de los datos, en este caso se usó con el fin de evaluar de manera simultánea con datos de entrenamiento y mediante validación cruzada.
muestra: este operador cumple la función de particionar un 70% de la muestra total con el fin de evaluar estos datos en un entrenamiento del algoritmo.
W-Ridor: este operador contiene el algoritmo para generar el modelo en base a los datos ingresados, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.
modelo_ent: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

evaluación_ent: este operador cumple la función de evaluar el rendimiento del algoritmo y muestra los resultados a través de una matriz de confusión, también se utilizó el mismo componente dentro del subproceso de validación cruzada denominado XValidacion.

XValidacion: este operador contiene los operadores necesarios para realizar la validación del modelo mediante el método de validación cruzada (ver figura 25).

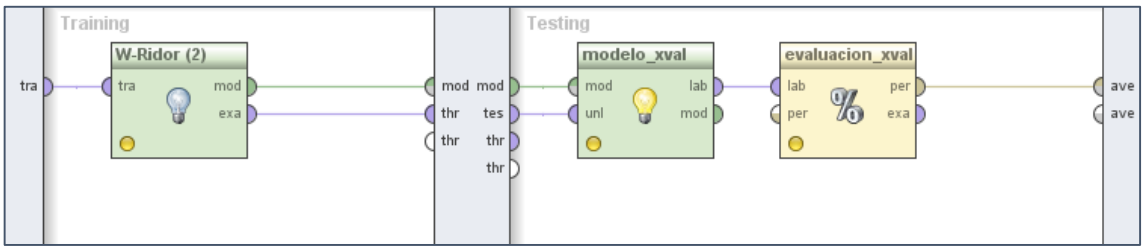


Figura 25: Proceso de Validación Cruzada para el algoritmo Ridor.

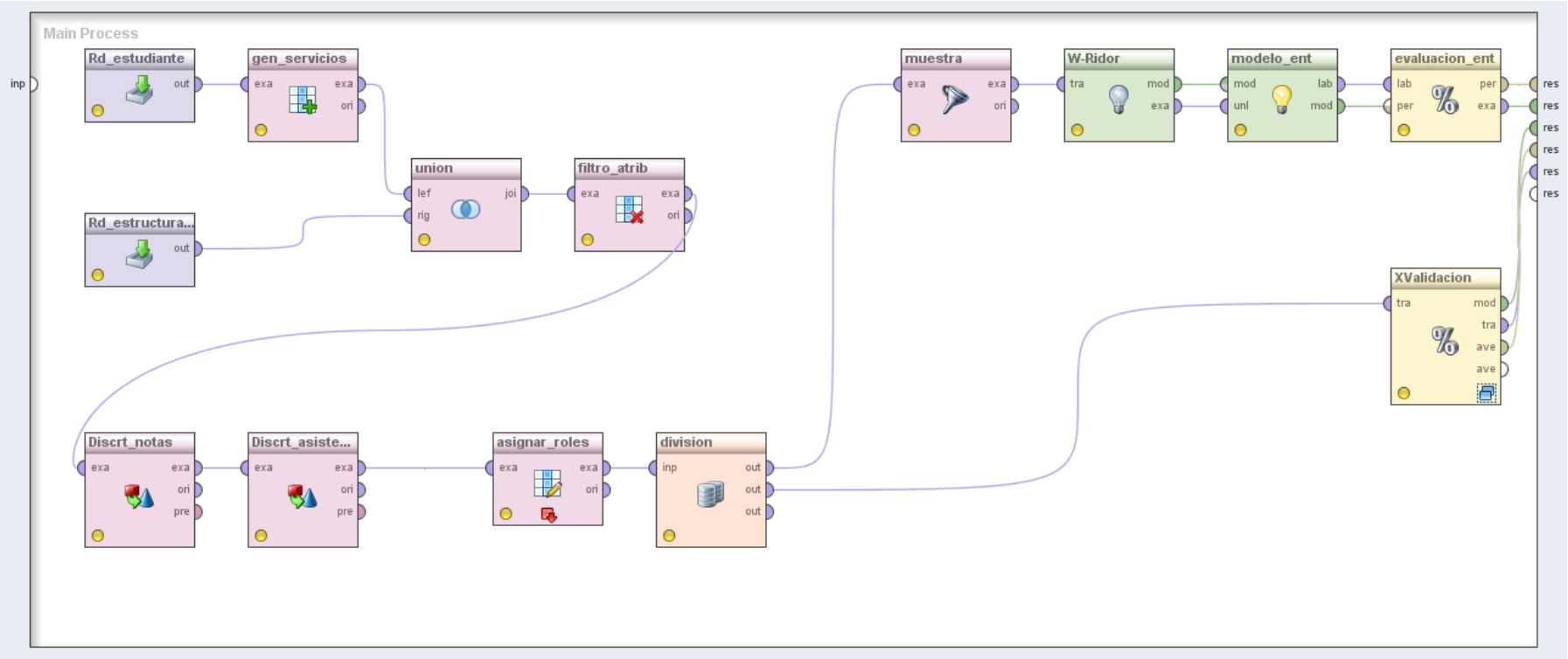


Figura 26: Conjunto de operadores que conforman el proceso para el algoritmo Ridor.

ANEXO 7: Procesos de Modelo con datos de Estudiantes Actuales

A continuación se describen los operadores utilizados para formar el proceso del algoritmo Decision Table para predecir estudiantes con probabilidades de desertar.

- **Clasificación mediante Decision Table**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 27):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md3: este operador se utilizó para leer los registros de la tabla estructura_md3 de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.
Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.
asignar_roles: este operador agrega los roles necesarios a la estructura de datos, el único rol asignado es al atributo numeroIdentificacion que se le asignó el rol de id.
Read model_decisiontable: en este operador se carga el modelo generado por el algoritmo Decision Table.
modelo_decisiontable: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo con los nuevos datos de prueba.

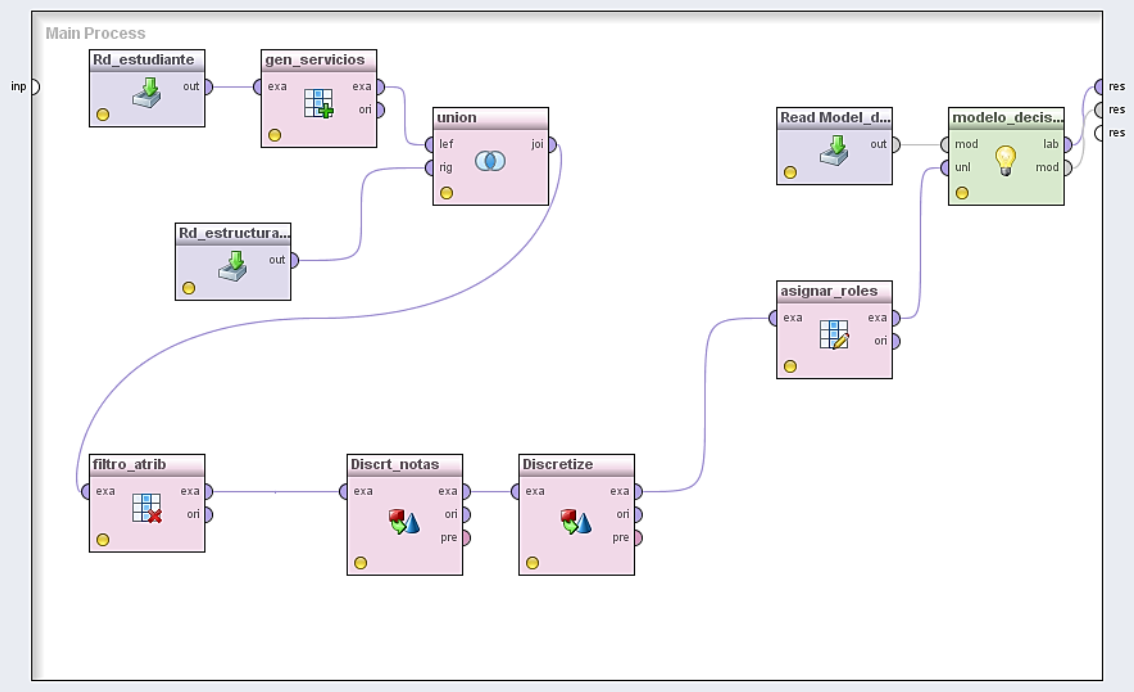


Figura 27: Conjunto de operadores que conforman el proceso para el algoritmo Deicision Table.

- **Clasificación mediante PART**

A continuación se describen los operadores necesarios para formar este proceso (ver figura 28):

rd_estudiante: este operador se utilizó con la finalidad de leer los registros de la tabla estudiante de la base de datos mineriacis que se generó previamente.
rd_estructura_md3: este operador se utilizó para leer los registros de la tabla estructura_md3 de la base de datos mineriacis que se generó previamente.
gen_servicios: este operador realiza la generación de un nuevo atributo denominado servicios a partir de los datos extraídos de la tabla estudiante mediante el operador rd_estudiante.
union: este operador se utilizó para unir datos de la tabla estudiante con la tabla estructura_md de la base de datos mineriacis con el objetivo de completar la estructura de minería.
filtro_atrib: este operador se utilizó con el fin de seleccionar los atributos considerados dentro de la estructura de minería dejando por fuera a las no necesarias.
Discrt_notas: este operador realiza la tarea de discretizar el atributo promedio_notas, con el fin de mejorar la calidad de la estructura.

Discrt_asistencia: este operador realiza la tarea de discretizar el atributo promedio_asistencias, con el fin de mejorar la calidad de la estructura.

asignar_roles: este operador agrega los roles necesarios a la estructura de datos, el único rol asignado es al atributo numeroIdentificacion que se le asignó el rol de id.

Read model_part: en este operador se carga el modelo generado por el algoritmo PART.

modelo_part: este operador se utiliza con el fin de consultar el modelo generado por el algoritmo y evaluarlo con los nuevos datos de prueba.

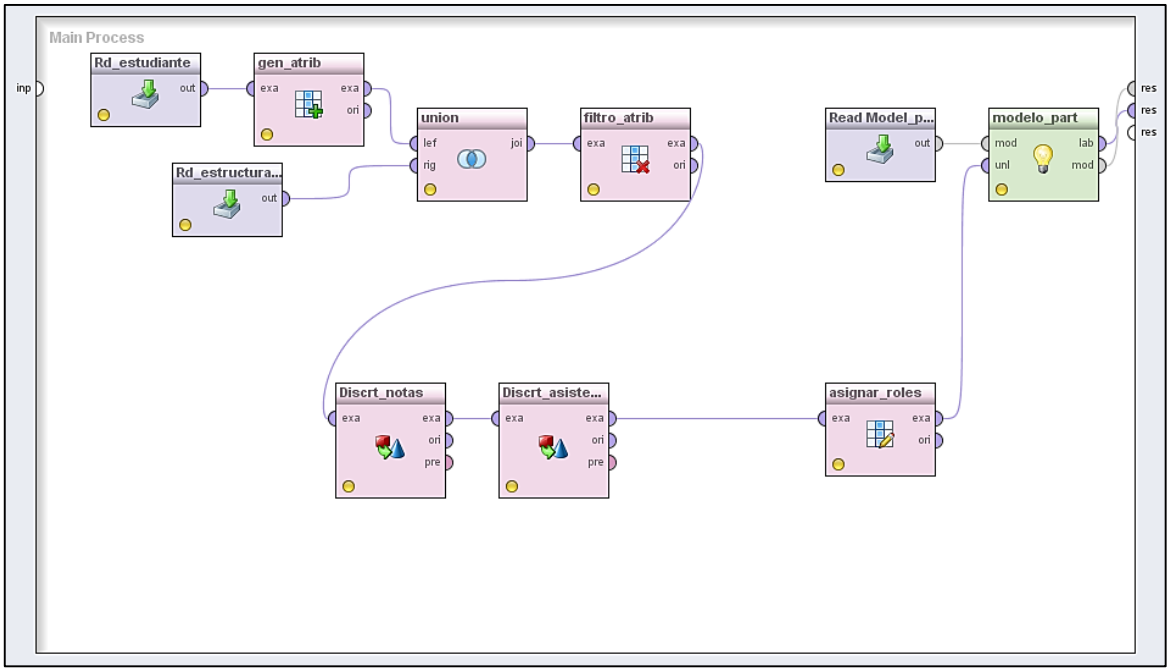


Figura 28: Conjunto de operadores que conforman el proceso para el algoritmo PART.

ANEXO 8: Modelos Generados por los mejores Algoritmos

A continuación se describen las reglas generadas por los algoritmos con el mejor rendimiento en deserción y reprobación.

Modelo 1: Reglas generadas por el algoritmo PART para predecir la Deserción.

- modulos_reprobados > 0 AND promedio_notas = malo AND cambio_carrera = no AND periodo_reprobacion = 1-3: desertor (745.0/2.0)
- modulos_reprobados <= 0: egresado (693.0) periodo_reprobacion = 1-3 AND sexo = m AND promedio_asistencia = alto: desertor (40.0)
- promedio_notas = malo AND periodo_reprobacion = 4-7 AND promedio_asistencia = bajo AND carrera = IS: desertor (145.0/5.0)
- periodo_reprobacion = 1-3 AND sexo = f: desertor (36.0)
- modulos_reprobados > 1 AND periodo_reprobacion = 1-3: desertor (54.0/5.0)
- promedio_notas = malo AND periodo_reprobacion = 4-7 AND estado_civil = soltero AND bienestar_servicios = no AND servicios = 1 AND carrera = IE: desertor (34.0/5.0)
- promedio_notas = malo AND periodo_reprobacion = 4-7 AND estado_civil = soltero AND bienestar_servicios = no AND servicios = 3: desertor (22.0)
- promedio_notas = malo AND distancia_origen <= 3 AND periodo_reprobacion = 4-7 AND estado_civil = soltero AND carrera = IE AND promedio_asistencia = bajo: desertor (17.0/2.0)
- modulos_reprobados > 1 AND periodo_reprobacion = 4-7 AND modulos_reprobados <= 2 AND carrera = IS: desertor (17.0/1.0)
- promedio_notas = malo AND distancia_origen <= 3 AND carrera = IET: desertor (14.0/1.0)
- periodo_reprobacion = 8-11 AND promedio_asistencia = medio: egresado (14.0/1.0)
- promedio_notas = malo AND distancia_origen > 3: desertor (12.0)
- promedio_notas = malo AND horario_estudio = matutino AND estado_civil = soltero AND servicios = 1 AND edad_ingreso <= 23 AND carrera = IS AND periodo_reprobacion = 4-7 AND sexo = m: desertor (11.0/4.0)

- promedio_notas = malo AND horario_estudio = matutino AND estado_civil = soltero AND servicios = 1 AND carrera = IGAOT: desertor (7.0/1.0)
- tipo_beca = nignuna AND promedio_notas = malo AND horario_estudio = matutino AND estado_civil = soltero AND distancia_origen <= 2: desertor (13.0/1.0)
- tipo_beca = nignuna AND periodo_reprobacion = 1-3 AND horario_estudio = matutino: desertor (12.0/1.0)
- promedio_notas = regular AND modulos_reprobados <= 2 AND tipo_beca = nignuna AND periodo_reprobacion = 4-7 AND sexo = m: egresado (39.0/7.0)
- tipo_beca = nignuna AND promedio_notas = regular AND carrera = IE: desertor (15.0/1.0)
- tipo_beca = nignuna AND servicios = 3 AND promedio_notas = malo: egresado (8.0/1.0)
- estado_gestancion = no AND tipo_beca = nignuna AND promedio_notas = malo AND estado_civil = casado AND periodo_reprobacion = 4-7 AND
  edad_ingreso > 19: desertor (9.0)
- tipo_beca = a: egresado (6.0/1.0)
- promedio_notas = muy_bueno: desertor (5.0)
- estado_civil = casado AND distancia_origen <= 2 AND carrera = IE: egresado (7.0/1.0)
- estado_civil = casado AND horario_estudio = vespertino: desertor (4.0)
- estado_civil = soltero AND servicios = 3: egresado (10.0/4.0)
- servicios = 1 AND estado_civil = soltero AND promedio_asistencia = medio AND sexo = m AND carrera = TECI AND edad_ingreso <= 22: desertor (8.0/3.0)
- Servicios = 1 AND estado_civil = union libre: egresado (4.0/1.0)
- estado_civil = soltero AND servicios = 4: egresado (4.0/1.0)
- Servicios = 1 AND estado_civil = casado: egresado (3.0/1.0)
- estado_civil = divorciado: desertor (3.0)
- Servicios = 1 AND carrera = TECI AND promedio_asistencia = bajo AND edad_ingreso <= 22: egresado (2.0)
- Servicios = 1 AND promedio_asistencia = medio AND edad_ingreso <= 22: desertor (8.0/1.0)
- Servicios = 1 AND distancia_origen <= 2 AND promedio_notas = malo: desertor (14.0/3.0)
- servicios = 1 AND promedio_notas = malo: egresado (6.0)

- Servicios = 1 AND promedio_notas = regular: egresado (5.0/2.0)
- promedio_notas = bueno: egresado (5.0/1.0)

Modelo 2: Generado por el algoritmo Ridor para encontrar Factores de Reprobación.

- reprobo = no (2821.0/1840.0)
- Except (promedio_notas = regular) and (promedio_asistencia = bajo) and (carrera = IE) and (edad_ingreso > 20.5) => reprobo = si (79.0/0.0) [19.0/0.0]
- Except (promedio_notas = regular) and (promedio_asistencia = bajo) and (carrera = IS) => reprobo = si (421.0/21.0) [102.0/7.0]
- Except (promedio_notas = regular) and (carrera = IE) and (distancia_origen > 1.5) and (servicios = 1) and (estado_civil = soltero) => reprobo = si (98.0/2.0) [15.0/1.0]
- Except (promedio_notas = regular) and (promedio_asistencia = bajo) and (edad_ingreso <= 19.5) and (distancia_origen > 1.5) and (servicios = 1) => reprobo = si (35.0/0.0) [11.0/3.0]
- Except (promedio_notas = regular) and (carrera = IE) and (estado_civil = soltero) and (edad_ingreso <= 17.5) => reprobo = si (34.0/1.0) [12.0/0.0]
- Except (promedio_notas = regular) and (promedio_asistencia = bajo) => reprobo = si (374.0/39.0) [95.0/13.0]
- Except (promedio_notas = regular) and (carrera = IS) => reprobo = si (134.0/16.0) [28.0/3.0]
- Except (promedio_notas = regular) and (horario_estudio = vespertino) and (promedio_asistencia = medio) and (distancia_origen > 1.5) and (bienestar_servicios = no) => reprobo = si (24.0/0.0) [7.0/0.0]
- Except (promedio_notas = regular) and (carrera = IE) and (edad_ingreso > 19.5) => reprobo = si (14.0/0.0) [4.0/1.0]
- Except (edad_ingreso <= 19.5) and (horario_estudio = vespertino) and (promedio_notas = regular) and (madre_trabaja = si) and (distancia_origen <= 2) and (carrera = IET) => reprobo = si (16.0/0.0) [5.0/1.0]
- Except (edad_ingreso <= 19.5) and (promedio_notas = bueno) and (madre_trabaja = si) and (bienestar_servicios = si) and (carrera = IE) => reprobo = si (8.0/0.0) [3.0/0.0]
- Except (promedio_asistencia = medio) and (promedio_notas = regular) and (edad_ingreso > 18.5) => reprobo = si (32.0/9.0) [5.0/2.0]

- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (madre_trabaja = si) and (carrera = IE) and (promedio_asistencia = alto) and (edad_ingreso <= 18.5) and (distancia_origen <= 2.5) and (servicios = 3) => reprobo = si (2.0/0.0) [0.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 20.5) and (carrera = IS) and (promedio_asistencia = medio) and (padre_trabaja = si) and (distancia_origen <= 1.5) and (bienestar_servicios = si) => reprobo = si (7.0/1.0) [3.0/1.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (madre_trabaja = si) and (tipo_beca = nignuna) and (edad_ingreso <= 18.5) and (promedio_asistencia = alto) and (distancia_origen <= 2.5) and (horario_estudio = vespertino) and (edad_ingreso > 17.5) => reprobo = si (13.0/3.0) [6.0/2.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (carrera = IE) and (edad_ingreso > 18.5) and (distancia_origen <= 1.5) and (promedio_asistencia = medio) and (servicios = 1) => reprobo = si (11.0/3.0) [5.0/2.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (madre_trabaja = si) and (promedio_asistencia = medio) and (carrera = IS) and (distancia_origen > 2.5) and (servicios = 3) => reprobo = si (7.0/1.0) [1.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (promedio_asistencia = medio) and (carrera = IS) and (servicios = 1) and (edad_ingreso > 17.5) and (edad_ingreso <= 18.5) => reprobo = si (54.0/26.0) [14.0/7.0]
- Except (carrera = IE) and (promedio_notas = regular) => reprobo = si (13.0/1.0) [4.0/2.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 20.5) and (carrera = IE) and (distancia_origen > 1.5) and (servicios = 1) and (edad_ingreso <= 18.5) and (edad_ingreso <= 17.5) and (distancia_origen > 3.5) => reprobo = si (5.0/1.0) [0.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 20.5) and (carrera = IE) and (promedio_asistencia = bajo) and (distancia_origen <= 2.5) => reprobo = si (16.0/5.0) [6.0/5.0]
- Except (promedio_notas = bueno) and (madre_trabaja = si) and (edad_ingreso <= 21.5) and (distancia_origen > 3.5) and (edad_ingreso > 18.5) and (horario_estudio = matutino) and (edad_ingreso > 19.5) => reprobo = si (6.0/0.0) [2.0/2.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (padre_trabaja = si) and (edad_ingreso > 17.5) and (carrera = IE) and (edad_ingreso > 18.5) and

(servicios = 3) and (promedio_asistencia = medio) and (distancia_origen <= 2.5) => reprobo = si (3.0/0.0) [2.0/2.0]

- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (promedio_asistencia = alto) and (distancia_origen > 1.5) and (horario_estudio = matutino) and (servicios = 1) and (edad_ingreso <= 18.5) and (distancia_origen <= 3.5) => reprobo = si (5.0/0.0) [1.0/1.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (madre_trabaja = si) and (horario_estudio = vespertino) and (servicios = 1) and (tipo_beca = nignuna) and (estado_civil = casado) and (distancia_origen <= 2) => reprobo = si (3.0/0.0) [2.0/1.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (madre_trabaja = si) and (horario_estudio = vespertino) and (servicios = 1) and (distancia_origen > 1.5) and (edad_ingreso > 18.5) and (distancia_origen > 3.5) and (promedio_asistencia = alto) => reprobo = si (2.0/0.0) [0.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (carrera = IE) and (edad_ingreso > 18.5) and (distancia_origen > 2.5) and (servicios = 1) and (promedio_asistencia = medio) and (distancia_origen <= 3.5) => reprobo = si (2.0/0.0) [1.0/1.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (padre_trabaja = si) and (distancia_origen <= 2.5) and (edad_ingreso <= 18.5) and (carrera = IGAOT) => reprobo = si (18.0/8.0) [10.0/6.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (horario_estudio = vespertino) and (servicios = 1) and (distancia_origen > 3.5) and (edad_ingreso <= 18.5) and (bienestar_servicios = no) and (promedio_asistencia = medio) => reprobo = si (5.0/1.0) [5.0/4.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (tipo_beca = nignuna) and (padre_trabaja = si) and (estado_civil = soltero) and (servicios = 3) and (distancia_origen <= 2.5) and (distancia_origen > 1.5) and (promedio_asistencia = alto) => reprobo = si (2.0/0.0) [0.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 20.5) and (carrera = IE) and (edad_ingreso > 18.5) and (promedio_asistencia = alto) and (distancia_origen > 1.5) => reprobo = si (5.0/2.0) [1.0/0.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 20.5) and (tipo_beca = nignuna) and (padre_trabaja = si) and (edad_ingreso <= 18.5) and (distancia_origen

<= 2.5) and (carrera = IET) and (promedio_asistencia = medio) => reprobo = si (3.0/0.0) [1.0/1.0]

- Except (promedio_notas = bueno) and (estado_civil = soltero) and (padre_trabaja = si) and (edad_ingreso <= 19.5) and (edad_ingreso > 17.5) and (horario_estudio = vespertino) and (servicios = 3) and (edad_ingreso > 18.5) and (distancia_origen > 3.5) => reprobo = si (3.0/0.0) [1.0/1.0]
- Except (promedio_notas = bueno) and (edad_ingreso <= 19.5) and (horario_estudio = vespertino) and (padre_trabaja = si) and (promedio_asistencia = alto) and (bienestar_servicios = si) and (edad_ingreso > 18.5) => reprobo = si (3.0/1.0) [0.0/0.0]
- Except (promedio_notas = bueno) and (servicios = 1) and (edad_ingreso <= 20.5) and (tipo_beca = nignuna) and (carrera = IE) and (edad_ingreso > 17.5) and (promedio_asistencia = alto) and (edad_ingreso <= 18.5) => reprobo = si (3.0/1.0) [1.0/1.0]

Total number of rules (incl. the default rule): 36

ANEXO 9: Resultados de Predicción con estudiantes

A continuación se describen los resultados de predicción para cada estudiante que cursa una carrera en el Área de Energía de las Industrias y los Recursos Naturales no Renovables de la Universidad Nacional de Loja (ver tabla I).

TABLA I:
VALORES DE PREDICCIÓN.

Nro.	Cedula	Confidence (desertor)	Confidence (egresado)	Prediction (estado)	Carrera
1	104646815	0.941176471	0.058823529	desertor	IS
2	105220347	0	1	egresado	IS
3	603951815	0.965517241	0.034482759	desertor	IS
4	703745349	1	0	desertor	IET
5	704300821	0.965517241	0.034482759	desertor	IS
6	704405596	0.997315436	0.002684564	desertor	IS
7	704406271	0	1	egresado	IGAOT
8	704412923	0	1	egresado	IGAOT
9	704418144	0	1	egresado	IS
10	704634427	0	1	egresado	IET
11	704655646	0.965517241	0.034482759	desertor	IS
12	704674704	0.965517241	0.034482759	desertor	IS
13	704813476	0.997315436	0.002684564	desertor	IS
14	705006369	0	1	egresado	IE
15	705048106	0.965517241	0.034482759	desertor	IS
16	705104982	0.852941176	0.147058824	desertor	IE
17	705114049	0	1	egresado	IE
18	705183986	0.965517241	0.034482759	desertor	IS
19	705184588	0.907407407	0.092592593	desertor	IE
20	705225746	0.852941176	0.147058824	desertor	IE
21	705232742	0.997315436	0.002684564	desertor	IE
22	705281939	0.997315436	0.002684564	desertor	IE
23	705286615	0.857142857	0.142857143	desertor	IGAOT
24	705342939	0.928571429	0.071428571	desertor	IET
25	705358596	1	0	desertor	IE
26	705363869	1	0	desertor	IET
27	705379287	0.071428571	0.928571429	egresado	IS
28	705383487	0	1	egresado	IS
29	705470656	0	1	egresado	IS
30	705497162	0	1	egresado	IS
31	705497576	0.2	0.8	egresado	IET
32	705559508	0	1	egresado	IGAOT
33	705602316	0	1	egresado	IE

34	705620946	0.997315436	0.002684564	desertor	IE
35	705634590	0	1	egresado	IET
36	705642734	0	1	egresado	IE
37	705649796	0	1	egresado	IET
38	705650448	0	1	egresado	IET
39	705707123	0	1	egresado	IET
40	705712248	0	1	egresado	IE
41	705738474	0	1	egresado	IS
42	705743383	0.179487179	0.820512821	egresado	IE
43	705810620	0.852941176	0.147058824	desertor	IE
44	705821965	0.997315436	0.002684564	desertor	IE
45	705846707	0.166666667	0.833333333	egresado	IE
46	705869527	0.179487179	0.820512821	egresado	IS
47	705877611	0	1	egresado	IET
48	705889814	0.179487179	0.820512821	egresado	IS
49	705921575	0.4	0.6	egresado	IS
50	705925402	0.997315436	0.002684564	desertor	IET
51	705933588	0	1	egresado	IE
52	706012937	0.997315436	0.002684564	desertor	IE
53	706030863	0.997315436	0.002684564	desertor	IE
54	706053279	0.875	0.125	desertor	IS
55	706220530	0	1	egresado	IE
56	706246436	0	1	egresado	IS
57	706332285	0	1	egresado	IE
58	706337854	0.907407407	0.092592593	desertor	TECI
59	706339884	0	1	egresado	IS
60	706411022	0.997315436	0.002684564	desertor	IET
61	706440229	0	1	egresado	IS
62	706440567	0.179487179	0.820512821	egresado	IE
63	706440674	0.179487179	0.820512821	egresado	IET
64	706474210	0.4	0.6	egresado	IET
65	706544517	0	1	egresado	IET
66	706557097	1	0	desertor	IGAOT
67	706575081	0.941176471	0.058823529	desertor	IS
68	706575123	0.997315436	0.002684564	desertor	IET
69	706579299	1	0	desertor	IET
70	706613320	1	0	desertor	IE
71	706897956	0	1	egresado	IE
72	922970538	0	1	egresado	IET
73	923609051	0	1	egresado	IE
74	924499593	0.2	0.8	egresado	IET
75	1003886221	0.2	0.8	egresado	IGAOT

76	1101362844	1	0	desertor	IGAOT
77	1102843172	0	1	egresado	IET
78	1103123467	1	0	desertor	IGAOT
79	1103314959	0.941176471	0.058823529	desertor	IS
80	1103473631	0.857142857	0.142857143	desertor	IGAOT
81	1103499545	0	1	egresado	IS
82	1103504757	0	1	egresado	IET
83	1103535991	0.965517241	0.034482759	desertor	IS
84	1103570840	0.333333333	0.666666667	egresado	IS
85	1103681340	0.179487179	0.820512821	egresado	IE
86	1103720122	0.852941176	0.147058824	desertor	IE
87	1103752513	0	1	egresado	IET
88	1103781496	0.923076923	0.076923077	desertor	IS
89	1103805519	0.928571429	0.071428571	desertor	IET
90	1103841456	0	1	egresado	IS
91	1103844740	0	1	egresado	IE
92	1103848790	0	1	egresado	IE
93	1103863732	0	1	egresado	IET
94	1103881353	0	1	egresado	IGAOT
95	1103890115	0.997315436	0.002684564	desertor	IET
96	1103899652	0.4	0.6	egresado	IS
97	1103900880	0.965517241	0.034482759	desertor	IS
98	1103916720	1	0	desertor	IE
99	1103931638	0.857142857	0.142857143	desertor	IGAOT
100	1104009871	0	1	egresado	IS
101	1104013824	0.857142857	0.142857143	desertor	IGAOT
102	1104029556	0	1	egresado	IS
103	1104036270	0.125	0.875	egresado	IS
104	1104038342	0.965517241	0.034482759	desertor	IS
105	1104041445	0.941176471	0.058823529	desertor	IS
106	1104063415	0.923076923	0.076923077	desertor	IS
107	1104067036	0.857142857	0.142857143	desertor	IGAOT
108	1104069081	0.965517241	0.034482759	desertor	IS
109	1104070717	0	1	egresado	IET
110	1104073034	0	1	egresado	IE
111	1104074479	0.636363636	0.363636364	desertor	IS
112	1104090384	0	1	egresado	IE
113	1104102601	0	1	egresado	IET
114	1104102759	0	1	egresado	IE
115	1104102999	0.997315436	0.002684564	desertor	IS
116	1104105026	0.852941176	0.147058824	desertor	IE
117	1104106933	0.071428571	0.928571429	egresado	IGAOT

118	1104107105	0.997315436	0.002684564	desertor	IS
119	1104107741	0	1	egresado	IE
120	1104110240	0	1	egresado	IE
121	1104116817	0	1	egresado	IS
122	1104120918	0	1	egresado	IE
123	1104121346	0	1	egresado	IS
124	1104122922	0	1	egresado	IGAOT
125	1104124456	0.179487179	0.820512821	egresado	IS
126	1104125511	0.875	0.125	desertor	IS
127	1104126089	0	1	egresado	IE
128	1104127079	0.179487179	0.820512821	egresado	IS
129	1104128168	0.179487179	0.820512821	egresado	IS
130	1104128275	0.852941176	0.147058824	desertor	IE
131	1104136625	0	1	egresado	IS
132	1104136880	0.965517241	0.034482759	desertor	IS
133	1104141260	0.179487179	0.820512821	egresado	IET
134	1104143035	0.179487179	0.820512821	egresado	IE
135	1104148307	0	1	egresado	IE
136	1104151285	0.636363636	0.363636364	desertor	IS
137	1104166374	0.997315436	0.002684564	desertor	IET
138	1104175367	0	1	egresado	IE
139	1104178437	0.928571429	0.071428571	desertor	IET
140	1104182348	0	1	egresado	IS
141	1104187180	0.142857143	0.857142857	egresado	IE
142	1104191158	0.179487179	0.820512821	egresado	IE
143	1104194608	0.928571429	0.071428571	desertor	IET
144	1104203672	0.907407407	0.092592593	desertor	IS
145	1104208275	0.928571429	0.071428571	desertor	IET
146	1104218910	0.997315436	0.002684564	desertor	IS
147	1104220007	0.852941176	0.147058824	desertor	IE
148	1104233117	0.857142857	0.142857143	desertor	IGAOT
149	1104257579	0.965517241	0.034482759	desertor	IS
150	1104267941	0.997315436	0.002684564	desertor	IGAOT
151	1104268899	0.333333333	0.666666667	egresado	IS
152	1104275134	0.997315436	0.002684564	desertor	IET
153	1104282734	0	1	egresado	IS
154	1104290943	0.907407407	0.092592593	desertor	IE
155	1104308653	0	1	egresado	TECI
156	1104311061	0.785714286	0.214285714	desertor	IE
157	1104314693	0	1	egresado	IET
158	1104316425	0	1	egresado	IE
159	1104317472	0	1	egresado	IS

160	1104322084	0.965517241	0.034482759	desertor	IS
161	1104332836	0.071428571	0.928571429	egresado	IE
162	1104338700	0.907407407	0.092592593	desertor	IE
163	1104339021	0.907407407	0.092592593	desertor	IE
164	1104346612	0.166666667	0.833333333	egresado	IGAOT
165	1104356082	0.941176471	0.058823529	desertor	IS
166	1104366735	0	1	egresado	IE
167	1104384787	0.071428571	0.928571429	egresado	IGAOT
168	1104387509	0.965517241	0.034482759	desertor	IS
169	1104402035	0	1	egresado	IGAOT
170	1104410368	0.965517241	0.034482759	desertor	IS
171	1104419641	0.933333333	0.066666667	desertor	IE
172	1104435738	0.071428571	0.928571429	egresado	IGAOT
173	1104437973	0.857142857	0.142857143	desertor	IGAOT
174	1104444235	0.882352941	0.117647059	desertor	IE
175	1104449903	0	1	egresado	TECI
176	1104456023	0	1	egresado	IE
177	1104457062	0	1	egresado	TECI
178	1104460751	0.941176471	0.058823529	desertor	IS
179	1104464753	0.179487179	0.820512821	egresado	IET
180	1104472566	0.928571429	0.071428571	desertor	IET
181	1104474521	1	0	desertor	TECI
182	1104478472	0.916666667	0.083333333	desertor	IS
183	1104481559	0.997315436	0.002684564	desertor	IE
184	1104486129	0	1	egresado	IE
185	1104492507	0.907407407	0.092592593	desertor	IS
186	1104494602	0.997315436	0.002684564	desertor	IGAOT
187	1104495252	0.852941176	0.147058824	desertor	IE
188	1104506892	0	1	egresado	IS
189	1104516107	0.179487179	0.820512821	egresado	IE
190	1104523467	0.923076923	0.076923077	desertor	IS
191	1104535305	0.907407407	0.092592593	desertor	IS
192	1104538960	1	0	desertor	IE
193	1104551732	0.997315436	0.002684564	desertor	IGAOT
194	1104552813	0.4	0.6	egresado	IET
195	1104566805	1	0	desertor	IS
196	1104567118	0.852941176	0.147058824	desertor	IE
197	1104570237	0.907407407	0.092592593	desertor	IS
198	1104577984	0	1	egresado	IS
199	1104578552	0	1	egresado	IGAOT
200	1104583040	1	0	desertor	IS
201	1104584238	1	0	desertor	IGAOT

202	1104589849	0	1	egresado	TECI
203	1104593015	0.997315436	0.002684564	desertor	IS
204	1104598113	0	1	egresado	IS
205	1104602535	0.923076923	0.076923077	desertor	IS
206	1104606965	0	1	egresado	IS
207	1104607187	0.636363636	0.363636364	desertor	IS
208	1104607278	0	1	egresado	IET
209	1104610611	0.907407407	0.092592593	desertor	IE
210	1104612302	0	1	egresado	IE
211	1104617053	0	1	egresado	IS
212	1104618671	0.933333333	0.066666667	desertor	IE
213	1104624430	0.997315436	0.002684564	desertor	IE
214	1104631518	0	1	egresado	IET
215	1104635089	0.179487179	0.820512821	egresado	IS
216	1104637721	0.142857143	0.857142857	egresado	IE
217	1104642473	0.071428571	0.928571429	egresado	IE
218	1104648199	0.941176471	0.058823529	desertor	IS
219	1104648777	0	1	egresado	IGAOT
220	1104650187	0.071428571	0.928571429	egresado	IET
221	1104650856	0.933333333	0.066666667	desertor	IE
222	1104657133	0	1	egresado	IS
223	1104657885	0.997315436	0.002684564	desertor	IE
224	1104659386	0.636363636	0.363636364	desertor	IS
225	1104660939	0	1	egresado	IET
226	1104664659	1	0	desertor	IE
227	1104666779	0.965517241	0.034482759	desertor	IS
228	1104668759	0.875	0.125	desertor	IE
229	1104668817	0.997315436	0.002684564	desertor	IE
230	1104671456	0	1	egresado	IET
231	1104671738	0	1	egresado	IS
232	1104675085	0.916666667	0.083333333	desertor	IS
233	1104675705	0.857142857	0.142857143	desertor	IGAOT
234	1104676141	0.071428571	0.928571429	egresado	IE
235	1104676778	0	1	egresado	IE
236	1104676893	0	1	egresado	IET
237	1104676901	0	1	egresado	IGAOT
238	1104677099	0	1	egresado	IS
239	1104678790	0	1	egresado	IS
240	1104679434	0	1	egresado	IET
241	1104679871	0.875	0.125	desertor	IS
242	1104680150	0.928571429	0.071428571	desertor	IET
243	1104680192	0.882352941	0.117647059	desertor	IE

244	1104681489	0.997315436	0.002684564	desertor	IE
245	1104681752	0.916666667	0.083333333	desertor	IGAOT
246	1104682149	1	0	desertor	IGAOT
247	1104683121	0	1	egresado	IGAOT
248	1104683238	0.179487179	0.820512821	egresado	IE
249	1104686645	0.907407407	0.092592593	desertor	IE
250	1104690365	0	1	egresado	IET
251	1104695190	0.997315436	0.002684564	desertor	IS
252	1104696453	0.997315436	0.002684564	desertor	IET
253	1104697865	0	1	egresado	IS
254	1104699531	0.941176471	0.058823529	desertor	IS
255	1104701741	0	1	egresado	IGAOT
256	1104706468	0.965517241	0.034482759	desertor	IS
257	1104708738	0.179487179	0.820512821	egresado	IE
258	1104722028	0.997315436	0.002684564	desertor	IS
259	1104722119	1	0	desertor	IGAOT
260	1104723182	0.965517241	0.034482759	desertor	IS
261	1104724750	0.965517241	0.034482759	desertor	IS
262	1104727084	1	0	desertor	IS
263	1104729148	0.875	0.125	desertor	IS
264	1104731144	0	1	egresado	IET
265	1104732100	0.965517241	0.034482759	desertor	IS
266	1104732209	0.965517241	0.034482759	desertor	IS
267	1104733645	0.965517241	0.034482759	desertor	IS
268	1104736663	0.965517241	0.034482759	desertor	IS
269	1104736762	1	0	desertor	IS
270	1104742240	0	1	egresado	IE
271	1104744006	0.857142857	0.142857143	desertor	IGAOT
272	1104744436	0	1	egresado	IGAOT
273	1104745045	0.965517241	0.034482759	desertor	IS
274	1104747058	0.907407407	0.092592593	desertor	IET
275	1104747462	0.997315436	0.002684564	desertor	IGAOT
276	1104748676	0	1	egresado	IET
277	1104749336	0	1	egresado	IE
278	1104752082	0	1	egresado	IET
279	1104752934	0	1	egresado	IET
280	1104753445	0	1	egresado	IS
281	1104755366	0	1	egresado	IE
282	1104757529	0.179487179	0.820512821	egresado	IS
283	1104759665	0	1	egresado	IS
284	1104761422	0.625	0.375	desertor	TECI
285	1104763261	0.071428571	0.928571429	egresado	IE

286	1104763592	0.2	0.8	egresado	IS
287	1104764871	0	1	egresado	IE
288	1104766348	0	1	egresado	IS
289	1104771777	0.997315436	0.002684564	desertor	IS
290	1104779044	0.923076923	0.076923077	desertor	IS
291	1104781180	0.965517241	0.034482759	desertor	IS
292	1104781610	0.997315436	0.002684564	desertor	IE
293	1104785942	0.179487179	0.820512821	egresado	IE
294	1104789852	0.2	0.8	egresado	IGAOT
295	1104793680	0.997315436	0.002684564	desertor	IS
296	1104797608	0	1	egresado	IE
297	1104798796	0	1	egresado	IS
298	1104799208	0	1	egresado	IET
299	1104804321	0	1	egresado	IET
300	1104805567	0.071428571	0.928571429	egresado	IGAOT
301	1104805591	0	1	egresado	IS
302	1104805963	0.916666667	0.083333333	desertor	IGAOT
303	1104806011	0.125	0.875	egresado	IS
304	1104806102	0	1	egresado	IE
305	1104807175	0	1	egresado	IGAOT
306	1104808785	0.997315436	0.002684564	desertor	IS
307	1104810971	0.997315436	0.002684564	desertor	IS
308	1104811920	0	1	egresado	IS
309	1104812274	0.785714286	0.214285714	desertor	IE
310	1104814502	0.997315436	0.002684564	desertor	IE
311	1104815335	0.997315436	0.002684564	desertor	IS
312	1104816580	0.928571429	0.071428571	desertor	IET
313	1104820004	0.4	0.6	egresado	IE
314	1104830268	0.916666667	0.083333333	desertor	IGAOT
315	1104835945	0	1	egresado	IS
316	1104836687	1	0	desertor	IE
317	1104850100	0.875	0.125	desertor	IS
318	1104856644	0	1	egresado	IE
319	1104857444	1	0	desertor	IET
320	1104859747	1	0	desertor	IS
321	1104861438	0.923076923	0.076923077	desertor	IS
322	1104865330	0.965517241	0.034482759	desertor	IS
323	1104865348	0.941176471	0.058823529	desertor	IS
324	1104870744	0.997315436	0.002684564	desertor	IET
325	1104871882	0	1	egresado	IS
326	1104872815	0.636363636	0.363636364	desertor	IS
327	1104875008	0	1	egresado	IET

328	1104876949	0.852941176	0.147058824	desertor	IE
329	1104880008	0.997315436	0.002684564	desertor	IE
330	1104881188	0.179487179	0.820512821	egresado	IE
331	1104882525	0	1	egresado	IGAOT
332	1104885262	0.997315436	0.002684564	desertor	IET
333	1104885411	0.965517241	0.034482759	desertor	IS
334	1104886542	0	1	egresado	IE
335	1104887847	0.179487179	0.820512821	egresado	IS
336	1104888241	0.965517241	0.034482759	desertor	IS
337	1104891484	0.071428571	0.928571429	egresado	IGAOT
338	1104891757	0.2	0.8	egresado	IE
339	1104892342	0.907407407	0.092592593	desertor	IGAOT
340	1104892821	0.997315436	0.002684564	desertor	IS
341	1104892839	0	1	egresado	IET
342	1104893019	0.916666667	0.083333333	desertor	IS
343	1104893118	0	1	egresado	IGAOT
344	1104893175	0.179487179	0.820512821	egresado	IE
345	1104893795	0.965517241	0.034482759	desertor	IS
346	1104897697	0.997315436	0.002684564	desertor	IE
347	1104898513	0.907407407	0.092592593	desertor	IET
348	1104898745	0	1	egresado	IS
349	1104899719	0	1	egresado	IE
350	1104900020	0	1	egresado	IGAOT
351	1104900780	0	1	egresado	IS
352	1104901895	0.071428571	0.928571429	egresado	IS
353	1104902273	0	1	egresado	IET
354	1104902554	1	0	desertor	IE
355	1104904667	0.179487179	0.820512821	egresado	IS
356	1104905524	0.997315436	0.002684564	desertor	IE
357	1104908197	0.852941176	0.147058824	desertor	IE
358	1104908353	0	1	egresado	IET
359	1104909286	1	0	desertor	IS
360	1104909906	0.997315436	0.002684564	desertor	IET
361	1104911084	0.997315436	0.002684564	desertor	IS
362	1104911118	0.965517241	0.034482759	desertor	IS
363	1104912595	1	0	desertor	IE
364	1104915820	0.997315436	0.002684564	desertor	IS
365	1104926256	1	0	desertor	IS
366	1104934052	0.997315436	0.002684564	desertor	IE
367	1104935018	0.941176471	0.058823529	desertor	IS
368	1104935166	0.907407407	0.092592593	desertor	IE
369	1104943624	0	1	egresado	IS

370	1104943632	0	1	egresado	IS
371	1104951866	0.928571429	0.071428571	desertor	IET
372	1104951890	0.997315436	0.002684564	desertor	IS
373	1104952328	0.636363636	0.363636364	desertor	IS
374	1104952708	0.636363636	0.363636364	desertor	IS
375	1104956949	0	1	egresado	IE
376	1104957194	0.179487179	0.820512821	egresado	IE
377	1104959547	0.875	0.125	desertor	IS
378	1104961873	0	1	egresado	IGAOT
379	1104962822	0.923076923	0.076923077	desertor	IS
380	1104962939	0.933333333	0.066666667	desertor	IE
381	1104964034	0.179487179	0.820512821	egresado	IS
382	1104967144	0.997315436	0.002684564	desertor	IE
383	1104967805	0	1	egresado	IGAOT
384	1104969033	1	0	desertor	IS
385	1104969124	0	1	egresado	IE
386	1104972144	0.4	0.6	egresado	IS
387	1104978612	0.997315436	0.002684564	desertor	IE
388	1104980493	0.071428571	0.928571429	egresado	IE
389	1104987969	0.071428571	0.928571429	egresado	IS
390	1104990591	0.965517241	0.034482759	desertor	IS
391	1104990666	0.997315436	0.002684564	desertor	IS
392	1104991573	0.179487179	0.820512821	egresado	IS
393	1104992118	0	1	egresado	IS
394	1104993157	1	0	desertor	IS
395	1104998396	0.852941176	0.147058824	desertor	IE
396	1104999147	0.965517241	0.034482759	desertor	IS
397	1104999634	0.907407407	0.092592593	desertor	IET
398	1105001646	1	0	desertor	IE
399	1105002180	0	1	egresado	IGAOT
400	1105003840	0	1	egresado	IS
401	1105004111	0.916666667	0.083333333	desertor	IGAOT
402	1105005480	0	1	egresado	IE
403	1105005522	0.997315436	0.002684564	desertor	IS
404	1105007353	0	1	egresado	IGAOT
405	1105007452	0.852941176	0.147058824	desertor	IE
406	1105007890	0	1	egresado	IS
407	1105009466	0.179487179	0.820512821	egresado	IE
408	1105012866	0	1	egresado	IS
409	1105014201	0	1	egresado	IE
410	1105021131	0	1	egresado	IET
411	1105023863	0.907407407	0.092592593	desertor	IS

412	1105026684	0.852941176	0.147058824	desertor	IE
413	1105027054	0.4	0.6	egresado	IS
414	1105027187	0.941176471	0.058823529	desertor	IS
415	1105027203	0.4	0.6	egresado	IET
416	1105028185	0.875	0.125	desertor	IS
417	1105028870	0.636363636	0.363636364	desertor	IS
418	1105029084	0	1	egresado	IET
419	1105029373	0.997315436	0.002684564	desertor	IE
420	1105031288	0.179487179	0.820512821	egresado	IE
421	1105032161	0.071428571	0.928571429	egresado	IS
422	1105032369	0.997315436	0.002684564	desertor	IGAOT
423	1105033557	0.907407407	0.092592593	desertor	IE
424	1105035396	0	1	egresado	IGAOT
425	1105035941	0.071428571	0.928571429	egresado	IS
426	1105039273	0.071428571	0.928571429	egresado	IE
427	1105041386	0.179487179	0.820512821	egresado	IS
428	1105042889	0.875	0.125	desertor	IS
429	1105044158	0	1	egresado	IS
430	1105046245	1	0	desertor	IS
431	1105047615	0.997315436	0.002684564	desertor	IE
432	1105048415	0	1	egresado	IS
433	1105048506	0.071428571	0.928571429	egresado	IS
434	1105050411	0	1	egresado	IGAOT
435	1105050775	0	1	egresado	IS
436	1105050890	0.125	0.875	egresado	IS
437	1105050981	0.997315436	0.002684564	desertor	IET
438	1105052755	0.997315436	0.002684564	desertor	IS
439	1105060535	0.997315436	0.002684564	desertor	IGAOT
440	1105067423	0.997315436	0.002684564	desertor	IE
441	1105078925	0.997315436	0.002684564	desertor	IS
442	1105081960	1	0	desertor	IE
443	1105086191	0	1	egresado	IET
444	1105097610	0.997315436	0.002684564	desertor	IGAOT
445	1105099277	0	1	egresado	TECI
446	1105104382	0	1	egresado	IE
447	1105104838	0	1	egresado	IET
448	1105106064	0	1	egresado	IS
449	1105109068	0	1	egresado	IET
450	1105112476	0.166666667	0.833333333	egresado	IE
451	1105113458	0	1	egresado	IS
452	1105114050	0.941176471	0.058823529	desertor	IS
453	1105116493	1	0	desertor	IGAOT

454	1105116733	0.179487179	0.820512821	egresado	IE
455	1105119851	0.179487179	0.820512821	egresado	IET
456	1105120289	0	1	egresado	IE
457	1105127508	0.916666667	0.083333333	desertor	IS
458	1105136129	0	1	egresado	IET
459	1105137143	0.923076923	0.076923077	desertor	IS
460	1105139248	0.965517241	0.034482759	desertor	IS
461	1105140105	0.997315436	0.002684564	desertor	IGAOT
462	1105145526	0	1	egresado	IGAOT
463	1105145708	0.907407407	0.092592593	desertor	IE
464	1105145773	1	0	desertor	IET
465	1105146862	0.179487179	0.820512821	egresado	IE
466	1105147944	0	1	egresado	IGAOT
467	1105148447	0	1	egresado	IGAOT
468	1105148942	0	1	egresado	IS
469	1105150393	0	1	egresado	IET
470	1105151185	0.997315436	0.002684564	desertor	IS
471	1105151888	0.4	0.6	egresado	IET
472	1105153454	0	1	egresado	IGAOT
473	1105153777	0	1	egresado	IE
474	1105154502	0	1	egresado	IS
475	1105154635	0.933333333	0.066666667	desertor	IE
476	1105156655	0	1	egresado	IS
477	1105156705	0.4	0.6	egresado	IGAOT
478	1105157075	0.636363636	0.363636364	desertor	IS
479	1105157463	0	1	egresado	IGAOT
480	1105157737	0	1	egresado	IE
481	1105158073	0	1	egresado	IGAOT
482	1105158784	0.179487179	0.820512821	egresado	IGAOT
483	1105159022	0.923076923	0.076923077	desertor	IS
484	1105159477	0.166666667	0.833333333	egresado	IS
485	1105159733	0	1	egresado	IS
486	1105161861	0	1	egresado	IS
487	1105162471	0.928571429	0.071428571	desertor	IET
488	1105163875	0.179487179	0.820512821	egresado	IS
489	1105165706	0.179487179	0.820512821	egresado	IET
490	1105166050	0	1	egresado	IGAOT
491	1105166068	0	1	egresado	IE
492	1105167199	0.923076923	0.076923077	desertor	IS
493	1105167496	0.997315436	0.002684564	desertor	IGAOT
494	1105167579	0	1	egresado	IET
495	1105168338	0.916666667	0.083333333	desertor	IGAOT

496	1105169278	0.997315436	0.002684564	desertor	IET
497	1105170755	0	1	egresado	IGAOT
498	1105171415	0.997315436	0.002684564	desertor	IS
499	1105171589	0	1	egresado	IET
500	1105172488	0.941176471	0.058823529	desertor	IS
501	1105175093	0.997315436	0.002684564	desertor	IE
502	1105175101	0.941176471	0.058823529	desertor	IS
503	1105180994	0.965517241	0.034482759	desertor	IS
504	1105182974	0.941176471	0.058823529	desertor	IS
505	1105183782	0	1	egresado	IGAOT
506	1105191645	0	1	egresado	IET
507	1105195976	0.997315436	0.002684564	desertor	IE
508	1105196412	0	1	egresado	IE
509	1105203358	0.916666667	0.083333333	desertor	IS
510	1105206088	0	1	egresado	IS
511	1105208969	0	1	egresado	IS
512	1105210742	0.179487179	0.820512821	egresado	IE
513	1105218273	0.875	0.125	desertor	IS
514	1105218281	0.875	0.125	desertor	IGAOT
515	1105219859	0	1	egresado	IE
516	1105220204	1	0	desertor	IET
517	1105224040	0.875	0.125	desertor	IS
518	1105228447	0.179487179	0.820512821	egresado	IE
519	1105228728	0.928571429	0.071428571	desertor	IET
520	1105228892	0.997315436	0.002684564	desertor	IET
521	1105229148	0	1	egresado	IET
522	1105233959	1	0	desertor	IS
523	1105235616	0	1	egresado	IS
524	1105242075	0	1	egresado	IE
525	1105260325	0.857142857	0.142857143	desertor	IGAOT
526	1105310526	0	1	egresado	IET
527	1105321770	0.179487179	0.820512821	egresado	IE
528	1105322257	0.997315436	0.002684564	desertor	IE
529	1105322596	0.4	0.6	egresado	IS
530	1105324899	0	1	egresado	IE
531	1105326886	0.997315436	0.002684564	desertor	IET
532	1105329922	0.965517241	0.034482759	desertor	IS
533	1105330268	0.997315436	0.002684564	desertor	IET
534	1105332256	0.997315436	0.002684564	desertor	IS
535	1105333734	0	1	egresado	IS
536	1105333908	0.941176471	0.058823529	desertor	IS
537	1105339640	0.997315436	0.002684564	desertor	IS

538	1105339848	0.941176471	0.058823529	desertor	IS
539	1105339947	0	1	egresado	IE
540	1105355679	0.997315436	0.002684564	desertor	IE
541	1105362063	0.179487179	0.820512821	egresado	IET
542	1105364077	1	0	desertor	IE
543	1105381063	0	1	egresado	IET
544	1105403701	0	1	egresado	IS
545	1105432163	0.179487179	0.820512821	egresado	IS
546	1105439267	0.997315436	0.002684564	desertor	IET
547	1105442261	0	1	egresado	IET
548	1105442311	0	1	egresado	IE
549	1105532707	0	1	egresado	IE
550	1105535205	0	1	egresado	IE
551	1105537532	0.25	0.75	egresado	IS
552	1105544967	0	1	egresado	IET
553	1105550600	0	1	egresado	IGAOT
554	1105572984	0.907407407	0.092592593	desertor	IE
555	1105573719	0.4	0.6	egresado	IE
556	1105574154	0	1	egresado	IS
557	1105578189	0.882352941	0.117647059	desertor	IE
558	1105578437	0.636363636	0.363636364	desertor	IS
559	1105578585	1	0	desertor	IET
560	1105579427	0.997315436	0.002684564	desertor	IE
561	1105580060	0.179487179	0.820512821	egresado	IE
562	1105580755	0	1	egresado	IET
563	1105581316	1	0	desertor	IS
564	1105582827	0	1	egresado	IS
565	1105582843	0.997315436	0.002684564	desertor	IE
566	1105585507	0.997315436	0.002684564	desertor	IS
567	1105589483	0.875	0.125	desertor	IET
568	1105593238	0	1	egresado	IS
569	1105594541	0.852941176	0.147058824	desertor	IE
570	1105600736	1	0	desertor	IS
571	1105604886	0	1	egresado	IET
572	1105614570	1	0	desertor	IS
573	1105632101	0	1	egresado	IE
574	1105632432	0.997315436	0.002684564	desertor	IS
575	1105637803	0	1	egresado	IET
576	1105638769	0	1	egresado	IGAOT
577	1105641151	0	1	egresado	IGAOT
578	1105645178	0.636363636	0.363636364	desertor	IS
579	1105647133	0.875	0.125	desertor	IGAOT

580	1105647497	0	1	egresado	IE
581	1105648016	0.875	0.125	desertor	IE
582	1105649758	0	1	egresado	IS
583	1105652190	1	0	desertor	IS
584	1105653792	0	1	egresado	IET
585	1105656712	0.997315436	0.002684564	desertor	IS
586	1105657645	0	1	egresado	IET
587	1105658213	0	1	egresado	IGAOT
588	1105661415	0.997315436	0.002684564	desertor	IET
589	1105663064	0	1	egresado	IE
590	1105665416	0	1	egresado	IGAOT
591	1105666471	0	1	egresado	IE
592	1105666836	0	1	egresado	IET
593	1105668485	0	1	egresado	IS
594	1105669186	0.907407407	0.092592593	desertor	IE
595	1105670879	0	1	egresado	IGAOT
596	1105675860	0.933333333	0.066666667	desertor	IE
597	1105676819	0.997315436	0.002684564	desertor	IS
598	1105679151	1	0	desertor	IET
599	1105679664	0.875	0.125	desertor	IS
600	1105680894	0.997315436	0.002684564	desertor	IS
601	1105681165	0	1	egresado	IGAOT
602	1105681710	0	1	egresado	IET
603	1105681769	0	1	egresado	IET
604	1105686008	0.179487179	0.820512821	egresado	IGAOT
605	1105687030	0.997315436	0.002684564	desertor	IE
606	1105688913	0	1	egresado	IE
607	1105691149	0.179487179	0.820512821	egresado	IS
608	1105692477	0	1	egresado	IGAOT
609	1105701724	0.997315436	0.002684564	desertor	IE
610	1105702946	0	1	egresado	IE
611	1105707499	0.997315436	0.002684564	desertor	IS
612	1105708067	0.997315436	0.002684564	desertor	IS
613	1105711624	0.997315436	0.002684564	desertor	IGAOT
614	1105728305	0.997315436	0.002684564	desertor	IE
615	1105739435	0.997315436	0.002684564	desertor	IS
616	1105763591	0	1	egresado	IET
617	1105764805	0.179487179	0.820512821	egresado	IE
618	1105766263	0.997315436	0.002684564	desertor	IE
619	1105772261	0.997315436	0.002684564	desertor	IS
620	1105786766	0.179487179	0.820512821	egresado	IS
621	1105794638	0	1	egresado	IE

622	1105804452	0.997315436	0.002684564	desertor	IS
623	1105822793	0	1	egresado	IET
624	1105826760	0.4	0.6	egresado	IET
625	1105832719	0.997315436	0.002684564	desertor	IET
626	1105836843	0.965517241	0.034482759	desertor	IS
627	1105844078	0.179487179	0.820512821	egresado	IS
628	1105852220	0	1	egresado	IGAOT
629	1105871501	0	1	egresado	IE
630	1105872897	0.997315436	0.002684564	desertor	IE
631	1105874422	0	1	egresado	IS
632	1105874851	0.179487179	0.820512821	egresado	IS
633	1105877250	0	1	egresado	IE
634	1105883894	0	1	egresado	IGAOT
635	1105884934	1	0	desertor	IE
636	1105886707	0	1	egresado	IET
637	1105888950	0	1	egresado	IET
638	1105889362	0.25	0.75	egresado	IGAOT
639	1105891814	0	1	egresado	IS
640	1105893059	0.166666667	0.833333333	egresado	IS
641	1105896805	0.928571429	0.071428571	desertor	IET
642	1105903437	0.997315436	0.002684564	desertor	IS
643	1105906299	0	1	egresado	IS
644	1105907230	0	1	egresado	IE
645	1105910044	0	1	egresado	IS
646	1105932097	0	1	egresado	IET
647	1105962904	0.997315436	0.002684564	desertor	IS
648	1105968141	0.179487179	0.820512821	egresado	IS
649	1105977019	0.875	0.125	desertor	IS
650	1105984254	0.997315436	0.002684564	desertor	IE
651	1105987919	1	0	desertor	IS
652	1105993065	1	0	desertor	IE
653	1105993701	0.997315436	0.002684564	desertor	IET
654	1105999369	0.875	0.125	desertor	IS
655	1106000852	0	1	egresado	IGAOT
656	1106000969	0.4	0.6	egresado	IET
657	1106008509	0	1	egresado	IET
658	1106015439	0	1	egresado	IET
659	1106015553	0.997315436	0.002684564	desertor	IE
660	1106020397	0	1	egresado	IET
661	1106021783	0	1	egresado	IET
662	1106032491	0	1	egresado	IET
663	1106043456	0	1	egresado	IGAOT

664	1106049289	0.997315436	0.002684564	desertor	IET
665	1106052481	0.997315436	0.002684564	desertor	IE
666	1106075730	0	1	egresado	IGAOT
667	1106079971	0	1	egresado	IET
668	1106085895	0.997315436	0.002684564	desertor	IE
669	1150009742	1	0	desertor	IE
670	1150019493	0	1	egresado	IGAOT
671	1150022406	0.997315436	0.002684564	desertor	IS
672	1150022570	0.997315436	0.002684564	desertor	IS
673	1150025110	0.997315436	0.002684564	desertor	IS
674	1150027504	0.997315436	0.002684564	desertor	IS
675	1150028882	0	1	egresado	IGAOT
676	1150029005	0	1	egresado	IGAOT
677	1150032603	0.997315436	0.002684564	desertor	IS
678	1150035648	0	1	egresado	IE
679	1150043444	0	1	egresado	IET
680	1150130480	0	1	egresado	IGAOT
681	1150130753	0	1	egresado	IE
682	1150131066	1	0	desertor	IS
683	1150137642	0.997315436	0.002684564	desertor	IS
684	1150150603	0	1	egresado	IS
685	1150178703	0.997315436	0.002684564	desertor	IS
686	1311077612	0.997315436	0.002684564	desertor	IGAOT
687	1312554619	0	1	egresado	IE
688	1400759237	0.997315436	0.002684564	desertor	IS
689	1400846208	1	0	desertor	IS
690	1600342925	0.997315436	0.002684564	desertor	IE
691	1600652380	0.4	0.6	egresado	IET
692	1717213183	0	1	egresado	IS
693	1718486689	1	0	desertor	IS
694	1718562919	0.941176471	0.058823529	desertor	IS
695	1718592171	0.997315436	0.002684564	desertor	IE
696	1719061523	0.928571429	0.071428571	desertor	IET
697	1719423830	0.997315436	0.002684564	desertor	IS
698	1719695494	0	1	egresado	IET
699	1720612868	0.166666667	0.833333333	egresado	IET
700	1720958345	0.923076923	0.076923077	desertor	IS
701	1720991817	0	1	egresado	IET
702	1722036199	0	1	egresado	IET
703	1722223755	0	1	egresado	IS
704	1722388434	0	1	egresado	IET
705	1722773064	0.179487179	0.820512821	egresado	IS

706	1723310080	0	1	egresado	IE
707	1724036890	0.941176471	0.058823529	desertor	IS
708	1724102429	0.179487179	0.820512821	egresado	IS
709	1724595242	0.179487179	0.820512821	egresado	IE
710	1725087637	0.997315436	0.002684564	desertor	IS
711	1725503872	0.916666667	0.083333333	desertor	IS
712	1726652397	0	1	egresado	IE
713	1726793969	0.997315436	0.002684564	desertor	IS
714	1727000943	0.179487179	0.820512821	egresado	IS
715	1803694973	1	0	desertor	IE
716	1900406594	0.25	0.75	egresado	IGAOT
717	1900415173	0	1	egresado	IET
718	1900481878	0	1	egresado	IS
719	1900482363	0	1	egresado	IGAOT
720	1900482884	0.852941176	0.147058824	desertor	IE
721	1900488832	0.997315436	0.002684564	desertor	IE
722	1900492214	0.997315436	0.002684564	desertor	IGAOT
723	1900503804	0.997315436	0.002684564	desertor	IE
724	1900516970	0.785714286	0.214285714	desertor	IE
725	1900519073	1	0	desertor	IE
726	1900523117	0	1	egresado	IET
727	1900524537	0	1	egresado	IE
728	1900535160	0	1	egresado	IE
729	1900536770	0.997315436	0.002684564	desertor	IGAOT
730	1900537059	0.941176471	0.058823529	desertor	IS
731	1900553817	0	1	egresado	IET
732	1900561323	0	1	egresado	IS
733	1900583327	0.997315436	0.002684564	desertor	IET
734	1900583335	0	1	egresado	IE
735	1900597830	0	1	egresado	IET
736	1900599190	0.071428571	0.928571429	egresado	IE
737	1900606656	0.997315436	0.002684564	desertor	IS
738	1900614064	0.923076923	0.076923077	desertor	IS
739	1900615491	0.179487179	0.820512821	egresado	IE
740	1900615608	0.907407407	0.092592593	desertor	IE
741	1900617075	0.907407407	0.092592593	desertor	IET
742	1900623685	1	0	desertor	IS
743	1900631761	0.4	0.6	egresado	IS
744	1900635200	0	1	egresado	IE
745	1900636828	0.997315436	0.002684564	desertor	IS
746	1900644236	0	1	egresado	IET
747	1900649128	0	1	egresado	IGAOT

748	1900653658	0.071428571	0.928571429	egresado	IS
749	1900674506	0.997315436	0.002684564	desertor	IE
750	1900683424	0.179487179	0.820512821	egresado	IE
751	1900708858	0	1	egresado	IE
752	1900717255	0	1	egresado	IE
753	1900745942	0.852941176	0.147058824	desertor	IE
754	1900763887	0.997315436	0.002684564	desertor	IS
755	1900766732	0.928571429	0.071428571	desertor	IET
756	1900787910	0	1	egresado	IS
757	1900790054	0.4	0.6	egresado	IS
758	1900801521	0.997315436	0.002684564	desertor	IS
759	1900802396	0	1	egresado	IET
760	1900840016	0	1	egresado	IGAOT
761	2000080289	0.875	0.125	desertor	IE
762	2100571609	0.997315436	0.002684564	desertor	IE
763	2100693239	0	1	egresado	IGAOT
764	2101126676	0.875	0.125	desertor	IET
765	2200081392	0	1	egresado	IGAOT
766	2200206262	0	1	egresado	IET
767	2300263445	0.875	0.125	desertor	IGAOT

- **Análisis de Resultados**

En base a los resultados de predicción que se describen en la tabla anterior, a continuación se describe la cantidad de estudiantes por Carreras con los cuales se trabajó (ver figura 1).

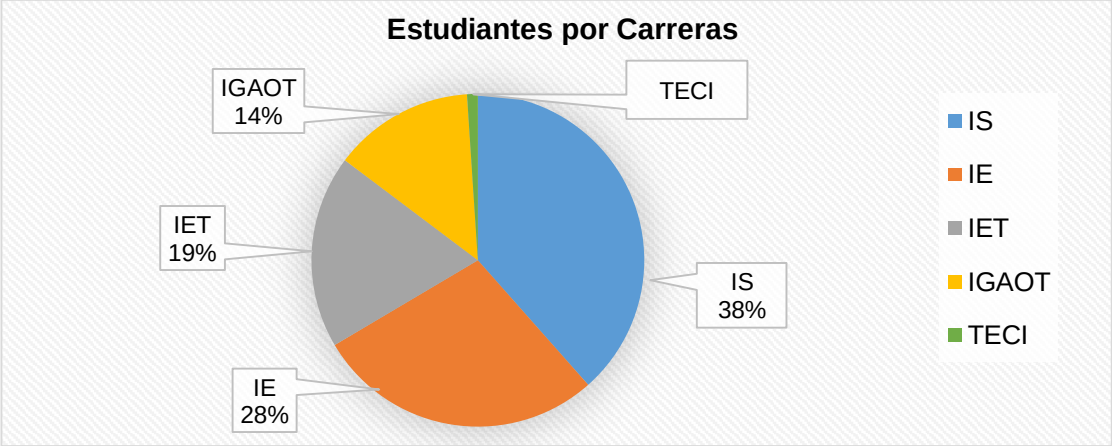


Figura 1: Distribución de estudiantes por carreras.

En la figura anterior se puede observar que de los 767 estudiantes, la carrera que mayor estudiantes tiene es Ingeniería en Sistemas con 38% de estudiantes (295), la carrera Ingeniería en Electromecánica tiene el 28% (215), la carrera de Ingeniería en Electrónica y Telecomunicaciones tiene un 19% (144), la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial tiene un 14% (105) y por último la carrera de Tecnología en Electricidad y Control Industrial tiene el 1% (8).

A continuación se describen los estudiantes con posibilidades de desertar en cada una de las carreras del Área de Energía según los resultados de predicción (ver figura 2).

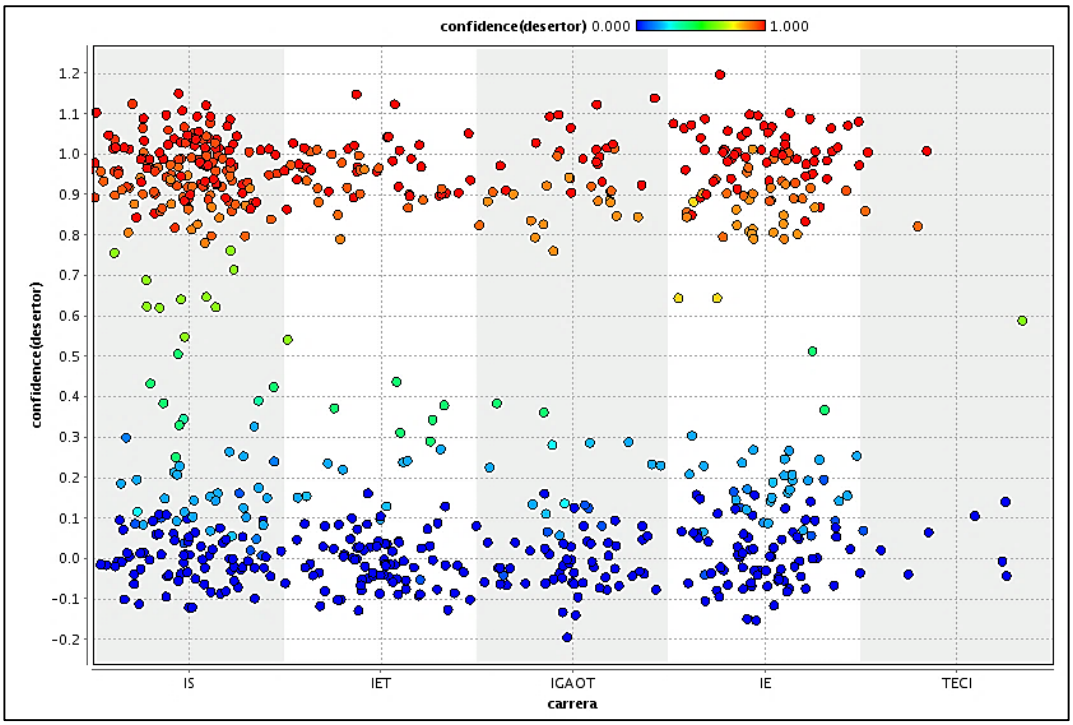


Figura 2: Distribución de estudiantes desertores por cada carrera.

La figura anterior se puede observar una considerable cantidad de estudiantes con probabilidades de desertar o abandonar los estudios (color rojo, amarillo y naranja), con respecto a las carreras que presentan la menor cantidad de estudiantes con probabilidades de desertar menor al 20% (color azul) está la Tecnología en Electricidad y Control Industrial (TECI) y la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial (IGAOT).

A continuación se describe la cantidad de estudiantes desertores por cada carrera del Área de la Energía de las Industrias y los Recursos Naturales No Renovables según los resultados de predicción (ver figura 3).

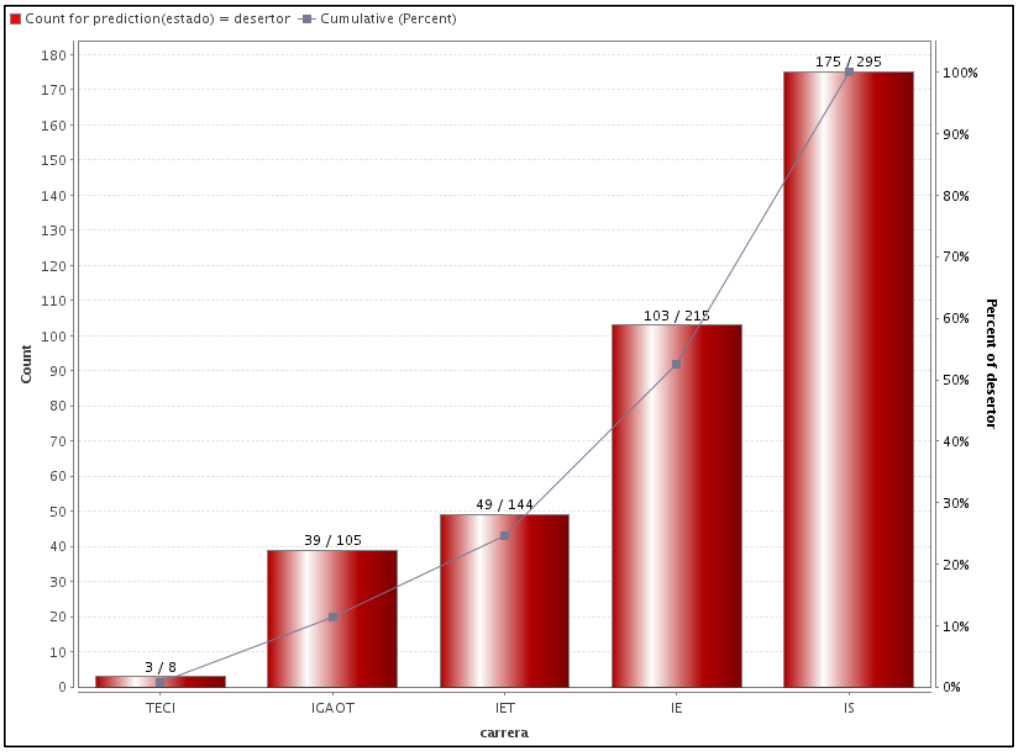


Figura 3: Cantidad de estudiantes desertores por cada carrera.

La figura describe que de los 767 estudiantes del Área de Energía 369 estudiantes presentan posibilidades de abandonar los estudios, es decir el 48.1%, la carrera que presenta mayor cantidad de estudiantes desertores es la carrera de Ingeniería en Sistemas (IS) con 175 estudiantes es decir el 47.43%, luego está la carrera de Ingeniería en Electromecánica (IE) con 103 estudiantes 27.91%, la carrera de Ingeniería en Electrónica y Telecomunicaciones (IET) con 49 estudiantes que corresponde al 13.28%, la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial (IGAOT) tiene 39 estudiantes que corresponde al 10.57% y por último la Tecnología en Electricidad y Control Industrial (TECI) que tiene 3 estudiantes que corresponde al 0.81%.

A continuación se describen los estudiantes con posibilidades de egresar en cada una de las carreras del Área de Energía las Industrias y los Recursos Naturales No Renovables según los resultados de predicción (ver figura 4).

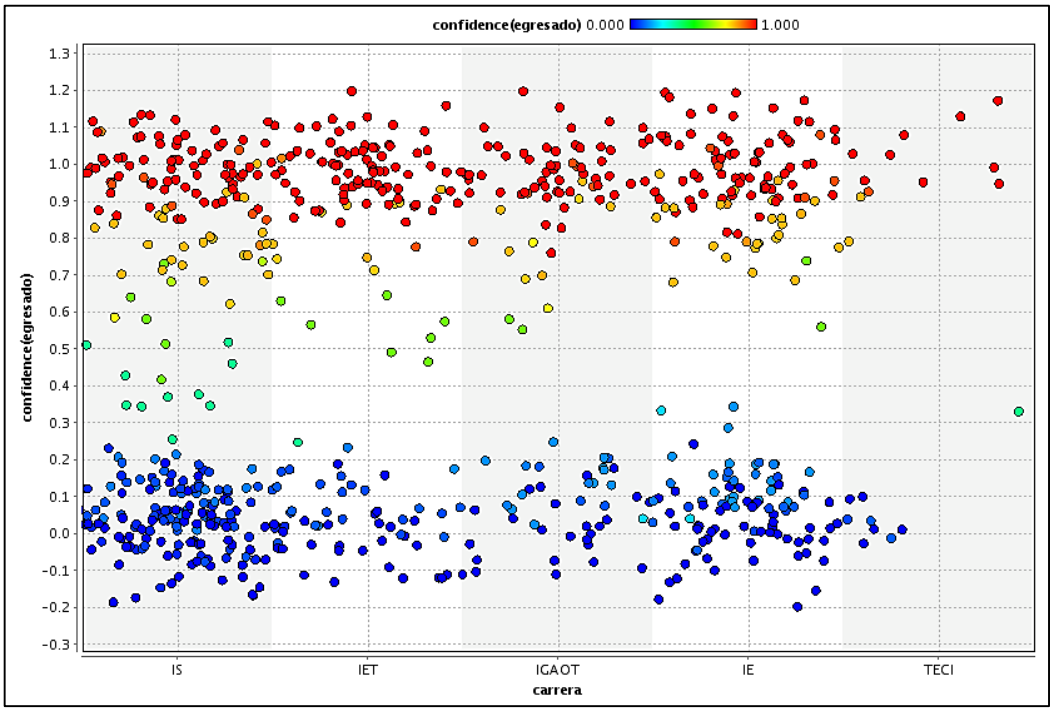


Figura 4: Distribución de estudiantes con posibilidades de egresar por cada carrera.

La figura anterior describe una considerable cantidad de estudiantes con el 70% probabilidades de egresar o terminar los estudios, con respecto a las carreras que presentan la menor cantidad de estudiantes con probabilidades de egresar esta la carrera Tecnología en Electricidad y Control Industrial (TECI) y la carrera Ingeniería en Electrónica y Telecomunicaciones (IET).

A continuación se describe la cantidad de estudiantes con probabilidades de egresar por cada carrera del Área de la Energía de las Industrias y los Recursos Naturales No Renovables, según los resultados de predicción (ver figura 5).

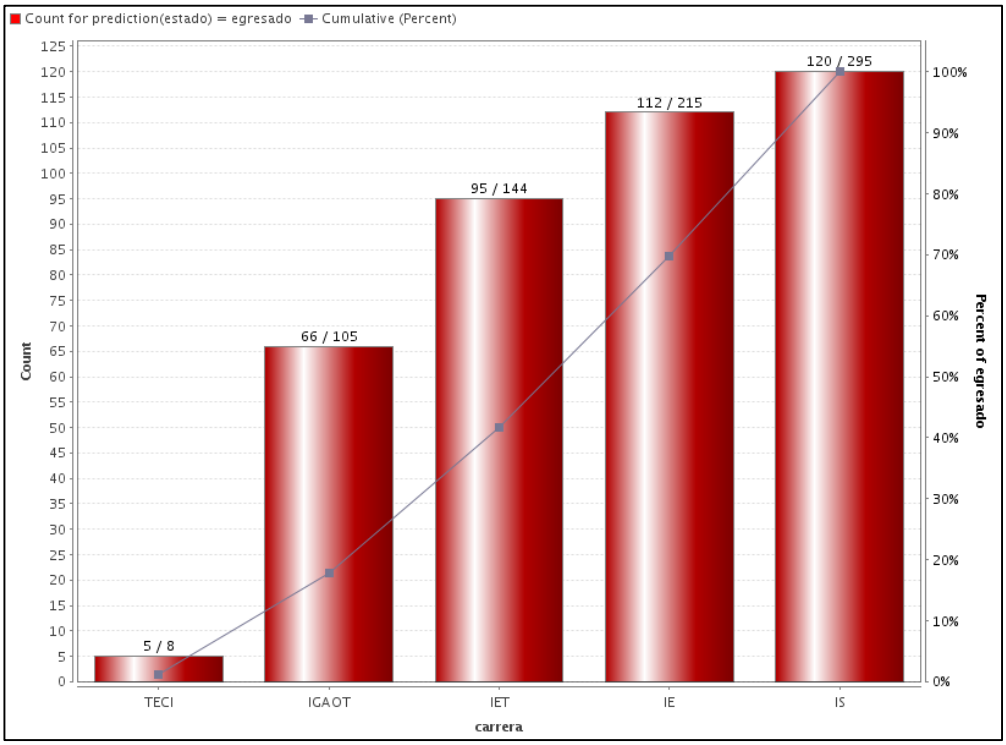


Figura 5: Cantidad de estudiantes con posibilidades de egresar por cada carrera.

La figura describe que de los 767 estudiantes del Área de Energía 398 estudiantes presentan posibilidades de egresar o terminar los estudios, es decir el 51.9%, mientras las carreras presentan los siguientes resultados: Ingeniería en Sistemas (IS) con 120, que corresponde al 30.15%, la carrera de Ingeniería en Electromecánica (IE) con 112, que corresponde al 28.14% y la carrera de Ingeniería en Electrónica y Telecomunicaciones (IET) con 95, que corresponde el 23.87%, estas carreras antes mencionadas presentan valores cercanos sin mucha diferencia mientras que la carrera de Ingeniería en Geología Ambiental y Ordenamiento Territorial (IGAOT) tiene 66 estudiantes que corresponde al 16.58% y la carrera de Tecnología en Electricidad y Control Industrial (TECI) tiene 5 estudiantes que corresponde al 1.26%.

A continuación se describe la cantidad de estudiantes con posibilidades de abandonar los estudios y de estudiantes con posibilidades de terminar los estudios, según los resultados de predicción (ver figura 6).

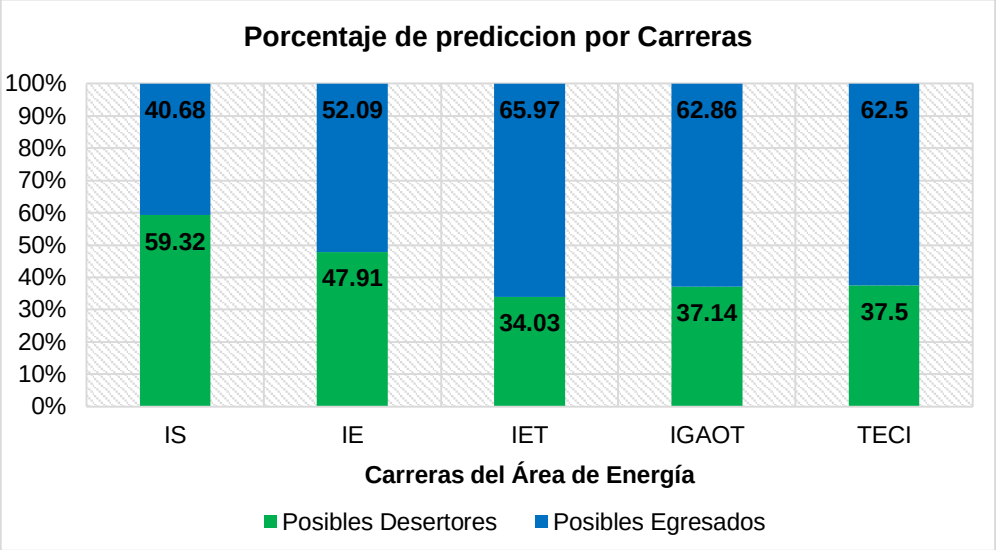


Figura 6: Distribución de estudiantes desertores y egresados por cada carrera.

Como se puede observar en la figura anterior, la carrera que supera el 50% de estudiantes con posibilidades de desertar es la carrera Ingeniería en Sistemas (IS) con el 59.32% mientras que las otras carreras tiene un porcentaje menor al 48%, como es la carrera de Ingeniería en Electromecánica (IE) con un porcentaje de 47.91% de estudiantes en peligro de desertar y las de Ingeniería en Electrónica y Telecomunicaciones (IET), Ingeniería en Geología Ambiental y Ordenamiento Territorial (IGAOT) y la Tecnología en Electricidad y Control Industrial (TECI) no superan el 38%.

ANEXO 10: Informe Ejecutivo

UNIVERSIDAD NACIONAL DE LOJA

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

CARRERA DE INGENIERÍA EN SISTEMAS

"Informe Ejecutivo"

Proyecto:

"Identificación de Factores en la Reprobación y Deserción mediante técnicas de Minería de Datos en el Área de la Energía de la Universidad Nacional de Loja"

Institución:

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

Autor:

- González Pineda, Anibal Israel

LOJA-ECUADOR
2014

Introducción:

El presente trabajo se centra en identificar los factores que inciden directa o parcialmente en el abandono y reprobación de los estudiantes que cursan las carreras del Área de Energía, las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja.

Para ello es necesaria la identificación de estos factores que obstruyen una adecuada formación académica, resultando en algunos casos que abandonen los estudios. Con el fin de identificar estos factores, se consideró indispensable el utilizar información almacenada por la institución como: datos personales, servicios brindados a estudiantes y notas académicas, esta información es obtenida en el transcurso que el estudiante estuvo matriculado en alguna carrera.

Problemática:

Actualmente el abandono y reprobación de los estudios es un problema que afronta el Área de Energía las Industrias y los Recursos Naturales No Renovables, es por ello que se han empleado varios esfuerzos que permitan al estudiante aumentar su compromiso académico y por consiguiente evitar que repruebe o abandone los estudios, sin embargo el problema aún persiste en cada período académico (ver figura 1).

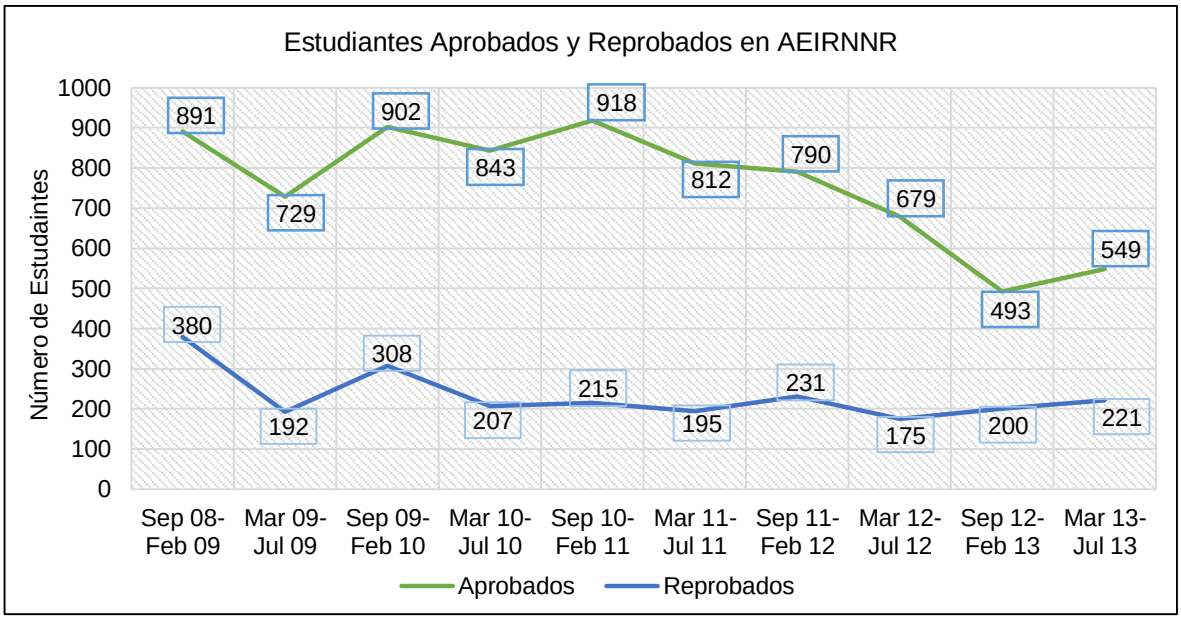


Figura 81: Estudiantes Aprobados y Reprobados en el Área de Energía de la Universidad Nacional de Loja

Por ello es importante abordar el problema, utilizando información asociada a los estudiantes y sus características, para descubrir información oculta en los mismos que

describirán situaciones o comportamientos; identificando además los factores que indicen en la deserción y reprobación de los estudiantes del Área de Energía, las Industrias y los Recursos Naturales No Renovables.

Objetivos del Negocio:

Los objetivos del negocio definidos en el presente trabajo son:

- Conocer los factores que determinan que un grupo de sus estudiantes, reprueben o abandonen sus estudios.
- Identificar los patrones de comportamiento de estudiantes abandonan los estudios.

Resultados y Aporte:

Por último se analizan los resultados obtenidos de la presente investigación realizada, enfocada en identificar los factores de deserción y reprobación en el Área de Energía, las Industrias y los Recursos Naturales No Renovables.

Los principales resultados encontrados al culminar el presente trabajo son:

- En el Área de Energía las Industrias y los Recursos Naturales No Renovables, en cada período académico en promedio el 23% de los estudiantes reprueban (ver figura 2).

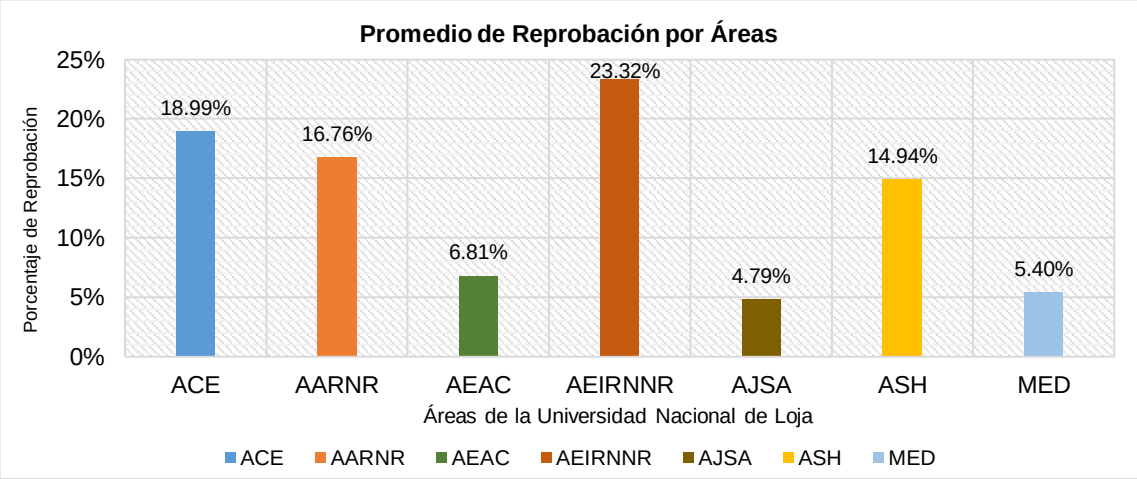


Figura 82: Porcentaje de Estudiantes reprobados en Áreas de la Universidad Nacional de Loja.

- En cuanto a los factores que inciden en la deserción se comprobó que el factor académico incide más en el abandono de los estudios, seguido por el factor institucional y en último lugar al factor individual. (ver figura 3).

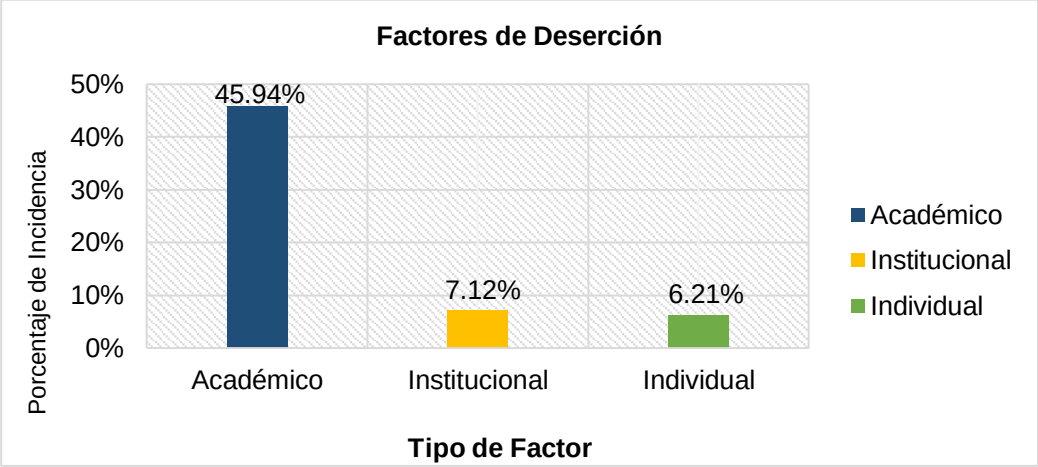


Figura 83: Porcentaje de incidencia de factores de deserción.

La destacada incidencia del factor académico en la figura anterior se debe a la relación que existe con atributos como: el promedio de notas, el promedio de asistencia, la decisión de cambiarse de carrera, el número de módulos reprobados y el período en el que ha reprobado; con respecto al factor institucional este se relaciona con atributos como: tipo de beca recibida por el Área de Bienestar Universitario, la carrera en la que se matriculo, el horario de estudio y servicios adquiridos por el área Bienestar Universitario.

Mientras que el factor individual este se relaciona con atributos como: la edad de ingreso a la universidad, el estado civil, la etnia, lugar de origen y genero del estudiante, si tiene hijos, si alguna vez estuvo período de gestación y los servicios de telefonía contratados por el estudiante.

- En cuanto a factores que inciden en la reprobación, se han identificado las siguientes reglas o patrones de comportamiento, los cuales se encuentran ordenadas por prioridad.

Regla 1: Estudiantes con un promedio de notas entre 7.5 y 8.4, con un promedio de asistencias menor al 85.0% y matriculados en la carrera de Ingeniería en Sistemas.

Regla 2: Estudiantes con un promedio de notas entre 7.5 y 8.4, matriculados en la carrera de Ingeniería en Electromecánica, provenientes del sector rural de la ciudad de Loja, poseen los servicios de teléfono y celular, con el estado civil soltero.

Regla 3: Estudiantes con un promedio de notas entre 7.5 y 8.4, con un promedio de asistencias menor al 85.0%, matriculados en la carrera de Ingeniería en Electromecánica y con una edad de ingreso a la universidad mayor a los 20 años.

Regla 4: Estudiantes con un promedio de notas entre 7.5 y 8.4, estudian por la tarde, tienen un promedio de asistencia entre 85.1% y 95.0%, provenientes del sector rural de la ciudad de Loja y no han recibido servicios de bienestar universitario.

- Los resultados obtenidos al predecir las probabilidades de deserción en estudiantes que cursan actualmente las carreras del Área de Energía las Industrias y los Recursos Naturales No Renovables, demuestran que la carrera de Ingeniería en Sistemas muestra el mayor porcentaje de estudiantes con peligro de abandonar los estudios en un 60% aproximadamente (ver figura 4).

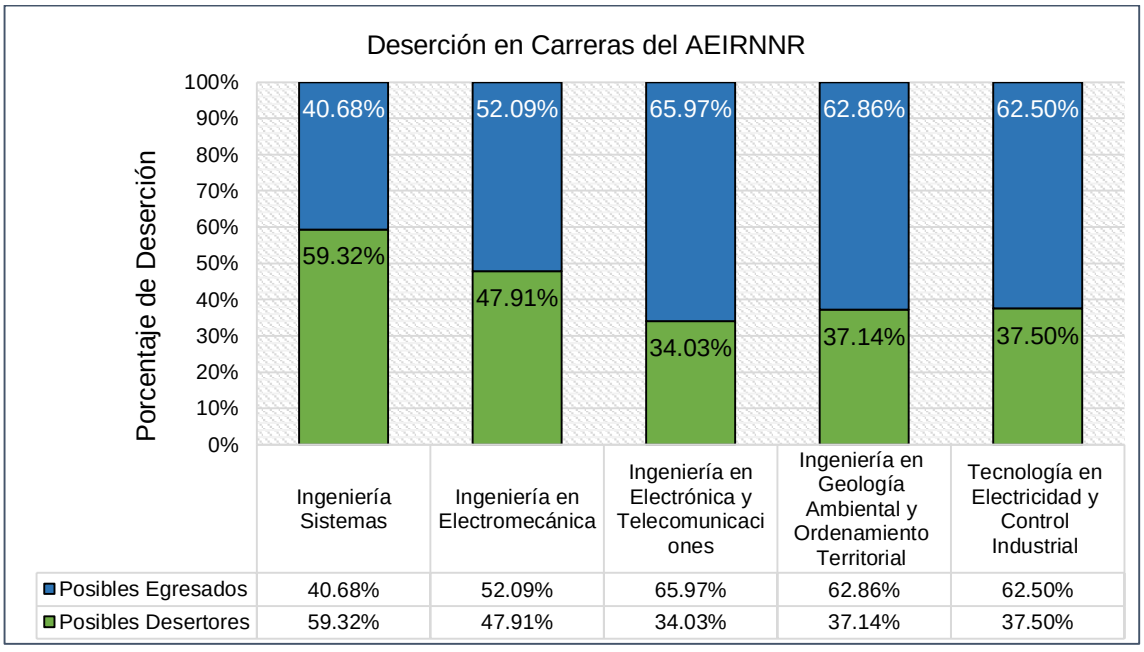


Figura 84: Distribución de posibles Desertores y Egresados del AEIRNNR.

A continuación se describe la cantidad de estudiantes con probabilidades de egresar y desertar en las carreras del Área de la Energía, las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja (ver tabla I).

Carreras el Área de Energía	Estudiantes con mayor probabilidad de Desertar	Estudiantes con mayor probabilidad de Egresar	Total de Estudiantes
Ingeniería en Sistemas	175	120	295
Ingeniería en Electromecánica	103	112	215
Ingeniería en Electrónica y Telecomunicaciones	49	95	144
Ingeniería en Geología Ambiental y Ordenamiento Territorial	39	66	105
Tecnología en Electricidad y Control Industrial	3	5	8

Tabla I: Número de estudiantes con probabilidades de desertar y egresar por Carreras del AEIRNNR.

Los índices de deserción en las carreras de Ingeniería en Sistemas y Ingeniería en Electromecánica, se dan porque la mayoría de estudiantes con posibilidades de abandonar los estudios provienen del sector rural de Loja, también un promedio de asistencia menor al 85% y un promedio de notas menor a 7,5. Además estas carreras presentan índices de reprobación en cada período académico, pero más significativamente en los períodos que van de primero a séptimo.

Mientras que las carreras de Ingeniería en Electrónica y Telecomunicaciones, Ingeniería en Geología Ambiental y Ordenamiento Territorial y Tecnología en Electricidad y Control Industrial muestran índices de reprobación solo en los primeros períodos académicos que van de primero a tercero y un promedio de asistencia entre 85% y 95%.

Con los resultados descritos anteriormente a continuación se detallan las estrategias de solución:

- Estrategias de Acompañamiento Académico: Tutorías Pares.
- Identificación del estilo de aprendizaje de los estudiantes, en los primeros días del período académico.

Se recomiendan las Estrategias de Acompañamiento Académico: estas se enfocan en ofrecer una oportunidad a los estudiantes de recibir retroalimentación y a su vez les permita tener una mejor comprensión acerca de los marcos teóricos y conceptuales y de esta forma hacer que el aprendizaje sea más efectivo, mejorando sus habilidades de pensamiento y con esto contribuir a la solución de problemas reales; entre las posibles estrategias están las Tutorías de Pares que consiste en la ayuda que un estudiante brinda a otro en tareas académicas generando un beneficio mutuo. El estudiante como tutor puede ser uno de los estudiantes con un alto rendimiento académico, pero también puede ser un estudiante en igualdad de condiciones de quien recibe la tutoría.

Esta estrategia de apoyo puede llevarse a cabo en grupos de dos o más estudiantes. Las Tutorías de Pares son experiencias de trabajo significativo que posibilitan un progreso y avance entre los estudiantes, los alientan a participar en actividades de enriquecimiento y los alientan a permanecer en el contexto Educativo.

Como estrategia para disminuir los índices de reprobación universitaria y con el propósito de incrementar los niveles de rendimiento académico de sus estudiantes, por medio del compromiso con su formación, se recomienda la identificación del estilo de aprendizaje de los estudiantes, en los primeros días del período académico, con el fin de estructurar las clases según el estilo de aprendizaje predominante en el conjunto de estudiantes.

La identificación de estilos de aprendizaje permitirá a los estudiantes comprender mejor las temáticas impartidas y generar un compromiso académico propios de estudios de tercer nivel.

Con todo lo descrito anteriormente, a continuación se mencionan los beneficios de emprender estas estrategias:

- Mejora en el desempeño académico de los estudiantes esencialmente en los que están cursando los primeros años de carrera.
- Incremento en el número de egresados en las carreras del Área de la Energía las Industrias y los Recursos Naturales No Renovables.

- Una mejor adaptación del estudiante a lo que se denomina como “vida universitaria” en aspectos como: valores, actitudes y hábitos de estudio propios del nivel en que se encuentra.

Para el desarrollo del presente proyecto y alcanzar los objetivos del mismo, se aplicaron técnicas de Minería de Datos para descubrir información oculta en los datos, además de herramientas Software como: DatAdmin para la manipulación y exploración de los datos recopilados; la herramienta Rapid Miner que contiene los algoritmos y técnicas de Minería de Datos y se utilizó la metodología CRISP-DM, para llevar a cabo un desarrollo ordenado del presente proyecto.

ANEXO 11: Certificado de Traducción y Certificado de Revisión de Estilo y Ortografía.

MAHOLY KATHERINE MOROCHO MERINO

LICENCIADA EN LENGUA EXTRANJERA DE LA ESCUELA DE EDUCACIÓN BÁSICA PARTICULAR "JUAN PABLO II"

CERTIFICA:

Que el resumen del Trabajo de Titulación denominado "IDENTIFICACIÓN DE FACTORES EN LA REPROBACIÓN Y DESERCIÓN MEDIANTE TÉCNICAS DE MINERÍA DE DATOS EN EL ÁREA DE LA ENERGÍA DE LA UNIVERSIDAD NACIONAL DE LOJA", realizado por el Señor egresado **ANIBAL ISRAEL GONZÁLEZ PINEDA**, 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, 16 de Septiembre del 2014.

MAHOLY KATHERINE MOROCHO MERINO
LICENCIADA EN LENGUA EXTRANJERA

Certificado Revisión de Estilo y Ortografía

HUGO ARTEMAN SANCHEZ GUAICHA

DOCTOR EN CIENCIAS DE LA EDUCACIÓN, ESPECIALIDAD EN LENGUA Y LITERATURA

CERTIFICA:

Haber revisado y corregido el estilo y ortografía del Trabajo de Titulación denominado "Identificación de Factores en la Reprobación y Deserción mediante Técnicas de Minería de Datos en el Área de la Energía de la Universidad Nacional de Loja", realizado por el Señor egresado **ANIBAL ISRAEL GONZÁLEZ PINEDA**, previo a la obtención del título de INGENIERO EN SISTEMAS, el mismo que cumple con las reglas establecidas por la RAE, y por el normativo de presentación dispuesto por la carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja.

Por lo cual autorizo su presentación, sustentación y defensa.

Loja, 29 de Agosto del 2014

HUGO ARTEMAN SANCHEZ GUAICHA
DOCTOR EN CIENCIAS DE LA EDUCACIÓN,

ANEXO 12: Artículo Científico

Identificación de Factores en la Deserción y Reprobación Universitaria

A. González
Carrera de Ingeniería en Sistemas
Universidad Nacional de Loja
Loja, Ecuador
aigonzalezp@unl.edu.ec

Ing. P. Ordoñez
Carrera de Ingeniería en Sistemas
Universidad Nacional de Loja
Loja, Ecuador
pfordonez@unl.edu.ec

***Abstract*—This article is based on the identification of factors affecting universitary desertion and reprobation. To run this process of data mining techniques were applied. The data used to carry out the identification of factors was obtained from the databases of Academic Management System (S.G.A.) through its web services, plus area data was compiled Wellness University, and later was integrated into a single database. In addition to identifying the factors desertion and failure rates, and in order to validate the results of a predictive model of attrition, which was evaluated with data from students currently pursuing careers energy area of the National University of Loja was generated.**

Keywords—data mining; modeling; web services; algorithm; data warehouse, desertion, reprobation.

I. INTRODUCCIÓN

En todas las instituciones de nivel superior se busca proporcionar una formación académica de excelencia, ya sea a través de docentes preparados, infraestructura adecuada, sin embargo la deserción y reprobación universitaria persiste como problema a disminuir. En ese aspecto es necesaria la identificación de factores de deserción y reprobación en los estudiantes cuando están cursando los primeros años de su carrera, siendo imprescindible para emprender las acciones oportunas y poder mitigar este índice, y no menos significativo, predecir su deserción en cualquier momento para el seguimiento respectivo [1-4].

En base a lo descrito, el presente trabajo se enfoca en identificar los factores de deserción y reprobación, en base a información proporcionada por la institución y luego mediante la aplicación de técnicas de minería de datos identificar estos factores.

Para el desarrollo de este trabajo se inició con la recopilación y estudio de las diferentes técnicas de minería de datos, en el cual se evidenció que las técnicas de árboles de decisión y técnicas de inducción de reglas son las que mejores resultados ofrecen. Posteriormente luego de estudiar las técnicas de minería de datos, se procedió a evaluar los modelos generados por las diferentes técnicas para identificar los factores de deserción y reprobación, además se generó un modelo predictivo de deserción para evaluar las probabilidades de estudiantes en abandonar los estudios. Cabe mencionar que la minería de datos aplicada a la educación ha dado como resultado de gran apoyo a predecir cualquier tipo de factor o característica de un caso, fenómeno o situación en la educación de nivel superior de un estudiante [5,6].

Posteriormente se aplicó los modelos con mejores resultados en cuanto a deserción sobre los estudiantes que cursan actualmente carreras en el Área de Energía de la Universidad Nacional de Loja.

La organización del trabajo es la siguiente: en la Sección II (ESTADO DEL ARTE) se documenta, conceptos y características de la minería de datos, herramienta escogidas para dicho estudio. La Sección III (CASO DE ESTUDIO) muestra el caso de estudio real, en el cual se detalla las diferentes temáticas tratadas y cuales han sido aplicadas en dicho caso de éxito. La Sección IV (RESULTADOS) se puede encontrar los resultados obtenidos de dicho estudio. La Sección V (CONCLUSIONES) se puede encontrar las conclusiones generadas a partir de la información recopilada en el presente artículo.

II. ESTADO DEL ARTE

La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos [7-11].

Fayyad define la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [10].

Otra definición por parte de Gartner Group, menciona que la minería de datos es el proceso de descubrir nuevas correlaciones significativas, patrones y tendencias por indagación a través de grandes cantidades de datos almacenados en repositorios, usando tecnologías así como técnicas matemáticas y estadísticas [12].

Con lo descrito anteriormente, las fases que permiten realizar un análisis y tomar decisiones adecuadas en base a los resultados generados son: recopilación e integración de datos; selección; limpieza y transformación; minería de datos; evaluación e interpretación; generación de nuevo conocimiento y finalmente toma de decisiones [13-14].

Para llevar un desarrollo ordenado en este trabajo se utilizó una metodología denominada CRISP-DM (Cross Industry Standard Process for Data Mining), esta decisión se basó en su amplio uso en el ámbito académico e industrial, la cual consta de las siguientes fases: Comprensión del negocio; Comprensión de los Datos; Preparación de los Datos; Modelado; Evaluación y por ultimo Implantación [22], la sucesión de estas fases no es necesariamente rígida, es decir cada fase es estructurada en varias tareas generales de segundo nivel.

Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

III. TÉCNICAS DE MINERÍA DE DATOS

La clasificación de las técnicas de minería de datos comprende las siguientes (ver tabla I): Técnicas Descriptivas en las cuales las variables tienen inicialmente el mismo estatus, Técnicas Predictivas en las cuales las variables pueden clasificarse inicialmente en dependientes e independientes [15,16].

TABLA I. TÉCNICAS DE MINERIA DE DATOS

Técnicas No Supervisadas o Descriptivas	**Técnicas Supervisadas o Predictivas**
Reglas de Asociación	Arboles de Decisión
Clustering (Agrupamiento)	Redes Neuronales
	Máquinas de Soporte Vectorial
	Clasificadores Bayesianos
	Reglas de Inducción

Para este caso de estudio las técnicas utilizadas fueron los arboles de decisión y las reglas de inducción [17-19]. El número de técnicas de minería de datos es muy grande y las mencionadas son algunas de las existentes. Además hay que señalar, que cualquiera sea el problema a resolver, no existe una sola técnica para solucionarlo, este puede ser abordado manejando aproximaciones distintas.

A. Herramientas de Minería de Datos

Existen algunas herramientas diseñadas para extraer conocimientos desde bases de datos que contienen grandes cantidades de información [20,21]. Se debe recordar que la minería de datos es una técnica compuesta por fases y la cual integra varias áreas por lo que no se debe confundir con un gran software [21].

En la actualidad existen aplicaciones o herramientas comerciales de minería de datos muy completas que permiten extraer conocimientos desde bases de datos y a su vez contienen diversas utilidades que facilitan el desarrollo de un proyecto. Sin embargo, en casi todo desarrollo de proyecto suelen complementarse con otras herramientas.

En este caso se determinó que Weka y RapidMiner son las herramientas convenientes para el trabajo, aunque finalmente se trabajó con la herramienta RapidMiner por sus amplias y flexibles características que ofrece a nuestras necesidades para realizar las actividades en el proceso de minería de datos, además las técnicas de procesamiento y algoritmos de la herramienta Weka se encuentran disponibles también en la herramienta RapidMiner.

IV. CASO DE ESTUDIO

La identificación de los factores de deserción y reprobación se lo realizo con la utilización de la versión 5.3.013 de la herramienta RapidMiner y el administrador de bases de datos DatAdmin.

Como caso de estudio se tomaron en cuenta datos de estudiantes del Área de la Energía las Industrias y los Recursos Naturales no Renovables por ser la que presenta el mayor índice de estudiantes que reprueban a nivel de toda la institución con un 23% en cada periodo académico.

Toda la información necesaria para el desarrollo de este proyecto se obtuvo de la base de datos del Sistema de Gestión Académica (S.G.A) a través de su Web Services y del Área de Bienestar Universitario almacenados en informes, libros y archivos digitales.

A. Fase Uno: Comprensión del Negocio

En esta fase se analizaron los recursos disponibles para el desarrollo del proyecto como recursos humanos, hardware, software y materiales y servicios, y principalmente se determinó los criterios de éxito del proyecto desde un punto de vista del negocio los cuales se describen a continuación:

- Como primer factor de éxito se definió la posibilidad de aplicar técnicas de minería de datos para predecir si un estudiante tiene probabilidades de abandonar los estudios, para ello se tiene en cuenta la utilización de datos personales así como notas académicas y el modulo y las materias en las cuales los estudiantes reprobaron.
- Como segundo factor de éxito se estableció, identificar al menos un factor que incida en la deserción y reprobación, en base a datos personales, académicos e institucionales los datos pertenecen a estudiantes del Área de Energía de la Universidad Nacional de Loja.

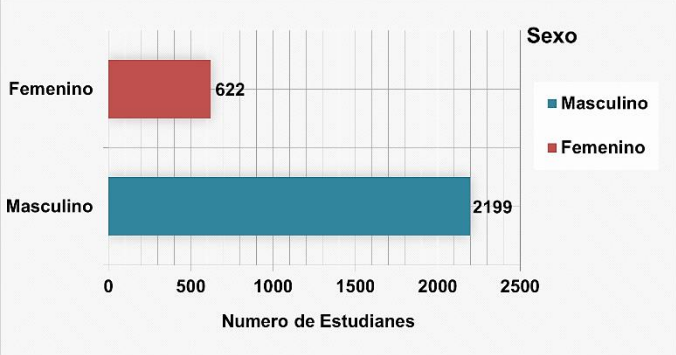
B. Fase Dos: Comprensión de los Datos

En esta fase se realizaron actividades con el fin de familiarizarse con los datos, comprender el problema y preparase para la fase posterior.

Además se describen los datos obtenidos en su estado original, la mayor parte de estos tuvieron que ser tratados para poder formar con ellos una base de datos coherente y consistente, la cual permita trabajar en todo el desarrollo del proyecto, los datos con los que se trabajó son del año 2008 hasta 2013.

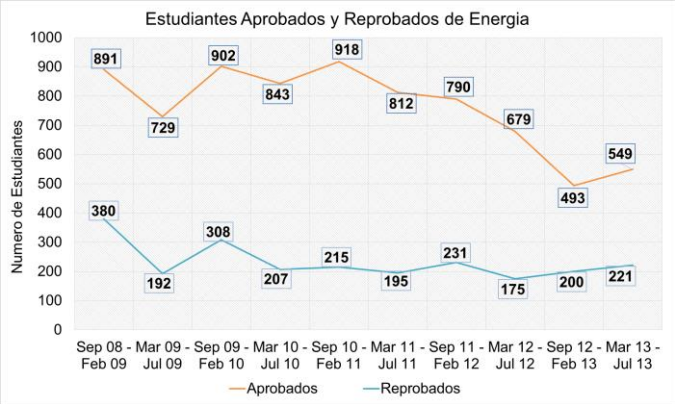
Luego de analizar los datos obtenidos y con el objetivo de familiarizarse con estos, a continuación de describen algunas características importantes: los estudiantes del Área de Energía, las Industrias y los Recursos Naturales No Renovables, contienen más estudiantes masculinos que estudiantes de sexo femenino (ver figura 1).

Fig. 1. Distribución del sexo en estudiantes del área de Energía.



El cuanto al número de estudiantes aprobados y reprobados del Área de Energía, el periodo que presenta más estudiantes reprobados es el periodo Septiembre 2008 – Febrero 2009 (ver figura 2).

Fig. 2. Distribución de estudiantes aprobados y reprobados en cada periodo.



C. Fase Tres: Preparación de los Datos

En esta fase se realizó la preparación final de datos, se eliminaron datos innecesarios e incorrectos, para luego proceder con la construcción de dos estructuras para identificar factores de deserción y de reprobación. Es decir se preparó todo lo relativo a la base de datos, para luego ingresarlos en la herramienta de modelado.

En cuanto a construcción de los datos se procedió a construir dos estructuras de atributos la primera enfocada a identificar factores de deserción la cual se emplean los siguientes atributos (ver tabla II):

TABLA II. ESTRUTURA PARA FACTORES DE DESERCIÓN

Atributo	Descripción
Cédula	Número de identificación del estudiante.
Edad de Ingreso	Edad de estudiante al momento de ingresar a la Universidad.
Estado	Situación del estudiante esta puede ser desertor o egresado.
Modulos Reprobados	Número de módulos reprobados de cada estudiante.
Periodo de Reprobación	Es el periodo en el cual el estudiante reprobó y los periodos pueden ser entre 1-3, 4-7 y 8-11.
Servicios	Los servicios que tiene el estudiante: teléfono fijo y móvil.
Sexo	Genero del estudiante.
Distancia de Origen	Distancia del lugar de procedencia (urbana, rural, cantón o provincia) hasta la ubicación de la Universidad.
Carrera	Describe la carrera a la que pertenece el estudiante estas pueden ser Ingeniería en Electromecánica, Ingeniería en Electrónica y Telecomunicaciones, Ingeniería en Geología Ambiental y Ordenamiento Territorial, Ingeniería en Sistemas, Tecnología en Electricidad y Control Industrial, Tecnología en Electrónica.
Cambio de Carrera	Describe si un estudiante se cambió de Carrera de las que posee el área de Energía.
Promedio de Asistencia	Describe si un estudiante tiene promedio de asistencias bajo (menor a 7.5), regular (de 7.5 8.5), bueno (de 8.5 a 9.0), muy Bueno (de 9.0 a 9.5) y excelente (mayor a 9.5).
Promedio de Notas	Describe el promedio de asistencia de un estudiante este puede ser bajo (menor a 80%), medio (entre 80% y 90%), alto (mayor a 90%).
Tipo de Beca	Describe el tipo de beca que recibe del área Bienestar Universitario.
Estado de Gestacion	Describe si una estudiante se encontró en estado de gestación mientras estudiaba.
Bienestar Servicios	Describe si recibió algún servicio de salud del área de Bienestar Universitario.
Estado Civil	Estado civil del estudiante y este puede ser: soltero, casado, divorciado, unión libre o viudo.
Etnia	Describe la etnia del estudiante esta puede ser: indígena, mestizo, blanco o montubio.
Hijos	Describe si un estudiante tiene hijos.
Horario de Estudio	Describe el horario de clases este puede ser matutino o vespertino.

A continuación se describen los atributos de la estructura enfocada a encontrar factores de reprobación (ver tabla III).

TABLA III. ESTRUTURA PARA FACTORES DE REPROBACIÓN

Atributo	Descripción
Cédula	Número de identificación del estudiante.
Edad de Ingreso	Edad de estudiante al momento de ingresar a la Universidad.
Servicios	Los servicios que tiene el estudiante: teléfono fijo y móvil.
Distancia de Origen	Distancia del lugar de procedencia (urbana, rural, cantón o provincia) hasta la ubicación de la Universidad.
Carrera	Describe la carrera a la que pertenece el estudiante estas pueden ser Ingeniería en Electromecánica, Ingeniería en Electrónica y Telecomunicaciones, Ingeniería en Geología Ambiental y Ordenamiento Territorial, Ingeniería en Sistemas, Tecnología en Electricidad y Control Industrial, Tecnología en Electrónica.
Promedio de Asistencia	Describe si un estudiante tiene promedio de asistencias bajo (menor a 7.5), regular (de 7.5 8.5), bueno (de 8.5 a 9.0), muy Bueno (de 9.0 a 9.5) y excelente (mayor a 9.5).
Promedio de Notas	Describe el promedio de asistencia de un estudiante este puede ser bajo (menor a 80%), medio (entre 80% y 90%), alto (mayor a 90%).

Atributo	Descripción
Tipo de Beca	Describe el tipo de beca que recibe el estudiante.
Bienestar Servicios	Describe si recibió algún servicio de salud del área de Bienestar Universitario.
Estado Civil	Estado civil del estudiante y este puede ser: soltero, casado, divorciado, unión libre o viudo.
Hijos	Describe si un estudiante tiene hijos.
Padre Trabaja	Describe si el padre del estudiante trabaja
Madre Trabaja	Describe si la madre del estudiante trabaja
Horario de Estudio	Describe el horario de clases: matutino o vespertino.
Reprobo	Describe si un estudiante reprobó o no alguna vez.

D. Fase Cuatro: Modelado

En esta fase se aplicaron algoritmos de árboles de decisión y reglas de inducción, también se aplicaron distintos parámetros para cada uno, además se evaluaron y compararon los resultados obtenidos.

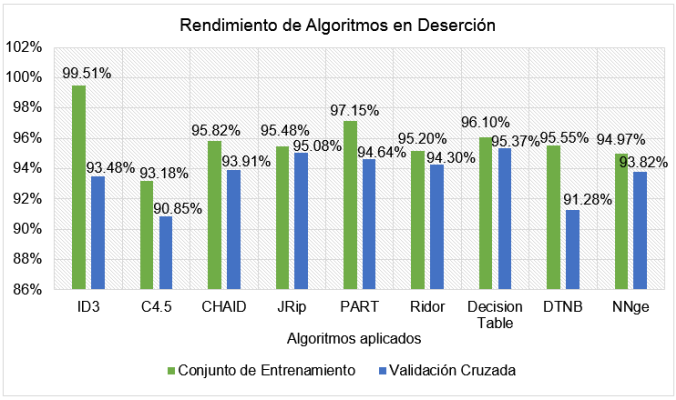
Los algoritmos seleccionados para identificar los factores de deserción en cuanto a los arboles de decisión son: ID3, C4.5, CHAID mientras que de los algoritmos basados en reglas de inducción son: Decision Table, DTNB, Ridor, JRip, NNge y PART. Los resultados del rendimiento se describen a continuación (ver tabla IV):

TABLA IV. RENDIMIENTO DE ALGORTIMOS PARA DESERCIÓN.

Clasificador	Modo de Prueba	Clasificados Correctamente %	Clasificados Incorrectamente %
ID3	Entrenamiento	99.51%	0.49%
	Validación Cruzada	93.48%	6.52%
C4.5	Entrenamiento	93.18%	6.82%
	Validación Cruzada	90.85%	9.15%
CHAID	Entrenamiento	95.82%	4.18%
	Validación Cruzada	93.91%	6.09%
JRIP	Entrenamiento	95.48%	4.52%
	Validación Cruzada	95.08%	4.92%
PART	Entrenamiento	97.15%	2.85%
	Validación Cruzada	94.64%	5.36%
RIDOR	Entrenamiento	95.20%	4.80%
	Validación Cruzada	94.30%	5.70%
DECISION TABLE	Entrenamiento	96.10%	3.90%
	Validación Cruzada	95.37%	4.63%
DTNB	Entrenamiento	95.55%	4.45%
	Validación Cruzada	91.28%	8.72%
NNGE	Entrenamiento	94.97%	5.03%
	Validación Cruzada	93.82%	6.18%

Entre los algoritmos que no obtuvieron los mejores resultados están el algoritmo C4.5, DTNB, NNge Chaid, si bien el rendimiento es bajo respecto a los demás algoritmos esto no significa que los resultados obtenidos fueron descartados de inmediato (ver figura 3).

Fig. 3. Resultados de Clasificacion correcta de algoritmos en Modelos de Deserción.



Según los datos mostrados, luego de aplicar las pruebas de Validación Cruzada los algoritmos que mejores resultados ofrecen son: Decision Table, Part, Ridor y JRip, con respecto a este último, las reglas generadas son pocas y carecen de significado y producto de eso los resultados que se presentan están sobreestimados.

Para identificar los factores de reprobación, se trabajó con la estructura de datos descrita en la tabla III y con la técnica de reglas de inducción específicamente los algoritmos: JRip, Part, Decision Table, Ridor, DTNB, NNge.

El rendimiento de los algoritmos aplicados se describe a continuación (ver tabla V):

TABLA V. RENDIMIENTO DE ALGORTIMOS PARA REPROBACIÓN.

Clasificador	Modo de Prueba	Clasificados Correctamente %	Clasificados Incorrectamente %
JRIP	Entrenamiento	82.62%	17.38%
	Validación Cruzada	82.03%	17.97%
PART	Entrenamiento	86.93%	13.07%
	Validación Cruzada	80.36%	19.64%
DECISION TABLE	Entrenamiento	82.17%	17.83%
	Validación Cruzada	82.13%	17.87%
RIDOR	Entrenamiento	82.57%	17.43%
	Validación Cruzada	80.50%	19.50%
DTNB	Entrenamiento	81.97%	18.03%
	Validación Cruzada	81.35%	18.65%
NNGE	Entrenamiento	77.42%	22.57%
	Validación Cruzada	75.96%	24.03%

Los algoritmos con mejor rendimiento son: Decision Table, JRip, DTNB y Ridor, estos presentan resultados similares, tanto

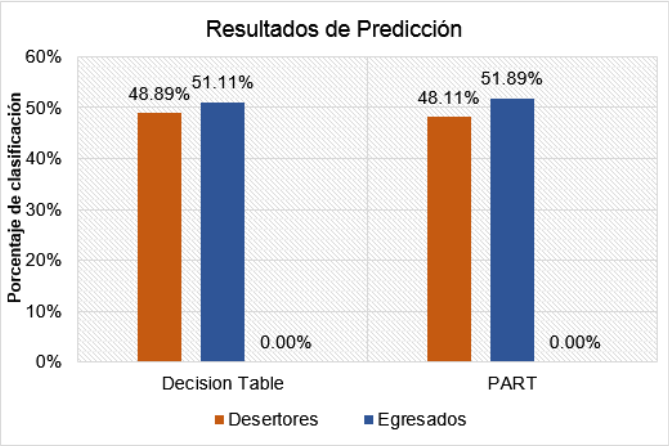
al evaluar con el conjunto entrenamiento, como la validación cruzada, es decir las reglas generadas obtenidas en la primera evaluación varían poco luego de aplicar la validación cruzada, sin embargo las reglas generadas por el algoritmo JRip se enfocan en describir reglas para estudiantes con que no han reprobado dificultando la tarea de análisis de factores de reprobación.

Por lo tanto en base a los resultados obtenidos se decidió elegir a los algoritmos: Decision Table y Ridor, ya que presentan los mejores resultados tanto en rendimiento, clasificación de la clase que reprueba (Si) y las medidas de error, no se tomó en cuenta el algoritmo DTNB, porque la tabla de decisión generada construyo reglas poco coherentes y escasamente descriptivas, además construyó reglas en base a tres de los 13 inicialmente seleccionados.

E. Fase Cinco: Evaluación

En esta fase se realizó una evaluación del rendimiento de los algoritmos Decision Table y Part enfocados a predecir los estudiantes desertores, para ello se trabajó con datos de estudiantes que cursan actualmente en las carreras del área de energía. Los resultados que se presentan a continuación reflejan el rendimiento de los modelos generados por cada algoritmo en donde se puede observar que ambos clasificaron en 100% de las instancias (ver figura 4).

Fig. 4. Resultados de Predicción con Decision Table y PART.



El algoritmo con mejores resultados fue Part debido a que ofrece las medidas de error más bajas, además este trabaja con todos los atributos seleccionados inicialmente, esto no ocurrió con el algoritmo Decision Table que seleccionó solo siete de los 17 atributos para generar la tabla de decisiones.

V. RESULTADOS

Como se estableció al comienzo del proyecto, se tiene como objetivo encontrar los factores de deserción y reprobación, es por ello que se ha planteado un conjunto de atributos asociados al problema de deserción universitaria [23], agrupados en factores individuales, académicos, institucionales como se describe en la siguiente tabla (ver tabla VI).

TABLA VI. ATRIBUTOS ASOCIADOS A DESERCIÓN.

Factores	Atributos
Individuales	Servicios, etnia, estado civil, hijos, sexo, distancia de origen, edad de ingreso, estado de gestación.
Académicos	Numero de módulos reprobados, periodo de reprobación, promedio de notas, promedio de asistencia, cambio de carrera
Institucionales	Tipo de beca, servicios de bienestar universitario, carrera, horario de estudio

Se realizó una evaluación de los atributos que más inciden en los modelos generados por los mejores algoritmos Decision Table y PART, obteniendo los siguientes resultados (ver tabla VII).

TABLA VII. PESOS DE ATRIBUTOS CON DECISION TABLE Y PART.

Algoritmo DECISION TABLE				
Factores	**Atributos**	**Peso**	**Suma**	**% de Proximidad**
Individuales	Servicios	0.010	0.497	6.21% / 100%
	Etnia	0.104		
	Estado civil	0.037		
	Hijos	0.010		
	Sexo	0.004		
	Distancia origen	0.007		
	Edad ingreso	0.185		
	Estado gestación	0.140		
Academicos	Módulos reprobados	1.000	2.306	46.12% / 100%
	Periodo reprobación	0.639		
	Promedio notas	0.396		
	Promedio asistencia	0.251		
	Cambio carrera	0.020		
Institucionales	Tipo beca	0.227	0.285	7.12% / 100%
	Bienestar servicios	0.007		
	Carrera	0.050		
	Horario estudio	0.001		
Algoritmo PART				
Individuales	Servicios	0.010	0.497	6.21% / 100%
	Etnia	0.104		
	Estado civil	0.037		
	Hijos	0.010		
	Sexo	0.004		
	Distancia origen	0.007		
	Edad ingreso	0.185		
	Estado gestación	0.140		
Academicos	Módulos reprobados	1.000	2.306	45.94% / 100%
	Periodo reprobación	0.639		
	Promedio notas	0.396		
	Promedio asistencia	0.251		
	Cambio carrera	0.020		
Institucionales	Tipo beca	0.227	0.285	7.12% / 100%
	Bienestar servicios	0.007		
	carrera	0.050		
	Horario estudio	0.001		

Luego de haber evaluado los atributos más relevantes con

cada uno de los atributos de mejor rendimiento (Decision Table y Part), se observó que el factor con más incidencia en la deserción es el académico seguido del factor institucional y por último el individual. Esto quiere decir que lo que más incide son atributos relacionados con el rendimiento académico del estudiante, seguido de este se encuentran los relacionados con ofrecidos por la institución y finalmente están los relacionados con el estudiante como lugar de procedencia, etnia sexo etc.

Como se estableció al comienzo del proyecto, se tiene como objetivo encontrar los factores de deserción y reprobación, es por ello que en la tarea de encontrar factores de reprobación se realizó de la siguiente manera en base a los resultados obtenidos mediante la aplicación de algoritmos basados en reglas de inducción, estas reglas permiten interpretar las situaciones en las que un estudiante reprueba considerablemente. El algoritmo con mejor rendimiento fue RIDOR y a continuación se describen las reglas o condiciones más relevantes.

- Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo y pertenecen a la carrera de Ingeniería en Sistemas entonces si reprueba.
- Si el estudiante tiene un promedio de notas regular, pertenecen a la carrera de Ingeniería en Electromecánica, son origen rural de la ciudad de Loja, poseen los servicios básicos y son solteros entonces si reprueba.
- Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo, pertenecen a la carrera de Ingeniería en Electromecánica con edad de ingreso a la universidad mayor a 20 años entonces si reprueba.
- Si el estudiante tiene un promedio de notas regular, estudia por la tarde, tiene un promedio de asistencia medio, son de origen rural y no han recibido servicios de bienestar universitario entonces si reprueba.

VI. CONCLUSIONES

Se han examinado distintos algoritmos de clasificación y una de las conclusiones que se obtiene del análisis de los resultados obtenidos es que en este caso los algoritmos que generaban modelos más sencillos son los que presentan mejores resultados. En el transcurso de la tarea de modelado se han obtenido excelentes resultados con los algoritmos basados en reglas, lo que implica que se obtuvieron modelos más simples.

En el trabajo realizado se llegó a la conclusión de que el factor que incide mayormente en la deserción de los estudiantes es el académico, seguido por el institucional dejando por ultimo al factor individual. Mientras que en el ambiente de reprobación los estudiantes que reprueban presentan las siguientes condiciones o comportamiento a través de las siguientes reglas:

Si el estudiante tiene un promedio de notas regular, un promedio de asistencia bajo y pertenecen a la carrera de Ingeniería en Sistemas entonces si reprueba; Si el estudiante tiene un promedio de notas regular, pertenecen a la carrera de Ingeniería en Electromecánica, son origen rural de la ciudad de Loja, poseen los servicios básicos y son solteros entonces si reprueba.

REFERENCIAS

[1] M. Boado, “Una aproximación a la deserción estudiantil universitaria en Uruguay,” Universidad de la República, Montevideo, Uruguay, en cooperación con el Instituto Internacional para la Educación Superior en América Latina y el Caribe, pp. 10–24, 2005.

[2] G. J. Paramo and C. A. C. Maya, “Deserción estudiantil universitaria. Conceptualización,” Revista Universidad EAFIT, vol. 35, no. 114, pp. 65–78, 2012.

[3] M. C. G. AFONSO, P. R. A. Pérez, L. C. PÉREZ, and J. T. B. Benítez, “El abandono de los estudios universitarios: factores determinantes y medidas preventivas,” Revista española de pedagogía, vol. 65, no. 236, pp. 71–88, 2007. 174

[4] E. Castaño, S. Gallón, and J. Vásquez, “Análisis de los factores asociados a la deserción estudiantil en la educación superior: un estudio de caso,” Revista de Educación, no. 345, pp. 255–280, 2008.

[5] E. B. Durán and R. N. Costaguta, “Minería de datos para descubrir estilos de aprendizaje,” Revista Iberoamericana de Educación, vol. 42, no. 2, p. 6, 2007.

[6] K. Gilbert, R. R. Sánchez, and J. C. R. Santos, “Mineria de datos: Conceptos y tendencias,” Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial, vol. 10, no. 29, pp. 11–18, 2006.

[7] RIQUELME, José C.; RUIZ, Roberto; GILBERT, Karina. Mineria de datos: Conceptos y tendencias. Revista Iberoamericana de Inteligencia Artificial, 2006, vol. 10, no 29, p. 11-18.

[8] LÓPEZ, César Pérez. Minería de datos: técnicas y herramientas. Editorial Paraninfo, 2007.

[9] I**.** K. Gilbert, R. R. Sánchez, and J. C. R. Santos, “Mineria de datos: Conceptos y tendencias,” *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, vol. 10, no. 29, pp. 11–18, 2006.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases”, AI magazine, vol. 17, no. 3, p. 37, 1996.

[11] GILBERT, Karina; SÁNCHEZ, Roberto Ruiz; SANTOS, José Cristobal Riquelme. Mineria de datos: Conceptos y tendencias. *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, 2006, vol. 10, no 29, p. 11-18.

[12] LUAN, Jing. Data mining and its applications in higher education. *New directions for institutional research*, 2002, vol. 2002, no 113, p. 17-36.

[13] I. Perversi and M. I. E. Fernández, Aplicación de Minería de Datos para la exploración y detección de patrones delictivos en Argentina. Proyecto de Tesis de Grado en Ingeniería Industrial. Instituto Tecnológico de Buenos Aires, 2007.

[14] WebMinnin Consultores, “KDD: Proceso de Extracción de conocimiento,” Enero de-2011. [En Línea]. Disponible: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/. [Acceso: 18-Sep-2013].

[15] Ballard C., Herreman D., Schau D., Bell R., Kim E., Valncic A.: "Data Modeling Techniques for Data Warehousing", IBM Red Book, 1998.

[16] A. J. CALLEJA GÓMEZ, “Minería de datos con WEKA para la predicción del precio de automóviles de segunda mano,” 2011.

[17] ORALLO, José Hernández; QUINTANA, Ma José Ramírez; RAMÍREZ, César Ferri. Introducción a la Minería de Datos. Pearson Prentice Hall, 2004.

[18] ALUJA BANET, Tomàs. La minería de datos, entre la estadística y la inteligencia artificial. 2001.

[19] BRITOS, Paola Verónica; GARCÍA MARTÍNEZ, Ramón. Propuesta de Procesos de Explotación de Información. En XV Congreso Argentino de Ciencias de la Computación. 2009.

[20] V. Valcárcel Asencios, “Data Mining y el descubrimiento del conocimiento,” Ind. data, vol. 7, no. 2, pp. 83–86, 2004.

[21] J. Alcalá-Fdez, M. J. del Jesus, J. M. Garrell, F. Herrera, C. Herbás, and L. Sánchez, “Proyecto KEEL: Desarrollo de una herramienta para el análisis e implementación de algoritmos de extracción de conocimiento evolutivos,” Tendencias de la Mineria de Datos en Espana, Red Espanola

[22] CHAPMAN, Pete, et al. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

[23] Universidad Nacional ICFES, "Estudio de la deserción estudiantil en la educación superior en Colombia," Bogotá, Documento sobre estado Del arte 2003.