UNIVERSIDAD NACIONAL DE LOJA

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

CARRERA DE INGENIERÍA EN SISTEMAS

"Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje"

"Tesis previa a la obtención del título de Ingeniera en Sistemas"

Autora:

Angélica Elizabeth Jaramillo Zhingre

Director:

Ing. Henry Patricio Paz Arias, Mg. Sc.

Loja-Ecuador

2015

Certificación de director

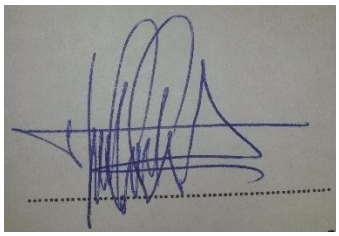
Ing. Henry Patricio Paz Arias, Mg. Sc.

DOCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS

CERTIFICA:

Que la egresada **Angélica Elizabeth Jaramillo Zhingre** autora del presente trabajo de tesis, cuyo tema versa sobre **"Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje"**, 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 autorizo su presentación y sustentación.

Loja, 03 de abril de 2015



Ing. Henry Patricio Paz Arias, Mg. Sc.

DIRECTOR DEL TRABAJO DE TITULACIÓN

Autoría

Yo, **ANGÉLICA ELIZABETH JARAMILLO ZHINGRE** declaro ser autora 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.

Autora: Angélica Elizabeth Jaramillo Zhingre.

Firma: 

Cédula: 1104999147

Fecha: 19 de junio del 2015

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

Yo **ANGÉLICA ELIZABETH JARAMILLO ZHINGRE** declaro ser autora de la tesis titulada: **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DETERMINAR LAS INTERACCIONES DE LOS ESTUDIANTES EN UN ENTORNO VIRTUAL DE APRENDIZAJE"**, como requisito para optar al grado de: **INGENIERA 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 la 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 realiza un tercero.

Para constancia de esta autorización, en la ciudad de Loja, a los diecinueve días del mes de junio del dos mil quince.

Firma: 

Autor: Angélica Elizabeth Jaramillo Zhingre

Cédula: 1104999147

Dirección: Loja (Bolonia, Av. Villonaco)

Correo Electrónico: aejaramilloz@unl.edu.ec

Teléfono: 073025830 **Celular:** 0981275667

DATOS COMPLEMENTARIOS

Director de Tesis: Ing. Henry Patricio Paz Arias, Mg. Sc.

Tribunal de Grado: Ing. Marco Augusto Ocampo Carpio, Mg. Sc.

Ing. Waldemar Victorino Espinoza Tituana, Mg. Sc.

Ing. Mario Andrés Palma Jaramillo, Mg. Sc.

Agradecimiento

Agradezco profundamente a DIOS por estar conmigo en cada paso que doy, guiándome en el sendero de la vida, fortaleciendo mi corazón e iluminando mi mente y por sus infinitas bendiciones.

A mi familia que la amo mucho, mis PADRES y HERMANOS por ser mi ejemplo para seguir adelante, quienes siempre han estado ahí para apoyarme incondicionalmente en cada momento de mi vida y motivándome en mi formación académica, creyeron en mí en todo momento, gracias por eso y por muchos más.

A los docentes quienes les debo gran parte de mis conocimientos, gracias a su paciencia y enseñanza y a mi director de tesis Mg. Sc. Henry Paz quién me ayudó en las tutorías para poder terminar con éxito mi Trabajo de Titulación,

A una persona especial por ser una parte muy importante en mi vida, estar en las buenas y en las malas, por su amor incondicional, apoyo continuo, comprensión y ayuda en todo momento.

Angélica Elizabeth Jaramillo Zhingre

Dedicatoria

Quiero dedicar este trabajo de titulación a:

Dios por ser el inspirador para cada uno de mis pasos dados en mí convivir diario, a mi familia por ser los guías en el sendero de cada acto que realizo, por haberme dado todas las herramientas para realizar este trabajo, por su paciencia, por su ayuda constante, por todo su amor, comprensión y por ser el incentivo para seguir adelante.

Angélica Elizabeth Jaramillo Zhingre

Cesión de derechos

Angélica Elizabeth Jaramillo Zhingre autora 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

"Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje"

b. Resumen

El presente trabajo de titulación está enfocado en determinar las interacciones de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia (MED) de la Universidad Nacional de Loja, para ello se realizó un análisis de la base de datos donde se encontraba la información correspondiente al periodo académico 2013 - 2014, que consta de datos personales, institucionales, socioeconómicos y de las interacciones (tareas, recursos y exámenes) de los estudiantes.

De igual forma se consideró realizar un estudio de las diferentes técnicas de minería de datos acordes al proyecto donde se seleccionó la técnica de clasificación para generar los modelos a través de los algoritmos Chaid, Prism, Knn, Decision Tree, ID3, J48, Jrip y Ridor, posteriormente se efectuó un análisis de las metodologías de minería de datos comparando cada una de ellas con el fin de seleccionar la que ayude al desarrollo del proyecto eligiendo la metodología CRISP-DM ya que contiene múltiples fases indicando cada una de sus actividades que se deben cumplir para obtener el modelo, convirtiéndose de esta forma en una guía práctica para cumplir con los objetivos establecidos.

Además se desarrolló un análisis comparativo tomando en cuenta características de las herramientas de minería de datos donde se eligió RAPIDMINER para realizar los procesos mediante los algoritmos conjuntamente con los datos de los estudiantes los mismos que se dividieron en dos conjuntos, para entrenamiento y validación, obteniendo como resultado que el mejor algoritmo fue el Decision Tree, ya que clasificó las instancias correctamente así mismo presentó un árbol con los diferentes atributos dando las mejores reglas de las interacciones de los estudiantes de tal forma se pudo generar el modelo que permitió establecer que en gran mayoría los estudiantes tienen un nivel de interacción medio en el curso virtual de inglés, donde los factores que más influyeron fueron las interacciones en las tareas, exámenes, recursos, situación laboral y estado civil del estudiante.

Summary

This research is focused on determining the interactions of virtual English course students who belong to the Distance Learning Mode (MED) of Loja National University, to do this I analyzed the database which contained data of the academic period 2013 - 2014, this information was about personal, institutional, socioeconomic data, and students interactions (tasks, resources and exams).

Also I considered a study of the different techniques of data mining and the classification technique was selected to generate models through Chaid, Prism, Knn, Decision Tree, ID3, J48, Jrip algorithms and Ridor, after that I do an analysis of mining data and I selected the CRISP-DM methodology, these are a practical guide to find results.

Finally I did a comparative analysis of data mining tools where RapidMiner was selected, it helps to do the processes by algorithms with data from the same students who were divided in two groups, for training and validation, and the best was the Decision Tree; it classified the correct instances and presented a tree with different attributes giving the best interactions rules of students so it could generate the model which established that most students have a medium level of interaction in virtual English course where most influential factors were interactions homework, tests, resources, employment status and marital status of the students.

Índice de Contenidos

Índice General

Índice de Figuras

Índice de Tablas

c. Introducción

La educación es la base del progreso de los países, por ello en la actualidad los sistemas educativos se enfrentan al desafío de utilizar las tecnologías de la información, teniendo un papel importante porque facilitan el aprendizaje en entornos virtuales preparando a los estudiantes a la adquisición del conocimiento en forma inmediata y amplia, sin que la distancia ni el tiempo sea un inconveniente en su formación académica [1].

Los avances en la educación a distancia se basan en el aprendizaje electrónico el mismo que brinda facilidades para la comunicación e interacción a través del internet, además los entornos virtuales almacenan una gran cantidad de datos sobre las actividades de los estudiantes cuando estos toman un curso y usualmente esta información es utilizada para monitorear características del mismo, la información se presenta en grandes volúmenes de datos por lo que resulta difícil su interpretación, esta es la razón por la cual el uso de la Minería de Datos es muy apropiada para descubrir información relevante en las bases de datos, los métodos pueden ser aplicados para explorar, visualizar y analizar datos con la finalidad de identificar patrones útiles de las actividades de los estudiantes durante la interacción en el entorno virtual [2].

Al respecto la Universidad Nacional de Loja cuenta con sistemas de información para brindar la facilidad de estudios a distancia ya sea en distintas carreras o cursos, estos sistemas almacenan grandes cantidades de información de los estudiantes como es el caso del curso virtual de inglés de la modalidad de estudios a distancia, el mismo que se ha tomado como objeto de estudio, pero tener numerosa información a disposición y no saber qué hacer con ella es un gran problema, es aquí donde interviene la Minería de Datos que contiene un conjunto de técnicas que se aplican para extraer conocimiento útil y comprensible, previamente desconocido, así mismo descubrir patrones para generar un modelo a través del análisis de la información de las interacciones en el curso, datos personales, institucionales y socioeconómicos del estudiante, que permita determinar las interacciones de los estudiantes en el curso virtual, para que de esta manera ayude a la toma de decisiones, y por tanto un beneficio a la institución.

CRISP-DM es la metodología utilizada para la creación del modelo ya que es una de las más usadas en la actualidad para la generación de proyectos de Minería de datos, con esta se pretende obtener un modelo de análisis de datos, conjuntamente con la implementación de algoritmos de Inteligencia Artificial, ya incorporados en la herramienta de pre-procesamiento de datos RapidMiner [3].

En el presente proyecto de trabajo de titulación se desarrollará un estado del arte donde se podrá comprender acerca de las interacciones de los estudiantes en entornos virtuales que permitió conocer la participación de los estudiantes en el curso virtual de inglés.

La minería de datos comprende un conjunto de procesos, técnicas, algoritmos y herramientas para analizar grandes cantidades de datos y obtener conocimientos que ayuden a la toma de decisiones, existen diferentes áreas en la actualidad en donde se emplean la minería de datos.

La implementación de la Metodología CRISP-DM durante el desarrollo del proyecto fue de gran importancia ya que presenta una guía práctica con cada una de las fases y las actividades para la elaboración del modelo que se deben llevar a cabo para conseguir el cumplimiento de los objetivos del proyecto, finalmente se presentan las conclusiones y recomendaciones respecto a los resultados obtenidos del proyecto.

d. Revisión de Literatura

CAPÍTULO I

1. HERRAMIENTAS PARA EL DESARROLLO DEL PROYECTO

1.1. Moodle

Moodle es una plataforma de aprendizaje a distancia basada en software libre que cuenta con una grande y creciente base de usuarios, también denominado " Ambiente de Aprendizaje Virtual o Educación en Línea " es decir, una aplicación diseñada para ayudar a los educadores a crear cursos de calidad en línea, se caracteriza por ser hoy en día el entorno más popular de formación virtual, que tiene una comunidad extensa de desarrolladores alrededor del mundo lo que la ha catalogado a ser la plataforma más extendida para la formación virtual y también como acompañamiento a la formación presencial [4].

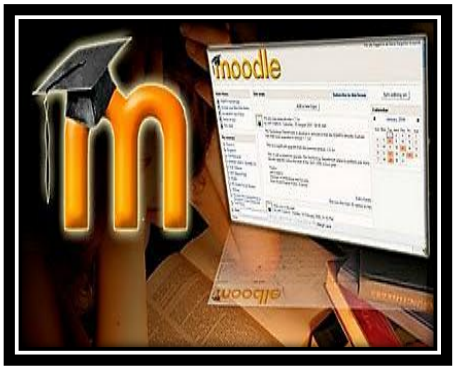


Figura 1: Moodle [4]

Moodle fue creado por el australiano Martin Dougiamas, esta herramienta ha venido evolucionando desde 1999, produciéndose nuevas versiones del producto, extendiéndose por más de 100 países y siendo traducida a más de 50 idiomas [4].

1.1.1. Estructura de Moodle

Permite el levantamiento de un centro capaz de gestionar distintos cursos a la vez a través de la red, que se caracteriza por poseer una estructura modular [4].

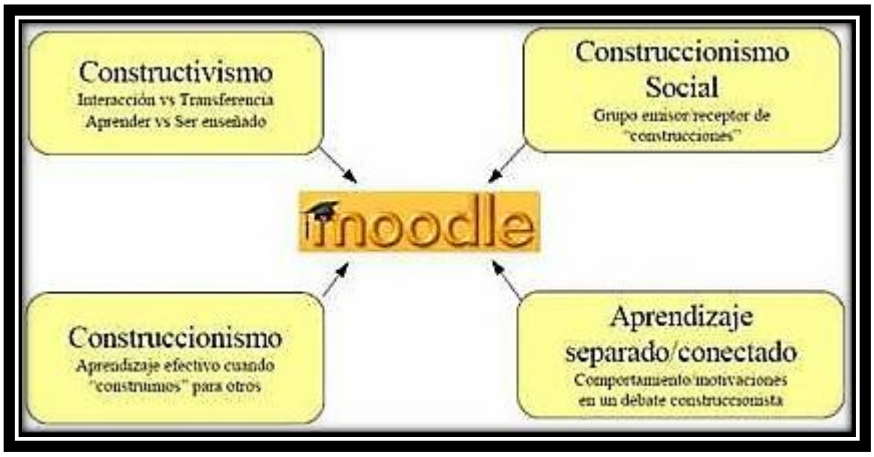


Figura 2: Estructura de Moodle [4].

A) Constructivismo: atribuida al filósofo Jean Piage, sostiene que las personas construyen nuevos conocimientos de manera activa al tiempo que interactúan con su entorno siguiendo un proceso de asimilación y acomodación. Una persona asimilará un concepto cuando las experiencias sean alineadas con respecto al conocimiento previo de la persona. Por otra parte el proceso de acomodación, es el proceso en el cual la persona debe acomodar los conocimientos previos a los nuevos conocimientos que ha adquirido [4].

B) Construccionismo: se afirma que el aprendizaje es más efectivo cuando se construyen cosas. El construccionismo explica que el aprendizaje es particularmente efectivo cuando se construye algo que debe llegar otros. Esto puede ir desde una frase hablada o enviar un mensaje en internet, a artefactos más complejos como una pintura, una casa o un paquete de software [4].

C) Construccionismo Social: Este concepto extiende las ideas anteriormente descritas a un grupo social, los individuos de este grupo construyen artefactos para los otros individuos del grupo, creando de manera colaborativa una pequeña cultura de artefactos compartidos con significados compartidos [4].

D) Conectado y Separado: Esta idea profundiza en las motivaciones de los individuos dentro de una discusión. Una persona aplica el comportamiento separado cuando intenta mantenerse “objetivo” y tiende a defender sus propias ideas utilizando la lógica y encontrando puntos débiles en las ideas del oponente [4].

1.1.2. Características básicas de Moodle

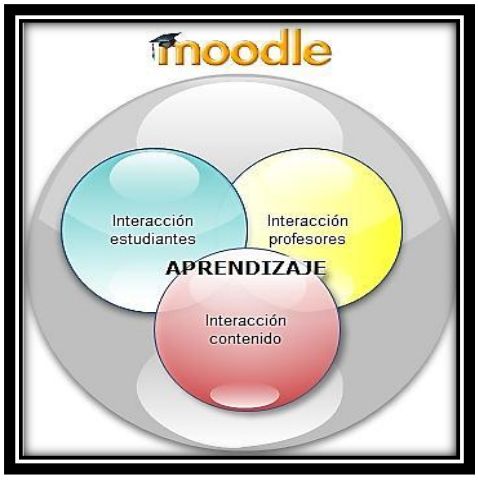


Figura 3: Interacciones en moodle [4]

A continuación las principales características que presenta Moodle en los tres niveles de relevancia [4]:

A nivel general [4]:

- **Escalable:** Se adapta a las necesidades que aparecen en el transcurso del tiempo.
- **Personalizable:** Moodle se puede modificar de acuerdo a los requerimientos específicos de una institución o empresa.
- **Económico:** En comparación a otros sistemas propietarios Moodle es gratuito, su uso no implica el pago de licencias u otro mecanismo de pago.

A nivel funcional [4]:

- **Facilidad de uso:** Permite la Gestión de Perfiles de Usuario, almacenar cualquier dato que se desee sobre el alumno o profesor.
- **Facilidad de Administración.** Cuenta con un panel de control central desde el cual se puede monitorear el correcto funcionamiento y configuración del sistema.
- **Realizar exámenes en línea:** es decir publicar una lista de preguntas dentro de un horario establecido y recibir las respuestas de los alumnos.

- **Presentación de cualquier contenido digital:** Se puede publicar todo tipo de contenido multimedia como texto, imagen, audio y video para su uso dentro de Moodle como material didáctico.
- **Gestión de tareas:** Los profesores pueden asignar tareas o trabajo prácticos de todo tipo, gestionar el horario y fecha de su recepción, evaluarlo y transmitir al alumno la retroalimentación respectiva.
- **Implementación de aulas virtuales:** Mediante el uso del chat o sala de conversación incorporada en Moodle, se pueden realizar sesiones o clases virtuales, en las cuales el profesor podría plantear y resolver interrogantes.
- **Implementación de foros de debate o consulta:** Esta característica se puede usar para promover la participación del alumnado en colectivo hacia el debate y reflexión.
- **Importación de contenidos de diversos formatos:** Se puede insertar dentro de Moodle, contenido educativo proveniente de otras plataformas.

Los principales beneficios son [4]:

- **Libertad:** Moodle no se encuentra atado a ninguna plataforma (Windows, Linux, Mac) especifica, brindando total libertad para escoger la que se ajuste a sus necesidades tanto en el presente como en el futuro.
- **Integración:** Moodle es un sistema abierto lo que significa que es posible integrarlo con otros sistemas.
- **Gestión del Conocimiento:** Permite el almacenamiento y recuperación de conocimiento producto de las actividades e interrelaciones del alumno.
- **Arquitectura Modular:** Moodle agrupa sus funciones o características a nivel de módulos. Estos módulos son independientes, configurables, además de poder ser habilitados o deshabilitados según sea conveniente.

1.1.3. Módulos principales en Moodle

- **Página principal del curso:** La página principal del curso se puede observar en la siguiente figura (ver Figura 4):

Figura 4: Página principal del curso

- **Usuarios del curso:** Los usuarios que conforman el curso se pueden observar de la siguiente grafica (ver Figura 5) [5]:

Figura 5: Usuarios del curso

- **Información Personal:** Contiene el perfil del estudiante como son sus datos personales los cales pueden ser modificados. En esta pantalla se indica las siguientes opciones: Perfil, Mensajes, Cambiar Contraseña y Editar información (ver Figura 6 y 7) [5].

Figura 6: Información del estudiante

Figura 7: Editar Información Personal del estudiante

- **Módulo de tareas:** Esta es la opción empleada para que el estudiante se informe acerca de la realización de una tarea, de acuerdo con los contenidos que se está desarrollando dentro del curso. El resultado de la misma, normalmente, consiste en la creación de un archivo en formato digital (documento, imagen, sonido, etc.) que se debe subir al curso. Esta actividad es calificable (ver Figura 8) [5].

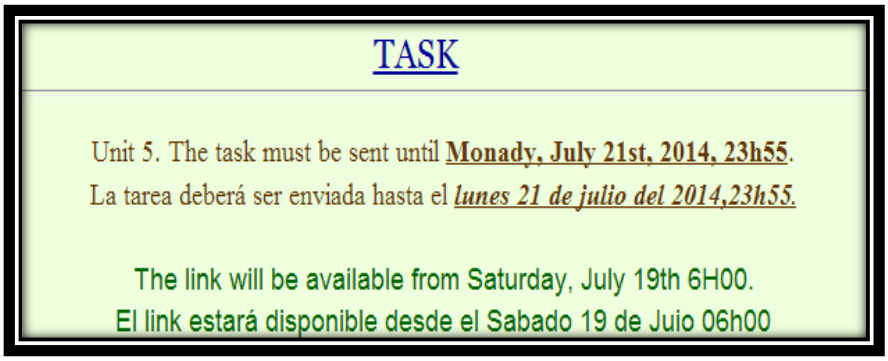


Figura 8: Información Tareas

Entre las actividades que se tiene en este módulo de tareas están:

- Puede especificarse la fecha final de entrega de una tarea y la calificación máxima que se le podrá asignar.
- Los estudiantes pueden subir sus tareas.
- Para cada tarea en particular, puede evaluarse a la clase entera (calificaciones y comentarios) en una única página con un único formulario.
- Las observaciones del profesor se adjuntan a la página de la tarea de cada estudiante y se le envía una notificación.

- **Módulo foro:** Los foros son un medio ideal para publicar pequeños mensajes y mantener discusiones públicas sobre la información u opiniones allí vertidas [5].

En todas las asignaturas, cursos o espacios existe un foro de forma predefinida, el foro de Novedades. Este foro se crea automáticamente para que participe la primera vez que ingrese (ver Figura 9) [5].

Figura 9: Foro

- **Módulo cuestionario:** Los cuestionarios consisten en opción múltiple, falso/verdadero y respuestas cortas que son calificadas por el profesor que puede decidir mostrar las respuestas correctas al finalizar el examen. Debe ser cauteloso con el tiempo límite, podrá visualizar un contador pero al finalizar el temporizador no desconecta al estudiante de la lección pero cualquier pregunta respondida después del límite no es contabilizada y puede tener inconvenientes [5].

Al ingresar al cuestionario le indicará la siguiente pantalla, ahí le enseña las recomendaciones, intentos permitidos, método de calificación y la opción para contestar el cuestionario (ver Figura 10) [5].

Figura 10: Ingreso a la evaluación

En la siguiente figura se indica la forma de evaluación del curso (ver Figura 11):

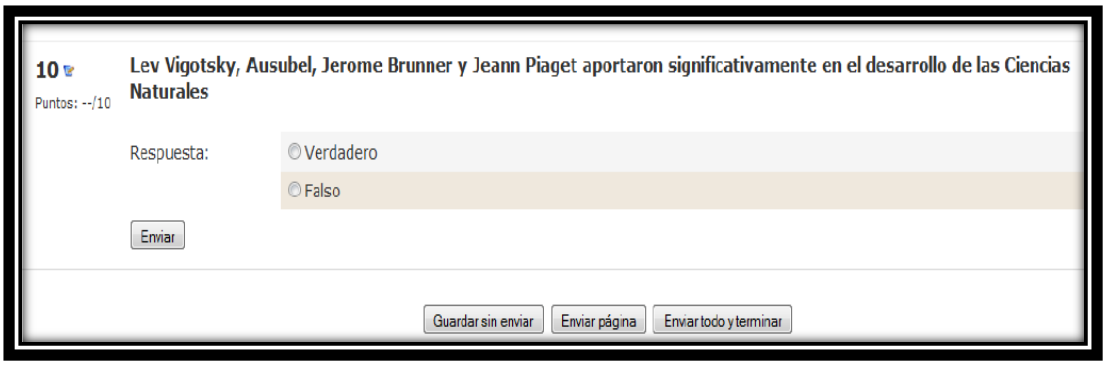


Figura 11: Evaluación

Entre la descripción de la evaluación se tiene:

- Las preguntas de opción múltiple pueden definirse con una única o múltiples respuestas correctas.
- Pueden crearse preguntas de respuesta corta (palabras o frases).
- Pueden crearse preguntas tipo verdadero/falso.
- Pueden crearse preguntas aleatorias.

❖ **Módulo recurso:** Admite la presentación de un importante número de contenido digital, Word, PowerPoint, Excel, Flash, vídeo, sonidos, etc. Los archivos pueden subirse o pueden ser creados sobre la marcha usando formularios web (ver Figura 12 y 13) [5].

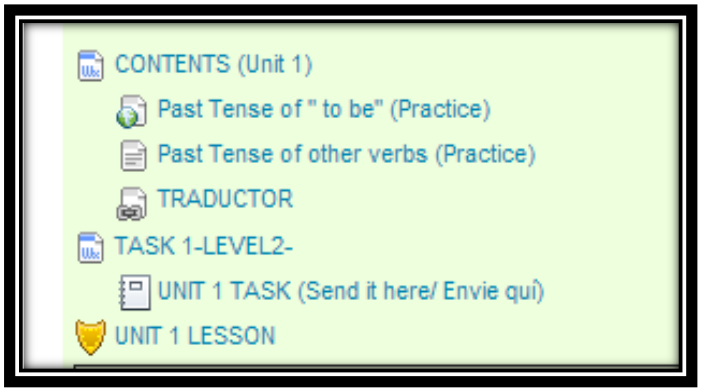


Figura 12: Recursos subidos por el docente

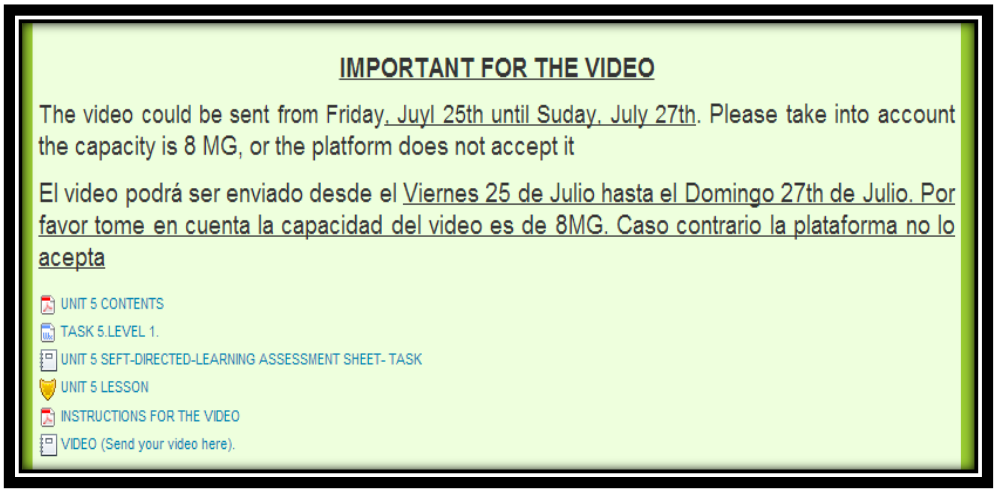


Figura 13: Recurso video.

1.2. MySQL

Figura 14: MySQL [6]

MySQL es un sistema gestor de bases de datos muy conocido y ampliamente usado por su simplicidad y notable rendimiento. El lenguaje de programación que utiliza MySQL es Structured Query Language (SQL) que fue desarrollado por IBM en 1981 y desde entonces es utilizado de forma generalizada en las bases de datos relacionales [6]-[7].

1.2.1. Características [7]:

- Está desarrollado en C/C++.
- Se distribuyen ejecutables para cerca de diecinueve plataformas diferentes.
- La API se encuentra disponible en C, C++, Java, Perl, PHP, Python, Ruby.
- Está optimizado para equipos de múltiples procesadores.

- Es muy destacable su velocidad de respuesta.
- Se puede utilizar como cliente-servidor o incrustado en aplicaciones.
- Su administración se basa en usuarios y privilegios.
- Se tiene constancia de casos en los que maneja cincuenta millones de registros, sesenta mil tablas y cinco millones de columnas.
- Sus opciones de conectividad abarcan TCP/IP, sockets UNIX y sockets NT, además de soportar completamente ODBC.
- Los mensajes de error pueden estar en español y hacer ordenaciones correctas con palabras acentuadas o con la letra 'ñ'.

1.2.2. Ventajas [7]:

- Velocidad al realizar las operaciones, lo que le hace uno de los gestores con mejor rendimiento.
- Bajo costo en requerimientos para la elaboración de bases de datos, ya que debido a su bajo consumo puede ser ejecutado en una máquina con escasos recursos sin ningún problema.
- Facilidad de configuración e instalación.
- Soporta gran variedad de Sistemas Operativos.
- Baja probabilidad de corromper datos, incluso si los errores no se producen en el propio gestor, sino en el sistema en el que está.
- Conectividad y seguridad.

CAPÍTULO II

2. INTERACCIONES DE LOS ESTUDIANTES EN ENTORNOS VIRTUALES

2.1. Educación

La educación es el proceso de facilitar el aprendizaje, conocimientos, habilidades, valores y hábitos de un grupo de personas son transferidos a otras personas, a través de la narración, la discusión, la enseñanza, la formación o la investigación. Además es aquella efectuada por maestros profesionales, esta se vale de las herramientas que postula la pedagogía para alcanzar sus objetivos, en general, esta educación suele estar dividida según las áreas del saber humano para facilitar la asimilación por parte del educando [5].

Figura 15: Página principal del curso MED

2.1.1. Modalidad del proceso de enseñanza aprendizaje

Modalidades de enseñanza los distintos escenarios donde tienen lugar las actividades a realizar por el profesorado y el alumnado a lo largo de un curso, y que se diferencian entre sí en función de los propósitos de la acción didáctica, las tareas a realizar y los recursos necesarios para su ejecución. Lógicamente diferentes modalidades de enseñanza reclaman tipos de trabajos distintos para profesores y estudiantes y exigen la utilización de herramientas metodológicas también diferentes [4].

Además, útil desde el punto de vista organizativo pues permite la asignación de tareas al profesorado (y, por consiguiente, su valoración en cuanto a volumen de trabajo), la distribución de espacios (aulas, laboratorios, seminarios) y la definición de horarios [5].

2.1.1.1. Tipos de modalidades de estudios [5],[7]:

-**Presencial**: es la forma de educación clásica. El alumno asiste a clase y allí es evaluado de las distintas capacidades.

-**A distancia**: esta modalidad permite a los alumnos compaginar sus estudios con otras actividades, así como estudiar desde una ciudad distinta a la del centro. En este caso, la universidad hace llegar a los alumnos el material que deben estudiar y a través de unos encuentros periódicos o del uso de sistemas online (como el Campus Virtual) se van marcando las pautas que el estudiante debe seguir y se le evalúa.

-**Online**: gracias al auge de las nuevas tecnologías esta modalidad se está imponiendo a la educación a distancia, ya que cada vez la red tiene un papel más importante en la comunicación, y por lo tanto en la relación de los alumnos con los profesores y el centro. Su funcionamiento es muy similar a la formación a distancia, con la diferencia que en los estudios online todo el proceso de formación y comunicación se realiza a través de la red.

-**Semipresencial**: esta modalidad es una combinación de la educación presencial con la educación a distancia u online. En ella se realizan algunas clases, temarios o actividades a distancia mientras que otras tienen lugar de forma presencial.

2.2. Entorno Virtual

Es un conjunto de disposiciones informáticas para la comunicación y el intercambio de información en el que se desarrollan procesos de enseñanza – aprendizaje, en el interactúan fundamentalmente profesores y estudiantes [8].

Se trata de aquellos espacios en donde se crean las condiciones para que el individuo se apropie de nuevos conocimientos, experiencias, además es un espacio o comunidad organizada con el propósito de aprender, en el cual deben estar presentes las funciones pedagógicas, tecnológicas y de organización social educativa [9].

Los Entornos Virtuales permiten el acceso a través de navegadores, protegido generalmente por contraseña, utilizan servicios de la web, disponen de un interface gráfico e intuitivo integrando de forma coordinada y estructurada los diferentes módulos, organización de cursos, calendario, materiales digitales, gestión de actividades, seguimiento del estudiante, evaluación del aprendizaje. Se adaptan a las características y necesidades del usuario. Para ello, disponen de diferentes roles en relación a la actividad que realizan, administrador, profesor, tutor y estudiante [9].

Posibilitan la comunicación e interacción entre los estudiantes y el profesor una de las virtudes que caracterizan a los entornos virtuales, es que el usuario se siente formando parte de un mundo generado por un ordenador, estableciendo contacto con los diferentes objetos que componen estos mundos [9].

2.3. Interacción

En los espacios virtuales son la base para generar instancias formativas basadas en el aprendizaje asistido por computador, es la clave de la calidad de los aprendizajes en línea [10].

Es el núcleo de la actividad de un ambiente de aprendizaje, ya que el conocimiento es generado y construido en conjunto [11].

Se dice que cuando se da el control de navegación a los usuarios para que exploren a voluntad el contenido, multimedia se convierte en interacción [11].

Por lo tanto la interacción es el elemento fundamental para propiciar la cooperación y la colaboración mediante la comunicación, dentro del ambiente de aprendizaje, caracterizándose entre otras cosas por la acción recíproca entre dos agentes: uno virtual (el material de autoaprendizaje) y el estudiante que aprende [11].

2.3.1. Tipos de Interacción

Se pueden definir tres tipos de interacción relacionadas al comportamiento o la forma en que el estudiante interactúa con la plataforma [11]:

- **Interacción conformista,** el nivel de interacción con los recursos de la plataforma es BAJO y se debería presentar recomendaciones básicas, con el fin de incentivar al estudiante a utilizar con mayor frecuencia la plataforma y así incrementar su nivel de participación y obtener más conocimientos.

- **Interacción consciente,** mayor utilización de los recursos, por lo que se concluye que el usuario tiene bien definidos cuáles son sus intereses en cuanto al material. Dicha interacción está relacionada con un nivel de interacción MEDIO.
- **Interacción autónoma,** se relaciona con un nivel de interacción ALTO, ya que los intereses del usuario están basados en la consulta de material novedoso, interacción frecuente en el curso.

2.4. Características de los ambientes interactivos [12]:

- Fomentar la participación de los miembros
- Generar preguntas enfocadas a los objetos de conocimiento
- Aclarar y resolver dudas
- Fomentar que los miembros construyan aportaciones para el grupo
- Proporcionar una navegación en forma natural
- Permitir mensajes instantáneos
- Controlar la cantidad de veces que participa un alumno
- Registrar el rango de actividades en que participa un alumno
- Contar con criterios y medios de evaluación y retroalimentación

Otros aspectos que logran la interacción y generan aprendizaje, son [12]:

- Crear escenarios para trabajos colaborativos, donde el profesor se encargue de facilitar el proceso de enseñanza, estimulando a los alumnos a participar en los foros.
- Crear espacios que permitan a los alumnos el fácil entendimiento de los materiales de estudio.

- Permitir al usuario navegar en el ambiente virtual y acceder a los materiales de estudio.

2.5. Análisis de las interacciones

Este análisis requiere de metodologías que proporcionen datos sobre la intervención de los participantes. Los datos pueden ser recogidos a partir de análisis cuantitativos (número de intervenciones, tareas, exámenes, etc.) pero también es preciso ir más allá y analizar los contenidos del discurso. En este sentido, se precisa combinar datos cualitativos y cuantitativos [10].

Un análisis adecuado de estos datos permite obtener una valiosa información para comprender dichas interacciones, la forma en que se producen, el tipo de interacciones, los factores que las afectan y mejorarlas a futuro de modo de explotar al máximo su valor pedagógico y social [10].

CAPÍTULO III

3. CASOS DE ÉXITO DE MINERÍA DE DATOS

La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, para encontrar modelos claros a partir de los datos, para que este proceso sea efectivo, debería ser automático o semiautomático y el uso de los patrones descubiertos debería ayudar a la toma de decisiones, y por tanto, un beneficio a la organización, convertir datos en conocimiento [17] - [18].

3.1. Casos de éxito aplicando las técnicas de Minería de Datos

3.1.1. Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL [13].

Resumen

La presente investigación está enfocada en la identificación de patrones de comportamiento relacionados con las acciones de los estudiantes que utilizan el entorno virtual de aprendizaje, se realizó un análisis previo de la base de datos correspondiente al periodo académico Abril2011/Agosto2011, para seleccionar exclusivamente las entidades que contienen información útil sobre las acciones que realizan los estudiantes. En la utilización de técnicas de minería se optó por la clasificación: arboles de decisión, reglas de decisión éstas se utilizaron para clasificar los indicadores: participación o interés en el curso y utilización de herramientas por carrera siendo estas informática y abogacía, mientras que el análisis de secuencias, mediante la agrupación o agrupación en clústeres, fue utilizado para conocer los grupos de los estudiantes con características similares en relación al indicador de la utilización de herramientas y para encontrar el estilo de aprendizaje dominante y conocer cómo aprenden los estudiantes mediante el análisis de los estilos.

Descripción del Escenario

La base de datos para la búsqueda de patrones, contiene información de todos los cursos, en las modalidades (Clásica y Abierta) de estudio en la UTPL del periodo correspondiente a Abril2011/Agosto2011, conviene indicar que lo que interesa es la información de los estudiantes de la modalidad Abierta la cual será el foco del desarrollo de la tesis para una representación de indicadores en un modelo de estudiante en base a los patrones de comportamiento en relación con las acciones.

Algoritmos para determinar los comportamientos del estudiante en el EVA

Una de las técnicas que se empleará en la minería para determinar el comportamiento de los estudiantes en base a las acciones que éste realiza sobre el EVA, es la de clasificación, como son los árboles de decisión, reglas de decisión, éstos se utilizan para el indicador de la participación del estudiante en el curso ya que según lo analizado ayudan a predecir una o más variables discretas, basándose en otros atributos del conjunto de datos, el algoritmo hace predicciones, el algoritmo de clústeres de secuencia permite explorar los datos que contienen eventos que pueden vincularse mediante rutas o secuencias. El algoritmo encuentra las secuencias más comunes mediante la agrupación, o agrupación en clústeres, de las secuencias que son idénticas, es utilizado para el indicador de la utilización de las herramientas analizando de manera individual cada atributo que representan las acciones de los estudiantes para de esta manera proceder a establecer los grupos y determinar cual hace más uso de las herramientas dentro de la plataforma.

TABLA I.

ALGORITMOS PARA EXPERIMENTACIÓN

Algoritmos	Opción
✓ C4.5 o J48 ✓ REPtree	Trees
✓ BayesNet ✓ NaiveBayes	Classifiers
✓ JRip	Rules
✓ SimpleKMeans ✓ FarthestFirst	Clústeres

Análisis de los Datos

En esta dimensión se enfocará concretamente a la interacción personal del estudiante con la plataforma y es usada para conocer los comportamientos de los estudiantes, según las acciones que realizan en el EVA, previo a esto se realiza un análisis de los atributos detectados en cada curso detallados en tabla I que en realidad sirven, ya que a partir de éstos se derivan algunas variables que pueden ser incluidas las cuales son:

- número de accesos al curso.
- número de accesos a las tareas enviadas por el profesor
- número de veces que revisa un cuestionario
- número de veces que descarga un recurso.
- número de veces que el usuario accede a un foro.
- número de mensajes que envía el usuario.
- número de veces que envía o sube una tarea.

Escenario de experimentación

Los algoritmos y datos que se utilizaron en las experimentaciones que se realizaron con la herramienta WEKA, dentro de la minería de datos se detallan a continuación (ver Figura 16):

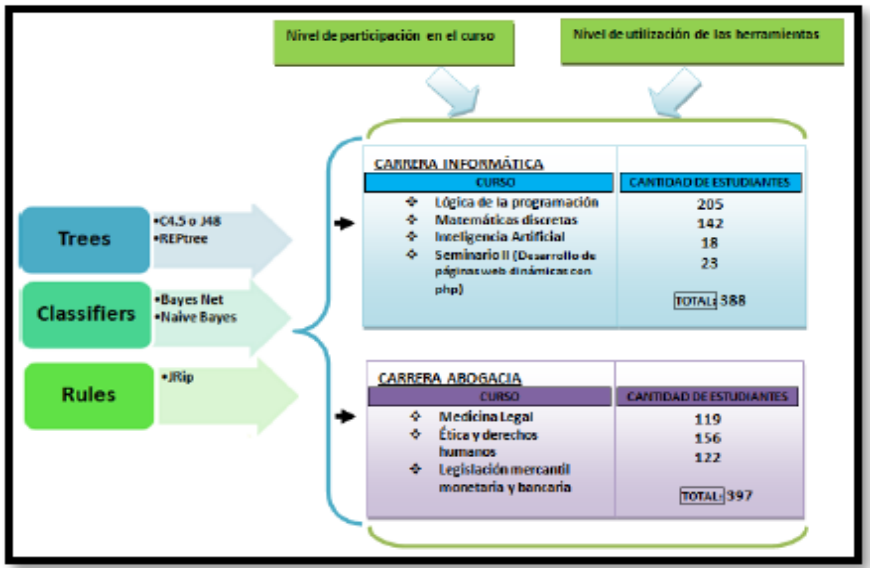


Figura 16: Escenario para el nivel de participación en el curso y el nivel de utilización de las herramientas por carrera.

Los algoritmos C4.5 o J48 (Trees) y JRip (Rules o reglas de decisión) para la experimentación, los cuales forman parte de la técnica de clasificación, estos serán aplicados a cada una de las carreras para encontrar dos indicadores: el nivel de participación en el curso y el nivel de utilización de las herramientas, en el caso de informática se cuenta con 388 instancias o cantidad de estudiantes mientras que en abogacía 397 instancias, a partir de la experimentación se seleccionará el algoritmo que presente los mejores resultados para ser sujetos a análisis e interpretación y posteriormente proceder a realizar comparaciones entre los resultados obtenidos de ambas carreras, previo a esto se realizó una descripción de los algoritmos que se utilizarán y un análisis de los datos de cada curso de ahí se obtuvieron algunas variables.

Minería de Datos

En la minería de datos, es necesario tener los datos, ya generados como atributos, además de que en esta fase se evaluaran 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.

Resultados

Para proceder con la experimentación, se la realiza con la aplicación de cada uno de los algoritmos que son aplicados para los indicadores del nivel de participación en el curso y la utilización de las herramientas, para los cuales se tomó los atributos de las acciones que realizan los estudiantes sobre la plataforma, para determinar cuál de los algoritmos resultó más eficiente para el indicador a evaluar, se lo hace bajo el parámetro de la razón de precisión, que es el número de clasificaciones correctas que nos arroja en los resultados del algoritmo. A continuación se muestran los resultados de los algoritmos que presentaron los mejores resultados, siendo escogidos en base a la mejor precisión en la clasificación de instancias.

TABLA II.

ALGORITMO SELECCIONADO PARA CADA INDICADOR

Indicador	Algoritmo Seleccionado
Nivel de participación en el curso	REPTree
Nivel de utilización de herramientas	REPTree

Resultados con Simplekmeans

Con el algoritmo caracterizado por su sencillez SimpleKMeans además de ser el más utilizado para hacer clustering y al extenso material bibliográfico con el que se cuenta sobre este algoritmo, teniendo en cuenta que es necesario realizar varios experimentos con este para obtener el resultado más óptimo posible, hacemos uso de éste ya que nos interesa evaluar cada atributo que conforma el nivel de utilización de herramientas para lo cual hemos elegido dos grupos de población, que hacen referencia a dos asignaturas en específico Lógica de la programación y Ética y derechos humanos a estas asignaturas se les aplicará el algoritmo, utilizando los valores reales obtenidos para cada atributo que conforman este indicador del nivel de utilización de las herramientas de la plataforma conforme las acciones que éstos realizan sobre las mismas, de tal manera para posteriormente poder proceder a realizar las comparaciones pertinentes entre estas dos asignaturas de la carrera de Informática.

TABLA III.

COMPARACIÓN DE GRUPOS

Nivel de utilización de las herramientas					
Herramienta	Grupos – Clúster	Lógica de la programación		Ética y derechos humanos	
		Instancias	Nivel	Instancias	Nivel
Foros	0	18%	Permanente	26%	Permanente
	1	40%	Moderado	14%	Escaso
	2	42%	Escaso	60%	Moderado
Recursos	0	18%	Permanente	26%	Moderado
	1	40%	Moderado	14%	Permanente
	2	42%	Escaso	60%	Escaso
Tareas	0	18%	Permanente	26%	Permanente
	1	40%	Moderado	14%	Moderado
	2	42%	Escaso	60%	Escaso
Mensajería	0	18%	Permanente	26%	Permanente
	1	40%	Moderado	14%	Moderado
	2	42%	Escaso	60%	Escaso
Twitter	0	18%	Permanente	26%	Moderado
	1	40%	Moderado	14%	Permanente
	2	42%	Escaso	60%	Escaso
Cuestionario	0	18%	Permanente	26%	Permanente
	1	40%	Moderado	14%	Moderado
	2	42%	Escaso	60%	Escaso

Los estudiantes que presentan más interacción con la herramienta (foros) pertenecen al curso de ética y derechos humanos, mientras que en los estudiantes del curso de lógica de la programación presentan más interacción con las herramientas (foros, recursos, tareas, mensajería, twitter y cuestionario). Denotando claramente a nivel general que de entre los dos cursos, la utilización de las herramientas permanentes y moderadas, representan mayor interacción los estudiantes que corresponden al curso lógica de la programación (informática).

Conclusiones:

- Con la minería de datos, a partir de la información se descubre y analiza nuevo conocimiento útil, para que un sistema cuente con una representación del estado

actual del usuario, que en este caso es la descripción de las acciones del estudiante, representado en un modelo que se obtiene a partir de la personalización y análisis de datos, que son aplicaciones propias del modelado de usuario.

- 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), y 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, tareas.

3.1.2. Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje [14].

Resumen.

La adopción de sistemas recomendadores en ambientes virtuales de aprendizaje se está convirtiendo en una alternativa; para lograr la adaptación automática requerida, para atender las necesidades de aprendizaje de los estudiantes. Con los datos de interacción, que proveen estos ambientes es posible encontrar indicadores que con la aplicación de técnicas de minería de datos y aprendizaje automático se pueda identificar información relevante, para la definición de recomendaciones. Se ha aplicado técnicas de aprendizaje no supervisado, para la identificación de patrones comunes de interacción con los foros

disponibles en un curso de la plataforma OpenACS/dotLRN. Esto facilitará la definición de recomendaciones que ayuden a mejorar la experiencia de aprendizaje de los estudiantes.

Descripción del escenario

El experimento fue desarrollado con 30 estudiantes del primer ciclo de la carrera de Ingeniería en Sistemas Informáticos y Computación de la UTPL, del total de estudiante enrolados 29 ingresaron en la plataforma. Esta experimentación fue desarrollada desde el 23 de noviembre hasta el 15 de diciembre de 2009.

Los estudiantes tuvieron acceso al curso de “Fundamentos Informáticos”, en el cuál realizaron las siguientes actividades:

- Interacción con las herramientas disponibles en la plataforma: foros, calendario, noticias, preguntas frecuentes, tareas, documentos compartidos.
- Durante el proceso de interacción, los estudiantes revisaron los archivos y enlaces subidos por el profesor/tutor a la plataforma.
- En base a las lecturas realizadas de los documentos compartidos, desarrollaron dos tareas que luego fueron subidas a la plataforma en la sección ¨Carpeta de tareas de la comunidad de Fundamentos Informáticos¨.
- Participaron de los foros propuestos en el curso.

Selección de los patrones de interacción

Para la selección de los patrones de interacción con los que definir las recomendaciones, fue necesario conocer en qué situación o condición es necesario sugerir la recomendación, aspecto que fue complementado con la aplicación de la técnica de clustering.

Aplicación de la técnica de clustering

Con el objeto de buscar semejanzas en el valor de ciertos atributos que son compartidos por los estudiantes en el momento de la interacción con la plataforma, es necesario trabajar con algoritmos que faciliten esta tarea.

En base a esto y considerando el estudio realizado se ha creído conveniente trabajar con la técnica de clustering que permite agrupar estudiantes en subclases de acuerdo a su nivel de participación y semejanza de acceso a la plataforma.

Con esta técnica se procedió a agrupar a los estudiantes del curso en diferentes grupos relacionados con las actividades realizadas en los foros, y así descubrir patrones que reflejen comportamientos análogos en los estudiantes. Para llevar a cabo este proceso se consideraron los indicadores estadísticos de la interacción de los estudiantes con los foros y mediante la aplicación de la técnica de clustering se pudo observar la relación entre estos indicadores y la participación/colaboración de los estudiantes en esta herramienta.

Selección de los patrones de interacción

Para la selección de los patrones de interacción con los que definir las recomendaciones, fue necesario conocer en qué situación o condición es necesario sugerir la recomendación, aspecto que fue complementado con la aplicación de la técnica de clustering.

El conjunto de datos con los indicadores estadísticos de las interacciones de los estudiantes, está formado por:

- Número de sesiones.
- Número de visitas a los foros.
- Número de visitas a mensajes.
- Número de mensajes enviados.
- Número de conversaciones en los que ha participado el estudiante y el profesor.
- El número de respuestas a los mensajes enviados por el usuario.
- El número de respuestas del usuario.

Selección del modelo

Se procedió a crear un modelo con la técnica de clustering usando el conjunto de datos tomados de la interacción con los foros. La herramienta seleccionada para esto fue WEKA3, que es un software de libre distribución de código abierto y que brinda la facilidad de aplicar diferentes técnicas de aprendizaje y minería de datos.

La evaluación del modelado se realizó comparando y analizando los resultados con varias técnicas e interpretando su significado. El proceso fue el siguiente:

Con el algoritmo SimpleKMeans se realizaron tres experimentaciones con diferentes valores de semilla (centros iniciales del clúster). En la primera experimentación se consideró el valor por defecto 10, en la segunda se incrementó a 20, y en la tercera a 100, esto con el propósito de ir mejorando los resultados obtenidos. El algoritmo fue configurado para obtener 3 clústeres.

Luego de haber aplicado este algoritmo se realizó una evaluación de los resultados obtenidos en la tres experimentaciones para determinar cuál es el mejor, con lo cual se determinó que los resultados de la primera experimentación no eran muy convenientes, ya que los grupos eran muy disparejos; en la segunda experimentación se obtuvieron mejores resultados, sin embargo la agrupación no era tan clara, ya que los clústeres no quedaban bien definidos; en la tercera experimentación es donde se obtuvieron mejores resultados, con un valor más bajo en la suma de los cuadrados de los errores.

Resultados de la experiencia

De igual forma, se realizaron tres aplicaciones del algoritmo EM con diferentes valores de semilla, obteniendo resultados similares en las tres experimentaciones, por ello, para la evaluación se presenta la aplicación realizada con los valores por defecto del modelo (semilla 100, clústeres=3), que coinciden con los de la tercera experimentación del algoritmo SimpleKMeans.

TABLA IV.

COMPARACIÓN DE RESULTADOS ENTRE SIMPLE KMEANS Y EM

Algoritmo	SimpleKMeans			EM		
Clúster	0	1	2	0	1	2
Instancias	12	12	6	13	15	2
% de instancias clasificadas	40	40	20	43	50	7

Al analizar en forma individual los valores encontrados para cada atributo con ambos algoritmos, y considerando la distribución de las instancias en cada clúster, se decidió trabajar con los centroides de los atributos obtenidos con SimpleKmeans (ver figura 17), que fue el que presentó una mejor clasificación en su tercera aplicación, ya que los clústeres obtenidos mediante este algoritmo presentaron mayor consistencia y similitud entre sus características. Los valores de estos centroides serán los que se consideren para la definición de las recomendaciones.

```
Clusterer output

kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 11.028945683637478
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                     Full Data        0          1          2
                                   (30)     (12)       (12)        (6)
=================================================================
num_sessions                        5     2.9167     8.5833          2
average_time_per_session      545.5478   343.9653   342.9874  1353.8333
num_forums_visited            13.4333     9.0833      19.25       10.5
num_msg_to_group_forum         1.2333          0     2.4167     1.3333
num_msg_visited                6.6333     4.3333     9.0833     6.3333
num_average_replies_threads      0.72          0     1.1333     1.3333
num_threads_started            1.5333     2.0833     0.9167     1.6667
num_threads_user_and_tutor     0.4333     0.3333     0.5833     0.3333
num_answers_to_user_msg        0.8667     0.1667     0.5833     2.8333
num_answered_threads           0.9667          0     1.8333     1.1667
```

Figura 17. Tercera aplicación del algoritmo SimpleKMeans

Los resultados obtenidos del procesamiento del archivo de log de la interacción con los foros se pueden observar en la figura 16, donde, de los 29 estudiantes que ingresaron a la plataforma, 28 interactuaron con los foros. El estudiante que resta sólo presentó accesos a la plataforma pero no registró visitas o envíos de mensajes a los foros propuestos.

TABLA V.

RESUMEN DE LAS INTERACCIONES CON LOS FOROS

Resultados de la interacción en los foros	
Interacción realizada	Número de estudiantes
Visitas a los foros	28
Visitas a los mensajes	27
Envío de mensajes	25
Inicio de conversaciones (Hilos)	17
Envío de mensajes dentro de los hilos	18
Interacción en conversaciones con el profesor	11
Respuestas a hilos	18

Definición de las recomendaciones

Para la definición de las recomendaciones hay que conocer las condiciones o situaciones en las que es necesario ofrecer las recomendaciones. Estas fueron determinadas teniendo en cuenta el análisis de los datos recogidos con los instrumentos de medición, el procesamiento del archivo de log y el análisis realizado de la aplicación de los algoritmos para los clústeres obtenidos.

En el cuadro siguiente se describe las recomendaciones identificadas durante este proceso, con la información obtenida para cada situación mediante las técnicas de clustering.

TABLA VI.

IDENTIFICACIÓN DE CUÁNDO ES NECESARIO OFRECER UNA RECOMENDACIÓN

Recomendación	Situaciones en la que es necesaria la recomendación	Información obtenida mediante clustering
• Postear un mensaje en el foro compartiendo un documento o un enlace.	• Cuando el estudiante no ha compartido ningún documento/ enlace.	El alumno se conecta a la plataforma más de 2 sesiones.
• Inicie una nueva conversación/hilo.	• En caso de que tenga un bajo número de contribuciones.	Número de contribuciones menor a 2.
• Ingresar mensajes en los hilos iniciados por los compañeros.	• Cuando el número de contribuciones del estudiante en el foro es bajo y el hilo tenga una tasa baja de contribuciones (menor a X).	Numero de contribuciones del estudiante menor a 2 y número de contribuciones que hay en el hilo X sea menor a 1.
• Dar respuesta a los mensajes comentados por los compañeros.	• Cuando otro estudiante haya comentado su mensaje.	Número de contribuciones menor a 2.
• Leer el mensaje del foro propuesto por el profesor.	• Cuando aún no haya leído el mensaje propuesto por el profesor en el foro.	Número bajo de contribuciones en el foro, menor a 2.
• Leer los mensajes de los compañeros.	• Cuando no haya visitado los mensajes posteados por los compañeros.	Número bajo de contribuciones en el foro, menor a 2.
• Visitar las conversaciones/hilos donde se han ingresado más contribuciones.	• Cuando un determinado hilo tenga un número alto de contribuciones.	El hilo cuenta con un número alto de contribuciones, igual a 9.

Conclusiones

- El propósito de esta investigación ha sido encontrar valores para las condiciones de aplicación de las recomendaciones, en base a la aplicación de técnicas de aprendizaje no supervisado considerando como fuente de datos las interacciones de los estudiantes en el curso de Fundamentos Informáticos de la carrera de Ingeriría en Sistemas Informáticos y Computación de la UTPL, con el fin de mejorar el soporte adaptativo ofrecido en un escenario de aprendizaje online. Estos estudiantes estuvieron inscritos en un curso en la plataforma

OpenACS/dotLRN, cuyo acceso fue proporcionado por los integrantes del grupo aDeNu de la UNED en la cual se encontraba instalada.

- Un conjunto de indicadores estadísticos se derivaron de las interacciones de los estudiantes en los foros propuestos, ya que son la principal herramienta para dar soporte al tipo de recomendación de "Leer y Postear mensajes en los foros", seleccionada para esta investigación.

- Una vez determinados los indicadores se tenía que seleccionar la técnica más apropiada para encontrar características similares en el comportamiento de los estudiantes. La técnica seleccionada fue la de clustering.

3.1.3. Uso de ambientes virtuales de aprendizaje en la enseñanza de la ingeniería [15].

Resumen

El aprendizaje de la ingeniería implica la adquisición de competencias en el modelado de fenómenos y procesos, así como en la comunicación mediante el uso adecuado del lenguaje, tanto en el contexto cotidiano como en el científico y de la profesión. La visión pedagógica de la ingeniería está dirigida principalmente a la resolución de problemas.

Este enfoque y la concepción de Ambientes Virtuales de Aprendizaje (AVAs) orientados a la ingeniería son apropiados para la construcción de espacios interactivos de enseñanza/aprendizaje, considerando principalmente la versatilidad determinada por la modalidad virtual. En el presente artículo se presentan experiencias llevadas a cabo en la Facultad de Ingeniería de la Universidad Nacional de Mar del Plata, se introducen las mejoras que se proponen a la plataforma Moodle con técnicas de la Minería de Datos.

Experiencia de Aula Extendida

En la Facultad de Ingeniería de la Universidad Nacional de Mar del Plata se están llevando a cabo experiencias en la plataforma Moodle. Diversas asignaturas están

implementando cursos con contenidos educativos originales, producidos por los docentes involucrados. A continuación, se presentan algunos ejemplos. La modalidad conocida como “extended learning” es decir, aprendizaje extendido, permite transformar el aula tradicional en un “aula extendida”, brindándole al alumno la posibilidad de complementar su capacitación presencial tradicional con el desarrollo de actividades en forma virtual, es decir, mediadas por tecnología.

En la presente experiencia, se hace uso de la modalidad de aula extendida para trabajar en temas del currículo de Electrotecnia, ciencia tecnológica básica imprescindible en varias carreras de ingeniería, en particular en Ingeniería Eléctrica y Electromecánica. Los destinatarios son estudiantes de segundo y tercer año de las carreras de Ingeniería Eléctrica y Electromecánica, que cursan las primeras materias asociadas directamente con la especialidad, en el agrupamiento de asignaturas denominado de Tecnologías Básicas.

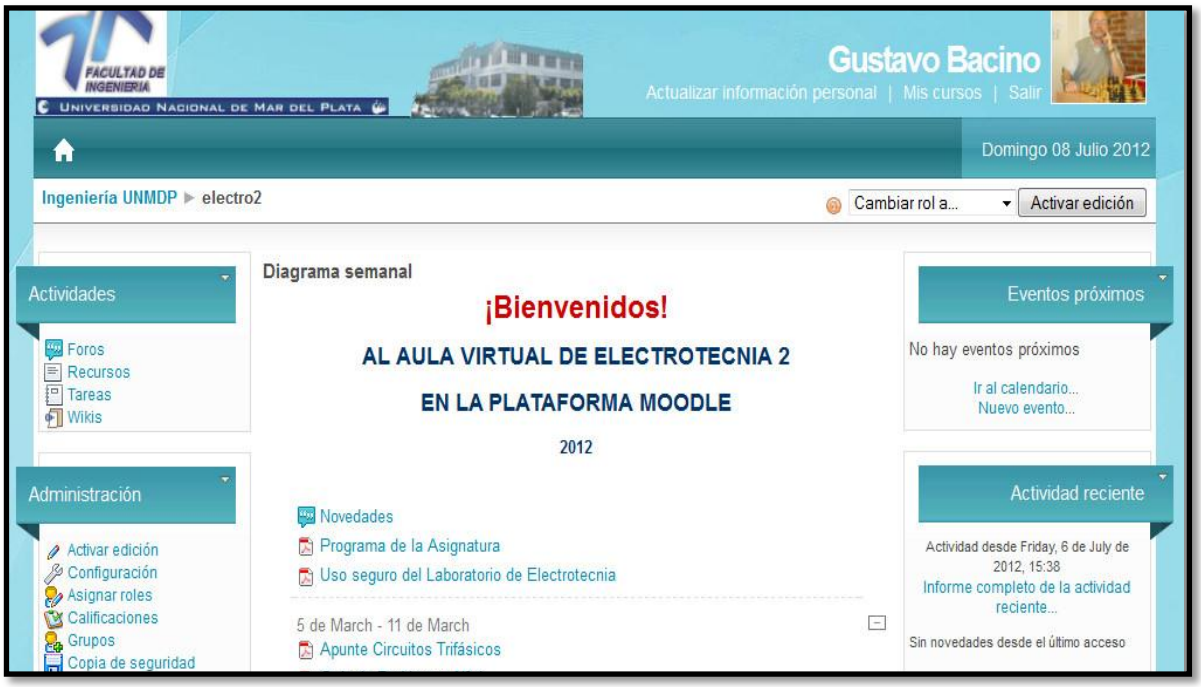


Figura 18. Presentación del aula extendida para el curso de Electrotecnia.

Experiencia con Técnicas de Minerías de Datos para Mejorar los AVAs en Ingeniería

Una experiencia llevada a cabo en la Facultad de Ingeniería comprendió el diseño y prueba de un modelo basado en las técnicas de análisis de clúster con el objetivo específico de aplicarlo al estudio de ciertas habilidades cognitivas de los alumnos desde un punto de vista grupal partiendo de sus características individuales.

En particular se aplicó en la asignatura Computación, común a todas las especialidades de Ingeniería, la asignatura se focaliza en el aprendizaje de la programación de computadoras, requiere que el alumno adquiera ciertas habilidades, como, construir un algoritmo, resolver un problema planteado, traducirlo a un lenguaje de programación específico y probarlo en la computadora. Los errores que el alumno puede cometer se clasifican en sintácticos y semánticos. La prueba en computadora detecta errores sintácticos, pero los semánticos deben ser analizados más profundamente pues abarcan una amplia gama de equivocaciones.

Dada la naturaleza del problema se eligió una técnica de clustering para investigar perfiles emergentes en grupos numerosos de alumnos a partir de diagnósticos cognitivos individuales. Inicialmente se probó uno de los algoritmos más conocidos de clustering, K-Medias. Este método separa los datos en un número determinado K (fijado a priori) de grupos independientes. Se definen K posibles centros y siguiendo algún criterio de proximidad preestablecido se sitúa cada objeto en un grupo según su ubicación con respecto al centro. Luego de esta primer partición se recalculan los centros de cada grupo y se redistribuyen los objetos por proximidad con este nuevo centro. El proceso se repite hasta que no ocurran cambios en los grupos de un paso al siguiente. La clasificación obtenida es de tipo dura, esto es, cada objeto pertenece a un grupo específico, logrando grupos disjuntos.

CAPÍTULO IV

4. MINERÍA DE DATOS

4.1. Minería de datos

Siempre se ha dicho que no tener información acerca de algún hecho es lo peor que puede ocurrir, pero esto no es cierto, peor es no poseer ningún tipo de información, es tener numerosa información a disposición y no saber qué hacer con ella. Para que este problema no ocurra, existe el concepto denominado Minería de Datos, es un conjunto de técnicas que se aplican para extraer información que se encuentran de manera implícita en los datos, pero que no se observan a simple vista, sino que es necesario preparar estos datos previamente para obtener este conocimiento [16].

Figura 19: Minería de Datos

Es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, para encontrar modelos claros a partir de los datos, para que este proceso sea efectivo, debería ser automático o semiautomático y el uso de los patrones descubiertos debería ayudar a la toma de decisiones, y por tanto, un beneficio a la organización, convertir datos en conocimiento [17] - [18].

El resumir datos para la toma de decisiones ha sido el campo tradicional de la estadística pero hoy en día existen nuevas técnicas, una de ella es la Minería de Datos [19].

Minería de Datos es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica y las Bases de Datos, principalmente usando como

materia prima las bases de datos, que tienen como propósito la identificación de un conocimiento que aporten un sesgo hacia la toma de decisiones [20] - [21].

4.2. Aplicaciones de minería de datos

El uso de la Minería de Datos puede ser beneficioso en el caso de poseer datos sobre sus procesos productivos, datos de seguimiento de clientes, datos externos de mercado, datos sobre la actividad de competidores, entre otros [22]-[23].

- **Educación**
  - Selección o captación de estudiantes
  - Detección de abandonos o fracasos
  - Estimación del tiempo de estancia en la institución

- **Análisis de Mercado y Administración**
  - Perfil del cliente
  - Descubrir las relaciones entre características personales y el tipo de productos que se compran.

- **Finanzas**
  - Compañías de inversión hacen transacciones en la bolsa de valores basándose en resultados de Minería de Datos.
  - Predicción de flujo de efectivo

- **Detección de fraude**
  - Utilizan bases de datos históricas para crear modelos de comportamiento fraudulento y utilizar Minería de Datos para identificar nuevos fraudes.

- **Deportes**
  - Para interpretar las estadísticas

- **Web**
  - Analizar log en general

- Analizar el comportamiento de los usuarios de un sitio

❖ **Medicina**

- Aplicaciones que buscan nuevos medicamentos
- Análisis de secuencias de genes
- Predecir si un compuesto causa cáncer
- Análisis de secuencias de proteínas
- Identificación de terapias médicas para diferentes enfermedades

4.3. Ventajas de la minería de datos [24]:

- ❖ Enormes bases de datos pueden ser analizadas.
- ❖ Descubrir información que no se esperaba obtener.
- ❖ El modelo es probado y comprobado usando técnicas antes de ser usado, luego las predicciones que se obtienen por el modelo son válidas y confiables.
- ❖ Contribuye a la toma de decisiones de forma estratégica.
- ❖ A largo plazo, ahorra dinero a la empresa.
- ❖ Genera modelos descriptivos, qué datos influyen en los resultados finales.

4.4. Diferencia entre Minería de Datos y Estadística

La Minería de Datos es el descendiente y el sucesor de la estadística, conducen al mismo objetivo, el de efectuar “modelos” compactos y comprensibles que ayuden a mejorar la toma de decisiones mediante un conocimiento del entorno [25].

Fundamentalmente, la diferencia entre ambas reside en que las técnicas del Minería de Datos construyen el modelo de manera automática mientras que las técnicas estadísticas necesitan ser manejadas y orientadas por un estadístico, proporcionando una gran libertad a los usuarios profesionales [25].

Pero es importante aclarar que la estadística se utiliza para validar un modelo sugerido y preexistente, no para generarlo [25].

La Minería de Datos aventaja a la estadística en los siguientes aspectos:

- Las técnicas estadísticas se centran generalmente en técnicas confirmatorias, mientras que las técnicas de Minería de Datos son generalmente exploratorias.

Así, cuando el problema al que se pretenda dar respuesta es refutar o confirmar una hipótesis, se podrá utilizar ambas ciencias [19].

- A mayor dimensión del problema la Minería de Datos ofrece mejores soluciones [19].
- Las técnicas de Minería de Datos son menos restrictivas que las estadistas [19].

Contextos en los que es más adecuado el análisis estadístico que el de Minería de Datos [26]:

- Si se pretende determinar cuáles son las causas de ciertos efectos, se debe utilizar técnicas de estadística (por ejemplo, ecuaciones estructurales).
- Se pretende generalizar sobre poblaciones desconocidas en su globalidad. Si las conclusiones han de ser extensibles a otros elementos de poblaciones similares habrán de utilizarse técnicas de inferencia estadística, muestras. En Minería de Datos, se generarán modelos y luego habrán de validarse con otros casos conocidos de la población, utilizando como significación el ajuste de la predicción sobre una población conocida, Minería de Datos y estadística son técnicas complementarias que permiten obtener conocimiento inédito en los almacenes de datos o dar respuestas a cuestiones concretas.

4.5. Técnicas de minería de datos

Las técnicas de la Minería de Datos son [16]:

- **Técnicas supervisadas o predictivas:** Utilizar algunas variables o campos en una base de datos para predecir valores desconocidos o futuros de tal manera que especifican el modelo para los datos en base a un conocimiento previo.

- **Técnicas no supervisadas o descriptivas:** Encontrar patrones que describan la información (interpretables por el hombre).

Las técnicas de Minería de Datos serán utilizadas con el objetivo de obtener la información oculta en grandes cantidades de datos, las cuales son descritas a continuación.

4.5.1. Agrupamiento o Clustering

Se agrupan datos dentro de un número de clases, se puede realizar mediante criterios de distancia o similitud, de forma que si las clases son similares entre sí estén agrupadas [27].

Consiste en ordenar, agrupar o dividir fenómenos complejos en pequeños basándose en la similitud de los valores de los atributos de los distintos datos, para permitir un mejor control o comprensión de la información [14].

Su aplicación a sistemas permite agrupar a los usuarios por su comportamiento de navegación, agrupar a las páginas por su contenido, tipo o acceso y agrupar los comportamientos de navegación similares [28].

Como función de la minería de datos, el análisis de clúster puede ser utilizado como una herramienta independiente para obtener una visión de la distribución de los datos, para observar las características de cada clúster y enfocar un análisis hacia un grupo o clúster determinado [29].

Al hacer clústeres, se puede identificar regiones densas y regiones dispersas en el espacio de características, y por lo tanto, descubrir distribución de patrones y correlaciones entre los atributos [30].

Pasos del análisis clúster [31]:

- Se tiene información de n casos y k variables.
- Se describen los grupos obtenidos y se comparan unos con otros.
- Validación del análisis.
- Se crean los grupos de acuerdo a la medida de similitud.

Métodos de Agrupamiento [32]:

- Jerárquicos: los datos se agrupan de manera arborescente.
- No jerárquicos: generar particiones a un nivel.
- Paramétricos: se asumen que las densidades condicionales de los grupos tienen cierta forma paramétrica conocida y se reduce a estimar los parámetros.
- No paramétricos: no asumen nada sobre el modo en el que se agrupan los objetos.

4.5.2. Clasificación

Permite una organización eficiente de un conjunto de datos, debido a que los árboles son construidos a partir de la evaluación del primer nodo raíz y de acuerdo a su evaluación o valor tomado se va descendiendo en las ramas hasta llegar al final del camino u hojas del árbol [15].

Los arboles de decisión son estructuras que representan conjuntos de decisiones que generan reglas para la clasificación de un conjunto de datos. Entre los algoritmos que aplica es el J48, ID3, entre otros [13].

Son útiles para explorar un conjunto de datos y entender cómo ciertas variables de las interacciones de los estudiantes con el Entorno Virtual de Aprendizaje inciden sobre otra [33].

4.5.3. Reglas de asociación

En minería de datos las reglas de asociación con base de datos se evalúan de acuerdo al soporte y a la confianza de las mismas, se utilizan para encontrar hechos que ocurren en común dentro de un conjunto de datos. Dicho de otra manera deben ocurrir ciertas condiciones para que se produzca cierta condición, también para buscar por medio de conjunto de datos reglas que revelan la naturaleza de las relaciones o asociaciones entre datos de las entidades [34].

Se aplican en el análisis de la canasta de mercado, marketing cruzado con correo, diseño de catálogos, segmentación de clientes respecto a las compras y el soporte para la toma de decisiones [35].

4.6. Algoritmos de la técnica de clasificación

4.6.1. Jrip

Construye un conjunto de reglas, las reduce usando la técnica heurística, con un conjunto de reglas de entrenamiento por separado y luego optimiza al mismo tiempo ese conjunto de reglas [36].

Utiliza un conjunto de reglas separadas para decidir podar reglas, utiliza ganancia de información, para crecer las reglas, la medida para podar reglas, esto es una medida basada en un conjunto global de reglas [36].

A continuación se presenta resultados arrojados por el presente algoritmo (ver Figura 20):

```
J Rip

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:                                                  meteorologia-
weka.filters.unsupervised.attribute.Remove-R1,3-5,9,14-15-
weka.filters.unsupervised.instance.Resample-S1-Z9.0
Instances:    10964
Time taken to build model: 35.45 seconds
=== Summary ===
Correctly Classified Instances          718               63.3157 %
Incorrectly Classified Instances        416               36.6843 %
Kappa statistic                           0.3863
K&B Relative Info Score               32617.5444 %
K&B Information Score                   769.1483 bits       0.6783
bits/instance
Class complexity | order 0             2415.7786 bits       2.1303
bits/instance
Class complexity | scheme              4961.5706 bits       4.3753
bits/instance
Complexity improvement       (Sf)      -2545.792   bits      -2.245
bits/instance
Mean absolute error                       0.0761
Root mean squared error                   0.1917
Relative absolute error                  78.6888 %
Root relative squared error              88.8366 %
Total Number of Instances             1134
Ignored Class Unknown Instances                  7
```

Figura 20: Resultados del algoritmo JRIP [37]

4.6.2. Ridor

Genera una regla por defecto y luego toma las excepciones para la regla predeterminando con la mínima tasa de error. Entonces genera la mejor excepción para cada excepción iterando hasta lograr disminuir el error. Luego genera una expansión similar a un árbol de excepciones. La excepción es un conjunto de reglas que predice clases. Este algoritmo es usado para generar dichas excepciones [36].

A continuación se presenta resultados arrojados por el presente algoritmo, en el cual se puede observar las reglas generados por el mismo (ver Figura 21):

```
W-Ridor

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

numerointeracciones = bajo  (352.0/254.0)
          Except (interaccionesexamen = IEM) and (interaccionestareas = ITM) => numerointeracciones = medio  (56.0/0.0) [19.0/0.0]
          Except (interaccionesexamen = IEM) and (ciudad = O) and (numeroHijos = No) => numerointeracciones = medio  (16.0/0.0) [5.0/0.0]
          Except (interaccionesexamen = IEA) => numerointeracciones = medio  (41.0/0.0) [15.0/0.0]
          Except (interaccionesexamen = IEM) => numerointeracciones = medio  (55.0/5.0) [35.0/7.0]
          Except (interaccionesrecurso = IRM) and (ciudad = L) => numerointeracciones = medio  (5.0/0.0) [1.0/0.0]
          Except (interaccionesrecurso = IRA) => numerointeracciones = medio  (3.0/0.0) [1.0/0.0]
                   Except (carrera = Derecho) and (interaccionesexamen = IEA) => numerointeracciones = alto  (3.0/0.0) [2.0/1.0]
          Except (interaccionesrecurso = IRM) => numerointeracciones = medio  (6.0/3.0) [2.0/0.0]
          Except (carrera = AdministraciÃƒÂ³n de Empresas) and (edad = c) => numerointeracciones = medio  (3.0/2.0) [1.0/0.0]

Total number of rules (incl. the default rule): 10
```

Figura 21: Resultados del algoritmo Ridor [37]

4.6.3. Part

Genera una lista de decisión sin restricciones usando el procedimiento de divide y vencerás. Además construye un árbol de decisión parcial para obtener una regla. Para poder podar una rama (una regla) es necesario que todas sus implicaciones sean conocidas. El PART evita la generalización precipitada, y usa los mismos mecanismos que el C4.5. La hoja con máxima cobertura se convierte en una regla y los valores ausentes de los atributos se tratan como en el C4.5, es decir, la instancia se divide en piezas. En cuanto al tiempo máximo para generar una regla, es el mismo que para construir un árbol podado, y esto ocurre cuando los datos tienen ruido. En el mejor de los casos el tiempo necesario es el mismo que para generar una regla sencilla, y esto se da cuando los datos no presentan ruido [37].

A continuación se presenta resultados arrojados por el presente algoritmo (ver Figura 22):

```
Number of Leaves  :    9

Size of the tree :    17


Time taken to build model: 0.11 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         100               99.0099 %
Incorrectly Classified Instances         1                0.9901 %
Kappa statistic                          0.987
Mean absolute error                      0.0047
Root mean squared error                  0.0486
Relative absolute error                  2.1552 %
Root relative squared error             14.7377 %
Total Number of Instances              101

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 20  0  0  0  0  0 |  b = bird
  0  0  5  0  0  0  0 |  c = reptile
  0  0  0 13  0  0  0 |  d = fish
  0  0  1  0  3  0  0 |  e = amphibian
  0  0  0  0  0  8  0 |  f = insect
  0  0  0  0  0  0 10 |  g = invertebrate
```

Figura 22: Resultados del algoritmo PART [38]

4.6.4. Chaid

Es un acrónimo de Chi-squared Automatic Interaction Detection (detector automático de interacciones mediante Ji cuadrado). Este algoritmo esta desde 1980 y fue desarrollado por Kass, aunque fue diseñado para trabajar sólo con variables categóricas, posteriormente se incluyó la posibilidad de trabajar con variables categóricas, nominales,

categóricas ordinales y variables continuas, permitiendo generar tantos árboles de decisión para resolver problemas de clasificación como árboles de regresión. En este algoritmo los nodos se pueden dividir en más de dos ramas. La construcción del árbol se basa en el cálculo de la significación de un acuerdo estadístico como criterio para definir la jerarquía de las variables predictores o de salida, al igual que para establecer las agrupaciones de valores similares respecto a las variables de salida a la vez que conserva inalterables todos los valores distintos [39].

A continuación se presenta resultados arrojados por el presente algoritmo, donde se puede observar el árbol generado por este algoritmo (ver Figura 23):

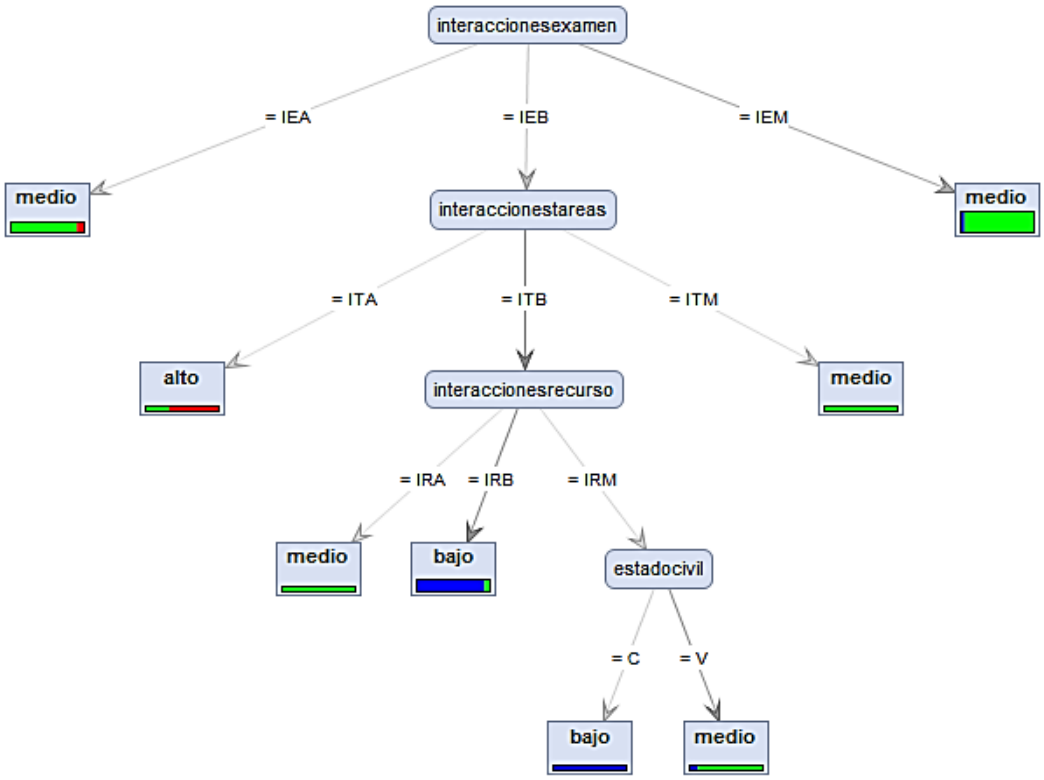


Figura 23: Resultados del algoritmo CHAID [37]

4.6.5. J48

Es un algoritmo de inducción que genera una estructura de reglas o árbol a partir de subconjuntos (ventanas) de casos extraídos del conjunto total de datos de "entrenamiento". En este sentido, su forma de procesar los datos es parecido al de Id3. El algoritmo genera una estructura de reglas y evalúa su "bondad" usando criterios que

miden la precisión en la clasificación de los casos. Emplea dos criterios principales para dirigir el proceso dados por [40]:

1. Calcula el valor de la información proporcionada por una regla candidata (o rama del árbol), con una rutina que se llama “info”.

2. Calcula la mejora global que proporciona una regla/rama usando una rutina que se llama gain (beneficio).

A continuación se presenta resultados arrojados por el presente algoritmo (ver Figura 24):

```
Number of Leaves  :    9
Size of the tree :    17

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          93               92.0792 %
Incorrectly Classified Instances         8                7.9208 %
Kappa statistic                          0.8955
Mean absolute error                      0.0225
Root mean squared error                  0.14
Relative absolute error                 10.2694 %
Root relative squared error             42.4999 %
Total Number of Instances              101

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 41  0  0  0  0  0  0 |  a = mammal
  0 20  0  0  0  0  0 |  b = bird
  0  0  3  1  0  1  0 |  c = reptile
```

Figura 24: Resultados del algoritmo J48 [37]

4.6.6. ID3

Algoritmo desarrollado por J. Ross Quinlan en 1983. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos [15].

ID3 realiza esta labor mediante la construcción de un árbol de decisión.

Los elementos son [15]:

- Nodos: Los cuales contendrán atributos.
- Arcos: Los cuales contienen valores posibles del nodo padre.

• Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

Permite construir un árbol de arriba abajo, de forma directa y no ejecuta vuelta atrás en su búsqueda. Una vez que el algoritmo selecciona un atributo, nunca reconsiderara esta elección. Los dominios de los atributos y de las clases deben ser discretos. Usa el concepto de ganancia de información para seleccionar el atributo más útil en cada paso [16].

Para decidir qué atributo es el más apropiado a usar en cada nodo del árbol se utiliza una propiedad estadísticas llamada ganancia de información, que mide que tan bien clasifica ese atributo a los datos de entrenamiento. Así que elige el nodo del árbol que tenga mayor ganancia de información y luego expande sus ramas utilizando la misma metodología [17].

A continuación se presenta resultados arrojados por el presente algoritmo de las instancias clasificadas correctamente (ver Figura 25):

accuracy: 98.32%

	true bajo	true medio	true alto
pred. bajo	199	9	0
pred. medio	3	486	0
pred. alto	0	0	19

Figura 25: Resultados del algoritmo ID3 [37]

4.6.7. PRISM

Es un algoritmo de aprendizaje de reglas que asume que no hay ruido en los datos, es decir que toda la información no presenta distorsión en los datos clasificados.

Una de sus principales ventajas es la no consideración de ruido en los datos y crea reglas que cubre la mayor parte de las observaciones, separando las instancias para analizarlas por separado y cumplir su cometido [36].

A continuación se presenta resultados arrojados por el presente algoritmo de las instancias clasificadas correctamente (ver Figura 26):

accuracy: 98.46%

	true bajo	true alto	true medio
pred. bajo	240	8	0
pred. alto	0	420	3
pred. medio	0	0	45

Figura 26: Resultados del algoritmo PRISM [37]

4.6.8. Decision Tree

Un árbol de decisión es un gráfico o modelo en forma de árbol. El objetivo es crear un modelo de clasificación que predice el valor de un atributo de destino (a menudo llamado clase o etiqueta) basado en varios atributos de entrada de la ExampleSet o data set. En RapidMiner un atributo de etiquetas está en relación con el operador de árbol de decisión. Cada nodo hoja representa un valor del atributo de la etiqueta dados los valores de los atributos de entrada representados por el camino desde la raíz a la hoja [36].

A continuación se presenta resultados arrojados por el presente algoritmo del árbol generado por el mismo (ver Figura 27):

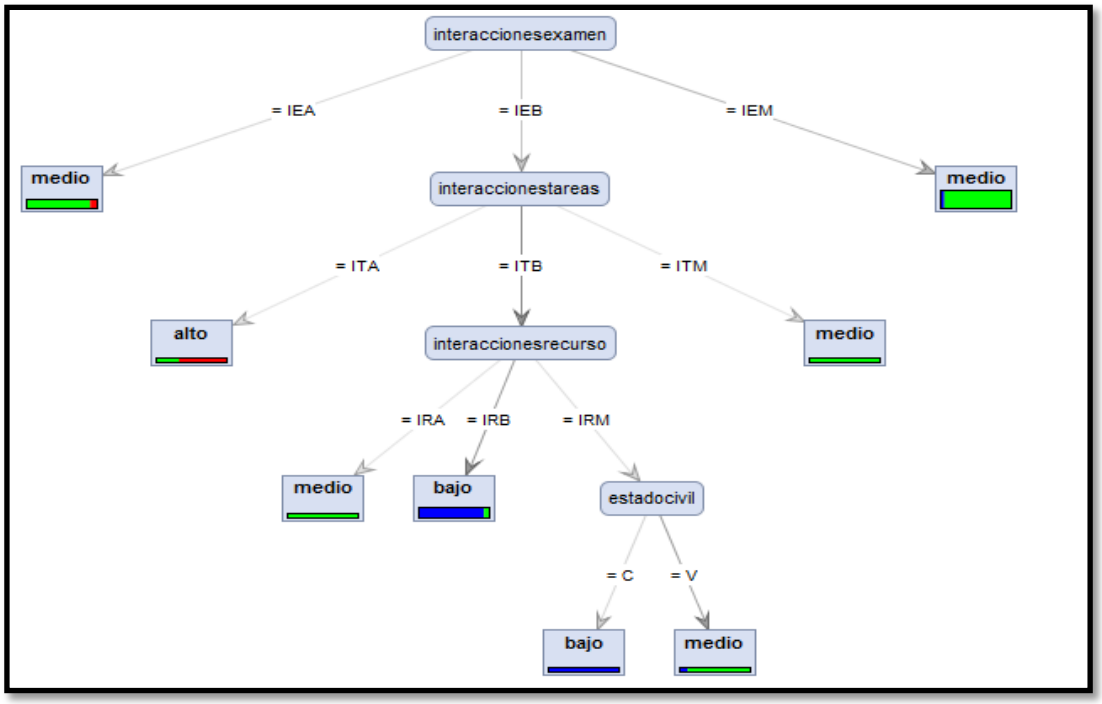


Figura 27: Resultados del algoritmo Decision Tree [38]

4.6.9. k-NN o K Nearest Neighbours (K vecinos más cercanos)

Este algoritmo almacena los datos o instancias utilizadas durante la fase de almacenamiento (conjunto de datos de entrenamiento) para realizar una estimación o clasificación basada en dicho conjunto [36].

Un objeto es clasificado por la clase más común de sus k vecinos más cercanos. K es un número entero positivo, típicamente pequeño y de valor impar. Si k=1, entonces el objeto obtienen la clase del vecino más cercano. Toman los vecinos de un sistema de objetos para los cuales se sabe la clasificación correcta [36].

Su funcionamiento almacena los ejemplos de entrenamiento de datos históricos y cuando se requiere clasificar a un nuevo objeto, se extraen los objetos más parecidos y se usa su clasificación para clasificar al nuevo objeto [37]

A continuación se presenta resultados arrojados por el presente algoritmo (ver Figura 28):

```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          40               53.3333 %
Incorrectly Classified Instances        35               46.6667 %
Kappa statistic                          0.3
Mean absolute error                      0.3168
Root mean squared error                  0.5461
Relative absolute error                 71.1423 %
Root relative squared error            115.6232 %
Coverage of cases (0.95 level)          53.3333 %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances               75

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
                   0.6       0.22       0.577      0.6       0.588      0.684    Alessandra-A
                   0.44      0.2        0.524      0.44      0.478      0.64     Jessica-A
                   0.56      0.28       0.5        0.56      0.528      0.642    Megan-F
Weighted Avg.      0.533     0.233      0.534      0.533     0.532      0.655
```

Figura 28: Resultados del algoritmo K-NN [38]

4.7. Algoritmos de la técnica de agrupamiento

4.7.1. K means

Se trata de un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar, sin etiquetar. Es el método más popular de los métodos de agrupamiento denominado por partición, en contraposición de los métodos jerárquicos, los cuales parten de tantos grupos como individuos haya y van agrupando hasta que todos los elementos se encuentren agrupados en un mismo conjunto [36].

Ejemplo: Se cuenta con las estaturas de 20 personas guatemaltecas y holandesas. Se quiere saber si clasifican como altas o bajas

TABLA VII.

DESCRIPCIÓN DE ESTATURAS

Descripción de Estaturas			
Número	**Altura**	**Número**	**Altura**
1	138	11	195
2	149	12	166
3	142	13	188
4	177	14	195
5	142	15	179
6	157	16	198
7	168	17	161
8	149	18	179
9	177	19	200
10	151	20	191

El algoritmo K means: Resultado

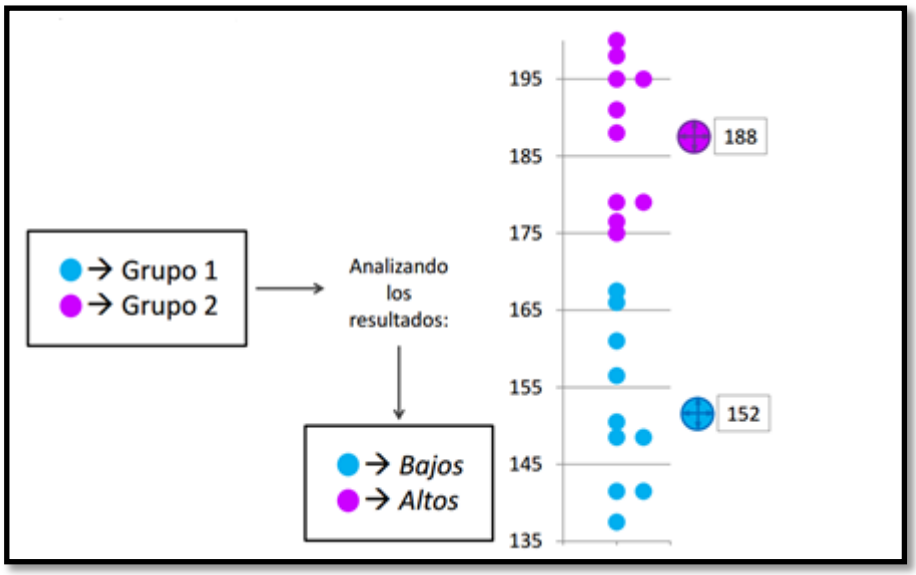


Figura 29. Resultado del algoritmo K-means

Analizando los resultados:

- 80% de los holandeses = Altos
- 80% de los guatemaltecos = Bajos

4.8. Algoritmos de la técnica de reglas de asociación

4.8.1. Algoritmo A Priori

Se usa en minería de datos para encontrar reglas de asociación en un conjunto de datos. Este algoritmo se basa en el conocimiento previo o “a priori” de los conjuntos frecuentes, esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia [37].

Ejemplo: Análisis de la canasta del mercado: Productos comprados por los clientes de un supermercado en los últimos días:

TABLA VIII.

DESCRIPCIÓN DE PRODUCTOS

Descripción de Productos	
TID	**PRODUCTO**
1	Leche, pan, huevos
2	Pan, azúcar
3	Leche, pan, azúcar
4	Leche, cereal, azúcar
5	Leche, cereal
6	Pan, cereal
7	Leche, cereal
8	Leche, pan, cereal, huevos
9	Leche, pan, cereal

Ejemplos de las reglas:

Si un cliente compra leche y pan entonces también compra huevos y azúcar

Si “Leche” y “Pan” => “Huevos” y “Azúcar”

Si un cliente compra leche entonces también compra pan y cereal

Si “Leche” => “Pan” y “Cereal”

4.9. Herramientas de minería de datos

4.9.1. Weka (Waikato environment for knowledge analysis)

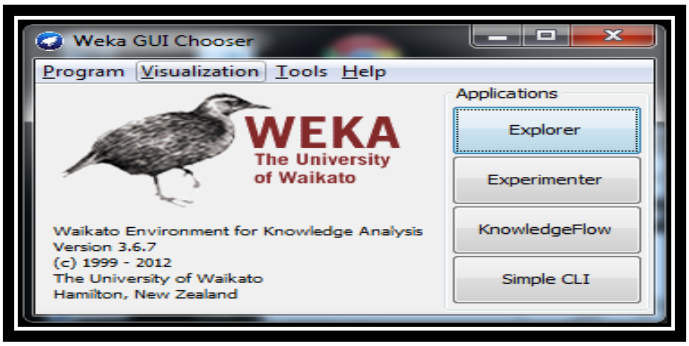


Figura 30: Ventana principal de Weka [44]

Es una herramienta de aprendizaje automático y Minería de Datos, permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos, escrita en lenguaje Java, gratuita y desarrollada en la Universidad de Waikato (WEKA = Waikato Environment for Knowledge Analysis), a continuación se tiene algunas caracteristicas [41].

- Diversas fuentes de datos (CSV, JDBC).
- Interfaz visual basado en procesos/flujos de datos.
- Distintas herramientas de minería de datos: reglas de asociación (a priori), agrupación/segmentación/conglomerado (Cobweb, EM y k-medias), clasificación (redes neuronales, reglas y árboles de decisión, aprendizaje Bayesiano).
- Manipulación de datos.
- Visualización anterior (datos en múltiples gráficas) y posterior (árboles).
- Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas (t-test).

WEKA dispone de 4 interfaces de usuario distintos, que se pueden elegir después de lanzar la aplicación completa, son [42]:

- Simple CLI: interfaz en modo texto.
- **Explorer:** interfaz gráfico básico.

  El Explorer permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos. Cada una de las tareas de minería de datos viene representada por una pestaña en la parte superior. Estas son:

  Preprocess: visualización y preprocesado de los datos (aplicación de filtros)

  - Classify: Aplicación de algoritmos de clasificación y regresión
  - Cluster: Agrupación
  - Associate: Asociación
  - Select Attributes: Selección de atributos
  - Visualize: Visualización de los datos por parejas de atributos
- **Experimenter:** interfaz gráfico con posibilidad de comparar el funcionamiento de diversos algoritmos de aprendizaje, sirve para aplicar varios algoritmos de

aprendizaje automático sobre distintos conjuntos de datos y determinar de manera estadística cual se comporta mejor [43].

- **Knowledge Flow:** interfaz gráfico que permite interconectar distintos algoritmos de aprendizaje en cascada, creando una red [44]-[45].

4.9.2. Spss Clementine

Es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente / servidor, se caracteriza por: Acceso a datos (fuentes de datos archivos ASCII); procesamiento de datos; aplicación de técnicas de aprendizaje como (redes neuronales, reglas de asociación), incorpora técnicas de evaluación de modelos visualización de resultados como (histogramas, diagramas de dispersión) [20].

4.9.3. KEPLER

Sistema desarrollador y transformado en una KEPLER herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis.

Sus principales herramientas de aprendizaje son [25].

- Árboles de decisión
- Redes neuronales
- Regresión no lineal
- Aplicaciones estadísticas

Así mismo permite el preprocesado de datos, la elección de un modelo o la manipulación de la representación gráfica de los modelos obtenidos [25].

4.9.4. Rapid miner

Figura 31: RapidMiner

Es una herramienta de aprendizaje automático implementado en Java por la Universidad de Dormund de libre distribución, incluye operaciones para Importación y pre-procesamiento de datos, aprendizaje automático, validación de modelos, permite la aplicación de técnica como (redes neuronales, reglas de asociación, arboles de decisión) [22].

- **RapidMiner Studio:** Es un GUI descargable para aprendizaje automático, minería de datos, minería de texto, análisis predictivo y análisis de negocios, permite extraer, transformar y analizar los datos de la A a la Z con sus funcionalidades básicas y plugins gratuitos [22].

Algunos servicios permiten extraer los tweets de forma automática desde Twitter, conecta las aplicaciones web que se utilizan para mover fácilmente sus datos y automatizar tareas tediosas. Un zap es una conexión entre dos servicios, que se puede configurar para automatizar tareas [22].

4.9.5. Odms

Herramienta comercial diseñado sobre una arquitectura cliente servidor; ofrece una gran versatilidad en cuanto al acceso a grandes volúmenes de información, se caracteriza principalmente por acceso a datos en diversos formatos, almacenes de datos, bases de datos relacionales como SQL, Oracle, archivos planos; preprocesador de datos: muestreo de datos, patrones de datos; posee modelos de aprendizaje como: redes neuronales, regresión lineal [46].

CAPÍTULO V

5. METODOLOGÍAS DE MINERÍA DE DATOS

5.1. Metodología para proyectos de minería de datos

Existen algunas metodologías para trabajar con minería de datos tanto comerciales como de código abierto, sin embargo en el presente estudio se ha considerado las más importantes.

5.1.1. Semma

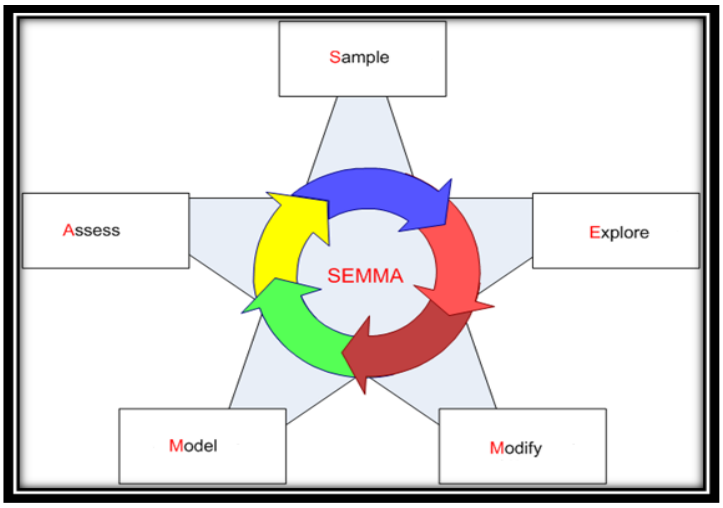


Figura 32: Fases de SEMMA [47]

Creada por el SAS Institute, se define como “el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos”. El nombre de esta terminología corresponde a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración) [47].

Se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando evidenciando que el modelo está orientado especialmente a aspectos técnicos [48].

5.1.2. KDD

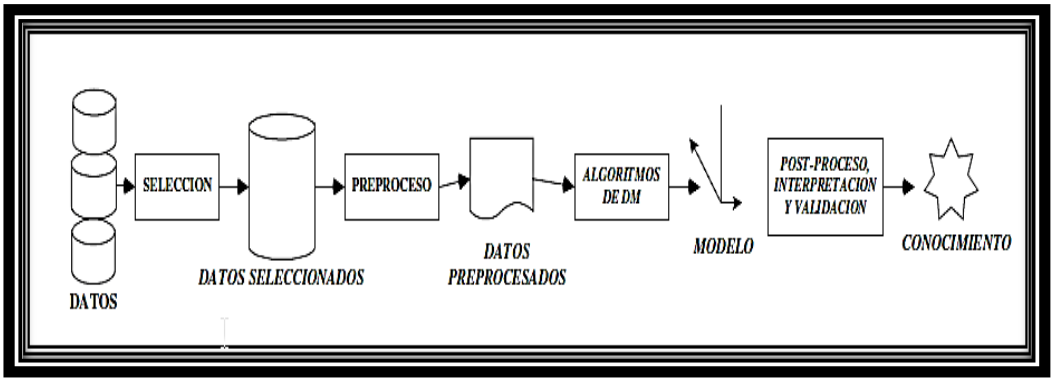


Figura 33: Fases de KDD [48]

Metodología KDD (Knowledge Discovery in Databases) constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información, está formado por nueve etapas. Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Sin embargo actualmente, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento [48].

5.1.3. Crisp-Dm

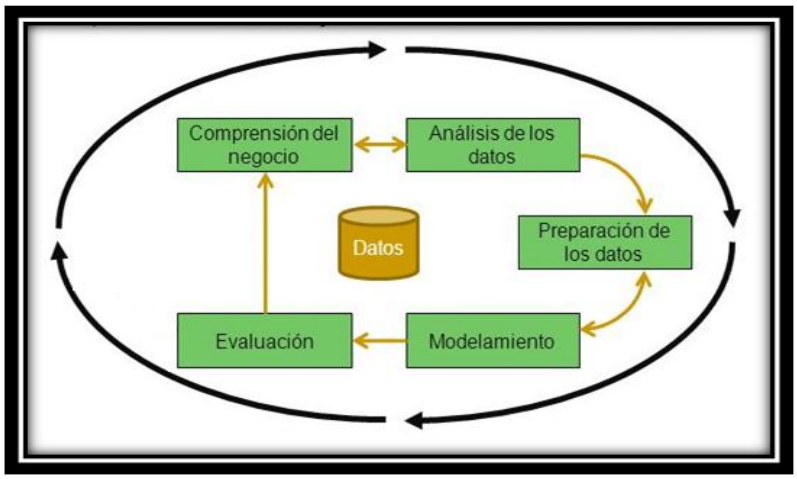


Figura 34: Fases de CRISP-DM [47]

Es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado y Evaluación. La sucesión de fases, no es necesariamente rígida. Cada fase es descompuesta en varias tareas generales que se proyectan a tareas específicas, pero en ningún momento se propone como realizarlas. Es decir, CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo [47].

5.1.4. Catalyst

Conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información [47].

La metodología Catalyst, en sus dos modelos, está compuesta por una serie de pasos llamados “boxes”, luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles [47].

Sobresale en su fase de Modelado del Negocio, contemplando cinco puntos de partida para el proyecto, que finalmente conducirán a la definición de un conjunto de requerimientos y a una situación organizacional que deberá ser abordada desde la minería de datos [48].

5.2. Comparación de las Metodologías de Minería de Datos

Algunos modelos profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA) [47].

KDD, CRISP-DM y Catalyst contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería. SEMMA excluye esta actividad del modelo [48].

En todos los modelos se contempla la selección y preparación de los datos esta situación se repite para la fase de modelado, donde se aplican las técnicas de minería para obtener los nuevos patrones [48].

La implementación de los resultados obtenidos es una fase que no está incluida en el modelo SEMMA. En CRISP-DM, se propone además una planificación para el control futuro y un análisis de cierre del proyecto [48].

SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Explotación de Datos en fases que se encuentran interrelacionadas entre sí [48].

SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Explotación de Datos siendo su distribución libre y gratuita [48].

5.3. Elección de la Metodología

La metodología a utilizar para el presente Trabajo de Titulación es CRISP-DM ya que cada una de sus fases se encuentra claramente estructurada definiendo de tal forma las actividades y tareas que se requieren para lograr el objetivo planteado es decir es la más completa entre las metodologías comparadas, es flexible por ende se puede hacer usos de cualquier herramienta de Minería de Datos.

5.3.1. Crisp-dm (CRoss-Industry Standard Process for Minería de Datos)

CRISP-DM es una metodología estándar para la construcción de proyectos de minería de datos con sus fases no necesariamente rígidas [49].

Puede ser integrada con una metodología de gestión de proyectos específica que complemente las tareas administrativas y técnicas, además es de libre distribución.

Organiza el desarrollo de un proyecto de Minería de Datos en una serie de fases o etapas que funcionan de manera cíclica e iterativa, cada una cuenta con tareas generales y específicas que permitan cumplir con los objetivos del proyecto (ver Figura 35) [50]-[52].

5.3.1.1. Descripción de las fases de CRISP–DM

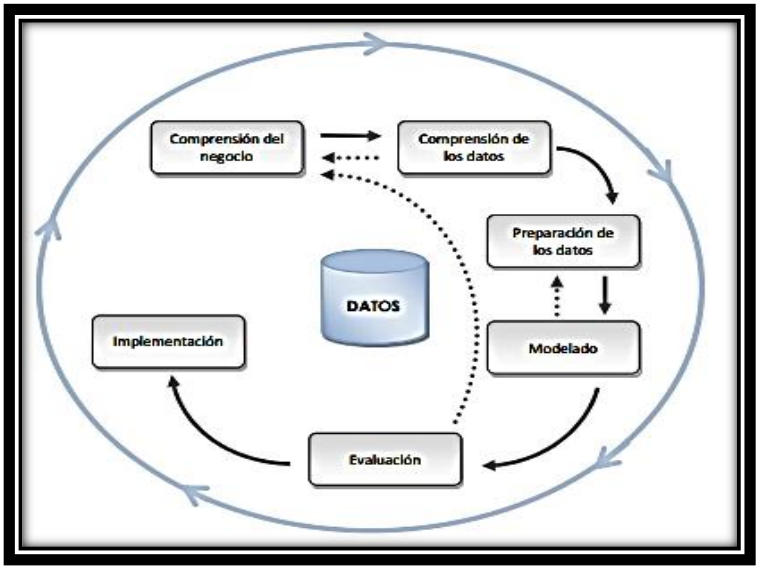


Figura 35: Fases de la Metodología CRISP – DM [51].

5.3.1.1.1. Comprensión del negocio

Comprender o definir el problema del negocio, lo cual es quizás el paso más importante de la metodología, permite entender los objetivos y requisitos que tendrá el proyecto [51].

Las tareas de esta fase es el establecimiento de los objetivos de negocio, evaluación de la situación mediante el inventario de recursos, requerimientos, suposiciones, restricciones, riesgos, contingencias, terminología, costes y beneficios, establecimiento de los objetivos de minería de datos, generación del plan del proyecto y evaluación inicial de herramientas y técnicas [51]-[53].

5.3.1.1.2. Comprensión de los datos

Comprende la búsqueda de la información y de las variables que se utilizarán para la generación de los indicadores del proceso a los cuales se aplicara minería de datos, contiene algunas tareas como es la recolección de datos, teniendo claro desde qué lugar fueron obtenidos. Descripción de los datos, estableciendo los volúmenes de información con que se trabajará, la cantidad de registros, y los significados de cada campo o variable y los formatos en los que se encuentran. Exploración de los datos, indicando una

estructura general de la información, comprobar frecuencia y distribución de los datos, verificación de la calidad de los datos, determinando la consistencia de los valores, comprobando la existencia de datos nulos y fuera de rango, identificando irregularidades para asegurar la completitud y exactitud de los datos [52]-[53].

5.3.1.1.3. Preparación de los datos

Preparación de los datos para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente, consta de algunas tareas como la selección de datos escogiendo un subconjunto de los datos recopilados en la etapa anterior. Limpieza de los datos, preparándolos para la fase de modelación, ya sea aplicando técnicas de normalización, discretización de campos numéricos, tratamiento de valores nulos, entre otros.

Estructuración de los datos con lo cual se pueden generar nuevos atributos a partir de los existentes o transformar valores de los atributos con que se cuenta. Integración de los datos, agrupar tablas o campos que se encuentren relacionadas, definiendo una estructura que las pueda contener. Formateo de los datos, transformar los datos sin modificar su significado, para que se puedan ajustar a las técnicas de minería de datos que se utilice [51]-[52].

5.3.1.1.4. Modelado

Se elige las técnicas de modelado que sean más apropiadas para resolver el problema, aplica algunas tareas que son en base al objetivo principal del proyecto. Generación del plan de prueba, diseñando un procedimiento para probar y validar el modelo. En general, se separa el conjunto de datos en dos: una parte de los datos destinada a entrenamiento del modelo y otra parte que será utilizada para las pruebas. Construcción del modelo a partir de la técnica de modelado seleccionada, se aplica sobre el conjunto de datos para generar uno o más modelos. En este punto se van ajustando los parámetros de la técnica seleccionada de forma iterativa para obtener mejores resultados. Evaluación del modelo, interpretando los modelos en base al conocimiento existente y los criterios de éxito ya establecidos [51].

5.3.1.1.5. Evaluación

Se evalúa el modelo en base al cumplimiento de los criterios de éxito del problema, revisar el proceso seguido teniendo en cuenta los resultados obtenidos, para poder repetir algún proceso en el que a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase [49].

Se evalúa el grado en el cual el modelo satisface los objetivos del negocio y busca determinar si hay alguna razón del negocio porque el modelo sería deficiente [50].

g. Materiales y Métodos

Para el desarrollo del Trabajo de Titulación fue necesario el empleo de algunos métodos y metodología para ello se realizó una búsqueda de información bibliográfica y algunos casos de éxito de tal forma que se emprendió un análisis comparativo entre algunas metodologías (sección 5.2. Comparación de las metodologías de minería de datos) con el fin de seleccionar la que más se adapte al trabajo de titulación denominado “Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje”.

En la recolección y organización de la información que se obtuvo para la sustentación del presente Trabajo de Titulación se hizo uso de los siguientes métodos y técnicas para cumplir con los objetivos planteados:

Observación activa.- A través de esta técnica se ha podido observar la realidad que se vive dentro de los Cursos Virtuales de Aprendizaje de la MED, los inconvenientes que se generan diariamente, así como también permitió seguir obteniendo información necesaria a lo largo del desarrollo del proyecto.

Estudio de Casos.- Sirvió para obtener un conocimiento más amplio de los casos reales actuales, los cuales ayudaron para tener una idea clara del problema, así mismo permitió realizar una exploración e investigación en profundidad de problemas específicos.

Entrevista.- Se obtuvo información concreta, relevante y confiable ya que fue dirigida a entes estratégicos, lo cual ayudo a sustentar y justificar la finalidad del desarrollo del Trabajo de Titulación. Se realizó la entrevista al Director de la MED para conocer algunos de los problemas que tienen los estudiantes en el acceso al curso virtual de inglés (ver anexo A), así mismo el encargado del departamento de la Unidad de Telecomunicaciones e Información de la Universidad Nacional de Loja, proporcionó información correspondiente a los estudiantes del curso virtual para realizar un análisis de cada uno de los parámetros obtenidos.

Revisión bibliográfica: Mediante esta técnica se sustentó la base teórica de la realización del proyecto, mediante consultas en fuentes bibliográficas, artículos científicos, libros, tesis de grado y casos de éxito.

Dando como resultado el desarrollo de la revisión literaria en donde se mencionó casos de éxito relacionados al tema tratado, minería de datos, técnicas de minería de datos con sus respectivos algoritmos, herramientas que son utilizables dentro de la minería de datos e interacciones de los estudiantes en entornos virtuales.

Metodología: Para el presente proyecto se estableció un plan de trabajo en el cual se utilizó la metodología CRISP-DM que es muy importante porque contiene etapas las cuales están compuestas por actividades o una secuencia de pasos ordenados estándar donde se tenga en cuenta las técnicas y herramientas de minería de datos, que permitieron cumplir con los objetivos del trabajo de titulación, por ende a continuación se explica detalladamente cada una de las fases de la metodología:

- **Comprensión del negocio:** Es el paso más importante de la metodología ya que se define el problema del negocio permitiendo de esta manera entender los objetivos y requisitos del proyecto, para lo cual se lleva a cabo las siguientes tareas: establecimiento de los objetivos de negocio, evaluación de la situación mediante el inventario de recursos, requerimientos, suposiciones, restricciones, riesgos, contingencias, terminología, costes y beneficios, establecimiento de los objetivos de minería de datos y generación del plan del proyecto.

- **Comprensión de los datos:** Se recolectó los datos iniciales, los mismos que fueron proporcionados por el encargado del departamento de la UTI, entre los datos se tiene el número de interacciones en el curso, accesos a las tareas, recursos, exámenes, datos personales, socioeconómicos, institucionales de los estudiantes, posteriormente se realizó la descripción de los datos con la finalidad de comprender el significado de cada campo o variable y el formato en que se encuentran, además se exploró los datos comprobando su frecuencia, distribución y calidad, determinando la consistencia de los valores, comprobando la existencia de datos nulos y fuera de rango para asegurar la mejora de los datos.

- **Preparación de los datos:** En la presente fase se realizó la selección de los datos que fueron recopilados en la etapa anterior para posteriormente efectuar la limpieza de los datos, preparándolos para la fase de modelación, con la ayuda de la estructuración de los datos se generó nuevos atributos a partir de los existentes, seguidamente se desarrolló la integración de los datos para agrupar tablas o campos que se encuentren relacionados, definiendo una estructura final o data set. Finalmente se hizo el formateo de los datos, para transformar los datos sin modificar su significado, de manera que se puedan ajustar a los algoritmos pertenecientes a las técnicas de minería de datos que se utilice.

- **Modelado:** Se realizó la construcción del modelo a partir de la técnica de minería de datos seleccionada, que se aplicó sobre el conjunto de datos para generar uno o más modelos, para ello se separa el conjunto de datos en dos: una parte de los datos destinada a entrenamiento del modelo y otra parte utilizada para las pruebas.

- **Evaluación:** Se evalúo el modelo en base al cumplimiento de los criterios de éxito del problema, se revisó los procesos teniendo en cuenta los resultados obtenidos, para poder repetir algún proceso en el que se hayan podido cometer errores.

h. Resultados

Para el desarrollo y obtención de los resultados del presente Trabajo de Titulación se ha establecido fases fundamentales que constan de tareas y actividades.

A continuación se presentan los resultas de cada objetivo planteado, para ello se utilizó algunas herramientas como de minería de datos y base de datos para la manipulación de la información de tal forma se pudo realizar el modelo.

1. FASE I: Investigar sobre las diversas técnicas de minería de datos que permitan determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.

Para la realización del primer objetivo se ha propuesto algunas actividades que son fundamentales para el cumplimiento del mismo, en las que cada una de ellas se las documentará para la constancia física, por ello la gran importancia de esta fase ya que aquí se realizará la búsqueda de casos de éxito relacionados con la aplicación de la minería de datos en plataformas virtuales, de manera que se pueda obtener información fidedigna, para utilizarla como un marco de referencia y realizar un análisis sobre los diferentes escenarios encontrados y ser persuasivo al momento de la selección de las técnicas de minería de datos, porque con el empleo de las mismas y otros recursos se podrá dar respuesta al problema, en el primer objetivo del proyecto se ha realizado las siguientes actividades:

1.1. Recolectar información de fuentes confiables sobre las diversas técnicas de Minería de Datos.

En esta actividad se ha investigado en diferentes repositorios sobre las distintas técnicas que se aplica en la Minería de Datos para obtener suficiente información. En primera instancia la minería de datos es un conjunto de procesos, algoritmos, herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información útil de base de datos, la misma que puede ser utilizada como soporte para la toma de decisiones [38].

Las técnicas de minería de datos (ver Revisión Literaria Capítulo IV) se aplican a un conjunto de datos para obtener resultados, los mismos que pueden ser utilizados para la toma de decisiones [39].

1.2. Realizar un análisis de las diversas técnicas de Minería de Datos.

Para el análisis de cada una de las técnicas de minería de datos se ha estudiado algunos casos de éxito donde estas han sido empleadas, obteniendo resultados confiables, además se realizó un análisis comparativo el cual se puede observar en la siguiente tabla (ver TABLA IX), cabe mencionar que dentro de cada una de las técnicas de minería de datos, estas constan con diferentes algoritmos que se pueden utilizar, el contenido de los mismos se encuentran descritos en la Revisión Litería del Capítulo IV: Minería de Datos, sección 4.5 técnicas de Minería de Datos, y sección 4.6 algoritmos de Minería de Datos.

Técnica	Descripción	Algoritmo	Casos de éxito
Agrupamiento O clustering	El análisis de clusters es utilizado para obtener una visión de la distribución de los datos, para observar las características de cada clúster [7]. A diferencia de la clasificación, el clustering no depende de clases [4].	Simple-Kmeans	- Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje [3], [5]. - Uso de ambientes virtuales de aprendizaje en la enseñanza de la ingeniería [11].
Clasificación	Son estructuras que representan conjuntos de decisiones [5]. Útiles para explorar un conjunto de datos [10]. Permite organización eficiente de conjunto de datos [11].	J48 REPTree ID3/C4.5	- Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL [5]. - Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL [2].

1.3. Determinar la técnica de Minería de Datos que se adapte al entorno a que se va a trabajar.

Para la selección de la técnica de minería de datos a utilizarla posteriormente en el desarrollo del trabajo de titulación se lo ha realizado mediante un cuadro comparativo que se lo puede observar en la TABLA IX, en donde se describe cada técnica y casos de éxito donde han alcanzado resultados idóneos, así mismo se ha podido identificar la técnica apropiada para el proyecto que es la clasificación para el análisis de las interacciones de los estudiantes en el entorno virtual de aprendizaje ya que es apropiada al problema a resolver y se tiene una comprensión de la misma.

Por otra parte se ha podido identificar una gran mayoría de casos de éxito relacionados con el estudio de entornos virtuales con diferentes fines en donde la técnica más aplicada es la antes mencionada, siendo esta eficiente en el momento de analizar grandes cantidades de datos y posteriormente el desarrollo de un modelo para que mediante este se pueda tomar decisiones y mejorar el uso de los entornos virtuales de aprendizaje.

2. FASE II: Diseñar un modelo computacional aplicando técnicas de minería de datos para determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.

En el presente objetivo se desarrollaron algunas actividades que se complementan con cada una de las fases de la minería de datos basadas en la metodología CRISP-DM, mediante la cual se obtuvo el modelo computacional con el que se dará respuesta al trabajo de titulación, las mismas se muestran a continuación.

2.1. Migración y alojamiento de los datos en una Base de Datos

Para el cumplimiento de esta actividad se recolectó los datos de los estudiantes los mismos que fueron proporcionados por la Unidad de Telecomunicaciones e Información en archivos .XML, se desarrolló algunas de las etapas de Minería de Datos que constan de una serie de actividades, las mismas se detallan a continuación:

2.1.1. ETAPA I. Comprensión del negocio

Esta fase se centra en comprender los objetivos y los requerimientos del proyecto desde una perspectiva del negocio, y luego en convertir este conocimiento en la definición de un problema de minería de datos y en un plan preliminar designado para alcanzar los objetivos.

Se presentan los objetivos del negocio (sección 2.1.1.1.), evaluación de la situación (sección 2.1.1.2.), costos (sección 2.1.1.3.), objetivos de la minería de datos (2.1.1.4.) y el plan de trabajo (sección 2.1.1.5.).

2.1.1.1. Determinar los objetivos del negocio

Se realiza una descripción del contexto (sección 2.1.1.1.1.), objetivos del negocio (sección 2.1.1.1.2.) y finalmente criterio de éxito (sección 2.1.1.1.3.).

2.1.1.1.1. Contexto del negocio

La Universidad Nacional de Loja es una Institución de nivel superior académico, consta de una Modalidad de Estudios a Distancia (MED) que tiene implementados cursos virtuales para la formación de los estudiantes, permitiéndoles de esta manera ser los actores

principales de su formación utilizando una plataforma virtual adecuada de la Institución, mejorando las guías de estudios en su forma y estructura didáctica, lo que beneficiará a los estudiantes universitarios en su formación. Por lo que se está realizando un estudio de los datos de interacción de los estudiantes con el entorno virtual de aprendizaje de la MED, para conocer si los estudiantes están aprovechando al máximo los recursos ofrecidos, mediante un seguimiento a los accesos que se registran a la Base de Datos y además será de gran apoyo para la toma de decisiones que permita implementar mejoras en el curso virtual de inglés.

El objetivo de la investigación es Desarrollar un Modelo Computacional Aplicando Técnicas de Minería de Datos Para Determinar la Interacción de los Estudiantes en un Entorno Virtual de Aprendizaje.

2.1.1.1.2. Objetivos del negocio

- Investigar sobre las diversas Técnicas de Minería de Datos que permitan determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.
- Diseñar un modelo computacional aplicando técnicas de Minería de Datos para determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.
- Evaluar el modelo computacional en un escenario real a través de los datos de interacción de los estudiantes en un Entorno Virtual de Aprendizaje.

2.1.1.1.3. Criterio de éxito

- Obtener una técnica de minería de datos adecuada para el proyecto en curso.
- Lograr un modelo mediante la utilización de técnicas y herramientas de minería de datos.
- Analizar las interacciones de los estudiantes del curso virtual de inglés de la MED.

2.1.1.2. Evaluación de la situación

Se llevó a cabo un análisis de recursos disponibles para el desarrollo del Trabajo de Titulación, en esta sección se describe el inventario de recursos (sección 2.1.1.2.1.), requerimientos (sección 2.1.1.2.2.), suposiciones (sección 2.1.1.2.3.), restricciones (sección 2.1.1.2.4.) y la terminología (sección 2.1.1.2.5.) que se muestran a continuación.

2.1.1.2.1. Inventario de recursos

Los recursos son presentados en cuatro categorías, recursos de software (sección 2.1.1.2.1.1.), recursos de hardware (sección 2.1.1.2.1.2.), fuentes de datos y conocimientos (sección 2.1.1.2.1.3.) y recursos humanos (sección 2.1.1.2.1.4.).

2.1.1.2.1.1. Recursos software

Los recursos software que han sido involucrados en el desarrollo proyecto se puede observar en la siguiente TABLA X.

TABLA X.

RECURSOS SOFTWARE

Recursos Software	
Software	**Descripción**
Paquete Office	Planilla de calculo
MySQL	Gestor de base de datos
RapidMiner	Software de minería de datos

2.1.1.2.1.2. Recursos hardware

Los recursos de hardware como son equipos computacionales, dispositivos de almacenamiento que han sido utilizados en el desarrollo del proyecto, se encuentran detallados en la siguiente TABLA XI.

TABLA XI.

RECURSOS HARDWARE

Recursos Hardware	
Hardware	**Descripción**
Ordenador	Disco duro 600GB
	Memoria RAM 8Gb
	Intel core i5
Disco Duro	1TB
Impresora	Canon MP250 Multifuncional
Memoria USB	8GB

2.1.1.2.1.3. Fuente de datos

Las fuentes de datos e información que han sido requeridas para el desarrollo del proyecto se pueden observar en la siguiente TABLA XII.

TABLA XII.

FUENTES DE DATOS

Datos	
Recurso	**Descripción**
Información	Información de las interacciones de los estudiantes del curso virtual de inglés de la modalidad de estudios a distancia de la Universidad Nacional de Loja.
	Información de las actividades que se realicen en el curso virtual de inglés de la modalidad de estudios a distancia de la Universidad Nacional de Loja.

2.1.1.2.1.4. Recursos humanos

En el presente Trabajo de Titulación se emplearon recursos humanos es decir son aquellos que intervienen en las actividades del proyecto, se puede observar en la siguiente tabla XIII.

TABLA XIII.

RECURSOS HUMANOS

Recursos Humanos	
Personal	**Descripción**
Investigadora	Encargada del desarrollo del proyecto para ello se tiene que realizar una investigación, manejo de la información, recolección de datos, manejar herramientas y la elaboración del modelo de Minería de Datos.
Director de tesis	Revisa constantemente el desarrollo del proyecto y verifica que se cumpla con los objetivos planificados.
Director de la MED	Persona encargada de la administración de la Modalidad de Estudios a Distancia.

2.1.1.2.2. Requerimientos

- Tener la información suficiente de las interacciones de los estudiantes para la obtención del modelo.
- Seleccionar técnicas de minería de datos adecuadas al problema a resolver.
- Disponer de herramientas de minería de datos para la realización del modelo.
- Contar con la colaboración continua por parte del director de tesis.
- Tener apoyo del Director de la MED para obtener información exacta de los procesos que se utilizan en el curso virtual.

2.1.1.2.3. Suposiciones

- Se supone que un estudiante pertenece al curso virtual ingles ya que se encuentra inscrito en el mismo.
- Para la determinación de que un estudiante haya ingresado al curso exista registros de esta interacción.

- Se supone que el estudiante tiene acceso a los diferentes recursos y actividades planteadas.
- Se supone que el estudiante al momento que ingresa al curso puede revisar toda su información personal y académica.

2.1.1.2.4. Restricciones

La información que se encuentra almacenada en la base de datos existen valores del estudiante que faltan.

La información de las interacciones de los estudiantes fue requerida en un tipo de formato adaptable a la base de datos que se los iba a integrar, por lo que se fue proporcionada en un formato diferente esto dificulta la manipulación de los mismos.

2.1.1.2.5. Riesgos y contingencias

Se presenta la identificación de los riesgos que se pueden presentar durante el desarrollo del proyecto y las medidas preventivas que se describen en un plan de contingencia para cada uno de ellos, para evitar situaciones peligrosas y minimizar su impacto en la planificación y el coste del proyecto.

A continuación, en la TABLA XIV se presenta la identificación de los principales riesgos que pueden suscitarse en el transcurso del proyecto de tal forma que se puede tener en cuenta dependiendo de cada riesgo existe alguna contingencia para dar solución a estos inconvenientes.

TABLA XIV.

RIESGOS Y CONTINGENCIAS DEL PROYECTO.

Riesgos	Contingencia
Fallo del equipo tecnológico de trabajo.	Sacar respaldos físicos de la documentación del proyecto como de la BD donde se encuentra toda la información de las interacciones de los estudiantes. Así mismo se puede alojar la información en la nube.
Selección inadecuada de las técnicas y herramientas de la minería de datos.	Elaborar un plan detallando los requerimientos del proyecto para aplicar la técnica y herramienta acorde a lo que se pretende realizar.
Mala estimación de la duración del proyecto.	Ajustar el tiempo requerido para cada fase del proyecto tomando en cuenta cada actividad a realizar y dejar un tiempo razonable en cada fase por si llegara a ocurrir algún contratiempo.
Identificación de actividades incompletas según las fases a desarrollar.	Analizar los requisitos del proyecto y la metodología a seguir para tomar en cuenta cada fase y las actividades que se deban realizar.
Estancamiento en alguna fase.	Buscar información de Minería de Datos correspondiente a las fases con el fin de tener en cuenta que actividades se debe realizar dependiendo de cada fase abordada.
Identificación de patrones inadecuados para la realización del modelo.	Analizar detalladamente cada uno de los datos y seleccionar los que más se adapten al problema para la generación del modelo computacional de minería de datos.

2.1.1.2.6. Terminología

Se describe la terminología del negocio (sección 2.1.1.2.6.1.) y la terminología de minería de datos (sección 2.1.1.2.6.2).

2.1.1.2.6.1. Terminología del negocio

- **EVA:** Entorno Virtual de Aprendizaje.
- **INTERACCIÓN:** Lineamiento de comunicación mediante las acciones que realiza cada persona con herramientas tecnológicas.
- **MED:** Modalidad de Estudios a Distancia.
- **UNL:** Universidad Nacional de Loja.
- **MOODLE:** Es una plataforma de aprendizaje diseñada para proporcionarle a educadores, administradores y estudiantes un sistema integrado único, robusto y seguro para crear ambientes de aprendizaje personalizados [12].

2.1.1.2.6.2. Terminología de minería de datos

- **Minería de datos:** Es un conjunto de procesos, algoritmos, herramientas y técnicas de análisis de datos que por medio de la identificación de patrones extrae información interesante, novedosa y potencialmente útil de bases de datos que puede ser utilizada como soporte para la toma de decisiones [4].
- **Modelo computacional:** Es un modelo matemático en las ciencias de la computación que requiere extensos recursos computacionales para estudiar el comportamiento de un sistema complejo por medio de la simulación por computadora.
- **Base de datos:** Son recursos que recopilan todo tipo de información, para atender las necesidades de un amplio grupo de usuarios.
- **Técnicas de minería de datos:** Provienen de la inteligencia artificial y la estadística, contienen conjunto de algoritmos que se pueden aplicar sobre un conjunto de datos y obtener resultados de alguna temática planteada.
- **Data Set (Almacén de datos):** Es un conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas.

2.1.1.3. Costos

A continuación se detalla el presupuesto que involucra el desarrollo del proyecto, así como los Recursos Humanos. En la siguiente TABLA XV se detalla los valores y las horas

a emplear por parte de cada personal, son valores en pago de honorarios que corresponden al trabajo realizado dependiendo al rol y el número de horas.

TABLA XV.

RECURSOS HUMANOS

Recursos Humanos			
Recurso	**Horas**	**Costo/H ($)**	**Total ($)**
Autora del TT	400	5.00	2,000.00
Director del TT	100	10.00	1,000.00
Director MED	30	10.00	300.00
		SUBTOTAL	**3,300**.00

Los Recursos Hardware como el ordenador se emplearan para la realización del proyecto en lo que respecta alojamiento y análisis de datos, así como también para la documentación (ver TABLA XVI).

TABLA XVI.

RECURSOS HARDWARE

Recursos Hardware			
Hardware	**Horas**	**Costo ($)**	**Total ($)**
Ordenador	400	2.00	800.00
Disco Duro	100	1.00	100.00
Impresora	60	0.90	54.00
Memoria USB	100	0.15	15.00
		SUBTOTAL	**969.00**

Los Recursos Software son necesarios para el tratamiento de los datos de las interacciones de los estudiantes en el curso virtual de la MED y para la obtención del modelo (ver TABLA XVII).

TABLA XVII.

RECURSOS SOFTWARE

Recursos Software			
Software	**Horas**	**Costo ($)**	**Total ($)**
Paquete office	200	0.00	0.00
MySQL	300	0.00	0.00
RapidMiner	350	0.00	0.00
		SUBTOTAL	**0.00**

Los recursos de servicios son de vital importancia como lo es el uso del internet ya que servirá para la obtención de información relacionada al proyecto en desarrollo (ver TABLA XVIII).

TABLA XVIII.

SERVICIOS

Recursos de Servicios			
Servicio	**Horas**	**Costo ($)**	**Total ($)**
Internet	300	1.00	300.00
		SUBTOTAL	300.00

Los Recursos Materiales sirvieron de apoyo para la evidencia de la documentación del proyecto de forma física y de tal forma realizar entregas de avances y finalmente la documentación final (ver TABLA XIX).

TABLA XIX.

RECURSOS MATERIALES

Recursos Materiales			
Recurso	**Cantidad**	**Costo ($)**	**Total ($)**
Resma de papel	5	3.50	17.50
Cartuchos de Tinta	2	21,00.00	42.00
Perfiles	3	0.50	1.50
		SUBTOTAL	**61.00**

El costo total del proyecto será asumido por la desarrolladora del mismo, a continuación se puede observar de forma detallada el presupuesto general, para ello se recopilo todos los gastos antes descritos. (Ver TABLA XX).

TABLA XX.

PRESUPUESTO TOTAL

PRESUPUESTO TOTAL ($)	
Recursos Humano	3,300.00
Recursos Hardware	969.00
Recursos Software	0.00
Recursos Servicios	300.00
Recursos Materiales	61.00
TOTAL	4,630.00
IMPREVISTOS (10% DEL TOTAL)	463.00
TOTAL PRESUPUESTO + IMPREVISTOS	**5,093.00**

2.1.1.4. Objetivos de la Minería de Datos

Aplicar técnicas y herramientas de minería de datos con la finalidad de realizar la explotación de la información para descubrir patrones de las interacciones por parte de los estudiantes del curso virtual de inglés de la MED.

2.1.1.5. Plan de Trabajo

Para la realización del presente proyecto se ha desarrollado un plan en el cual consta cada una de las fases abordadas con sus respectivas tareas para la generación del modelo, así mismo permitió detallar cada actividad llevada a cabo con su duración en horas, las dependencias, entradas y salidas. Las tareas del plan de proyecto se detallan a continuación (ver TABLA XXI).

TABLA XXI.

PLAN DEL PROYECTO

#	Fase	Dependencia	Actividades	Duración (horas)	Entradas	Salidas
1	**Comprensión del negocio**		Objetivos del negocio	4		Objetivos del proyecto
		1	Evaluación de situación	10	Información de la Institución	Documentación
		2	Objetivos de la minería	7	Información de la minería datos	Redacción de objetivos
		3	Plan de proyecto	15	Objetivos del proyecto	Planificación del proyecto
2	**Comprensión de los Datos**	4	Recolectar datos	25	Plantilla en Excel	Reporte de los datos obtenidos
		5	Descripción de datos.	10	Reporte de datos	Reporte de descripción de datos
		6	Exploración de datos.	18	Base de datos.	Reporte de exploración de datos
		7	Verificación de datos.	6	Reporte de exploración	Reporte de la calidad de los datos
3	**Preparación de los Datos**	8	Seleccionar los Datos	50	Base de datos	Información fundamentada de la inclusión/exclusión de los datos
		9	Limpiar los Datos	25	Base de datos	Reporte de limpieza de datos
		10	Estructurar los Datos	40	Base de datos	Reporte de la estructuración datos
		11	Integrar los Datos	15	Base de datos	Reporte de los datos integrados
		12	Formateo de los Datos	10	Base de datos	Reporte de formateo de datos
4	**Modelado**	13	Selección de la técnica	30	Técnicas de minería de datos	Técnica de modelado
		14	Construcción del Modelo	60	Set de datos.	Modelo
5	**Evaluación**	15	Evaluar los Resultados	75	Modelo	Resultados del modelo

2.2. Realizar un estudio de los datos obtenidos que permitan determinar las interacciones de los estudiantes en el curso virtual de inglés.

Para el desarrollo de esta actividad fue necesario la recolección de los datos que fueron proporcionados por la Unidad de Telecomunicaciones e Información, los cuales sirvieron para poder realizar un estudio de los mismos, para ello fueron necesarias llevar a cabo la continuidad de las etapas de la metodología CRISP-DM, que se detallan a continuación:

2.2.1. Etapa II: Comprensión de los Datos

En esta etapa se recolectó los datos relacionados con las interacciones de los estudiantes para una mejor comprensión de los mismos, de manera que es el primer acercamiento que se tiene para posteriormente realizar el análisis y de esta manera identificar algún inconveniente que exista, de tal forma que se analizó la estructura de los datos mediante consultas ejecutadas en la base de datos.

Así mismo las secciones cuentan la recolección de datos iniciales (sección 2.2.1.1.), descripción de los datos (sección 2.2.1.2.) y la exploración de los datos (sección 2.2.1.3.).

2.2.1.1. Recolección de datos iniciales

Los datos recolectados pertenecen a las interacciones de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia perteneciente a la Universidad Nacional de Loja del periodo académico 2013 - 2014, para ello la información fue proporcionada por la Unidad de Telecomunicaciones e Información, entre la información con la que se trabajó se tiene las interacciones de los estudiantes según las actividades que desarrollaron en el curso, las mismas que se describen a continuación:

- Interacción con los archivos compartidos con las temáticas del curso.
- Realizaron tareas y evaluaciones, para la aprobación del curso.
- Leer o imprimir los contenidos y actividades del curso.
- Enviar las actividades al docente para su corrección y recibir sus calificaciones.
- Evaluaciones On-Line y calificaciones.

De esta manera se ofrece a los estudiantes en formación la oportunidad de reforzar el aprendizaje brindando a través de contenidos y evaluaciones, para posibilitar la interacción estudiante-profesor y estudiante-herramientas.

Los datos se encuentran estructurados en archivos XML que consta de las interacciones, datos personales, institucionales y socioeconómicos de los estudiantes del curso, como son:

- número de accesos al curso
- número de accesos a las tareas
- número de veces que accede a un recurso
- número de accesos a exámenes
- datos personales de los estudiantes
- datos socioeconómicos de los estudiantes
- datos institucionales de los estudiantes

En la siguiente figura (ver Figura 36) se puede observar las tablas que conforman la Base de Datos donde esta almacenada la información de las interacciones de cada uno de los estudiantes de la Modalidad de Estudios a Distancia:

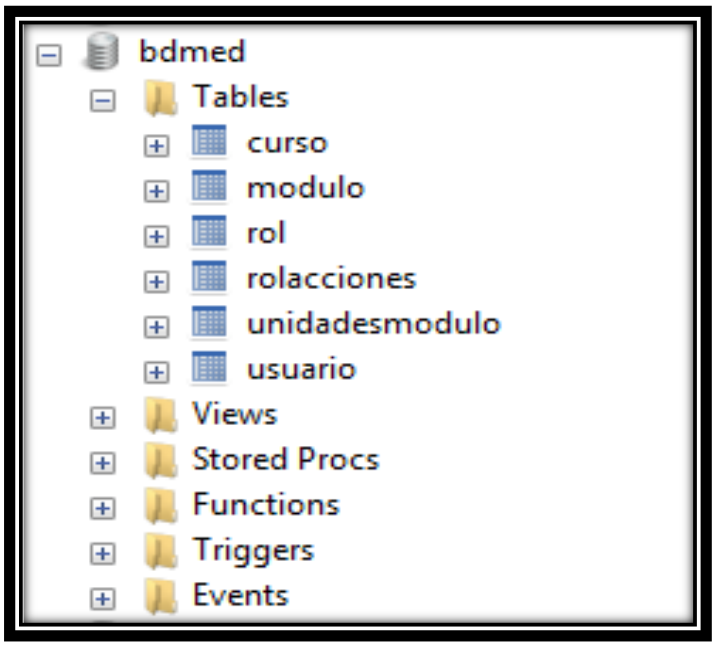


Figura 36. Estructura de la Base de datos

Además el modelo relacional de la BD integrada por seis tablas, conformada por entidades, atributos y las relaciones existentes entre ellas se puede observar en la siguiente figura (ver Figura 37):

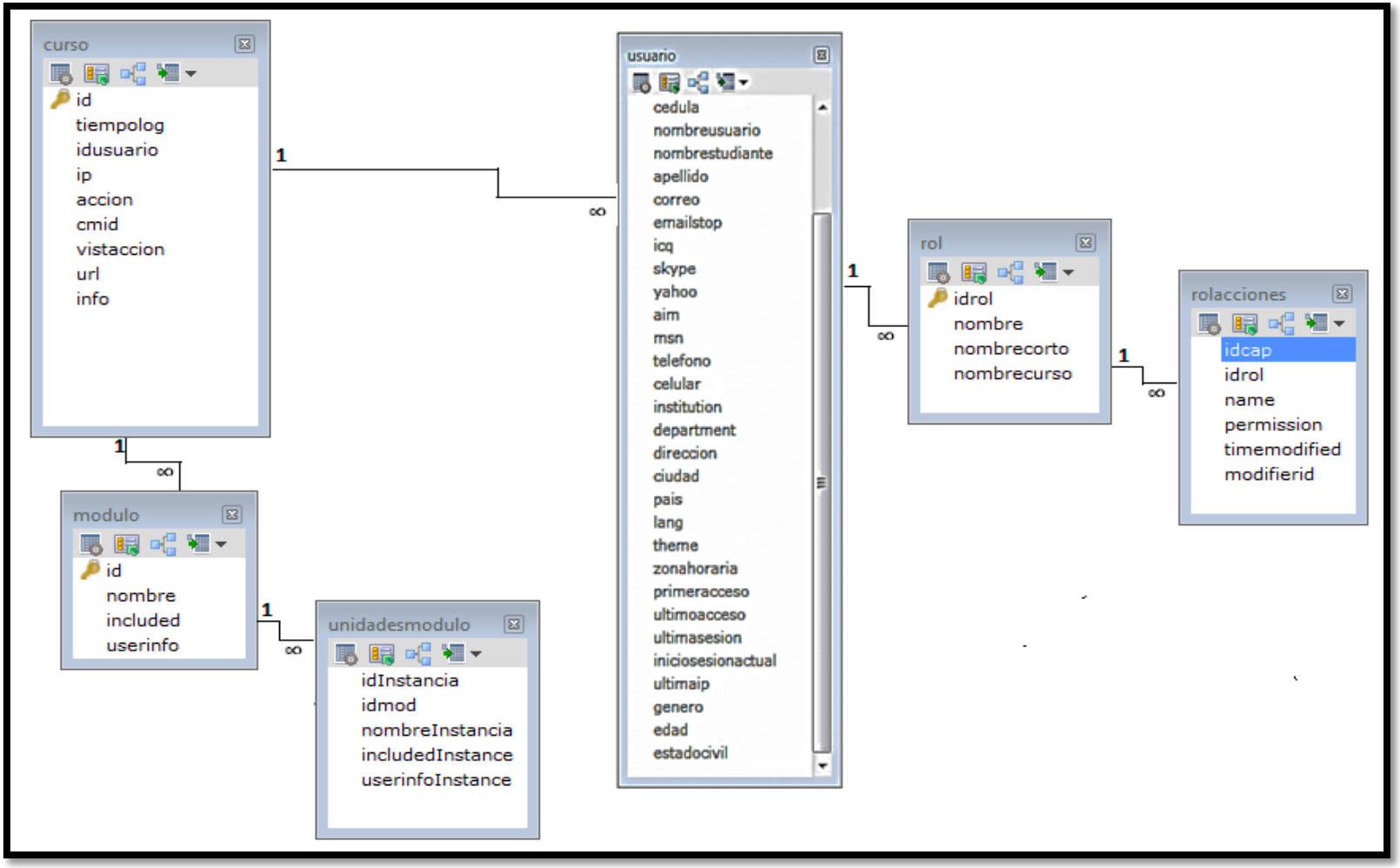


Figura 37. Modelo Entidad-Relación de la BD

2.2.1.2. Descripción de los datos

Los datos se recopilaron de la plataforma Moodle de la Modalidad de Estudios a Distancia de la Universidad Nacional de Loja en formato XML y posteriormente se almacenaron en tablas relacionadas en la Base de Datos para poder manipular la información, la cual está integrada por 6 tablas, las mismas que se describen a continuación:

En la siguiente figura se puede observar un fragmento de cómo está integrada la tabla curso y cada uno de sus atributos con sus respectivos datos pertenecientes a los estudiantes del curso virtual inglés (ver Figura 38):

id	timeLog	userid	ip	accion	cmid	action	url
4736514	Fri Jan 16 14:13:25	3	172.16.32.28	course	0	view	view.php?id=10
4736518	Fri Jan 16 14:13:25	3	172.16.32.28	course	0	view	view.php?id=10
4736527	Fri Jan 16 14:13:25	3	172.16.32.28	course	0	view	view.php?id=10
4736528	Fri Jan 16 14:13:25	3	172.16.32.28	course	0	view	view.php?id=10
4736531	Fri Jan 16 14:13:26	7105	186.42.182.66	course	0	view	view.php?id=10
4736532	Fri Jan 16 14:13:26	7105	186.42.182.66	forum	214	view forum	view.php?id=214
4736681	Fri Jan 16 14:13:33	4173	186.42.72.104	course	0	view	view.php?id=10
4736682	Fri Jan 16 14:13:33	4173	186.42.72.104	forum	214	view forum	view.php?id=214
4736685	Fri Jan 16 14:13:33	4173	186.42.72.104	course	0	view	view.php?id=10
4736686	Fri Jan 16 14:13:33	4173	186.42.72.104	forum	215	view forum	view.php?id=215
4736688	Fri Jan 16 14:13:33	4173	186.42.72.104	course	0	view	view.php?id=10
4736715	Fri Jan 16 14:13:39	8815	186.178.139.135	course	0	view	view.php?id=10
4736726	Fri Jan 16 14:13:39	8815	186.178.139.135	forum	214	view forum	view.php?id=214
4736729	Fri Jan 16 14:13:39	8815	186.178.139.135	forum	215	view forum	view.php?id=215

Figura 38. Tabla curso

En la siguiente tabla se representa una descripción de la clase curso con sus respectivos atributos, así mismo se tiene el tipo de dato que corresponde a cada uno (ver TABLA XXII):

TABLA XXII.

ATRIBUTOS DE LA TABLA CURSO

Tabla curso		
Atributos	**Descripción**	**Tipo de Dato**
Id	Identificador del log	Varchar
tiempoLog	Tiempo de ingreso al log	Varchar
idusuario	Identificador del usuario.	Varchar
ip	Dirección ip de donde ingreso el usuario.	Varchar
vistaccion	Vista realizada en la acción del módulo.	Varchar
cmid	Identificador de los modulos del curso.	Varchar
accion	Acciones que realiza el usuario en el curso.	Varchar
url	Dirección de internet.	Varchar

En la siguiente figura se presenta como está integrada la tabla unidadesmodulo, como las unidades del módulo con su respectivo título y las tareas, lecciones y el video según la materia de cada contenido (ver Figura 39):

idInstance	idmod	nameInstance	includedInstance	userinfoInstance
28	1	VIDEO (Send your video here)	true	true
12	2	UPDATE PERSONAL INFORMATION	true	true
321	5	UNIT THREE TASK	true	true
27	1	UNIT 5 TASK (Send your task here)	true	true
13	4	UNIT 5 LESSON	true	true
294	5	UNIT 5 CONTENTS	true	true
25	1	UNIT 4 TASK (Send your task here)	true	true
12	4	UNIT 4 LESSON	true	true
293	5	UNIT 4 CONTENTS	true	true
23	1	UNIT 3 TASK (Send here your task)	true	true
11	4	UNIT 3 LESSON	true	true
286	5	UNIT 3 CONTENTS	true	true
85	1	UNIT 2 TASK (Send here your task)	true	true
10	4	UNIT 2 LESSON	true	true
279	5	UNIT 2 CONTENTS	true	true
19	1	UNIT 1 TASK (Send here your task)	true	true
9	4	UNIT 1 LESSON	true	true

Figura 39. Tabla unidadesmodulo

Descripción de la tabla unidadesmodulo con lo que representa cada uno de sus atributos y el tipo de dato que les corresponde (ver TABLA XXIII):

TABLA XXIII.

ATRIBUTOS DE LA TABLA UNIDADESMODULO

Tabla **unidadesmodulo**		
Atributos	**Descripción**	**Tipo de Dato**
IdInstancia	Identificador de la instancia del módulo	Varchar
Idmod	Identificador de las actividades planteadas del módulo	Varchar
nombreInstancia	Nombre de la instancia	Varchar

En la siguiente figura se presenta como está integrada la tabla modulo es decir representa todas las actividades de trabajo disponibles para el estudiante en el curso virtual de inglés (ver Figura 40):

	id	nombre	included	userinfo
	1	assignment	true	true
	2	forum	true	true
	3	label	true	true
	4	quiz	true	true
	5	resource	true	true

Figura 40. Tabla modulo

Descripción de la tabla modulo, donde se detalla cada uno de sus atributos como es el tipo de dato al que pertenece cada uno (ver TABLA XXIV):

TABLA XXIV.

ATRIBUTOS DE LA TABLA MODULO

Tabla modulo		
Atributos	**Descripción**	**Tipo de Dato**
id	Identificador de las actividades planteadas del módulo	Varchar
Nombre	Nombre de las actividades planteadas	Varchar

En la siguiente figura se presenta como está integrada la tabla rol, es decir los permisos o privilegios que tiene cada estudiante al ingresar al curso virtual de inglés (ver Figura 41):

	idrol	nombre	nombrecorto	nombrecurso
	3	Profesor	editingteacher	Docente
	4	Profesor sin permiso de edición	teacher	Docente
	5	Estudiante	student	

Figura 41. Tabla rol

Descripción de la tabla rol con sus respectivos atributos que conforman la misma, de manera que se detalla cada uno ellos (ver TABLA XXV):

TABLA XXV.

ATRIBUTOS DE LA TABLA ROL

Tabla rol		
Atributos	**Descripción**	**Tipo de Dato**
Idrol	Identificador del rol	Varchar
nombre	Nombre del rol	Varchar
nombrecorto	Nombre corto del rol en inglés	Varchar
nombrecurso	Nombre del rol en el módulo	Varchar

En la siguiente figura se presenta como está distribuida la tabla rolacciones, donde se encuentra el identificador de cada uno de los roles con sus respectivos permisos o privilegios que puede ser del docente o estudiante (ver Figura 42):

	idcap	idrol	name	permission
☐	1	3	report/cpd:userview	1
☐	2	3	report/courseoverview:view	1
☐	3	3	moodle/user:viewhiddendetails	1
☐	4	3	moodle/user:viewdetails	1
☐	5	3	moodle/user:readuserposts	1
☐	6	3	moodle/user:readuserblogs	1
☐	7	3	moodle/tag:manage	1
☐	8	3	moodle/tag:editblocks	1
☐	9	3	moodle/site:viewreports	1
☐	10	3	moodle/site:viewfullnames	1
☐	11	3	moodle/site:trustcontent	1
☐	12	3	moodle/site:restore	1

Figura 42. Tabla rolacciones

Descripción de la tabla rolacciones, donde se detalla cada uno de sus atributos como el tipo de dato al que pertenecen (ver TABLA XXVI):

TABLA XXVI.

ATRIBUTOS DE LA TABLA ROLACCIONES

Tabla rolacciones		
Atributos	**Descripción**	**Tipo de Dato**
idrol1	Identificador del rol	Varchar
timemodified	Tiempo de modificación	Varchar
name	Corresponde a la ruta de acceso de acuerdo al rol asignado	Varchar
permission	Permiso de acuerdo al tipo de usuario	Varchar

En la siguiente figura se presenta como está integrada la tabla usuario y un fragmento de los datos que corresponden a la misma como puede ser cedula, nombres, dirección, entre otros (ver Figura 43):

	id	cedula	nombreusuario	nombrestudiante	apellido	correo	telefono
☐	10037	0701181505	0701181505	Gilber Ruperto	Loayza Rodriguez	gilberloayza@gmail.com	2976273 2974109
☐	10054	1900758390	1900758390	betty alexandra	valle espinosa	alexa92@live.com	2300227
☐	10062	1104871296	1104871296	inti pachacutic	guaman puchaicela	intipachapenku@hotmail.co	073029603
☐	10143	1104606643	1104606643	karla patricia	ruiz placencia	kpr_romi@hotmail.com	2581519
☐	10166	1104459761	1104459761	javier israel	mora medina	javier198599@hotmail.com	2576-843
☐	10184	1103969430	1103969430	raul alexander	tocto gonzalez	raulalexa_t200806@yahoo.e	2585087
☐	10221	1900650944	1900650944	Maritza Alexandra	Olmedo Salinas	mari_alexa06@hotmail.com	073606910

Figura 43. Tabla usuario

Descripción de la tabla usuario con sus respectivos atributos, es decir los campos con la información que pertenecen al estudiante como género, edad, estado civil, entre otros (ver TABLA XXVII):

TABLA XXVII.

ATRIBUTOS DE LA TABLA USUARIO

Tabla usuario		
Atributos	**Descripción**	**Tipo de Dato**
id	Identificador del usuario	Varchar
nombreusuario	Nombre del usuario	Varchar
cedula	Número de cédula del usuario	Varchar
nombresestudiante	Nombres del estudiante del módulo	Varchar
apellido	Apellidos del estudiante del módulo	Varchar
correo	Email del estudiante del módulo	Varchar
telefono	Número de teléfono del estudiante del módulo	Varchar
celular	Número de celular del estudiante del módulo	Varchar
direccion	Dirección del estudiante del módulo	Varchar
ciudad	Ciudad donde nació el estudiante del módulo	Varchar
pais	País donde nació el estudiante del módulo	Varchar
ultimaip	Direccion ip de donde el usuario accede al módulo	Varchar
genero	Género del usuario del curso	Varchar
edad	Edad del usuario del curso	Varchar
Estadocivil	Estado civil del usuario del curso	Varchar

2.2.1.3. Exploración de los datos

Luego de haber realizado una descripción de los datos que conforman la base de datos de los estudiantes del Curso Virtual de Inglés de la MED, es necesario realizar una exploración de los mismos que contiene cada una de las variables seleccionadas en la base de datos, con el fin de hacer un análisis estadístico y además conocer la distribución que existe entre los datos, tomando en cuenta que la exploración que se ha realizado, se incluyeron los datos de todos los estudiantes.

a. **Distribución según el Género:** El Curso Virtual de Inglés posee más estudiantes en el género femenino con 646 en comparación con el género masculino por lo que se puede observar en la siguiente tabla (TABLA XXVIII) y ver ANEXO G.

TABLA XXVIII.

DISTRIBUCIÓN DE ESTUDIANTES POR GÉNERO

Género	Número de Estudiantes
Masculino	423
Femenino	646

Así mismo se ha representado los resultados mediante porcentajes la distribución del género por lo que se puede observar en la siguiente figura (Figura 44).

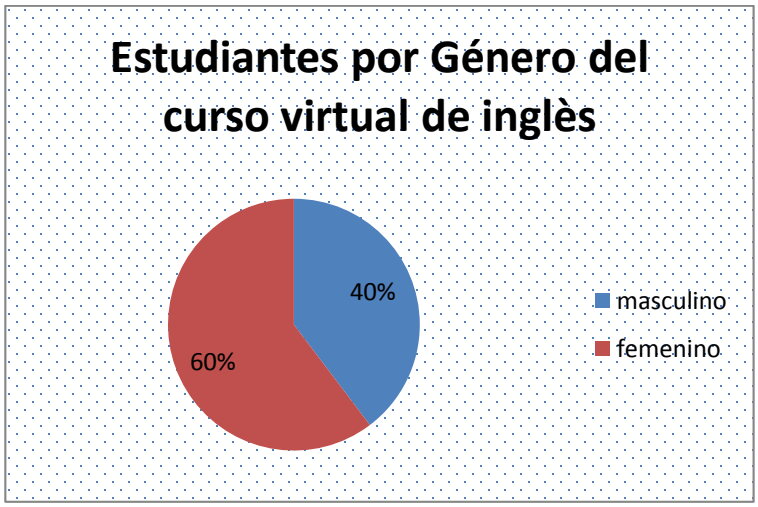


Figura 44. Estudiantes por Género del curso virtual de inglés

b. **Distribución según la Edad.-** El Curso Virtual de Inglés posee más estudiantes en la edad comprendida entre el rango de 16 a 26 años con 412 estudiantes, por lo que se puede observar en la siguiente tabla (TABLA XXIX) y ver ANEXO G.

TABLA XXIX.

DISTRIBUCIÓN DE ESTUDIANTES POR EDAD

Edad	Número de Estudiantes
16-26	412
27-37	395
38-48	262

Así mismo se ha representado los resultados mediante porcentajes la distribución de la edad por lo que se puede observar en la siguiente figura (Figura 45).

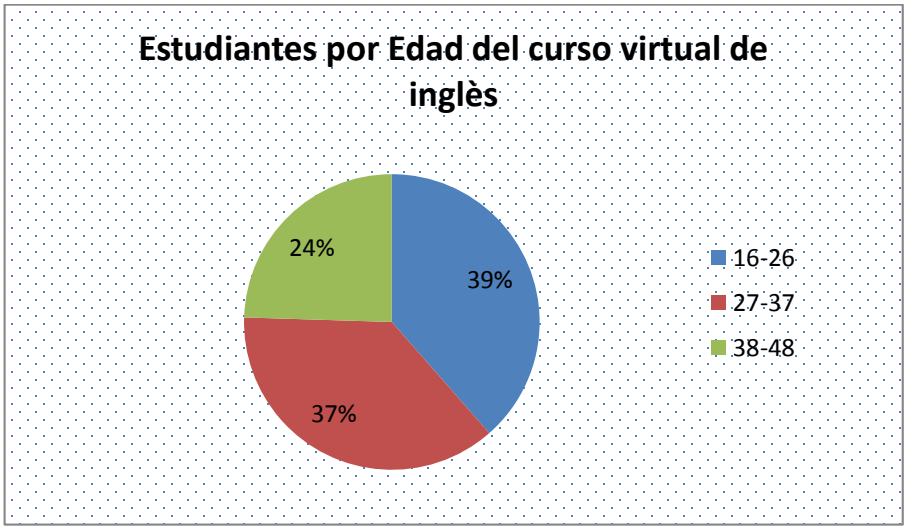


Figura 45. Estudiantes por Edad del curso virtual de inglés

c. **Distribución según las Acciones:** El Curso Virtual de Inglés posee 148472 acciones por parte de los estudiantes, dando como resultado que posee mayor cantidad que corresponde a la acción de tareas con 62486 como se puede ver en la siguiente tabla (ver TABLA XXX) y ver ANEXO G.

TABLA XXX.

DISTRIBUCIÓN DE LAS ACCIONES REALIZADAS POR LOS ESTUDIANTES

Acciones	Número de Acciones
Acción Tareas	62486
Acción Examen	43080
Acción Recursos	42906

Así mismo se ha representado los resultados mediante porcentajes de la distribución de las acciones realizadas por parte de los estudiantes por lo que se puede observar en la siguiente figura (Ver Figura 46).

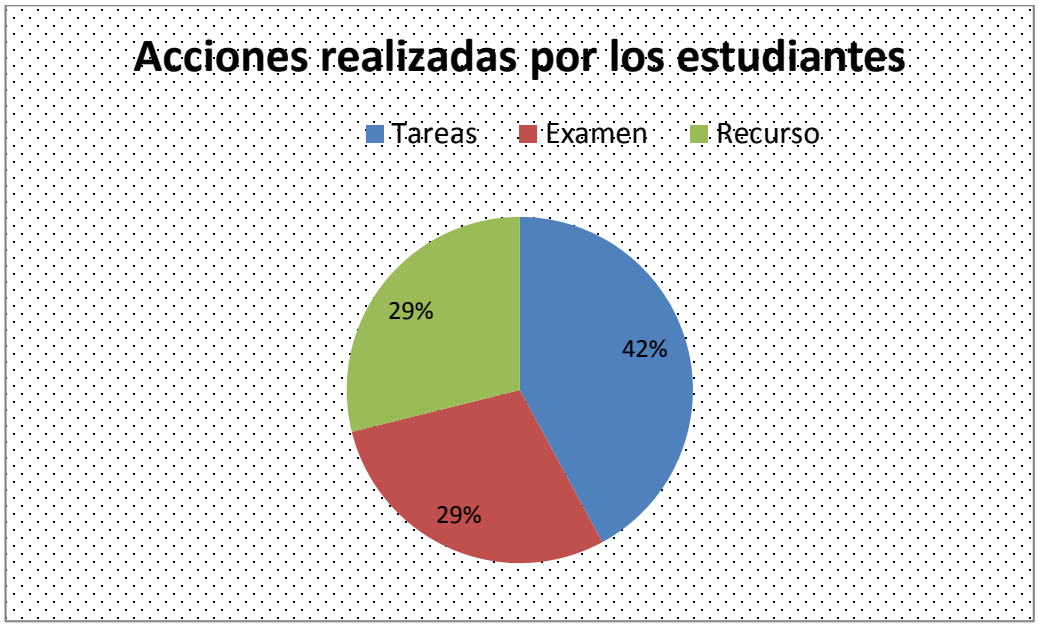


Figura 46. Acciones realizadas por los estudiantes

d. **Distribución según el estado civil:** En el Curso Virtual de Inglés el estado civil soltero posee más estudiantes con un número de 538, como se puede ver en la siguiente tabla (ver TABLA XXXI) y ver ANEXO G.

TABLA XXXI.

DISTRIBUCIÓN SEGÚN EL ESTADO CIVIL

Estado civil	Número de Estudiantes
Soltero	538
Casado	437
Viudo	14
Divorciado	80

Así mismo se ha representado los resultados mediante porcentajes de la distribución del estado civil de los estudiantes, por lo que se puede observar en la siguiente figura (Ver Figura 47).

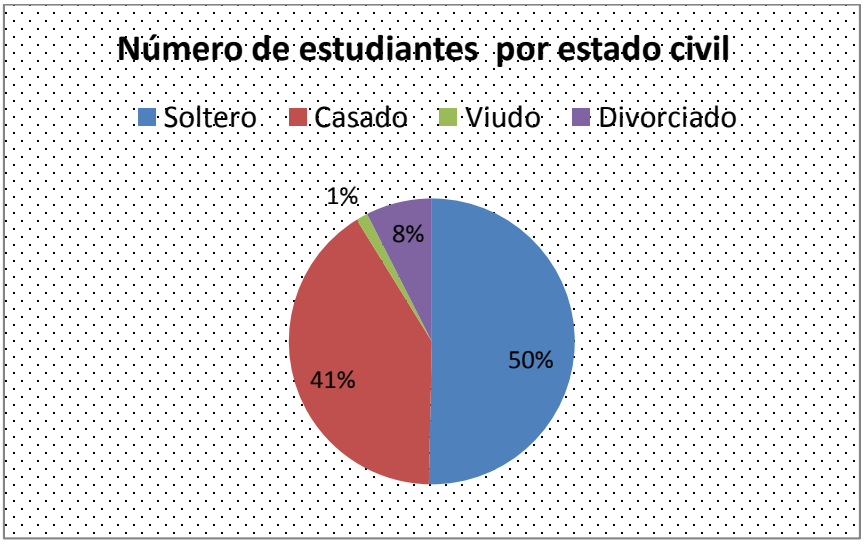


Figura 47. Distribución según el estado civil de los estudiantes

2.3. Seleccionar los parámetros pertinentes para construir el modelo computacional.

En esta actividad luego de realizar un análisis de los datos obtenidos, se seleccionó cada uno de los parámetros que servirán para la construcción del modelo, para ello se desarrolló las siguientes etapas:

2.3.1. ETAPA III: Preparación de los datos

La preparación de los datos se realizó para trabajar con la técnica de minería de datos seleccionada, consta de algunas tareas como la selección de datos donde se eligió una estructura de datos recopilados en la etapa anterior, por otra parte se realizó la limpieza de los datos para poder generar el modelo de minería de datos, de tal manera que no debe contener valores nulos para obtener mejores resultados.

Además se realizara la estructura de los datos con lo cual se generaran nuevos atributos a partir de los existentes, así mismo se desarrollara la integración de los datos que consiste en agrupar tablas o campos que se encuentren relacionadas, finalmente se hizo el formateo de los datos que consiste en transformar los datos sin modificar su significado, para que se puedan ajustar a la técnica de minería de datos.

Por lo tanto el objetivo de la presente etapa fue generar la estructura de datos final, con el propósito de obtener el modelo a través de técnica de Minería de Datos.

Entre las siguientes secciones se tiene seleccionar los datos (sección 2.3.1.1.), limpiar los datos (sección 2.3.1.2.), estructura de los datos (sección 2.3.1.3.) y finalmente la integración de los datos (sección 2.3.1.4.).

2.3.1.1. Seleccionar los datos

Para realizar la selección de los datos se tomó en cuenta el objetivo general del Trabajo de Titulación que es Determinar las Interacciones de los Estudiantes, por ello se procedió a eliminar tablas que no son necesarias para poder llegar a cumplir el objetivo principal, dentro de las tablas eliminadas se encuentran unidadesmodulo y rolacciones.

- Tabla unidadesmodulo: Contiene información de los nombres de etiquetas que no es útil para cumplir con el objetivo.
- Tabla rolacciones: Contiene datos de direcciones web de moodle es decir esta tabla es eliminada porque todos los datos que se encuentran no son relevantes para el proceso de Minería de Datos.

2.3.1.2. Limpiar los datos

Esta tarea se realizó con la finalidad de eliminar datos que no contienen todos los valores o que contengan campos vacíos o completarlos, de tal forma que permita obtener un modelo de calidad.

Se eliminaron atributos de algunas tablas, se los menciona a continuación:

- Tabla usuario: Los campos eliminados son:
  - **numeroid:** contiene información de la cédula del estudiante y debido a que este campo está repetido en la misma tabla.
  - **correo:** presenta información de los correos del estudiante.
  - **dirección:** corresponde a la información de calles de donde vive el estudiante.
  - **ciudad:** información del país de donde vive el estudiante.
  - **ultimaip:** contiene información de la última dirección ip de donde accedió el estudiante al curso.
- Tabla curso: Los campos eliminados son:
  - **url:** campo que contiene información de la página de internet en base a la acción que realiza el estudiante, la misma que no es necesaria para el modelo.

Los atributos que se han tomado en cuenta para realizar el modelo son datos personales, institucionales, socioeconómicos de los estudiantes y las interacciones en tareas, exámenes y recursos.

2.3.1.3. Estructura de los datos

En esta tarea se realizó la construcción del data set final o estructura de datos que es útil para poder desarrollar el modelo computacional, donde se ha tomado en cuenta información personal, socioeconómicas y las interacciones de los estudiantes, para lo cual se trabajó con los siguientes campos para cumplir con el objetivo planteado, los mismos se pueden observar en la siguiente tabla (ver TABLA XXXII):

TABLA XXXII.

ATRIBUTOS DE MINERÍA DE DATOS PARA DETERMINAR LAS INTERACCIONES DE LOS ESTUDIANTES.

Atributo	Tipo de Datos	Categorización
cedula	Nominal	
interaccionesrecurso	Int	- IRB - IRM - IRA
interaccionesexamen	Int	- IEB - IEM - IEA
interaccionestareas	Int	- ITB - ITM - ITA
numerointeracciones	Int	- bajo - medio - alto
servicios	Nominal	- 1 - 2 - 3
ciudad	Nominal	- L - O
edad	Int	- a - b - c
genero	Nominal	- 0 - 1
estadocivil	Nominal	- S - C - D - V
numeroHijos	Nominal	- Si - No
Trabajo	Nominal	- Si - No

La data set final que se encuentra en la Tabla anterior (ver TABLA XXXII) se utilizó para construir el modelo, a continuación se describe cada uno de los atributos que la conforman:

- cedula: contiene el número del documento de identificación del estudiante del curso virtual.
- interaccionesrecurso: número de interacciones que el estudiante ha realizado sobre los recursos del curso. Para una mejor comprensión en la siguiente tabla (ver TABLA XXXIII) se indica este atributo.

TABLA XXXIII.

ATRIBUTO INTERACCIONESRECURSO.

Siglas	Descripción
IRB	Interacciones del estudiante sobre los recursos es bajo con un rango menor a 40.
IRM	Interacciones del estudiante sobre los recursos es medio con un rango entre 41 y 80.
IRA	Interacciones del estudiante sobre los recursos es alto mayores a 81.

- interaccionesexamen: Atributo que contiene el número de interacciones que el estudiante ha realizado sobre los exámenes del curso virtual. En la siguiente tabla (ver TABLA XXXIV) se describe este atributo.

TABLA XXXIV.

ATRIBUTO INTERACCIONESEXAMEN

Siglas	Descripción
IEB	Interacciones del estudiante sobre los exámenes es bajo con un rango menor 30
IEM	Interacciones del estudiante sobre los exámenes es medio con un rango entre 31 y 60.
IEA	Interacciones del estudiante sobre los exámenes es alto mayores 61.

- interaccionestareas: Atributo que contiene el número de interacciones que el estudiante ha realizado sobre las tareas del curso. En la siguiente tabla (ver TABLA XXXV) se indica este atributo.

TABLA XXXV.

ATRIBUTO INTERACCIONESTAREAS.

Siglas	Descripción
ITB	Interacciones del estudiante sobre las tareas es bajo con un rango menor a 50.
ITM	Interacciones del estudiante sobre las tareas es medio con un rango entre 51 y 100.
ITA	Interacciones del estudiante sobre las tareas es alto mayores a 101.

- numerointeracciones: Atributo que contiene el número de interacciones totales que el estudiante ha realizado durante el transcurso del curso virtual. En la siguiente tabla (ver TABLA XXXVI) se describe este atributo.

TABLA XXXVI.

ATRIBUTO NUMEROINTERACCIONES.

Siglas	Descripción
bajo	Interacciones del estudiante durante el curso es bajo con un rango menor a 80.
medio	Interacciones del estudiante durante el curso es medio con un rango entre 81 a 160.
alto	Interacciones del estudiante durante el curso es alto mayores a 161.

- servicios.- Atributo que contiene los servicios de comunicación como es el teléfono o celular de cada uno de los estudiantes del curso. En la siguiente tabla (ver TABLA XXXVII) se indica la distribución de este atributo.

TABLA XXXVII.

ATRIBUTO SERVICIOS.

Siglas	Descripción
1	Si el estudiante posee número de celular.
2	Si el estudiante posee número de teléfono convencional.
3	Si el estudiante posee ambos servicios (teléfono y celular).

- ciudad.- Atributo que contiene la ciudad donde pertenece cada estudiante del curso virtual. En la siguiente tabla (ver TABLA XXXVIII) se indica la distribución de este atributo.

TABLA XXXVIII.

ATRIBUTO CIUDAD.

Siglas	Descripción
L	Si es estudiante pertenece a la ciudad de Loja.
O	Si es estudiante pertenece a otra ciudad.

- edad.- Atributo que contiene los rangos de valores de la edad de los estudiantes del curso. En la siguiente tabla (ver TABLA XXXIX) se indica la distribución de este atributo.

TABLA XXXIX.

ATRIBUTO EDAD.

Siglas	Descripción
a	Estudiantes menores a 24 años
b	Estudiantes entre 25 y 29 años
c	Estudiantes mayores a 30 años

- genero.- Atributo que contiene los valores de masculino y femenino del estudiante del curso virtual. En la siguiente tabla (ver TABLA XL) se indica este atributo.

TABLA XL.

ATRIBUTO GENERO

Valor	Descripción
0	Género masculino del estudiante del curso.
1	Género femenino del estudiante del curso.

- estado_civil: Campo que contiene el estado civil que tiene cada estudiante del curso virtual. En la siguiente tabla (ver TABLA XLI) se indica este atributo.

TABLA XLI.

ATRIBUTO ESTADO_CIVIL.

Siglas	Descripción
S	Soltero
C	Casado
D	Divorciado
V	Viudo

- numeroHijos: Campo que describe si el estudiante del curso virtual tiene hijos o no. En la siguiente tabla (ver TABLA XLII) se indica este atributo.

TABLA XLII.

ATRIBUTO NUMEROHIJOS.

Siglas	Descripción
si	El estudiante tiene hijos
no	El estudiante no tiene hijos

- trabajo: Campo que describe si el estudiante del curso virtual trabaja o no. En la siguiente tabla (ver TABLA XLIII) se indica este atributo.

TABLA XLIII.

ATRIBUTO TRABAJO.

Siglas	Descripción
si	El estudiante trabaja
no	El estudiante no trabaja

- carrera: campo que contiene la carrera de cada estudiante, las mismas que se encuentran especificadas en la tabla anterior (ver TABLA XXXII).

2.3.1.4. Integración de los datos

En esta tarea se realizó la integración de la información en una sola base de datos que contiene la información de las interacciones de los estudiantes del curso de inglés de la MED los mismos que fueron proporcionados por el Departamento de Telecomunicación e Información.

Luego de realizar la integración de los datos, los cuales al inicio estuvieron en un archivo .XML quedaron recopilados en una sola BD, finalmente se integró todas las tablas con los atributos en una sola estructura de datos (ver TABLA XXXII) que fueron utilizados para crear el modelo de minería de datos que permitió cumplir el objetivo planteado.

En la siguiente figura se puede observar un fragmento del data set final, en otras palabras es la integración de todos los datos utilizados (ver Figura 48).

cedula	ciudad	genero	edad	estadocivil	numerointe...	servicios	interaccionesrec...	interaccionesexa...	interaccione...	carrera	modalidad...	horario	sectorVivien...	trabajo	numeroHijos
070118150E	PIÑAS	masculir	40	casado	417	3	58	92	84	Derecho	distancia	nocturnc	Urbano	No	3
190075839C	Loja	femenin(	22	soltero	202	3	34	54	27	Contabilidac	presencial	matutinc	Urbano	No	9
1104871296	SAN LUCA	masculir	25	casado	518	3	60	89	122	EducaciÃƒÂ³	semipresen	matutinc	Urbano	No	1
1104606643	Loja	femenin(	27	divorciado	169	3	25	44	32	Derecho	presencial	nocturnc	Urbano	No	1
1104459761	Loja	masculir	29	casado	14	3	0	0	0	Derecho	distancia	nocturnc	Urbano	No	0
1103969430	LOJA	masculir	33	casado	297	3	61	50	50	Derecho	distancia	nocturnc	Urbano	No	1
1900650944	Loja	femenin(	25	soltero	83	3	4	27	16	Contabilidac	presencial	matutinc	Urbano	No	1
1104894124	Loja	femenin(	23	casado	172	3	32	30	33	PsicologÃƒÂ	presencial	vespertii	Urbano	No	0
1104700057	Loja	masculir	27	soltero	216	3	25	27	36	Administraci	presencial	vespertii	Rural	No	1
0703379784	Santa Ros	femenin(	37	soltero	334	3	52	39	70	Contabilidac	distancia	matutinc	Urbano	Si	1
1104061377	loja	masculir	30	divorciado	87	3	29	25	16	Cultura FÃƒÂ	presencial	matutinc	Urbano	No	2
1104028038	LOJA	femenin(	33	soltero	112	3	14	38	18	QuÃƒÂ-mico	presencial	vespertii	Urbano	No	1
1722372081	LOJA	femenin(	27	soltero	404	3	95	77	80	Administraci	distancia	matutinc	Urbano	No	0
1104770654	Loja	femenin(	27	casado	76	2	0	31	21	FÃƒÂ-sico M	presencial	vespertii	Urbano	No	2
1104299019	Loja	masculir	28	soltero	89	3	0	28	18	IngenierÃƒÂ-	presencial	matutinc	Urbano	No	0
1715530422	Santo Don	masculir	36	soltero	243	3	7	56	42	Contabilidac	distancia	matutinc	Urbano	No	1
0705202299	PIÑAS	femenin(	26	soltero	510	3	36	111	137	Derecho	distancia	nocturnc	Urbano	No	1
1104673825	Piñas	masculir	27	soltero	210	3	27	47	38	Derecho	distancia	matutinc	Urbano	No	0
1715632202	Quito	masculir	35	soltero	232	3	51	34	32	Derecho	distancia	nocturnc	Urbano	No	2
1104190887	Loja	masculir	31	divorciado	257	3	25	55	73	Comunicaci	presencial	vespertii	Urbano	No	1
2100413992	QUITO	femenin(	27	soltero	181	2	31	36	25	Administraci	distancia	nocturnc	Urbano	No	0
1103764617	loja	femenin(	35	casado	4	3	0	0	0	InformÃƒÂ¡ti(	distancia	matutinc	Urbano	Si	1
1105218083	Loja	femenin(	23	soltero	219	3	21	46	46	PsicologÃƒÂ	presencial	nocturnc	Urbano	No	0
1104265382	Loja	femenin(	28	soltero	397	1	82	61	94	PsicologÃƒÂ	presencial	vespertii	Urbano	No	1
1104071640	Loja	masculir	27	casado	398	3	51	105	102	Administraci	presencial	null	Urbano	No	2

Figura 48: Integración de los datos

2.4. Plantear un modelo computacional mediante la técnica de Minería de Datos seleccionada anteriormente para determinar las interacciones de los estudiantes en el curso virtual de inglés.

En la presente actividad se realizará el modelo para determinar las interacciones de los estudiantes para ello se llevó a cabo la fase de modelado se encuentra a continuación:

2.4.1. Etapa IV: Modelado

En la presente etapa se realizó el modelado de minería de datos para ello se empleó el data set final de la etapa de preparación de los datos (sección 2.3.1.3.) conjuntamente con la técnica seleccionada y los algoritmos que pertenecen a dicha técnica, finalmente la selección de la herramienta de minería de datos para realizar los distintos procedimientos con la finalidad de obtener los resultados del modelo.

Entre las secciones está el cuadro comparativo de herramientas de minería de datos (sección 2.4.1.1.), seleccionar la herramienta de minería de datos (sección 2.4.1.2.), selección de la técnica de modelado (sección 2.5.1.), generar el plan de prueba (sección 2.5.2.) y construir el modelo.

2.5. Recolectar información de fuentes confiables sobre herramientas de minería de datos y seleccionar la que más se adapte al modelo computacional.

Existen numerosas herramientas de minería de datos tanto comerciales como de código abierto para realizar el procesamiento de los datos, sin embargo en el presente estudio se ha considerado algunas de ellas que son las más empleadas según los casos de éxito analizados anteriormente (Capítulo III: Casos de éxito de minería de datos).

- **SAS Enterprise Miner**

Es una herramienta de minería de datos comercializada, crea modelos predictivos y descriptivos precisos sobre grandes volúmenes de datos a través de diferentes fuentes mediante un proceso transparente, lo que permite

colaborar de manera más eficiente, incluye una interfaz de usuario intuitiva que incorpora los principios de diseño comunes establecidos para el software de SAS y herramientas de navegación adicionales para mover fácilmente alrededor del área de trabajo [1] - [4].

- **RapidMiner**

Es una herramienta de minería de datos desarrollado en Java, permite el tratamiento de procesos de análisis de datos mediante el encadenamiento de 500 operadores a través de un entorno gráfico, permite utilizar los algoritmos incluidos en Weka, contiene técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos, aprendizaje automático, etc. [5].

- **Weka**

Es una herramienta para el aprendizaje automático y minería de datos diseñado en Java, es de distribución de licencia GNU-GLP, contiene una colección de algoritmos para el análisis de datos y modelado predictivo, permite la visualización de datos, provee una interfaz gráfica. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un repositorio de datos [6].

- **Knime**

Está desarrollado sobre la plataforma Eclipse y programado en Java, su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos y predecir posibles resultados [7].

Es una plataforma de código abierto de fácil uso y comprensible para integración de datos, procesamiento, análisis, y exploración. Ofrece a los usuarios la capacidad de crear de forma visual flujos de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas [8-9].

2.5.1. Cuadro comparativo de Herramientas de Minería de Datos

Se realizó una recolección de información con referente a las herramientas de minería de datos en los cuales sobresalen cada una de las características de las herramientas más utilizadas en proyectos similares según el estudio de algunos casos de éxito, entre las que se ha seleccionado se tiene SAS Enterprise Miner, RapidMiner, Weka, Knime, mismas que se puede observar en la siguiente Tabla (ver Tabla XLIV):

TABLA XLIV.

CUADRO COMPARATIVO DE HERRAMIENTAS DE MINERÍA DE DATOS [5]- [9]

CARACTERÍSTICAS	HERRAMIENTAS			
	SAS Enterprise Miner	RapidMiner	Weka	Knime
Licencia libre	X	✓	✓	✓
Multiplataforma	✓	✓	✓	✓
Técnicas Descriptivas(agrupación)	✓	✓	✓	✓
Técnicas Predictivas(clasificación)	✓	✓	✓	X
Interfaz amigable	✓	✓	X	X
Permite visualización de datos	✓	✓	✓	✓
Flexibilidad	X	✓	✓	✓
Fácil de Configurar	X	✓	✓	X
Fácil de Instalar	✓	✓	✓	✓
Conversión de datos	✓	✓	✓	X
Dispone de un módulo de integración con Weka y R	X	✓	X	X
Procesamiento de datos	✓	✓	✓	✓
Validación del modelo	✓	✓	✓	X

En la tabla anterior (ver TABLA XLIV) se han mencionado algunas características que han sido tomadas en cuenta para la comparación de las mismas y posteriormente seleccionar según cumpla cada herramienta con dicha característica de tal forma se eligió la que más se adaptó al trabajo, cabe mencionar que el contenido de las mismas se elaboró en la Revisión Literaria del Capítulo IV: Minería de Datos sección 4.7 Herramientas de MD.

2.5.2. Selección de la Herramienta de Minería de Datos

Al realizar el análisis de las herramientas seleccionadas anteriormente (ver TABLA XLIV) en base a algunas de sus características se pudo seleccionar a la herramienta RAPIDMINER para llevar a cabo cada una de las actividades acerca del modelado del presente proyecto, ya que se adapta al trabajo de titulación, posee una licencia libre, combinación de modelos, interfaz amiga, multiplataforma, empleo de técnicas, además permite aplicar varios algoritmos de minería de datos, con la integración de algunos plugins se puede utilizar los algoritmos incluidos en Weka, para poder generar el modelo y validarlo. Así mismo se puede mencionar que no existe una herramienta que contenga todas las funcionalidades pero RapidMiner es ampliamente usada y probada a nivel internacional en aplicaciones empresariales, de gobierno y academia, posee gran cantidad de operadores que permiten generar el modelo para determinar las interacciones en entornos virtuales.

2.6. Implementar el modelo computacional en la herramienta seleccionada.

Una vez que se seleccionó la herramienta de minería de datos para la realización del presente proyecto se procedió a seleccionar la técnica de minería de datos adecuada para la realización del modelo.

2.6.1. Seleccionar técnica de modelado

Es muy importante la selección de la técnica de minería de datos ya que mediante esta y conjuntamente con los algoritmos que se seleccione según la clasificación de dicha técnica se podrá realizar el modelo para determinar las interacciones de los estudiantes, es por ende que se recolecto información de algunas de las técnicas y casos de éxito que hayan utilizado en sus proyectos dichas técnicas, a continuación se puede observar en la siguiente tabla (ver TABLA XLV).

TABLA XLV.

TÉCNICAS PARA LA GENERACIÓN DEL MODELO.

Técnica	Descripción	Algoritmo	Casos de éxito
Agrupamiento O clustering	El análisis de clusters es utilizado para obtener una visión de la distribución de los datos, para observar las características de cada clúster [7]. A diferencia de la clasificación, el clustering no depende de clases [4].	Simple-Kmeans	- Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje [3], [5]. - Uso de ambientes virtuales de aprendizaje en la enseñanza de la ingeniería [11].
Clasificación	Son estructuras que representan conjuntos de decisiones [5]. Útiles para explorar un conjunto de datos [10]. Permite organización eficiente de conjunto de datos [11].	J48 REPTree ID3/C4.5	- Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL [5]. - Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL [2].

Para la selección de la técnica de minería de datos para el desarrollo del proyecto se lo realizó mediante un cuadro comparativo que se lo puede observar en la TABLA XLV en donde se describe cada técnica y algunos casos de éxito que han empleado estas técnicas para alcanzar resultados idóneos, así mismo se pudo identificar la técnica apropiada para el proyecto en ejecución es la de clasificación para el análisis de las interacciones de los estudiantes en el entorno virtual de aprendizaje ya que es apropiada al problema a resolver y se tiene una comprensión de la técnica.

Por otra parte se ha podido identificar una gran mayoría de casos de éxito relacionados con el estudio de entornos virtuales con diferentes fines en donde la técnica más aplicada es la antes mencionada, siendo esta eficiente en el momento de analizar grandes cantidades de datos y posteriormente el desarrollo de un modelo para mediante este se pueda tomar decisiones y mejorar el uso de los entornos virtuales de aprendizaje.

2.6.2. Generar el plan de prueba

La generación del plan de pruebas consiste en probar la calidad y validez de los resultados obtenidos por el modelo**,** por ende es necesario generar un plan de pruebas mediante el cual se pueda probar la validez del modelo generado, para ello se trabajó con los datos pertenecientes a los estudiantes del curso virtual de inglés de la MED los cuales se los dividió en dos grupos uno para entrenamiento y el otro para emplearlo en la validación del modelo.

El conjunto de datos para entrenamiento es del 67% y el conjunto de datos restantes se los utilizó para realizar la validación de tal manera que da un resultado del 100% de datos utilizados para el modelado.

A continuación, se describe el plan de pruebas realizado con los diferentes algoritmos clasificados de la siguiente forma:

- **Algoritmos de Reglas de decisión**

  Los algoritmos utilizados dentro de esta clasificación corresponden a JRip, Ridor, Prism, K-NN, en donde se utilizó el 67% del conjunto de datos para entrenamiento y el 33% para la validación.

- **Algoritmos de Arboles de decisión**

Los algoritmos utilizados dentro de esta clasificación corresponden a CHAID, Decision Tree, ID3, J48, en donde se utilizó el 67% del conjunto de datos para entrenamiento y el 33% para la validación.

2.6.3. Construir el modelo

Para generar el modelo se empleó cada uno de los algoritmos seleccionados anteriormente pertenecientes a la técnica de clasificación de minería de datos, los mismos que se encuentran en la herramienta RapidMiner en donde se utilizó la base de datos donde se encuentra la información de los estudiantes del curso virtual de inglés de la MED con la finalidad de determinar las interacciones de los estudiantes.

Los parámetros a evaluar en los modelos generados son los siguientes: instancias clasificadas correctamente (accuracy), instancias clasificadas incorrectamente (clasification_error), estadística de Kappa (Kappa), error cuadrático (squared_error), error relativo (relative_error), error absoluto (absolute_error).

Las secciones que se tienen están los algoritmos pertenecientes a las reglas de decisión (sección 2.6.3.1.) y los algoritmos pertenecientes a los arboles de decisión (sección 2.6.3.2.)

2.6.3.1. Algoritmos pertenecientes a las Reglas de decisión

Existen varios algoritmos pertenecientes a la presente técnica pero solo se han tomado en cuenta 4 como es el algoritmo JRip (sección 2.6.3.1.1.), Ridor (sección 2.6.3.1.2.), K-NN (sección 2.6.3.1.3.) y Prism (sección 2.6.3.1.4.).

2.6.3.1.1. Algoritmo JRip

Mediante las reglas de clasificación se pudo observar de manera estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta es el número minimo de divisiones que se puede dar por cada nodo (minimal size for Split=25), el tamaño minimo de cada hoja (minimal leaf size=25), la ganacia mínima (minimal gain=0.01), profundidad máxima (maximal depth= 3), confidencia de clasificación (confidence=0.1),

criterio de confidencialidad (criterion= information gain) y clasificación correcta (accuracy) que maximiza la precisión.

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 49 y 51).

❖ Proceso de entrenamiento

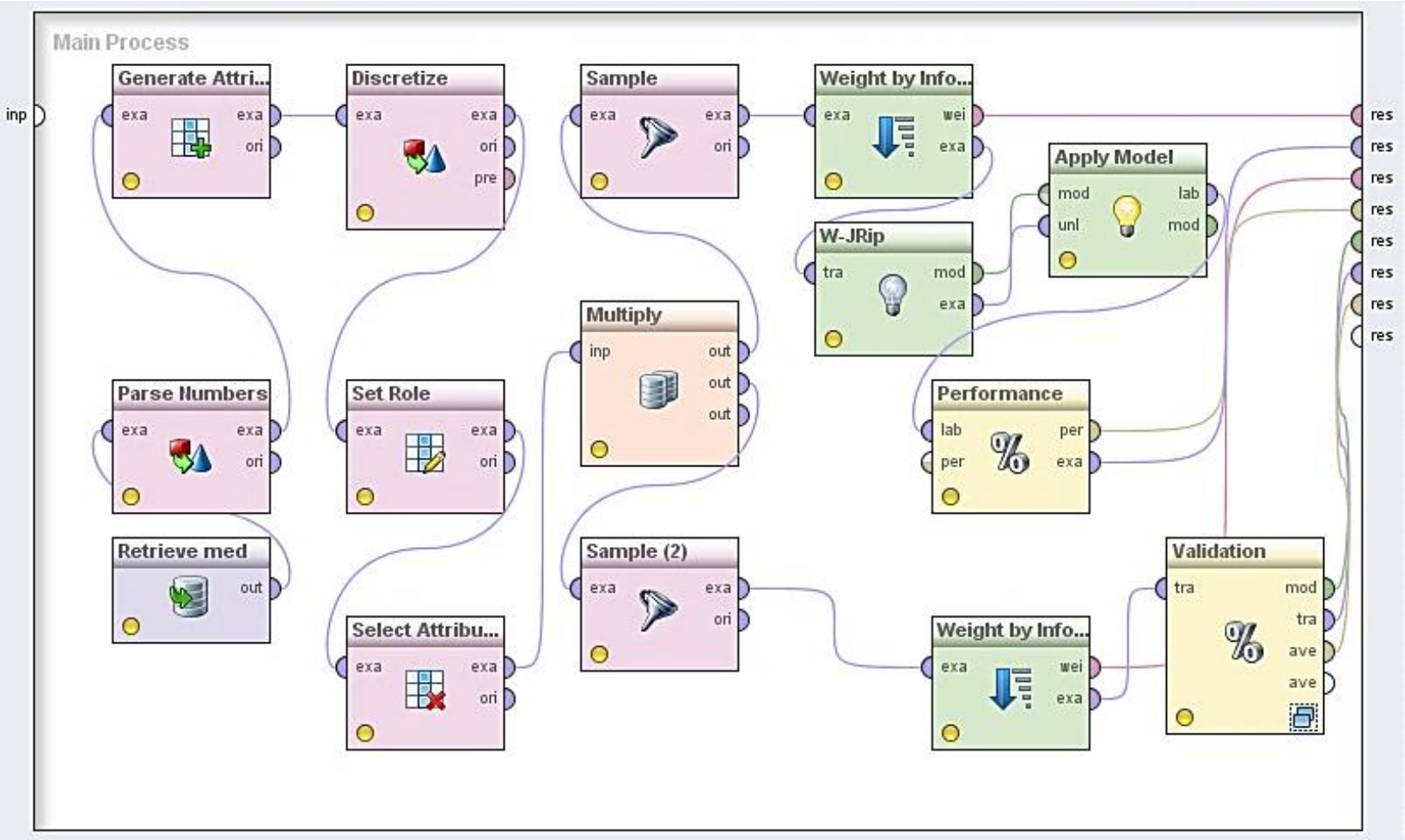


Figura 49: Proceso de Entrenamiento algoritmo JRip

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA XLVI):

TABLA XLVI.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO JRIP

JRip Entrenamiento	
Criterios	**Valores**
Accuracy	94.41%
Classification_error	5.59%
Kappa	0.880
Absolute_error	0.102
Relative_error	10.18
Root_mean_squared_error	0.226
Root_relative_squared_error	0.802
Squered_error	0.051

- **Matriz de confusión del entrenamiento**

En la matriz de confusión del proceso de entrenamiento (ver Figura 50) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 94.41%

	true bajo	true medio	true alto
pred. bajo	206	19	0
pred. medio	15	453	2
pred. alto	0	4	17

Figura 50: Matriz de confusión del Entrenamiento algoritmo JRip

❖ Proceso de Validación

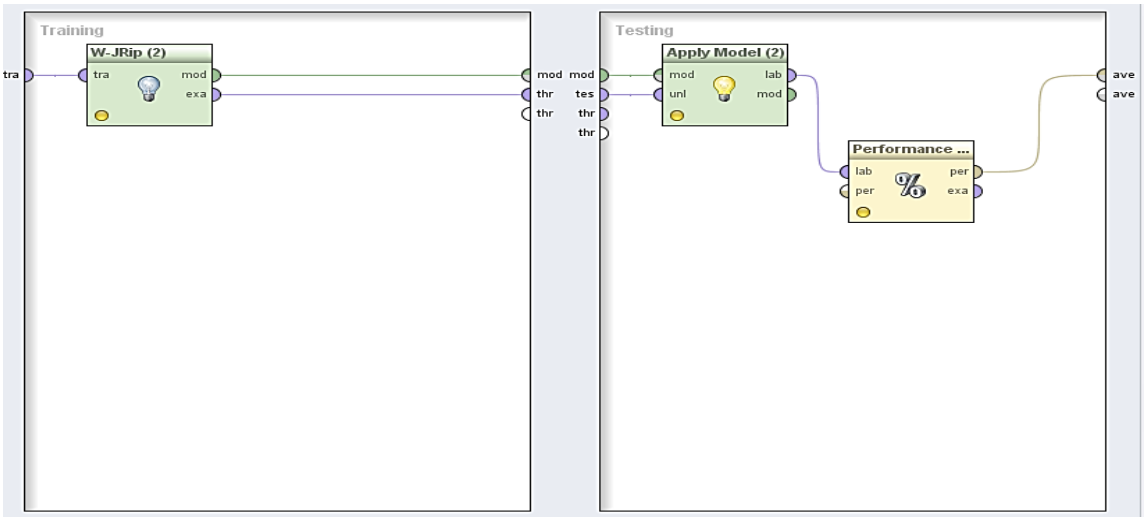


Figura 51: Proceso de validación algoritmo JRip

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA XLVII):

TABLA XLVII.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO JRIP

JRip Validación	
Criterios	**Valores**
Accuracy	92.63%
Classification_error	7.37%
Kappa	0.820
Absolute_error	0.132
Relative_error	13.17
Root_mean_squared_error	0.247
Root_relative_squared_error	0.980
Squered_error	0.071

❖ Matriz de confusión de la validación

En la matriz de confusión del proceso de validación (ver Figura 52) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 92.63% +/- 5.77% (mikro: 92.61%)

	true bajo	true medio	true alto
pred. bajo	83	7	1
pred. medio	15	239	2
pred. alto	0	1	4

Figura 52: Matriz de confusión de la validación del algoritmo JRip

❖ Reglas del Algoritmo JRip

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar las interacciones de los estudiantes, se puede observar en la siguiente figura (ver Figura 53).

W-JRip

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

(interaccionesexamen = IEA) and (carrera = Derecho) and (interaccionesrecurso = IRA) => numerointeracciones=alto (5.0/1.0)
(interaccionesexamen = IEB) and (interaccionesrecurso = IRB) and (interaccionestareas = ITB) => numerointeracciones=bajo (89.0/6.0)
 => numerointeracciones=medio (258.0/18.0)

Number of Rules : 3
```

Figura 53: Reglas generadas por el algoritmo JRip

Las reglas generadas por el algoritmo JRip (ver figura 53) se describen a continuación:

- Cuando los estudiantes tienen interacciones altas con los exámenes y los recursos, entonces las interacciones en el curso virtual son altas.
- Cuando los estudiantes tienen interacciones bajas con los exámenes, los recursos y las tareas, entonces las interacciones son medias.

2.6.3.1.2. Algoritmo Ridor

Mediante el árbol de clasificación se pudo observar de manera gráfica y estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta para la utilización del presente algoritmo fue establecer el número de pliegues para las excepciones (F=3), número aleatorio para los datos con el fin de obtener una mejor regla (S=1), una banderas para la tasa de error de los datos(A=false); (M=false) y se estableció los pesos mínimos (N=2), de tal forma se puede obtener algunas reglas de interés en referente a las interacciones de los estudiantes.

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 54 y 56).

❖ Proceso de entrenamiento

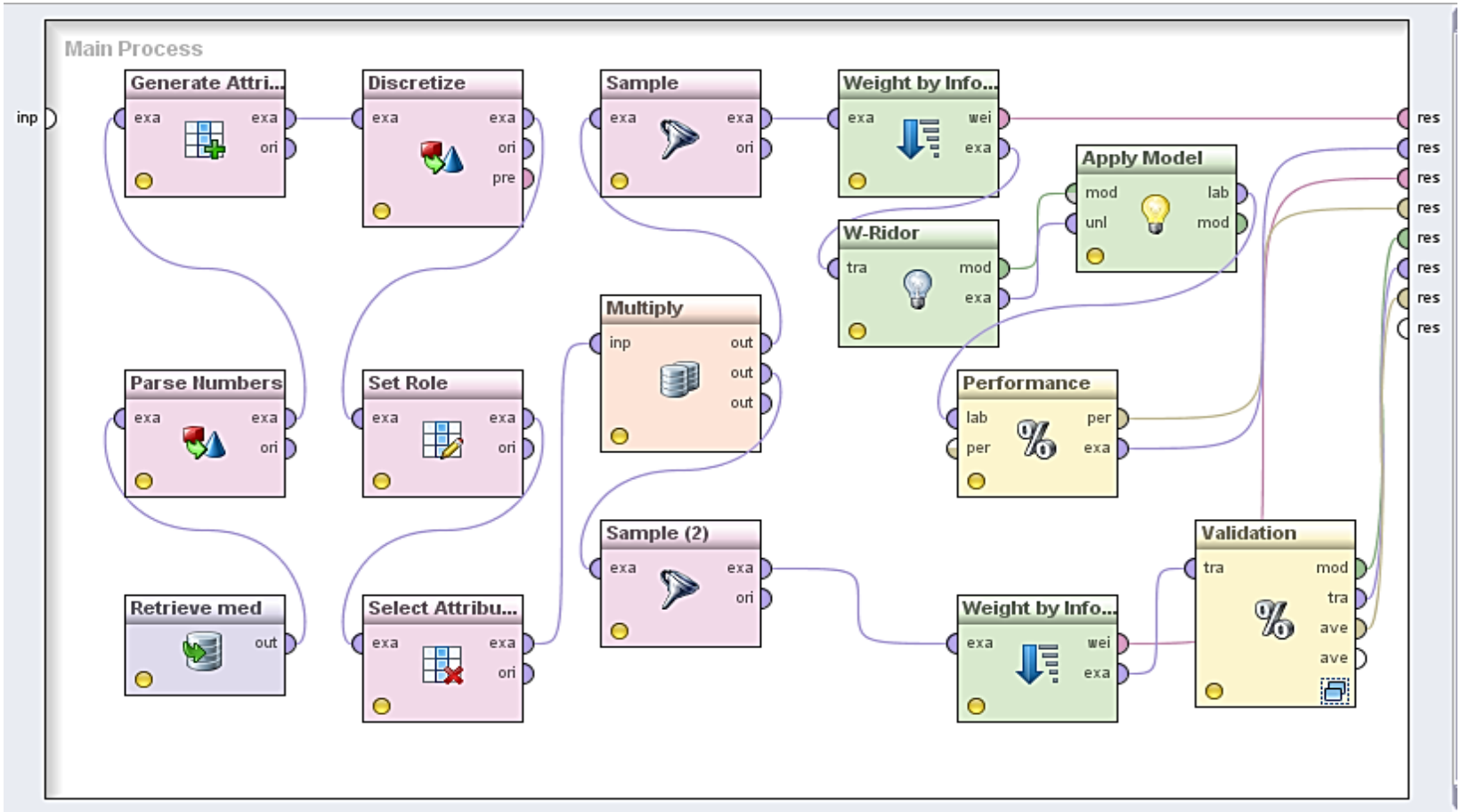


Figura 54: Proceso de Entrenamiento algoritmo RIDOR

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA XLVIII):

TABLA XLVIII.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO RIDOR

Ridor Entrenamiento	
Criterios	**Valores**
Accuracy	89.66%
Classification_error	10.34%
Kappa	0.770
Absolute_error	0.103
Relative_error	10.34
Root_mean_squared_error	0.321
Root_relative_squared_error	1.140
Squered_error	0.103

- **Matriz de confusión del entrenamiento**

En la matriz de confusión del proceso de entrenamiento (ver Figura 55) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 89.66%

	true bajo	true medio	true alto
pred. bajo	188	28	0
pred. medio	33	447	12
pred. alto	0	1	7

Figura 55: Matriz de confusión del Entrenamiento algoritmo RIDOR

❖ Proceso de Validación

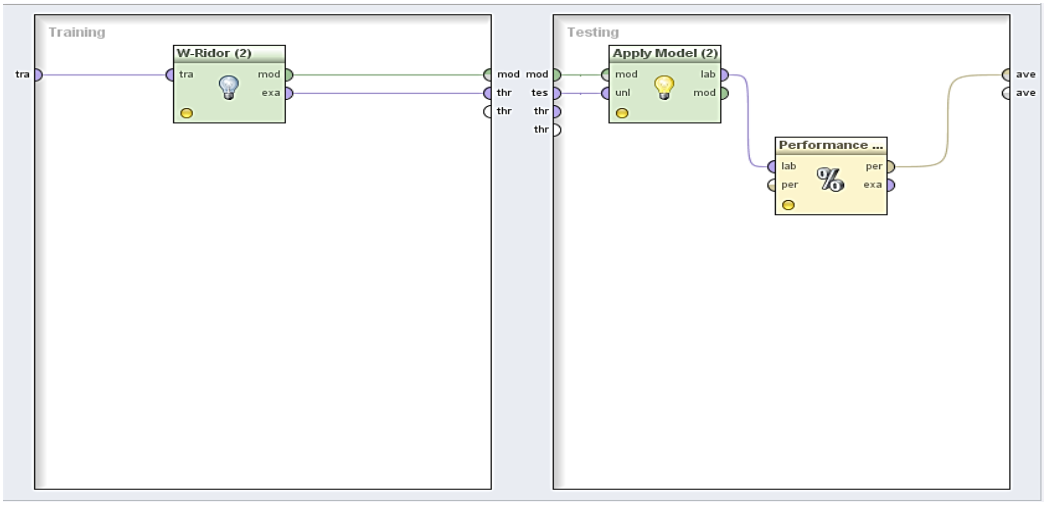


Figura 56: Proceso de validación algoritmo RIDOR

En el proceso de validación que consta cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del mismo se ha obtenido los siguientes resultados (ver TABLA XLIX):

TABLA XLIX.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO RIDOR

Ridor Validación	
Criterios	**Valores**
Accuracy	88.66%
Classification_error	11.34%
Kappa	0.741
Absolute_error	0.113
Relative_error	11.34
Root_mean_squared_error	0.309
Root_relative_squared_error	1.229
Squered_error	0.113

❖ Matriz de confusión de la validación

En la matriz de confusión del proceso de validación (ver Figura 57) se puede observar sobre la diagonal principal las instancias clasificadas correctamente o datos clasificados correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 88.66% +/- 6.79% (mikro: 88.64%)

	true bajo	true medio	true alto
pred. bajo	83	8	2
pred. medio	15	228	4
pred. alto	0	11	1
class recall	84.69%	92.31%	14.29%

Figura 57: Matriz de confusión de la validación del algoritmo RIDOR

❖ Reglas del Algoritmo RIDOR

Mediante el uso del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar las interacciones de los estudiantes, se puede observar en la siguiente figura (ver Figura 58).

W-Ridor

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

numerointeracciones = bajo  (352.0/254.0)
          Except (interaccionesexamen = IEM) and (interaccionestareas = ITM) => numerointeracciones = medio  (56.0/0.0) [19.0/0.0]
          Except (interaccionesexamen = IEM) and (ciudad = O) and (numeroHijos = No) => numerointeracciones = medio  (16.0/0.0) [5.0/0.0]
          Except (interaccionesexamen = IEA) => numerointeracciones = medio  (41.0/0.0) [15.0/0.0]
          Except (interaccionesexamen = IEM) => numerointeracciones = medio  (55.0/5.0) [35.0/7.0]
          Except (interaccionesrecurso = IRM) and (ciudad = L) => numerointeracciones = medio  (5.0/0.0) [1.0/0.0]
          Except (interaccionesrecurso = IRA) => numerointeracciones = medio  (3.0/0.0) [1.0/0.0]
                   Except (carrera = Derecho) and (interaccionesexamen = IEA) => numerointeracciones = alto  (3.0/0.0) [2.0/1.0]
          Except (interaccionesrecurso = IRM) => numerointeracciones = medio  (6.0/3.0) [2.0/0.0]
          Except (carrera = AdministraciÃƒÂ³n de Empresas) and (edad = c) => numerointeracciones = medio  (3.0/2.0) [1.0/0.0]

Total number of rules (incl. the default rule): 10
```

Figura 58: Reglas generadas por el algoritmo RIDOR

Las reglas generadas por el algoritmo RIDOR (ver figura 58) se describen a continuación:

- Las interacciones de los estudiantes en el curso virtual son bajas excepto cuando:
  - Las interacciones en los exámenes y en las tareas son medias, entonces las interacciones en el curso virtual es media.
  - Las interacciones en los exámenes es media y pertenece a otra ciudad y no tiene hijos, entonces las interacciones en el curso virtual es media.
  - Las interacciones con los exámenes es alta, entonces las interacciones en el curso virtual es media.
  - Las interacciones con los recursos es media y pertenece a la ciudad de Loja, entonces las interacciones en el curso virtual es media.
  - Las interacciones con los recursos es media, entonces las interacciones en el curso virtual es media.

2.6.3.1.3. Algoritmo K-NN

Mediante la utilización del algoritmo perteneciente a la técnica de clasificación se generó el árbol de clasificación que se puede observar de manera gráfica y estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta se tiene la clasificación de los vecinos más cercanos (K=1), bandera que se establece para el peso de los vecinos (weighted vote=false), tipo de medidencontrar el vecino más cercano (measure types=mixed measures), (mixed measure=Mixed Euclia para dean Distance).

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen algunos operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 59 y 61).

Proceso de entrenamiento

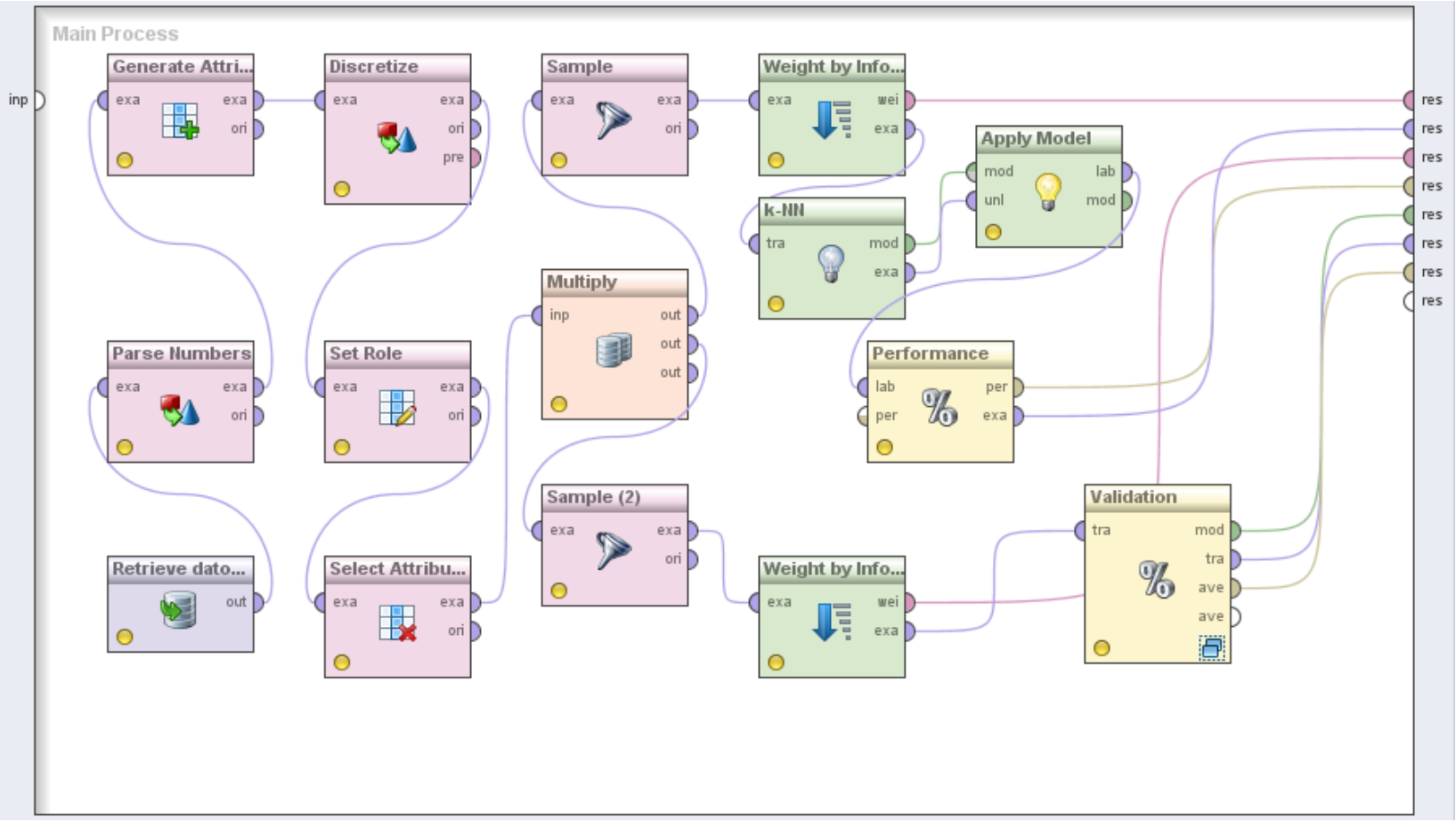


Figura 59: Proceso de Entrenamiento algoritmo K-NN

En el proceso de entrenamiento que consta cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se ha establecido el valor de 10 para que internamente el algoritmo realice dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA L):

TABLA L.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO K-NN

K-NN **Entrenamiento**	
Criterios	**Valores**
Accuracy	98.74%
Classification_error	1.26%
Kappa	0.973
Absolute_error	0.013
Relative_error	1.26
Root_mean_squared_error	0.112
Root_relative_squared_error	0.397
Squered_error	0.013

❖ **Matriz de confusión del Entrenamiento algoritmo K-NN**

En la matriz de confusión del proceso de entrenamiento (ver Figura 60) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 98.74%

	true bajo	true medio	true alto
pred. bajo	217	5	0
pred. medio	4	471	0
pred. alto	0	0	19

Figura 60: Matriz de confusión del Entrenamiento algoritmo K-NN

❖ Proceso de Validación

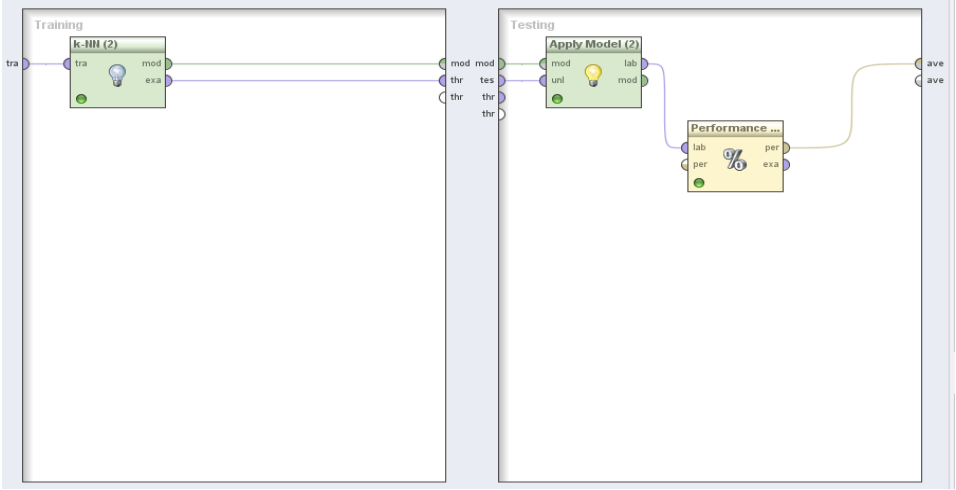


Figura 61: Proceso de validación algoritmo K-NN

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LI):

TABLA LI.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO K-NN

K-NN Validación	
Criterios	**Valores**
Accuracy	84.15%
Classification_error	15.85%
Kappa	0.635
Absolute_error	0.158
Relative_error	15.85
Root_mean_squared_error	0.388
Root_relative_squared_error	1.511
Squered_error	0.158

❖ **Matriz de confusión de la validación del algoritmo K-NN**

En la matriz de confusión del proceso de validación (ver Figura 62) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 84.15% +/- 7.14% (mikro: 84.09%)

	true bajo	true medio	true alto
pred. bajo	81	30	0
pred. medio	17	213	5
pred. alto	0	4	2

Figura 62: Matriz de confusión de la validación del algoritmo K-NN

❖ **Reglas del Algoritmo K-NN**

Mediante el uso del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar las interacciones de los estudiantes, se puede observar en la siguiente figura (ver Figura 63).

KNNClassification

```
1-Nearest Neighbour model for classification.
The model contains 352 examples with 11 dimensions of the following classes:
  bajo
  medio
  alto
```

Figura 63: Reglas generadas por el algoritmo K-NN

2.6.3.1.4. Algoritmo Prism

Mediante el árbol de clasificación se pudo observar de manera gráfica y estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta es el modo de depuración (D=false).

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 64 y 66).

Proceso de entrenamiento

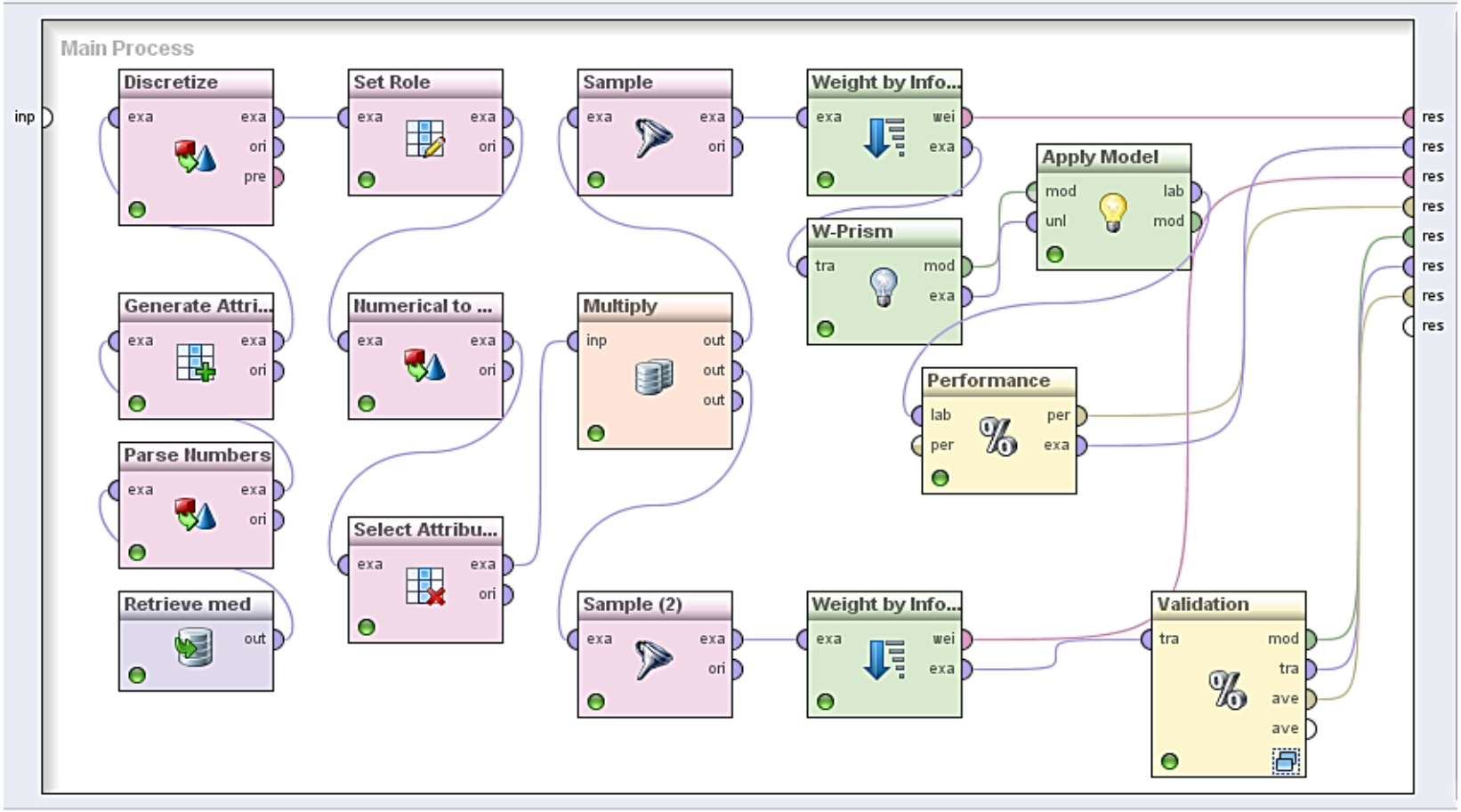


Figura 64: Proceso de Entrenamiento algoritmo PRISM

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LII):

TABLA LII.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO PRISM

Prism Entrenamiento	
Criterios	**Valores**
Accuracy	98.46%
Classification_error	1.54 %
Kappa	0.971
Absolute_error	0.015
Relative_error	1.54
Root_mean_squared_error	0.124
Root_relative_squared_error	0.462
Squered_error	0.015

- **Matriz de confusión del Entrenamiento algoritmo PRISM**

En la matriz de confusión del proceso de entrenamiento (ver Figura 65) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 98.46%

	true bajo	true alto	true medio
pred. bajo	240	8	0
pred. alto	0	420	3
pred. medio	0	0	45

Figura 65: Matriz de confusión del Entrenamiento algoritmo PRISM

❖ Proceso de Validación

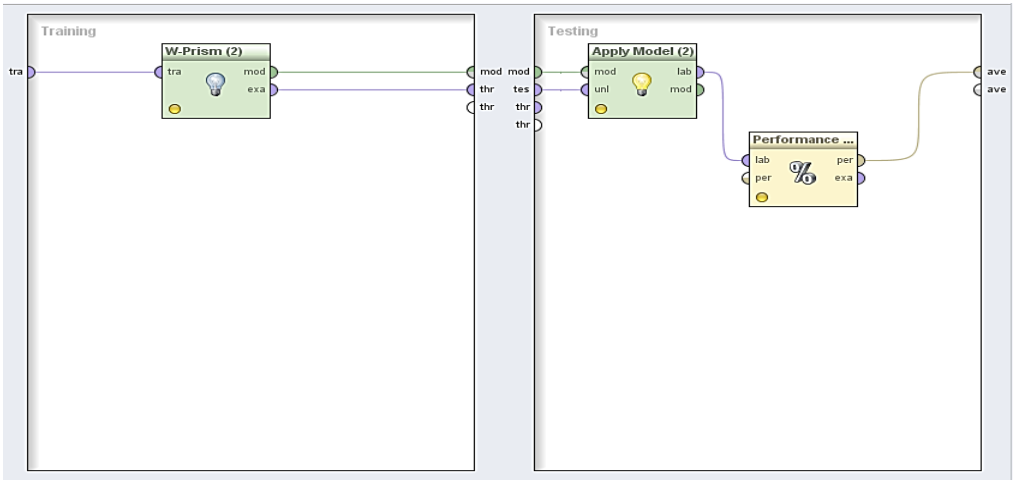


Figura 66: Proceso de validación algoritmo PRISM

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LIII):

TABLA LIII.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO PRISM

Prism Validación	
Criterios	**Valores**
Accuracy	77.06%
Classification_error	22.94%
Kappa	0.576
Absolute_error	0.261
Relative_error	26.06
Root_mean_squared_error	0.504
Root_relative_squared_error	2.374
Squered_error	0.261

❖ Matriz de confusión de la validación del algoritmo PRISM

En la matriz de confusión del proceso de validación (ver Figura 67) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 77.06% +/- 8.38% (mikro: 76.99%)

	true bajo	true alto	true medio
pred. bajo	91	41	5
pred. alto	14	168	12
pred. medio	0	9	12

Figura 67: Matriz de confusión de la validación del algoritmo PRISM

❖ Reglas del Algoritmo PRISM

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar las interacciones de los estudiantes, se puede observar en la siguiente figura (ver Figura 68).

W-Prism

```
Prism rules
----------
If interaccionesexamen = IEB
   and interaccionesrecurso = IRB
   and trabajo = Si then bajo
If interaccionesexamen = IEB
   and interaccionesrecurso = IRB
   and carrera = Derecho
   and numeroHijos = No then bajo
If interaccionesexamen = IEB
   and interaccionesrecurso = IRB
   and carrera = PsicologÃƒÂ-a Infantil y EducaciÃƒÂ³n Parvularia
   and interaccionestareas = ITB then bajo
If interaccionesexamen = IEB
   and interaccionesrecurso = IRB
   and genero = 1
   and numeroHijos = No then bajo
```

```
If interaccionesexamen = IEB
   and interaccionesrecurso = IRB
   and numeroHijos = Si
   and edad = a then bajo

If interaccionesexamen = IEB
   and carrera = Derecho
   and ciudad = L
   and servicios = 3
   and trabajo = No
   and genero = 2
   and estadocivil = V
   and edad = c
   and interaccionesrecurso = IRB
   and interaccionestareas = ITB
   and numeroHijos = Si then bajo

If genero = 1
   and carrera = AdministraciÃƒÂ³n de Empresas
   and interaccionestareas = ITB
   and numeroHijos = Si
   and ciudad = O
   and servicios = 3
   and trabajo = No
   and estadocivil = V
   and edad = c
   and interaccionesrecurso = IRB
   and interaccionesexamen = IEM then bajo
If interaccionesrecurso = IRM
   and edad = b
   and genero = 1 then alto

If interaccionesexamen = IEM
   and edad = c
   and trabajo = Si then alto
If interaccionesrecurso = IRM
   and numeroHijos = No
   and interaccionestareas = ITM then alto

If interaccionesexamen = IEM
   and ciudad = O
   and interaccionesrecurso = IRM
   and numeroHijos = Si then alto
```

```
If interaccionesexamen = IEM
   and ciudad = O
   and edad = c
   and interaccionesrecurso = IRA
   and interaccionestareas = ITB then alto
If interaccionesexamen = IEM
   and interaccionestareas = ITA
   and carrera = Derecho
   and servicios = 3
   and trabajo = No
   and genero = 1
   and ciudad = O
   and estadocivil = V
   and edad = c
   and interaccionesrecurso = IRA
   and numeroHijos = Si then alto
If interaccionesrecurso = IRA
   and interaccionestareas = ITA
   and interaccionesexamen = IEM
   and servicios = 3
   and carrera = Derecho
   and trabajo = No
   and genero = 1
   and ciudad = O
   and estadocivil = V
   and edad = c
   and numeroHijos = Si then medio
If interaccionesrecurso = IRA
   and genero = 1
   and interaccionestareas = ITM
   and interaccionesexamen = IEM then medio
```

Figura 68: Reglas generadas por el algoritmo PRISM

Las reglas generadas por el algoritmo PRISM (ver figura 68) se describen a continuación:

- Si las interacciones con los exámenes y los recursos son bajas y trabaja, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes y los recursos son bajas y trabaja, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes y los recursos son bajas y el género es femenino y no tiene hijos, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes y los recursos son bajas y tiene hijos y es menor a 25 años, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes, las tareas y los recursos son bajas y pertenece a la ciudad de Loja y no trabaja y posee todos los servicios y el género es femenino y es mayor a 29 años y tiene hijos, entonces las interacciones en el curso virtual es bajo.

- Si tiene entre 25 y 29 años y pertenece a la ciudad de Loja y el género es femenino y posee todos los servicios, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los recursos es medio y tiene entre 25 y 29 años y el género es femenino, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes es medio y es mayor a 29 años y no trabaja, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes y los recursos es medio y no trabaja, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes y los recursos es medio y no tiene hijos, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes y los recursos es medio y pertenece a otra ciudad y no tiene hijos, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes es medio y las interacciones con los recursos es alto y pertenece a otra ciudad y es mayor a 29 años, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y tiene todos los servicios y no trabaja y pertenece a otra ciudad y es mayor a 29 años, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y tiene todos los servicios y no trabaja y el género es masculino y pertenece a otra ciudad y es mayor a 29 años, entonces las interacciones en el curso virtual es medio.
- Si las interacciones con los recursos es alto y las interacciones con los exámenes y las tareas es medio y el género es masculino, entonces las interacciones en el curso virtual es medio.
- Si las interacciones con los recursos es alto y las interacciones con las tareas es medio y el género es masculino y posee todos los servicios y es mayor a 29 años, entonces las interacciones en el curso virtual es medio.

2.6.3.2. Algoritmos pertenecientes a los Árboles de decisión

Los algoritmos que se ha tomado en cuenta se tienen Chaid (sección 2.6.3.2.1.), Decision Tree (sección 2.6.3.2.2.), ID3 (sección 2.6.3.2.3.) y el J48 (sección 2.6.3.2.4.).

2.6.3.2.1. Algoritmo CHAID

Mediante la utilización del algoritmo se estableció valores a los parámetros los cuales se detallan a continuación:

La longitud mínima de cada nodo (Minimal size for Split=4), tamaño mínimo de un nodo hoja (Minimal leaf size=2), ganancia mínima de un nodo (Minimal gain=0.1), profundidad máxima o tamaño del árbol de decisión (Maximal deph=20), nivel de confianza utilizada para el cálculo de error (Confidence=0.25), número de nodos probados (Number of prepruning=3).

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes gráficas (ver Figura 69 y 71).

❖ Proceso de entrenamiento

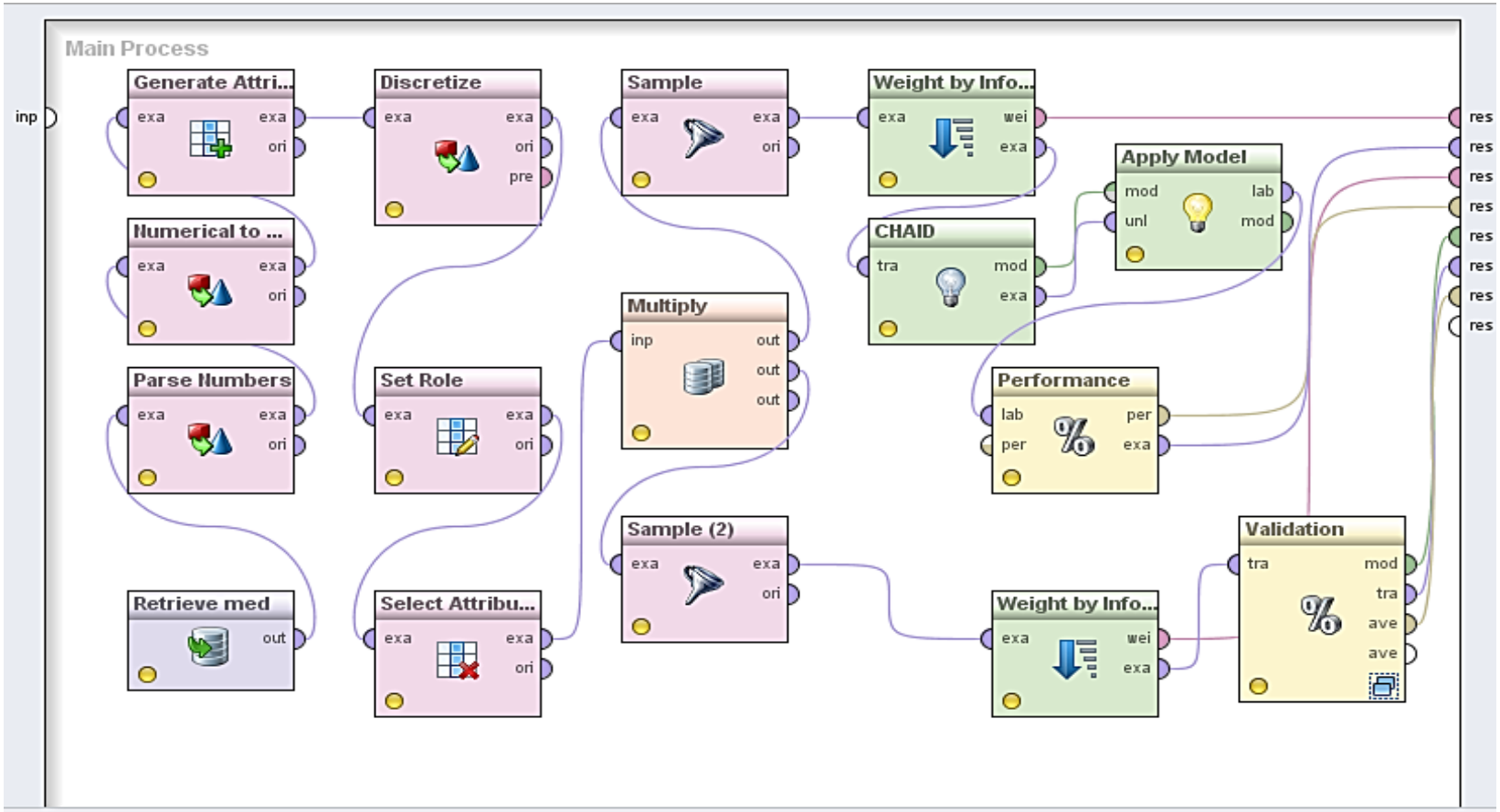


Figura 69: Proceso de Entrenamiento algoritmo CHAID

En el proceso de entrenamiento que consta cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LIV):

TABLA LIV.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO CHAID

CHAID Entrenamiento	
Criterios	**Valores**
Accuracy	91.06%
Classification_error	8.94%
Kappa	0.806
Absolute_error	0.124
Relative_error	12.41
Root_mean_squared_error	0.252
Root_relative_squared_error	0.893
Squered_error	0.063

- **Matriz de confusión del Entrenamiento algoritmo CHAID**

En la matriz de confusión del proceso de entrenamiento (ver Figura 70) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 91.06%

	true bajo	true medio	true alto
pred. bajo	196	30	2
pred. medio	24	445	6
pred. alto	1	1	11

Figura 70: Matriz de confusión del Entrenamiento algoritmo CHAID

❖ Proceso de Validación

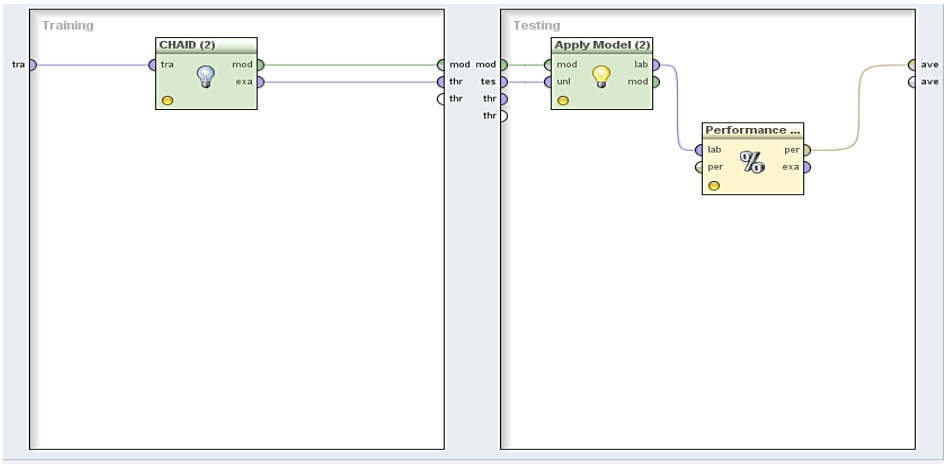


Figura 71: Proceso de validación algoritmo CHAID

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LV):

TABLA LV.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO CHAID

CHAID Validación	
Criterios	**Valores**
Accuracy	73.58%
Classification_error	26.42 %
Kappa	0.407
Absolute_error	0.274
Relative_error	27.35
Root_mean_squared_error	0.410
Root_relative_squared_error	5.456
Squered_error	0.174

❖ Matriz de confusión de la validación del algoritmo CHAID

En la matriz de confusión del proceso de validación (ver Figura 72) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 73.58% +/- 9.06% (mikro: 73.58%)

	true bajo	true medio	true alto
pred. bajo	61	47	1
pred. medio	35	196	4
pred. alto	2	4	2

Figura 72: Matriz de confusión de la validación del algoritmo CHAID

❖ Reglas del Algoritmo CHAID

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo, se puede observar en la siguiente figura (ver Figura 73).

Tree

```
trabajo = No
|   genero = 1
|   |   ciudad = L
|   |   |   servicios = 0: medio {bajo=2, medio=2, alto=0}
|   |   |   servicios = 1
|   |   |   |   edad = a
|   |   |   |   |   interaccionesrecurso = IRB
|   |   |   |   |   |   interaccionestareas = ITB: bajo {bajo=1, medio=1, alto=0}
|   |   |   |   |   |   interaccionestareas = ITM: medio {bajo=0, medio=2, alto=0}
|   |   |   |   edad = b: medio {bajo=0, medio=6, alto=0}
|   |   |   |   edad = c: medio {bajo=0, medio=3, alto=0}
|   |   |   servicios = 2: medio {bajo=0, medio=5, alto=0}
|   |   |   servicios = 3
|   |   |   |   estadocivil = C
|   |   |   |   |   edad = b: bajo {bajo=1, medio=1, alto=0}
|   |   |   |   |   edad = c: medio {bajo=1, medio=1, alto=0}
|   |   |   |   estadocivil = S
|   |   |   |   |   edad = a
|   |   |   |   |   |   interaccionesrecurso = IRB
|   |   |   |   |   |   |   interaccionestareas = ITB
|   |   |   |   |   |   |   |   numeroHijos = No: bajo {bajo=2, medio=0, alto=0}
```

```
|   |   ciudad = 0
|   |   |   estadocivil = C: medio {bajo=1, medio=3, alto=0}
|   |   |   estadocivil = D: medio {bajo=1, medio=1, alto=0}
|   |   |   estadocivil = S: medio {bajo=0, medio=3, alto=0}
|   genero = 2
|   |   servicios = 0: bajo {bajo=5, medio=1, alto=0}
|   |   servicios = 1
|   |   |   ciudad = L
|   |   |   |   estadocivil = C: bajo {bajo=3, medio=1, alto=0}
|   |   |   |   estadocivil = S
|   |   |   |   |   edad = a
|   |   |   |   |   |   interaccionestareas = ITB: bajo {bajo=2, medio=0, alto=0}
|   |   |   |   |   |   interaccionestareas = ITM: medio {bajo=0, medio=2, alto=0}
|   |   |   |   estadocivil = V
|   |   |   |   |   edad = a
|   |   |   |   |   |   interaccionestareas = ITB
|   |   |   |   |   |   |   interaccionesrecurso = IRA: medio {bajo=0, medio=2, alto=0}
|   |   |   |   |   |   |   interaccionesrecurso = IRB: bajo {bajo=5, medio=2, alto=0}
|   |   |   |   |   |   interaccionestareas = ITM: medio {bajo=0, medio=3, alto=0}
|   |   |   |   |   edad = b
|   |   |   |   |   |   interaccionesexamen = IEA: medio {bajo=0, medio=2, alto=0}
|   |   |   |   |   |   interaccionesexamen = IEB: bajo {bajo=1, medio=1, alto=0}
|   |   |   |   |   |   interaccionesexamen = IEM: medio {bajo=1, medio=4, alto=0}
|   |   |   |   |   edad = c: medio {bajo=0, medio=5, alto=0}
|   |   servicios = 2
|   |   |   ciudad = L
|   |   |   |   estadocivil = V
|   |   |   |   |   interaccionesexamen = IEB: bajo {bajo=2, medio=0, alto=0}
|   |   |   |   |   interaccionesexamen = IEM: medio {bajo=1, medio=2, alto=0}
|   |   |   ciudad = O: bajo {bajo=1, medio=1, alto=0}
|   |   servicios = 3
|   |   |   ciudad = L
|   |   |   |   estadocivil = C: medio {bajo=0, medio=7, alto=0}
|   |   |   |   estadocivil = S
|   |   |   |   |   edad = a
|   |   |   |   |   |   interaccionesrecurso = IRB: medio {bajo=1, medio=3, alto=0}
|   |   |   |   |   |   interaccionesrecurso = IRM: alto {bajo=0, medio=1, alto=1}
|   |   |   |   |   edad = b: medio {bajo=1, medio=5, alto=0}
```

```
|   |   |   ciudad = O
|   |   |   |   estadocivil = C
|   |   |   |   |   interaccionesexamen = IEB: bajo {bajo=5, medio=0, alto=0}
|   |   |   |   |   interaccionesexamen = IEM: medio {bajo=0, medio=3, alto=0}
trabajo = Si
|   servicios = 0
|   |   genero = 1: medio {bajo=0, medio=2, alto=0}
|   |   genero = 2
|   |   |   ciudad = L
|   |   |   |   estadocivil = V
|   |   |   |   |   edad = c
|   |   |   |   |   |   interaccionesrecurso = IRB
|   |   |   |   |   |   |   interaccionesexamen = IEB
|   |   |   |   |   |   |   |   interaccionestareas = ITA: alto {bajo=0, medio=0, alto=2}
|   |   |   |   |   |   |   |   interaccionestareas = ITB: bajo {bajo=2, medio=0, alto=0}
|   servicios = 3
|   |   genero = 1
|   |   |   estadocivil = V
|   |   |   |   interaccionesrecurso = IRA: medio {bajo=0, medio=3, alto=0}
|   |   |   |   interaccionesrecurso = IRB: bajo {bajo=2, medio=0, alto=0}
|   |   genero = 2
|   |   |   ciudad = L
|   |   |   |   edad = a: medio {bajo=1, medio=1, alto=0}
|   |   |   |   edad = c: bajo {bajo=2, medio=1, alto=0}
|   |   |   ciudad = O: medio {bajo=0, medio=5, alto=0}
```

Figura 73: Reglas generadas por el algoritmo CHAID

Las reglas generadas por el algoritmo CHAID (ver figura 73) se describen a continuación:

- El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y no posee ningún servicio, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y posee un servicio y es menor a 25 años y las interacciones con los recursos y las tareas es baja, entonces las interacciones en el curso virtual es bajo.
- El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y posee todos los servicios y es casado y tiene entre 24 y 29 años, entonces las interacciones en el curso virtual es bajo.
- El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y posee todos los servicios y es soltero y es menor a 25 años y las interacciones con

los recursos y las tareas es baja y no tiene hijos es baja, entonces las interacciones en el curso virtual es bajo.

- El estudiante no trabaja y el género es masculino y pertenece a otra ciudad y es soltero, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es femenino y pertenece a la ciudad de Loja y no posee ningún servicio y es soltero y es menor a 25 años y las interacciones con las tareas y los recursos es bajo, entonces las interacciones en el curso virtual es bajo.
- El estudiante no trabaja y el género es femenino y no posee un servicio de telefono y pertenece a la ciudad de Loja y las interacciones con el examen es bajo, entonces las interacciones en el curso virtual es bajo.
- El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a la ciudad de Loja y es casado, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a la ciudad de Loja y es soltero y es menor a 25 años y las interacciones con los recursos es medio, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a otra ciudad y es casado y las interacciones con los exámenes es medio, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a otra ciudad y es casado y las interacciones con los exámenes es bajo, entonces las interacciones en el curso virtual es bajo.
- El estudiante trabaja y no posee servicios y el género es femenino, entonces las interacciones en el curso virtual es medio.
- El estudiante trabaja y no posee servicios y el género es masculino y pertenece a la ciudad de Loja y es mayor a 29 años y las interacciones con los exámenes, con los recursos y con las tareas es bajo, entonces las interacciones en el curso virtual es bajo.

- El estudiante trabaja y posee todos los servicios y el género es femenino y pertenece a la ciudad de Loja y es mayor a 29 años, entonces las interacciones en el curso virtual es bajo.

2.6.3.2.2. Algoritmo Decision Tree

Mediante el árbol de clasificación se pudo observar de manera gráfica y estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta es el número mínimo de divisiones que se puede dar por cada nodo (minimal size for Split=25), el tamaño mínimo de cada hoja (minimal leaf size=25), la ganancia mínima de un nodo (minimal gain=0.01), profundidad máxima o tamaño del árbol de decisión (maximal depth= 3), (confidence=0.1),criterio de selección de atributos (criterion= information gain) y las instancias clasificadas correctamente (accuracy) que maximiza la precisión de todo el árbol.

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 74 y 76).

❖ Proceso de entrenamiento

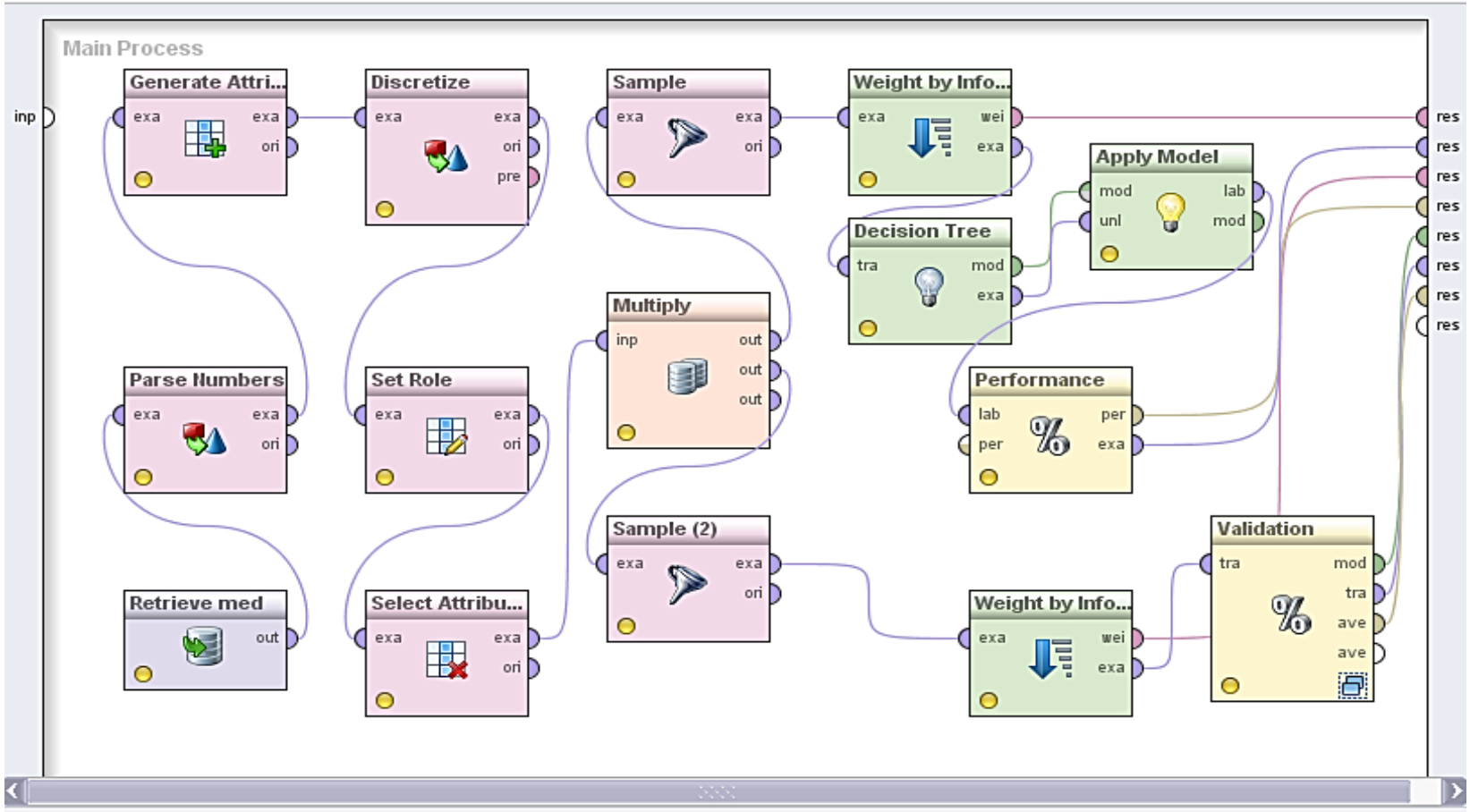


Figura 74: Proceso de Entrenamiento algoritmo Decision Tree

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LVI):

TABLA LVI.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO DECISION TREE

Decision Tree Entrenamiento	
Criterios	**Valores**
Accuracy	87.71%
Classification_error	12.29 %
Kappa	0.728
Absolute_error	0.214
Relative_error	21.44
Root_mean_squared_error	0.333
Root_relative_squared_error	1.179
Squered_error	0.111

- **Matriz de confusión del Entrenamiento algoritmo Decision Tree**

En la matriz de confusión del proceso de entrenamiento (ver Figura 75) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 87.71%

	true bajo	true medio	true alto
pred. bajo	190	38	5
pred. medio	31	438	14
pred. alto	0	0	0

Figura 75: Matriz de confusión del Entrenamiento algoritmo Decision Tree

❖ Proceso de Validación

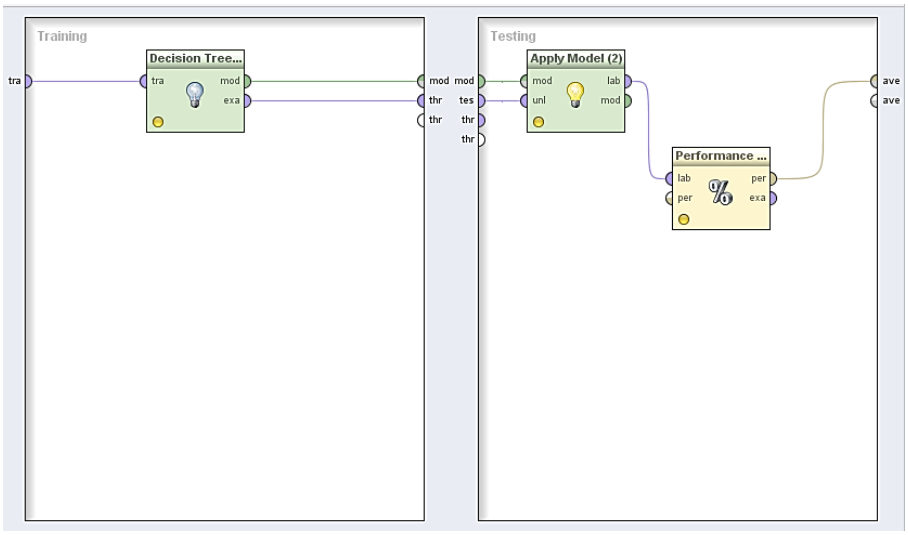


Figura 76: Proceso de validación algoritmo Decision Tree

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LVII):

TABLA LVII.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO DECISION TREE

Decision Tree Validación	
Criterios	**Valores**
Accuracy	92.90%
Classification_error	7.10%
Kappa	0.823
Absolute_error	0.133
Relative_error	13.33
Root_mean_squared_error	0.253
Root_relative_squared_error	4.285
Squered_error	0.068

❖ Matriz de confusión de la validación del algoritmo Decision Tree

En la matriz de confusión del proceso de validación ver Figura 77) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 92.90% +/- 4.24% (mikro: 92.90%)

	true bajo	true medio	true alto
pred. bajo	86	7	0
pred. medio	12	239	5
pred. alto	0	1	2

Figura 77: Matriz de confusión de la validación del algoritmo Decision Tree

❖ Reglas del Algoritmo Decision Tree

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo, se puede observar en la siguiente figura (ver Figura 78).

Tree

```
interaccionesexamen = IEA: medio {bajo=0, medio=51, alto=5}
interaccionesexamen = IEB
|   interaccionestareas = ITA: alto {bajo=0, medio=1, alto=2}
|   interaccionestareas = ITB
|   |   interaccionesrecurso = IRA: medio {bajo=0, medio=4, alto=0}
|   |   interaccionesrecurso = IRB: bajo {bajo=83, medio=6, alto=0}
|   |   interaccionesrecurso = IRM
|   |   |   estadocivil = C: bajo {bajo=2, medio=0, alto=0}
|   |   |   estadocivil = V: medio {bajo=1, medio=7, alto=0}
|   interaccionestareas = ITM: medio {bajo=0, medio=4, alto=0}
interaccionesexamen = IEM: medio {bajo=12, medio=174, alto=0}
```

Figura 78: Reglas generadas por el algoritmo Decision Tree

Las reglas generadas por el algoritmo Decision Tree (ver figura 78) se describen a continuación:

- Las interacciones con los exámenes es alto, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas es alto, entonces las interacciones en el curso virtual es alto.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas y los recursos es bajo, entonces las interacciones en el curso virtual es bajo.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas y los recursos es bajo y es casado, entonces las interacciones en el curso virtual es bajo.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas es bajo y las interacciones con los recursos es medio y es viudo, entonces las interacciones en el curso virtual es medio.

2.6.3.2.3. Algoritmo ID3

Mediante la utilización del algoritmo se genera el árbol de clasificación que se puede observar de manera gráfica y estructurada la existencia de criterios de interés o las interacciones de los estudiantes en el entorno virtual de aprendizaje, los parámetros que se han tomado en cuenta es el número mínimo de divisiones que se puede dar por cada nodo (minimal size for Split=4), el tamaño mínimo de cada hoja (minimal leaf size=2), la ganancia mínima (minimal gain=0.1), criterio de evaluación (criterion= gain ratio) y las instancias clasificadas correctamente (accuracy) que maximiza la precisión de todo el árbol.

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 79 y 81).

❖ Proceso de entrenamiento

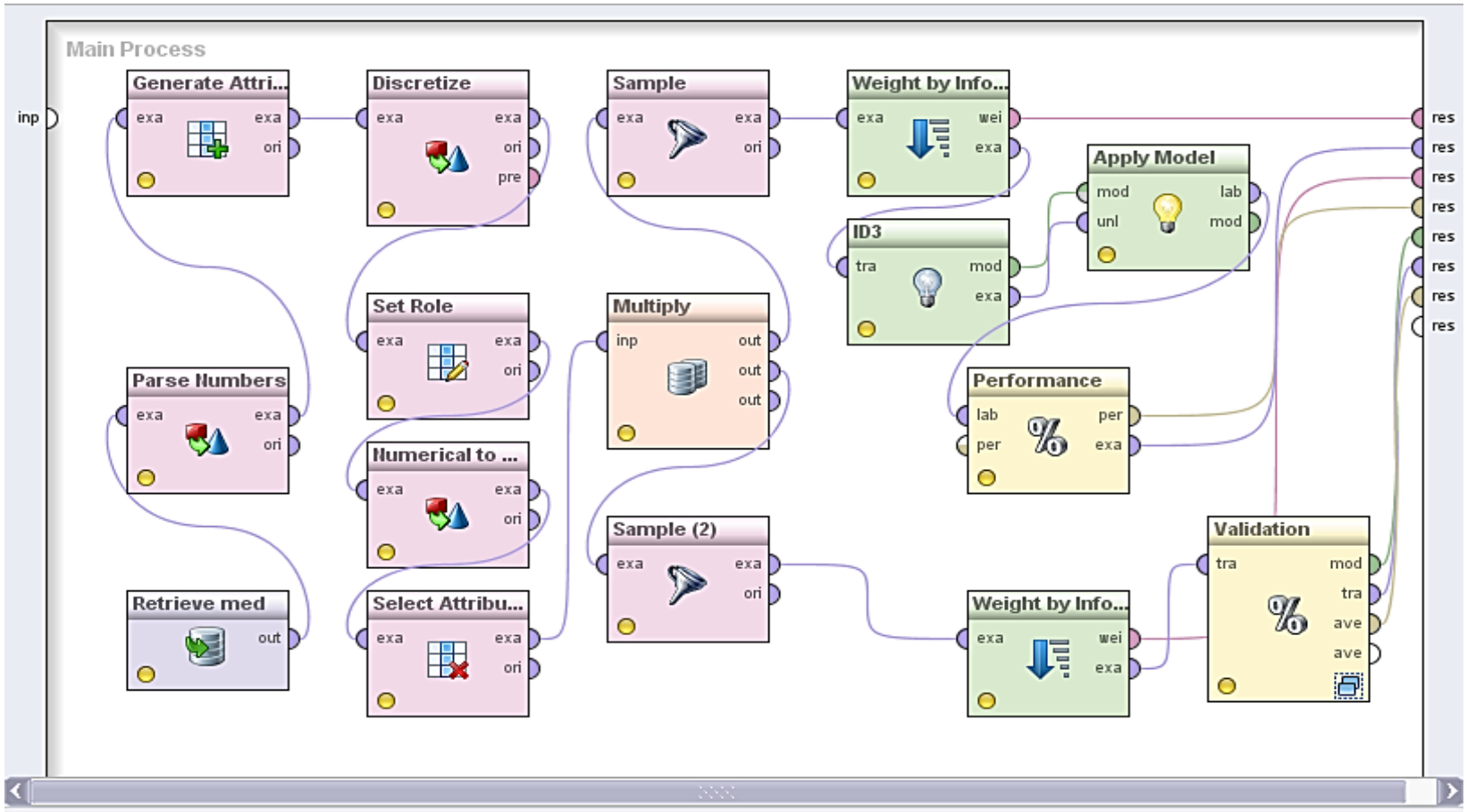


Figura 79: Proceso de Entrenamiento algoritmo ID3

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LVIII):

TABLA LVIII.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO ID3

ID3 Entrenamiento	
Criterios	**Valores**
Accuracy	98.32 %
Classification_error	1.68 %
Kappa	0.962
Absolute_error	0.020
Relative_error	1.97
Root_mean_squared_error	0.099
Root_relative_squared_error	0.389
Squered_error	0.010

- **Matriz de confusión del Entrenamiento algoritmo ID3**

En la matriz de confusión del proceso de entrenamiento (ver Figura 80) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 98.32%

	true bajo	true medio	true alto
pred. bajo	199	9	0
pred. medio	3	486	0
pred. alto	0	0	19

Figura 80: Matriz de confusión del Entrenamiento algoritmo ID3

❖ Proceso de Validación

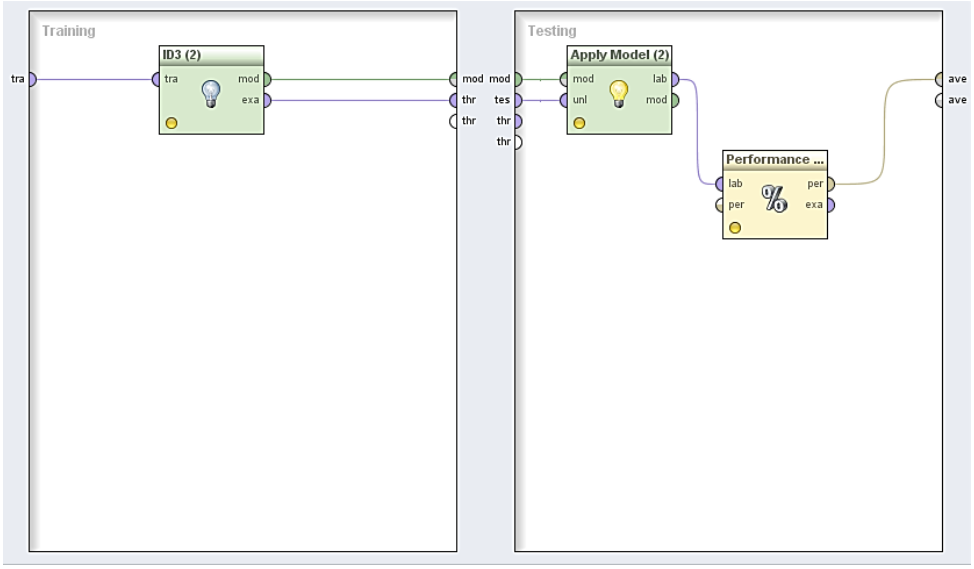


Figura 81: Proceso de validación algoritmo ID3

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LIX):

TABLA LIX.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO ID3

ID3 Validación	
Criterios	**Valores**
Accuracy	82.10%
Classification_error	17.90 %
Kappa	0.587
Absolute_error	0.130
Relative_error	13.03
Root_mean_squared_error	0.356
Root_relative_squared_error	0.345
Squered_error	0.130

❖ Matriz de confusión de la validación del algoritmo ID3

En la matriz de confusión del proceso de validación (ver Figura 82) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 82.10% +/- 5.11% (mikro: 82.10%)

	true bajo	true medio	true alto
pred. bajo	70	38	0
pred. medio	18	215	3
pred. alto	0	4	4

Figura 82: Matriz de confusión de la validación del algoritmo ID3

❖ Reglas del Algoritmo ID3

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo, se puede observar en la siguiente figura (ver Figura 83).

```
interaccionesexamen = IEA
|   carrera = AdministraciÃƒÂ³n de Empresas: medio {bajo=0, medio=21, alto=0}
|   carrera = ComunicaciÃƒÂ³n Social: medio {bajo=0, medio=1, alto=0}
|   carrera = Contabilidad y AuditorÃƒÂ-a
|   |   estadocivil = C: medio {bajo=0, medio=1, alto=0}
|   |   estadocivil = S: alto {bajo=0, medio=0, alto=1}
|   |   estadocivil = V: medio {bajo=0, medio=5, alto=0}
|   carrera = Derecho
|   |   interaccionesrecurso = IRA
|   |   |   numeroHijos = No
|   |   |   |   edad = b: alto {bajo=0, medio=0, alto=1}
|   |   |   |   edad = c: medio {bajo=0, medio=1, alto=0}
|   |   |   numeroHijos = Si: alto {bajo=0, medio=0, alto=3}
|   |   interaccionesrecurso = IRB: medio {bajo=0, medio=3, alto=0}
|   |   interaccionesrecurso = IRM: medio {bajo=0, medio=3, alto=0}
|   carrera = InformÃƒÂ¡tica Educativa: medio {bajo=0, medio=1, alto=0}
|   carrera = IngenierÃƒÂ-a en AdministraciÃƒÂ³n y ProducciÃƒÂ³n: medio {bajo=0, medio=2, alto=0}
|   carrera = IngenierÃƒÂ-a en Manejo y ConservaciÃƒÂ³n del Medi: medio {bajo=0, medio=2, alto=0}
|   carrera = IngenierÃƒÂ-a en Sistemas: medio {bajo=0, medio=1, alto=0}
|   carrera = Medicina Humana: medio {bajo=0, medio=3, alto=0}
|   carrera = PsicologÃƒÂ-a Infantil y EducaciÃƒÂ³n Parvularia: medio {bajo=0, medio=4, alto=0}
|   carrera = Psicorrehabilitacion y EducaciÃƒÂ³n Especial: medio {bajo=0, medio=1, alto=0}
|   carrera = Trabajo Social: medio {bajo=0, medio=2, alto=0}
```

```
|   |   servicios = 3: medio {bajo=0, medio=1, alto=0}
|   interaccionestareas = ITB
|   |   interaccionesrecurso = IRA: medio {bajo=0, medio=4, alto=0}
|   |   interaccionesrecurso = IRB
|   |   |   carrera = AdministraciÃƒÂ³n PÃƒÂºblica: bajo {bajo=1, medio=0, alto=0}
|   |   |   carrera = AdministraciÃƒÂ³n de Empresas
|   |   |   |   edad = a: bajo {bajo=3, medio=0, alto=0}
|   |   |   |   edad = b
|   |   |   |   |   ciudad = L: medio {bajo=0, medio=2, alto=0}
|   |   |   |   |   ciudad = O: bajo {bajo=2, medio=0, alto=0}
|   |   |   |   edad = c: bajo {bajo=2, medio=0, alto=0}
|   |   |   carrera = ComunicaciÃƒÂ³n Social: bajo {bajo=2, medio=0, alto=0}
|   |   |   carrera = Contabilidad y AuditorÃƒÂ-a
|   |   |   |   numeroHijos = No: bajo {bajo=5, medio=0, alto=0}
|   |   |   |   numeroHijos = Si
|   |   |   |   |   edad = a: medio {bajo=0, medio=1, alto=0}
|   |   |   |   |   edad = b
|   |   |   |   |   |   estadocivil = C: medio {bajo=0, medio=1, alto=0}
|   |   |   |   |   |   estadocivil = V: bajo {bajo=2, medio=0, alto=0}
|   |   |   |   |   edad = c: bajo {bajo=4, medio=0, alto=0}
|   |   |   carrera = Cultura FÃƒÂ-sica y Deportes: bajo {bajo=4, medio=0, alto=0}
|   |   |   carrera = Derecho
|   |   |   |   servicios = 0: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   servicios = 1: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   servicios = 2: medio {bajo=0, medio=1, alto=0}
|   |   |   |   servicios = 3
|   |   |   |   |   numeroHijos = No: bajo {bajo=8, medio=0, alto=0}
|   |   |   |   |   numeroHijos = Si
|   |   |   |   |   |   edad = a: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   |   |   edad = b: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   |   |   edad = c
|   |   |   |   |   |   |   trabajo = No
|   |   |   |   |   |   |   |   genero = 1: bajo {bajo=3, medio=0, alto=0}
|   |   |   |   |   |   |   |   genero = 2
|   |   |   |   |   |   |   |   |   ciudad = L
|   |   |   |   |   |   |   |   |   |   estadocivil = V: bajo {bajo=1, medio=1, alto=0}
|   |   |   |   |   |   |   trabajo = Si
|   |   |   |   |   |   |   |   genero = 1
|   |   |   |   |   |   |   |   |   ciudad = O
|   |   |   |   |   |   |   |   |   |   estadocivil = V: medio {bajo=1, medio=1, alto=0}
|   |   |   carrera = DiseÃƒÂto de Interiores y DecoraciÃƒÂ³n de Ambient: bajo {bajo=1, medio=0, alto=0}
|   |   |   carrera = EducaciÃƒÂ³n BÃƒÂ¡sica
|   |   |   |   genero = 1: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   genero = 2
|   |   |   |   |   edad = a: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   |   edad = b: medio {bajo=0, medio=1, alto=0}
```

```
|   |   |   carrera = IngenierÃƒÂ-a en Sistemas
|   |   |   |   genero = 1: bajo {bajo=2, medio=0, alto=0}
|   |   |   |   genero = 2: medio {bajo=0, medio=1, alto=0}
|   |   |   carrera = Medicina Humana
|   |   |   |   servicios = 1
|   |   |   |   |   estadocivil = S: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   |   estadocivil = V: medio {bajo=0, medio=1, alto=0}
|   |   |   |   servicios = 3: bajo {bajo=1, medio=0, alto=0}
|   |   |   carrera = PsicologÃƒÂ-a ClÃƒÂ-nica: bajo {bajo=2, medio=0, alto=0}
|   |   |   carrera = PsicologÃƒÂ-a Infantil y EducaciÃƒÂ³n Parvularia: bajo {bajo=13, medio=0, alto=0}
|   |   |   carrera = Psicorrehabilitacion y EducaciÃƒÂ³n Especial: medio {bajo=0, medio=1, alto=0}
|   |   |   carrera = QuÃƒÂ-mico BiolÃƒÂ³gicas
|   |   |   |   genero = 1: medio {bajo=0, medio=1, alto=0}
|   |   |   |   genero = 2: bajo {bajo=1, medio=0, alto=0}
|   |   |   carrera = Trabajo Social: bajo {bajo=1, medio=0, alto=0}
|   |   interaccionesrecurso = IRM
|   |   |   estadocivil = C: bajo {bajo=2, medio=0, alto=0}
|   |   |   estadocivil = V
|   |   |   |   carrera = AdministraciÃƒÂ³n de Empresas: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   carrera = Contabilidad y AuditorÃƒÂ-a: medio {bajo=0, medio=3, alto=0}
|   |   |   |   carrera = Derecho: medio {bajo=0, medio=1, alto=0}
|   |   |   |   carrera = Medicina Humana: medio {bajo=0, medio=1, alto=0}
|   |   |   |   carrera = PsicologÃƒÂ-a Infantil y EducaciÃƒÂ³n Parvularia: medio {bajo=0, medio=2, alto=0}
|   interaccionestareas = ITM: medio {bajo=0, medio=4, alto=0}
interaccionesexamen = IEM
|   interaccionestareas = ITA: medio {bajo=0, medio=5, alto=0}
|   interaccionestareas = ITB
|   |   carrera = AdministraciÃƒÂ³n PÃƒÂºblica: medio {bajo=0, medio=2, alto=0}
|   |   carrera = AdministraciÃƒÂ³n de Empresas
|   |   |   interaccionesrecurso = IRA: medio {bajo=0, medio=2, alto=0}
|   |   |   interaccionesrecurso = IRB
|   |   |   |   servicios = 1: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   servicios = 3
|   |   |   |   |   estadocivil = S: medio {bajo=0, medio=2, alto=0}
|   |   |   |   |   estadocivil = V
|   |   |   |   |   |   edad = b: bajo {bajo=1, medio=0, alto=0}
|   |   |   |   |   |   edad = c
|   |   |   |   |   |   |   ciudad = L: medio {bajo=0, medio=3, alto=0}
|   |   |   |   |   |   |   ciudad = O
|   |   |   |   |   |   |   |   numeroHijos = No: medio {bajo=0, medio=1, alto=0}
|   |   |   |   |   |   |   |   numeroHijos = Si
|   |   |   |   |   |   |   |   |   trabajo = No
|   |   |   |   |   |   |   |   |   |   genero = 1: bajo {bajo=1, medio=1, alto=0}
|   |   |   interaccionesrecurso = IRM: medio {bajo=0, medio=7, alto=0}
|   |   carrera = ComunicaciÃƒÂ³n Social: medio {bajo=0, medio=3, alto=0}
```

Figura 83: Reglas generadas por el algoritmo ID3

Las reglas generadas por el algoritmo ID3 (ver figura 83) se describen a continuación:

- Las interacciones con los exámenes es alto y es soltero, entonces las interacciones en el curso virtual es alto.
- Las interacciones con los exámenes y con los recursos es alto y pertenece a la carrera de Derecho y no tiene hijos y tiene entre 25 y 29 años, entonces las interacciones en el curso virtual es alto.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas es alto y no posee ningún tipo de servicio, entonces las interacciones en el curso virtual es alto.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y es menor de 25 años, entonces las interacciones en el curso virtual es bajo.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y tiene entre 25 y 29 años y pertenece a la ciudad de Loja, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es masculino y pertenece a otra ciudad y es viudo, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y pertenece a la carrera de Derecho y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es femenino y pertenece a la ciudad de Loja y es viudo, entonces las interacciones en el curso virtual es bajo.
- Las interacciones con los exámenes es medio y las interacciones con las tareas y recursos es baja y tiene los servicios y es soltero, entonces las interacciones en el curso virtual es medio.

2.6.3.2.4. Algoritmo J48

Mediante la utilización del presente algoritmo se establecieron valores a los parámetros como es el uso de árboles sin podar (T=false), umbral de confianza para la poda (C=0.25), número mínimo de instancias por hoja (M=2), error de poda (R= false), número de pliegues de la poda (N=3), divisiones binarias (B= false), limpieza del árbol (L= false), Semilla de datos aleatorios (Q=1).

Además se estableció el atributo objetivo que es el número de interacciones para la generación del modelo, las condiciones generadas por el árbol se basan en el objetivo antes mencionado.

Para la generación del modelo se llevó a cabo una serie de procesos en los que intervienen operadores los mismos que se pueden visualizar en las siguientes graficas (ver Figura 84 y 86).

❖ **Proceso de entrenamiento**

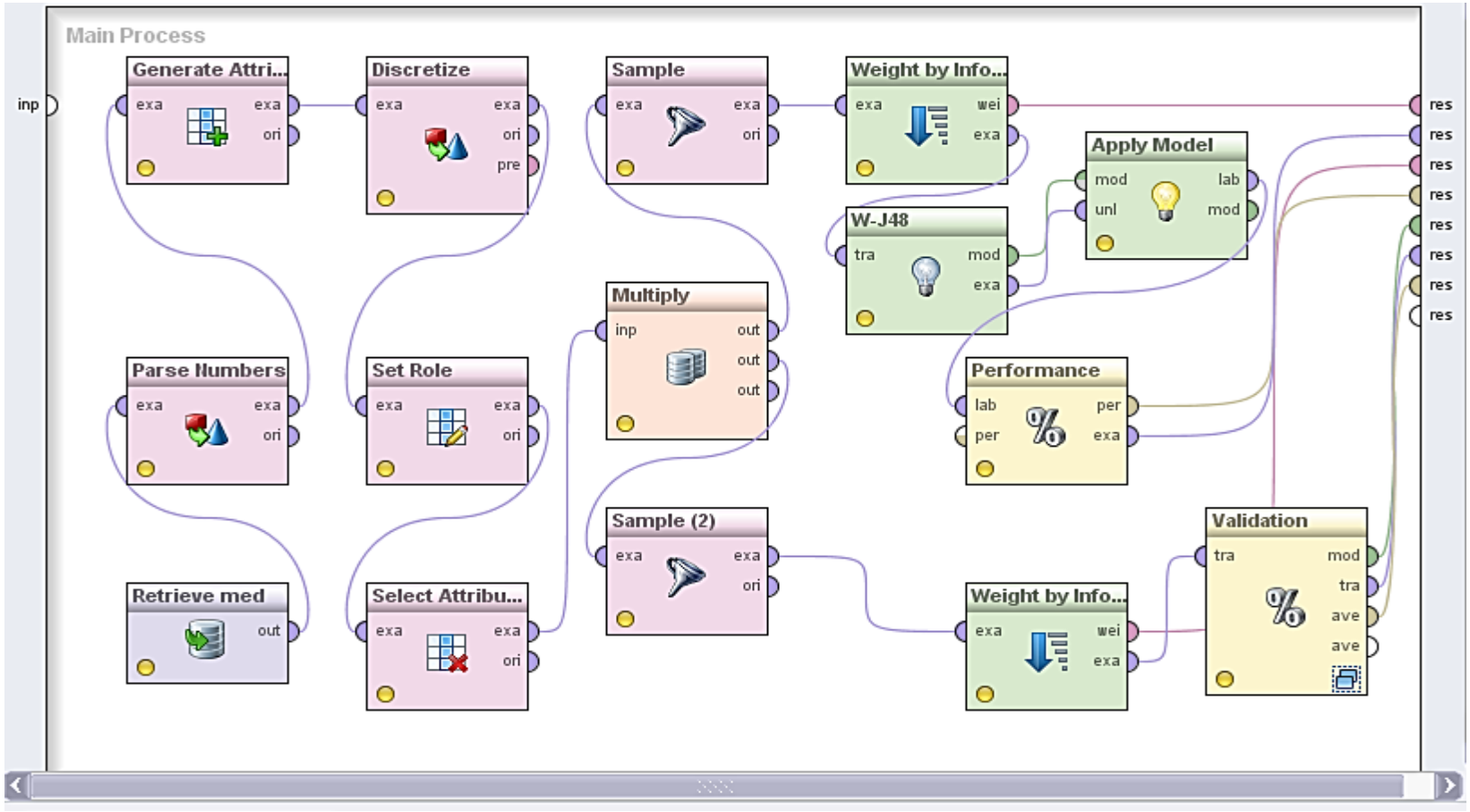


Figura 84: Proceso de Entrenamiento algoritmo J48

En el proceso de entrenamiento que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LX):

TABLA LX.

RESULTADOS OBTENIDOS EN EL PROCESO DE ENTRENAMIENTO DEL ALGORITMO J48

J48 Entrenamiento	
Criterios	**Valores**
Accuracy	91.06%
Classification_error	8.94 %
Kappa	0.797
Absolute_error	0.161
Relative_error	16.07
Root_mean_squared_error	0.285
Root_relative_squared_error	1.009
Squered_error	0.081

- **Matriz de confusión del Entrenamiento algoritmo J48**

En la matriz de confusión del proceso de entrenamiento (ver Figura 85) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 91.06%

	true bajo	true medio	true alto
pred. bajo	187	16	0
pred. medio	34	460	14
pred. alto	0	0	5

Figura 85: Matriz de confusión del Entrenamiento algoritmo J48

❖ Proceso de Validación

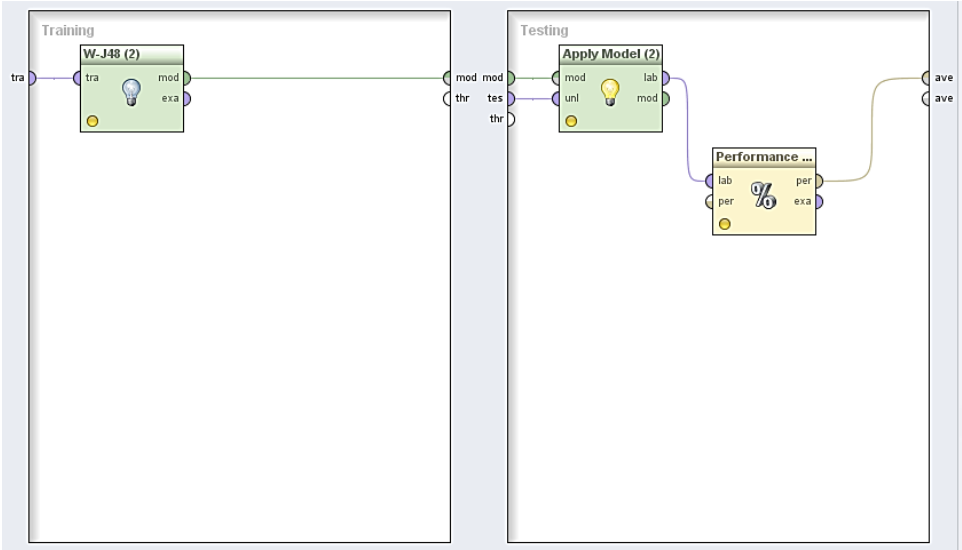


Figura 86: Proceso de validación algoritmo J48

En el proceso de validación que consta en cada algoritmo se ha fijado un parámetro muy importante que es el número de validaciones en este caso se le ha dado un valor de 10 es decir que internamente el algoritmo realiza dicho proceso para en si evaluar al algoritmo, según los parámetros del algoritmo se ha obtenido los siguientes resultados (ver TABLA LXI):

TABLA LXI.

RESULTADOS OBTENIDOS EN EL PROCESO DE VALIDACIÓN DEL ALGORITMO J48

J48 Validación	
Criterios	**Valores**
Accuracy	91.49%
Classification_error	8.51%
Kappa	0.153
Absolute_error	0.145
Relative_error	14.50
Root_mean_squared_error	0.269
Root_relative_squared_error	1.064
Squered_error	0.080

❖ Matriz de confusión de la validación del algoritmo J48

En la matriz de confusión del proceso de validación (ver Figura 87) se puede observar sobre la diagonal principal las instancias clasificadas correctamente y las instancias sobrantes son clasificadas incorrectamente.

accuracy: 91.49% +/- 5.76% (mikro: 91.48%)

	true bajo	true medio	true alto
pred. bajo	83	7	2
pred. medio	15	239	5
pred. alto	0	1	0

Figura 87: Matriz de confusión de la validación del algoritmo J48

❖ Reglas del Algoritmo J48

Mediante el empleo del algoritmo y conjuntamente con la base de datos que contiene información acerca de las interacciones de los estudiantes del curso de inglés de la Modalidad de Estudios a Distancia se han generado las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo, se puede observar en la siguiente figura (ver Figura 88).

W-J48

```
J48 pruned tree
------------------

interaccionesexamen = IEA: medio (56.0/5.0)
interaccionesexamen = IEM: medio (186.0/12.0)
interaccionesexamen = IEB
|   interaccionestareas = ITM: medio (4.0)
|   interaccionestareas = ITB
|   |   interaccionesrecurso = IRM
|   |   |   estadocivil = V: medio (8.0/1.0)
|   |   |   estadocivil = C: bajo (2.0)
|   |   |   estadocivil = S: medio (0.0)
|   |   |   estadocivil = D: medio (0.0)
|   |   interaccionesrecurso = IRB: bajo (89.0/6.0)
|   |   interaccionesrecurso = IRA: medio (4.0)
|   interaccionestareas = ITA: alto (3.0/1.0)

Number of Leaves  :     10

Size of the tree :      14
```

Figura 88: Reglas generadas por el algoritmo J48

Las reglas generadas por el algoritmo J48 (ver figura 88) se describen a continuación:

- Las interacciones con los exámenes es alto o medio, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas es medio, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes y las tareas es bajo y las interacciones con las recursos es medio y es viudo, soltero o divorciado, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes y las tareas es bajo y las interacciones con las recursos es medio y es casado, entonces las interacciones en el curso virtual es medio.

3. FASE III. Evaluar el modelo computacional en un escenario real a través de los datos de interacción de los estudiantes en un Entorno Virtual de Aprendizaje.

Para el cumplimiento del presente objetivo se utilizó los datos de los estudiantes para realizar el entrenamiento y la validación del modelo.

3.1. Evaluar el modelo computacional en un escenario real con datos de los estudiantes sobre la interacción con el curso virtual de inglés de la MED.

Mediante los datos de los estudiantes utilizados para evaluar el modelo se realizó las siguientes actividades:

3.1.1. Evaluar el modelo

En esta tarea se realizó la evaluación del modelo generado por cada uno de los algoritmos pertenecientes a la técnica de clasificación, comparando cada uno de los resultados obtenidos, por ende se consideraron diferentes parámetros que permitieron evaluar el modelo, para ello se utilizó los datos de entrenamiento en el porcentaje del 67% y los demás datos en pruebas con el 33%.

Entre los parámetros que se tomó en cuenta para evaluar los modelos generados son los siguientes: instancias clasificadas correctamente (accuracy), instancias clasificadas incorrectamente (clasification_error), estadística de Kappa que mide la coincidencia de la predicción con la clase real (Kappa), error cuadrático (squared_error), error relativo (relative_error), error absoluto (absolute_error), presentando los resultados obtenidos en la siguiente tabla (ver TABLA LXII):

TABLA LXII. EVALUACIÓN DE LOS MODELOS GENERADOS POR LOS ALGORITMOS

Algoritmo	Conjunto de datos	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de Kappa	Error Cuadrático	Error Relativo (%)	Error Cuadrático Medio	Error Cuadrático Relativo
DECISION TREE	Entrenamiento	87.71	12.29	0.728	0.111	21.44	0.333	1.179
	Validación	92.90	7.10	0.823	0.068	13.33	0.253	4.285
JRip	Entrenamiento	94.41	5.59	0.880	0.051	10.18	0.226	0.802
	Validación	92.63	7.37	0.820	0.071	13.17	0.247	0.980
RIDOR	Entrenamiento	89.66	10.34	0.770	0.103	10.34	0.321	1.140
	Validación	88.66	11.34	0.741	0.113	11.34	0.309	1.229
K-NN	Entrenamiento	98.74	1.26	0.973	0.013	1.26	0.112	0.397
	Validación	84.15	15.85	0.635	0.158	15.85	0.388	1.511
PRISM	Entrenamiento	98.46	1.54	0.971	0.015	1.54	0.124	0.462
	Validación	77.06	22.94	0.576	0.261	26.06	0.504	2.374
CHAID	Entrenamiento	91.06	8.94	0.806	0.063	12.41	0.252	0.893
	Validación	73.58	26.42	0.407	0.174	27.35	0.410	5.456
ID3	Entrenamiento	98.32	1.68	0.962	0.010	1.97	0.099	0.389
	Validación	82.10	17.90	0.587	0.130	13.03	0.356	0.345
J48	Entrenamiento	91.06	8.94	0.797	0.081	16.07	0.285	1.009
	Validación	91.49	8.51	0.153	0.080	14.50	0.269	1.064

En la tabla anterior (ver TABLA LXII) se puede observar el resultado de cada algoritmo obtenido mediante la utilización de la herramienta RapidMiner conjuntamente con los datos de los estudiantes del curso virtual inglés de la Modalidad de Estudios a Distancia, donde existe un porcentaje mínimo de error de clasificación en cada uno de los algoritmos, además se puede indicar que con el conjunto de entrenamiento de los datos la mayoría de los resultados obtenidos de los algoritmos son favorables es decir que sobrepasan el 90% de los datos han sido clasificados correctamente, los algoritmos que presentan mejores resultados se tiene el JRip 94.41%, K-NN 98.74%, Prism 98.46%, Chaid 91.06%, ID3 98.32 y el J48 91.06% . Así mismo con el conjunto de datos de validación los algoritmos que presentan los mejores resultados de datos clasificados correctamente se tiene el Decision Tree 92.90%, Jrip 92.63% y J48 91.49%.

Con el conjunto de datos utilizados en el entrenamiento el mejor algoritmo que presenta es el K-NN con un porcentaje de 98.74% es decir las instancias se han clasificado correctamente y un 1.26% de las instancias clasificadas incorrectamente, el índice de Kappa 0.973, el error cuadrático 0.013, error relativo 1.26, error cuadrático medio 0.11 y finalmente el error cuadrático relativo con el 0.397; y con el conjunto de datos utilizada en la validación el algoritmo que ha arrojado el mejor resultado es el Decision Tree con el 92.90% que significa que dicho porcentaje es el de las instancias clasificadas correctamente, el 7.10% es el de las instancias clasificadas incorrectamente, el índice de Kappa 0,823, el error cuadrático 0.068, error relativo 13.33, error cuadrático medio 0.253 y finalmente el error cuadrático relativo con el 4.285.

En la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos ya sea tanto en el entrenamiento como en la validación en donde se muestra las instancias clasificadas correctamente (ver Figura 89).

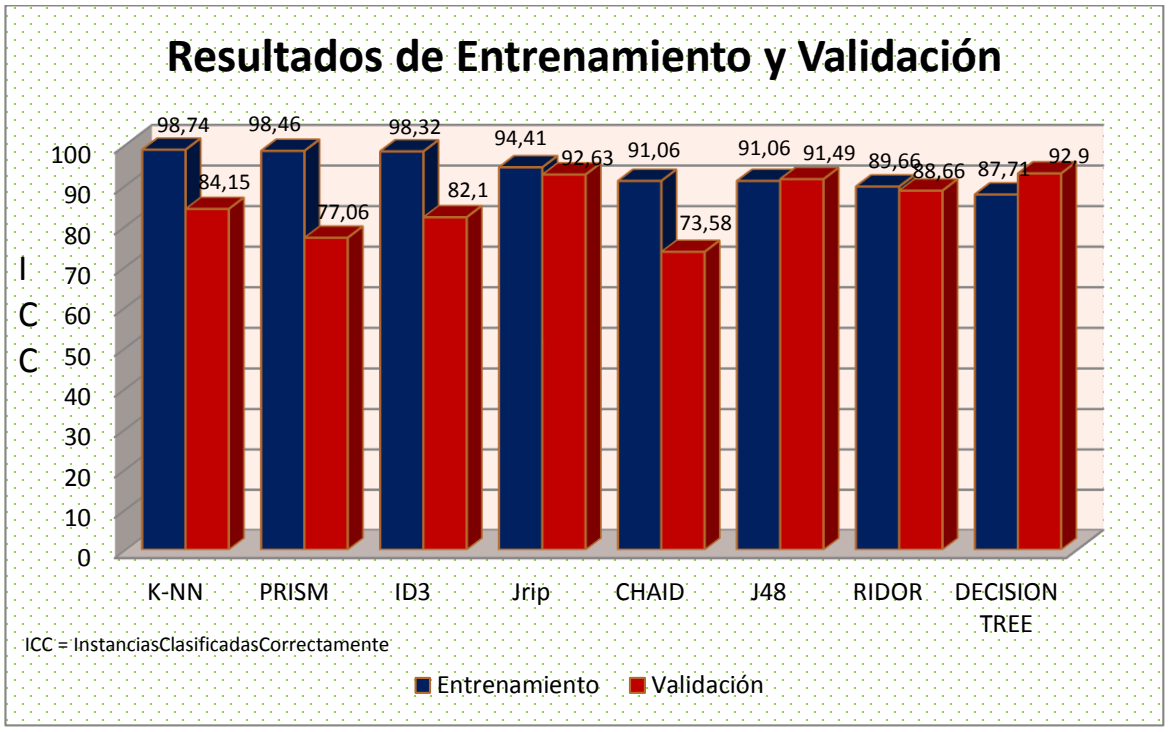


Figura 89: Resultados por cada algoritmo

Así mismo en la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos en lo que respecta en la evaluación en donde se muestra las instancias clasificadas correctamente y las instancias clasificadas incorrectamente, donde se puede mencionar que el algoritmo que tiene mayor porcentaje es el Decision Tree con el 92,90% con un margen de error del 7,10%, el mismo se muestra a continuación (ver Figura 90).

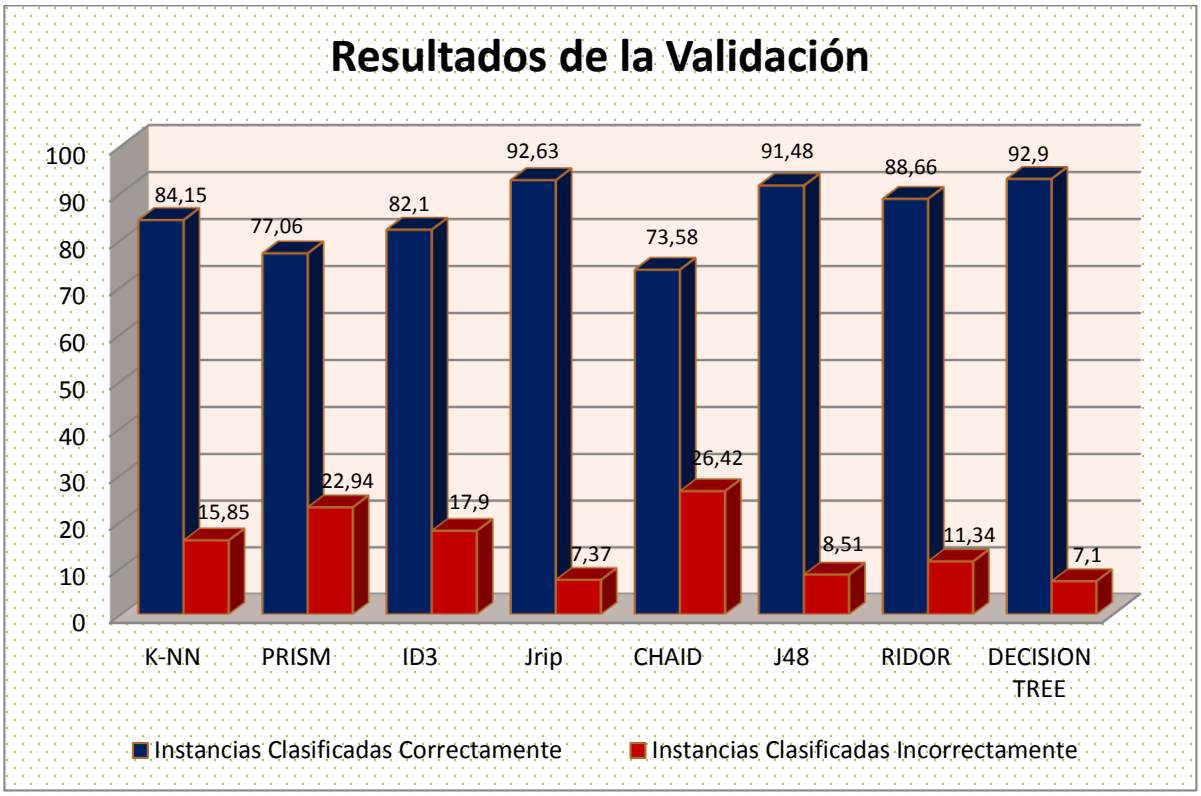


Figura 90: Resultados de algoritmos de instancias clasificadas correcta e incorrectamente

3.2. Interpretar los resultados arrojados por la herramienta de Minería de Datos acerca del modelo computacional.

Una vez obtenido los resultados mediante la herramienta de minería de datos RapidMiner con la técnica de clasificación y con el empleo de algunos algoritmos se procedió a realizar el análisis de cada resultado para ello se desarrolló la siguiente etapa.

3.2.1. ETAPA V: Evaluación

En la presente etapa se realizó la evaluación del modelo para determinar si los resultados obtenidos cumplen con los objetivos planteados inicialmente, de tal forma que mediante la generación del modelo a través de los atributos (ver TABLA XXXII) se puede determinar el nivel de interacción de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia.

3.2.1.1. Evaluar los resultados

Se realizó un análisis de los resultados obtenidos mediante la Minería de Datos, con una evaluación de los modelos obtenidos a través de la implementación de los algoritmos de la técnica de clasificación, los cuales fueron analizados en la fase anterior, para ello se utilizó datos de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia como son el número de accesos al curso, número de accesos a las tareas, número de veces que accede a un recurso, numero de accesos a exámenes, datos personales, socioeconómicos e institucionales de los estudiantes.

3.2.1.1.1. Determinar las interacciones de los estudiantes del curso virtual de inglés mediante una técnica de minería de datos

Para poder determinar las interacciones de los estudiantes, se tomó en cuenta el mejor resultado de los algoritmos utilizados, que fueron analizados en la evaluación del modelo (TABLA LXII y Figura 90) obteniendo el mejor resultado el algoritmo Decision Tree el cual presenta una buena clasificación con un 92.9% y menor margen de error del 7.1% en la validación del modelo.

Para la obtención de los resultados de este algoritmo se estableció el atributo objetivo determinar el nivel de las interacciones de los estudiantes (numerointeracciones), mediante este parámetro se clasifica el tipo de interacción ya sea alto, medio o bajo, según el valor que el algoritmo asigna, este puede estar comprendido entre 0 a 1, dando la sumatoria de estas tres clases el valor de 1, y el resultado de la clasificación se da según el valor más alto de los tres, como se puede observar en la siguiente figura (ver Figura 91) .

numerointeracciones	confidence(bajo)	confidence(medio)	confidence(alto)	prediction(numerointeracciones)
bajo	0.815	0.163	0.021	bajo
medio	0.081	0.913	0.005	medio
medio	0.081	0.913	0.005	medio
medio	0.081	0.913	0.005	medio
bajo	0.815	0.163	0.021	bajo
bajo	0.815	0.163	0.021	bajo
medio	0.081	0.913	0.005	medio
bajo	0.815	0.163	0.021	bajo

Figura 91: Clasificación de las interacciones de los estudiantes

Mediante el algoritmo Decision Tree se pudo determinar el nivel de interacción de los estudiantes conjuntamente con la utilización de cada uno de los atributos seleccionados que conforman el data set final (ver TABLA XXXII) de tal manera se obtuvieron los siguientes resultados durante la fase de entrenamiento del algoritmo el nivel de interacción es 19 altas, 438 medias y 190 bajas, así mismo en la fase de validación 7 altas, 239 medias y 86 bajas, los resultados finales del nivel de interacciones es 26 estudiantes se clasificaron dentro de la interacción autónoma o alta, 677 estudiantes tienen el nivel de interacción consiente o medio y 276 estudiantes su nivel de interacción es conformista o bajo que se encuentran representados en la siguiente figura (ver Figura 92).

Según con los resultados obtenidos se pudo determinar que en gran mayoría los estudiantes tienen un nivel de interacción consciente o medio en el curso virtual de inglés con 677 estudiantes equivalente a un porcentaje del 69%.

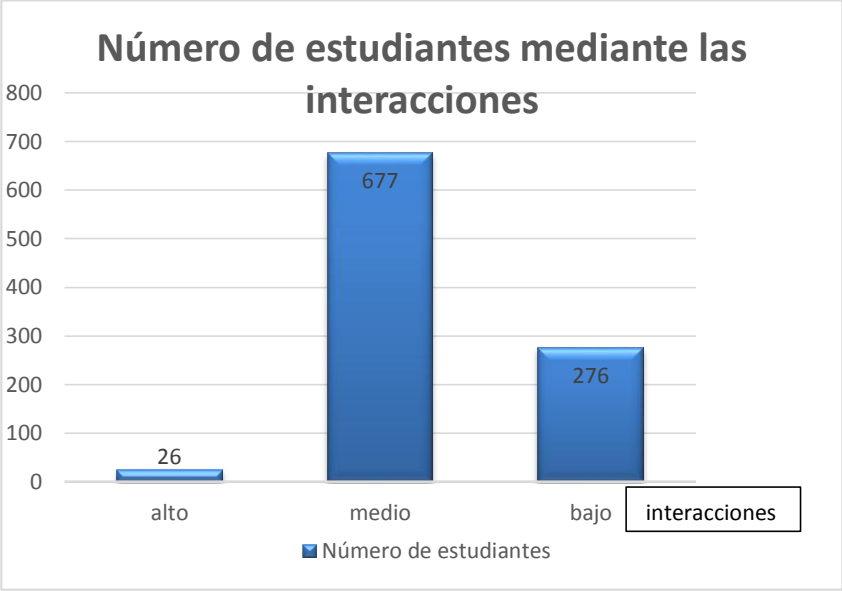


Figura 92: Interacciones de los estudiantes

Además de obtener los resultados mediante el algoritmo Decision Tree se pudo evidenciar que de los 1069 datos de los estudiantes obtenidos para generar el modelo, 90 estudiantes es decir el 9% de los datos no han sido clasificados en ningún tipo del nivel de interacción debido a que no cumplen con las reglas definidas por los algoritmos los mismos que se puede observar en la siguiente figura (ver Figura 93).

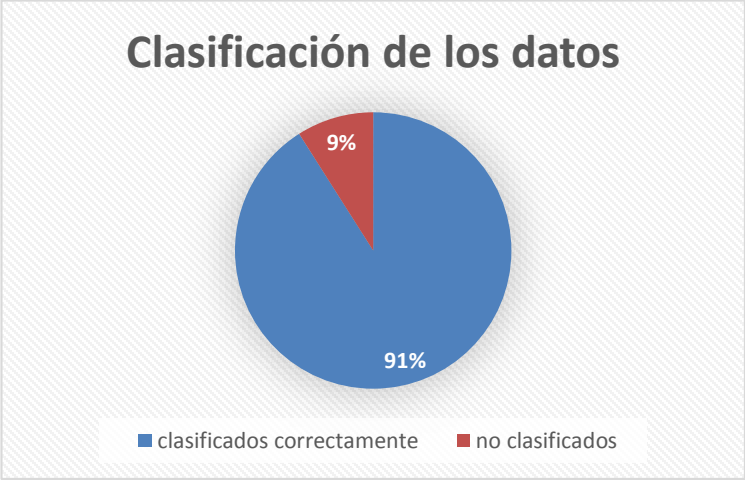


Figura 93: Clasificación de los datos

3.2.1.1.2. Reglas según el nivel de interacción

Nivel alto de interacción en el curso virtual

- Cuando los estudiantes tienen interacciones altas en los exámenes y los recursos y tiene entre 25 a 29 años y es soltero, entonces las interacciones en el curso virtual son altas.
- Si las interacciones con los recursos es medio y tiene entre 25 a 29 años y el género es femenino, entonces las interacciones en el curso virtual es alto.
- Si las interacciones en los exámenes es medio, es mayor a 29 años y no trabaja, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes y los recursos es medio y pertenece a otra ciudad y no tiene hijos, entonces las interacciones en el curso virtual es alto.
- Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y el género es femenino y no trabaja y pertenece a otra ciudad y es mayor a 29 años, entonces las interacciones en el curso virtual es alto.
- Las interacciones con los exámenes y con los recursos es alto y no posee ningún tipo de servicio y no tiene hijos y tiene entre 25 a 29 años, entonces las interacciones en el curso virtual es alto.

Nivel medio de interacción en el curso virtual

- Las interacciones en los exámenes es media y pertenece a otra ciudad y no tiene hijos y es casado, entonces las interacciones en el curso virtual es media.
- Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y tiene todos los servicios, no trabaja, el género es masculino, pertenece a otra ciudad y es mayor a 25 años, entonces las interacciones en el curso virtual es medio.
- Si las interacciones con los recursos es alto y las interacciones con las tareas es medio y el género es masculino y posee todos los servicios y es mayor a 29 años, entonces las interacciones en el curso virtual es medio.

- El estudiante no trabaja y el género es masculino y pertenece a otra ciudad y es soltero y solo posee un servicio que es número de celular, entonces las interacciones en el curso virtual es medio.
- El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a la ciudad de Loja y es soltero y es menor a 25 años y las interacciones con los recursos es medio, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes es bajo y las interacciones con las tareas es bajo y las interacciones con los recursos es medio y es casado, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es masculino y pertenece a otra ciudad y es casado, entonces las interacciones en el curso virtual es medio.
- Las interacciones con los exámenes es medio y las interacciones con las tareas y recursos es baja y tiene todos los servicios y es soltero, entonces las interacciones en el curso virtual es medio.

Nivel bajo de interacción en el curso virtual

- Si las interacciones con los exámenes y los recursos o tareas son bajas y trabaja, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes y los recursos son bajas y el género es femenino y tiene hijos, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes y los recursos son bajas y tiene hijos y es menor a 25 años, entonces las interacciones en el curso virtual es bajo.
- Si las interacciones con los exámenes, las tareas y los recursos son bajas y pertenece a la ciudad de Loja y no trabaja y posee todos los servicios y el género es masculino y es mayor a 29 años y tiene hijos, entonces las interacciones en el curso virtual es bajo.
- El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y posee un servicio y es menor a 25 años y las interacciones con los recursos y las tareas es baja, entonces las interacciones en el curso virtual es bajo.

- El estudiante trabaja y el género es masculino y pertenece a la ciudad de Loja y posee todos los servicios y es soltero y es menor a 25 años y las interacciones con los recursos y las tareas es baja y tiene hijos, entonces las interacciones en el curso virtual es bajo.
- El estudiante trabaja y el género es femenino y posee todos los servicios y pertenece a otra ciudad y es casado y las interacciones con los exámenes es bajo, entonces las interacciones en el curso virtual es bajo.
- Las interacciones con los exámenes y con las tareas y con los recursos es bajo y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es femenino y pertenece a la ciudad de Loja y es viudo, entonces las interacciones en el curso virtual es bajo.

3.2.1.1.3. Factores para determinar las interacciones de los estudiantes

Los factores que influyeron para determinar las interacciones de los estudiantes por lo tanto en la realización del modelo se encuentran asociados entre sí, los cuales son datos: personales, institucionales, socioeconómicos e interacciones del estudiante los cuales se detallan a continuación:

- **Interacciones en el curso:** interaccionestareas (número de accesos a las tareas), interaccionesrecurso (número de veces que accede a un recurso), interaccionesexamenes (número de accesos a exámenes).
- **Datos personales de los estudiantes:** género, estadocivil, edad, servicios (teléfono, celular), ciudad (estudiantes que residen en Loja o en otra ciudad del país).
- **Datos socioeconómicos de los estudiantes:** numeroHijos, trabajo (si el estudiante trabaja o no).
- **Datos institucionales de los estudiantes:** carrera (a que carrera pertenece el estudiante).

A continuación se presenta cada uno de los atributos con sus respectivos pesos según los resultados obtenidos mediante el algoritmo Decision Tree de tal forma que se pueda determinar el que más influye en el modelo (ver TABLA LXIII).

TABLA LXIII.

PESO DE ATRIBUTOS

Atributos con sus respectivos pesos	
Atributo	**Porcentaje del atributo (%)**
interaccionestareas	12.196
interaccionesrecurso	10.946
interaccionesexamenes	13.299
genero	4.562
estadocivil	9.509
edad	8.671
servicios	8.299
carrera	9.137
numeroHijos	5.346
trabajo	8.126
ciudad	9.908

Luego de obtener el peso de cada uno de los atributos pertenecientes a los datos de los estudiantes, los que más inciden en el objetivo principal del presente trabajo de titulación que es determinar el nivel de interacción de los estudiantes del curso virtual inglés de la Modalidad de Estudios a Distancia son las interacciones en las tareas con un 12%, en los recursos con el 11% y en los exámenes el 13% como se puede observar en la siguiente figura (ver Figura 94).

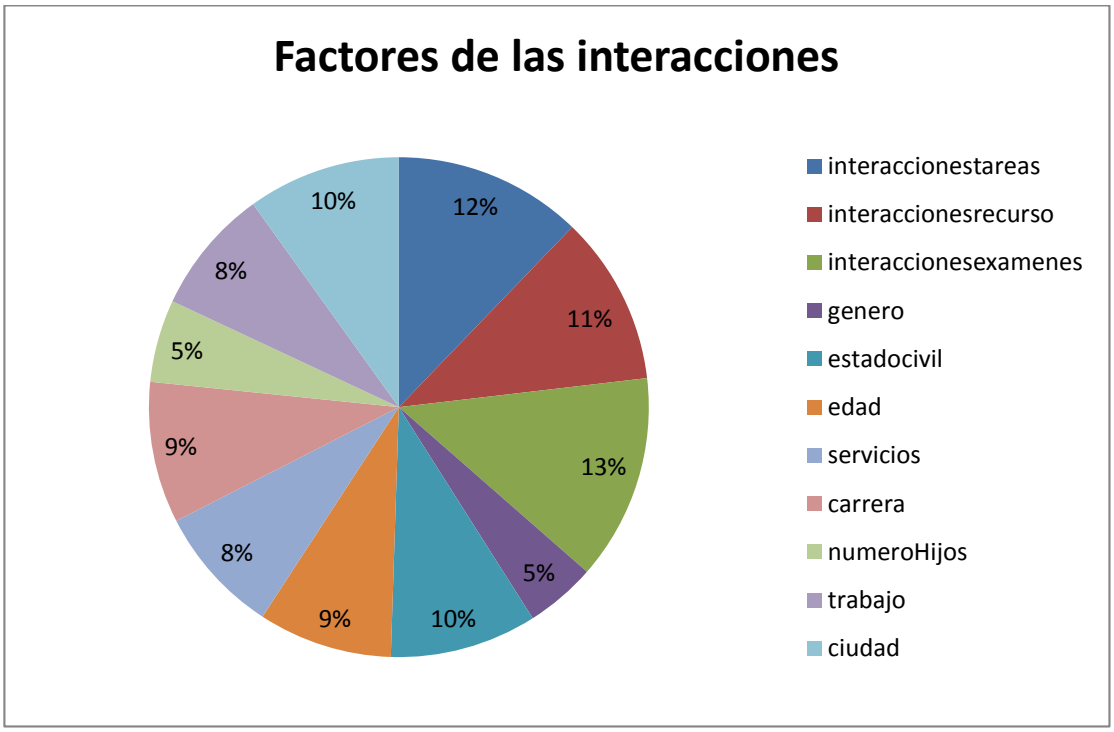


Figura 94: Factores de las interacciones de los estudiantes.

3.2.1.1.4. Análisis de los Resultados

La Universidad Nacional de Loja a través de la Modalidad de Estudios a Distancia brinda cursos virtuales de aprendizaje entre estos se tiene el curso de inglés, el mismo que fue tomado como objeto de estudio para el presente trabajo de titulación, ya que este idioma es fundamental porque se lo utiliza como un medio de comunicación, así mismo para laborar, además la mayoría de la información que se encuentra disponible ya sea en digital o impresa está en inglés, entre otras utilidades, es por ello que se realizó el análisis para determinar la interacción de los estudiantes en este curso virtual de aprendizaje del periodo académico 2013–2014, tomando dos conjuntos de datos los cuales estuvieron conformados por 147 alumnos obteniendo un total de 32029 interacciones que se encuentran distribuidas en tareas, recursos y exámenes, también se obtuvo datos personales como edad, estado civil, servicios, género, dirección, de cada uno de los estudiantes, los cuales conformaron el data set.

Con estos datos se realizó el análisis mediante la minería de datos donde surgieron algunos inconvenientes que los resultados obtenidos de los modelos fueron muy bajos es decir no cumplían con un rango de aceptación, motivo por el cual fue necesario la incrementación de nuevos datos, mismos que fueron proporcionados por la Unidad de Telecomunicaciones e Información, ya una vez que se contaba con 1069 datos correspondientes a 148472 interacciones se procedió a integrar nuevos atributos de estos estudiantes como son datos socioeconómicos (tienen hijos o no, situación laboral), datos institucionales (carrera a la que pertenecen), con estos parámetros se realizó nuevamente el análisis conjuntamente con la técnica de minería de datos y los algoritmos seleccionados obteniendo un mejoramiento en los resultados, porque el mayor porcentaje que arrojaron los algoritmo con los datos anteriores fue del 77%, mientras que con los nuevos datos y atributos incorporados los resultados mejoraron obteniendo los porcentajes mayores al 90%, por ende mediante el modelo obtenido los resultados son confiables.

Cabe indicar que existen tres tipos de interacción relacionadas al comportamiento o la forma en que el estudiante interactúa en la plataforma. La interacción conformista es cuando el nivel de interacción con los recursos de la plataforma es BAJO por lo que se debería incentivar al estudiante a utilizar con mayor frecuencia la plataforma y así incrementar su nivel de participación y obtener más conocimientos. La interacción consciente presenta mayor utilización de la plataforma, por lo que se concluye que el usuario tiene bien definidos cuáles son sus intereses en cuanto al material, está relacionada con un nivel de interacción MEDIO. La interacción autónoma se relaciona con el nivel de interacción ALTO, ya que los intereses del usuario están basados en la consulta de material e interactúan frecuentemente en el curso.

Consecuentemente se pudo determinar que 26 estudiantes pertenecen a la clasificación de la interacción alta, 677 estudiantes dentro de las interacciones medias y 276 estudiantes en interacciones bajas, estos resultados se obtuvieron a través de las reglas generadas por los algoritmos que conforman el modelo, presentando diferentes situaciones para cada nivel de interacción, las cuales se presentan a continuación:

- El nivel de interacción de los estudiantes es alto en el curso virtual de inglés, cuando las interacciones en los exámenes y los recursos son altas, su edad es mayor a 25

años, su estado civil es soltero, no trabaja, no tienen hijos, pertenecen a otra ciudad y su género es femenino.

- El nivel de interacción de los estudiantes es medio en el curso virtual de inglés, cuando las interacciones en los exámenes, tareas y los recursos son medias, pertenecen a cualquier ciudad, puede tener cualquier edad, su estado civil es soltero o casado, puede poseer un trabajo o no, puede poseer todos los servicios o uno de ellos, puede tener hijos o no, puede ser masculino o femenino.
- El nivel de interacción de los estudiantes es bajo en el curso virtual de inglés, cuando las interacciones en los exámenes, tareas y los recursos son bajas, pertenecen a la ciudad de Loja, es mayor a 29 años, su estado civil es casado, posee un trabajo, puede poseer todos los servicios o uno de ellos, tiene hijos, puede ser masculino o femenino.

Por lo tanto se puede indicar que las interacciones de los estudiantes que mayor prevalecen es el nivel medio con un porcentaje del 69% y el nivel bajo con el porcentaje del 25% en el curso virtual de inglés, significando que el estudiante utiliza los recursos conscientemente teniendo definido sus intereses en cuanto al material, es decir que no acceden con frecuencia al curso, por lo tanto realizan pocas consultas a las tareas, recursos y exámenes que deben cumplir y los factores que más influyen son las interacciones en el curso en general como es en las tareas, exámenes y recursos, así mismo los datos donde consta su situación laboral, si tiene hijos y el estado civil del estudiante.

Una situación preocupante se ve reflejado al momento de la obtención de los resultados pertenecientes al nivel de interacción alto donde una mínima parte de estudiantes para ser más exactos el 6% representando a solo 26 estudiantes interactúan de manera autónoma en el curso virtual de inglés es decir estos estudiantes demuestran una participación constante y un gran interés en el curso, basado en estos resultados se debería incentivar al estudiante a utilizar con mayor frecuencia la plataforma y así incrementar su nivel de participación en aquellos estudiantes que tiene un nivel de interacción bajo y medio, con el fin de que la mayoría de estudiantes tengan una mejora continua en referente a la participación en la plataforma y que los cursos sean tomados con responsabilidad de tal forma se pueda explotar todo el conocimiento brindado.

i. Discusión

1. Desarrollo de la propuesta alternativa

- **OBJETIVO ESPECÍFICO 1: Investigar sobre las diversas técnicas de minería de datos que permitan determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.**

  En el desarrollo del presente objetivo se realizó una búsqueda de información referente a las técnicas de minería de datos conjuntamente con los algoritmos que se puedan aplicar dependiendo de la clasificación de las mismas, así mismo se investigó algunos casos de éxito relacionados al trabajo de titulación donde aplicaban dichas técnicas, con la obtención de la información se realizó un análisis comparativo con la finalidad de seleccionar la técnica que más se adapte para el desarrollo del trabajo de titulación.

- **OBJETIVO ESPECÍFICO 2: Diseñar un modelo computacional aplicando técnicas de minería de datos para determinar la interacción de los estudiantes en un Entorno Virtual de Aprendizaje.**

  En la presente fase en primera instancia se realizó la recolección de los datos con los que se trabajó que pertenecen a las interacciones de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia, posteriormente se efectuó un análisis de los datos que se encuentran almacenados en la base de datos que fue proporcionada por la unidad de Telecomunicaciones e Información con la finalidad de conocer como están estructurados, relacionados y seleccionar los datos, de tal manera formar el data set final para seguidamente trabajar en lo que respecta la minería a de datos es decir que con la ayuda de la herramienta RapidMiner y la técnica de clasificación de minería de datos realizar el modelado y cumplir con el trabajo de titulación.

- **OBJETIVO ESPECÍFICO 3: Evaluar el modelo computacional en un escenario real a través de los datos de interacción de los estudiantes en un Entorno Virtual de Aprendizaje.**

Una vez obtenido el modelo se precedió a la evaluación del mismo generado por cada uno de los algoritmos pertenecientes a la técnica de clasificación, comparando cada uno de los resultados obtenidos, por ende se consideraron diferentes parámetros que permitieron evaluarlo, para ello se utilizó los datos de entrenamiento en el porcentaje de 67% y los demás datos para pruebas, es decir el 33%, el algoritmo que arrojó el mejor resultado fue el Decision Tree con el 92.90% que significa que dicho porcentaje es el de las instancias clasificadas correctamente, se pudo determinar el nivel de interacción de los estudiantes conjuntamente con la utilización de cada uno de los atributos seleccionados, se obtuvieron los resultados del algoritmo donde el nivel de interacción es 26 altas, 677 medias y 276 bajas, dando consigo la generación de algunas reglas en las que se pudo verificar los factores que más influyeron en este estudio y finalmente se realizó un análisis de cada una de las reglas con las cuales se pudo determinar que el nivel de interacción de los estudiantes en el curso virtual de inglés es medio y que los factores que más influyeron fueron las interacciones en las tareas, exámenes, recursos, situación laboral del estudiante, si tiene hijos y estado civil.

Con lo argumentado anteriormente se puede concluir que el presente trabajo de titulación obtuvo una culminación exitosa, debido al cumplimiento de cada uno de los objetivos planteados.

2. Valoración técnica económica ambiental

Mediante el desarrollo del trabajo de titulación denominado “Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje” se pudo analizar diferentes tipos de datos como son de tipo personal, institucional, socioeconómicos e interacciones de los estudiantes.
En el ámbito académico el tema tratado es de gran interés mediante el cual se pudo adquirir capacidades y habilidades para emplearlos a futuro, además con ello se podrá obtener experiencia en la resolución de problemas que están inmersos constantemente dentro de la sociedad.

Para la recolección de la información de los estudiantes se hizo uso de algunas técnicas de investigación también se utilizó la técnica de clasificación conjuntamente con la herramienta RapidMiner de minería de datos para analizar la información obtenida y extraer conocimiento de la misma para evaluar la interacción de los estudiantes y finalmente obtener el modelo computacional.

Así mismo mediante los recursos humanos y económicos se pudo llevar a cabo la efectiva realización del proyecto, así como el tiempo necesario que involucra el desarrollo del mismo, conjuntamente se contó con la disponibilidad y apoyo del tutor y del director de la Modalidad de Estudios a Distancia quien considera que es fundamental el estudio de estos datos, ya que les permitirá tomar decisiones que les ayude a mejorar el entorno virtual de Inglés.

En las siguientes tablas se indica detalladamente cada uno de los recursos empleados en el presente proyecto así como el talento humano (ver TABLA LXIV) se detalla los valores y las horas empleadas por parte de cada persona, recursos hardware (ver TABLA LXV), software (ver TABLA LXVI), servicios (ver TABLA LXVII) y recursos materiales (ver TABLA LXVIII).

TABLA LXIV.

TALENTO HUMANO

Talento Humano			
Recurso	**Horas**	**Costo/H ($)**	**Total ($)**
Autora del TT	400	5.00	2,000.00
Director del TT	100	10.00	1,000.00
Director MED	30	10.00	300.00
		SUBTOTAL	**3,300**.00

TABLA LXV.

HARDWARE

Hardware			
Hardware	**Horas**	**Costo ($)**	**Total ($)**
Ordenador	400	2.00	800.00
Disco Duro	100	1.00	100.00
Impresora	60	0.90	54.00
Memoria USB	100	0.15	15.00
		SUBTOTAL	**969.00**

TABLA LXVI.

SOFTWARE

Software			
Software	**Horas**	**Costo ($)**	**Total ($)**
Paquete office	200	0.00	0.00
MySQL	300	0.00	0.00
RapidMiner	350	0.00	0.00
		SUBTOTAL	**0.00**

TABLA LXVII.

SERVICIOS

Servicios			
Servicio	**Horas**	**Costo ($)**	**Total ($)**
Internet	300	1.00	300.00
		SUBTOTAL	300.00

TABLA LXVIII.

MATERIALES

Materiales			
Recurso	**Cantidad**	**Costo ($)**	**Total ($)**
Resma de papel	5	3.50	17.50
Cartuchos de Tinta	2	21,00.00	42.00
Perfiles	3	0.50	1.50
		SUBTOTAL	**61.00**

A continuación se presenta una tabla general de todos los recursos utilizados con sus respectivos valores que es el presupuesto final (Ver TABLA LXIX).

TABLA LXIX.

PRESUPUESTO TOTAL

PRESUPUESTO TOTAL ($)	
Recursos Humano	3,300.00
Recursos Hardware	969.00
Recursos Software	0.00
Recursos Servicios	300.00
Recursos Materiales	61.00
TOTAL	4,630.00
IMPREVISTOS (10% DEL TOTAL)	463.00
TOTAL PRESUPUESTO + IMPREVISTOS	**5,093.00**

Con lo anteriormente expuesto se afirma que el proyecto fue viable, debido a que desde el ámbito académico, tecnológico y económico se han cumplido según lo requerido, así se puede asegurar que se obtuvo los resultados esperados en cuanto al desarrollo del modelo para determinar las interacciones de los estudiantes del curso virtual de Inglés de la Modalidad de Estudios a Distancia.

j. Conclusiones

A través de los resultados obtenidos se ha podido concluir lo siguiente:

- El curso virtual de inglés tiene una gran acogida por parte de los estudiantes ya que es muy importante en su formación profesional, así mismo es fundamental porque la mayoría de la información que se encuentra disponible ya sea en digital o impresa está en este idioma, además se lo emplea como un medio de comunicación.

- La minería de datos es muy importante dentro del campo de la educación ya que ayudó a extraer información que se encuentra oculta en los datos de tal forma permitió el análisis y la generación de nuevo conocimiento para poder determinar en nivel de interacción de los estudiantes.

- RapidMiner es una herramienta de minería de datos potente ya que contiene complementos que permite hacer uso de diferentes algoritmos tanto de esta herramienta como de otras herramientas, además tiene operadores que ayudan a facilitar el desarrollo de los procesos para crear los modelos aplicables para el análisis de los datos.

- Para determinar el nivel de interacción en el entorno virtual de aprendizaje se aplicó diferentes algoritmos de clasificación, presentando los mejores resultados el Decision Tree, ya que obtuvo el menor margen de error durante la clasificación de los datos de las interacciones en el curso (tareas, exámenes, recursos), datos personales, institucionales y socioeconómicos.

- Mediante el modelo de minería de datos obtenido se pudo determinar que las interacciones de los estudiantes en el entorno virtual de aprendizaje que mayor prevalecen es el nivel medio con un porcentaje del 69% y los factores que más influyeron en el modelo fueron las interacciones en los exámenes, tareas, recursos, el estado civil y la situación laboral del estudiante.

k. Recomendaciones

Luego de concluir el trabajo de titulación se ha llegado a las siguientes recomendaciones:

- Para poder desarrollar un modelo de minería de datos de forma ordenada es importante hacer uso de una metodología siendo en la actualidad CRISP–DM una de las más utilizadas, ya que propone las fases necesarias para generar un modelo de calidad y poder cumplir con los objetivos de un proyecto.

- Cuando se realiza el análisis de la información requerida para llevar a cabo los procesos de minería de datos es necesario analizar las diferentes técnicas de minería de datos e implementar más de un algoritmo de la técnica seleccionada, para comparar los resultados y verificar cual es el algoritmo idóneo según las necesidades del proyecto.

- Trabajar con grandes cantidades de datos para los procesos que se desarrollan en la minería de datos con la finalidad de asegurar la calidad de los modelos obtenidos.

- Utilizar RapidMiner porque es una de las herramientas más completas para minería de datos, contiene varios operadores que permiten llevar procesos generando modelos, además cuenta con complementos que permiten la conexión con otras herramientas de tal forma que se puede utilizar los algoritmos que contienen estas.

- Después de obtener los niveles de interacción predominantes en los estudiantes es importante incluir estrategias de incentivo para que puedan utilizar todos los recursos establecidos dentro del entorno virtual de aprendizaje consiguiendo que mejore el cumplimiento de la planificación de tal forma se

pueda lograr que las interacciones dentro de la plataforma alcance el nivel ideal (alto).

- Implementar en el curso virtual de inglés actividades para la comunicación como foros y videoconferencias para incrementar el nivel de interacción de los estudiantes dentro del entorno virtual de aprendizaje e incentivar que el estudiante se interese en continuar en los cursos virtuales de la oferta académica de la Modalidad de Estudios a Distancia.

I. Bibliografía

[1]. Luz Marina Gómez y Julio Macelo Buleje, Importancia de las TIC en la Educación Básica Regular, Universidad Nacional Mayor de San Marcos – Facultad de Educación.

[2]. Minería de Datos, Universidad de Extremadura - Campus Libre y Abierto, En línea: http://cala.unex.es/cala/epistemowikia/index.php?title=Miner%C3%ADa_de_Datos

[3]. Karla Fernanda Ordoñez, Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL, Universidad Técnica Particular de Loja – Área Técnica, En línea: http://dspace.utpl.edu.ec/bitstream/123456789/7897/1/Ordonez%20Brice%C3%B1o%20Karla-%20Informatica.pdf

[4]. Moodle, ¿Qué es Moodle? ¿Para qué?, Universidad Luterana Salvadoreña - Unidad de Informática y Comunicaciones, En línea: http://www.uls.edu.sv/pdf/manuales_moodle/queesmoodle.pdf

[5]. Moodle, Acerca de Moodle, En línea: https://docs.moodle.org/all/es/Acerca_de_Moodle

[6]. Adrián Villegas Dianta, ¿Qué es Moodle?, En línea: http://www.e-historia.cl/e-historia-2/%C2%BFque-es-moodle/

[7]. Coordinación Tecnológica de la Modalidad de Estudios a Distancia, Manual de Usuario Plataforma Virtual, Universidad Nacional de Loja - Modalidad de Estudios a Distancia.

[8]. Enríquez Toledo Alma, Maldonado Ayala Jesús, Nakamura Ortega Yunko, Nogueron Toledo Goretty, MySQL, En línea: http://www.gridmorelos.uaem.mx/~mcruz//cursos/miic/MySQL.pdf

[9]. Luis Alberto Casillas Santillán, Marc Gibert Ginestà, Óscar Pérez Mora, Bases de datos en MySQL, Universidad Abierta de Cataluña, En línea: http://ocw.uoc.edu/computer-science-technology-and-multimedia/bases-de-datos/bases-de-datos/P06_M2109_02151.pdf

[10]. Alberto Méndez Barceló, Aramís Rivas Diéguez y Marlene del Toro Borrego, Entornos virtuales de enseñanza aprendizaje, En línea: http://bibliotecalibre.org/bitstream/001/251/8/978-959-16-0637-2.pdf

[11]. Consuelo Belloch, Entornos Virtuales de Aprendizaje, Universidad de Valencia - Unidad de Tecnología Educativa, En línea: http://www.uv.es/bellochc/pedagogia/EVA3.pdf

[12]. Universidad de Salamanca, Juan Silva y Begoña Gros, Una propuesta para el análisis de interacciones en un espacio virtual de aprendizaje para la formación continua de los docentes, En línea: http://campus.usal.es/~teoriaeducacion/rev_numero_08_01/n8_01_silva_gros.pdf

[13]. Estela Lizbeth Muñoz Andrade y Jaime Muñoz Arteaga, Interactividad en ambientes virtuales de aprendizaje: Características, , Universidad Autónoma de Aguascalientes, En línea: http://ingsw.ccbas.uaa.mx/sitio/images/publicaciones/7%20Interactividad%20en%20AA_141005.pdf

[14]. Patricia Ávila M. y Martha Diana Bosco H. Ambientes Virtuales de Aprendizaje una nueva experiencia, Instituto Ecuatoriano de Comunicación Educativa, En línea: http://investigacion.ilce.edu.mx/panel_control/doc/c37ambientes.pdf

[15]. Sarango Sedamanos Marcia Yudy, 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 – Escuela de Ciencias de la Computación.

[16]. Priscila M. Valdivieso, Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje.

[17]. Constanza Huapaya, Francisco Lizarralde, Graciela Arona, Jorge Vivas, Stella Massa, Gustavo Bacino, Carlos Rico, Felipe Evans, Uso de ambientes virtuales de aprendizaje en la enseñanza de la ingeniería, Universidad Nacional de Mar del Plata - Facultad de Ingeniería.

[18]. Minería de Datos, Universidad de Extremadura - Campus Libre y Abierto, En línea: http://cala.unex.es/cala/epistemowikia/index.php?title=Miner%C3%ADa_de_Datos

[19]. María del Carmen, Galán San José, Definición de Minería de Datos, Universidad Carlos III de Madrid, En línea: http://www.oocities.org/es/mineria.datos/definicion_tecnicas_mineria_datos.pdf

[20]. Instituto de Investigación en Inteligencia Artificial, MINERÍA DE DATOS O DATA MINING, Consejo Superior de Investigaciones Científicas - Instituto de Investigación en Inteligencia Artificial, En línea: http://www.iiia.csic.es/udt/files/DataMining.pdf

[21]. Ramón David Lezcano, Minería de Datos, Universidad Nacional del Nordeste – Departamento de Informática, En línea: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosLezcano.pdf

[22]. Javier Román Carrillo y Fernando Virseda Benito, Minería de datos y aplicaciones, Universidad Carlos III – Departamento de Ingeniería Telemática, En línea: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/22.pdf

[23]. Jiawei Han y Micheline Kamber, Data Mining, University of Illinois at Urbana-Champaign, En línea: http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf

[24]. Viaani Lily Álvarez Prados, Bases de Datos, Universidad Veracruzana, En línea: http://cdigital.uv.mx/bitstream/123456789/29380/1/Alvarez%20Prados.pdf

[25]. Juan Camilo Giraldo Mejía y Jovani Alberto Jiménez Builes, Caracterización del Proceso de Obtención de Conocimiento y Algunas Metodologías para Crear Proyectos de Minería de Datos, Universidad Nacional de Colombia, En línea: http://www.unla.edu.ar/sistemas/redisla/ReLAIS/relais-v1-n2-p-42-44.pdf

[26]. Carlos Cobos, John Zuñiga, Juan Guarín, Elizabeth León y Martha Mendoza, CMIN - herramienta case basada en CRISP-DM para el soporte de proyectos de minería de datos, En línea: http://www.scielo.org.co/pdf/iei/v30n3/v30n3a04

[27]. Jesús Antonio González Bernal, Minería de Datos, Universidad Politécnica de Puebla, En línea: http://ccc.inaoep.mx/~jagonzalez/AI/Sesion13_Data_Mining.pdf

[28]. Jason Frand's, Data Mining: What is Data Mining?, Universidad Centroccidental "Lisandro Alvarado", En línea: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[29]. Cristóbal Romero Morales, Sebastián Ventura Soto, Cesar Hervás Martínez, Escuela Politécnica Superior Universidad de Córdoba, Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web.

[30]. Miguel Cárdenas Montes, Clustering: Clasificación no Supervisada, Centro de Investigaciones Energéticas Medioambientales y Tecnológicas, En línea: http://wwwae.ciemat.es/~cardenas/curso_MD/clustering.pdf

[31]. Gutiérrez Rüegg, P., Merlino, H., Rancan, C., Procopio, C., Rodríguez, D., Britos, P., García-Martínez, R., Identificación de patrones característicos de la población carcelaria mediante minería de datos, Universidad Nacional de la Plata – Facultad de Ingeniería, En línea:

http://sedici.unlp.edu.ar/bitstream/handle/10915/20617/Documento_completo.pdf?sequence=1

[32]. José Perea, Análisis clúster, Universidad de Córdoba, En línea: http://www.uco.es/zootecniaygestion/img/pictorex/09_13_25_sesion_8.pdf

[33]. José Hernández Orallo, Técnicas de Minería de Datos, Universidad Politécnica de Valencia, En línea: http://users.dsic.upv.es/~jorallo/cursoDWDM/dwdm-III-3.pdf

[34]. Román García Ayesha Sagrario, Minería de Datos en encuestas de profesores al fin de semestre de la Facultad de Ingeniería, UNAM, Universidad Nacional Autónoma de México, México - Facultad de Ingeniería, En línea: http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/227/tesis.pdf?sequence=1

[35]. Ansel Yoan Rodríguez González, José Francisco Martínez Trinidad, Jesús Ariel Carrasco Ochoa, José Ruiz Shulcloper, Minería de Reglas de Asociación sobre Datos Mezclados, Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, En línea: http://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf

[36]. José Antonio García Bermúdez y Ángela María Acevedo Ramírez, Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies, Universidad Tecnológica de Pereira - Facultad de ingenierías: eléctrica, electrónica, física y Ciencias de la computación - Ingeniería de sistemas y computación, En línea: http://repositorio.utp.edu.co/dspace/bitstream/11059/1339/1/006312G216.pdf

[37]. Ronald Augusto Velandia y Fredy Leonardo Hernández, Evaluación de Algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión, Universidad Industrial de Santander, En línea: http://tangara.uis.edu.co/biblioweb/tesis/2010/134742.pdf

[38]. Julio Alberto Polo Fernández, Raúl Muñoz Martínez, Predicción Meteorológica, Universidad Carlos III de Madrid, En línea: http://www.it.uc3m.es/~jvillena/irc/practicas/05-06/11mem

[39]. Julio Alberto Polo Fernández, Raúl Muñoz Martínez, Madrid, Minería de datos , Universidad Carlos III de, En línea: http://www.it.uc3m.es/jvillena/irc/practicas/03-04/18.mem.pdf

[40]. M. Beltrán Pascual, Á. Muñoz Alamillos, A. Muñoz Martínez, Un nuevo clasificador de préstamos bancarios a través de la minería de datos, Universidad Nacional de Educación a Distancia, En línea: http://www.uned.es/dpto-economia-aplicada-y-estadistica/SEIO2012.pdf

[41]. Paula Andrea Vizcaino Garzón, Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de weka (waikato environment for knowledge analysis), Fundación Universitaria Konrad Lorenz - Facultad de Ingeniería de Sistemas, En línea: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf

[42]. María García Jiménez y Aránzazu Álvarez Sierra, Análisis de Datos en WEKA – Pruebas de Selectividad, Universidad Carlos III - Ingeniería de Telecomunicación, En línea: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

[43]. José Hernández Orallo y Cèsar Ferri Ramírez, Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software, Práctica de Minería de Datos, Universidad Politécnica de Valencia, En línea: http://users.dsic.upv.es/~cferri/weka/CursDoctorat-weka.pdf

[44]. Ricardo Aler, Tutorial Weka 3.6.0, Universidad Carlos III de Madrid, En línea: http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/TutorialWeka.pdf

[45]. Programa de Doctorado Tecnologías Industriales, Aplicaciones de la Inteligencia Artificial en Robótica, Universidad Michoacana - Facultad de Ingeniería Eléctrica, En línea: http://lsc.fie.umich.mx/~juan/Materias/Cursos/UDM/InformationSystems/Tareas/Tarea3Weka/P1.pdf

[46]. Técnicas de Análisis de Datos en WEKA, Universidad Miguel Hernández - Ingeniería de Sistemas y Automática, En línea: http://isa.umh.es/asignaturas/crss/turorialWEKA.pdf

[47]. Abdelmalik Moujahid, Inaki Inza y Pedro Larrañaga, Introducción a la Minería de Datos, Universidad del País Vasco - Departamento de Ciencias de la Computación e Inteligencia Artificial, En línea: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/mineria-datos0708.pdf

[48]. Juan Miguel Moine, Ana Silvia Haedo y Silvia Gordillo, Estudio comparativo de metodologías para minería de datos, Universidad Nacional de La Plata - Facultad de Informática, En línea: http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

[49]. Juan Miguel Moine, Ana Silvia Haedo y Silvia Gordillo, Análisis comparativo de metodologías para la gestión de proyectos de minería de datos, Universidad Nacional de La Plata - Facultad de Informática, En línea: http://sedici.unlp.edu.ar/bitstream/handle/10915/18749/Documento_completo.pdf?sequence=1

[50]. Juan Miguel Moine, Cristian Germán Bigatti, Guillermo Leale, Graciela Carnevali y Esther Francheli, Un modelo predictivo para reducir la tasa de ausentismo en atenciones médicas programadas, Universidad Técnica del Norte - Facultad Regional Rosario En línea: http://www.42jaiio.org.ar/proceedings/simposios/Trabajos/CAIS/20.pdf

[51]. Rendón Herrera Melina Alejandra y Acosta Vásquez Juan David, Estudio sobre el estado de las soluciones ict y de los casos prácticos de aplicación de la minería de datos a nivel mundial en al menos 5 casos representativos, Universidad EAFIT, En línea: http://bdigital.eafit.edu.co/bachelorThesis/005.74CDR397/marcoTeorico.pdf

[52]. Jordi Gironés Roig, Metodologías y estándares, Universidad Abierta de Cataluña, En línea: http://www.exabyteinformatica.com/uoc/Administracio_i_direccio_dempreses/Business_analytics/Business_analytics_(Modulo_3).pdf

[53]. Clemente Antonio Martínez Álvarez, Aplicación de técnicas de minería de datos para mejorar el proceso de Control de Gestión en Entel, universidad de chile - Facultad de Ciencias Físicas y Matemáticas, En línea: http://www.tesis.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1

[54]. Erwin Sergio Fischer Angulo, Modelo para la Automatización del Proceso de Determinación de Riesgo de Deserción de Estudiantes Universitarios, Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, En línea: http://preu.unillanos.edu.co/sites/default/files/fields/documentos/PREDICION%20DESERCI.pdf

[55]. José Hernández Orallo, Minería de Datos, Universidad Politécnica de Valencia, En línea: http://users.dsic.upv.es/~jorallo/master/dm5.pdf

m. Anexos

Anexo A: Certificado de la Directora de la Modalidad de Estudios a Distancia

UNIVERSIDAD NACIONAL DE LOJA

MODALIDAD DE ESTUDIOS ADISTANCIA

Ing. Carmen Elizabeth Cevallos Cueva.- DIRECTORA DE LA MODALIDAD DE ESTUDIOS A DISTANCIA.

CERTIFICA:

Que luego de haber autorizado el desarrollo de Tesis, a la Señorita Angélica Elizabeth Jaramillo Zhingre, alumna del Área de la Energía, las Industrias y los Recursos Naturales No renovables, intitulado: "**APLICACIÓN DE TECNICAS DE MINERIA DE DATOS PARA DETERMINAR LAS INTERACCIONES DE LOS ESTUDIANTES EN UN ENTORNO VIRTUAL DE APRENDIZAJE",.** Se recibe, el Informe Ejecutivo, en donde se indica cada uno de los resultados del tema de Investigación.

Lo Certifico.-

Loja, 10 de abril de 2015

Ing. Carmen Cevallos Cueva.
DIRECTORA DE LA MED
cc. Archivo
Elab./Marlene Valdez Pardo.

Anexo B: Resultado preliminar para generación del modelo de minería de datos

Se realizó la evaluación del modelo generado por cada uno de los algoritmos pertenecientes a la técnica de clasificación, comparando cada uno de los resultados obtenidos, por ende se consideraron diferentes parámetros que permitieron evaluar el modelo, para ello se utilizó los datos para entrenamiento en el porcentaje de 67% y los demás datos para pruebas, es decir el 33%.

Entre los parámetros que se tomó en cuenta para evaluar los modelos generados son los siguientes: instancias clasificadas correctamente (accuracy), instancias clasificadas incorrectamente (clasification_error), estadística de Kappa que mide la coincidencia de la predicción con la clase real (Kappa), error cuadrático (squared_error), error relativo (relative_error), error absoluto (absolute_error), presentando los resultados obtenidos en la siguiente tabla (ver TABLA LXX):

TABLA LXX.

EVALUACIÓN DE LOS MODELOS GENERADOS POR LOS ALGORITMOS

Algoritmo	Conjunto de datos	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de Kappa	Error Cuadrático	Error Relativo (%)	Error Cuadrático Medio	Error Cuadrático Relativo
DECISION TREE	Entrenamiento	87.76	12.24	0.769	0.080	16.00	0.283	0.645
	Validación	72.50	27.50	0.441	0.215	28.46	0.413	
JRip	Entrenamiento	88.78	11.22	0.784	0.109	19.63	0.330	0.752
	Validación	77.00	23.00	0.504	0.208	30.53	0.415	1.223
RIDOR	Entrenamiento	76.53	23.47	0.629	0.235	23.47	0.484	1.249
	Validación	56.50	43.50	0.229	0.435	43.50	0.610	2.350
K-NN	Entrenamiento	96.94	3.06	0.948	0.031	3.06	0.175	0.451
	Validación	60.50	39.50	0.321	0.395	39.50	0.607	2.214
PRISM	Entrenamiento	96.94	3.06	0.947	0.031	3.06	0.175	0.451
	Validación	62.00	38.00	0.312	0.405	40.50	0.610	2.138
CHAID	Entrenamiento	76.53	23.47	0.614	0.147	27.55	0.383	0.988
	Validación	33.00	67.00	0.128	0.391	53.32	0.609	
ID3	Entrenamiento	96.94	3.06	0.948	0.017	3.40	0.130	0.336
	Validación	58.00	42.00	0.319	0.371	38.33	0.557	
J48	Entrenamiento	79.59	20.41	0.657	0.177	32.01	0.421	1.086
	Validación	71.00	29.00	0.540	0.266	39.84	0.490	1.815

En la tabla anterior (ver TABLA LXX) se puede observar el resultado de cada algoritmo obtenido mediante la utilización de la herramienta RapidMiner conjuntamente con los datos de los estudiantes del curso virtual inglés de la Modalidad de Estudios a Distancia, donde existe un alto porcentaje de error de clasificación en cada uno de los algoritmos, además se puede indicar que con el conjunto de entrenamiento de los datos la mayoría de los resultados obtenidos de los algoritmos son desfavorables es decir que no sobrepasan el 77% de los datos han sido clasificados correctamente en la validación, los algoritmos que presentan mejores resultados se tiene el K-NN 96.94%, PRISM 96.94%, ID3 96.94% . Así mismo con el conjunto de datos de validación los algoritmos que presentan los mejores resultados de datos clasificados correctamente se tiene el JRip 77%, Decision Tree 72.50%, J48 71%.

Con el conjunto de datos utilizados en el entrenamiento el mejor algoritmo que presenta es el K-NN, ID3 Y PRISM con un porcentaje de 96.94% es decir las instancias se han clasificado correctamente y un 3.06% de las instancias clasificadas incorrectamente, el índice de Kappa 0.948, el error cuadrático 0.013, error relativo 1.26, error cuadrático medio 0.11 y finalmente el error cuadrático relativo con el 0.397; y con el conjunto de datos utilizada en la validación el algoritmo que ha arrojado el mejor resultado es el JRip con un 77% que significa que dicho porcentaje es el de las instancias clasificadas correctamente, el 23% es el de las instancias clasificadas incorrectamente, el índice de Kappa 0.504, el error cuadrático 0.208, error relativo 30.53, error cuadrático medio 0.415 y finalmente el error cuadrático relativo con el 1.223.

En la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos ya sea tanto en el entrenamiento como en la validación en donde se muestra las instancias clasificadas correctamente (ver Figura 95).

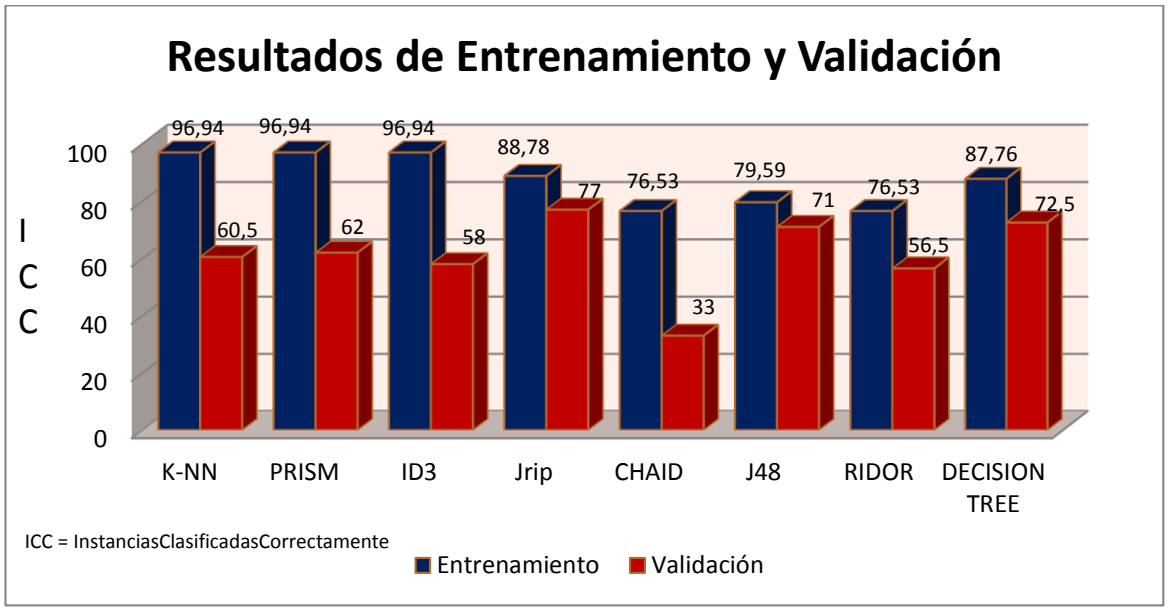


Figura 95: Resultados por cada algoritmo

Así mismo en la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos en lo que respecta en la evaluación en donde se muestra las instancias clasificadas correctamente y las instancias clasificadas incorrectamente, donde se puede mencionar que el algoritmo que tiene mayor porcentaje es el JRip con el 77,00% con un margen de error del 23,00%, el mismo se muestra a continuación (ver Figura 96).

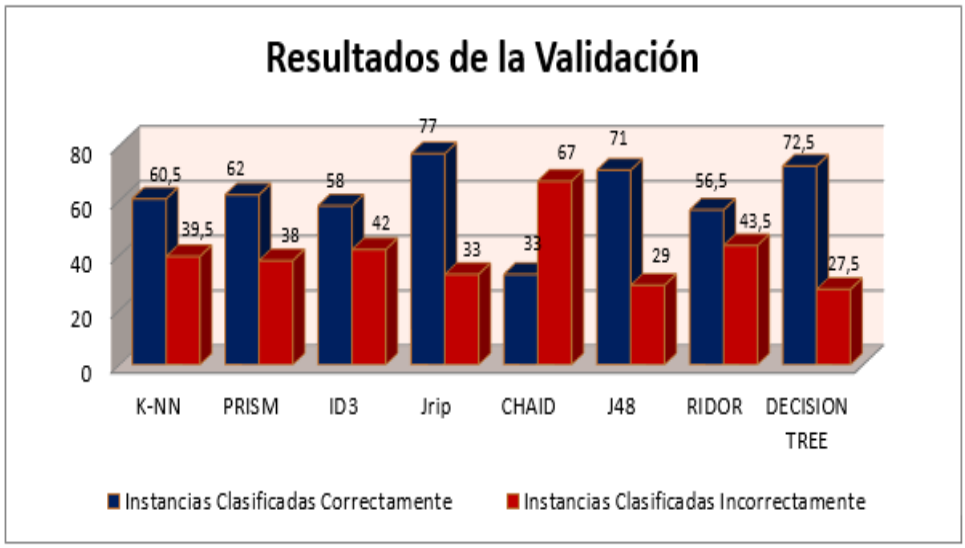


Figura 96: Resultados de algoritmos de instancias clasificadas correcta e incorrectamente

Anexo C:

1. Migración de la Base de Datos a la Herramienta de minería de datos

Para efectuar la migración de los datos se utilizó MySQL, la misma que permitió la conexión con la herramienta RAPIDMINER, para luego poder manipular los datos y llevar a cabo los procesos de minería de datos que permitan generar el modelo (ver Figura 97).

Figura 97: Ventana principal de RAPIDMINER

En la siguiente figura se puede indicar el primer paso para realizar la conexión con la base de datos e importarlos a la herramienta RAPIDMINER (ver Figura 98).

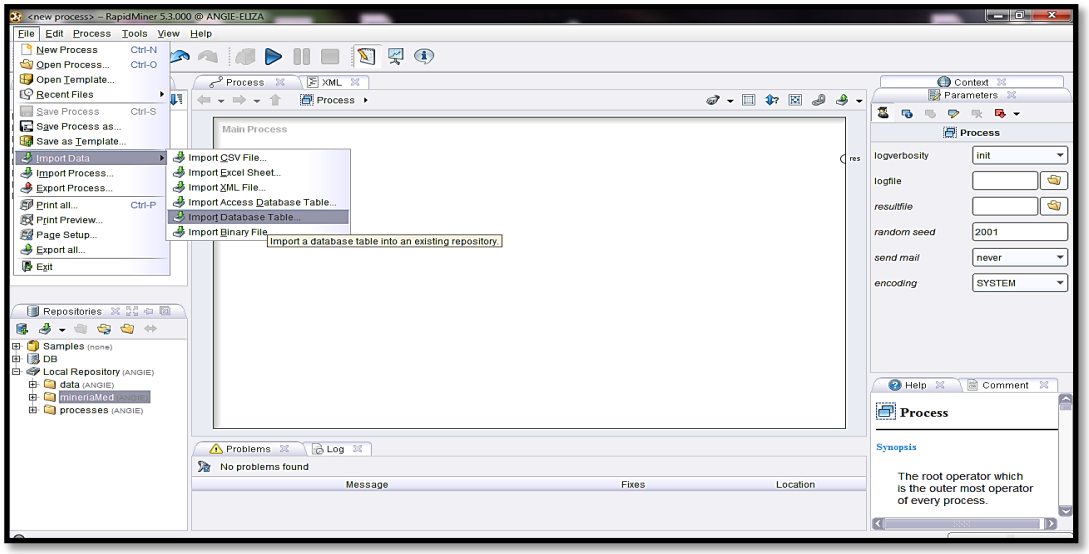


Figura 98: Importar los datos de la base de datos

En la siguiente figura se indica las configuraciones que permiten la conexión con la base de datos (ver Figura 99).

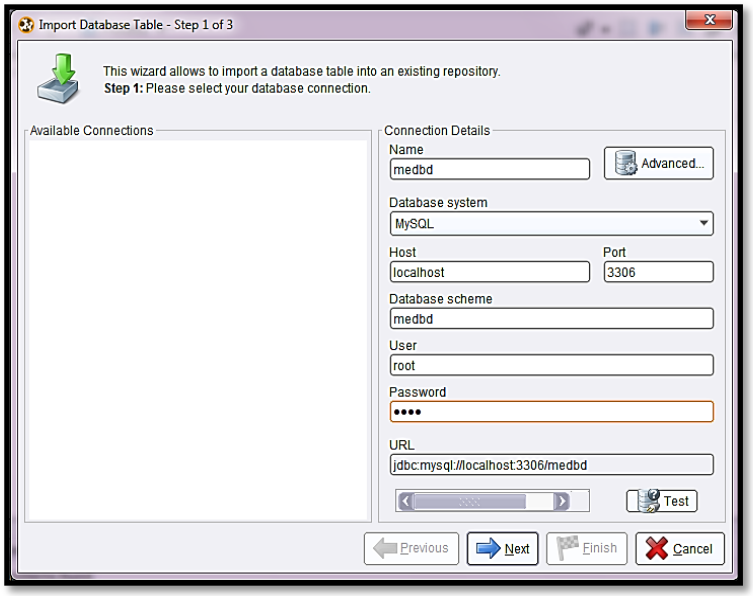


Figura 99: Configuración para la conexión a la base de datos

En la siguiente figura se observa (ver Figura 100) las tablas que pertenecen a la base de datos con la que se conectó, luego de la configuración anterior.

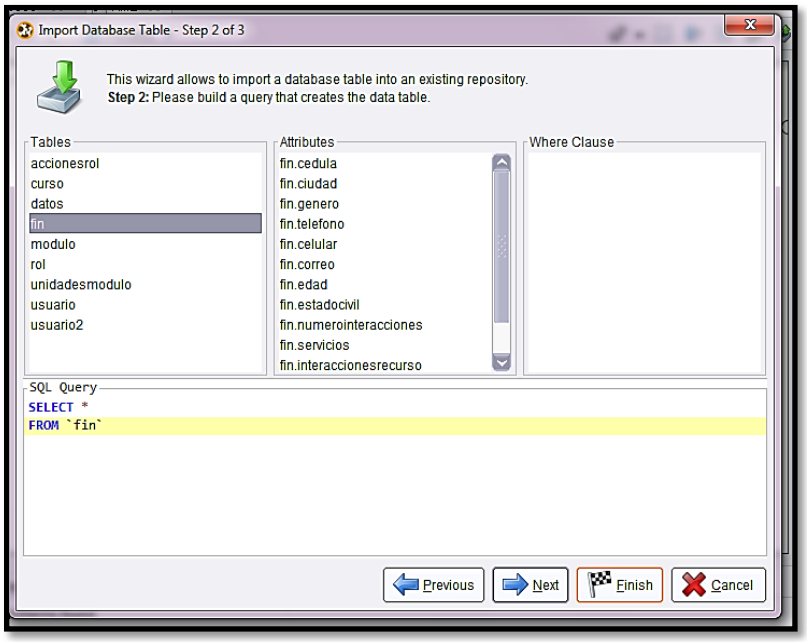


Figura 100: Tablas de las base de datos

En la siguiente figura se utiliza un operador de la herramienta RAPIDMINER para comprobar que los datos se importaron correctamente (ver Figura 101).

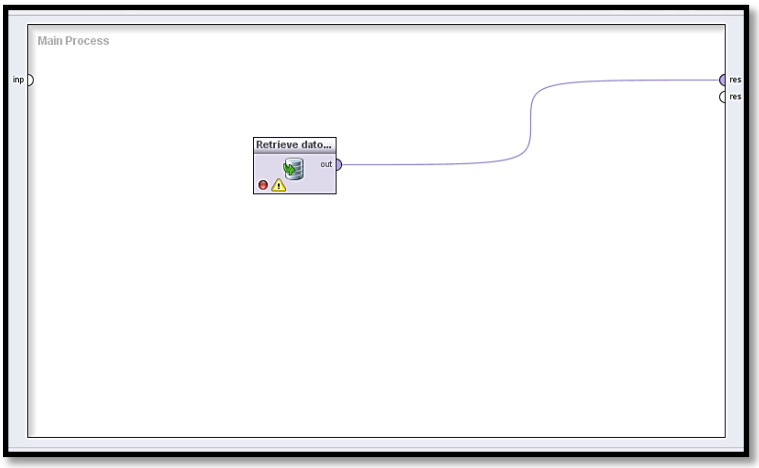


Figura 101: Operador de la base de datos

En la siguiente figura se indica los resultados donde se indica los datos de lo que contiene la base de datos (ver Figura 102).

cedula	ciudad	genero	edad	estadocivil	numerointe...	serv...	interaccion...	interaccion...	interaccion...	carrera	modalidad...	horario	...	trabajo	n
701181505	PIÑAS	masculino	40	casado	417	3	58	92	84	Derecho	distancia	nocturno	U	No	3
000758390	Loja	femenino	22	soltero	202	3	34	54	27	Contabilidac	presencial	matutino	U	No	9
104871296	SAN LUCAS	masculino	25	casado	518	3	60	89	122	EducaciÃƒÂ³	semipresen	matutino	U	No	1
104606643	Loja	femenino	27	divorciado	169	3	25	44	32	Derecho	presencial	nocturno	U	No	1
104459761	Loja	masculino	29	casado	14	3	0	0	0	Derecho	distancia	nocturno	U	No	0
103969430	LOJA	masculino	33	casado	297	3	61	50	50	Derecho	distancia	nocturno	U	No	1
000650944	Loja	femenino	25	soltero	83	3	4	27	16	Contabilidac	presencial	matutino	U	No	1
104894124	Loja	femenino	23	casado	172	3	32	30	33	PsicologÃƒÂ	presencial	vespertino	U	No	0
104700057	Loja	masculino	27	soltero	216	3	25	27	36	Administraci	presencial	vespertino	R	No	1
703379784	Santa Rosa	femenino	37	soltero	334	3	52	39	70	Contabilidac	distancia	matutino	U	Si	1
104061377	loja	masculino	30	divorciado	87	3	29	25	16	Cultura FÃƒÂ	presencial	matutino	U	No	2
104028038	LOJA	femenino	33	soltero	112	3	14	38	18	QuÃƒÂ-mico	presencial	vespertino	U	No	1
722372081	LOJA	femenino	27	soltero	404	3	95	77	80	Administraci	distancia	matutino	U	No	0
104770654	Loja	femenino	27	casado	76	2	0	31	21	FÃƒÂ-sico M	presencial	vespertino	U	No	2
104299019	Loja	masculino	28	soltero	89	3	0	28	18	IngenierÃƒÂ-	presencial	matutino	U	No	0
715530422	Santo Domir	masculino	36	soltero	243	3	7	56	42	Contabilidac	distancia	matutino	U	No	1
705202299	PIÑAS	femenino	26	soltero	510	3	36	111	137	Derecho	distancia	nocturno	U	No	1
104673825	Piñas	masculino	27	soltero	210	3	27	47	38	Derecho	distancia	matutino	U	No	0

Figura 102: Información de la base de datos

2. Conexión de RapidMiner Studio con Twitter

El conector de Twitter permite acceder fácilmente a los datos de Twitter directamente desde RapidMiner Studio, el conector puede buscar frases, mensajes de twitter o información del perfil de usuario. El conector de Twitter utiliza un mecanismo de autenticación llamado OAuth 2.0, en lugar de dar RapidMiner su nombre de usuario y contraseña, se genera un token de acceso que puede ser utilizado por RapidMiner Studio para conectarse a su cuenta de Twitter, esta ficha no puede ser utilizado por otras aplicaciones y ayuda a mantener sus credenciales de Twitter seguro [22].

❖ **Pasos para conectar RapidMiner Studio con Twitter:**

1. Crear un nuevo proceso New Process en RapidMiner Studio, arrastre la búsqueda Twitter operador en el Editor de procesos, y hacer clic en el operador. Haga clic en el icono de Twitter en los parámetros del operador para abrir el cuadro de diálogo Administrar conexiones. También puede abrir la ventana Administrar conexiones a través de Herramientas> Gestionar conexiones (ver Figura 103).

Operador para Twitter

Figura 103: Operador Twitter

Con la búsqueda de Twitter operador, puede especificar una consulta y obtener estados de Twitter que contienen esta consulta. La lista de estados contiene datos adicionales con el contexto de los estados. En el modo experto, puede especificar restricciones adicionales de búsqueda.

1. Seleccione una conexión de Twitter para especificar la cuenta de Twitter para el acceso a la API de Twitter. Especificar por lo menos una consulta para buscar Twitter por ello. En el modo experto, puede especificar restricciones adicionales de búsqueda (ver Figura 104).

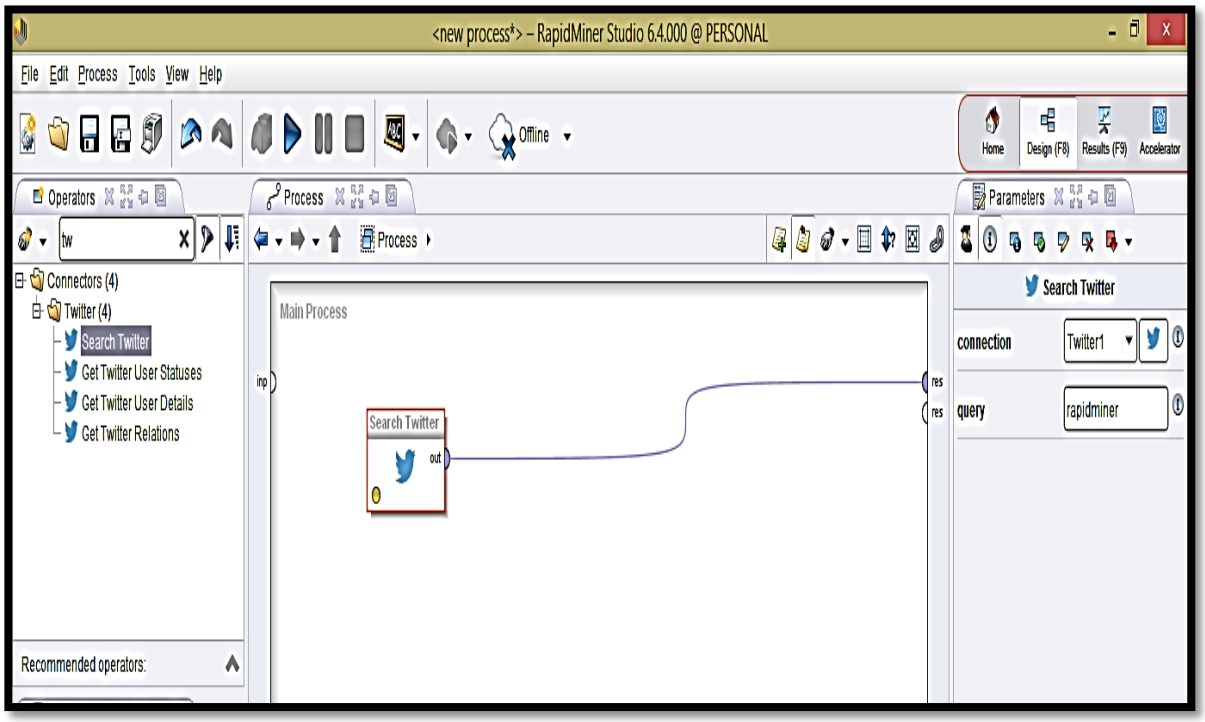


Figura 104: Conexión del operador para Twitter.

2. Haga clic en el botón de conexión add en la parte inferior izquierda de la ventana, darle un nombre para la nueva conexión y seleccione Twitter Conexión para Tipo de conexión (ver Figura 105):

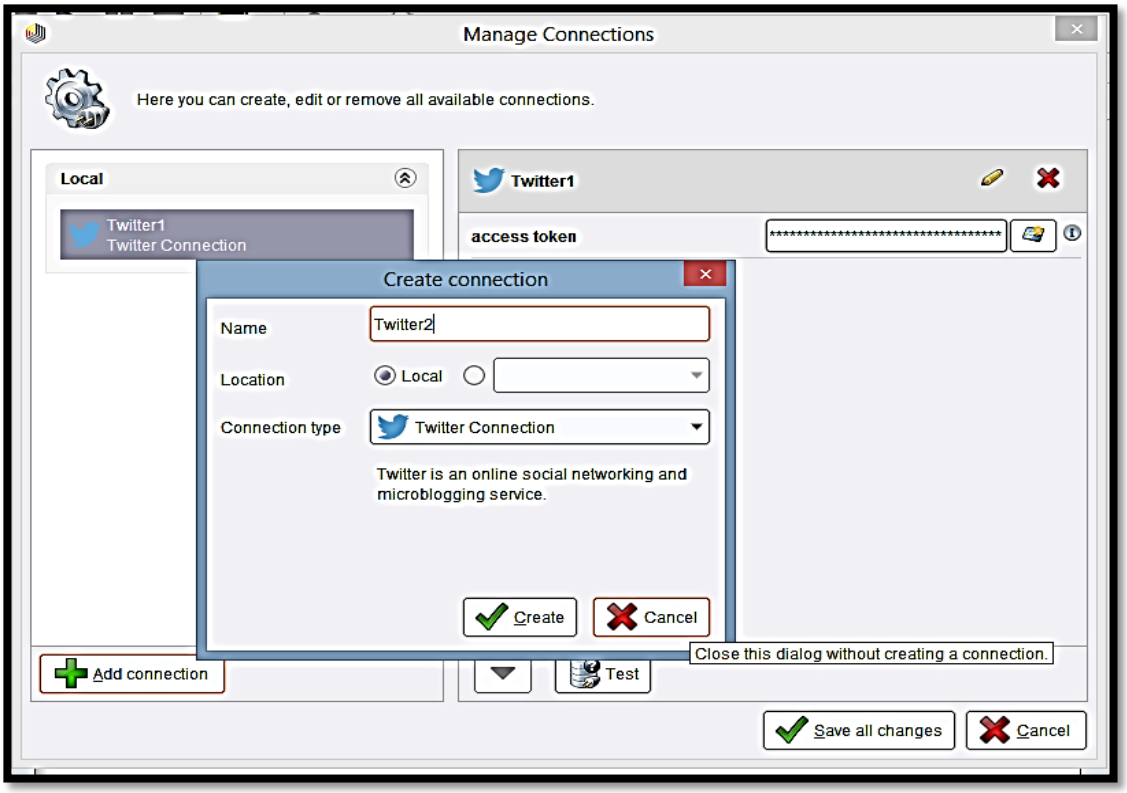


Figura 105: Establecer nueva conexión twitter

3. Haga clic en Crear y seleccione su nueva conexión de Twitter.
4. Haga clic en el botón de autenticación a la derecha del campo de token de acceso.
5. Haga clic en Solicitud token de acceso para abrir el sitio web de Twitter. Puede que tenga que iniciar sesión en su cuenta de Twitter. Puede copiar manualmente la URL haciendo clic en Mostrar URL en su lugar (ver Figura 106).

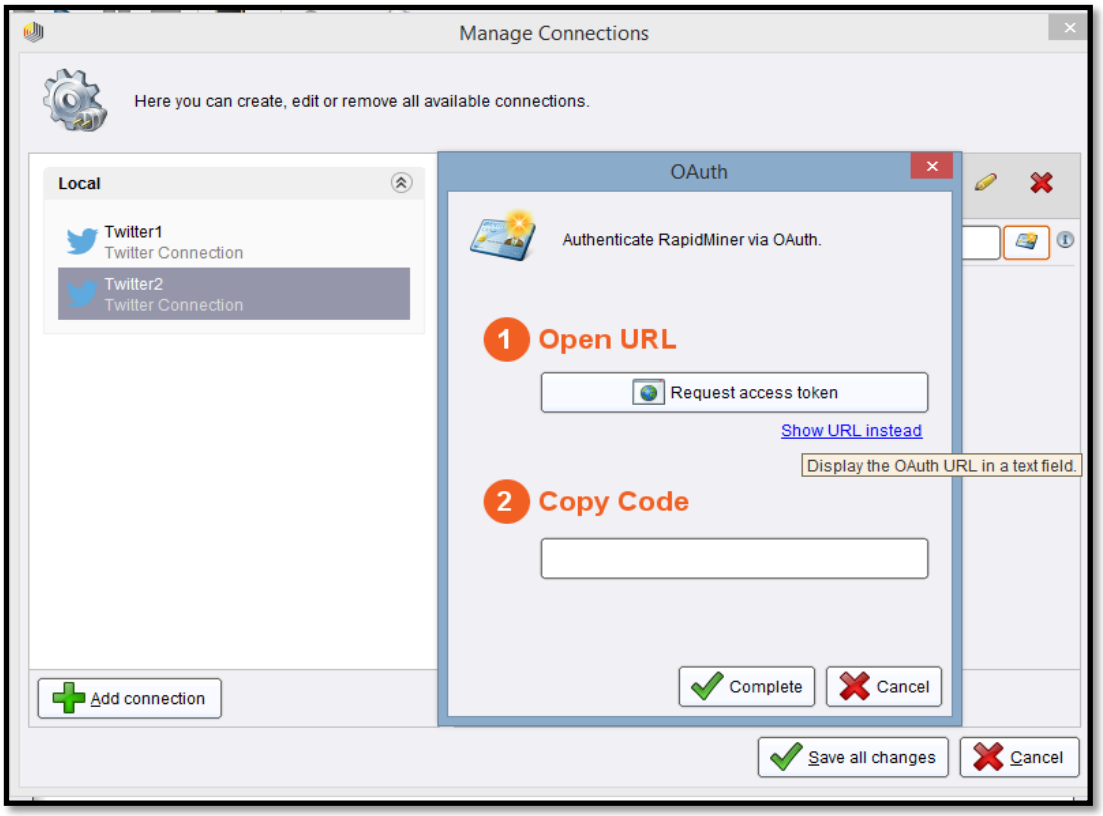


Figura 106: Autenticación de rapidminer

6. Permitir a RapidMiner para acceder a su cuenta de Twitter haciendo clic en Autorizar aplicación:
7. Copiar el token de acceso se muestra en la siguiente página (ver Figura 107):

Figura 107: Token de acceso

8. Regresar a RapidMiner Estudio, introduzca el token de acceso, y haga clic en Completo (ver Figura 108):

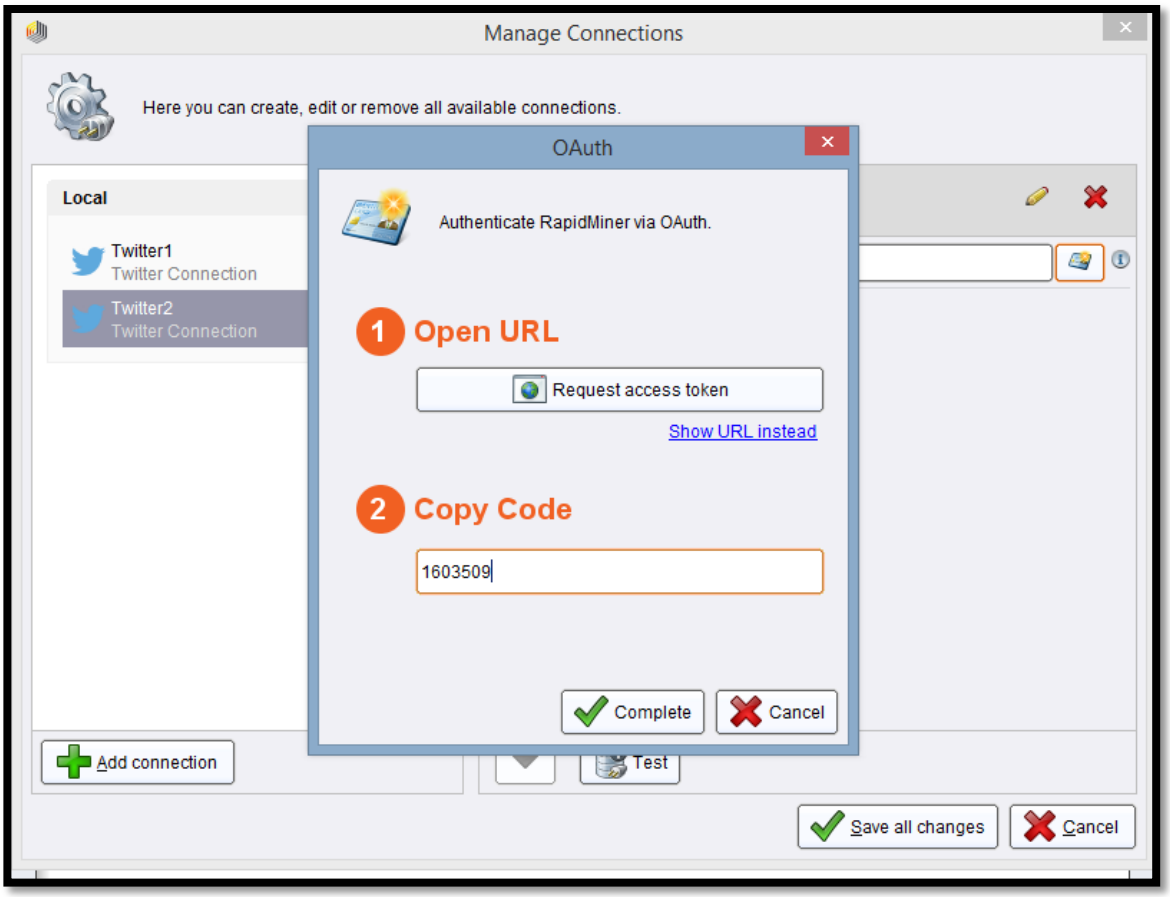


Figura 108: Configurando el token de twitter

9. Si bien no es obligatorio, se recomienda probar la nueva conexión Twitter haciendo clic en el botón Prueba en la parte inferior de la ventana Administrar (ver Figura 109).

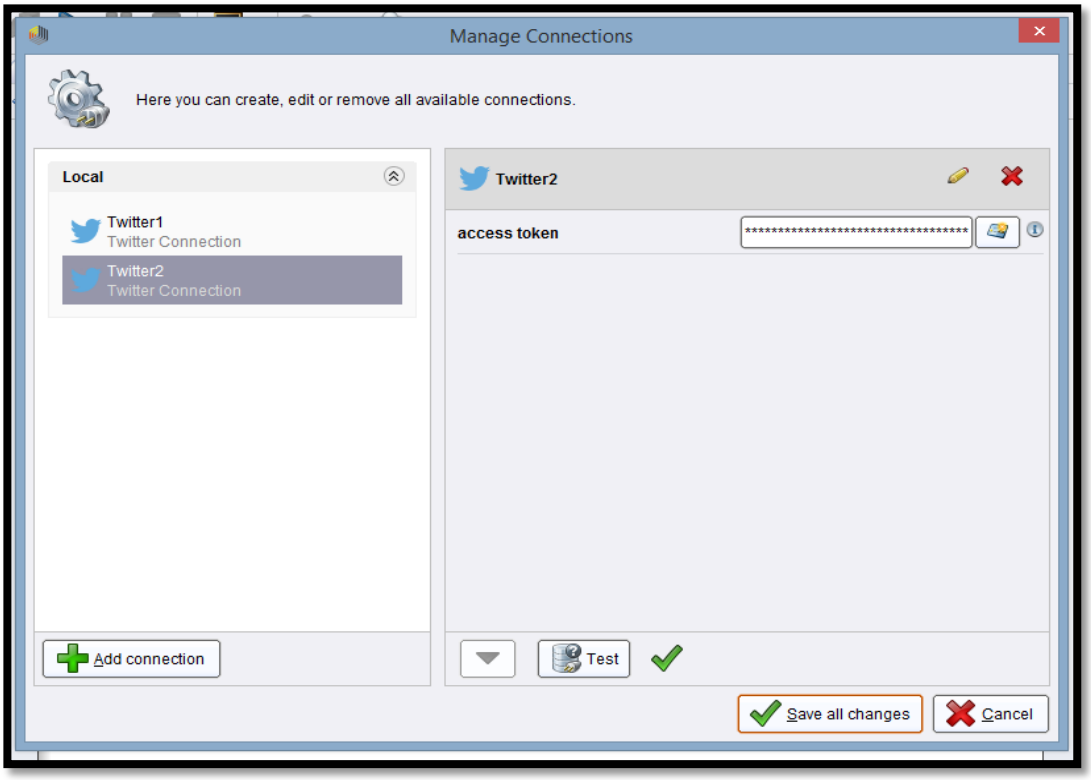


Figura 109: Probando configuración con twitter

10. Cierre el cuadro de diálogo Administrar Conexiones, haga clic en Guardar todos los cambios ✔ .
11. Ha conectado con éxito RapidMiner Studio para su cuenta de Twitter! Las siguientes secciones le mostrarán cómo utilizar este conector para buscar diferentes tipos de información de Twitter (ver Figura 110).

❖ **Búsqueda de tweets que contengan una frase**

Con la búsqueda de Twitter operador podrá encontrar todos los tweets que contienen una frase especificada, junto con los metadatos tweets`.

1. Haga clic en la búsqueda de Twitter operador en el Editor de Proceso.
2. Seleccione su conexión a Twitter desde la conexión del menú desplegable en los parámetros del operador y rellene el campo de consulta. Esta es la frase vamos a buscar para Twitter.

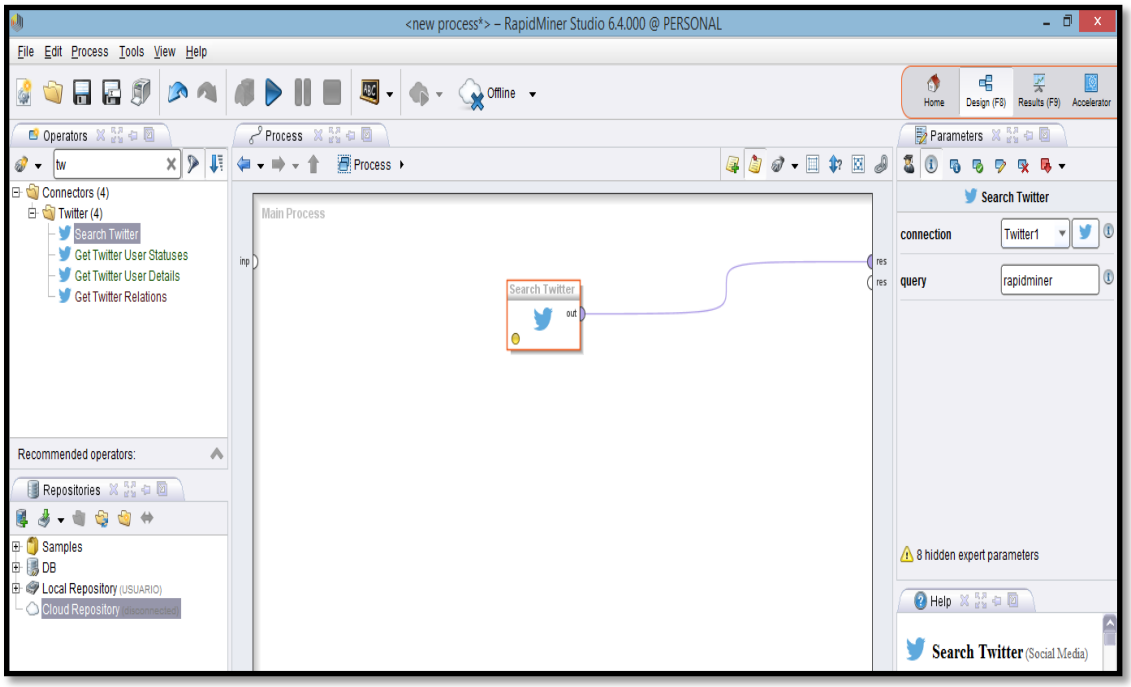


Figura 110: Establecer conexión con rapidminer

3. Ejecutar el proceso y ver los resultados (ver Figura 111):

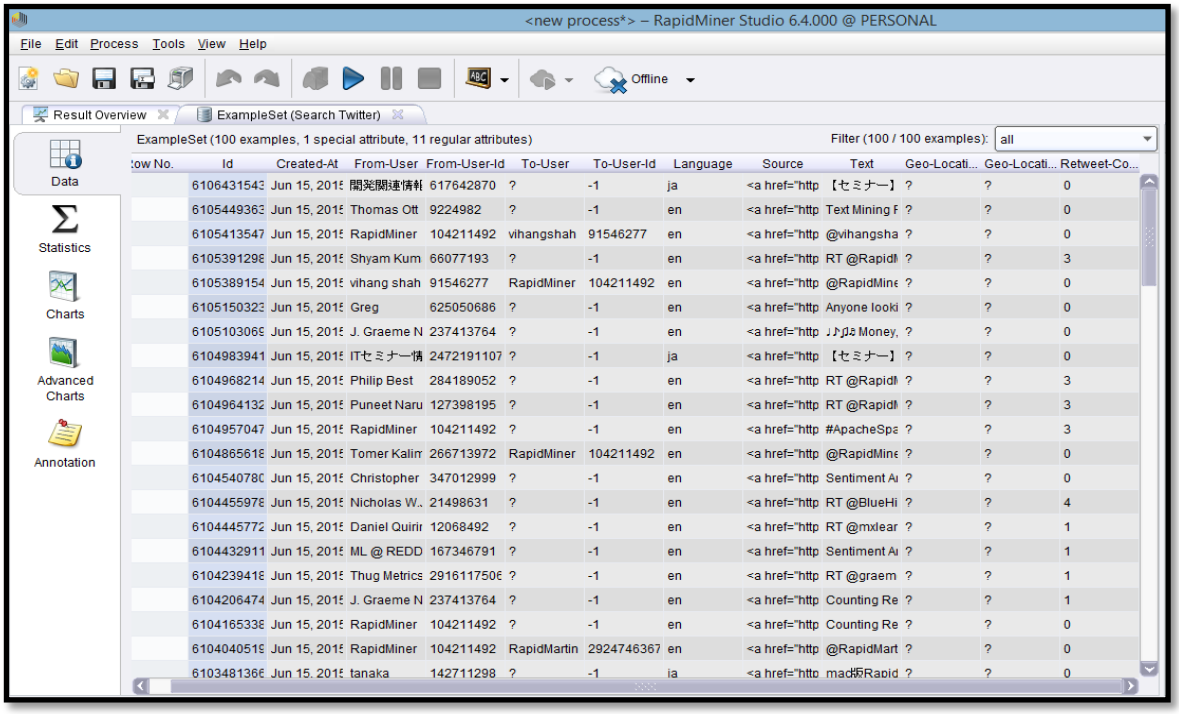


Figura 111: Resultados de Rapidminer

❖ Obtener cuenta detalles de un usuario de Twitter

El Twitter Detalles de usuario operador se puede utilizar para encontrar la información de perfil de un nombre de usuario especificado.

1. Haga clic en el Twitter Detalles de usuario operador, seleccionar la conexión de Twitter en los parámetros del operador, y especificar el usuario por nombre de usuario o número de identificación (ver Figura 112):

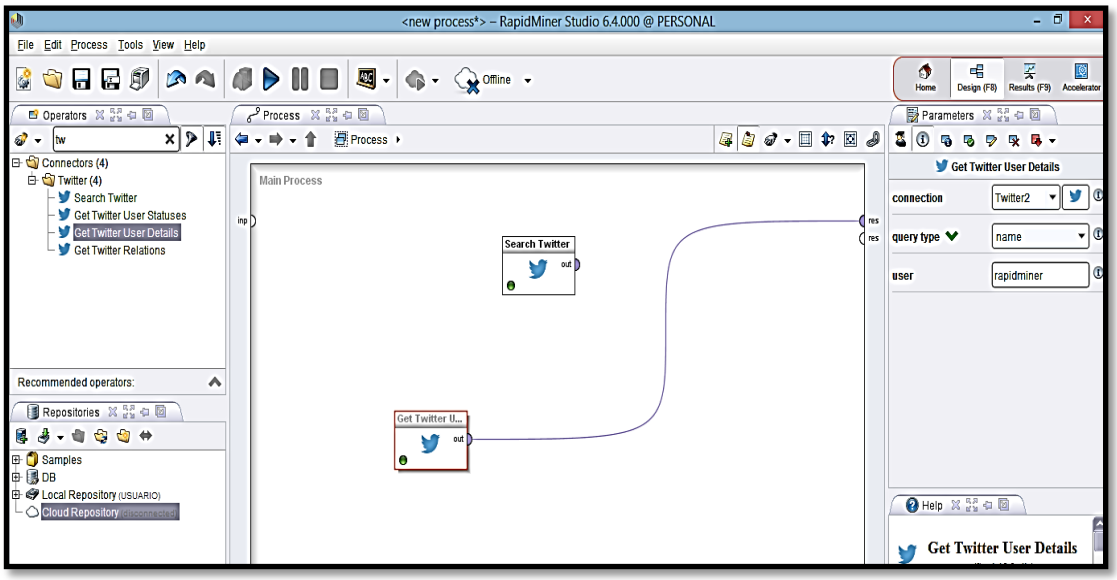


Figura 112: configuración del operador de Twitter user

2. Ejecutar el proceso (ver Figura 113).

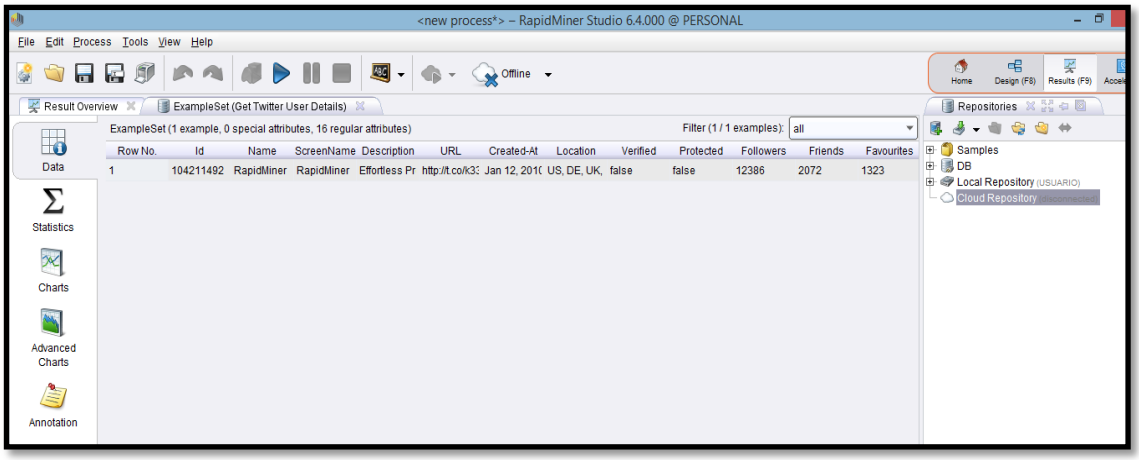


Figura 113: Resultados de perfiles de usuarios

La tabla contiene información tal como datos de creación de la cuenta y el número de seguidores. Para obtener más información sobre los seguidores del usuario, consulte la siguiente sección.

❖ Obtener una Lista de amigos o seguidores

El operador Obtén Relaciones de Twitter se puede utilizar para obtener una lista de todos los amigos o seguidores de un usuario.

1. Haga clic en el operador Obtener Relaciones Twitter, selecciona la conexión de Twitter en los parámetros del operador y especifique el usuario por nombre de usuario o número de identificación (ver Figura 114):

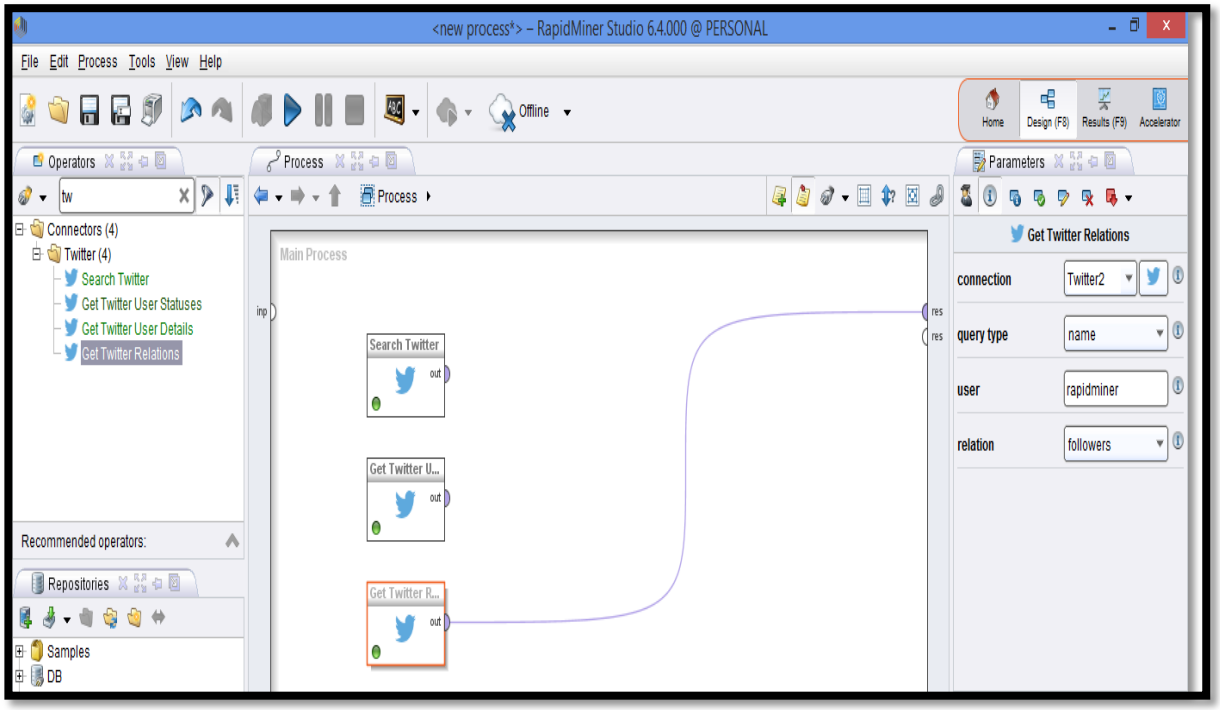


Figura 114: conexión del operador relaciones twitter

2. Ejecutar el proceso (ver Figura 115).

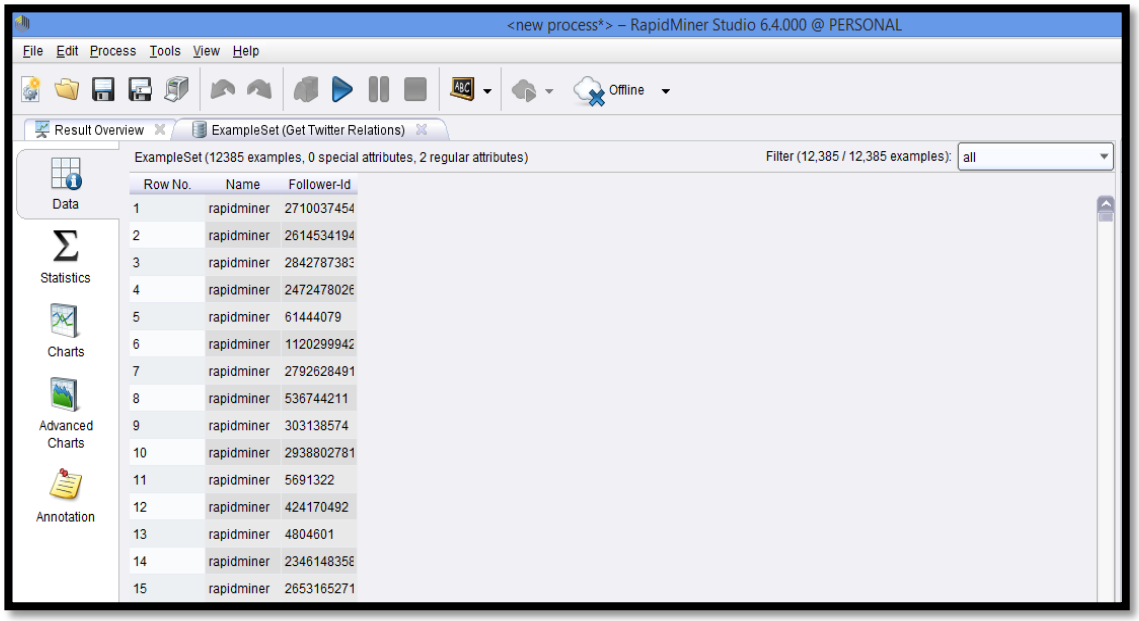


Figura 115: usuarios de twitter

La primera columna contiene el usuario que ha especificado y la segunda columna muestra el numberss Identificación de sus seguidores. Con su conocimiento de la Obtén Detalles Twitter del usuario operador anterior.

Anexo D: Operadores Utilizados en la Herramienta RapidMiner

- **Base de datos:** Mediante este operador se puede utilizar los datos que se tiene en la base de datos (ver Figura 116).

Figura 116: Componente de la base de datos

- **Parse Numbers:** Se utiliza para cambiar el tipo de atributos nominales a un tipo numérico es decir convertir algún atributo que se requiera como puede ser de nominal a entero para posteriormente se puede utilizar con los siguientes operadores y no se tenga ningún inconveniente (ver Figura 117).

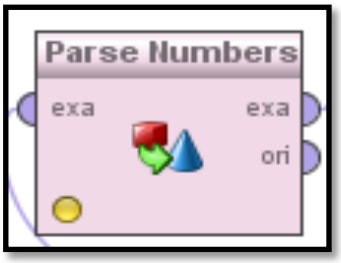


Figura 117: Componente Parse Numbers

- **Generate Attributes:** Este operador construye nuevos atributos a partir de los atributos del conjunto de datos de entrada o data set. Los nombres de los nuevos atributos y sus descripciones de construcción se definen en el parámetro functions descriptions, permite la escritura de expresiones con diferentes operaciones o funciones (ver Figura 118).

Figura 118: Componente Generate Attributes

En el presente trabajo se construyó algunos atributos que requerían ser programadas para ello se utilizó cadenas if () como se puede observar en las siguiente figura (ver Figura 119).

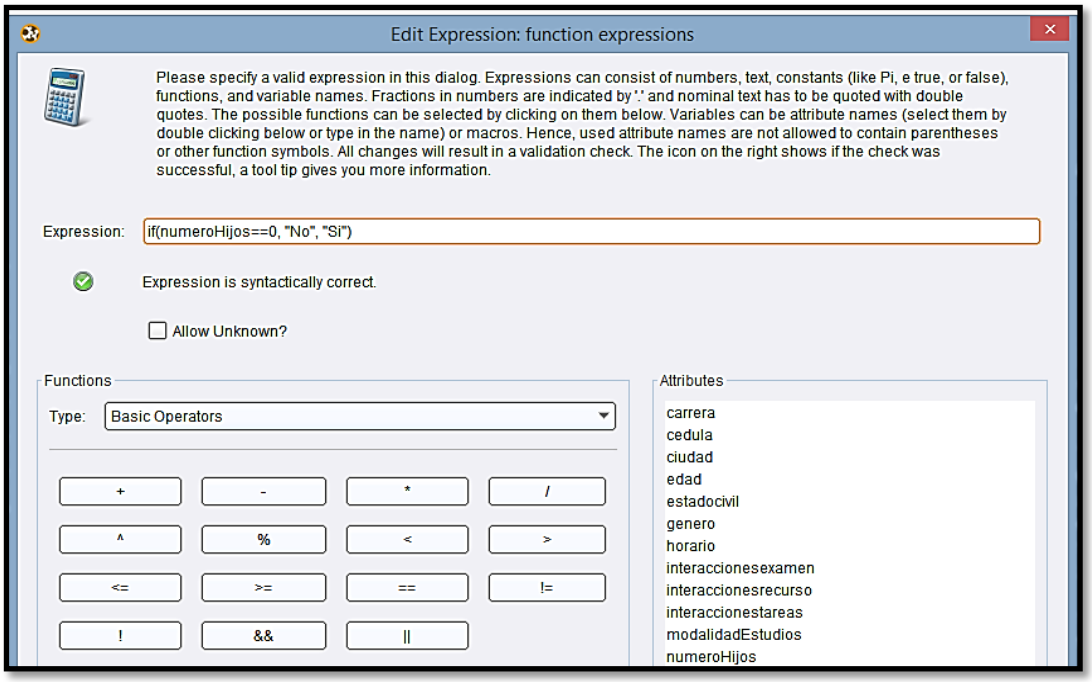


Figura 119: Panel de configuración de atributos

- **Discretize:** Mediante este operador se establece el atributo objetivo que es el número de interacciones, además discretiza los atributos numéricos seleccionados, bien con atributos nominales u ordinales. El límite inferior de cada clase se define automáticamente como el límite superior de la clase anterior, como se puede observar en la siguiente figura (ver Figura 120).

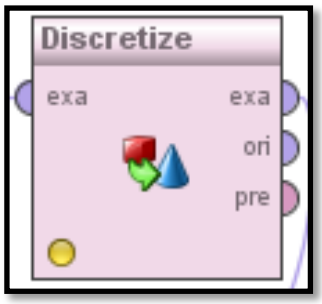


Figura 120: Componente Discretize

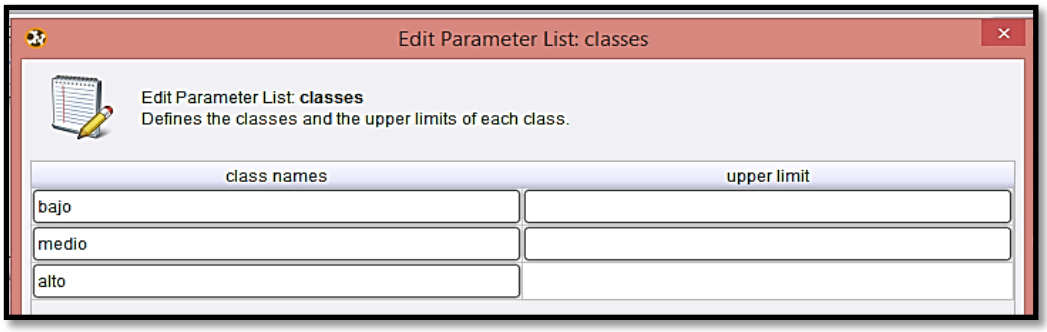


Figura 121: Panel de configuración del atributo objetivo

- **Set Role:** Este operador se utilizó para determinar el rol de un atributo del conjunto de datos, el rol objetivo indica si el atributo es un atributo regular o un atributo especial, en este caso es el número de interacciones de los estudiantes que es de tipo label (ver Figura 122).

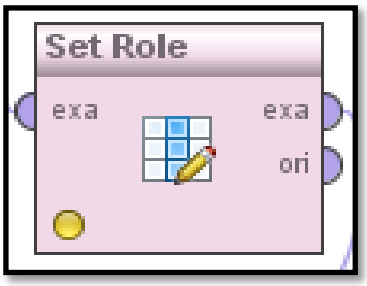


Figura 122: Componente Set Role

- **Select Attribute:** Mediante este operador se selecciona cada uno los atributos que se va utilizar para el desarrollo del modelo, se puede observar en la siguiente figura (ver Figura 123).

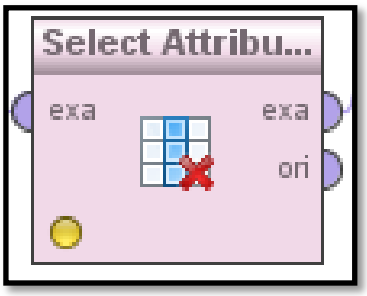


Figura 123: Componente Discretiza

- **Multiply:** Permite dividir el conjunto de datos que se seleccionó inicialmente para hacer un conjunto de entrenamiento y otro para validación (ver Figura 124).

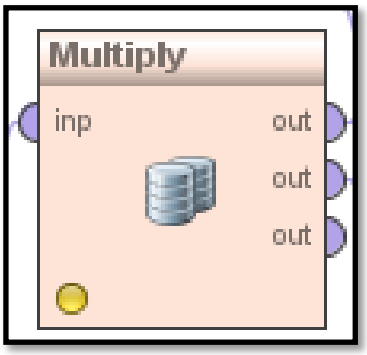


Figura 124: Componente Multiply

- **Sample:** Permite obtener una muestra del conjunto de datos, estableciendo un porcentaje de datos que se va a utilizar para entrenamiento y validación (ver Figura 125).

Figura 125: Componente Sample

- **Weight by information gain:** Sirve para obtener el peso de cada atributo con que se está trabajando, así mismo este operador permite calcular el nivel de contribución del atributo para predecir la clase, debe considerarse que el peso es un cálculo a priori, es decir que se calcula considerando que cada atributo influye en la clase (Figura 126).

Figura 126: Componente Weight by information gain

- **Apply Model:** Este operador aplica el modelo al conjunto de datos, los modelos suelen contener información sobre los datos con los han sido entrenados. La cantidad de atributos, el orden, el tipo y el rol son consistentes durante el entrenamiento y la aplicación (ver Figura 127).

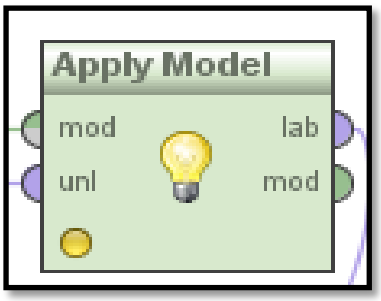


Figura 127: Componente Apply Model

- **Performance:** Mediante este operador se selecciona los criterios para evaluar el modelo, como es la matriz de confusión donde indica las instancias clasificadas correctamente y las que no se clasificaron correctamente, otros (ver Figura 128).

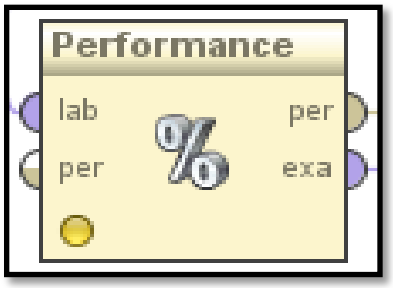


Figura 128: Componente Performance

- **Validation:** Permite realizar la validación cruzada, donde se establece el número de validaciones, que es el número de subconjuntos que se crean para evaluar el modelo (ver Figura 129).

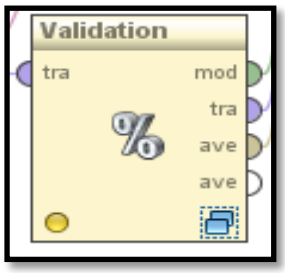


Figura 129: Componente Validation

Anexo E: Algoritmos de la técnica de clasificación en Rapidminer

- **Algoritmo JRip**

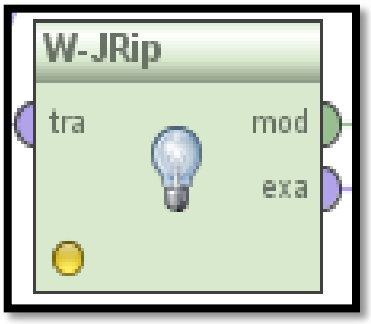


Figura 130: Componente Algoritmo JRIP

- F: Establecer el número de pliegues para las excepciones.
- N: Pesos mínimos
- D: Depuración
- S: Número aleatorio para los datos con el fin de obtener una mejor regla.

- **Algoritmo Ridor**

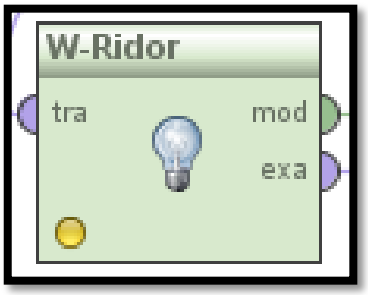


Figura 131: Componente Algoritmo Ridor

- F: Establecer el número de pliegues para las excepciones.
- S: Número aleatorio para los datos con el fin de obtener una mejor regla.
- A: Banderas para la tasa de error de los datos
- N: Pesos mínimos

- **Algoritmo K-NN**

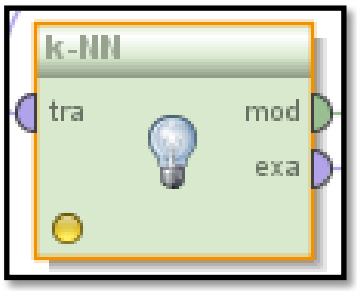


Figura 132: Componente Algoritmo K-NN

- K: Clasificación de los vecinos más cercanos.
- Weighted vote: Bandera que se establece para el peso de los vecinos.
- Measure types: Tipo de medida para encontrar el vecino más cercano.
- Mixed measure

- **Algoritmo Prism**

Figura 133: Componente Algoritmo Prism

- **D:** Depuración

- **Algoritmo Chaid**

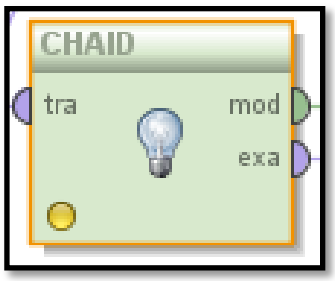


Figura 134: Componente Algoritmo CHAID

- Minimal size for Split: La longitud mínima de cada nodo.
- Minimal leaf size: Tamaño mínimo de un nodo hoja.
- Minimal gain: Ganancia mínima de un nodo.
- Maximal deph: Profundidad máxima o tamaño del árbol de decisión.
- Confidence: Nivel de confianza utilizada para el cálculo de error.
- Number of prepruning: Número de nodos probados.

➢ **Algoritmo Decision Tree**

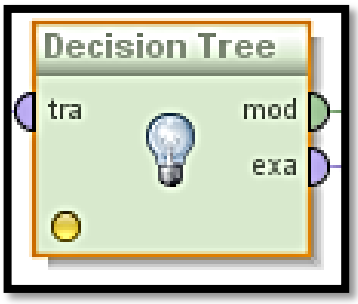


Figura 135: Componente Algoritmo Decision Tree

Este operador entrena arboles de decisión a partir de datos nominales y numéricos. Cada vez que un nuevo nodo se crea en un determinado momento, un atributo se escoge para maximizar el poder de discriminación de ese nodo con respecto a las demás alternativas asignadas al subárbol en particular (ver Figura 135).

- Criterion: Especifica el criterio de selección de atributos y de divisiones numéricas (ganancia de información, precisión, proporción de ganancia).
- Minimal size for Split: Tamaño mínimo de divisiones que se pueden dar en cada nodo.
- Minimal leaf size: Tamaño mínimo de cada hoja.
- Minimal gain: La ganancia mínima que debe lograrse con el fin de producir una división.
- Maximal depth: Profundidad máxima o tamaño del árbol de decisión.
- Confidence: El nivel de confianza utilizado para el cálculo del error pesimista de la poda.

- Number of prepruning: El número de nodos alternativos probados cuando la técnica de la poda evitaría una división.

➢ **Algoritmo ID3**

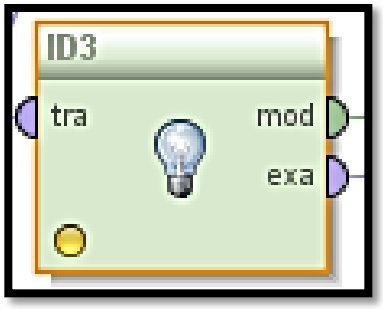


Figura 136: Componente Algoritmo ID3

- Criterion: Criterio de evaluación.
- Minimal size for Split: Número mínimo de divisiones que se puede dar por cada nodo.
- Minimal leaf size: Tamaño mínimo de cada hoja.
- Minimal gain: Ganancia mínima

➢ **Algoritmo J48**

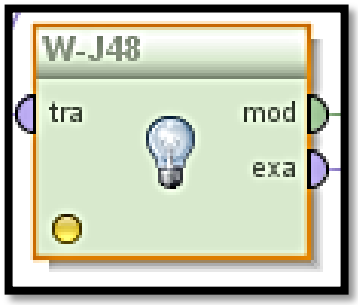


Figura 137: Componente Algoritmo J48

- C: Umbral de confianza para la poda.
- M: Número mínimo de instancias por hoja.
- N: Número de pliegues de la poda.
- B: Divisiones binarias.
- L: Limpieza del árbol.
- Q: Semilla de datos aleatorios.

Anexo F: Migración y alojamiento de los datos en una Base de Datos.

La información de los estudiantes que fue proporcionada por la Unidad de Telecomunicaciones e Información de la Universidad Nacional de Loja, se encontraba en archivos .xml para ello fue necesario la realización de una aplicación para que facilite el transformar estos datos a una base de datos y de tal forma poder utilizarla para llevar a cabo los procesos que se requiera para la realización del modelo de minería de datos, a continuación se presenta cada una de las clases java con una descripción de las funciones de las mismas:

Clase conexión: El siguiente fragmento de código (ver Figura 138) contiene la conexión con la base de datos a la cual se va almacenar la información.

```
 * @author ANGIE
 */
public class Conexion {
    public Connection getconnection() throws SQLException, ClassNotFoundException {
        Connection con = null;
        String url = "", driver = "";
        try {
            url = "jdbc:mysql://localhost/data";
            driver = "com.mysql.jdbc.Driver";
            Class.forName(driver);
            con = DriverManager.getConnection(url, "root", "");
        } catch (Exception e) {
            System.out.println("Error al conectarse con la BD\nError:" + e);
        }
        return con;
    }
}
```

Figura 138. Conexión con la Base de Datos

Clase rol: Sirvió para extraer la información del archivo .xml y pasarla a la base de datos, contiene el rol del participante en el curso como es nombre del rol que ocupa el estudiante (ver figura 139).

```
public class ROL {
    public static void main(String[] args) throws SQLException, ClassNotFoundException {
        Conexion conexion = new Conexion();
        Connection conMysql = conexion.getconnection();
        Statement st = conMysql.createStatement();
        st = conMysql.createStatement();
        SAXBuilder builder = new SAXBuilder();
        File xmlFile = new File("D:\\10 modulo 2014\\Anteproyectos\\ANTEPROYECTO\\DATOS\\moodle.xml");
        try {
            Document document = (Document) builder.build(xmlFile);
            Element rootNode = document.getRootElement();
            List moodle = rootNode.getChildren();
            Element roles = (Element) moodle.get(1);
            List roleslist = roles.getChildren();
            for (int i = 0; i < 5; i++) {
                Element rol = (Element) roleslist.get(i);
                List rollist = rol.getChildren();
                Element id = (Element) rollist.get(0);
                Element name = (Element) rollist.get(1);
                Element shortname = (Element) rollist.get(2);
                Element nameincourse = (Element) rollist.get(3);
                int auxid = 0;
                String aux = "";
                if(nameincourse.getName().equals("NAMEINCOURSE")){
                    aux = nameincourse.getValue();
                    auxid = 4;
                }else{
                    aux = "";
                    auxid = 3;
                }
                String query = "insert into rol "
                        + "values ('" + id.getValue() + "','" + name.getValue() + "','"
                        + shortname.getValue() + "','" + aux + "')";
```

Figura 139. Importación de los roles a la Base de Datos

Clase usuario: sirvió para extraer la informacion del estudiante que se encuentra en el archivo .xml y pasarla a la base de datos, entre los datos se tienen nombres, cédula, dirección, entre otros, (ver figura 140).

```
public class USER {
    public static void main(String[] args) throws SQLException, ClassNotFoundException {
        Conexion conexion = new Conexion();
        Connection conMysql = conexion.getconnection();
        Statement st = conMysql.createStatement();
        st = conMysql.createStatement();
        SAXBuilder builder = new SAXBuilder();
        File xmlFile = new File("D:\\10 modulo 2014\\Anteproyectos\\ANTEPROYECTO\\DATOS\\moodle.xml");
        try {
            Document document = (Document) builder.build(xmlFile);
            Element rootNode = document.getRootElement();
            List moodle = rootNode.getChildren();
            Element course = (Element) moodle.get(2);
            List courselist = course.getChildren();
            Element users = (Element) courselist.get(3);
            List userslist = users.getChildren();
//            System.out.println(logslist);
            for (int i = 0; i < userslist.size(); i++) {
                Element user = (Element) userslist.get(i);
                List userlist = user.getChildren();
                Element iduser = (Element) userlist.get(0);
                Element auth = (Element) userlist.get(1);
                Element confirmed = (Element) userlist.get(2);
                Element policyagreed = (Element) userlist.get(3);
                Element deleted = (Element) userlist.get(4);
                Element username = (Element) userlist.get(5);
                Element idnumber = (Element) userlist.get(6);
                Element firstname = (Element) userlist.get(7);
                Element lastname = (Element) userlist.get(8);
                Element email = (Element) userlist.get(9);
                Element emailstop = (Element) userlist.get(10);
                Element icq = (Element) userlist.get(11);
                Element skype = (Element) userlist.get(12);
                Element yahoo = (Element) userlist.get(13);
                Element aim = (Element) userlist.get(14);
                Element msn = (Element) userlist.get(15);
```

Figura 140. Importación de los usuarios a la Base de Datos

Clase Accion: Permitió extraer la información del archivo .xml a la base de datos de las acciones realizadas por los estudiantes en el curso como puede ser acciones en las tareas, examenes o recursos (ver Figura 141).

```
public class MOD {
    public static void main(String[] args) throws SQLException, ClassNotFoundException {
        Conexion conexion = new Conexion();
        Connection conMysql = conexion.getconnection();

        Statement st = conMysql.createStatement();
        st = conMysql.createStatement();

        SAXBuilder builder = new SAXBuilder();
        File xmlFile = new File("D:\\10 modulo 2014\\Anteproyectos\\ANTEPROYECTO\\DATOS\\moodle.xml");
        try {
            Document document = (Document) builder.build(xmlFile);
            Element rootNode = document.getRootElement();
            List moodle = rootNode.getChildren();
            Element info = (Element) moodle.get(0);
            List detailslist = info.getChildren();
            Element details = (Element) detailslist.get(9);
            List modlist = details.getChildren();
            for (int i = 0; i < 5; i++) {
                Element mod = (Element) modlist.get(i);
                List modlist1 = mod.getChildren();
                Element name = (Element) modlist1.get(0);
                Element included = (Element) modlist1.get(1);
                Element userinfo = (Element) modlist1.get(2);
                Element instances = (Element) modlist1.get(3);
                String query = "insert into mods "
                        + "values ('" + (i+1) + "','" + name.getValue() + "','"
                        + included.getValue() + "','" + userinfo.getValue() + "')";
                System.out.println(query);
                st.executeUpdate(query);
//              System.out.println("\nname: " + name.getValue() + "\tincluded: " + included.getValue(
                List instanceslist = instances.getChildren();
                for (int j = 0; j < instanceslist.size(); j++) {
                    Element instance = (Element) instanceslist.get(j);
                    List instancelist = instance.getChildren();
```

Figura 141. Importación a las acciones a la Base de Datos

Anexo G: SENTENCIAS SQL

Código sql utilizado para consultar en la base de datos acerca del número de estudiantes pertenecientes al género masculino (ver Figura 142) y femenino (ver Figura 143) en el curso.

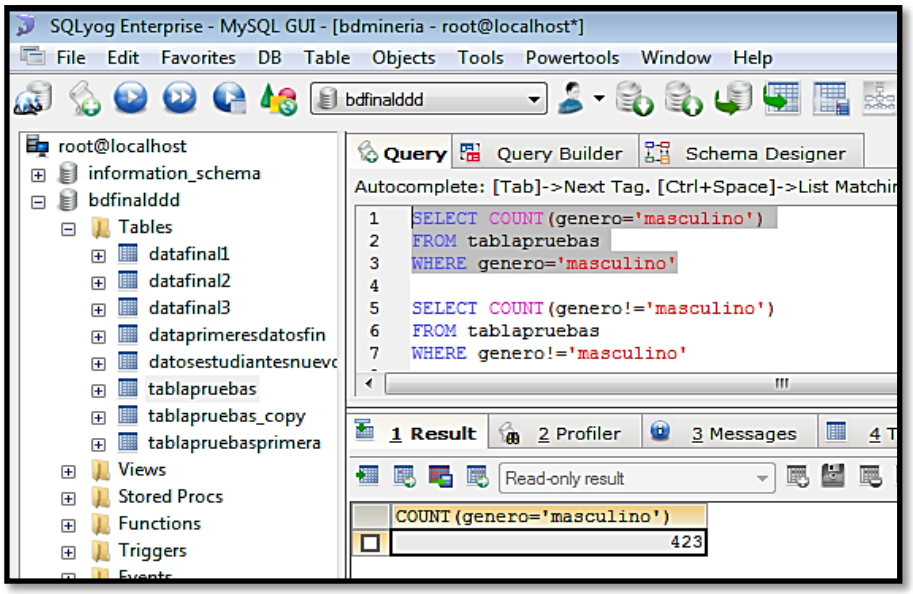


Figura 142: Consulta género masculino

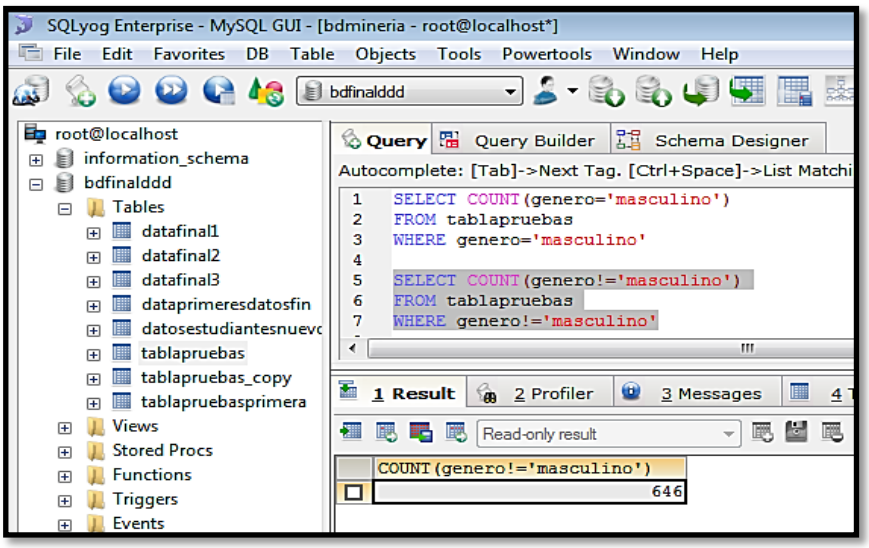


Figura 143: Consulta género femenino

Código sql utilizado para consultar en la base de datos acerca del número de estudiantes pertenecientes al rango de edades menores a 27 años (ver Figura 144), rango de edades comprendidas entre 27 a 37 años (ver Figura 145) y rango de edades de 38 a 48 años en el curso (ver Figura 146).

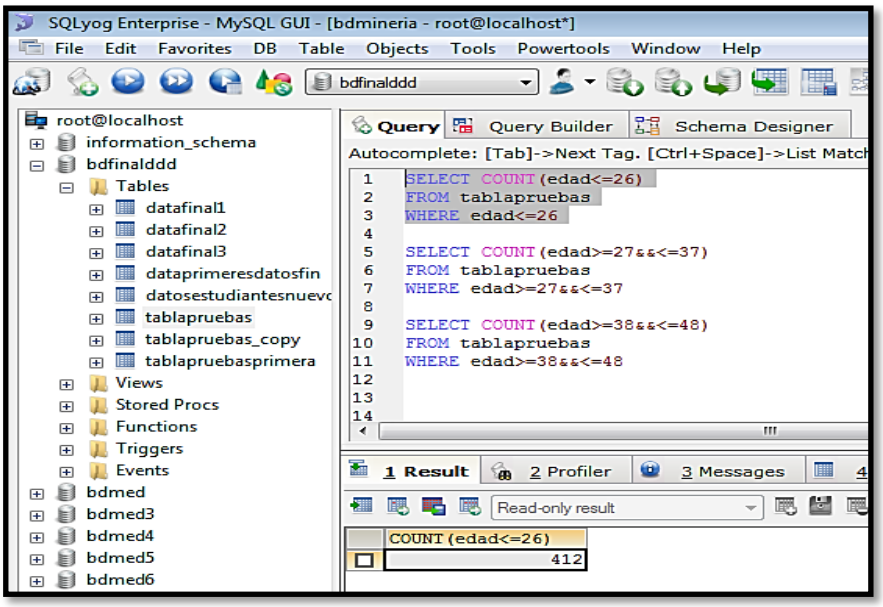


Figura 144: Consulta edad menores a 27 años

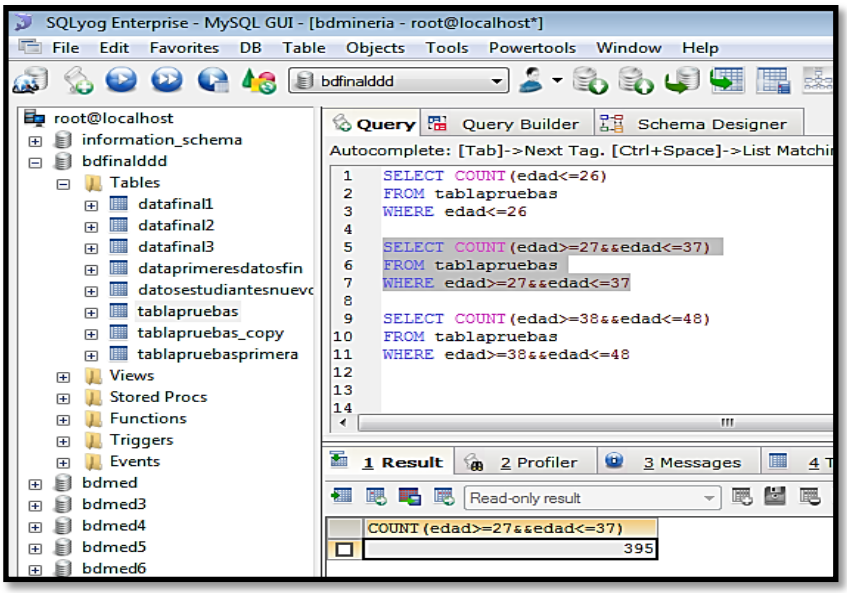


Figura 145: Consulta del rango de edades entre 27 a 37 años

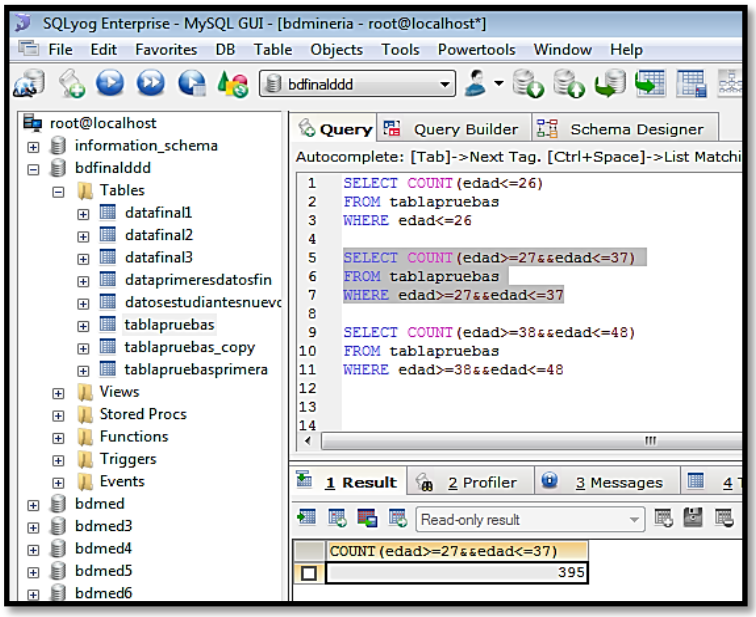


Figura 146: Consulta del rango de edades entre 38 a 48 años

Código sql utilizado para consultar en la base de datos acerca del número de estudiantes pertenecientes al estado civil casado (ver Figura 147), soltero (ver Figura 148), divorciado (ver Figura 149) y viudo (ver Figura 150) en el curso.

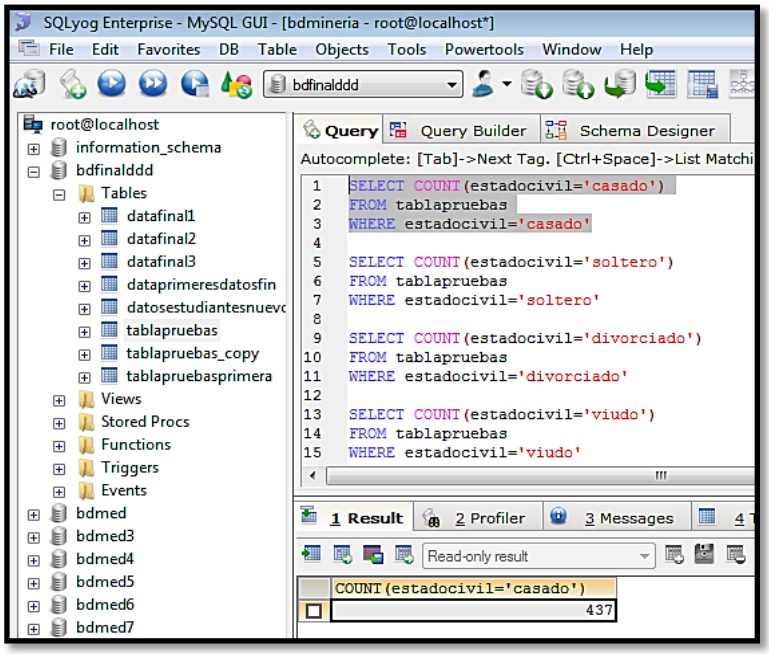


Figura 147: Consulta del estado civil casado

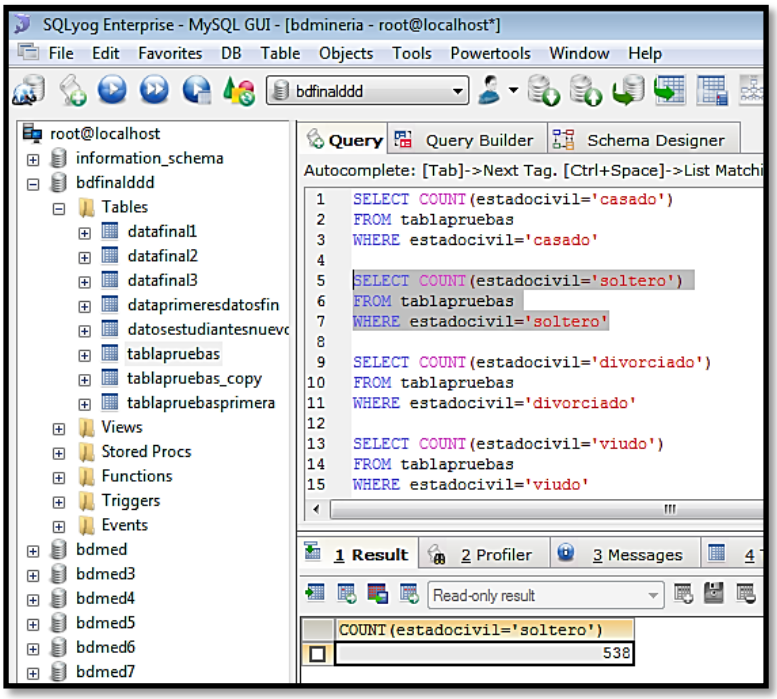


Figura 148: Consulta del estado civil soltero

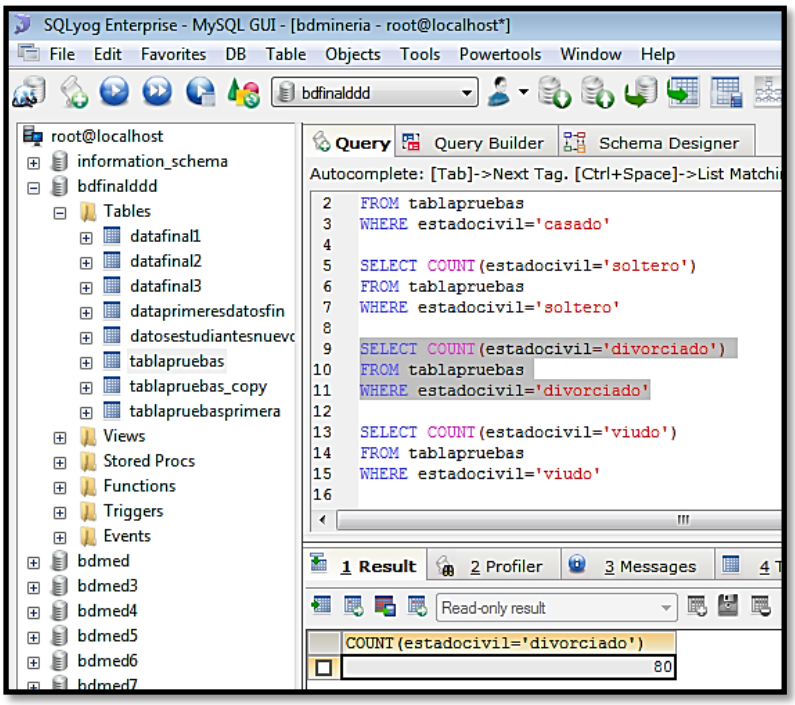


Figura 149: Consulta del estado civil divorciado

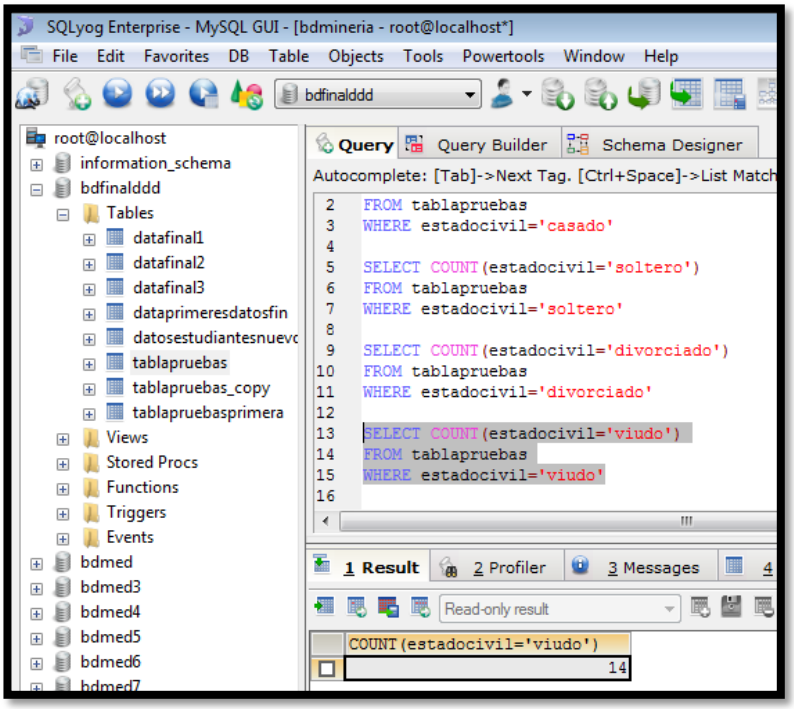


Figura 150: Consulta del estado civil viudo

Anexo H: Autorización de la utilización de los datos de los estudiantes de la Modalidad de Estudios a Distancia

Loja, 27 de Enero del 2015

Sr. Ing.

Milton Ricardo Palacios Morocho

DIRECTOR DEL DEPARTAMENTO DE TELECOMUNICACIONES E INFORMACIÓN - UTI

Ciudad.

De mis consideraciones:

Yo, Angélica Elizabeth Jaramillo Zhingre, portadora de la cédula de ciudadanía 1104999147, egresada de la Carrera Ingeniería en Sistemas me dirijo a usted muy comedidamente para solicitarle se me proporcione información que será detallada más adelante correspondiente a los estudiantes del curso Virtual de Ingles de la Modalidad de Estudios a Distancia, debido que son de suma importancia para poder desarrollar el Trabajo de Titulación denominado "Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje".

A continuación se detallan los datos a solicitar:

- Número de cédula del estudiante.
- Nombre de la carrera a la que pertenece el estudiante.
- Horario de clases del estudiante
- Número de hijos del estudiante.
- Situación laboral del estudiante.
- Situación económica del estudiante.

Con la certeza de ser atendido favorablemente me suscribo de usted con los más sinceros agradecimientos.

Atentamente:

Angélica Elizabeth Jaramillo Zhingre
Egresada
C.I 1104999147

Recibido 30-01-2015

ANEXO I: Artículo Científico

Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje

Angélica Jaramillo [a1], Henry Paz[a]

[a] Carrera de Ingeniería en Sistemas, Universidad Nacional de Loja, La Argelia, Loja, Ecuador
aejaramilloz@unl.edu.ec, hpaz@unl.edu.ec

Resumen. El presente artículo está enfocado en determinar las interacciones de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia (MED) de la Universidad Nacional de Loja, para ello se realizó un análisis a la base de datos donde se encontraban los datos de los estudiantes correspondiente al periodo académico 2013 - 2014, para seleccionar los atributos necesarios que permitieron generar un modelo de minería de datos. De igual forma se consideró realizar un estudio de las diferentes técnicas de minería de datos donde se seleccionó la que más se adapta al proyecto, eligiendo la técnica de clasificación para generar los modelos a través de los algoritmos pertenecientes a la misma, posteriormente se efectuó un análisis de las metodologías de minería de datos comparando cada una de ellas con el fin de seleccionar la que ayude al desarrollo del proyecto eligiendo la metodología Crisp-dm porque contiene múltiples fases indicando cada una de las actividades que se deben cumplir, convirtiéndose de esta forma en una guía práctica. Además se desarrolló un análisis comparativo tomando en cuenta características de las herramientas de minería de datos donde se seleccionó RapidMiner para realizar los procesos mediante algunos algoritmos conjuntamente con los datos de los estudiantes los mismos que se dividieron en dos conjuntos, para entrenamiento y validación, obteniendo como resultado que el mejor algoritmo fue el decisión tree, ya que clasificó las instancias correctamente con un margen de error mínimo, así mismo presentó un árbol con los diferentes atributos dando las mejores reglas de las interacciones de los estudiantes, de tal forma se pudo generar el modelo mediante el cual se determinó que en gran mayoría los estudiantes tienen un nivel de interacción medio en el curso virtual de inglés, donde los factores que más influyen son las interacciones en las tareas, exámenes, recursos, situación laboral y estado civil del estudiante.

Palabras Clave: minería de datos, técnica de clasificación, modelo, algoritmo, metodología.

1 Introducción

La educación es la base del progreso de cualquier país, por ello en la actualidad los sistemas educativos de todo el mundo se enfrentan al desafío de utilizar las tecnologías de la información, teniendo un papel importante porque facilitan el aprendizaje en entornos virtuales prestando a los estudiantes la adquisición del conocimiento en forma inmediata y amplia, sin que la distancia ni el tiempo sea un inconveniente en su formación académica [1].

Al respecto la Universidad Nacional de Loja cuenta con sistemas de información para brindar la facilidad de estudios a distancia ya sea en distintas carreras o cursos, estos sistemas almacenan grandes cantidades de información de los estudiantes como es el caso del curso virtual de inglés de la modalidad de estudios a distancia, el mismo que se ha tomado como objeto de estudio, pero tener numerosa información a disposición y no saber qué hacer con ella es un gran problema, es aquí donde interviene la minería de datos que contiene un conjunto de técnicas que se aplican para extraer conocimiento útil y comprensible, previamente desconocido, así mismo descubrir patrones para generar un modelo a través del análisis de la información de las interacciones en el curso, datos personales, institucionales y socioeconómicos del estudiante, que permitió determinar las interacciones de los estudiantes en el curso virtual, para que de esta manera ayude a la toma de decisiones, y por tanto un beneficio a la institución [2].

Crisp-dm es una metodología de las más usadas en la actualidad para la generación de proyectos de minería de datos, se obtuvo un modelo de análisis de datos, conjuntamente con la implementación de algoritmos de inteligencia artificial, ya incorporados en la herramienta de pre-procesamiento de datos RapidMiner [3].

El artículo está estructurado de la siguiente forma resumen, introducción, materiales y métodos, estado del arte, resultados alcanzados, conclusiones y referencias bibliográficas.

2 Materiales y Métodos

Para el desarrollo del artículo fue necesario el empleo de algunos métodos y metodología para ello se realizó una búsqueda de información bibliográfica y algunos casos de éxito.

En la recolección y organización de la información que se obtuvo se utilizó los siguientes métodos y técnicas:

Estudio de Casos: Sirvió para obtener un conocimiento más amplio de los casos reales actuales, los cuales ayudaron para tener una idea clara del problema, así mismo permitió realizar una exploración e investigación en profundidad de problemas específicos.

Revisión bibliográfica: Se sustentó la base teórica de la realización del proyecto, mediante consultas en fuentes bibliográficas, textos, artículos científicos, libros, tesis de grado y casos de éxito.

Dando como resultado el desarrollo del estado del arte en donde se menciona en que consiste la minería de datos, técnicas de minería de datos, herramientas que son utilizables dentro de la minería de datos.

Metodología: Se utilizó la metodología Crisp-dm que es muy importante porque contiene etapas las cuales están compuestas por actividades o una secuencia de pasos ordenados.

3 Estado del Arte

Dentro del estado del arte se han considerado temas relevantes concernientes acerca de la minería de datos.

3.1 Minería de Datos

Proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, para encontrar modelos a partir de los datos, para que este proceso sea efectivo, debería ser automático y el uso de los patrones descubiertos debería ayudar a la toma de decisiones, y por tanto, un beneficio a la organización, convertir datos en conocimiento [2], [3].

3.2 Técnicas de Minería de Datos

Las tareas de la minería de datos son [2]:

Técnicas supervisadas o predictivas: Utilizar algunas variables o campos en una base de datos para predecir valores desconocidos o futuros de tal manera que especifican el modelo para los datos en base a un conocimiento previo [2].

Técnicas no supervisadas o descriptivas: Encontrar patrones que describan la información que puedan ser interpretadas.

Las técnicas de minería de datos serán utilizadas con el objetivo de obtener la información oculta en grandes cantidades de datos, las cuales son descritas a continuación [2].

Agrupamiento: Se agrupan datos dentro de un número de clases, se puede realizar mediante criterios de distancia o similitud, de forma que si las clases son similares entre sí estén agrupadas, la agrupación o clustering consiste en agrupar un conjunto de datos basándose en la similitud de los valores de sus atributos [6].

Clasificación: Árboles de decisión: Son estructuras que representan conjuntos de decisiones que generan reglas para la clasificación de un conjunto de datos. Entre los algoritmos que aplica es el J48, ID3, entre otros [7].

Son útiles para explorar un conjunto de datos y entender cómo ciertas variables de las interacciones de los estudiantes con el entorno virtual de aprendizaje inciden sobre otra [8].

Permite una organización eficiente de un conjunto de datos, debido a que los árboles son construidos a partir de la evaluación del primer nodo raíz y de acuerdo a su evaluación o valor tomado se va descendiendo en las ramas hasta llegar al final del camino u hojas del árbol [8].

Reglas de asociación: En minería de datos las reglas de asociación en base de datos se evalúan de acuerdo al soporte y a la confianza de las mismas, se utilizan para encontrar hechos que ocurren en común dentro de un conjunto de datos. Dicho de otra manera deben ocurrir ciertas condiciones para que se produzca cierta condición, también para buscar por medio de conjunto de datos reglas que revelan la naturaleza de las relaciones o asociaciones entre datos de las entidades [9], [10].

Se aplican en el análisis de la canasta de mercado, marketing cruzado con correo, diseño de catálogos, segmentación de clientes respecto a las compras y el soporte para la toma de decisiones [9].

3.3 Herramientas de Minería de Datos

SAS Enterprise Miner: Es una herramienta de minería de datos comercializada, crea modelos predictivos y descriptivos precisos sobre grandes volúmenes de datos a través de diferentes fuentes mediante un proceso transparente, lo que permite colaborar de manera más eficiente, incluye una interfaz de usuario intuitiva que incorpora los principios de diseño comunes establecidos para el software de SAS y herramientas de navegación adicionales para mover fácilmente alrededor del área de trabajo [11],[12].

RapidMiner: Es una herramienta de minería de datos desarrollado en Java, permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de 500 operadores a través de un entorno gráfico, permite utilizar los algoritmos incluidos en weka, contiene técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos, aprendizaje automático [9],[13].

Weka: Es una herramienta para el aprendizaje automático y minería de datos diseñado en Java, es de distribución de licencia GNU-GLP, contiene una colección de algoritmos para el análisis de datos y modelado predictivo, permite la visualización de datos, provee una interfaz gráfica [14].

Este programa se desarrolló en Java y dispone de tres entornos de trabajo gráficos y un entorno en modo co nsola, permitiendo la implementación de algoritmos para preprocesamiento de datos, clasificación, clustering, selección de atributos, reglas de asociación [15].

Knime: Está desarrollado sobre la plataforma Eclipse y programado en Java, su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos y predecir posibles resultados [11], [16].

Es una plataforma de código abierto de fácil uso y comprensible para integración de datos, procesamiento, análisis, y exploración. Ofrece a los usuarios la capacidad de crear de forma visual flujos de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas [16].

3.4 Metodología de Minería de Datos

Se ha considera el estudio de dos metodologías de minería de datos como lo es Semma, Crisp-dm, ya que estas son las más utilizadas en distintos proyectos.

3.4.1 SEMMA

Creada por el SAS Institute, se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". El nombre de esta terminología corresponde a las cinco fases básicas del proceso: sample (muestreo), explore (exploración), modify (modificación), model (modelado), assess (valoración) [17].

Se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando evidenciando que el modelo está orientado especialmente a aspectos técnicos [18].

3.4.2 Crisp-dm

Es una metodología estándar para la construcción de proyectos de minería de datos con sus fases no necesariamente rígidas [19].

Puede ser integrada con una metodología de gestión de proyectos específica que complemente las tareas administrativas y técnicas, además es de libre distribución [19].

Organiza el desarrollo de un proyecto de minería de datos en una serie de fases o etapas que funcionan de manera cíclica e iterativa, cada una cuenta con tareas generales y específicas que permitan cumplir con los objetivos del proyecto [20], [21], [22].

3.4.2.1 Fases de Crisp-dm

Comprensión del negocio: Comprender o definir el problema del negocio, lo cual es quizás el paso más importante de la metodología, permite entender los objetivos y requisitos que tendrá el proyecto [19].

Las tareas de esta fase es el establecimiento de los objetivos de negocio, evaluación de la situación mediante el inventario de recursos, requerimientos, suposiciones, restricciones, riesgos, contingencias, terminología, costes y beneficios, establecimiento de los objetivos de minería de datos, generación del plan del proyecto y evaluación inicial de herramientas y técnicas [19], [20], [21].

Comprensión de los datos: Comprende la búsqueda de la información y de las variables que se utilizarán para la generación de los indicadores del proceso a los cuales se aplicara minería de datos, contiene algunas tareas como es la recolección de datos, teniendo claro desde qué lugar fueron obtenidos. Descripción de los datos, estableciendo los volúmenes de información con que se trabajará, la cantidad de registros, y los significados de cada campo o variable y los formatos en los que se encuentran. Exploración de los datos, indicando una estructura general de la información, comprobar frecuencia y distribución de los datos, verificación de la calidad de los datos, determinando la consistencia de los valores, comprobando la existencia de datos nulos y fuera de rango, identificando irregularidades para asegurar la completitud y exactitud de los datos [20],[21].

Preparación de los datos: Preparación de los datos para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente, consta de algunas tareas como la selección de datos escogiendo un subconjunto de los datos recopilados en la etapa anterior. Limpieza de los datos, preparándolos para la fase de modelación, ya sea aplicando técnicas de normalización, discretización de campos numéricos, tratamiento de valores nulos, entre otros [20].

Estructuración de los datos con lo cual se pueden generar nuevos atributos a partir de los existentes o transformar valores de los atributos con que se cuenta. Integración de los datos, agrupar tablas o campos que se encuentren relacionadas, definiendo una estructura que las pueda contener. Formateo de los datos, transformar los datos sin modificar su significado, para que se puedan ajustar a las técnicas de minería de datos que se utilice [20], [22].

Modelado: Se elige las técnicas de modelado que sean más apropiadas para resolver el problema, aplica algunas tareas que son en base al objetivo principal del proyecto [19].

Generación del plan de prueba, diseñando un procedimiento para probar y validar el modelo. En general, se separa el conjunto de datos en dos: una parte de los datos destinada a entrenamiento del modelo y otra parte que será utilizada para las pruebas [21].

Construcción del modelo a partir de la técnica de modelado seleccionada, se aplica sobre el conjunto de datos para generar uno o más modelos. En este punto se van ajustando los parámetros de la técnica seleccionada de forma iterativa para obtener mejores resultados. Evaluación del modelo, interpretando los modelos en base al conocimiento existente y los criterios de éxito ya establecidos [19].

Evaluación: Se evalúa el modelo en base al cumplimiento de los criterios de éxito del problema, revisar el proceso seguido teniendo en cuenta los resultados obtenidos, para poder repetir algún proceso en el que a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase [21].

Se evalúa el grado en el cual el modelo satisface los objetivos del negocio y busca determinar si hay alguna razón del negocio del porque el modelo sería deficiente [22].

4 Resultados Alcanzados

A lo largo del desarrollo del proyecto se ha considerado cada una de las etapas de la metodología llegado a los siguientes resultados.

4.1 Comparación de las Metodologías de Minería de Datos

Algunos modelos profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como Crisp-dm), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA) [16], [19].

SEMMA inicia el proyecto de minería a partir del conjunto de datos (la primera fase es el muestreo de los datos). Crisp-dm y KDD comienzan con un análisis del negocio y del problema organizacional. Catalyst considera cinco escenarios posibles como punto de partida, entre los cuales se encuentra el inicio desde un problema u oportunidad de negocio [16], [20].

KDD, Crisp-dm y Catalyst contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería. SEMMA excluye esta actividad del modelo [17], [21].

En todos los modelos se contempla la selección y preparación de los datos esta situación se repite para la fase de modelado, donde se aplican las técnicas de minería para obtener los nuevos patrones [17].

La implementación de los resultados obtenidos es una fase que no está incluida en el modelo SEMMA. En Crisp-dm, se propone además una planificación para el control futuro y un análisis de cierre del proyecto [19], [20].

Los modelos Crisp-dm y Catalyst cuentan con un nivel de detalle con el que describen las tareas en cada fase del proceso, y porque incorporan actividades para la gestión del proyecto [19], [21].

La metodologías para la gestión de un proyecto de minería de datos, el modelo a tener en cuenta debería ser Crisp-dm ya que posee importantes características de las ya se hablado anteriormente [20], [22].

SEMMA y Crisp-dm comparten la misma esencia, estructurando el proyecto de explotación de datos en fases que se encuentran interrelacionadas entre sí [19].

SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Crisp-dm ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de explotación de datos siendo su distribución libre y gratuita [22].

4.1.1 Elección de la Metodología

La metodología a utilizar es Crisp-dm ya que cada una de sus fases se encuentra claramente estructurada definiendo de tal forma las actividades y tareas que se requieren para lograr el objetivo planteado es decir es la más completa entre las metodologías comparadas, es flexible por ende se puede hacer usos de cualquier herramienta de minería de datos.

4.2 Desarrollo de la Metodología CRISP-DM

En el presente artículo se describe la aplicación de técnicas de minería de datos conjuntamente con la herramienta RapidMiner para determinar las interacciones de los estudiantes en el curso virtual de inglés, para lo cual fue necesario aplicarlo en un escenario real con datos personales, institucionales, socioeconómicos y las interacciones de los estudiantes de la modalidad de estudios a distancia, para ello se empleó la metodología Crisp-dm como una guía que permita desarrollar el proyecto, a continuación cada una de sus fases:

4.2.1 Compresión del Negocio

Esta fase se centra en comprender los objetivos y los requerimientos del proyecto desde una perspectiva del negocio, y luego en convertir este conocimiento en la definición de un problema de minería de datos y en un plan preliminar designado para alcanzar los objetivos.

4.2.1.1 Objetivos del Negocio

Investigar sobre las diversas técnicas de minería de datos que permitan determinar la interacción de los estudiantes en un entorno virtual de aprendizaje.

Diseñar un modelo computacional aplicando técnicas de minería de datos para determinar la interacción de los estudiantes en un entorno virtual de aprendizaje.

Evaluar el modelo computacional en un escenario real a través de los datos de interacción de los estudiantes en un entorno virtual de aprendizaje.

4.2.1.2 Requerimientos

Tener la información suficiente de las interacciones de los estudiantes para la obtención del modelo.

Seleccionar técnicas de minería de datos adecuadas al problema a resolver.

Disponer de herramientas de minería de datos para la realización del modelo.

Contar con la colaboración continua por parte del director de tesis.

Tener apoyo del Director de la MED para obtener información exacta de los procesos que se utilizan en el curso virtual.

4.2.2 Comprensión de los Datos

En esta etapa se recolectó los datos relacionados con las interacciones de los estudiantes para una mejor comprensión de los mismos, de manera que es el primer acercamiento que se tiene para posteriormente realizar el análisis y de esta manera identificar algún inconveniente que exista, de tal forma que se analizó la estructura de los datos mediante consultas ejecutadas en la base de datos.

4.2.2.1 Recolección de Datos Iniciales

Los datos recolectados pertenecen a las interacciones de los estudiantes del curso virtual de inglés de la Modalidad de Estudios a Distancia perteneciente a la Universidad Nacional de Loja del periodo académico 2013 - 2014, para ello se descargó la información de la plataforma moodle perteneciente a la MED, entre la información con la que se trabajó se tiene las interacciones de los estudiantes según las actividades que desarrollaron en el curso, las mismas que se describen a continuación:
Interacción con los archivos compartidos con las temáticas del curso.
Realizaron tareas y evaluaciones, para la aprobación del curso.
Leer o imprimir los contenidos y actividades del curso.
Enviar las actividades al docente para su corrección y recibir sus calificaciones.
Evaluaciones On-Line y calificaciones.

De esta manera se ofrece a los estudiantes en formación la oportunidad de reforzar el aprendizaje brindando a través de contenidos y evaluaciones, para posibilitar la interacción estudiante-profesor y estudiante-herramientas.

Los datos se encuentran estructurados en archivos XML que consta de las interacciones, datos personales, institucionales y socioeconómicos de los estudiantes del curso, como son: número de accesos al curso, número de accesos a las tareas, número de veces que accede a un recurso, numero de accesos a exámenes, descripción de los módulos, datos personales de los estudiantes, datos socioeconómicos e institucionales de los estudiantes.

En la siguiente figura (ver Figura 1) se puede observar las tablas que conforman la base de datos donde esta almacenada la información de las interacciones de cada uno de los estudiantes de la MED:

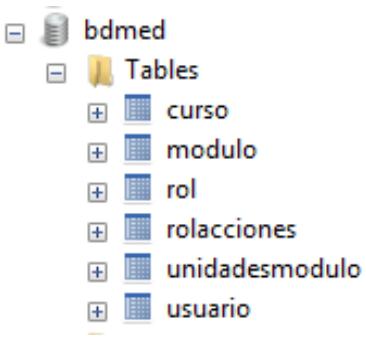


Fig. 1. Base de datos

4.2.3 Preparación de los Datos

La preparación de los datos se realizó para trabajar con las técnicas de minería de datos seleccionada, consta de algunas tareas como la selección de datos donde se eligió una estructura de datos recopilados en la etapa anterior, por otra parte se realizó la limpieza de los datos para poder generar el modelo de minería de datos, de tal manera que no debe contener valores nulos que permitieron obtener mejores resultados.

Además se realizó la estructura de los datos con lo cual se generaron nuevos atributos a partir de los existentes, así mismo se desarrolló la integración de los datos que consistió en agrupar tablas o campos que se

encuentren relacionadas, finalmente se hizo el formateo de los datos que consiste en transformar los datos sin modificar su significado, para que se puedan ajustar a las técnicas de minería de datos.

Por lo tanto el objetivo de la presente etapa fue generar la estructura de datos final, con el propósito de obtener el modelo a través de técnicas de minería de datos.

4.2.3.1 Estructura de los Datos

En esta tarea se realizó la construcción del data set final o estructura de datos que es útil para poder desarrollar el modelo computacional, donde se ha tomado en cuenta información personal, socioeconómicas y las interacciones de los estudiantes, para lo cual se trabajó con los siguientes campos para cumplir con el objetivo planteado, los mismos se pueden observar en la siguiente tabla (ver Tabla 1):

Tabla 1. Atributos de minería de datos para determinar las interacciones de los estudiantes.

Atributo	Tipo de Datos	Categorización
cedula	Nominal	
interaccionesrecurso	Int	- IRB - IRM - IRA
interaccionesexamen	Int	- IEB - IEM - IEA
interaccionestareas	Int	- ITB - ITM - ITA
numerointeracciones	Int	- Bajo - medio - alto
servicios	Nominal	- 1 - 2 - 3
ciudad	Nominal	- L - O
edad	Int	- a - b - c
genero	Nominal	- 0 - 1
estadocivil	Nominal	- S - C - D - V
carrera	Nominal	- Administración Pública - Artes Plásticas - Banca Y Finanzas - Bibliotecología e Información Cientifico - Comunicación Social - Contabilidad Y Auditoría

		- Cultura Física Y Deportes - Derecho - Diseño de Interiores Y Decoración - Economía - Educación Básica - Educación Musical - Enfermería - Físico Matemáticas - Informática Educativa - Ingeniería Agrícola - Ingeniería Agronómica - Ingeniería en Administración Y Producción - Ingeniería en Electrónica Y Telecomunicaciones - Ingeniería en Electromecánica - Ingeniería en Geología Ambiental - Ingeniería en Manejo Y Conservación - Ingeniería en Sistemas - Ingeniería Forestal - Laboratorio Clínico - Lengua Castellana Y Literatura - Medicina Humana - Medicina Veterinaria Y Zootecnia - Odontología - Psicología Clínica - Psicología Educativa Y Orientación - Psicología Infantil Y Educación Parvularia - Psicorrehabilitacion Y Educación Especial - Químico Biológicas - Radiología e Imagen Diagnóstica - Trabajo Social
numeroHijos	Nominal	- Si - No
trabajo	Nominal	- Si - No

4.2.4 Modelado

En la presente etapa se realizó el modelado de minería de datos para ello se empleó el data set final de la etapa anterior (ver Tabla 1) conjuntamente con la técnica seleccionada los algoritmos que pertenecen a dicha técnica y finalmente la herramienta elegida de minería de datos para realizar los distintos procedimientos con la finalidad de obtener los resultados del modelo.

4.2.4.1 Análisis de las Diversas Técnicas de Minería de Datos

Para el análisis de cada una de las técnicas de minería de datos se ha estudiado algunos casos de éxito donde estas han sido empleadas, obteniendo resultados confiables, además se realizó un análisis comparativo el cual se puede observar en la siguiente tabla (ver Tabla 2), cabe mencionar que dentro de cada una de las técnicas de minería de datos, estas constan con diferentes algoritmos que se pueden utilizar.

Tabla 2. Técnicas de minería de datos para determinar las interacciones de los estudiantes.

Técnica	Algoritmo	Casos de éxito
Agrupamiento O clustering	Simple-Kmeans	- Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje [23]. El estudio fue desarrollado con 30 estudiantes del primer ciclo de la carrera de Ingeniería en Sistemas Informáticos y Computación de la UTPL, con la técnica del clustering se procedió a agrupar a los estudiantes del curso en diferentes grupos de acuerdo a su nivel de participación y semejanza de acceso a la plataforma, para descubrir patrones que reflejen comportamientos análogos en los estudiantes, con la ayuda del algoritmo SimpleKMeans se realizaron tres experimentaciones y también la construcción del modelo [23].
Clasificación	J48 REPTree ID3/C4.5	- Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de la UTPL [7]. Una de las técnicas que se empleará en la minería para determinar el comportamiento de los estudiantes en base a las acciones que éste realiza sobre el EVA, es la de clasificación, como son los árboles de decisión, reglas de decisión, éstos se utilizan para el indicador de la participación del estudiante en el curso ya que según lo analizado ayudan a predecir una o más variables discretas, basándose en otros atributos del conjunto de datos, el algoritmo hace predicciones. Los algoritmos utilizados fueron C4.5 o J48 y JRip para la experimentación [7].

4.2.4.2 Determinar la Técnica de Minería de Datos

Para la selección de la técnica de minería de datos a utilizarla posteriormente se lo ha realizado mediante un cuadro comparativo que se lo puede observar en la Tabla 2, en donde se describe cada caso de éxito donde han alcanzado resultados idóneos, así mismo se ha podido identificar la técnica apropiada para el proyecto

que es la clasificación para el análisis de las interacciones de los estudiantes en el entorno virtual de aprendizaje ya que es apropiada al problema a resolver y se tiene una comprensión de la misma.

Por otra parte se ha podido identificar una gran mayoría de casos de éxito relacionados con el estudio de entornos virtuales con diferentes fines en donde la técnica más aplicada es la antes mencionada, siendo esta eficiente en el momento de analizar grandes cantidades de datos y posteriormente el desarrollo de un modelo para que mediante este se pueda tomar decisiones y mejorar el uso de los entornos virtuales de aprendizaje.

4.2.4.3 Cuadro Comparativo de Herramientas de Minería de Datos

Se realizó una recolección de información referente a las herramientas de minería de datos en los cuales sobresalen cada una de las características de las herramientas más utilizadas en proyectos similares según el estudio de algunos casos de éxito, entre las que se ha seleccionado se tiene SAS Enterprise Miner, RapidMiner, Weka, Knime, mismas que se puede observar a continuación:

Tabla 3. Herramientas de minería de datos [12], [13], [14], [15], [16], [17], [18].

CARACTERÍSTICAS	HERRAMIENTAS			
	SAS Enterprise Miner	RapidMiner	Weka	Knime
Licencia libre	X	✓	✓	✓
Multiplataforma	✓	✓	✓	✓
Puede combinar modelos	✓	✓	✓	X
Técnicas Descriptivas(agrupación)	✓	✓	✓	✓
Técnicas Predictivas(clasificación, otras)	✓	✓	✓	X
Interfaz amigable	✓	✓	X	X
Permite visualización de datos	✓	✓	✓	✓
Flexibilidad	X	✓	✓	✓
Fácil de Configurar	X	✓	✓	X
Fácil de Instalar	✓	✓	✓	✓
Conversión de datos	✓	✓	✓	X
Filtros	✓	✓	✓	X
Dispone de un módulo de integración con Weka y R	X	✓	X	X
Procesamiento de datos	✓	✓	✓	✓
Validación del modelo	✓	✓	✓	X

En la tabla anterior (ver Tabla 3.) se han mencionado algunas características que han sido tomadas en cuenta para la comparación de las mismas y posteriormente seleccionar según cumpla cada herramienta con dicha característica, donde se eligió la que más se adaptó al trabajo realizado.

4.2.4.4 Selección de la Herramienta de Minería de Datos

Al realizar el análisis de las herramientas seleccionadas anteriormente (ver Tabla 3.) en base a algunas de sus características se pudo seleccionar a la herramienta RapidMiner para llevar a cabo cada una de las actividades acerca del modelado del presente proyecto, ya que se adapta al trabajo de titulación, posee una licencia libre, combinación de modelos, interfaz amiga, multiplataforma, empleo de técnicas, además permite aplicar varios algoritmos de minería de datos, con la integración de algunos complementos se puede utilizar los algoritmos incluidos en Weka, para poder generar el modelo y validarlo.

Así mismo se puede mencionar que no existe una herramienta que contenga todas las funcionalidades pero RapidMiner es ampliamente usada y probada a nivel internacional en aplicaciones empresariales, de gobierno y academia, posee gran cantidad de operadores que permiten generar el modelo para determinar las interacciones en entornos virtuales.

4.2.4.5 Generar el Plan de Prueba

La generación del plan de pruebas consiste en probar la calidad y validez de los resultados obtenidos por el modelo**,** por ende es necesario generar un plan de pruebas mediante el cual se pueda probar la validez del modelo generado, para ello se trabajó con los datos pertenecientes a los estudiantes de la MED los cuales se los dividió en dos grupos uno para entrenamiento y el otro para emplearlo en la validación del modelo.

El conjunto de datos para entrenamiento es el 67% y el conjunto de datos restantes se los utilizó para realizar la validación de tal manera que da un resultado del 100% de datos utilizados para el modelado.

A continuación, se describe el plan de pruebas realizado con los diferentes algoritmos clasificados de la siguiente forma:

Algoritmos de Reglas de decisión: Los algoritmos utilizados dentro de esta clasificación corresponden a JRip, Ridor, Prism, K-NN, en donde se utilizó el 67% del conjunto de datos para entrenamiento y el 33% para la validación.

Algoritmos de Arboles de decisión: Los algoritmos utilizados dentro de esta clasificación corresponden a CHAID, Decision Tree, ID3, J48, en donde se utilizó el 67% del conjunto de datos para entrenamiento (E) y el 33% para la validación (V).

Entre los parámetros que se tomó en cuenta para evaluar los modelos generados son los siguientes: instancias clasificadas correctamente (accuracy), instancias clasificadas incorrectamente (clasification_error), estadística de Kappa que mide la coincidencia de la predicción con la clase real (Kappa), error cuadrático (squared_error), error relativo (relative_error), error absoluto (absolute_error), presentando los resultados obtenidos en la siguiente tabla (ver Tabla 4 y Tabla 5):

Tabla 4. Resultados de Algoritmos.

Algoritmo	Datos	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de Kappa
DECISION TREE	E	87.71	12.29	0.73
	V	92.90	7.10	0.82
JRip	E	94.41	5.59	0.88
	V	92.63	7.37	0.82
RIDOR	E	89.66	10.34	0.77
	V	88.66	11.34	0.74
K-NN	E	98.74	1.26	0.97
	V	84.15	15.85	0.64
PRISM	E	98.46	1.54	0.97
	V	77.06	22.94	0.58
CHAID	E	91.06	8.94	0.81
	V	73.58	26.42	0.41
ID3	E	98.32	1.68	0.96
	V	82.10	17.90	0.59
J48	E	91.06	8.94	0.80
	V	91.49	8.51	0.15

Tabla 5. Resultados de Algoritmos.

Algoritmo	Datos	Error Cuadrático	Error Relativo (%)	Error Cuadrático Medio	Error Cuadrático Relativo
DECISION TREE	E	0.11	21.44	0.33	1.18
	V	0.07	13.33	0.25	4.29
JRip	E	0.05	10.18	0.23	0.80
	V	0.07	13.17	0.25	0.98
RIDOR	E	0.10	10.34	0.32	1.14
	V	0.11	11.34	0.31	1.23
K-NN	E	0.01	1.26	0.11	0.39
	V	0.16	15.85	0.39	1.51
PRISM	E	0.02	1.54	0.12	0.46
	V	0.26	26.06	0.50	2.37
CHAID	E	0.06	12.41	0.25	0.89
	V	0.17	27.35	0.41	5.46
ID3	E	0.01	1.97	0.09	0.39
	V	0.13	13.03	0.36	0.35
J48	E	0.08	16.07	0.29	1.01
	V	0.08	14.50	0.27	1.06

En las tablas anteriores (ver Tabla 4 y Tabla 5) se puede observar el resultado de cada algoritmo obtenido mediante la utilización de la herramienta RapidMiner conjuntamente con los datos de los estudiantes del curso virtual inglés de la MED, donde existe un porcentaje mínimo de error de clasificación en cada uno de los algoritmos, además se puede indicar que con el conjunto de entrenamiento de los datos la mayoría de los resultados obtenidos de los algoritmos son favorables es decir que sobrepasan el 90% de los datos han sido clasificados correctamente, los algoritmos que presentan mejores resultados se tiene el JRip 94.41%, K-NN 98.74%, Prism 98.46%, Chaid 91.06%, ID3 98.32 y el J48 91.06% . Así mismo con el conjunto de datos de validación los algoritmos que presentan los mejores resultados de datos clasificados correctamente se tiene el Decision Tree 92.90%, Jrip 92.63% y J48 91.49%.

Con el conjunto de datos utilizados en el entrenamiento el mejor algoritmo que presenta es el K-NN con un porcentaje de 98.74% es decir las instancias se han clasificado correctamente y un 1.26% de las instancias clasificadas incorrectamente, el índice de kappa 0.973, el error cuadrático 0.013, error relativo 1.26, error cuadrático medio 0.11 y finalmente el error cuadrático relativo con el 0.397; y con el conjunto de datos utilizada en la validación el algoritmo que ha arrojado el mejor resultado es el Decision Tree con el 92.90% que significa que dicho porcentaje es el de las instancias clasificadas correctamente, el 7.10% es el de las instancias clasificadas incorrectamente, el índice de kappa 0.823, el error cuadrático 0.068, error relativo 13.33, error cuadrático medio 0.253 y finalmente el error cuadrático relativo con el 4.285.

4.2.5 Evaluación

En la presente etapa se realizó la evaluación del modelo para determinar si los datos obtenidos cumplen con el problema planteado, de tal forma que mediante la generación del modelo a través de los atributos (ver Tabla 1) se puede determinar las interacciones de los estudiantes del curso virtual de inglés de la modalidad de estudios a distancia.

En la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos ya sea tanto en el entrenamiento como en la validación en donde se muestra las instancias clasificadas correctamente (ver Fig. 2).

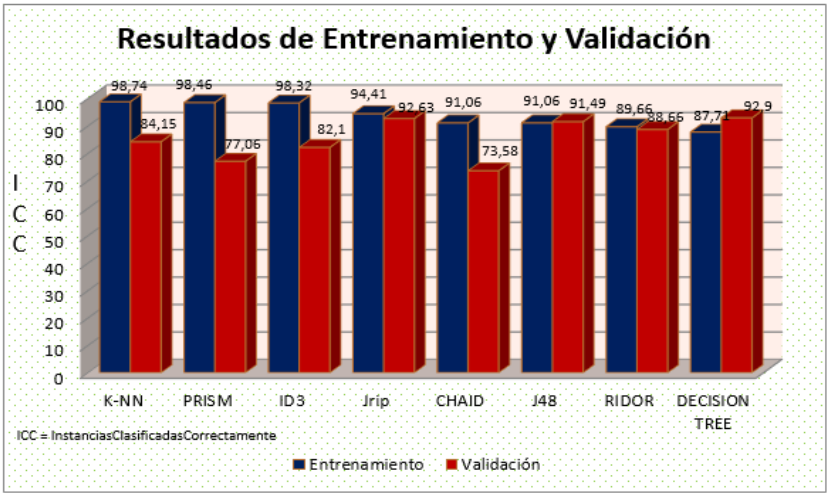


Fig. 2. Resultados por cada algoritmo

Así mismo en la siguiente figura se indican los resultados obtenidos de cada uno de los algoritmos en lo que respecta en la evaluación en donde se muestra las instancias clasificadas correctamente y las instancias clasificadas incorrectamente, donde se puede mencionar que el algoritmo que tiene mayor porcentaje es el Decision Tree con el 92,90% con un margen de error del 7,10%, el mismo se muestra a continuación (ver Fig. 3).

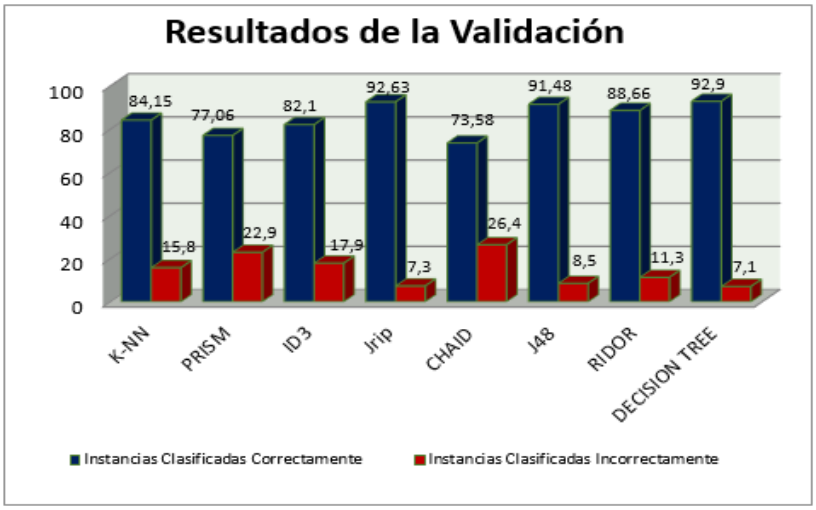


Fig. 3. Resultados de algoritmos de instancias clasificadas correcta e incorrectamente

Se realizó un análisis de los resultados obtenidos en la Minería de Datos, mediante una evaluación de los modelos obtenidos a través de la implementación de los algoritmos de la técnica de clasificación, los cuales fueron analizados en la fase anterior, para ello se utilizó datos de los estudiantes del curso virtual de inglés de la modalidad de estudios a distancia como son el número de accesos al curso, número de accesos a las tareas, número de veces que accede a un recurso, numero de accesos a exámenes, datos personales de los estudiantes, datos socioeconómicos, datos institucionales.

4.2.5.1 Determinar las Interacciones de los Estudiantes del Curso Virtual de Inglés Mediante Técnicas de Minería de Datos

Para poder determinar las interacciones de los estudiantes, se tomó en cuenta el mejor resultado de los algoritmos utilizados, que fueron analizados en la evaluación del modelo (Tabla 4 y Tabla 5 y Fig. 3) obteniendo el mejor resultado el algoritmo Decision Tree el cual presenta una buena clasificación con un 92.9% y menor margen de error del 7.1% en la validación del modelo.

Mediante el algoritmo Decision Tree se pudo determinar el nivel de interacción de los estudiantes conjuntamente con la utilización de cada uno de los atributos seleccionados que conforman el data set final (ver Tabla 1) de tal manera se obtuvieron los siguientes resultados durante la fase de entrenamiento del algoritmo el nivel de interacción es 19 altas, 438 medias y 190 bajas, así mismo en la fase de validación 7 altas, 239 medias y 86 bajas, los resultados finales del nivel de interacciones es 26 altas, 677 medias y 276 bajas (ver Fig. 4).

Según con los resultados obtenidos se pudo determinar que 677 corresponden al nivel de interacción medias de los estudiantes en el curso virtual de inglés equivalente a un porcentaje del 69%, el nivel de interacción media significa que el estudiante utiliza los recursos conscientemente teniendo definido sus intereses en cuanto al material.

Fig. 4. Interacciones de los estudiantes

Además de obtener los resultados mediante el algoritmo Decision Tree se pudo evidenciar que de los 1069 datos de los estudiantes obtenidos para generar el modelo, 90 estudiantes es decir el 9% de los datos no han sido clasificados en ningún tipo del nivel de interacción (ver Fig. 5).

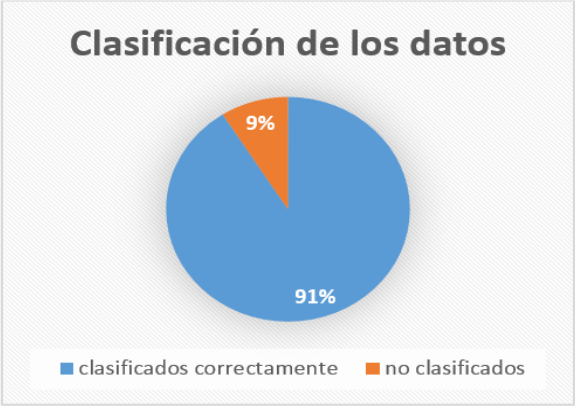


Fig. 5. Clasificación de los datos

4.2.5.2 Reglas Obtenidas Mediante los Algoritmos de Minería de Datos

Nivel alto de interacción en el curso virtual

Cuando los estudiantes tienen interacciones altas con los exámenes y los recursos y tiene entre 25 y 29 años y es soltero, entonces las interacciones en el curso virtual son altas.

Si las interacciones con los recursos es medio y tiene entre 25 y 29 años y el género es femenino, entonces las interacciones en el curso virtual es alto.

Si las interacciones con los exámenes es medio y es mayor a 29 años y no trabaja, entonces las interacciones en el curso virtual es alto.

Si las interacciones con los exámenes y los recursos es medio y pertenece a otra ciudad y no tiene hijos, entonces las interacciones en el curso virtual es alto.

Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y el género es femenino y no trabaja y pertenece a otra ciudad y es mayor a 29 años, entonces las interacciones en el curso virtual es alto.

Las interacciones con los exámenes y con los recursos es alto y no posee ningún tipo de servicio y no tiene hijos y tiene entre 25 y 29 años, entonces las interacciones en el curso virtual es alto.

Nivel medio de interacción en el curso virtual

Las interacciones en los exámenes es media y pertenece a otra ciudad y no tiene hijos y es casado, entonces las interacciones en el curso virtual es media.

Si las interacciones con los exámenes es medio y las interacciones con los recursos y las tareas es alto y tiene todos los servicios, no trabaja, el género es masculino, pertenece a otra ciudad y es mayor a 25 años, entonces las interacciones en el curso virtual es medio.

Si las interacciones con los recursos es alto y las interacciones con las tareas es medio y el género es masculino y posee todos los servicios y es mayor a 29 años, entonces las interacciones en el curso virtual es medio.

El estudiante no trabaja y el género es masculino y pertenece a otra ciudad y es soltero y solo posee un servicio que es número de celular, entonces las interacciones en el curso virtual es medio.

El estudiante no trabaja y el género es femenino y posee todos los servicios y pertenece a la ciudad de Loja y es soltero y es menor a 25 años y las interacciones con los recursos es medio, entonces las interacciones en el curso virtual es medio.

Las interacciones con los exámenes es bajo y las interacciones con las tareas es bajo y las interacciones con los recursos es medio y es casado, entonces las interacciones en el curso virtual es medio.

Las interacciones con los exámenes y con las tareas y con los recursos es bajo y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es masculino y pertenece a otra ciudad y es casado, entonces las interacciones en el curso virtual es medio.

Las interacciones con los exámenes es medio y las interacciones con las tareas y recursos es baja y tiene todos los servicios y es soltero, entonces las interacciones en el curso virtual es medio.

Nivel bajo de interacción en el curso virtual

Si las interacciones con los exámenes y los recursos o tareas son bajas y trabaja, entonces las interacciones en el curso virtual es bajo.

Si las interacciones con los exámenes y los recursos son bajas y el género es femenino y tiene hijos, entonces las interacciones en el curso virtual es bajo.

Si las interacciones con los exámenes y los recursos son bajas y tiene hijos y es menor a 25 años, entonces las interacciones en el curso virtual es bajo.

Si las interacciones con los exámenes, las tareas y los recursos son bajas y pertenece a la ciudad de Loja y no trabaja y posee todos los servicios y el género es masculino y es mayor a 29 años y tiene hijos, entonces las interacciones en el curso virtual es bajo.

El estudiante no trabaja y el género es masculino y pertenece a la ciudad de Loja y posee un servicio y es menor a 25 años y las interacciones con los recursos y las tareas es baja, entonces las interacciones en el curso virtual es bajo.

El estudiante trabaja y el género es masculino y pertenece a la ciudad de Loja y posee todos los servicios y es soltero y es menor a 25 años y las interacciones con los recursos y las tareas es baja y tiene hijos, entonces las interacciones en el curso virtual es bajo.

El estudiante trabaja y el género es femenino y posee todos los servicios y pertenece a otra ciudad y es casado y las interacciones con los exámenes es bajo, entonces las interacciones en el curso virtual es bajo.

Las interacciones con los exámenes y con las tareas y con los recursos es bajo y posee los servicios y tiene hijos y es mayor a 29 años y trabaja y el género es femenino y pertenece a la ciudad de Loja y es viudo, entonces las interacciones en el curso virtual es bajo.

4.2.5.3 Factores Para Determinar las Interacciones de los Estudiantes

Los factores que influyeron en la realización del modelo se encuentran asociados entre sí, los cuales son datos: personales, institucionales, socioeconómicos e interacciones del estudiante los cuales se detallan a continuación:

Interacciones en el curso: interaccionestareas (número de accesos a las tareas), interaccionesrecurso (número de veces que accede a un recurso), interaccionesexamenes (número de accesos a exámenes).

Datos personales de los estudiantes: género, estadocivil, edad, servicios (teléfono, celular), ciudad (estudiantes que residen en Loja o en otra ciudad del país).

Datos socioeconómicos de los estudiantes: numeroHijos, trabajo (si el estudiante trabaja o no).

Datos institucionales de los estudiantes: carrera (a que carrera pertenece el estudiante).

A continuación se presenta cada uno de los atributos con sus respectivos pesos según los resultados obtenidos mediante el algoritmo Decision Tree de tal forma que se pueda determinar el que más influye en el modelo (ver Tabla 4).

Tabla 6. Porcentaje de los factores, atributos con sus respectivos pesos.

Atributo	Porcentaje del atributo (%)
Interacciones tareas	12.196
Interacciones recurso	10.946
Interacciones exámenes	13.299
Genero	4.562
Estado civil	9.509
Edad	8.671
Servicios	8.299
Carrera	9.137
Número hijos	5.346
Trabajo	8.126
Ciudad	9.908

Luego de obtener el peso de cada uno de los atributos pertenecientes a los datos de los estudiantes, los que más inciden en el objetivo principal del presente proyecto es determinar el nivel de interacción de los estudiantes del curso virtual inglés de la MED son las interacciones en las tareas con un 12%, en los recursos con el 11% y en los exámenes el 13% como se puede observar en la siguiente figura (ver Fig. 6).

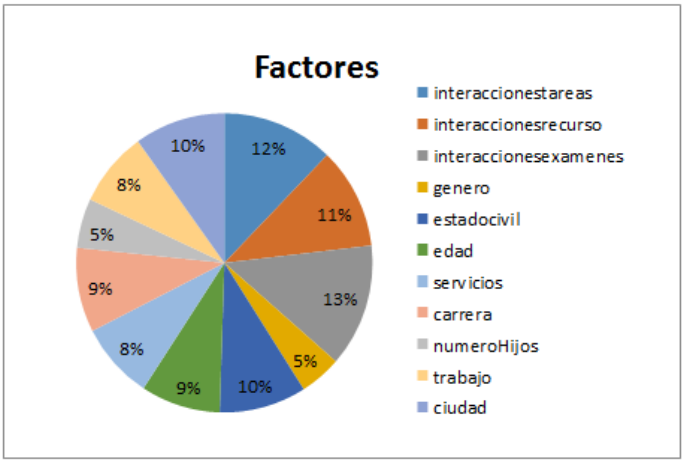


Fig. 6. Factores

4.2.5.4 Análisis de los Resultados

La Universidad Nacional de Loja a través de la MED brinda cursos virtuales de aprendizaje entre estos se tiene el curso de inglés, el mismo que fue tomado como objeto de estudio para el presente proyecto, ya que este idioma es fundamental porque se lo utiliza como un medio de comunicación, así mismo para poder obtener becas e ir a otro país a especializarse, para laborar, además la mayoría de la información que se encuentra disponible ya sea en digital o impresa está en inglés, entre otras utilidades, es por ello que se realizó el análisis para determinar la interacción de los estudiantes en este curso virtual, tomando dos conjuntos de datos los cuales estuvieron conformados por 147 alumnos obteniendo un total de 32029 interacciones que se encuentran distribuidas en tareas, recursos y exámenes, también se obtuvo datos personales como edad, estado civil, servicios, género, dirección, de cada uno de los estudiantes, los cuales conformaron el data set.

Con estos datos se realizó el análisis mediante la minería de datos donde se obtuvo algunos inconvenientes que los resultados obtenidos de los modelos fueron muy bajos es decir no cumplían con un rango de aceptación, motivo por el cual fue necesario la incrementación de nuevos datos, mismos que fueron proporcionados por la unidad de telecomunicaciones, ya una vez que se contaba con 1069 datos se procedió a integrar nuevos atributos de estos estudiantes como son datos socioeconómicos (tienen hijos o no, situación laboral), datos institucionales (carrera a la que pertenecen), con estos parámetros se realizó nuevamente el análisis conjuntamente con la técnica de minería de datos y los algoritmos seleccionados obteniendo un mejoramiento en los resultados, porque el mayor porcentaje que arrojaron los algoritmo con los datos anteriores fue del 77%, mientras que con los nuevos datos y atributos incorporados los resultados mejoraron obteniendo los porcentajes mayores al 90%, por ende mediante el modelo obtenido los resultados son confiables.

Consecuentemente se pudo determinar que 26 estudiantes pertenecen a la clasificación de la interacción alta, 677 estudiantes dentro de las interacciones medias y 276 estudiantes en interacciones bajas, estos resultados se obtuvieron a través de las reglas generadas por los algoritmos que conforman el modelo, presentando diferentes situaciones para cada nivel de interacción, las cuales se presentan a continuación:

El nivel de interacción de los estudiantes es alta en el curso virtual de inglés, cuando las interacciones en los exámenes y los recursos son altas, su edad es mayor a 25 años, su estado civil es soltero, no trabaja, no tienen hijos, pertenecen a otra ciudad y su género es femenino.

El nivel de interacción de los estudiantes es media en el curso virtual de inglés, cuando las interacciones en los exámenes, tareas y los recursos son medias, pertenecen a cualquier ciudad, puede tener cualquier edad, su estado civil es soltero o casado, puede poseer un trabajo o no, puede poseer todos los servicios o uno de ellos, puede tener hijos o no, puede ser masculino o femenino.

El nivel de interacción de los estudiantes es bajo en el curso virtual de inglés, cuando las interacciones en los exámenes, tareas y los recursos son bajas, pertenecen a la ciudad de Loja, es mayor a 29 años, su estado civil es casado, posee un trabajo, puede poseer todos los servicios o uno de ellos, tiene hijos, puede ser masculino o femenino.

Cabe indicar que las interacciones en las tareas se refiere a la revisión y envío de tareas, consulta de las sugerencias de las actividades que contiene las tareas; en las interacciones de los exámenes contiene revisión de los horarios de los que se puede rendir el examen, además saber la calificación y soluciones a las preguntas incorrectas que se han cometido y las interacciones de recursos se refiere cuando el estudiante descarga el material de apoyo que contiene cada unidad que se va a revisar en el curso.

Por lo tanto se puede indicar que las interacciones de los estudiantes que mayor prevalecen es el nivel medio con un porcentaje del 69% y el nivel bajo con el porcentaje del 25% en el curso virtual de inglés, significando que el estudiante utiliza los recursos conscientemente teniendo definido sus intereses en cuanto al material, es decir que no acceden con frecuencia al curso, por lo tanto realizan pocas consultas a las tareas, recursos y exámenes que deben cumplir.

5 Conclusiones

A través de los resultados obtenidos se ha podido concluir lo siguiente:

La minería de datos es muy importante dentro del campo de la educación ya que ayudó a extraer información que se encuentra oculta en los datos de tal forma permitió el análisis y la generación de nuevo conocimiento para poder determinar en nivel de interacción de los estudiantes.

RapidMiner es una herramienta de minería de datos potente ya que contiene complementos que permite hacer uso de diferentes algoritmos tanto de esta herramienta como de otras herramientas, además tiene operadores que ayudan a facilitar el desarrollo de los procesos para crear los modelos aplicables para el análisis de los datos.

Para determinar el nivel de interacción en el curso de inglés se aplicó diferentes algoritmos de clasificación, presentando los mejores resultados el Decision Tree, ya que este algoritmo obtuvo el menor margen de error durante la clasificación de los datos de las interacciones en el curso (tareas, exámenes, recursos), datos personales, institucionales y socioeconómicos.

Mediante el modelo de minería de datos obtenido se pudo determinar que las interacciones de los estudiantes en el curso virtual de inglés que mayor prevalece es el nivel medio con un porcentaje del 69% y los factores que más influyeron en el modelo fueron las interacciones en los exámenes, tareas, recursos, el estado civil y la situación laboral del estudiante.

Agradecimientos

Este trabajo ha sido auto-financiado por la autora Angélica Elizabeth Jaramillo Zhingre y contando con el apoyo de la carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja. El presente artículo forma parte del trabajo de titulación denominado “Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de aprendizaje”.

La autora desea expresar su agradecimiento particular a la carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja, y al Ing. Henry Paz por la tutoría prestada a lo largo del desarrollo del trabajo de titulación.

Referencias

1. Gómez, LM, Macedo, JC: Importancia de las TIC en la Educación Básica Regular, Universidad Nacional Mayor de San Marcos – Facultad de Educación.

2. Minería de Datos, Universidad de Extremadura - Campus Libre y Abierto, http://cala.unex.es/cala/epistemowikia/index.php?title=Miner%C3%ADa_de_Datos
3. Ordoñez, KF,: Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL, Universidad Técnica Particular de Loja – Area Técnica, http://dspace.utpl.edu.ec/bitstream/123456789/7897/1/Ordonez%20Brice%C3%B1o%20Karla-%20Informatica.pdf
4. Carmen, M., Galán, SJ: Definición de Minería de Datos, Universidad Carlos III de Madrid, En línea: http://www.oocities.org/es/mineria.datos/definicion_tecnicas_mineria_datos.pdf
5. Instituto de Investigación en Inteligencia Artificial, Minería de Datos o Data Mining, Consejo Superior de Investigaciones Científicas - Instituto de Investigación en Inteligencia Artificial, http://www.iiia.csic.es/udt/files/DataMining.pdf
6. Romero, C., Ventura, S., Hervás, C.: Escuela Politécnica Superior Universidad de Córdoba, Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web, http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CICA_2009_Aplicacion_Mineria_Datos_basada_ense%F1anza_web.pdf
7. Sarango, MY,: 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 – Escuela de Ciencias de la Computación, http://dspace.utpl.edu.ec/bitstream/123456789/2387/1/MarciaSarangoTsis.pdf
8. Solarte, GR, Soto, JA: Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares, Universidad Tecnológica de Pereira, http://revistas.utp.edu.co/index.php/revistaciencia/article/viewFile/1487/947
9. García, JA, Acevedo AM: Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies, Universidad Tecnológica de Pereira - Facultad de ingenierías: eléctrica, electrónica, física y Ciencias de la computación - Ingeniería de sistemas y computación, http://repositorio.utp.edu.co/dspace/bitstream/11059/1339/1/006312G216.pdf
10. Velandia, RA, Hernández, FL: Evaluación de Algoritmos de extracción de reglas de decisión para el diagnóstico de huecos de tensión, Universidad Industrial de Santander, http://tangara.uis.edu.co/biblioweb/tesis/2010/134742.pdf
11. Cubero, JC, Berzal, F.: Herramientas de Minería de Datos, Universidad de Granada – Departamento de Ciencias de la Computación e Inteligencia Artificial, http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf
12. Qualex Consulting Services, SAS Enterprise Miner, http://www.qlx.com/Software_Sales/enterprise_miner.html
13. Beltrán, D. Poveda, D.: RAPIDMINER, Universidad Nacional de Colombia- Facultad de Ciencias Económicas - Unidad de Informática y Comunicaciones, http://www.fce.unal.edu.co/uifce/pdf/Rapid_Miner.pdf
14. WEKA: Software de minería de datos en JAVA, Universidad de Waikato, http://www.cs.waikato.ac.nz/ml/weka/
15. García FJ: Aplicación de Técnicas de minería de datos a datos obtenidos por el centro de Andaluz de Medio Ambiente, Universidad de Granada – Master Universitario en Estadística Aplicada, http://masteres.ugr.es/moea/pages/tfm-1213/tfm_garciagonzalezfrancisco_1/!
16. KNIME, KNIME, http://www.knime.org/
17. Moine, JM, Haedo, AS, Gordillo, S.: Estudio comparativo de metodologías para minería de datos, Universidad Nacional de La Plata - Facultad de Informática, http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf%3Fsequence%3D1
18. Moine, JM, Haedo, AS, Gordillo, S.: Análisis comparativo de metodologías para la gestión de proyectos de minería de datos, Universidad Nacional de La Plata - Facultad de Informática, http://sedici.unlp.edu.ar/bitstream/handle/10915/18749/Documento_completo.pdf%3Fsequence%3D1
19. Gironés, J.: Metodologías y estándares, Universidad Abierta de Cataluña, http://www.exabyteinformatica.com/uoc/Administracio_i_direccio_dempreses/Business_analytics/Business_analytics_(Modulo_3).pdf
20. Herrera, MA, Acosta, JD: Estudio sobre el estado de las soluciones ict y de los casos prácticos de aplicación de la minería de datos a nivel mundial en al menos 5 casos representativos, Universidad EAFIT, https://repository.eafit.edu.co/bitstream/handle/10784/2457/AcostaVasquez_JuanDavid_2006.pdf?sequence=1&isAllowed=y
21. Martínez, CÁ: Aplicación de técnicas de minería de datos para mejorar el proceso de Control de Gestión en Entel, universidad de chile - Facultad de Ciencias Físicas y Matemáticas, http://www.tesis.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1
22. Fischer, ES: Modelo para la Automatización del Proceso de Determinación de Riesgo de Deserción de Estudiantes Universitarios, Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, http://preu.unillanos.edu.co/sites/default/files/fields/documentos/PREDICION%20DESERCI.pdf
23. Valdivieso, PM: Aplicación de métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje, http://www.redalyc.org/pdf/3314/331427213010.pdf

Anexo J: Certificado de Traducción

Lic.

Carlos Eduardo Zurita Valencia

LICENCIADO EN CIENCIAS DE LA EDUCACIÓN EN LA ESPECIALIDAD DEL IDIOMA INGLÉS

CERTIFICA:

Que la traducción del resumen del Trabajo de Titulación cuyo tema es **"Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje"** es fiel traducción, por lo que su contenido puede ser interpretado de forma correcta.

Atentamente:

…….

Lic. Carlos Eduardo Zurita Valencia

Anexo K: Licencia Creative Commons

Figura 151: Licencia Creative Commons