UNIVERSIDAD TECNOLÓGICA EQUINOCCIAL

FACULTAD DE CIENCIAS DE LA INGENIERÍA E INDUSTRIAS

CARRERA DE INGENIERÍA INFORMÁTICA

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

EVALUACIÓN DE MÉTODOS Y ALGORITMOS PARA REALIZAR MINERÍA DE DATOS EDUCACIONALES.

AUTOR:
MAYRA ELIZABETH VIZCAÍNO RUIZ

DIRECTOR:
ING. OSWALDO MOSCOSO ZEA

Quito - 2017

DERECHOS DE AUTOR

FORMULARIO DE REGISTRO BIBLIOGRÁFICO
PROYECTO DE TITULACIÓN

DATOS DE CONTACTO	
CÉDULA DE IDENTIDAD:	0502840515
APELLIDO Y NOMBRES:	VIZCAÍNO RUIZ MAYRA ELIZABETH
DIRECCIÓN:	OFELIA MOLLES Y JAIME ALBUJA
EMAIL:	mayeliza2008@hotmail.com
TELÉFONO FIJO:	022293377
TELÉFONO MOVIL:	0983479911

DATOS DE LA OBRA	
TITULO:	EVALUACIÓN DE MÉTODOS Y ALGORITMOS PARA REALIZAR MINERÍA DE DATOS EDUCACIONALES
AUTORA:	VIZCAÍNO RUIZ MAYRA ELIZABETH
FECHA DE ENTREGA DEL PROYECTO DE TITULACIÓN:	27 MARZO 2017
DIRECTOR DEL PROYECTO DE TITULACIÓN:	ING.OSWALDO MOSCOSO
PROGRAMA	PREGRADO ● POSGRADO ☐
TITULO POR EL QUE OPTA:	INGENIERA EN INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÒN
RESUMEN: Mínimo 250 palabras	Con el desarrollo de las técnicas de minería de datos, en los últimos años se ha intensificado el análisis de información académica para determinar las causas que ocasionan la deserción, la permanencia y la culminación exitosa en las carreras a nivel universitario. Muchos trabajos, estudios y artículos se han realizado sobre este tema, los cuales en su generalidad, concluyen que es posible predecir el porcentaje de graduación, deserción y el nivel de permanencia de estudiantes en las instituciones de educación superior. El presente proyecto de titulación presenta la minería de datos educacionales, evaluación de métodos y algoritmos existentes. Se utiliza la metodología KDD por sus siglas en inglés (Knowledge Discovery in Databases) que realiza

el proceso de descubrimiento del conocimiento en bases de datos. Como primeros pasos se da un especial énfasis al estudio de los datos mediante la selección, limpieza y transformación de los mismos creando un dataset que representa el conjunto completo de los datos de una universidad particular ubicada en la ciudad de Quito, Ecuador.

Sobre el dataset se aplican técnicas de predicción utilizando diferentes criterios de representación y aplicación de algoritmos de clasificación como árboles de decisión, redes bayesianas, reglas de decisión y metaclasificadores. Para determinar cuáles son los más óptimos se realiza experimentos utilizando dos métodos de clasificación la validación cruzada y división porcentual con las clases de graduación y deserción estudiantil. Las herramientas de minería de datos usadas en este trabajo son: WEKA, ORANGE 3 Y RAPID MINER.

Los resultados obtenidos permiten predecir u obtener patrones con datos educativos con un modelo que indica los porcentajes de la deserción y graduación de estudiantes en la universidad, para que las autoridades puedan predecir tendencias que permitirán tomar acciones correctivas y preventivas, mejorando los procesos de enseñanza de los nuevos estudiantes que ingresen a la carrera de Ingeniería Informática.

PALABRAS CLAVES:

DBMS Sistema de gestión de base de datos

KDD Descubrimiento de conocimiento de bases de datos

MD Minería de datos

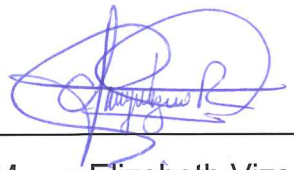
ABSTRACT:

With the development of data mining techniques, the analysis of academic information has been enhanced in recent years to determine the causes that lead to dropout, permanence, and successful completion of college-level careers. Many papers, studies and articles have been carried out on this subject, which in general, conclude that it is possible to predict the percentage of graduation, dropout and the level of permanence of students in higher education institutions.

The present project presents the mining of educational

	data, evaluation of existing methods and algorithms. It uses the KDD methodology (Knowledge Discovery in Databases) that performs the process of knowledge discovery in databases. As a first step, a special emphasis is given to the study of data by selecting, cleaning and transforming them, creating a dataset that represents the complete set of data of a private university located in the city of Quito, Ecuador. On the dataset, prediction techniques are applied using different criteria for the representation and application of classification algorithms such as decision trees, Bayesian networks, decision rules and metaclasificators. To determine which are the most optimal experiments are performed using two methods of classification cross-validation and percentage division with classes of graduation and dropout. The data mining tools used in this work are: WEKA, ORANGE 3 AND RAPID MINER. The obtained results allow to predict or obtain patterns with educational data with a model that indicates the percentages of dropout and graduation of students in the university, so that the authorities can predict trends that will allow corrective and preventive actions, improving the teaching processes of The new students who enter the career of Computer Engineering.
KEYWORDS	**DBMS** Database management system **KDD** Knowledge Discovery from Databases **MD** Minería de datos

Se autoriza la publicación de este Proyecto de Titulación en el Repositorio Digital de la Institución.



Mayra Elizabeth Vizcaíno Ruiz

C.I. 0502840515

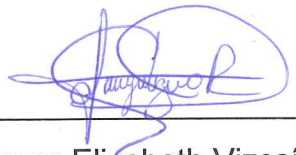
DECLARACIÓN Y AUTORIZACIÓN

Yo, **VIZCAÍNO RUIZ MAYRA ELIZABETH** CI 0502840515 autora del proyecto titulado: **Evaluación de métodos y algoritmos para realizar minería de datos educacionales**, previo a la obtención del título de Ingeniera Informática y Ciencias de la Computación en la Universidad Tecnológica Equinoccial.

1. Declaro tener pleno conocimiento de la obligación que tienen las Instituciones de Educación Superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de graduación para que sea integrado al Sistema Nacional de información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2. Autorizo a la BIBLIOTECA de la Universidad Tecnológica Equinoccial a tener una copia del referido trabajo de graduación con el propósito de generar un Repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Quito, marzo 2017

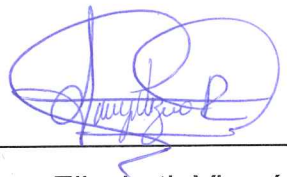


Mayra Elizabeth Vizcaíno Ruiz
C.I. 0502840515

DECLARACIÓN

Yo **Mayra Elizabeth Vizcaíno Ruiz** declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Tecnológica Equinoccial puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.



Mayra Elizabeth Vizcaíno Ruiz
C.I. 0502840515

CERTIFICACIÓN

Certifico que el presente trabajo que lleva por título **"Evaluación de métodos y algoritmos para realizar minería de datos educacionales."**, que, para aspirar al título de **Ingeniera Informática y Ciencias De La Computación** fue desarrollado por **Mayra Elizabeth Vizcaíno Ruiz**, bajo mi dirección y supervisión, en la Facultad de Ingeniería e Industrias; y cumple con las condiciones requeridas por el reglamento de Trabajos de Titulación artículos 19 ,27 y 28.

ING. OSWALDO MOSCOSO ZEA

DIRECTOR DEL TRABAJO

C.I 1709629651

DEDICATORIA

El presente trabajo de titulación va dedicado a Dios por darme salud y fuerzas para seguir adelante pudiendo resolver todos los problemas que se presentaron.

A mi familia quienes por ellos soy lo que soy.

A mis padres quienes me han guiado por el camino del bien siendo un pilar muy importante en mi vida, a mi esposo quien me ha dado amor, confianza y su apoyo incondicional para que el camino universitario sea más llevadero y finalmente quiero dedicarle mi esfuerzo, motivación y todas las ganas de seguir superándome a mi hijo Gorky Johan.

Mayra Elizabeth Vizcaíno Ruiz

AGRADECIMIENTOS

Agradezco a Dios por darme salud y vida permitiéndome culminar mi carrera universitaria, a mi familia por brindarme su apoyo incondicional en especial a mi esposo e hijo quienes supieron comprender mi ausencia mientras culminaba mis estudios universitarios ofreciéndome siempre su amor y confianza.

A mis maestros de Universidad Tecnología Equinoccial quienes desde un inicio compartieron sus conocimientos de la mejor manera para seguir formándome día a día. Y como no agradecer a mi tutor de tesis Ing. Oswaldo Moscoso por su esfuerzo, dedicación y apoyo constante.

Mayra Elizabeth Vizcaíno Ruiz

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

PÁGINA

ÍNDICE DE FIGURAS

PÁGINA

INDICE DE ANEXOS

PÁGINA

RESUMEN

Con el desarrollo de las técnicas de minería de datos, en los últimos años se ha intensificado el análisis de información académica para determinar las causas que ocasionan la deserción, la permanencia y la culminación exitosa en las carreras a nivel universitario. Muchos trabajos, estudios y artículos se han realizado sobre este tema, los cuales en su generalidad, concluyen que es posible predecir el porcentaje de graduación, deserción y el nivel de permanencia de estudiantes en las instituciones de educación superior.

El presente proyecto de titulación presenta la minería de datos educacionales, evaluación de métodos y algoritmos existentes. Se utiliza la metodología KDD por sus siglas en inglés (Knowledge Discovery in Database) que realiza el proceso de descubrimiento del conocimiento en bases de datos. Como primeros pasos se da un especial énfasis al estudio de los datos mediante la selección, limpieza y transformación de los mismos creando un dataset que representa el conjunto completo de los datos de una universidad particular ubicada en la ciudad de Quito, Ecuador.

Sobre el dataset se aplican técnicas de predicción utilizando diferentes criterios de representación y aplicación de algoritmos de clasificación como árboles de decisión, redes bayesianas, reglas de decisión y metaclasificadores. Para determinar cuáles son los más óptimos se realiza experimentos utilizando dos métodos de clasificación la validación cruzada y división porcentual con las clases de graduación y deserción estudiantil. Las herramientas de minería de datos usadas en este trabajo son: WEKA, ORANGE 3 Y RAPID MINER.

Los resultados obtenidos permiten predecir u obtener patrones con datos educativos con un modelo que indica los porcentajes de la deserción y graduación de estudiantes en la universidad, para que las autoridades puedan predecir tendencias que permitirán tomar acciones correctivas y preventivas, mejorando los procesos de enseñanza de los nuevos estudiantes que ingresen a la carrera de Ingeniería Informática.

ABSTRACT

With the development of data mining techniques, the analysis of academic information has been enhanced in recent years to determine the causes that lead to dropout, permanence, and successful completion of college-level careers. Many papers, studies and articles have been carried out on this subject, which in general, conclude that it is possible to predict the percentage of graduation, dropout and the level of permanence of students in higher education institutions.

The present project presents the mining of educational data, evaluation of existing methods and algorithms. It uses the KDD methodology (Knowledge Discovery in Database) that performs the process of knowledge discovery in databases. As a first step, a special emphasis is given to the study of data by selecting, cleaning and transforming them, creating a dataset that represents the complete set of data of a private university located in the city of Quito, Ecuador.

On the dataset, prediction techniques are applied using different criteria for the representation and application of classification algorithms such as decision trees, Bayesian networks, decision rules and metaclasificators. To determine which are the most optimal experiments are performed using two methods of classification cross-validation and percentage division with classes of graduation and dropout. The data mining tools used in this work are: WEKA, ORANGE 3 AND RAPID MINER.

The obtained results allow to predict or obtain patterns with educational data with a model that indicates the percentages of dropout and graduation of students in the university, so that the authorities can predict trends that will allow corrective and preventive actions, improving the teaching processes of The new students who enter the career of Computer Engineering.

INTRODUCCIÓN

1 INTRODUCCIÓN

La minería de datos (MD) permite el análisis de grandes volúmenes de datos mediante el uso de herramientas desarrolladas por científicos y académicos. Estas herramientas permiten la aplicación de algoritmos para la obtención de conocimiento a partir de datos preprocesados.

Tanto la preparación de los datos como la minería de datos educacionales (MDE) forman parte de un proceso mayor denominado “Desubrimiento del conocimiento en bases de datos” conocido en inglés como Knowledge Discovery in Databases (KDD) el cual permite obtener conocimiento útil a partir de la información de las bases de datos institucionales, ver Figura 1.

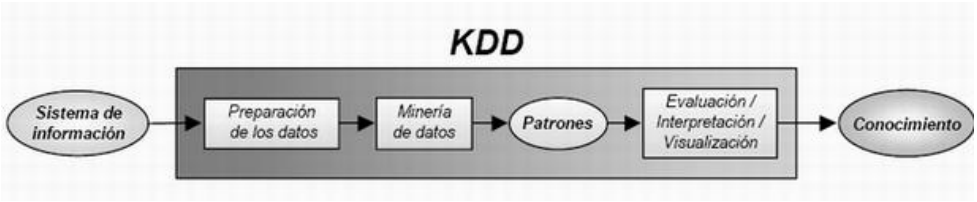


Figura 1. KDD
(Han, J., Kamber, M.,& Mark, A., 2010)

En la etapa de preparación de los datos se realiza la integración o selección desde fuentes de información para la creación del conjunto de datos (dataset). En la transformación de datos se generan nuevas variables a partir del procesamiento de la información original. La sección 3.1 presenta la descripción detallada de esta etapa.

En la etapa de minería de datos se realiza un análisis del dataset con el fin de intentar descubrir el algoritmo que tiene mejor rendimiento para analizar y predecir tendencias en ambientes educativos.

Las instituciones de educación superior tienen problemas relacionados con altos índices de deserción, bajas tasas de graduación y de retención de estudiantes. La MDE actualmente es reconocida como una disciplina emergente ya que proporciona nuevos mecanismos para tomar acciones correctivas y preventivas en favor de los estudiantes y los docentes de las

instituciones educativas. Últimamente se ha incrementado la aplicación de la MDE, con el objetivo de conocer métodos de descubrimiento donde se usen datos educacionales y se implementen métodos para comprender mejor a los estudiantes y el entorno en el que aprenden, de esta forma se puede crear estrategias para incrementar las tasas de graduación y disminuir las tasas de deserción estudiantil. El KDD ha sido visto como un proceso válido para determinar las posibles causas de estos graves problemas que se han detectado en las instituciones educativas.

En el presente trabajo de titulación, se ha elegido analizar la situación de una universidad particular, ubicada en la ciudad de Quito, Ecuador. La base de datos de la institución educativa presenta información de estudiantes desde el año 2002 hasta el año 2015.

Se usan 3 herramientas de MDE: Weka, Orange 3 y Rapid Miner. La mejor se elige luego de un análisis de factibilidad técnica con el fin de determinar las diferencias y el mejor desempeño de cada una.

La estructura de este documento es la siguiente:

El capítulo 2 muestra el marco teórico de este trabajo con las definiciones existentes de MDE, métodos y algoritmos, en el capítulo 3 se explica cuál fue la metodología del procesamiento de datos y el estado del arte para posteriormente obtener conocimiento mediante la exploración, desarrollo de métodos y algoritmos de minería de datos, en el capítulo 4 mediante el uso de la herramienta Weka y el dataset creado se muestran los resultados de MDE con 2 métodos de clasificación la validación cruzada y división porcentual, los resultados se comparan para sugerir al que tiene mejor precisión y mejor comprensión de resultados pudiendo identificar a estudiantes con tendencia a la deserción (abandono) y con mayor posibilidad de graduación lo que permitirá a la institución tomar acciones preventivas y correctivas.

Finalmente se presentan la generación del conocimiento y las conclusiones del trabajo de titulación.

1.1 PROBLEMA

Actualmente en la Facultad de Ciencias de la Ingeniería e Industrias de una universidad particular no se analiza la información académica de docentes y estudiantes con una metodología adecuada que permita descubrir conocimiento e información potencialmente útil. Para la extracción del conocimiento es importante la aplicación de métodos y algoritmos de minería de datos. En este trabajo se define un caso de estudio para evaluar dos indicadores importantes tasa de graduación y tasa de deserción. El caso de estudio presenta diferentes escenarios de análisis y evaluación de algoritmos que permiten dar respuestas óptimas para la toma de decisiones de los funcionarios de la universidad.

Este proyecto de investigación da lugar a las siguientes preguntas:

¿Cuáles serían los beneficios que tendría la universidad con la implementación de métodos y algoritmos para la MDE?

¿Qué aspectos son necesarios para que la MDE sea exitosa en base a los algoritmos definidos en el caso de estudio?

1.2 JUSTIFICACIÓN

Se propone realizar este proyecto de titulación debido a la necesidad de contar con una herramienta de minería de datos y con un proceso que intente descubrir patrones en grandes volúmenes de datos utilizando métodos de inteligencia artificial, aprendizaje automático, estadísticas y sistemas de bases de datos.

Al realizar una MDE en una institución de educación superior se puede encontrar un modelo comprensible el cual a partir del análisis de datos y el uso de algoritmos presente resultados con métricas y consideraciones claras del porcentaje de deserción y graduación de estudiantes y así finalmente tomar decisiones oportunas.

1.3 OBJETIVOS

1.3.1 OBJETIVO GENERAL

Evaluación de métodos y algoritmos con herramientas de minería de datos para determinar cuáles son los más óptimos estudiando los indicadores de graduación y deserción estudiantil.

1.3.2 OBJETIVOS ESPECÍFICOS

1. Definir el estado del arte de la inteligencia de negocios y la minería de datos educacionales.
2. Mediante experimentos analizar las técnicas, algoritmos y herramientas de evaluación de minería de datos existentes.
3. Definir un caso de estudio para la realización de minería de datos.
4. Evaluar métodos o algoritmos de minería de datos sobre deserción y graduación.

MARCO TEÓRICO

2 MARCO TEÓRICO

La investigación propuesta en este proyecto es la de realizar una evaluación de métodos y algoritmos de minería de datos educacionales para elegir los que se ajusten mejor de acuerdo a las necesidades y características de la institución educativa. Además se define el enfoque usado, detallando cada uno de los procesos para el descubrimiento de conocimiento mediante el uso de minería de datos. La metodología que se detalla en este trabajo es KDD. A continuación se presenta el estado de arte y la revisión bibliográfica para la realización de este trabajo.

2.1 SISTEMA DE INFORMACIÓN

Los sistemas de información (SI), son el "conjunto formal de procesos que operan sobre una colección de datos estructurada de acuerdo con las necesidades de una empresa que recopila, elabora y distribuye la información necesaria para la operación de dicha empresa y para las actividades de dirección, control correspondientes, apoyando, al menos en parte, los procesos de toma de decisiones necesarios para desempeñar las funciones de negocio de la empresa de acuerdo con su estrategia" (Andrew, 2007, pág. 243).

Actualmente, en una empresa un SI trata una gran cantidad de datos y proporciona información con diferentes estructuras que engloban equipos, programas informáticos, telecomunicaciones, bases de datos, recursos humanos y procedimientos (Andrew, 2007).

2.1.1 ADMINISTRACIÓN DE DATOS E INFORMACIÓN

El tratamiento de la información tiene como objetivo transformarla en conocimiento útil e importante para quienes lo requieran.

La evolución de los ordenadores ha hecho posible que, por un lado, el volumen de datos almacenados y procesados se incremente cada vez más y, por otro lado, que al disminuir el coste de los equipos informáticos sea posible establecer mecanismos de analítica de datos (Arjonilla, D.,& Medina, G., 2007).

2.2 TIPOS DE SI

En los SI cada uno cuenta con su nivel estratégico y a su vez, los sistemas de cada nivel se especializan en apoyar a cada una de las principales áreas funcionales.

Tenemos diferentes tipos de sistemas de información los cuales van ejecutándose según su aplicación entre ellos tenemos.

- Sistemas transaccionales
- Sistemas para planificación de una empresa
- Sistemas de información gerencial
- Sistemas de apoyo a las decisiones
- Sistemas de trabajo con conocimiento

2.3 ACTIVIDADES DE UN SI

Existen tres actividades en un SI las cuales producen información para que las organizaciones puedan tomar decisiones, controlar operaciones, analizar problemas y crear nuevos productos o servicios. Estas actividades son: entrada, procesamiento y salida.

2.3.1 ENTRADA

Es el conjunto de elementos tangibles o intangibles que serán procesados, dichos datos son capturados tanto en el interior de la organización como de su entorno externo.

2.3.2 PROCESAMIENTO

Es la interacción de los elementos de entrada que se procesarán de una forma significativa.

2.3.3 SALIDA

Es el resultado o respuesta del procesamiento que se transfiere a las personas que lo usarán o a las actividades para las que se utilizará.

2.4 BDD

Las bases de datos son un sistema formado por un conjunto de datos almacenados en discos. Estos datos u archivos pueden manipularse mediante programas que permiten el acceso directo a ellos. Los usuarios que usen el sistema pueden agregar nuevos archivos, insertar, recuperar, modificar, eliminar datos dentro de estos archivos y eliminar los archivos existentes dentro de la base de datos.

Un sistema de bases de datos comprende cuatro componentes principales: datos, hardware, software y usuarios (Hernández, J., Ramirez, M.,& Ferri , C., 2004).

2.5 DATA WAREHOUSE

Es una BDD corporativa que se caracteriza por integrar y depurar información de una o más fuentes operacionales, para luego procesarla permitiendo su análisis desde diferentes perspectivas y con grandes velocidades de respuesta.

El término data warehouse (DW) fue acuñado por primera vez por Bill Inmon, y se traduce como almacén de datos integrado, temático, histórico, no volátil sus definiciones son las siguientes:

Integrado: los datos almacenados se integran en una estructura consistente según las necesidades de los usuarios.

Temático: Se integran los datos por temas para fácil acceso y comprensión este proceso es la generación del conocimiento del negocio.

Histórico: El tiempo es parte implícita de la información contenida en un DW el cual realiza comparaciones con una variable en distintos tiempos. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente.

No volátil: El almacén de información de un DW puede ser leído, pero no modificado. La información es por tanto permanente, significando que la actualización e incorporación de los últimos valores tomarán distintas variables contenidas sin ningún tipo de acción sobre lo que ya existía (Guevara, J.,& Valencia, J., 2007).

2.5.1 ARQUITECTURA DE UN DW

A diferencia de un sistema tradicional de BDD, un DW presenta al usuario final la estructura de los datos con sus respectivos procesos y comunicaciones. Una de sus principales características es ejecutar consultas para facilitar el análisis de datos.

La Figura 2 muestra la arquitectura básica de un DW en la parte inferior se puede visualizar las fuentes externas conectadas a un monitor el cual es responsable de homogenizar la información y además detecta los cambios que se pueden realizar en las mismas.

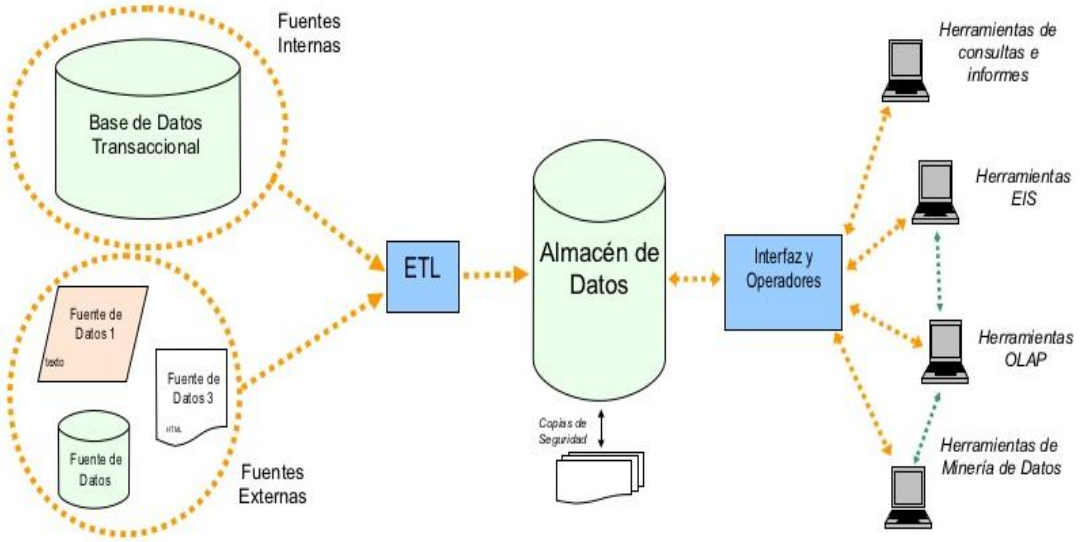


Figura 2. Arquitectura básica de un DW
(Inmon, 2002)

El integrador recibe los resultados de los extractores y después de integrarlos, los carga al DW. El administrador de consultas se encarga de organizar las consultas y seleccionar los operadores para permitir su análisis.

Para explicar la actividad de un DW se puede identificar dos grandes fases: construcción y explotación.

En la fase de diseño e implementación de las herramientas encargadas de llevar los datos de las fuentes al repositorio.

En la fase de explotación se realiza el análisis de los datos almacenados dentro del DW a través de técnicas que facilitan y son eficientes en una consulta. Ya con el DW poblado lo último es diseñar e implementar una interfaz que le permita al usuario final interactuar con el repositorio, brindándole todas las ventajas del análisis de la información (Lakshman, 2013).

2.5.1.1 Integración de datos

Consiste en integrar datos recolectados de diferentes sistemas operacionales de la organización y o fuentes externas. Las inconsistencias

encontradas en los diversos sistemas operacionales deben eliminarse. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.

En la Figura 3 se muestra el proceso de manejar información en distintas aplicaciones, para la integración se escoge un estándar de datos uniformes y se introduce al repositorio.

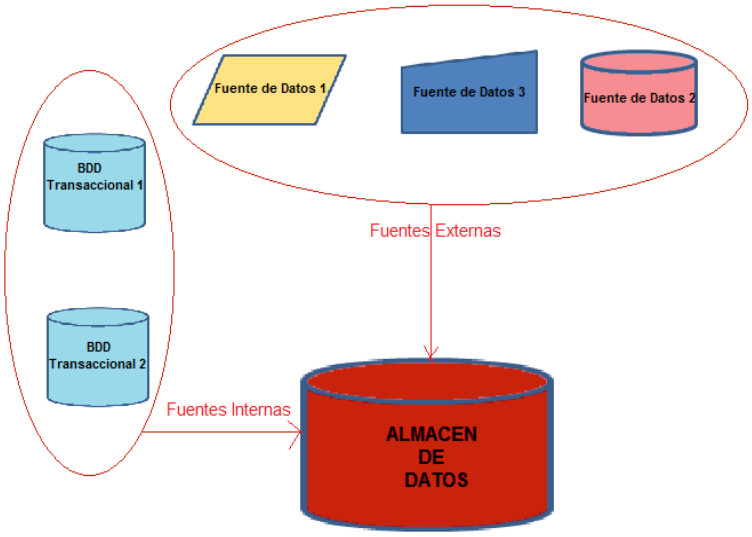


Figura 3. Integración de datos DW

La Figura 4 muestra la actualización de la BDD (insertar, borrar y modificar) en un ambiente operacional, pero la manipulación básica de los datos que ocurre en el DW es mucho más simple.

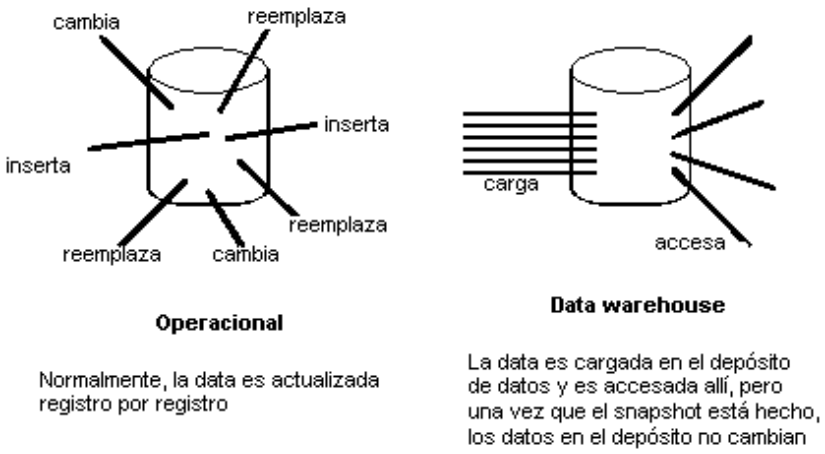


Figura 4. Datawarehouse no volátil
(Visual Studio .VisualBasic.net, 2016)

2.5.2 FUNCIONES ETL

En inglés Extract Transform Load (ETL) que significa extraer, transformar y cargar. Este proceso organiza el flujo de los datos entre diferentes sistemas además aporta con métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos. Lo principal es que la aplicación ETL lea los datos primarios de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos para su disponibilidad (Inmon,W.,& Dan, L., 2014).

2.5.2.1 Extracción de datos

Los datos con los que generalmente se trabaja son extraídos de diferentes sistemas de origen, generalmente los formatos se encuentran en bases de datos relacionales o ficheros planos y con estructuras diferentes. El proceso de extracción de datos los convierte en un formato uniforme para iniciar con el proceso de transformación.

2.5.2.2 Transformación de datos

Sobre los datos cargados se aplica diferentes reglas de negocio para transformarlos, a continuación se detallan algunas reglas:

- Selección de datos en donde se escoge ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
- Traducir códigos (por ejemplo, si la fuente almacena una “M” para Masculino y “F” para Femenino pero el destino tiene que guardar “1″ para Masculino y “2″ para Femenino).
- Codificar valores libres (por ejemplo, convertir “Femenino” en “F” o “Sra” en “1″) (Inmon,W.,& Dan, L., 2014).

2.5.2.3 Carga de datos

En esta fase, los datos procedentes de la fase de transformación son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes (Inmon,W.,& Dan, L., 2014).

2.5.3 APLICACIONES ETL

Las aplicaciones ETL más usadas en el mercado son:

- IBM Web sphere Data Stage
- Kettle ETL integration
- Oracle Database
- Informática Power Center
- Business Objects (BODI)
- SQL Server Integration (SSIS)

2.5.4 DESARROLLO DEL PROCESO DE CARGA DE DATOS

Según los requerimientos de la organización, este proceso puede tener una variedad de acciones. Por ejemplo se puede actualizar los datos donde se sobrescribe la información antigua. A continuación se detallan las dos formas básicas para el proceso de carga de datos:

2.5.4.1 Acumulación simple

El procedimiento más sencillo es realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y se transporta el resultado como una única transacción hacia el DW, este valor calculado es un sumatorio o un promedio de la magnitud considerada.

2.5.4.2 Rolling

Esta fase interactúa directamente con la BDD destino y, al realizar el análisis de la información muestra las operaciones con las restricciones que se aplicaron. Se garantiza la calidad de los datos en el proceso ETL con restricciones de valores únicos, integridad referencial, campos obligatorios y rangos de valores.

2.6 EL PROCESAMIENTO ANALÍTICO EN LÍNEA OLAP

En inglés Online Analytical Processing (OLAP) se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL), esta tecnología que se usa para organizar grandes BDD en las empresas aplicando la inteligencia de negocios, se dividen en uno o más cubos. Cada cubo tiene una forma para recuperar y analiza los datos con el fin de que sea más fácil crear y usar los informes de las tablas dinámicas con gráficos. El acceso a los datos es sólo de lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones (Jarke, M., Lenzerine, M., Vassiliu, Y.,& Vassiliadis,P., 2002).

2.7 MINERÍA DE DATOS

Es un conjunto de técnicas y tecnologías que permiten explorar de manera automática grandes bases de datos, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

La minería de datos ayuda a comprender el contenido de un repositorio de datos con el uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la inteligencia artificial (Hernández, J., Ramirez, M.,& Ferri , C., 2004).

2.7.1 ETAPAS PRINCIPALES DE LA MINERÍA DE DATOS

En cada caso es radicalmente distinto al anterior, el proceso común se compone de cuatro etapas principales.

2.7.1.1 Determinación de los objetivos

Se crean mediante la delimitación de los objetivos que el cliente desea con una orientación del especialista en minería de datos.

2.7.1.2 Pre procesamiento de los datos

En esta etapa se realiza la selección, limpieza, reducción y la transformación de las bases de datos, es generalmente el setenta por ciento del tiempo total de un proyecto de MD.

2.7.1.3 Determinación del modelo

Empieza con un análisis estadístico de los datos, para luego llevar a cabo una visualización gráfica de los mismos.

2.7.1.4 Análisis de los resultados

Se verifica los resultados obtenidos revisando si están correctos comparando con los obtenidos mediante análisis estadísticos y de visualización gráfica. El cliente determina si los resultados son novedosos y si le aportan un nuevo conocimiento para tomar decisiones.

2.7.2 ARQUITECTURA DEL MODELO DE MD

Un modelo de MD tiene una estructura independiente, la arquitectura del modelo de la MD ayuda a comprender el contenido de los datos los cuales se procesan y se analizan utilizando algoritmos de búsqueda. La información que almacena se define desde el origen de datos derivada del procesamiento estadístico con patrones encontrados mediante el análisis (Silberschatz, 2007).

En la Figura 5 se presenta el contenido de MD.

Figura 5. Arquitectura modelo de minería de datos
(Silva, 2007)

Una vez procesado el modelo, contiene los metadatos, resultados y enlaces de la estructura. Los metadatos especifican el nombre del modelo y el servidor donde está almacenado, para el análisis de los resultados se incluye las columnas, definiciones de los filtros, algoritmos empleados.

2.7.3 TAREAS DE MD

Se tiene 2 tareas: predictivas y descriptivas:

- Predictivas: tratan de problemas y tareas para predecir uno o más valores mediante la clasificación y probabilidad de clasificación, categorización, preferencias o priorización y regresión.
- Descriptivas: buscan describir los datos existentes mediante el agrupamiento, correlaciones, reglas de asociación, dependencias funcionales y detección de valores e instancias anómalas (Korth, H.,& Silberschatz, A., 2006).

2.8 MINERÍA DE DATOS EDUCACIONALES

La minería de datos educacionales (MDE) hace referencia a los métodos, algoritmos y herramientas para extraer y analizar datos generados o relacionados con las actividades de aprendizaje en las instituciones educativas. Muy a menudo, estos datos son extensos y no muy precisos.

Actualmente existe mucho interés en utilizar la MDE, centrándose en el descubrimiento del conocimiento donde se analice los datos de plataformas educacionales con el uso de métodos para comprender mejor a los estudiantes o a su vez encontrar cuál es el entorno en el que aprenden. Por ejemplo, temas como técnicas analíticas para predecir la tasa de graduación y abandono de los estudiantes juegan papeles importantes en el estudio de los datos educacionales con el objetivo de aportar información que contribuya a determinar cuáles son las causas.

Estas técnicas son usadas para la comprensión del comportamiento de los estudiantes. Algunos de los análisis que se pueden realizar son:

- Análisis del nivel participativo de los estudiantes (aprendizaje – enseñanza).
- Análisis de indicadores académicos como graduación, retención, deserción.
- Análisis de procesos organizacionales como investigación, vinculación y docencia.

Cuando se finalice el estudio de MDE se presentará analíticamente ambientes de aprendizaje, lo que permitirá tomar decisiones de mejora para la institución educativa sabiendo como interactuar con los estudiantes, profesores, administradores escolares.

2.8.1 APLICACIONES DE MDE

A continuación se presenta una lista de aplicaciones primarias de la MDE elaboradas en España por (Cristobal, R.,& Ventura, S., 2012).

- Realizar el análisis y visualización de datos
- Conocer datos relevantes y presentar a los instructores
- Realizar un análisis previo para predecir el desempeño del estudiante
- Presentar los comportamientos indeseables de estudiantes
- Realizar análisis de uso de las redes sociales
- Presentar recomendaciones para navegadores web

Las nuevas aplicaciones de MDE se centrarán en que los usuarios no técnicos utilicen estas herramientas y mediante la recopilación de datos su procesamiento sea más accesible y de fácil uso, algunos ejemplos incluyen las herramientas estadísticas y de visualización que analizan las redes sociales y su influencia en los resultados del aprendizaje y la productividad (Korth, H.,& Silberschatz, A., 2006).

2.8.2 VENTAJAS DE LA MDE

Se han realizado varias investigaciones y experimentos de laboratorio con datos reales que favorecen la MDE. Los experimentos se listan a continuación:

Aplicación de técnicas de minería de datos obtenidos por el centro de Andaluz CEAMA (García, 2008).

Diseño de un nuevo clasificador supervisado para minería de datos (Piorno, 2010).

Estos estudios mencionados permiten a los investigadores ahorrar mucho tiempo en tareas como la búsqueda de individuos o patrones en las bases de datos que por lo general son grandes también les ayuda a comprender el contenido del repositorio de los datos.

La MDE posee el potencial de extender un conjunto de herramientas mucho más amplio para el análisis de cuestiones importantes sobre diferencias individuales (Peña, 2014).

2.8.3 BENEFICIOS DE LA MDE

Para descubrir estudiantes con posibilidades de deserción en la institución universitaria se puede realizar la identificación temprana de los estudiantes que tienen problemas de aprendizaje, problemas económicos, entre otros, proporcionando algún tipo de atención personalizada con el fin de evitar que los estudiantes dejen sus estudios. La MDE da lugar a patrones interesantes que ayudan a hacer modelos predictivos para predecir el triunfo o fracaso de un estudiante, por lo que es muy importante intervenir lo más pronto posible para facilitar la retención estudiantil.

2.9 METODOLOGÍA

La MDE proviene de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, se aplican sobre un conjunto de datos creando un modelo que busca patrones y tendencias específicas para obtener unos resultados (Hernández, J., Ramirez, M.,& Ferri , C., 2004).

2.9.1 METODOLOGÍA KDD

La Figura 6 muestra el proceso KDD con las siguientes etapas: selección de datos, pre-procesamiento, transformación, minería de datos, interpretación y evaluación.

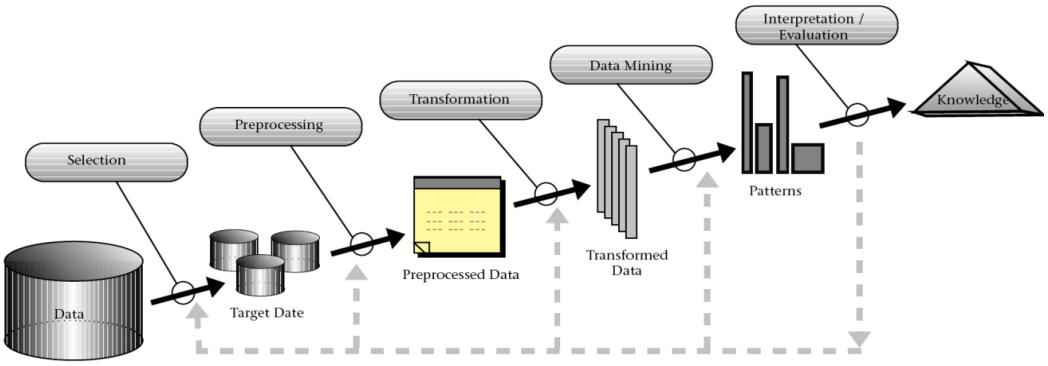


Figura 6. KDD

(Vieira, L., Ortiz, L.,& Ramirez, S., 2001)

Para identificar los métodos, técnicas, algoritmos y herramientas que existen en la rama de MDE, en el presente proyecto se va aplicar la metodología KDD, con los resultados se realizará tabulaciones y análisis donde la respuesta del sistema será estructurada según las condiciones y requerimientos de la institución para finalmente proponer un plan de implementación.

La Tabla 1 muestra el proceso KDD.

Tabla 1. Metodología KDD

(Santa Cruz, 2016)

Preparación de los datos	Selección	Recopilar e integrar las fuentes de datos Integrar y seleccionar las variables relevantes Aplicar técnicas de muestreo
	Exploración	Uso de técnicas de análisis exploratorio Deducir la distribución de los datos, simetría y correlaciones
	Limpieza	Detectar y trazar la presencia de los valores atípicos Imputar valores faltantes o perdidos Eliminar datos incorrectos e irrelevantes
	Transformación	Uso de técnicas de reducción y aumento de las dimensiones Uso de técnicas de descretización y numeración Escalado simple y multidimensional
Análisis de los datos	Técnicas predictivas	Regresión y series temporales Análisis discriminatorios Análisis de la varianza Árboles de decisión Redes neuronales
	Técnicas descriptivas	Clustering y segmentación Asociación Dependencia Análisis exploratorio
Evaluación e interpretación de datos	Intervalos de confianza Evaluación de los modelos	
Difusión y uso de modelos	Visualización y simulación	

2.9.2 METODOLOGÍA CRISP

La metodología CRISP es muy usada para la MD (Marques, 2014). En la Figura 7 se presenta el proceso de la metodología.

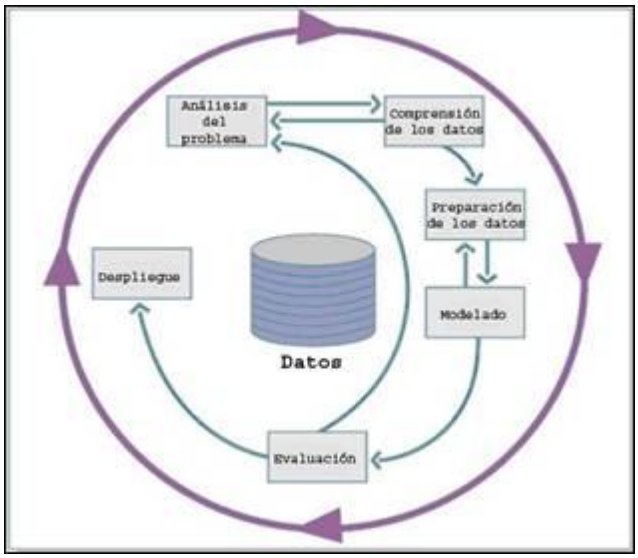


Figura 7. Metodología CRISP
(Marques, 2014)

Los pasos del proceso de la metodología CRISP son los siguientes:

Entendimiento del negocio: se establece de los objetivos y se realiza evaluación de la situación actual del negocio.

Entendimiento de los datos: se inicia con la descripción, exploración y verificación de calidad de datos.

Preparación de los datos: se crea el dataset mediante la selección, limpieza, construcción, e integración de datos.

Modelado: se realiza el modelado mediante el diseño, construcción y evaluación del modelo,

Evaluación de resultados finalmente se revisa el proceso realizado.

2.10 CLASIFICACIÓN DE ALGORITMOS

Una vez entrenado el modelo, se tiene a los algoritmos predictivos que predicen un dato o un conjunto de ellos, los algoritmos no supervisados

descubren patrones y tendencias en los datos pudiendo tomar acciones y obtener un beneficio.

La Conferencia Internacional del IEEE (siglas en inglés del Instituto de Ingeniería Eléctrica y Electrónica) de 2006 sobre minería de datos puntuó los mejores 8 algoritmos del campo. A continuación los describimos:

2.10.1 ÁRBOLES DE DECISIÓN

Los algoritmos de árbol de decisión consisten en organizarlos con ramas de influencia después de una decisión inicial. El tronco del árbol representa la decisión inicial, y empieza con una pregunta de sí o no, como por ejemplo si el estudiante se gradúa o no, por lo tanto las dos ramas divergentes del árbol, y cada elección posterior tendría sus propias ramas divergentes que llevan a un punto final.

Un árbol de decisión es un modelo de predicción para construir diagramas lógicos basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos: algoritmo ID3 y algoritmo C4.5.

2.10.2 ALGORITMO K-MEANS

Se basa en el análisis de grupos que divide los datos recogidos separándolos según las características con "bloques" o "clústeres".

2.10.3 ALGORITMO APRIORI

Este algoritmo normalmente controla los datos de transacciones. Por ejemplo, en una tienda de ropa, el algoritmo podría controlar qué camisas suelen comprar juntas los clientes.

2.10.4 ALGORITMO EM

Define parámetros analizando los datos y predice la posibilidad de una salida futura o evento aleatorio. Por ejemplo, el algoritmo EM podría intentar predecir el momento de una siguiente erupción de un géiser según los datos de tiempo de erupciones pasadas.

2.10.5 ALGORITMO ADABOOST

Se lo conoce como un clasificador fuerte que entrena a los clasificadores débiles de manera iterativa, funciona dentro de otros algoritmos de aprendizaje con la finalidad de encontrar una hipótesis fuerte a partir de hipótesis débiles o simples las cuales pueden anticipar un comportamiento según los datos observados para que sean sensibles a extremos estadísticos.

2.10.6 ALGORITMO K VECINO MÁS CERCANO

Es un algoritmo simple en la clasificación de patrones ya que realiza una probabilidad sobre la cercanía de sus vecinos, va reconociendo patrones en la ubicación de los datos y los asocia con un identificador mayor. Para los resultados establece las distancias de menor valor a *n* vecinos para finalmente elegir la clase a la que pertenezcan el mayor número de los *n* vecinos involucrados.

2.10.7 ALGORITMO NAIVE DE BAYES

Predice la salida de una identidad basándose en los datos de observaciones conocidas. La metodología bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el teorema de Bayes. Estos modelos primordialmente incorporan conocimiento previo para poder estimar modelos útiles dentro de un espacio muestral y de este modo poder

estimar parámetros que provengan de la experiencia o de una teoría probabilística (Han, J., Kamber, M.,& Mark, A., 2010).

2.10.8 ALGORITMOS BASADOS EN REDES NEURONALES

Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Ejemplos de red neuronal son:

- Construcción de mapas topográficos.
- Mapas auto organizados, también conocidos como redes de Kohonen.
- Modelamiento de la densidad de los datos

En la Figura 8 se muestra una red neuronal donde se puede observar los múltiples nodos que constituyen puntos de entrada de los datos estos están agrupados y sometidos a un tratamiento mediante un algoritmo que da lugar a la obtención de resultados esperados.

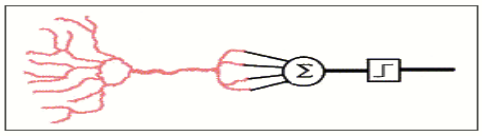


Figura 8. Red neuronal
(Riquelme, 2010)

2.11 HERRAMIENTAS PARA EVALUAR LA MDE

Predicen futuras tendencias y comportamientos, permitiendo a las empresas tomar decisiones. A continuación se presentan algunas herramientas de MD comerciales y de software libre.

- Intelligent Miner / DB2 Datawarehouse
- Clementine (SPSS)
- Enterprise Miner (SAS)
- DataEngine
- Orange 3
- Rapid Miner Studio
- Weka

2.11.1 IBM INTELLIGENT MINER

Es un software que comprende un conjunto de funciones: estadísticas, pre-proceso y minería que se utilizan para analizar grandes volúmenes de datos. En la Figura 9 se presenta un ejemplo de minería que comunica funciones en el servidor, y presenta la visualización de datos al cliente.

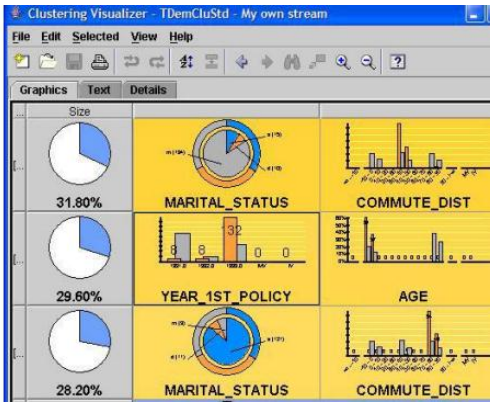


Figura 9. IBM Intelligent Miner
(IBM, 2012)

2.11.2 SPSS MODELER

Es una plataforma de análisis predictiva diseñada para aportar inteligencia predictiva a decisiones llevadas a cabo por personas, grupos, y sistemas proporciona un rango de algoritmos y técnicas avanzados, incluidos el análisis de texto, el análisis de entidad, la gestión y optimización de decisiones, SPSS. En la Figura 10 se presenta el diseño para mejorar las decisiones e integrar los resultados extrayendo el valor de datos.

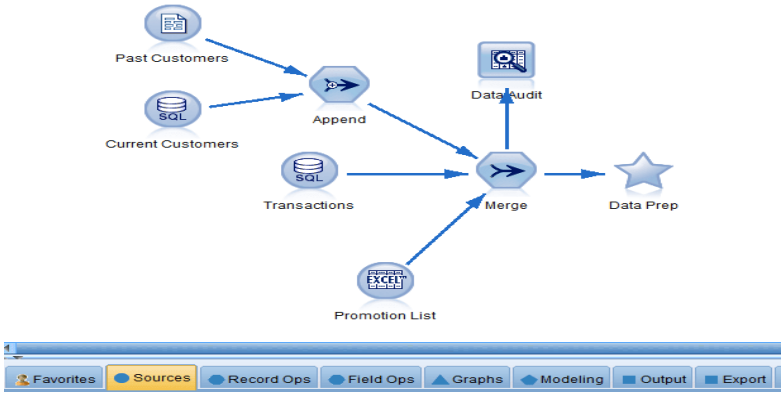


Figura 10. SPSS Modeler

2.11.3 SAS ENTERPRISE MINER

Esta herramienta realiza modelados descriptivos y predictivos que impulsa la mejor toma de decisiones. En la Figura 11 se presenta un modelo que es desarrollado para simplificar el proceso de minería.

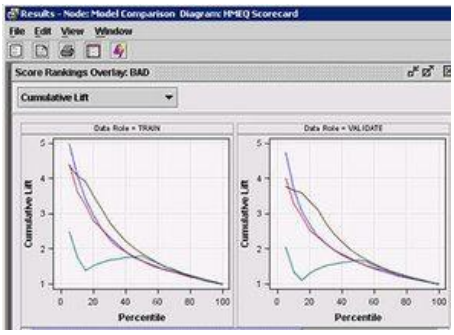


Figura 11. SAS Enterprise Miner
(SAS Institute Inc, 2010)

2.11.4 WEKA

Es un software de código abierto publicado bajo la licencia pública, tiene un conjunto de algoritmos de aprendizaje automático para tareas de minería que se pueden aplicar directamente a un conjunto de datos o llamadas de su propio código Java. En la Figura 12 se presenta a la herramienta Weka con el proceso de datos, clasificación, regresión, clustering, reglas de asociación, y visualización (Bouckaert, C.,& Peter, R., 2011).

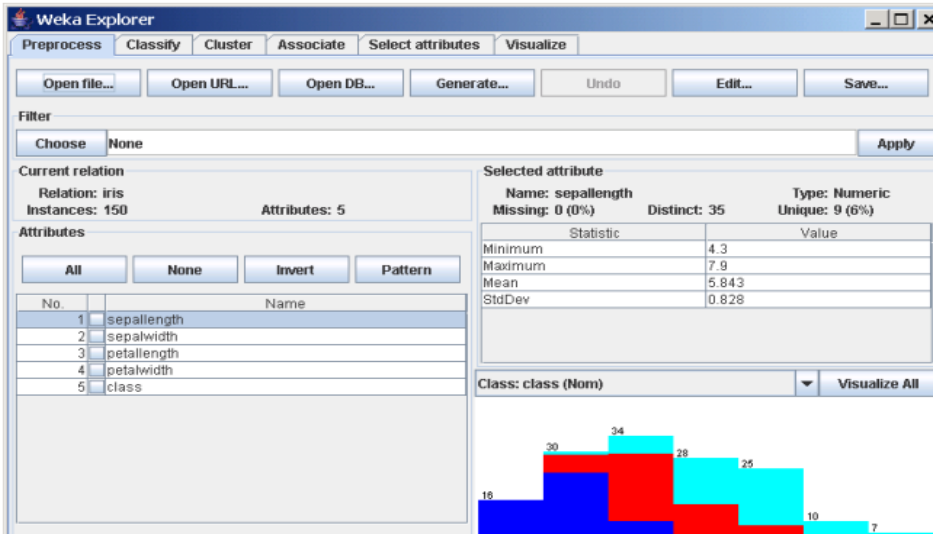


Figura 12. Weka Knowledge Explorer

2.11.4.1 Pestaña explorer en Weka

Se presenta en la interfaz principal de Weka donde se lleva a cabo las tareas, de ejecución de los algoritmos de análisis implementados sobre los ficheros de entrada. A estas funcionalidades se ingresa por las siguientes pestañas “preprocess” que permite la visualización y pre procesado de los datos, “classify”: para la aplicación de algoritmos de clasificación y regresión, “clúster” que muestra las técnicas de agrupación y “associate” que muestra los métodos de asociación.

Al entrar en la aplicación se presenta la interfaz vacía, tal como se muestra en la Figura 13.

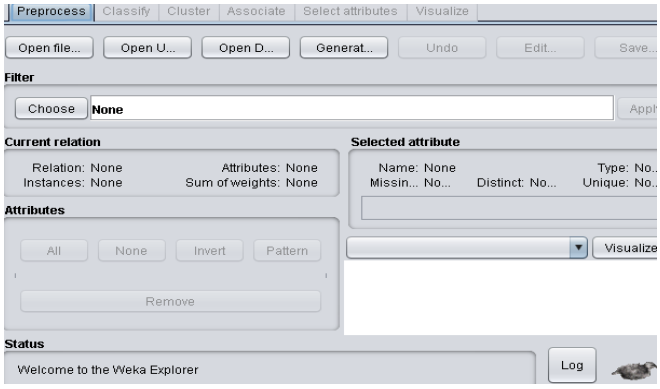


Figura 13. Interfaz de Explorer WEKA

2.11.4.2 Pestaña preprocess en Weka

Es la primera pestaña de la aplicación en Weka se carga el dataset tal como se muestra en la Figura 14.

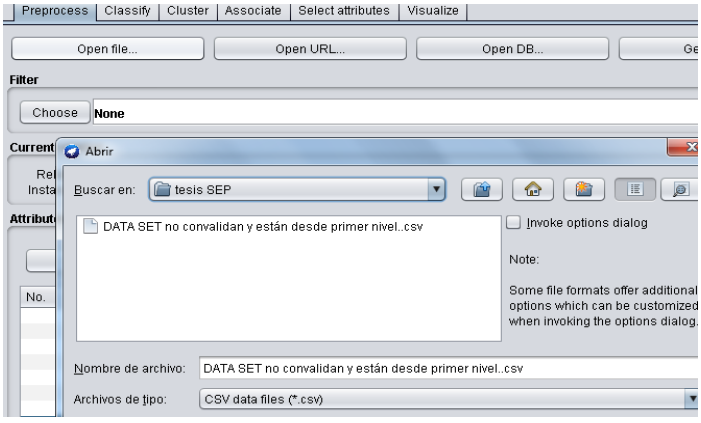


Figura 14. Preprocess en WEKA

Las herramientas de preprocesamiento se denominan filtro (Filter) y cada filtro actúa en uno de los siguientes niveles:

- Atributos: Actúan “en vertical” en la base de datos, modificando un atributo completo. Ejemplo: Filtro de discretización.
- Instancias: Actúan en horizontal, seleccionando un grupo de registros (instancias). Ejemplo: Filtro de selección aleatoria (Bouckaert, C.,& Peter, R., 2011).

2.11.4.3 Pestaña classify

En esta pestaña se puede definir y resolver un problema de clasificación aunque en ocasiones, el problema de clasificación se modifique con un análisis donde se aplican algoritmos no supervisados de agrupamiento y asociación para describir relaciones de interés en los datos.

En el caso de WEKA, la clase es uno de los atributos simbólicos disponibles, que se convierte en la variable objetivo a predecir. Por defecto, es el último atributo de la última columna, en este caso se va a escoger la clase grado y deserción. Como se muestra en la Figura 15 al pulsar sobre el botón “choose” se podrá escoger el algoritmo con el que se desea evaluar el dataset.

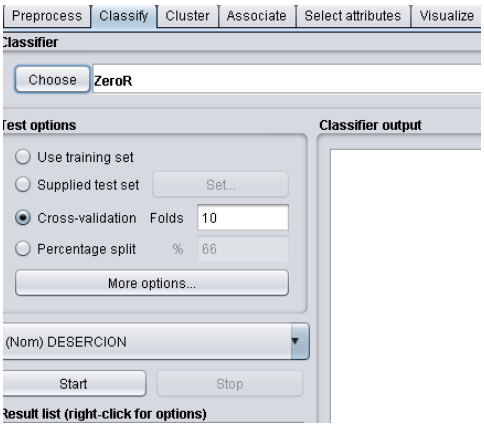


Figura 15. WEKA classiffy

Además en la parte inferior se presenta “test option” donde se puede incluir y modificar los parámetros asociados al clasificador entre ellos tenemos los siguientes:

- **Use training set:** En esta opción se entrenará al método con todos los datos disponibles posteriormente para presentar los resultados se realiza una evaluación de los mismos.

- **Supplied test set**: Con esta opción se puede seleccionar un fichero de datos con el que se probará el clasificador con el método de clasificación usado con datos iniciales.

- **Cross-validation:** Realizará una validación cruzada con el número de particiones.

- **Percentage split:** Con un porcentaje de los datos se realiza la clasificación y con la otra parte de datos se realizarán las pruebas.

2.11.4.4 Estructura de la matriz de confusión

En Weka presenta los resultados con la siguiente estructura

	a	**b**
Actual a = 0	TP	FN
Actual b = 1	FN	TP

Para realizar la lectura de la matriz de confusión se debe comprender la siguiente Tabla 6.

Tabla 2. Presentación de la matriz de confusión

(Bouckaert, C.,& Peter, R., 2011)

		Clase actual	
Categorías		0	1
Clase hipotética	0	**TN** Verdaderos negativos	**FN** Falsos negativos
	1	**FP** Falsos positivos	**TP** Verdaderos positivos
Columnas totales		N=FP+TN	P=TP+FN

A continuación se presentan valores que calculan los indicadores de precisión para cada clase, que se definen como:

- **Tasa de verdaderos positivos:**

$$Tp\ rate = \frac{TP}{TP + FN}$$

- **Tasa de falsos positivos:**

$$Fp\ rate = \frac{FP}{FP + TN}$$

- **Medida de precisión:**

$$Precision = \frac{TP}{TP + FP}$$

2.11.5 RAPID MINER

Es una herramienta de inteligencia de negocios que brinda a los usuarios un entorno gráfico, el cual está integrado de aprendizaje automático para la minería, es compatible con todos los pasos del proceso de minería de datos. Implementa más de 500 técnicas de pre-procesamiento de datos, visualización, modelación predictiva, métodos estadísticos de evaluación y despliegue (RapidMiner, Community, 2007).

En la figura 16 se presenta un ejemplo de MD en la herramienta Rapid Miner

CEDULA_ES...	NOMBRE_E...	PRIMER_AP...	CAMPUS_ID	COLEGIO_ID	ESTADO_CI...	NOMBRE_E...	CARRERA_ID	DISCAPACID...	PRI_NIVEL	EGRESA
1717277634	JOSE LEONA...	CEVALLOS	1	1064	5	SOLTERO	4	7	SI	NO
1714489158	SERGIO DAVID	PAZMInO	1	1431	5	SOLTERO	4	7	SI	NO
1725580029	DAYSI PAMELA	CASTELLAN...	1	121	5	SOLTERO	4	7	SI	NO
1723562391	DANILO ALE...	SALAZ	1	1377	5	SOLTERO	4	7	SI	NO
1718266842	JORGE STAL...	QUIMBIULCO	1	1689	5	SOLTERO	4	7	SI	NO
1722318548	MARCELO AL...	ROSERO	1	263	5	SOLTERO	4	7	SI	NO
1804471322	JAIME ABRA...	BERMEO	1	2009	5	SOLTERO	4	7	SI	NO
1726818618	JUAN JOSE	ESTRELLA	1	1230	5	SOLTERO	4	7	SI	NO
401481007	NELSON AN...	GUERRON	1	1837	5	SOLTERO	4	7	SI	NO
1722720784	JONATHAN A...	CORRALES	1	697	5	SOLTERO	4	7	SI	NO
1712987831	MIGUEL IGN...	CARDENAS	1	434	5	SOLTERO	4	7	SI	NO

Figura 16. Rapid Miner

A continuación en la figura 17 se presenta los datos estadísticos de la MDE realizada en la herramienta Rapid Miner.

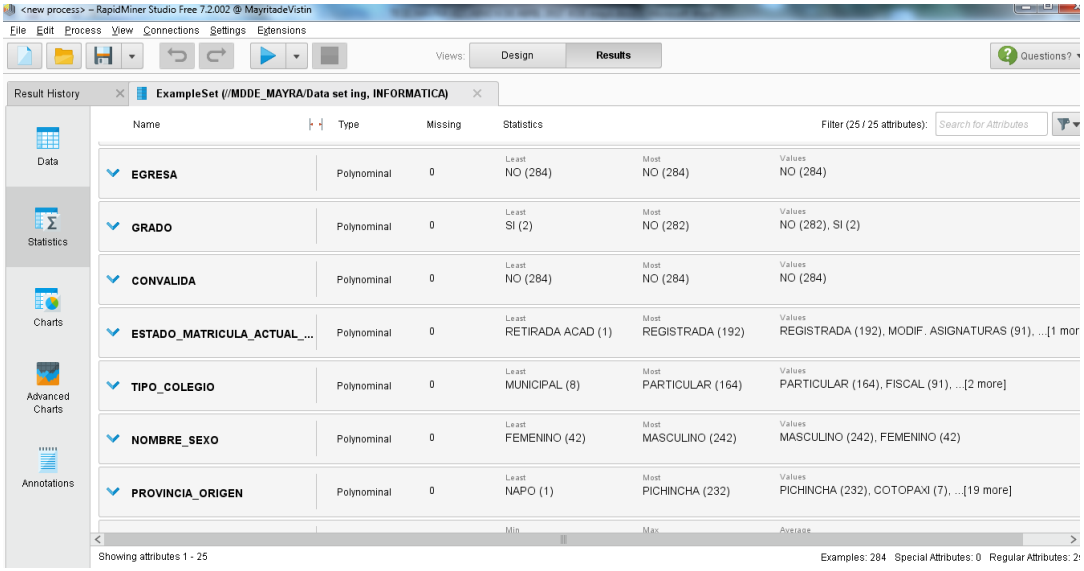


Figura 17. Datos estadísticos Rapid Miner

2.11.6 ORANGE 3

Permite analizar detalladamente gran cantidad de datos con una programación visual, rápida y versátil para un análisis exploratorio de datos, de manera automática con el objetivo de explicar el comportamiento de los datos en un determinado contexto. Esto a su vez permite reunir y transformar los datos en información, de forma que se pueda optimizar el proceso en la toma de decisiones de los negocios.

En la Figura 18 se presenta un ejemplo de MDE realizado en la herramienta Orange.

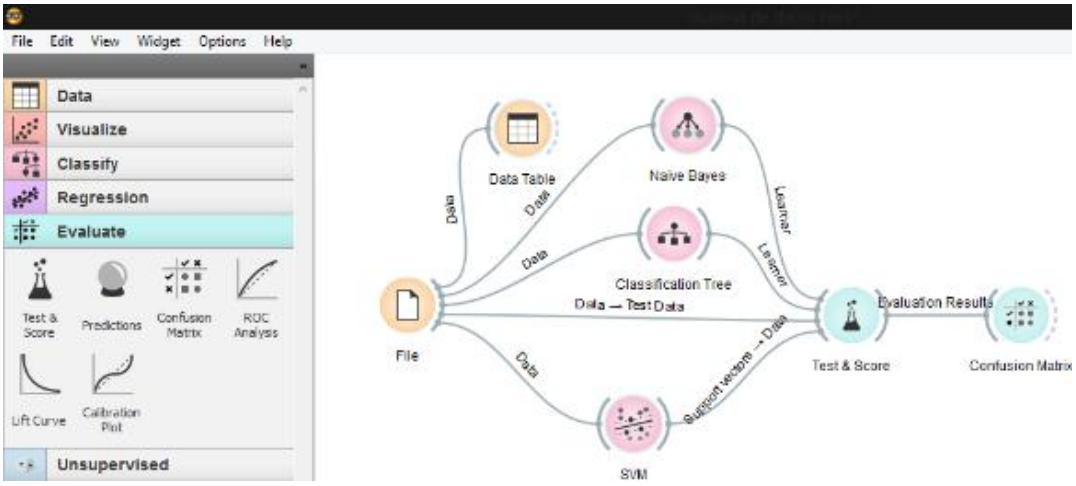


Figura 18. Orange 3

En la figura 19 se presenta la tabla de MDE que fue realizada en Orange.

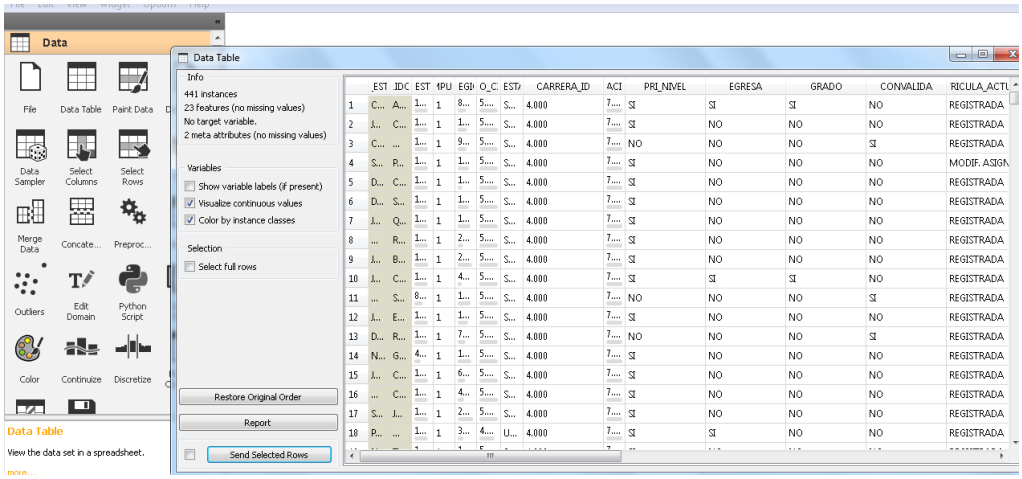


Figura 19. Tabla de MDE en la herramienta Orange

METODOLOGÍA

3 METODOLOGÍA

En este proyecto se va a realizar experimentos de MDE, con las clases deserción y graduación de los estudiantes de la carrera de Ingeniería Informática y Ciencias de la Computación con registros de estudiantes desde el año 2002 hasta el año 2015 además se realizó un análisis de factibilidad técnica de 3 herramientas de MDE para elegir la que permita realizar un mejor análisis de métodos y algoritmos para este estudio. En el caso de estudio se presenta un escenario establecido para el análisis de datos. En este caso de estudio se compara resultados de evaluación de algoritmos para presentar aquel que tenga el mejor desempeño para realizar MDE.

Se utiliza diferentes criterios de representación y aplicación de algoritmos de clasificación como árboles de decisión, redes bayesianas, metaclasificadores y reglas de decisión. Se escogen estas alternativas de análisis por ser las más usadas en diferentes artículos científicos, uno de ellos es el artículo de análisis de deserción y permanencia (Suénaga, K.,& Eckert, R., 2015).

La fuente de datos primaria del DW contiene información proporcionada por los estudiantes cuando ingresan a la universidad. Además, contiene datos que se generan durante el periodo de estudios.

Las clases influyentes son la graduación y la deserción de estudiantes. La deserción se clasifica según la cantidad de semestres matriculados, es decir si el estudiante no se matricula por más de 2 semestres consecutivos entra en estado de deserción tal como indica el reglamento de régimen académico del Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES) entidad pública que realiza procesos continuos de evaluación y acreditación.

3.1 DESARROLLO DE LA METODOLOGÍA

En el presente proyecto de titulación se va a usar la metodología KDD para realizar la MDE, con un análisis automático de grandes cantidades de datos, presentando los resultados con una, revisión de anomalías y reglas de asociación. “El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, BDD, y visualización donde se extrae conocimiento o información” (Usama, F.,& Evangelos, S., 1997, pág. 54).

La metodología KDD permite realizar la evaluación de los algoritmos mediante las siguientes etapas:

3.1.1 SELECCIÓN DE DATOS

La información que se quiere investigar está sobre un dominio de la organización que se encuentra almacenada en un DW de Microsoft SQL Server 2008 R2.

Este proceso tiene como objetivo, elaborar una lista de los datos para identificar los atributos e instancias. Se utiliza diferentes estrategias para manejar los datos inconsistentes o que están fuera de rango para finalmente tener una estructura adecuada.

3.1.2 PRE-PROCESAMIENTO

Se analiza los datos de estudiantes de ingeniería desde el año 2002 hasta el año 2015. En este repositorio existen 441 estudiantes, de los cuales 350 estudiantes ingresaron a la universidad desde el primer semestre y 91 ingresaron convalidando materias.

En la BDD se tiene varios atributos con sus respectivos id los cuales ayudan con el proceso de identificar los casos de deserción y graduación.

Los datos que nos presenta la BDD son datos reales pero los mismos en varias tablas son datos impuros esto quiere decir que para la construcción del dataset los datos necesitan ser preparados y analizados con eficiencia para mejorar el proceso de MDD. Por ejemplo los datos en la tabla año ingreso se presentan datos del año que ingreso el estudiante pero si el estudiante convalido materia para ingresar a la universidad el campo muestra como año de ingreso 1900. Por lo tanto en la fase de transformación se crea una regla de calidad que aporte mayor información en el dataset con varios atributos como primer nivel (PRI_NIVEL) el cual muestra SI o NO ingreso desde primer nivel a la universidad, del mismo modo se crea el campo egresado y el grado.

Una vez que se tenga una estructura de datos adecuada se procede a realizar la limpieza y pre-procesamiento.

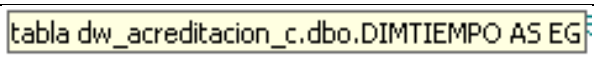
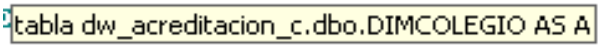
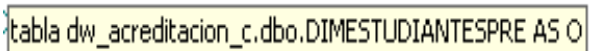
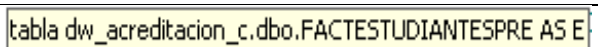
3.1.3 TRANSFORMACIÓN

Con la estructura de datos existente se realiza el tratamiento preliminar de los datos, transformación y generación de nuevas variables evaluando el desempeño de los métodos para que se ajusten al escenario propuesto realizando una comparación con el caso de estudio y experimento.

Para la creación del dataset la tarea principal es identificar la estructura natural de los datos, posteriormente se crean nuevas tablas donde se realizan operaciones de agregación, normalización y consolidacion de datos. Por ejemplo se crea una tabla con el nombre numero_períodos que tiene el atributo num el cual indica el número de períodos matriculados según el id del estudiante. El procedimiento que se realiza para mostrar la deserción de estudiantes es calcular desde que el estudiante ingresa a primer nivel o convalidando materias. En caso de que deje de estudiar más de dos semestres el estudiante se considera en deserción.

Finalmente se crea una sola tabla general con el nombre UTE_ING_INFORM_QUITO que tiene relación con otras tablas con diferentes columnas unas de ellas se detallan a continuación en la Tabla 3.

Tabla 3. BDD dw_acreditacion_c de la UTE

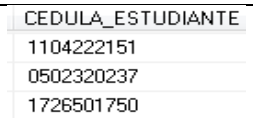
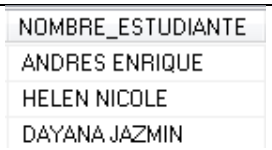
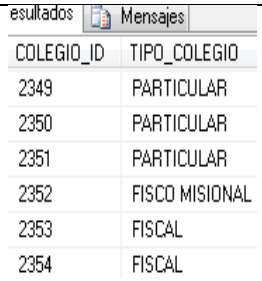
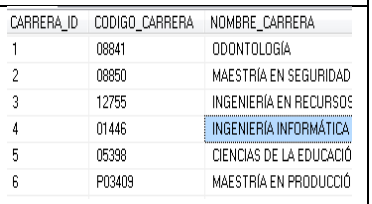
Base de datos dw_acreditacion_c de la UTE	
Nombre de la tabla	**Imagen de la relación con la BDD**
dbo.DIMSEXO	
dbo.DIMTIEMPO GRADO	
dbo.DIMTIEMPO EGRESA	
dbo.DIMTIEMPO INGRESO	
dbo.DIMCOLEGIO	
dbo.DIMESTUDIANTESPRE	
dbo.FACTESTUDIANTESPRE	

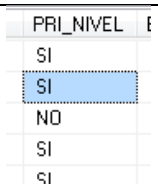
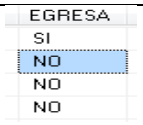
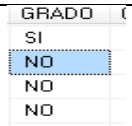
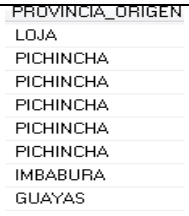
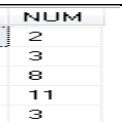
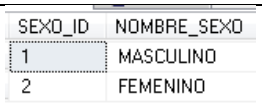
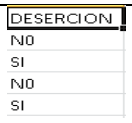
Los escenarios para la MDE son 2: grado y deserción. Por lo tanto, el dataset debe contener estos datos para el análisis de MDE.

El dataset contiene 19 atributos se trabaja con 18 atributos porque si analizo la deserción de estudiantes se elimina a los estudiantes graduados y al analizar la graduación de estudiantes se elimina a los que están en deserción.

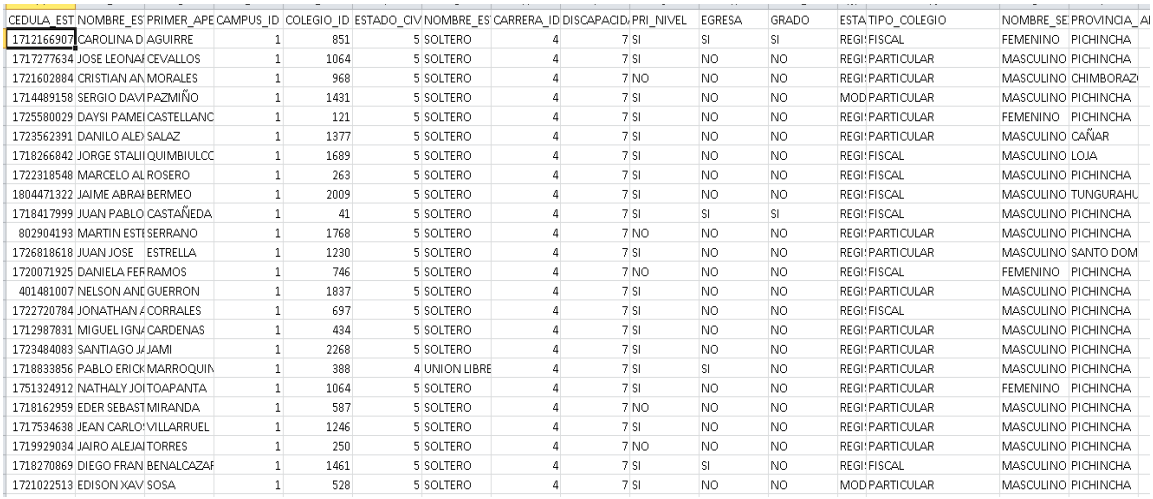
En la tabla 4 se presenta los atributos del dataset.

Tabla 4 Atributos del dataset

N°	**Atributo**	**Estandarización**	**Descripción del campo en la BDD**
1	Cedula del estudiante	CEDULA_ESTUDIANTE	
2	Nombres del estudiante	NOMBRE_ESTUDIANTE	
3	Apellido del estudiante	PRIMER_APELLIDO_ESTUDIANTE	
4	ID del campus universitario	CAMPUS_ID	
5	Colegio ID	COLEGIO_ID	
6	Tipo colegio	TIPO_COLEGIO	
7	Estado civil	ESTADO_CIVIL_ID	
8	Nombre del estado civil	NOMBRE_ESTADO_CIVIL	
9	ID de la carrera	CARRERA_ID	
10	Discapacidad	DISCAPACIDAD	

11	Condición de ingreso a primer nivel	PRI_NIVEL	
12	Condición de egresado	EGRESA	
13	Condición de graduado	GRADO	
14	Estado matricula	ESTADO_MATRICULA_ACTUAL_ESTUDIANTE	
15	Provincia de origen	PROVINCIA_ORIGEN	
16	Número de períodos matriculados	NUM	
17	Nombre del sexo	NOMBRE_SEXO	
18	Año ingreso	ANIO_INGRESO	
19	Deserción	DESERCION	

A continuación en la Figura 20 se presenta el dataset creado que contiene 441 registros de todos los estudiantes de Ingeniería Informática que ingresaron a la universidad entre el año 2002 y 2015.



CEDULA_EST	NOMBRE_ES	PRIMER_APE	CAMPUS_ID	COLEGIO_ID	ESTADO_CIV	NOMBRE_ES	CARRERA_ID	DISCAPACID	PRI_NIVEL	EGRESA	GRADO	ESTA	TIPO_COLEGIO	NOMBRE_SE	PROVINCIA_AI
1712166907	CAROLINA D	AGUIRRE	1	851	5	SOLTERO	4	7	SI	SI	SI	REGI	FISCAL	FEMENINO	PICHINCHA
1717277634	JOSE LEONAI	CEVALLOS	1	1064	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1721602884	CRISTIAN AN	MORALES	1	968	5	SOLTERO	4	7	NO	NO	NO	REGI	PARTICULAR	MASCULINO	CHIMBORAZ
1714489158	SERGIO DAVI	PAZMIÑO	1	1431	5	SOLTERO	4	7	SI	NO	NO	MOD	PARTICULAR	MASCULINO	PICHINCHA
1725580029	DAYSI PAMEI	CASTELLANC	1	121	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	FEMENINO	PICHINCHA
1723562391	DANILO ALE	SALAZ	1	1377	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	CAÑAR
1718266842	JORGE STALI	QUIMBIULCC	1	1689	5	SOLTERO	4	7	SI	NO	NO	REGI	FISCAL	MASCULINO	LOJA
1722318548	MARCELO AL	ROSERO	1	263	5	SOLTERO	4	7	SI	NO	NO	REGI	FISCAL	MASCULINO	PICHINCHA
1804471322	JAIME ABRA	BERMEO	1	2009	5	SOLTERO	4	7	SI	NO	NO	REGI	FISCAL	MASCULINO	TUNGURAHU
1718417999	JUAN PABLO	CASTAÑEDA	1	41	5	SOLTERO	4	7	SI	SI	SI	REGI	FISCAL	MASCULINO	PICHINCHA
802904193	MARTIN EST	SERRANO	1	1768	5	SOLTERO	4	7	NO	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1726818618	JUAN JOSE	ESTRELLA	1	1230	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	SANTO DOM
1720071925	DANIELA FER	RAMOS	1	746	5	SOLTERO	4	7	NO	NO	NO	REGI	FISCAL	FEMENINO	PICHINCHA
401481007	NELSON ANI	GUERRON	1	1837	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1722720784	JONATHAN A	CORRALES	1	697	5	SOLTERO	4	7	SI	NO	NO	REGI	FISCAL	MASCULINO	PICHINCHA
1712987831	MIGUEL IGN	CARDENAS	1	434	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1723484083	SANTIAGO J	JAMI	1	2268	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1718833856	PABLO ERIC	MARROQUIN	1	388	4	UNION LIBRE	4	7	SI	SI	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1751324912	NATHALY JOI	TOAPANTA	1	1064	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	FEMENINO	PICHINCHA
1718162959	EDER SEBAST	MIRANDA	1	587	5	SOLTERO	4	7	NO	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1717534638	JEAN CARLO	VILLARRUEL	1	1246	5	SOLTERO	4	7	SI	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1719929034	JAIRO ALEJA	TORRES	1	250	5	SOLTERO	4	7	NO	NO	NO	REGI	PARTICULAR	MASCULINO	PICHINCHA
1718270869	DIEGO FRAN	BENALCAZAF	1	1461	5	SOLTERO	4	7	SI	SI	NO	REGI	FISCAL	MASCULINO	PICHINCHA
1721022513	EDISON XAV	SOSA	1	528	5	SOLTERO	4	7	SI	NO	NO	MOD	PARTICULAR	MASCULINO	PICHINCHA

Figura 20. Dataset

3.1.4 MINERÍA DE DATOS EDUCACIONALES

Con la MDE se va a explorar los datos de interés del DW con el objetivo de encontrar patrones que se repiten y reglas que expliquen el comportamiento de los datos. Durante el desarrollo del proyecto se usan diferentes aplicaciones del software donde en cada etapa se puede analizar los resultados mediante estadísticas de visualización de datos o de inteligencia artificial. Básicamente este proceso se realiza para comprender el repositorio de los datos atribuyendo un significado especial para convertirlo en conocimiento.

Se va a trabajar con el dataset creado el cual tiene el conjunto de datos que posteriormente nos presentará información clasificada y comprensible para la toma de decisiones.

3.1.5 INTERPRETACIÓN Y EVALUACIÓN

Se realiza una evaluación de los resultados interpretando e identificando cada uno de los patrones más interesantes.

En Weka los parámetros asociados del clasificador que se usan para la evaluación del dataset son los siguientes:

- **Cross-validation:** se evalúa al dataset con todos los atributos y con la selección de 8 atributos se realizará una validación cruzada con el número de particiones.

- **Percentage split:** para el porcentaje de división se evalúa los datos del dataset con el 66% con todos los atributos y con 7 atributos, se realiza la clasificación y con la otra parte de datos se realizarán las pruebas.

3.2 DATOS GENERALES PARA MDE

En la Tabla 5 se presenta el dataset creado para el estudio de la graduación y deserción de estudiantes, se incluye la descripción y la denominación estandarizada de cada atributo con el tipo de dato y valores posibles para los nominales.

Se realiza el análisis de los datos con todos los estudiantes de la carrera de Ingeniería.

Tabla 5. Atributos seleccionados y estandarizados del dataset

ATRIBUTOS SELECCIONADOS	ESTANDARIZACIÓN	OPCIÓN DE RESULTADO
Condición de deserción	DESERCION	SI / NO
Condición de graduado	GRADO	SI / NO
Condición de egresado	EGRESA	SI / NO
Condición de ingreso a primer nivel	INGRESO	SI / NO
Número de períodos matriculados	NUM	Numérico
Estado matricula actual del estudiante	ESTADO_MATRICULA_ACTUAL_ESTUDIANTE	Registrada / modif. Asignaturas
Campus ID	CAMPUS_ID	Numérico Quito (1)
Discapacidad	DISCAPACIDAD	Numérico

Para la explicación detallada de la creación del dataset se anexa un manual de construcción del dataset. Los datos están separados por provincias, estudiantes que convalidan y estudiantes que están desde primer nivel de la carrera de Ingeniería Informática.

3.3 ANÁLISIS DE FACTIBILIDAD TÉCNICA

El análisis de factibilidad de las herramientas es de vital importancia porque nos permite identificar que herramienta de MDE es más efectiva indicando los criterios de evaluación. Se realizan varias pruebas con validez estadística validando las características del modelo y de la herramienta.

Se evalúa el dataset con todos los atributos y también se separa los atributos escogiendo solo ocho, los cuales son los más indicados para la predicción de la deserción y graduación estudiantil.

Se realiza la MDE con las tres herramientas Weka, Orange 3 y Rapid Miner las cuales fueron seleccionadas en base a estudios realizados y la representación del conocimiento que muestra cada una de ellas.

3.4 COMPARACIÓN DE HERRAMIENTAS DE MDE

Para la interpretación de los resultados se realiza varias pruebas con el dataset por lo tanto es necesario equilibrar los resultados con el costo, beneficios, desventajas y formas de uso de cada una de las herramientas. Los resultados se presentan en base al uso de las herramientas de minería de datos en una institución educativa realizada a los estudiantes de bachillerato, el tiempo de uso fue por tres meses y los resultados fueron satisfactorios indicando que no todas las herramientas tienen el mismo rendimiento.

En la Tabla 6 se presenta la comparación de 3 herramientas de minería de datos.

Tabla 6. Comparación de 3 herramientas de MDE

MDE	Costo	Beneficios	Desventajas	Formas de uso
Weka	Licencia GPL General Public License	Reducción de las necesidades de almacenamiento Visualización y comprensión de datos. Funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible.	Existe poca documentación sobre el uso de Weka dirigido al usuario.	Interfaz gráfica por algoritmos y se puede acceder fácilmente
Orange	Licencia GPL General Public License	Beneficios creados en C++ Multilenguaje Implementa algoritmos.	Tiene como ejemplos videos tutoriales pero no son tan claros ya que al momento de cargar el data set solo adjunta archivos .tab. Los ejemplos de uso de la herramienta son pocos	Interfaz gráfica por algoritmos
Rapid Miner	Licencia AGPL Affero General Public License Varias versiones con pago	Más de 500 operadores Compatible con algoritmos de Weka. Existen varios tutoriales de uso básico de la herramienta	Muchas veces los valores del dataset no todos son numéricos y da problemas al momento de presentar la matriz de correlación.	Interfaz gráfica para diseñar y ejecutar flujos de trabajo de análisis.

EVALUACIÓN DE HERRAMIENTAS DE MINERÍA DE DATOS

Este proyecto de titulación justifica su procedimiento de acuerdo a lo encontrado en la factibilidad técnica de la Tabla 7.la cual presenta el análisis de las herramientas MD sobre 10 puntos.

Tabla 7. Factibilidad técnica de las herramientas de MDE

Descripción de la herramienta MD		**WEKA**		**RAPID MINER**		**ORANGE 3**	
Criterio	Peso General 100%	WEKA Puntaje sobre 10 puntos	% WEKA	RAPID MINER Puntaje sobre 10 puntos	% Rapid Miner	ORANGE Puntaje sobre 10 puntos	% Orange
Usuarios que saben usar la plataforma siguiendo un manual de usuario.	20%	7	14%	6	12%	7	14%
Costo del software 10 p si es gratis y 5 p si es pagado.	10%	10	10%	5	5%	10	10%
Facilidad de mantenimiento.	10%	8	8%	8	8%	8	8%
Comprensión y presentación de los datos en la herramienta.	10%	8	8%	7	7%	9	9%
Compatibilidad con archivos CSV y xls.	10%	9	9%	9	9%	9	9%
Presentación de ejemplos para el uso de la herramienta.	10%	8	8%	9	9%	7	7%
Facilidad de resultados.	10%	9	9%	9	9%	8	8%
Visualización de resultados (interfaz gráfica).	20%	9	18%	8	16%	9	18%
	100%		**84%**		**75%**		**83%**

Según los resultados obtenidos se considera que la herramienta más adecuada para minería de datos es WEKA.

A pesar de que existe une evaluación muy similar entre WEKA y ORANGE se optó por trabajar con WEKA debido a que esta herramienta tiene una menor curva de aprendizaje para el responsable de los experimentos por cuanto esta herramienta fue usada como parte de las herramientas de clase.

3.5 IMPLEMENTACIÓN DE LA MDE

Se realiza el proceso de MDE en las 3 herramientas Weka, Rapid Miner, Orange 3. Según los resultados obtenidos anteriormente en la tabla 7 se considera que la herramienta más adecuada para MDE es Weka por lo tanto se va a explicar el procedimiento en la herramienta Weka. Finalmente, se podrá determinar con qué algoritmos se obtienen los mejores resultados.
A continuación se presenta los clasificadores con los cuales se realiza la MDE.

3.5.1 SELECCIÓN DE CLASIFICADORES

Existen ocho familias de clasificadores, en este trabajo y de acuerdo a un estudio realizado por tres profesores especializados en MDE concluyen que los clasificadores más utilizados son cuatro: los bayesianos, los metaclasificadores, las reglas y los árboles de decisión (Marquez, C.,Romero, C. & Ventura,S., 2012). Es por esto que en este trabajo se realiza experimentos con estos clasificadores.

A continuación se explicará cada uno de estos clasificadores y se pondrá un ejemplo de la clasificación para la comparación de los resultados.

3.5.1.1 Bayesianos

Maximiza la probabilidad de que una nueva instancia del dataset se clasifique correctamente presentando una medida probabilística en los resultados de la clasificación.

Naïve Bayes: Inicia con la hipótesis de que todos los atributos son independientes entre sí, por ejemplo la clase deserción y grado siendo los otros atributos las hojas o nodos que tienen como único origen a la variable clase.

3.5.1.2 Metaclasificadores

En Weka se incluye todos aquellos clasificadores complejos, es decir, aquellos que se obtienen mediante composición de clasificadores simples o que incluyen algún pre procesamiento de los datos.

Stacking: Se basa en la combinación de modelos, el cual construye un conjunto de diferentes algoritmos de aprendizaje presentando conjuntos de aprendizaje distintos.

3.5.1.3 Reglas

Existen diversos métodos para generar reglas de clasificación con diferentes conjuntos de entrenamiento.

One R: Este es uno de los clasificadores más sencillos y rápidos, aunque en ocasiones sus resultados son sorprendentemente buenos en comparación con algoritmos mucho más complejos. Genera una regla por cada atributo y escoge la del menor error. Si hay atributos numéricos, busca los umbrales para hacer reglas con mejor tasa de aciertos (Bouckaert, C.,& Peter, R., 2011).

3.5.1.4 Árboles de decisión

Los árboles son una manera práctica para visualizar la clasificación de un conjunto de datos. Entre ellos tenemos:

J48: Es una implementación del algoritmo C4.5, uno de los algoritmos de MDE que más se ha utilizado en multitud de aplicaciones. Uno de los parámetros más importantes de este algoritmo es el factor de confianza para la poda (confidence level). Una explicación simplificada es la siguiente: para cada operación de poda, define la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en

los errores de predicción antes y después de podar sea más significativa para no podar (Bouckaert, C.,& Peter, R., 2011).

Randomtree: Presenta árboles uniformes dibujados "al azar" que significa que cada árbol tiene una posibilidad igual de ser probado con permutaciones arbitrarias.

3.5.2 ÍNDICE KAPPA

Este índice kappa es una medida entre las categorías pronosticadas por el clasificador y las categorías observadas, que tiene en cuenta las posibles concordancias debidas al azar.

- Si el valor es 1: lista ordenada y clasificada perfecta.
- Si el valor es mayor que 0: lista ordenada y clasificada con un grado de concordancia.
- Si el valor es 0: lista ordenada y clasificada al azar.
- Si el valor es negativo: lista ordenada y clasificada menor.

3.5.3 INDICADORES DE ERRORES

Los indicadores asociados al error de la clasificación son:

- Mean absolute error:

$$\frac{1}{N}\sum_i |di|$$

- Root mean squared error:

$$\sqrt{\frac{1}{N}\sum_I di^2}$$

- Relative absolute error

$$\frac{\textbf{Mean absolute error}}{\textbf{Root mean squared error (Zero R)}}\ \textbf{X 100}$$

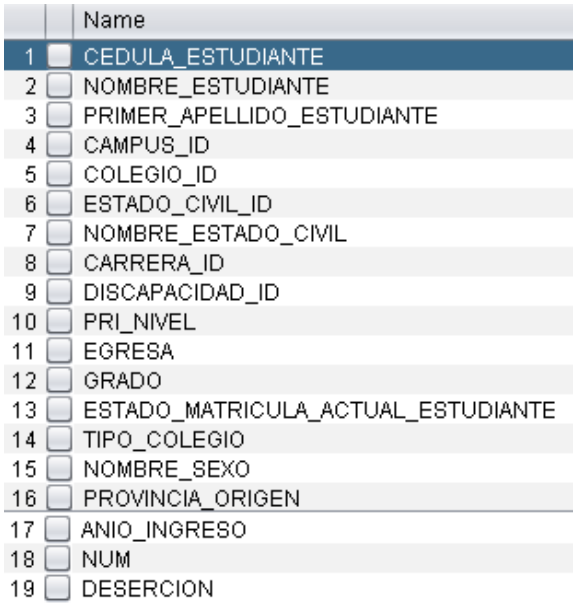
Mientras el resultado del error se aproxime a cero mejor será la clasificación.

RESULTADOS

4 RESULTADOS

En el presente capítulo se muestra el desarrollo del trabajo usando la metodología KDD. Se analiza el dataset creado con la información de la Universidad Tecnológica Equinoccial, institución universitaria creada en 1971, cuya misión es formar con excelencia a profesionales íntegros.

En la herramienta Weka se utiliza el dataset con un formato de datos denominado arff o archivos csv que contienen los atributos, cada fichero consta de 3 partes: cabecera con el nombre del dato, declaración de atributos o variables indicando su tipo y la sección de datos @data. Al abrir el dataset en Weka se puede observar los 19 atributos tal como se muestra a continuación.



	Name
1	CEDULA_ESTUDIANTE
2	NOMBRE_ESTUDIANTE
3	PRIMER_APELLIDO_ESTUDIANTE
4	CAMPUS_ID
5	COLEGIO_ID
6	ESTADO_CIVIL_ID
7	NOMBRE_ESTADO_CIVIL
8	CARRERA_ID
9	DISCAPACIDAD_ID
10	PRI_NIVEL
11	EGRESA
12	GRADO
13	ESTADO_MATRICULA_ACTUAL_ESTUDIANTE
14	TIPO_COLEGIO
15	NOMBRE_SEXO
16	PROVINCIA_ORIGEN
17	ANIO_INGRESO
18	NUM
19	DESERCION

Los resultados se presentan mediante una matriz de confusión la cual indica las categorías y la clase actual de cada algoritmo.

Para la evaluación de la deserción de estudiantes se elimina el atributo grado y para la graduación de estudiantes se elimina el atributo deserción.

4.1 INTERPRETACIÓN DE RESULTADOS DE LA DESERCIÓN DE ESTUDIANTES

Las siguientes interpretaciones son resultados de evaluaciones que fueron usadas con las metodologías y algoritmos para la clase deserción de estudiantes con los siguientes atributos.

Se evalúa a los algoritmos con todos los atributos del dataset y para la mejora de precisión (selección de atributos) se selecciona los siguientes atributos del dataset.

No.	Name
1	NOMBRE_ESTADO_CIVIL
2	PRI_NIVEL
3	EGRESA
4	TIPO_COLEGIO
5	NOMBRE_SEXO
6	ANIO_INGRESO
7	NUM
8	DESERCION

Finalmente, es importante resaltar que estas técnicas no son las únicas que existen, pero si son parte de las más utilizadas.

Se realiza la clasificación de los datos presentándolos en las siguientes tablas:

En la Tabla 8 se presenta los resultados del algoritmo Naïve Bayes.

Tabla 8. Deserción de estudiantes clasificador Naïve Bayes

DESERCIÓN DE ESTUDIANTES					
Clasificador NAÏVE BAYES	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**(%)Error absoluto**
Con todos los atributos del dataset	Validación cruzada	78%	21,99	0,50	0,25
Mejora de precisión (selección de atributos)	Validación cruzada	82,31%	17,68%	0,59	0,23
Con todos los atributos del dataset	División porcentual	82%	18%	0,59	0,23
Mejora de precisión (selección de atributos)	División porcentual	86,66%	13,33%	0,69	0,21

Por los resultados que se tiene con el algoritmo Naive Bayes es la evaluación con selección de los atributos del dataset con el modo de prueba división porcentual.

En la Tabla 9 se presenta los resultados del algoritmo Stacking.

Tabla 9. Deserción de estudiantes Clasificador Stacking

DESERCIÓN DE ESTUDIANTES					
Clasificador STACKING	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**(%)Error absoluto**
Con todos los atributos del dataset	Validación cruzada	60,71%	39,22%	0	0,47%
Mejora de precisión (selección de atributos)	Validación cruzada	60,71%	39,22%	0	0,47%
Con todos los atributos del dataset	División porcentual	62,66%	37,33%	0	0,47%
Mejora de precisión (selección de atributos)	División porcentual	62,66%	37,33%	0	0,47%

Realizando el análisis con el algoritmo Stacking los mejores resultados son con el modo prueba división porcentual y validación cruzada con todos los atributos y con la selección de atributos del dataset.

Para la selección del mejor resultado se recomienda hacer el cuadro de diferencia significativa.

En la Tabla 10 se presenta los resultados del algoritmo One R.

Tabla 10. Deserción de estudiantes Clasificador One R

DESERCIÓN DE ESTUDIANTES					
Clasificador **ONE R**	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**(%)Error absoluto**
Con todos los atributos del dataset	Validación cruzada	39,45%	60,54%	-0,02	0,60
Mejora de precisión (selección de atributos)	Validación cruzada	76,87%	23,12%	0,47	0,23
Con todos los atributos del dataset	División porcentual	36%	64%	-0,05	0,64
Mejora de precisión (selección de atributos)	División porcentual	76,66	23,33	0,48	0,23

En el algoritmo ONE R el mejor resultado es con el modo de división porcentual con la selección de los atributos del dataset.

En la Tabla 11 se presenta los resultados del algoritmo J48.

Tabla 11. Deserción de estudiantes algoritmo J48

DESERCIÓN DE ESTUDIANTES					
Clasificador J48	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**(%)Error absoluto**
Con todos los atributos del dataset	Validación cruzada	79,13%	20,86	0,51	0,30
Mejora de precisión (selección de atributos)	Validación cruzada	94,55%	5,44%	0,88	0,08
Con todos los atributos del dataset	División porcentual	74,66%	25,33%	0,38	0,37
Mejora (selección de atributos)	División porcentual	94%	6%	0,85	0,07

Realizando el análisis de clasificación con el árbol de decisión J48 el mejor resultado es con la selección de atributos y validación cruzada obteniendo un porcentaje del 94,55% de instancias bien clasificadas.

En la Tabla 12 se presenta los resultados del algoritmo Randomtree.

Tabla 12. Deserción de estudiantes algoritmo Randomtree

DESERCIÓN DE ESTUDIANTES					
Clasificador **RANDOMTREE**	Modo de prueba	Instancias bien clasificadas	Instancias mal clasificadas	Índice Kappa	(%)Error absoluto
Con todos los atributos del dataset	Validación cruzada	65,07%	34,92%	0,19	0,39
Mejora de precisión (selección de atributos)	Validación cruzada	91,83%	8,16%	0,82	0,07
Con todos los atributos del dataset	División porcentual	56,66%	43,33%	0,22	0,4
Mejora de precisión (selección de atributos)	División porcentual	91,33%	8,66%	0,81	0,08

El mejor resultado con este algoritmo se da con validación cruzada seleccionando los atributos del dataset.

4.1.1 MEJOR ALGORITMO DE CLASIFICACIÓN PARA LA DESERCIÓN

En la siguiente Tabla 13 se presenta que el mejor algoritmo para la deserción de estudiantes es el J48.

Tabla 13. Mejor algoritmo para la deserción de estudiantes

CLASE DESERCIÓN					
DETALLE	**NAÏVE BAYES**	**STACKING**	**ONE R**	**J48**	**RANDOM TREE**
Instancias bien clasificadas (%)	86,66	62,66	76,87	94,55	91,83
Instancias mal clasificadas %	13,33	37,33	23,12	5,44	8,16
Índice Kappa	0,69	0	0,47	0,88	0,82
(%)Error absoluto	0,21	0,47	0,23	0,08	0,07
Parámetro de clasificación	División porcentual	División porcentual	División porcentual	Validación cruzada	Validación cruzada
Número de atributos del dataset	8	Todos	8	8	8

El mejor algoritmo para la deserción de estudiantes es el J48 con el parámetro de clasificación validación cruzada y evaluando al dataset con 8 atributos además el algoritmo J48 presenta el árbol de decisión que permite al usuario final comprender mejor la clasificación.

4.2 INTERPRETACIÓN DE RESULTADOS DE LA GRADUACIÓN DE ESTUDIANTES

Para la evaluación de los algoritmos de MDE en Weka se realiza el mismo proceso presentado anteriormente con el mismo dataset, en este caso se debe eliminar el campo deserción y se trabaja con el escenario de la clase grado.

Para mejora de precisión de resultados se escoge los siguientes atributos del dataset:

Se realiza la clasificación de los datos con cada uno de los algoritmos presentando los resultados en las siguientes tablas:

En la Tabla 14 se presenta los resultados del algoritmo J48 con la clase grado.

Tabla 14. Graduación de estudiantes árbol de decisión J48

GRADUACIÓN DE ESTUDIANTES					
Clasificador **J48**	Modo de prueba	Instancias bien clasificadas	Instancias mal clasificadas	Índice Kappa	(%)Error absoluto
Con todos los atributos del dataset	Validación cruzada	95,23%	4,76%	0,68	0,06
Mejora de precisión (selección de atributos)	Validación cruzada	95,69%	4,30%	0,70	0,06
Con todos los atributos del dataset	División porcentual	93,33%	6,66%	0,40	0,07
Mejora de precisión	División porcentual	94,66%	5,33%	0,57	0,06

Con el algoritmo J48 como mejor modo de evaluación es la división porcentual con los 7 atributos del dataset.

En la Tabla 15 se presenta los resultados del algoritmo Randomtree con la clase grado.

Tabla 15. Graduación de estudiantes árbol de decisión Randomtree

GRADUACIÓN DE ESTUDIANTES					
Clasificador **RANDOMTREE**	Modo de prueba	Instancias bien clasificadas	Instancias mal clasificadas	Índice Kappa	(%)Error absoluto
Con todos los atributos del dataset	Validación cruzada	93,42%	6,57%	0,32	0,10
Mejora de precisión 7 atributos	Validación cruzada	94,78%	5,21%	0,63	0,05%
Con todos los atributos del dataset	División porcentual	96%	4%	0,67	0,05
Mejora de precisión 7 atributos)	División porcentual	96%	4%	0,64	0,04

El mejor resultado se da con el modo de prueba división porcentual con todos los atributos del dataset y escogiendo los 7 atributos del dataset, al visualizar los resultados en el árbol de decisión se observa que la figura es muy extensa.

En la Tabla 16 se presenta los resultados del algoritmo Naive Bayes con la clase grado.

Tabla 16. Graduación de estudiantes algoritmo Naive Bayes

GRADUACIÓN DE ESTUDIANTES					
Clasificador NAÏVE BAYES	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**(%)Error absoluto**
Con todos los atributos del dataset	Validación cruzada	95,46%	4,53%.	0.69	0,06
Mejora de precisión 7 atributos	Validación cruzada	94,10%	5,89%	0,64	0,06
Con todos los atributos del dataset	División porcentual	94,66	5,33	0,66	0,06
Mejora de precisión 7 atributos	División porcentual	94,66	5,33	0,66	0,06

En este análisis se muestra que el modo de prueba validación cruzada con todos los atributos del dataset tiene los mejores resultados.

En la Tabla 17 se presenta los resultados del algoritmo Stacking con la clase grado.

Tabla 17. Graduación de estudiantes clasificación Stacking

GRADUACIÓN DE ESTUDIANTES					
Clasificador STACKING	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**Error absoluto**
Con todos los atributos del dataset	Validación cruzada	92,74%	7,25%	0	0,13%
Mejora de precisión 7 atributos	Validación cruzada	92,74%	7,25%	0	0,13%
Con todos los atributos del dataset	División porcentual	93,33%	6,66%	0	0,13
Mejora de precisión (selección de atributos)	División porcentual	93,33%	6,66%	0	0,13

Con la clasificación Stacking en los 2 tipos de modo prueba da los mismos resultados.

En la Tabla 18 se presenta los resultados del algoritmo ONE R con la clase grado.

Tabla 18. Graduación de estudiantes clasificación One R

GRADUACIÓN DE ESTUDIANTES					
Clasificador **ONE R**	**Modo de prueba**	**Instancias bien clasificadas**	**Instancias mal clasificadas**	**Índice Kappa**	**Error absoluto**
Con todos los atributos del dataset	Validación cruzada	19,95%	80,04%	0	0,80
Mejora de precisión (selección de atributos)	Validación cruzada	94,78%	5,21%	0,68	0,05
Con todos los atributos del dataset	División porcentual	17,33%	82,66%	0	0,82
Mejora de precisión (selección de atributos)	División porcentual	94,66	5,33	0,66	0,05

Realizando el análisis con la clasificación One R el mejor resultado es validación cruzada con la selección de atributos, en la matriz de confusión muestra el número total de instancias clasificadas que son 441.

4.2.1 MEJOR ALGORITMO DE CLASIFICACIÓN PARA LA GRADUACIÓN

En la Tabla 19 se presenta los mejores resultados de cada algoritmo, para realizar la comparación e indicar cuál es el mejor algoritmo para el análisis del data set con la clase grado.

Tabla 19. Mejores resultados de cada algoritmo con la clase grado

CLASE GRADO					
DETALLE	**NAÏVE BAYES**	**STACKING**	**ONE R**	**RANDOMTREE**	**J48**
Instancias bien clasificadas (%)	95,46	93,33	94,78	96	95,69
Instancias mal clasificadas	4,53	6,66	5,21	4	4,30
Índice Kappa	0,69	0	0,68	0,67	0,70
(%)Error absoluto	0,06	0,13	0,05	0,05	0,05
Parámetro de clasificación	División porcentual	División porcentual	Validación cruzada	Validación cruzada	División porcentual
Número de atributos del dataset	Todos	Todos	7	7	7

Para la graduación de estudiantes el mejor clasificador es el algoritmo J48, debido a que una vez realizado el experimento de evaluación se procede con el descubrimiento de patrones y relaciones en los datos con el árbol de decisión que no es muy extenso los resultados se pueden presentar al usuario de una manera comprensible.

4.3 GENERACIÓN DEL CONOCIMIENTO

Para la generación del conocimiento, los resultados que se obtuvieron después del análisis de los métodos y algoritmos se detallan a continuación.

El dataset de la carrera de Ingeniería en Informática contiene datos desde el año 2002 hasta del año 2015, dentro del dataset existen 441 estudiantes, de los cuales existen 91 estudiantes que ingresan por convalidación y el resto ingresan a la carrera desde el primer nivel.

Dentro de esta muestra de datos se observa lo siguiente:

- Existen 32 estudiantes graduados y 409 están cursando la carrera.
- Se tiene 235 estudiantes que se han matriculado más de 8 niveles por lo tanto, se observa que existen pocos estudiantes en los primeros niveles de la carrera.
- Existen 11 estudiantes de sexo masculino que no se gradúan y están matriculados 11 períodos por lo tanto se recomienda hacer un seguimiento a estos estudiantes para ver si ya van a culminar la carrera.
- Existen 19 estudiantes de los primeros niveles en deserción que han ingresado desde el primer nivel.
- Existen 25 estudiantes que han convalidado y han ingresado a los últimos niveles de la carrera, estos estudiantes han estudiado menos de 5 períodos. Además existen 3 estudiantes graduados que convalidan y estudian más de 5 períodos.
- Existen 123 estudiantes que desertan e ingresaron en primer nivel.
- Todos los estudiantes con estado civil casado desertan.
- En el año 2014 existen 2 estudiantes graduados que ingresan desde 1er nivel.
- En la carrera de Ingeniería Informática existen más estudiantes de colegios particulares en total 249 y de ellos 108 estudiantes están en deserción.
- Existen 156 estudiantes de colegios fiscales, 26 estudiantes de colegios fiscomisionales y 10 estudiantes de colegios municipales ninguno de estos colegios tiene algún caso de deserción.
- La mayoría de estudiantes de la carrera son de la provincia de Pichincha en total 354, de la provincia de Chimborazo 5, de la provincia de Cañar 6 y de la

provincia de Loja 6, Azuay 1, Los Ríos 2, Napo 2, Orellana 2, Zamora Chinchipe 1, El Oro 2, Esmeraldas 7, Carchi 13, Guayas 3, Cotopaxi 9, Manabí 3, Imbabura 7, Galápagos 1, Bolívar 7, Santo Domingo 5, Tungurahua 4.

- Se observa que no existen muchos estudiantes extranjeros de España existe 1 estudiante de Colombia 3 estudiantes y de Estados Unidos 2 estudiantes.

5 CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

Se realizó la minería de datos educacionales con la implementación de la metodología KDD usando métodos y algoritmos creando un dataset con estructura comprensible

El análisis de factibilidad técnica realizado muestra que la herramienta Weka es la más recomendada.

Los resultados obtenidos de la MDE son de mucha utilidad para la institución porque se podrá mejorar los procesos de: enseñanza, aprendizaje, matriculación, evaluación.

Los resultados del análisis de graduación y deserción (abandono) permitirán a la institución tomar acciones preventivas y correctivas.

Mediante la realización de varios experimentos se evaluó las técnicas, algoritmos y herramientas de MDE existentes. Estos análisis se realizaron sobre la clase deserción y graduación usando 5 algoritmos de clasificación.

Para la graduación de estudiantes el mejor clasificador es el algoritmo Randomtree pero presentar un árbol muy extenso se escoge al algoritmo J48 que tiene una diferencia significativa en sus resultados.

Para la deserción de estudiantes el mejor clasificador es el algoritmo J48.

Con los modelos de los árboles de decisión Randomtree y J48 se pudo obtener conocimiento para estimar modelos útiles y de este modo presentar parámetros secuenciales de decisión con resultados y probabilidades acertadas, el escenario de evaluación que se tomó como base es la tasa de graduación y deserción de los estudiantes de la carrera de Ingeniería Informática y Ciencias de la UTE.

RECOMENDACIONES

Se debe estructurar un dataset consistente para la implementación analítica de métodos y algoritmos de MDE.

Definir el caso de estudio es un aspecto necesario para que la MDE sea exitosa.

En la creación del dataset se debe ser muy minuciosos con el proceso de interpretación y transformación de datos, para no tener problemas al momento de cargar los datos en la herramienta Weka.

Para el descubrimiento del conocimiento en grandes bases de datos se recomienda la metodología KDD.

Para realizar MDE con la herramienta Weka y analizar varios métodos de combinación para obtener diferentes resultados se necesita más recursos de tiempo y memoria del computador ya que el modelo sería más grande y se perdería la comprensibilidad del mismo.

Para futuros estudios de minería de datos para trabajos similares se recomienda la herramienta Weka y para trabajos que requieran un mayor análisis gráfico se recomienda usar la herramienta Orange 3 por presentar una estructura gráfica de los datos la cual permite al usuario interpretar las respuestas de forma más clara.

Se recomienda realizar varias pruebas de los resultados obtenidos para tener respuestas consistentes ya que de eso dependen las decisiones que la institución podría tomar.

6 BIBLIOGRAFÍA

Andrew, R. (2007). Estrategia y sistemas de información. España: McGraw Hill.

Arjonilla, D.,& Medina, G. (2007). La gestión de los sistemas de información. Madrid: Pirámide.

Bouckaert, C.,& Peter, R. (2011). WEKA Manual for versión 3.7.8. *Manual WEKA*, 38.

Cristobal, R.,& Ventura, S. (2012). Data Mining and Knowledge Discovery. En *Data mining in education* (págs. 12-27). Cordoba España: Witold Pedrycz.

García, A. (2008). Estadística aplicada. En *conceptos básicos y ejemplos* (pág. 298). UNED.

Guevara, J.,& Valencia, J. (2007). *Data Warehouse para el análisis académico.* España: Pearson.

Han, J., Kamber, M.,& Mark, A. (2010). Dataminig. En *Data Mining conceptos y técnicas* (pág. 61). Madrid: Third edition.

Hernández, J., Ramirez, M.,& Ferri , C. (2004). *Introducción a la minería de datos.* España: Pearson.

IBM. (2012). *IBM Intelligent Miner.* Recuperado el 18 de 05 de 2016, de https://www.ibm.com/ec-es/

Inmon, W. H. (2002). Building the data warehouse. John Wiley and Son.

Inmon,W.,& Dan, L. (2014). *Data Scientist: Big Data, Data Warehouse and Data Vault.* Dallas USA: Kindle.

Jarke, M., Lenzerine, M., Vassiliu, Y.,& Vassiliadis,P. (2002). Fundamentals of Data Warehouses. Verlag - New York: Springer Science & Business Media.

Korth, H.,& Silberschatz, A. (2006). BDD. En A. D. Mining. Mexico: Interamericana.

Lakshman, B. (2013). Open source data warehousing and business intelligent. En *Data warehouse.* Canadá: Wiley.: CRC Press.

Marques, M. (11 de 08 de 2014). *CRISP-DM, Una metodología para proyectos de Minería de Datos.* España: Createspace.

Marquez, C.,Romero, C. & Ventura,S. (3 de Nov de 2012). Predicción del fracaso escolar mediante técnicas de minería de datos. pág. 9.

Peña, A. (2014). Educational data mining based analysis of recent works. Pergamon: Expert Systems.

Piorno, J. (15 de 02 de 2010). *Diseño de un nuevo clasificador supervisado para la mineria de datos*, pág. 80.

RapidMiner, Community. (2007). *https://my.rapidminer.com/nexus/account/index.html#downloads*. Recuperado el 08 de 08 de 2016, de https://my.rapidminer.com

Riquelme, J. C. (2010). La informática del futuro. *Jornadas Imaginática*, 26-29.

Santa Cruz, R. (2016). *Técnicas y Herramientas de la minería de datos*. Recuperado el 01 de 06 de 2016, de https://santacruzramos.wikispaces.com/3.4.5+T%C3%A9cnicas+y+herramientas+de+la+miner%C3%ADa+de+datos.

SAS Institute Inc. (2010). *sas.com*. Recuperado el 05 de 10 de 2016, de https://www.sas.com/en_us/software/enterprise-miner.html

Silberschatz, A. (2007). Arquitectura de base de datos. En *Arquitectura de base de datos* (págs. 83-107). España: Mcgraw-hill 6ta Edition.

Silva, M. (10 de 08 de 2007). Recuperado el 12 de 05 de 2016, de Minería de datos: http://exa.unne.edu.ar/informatica/SO/Mineria_de_Datos_y_KDD.pdf

Suénaga, K.,& Eckert, R. (2015). *http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062015000500002.* Recuperado el 10 de 11 de 2016, de http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062015000500002

Usama, F.,& Evangelos, S. (1997). Recuperado el 29 de 04 de 2016, de Data Mining and Knowledge Discovery: http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1

Vieira, L., Ortiz, L.,& Ramirez, S. (2001). *Metodología KDD y minería de datos.* Rio de Janeiro: E papers.

Visual Studio .VisualBasic.net. (10 de 10 de 2016). *Canal visual basic .net*. Recuperado el 02 de 02 de 2017, de http://www.canalvisualbasic.net/otros/data-warehousing/

ANEXOS

ANEXO 1. CASO 1 SQL SERVER

El siguiente caso permite evaluar una serie de condiciones y devolviendo en uno de los resultados la deserción de estudiantes que existe hasta el año 2016.

```
CASE
(MES_INGRESO == 2 OR MES_INGRESO = 3) AND 2016 - ANIO_INGRESO * 2 -
N.NUM + 1 >2
WHEN TRUE THEN 'SI'
IIF((MES_INGRESO = 9 OR MES_INGRESO = 10) AND 2016 - ANIO_INGRESO *
2 - N.NUM  >2, 'SI','NO')
AS DESERCION
FROM dbo.ING_INFORM_QUITO GQ,
dbo.NUMEROS_PERÍODOS N
WHERE GQ.ESTUDIANTE_ID = N.ESTUDIANTE_ID
AND EGRESA  ='NO'
AND ANIO_INGRESO > 1900
AND CONVALIDA = 'NO'
```

ANEXO 2. CASO 2 SQL SERVER

Dataset con registros de estudiantes hasta el año 2015

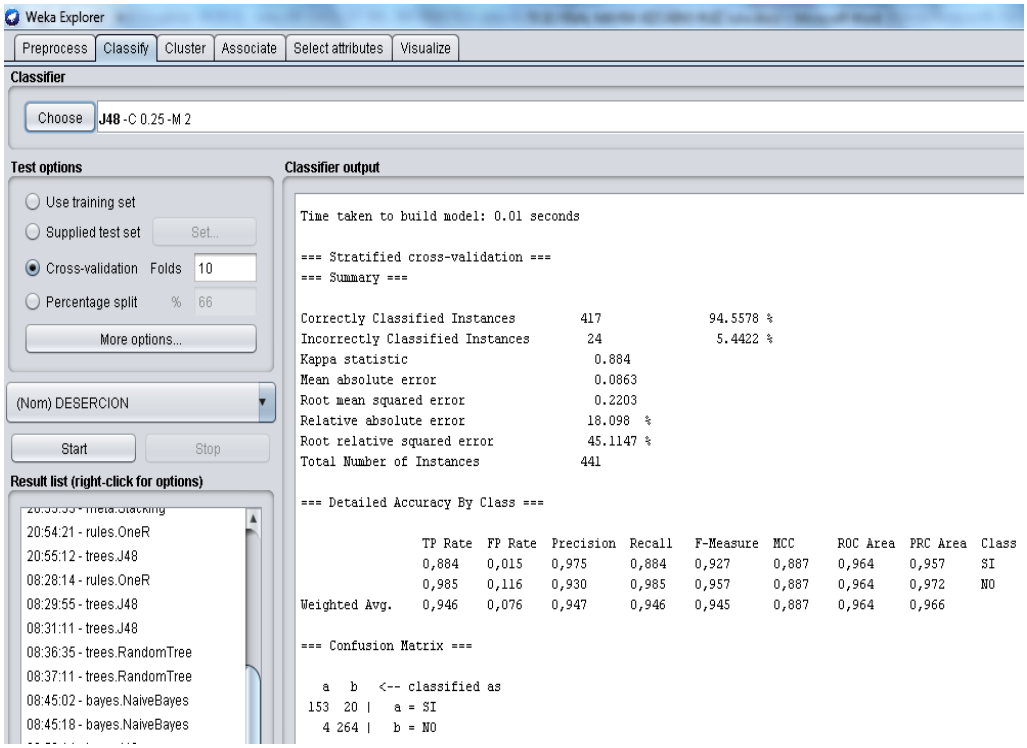
```
SELECT
GQ.CEDULA_ESTUDIANTE,
GQ.NOMBRE_ESTUDIANTE,
GQ.PRIMER_APELLIDO_ESTUDIANTE,
GQ.CAMPUS_ID,GQ.COLEGIO_ID,GQ.ESTADO_CIVIL_ID,
GQ.NOMBRE_ESTADO_CIVIL,GQ.CARRERA_ID,GQ.DISCAPACIDAD_ID,
GQ.PRI_NIVEL,
GQ.EGRESA,GQ.GRADO,GQ.CONVALIDA,
GQ.ESTADO_MATRICULA_ACTUAL_ESTUDIANTE,
GQ.TIPO_COLEGIO,GQ.NOMBRE_SEXO, GQ.PROVINCIA_ORIGEN,
GQ.ANIO_INGRESO,GQ.MES_INGRESO,GQ.ANIO_EGRESO,GQ.MES_EGRESO,GQ.ANIO_
GRADO,GQ.MES_GRADO,N.NUM,

CASE
WHEN (MES_INGRESO >= 2 AND MES_INGRESO <= 3 AND (2015 -
ANIO_INGRESO) * 2 - N.NUM + 1 >= 2) OR
     (MES_INGRESO >= 9 AND (2015 - ANIO_INGRESO) * 2 - N.NUM >=2)
THEN 'SI'
     ELSE 'N0'
END AS DESERCION
FROM dbo.ING_INFORM_QUITO GQ, dbo.NUMEROS_PERÍODOS N
WHERE GQ.ESTUDIANTE_ID =N.ESTUDIANTE_ID
AND
CONVALIDA ='no' AND
--ANIO_INGRESO > 1900 and
GRADO ='no'
AND
EGRESA      ='no'
```

ANEXO 3. ALGORITMO J48

DESERCIÓN DE ESTUDIANTES

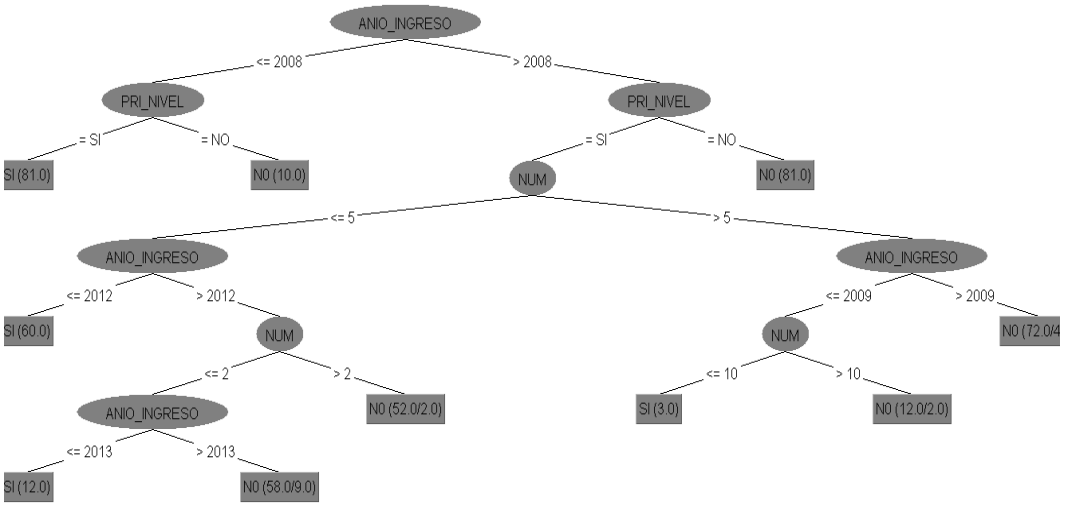
En la opción de evaluación “test option” se selecciona “cross validation” y se escoge la clase que puede ser deserción o grado. Se presiona “start” se inicia la evaluación de los algoritmos presentando los siguientes resultados.



Se presenta que las instancias clasificadas correctas 94,55%.

También se puede visualizar el árbol de forma gráfica si pulsamos el botón derecho sobre el texto trees. J48 de la caja result list y seleccionamos la opción visualize tree. A continuación se puede visualizar el árbol.

A continuación se presenta el árbol de decisión:



Árbol de decisión J48 para la deserción de estudiantes

La clase del árbol es la deserción a continuación se describe los resultados, estudiantes que ingresaron desde el primer nivel con el campo PRI_NIVEL, a la derecha del árbol existen mayor del año 2008 estudiantes que NO están en deserción son estudiantes que han convalidado materias.

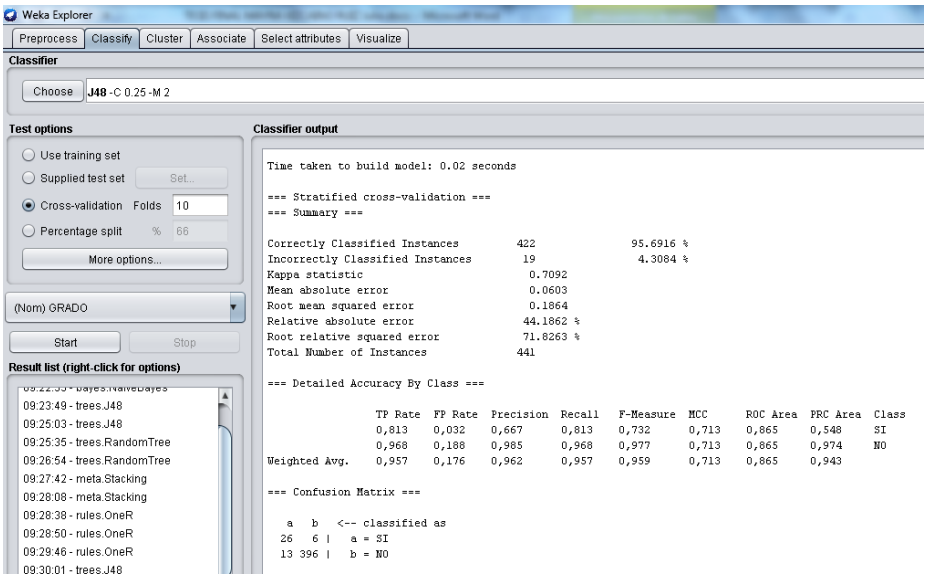
A la izquierda la opción SI que indica estudiantes que ingresaron desde primer nivel a continuación a la izquierda se clasifica por el año de ingreso menor o igual al año 2012 indicando con el campo NUM que es el número de periodos matriculados menor e igual a 5 periodos, el campo grado que muestra que se gradúan 15 estudiantes, a la derecha del atributo grado se muestra que no se gradúan 89 estudiantes y que si están en deserción. Número de periodos matriculados mayor a 5 periodos indicando la siguiente clasificación con el campo año ingreso con el año mayor al 2008 se tiene que 87 estudiantes no estan en desercion y en el año menor e igual al año 2008 con el atributo grado indica que existen 10 estudiantes que se han graduado y por lo tanto no están en deserción y 28 estudiantes que no se han graduado que si están en deserción.

A la derecha del árbol en la sección mayor al año 2012 indica que en el año 2013 existen 8 estudiantes que están en deserción se han matriculado un periodo y 40 estudiantes que no están en deserción y se han matriculado más de un periodo. Mayor que el año 2013 existe 73 estudiantes no está en deserción. Para finalizar el análisis se realiza la suma de las ramas del árbol J48 dando un total de 441 estudiantes.

ALGORITMO J48

GRADUACIÓN DE ESTUDIANTES

A continuación se presenta el análisis del dataset con 8 atributos con la clase grado y validación cruzada.



Los resultados que se obtienen en la matriz de decisión son:

Verdaderos positivos (TP): 26

Verdaderos negativos (TN) 396.

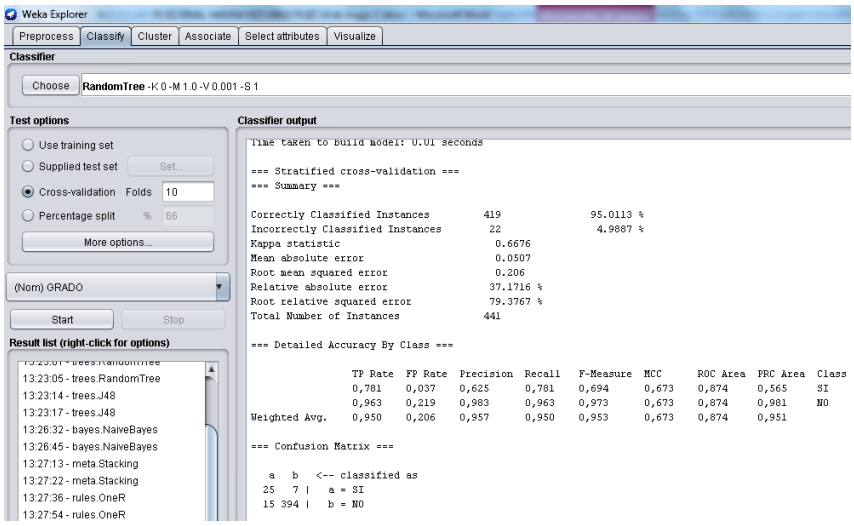
Falsos positivos (FP): 13

Falsos negativos (TN):6

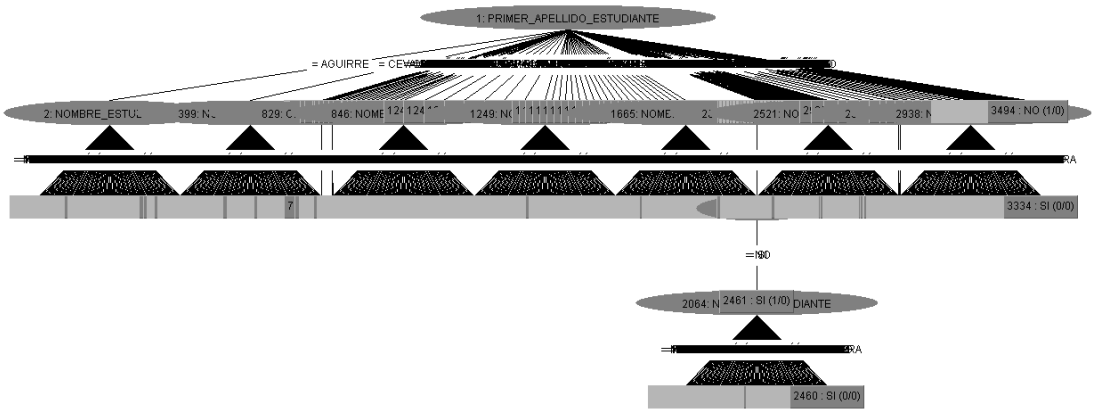
Los resultados para la graduación de estudiantes son 409 estudiantes no están graduados, 32 estudiantes solteros graduados y 7 estudiantes casados no se han graduado.

ANEXO 4 ALGORITMO RANDOMTREE

En datos numéricos el mejor algoritmo para la graduación es el algoritmo Randomtree con el método de validación cruzada con 7 atributos del dataset a continuación se presenta los resultados y en el árbol de decisión se muestra que es muy extenso para la presentación de los resultados.

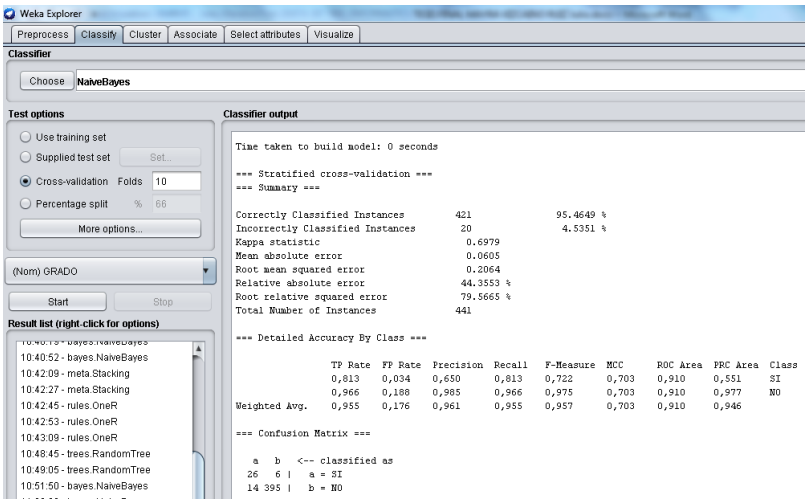


Árbol Random tree



ANEXO 5 ALGORITMO NAÏVE BAYES

Evaluación del algoritmo NaiveBayes con el método validación cruzada para la clase grado.con 7 atributos del dataset. Los resultados se presentan a continuación.



Los resultados aplicando la validación cruzada con el algoritmo Naive Bayes son: instancias clasificadas como correctas ha sido del 93,42%, mientras que el de clasificadas incorrectamente es del 6,57%. El índice Kappa toma el valor 0,54 por lo que no existe un alto grado de concordancia.

GLOSARIO DE TÉRMINOS

SGBD Un Sistema de Gestión de Bases de Datos es un conjunto de programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos, además de proporcionar herramientas para añadir, borrar, modificar y analizar los datos.

Mainframe es un ordenador de grandes dimensiones pensado principalmente para el tratamiento de grandísimos volúmenes de datos.

DBMS Database Management System (DBMS), es un conjunto de programas que se encargan de manejar la creación y todos los accesos a las bases de datos.

DDL Lenguaje de Definición de Datos.

MDL Lenguaje de Manipulación de Datos.

KDD (Knowledge Discovery from Databases) es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia, comprensibles a partir de los datos.