REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA

# UNIVERSIDAD CATÓLICA DE SANTA MARÍA

## FACULTAD DE CIENCIAS E INGENIERIAS FISICAS Y FORMALES

## SEGUNDA ESPECIALIDAD EN INGENIERIA DE SOFTWARE

Tesis

# Aplicación de técnicas supervisadas de minería de datos para determinar la predicción de deserción académica

Presentado por:

SULLA TORRES, José Alfredo

Para optar el Título de Segunda Especialidad en Ingeniería de Software

Arequipa – Perú
2015

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA



*Dedicado a mi familia:*

*Zayda, Andrea y Alejandro.*

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA



# Agradecimientos

Un agradecimiento especial a los docentes y amigos de la primera promoción de la Segunda Especialidad de Ingeniería de Software, quienes han sabido compartir sus conocimientos y experiencias para acrecentar la formación en ésta especialidad.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA



# Índice general

Pág.

Índice de figuras ..... 6

Resumen ..... 8

Abstract ..... 9

Capítulo 1- Introducción y objetivos ..... 10
1.1. Introducción ..... 10
1.2. Descripción del problema ..... 12
1.3. Definición del problema ..... 12
1.4. Objetivo General ..... 13
1.5. Objetivos Específicos ..... 13
1.6. Justificación ..... 13
1.7. Trabajos relacionados ..... 14
1.8. Metodología de desarrollo del trabajo ..... 16
1.9. Estructura del trabajo ..... 17

Capítulo 2- Marco Teórico ..... 18
2.1. Minería de datos ..... 18
2.2. Técnicas de Minería de datos ..... 20
2.2.1. Árboles de decisión. ..... 21
2.2.2. Reglas de decisión. ..... 24
2.2.3. Redes Neuronales. ..... 25
2.2.4. Redes Bayesianas ..... 25
2.2.5. Reglas de Asociación. ..... 26
2.2.6. Algoritmos Genéticos. ..... 26
2.2.7. Modelización estadística paramétrica. ..... 27
2.2.8. Modelización estadística no paramétrica. ..... 27
2.3. Análisis predictivo ..... 27
2.3.1. Técnicas de análisis predictivo ..... 29
2.3.2. Técnicas avanzadas de análisis predictivo ..... 31
2.4. Minería de datos educativo ..... 33
2.4.1. Componentes MDE ..... 33

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



2.4.2. Principales métodos de Minería de Datos utilizados en EDM ..... 38
2.5. CRISP-DM ..... 46
2.5.1. Comprensión del negocio: ..... 46
2.5.2. Comprensión de los datos: ..... 46
2.5.3. Preparación de los datos: ..... 46
2.5.4. Modelado: ..... 47
2.5.5. Evaluación: ..... 47
2.5.6. Implantación: ..... 47

Capítulo 3- Desarrollo de la Propuesta ..... 49
3.1. Comprensión del negocio ..... 49
3.2. Comprensión de los datos ..... 50
3.3. Preparación de datos ..... 54
3.4. Modelado ..... 58

Capítulo 4- Pruebas y resultados ..... 60
4.1. Pruebas ..... 60
4.1.1. De la precisión del algoritmo ..... 60
4.1.2. De la eficiencia del algoritmo ..... 68

Conclusiones ..... 72

Recomendaciones ..... 74

Bibliografía ..... 75

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

# Índice de figuras

Pág.

Figura 1.1. Ciclo de aplicación de minería de datos al sistema educativo…… 11
Figura 1.2. Número de investigaciones publicadas por años…… 14
Figura 2.1. Pasos de la Minería de datos…….. 20
Figura 2.2. Clasificación de las técnicas de Minería de Datos…….. 22
Figura 2.3. Ciclo de un árbol de decisión: Aprendizaje……..23
Figura 2.4. Ciclo de un árbol de decisión: Clasificación…….. 23
Figura 2.5. Los estados de la Minería de datos predictivo……..28
Figura 2.6. Regresión lineal. …… 29
Figura 2.7. Componentes MDE. …….. 34
Figura 2.8. Fases de la Metodología CRISP-DM. …… 48
Figura 3.1. Hoja de Cálculo Alumnos Ing. De Sistemas – 2004 – 2008 – Pestaña ALUMNOS SISTEMAS……. 52
Figura 3.2. Hoja de Cálculo Alumnos Ing. De Sistemas – 2004 – 2008 – Pestaña CALIFICACIONES. …….. 53
Figura 3.3. ETL para la conversión de datos…… 55
Figura 3.4. ETL para determinar las notas Aprobados y Desaprobados……... 56
Figura 3.5. ETL para generar el Promedio Final y la cantidad de Aprobados y Desaprobados……... 56
Figura 3.6. Diagrama de la Base Datos Académico……. 57
Figura 3.7. Ventana de inicio de Weka…… 59
Figura 3.8. Archivo .csv con los atributos cargados en Weka……59
Figura 4.1. Selección del algoritmo J48……. 60
Figura 4.2. Ejecución del clasificador J48 con los datos Académicos……... 61
Figura 4.3. Árbol creado con el algoritmo J48……..…… 63
Figura 4.4. Abandono de estudiantes en relación a la Categoría……..…… 64
Figura 4.5. Abandono de estudiantes en relación al Puntaje de ingreso……..…. 64
Figura 4.6. Abandono de estudiantes en relación a los cursos Aprobados……..…. 65

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA



Figura 4.7. Abandono de estudiantes en relación a los cursos Desaprobados…….. 65
Figura 4.8. Abandono de estudiantes en relación al Total de cursos…….… 66
Figura 4.9. Abandono de estudiantes en relación al promedio final……..… 66
Figura 4.10. Ejecución del clasificador J48 con los datos de prueba……. 67
Figura 4.11. Ejecución del clasificador J48 con los datos Académicos……. 68
Figura 4.12. Árbol creado con el algoritmo RandomTree…….. 69

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA

# Resumen

La Minería de Datos Educacional (MDE) es usado para estudiar los datos disponibles en el campo educativo y descubrir el conocimiento oculto en ella. Los métodos de clasificación como árboles de decisión, las reglas minadas, red bayesiana, etc. se pueden aplicar a los datos educativos para predecir el comportamiento de los estudiantes, el rendimiento en los exámenes, deserción estudiantil, etc. Esta predicción ayudará a las autoridades para identificar la deserción de estudiantes y poder determinar la proyección de apertura de grupos o secciones y otras acciones. El algoritmo de árbol de decisión C4.5 (J48) se aplica en los datos de evaluación original de los estudiantes para predecir si abandona o no los estudios. El resultado del árbol de decisión predijo el número de estudiantes que son propensos a abandonar la carrera. El resultado lo pueden utilizar a las autoridades para que puedan tomar las medidas para mejorar la toma de decisiones. Después de la evaluación con los datos originales se introducen un conjunto de datos de prueba en el sistema para analizar los resultados. El análisis comparativo de los resultados indica que la predicción ha ayudado determinar con mayor precisión el mejoramiento en el resultado. Para analizar la exactitud del algoritmo, se compara con el algoritmo RandomTree y se encontró que es tan eficiente en términos de precisión de los resultados académicos del estudiante y el tiempo tomado para crear el árbol.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA

# Abstract

Educational Data Mining (EDM) is used to study the data available in the educational field and discover the hidden knowledge in it. The classification methods such as decision trees, mined rules, Bayesian network, etc. can be applied to educational data to predict the behavior of students, examination performance, dropout, etc. This prediction will help the authorities to identify the dropout students and to determine the projected opening of groups or sections and other actions. The C4.5 (J48) decision tree algorithm applies original data evaluating students to predict whether or not drop out. The result of the decision tree predicted the number of students who are likely to leave the race. The result can be used by the authorities so they can take steps to improve decision making. After evaluation with the original data set of test data is entered into the system to analyze the results. The comparative analysis of the results indicates that the prediction has helped determine more precisely the improvement in the result. To analyze the accuracy of the algorithm, compared with the algorithm RandomTree and was found to be as efficient in terms of accuracy of student academic results and the time taken to create the tree.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM


# Capítulo 1
# Introducción y objetivos

## 1.1. Introducción

La minería de datos se ha implementado con éxito en el mundo de los negocios desde hace algún tiempo, pero su uso en la educación superior es todavía relativamente nuevo, su uso está destinado a la identificación y extracción de conocimiento nuevo y potencialmente valioso de los datos. EDM es el proceso de transformación de datos originales reunidos por los sistemas educativos en información útil que pudiera ser usado para la toma de decisiones y respuesta a las preguntas de investigación. En el ambiente tradicional de enseñanza/aprendizaje, el rendimiento y análisis del comportamiento se llevan a cabo sobre la base de registros de observación y de papel utilizados. Este sistema tiene los inconvenientes tales como que no puede satisfacer las necesidades de cada alumno, así como falta de aprendizaje dinámico que puede ser mejorada mediante el uso de cinco pasos de un proceso de análisis académico, tales como la captura, informe, predecir, actuar y refinar. El Análisis Predictivo es el proceso de tratar con variedad de datos y aplicar diversas fórmulas matemáticas para descubrir la mejor decisión para una situación dada. El análisis predictivo da a la institución una ventaja competitiva. Para las instituciones de educación, cuyo objetivo es contribuir a la mejora de la calidad de la educación superior, el éxito de la creación de capital humano es el sujeto de un análisis continuo. Por lo tanto, la predicción de éxito de los estudiantes es fundamental para las instituciones de educación superior, porque la calidad del proceso de enseñanza es la capacidad de satisfacer las necesidades de los estudiantes. El análisis predictivo abarca una variedad de técnicas de minería de datos, las estadísticas, que analizan los hechos actuales e históricos para hacer predicciones sobre eventos futuros. El término predictivo de minería de datos se

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM
UNIVERSIDAD CATÓLICA DE SANTA MARÍA

aplica generalmente para identificar los proyectos de minería de datos con el objetivo de identificar un modelo de red neuronal o estadística o un conjunto de modelos que se puede utilizar para predecir algunas respuestas de interés. Por ejemplo, una institución educativa que desea, puede utilizar los datos con predicción minera, para obtener un modelo que pueda identificar rápidamente el índice de deserción académica. La calidad de la educación superior implica proporcionar servicios que probablemente satisfagan las necesidades de los estudiantes y otros participantes en el sistema del proceso educativo (Romero and Ventura, 2007). Los participantes en el proceso educativo a través de sus actividades crean una gran cantidad de datos que son almacenados e integrados y utilizados. Todos los participantes en el proceso educativo podrían beneficiarse aplicando minería de datos a la enorme cantidad de datos académicos almacenados, como se aprecia en la Figura 1.1. Dado que la minería de datos representa el proceso de datos computacional desde diferentes perspectivas, con el objetivo de extraer muestras implícitas e interesantes (Witten and Frank, 2005a), las tendencias y la información de los datos, pueden ayudar mucho a todos los participantes en el proceso educativo para mejorar el entendimiento del proceso de enseñanza y descubrir, detectar y explicar el fenómeno de la educación (El-Halees, 2009).

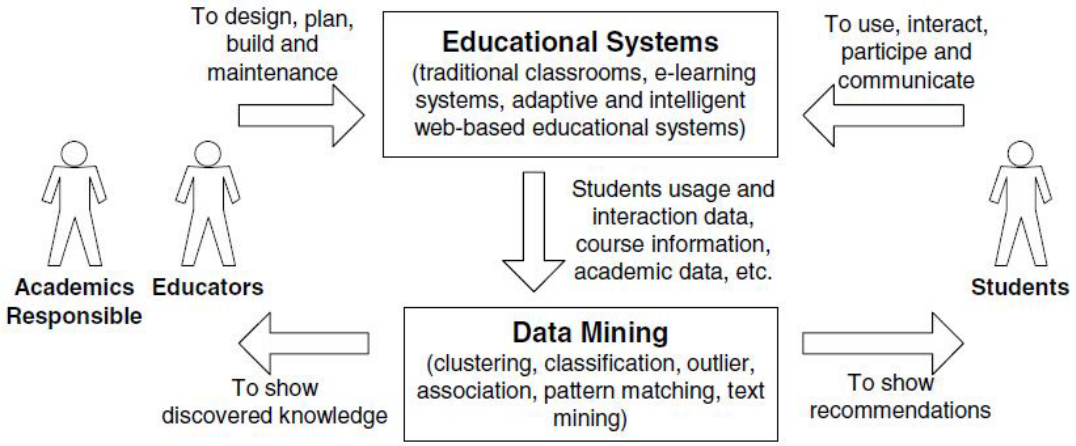



Figura 1.1: Ciclo de aplicación de Minería de Datos al sistema educativo. (Romero and Ventura, 2007)

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


## 1.2. Descripción del problema

La deserción universitaria es uno de los problemas que aborda la mayoría de las instituciones de educación superior, que afecta tanto a los estudiantes como a la institución que los acogen. Esto adquiere una importancia relevante según las cifras sobre deserción universitaria expuestas por la Organización de Cooperación Económica y el Desarrollo (OECD) que agrupa a los principales países desarrollados y que señala que las cifras de deserción en estos países vienen en franco crecimiento.

El estudio señala que el índice de deserción en países como EEUU llega al 35% en los college. Este mismo estudio señala también que en España la deserción llega a más del 20%, mientras que en Italia el abandono de las carreras llega al 60%. Las causas en los países desarrollados se deben básicamente a una inadecuada elección de la carrera. Pero, sin duda, la mayor causal es la inconformidad, la insatisfacción con la carrera y con la universidad.

En el Perú, alcanza una tasa de deserción del 17% del problema de educación, según el portal del postulante Logros (Logros, 2011), donde se señala que cada año en el Perú se pierden más de 100 millones de dólares por el abandono de las aulas y además indica que en la próxima década se podrían perder más de 2 mil 100 millones de dólares si no se toman las medidas adecuadas.

Por otro lado, las instituciones que imparten asignaturas de diferentes especialidades no pueden predecir de manera precisa la cantidad de alumnos que van a abandonar los estudios en un futuro cercano, lo cual impide determinar, entre otros aspectos, la apertura de grupos o secciones, así como, la correspondiente asignación docentes y de aulas a los grupos respectivos; esta deserción de alumnos es indeterminado por diferentes factores e incluye diversas causas..

## 1.3. Definición del problema

Predecir el comportamiento académico en deserción de estudios mediante técnicas supervisadas de minería de datos.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


## 1.4. Objetivo General

Aplicar técnicas supervisada de minería de datos que permita determinar la predicción categórica de deserción académica.

## 1.5. Objetivos Específicos

1. Analizar distintos tipos de técnicas supervisadas para la predicción de deserción académica.
2. Seleccionar las técnicas más adecuadas para un caso académico.
3. Elegir un caso de prueba de gestión académica.
4. Aplicar las técnicas supervisada en el caso de prueba.

## 1.6. Justificación

En estos tiempos donde la tecnología ocupa un rol importante dentro de las actividades humanas, los modelos de predicción de minería de datos dan solución a problemas de clasificación, agrupamientos, tendencias, etc. no solo en el ámbito empresarial sino también en el ámbito académico logrando realizar este proceso de manera razonable, no costosa, facilitando el análisis de los planes de trabajo o las políticas de operación. El trabajo propuesto es necesario, por la gran cantidad de datos que se manejan en el ámbito académico, es conocido que el aspecto educativo informático tiene uno de los más altos porcentajes de interés de las personas, por lo que es importante conocer sus tendencias en cuanto a la deserción de estudiantes para las instituciones educativas, para ello se cuenta con la existencia de diferentes tecnologías, técnicas y algoritmos que pueden ayudar a ser más eficiente las operaciones en este aspecto académico.

De otro lado, de acuerdo a nuestra Constitución y la Ley Universitaria, la educación es un derecho de todos por lo que todos los alumnos que ingresan a estudiar deberían de culminar sus estudios con su respectivo Grado Profesional.

El grupo de trabajo internacional *Educational Data Mining* (EDM) (EducationalDataMining, 2010) realiza conferencias internacionales anuales en

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


EDM. El número de publicaciones sobre EDM ha crecido exponencialmente en los últimos años como se aprecia en la Fig. 1.2.

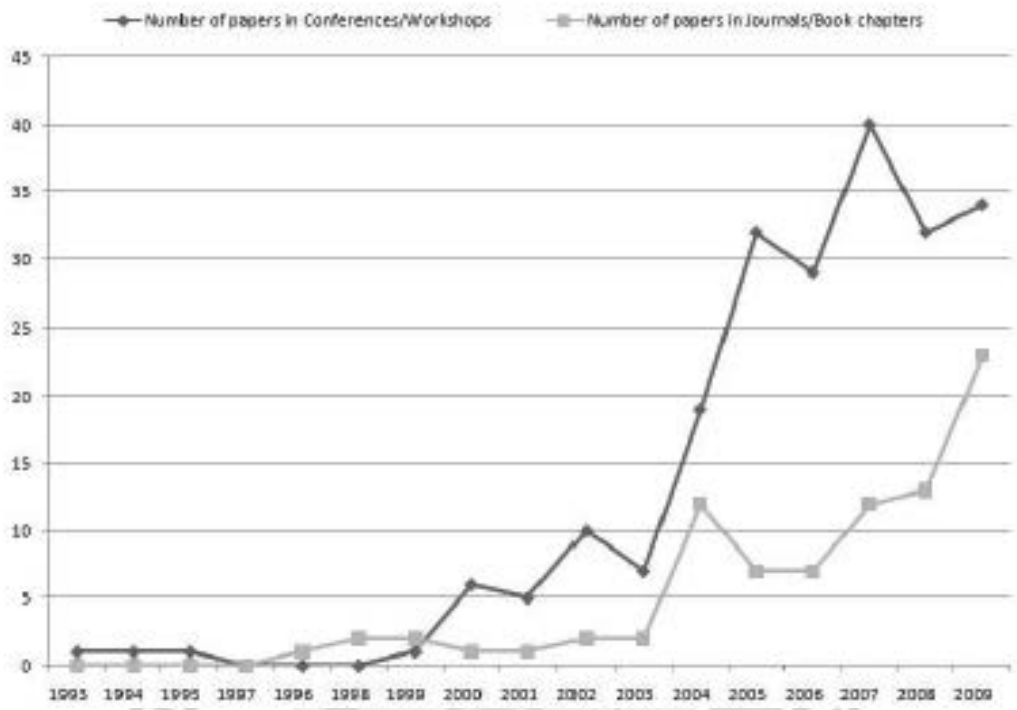



Figura 1.2. Número de investigaciones publicadas por años. (Romero and Ventura, 2009)

## 1.7. Trabajos relacionados

En esta sección se muestra de manera resumida algunos documentos y publicaciones sobre las temáticas del proceso Minería de Datos Educacional con el fin de entregar un conocimiento base para el entendimiento de este trabajo de tesis.

El trabajo de S. Anupama Kumar (Kumar and Vijayalakshmi, 2011) la predicción académica lo realiza aplicando el algoritmo de árbol de decisión C4.5 a las calificaciones iniciales de los alumnos para predecir el número de alumnos que aprobaran los cursos, y luego lo aplica a las calificaciones finales. El análisis comparativo de los resultados lo realiza de acuerdo a la precisión del algoritmo comparándolo con el algoritmo ID3 encontrando cuál es el más eficiente en términos de precisión predictiva y tiempo que toma construir el árbol.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


Edin Osmanbegović y Mirza Suljić (Osmanbegović and Suljic, 2012) compara las diferentes técnicas y métodos de predicción en la aprobación de los estudiantes, la aplicación de los datos obtenidos de las encuestas realizadas durante el semestre de verano de la Universidad de Tuzla, en la Facultad de Economía, durante el año académico 2010-2011, los datos fueron tomados durante la inscripción entre los estudiantes de primer año. La prueba se evaluó con la calificación aprobatoria en el examen. El impacto de las variables socio-demográficas de los estudiantes y las actitudes hacia el estudio puede tener un efecto en aprobación, todos fueron investigados.

En el trabajo de Minaei-Bidgoli y B. Kashy (Minaei-Bidgoli et al., 2003) presenta un enfoque para clasificar a los alumnos con el fin de predecir su calificación final en base a las características extraídas de los datos registrados en un sistema educativo basado en la Web. Se diseñó, implementó y evaluó una serie de clasificadores de patrones y comparó su rendimiento en un conjunto de datos de un curso en línea. Usa una combinación de múltiples clasificadores que conduce a una mejora significativa en el rendimiento de clasificación. Además, al aprender una ponderación adecuada de las funciones utilizadas a través de un algoritmo genético (GA), mejorando aún más la precisión de la predicción. El GA demuestra que mejora con éxito la exactitud del rendimiento clasificador combinado, alrededor del 10 al 12% cuando se compara con un clasificador que no utiliza GA. Este método puede ser de gran utilidad en la identificación de los estudiantes en situación de riesgo inicial, especialmente en las clases muy grandes.

Mohammed M. Abu Tair y Alaa M. El Halees (Tair and El-Halees, 2012) en su trabajo utiliza la minería de datos para mejorar el rendimiento educativo de los estudiantes graduados y superar el problema de las bajas calificaciones de los estudiantes de posgrado. En este caso de estudio, extrae conocimiento útil a partir de los datos recogidos de estudiantes graduados de la universidad de Ciencia y Tecnología - Khanyounis. Los datos incluyen período de quince años [1993-2007]. Después del pre-procesamiento de los datos se aplicó las técnicas de minería de datos para descubrir la asociación, clasificación, clustering y reglas de detección de valores atípicos; en este caso la inclusión de otras técnicas de minería de datos hace

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


que tenga un espectro más amplio de los resultados y pueda tener otros puntos adicionales sobre sus resultados.

En el trabajo de Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás (Romero et al, 2008), en su trabajo compara los diferentes métodos y técnicas de minería de datos para clasificar a los alumnos en función de sus datos de uso en el Moodle y la calificación final obtenida en sus respectivos cursos. Desarrolló una herramienta de minería específica para hacer la configuración y ejecución de técnicas de minería de datos más fácil para los instructores. Utilizó los datos reales de siete cursos del Moodle con estudiantes de la Universidad de Córdoba. También se ha aplicado discretización y técnicas de pre-procesamiento para rebalancear los datos numéricos originales con el fin de verificar si se obtienen mejores modelos de clasificadores. Un modelo clasificador apropiado para uso educativo tiene que ser preciso y comprensible para los instructores con el fin de ser de utilidad para la toma de decisiones.

José Hernández y Cesar Ferri (FERRI and HERNANDEZ, 2006) en el trabajo "Practica de Minería de Datos" presenta el uso de la herramienta de Minería de Datos Weka (*The Waikato Environment for Knowledge Analys*), utilizando diferentes algoritmos en base a ejemplos sencillos y prácticos para dar al usuario, una introducción básica de la herramienta. Weka es un conjunto de librerías Java para la extracción de conocimiento desde las bases de datos, esta incluye: Diversas fuentes de datos (ASCII, JDBC), interfaz visual basado en procesos/flujos, distintos algoritmos de Minería de Datos, manipulación de datos, combinación de modelos y entorno de experimento del cual sólo presentaremos el módulo explorer que es el que se va a utilizar en el presente trabajo.

## 1.8. Metodología de desarrollo del trabajo

La metodología a utilizar va a ser CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman et al, 2000).

Otras metodologías como SEMMA, está ligada a los productos de SAS Institute donde se encuentra implementada. La metodología CRISP-DM es una metodología

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


libre y gratuita que no depende de la herramienta que se utilice para el desarrollo del proyecto de data mining. La metodología CRISP-DM se organiza en seis etapas:

- Comprensión del negocio.
- Comprensión de datos.
- Preparación de datos
- Modelado.
- Evaluación.
- Desarrollo

## 1.9. Estructura del trabajo

El presente trabajo está organizado de la siguiente manera:

En el Capítulo I se realiza la introducción y objetivos del trabajo, indicando la descripción del problema, los objetivos planteados, justificación, metodología y trabajos relacionados. En el Capítulo II se presenta el Marco Teórico, resaltando los temas más importantes sobre los que descansa la base conceptual del trabajo con sus respectivas referencias bibliográficas. En el Capítulo III se desarrolla la propuesta de trabajo considerando las diferentes etapas de que está conformada la metodología utilizada. En el Capítulo IV se presenta las pruebas realizadas y los resultados obtenidos según el caso de estudio y evaluaciones respectivas. Finalmente de enumera las conclusiones y recomendaciones propuestos.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


# Capítulo 2
# Marco Teórico

En este capítulo se presentan los fundamentos teóricos para el desarrollo del trabajo. Se describen los temas de Minería de Datos, las técnicas de minería de datos, el análisis predictivo, la minería de datos educativos y la metodología de minería de datos Crisp.

## 2.1. Minería de datos

Minería de Datos o *Data Mining* es el proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, data warehouses u otros repositorios de información (Han and Kamber, 2006).

Las herramientas de Minería de datos inspeccionan las bases de datos e identifican modelos previamente escondidos en un sólo paso, posteriormente realizan un proceso de la limpieza, o preparación de los datos, aquí se lleva a cabo la reducción y la transformación de las bases de datos, ya limpia se lleva a cabo la clasificación de las bases de datos, una valoración de la información y finalmente se conforma la base del conocimiento y se pueden tomar decisiones en base a la información clasificada. (Witten and Frank, 2005b). Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos. Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos.

Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- **Pronóstico**: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


- **Riesgo y probabilidad**: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.

- **Recomendaciones**: determinación de los productos que se pueden vender juntos y generación de recomendaciones.

- **Búsqueda de secuencias**: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.

- **Agrupación**: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

La Gestión de conocimiento (*Knowledge Discovery*) abarca todas aquellas tecnologías relativamente nuevas que surgen de la necesidad de procesar, analizar y aprovechar la información "escondida" en grandes volúmenes de datos. La Gestión de Conocimiento requiere una captación, estructuración y transmisión de conocimiento. Permite a los que la usen obtener información útil de la forma más eficiente posible a partir de los datos.

La minería de datos es un área dentro de la Gestión de Conocimiento y se puede definir como un conjunto de metodologías y herramientas que permiten extraer el conocimiento útil para ayuda a la toma de decisiones, comprensión y mejora de procesos o sistemas partiendo de grandes volúmenes de datos. En la Figura 2.1. se puede apreciar los pasos de la Minería de datos comprendida dentro de los estados del KDD.

Por tanto, y según (Hernández Orallo et al., 2004), dos son los retos de la Minería de datos: trabajar con grandes volúmenes de datos y usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado tiene una relación directa con la comprensibilidad del modelo inferido. Dado que el usuario en muchas ocasiones no es un experto en técnicas de Minería de Datos, este modelo generado deber ser comprensible. Para poder llevar a cabo las técnicas de Minería de Datos de una

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


manera eficiente, se necesita un sistema de adquisición, almacenamiento y manejo de la información eficiente. El objetivo final de la Minería de Datos es transformar datos en conocimiento.

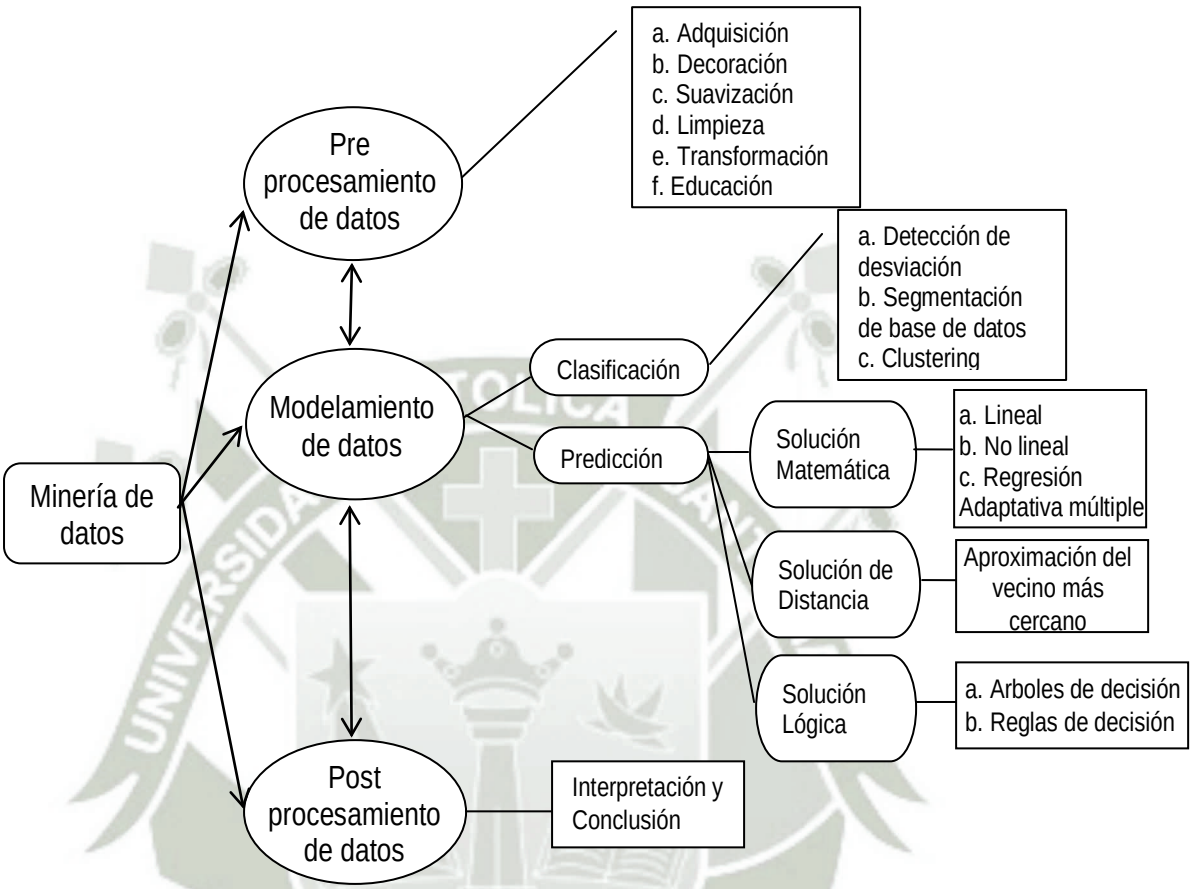



Figura 2.1. Pasos de la Minería de datos.

## 2.2. Técnicas de Minería de datos

La minería de datos tiene su enfoque de análisis de datos dirigido al descubrimiento de conocimiento donde se descubre información sin necesidad de formular previamente una hipótesis. La aplicación automatizada de algoritmos de minería de datos permite detectar patrones en los datos por lo que es más eficiente su análisis. Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Weis, 1998)

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


Los algoritmos supervisados o predictivos predicen el valor del atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento y prueba. Las técnicas más utilizadas son: árboles de decisión, inducción neuronal, regresión y series temporales.

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o de descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales.

La clasificación de las técnicas de Minería de Datos se distinguen en:

- **Predictivas**: Aquellas que las variables pueden clasificarse inicialmente en dependientes e independientes. Especifican el modelo para los datos en base -a un conocimiento teorico previo, este modelo es validado y aceptado una vez terminado el proceso de Mineria de Datos.
- **Descriptivas**: Aquellas que todas las variables tienen inicialmente el mismo estatus, es decir no se les asigna ningún papel predeterminado, además están exentos de variables dependientes e independientes de un modelo previo de datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones.
- **Auxiliares**: Básicamente son herramientas de apoyo mas superficiales y limitadas. Son nuevos métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la verificación.

En la Figura 2.2 se muestra de manera clasificada las técnicas de Minería de Datos.

### 2.2.1. Árboles de decisión.

Es un conjunto de condiciones o reglas organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


condiciones que se cumplen desde la raíz hasta alguna de sus hojas. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

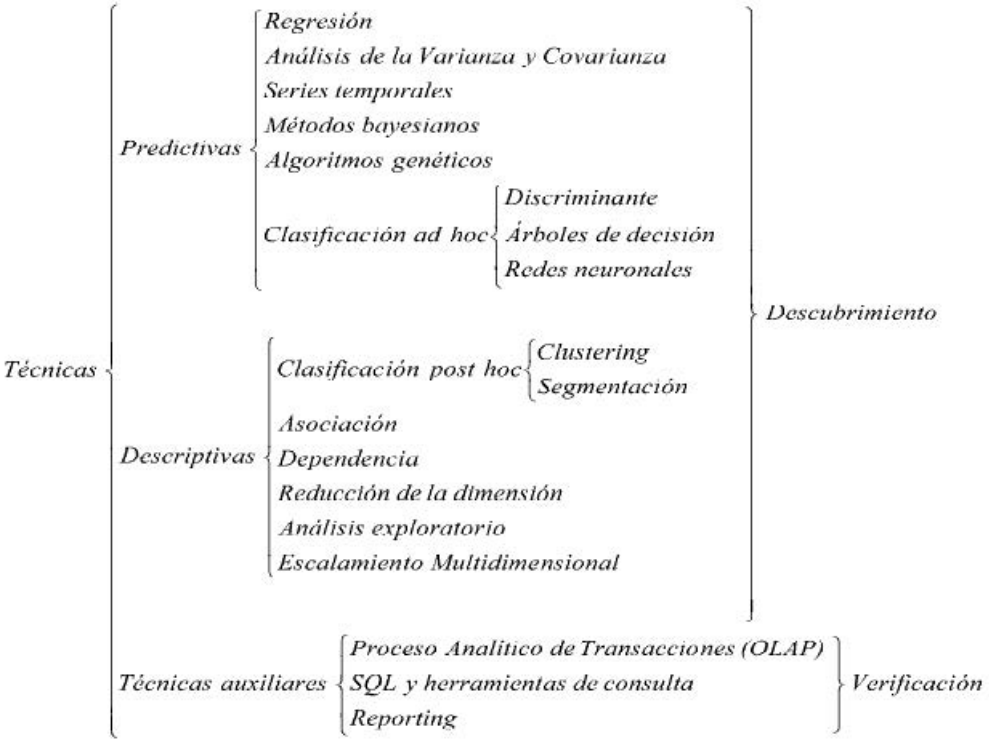



Figura 2.2: Clasificación de las técnicas de Minería de Datos. (López, 2007)

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta la cual en últimas es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión.

Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión. El árbol de decisión suele contener

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un test sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema, este tipo de nodos es redondo, los demás son cuadrados. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada.

a) **Ciclo de un árbol de decisión:** De acuerdo al ciclo que debe ser aplicado a un árbol de decisión, se tiene:

1. Aprendizaje: Crea un modelo predictivo

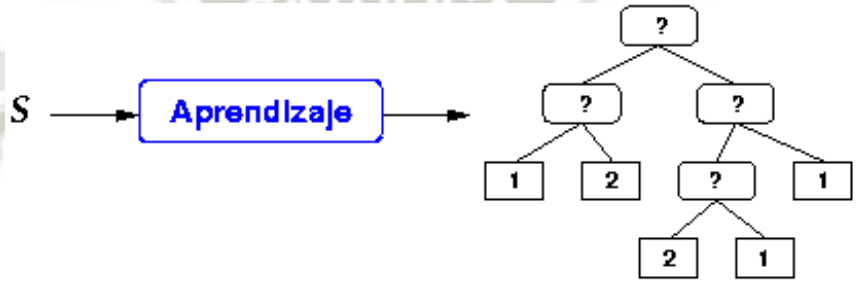



Figura 2.3. Ciclo de un árbol de decisión: Aprendizaje

2. Clasificación: Las hojas representan la clasificación y las ramas representan conjunciones de características.

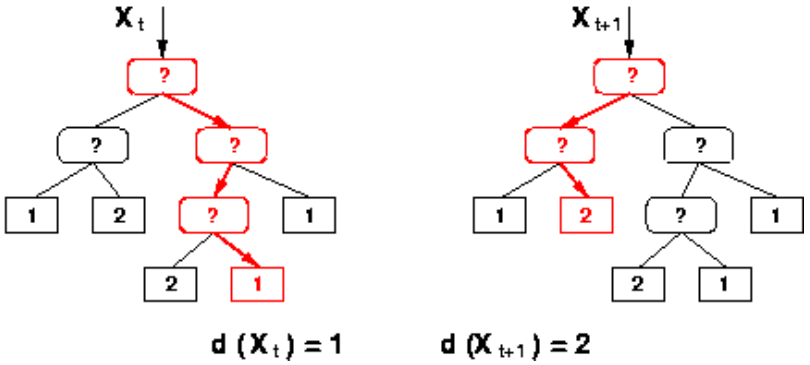



Figura 2.4. Ciclo de un árbol de decisión: Clasificación.

**b) Construcción de un árbol de decisión:** Para la construcción de árboles de decisión se deben tener en cuenta ciertas etapas, estas son:

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


1. Construir el árbol (Reglas de división)
   - Al inicio todos los ejemplos de entrenamiento están a la raíz.
   - Los atributos deben ser categóricos (si son continuos ellos deben ser discretizados)
   - El árbol es construido recursivamente de arriba hacia abajo con una visión de divide y conquista.
   - Los ejemplos son particionados en forma recursiva basado en los atributos seleccionados
   - Los atributos son seleccionados basado en una medida heurística o estadística (ganancia de información)
   - La ganancia de información se calcula desde el nivel de entropía de los datos.
2. Detener la construcción (Reglas de parada), se tienen en cuenta las siguientes condiciones:
   - Todas las muestras para un nodo dado pertenecen a la misma clase.
   - No existe ningunos atributos restantes para ser particionados (el voto de la mayoría es empleada para clasificar la hoja).
   - No existe más ejemplo para la hoja.
3. Podar el árbol (Reglas de poda)
   - Identificar y eliminar ramas que reflejen ruido o valores atípicos.

### 2.2.2. Reglas de decisión.

Son una generalización de los árboles de decisión. La principal diferencia con los anteriores es la filosofía del algoritmo que utilizan: partición (arboles de decisión) o cobertura (sistemas de reglas). Hay conjuntos de reglas que no se derivan de particiones y son capaces de clasificar la evidencia de una manera conveniente.

Existen métodos que generan conjuntos de reglas que podrían ser en algunos casos contradictorias para algunos ejemplos. Por ello, se da una orden de reglas, y estas se aplican en ese orden. Estos métodos van generando reglas, unas detrás

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

de otra, mientras vayan cubriendo ejemplos de manera consistente y son los llamados métodos por cobertura.

En los algoritmos de aprendizaje de reglas el criterio de selección de condiciones es determinante y debe evaluar muy bien y a priori que particiones son mejores. Este criterio debe incluir la creación de reglas por defecto, los criterios de parada o de evaluación, los tipos de condiciones, etc.

### 2.2.3. Redes Neuronales.

Son modelos no lineales, inspirados en el funcionamiento del cerebro, que fueron diseñados para resolver una gran variedad de problemas. Se emplea normalmente un conjunto de ejemplos representativos de la transformación deseada para "entrenar" el sistema, que, a su vez, se adapta para producir las salidas deseadas cuando se lo evalúa con las entradas "aprendidas".

Los perceptrones multi-capa son algoritmos de regresión que construyen un modelo determinista $y=f(x)$, relacionando un conjunto de predictores, $x$, y predictandos. Las redes auto-organizativas (SOM) son redes competitivas diseñadas para problemas de agrupación (clustering) y visualización.

### 2.2.4. Redes Bayesianas.

Es un grafo dirigido acíclico cuyos nodos representan variables y los arcos que los unen codifican dependencias condicionales entre las variables. El grafo proporciona una forma intuitiva de describir las dependencias del modelo y define una factorización sencilla de la distribución de probabilidad conjunta consiguiendo un modelo manejable que es compatible con las dependencias codificadas. Existen algoritmos eficientes para aprender modelos gráficos probabilísticos a partir de datos, permitiendo así la aplicación automática de esta metodología en problemas complejos. Las Redes Bayesianas que modelizan secuencias de variables (por ejemplo, series temporales de observaciones) se denominan Redes Bayesianas Dinámicas. Una generalización de las Redes

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

Bayesianas que permiten representar y resolver problemas de decisión con incertidumbre son los Diagramas de Influencia.

### 2.2.5. Reglas de Asociación.

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

Para encontrar estas reglas de debe considerar cada posible combinación de condiciones para que haya una consecuencia. Al hablar de reglas de asociación también cumple un rol importante la cobertura o también llamado soporte que es el número de instancias predichas correctamente y la precisión o confianza que es la proporción de numero de instancias que es aplicada la regla.

$$\{Confianza\} \rightarrow \{Soporte\}$$
$$\{X\} \rightarrow \{Y\}$$

### 2.2.6. Algoritmos Genéticos.

Los Algoritmos Genéticos (AGs) son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza de acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin. Por imitación de este proceso, los Algoritmos Genéticos son capaces de ir creando soluciones para problemas del mundo real. La evolución de dichas soluciones hacia valores óptimos del problema depende en buena medida de una adecuada codificación de las mismas.

Un algoritmo genético consiste en una función matemática o una rutina de software que toma como entradas a los ejemplares y retorna como salidas cuáles de ellos deben generar descendencia para la nueva generación. Versiones más

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


complejas de algoritmos genéticos generan un ciclo iterativo que directamente toma a la especie (el total de los ejemplares) y crea una nueva generación que reemplaza a la antigua una cantidad de veces determinada por su propio diseño. Una de sus características principales es la de ir perfeccionando su propia heurística en el proceso de ejecución, por lo que no requiere largos períodos de entrenamiento especializado por parte del ser humano, principal defecto de otros métodos para solucionar problemas, como los Sistemas Expertos.

### 2.2.7. Modelización estadística paramétrica.

Trata tanto problemas con variables numéricas (regresión) como en problemas de variables categóricas, dando lugar a la clasificación supervisada. La regresión y la clasificación son dos de las tareas más habituales en Minería de Datos. Son los modelos que dependen de un número finito de parámetros. Este es el nombre para definir en estadísticas las técnicas de clasificación supervisada.

### 2.2.8. Modelización estadística no paramétrica.

Es la que hace uso de métodos núcleo (kernel) para poder construir modelos más flexibles que en la modelización paramétrica, al hacer ajustes de manera local (ajustes a los diferentes puntos y no de manera conjunta al problema). Los modelos paramétricos presentan el problema de tener una estructura muy rígida para ajustarse los problemas, por ello surge la estadística no paramétricas.

## 2.3. Análisis predictivo

El análisis predictivo es la rama de la minería de datos que trata con la predicción de las probabilidades y las tendencias futuras. El elemento central de análisis predictivo es el predictor, una variable que puede ser medida por un individuo u otra entidad para predecir el comportamiento futuro (Mishra et al., 2010).

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


Análisis predictivo (Lutu and Engelbrecht, 2010) es el proceso de tratar con variedad de datos y aplicar diversas fórmulas matemáticas para descubrir la mejor decisión para una situación dada. El análisis predictivo da a la organización una ventaja competitiva. Es la ciencia la decisión que elimina las conjeturas fuera del proceso de toma de decisiones y aplica las directrices científicas probadas para encontrar la solución correcta en el menor tiempo posible. El análisis predictivo pasa por una serie de estados como se aprecia en la Figura 2.5.

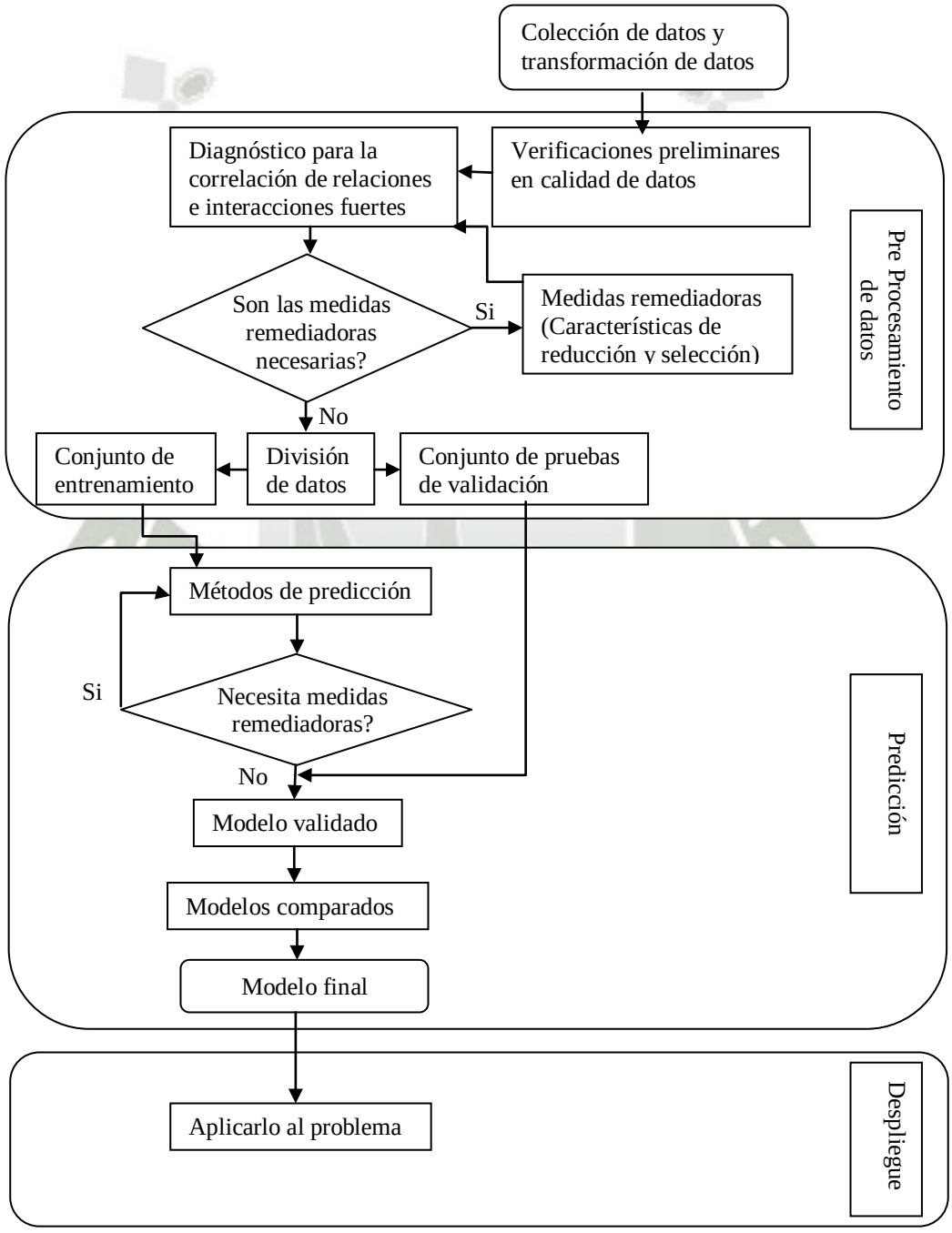



Figura 2.5. Los estados de la Minería de datos predictivo

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

### 2.3.1. Técnicas de análisis predictivo

Como en todo procedimiento automático, las predicciones estadísticas de Minería de Datos deben ser inspeccionadas por técnicas de soporte, de manera que se comprendan y verifiquen lo que se ha producido.

Hay que encontrar un término medio entre la capacidad explicativa del modelo (claridad) y su poder de predicción. En general, conforme aumenta el poder de predicción del modelo baja su capacidad de interpretar el fenómeno objeto de estudio. Mientras más sencilla sea la forma del modelo, más fácil será su comprensión, pero tendrá menor capacidad para tener en cuenta dependencias sutiles o demasiado variadas (no lineales).

- **Regresión lineal:** Llamado como el análisis de regresión línea recta es una regresión para estimar el efecto desconocido de cambio de una variable sobre otra. Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables. El coeficiente de correlación lineal nos permite determinar si, efectivamente, existe relación entre las dos variables. Una vez que se concluye que sí existe relación, la regresión nos permite definir la recta que mejor se ajusta a esta nube de puntos.

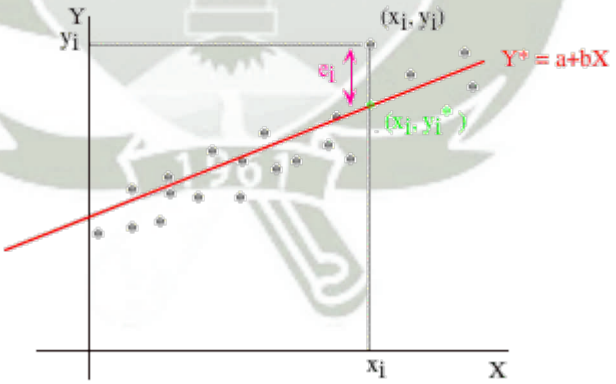



Figura 2.6. Regresión lineal.

Una recta viene definida por la siguiente fórmula:

$$y = a + bx$$

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

Donde *y* sería la variable dependiente, es decir, aquella que viene definida a partir de la otra variable *x* (variable independiente). Para definir la recta hay que determinar los valores de los parámetros *a* y *b*:

El parámetro *a* es el valor que toma la variable dependiente *y*, cuando la variable independiente *x* vale 0, y es el punto donde la recta cruza el eje vertical.

El parámetro b determina la pendiente de la recta, su grado de inclinación.

La regresión lineal nos permite calcular el valor de estos dos parámetros, definiendo la recta que mejor se ajusta a esta nube de puntos, como se muestra en la Figura 2.6.

- **Regresión lineal múltiple (MLR):** Es una técnica matemática que utiliza un número de variables para predecir alguna variable desconocida. Se trata de un estudio sobre la relación entre una única variable dependiente y una o más variables independientes.

- **Regresión logística:** Es un tipo de modelo predictivo que se puede utilizar cuando la variable objetivo es una variable categórica que tiene exactamente dos categorías como, ganar el juego/no ganar. Técnicamente se puede decir que la regresión logística se utiliza para la regresión binomial. Al mismo tiempo también se aplica a la variable objetivo continuo que modela la probabilidad de un suceso que ocurre como una función lineal de un conjunto de variables predictores.

- **Regresión Poisson:** Es otra forma de regresión donde la variable dependiente es un *contador* donde se limita la regresión logística en los términos de la distribución binomial. Esto significa que el número de ensayos se hace muy grande, mientras que la expectativa se mantiene estable es decir, la probabilidad de éxito es relativamente pequeña.

- **Árbol de Regresión:** A veces los árboles de regresión se consideran como una variante de árboles de decisión. Toma como entrada una mezcla de

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


variables continuas y categóricas y emite una variable numérica. El árbol está construido básicamente por un proceso de partición recursiva es decir, que divide los datos de forma iterativa en particiones y otra vez los divide más a fondo en cada uno de la rama. Inicialmente, el conjunto de entrenamiento están juntos y luego el algoritmo elige un punto de división que divide los datos. El proceso de división o partición continúa hasta que cada nodo se convierte en nodo terminal. Los árboles de regresión hacen la predicción rápidamente ya que no trata con cálculos complicados e incluso fáciles de entender. Con sólo mirar el árbol se puede inferir que de la variable están haciendo la predicción.

- **Análisis de factor:** Es un método que es usado para determinar cuales variables se combinan para generar un factor dado. Esto es básicamente una técnica estadística multivariante en el que se examina el conjunto de relación independiente.

### 2.3.2. Técnicas avanzadas de análisis predictivo

- **Predicción de series temporales** (Enders, 2004) predice el valor futuro de una medida basada en los valores del pasado. El pronóstico de las series temporales utiliza un modelo para predecir hechos futuros basados en hechos pasados conocidos. Ejemplos incluyen precios de acciones y los ingresos por ventas.

- **Transformación y perfiles y de datos** (Greene, 2003) utiliza funciones que analizan las filas y las columnas, los atributos y dependencias, formatos de datos, cambio de campos, agregar registros, y unir filas y columnas.

- **Análisis bayesiano** captura los conceptos utilizados en la predicción de la probabilidad. Se trata de un procedimiento estadístico que estima parámetros de una distribución subyacente sobre la base de la distribución observada.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM

UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA

- **Análisis de regresión,** es una herramienta estadística para la investigación de las relaciones entre las variables. Por lo general, el investigador pretende saber el efecto causal de una variable sobre otra, por ejemplo, el efecto de un aumento de precios a la demanda, o el efecto de los cambios en la oferta de dinero en la tasa de inflación.

- **Clasificación,** utiliza los atributos de datos para asignar un objeto a una clase predefinida o predecir el valor de una variable numérica de interés. Los ejemplos incluyen el análisis del riesgo de crédito, la probabilidad de compra.

- **Agrupación o segmentación,** separa los datos en subgrupos homogéneos basados en atributos. Clustering asigna un conjunto de observaciones en subconjuntos (clusters) de modo que las observaciones en el mismo grupo son similares. Un ejemplo es cliente segmentación demográfica.

- **Análisis de dependencia o asociación** describe asociaciones significativas entre elementos de datos. Un ejemplo es el análisis de mercado. Análisis de mercado es una técnica de modelado en base a la teoría de que si usted compra un determinado grupo de artículos, que son más (o menos) probables de comprar otro grupo de artículos.

- **Modelos de simulación,** es una estructura para estimar el impacto de las decisiones de gestión o cambios. El comportamiento del modelo de simulación cambiará en cada simulación según el conjunto de parámetros iniciales asumidos por el medio ambiente. Los ejemplos incluyen las políticas de reordenamiento de inventarios, entrenamiento militar.

- **Modelos de optimización** (Hernández Orallo et al., 2004), es una estructura en términos de restricciones para encontrar la mejor solución posible. Los modelos de optimización forman parte de un sistema mayor que usa la gente para ayudarles a tomar decisiones. El usuario es capaz de

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


influir en las soluciones que el modelo produce y los comentarios antes de tomar una decisión final en cuanto a lo que debe hacer. Ejemplos incluyen la programación de los trabajadores por turnos, enrutamiento de trenes de carga, y los precios por los sitios en los aviones.

- **Logro de servicios de análisis predictivo** (Kannan et al, 2010) el aumento de las tasas de respuesta de las ofertas, mailings, promociones y, así como, la venta cruzada y hasta el modo de venta para una mayor participación en el gasto que permitan:
  - Reconocer los factores desencadenantes y los conductores que maximizan la satisfacción del cliente y actualizan el valor potencial no realizado.
  - Fortalecimiento de la lealtad del cliente.
  - Retener a los clientes más rentables.
  - Optimizar las características del producto y los precios.
  - Optimizar los procesos de servicio al cliente y punto de contacto.

## 2.4. Minería de datos educativo

La Minería de Datos Educativos (MDE) es un campo emergente exploración de datos en el contexto educativo, aplicando diferentes técnicas de minería de datos. MDE hereda propiedades de áreas como análisis de aprendizaje, Psicometría, Inteligencia Artificial, Tecnología de Información, Máquinas aprendizaje, Sistema de Gestión de Base de Datos, Informática y Minería de Datos. Se puede considerar como campo de investigación interdisciplinar que proporciona el conocimiento intrínseco del proceso de enseñanza y aprendizaje para la educación efectiva.

### 2.4.1. Componentes MDE

Los componentes clave de MDE son los *Stakeholder* de Educación, las técnicas, herramientas/métodos de Minería de datos, los datos educativos, las

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


tareas y salidas las cuales reúnen los objetivos educativos, como se aprecia en la Figura 2.7.

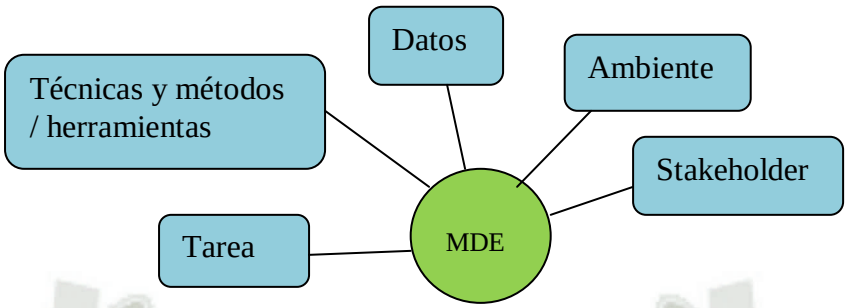



Figura 2.7. Componentes MDE. (Jindal and Borah, 2013) Adaptado

a. **Stakeholders:** Considerando la educación universitaria, se puede dividir en tres grupos:
   - **Grupo primario:** Es el grupo que está directamente involucrado con el proceso de enseñanza/aprendizaje: estudiantes y docentes.
   - **Grupo secundario:** Es el grupo que está relacionado indirectamente con el crecimiento académico: padres y egresados.
   - **Grupo hibrido:** Es el grupo que esté conformado por los administradores, empleados, tomadores de decisión y expertos.

b. **Ambiente:** Según como se da la interacción con los *stakeholders*:
   - **Ambiente formal:** Es el ambiente de interacción directa con el grupo primario de los *stakeholders*: salón de clases.
   - **Ambiente informal:** Es el ambiente de interacción indirecta con el grupo primario de los *stakeholders*: aula virtual, etc.
   - **Ambiente con soporte computacional:** Interacción directa o indirecta con todos los grupos: Sistemas Inteligentes de tutoría, Herramientas de apoyo de aprendizaje.

c. **Datos:** La toma de decisiones en el campo académico involucra el análisis de grandes volúmenes de datos educativos. Los datos son generados de

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


diferentes fuentes heterogéneas pudiendo ser diversas y distribuidas, estructuradas y no estructuradas, etc.

La mayoría de datos pueden ser generados de fuentes *offline* u *online:*

- **Datos *Offline*:** Los datos se generan a partir de la interacción tradicional en los ambientes interactivos de enseñanza/aprendizaje, información de educador/estudiante, información de cursos, etc.
- **Datos *Online*:** Los datos se generan a partir de participación distante de los *stakeholders* como los sistemas educativos basados en la web, aprendizajes colaborativos, redes sociales, etc.

d. **Tarea educativa:** Este es un proceso continuo para la formación de la Visión y Misión de una institución. Para nutrir el talento de los estudiantes que lo direccionen de una manera responsable, ética e innovadora para alcanzar los objetivos académicos y administrativos. Esta tarea se puede dividir en dos tipos:
   - **Tarea para la toma de decisiones:** Participación activa del grupo hibrido de los *stakeholders* para cumplir los objetivos administrativos.
   - **Tarea basada en el aprendizaje:** Participación activa del grupo primario de los *stakeholders* para cumplir los objetivos académicos.

e. **Métodos MD:** Los métodos MD son uno de los principales componentes en MDE, por su amplio propósito se pueden dividir en dos grupos
   - **Orientado a la verificación:** Estadísticas tradicionales: Análisis de varianza, pruebas de hipótesis.
   - **Orientado al descubrimiento:** Descripción y Predicción: Clasificación, Clustering, Predicción, Redes Neuronales, Minería web, etc.

f. **Herramientas:** Debido al rápido crecimiento de los datos educativos, hay una necesidad de resumir las herramientas de acuerdo a sus funciones y/o características, técnicas integradas y plataformas de trabajo. EPRules, GISMO, TADAED, O3R, Synergo/COLAT, LISTEN Mining Tool, MINEL, LOCO, CIECoF, PDinamet, Meerkat, MMT son ejemplos de herramientas

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


de EDM (Romero and Ventura, 2013). Aparte de estos, existen algunas herramientas más que sean útiles para la comunidad de investigadores a los datos educativos según los requisitos (ver Tabla 2.1)

| **Herramienta y desarrollador** | **Código Abierto / Libre / Comercial** | **Funciones / Características** | **Técnicas / Herramientas** | **Ambiente** |
|---|---|---|---|---|
| Intelligent Miner ( IBM ) | Comercial | Proporciona una estrecha integración con el sistema de bd relacional DB2, escalabilidad del algoritmo de Minería | Asociación Minera, Clasificación, Regresión, Predictivo Modelado, Detección de Desviación, Clustering. | Windows, Solaris, Linux |
| MSSQL Server (Microsoft) | Comercial | Proporciona funciones de DM, tanto en el sistema de bd relacional y Data Warehouse (DWH) | Integra los algoritm desarrollados por terceros proveedores y usuarios de la aplicación. | Windows, Linux |
| MineSet (SGI ) | Comercial | Proporciona Herramientas gráficas robustas como árbol de visualización, Mapa visualizar | Asociación Minera, Clasificación, Estática avanzadas y visualización | Windows, Linux |
| Oracle Data Mining (Oracle Corporation) | Comercial | Proporciona una infraestructura DWH integrado para el análisis de datos multidimensional | Asociación Minera, Clasificación, Predicción, regresión, Clustering, Secuencia de similitudes. | Windows, Mac, Linux |
| SPSS Clementine (IBM) | Comercial | Proporciona un entorno de desarrollo integrado de minería de datos para los usuarios finales y desarrolladores. | Asociación Minera, Clustering, Clasificación, Predicción y visualización | Windows, Solaris, Linux |
| Enterprise Miner (SAS Institute ) | Comercial | Ofrece gran variedad de herramientas de análisis estadístico | Asociación Minera, Clasificación, Regresión, Series de tiempo, análisis estadist., Clustering | Windows, Solaris, Linux |
| Insightful Miner (Insightful Incorporation) | Comercial | Proporciona una interfaz visual, que permite a los usuarios juntar componentes para crear programas auto-documentado | La limpieza de datos, Clustering, Clasificación, Predicción, Análisis estadist. | Windows, Solaris, Linux |
| CART (Salford Systems) | Comercial | Proporciona división binaria para Clasificación (árbol de decisión) y para la Predicción (árboles de regresión). | Clasificación-Decisión y árbol de regresión | Windows, Linux |
| TreeNet(R) (Salford Systems) | Comercial | Proporciona una selección automática de predictores candidatos, Capacidad para manejar datos sin procesamiento previo. | Clasificación, Regresión | Windows, Linux |

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


| | | | | |
|---|---|---|---|---|
| RandomForests (Salford Systems) | Comercial | Proporciona un alto nivel de exactitud de predicción y un conjunto innovador de pantallas gráficas para revelar patrones inesperados en los datos | Clustering | Windows, Linux |
| GeneSight (Inc. of EI Segundo,CA) | Comercial | Explora grandes conjuntos de datos de varios grupos experimentales utilizando la normalización avanzada, visualización y apoyo a la decisión estadística | Visualización, Kmeans, y de redes neuronales de Clustering, Análisis de series temporales | Windows, Mac, Linux |
| PolyAnalyst (Megaputer Intelligence) | Comercial | Proporciona toma de decisiones para obtener conocimiento a partir de grandes volúmenes de texto y datos estructurados | Clustering, Clasificación, Predicción, Asociación Minera | Windows, |
| iData Analyzer (Microsoft) | Abierto / Libre | Proporciona plataforma para el ambiente de aprendizaje visual | Procesador de Pre, ESX, Heurística de agente, Red neuronal, Regla Maker y reportes | Windows, Linux, Solaris |
| See5 and C5.0 (RuleQuest) | Abierto / Libre | Proporciona análisis de árbol de decisión, la versión comercial del algoritmo C4.5 DT | Árbol de decisión. | Windows, Unix |
| TANAGRA (SPAD) | Abierto / Libre | Proporciona análisis de datos de minería de datos para fines académicos. Capacidad para diseñar la interfaz gráfica de usuario, Agregar nuevo algoritmo | Análisis factorial de los datos mixtos, Algoritmos de Soporte de Máquinas de Vectores, Análisis de factor principal no iterativo. | Windows, Linux, Mac OS, Solaris |
| SIPINA (Ricco Rakotomalala Lyon, France) | Abierto / Libre | Proporciona un entorno para los algoritmos de aprendizaje supervisado, maneja datos tanto continuos como discretos | Algoritmo C4.5, ID3 | Windows, Linux |
| ORANGE (University of Ljubljana, Slovenia.) | Abierto / Libre | Proporciona visualización de datos de código abierto y de análisis para el principiante y expertos | Minería de textos y complementos de Bioinformática | Windows, Linux |
| ALPHA MINER (E-Business Technology Institute) | Abierto / Libre | Proporciona la mejor relación costo y rendimiento para aplicaciones de minería de datos | Funciones de minería de datos versátiles | Windows, Linux, Mac |
| WEKA ( University of Waikato, New Zealand) | Abierto / Libre | Proporciona algoritmos de aprendizaje automático para tareas de minería de datos. Muy adecuado para el desarrollo de nuevos sistemas de aprendizaje automático. | Pre procesamiento de datos, Clasificación, Regresión, Clustering, Reglas de asociación, y Visualización. | Windows, Linux |
| Carrot | Abierto / Libre | Proporciona componentes listos para usar para traer los resultados de la búsqueda de diversas fuentes | Clustering | Windows, Linux |

Tabla 2.1. Herramientas usadas para MDE

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

### 2.4.2. Principales métodos de Minería de Datos utilizados en EDM

#### 2.4.2.1. Clasificación

Los problemas de clasificación tienen como objetivo identificar las características que indican el grupo al que pertenece cada caso. Este patrón se puede utilizar tanto para comprender los datos existentes y para predecir cómo los nuevos casos se comportarán. La minería de datos crea modelos de clasificación mediante el examen de datos ya clasificados (casos) e inductivamente encuentra el patrón predictivo. Así la clasificación proporciona un modelo para describir datos futuros (Duda et al., 2001).

Los algoritmos de Clasificación utilizan arboles de decisión y los más representativos son:

**a) CLS (*Concept Learning Systems*)** (Hunt, 1966). El método de Hunt para construir un árbol de decisión a partir de un conjunto $T$ de casos de entrenamiento es muy simple. Siendo las clases $\{C_1, C_2,\ldots, C_k\}$, hay tres posibilidades:

- Si $T$ contiene uno o más casos y todos ellos son de la misma clase $C_j$, el árbol de decisión para el conjunto $T$ será una hoja etiquetada con la clase $C_j$.
- Si $T$ no contiene casos, es decir, si el conjunto de casos de entrenamiento queda vacío, el árbol de decisión será de nuevo una hoja, pero la clase asociada con dicha hoja tendrá que ser determinada utilizando información adicional. Por ejemplo, la etiqueta de la hoja podría ser escogida de acuerdo con algún conocimiento del dominio, como la clase mayoritariamente en el conjunto.
- Si $T$ contiene casos de distintas clases, la idea es dividir $T$ en subconjuntos que sean o conduzcan a agrupaciones uniformes de casos, esto es, conjuntos de casos correspondientes a una misma clase. Utilizando los casos de entrenamiento disponibles, se escoge

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

una pregunta para ramificar el árbol. Dicha pregunta, basada en uno de los atributos predictivos, tendrá una o más respuestas mutuamente excluyentes $\{O_1, O_2,\ldots O_n\}$. Entonces $T$ se divide en los subconjuntos $T_1, T_2,\ldots T_n$, donde $T_i$ contiene todos los casos de $T$ con respuesta $O_i$ a la pregunta escogida. El árbol de decisión para $T$ consistirá entonces en un nodo que identifica la pregunta realizada del que cuelgan tantas ramas como respuestas alternativas. El mismo método se aplica recursivamente para construir los subárboles correspondientes a cada hijo del nodo, teniendo en cuenta que a cada rama $i$, a cada hijo del nodo anterior, se le asigna el subconjunto de casos de entrenamiento $T_i$ correspondientes a la respuesta $O_i$.

La sucesiva división del conjunto de casos de entrenamiento prosigue hasta que los subconjuntos están formados por casos pertenecientes a la misma clase, o hasta que ya no se encuentre ninguna pregunta para seguir ramificando el árbol que suponga alguna mejora.

**b) ID3 (*Interactive Dichotomizer Version* 3)** (Quinlan, 1986). Basado en el algoritmo CLS (*Concept Learning Systems*). Su uso se engloba en la búsqueda de hipótesis o reglas, dado un conjunto de ejemplos. El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.). realiza un proceso recursivo sobre todas las ramas del árbol generado. Inicialmente, se le llama con el conjunto de ejemplos, el de atributos, y el nodo raíz del árbol de decisión que estará vacío. El proceso se realiza hasta que todos los ejemplos de la rama en cuestión pertenecen a una única clase. Sin embargo, esto no siempre es posible. En primer lugar, puede ocurrir que se hayan agotado todos los atributos y, sin embargo, sigan existiendo ejemplos con distintos valores de clase. De otra parte, también puede suceder que, una vez elegido un atributo para un nodo de decisión, no exista ningún ejemplo para una de las ramas generadas por dicho

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

atributo. En esos casos, se etiqueta el nodo hoja con la clase mayoritaria. En los otros casos, se selecciona un atributo de acuerdo a la heurística definida anteriormente. Por cada valor del atributo escogido se crea un nodo sucesor de N etiquetado con el valor correspondiente del atributo, a través de la función crear-nodo. Ese nodo se añade al conjunto de nodos sucesores del nodo raíz N y, finalmente, se llama recursivamente a ID3. La llamada recursiva se hace con el conjunto de ejemplos que tienen ese valor concreto para el atributo escogido $A_i$, el conjunto de atributos menos el que acabamos de utilizar, y el nodo que acabamos de crear.

Nunca produce árboles demasiado grandes.

**c) C4.5**: (Quinlan, 1993): Es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan. C4.5 y su versión comercial C5.0 son una extensión del algoritmo ID3 desarrollado anteriormente por Quinlan. Los árboles de decisión generador por C4.5 pueden ser usados para clasificación, y por esta razón, C4.5 está casi siempre referido como un clasificador estadístico.

C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que lo hace ID3, usando el concepto de entropía de información. Los datos de entrenamiento son un grupo $S=s_1,s_2,\ldots$ de ejemplos ya clasificados. Cada ejemplo $si = x_1,x_2,\ldots$ es un vector donde $x_1,x_2,\ldots$ representan los atributos o características del ejemplo. Los datos de entrenamiento son aumentados con un vector $C=c_1,c_2,\ldots$ donde $c_1,c_2,\ldots$ representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C4.5 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

decisión. El algoritmo C4.5 divide recursivamente en sublistas más pequeñas.

Este algoritmo tiene unos pocos casos base.

- Todas las muestras en la lista pertenecen a la misma clase. Cuando esto sucede, simplemente crea un nodo de hoja para el árbol de decisión diciendo que elija esa clase.
- Ninguna de las características proporciona ninguna ganancia de información. En este caso, C4.5 crea un nodo de decisión más arriba del árbol utilizando el valor esperado de la clase.
- Instancia de la clase previamente no vista encontrada. Una vez más, C4.5 crea un nodo de decisión más arriba en el árbol con el valor esperado.

**d) ADTree (*Alternating Decision Tree*)** (Freund and Mason, 1999)**:** Un Árbol de decisión alternativo es un método de clasificación proveniente del aprendizaje automático conocido en inglés como Alternating Decision Tree (ADTree). Las estructuras de datos y el algoritmo son una generalización de los árboles de decisión.

Los ADTree contienen nodos splitter y nodos de predicción. El primero es un nodo que es asociado con una prueba, mientras que un nodo de predicción es asociado con una regla. Una instancia define una serie de caminos en un ADTree. La clasificación es asociada con una instancia que es el signo de la suma de las predicciones cercanas al camino en el que es definido por esta instancia.

**e) Decision Stump o árbol de decisión de un nivel:** Como bien dice su nombre se trata de árboles de decisión de un solo nivel.

Funcionan de forma aceptable en problemas de dos clases. No obstante, para problemas de más de dos clases es muy difícil encontrar tasas de error inferiores a 0.5. El propósito del algoritmo es construir un modelo de cada caso que será clasificada, tomando únicamente un subconjunto de casos de entrenamiento. Este subconjunto es escogido en base a la

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


distancia métrica entre las pruebas del caso y las pruebas de los casos dentro del espacio. Por cada caso de prueba, se hace una empaquetación conjunta de un árbol de un nivel clasificando así el aprendizaje de los puntos de entrenamiento cerrando la prueba actual del caso.

Los árboles de decisión de un nivel o decisión stump (DS) son árboles que clasifican casos, basados en valores característicos. Cada nodo en un árbol de decisión de un nivel representa una característica de un caso para ser clasificado, y cada rama representa un valor que el nodo puede tomar. Los casos son clasificados comenzando en el nodo raíz y se cataloga basándose en sus valores característicos. En el peor de los casos un árbol de decisión de un nivel puede reproducir el sentido más común, y puede hacerse mejor si la selección característica es particularmente informativa.

**f) LMT (*Logistic Model Tree*):** Un LMT consiste básicamente en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. Como en los árboles de decisión ordinarios, una prueba sobre uno de los atributos es asociado con cada nodo interno. Para enumerar los atributos con k valores, el nodo tiene k nodos hijos, y los casos son clasificados en las k ramas dependiendo del valor del atributo. Para atributos numéricos, el nodo tienen dos nodos hijos y la prueba consiste en comparar el valor del atributo con un umbral: un caso puede ser clasificar los datos menores en la rama izquierda mientras que los valores mayores en la rama derecha.

**g) M5P (Árbol de regresión):** Es un método de aprendizaje mediante árboles de decisión, utiliza el criterio estándar de poda M5. Es un árbol basado en árbol de decisión numérico tipo model tree. Las características son:

- Construcción de árbol mediante algoritmo inductivo de árbol de decisión.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


- Decisiones de enrutado en nodos tomadas a partir de valores de los atributos.
- Cada hoja tiene asociada una clase que permite calcular el valor estimado de la instancia mediante una regresión lineal.

**h) RandomForest:** Se basan en el desarrollo de muchos árboles de clasificación. Para clasificar un objeto desde un vector de entrada, se pone dicho vector bajo cada uno de los árboles del bosque.

Cada árbol genera una clasificación, el bosque escoge la clasificación teniendo en cuenta el árbol más votado sobre todos los del bosque.

Cada árbol se desarrolla como sigue:

- Si el número de casos en el conjunto de entrenamiento es *N*, prueba *N* casos aleatoriamente, pero con sustitución, de los datos originales. Este será el conjunto de entrenamiento para el desarrollo del árbol.
- Si hay *M* variables de entrada, un número $m<<M$ es especificado para cada nodo, m variables son seleccionadas aleatoriamente del conjunto *M* y la mejor participación de este *m* es usada para dividir el nodo. El valor de m se mantienen constante durante el crecimiento del bosque.
- Cada árbol crece de la forma más extensa posible, sin ningún tipo de poda.

**i) RandomTree:** Es un árbol dibujado al azar de un juego de árboles posibles. En este contexto "al azar" significa que cada árbol en el juego de árboles tiene una posibilidad igual de ser probado. Otro modo de decir esto consiste en que la distribución de árboles es "uniforme".

El proceso del RandomTree es un proceso que produce random trees de permutaciones arbitrarias.

Siguiendo la conceptualización, primero se marcan los vértices *n* por número 1 a través de $\binom{n}{2}$ de una manera aleatoria para cada que cada vértice tenga la misma probabilidad (este vértice es la permutación aleatoria).

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


Usando esta permutación, se comienza a construir un árbol sobre vértices de n: inicialmente, se tiene vértices de n y ninguna marca. En el paso *k-th* se intenta agregar el borde de *k-th* y ver si el gráfico resultante contiene un ciclo. Si es así, se salta el borde o línea de marca, además se agrega al gráfico y se repite para *k*+1. Durante este proceso el gráfico almacenará un bosque. Después de al menos $\binom{n}{2}$ pasos se obtendrá un árbol (un bosque conectado).

### 2.4.2.2. Regresión

La Regresión utiliza valores existentes para pronosticar cuáles serán los demás valores. En el caso más simple, la regresión utiliza técnicas estadísticas estándar, tales como la regresión lineal. Por desgracia, muchos de los problemas del mundo real no son simplemente proyecciones lineales de los valores anteriores. Por lo tanto, las técnicas más complejas (por ejemplo, de regresión logística, árboles de decisión, redes neuronales) o pueden ser necesarios para pronosticar valores futuros.

Los algoritmos de regresión están basados en árboles de regresión y árboles de modelos y los más representativos son:

**a) CART** (***Classification And Regression Tree***) (Breiman et al., 1984): El árbol de decisión CART es un procedimiento de partición recursiva binaria capaz de procesar atributos continuos y nominales, con una estrategia de poda basada en un criterio de coste-complejidad. Las particiones binarias son el resultado de evaluar una condición que tiene dos únicas respuestas. Se define el conjunto estándar de preguntas de la siguiente forma: Cada partición depende de solo un atributo:

- Si $X_i$ es un atributo categórico con valores $\{c_1, c_2,\ldots,c_n\}$, se pregunta si $X_i$ está entre un subconjunto de los valores categóricos, siendo la respuesta sí o no.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

- Si $X_i$ es un atributo continuo, *Q* incluye preguntas del tipo: ¿$X_i$ <=v? , siendo v un valor real cualquiera. Para simplificar, CART toma *v* como punto medio entre dos valores consecutivos de $X_i$.

Cada pregunta da lugar a una partición con una medida de impureza asociada. Esta medida de impureza es el método Gini.

El coeficiente de Gini se puede interpretar como un coeficiente donde se clasifica cada observación de un nodo a una clase *j* con probabilidad *p(j/t)* en lugar de clasificar todas las observaciones a la clase con mayor número de observaciones. Luego se selecciona la mejor puntuación (la mayor reducción de impureza). Así, se obtiene un árbol muy grande ($T_{max}$) donde los nodos terminales son muy homogéneos o están asociados a pocos patrones. Cuando se ha obtenido el árbol $T_{max}$ se procede a la poda mediante la construcción de una secuencia decreciente y anidada de árboles: $T_{max} = T_0 > T_1 > \ldots > \{t_i\}$ de manera que el árbol $\{t_i\}$ conste de un único nodo.

Cada árbol podado de $T_{i+1}$ se construye seleccionando de entre todos los subárboles de $T_i$, el subárbol con una menor medida de coste-complejidad asociada. Luego se calcula sobre cada árbol la tasa de error de clasificación para un conjunto de patrones de prueba independiente del conjunto de patrones utilizado en el entrenamiento, y así seleccionar el árbol con menor tasa de error asociada.

**b) M5** (Quinlan, 1993): Es una variación de CART que genera modelos lineales en las hojas, árbol de modelos, en lugar de valores numéricos. El nuevo heurístico para separar ejemplos no es la entropía de clases, sino la varianza del error en cada hoja.

Para la poda del árbol con modelos, primero se obtiene un modelo de regresión lineal para cada uno de los nodos interiores del árbol no podado*:* $a_0+a_1x_1++a_2x_2+\ldots+a_kx_k$, $x_1$, $x_2,\ldots,x_k$ son los atributos que aparecen únicamente en el subárbol que hay debajo.

El subárbol se poda si el error sobre nuevas instancias es inferior con el modelo que con él heurístico: $(n+v)/(n-v)$, *n* es el número de instancias

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

que alcanzan el nodo, *v* es el número de parámetros en el modelo lineal, permite simplificar el modelo quitando atributos.

## 2.5. CRISP-DM

La metodología de CRISP-DM (*Cross - Industry Standard Process for Data Mining*) describe los enfoques de uso común que los expertos usan para enfrentar a los problemas de minería de datos (Chapman et al., 2000).

CRISP-DM divide el proceso de minería de datos en seis fases principales, como se aprecia en la Figura 2.8:

### 2.5.1. Comprensión del negocio:

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

### 2.5.2. Comprensión de los datos:

La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

### 2.5.3. Preparación de los datos:

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas de

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

### 2.5.4. Modelado:

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

### 2.5.5. Evaluación:

En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.
Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.

### 2.5.6. Implantación:

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

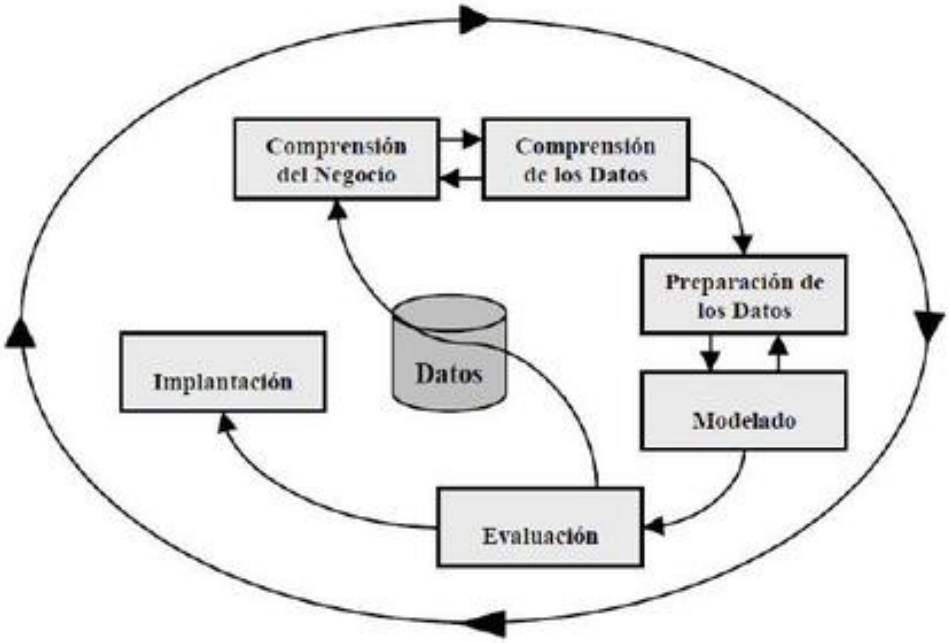



Fig. 2.8. Fases de la Metodología CRISP-DM. (Chapman et al., 2000)

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

# Capítulo 3

# Desarrollo de la Propuesta

Desarrollaremos la metodología de CRISP-DM siguiendo las fases y tareas que propone.

## 3.1. Comprensión del negocio

Esta fase inicial se enfoca en entender los objetivos y requerimiento del proyecto convirtiendo esto en la definición del problema de minería de datos.

*a) Determinación de objetivos de negocio*

La Universidad Católica de Santa María (UCSM) es una comunidad integrada por profesores, estudiantes, y graduados. Se dedica al estudio, la investigación, la educación, la difusión del saber, la cultura, y a la extensión y proyección social, bajo la inspiración de la fe cristiana.

Tiene como objetivos, entre otros, de docencia y de gestión integral:

- Índices adecuados de postulante/ingresante,
- Programas profesionales acreditados,
- Mejora permanente de los servicios de apoyo académico.
- Racionalización de procesos y funciones.
- Presupuesto equilibrado y sostenible

*b) Evaluación de la situación*

En el ámbito universitario, la UCSM adolece de la problemática de una tasa significativa de deserción estudiantil por diferentes motivos, no todos conocidos, lo que ocasiona que se enfrente a esa situación cada vez que se presente, afectando la planificación realizada por cada unidad académica.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

Los datos de los alumnos son capturados desde el momento en que ingresa a la Universidad, llenado una ficha de matrícula donde figura los siguientes atributos: Nombre y Apellidos, Dirección de vivienda, Numero de hermanos,

De la Oficina de Admisión se recaban los atributos: Puntaje de Ingreso, Opción de ingreso, Categoría.

Estos atributos han sido almacenados en diferentes fuentes de datos a lo largo del tiempo y se ha centralizado en la oficina de Informática de la Universidad.

Los recursos inicialmente disponibles son los siguientes:

- Materiales: Weka versión 3.7.11 como herramienta de Minería de datos, SQL Server Business Intelligence para el ETL.
- Humanos: El autor de la investigación, El Jefe de la Oficina de Informática, El Asesor de la investigación,
- Datos del trabajo: Se dispone de una serie de datos académicos de los alumnos del programa profesional de Ingeniería de Sistemas de la UCSM comprendida entre los años 2004 y 2008 que hacen un total de 651 registros de alumnos y 40067 registros de calificaciones.

*c) Determinación de los objetivos de la Minería de datos*

Objetivo de minería de datos: Dar soporte mediante técnicas de Minería de datos a los objetivos de la investigación

Objetivos específicos en la investigación:

- Realizar un estudio estadístico genérico de los datos.
- Encontrar la cantidad de alumnos que pertenecen a la categoría de abandono o no abandono.

Conocer estos objetivos permitirá determinar de una mejor manera la planificación en la apertura y distribución de secciones y grupos de las asignaturas.

## 3.2. Comprensión de los datos

*a) Recolección de los datos iniciales*

Los datos iniciales fueron proporcionados por la Oficina de Informática de la Universidad Católica de Santa María, los cuales se recogieron en diferentes

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


modalidades y momentos. Cuando un postulante desea ingresar a la Universidad se recaba cierta información, cuando el postulante alcanza un vacante y se matricula en un Programa Profesional se recoge otro tipo de información mediante un formato de ficha de matrícula.

Estos datos fueron entregados en el formato de una hoja de cálculo Excel como sigue:

Pestaña ALUMNOS_SISTEMAS:

Campos:

- Código
- Apellidos
- Nombres
- Categoría
- Opción
- Puntaje
- Fech. Nac. (Fecha de Nacimiento)
- Sexo
- Dirección
- Departamento
- Provincia
- Distrito
- Est. Civil (Estado Civil)
- Hijos
- Vive solo

Pestaña CALIFICACIONES:

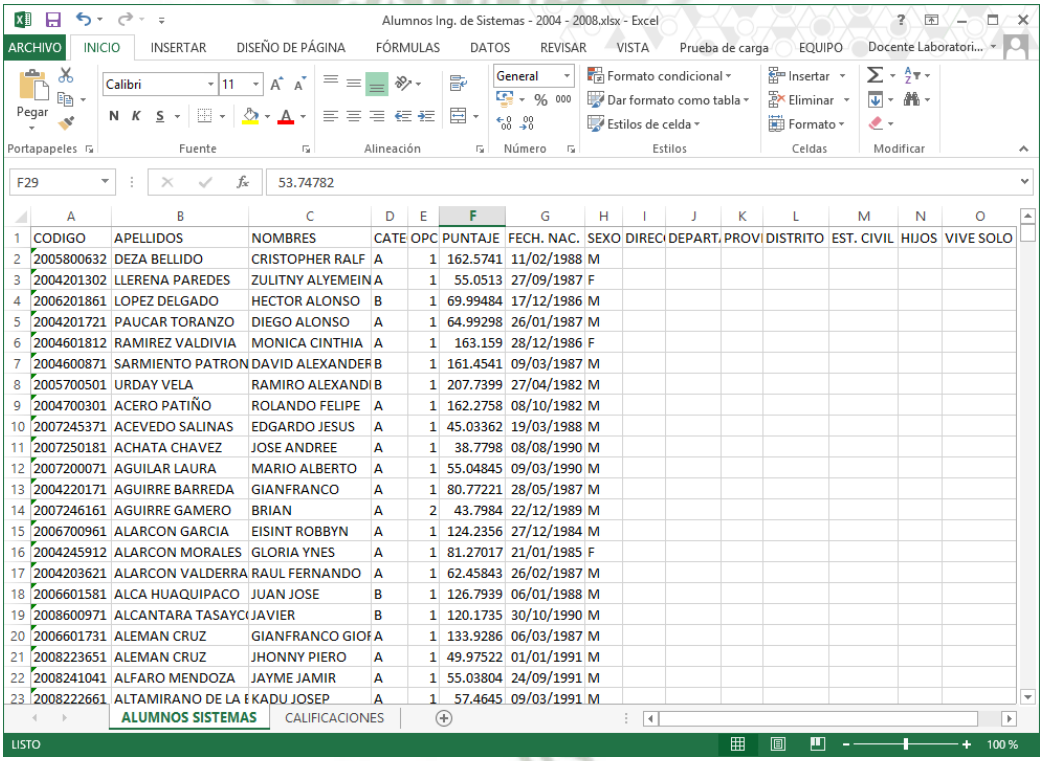
Campos:

- Código (Alumno)
- Cod_Asi (Código de la Asignatura)
- Asignatura
- Cre_Teo (Créditos Teóricos)
- Cre_Pra (Créditos Prácticos)
- Semestre

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


- Nota1 (Nota de Teoría 1)
- Nota2 (Nota de Teoría 1)
- Nota3 (Nota de Teoría 1)
- Pract1 (Nota de Práctica 1)
- Pract2 (Nota de Práctica 2)
- Pract3 (Nota de Práctica 3)
- Aplazados (Nota de Aplazados)
- Pro_Pra (Promedio de Práctica)
- Promedio (Promedio de la Asignatura)




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CODIGO | APELLIDOS | NOMBRES | CATE | OPC | PUNTAJE | FECH. NAC. | SEXO | DIREC | DEPART. | PROVI | DISTRITO | EST. CIVIL | HIJOS | VIVE SOLO |
| 2 | 2005800632 | DEZA BELLIDO | CRISTOPHER RALF | A | 1 | 162.5741 | 11/02/1988 | M | | | | | | | |
| 3 | 2004201302 | LLERENA PAREDES | ZULITNY ALYEMEIN | A | 1 | 55.0513 | 27/09/1987 | F | | | | | | | |
| 4 | 2006201861 | LOPEZ DELGADO | HECTOR ALONSO | B | 1 | 69.99484 | 17/12/1986 | M | | | | | | | |
| 5 | 2004201721 | PAUCAR TORANZO | DIEGO ALONSO | A | 1 | 64.99298 | 26/01/1987 | M | | | | | | | |
| 6 | 2004601812 | RAMIREZ VALDIVIA | MONICA CINTHIA | A | 1 | 163.159 | 28/12/1986 | F | | | | | | | |
| 7 | 2004600871 | SARMIENTO PATRON | DAVID ALEXANDER | B | 1 | 161.4541 | 09/03/1987 | M | | | | | | | |
| 8 | 2005700501 | URDAY VELA | RAMIRO ALEXAND | B | 1 | 207.7399 | 27/04/1982 | M | | | | | | | |
| 9 | 2004700301 | ACERO PATIÑO | ROLANDO FELIPE | A | 1 | 162.2758 | 08/10/1982 | M | | | | | | | |
| 10 | 2007245371 | ACEVEDO SALINAS | EDGARDO JESUS | A | 1 | 45.03362 | 19/03/1988 | M | | | | | | | |
| 11 | 2007250181 | ACHATA CHAVEZ | JOSE ANDREE | A | 1 | 38.7798 | 08/08/1990 | M | | | | | | | |
| 12 | 2007200071 | AGUILAR LAURA | MARIO ALBERTO | A | 1 | 55.04845 | 09/03/1990 | M | | | | | | | |
| 13 | 2004220171 | AGUIRRE BARREDA | GIANFRANCO | A | 1 | 80.77221 | 28/05/1987 | M | | | | | | | |
| 14 | 2007246161 | AGUIRRE GAMERO | BRIAN | A | 2 | 43.7984 | 22/12/1989 | M | | | | | | | |
| 15 | 2006700961 | ALARCON GARCIA | EISINT ROBBYN | A | 1 | 124.2356 | 27/12/1984 | M | | | | | | | |
| 16 | 2004245912 | ALARCON MORALES | GLORIA YNES | A | 1 | 81.27017 | 21/01/1985 | F | | | | | | | |
| 17 | 2004203621 | ALARCON VALDERRA | RAUL FERNANDO | A | 1 | 62.45843 | 26/02/1987 | M | | | | | | | |
| 18 | 2006601581 | ALCA HUAQUIPACO | JUAN JOSE | B | 1 | 126.7939 | 06/01/1988 | M | | | | | | | |
| 19 | 2008600971 | ALCANTARA TASAYC | JAVIER | B | 1 | 120.1735 | 30/10/1990 | M | | | | | | | |
| 20 | 2006601731 | ALEMAN CRUZ | GIANFRANCO GIOF | A | 1 | 133.9286 | 06/03/1987 | M | | | | | | | |
| 21 | 2008223651 | ALEMAN CRUZ | JHONNY PIERO | A | 1 | 49.97522 | 01/01/1991 | M | | | | | | | |
| 22 | 2008241041 | ALFARO MENDOZA | JAYME JAMIR | A | 1 | 55.03804 | 24/09/1991 | M | | | | | | | |
| 23 | 2008222661 | ALTAMIRANO DE LA F | KADU JOSEP | A | 1 | 57.4645 | 09/03/1991 | M | | | | | | | |

Figura 3.1. Hoja de Cálculo: Alumnos Ing. de Sistemas, 2004 – 2008, Pestaña ALUMNOS SISTEMAS

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM



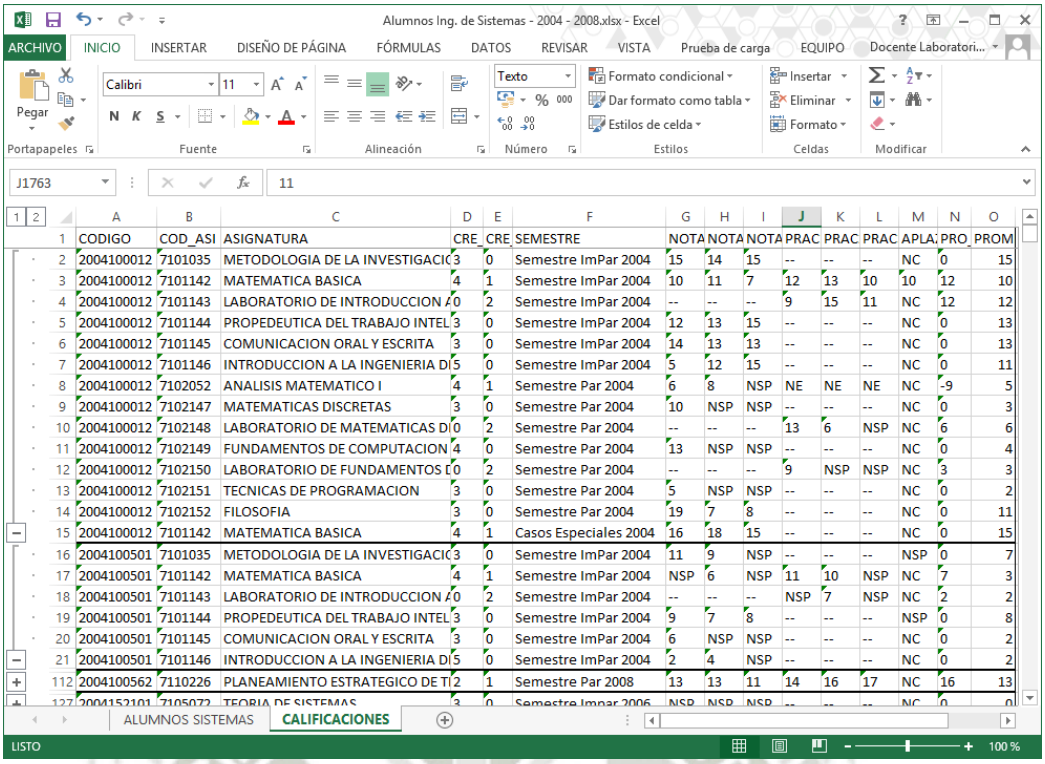



Figura 3.2. Hoja de Cálculo: Alumnos Ing. de Sistemas, 2004 – 2008, Pestaña CALIFICACIONES

*b) Describir los datos*

Los datos proporcionados por la Oficina de Informática, fue entregado en un archivo de hoja cálculo Excel correspondiente a los alumnos desde el año 2004 hasta el año 2008, esto con motivo de manejar sus notas de los que debieron acabar sus estudios en el lapso de los 5, 6, 7 años en algunos casos.

Todos los campos listados en el punto a) tienen un formato de campo del tipo TEXTO, ya que al parecer los datos solicitados a la Oficina de Informática fueron recabándose de diferentes tablas y se juntaron en una hoja de cálculo.

Los datos solicitados fueron en una mayor cantidad a los listados en el punto a) en muchos de los casos no se cuenta con algunos datos como el Estado Civil, si tiene Hijos, Vive solo, como se aprecia en la Figura 3.1.

*c) Explorar los datos.*

Al realizar la exploración de los datos, se apreció que se necesitaba algunos campos que sirvan como referencia significativa del alumno y luego se pueda

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

utilizar como fuente de entrada para los modelos en Weka, para ello se tuvo que realizar las siguientes actividades:

1. Transformar los campos necesarios del tipo Texto a su correspondiente tipo real, como por ejemplo el Promedio de cada asignatura convertirlo a tipo Numérico.
2. Generar nuevos campos como **Promedio Final** de todos los cursos para cada alumno, asignaturas **Aprobadas** y asignaturas **Desaprobadas**.

*d) Verificar la calidad de los datos*

En esta tarea se observó que existían algunos registros de alumnos:

- No tenían notas ni promedio alguno por lo que se procedió a identificarlos y omitirlos.
- Algunas notas figuran con un símbolo de -- o NC (Aplazados) que representa que no tienen nota, por tal motivo no se podían procesar numéricamente.

## 3.3. Preparación de datos

En esta etapa se convierte todas las actividades para construir la base de datos final (datos que alimentara a la herramienta de modelamiento). Para ello se procedió a crear una base de datos en Microsoft SQL Server y a migrar los datos de la Hoja de Cálculo a la nueva base de datos, además, la preparación tratará de:

- Eliminar valores anómalos, inconsistencias, valores ausentes, etc.
- Seleccionar datos a tratar.
- Modificar valores para su mejor tratamiento en función del algoritmo.

*a) Selección de datos*

La selección de los datos permitió distinguir: primero, que los campos que no tenían valor alguno (Dirección, EstadoCivil, Hijos, ViveSolo) se omitan, segundo, para aquellos registros que no están relacionados (alumnos que no tiene notas) no se consideren, de tal forma que solo se utilicen los datos que nos sirvan para el minado.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


*b) Limpieza de datos*

Las tareas de limpieza de datos a realizar son:

- Conversión de tipos: El campo de Promedio se convirtió de tipo de dato Texto a numérico.
- Modificar datos erróneos: algunos símbolos especiales por el alfabeto utilizado en la fuente de datos.
- Añadir datos faltantes.
- Trimming: Eliminar espacios sobrantes antes y después de las cadenas.

**Transformación ETL para la obtención de la Vista minable.**

Para lograr ello se utilizó el proceso ETL (Extraer, Transformar y Cargar) que permite mover datos desde diferentes fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, para analizar, o en otro sistema operacional para apoyar un proceso de negocio. La herramienta utilizada fue SQL Server Business Intelligence Development Studio como se aprecia en las figuras 3.3, 3.4 y 3.5

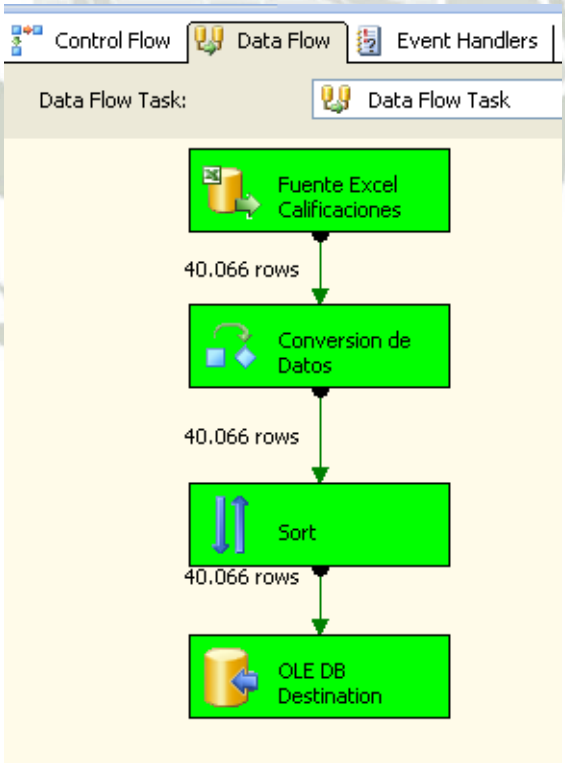



Figura 3.3. ETL para la conversión de datos.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



UNIVERSIDAD
CATÓLICA
DE SANTA MARÍA

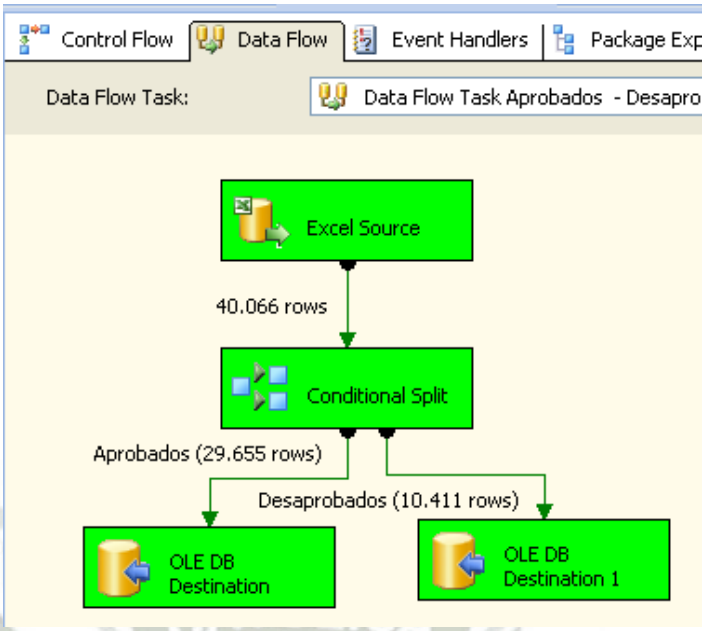



Figura 3.4. ETL para determinar las notas Aprobados y Desaprobados

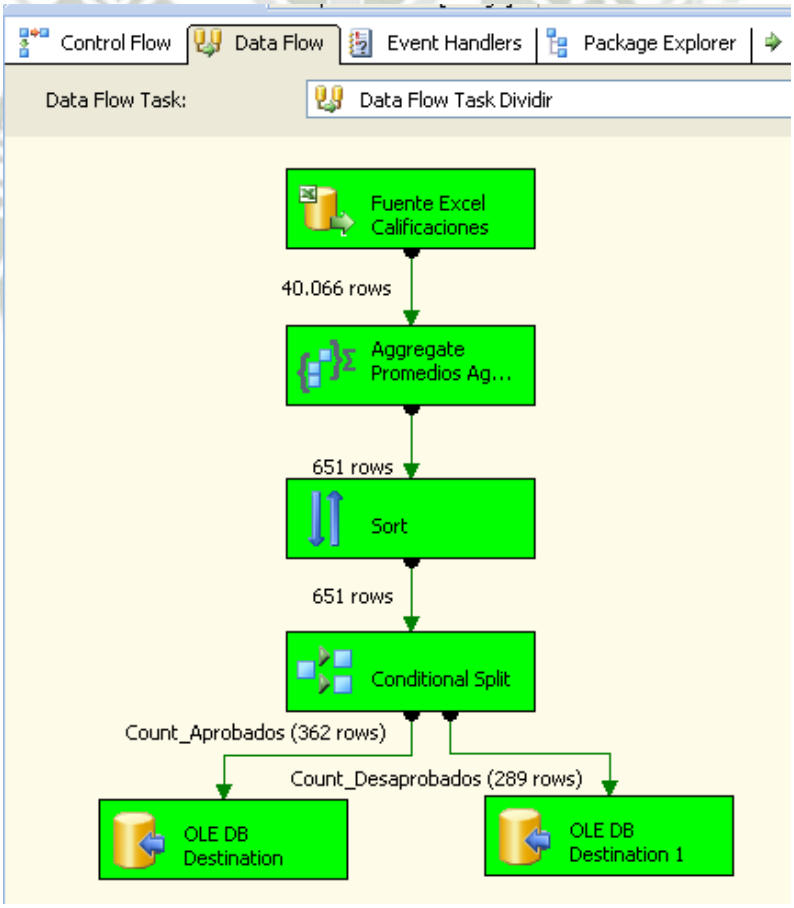



Figura 3.5. ETL para generar el Promedio Final y la cantidad deAprobados y Desaprobados

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


*c) Construir datos*

La construcción de los datos ha permitido generar las cuatro tablas que permitan organizar mejor lo datos quedando como sigue: ASIGNATURAS, CALIFICACIONES, ALUMNOS y ALUMNOS_DETALLE con sus respectivos atributos (campos) como se observa en la Figura 3.6:

Figura 3.6. Diagrama de la Base Datos Académico

La tabla ALUMNOS_DETALLE es la que se va a utilizar en la Herramienta de Minería de Datos WEKA, para ello se ha generado el correspondiente archivo .cvs (comma-separated values) con los siguientes campos: Codigo, Categoría, Puntaje, Aprobados, Desaprobados, TotalCursos, PromFinal, Abandono.

*d) Integrar datos*

La integración de los datos se ha realizado para obtener la tabla ALUMNOS_DETALLE que es la que finalmente se ha utilizado para aplicar la Minería de datos, en esta se encuentra los siguientes atributos: Categoría, Puntaje (Puntaje de ingreso), Aprobados (Asignaturas aprobadas), Desaprobados (Asignaturas desaprobadas), TotalCursos (Total de Asignaturas), PromFinal (PromedioFinal) y Abandono (Si / No)

*e) Formatear datos*

El formateo de los campos se ha realizado para tener una mejor comprensión en el aspecto sintáctico sin cambiar su significado, dándole nuevas denominaciones y etiquetas más comprensibles.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

## 3.4. Modelado

*a) Selección de la técnica de modelado*

Antes de seleccionar el modelo apropiado, debemos de enfocarnos en nuestro objetivo: ¿Cuál el propósito de buscamos?, el propósito que buscamos es la predicción de la deserción académica.

A continuación decidimos el tipo de predicción más apropiado, que será el de clasificación que predice en que categoría o clase caerá el resultado, en nuestro caso será de Abandona y No abandona. Entonces nuestro modelo elegido será un árbol de decisión para la clasificación.

Para ello comparamos algunos clasificadores populares: J48 es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta Weka de minería de datos. Luego también utilizaremos RandomTree para realizar la tarea comparativa.

*b) Generación de la prueba de diseño*

Se ha generado un archivo de prueba con registros adicionales, para las tareas de Minería de datos supervisadas como clasificación, por lo tanto se separa el conjunto de datos en datos originales y datos de prueba, se construye el modelo sobre el conjunto originales, y se evalúa el resultado del conjunto de prueba.

*c) Construcción del modelo*

Para llevar a cabo el objetivo de esta investigación, aplicaremos minería de datos usando el paquete de software Weka.

Weka es un acrónimo de *Waikato Environment for Knowledge Analysis*, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

WEKA se distribuye como software de libre distribución desarrollado en Java.

Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación, y

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados.

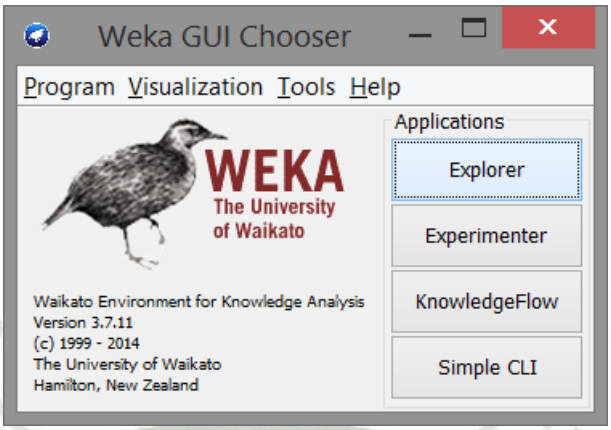



Figura 3.7. Ventana de inicio de Weka

Los datos de entrada a la herramienta, sobre los que operarán las técnicas implementadas, deben estar codificados en un formato específico, para nuestro caso fue *.csv*, que es un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en la que las columnas se separan por comas y las filas por saltos de línea.

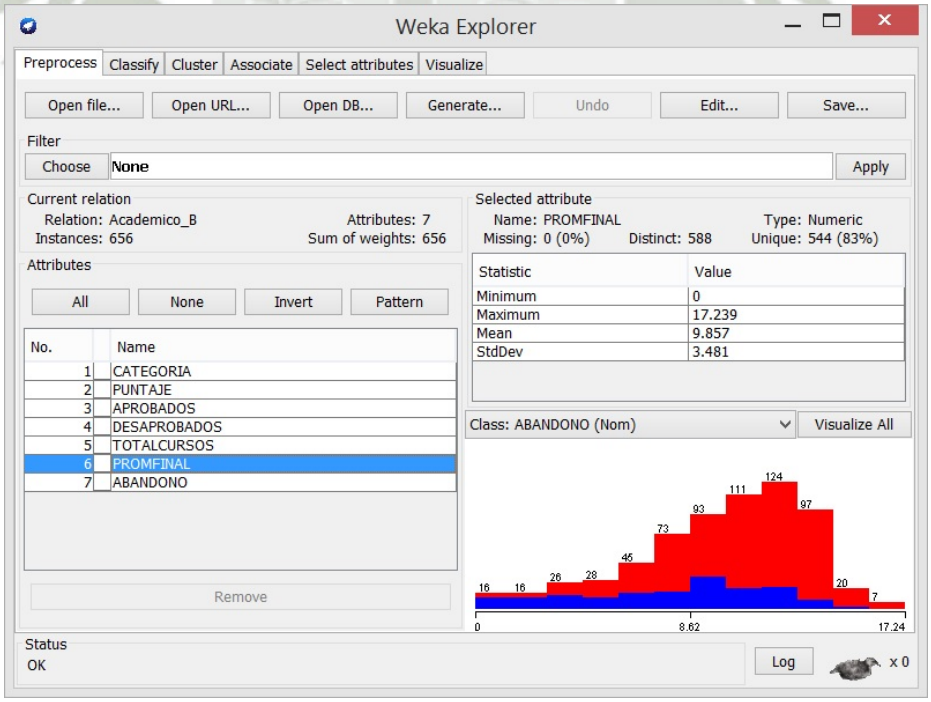



Figura 3.8. Archivo .csv con los atributos cargados en Weka

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM


# Capítulo 4

# Pruebas y resultados

## 4.1. Pruebas

El principal objetivo del análisis predictivo es la mejora de la actividad académico del estudiante. El algoritmo de clasificación C4.5 (J48) se analiza usando los siguientes métodos (Kumar and Vijayalakshmi, 2011):

a) La precisión del algoritmo es medido usando la comparación de los datos originales con los de prueba de los estudiantes.
b) La eficiencia del algoritmo es medido comparando el algoritmo C4.5 (J48) con el algoritmo RandomTree.

### 4.1.1. De la precisión del algoritmo

Los datos originales generados en el archivo *.cvs* se cargan en la herramienta de minería de datos Weka y luego se selecciona el algoritmo J48.

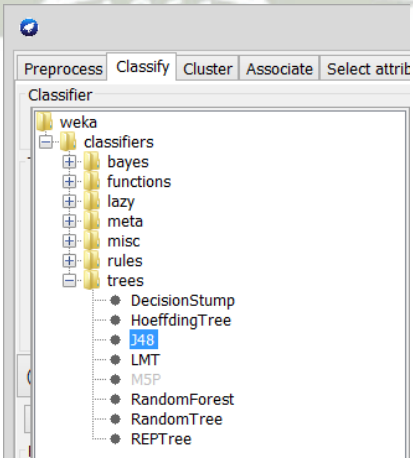



Figura 4.1. Selección del algoritmo J48 en Weka.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

Una vez seleccionado el algoritmo J48, se carga el archivo de Alumnos de Ing. de Sistemas con lo cual se aprecia como en la Figura 4.2.

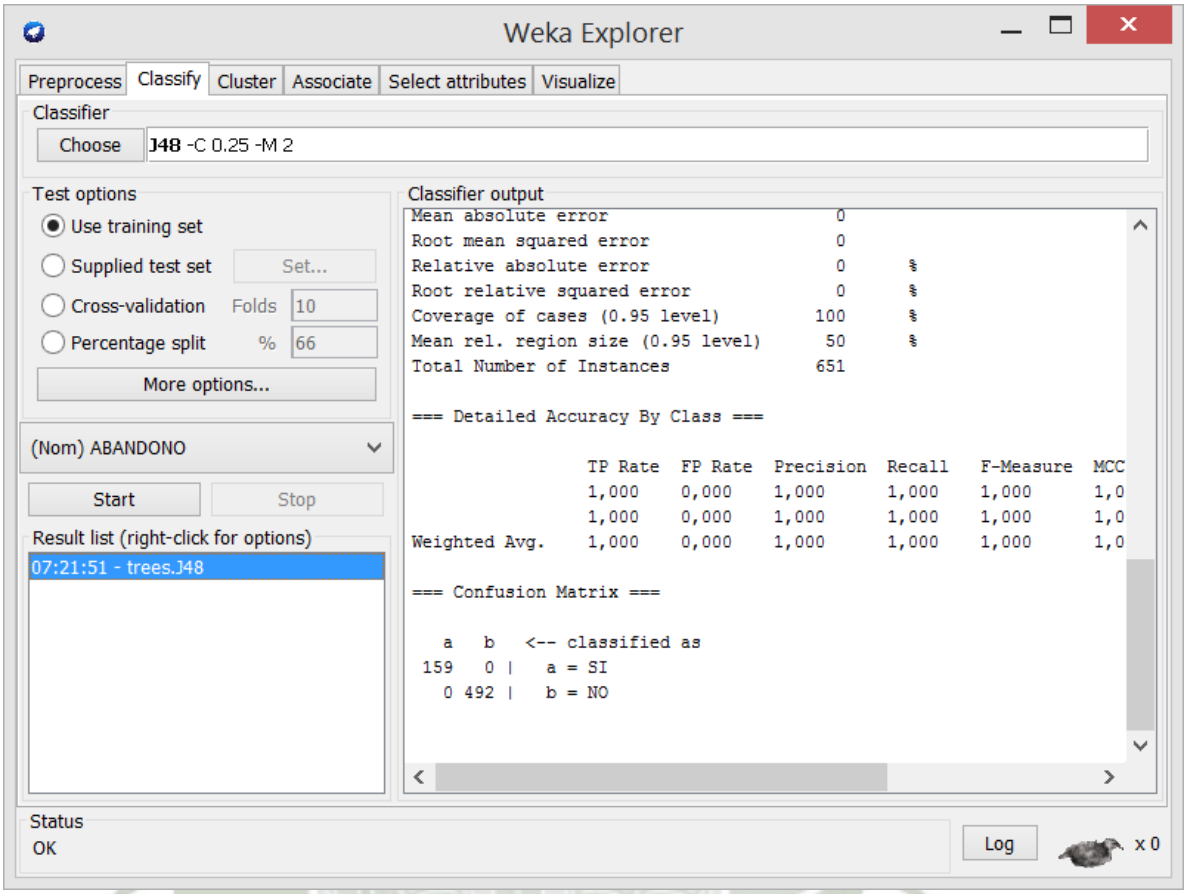



Figura 4.2. Ejecución del clasificador J48 con los datos Académicos

El resultado de salida de la ejecución del algoritmo se muestra a continuación:

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Academico

Instances: 651

Attributes: 7

CATEGORIA

PUNTAJE

APROBADOS

DESAPROBADOS

TOTALCURSOS

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

PROMFINAL

ABANDONO

Test mode: evaluate on training data

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

J48 pruned tree

---

TOTALCURSOS <= 20: SI (159.0)

TOTALCURSOS > 20: NO (492.0)

Number of Leaves : 2

Size of the tree : 3

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 651 | 100 % |
| Incorrectly Classified Instances | 0 | 0 % |
| Kappa statistic | 1 | |
| Mean absolute error | 0 | |
| Root mean squared error | 0 | |
| Relative absolute error | 0 % | |
| Root relative squared error | 0 % | |
| Coverage of cases (0.95 level) | 100 % | |
| Mean rel. region size (0.95 level) | 50 % | |
| Total Number of Instances | 651 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | SI |
| | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | NO |
| Weighted Avg. | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 159 | 0 | a = SI |
| 0 | 492 | b = NO |

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM


La matriz de confusión obtenida con el algoritmo J48:

| Instancias clasificadas como Abandona | Instancias clasificadas como No Abandona |
|---|---|
| 159 | 0 |
| 0 | 492 |

Tabla 4.1. Matriz de confusión obtenido por el algoritmo J48

El árbol obtenido se observa en la Figura 4.3

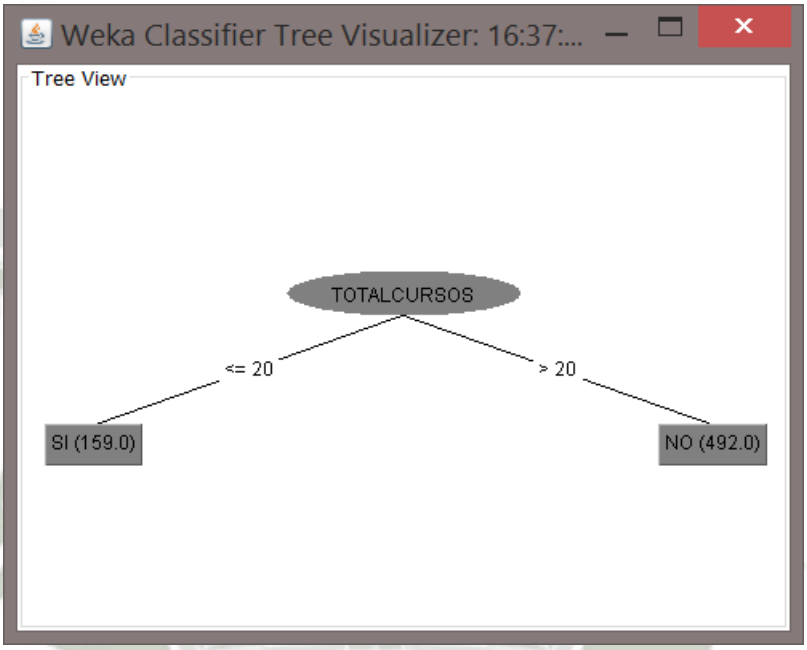



Figura 4.3. Arbol creado con el algoritmo J48.

Las gráficas obtenidas muestran las diferentes relaciones del atributo Abandono con los demás atributos considerados: Categoría, Puntaje de ingreso, Aprobados, Desaprobados, Total de cursos, y Promedio _nal, como se aprecia desde la Figura 4.4 hasta la Figura 4.9.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



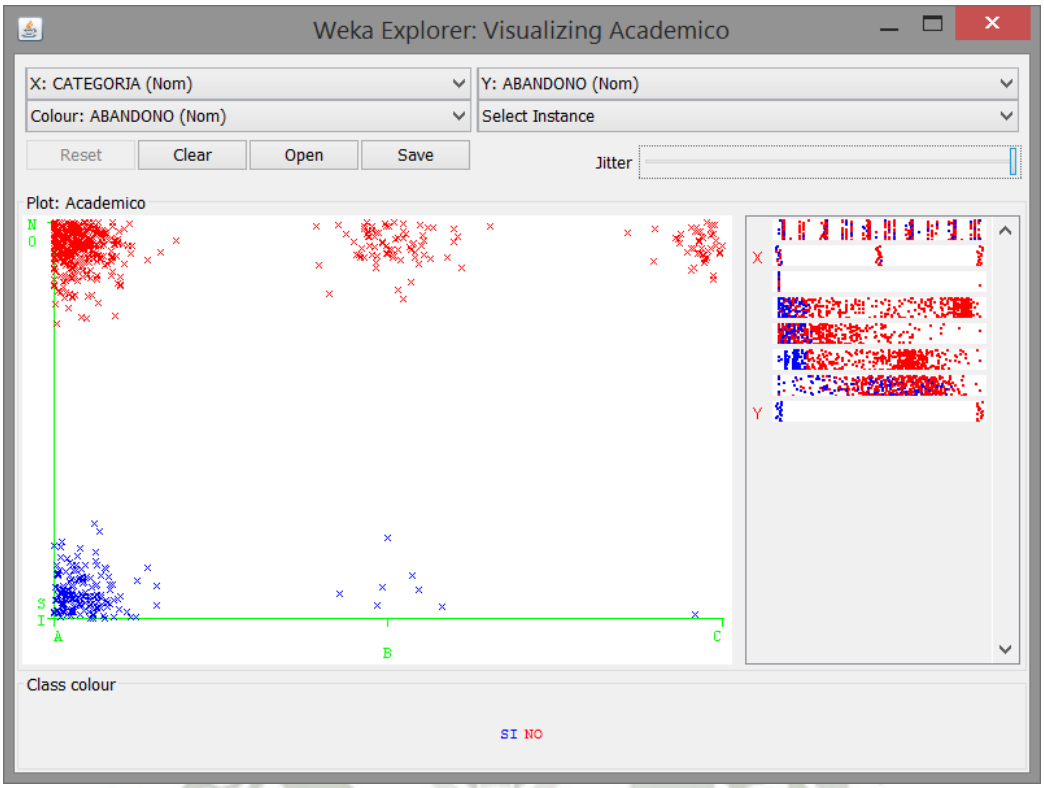



Figura 4.4. Abandono de de estudiantes en relación a la Categoría.

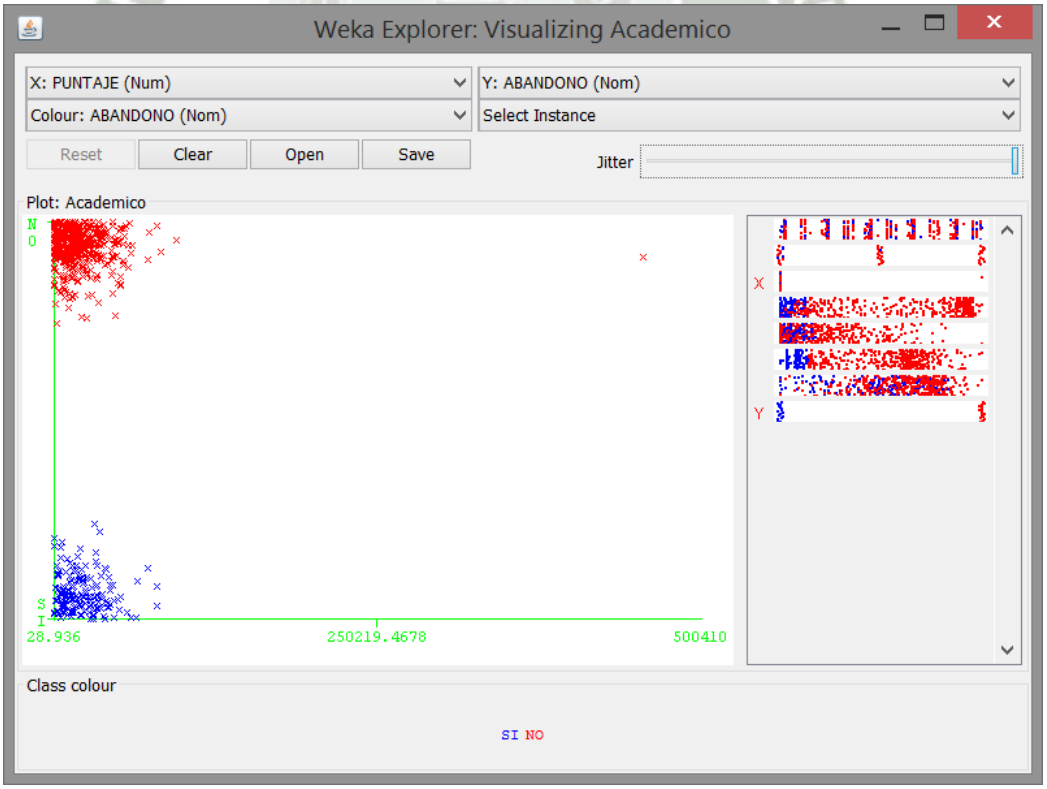



Figura 4.5. Abandono de de estudiantes en relación al Puntaje de ingreso.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



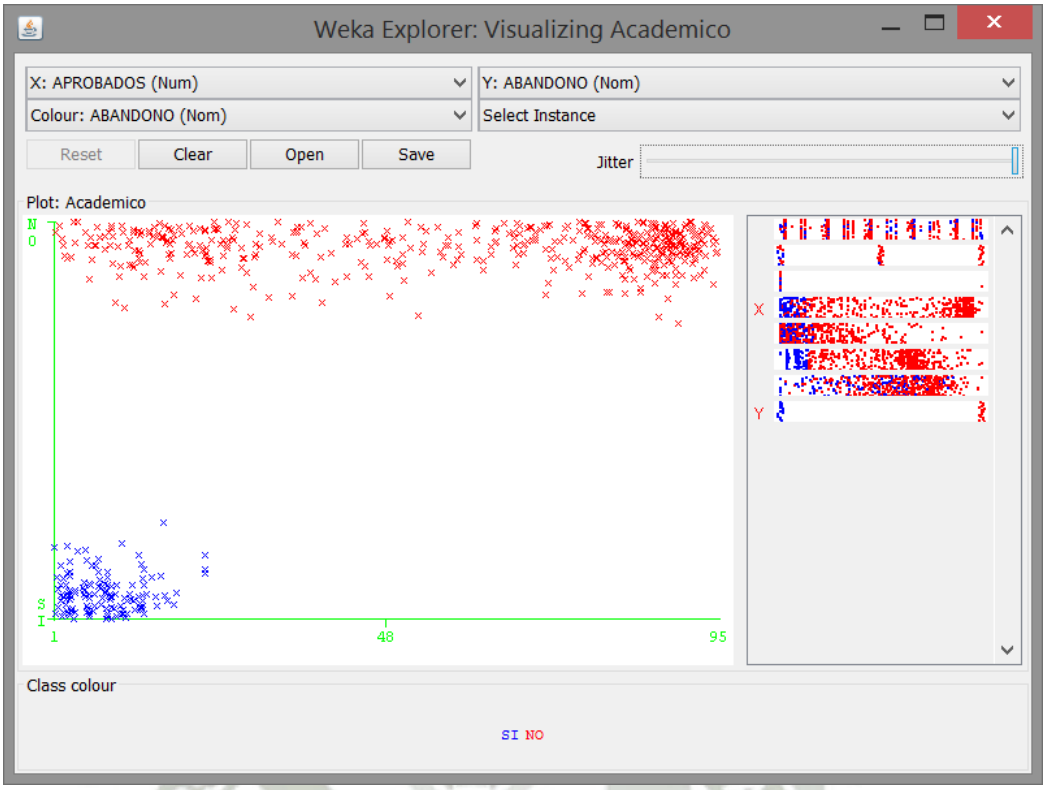



Figura 4.6. Abandono de de estudiantes en relación a los cursos Aprobados.

Figura 4.7. Abandono de de estudiantes en relación a los cursos Desaprobados.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



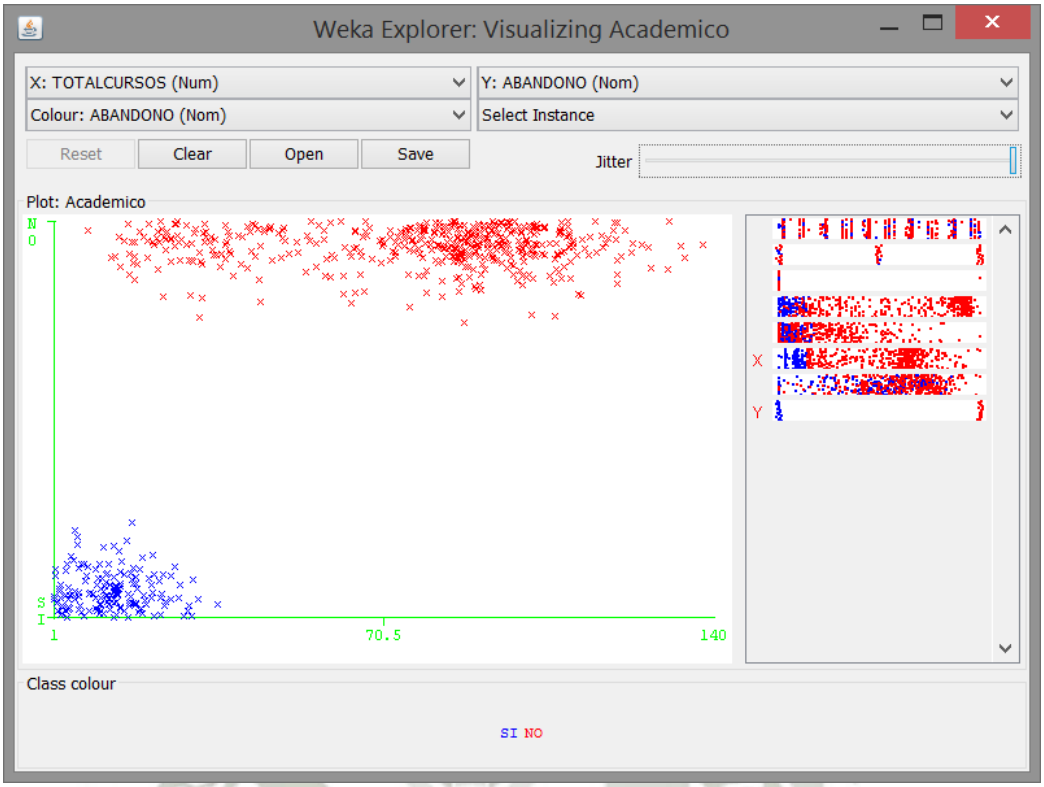



Figura 4.8. Abandono de de estudiantes en relación al Total de cursos.

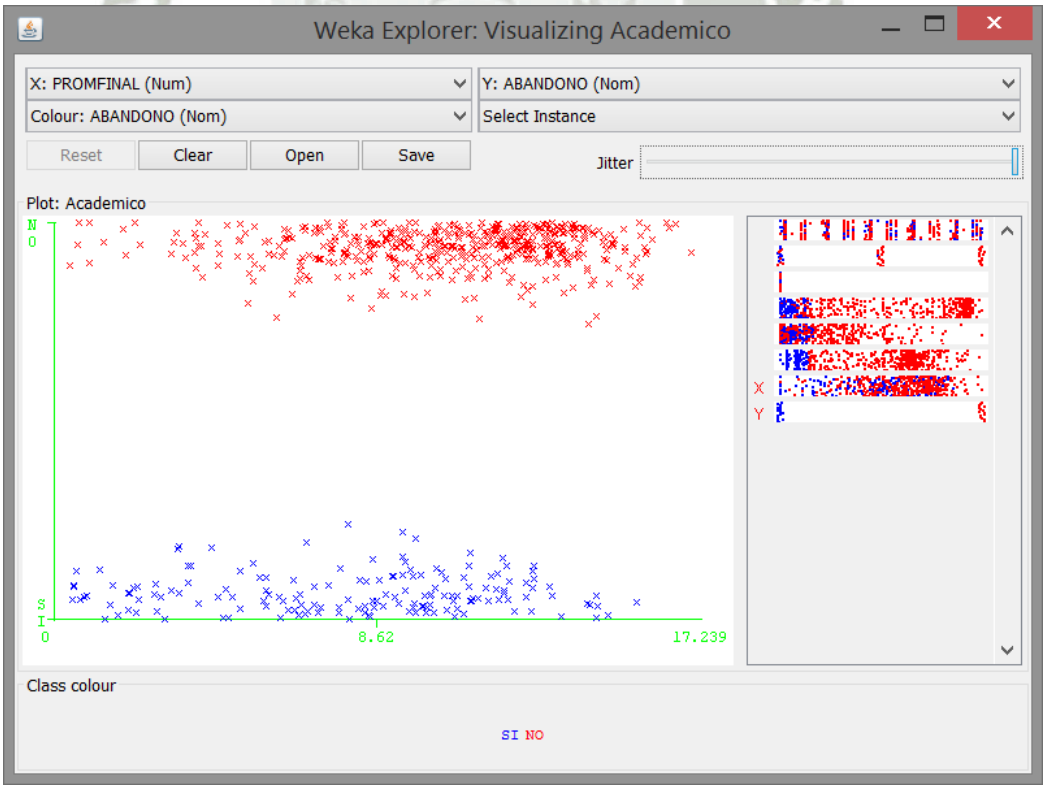



Figura 4.9. Abandono de de estudiantes en relación al Promedio final.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


Para los datos de prueba se generaron el doble de registros adicionales en un archivo *.cvs*, que se cargaron en la herramienta de minería de datos Weka y luego se aplicó el algoritmo J48.

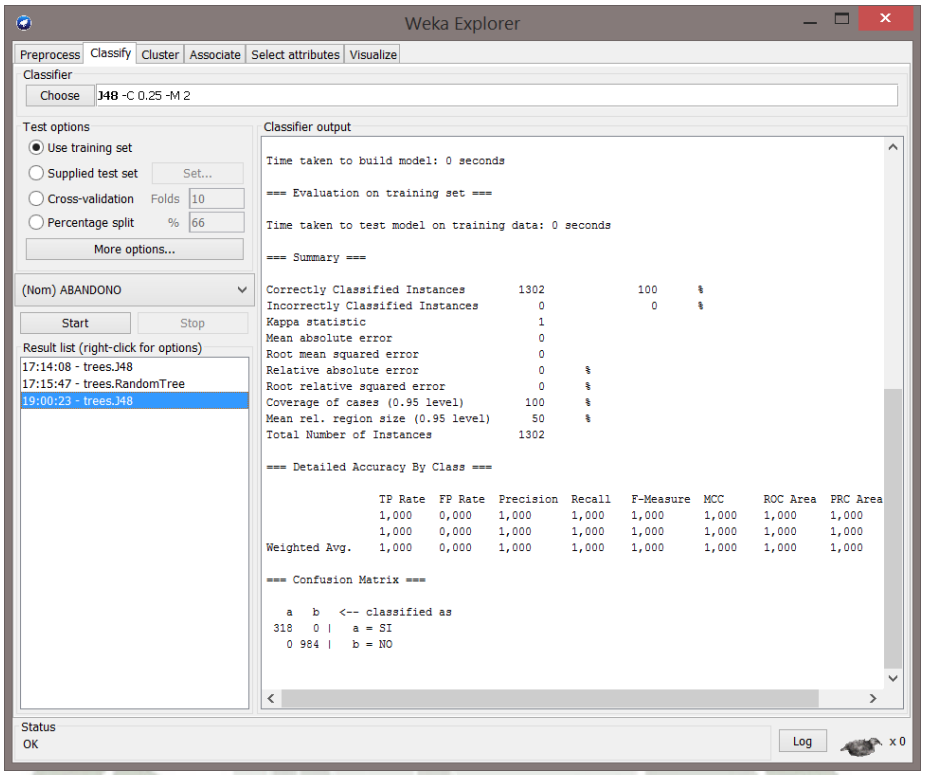



Figura 4.10. Ejecución del clasificador J48 con los datos de prueba

La matriz de confusión obtenida con los datos de prueba y el algoritmo J48:

| Instancias clasificadas como Abandona | Instancias clasificadas como No Abandona |
|---|---|
| 318 | 0 |
| 0 | 984 |

Tabla 4.2. Matriz de confusión obtenido por el algoritmo J48 con los datos de prueba

De las visualizaciones obtenidas se puede observar los resultados de utilizar el algoritmo J48 realizada en función del atributo Abandono del cual se puede reflejar lo siguiente:

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


- *Abandono y Categoria*: La mayor cantidad de estudiantes que abandonan la carrera es de la categoría A.
- *Abandono y Puntaje de Ingreso*: La mayor cantidad de estudiantes que abandonan tiene un puntaje de ingreso promedio de 80.77.
- *Abandono y cursos Aprobados*: La mayor cantidad de estudiantes que abandonan tienen un promedio de 6.23 cursos Aprobados.
- *Abandono y cursos Desaprobados*: La mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 6.94 cursos Desaprobados.
- *Abandono y Total de cursos*: La mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 11.25 cursos llevados.
- *Abandono y Promedio final*: La mayor cantidad de estudiantes que abandonan la carrera tienen un promedio final de 7.84.

### 4.1.2. De la eficiencia del algoritmo

La eficiencia del algoritmo se ha medido comparando el algoritmo J48 con el algoritmo RandomTree.

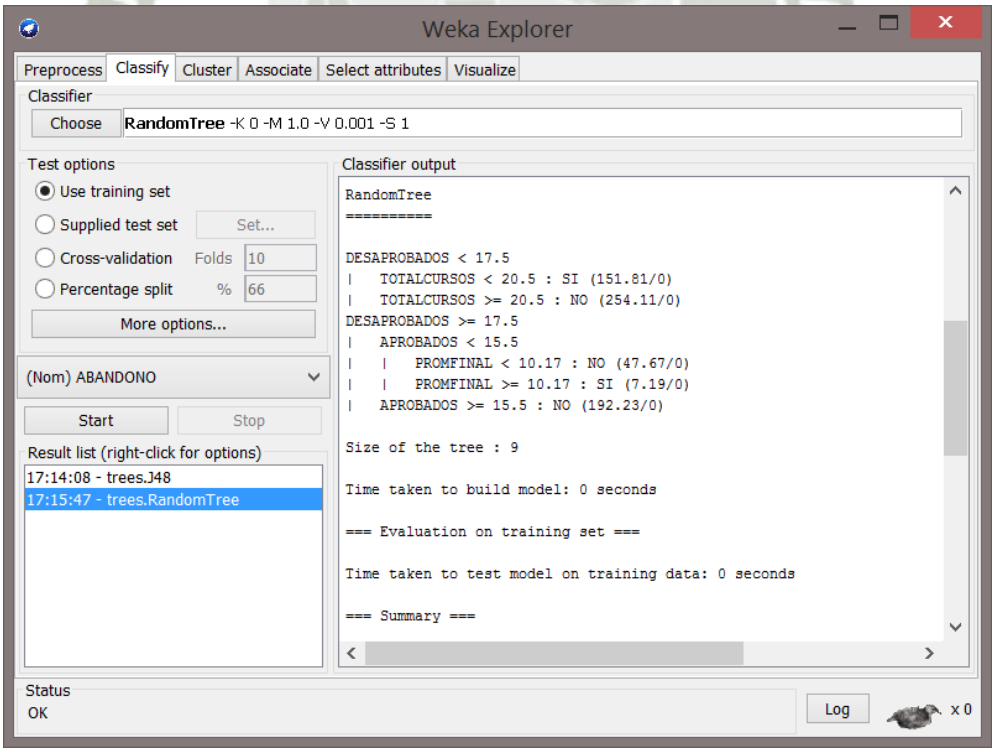



Figura 4.11. Ejecución del clasificador RandomTree con los datos Académicos.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


El árbol generado de la ejecución del algoritmo RandomTree con los datos Académicos de los alumnos de Ingeniería de Sistemas es:

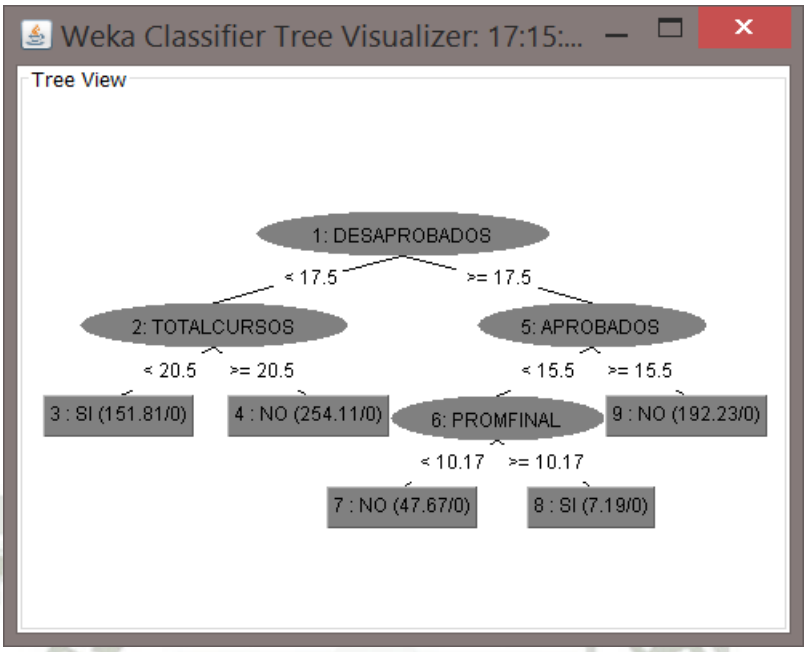



Figura 4.12. Arbol creado con el algoritmo RandomTree en Weka.

El resultado de salida de la ejecución del algoritmo se muestra a continuación:

```
=== Run information ===
Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     Academico
Instances:    651
Attributes:   7
              CATEGORIA
              PUNTAJE
              APROBADOS
              DESAPROBADOS
              TOTALCURSOS
              PROMFINAL
              ABANDONO
Test mode:    evaluate on training data

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

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

RandomTree

==========

DESAPROBADOS < 17.5

| TOTALCURSOS < 20.5 : SI (151.81/0)

| TOTALCURSOS >= 20.5 : NO (254.11/0)

DESAPROBADOS >= 17.5

| APROBADOS < 15.5

| | PROMFINAL < 10.17 : NO (47.67/0)

| | PROMFINAL >= 10.17 : SI (7.19/0)

| APROBADOS >= 15.5 : NO (192.23/0)

Size of the tree : 9

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 651 | 100 % |
| Incorrectly Classified Instances | 0 | 0 % |
| Kappa statistic | 1 | |
| Mean absolute error | 0 | |
| Root mean squared error | 0 | |
| Relative absolute error | 0 | % |
| Root relative squared error | 0 | % |
| Coverage of cases (0.95 level) | 100 | % |
| Mean rel. region size (0.95 level) | 50 | % |
| Total Number of Instances | 651 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|---|
| | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | SI |
| | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | NO |
| Weighted Avg. | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | |

=== Confusion Matrix ===

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM


```
  a   b   <-- classified as
159   0 |  a = SI
  0 492 |  b = NO
```

La matriz de confusión obtenida con los datos de prueba y el algoritmo RandomTree:

| Instancias clasificadas como Abandona | Instancias clasificadas como No Abandona |
|---|---|
| 159 | 0 |
| 0 | 492 |

Tabla 4.3. Matriz de confusión obtenido por el algoritmo RandomTree

Al mantener todas las instancias común, la tabla anterior especifica claramente que el número de instancias declaradas como Abandono es igual en ambos algoritmos y lo mismo en las instancias declarados de No Abandono, pro en el proceso de construcción del árbol el algoritmo RandomTree es más detallado para obtener los resultados. La Tabla 4.4. muestra el análisis comparativo de ambos los algoritmos.

| Algoritmo | J48 | RandomTree |
|---|---|---|
| Instancias clasificadas como Abandona | 159 | 159 |
| Instancias clasificadas como No Abandona | 492 | 492 |
| Tiempo tomado | 0.02 segundos | 0 segundos |

Tabla 4.4. Comparación del algoritmo C4.5 (J48) y RandomTree

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM


# Conclusiones

1. Las diversas técnicas supervisadas de minería de datos se pueden aplicar de manera efectiva en los datos educativos prediciendo la deserción académica.
2. Del análisis de las técnicas de clasificación se puede aplicar en los datos educativos para predecir el resultado de abandono o no del estudiante a los estudios y conocer la cantidad tentativa de grupos o secciones que se pueden asignar para su planificación.
3. Las pruebas realizadas en el caso de estudio con el programa profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María, en cuanto a la relación de atributos Abandono y Categoría: demuestran que la mayor cantidad de estudiantes que abandonan la carrera es de la categoría A. La relación de los atributos Abandono y Puntaje de Ingreso demuestran que la mayor cantidad de estudiantes que abandonan tiene un puntaje de ingreso promedio de 80.77. La relación de los atributos Abandono y cursos Aprobados demuestran que la mayor cantidad de estudiantes que abandonan tienen un promedio de 6.23 cursos Aprobados. La relación de los atributos Abandono y cursos Desaprobados demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 6.94 cursos desaprobados. La relación de los atributos Abandono y Total de cursos demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 11.25 cursos llevados. La relación de los atributos Abandono y Promedio Final demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio final de 7.84.
4. La eficiencia de los algoritmos de árboles de decisión C4.5 (J48) y RandomTree se puede analizar en función de su precisión comprobando que ambos algoritmos obtienen los mismos resultados.
5. En cuanto al tiempo tomado para derivar el árbol, se observa que el algoritmo RandomTree es ligeramente más veloz que el algoritmo J48 en la construcción del árbol de decisión.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



6. El análisis de los resultados realizados en la Universidad es una prueba para su implementación.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



# Recomendaciones

1. Ampliar los atributos determinantes como los aspectos psicológicos y otros para realizar la predicción categórica y obtener resultados más afinados.
2. Realizar la aplicación de otros algoritmos de clasificación para comparar los resultados y comprobar su eficiencia.
3. Utilizar predictores con diferentes variables de estudio así como en otras muestras temporales para observar si se encuentran resultados similares a los de la presente muestra.
4. Incorporar criterios difusos a los atributos de tal forma que pueda considerar cierto grado de verdad.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



# Bibliografía

Breiman, L., Friedman, J., and Olshen, R. (1984). Stone. *CJ: Classi_cation and Regression Trees, Wadsworth*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). Pattern classification. 2nd. *Edition. New York*.

EducationalDataMining (2010). http://www.educationaldatamining.org/.

El-Halees, A. (2009). Mining students data to analyze e-learning behavior: A case study. In *Proceedings of the 2008 International Arab Conference of Information Technology (ACIT2008), University of Sfax, Tunisia, Dec*, pages 15-18.

Enders, W. (2004). Applied time series econometrics. *Hoboken: John Wiley and Sons. ISBN X*, 52183919.

FERRI, H. and HERNANDEZ, J. (2006). Práctica de minería de datos, introducción al weka. *Universidad Politécnica de Valencia*.

Freund, Y. and Mason, L. (1999). The alternating decision tree learning algorithm. In *ICML*, volume 99, pages 124{133.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Han, J. and Kamber, M. (2006). *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann.

Hernández Orallo, J., Ramírez Quintana, M., and Ferri Ramírez, C. (2004). Et al.]: Introducción a la minería de datos, departamento de sistemas informáticos y computación. universidad politécnica de valencia.

Hunt, E. (1966). J.. marin and pt stone, experiments in induction.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

Jindal, R. and Borah, M. D. (2013). Asurvey on educational data mining and.

Kannan, K. S., Sekar, P. S., Sathik, M. M., and Arumugam, P. (2010). Financial stock market forecast using data mining techniques. In *Proceedings of the International Multiconference of Engineers and computer scientists*, volume 1.

Kumar, S. A. and Vijayalakshmi, M. (2011). E_ciency of decision trees in predicting student's academic performance. In *First International Conference on Computer Science, Engineering and Applications, CS and IT*, volume 2, pages 335-343.

Logros (2011). http://www.logros.edu.pe/.

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. Editorial Paraninfo.

Lutu, P. E. and Engelbrecht, A. P. (2010). A decision rule-based method for feature selection in predictive data mining. *Expert Systems with Applications*, 37(1):602-609.

Minaei-Bidgoli, B., Kashy, D. A., Kortmeyer, G., and Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in Education, 2003. FIE 2003 33rd Annual*, volume 1, pages T2A-13. IEEE.

Mishra, D. et al. (2010). Predictive data mining: Promising future and applications. *Int. J. of Computer and Communication Technology*, 2(1):20-28.

Osmanbegovi_c, E. and Sulji_c, M. (2012). Data mining approach for predicting student performance. *Economic Review*, 10(1).

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1):81-106.

Quinlan, J. R. (1993). C4. 5: *programs for machine learning*, volume 1. Morgan kaufmann.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135-146.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601-18.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12-27.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis

REPOSITORIO DE
TESIS UCSM



Romero, C., Ventura, S., Espejo, P. G., and Herv_as, C. (2008). Data mining algorithms to classify students. In *EDM*, pages 8-17.

Tair, M. and El-Halees, A. M. (2012). Mining educational data to improve students performance: A case study. *International Journal of Information and Communication Technology Research*, 2(2).

Weiss, S. M. (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann.

Witten, I. H. and Frank, E. (2005a). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Publicación autorizada con fines académicos e investigativos
En su investigación no olvide referenciar esta tesis