UNIVERSIDAD PÚBLICA DE EL ALTO
CARRERA INGENIERÍA DE SISTEMAS

TESIS DE GRADO

"MODELO DE PREDICCIÓN BASADO EN MINERÍA DE DATOS SOBRE ÍNDICES DE DESERCIÓN DE ALUMNOS"
CASO: (UNIVERSIDAD PÚBLICA DE EL ALTO)

Para optar al título de Licenciatura en Ingeniería de Sistemas

MENCIÓN: INFORMÁTICA Y COMUNICACIONES

Postulante : Ivan Rodrigo Hidalgo Mamani
Tutor Metodológico : M. Sc. Ing. Enrique Flores Baltazar
Tutor Especialista : Ing. Juan Regis Muñez Sirpa
Tutor Revisor : Ing. Elías Carlos Hidalgo Mamani

EL ALTO – BOLIVIA
2020

DEDICATORIA

A Daniel, José, David y Sarahy mis hijos, que son la alegría y principal motivación en mi vida para no rendirme en los estudios.

A Mónica mi amada esposa, por su apoyo y comprensión en todo este sueño alcanzado.

A mis padres Tomas y Teresa (+), por mostrarme el camino hacia a la superación.

A mis hermanos, por brindarme su tiempo y apoyo**.**

Ivan Rodrigo Hidalgo Mamani

AGRADECIMIENTOS

A Dios, por permitirme la dicha de la vida, guiarme y darme la fortaleza en los momentos de dificultad y debilidad.

A mi familia, por la comprensión y apoyo incondicional, en todo este camino. Cada uno de ustedes representa un cimiento en el desarrollo de mi vida y en particular de este proyecto.

Al Ing. Enrique Flores, por haberme brindado su tiempo, guía y consejos a través de cada una de las etapas de este proyecto.

Al Ing. Regis Muñes, por su apoyo incondicional e interés prestado en la conclusión del presente proyecto.

Al Ing. Elías Hidalgo, por su colaboración, guía y ayuda durante este proyecto.

A los docentes de la carrera de Ingeniería de Sistemas de la Universidad Pública de El Alto, por haber compartido sus conocimientos y oportunidades de enriquecer mis conocimientos.

RESUMEN

El presente trabajo de investigación se elabora a partir de la inquietud de obtener conocimiento sobre el índice de deserción de los alumnos de la Universidad Pública de El Alto. Para lo cual se recurrió a la unidad de Sistemas de Información y Estadísticas (SIE). Son en base a los Datos proporcionados por esta unidad dependiente de la Universidad Pública de El Alto, la aplicación de Minería de Datos, como el desarrollo del prototipo PREDESMIN.

En la primera parte corresponde a la Introducción del trabajo de investigación; se realiza el planteamiento del problema, se fija el objetivo general y los objetivos específicos; se formuló la hipótesis y sus variables dependientes e independientes; se hace una referencia a justificación en aspecto científica, técnico, económica y social; finalmente se muestra las herramientas disponibles para la Minería de Datos y se ven los límites y alcances del proyecto.

En la segunda parte se proporciona la información y definiciones que se hacen necesarios para la comprensión del trabajo e investigación, como ser: los conceptos básicos, la Minería de Datos, la ingeniería de software, el lenguaje de modelado unificado, la deserción universitaria y las métricas de calidad.

En el capítulo tercero nos muestra el modelo de predicción en base a Minería de Datos; el lenguaje, la arquitectura que son necesarias para su implementación y las métricas de calidad usadas en el prototipo del presente trabajo de investigación.

En el cuarto capítulo se ven las pruebas y resultados de la aplicación del prototipo de predicción de índices de deserción, como la interpretación del mismo.

En el capítulo quinto se arriba a las conclusiones, de acuerdo a los objetivos planteados en la primera parte y las recomendaciones para futuros trabajos.

Palabras claves: Minería de Datos, Índices de deserción, Predicción

ABSTRACT

The present research work is elaborated from the concern to obtain knowledge about the dropout rate of the students of the Public University of El Alto. For this, the Information Systems and Statistics unit (SIE) was used. Based on the Data provided by this unit dependent on the Public University of El Alto, the Data Mining application, such as the development of the PREDESMIN prototype.

The first part corresponds to the Introduction of the research work; the problem statement is made, the general objective and specific objectives are set; the hypothesis and its dependent and independent variables were formulated; a reference is made to justification in scientific, technical, economic and social aspects; Finally, the tools available for Data Mining are shown and the limits and scope of the project are seen.

The second part provides the information and definitions that are necessary for the understanding of the work and research, such as: the basic concepts, Data Mining, software engineering, the unified modeling language, university dropouts and the quality metrics.

In the third chapter he shows us the prediction model based on Data Mining; the language, the architecture that are necessary for its implementation and the quality metrics used in the prototype of this research work.

In the fourth chapter, the tests and results of the application of the prototype for predicting dropout rates are seen, as well as its interpretation.

In the fifth chapter the conclusions are reached, according to the objectives set out in the first part and the recommendations for future work.

Keywords: Data Mining, Attrition rates, Prediction

ÍNDICE GENERAL

ÍNDICE ESPECIFICO

CAPÍTULO I

MARCO INTRODUCTORIO

CAPÍTULO II

MARCO TEÓRICO

CAPITULO III

MARCO APLICATIVO

ÍNDICE DE FIGURAS

CAPÍTULO I

CAPÍTULO II

CAPITULO III

CAPITULO IV

ÍNDICE DE TABLAS

CAPÍTULO I

CAPÍTULO II

CAPITULO III

CAPITULO IV

CAPITULO I

MARCO INTRODUCTORIO

MARCO INTRODUCTORIO

El presente capítulo se da a conocer las referencias de investigación relacionadas con la Minería de Datos y el índice de deserción universitaria respecto de la investigación, la definición de los objetivos y la hipótesis a ser comprobadas, también se describen los métodos de investigación y de ingeniería, las herramientas, los limites y los alcances del presente trabajo.

1.1. INTRODUCCIÓN

En el ámbito de la ciencia, una predicción es un anticipo de lo que ocurrirá de acuerdo al análisis de las condiciones existentes. Es frecuente que las predicciones surjan tras experimentos o investigaciones que permiten conocer las condiciones y estimar que, si se repiten el resultado será el mismo.

Los índices son indicios o señales de algo, pueden tratarse de la expresión de la relación entre dos cantidades o distintos tipos de indicadores.” Dato o información que sirve para conocer o valorar las características y la intensidad de un hecho o para determinar su evolución futura” Lexico (2020).

“Notoriamente los estudiantes que inician una carrera universitaria son categorizados según su promedio ponderado siendo así la clase irregular la más abordada en deserción durante los primeros años de estudios” Mamani (2019). La Minería de Datos es una de las principales herramientas que se utilizan dentro de los programas de gestión del conocimiento como base para la toma de decisiones.

La Minería de Datos o también llamada explotación de Datos, es el proceso que intenta encontrar patrones ocultos en grandes bases de datos. Apaza (2009), describe a la Minería de Datos como un proceso no trivial de identificación válida, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los Datos. Marquez (2006), menciona que “La Minería de Datos surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos, de forma general, los Datos son la materia prima bruta”.

En la actualidad los sistemas de información son una necesidad en el ámbito académico y administrativo, lo cual ha llevado a las universidades contar con una gran cantidad de información sobre la población estudiantil. La Universidad Pública de El Alto desde su creación hasta la fecha cuenta con un crecimiento considerable, tanto en su infraestructura como en su población de estudiantes, lo que significa que el volumen de información que maneja respecto a los estudiantes es relativamente grande. Actualmente la Universidad Pública de El Alto cuenta con un sistema de información respecto al registro de estudiantes activos desde el 2001 pero no así de las bajas y abandonos.

El presente trabajo pretende implementar un modelo de predicción basado en Minería de Datos para encontrar índices de deserción de los alumnos, el mismo que nos permitirá la toma de decisiones para los encargados de la planificación del área social de la universidad, entre ellos las autoridades que tiene esta casa superior de estudios. Ellos (los directivos) podrán implementar políticas para la reducción de los índices de deserción de los alumnos.

1.2. ANTECEDENTES

1.2.1. Antecedentes institucionales

La Universidad Pública de El Alto (UPEA) es una universidad pública y autónoma de Bolivia, con sede en la ciudad de El Alto, la cual ofrece 35 carreras en las áreas socio-político-económica, salud y tecnología. En el 2011 se implementaron 13 carreras sumando un total de 35 la oferta académica de esta

casa superior de estudios. Su accionar se enmarca en el estatuto orgánico aprobado en 2007.

Tiene como Visión:

La UPEA es una institución que se proyecta al desarrollo de sus actividades académico-productivas, científicas, tecnológicas de interacción social contemporáneo, para priorizar la investigación científica en todos los campos del conocimiento relacionando la teoría con la práctica para transformar la estructura económica, social, cultural y política vigente en favor de las naciones originarias y clases populares.

Tiene como Misión:

Formar profesionales integrales altamente calificados en todas las disciplinas del conocimiento científico-tecnológico, con conciencia crítica y reflexiva; capaz de crear, adaptar y transformar la realidad en que vive; desarrollar la investigación productiva para fomentar el desarrollo local, regional y nacional para que responda al encargo social y las necesidades de las nacionalidades de manera eficiente y oportuna hacia la transformación revolucionaria de la sociedad.

Y tiene como Objetivos:

Formar Profesionales idóneos a partir del desarrollo de la ciencia, la tecnología y el conocimiento científico, en un proceso único de interacción entre la teoría y la práctica, que permita transformar y desarrollar la realidad local, regional, nacional promoviendo de múltiples formas del bienestar del pueblo boliviano.

Formar profesionales con una concepción crítica contra hegemónica para el logro del poder político "de" y "para" las mayorías nacionales.

Desarrollar y difundir ciencia, tecnología y cultura dentro y fuera de la universidad.

1.2.2. Antecedentes Internacionales

Mamani (2019), nos plantea a través de su tesis titulado **"MODELO DE MINERÍA DE DATOS BASADO EN FACTORES ASOCIADOS PARA LA PREDICCIÓN DE DESERCIÓN ESTUDIANTIL UNIVERSITARIA**"**,** su objetivo general fue el de desarrollar e implementar un modelo de Minería de Datos para la predicción de deserción estudiantil universitaria.

Para el desarrollo de su trabajo utilizo una metodología cuantitativa de tipo descriptivo y experimental, así mismo empleo la metodología CRISP-DM para desarrollar el modelo predictivo, la Tesis se realizó en la Universidad Nacional de Moquegua, Moquegua-Perú.

Marcano y Rodríguez (2014), los aportes de estos investigadores respecto de su trabajo de investigación **"MINERÍA DE DATOS APLICADO A LA DESERCIÓN",** nos muestra tres elementos fundamentales; uno referido a la inferencia de los conocimientos adquiridos durante la permanencia en la educación media, es decir en la etapa de colegiatura; un segundo aspecto hace referencia a la falta de recursos económicos, los cuales no le permiten al alumno contar con los recursos didácticos acordes al modelo educativo; finalmente un tercer aspecto es la concentración en los estudios, es decir la horas que el alumno dedica al desempeño académico tanto en el hogar como en las casas superiores de estudio. El objetivo de la investigación hace uso de la aplicación de la Minería de Datos para obtener los patrones sobre los estudiantes que no han podido concluir sus estudios universitarios. La investigación fue de tipo descriptiva de campo y utilizo la metodología CRISP-DM, realizada en Universidad de Zulia, Barquisimeto-Venezuela.

Quintero (2016), presenta su investigación titulada "**ANÁLISIS DE LAS CAUSAS DE DESERCIÓN UNIVERSITARIA**", su objetivo general de la investigación fue el de tratar de identificar las causas que llevan al estudiante a la deserción universitaria, para luego proponer estrategias que ayuden a enfrentar la deserción en base a estos estudios de cada estudiante. La

investigación utilizo una metodología cualitativa, teniendo en cuenta las características de problema y lo que se buscaba indagar a través de él, la investigación se realizado en la Universidad Nacional Abierta y a Distancia UNAD, Bogotá-Colombia.

1.2.3. Antecedentes Nacionales

Apaza (2009), presenta su trabajo titulado **"MODELO DE PREDICCIÓN DE PREVALENCIA DE ENFERMEDADES PARA CENTROS DE SALUD BASADOS EN MINERÍA DE DATOS",** su objetivo mediante el uso de técnicas de Minería de Datos desarrollar un modelo de predicción de prevalencia de enfermedades para centros de salud de Bolivia. La fue desarrollar un modelo de Minería de Datos empleando una Red Neuronal Evolutiva para realizar análisis de datos en una Base de Datos de gestión académica para generar información útil, confiable y oportuna en el entorno.

Para el desarrollo de la investigación se utilizó en método científico en base al lineamiento de Roberto Hernández y Carlos Fernández, la tesis se realizó en la Universidad Mayor de San Andrés, La Paz- Bolivia.

Márquez (2006), da a conocer su tesis titulada "**DESARROLLO DE UN MODELO DE MINERÍA DE DATOS ACADÉMICO**", su objetivo general fue desarrollar un modelo de Minería de Datos empleando una Red Neuronal Evolutiva para realizar análisis de datos en una Base de Datos de gestión académica para generar información útil, confiable y oportuna en el entorno.

Para el desarrollo de la investigación se utilizó en método científico en base al lineamiento de Roberto Hernández y Carlos Fernández, la tesis se realizó en la Universidad Mayor de San Andrés, La Paz- Bolivia.

1.2.4. Antecedente Locales

Hidalgo (2014), nos presenta la investigación titulada **"APLICACIÓN DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL RENDIMIENTO ACADÉMICO EN ESTUDIANTES DE SECUNDARIA",** su objetivo es aplicar herramientas de

Minería de Datos, para conocer la influencia de los factores sociales, económicos contenidos en el RUDE sobre el rendimiento académico de los estudiantes de secundaria del distrito dos de la ciudad de El Alto, en ella nos muestra la obtención de estos patrones, como la reducción de tiempo como el uso de metodologías CRISP-DM Minería de Datos, como herramientas muy potentes para el análisis en grandes cantidades almacenadas en el RUDE.

1.3. PLANTEAMIENTO DEL PROBLEMA

1.3.1. Problema Principal

En los últimos años la base de datos de la Universidad Pública de El Alto ha crecido de manera significativa en todos sus ámbitos. Los aspectos socioeconómicos, dan cuenta de ello en los registros iniciales de los universitarios al inicio de su carrera académica.

Sin embargo, estos factores socioeconómicos no reflejan índices de deserción académica a través de herramientas tecnológicas, menos sobre modelos elaborados para tal efecto. Las bases de datos del SIE[1] y unidades descentralizadas, hacen uso de gestores de bases de datos, no muestran proyecciones explicitas sobre los índices de deserción estudiantil en la U.P.E.A.

1.3.2. Problemas Secundario

- Falta de un modelo de predicción en base a la Minería de Datos.
- No se cuenta con la aplicación de algoritmos de Minería de Datos para la deserción de alumnos en la UPEA.
- Ausencia de un modelo sobre índices de deserción.

[1] SIE, Sistema de Información y Estadística

1.3.3. Formulación del problema

Si bien la Minería de Datos es eficiente en procesar y analizar grandes bases de datos, en qué medida es extensible al análisis de índices de deserción de alumnos. Teniendo en cuenta lo mencionado anteriormente surge la pregunta:

¿Cuáles son los factores que influyen en el índice de deserción de alumnos de la Universidad Pública de El Alto?

1.4. OBJETIVOS

1.4.1. General

Desarrollar un modelo de predicción en base a la Minería de Datos y los factores socioeconómicos, sobre el índice de deserción de alumnos de la Universidad Pública de El Alto.

1.4.2. Específicos

- Analizar los algoritmos de minería de datos que sean útiles para el modelo de predicción.
- Plantear un modelo de predicción en base a la Minería de Datos.
- Implementar un prototipo en base a algoritmos de Minería de Datos.
- Identificar los factores de deserción en base a la aplicación de algoritmos de Minería de Datos.

1.5. HIPÓTESIS

Debido a la aplicación de técnicas de Minería de Datos y la ingeniería de Software se tiene el modelo predictivo del índice de deserción en base a factores del alumno, teniendo una eficacia del 90% en la población estudiantil de la Universidad Pública de El Alto.

1.5.1. Identificación de Variables

Variable Independiente: Modelo Predictivo.

Variable Dependiente: Deserción Universitaria.

1.5.2. Operacionalización de Variables

Tabla 1.1
Operacionalización de Variables

VARIABLE	TIPO DE VARIABLE	DIMENSIÓN	INDICADORES
Modelo Predictivo.	Variable Independiente	Conceptos Reglas Patrones Comprensión	Numero de conceptos. Numero de reglas. Numero de patrones. Índice de pruebas del sistema.
Deserción Universitaria.	Variable Dependiente	Aspectos Socioeconómicos	Tipo de unidad educativa. Área de la unidad educativa. Actividad laboral del alumno. Cantidad de hermanos Tipo de vivienda. Característica de vivienda.

Nota: Elaboración propia.

1.5.3. Conceptualización de Variables

Tabla 1.2
Conceptualización de Variables

Variable	Definición Conceptual
Modelo Predictivo.	Un modelo es la representación simplificada de la realidad por medio de un conjunto de hipótesis, las cuales son usados para la explicación de algunos patrones de comportamiento que se observa en el mundo real (Carrillo & Gimenez, 2014).
Deserción Universitaria.	"la deserción es un estado en el que el estudiante se enfrenta a una situación en la que no logra concluir su proyecto educativo" (Quintero,2016). Según Himmel (2002) define a la deserción como el "abandono prematuro de un programa de estudios antes de alcanzar el título o grado".

Nota: Elaboración propia.

1.6. JUSTIFICACIÓN

Las justificaciones son desarrolladas de acuerdo a cuatro aspectos técnica, económica, social y científica.

1.6.1. Científica

La Minería de Datos es una herramienta que en los últimos tiempos se ha vuelto muy importante para la toma de decisiones. El modelo de predicción en base a la Minería de Datos de índices de deserción de estudiantes es importante

ya que aporta conocimiento al área de minería de datos y la toma de decisiones ya que la MD[2] permite crear modelos de acuerdo a las necesidades planteadas.

1.6.2. Técnica

Actualmente la universidad no cuenta con datos sobre la deserción de alumnos en consecuencia no tiene proyección de los mismos. La construcción del modelo de predicción en base a la Minería de Datos de índices de deserción de estudiantes de la Universidad Pública de El Alto será útil, para conocer la cantidad de alumnos que abandonan y su respectiva proyección, por lo que ayudará a la toma de decisiones al respecto de estos datos, también servirá como base para futuras investigaciones destinas a reducir la deserción de alumnos.

1.6.3. Económica

La aplicación del modelo significa un gran ahorro de dinero al utilizar software libre, además de que con los datos obtenidos por el modelo podrán ayudar a planificar un mejor uso de los recursos de la universidad.

1.6.4. Social

El modelo en base a la minera de datos de índices de deserción será importante para el desarrollo de planificación de apoyo a los estudiantes para que se reduzcan estos índices y así la universidad pueda aportar con más profesionales, beneficiándose tanto los alumnos como a la sociedad misma.

1.7. METODOLOGÍA

Para la presente investigación se utilizará un método correlacional y un enfoque cuantitativo. Para lo cual se divide en dos partes, la primera enfocada al desarrollo de la investigación y la segunda al desarrollo del modelo.

[2] MD, Minería de Datos

1.7.1. Método investigación científica

El trabajo de investigación a desarrollar utiliza el método científico bajo el enfoque cuantitativo establecidos por Hernández (2006) del siguiente modo:

- **Concebir la idea a investigar,** la investigación responde a una observación que se le da a una base de datos, de la cual se delimitaran en espacio y tiempo, identificando los beneficios o perjuicios que provoca la relación entre estos datos y el alumno para identificar el problema de índice de deserción.
- **Planteamiento del problema,** el problema de la investigación delimitado muestra la relación entre los objetivos que se espera alcanzar y la justificación de la investigación.
- **Elaboración del marco teórico,** es la parte teórica textual o de referencia del tema a investigar, muestra la recopilación y extracción de información para construir el marco teórico del trabajo.
- **Definición de la investigación,** se define el tipo de investigación más apropiada para el desarrollo del tema, en el que se describe una investigación exploratoria, descriptiva, correlacional o explicativa.
- **Establecimiento de hipótesis,** la formulación de Hipótesis está estrechamente relacionada con el planteamiento del problema, el marco teórico y el tipo de investigación. los mismos son definidos antes de la recolección de datos.
- **Selección de la muestra,** se fijan una muestra y procedimiento de selección, para poder definir los objetos que van a ser medidos.
- **Recolección de los datos,** los datos a recolectar se enmarcan de acuerdo al contexto de la investigación, los instrumentos de medida, la codificación y archivo de los datos.
- **Análisis de los datos,** la definición de las técnicas para el análisis de los datos recolectados, dependen de la hipótesis formulada y los niveles de medición de otras variables.

- **Elaboración del reporte de investigación,** se elabora el reporte final a presentar, basándose en los resultados de la investigación

1.7.2. Método de ingeniería

1.7.2.1. Ingeniería de software

La ingeniería de software busca el desarrollo del software con calidad. La ingeniería de software es una tecnología de varias capas con enfoque sistemático, disciplinado y de compromiso organizacional con calidad (Pressman, 2010).

Figura 1.1 Capas de la Ingeniería de Software.

Fuente: Choque (2002).

La ingeniera de software es una tecnología multicapa. Los cimientos que son la base de ingeniería de software están orientados hacia la calidad. La gestión de calidad total y las filosofías similares fomentan una cultura continua de mejora de proceso, y es esta cultura que conduce últimamente al desarrollo de enfoques cada vez más robustos para la ingeniería de software (Choque, 2002).

1.7.2.2. Metodología CRISP - DM

La metodología seleccionada fue CRISP - DM[3] , por medio de comparaciones conceptuales, aplicaciones y viendo la diferencia entre metodologías existentes para la Minería de Datos. La metodología CRISP-DM, es una metodología relativamente joven nace después de que la Minería de Datos se forje como disciplina importante dentro el análisis de la información.

Consta de cuatro niveles de abstracción organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. Cada una de ellas a su vez se divide en varias tareas (Figura 1.1), las flechas muestran las relaciones más habituales entre las etapas, aunque se debe aclarar que pueden establecer relaciones entre cualquiera de las fases. El círculo exterior ilustra la naturaleza cíclica del proceso de modelado.

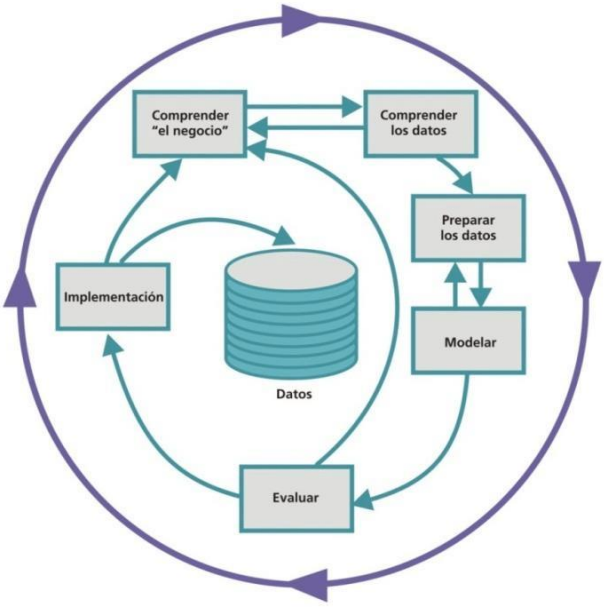


Figura 1.2 Fases del proceso de modelado metodología CRISP-DM.

Fuente: CRISP-DM (2015).

[3] CRISP-DM, Cross Industry Standard for Data Mining

1.7.2.3. Metodología de desarrollo RUP

Es una metodología cuyo fin es entregar un producto de software de alta calidad que satisfaga las necesidades de sus usuarios finales dentro de un presupuesto y tiempos predecibles. Junto a el Lenguaje Unificado de Modelado UML, constituyen la metodología estándar más utilizada para el análisis, implantación y documentación de sistemas orientado a objetos.

Esta metodología permite que todos los integrantes de un equipo de trabajo, conozcan y compartan el proceso de desarrollo, una base de cenicientos y los distintos modelos de cómo desarrollar el software utilizando un lenguaje de modelado en común.

1.7.2.4. Métrica de calidad

La ISO 9126 es un estándar internacional para evaluar la calidad de los atributos relacionados al conjunto de las funciones y propiedades específicas, donde estas funciones cumplen con las necesidades específicas como aquellos aspectos relacionados con la capacidad del software para mantener su usabilidad en función del tiempo son criterios sobre los cuales están estructurados las métricas de calidad ISO/IEC 9126, que nos permite especificar y evaluar la calidad del software desde diferentes criterios.

1.7.2.5. Evaluación de costos

El Modelo Constructivo de Costes COCOMO II, permite realizar estimaciones en función del tamaño del software y de un conjunto de factores de costos y de escala. Con el cual se estimará el esfuerzo de desarrollo, el tiempo de desarrollo y la cantidad de personas que se requieren para la realización del proyecto.

1.8. HERRAMIENTAS

En la actualidad se tiene varias herramientas para realizar la Minería de Datos. A continuación, se listan las herramientas que se utilizaran para el desarrollo e implementación de la presente propuesta:

1.8.1. Minería de Datos

- **Weka** [4](Entorno para Análisis del Conocimiento de la Universidad de Waikato), es una plataforma de software para aprendizaje automático y Minería de Datos escrito en Java. Weka es un software libre distribuido bajo licencia GNU-GPL, servirá para realizar la Minería de Datos.

1.8.2. Sistema operativo

- **Windows 10,** es una edición súper completa diseñado para toda la familia de los productos Microsoft, es un sistema operativo propietario que funciona bajo la arquitectura de x86 bits. y x64 bits. Es uno de los sistemas más comercial en el mundo. Su modelo es de desarrollo privativo, Share Source; tipo de núcleo monolítico, Uno de los aspectos más importantes de Windows 10 es el enfoque en la armonización de experiencias de usuario y funcionalidad entre diferentes tipos de dispositivos.

1.8.3. Base de datos

- **PostgreSQL,** es un servidor de Bases de Datos relacionales Orientadas a Objetos, de software libre bajo licencia BSD. Se opta por postgreSQL es un magnifico gestor de bases de datos. Tiene prácticamente todo lo que tienen los gestores comerciales, haciendo de él una muy buena alternativa GPL.

1.8.4. Lenguaje de programación

- **Java,** es un lenguaje de programación y una plataforma informática comercializada por primera vez en 1995 por Sun Microsystems. Hay muchas aplicaciones y sitios web que no funcionarán a menos que tenga Java instalado y cada día se crean más. Java es rápido, seguro

[4] WEKA, por sus siglas en ingles Waikato Environment for Knowledge Analysis

y fiable. Desde portátiles hasta centros de datos, desde consolas para juegos hasta súper computadoras, desde teléfonos móviles hasta Internet, Java está en todas partes.

1.8.5. Herramienta IDE

- **NetBeans,** es un programa que sirve como IDE (un entorno de desarrollo integrado) que nos permite programar en diversos lenguajes. ofrece herramientas de primera clase para el desarrollo de aplicaciones web, corporativas, de escritorio y móviles con Java. Siempre es el primer IDE en ofrecer soporte para las últimas versiones de JDK, Java EE y JavaFX. Proporciona descripciones generales inteligentes para ayudarle a comprender y gestionar sus aplicaciones, lo que incluye el soporte inmediato para tecnologías populares, como Maven.

NetBeans contiene tecnologías innovadoras listas para usar y es el estándar en el desarrollo de aplicaciones, gracias a sus características integrales para el desarrollo de aplicaciones, las constantes mejoras en el editor de Java y el perfeccionamiento del rendimiento y la velocidad.

1.8.6. Herramienta case

- **MagicDraw UML**, es una herramienta CASE desarrollada por No Magic. La herramienta es compatible con el estándar UML 2.3, desarrollo de código para diversos lenguajes de programación (Java, C++ y C#, entre otros) así como para modelar datos.

1.9. LÍMITES Y ALCANCES

1.9.1. Limites

- El actual trabajo de investigación no comprende la construcción del Data Warehouse.
- Se limita a los datos de la base de datos del registro de la universidad.

- El actual trabajo de investigación no es el producto de un análisis estadístico.

1.9.2. Alcances

- EL Presente trabajo de investigación contempla el descubrimiento de conocimiento en la base de datos Registro de Alumnos de la Universidad Pública de El Alto.
- Diseñar y aplicar un modelo de Minería de Datos, aplicado al índice de deserción de universitaria.

1.10. APORTES

Los aportes que ofrecerá el modelo serán los siguientes:

- ✓ La implementación del modelo nos ayuda a conocer los índices de deserción de los alumnos.
- ✓ Realizar una proyección sobre el comportamiento de los alumnos en términos de deserción.

CAPITULO II

MARCO TEÓRICO

MARCO TEÓRICO

En este capítulo se representa la teoría relacionada con la Minería de Datos, su definición, objetivos, modelos, métodos y técnicas, también se describe la relación con el descubriendo de conocimiento en base de datos. Además, se describen las herramientas de desarrollo para la propuesta de la solución, descubriendo sus componentes, estructura, aprendizaje, ventajas y aplicaciones.

2.1. INTRODUCCIÓN

El almacenamiento de datos a lo largo de la historia se a vuelo algo necesario, pero al mismo tiempo complicado, con el avance de la tecnología y la posibilidad de almacenamiento de mayor información en bases de datos ha vuelto a es esta tarea algo rutinario, más aún en aquellas entidades u organizaciones que aplicaron sistemas de información para dicha tarea. La enorme cantidad de datos que se almacenan en estas bases de datos no tienen la capacidad de soportar el análisis y toma de decisiones mediante la aplicación del lenguaje estructurado de consultas (SQL), entonces la información queda oculta en espera de nuevas técnicas que pueda transformar los datos en información y en base a estos se genere nuevos conocimientos.

La Minería de Datos y su búsqueda de conocimiento se aplica a una gran cantidad de problemas en diferentes áreas relacionadas con la predicción, caracterización o clasificación de datos. Daza (2004), menciona que la Minería de Datos ha sido reconocida como un tópico importante en la investigación de

base de datos, inteligencia artificial, redes neuronales, estadística, reconocimiento de patrones, sistema basados en conocimiento, recuperación de información y visualizaron de datos.

2.2. MINERÍA DE DATOS

La digitalización de la información en instituciones se ha vuelto algo muy común y necesario, lo que conlleva a que las bases de datos crezcan de manera significativa. Estas bases de datos contienen todo tipo de información y más, algunas que se desconocen pero que se pueden aprovechar para generar conocimiento útil y tomar decisiones en base a estos.

La Minería de Datos busca el procesamiento de información de forma que para el usuario o cliente sea más entendible, clasificando información a partir de parámetros establecidos y de acuerdo a las necesidades del caso, por medio de la Minería de Datos se busca que la persona pueda comprender los resultados de una manera entendible (Vizcaino, 2008).

Según Landa (2016) hacer una un enfoque muy interesante al afirmar que la Minería de Datos es el núcleo de todo un proceso metodológico para encontrar un modelo valido, útil y entendible que describe patrones de acuerdo a la información, también aclara que un modelo es la representación que intenta explicar ese patrón en los datos.

2.2.1. Historia

El concepto de Minería de Datos o Data Mining no es un nuevo si nomas bien un concepto de moda que muchas veces se confunde, términos como Data Fishing, Data Mining o Data Archaeology se empiezan a usar ya en los años 60 en el área de estadística, pero no es hasta los años 90 que se consolida como tal.

La idea de la Minería de Datos en sus inicios era la de encontrar correlaciones sin tomar en cuanta aun lo que era una base de datos. Ya en los años 80, Rakesh

Agrawal y otros empezaron a consolidar el término de Minería de Datos y el Descubrimiento de Conocimiento en Bases de Datos.

Las evoluciones de sus herramientas en el transcurso del tiempo pueden dividirse en cuatro etapas principales:

- Colección de Datos (1960).
- Acceso de Datos (1980).
- Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990).
- Minería de Datos Inteligente (1990).

2.2.2. Definición de Minería de Datos

La Minería de Datos es el proceso trivial de extraer conocimiento humanamente entendible y útil, mediante técnicas y herramientas, con el objetivo de predecir de forma automatizada y generar modelos (Piatestsky-Shapiro y Frawley,1991 citado por Rodríguez Suárez & Díaz Amador, 2009).

La Minería de Datos consiste en el la exploración y el análisis de grandes bases de datos, mediante métodos ya sean automáticos o semiautomáticos, en busca de patrones útiles (Berry & Linoff, 2004).

Según (Gonzáles, 2006), la Minería de Datos es el proceso de encontrar y extraer conocimiento útil y comprensible en grandes volúmenes de bases de datos previamente desconocidos.

La Minería de Datos es la extracción de información desconocida pero potencialmente útil de base de datos, refiriéndose al volumen de los datos almacenados en los equipos de computación que almacenan enormes cantidades de datos (Witten & Frank, 2005).

Para Thuraisingham (1999), la Minería de Datos consiste en el análisis de series de datos en busca de relaciones inesperadas y resumir la información de manera que sea útiles y entendible para el propietario de los datos.

La Minería de Datos combina técnicas de estadística, inteligencia artificial y bases de datos entre otras ramas, de manera que se puedan obtener modelos o patrones de forma automática o semiautomática en base a los datos almacenados en las bases de datos (Siebes, 2020).

La Minería de Datos es la búsqueda de patrones de interés mediante diferentes algoritmos entre las cuales están los arboles de clasificación, regresión, clusterización, dependencias entre otras más (Wang, 1999).

Por su parte Rodríguez Suárez y Díaz Amador (2009), sostienen que la Minería de Datos es una tecnología de apoyo que tiene por objeto el explorar, comprender y buscar patrones, relaciones o excepciones en grandes bases de datos para luego aplicar en conocimiento adquirido.

Fayyad, Pieatetsky-Shapiro y Smyth (1996), concluyen que la Minería de Datos es un paso en el KDD, que implica en la aplicación de algoritmos destinados a extraer patrones de datos, ya que esta depende de otros pasos y que sin ellos puede ser una actividad peligrosa.

La Minería de Datos consiste en la extracción de información de manera automática o semiautomáticas de grandes bases de datos, pero es un paso importante dentro del descubrimiento de conocimiento en bases de datos (Hidalgo, 2014).

“La integración de un conjunto de áreas que tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión” (Molina y otros, 2001).

Por su parte Pautsch (2009) afirma que la “Minería de Datos permite extraer la información oculta, descubriendo patrones y relaciones entre los datos y así crear de modelos”, pero es el KDD el encargado de la preparación de los datos y la

interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones hallados.

La Minería de Datos consiste en la aplicación de algoritmos en grandes volúmenes de datos mediante técnicas y herramientas en busca de nuevos patrones o excepciones que sean entendibles y útiles para los humanos, de manera que estos puedan ser utilizados en la toma de decisiones.

La Minería de Datos no es un software sino más bien un proceso metodológico el cual busca dentro de las bases de datos el conocimiento nuevo, valido, útil y entendible para el ser humano.

2.2.3. Modelos de Minería de Datos

Dentro la Minería de Datos existe varios modelos que varían dependiendo de la aplicación que se le quiera dar, para encontrar el conocimiento que se encuentran en las grandes bases de datos.

Según Agrawal y Shafer (1996), los modelos que pueden ser descriptivos o predictivos:

- **Descriptivos o No supervisados**: este modelo aspira a descubrir patrones y tendencias sobre el conjunto de datos sin tener ningún tipo de conocimiento previo de la situación a la cual se quiere llegar. Descubre patrones en los datos analizados y proporciona información sobre las relaciones de los mismos.
- **Predictivos o Supervisados:** crean un modelo de una situación donde las respuestas son conocidas y luego, lo aplica en otra situación de la cual se desconoce la respuesta. Conociendo y analizando un conjunto de datos, intentan predecir el valor de un atributo (etiqueta), estableciendo relaciones entre ellos.

Los OLAP verifican patrones hipotéticos y la Minería de Datos usa los datos para descubrir dichos patrones. Aguilar (2003), afirma que la Minería de Datos es un proceso y como resultado es el conocimiento, y que no se lo debe confundir con

los sistemas OLAP[5] y cuadros de mandos ya que estos solo son herramientas que ayudan a la dirección y gestión de las empresas.

Un patrón es un suceso que se repite en una base de datos es así que Aruquipa (2015), afirma los siguiente:

> Un patrón, es algo que se repite, una tendencia, como una representación de los datos e información obtenidos de una fuente de información, como puede ser una base de datos. Un patrón, ha de cumplir una serie de características para que nos resulte de utilidad a la hora de trabajar con él y obtener información de utilidad. (p.19)

La Minería de Datos dispone de diferentes métodos y algoritmos para ser aplicados en grandes bases de datos y así poder hallar nuevos patrones o tendencias significativamente útiles capaces de generar nuevo conocimiento (Molina & Garcia, 2006).

2.2.3.1. Modelos Descriptivos

Este tipo de me modelo busca encontrar y describir la relación entre los datos disponibles. Buscan información actual para obtener beneficios de ellas, este tipo de modelo consta de varios algoritmos dependiendo de su aplicación y se pueden agrupar en:

a. Clustering

Un clúster es un conjunto de datos con características similares. Esta similitud puede medirse con funciones de distancia. La Minería de Datos intenta encontrar clústeres de buena calidad para que luego puedan ser utilizados en grandes bases de datos.

Existen una variedad de algoritmos usados en el clustering, entre los cuales citaremos algunos:

[5] OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line-Analytical-Processing)

- *Algoritmo K-medias,* trata de encontrar los puntos k más densos en un conjunto de puntos arbitrarios. Primero divide en conjuntos y luego calcula la media, reagrupa los puntos de acuerdo al resultado obtenido hasta que no varíen.

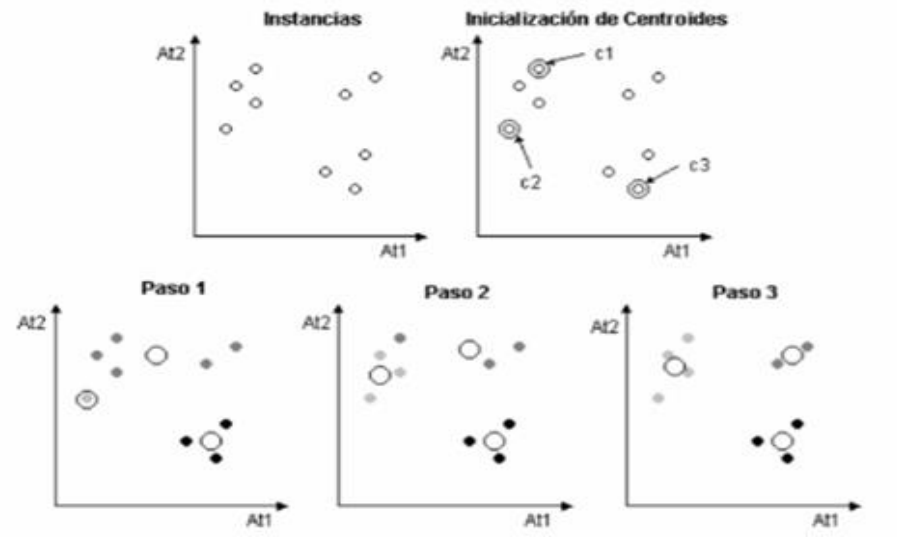


Figura 2.1 **Ejemplo de clustering con k-medias.**

Fuente: Molina & Garcia (2006).

- *Clustering Conceptual (COBWEB), e*ste algoritmo surge a partir de que el algoritmo de k-medias no puede con datos que no son numéricos, es decir datos conceptuales. Este algoritmo fue presentado por Michaslki para justificar la necesidad de clustering cualitativo frente al clustering cuantitativo. A semejanza de los humanos, COBWEB forma los conceptos por agrupación de ejemplos con atributos similares.

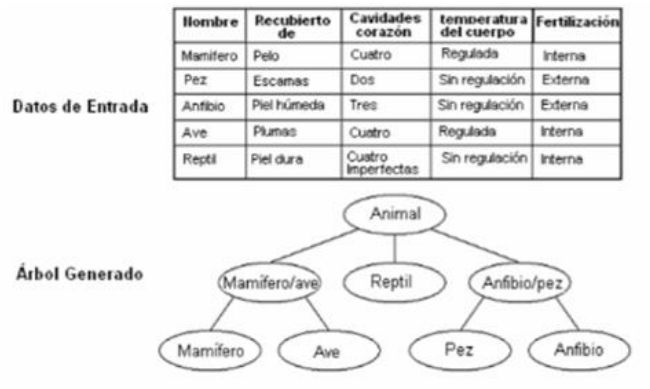


Figura 2.2 **Ejemplo de árbol generado por COBWEB.**

Fuente: Molina & Garcia (2006).

b. Asociación

Este tipo de modelos es empleado para establecer posibles relaciones entre las acciones aparentemente independiente, buscan describir sucesos llevados a partir de información disponible. Este modelo basa sus fundamentos en técnicas estadísticas como los análisis de correlación y variación. Como otros modelos este dispone de varios algoritmos entre los cuales están:

- *Algoritmo A Priori, b*usca generar reglas de asociación en base a procedimientos de reglas de covering, este algoritmo se aplica a generalmente a transacciones comerciales y en problemas de predicción. Asocia los datos frecuentes y los asocia en conjuntos, para generar las reglas combina cada posible combinación entre pares de los atributos que serán llamados ítem y luego se le asigna un nivel de confianza. Las reglas que interesan son aquellas que tiene un nivel de confianza alto, las otras con nivel bajo de confianza serán descartadas.

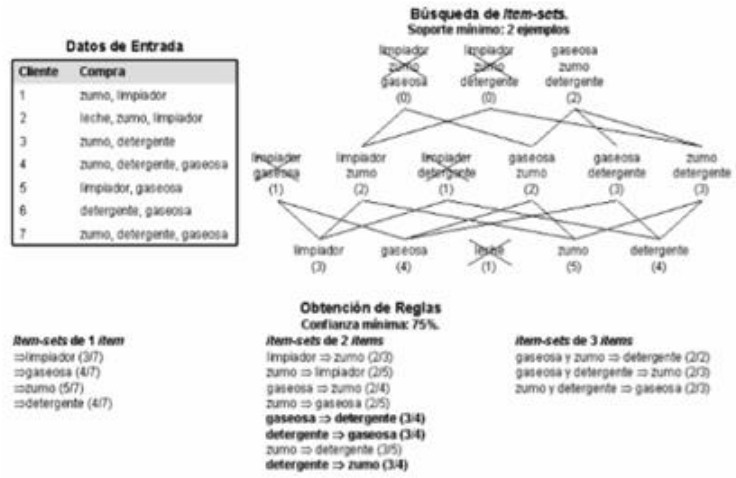


Figura 2.3 **Ejemplo de obtención de reglas de Asociación A Priori.**

Fuente: Molina & Garcia (2006).

2.2.3.2. Modelos Predictivos

Los modelos predictivos buscan predecir atributos de un conjunto de datos a partir de otros estableciendo relaciones entre sí. Estos modelos se pueden agrupar en:

a. Regresión

Dentro de la regresión existen dos tipos, la regresión lineal y no lineal, estos métodos son uno de los más comunes para estudio de la correlación de datos. Una de las ventajas que tienen estos métodos es que a partir del conocimiento de las sus ecuaciones se puede obtener conocimiento cualitativo (Wang, 1999).

Es similar a los algoritmos de clasificación, nos permiten proyectar métricas tanto a futuro como al pasado. El modelo generado intenta a partir de los datos

obtenidos predecir el valor más probable para una situación observada. Dentro de la regresión existen varios entre los cuales podemos mencionar algunos:

- *Regresión lineal simple, e*s la forma más básica de la regresión, en cual se modelan los datos mediante el uso de una recta. Utiliza dos variables aleatorias (x, y), y también llamada variable de respuesta y x llamada variable predictoria.

Sus principales ecuaciones son:

$$y = a + bx$$

$$b = \frac{S_{xy}}{S_x^2}$$

$$a = y - bx$$

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

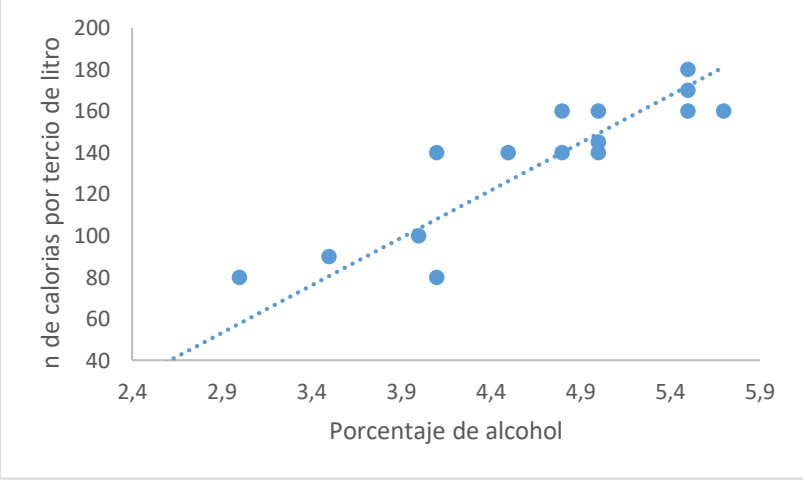


Figura 2.4 **Regresión lineal simple.**

Fuente: Molina & Garcia (2006).

- *Regresión lineal múltiple, b*usca los coeficientes de una ecuación lineal mediante las siguientes ecuaciones:

$$f(\mathrm{x}) = bo + b1x1 + \cdots + b_{\mathrm{n}}\mathrm{x}_{\mathrm{n}}$$

si es lineal simple en dos dimensiones (x,y) seria:

$$\mathrm{b}_1 = \frac{n(\Sigma \mathrm{xy})(\Sigma x)(\Sigma y)}{\mathrm{n}(\Sigma x^2) - (\Sigma x)^2}$$

$$\mathrm{b}_0 = \frac{(\Sigma \mathrm{y})(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{\mathrm{n}(\Sigma x^2) - (\Sigma x)^2}$$

De donde resulta:

$$Y = b_0 + b_1 X$$

- *Regresión no lineal, e*xisten casos en los cuales los datos no son dependientemente lineales, esto sucede en ecuaciones polinómicas en donde para resolver se aplican ciertas transformaciones al modelo no lineal y así este se vuelva en uno lineal, pudiéndose resolver por el método de mínimos cuadrados. Cabe recalcar que no siempre se puede realizar ya que existen algunos casos son especialmente no lineales.

Ecuación original:

$$y = \mathrm{a} + b_1 x + b_2 x^2 + b_3 \mathrm{x}^3$$

cambio de variables:

$$x_1 = x \qquad x_2 = x^2 \qquad x_3 = x^3$$

ecuación final:

$$y = \mathrm{a} + b_1 x_1 + b_2 x_2 + b_3 \mathrm{x}_3$$

b. Clasificación

El modelo de clasificación también pertenece al aprendizaje automático supervisado porque dicho modelo siempre requiere ser entrenado con patrones conocidos, para que a partir de dicho entrenamiento se pueda predecir nuevos patrones, se basan en las redes neuronales (Wang, 1999).

Tabla 2.1
Ejemplo de estructura de dato

Instances	**Attributes**					
	1	2	… …	j	… …	m
X_1	X_{11}	X_{12}	… …	X_{1j}	… …	X_{1m}
X_2	X_{21}	X_{22}	… …	X_{2j}	… …	X_{2m}
.						
X_j	X_{j1}	X_{j2}	… …	X_{jj}	… …	X_{jm}
.						
X_n	X_{n1}	X_{n2}	… …	X_{nj}	… ….	X_{nm}

Nota: Tomado de Data Mining and Knowledge Discovery for Process Monitoring and Control. *Wang (1999).*

El modelo de clasificación analiza un conjunto de datos clasificados que se conocen para que partir de este el modelo pueda realizar clasificaciones futuras, este tipo de modelos son utilizados en la detección de fraudes, análisis de riesgos de créditos y otros (Pautsch, 2009).

La clasificación consiste en clasificar los datos en grupos que estén lo más cerca posible y alejado de los otros grupos, mediante patrones que los asocien en dichos conjuntos. Entre algunos algoritmos podemos mencionar a:

➢ ***Arboles de decisión,*** entre algunos del algoritmo de árboles de decisión están:

- *Arboles de decisión simple,* Los arboles de decisión tiene el enfoque de “divides y vencerás”, por lo general consiste en comparar en un nodo un atributo con una constante, pero hay casos en que se compran dos atributos con cada uno. Para clasificar la instancia desconocida se en ruta hacia abajo, esto da a la formación de la forma de un árbol (Witten & Frank, 2005).

  Los árboles de decisión son reglas compactas para ser representadas en forma de árbol, es un aprendizaje supervisados o entrada que parte de un nodo u hoja y se ramificando en forma de árbol.

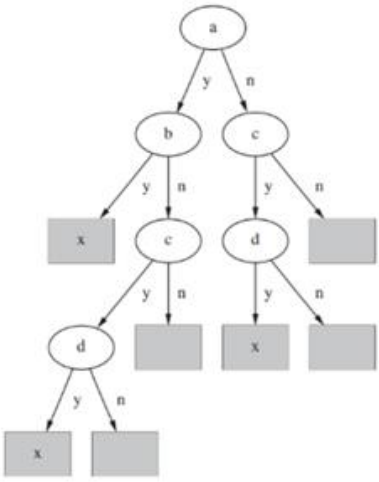


Figura 2.5 **Árbol de decisión para una simple disyunción.**

Fuente: Witten & Frank (2005).

- *El sistema ID3,* Un algoritmo simple pero potente, cuyo propósito es generar un árbol de decisión. El proceso consiste en seleccionar un atributo y a partir de este generar con cada uno de los posibles valores una rama y repetir el proceso con cada nodo resultante, es así hasta que se agoten los ejemplo.

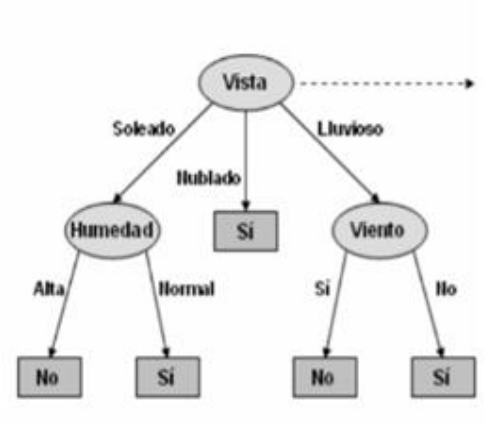


Figura 2.6 **Ejemplo de clasificación con ID3.**

Fuente: Molina & Garcia (2006).

- *Sistema C4.5,* Molina & García (2006), describe al algoritmo ID3 como capaz de tratar valores continuos, tendrá tantas como valores posibles atributos. Por ello, Quinlan propuso el C4.5, como extensión del ID3, que permite:

    1. Empleo del concepto razón de ganancia (GR)
    2. Construir árboles de decisión cuando algunos de los ejemplos presentan valores desconocidos para algunos de los atributos.
    3. Trabajar con atributos que presenten valores continuos.
    4. La poda de los árboles de decisión.
    5. Obtención de Reglas de Clasificación.

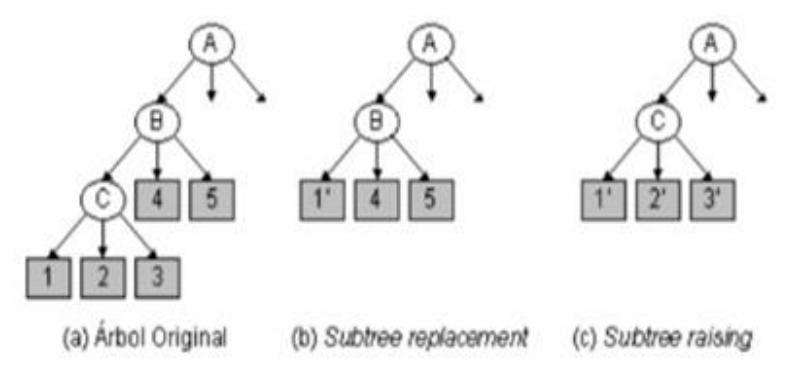


Figura 2.7 **Tipos de operaciones de poda en C4.5.**

Fuente: Molina & Garcia (2006).

- *Arboles de predicción,* los arboles de predicción numérica son similares a los arboles de decisión, pero estos en vez de usar la entropía para definir un atributo estos utilizan la varianza del error en cada hoja. Una vez construido el árbol completo se poda realiza la mejora de este, obteniendo una constante en cada nodo.

$$\mathrm{p}' = \frac{\dot{\mathrm{n}}\mathrm{p} + kq}{\mathrm{n} + \mathrm{k}}$$

En esta ecuación, p es la predicción que llega al nodo (desde abajo), p' es la predicción filtrada hacia el nivel superior, q el valor obtenido por el modelo lineal de este nodo, n es el número de ejemplos que alcanzan el nodo inferior y k el factor de suavizado. Para construir el árbol se emplea como heurística el minimizar la variación interna de los valores de la clase dentro de cada subconjunto. Se trata de seleccionar aquel atributo que maximice la reducción de la desviación estándar de error.

$$\mathrm{SD} = \mathrm{SD}(\mathrm{E}) - \sum_{\mathrm{i}} \frac{|\mathrm{E}|}{|\mathrm{E}|} \mathrm{SD}(\mathrm{E_i})$$

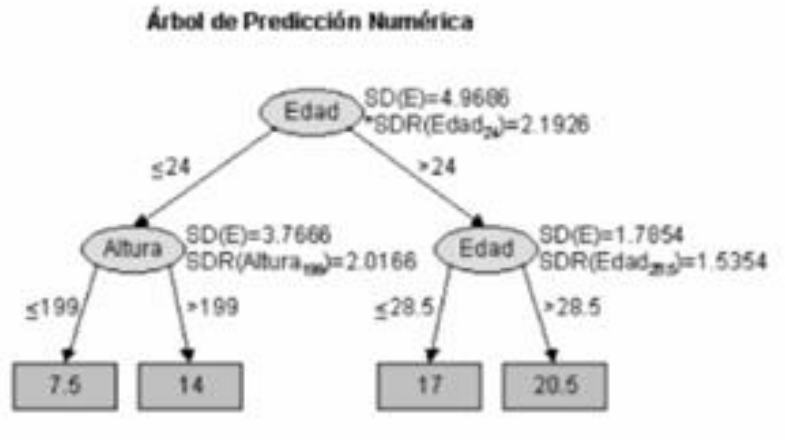


Figura 2.8 **Ejemplo de generación de árbol de predicción.**

Fuente: Molina & Garcia (2006).

- ***Reglas de clasificación,*** son una alternativa a los árboles de decisión, son de tipo inductivas que parten de los datos de entrada generando arboles de decisión y se extrayendo un conjunto de reglas para luego podar y volver a buscar nuevas reglas, la otra forma seria el de covering que consta en cubrir todos los ejemplos de esa clase y cuando se obtiene una regla se eliminan todos los ejemplos que cubre y a continuación se busca más reglas hasta que no queden ejemplos. Podemos citar algunos algoritmos dentro de las reglas de clasificación:

- *Algoritmo ZeroR,* es el método de clasificación más simple, este se basa en el objetivo e ignora todos los predictores. El clasificador ZeroR predice simplemente la categoría mayoritaria (clase) (Vijayarani & Muthulakshmi, 2013). Aunque no hay poder de predictibilidad en ZeroR, es útil para determinar un desempeño como un punto de referencia para otros métodos de clasificación. Este Algoritmo es el más primitivo en Weka. Modela el conjunto de datos con una sola regla. Dado un nuevo elemento de datos para la clasificación, ZeroR siempre predice el valor de categoría más frecuente en los datos de

entrenamiento para problemas con un valor de clase nominal o el valor de clase promedio para problemas de predicción numérica. En algunos conjuntos de datos es posible que otros esquemas de aprendizaje induzcan a modelos que presentan peores resultados en los 42 nuevos datos que ZeroR, que es un claro indicador de sobre ejecución grave.

- *Algoritmo PART, u*no de los más importantes dentro del aprendizaje por reglas de clasificación dado por C4.5. Como las anteriores primero genera el árbol de decisiones y posteriormente aplica un proceso de refinamiento en busca de la optimización global. PART [6]recibe ese nombre por su forma de trabajo la obtención de reglas de los árboles de decisión PARCIAL, fue desarrollado por los mismos que crearon WEKA.

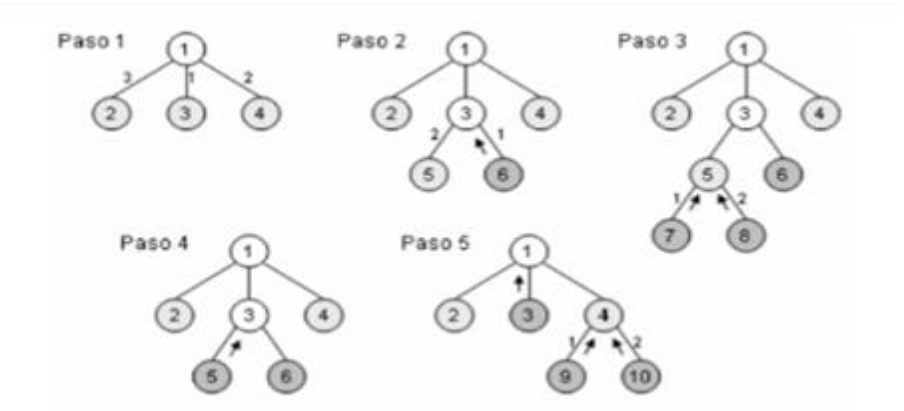


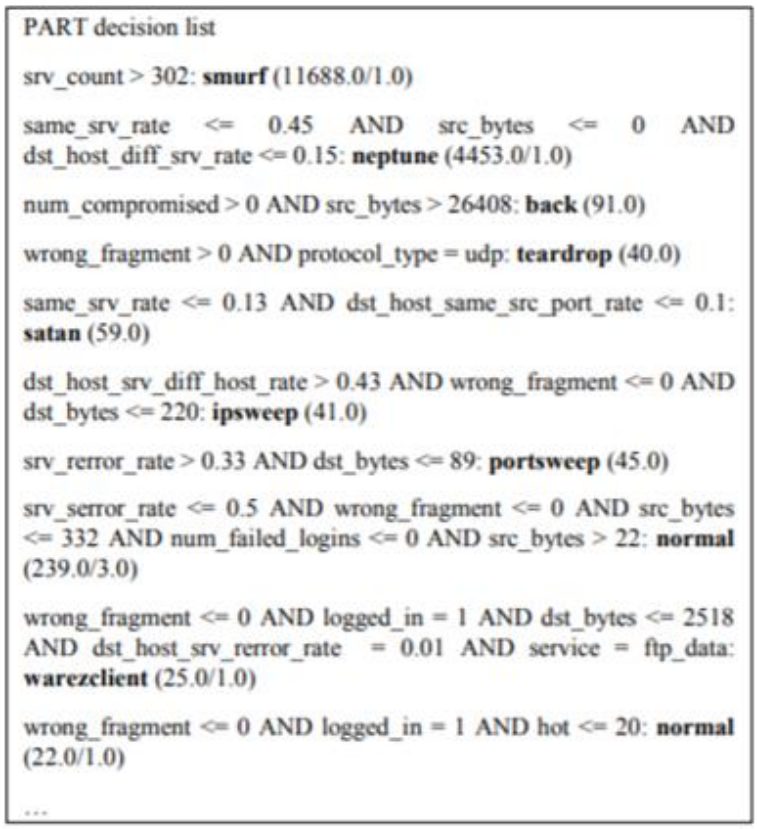
Figura 2.9 **Ejemplo de generación de árbol parcial con PART.**

Fuente: Molina & Garcia (2006).

Este algoritmo genera una lista de reglas de clasificación, es decir, una secuencia ordenada de reglas que, a la hora de aplicarse a un ejemplo, deben ejecutarse en el orden en que fueron generadas hasta encontrar

[6] PART obtaining rules from PARTial decision trees

la primera que cubra el ejemplo en cuestión. Dicha regla tendrá en su consecuente la clase a asignar. El resultado obtenido es una lista y no un conjunto porque luego de obtener cada regla se retira del conjunto


PART decision list

srv_count > 302: **smurf** (11688.0/1.0)

same_srv_rate <= 0.45 AND src_bytes <= 0 AND dst_host_diff_srv_rate <= 0.15: **neptune** (4453.0/1.0)

num_compromised > 0 AND src_bytes > 26408: **back** (91.0)

wrong_fragment > 0 AND protocol_type = udp: **teardrop** (40.0)

same_srv_rate <= 0.13 AND dst_host_same_src_port_rate <= 0.1: **satan** (59.0)

dst_host_srv_diff_host_rate > 0.43 AND wrong_fragment <= 0 AND dst_bytes <= 220: **ipsweep** (41.0)

srv_rerror_rate > 0.33 AND dst_bytes <= 89: **portsweep** (45.0)

srv_serror_rate <= 0.5 AND wrong_fragment <= 0 AND src_bytes <= 332 AND num_failed_logins <= 0 AND src_bytes > 22: **normal** (239.0/3.0)

wrong_fragment <= 0 AND logged_in = 1 AND dst_bytes <= 2518 AND dst_host_srv_rerror_rate = 0.01 AND service = ftp_data: **warezclient** (25.0/1.0)

wrong_fragment <= 0 AND logged_in = 1 AND hot <= 20: **normal** (22.0/1.0)

...

Figura 2.10 **Modelo PART**

Fuente: Minería de Datos: Intrusiones de Red. Antolin & Barcenilla (2008).

de ejemplos los correctamente cubiertos por dicha regla.

Es importante tener en cuenta la diferencia entre un conjunto de reglas independientes y una lista de clasificación. Si se trata de reglas independientes, podrán superponerse en el espacio de entrada permitiendo que un mismo ejemplo pertenezca a dos clases distintas. Cuando se trata de listas de clasificación, las reglas se analizan una por una comenzando por la que se generó primero hasta encontrar aquella que cubra el ejemplo. La lista incluye una clase por defecto, la misma

que es asignada al conjunto de ejemplos que no cumplen con ninguna regla. En otras palabras, al utilizar una lista de clasificación, un ejemplo dado sólo puede ser asignado a una clase. El método PART requiere de la definición previa de un árbol de clasificación parcialmente construido. Esto implica que, en cierto punto del proceso, el árbol deja de construirse y a partir de las ramas que terminan en hoja se decide la regla que se va a generar. Cuando la construcción se detiene, se agrega a la lista de reglas la rama con mayor cobertura y se repite el proceso hasta que la cantidad de ejemplos por cubrir se encuentre por debajo de un cierto umbral (o se hayan cubierto todos). La expansión del árbol se realiza en orden comenzando por el subconjunto de menor valor de entropía, repitiéndose recursivamente hasta que todos los subconjuntos expandidos sean hojas. Este proceso se puede observar en el Algoritmo luego que se ha finalizado la construcción del árbol parcial, se selecciona la rama con mayor cobertura, que finaliza en una hoja. Después de obtener una regla, los ejemplos correctamente cubiertos por ella son eliminados de los datos de entrada, y el árbol se descarta. Este proceso se realiza de forma iterativa hasta lograr llegar hasta la cobertura deseada.

- *Algoritmo RIDOR*[7]*,* genera primero una regla por defecto (predeterminada) y luego toma las excepciones para la regla predeterminada con la mínima tasa de error. Entonces genera la mejor excepción para cada excepción iterando hasta logras disminuir el error. Luego genera una expansión similar a un árbol de excepciones. La excepción es un conjunto de reglas que predice clases. Este algoritmo es usado para genera dichas excepciones.

[7] RIDOR por sus siglas en ingles Ripple Down Rule

- *Algoritmo KNN,* el método KNN pertenece al grupo de métodos para tareas de clasificación de datos que se pueden encontrar dentro de Minería de Datos, estos son fundamentalmente dependientes de la distancia y en consecuencia poseen características propias; como la cercanía, la lejanía y la magnitud de longitud, entre otras (Rodriguez, Rojas, & Franco, 2012). El objetivo de la clasificación es encontrar un modelo, para predecir la clase a la que pertenecería cada registro, esta asignación es una clase que se debe hacer con la mayor precisión posible. Por lo general, el conjunto de datos, se divide en dos conjuntos al azar, uno para entrenamiento y el otro de prueba, Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo.

2.2.4. Etapas de la Minería de Datos

Dentro de la Minería de Datos se debe tener en cuenta las siguientes etapas:

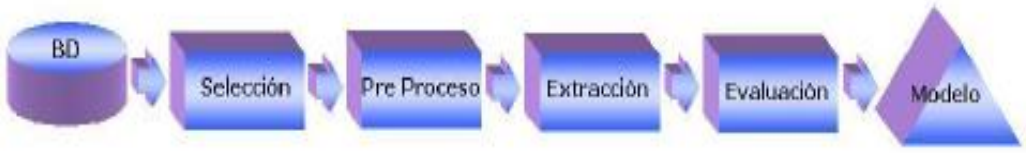


Figura 2.11 **Proceso de la Minería de Datos.**

Fuente: Hidalgo (2014).

2.2.4.1. Selección de Datos

Las selecciones de datos pueden tener un gran volumen y contener una cantidad ingente de datos. Según (García, 2005), en esta etapa se reduce considerablemente el volumen de los datos seleccionando solo los atributos y

tuplas que aporten la información y sea más influyentes sobre el tema a tratar. Existen varios métodos para la selección de este subconjunto de atributos. Entre algunos de ellos se pueden citar:

- *Selección por Pasos Hacia Adelante,* se comienza con un conjunto vacío de atributos, en cada paso se agrega al conjunto el mejor atributo del conjunto original.
- *Eliminación por Pasos Hacia Atrás,* se comienza con un conjunto que posee todos los atributos originales, en cada paso se elimina del conjunto el peor atributo.
- *Combinación de Selección por Pasos Hacia Adelante y Eliminación por Pasos Hacia Atrás,* es una combinación de los dos anteriores. Se puede utilizar un umbral de medición para establecer cuándo detener la eliminación y agregación de los atributos.
- *Inducción con árboles de decisión,* se utilizan algoritmos como ID3 y C4.5. Los atributos que no son representados en el árbol se consideran irrelevantes y se los descarta. Por el contrario, los atributos que aparecen en el árbol son los elegidos para conformar el subconjunto de atributos.

2.2.4.2. Pre Procesamiento de Datos

El formato de los datos de las distintas fuentes (OLPT, Fuentes Externas, etc.) por lo general no suele ser apropiado. Esto dificulta que los algoritmos de minería obtengan buenos modelos trabajando sobre estos datos en bruto.

El objetivo del pre procesado es adecuar los datos para que la aplicación a los algoritmos de minería sea óptima. Para esto hay que filtrar, eliminar datos incorrectos, no válidos, crear nuevos valores y categorías para los atributos e intentar completar o descartar los valores desconocidos e incompletos (García, 2005).

2.2.4.3. Extracción del Conocimiento

Es la aplicación de diferentes algoritmos sobre los datos ya pre procesados, para extraer patrones. Una vez que ya sean normalizado los datos comienza ahora la búsqueda y extracción de patrones o conocimiento mediante la aplicación de algunos de los algoritmos ya mencionados anteriormente, estos algoritmos serán elegidos de acuerdo a los objetivos que se plantea el problema.

2.2.4.4. Evaluación e Interpretación de Patrones

Una vez obtenidos los patrones se debe comprobar su validez. Si los modelos son varios, se debe elegir el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, se debe volver a las etapas anteriores y modificar alguna entrada para, de esta manera, generar nuevos modelos.

2.2.5. Aplicaciones de la Minería de Datos

Hoy en día la Minería de Datos como búsqueda de patrones tiene mucha acogida en las diferentes áreas es así que su aplicación se ha diversificado mucho, entre algunas de estas aplicaciones podemos citar:

- La utilización de árboles de decisión en la construcción de modelos de clasificación de diferentes características del desarrollo de software.
- Aspectos climatológicos, predicción de tormentas, clima, lluvias, etc.
- Medicina, encontrar la probabilidad de una respuesta satisfactoria de un tratamiento médico.
- Mercadotecnia, identificación de clientes susceptibles a responder a ofertas de productos y servicios por correo, fidelidad de clientes, afinidad de producto.
- Industria y manufactura, los fabricantes pueden predecir el desgaste de activos de producción y anticipar su mantenimiento, lo cual puede maximizar el tiempo en operación y mantener la línea de producción acorde a lo programado.

- Análisis de las canastas de mercado para mejorar la organización de tiendas, segmentación de mercado (Clustering). A través de modelos de datos más precisos, las compañías detallistas pueden ofrecer campañas más enfocadas y encontrar la oferta que tenga el mayor impacto en el cliente.
- Educación, la Minería de Datos ayuda a los educadores a acceder a datos de los estudiantes, a predecir niveles de logro y a detectar estudiantes o grupos de estudiantes que necesitan atención extra.
- Comunicaciones, las compañías de multimedia y telecomunicaciones pueden utilizar modelos analíticos para entender montañas de datos de clientes, ayudándoles así a predecir el comportamiento de sus clientes y ofrecer campañas altamente dirigidas y relevantes.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, tarjetas de crédito, evasión fiscal, electricidad.
- Determinación de niveles de audiencia de programas de televisión.

2.3. CONCEPTOS BÁSICOS

2.3.1. Dato

Los datos son hechos, medidas u observaciones que pueden presentarse, descritos al interior de un contexto. La validez y la efectividad de los datos vienen determinadas principalmente por su exactitud (Molina, 2002, citado por Hidalgo, 2014).

Dato es una descripción numérica o verbal sobre algo que no fue analizada o resumida, que pueden ser presentados en diversas formas como ser patrones o letras y son almacenados en una memoria electrónica o en hechos en la mente de una persona (Ferrel, Geoffey, & Ferrel, 2009).

Los datos son la materia prima con se cuenta ya sea en forma digital o tradicional, que después de será analizada, tratada tomara sentido y servirá para la generación de conocimiento.

2.3.2. Información

Chiavenato (2006), define la información como un conjunto de datos significativos que aumenta el conocimiento de algo. La información es un mensaje significativo que reduce la incertidumbre y puede ser utilizada inmediatamente para la orientación de las acciones y decisiones.

La información es la interpretación de datos y cocimiento con sentido que sirven para la toma de decisiones. El riesgo para una buena decisión dependerá de cuanta información se tiene, es decir a mayor información menor riesgo y además mayor información simplifica y mejora las decisiones futuras. “Entonces, la información entraña una interpretación de datos y conocimientos que tienen sentido y que sirven para tomar decisiones” (Ferrel, Geoffey, & Ferrel, 2009, p. 121).

La información es el tratamiento de los datos, son datos contextualizados, datos con sentido y con un mensaje claro y orientado hacia un determinado algo, aumentado el conocimiento y que ayuda a la reducción de la incertidumbre para la toma de decisiones.

2.3.3. Conocimiento

Según Ferrel, Geoffey y Ferrel (2009), definen al conocimiento como a la comprensión de datos mediante un estudio o mediante la experiencia. Por su parte Alavi y Leidner (2003) citado por (Flores, 2005) definen el conocimiento como la información que posee un individuo en su mente de manera personalizada y subjetiva adquirida y relacionada a hechos y procedimientos, ideas que pueden o no ser útiles.

“Se puede decir que el conocer es un proceso a través de cual un individuo se hace consciente de su realidad y en éste se presenta un conjunto de representaciones sobre las cuales no existe duda de su veracidad” (Martinez & Rios, 2006).

Entonces el conocimiento es el proceso en el cual un individuo relaciona sus experiencias con la información que este tiene sobre algo y lo guarda en su mente o en un texto escrito, es así que el conocimiento dependiendo del individuo será personal y cambiante a su realidad.

2.4. MÉTODO CIENTÍFICO

Al hablar del método científico es referirse a la ciencia (básica y aplicada) como un conjunto de pensamientos universales y necesarios, y que en función de esto surgen algunas cualidades importantes, como la de que está constituida por leyes universales que conforman un conocimiento sistemático de la realidad.

Etimológicamente, la palabra método está formada por dos raíces griegas: META= camino, o lo largo de, ODOS= camino. Por lo tanto, el método científico es una forma de ordenar y estructurar el trabajo y si esto no se realiza de forma eficaz, obviamente se desperdiciarán tiempo y recursos.

Mas la ciencia no está en modo alguno circunscrita a lo mensurable. “El papel desempeñado por la medición y por la cantidad (cualidades cuantitativas) en la ciencia –dice Bertrand Russell- es en realidad muy importante, pero creo que a veces se le supervalora. Las leyes cualitativas pueden ser tan científicas como la ley cuantitativa.” Tampoco la ciencia está reducida a la física y a la química; más a los defensores del “elevado camino hacia la verdad” les conviene creer que ello es así. Para ellos es necesario, en efecto, presentar a la ciencia como estando limitada, por su misma naturaleza, a la tarea de preparar el escenario para que la entrada en él una forma más elevada de conocimiento.

La investigación científica es esencialmente como cualquier tipo de investigación, sólo que más rigurosa y cuidadosamente realizada. Podemos definirla como un tipo de investigación “sistemática, controlada, empírica, y crítica, de proposiciones hipotéticas sobre las presumidas relaciones entre fenómenos naturales” (Kerlinger, 1975, p. 11).

La investigación puede cumplir dos propósitos fundamentales: a) producir conocimiento y teorías (investigación básica) y b) resolver problemas prácticos (investigación aplicada). Gracias a estos dos tipos de investigación la humanidad ha evolucionado. La investigación es la herramienta para conocer lo que nos rodea y su carácter es universal. Como señala uno de los científicos de nuestros tiempos, Carl Sagan (1998).

Según Hernández los pasos para una investigación son:

- **Concebir la idea a investigar,** la investigación responde a una observación que se le da a una base de datos, de la cual se delimitaran en espacio y tiempo, identificando los beneficios o perjuicios que provoca la relación entre estos datos y el alumno para identificar el problema de índice de deserción para lo cual la idea de investigación debe ser:
  - Debe ser algo novedoso y propositivo
  - Dar un enfoque diferente a lo ya existente
- **Planteamiento del problema,** el problema de la investigación delimitado muestra la relación entre los objetivos que se espera alcanzar y la justificación de la investigación. Para la cual:
  - Planteamiento el problema de manera clara
  - Desarrollar la pregunta de investigación
  - Justificar la investigación
- **Elaboración del marco teórico,** es la parte teórica textual o de referencia del tema a investigar, muestra la recopilación y extracción de información para construir el marco teórico del trabajo.
  - Integrar teorías, enfoque, estudios y antecedentes
  - Revisión de la literatura
  - Recopilación y ordenar la información
  - Evaluar la investigación

- **Definición de la investigación,** se define el tipo de investigación más apropiada para el desarrollo del tema, en el que se describe que tipo de investigación entre las cuales se tienen los tipos de investigación:
  - Estudio exploratorio
    - Es la base de una nueva investigación
    - Parte desde cero
    - No hay referencias en ninguna fuente
    - El tema generalmente no ha sido tocando antes
  - Estudio descriptivo
    - Buscan especificar las características
    - Describir es medir
    - Miden el fenómeno en cualquier dimensión
    - Pueden medir con especificidad cada parte por separado de la investigación
  - Estudio correlacional
    - Se encargan de relacionar dos o más variables dentro de un mismo contexto
    - Evalúan el grado de relación entre las variables
  - Estudio explicativo
    - Están destinados a responder las causas de los eventos físicos y sociales
    - Explican los fenómenos
- **Establecimiento de hipótesis,** Es la suposición de algo posible o imposible para sacar de ello una consecuencia, la hipótesis de investigación se clasifica en:
  - Descriptiva, del valor de variables a observar
  - Correlacionales
  - De las diferencias de grupos
  - Causales,
  - Estadística de estimación

- Estadística de correlación
- Estadísticas de la diferencia de grupos

- **Selección de la muestra,** se fijan una muestra y procedimiento de selección, para poder definir los objetos que van a ser medidos.
  - La fijación del universo
  - Determinar cuál ha de ser la unidad de muestra
  - Determinar el tamaño de la muestra
  - Determinar el método a seguir para selección de los elementos que han de integrar la muestra
- **Recolección de los datos** los datos a recolectar se enmarcan de acuerdo al contexto de la investigación.
  - Elaborar el instrumento de medición y administrarlo
  - Calcular la validez y confiabilidad del instrumento de medición
- **Análisis de los datos,** la definición de las técnicas para el análisis de los datos recolectados, dependen de la hipótesis formulada y los niveles de medición de otras variables.
  - Selecciona pruebas estadísticas
  - Elaborar el problema de análisis
  - Elaborar del reporte final
  - Elaborar el reporte de investigación
  - Presentar el reporte de investigación

2.5. METODOLOGÍA CRISP-DM

Una metodología consiste en un conjunto de actividades organizadas que tiene por objetivo la realización de un trabajo. Es así que, para una correcta ejecución sistemática de un proyecto de Minería de Datos, existen varias metodologías de las cuales podemos citar a KDD, SEMMA, CRISP-DM que actualmente son las más conocidas.

Según un estudio publicado por Nuggets en agosto del 2007, CRISP-DM se ha convertido en la metodología más utilizada por las personas a la hora de afrontar procesos de Minería de Datos.

En el año 1999, empresas europeas como NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), desarrollan la metodología de libre distribución CRISP-DM (Cross-Industry Standard Process for Data Mining).

Según CRIS-DM (2015), la metodología se describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción. De general a especificas serian fase, tarea genérica, tarea especializada e instancia de proyecto.

Shearer (2000) afirma que “CRISP-DM es una Minería de Datos integral metodología y modelo de proceso que proporciona a cualquier persona, desde principiantes hasta expertos en Minería de Datos”.

CRISP-DM, es un método probado para orientar los trabajos de Minería de Datos es así que:

- Como metodología incluye descripciones de las fases normales de un proyecto, las tares necesarias en cada fase y una explicación de las elaciones entre las tareas.
- Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de Minería de Datos.

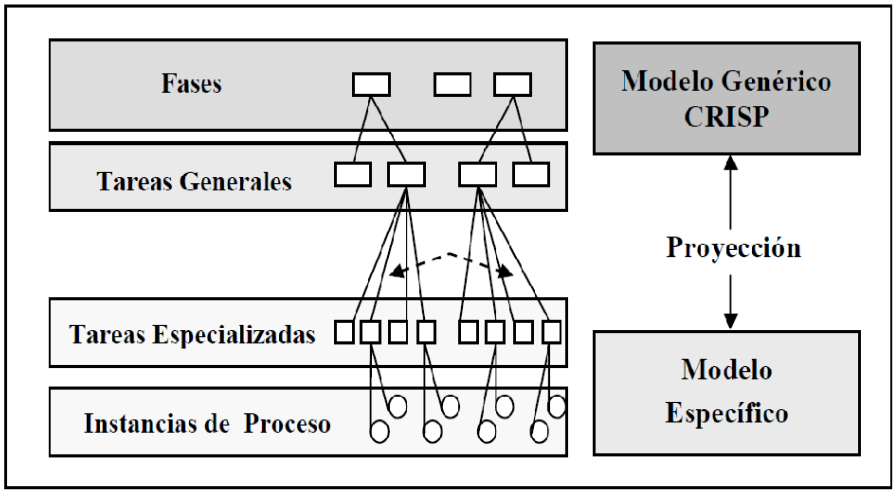


Figura 2.12 **Niveles de la metodología CRISP-DM.**

Fuente: CRIS-DM (2015).

Sus niveles están organizados de la siguiente manera:

- El primer nivel superior organiza a su vez en fases, estas cada una con tareas específicas.
- En el segundo genérico descrito así porque busca cubrir todas las situaciones posibles de la Minería de Datos. Estas tareas tienes que ser lo más completas y estables posibles. “Estable significa que el modelo debe ser válido para desarrollos aún imprevistos como las nuevas técnicas de modelado” (CRIS-DM, 2015).
- En el tercer nivel están las tareas especializadas y tareas específicas.
- En el cuarto nivel esta la instancia del proyecto en sí.

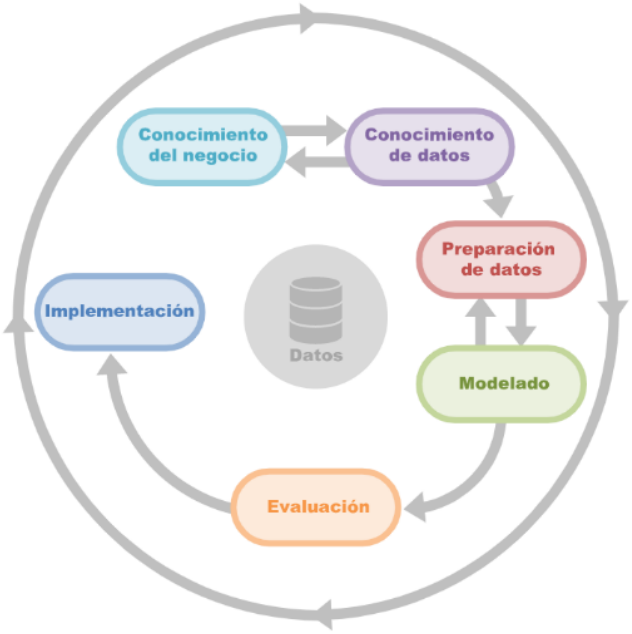


Figura 2.13 **Fases del modelo referencial CRISP-DM.**

Fuente: CRISP-DM (2015).

La sucesión de las fases no es necesariamente tiene que ser rígida, las tareas generales como especificas se deben adecuar y desarrollar para cada situación específica que se encuentre.

Explicado los niveles CRISP-DM se estructura en 6 fases y sus respectivas tareas para cada una de ellas.

2.5.1. Fase de comprensión del problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la Minería de Datos, es necesario entender de la manera más completa el problema que se desea

resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados.

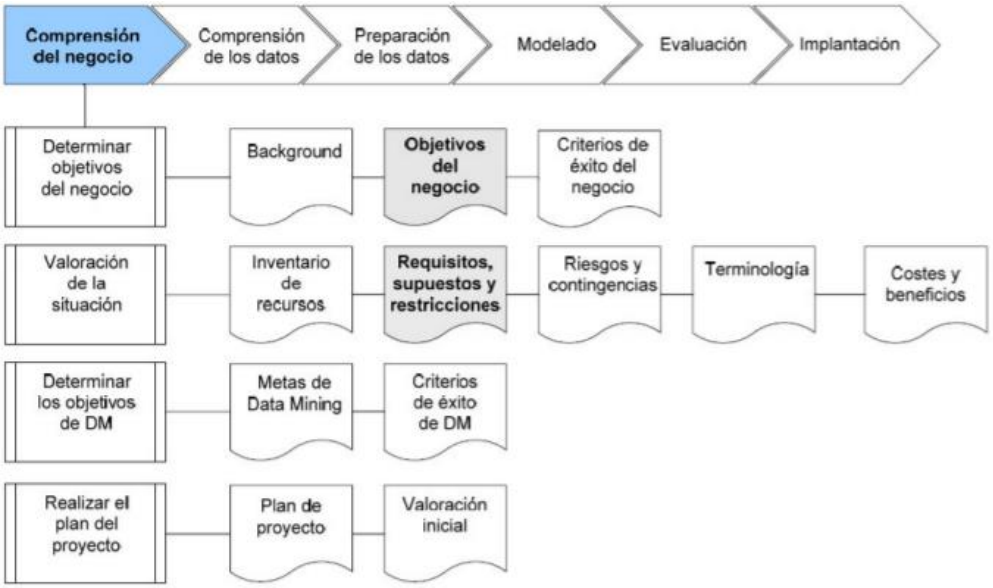


Figura 2.14 **Fase de comprensión del problema o negocio.**

Fuente: CRISP-DM (2015).

En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Minería de Datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

- *Determinar los objetivos del negocio*, esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Minería de Datos y definir los criterios de éxito. Los problemas pueden ser diversos, por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del

proceso de MD, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

- *Evaluación de la situación*, en esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de MD, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de Minería de Datos?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Minería de Datos.
- *Determinación de los objetivos de Minería de Datos,* esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de MD, por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, determinar el perfil de los clientes respecto de su capacidad de endeudamiento.
- *Desarrollar un plan para el proyecto,* que describa los pasos a seguir y las técnicas a emplear en cada paso.

2.5.2. Fase de comprensión de los datos

La fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de MD. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de MD, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar

consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

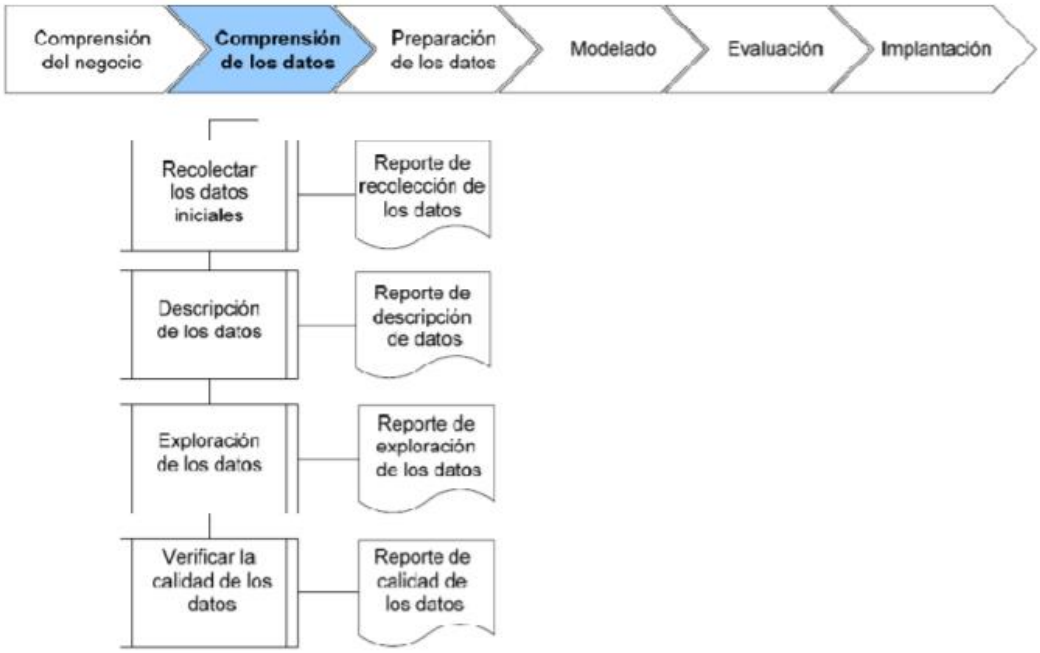


Figura 2.15 **Fase de comprensión de los datos.**

Fuente: CRISP-DM (2015).

Las principales tareas a desarrollar en esta fase del proceso son:

- *Recolección de datos iniciales,* la primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.
- *Descripción de los datos,* después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

- *Exploración de datos,* se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.
- *Verificación de la calidad de los datos,* en esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos.

2.5.3. Fase de preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Minería de Datos que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. Esta fase se encuentra relacionada con la fase de modelado puesto que, en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente.

Una descripción de las tareas involucradas en esta fase es la siguiente:

- *Selección de datos,* se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y

corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de MD seleccionadas.

- *Limpieza de los datos, e*sta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.
- *Estructuración de los datos,* esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.
- *Integración de los datos,* esta tarea involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
- *Formateo de los datos,* esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de MD en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

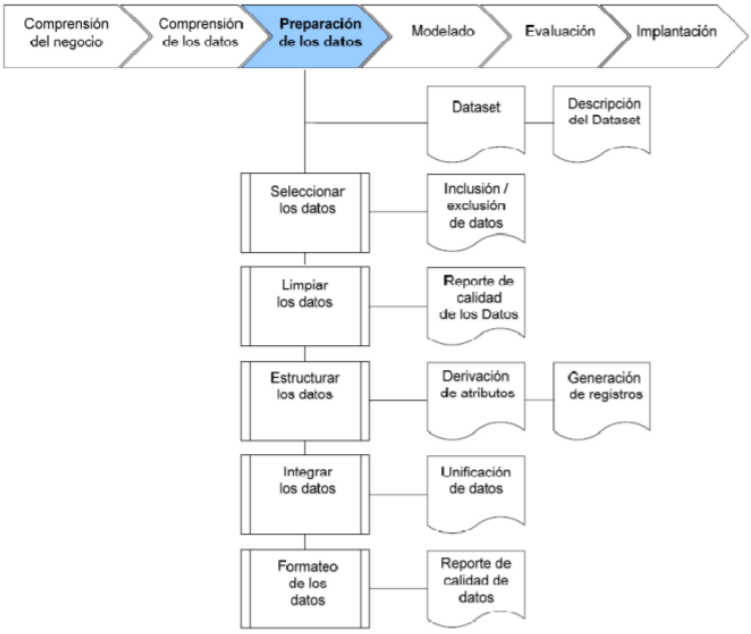


Figura 2.16 **Fase de preparación de los datos.**

Fuente: CRISP-DM (2015).

2.5.4. Fase de modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos.

Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo.

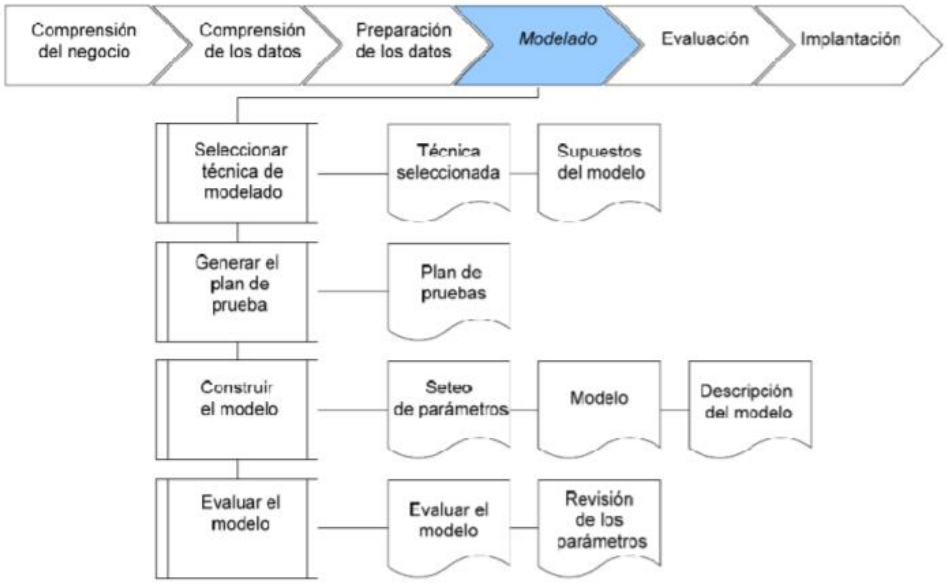


Figura 2.17 **Fase de modelado.**

Fuente: CRISP-DM (2015).

Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

Una descripción de las principales tareas de esta fase es la siguiente:

- *Selección de la técnica de modelado,* esta tarea consiste en la selección de la técnica de MD más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de MD existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el

problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

- *Generación del plan de prueba,* una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de MD como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.
- *Construcción del Modelo,* después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.
- *Evaluación del modelo,* en esta tarea, los ingenieros de MD interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Minería de Datos aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc…).

2.5.5. Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe además considerarse, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se

realizó el análisis. Los modelos de MD, necesariamente están relacionados con los objetivos originales y todos los demás hallazgos.

Después de evaluar los modelos con respecto a los criterios de éxito empresarial, los modelos generados que cumplen los criterios seleccionados se convierten en modelos aprobados.

Las matrices de confusión es una herramienta fundamental a la hora de evaluar el desempeño de un algoritmo de clasificación, ya que dará una mejor idea de cómo se está clasificando dicho algoritmo, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Así se puede comprobar si el algoritmo está clasificando mal las clases y en qué medida.

Tabla 2.2
Tabla de matriz de confusión para un clasificador de dos clases.

		Clasificador	
		Negativos	Positivos
Valores Reales	Negativos	Negativos Reales	Falsos Positivos
	Positivos	Falsos Negativos	Positivos Reales

Nota: Elaboración propia.

Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

Las tareas involucradas en esta fase del proceso son las siguientes:

- *Evaluación de los resultados,* esta tarea involucra la evaluación del modelo en relación a los objetivos del problema y busca determinar si hay alguna razón del problema para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente

relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

- *Proceso de revisión,* se refiere a calificar al proceso entero de MD, a objeto de identificar elementos que pudieran ser mejorados.
- *Determinación de futuras fases, s*i se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de MD.

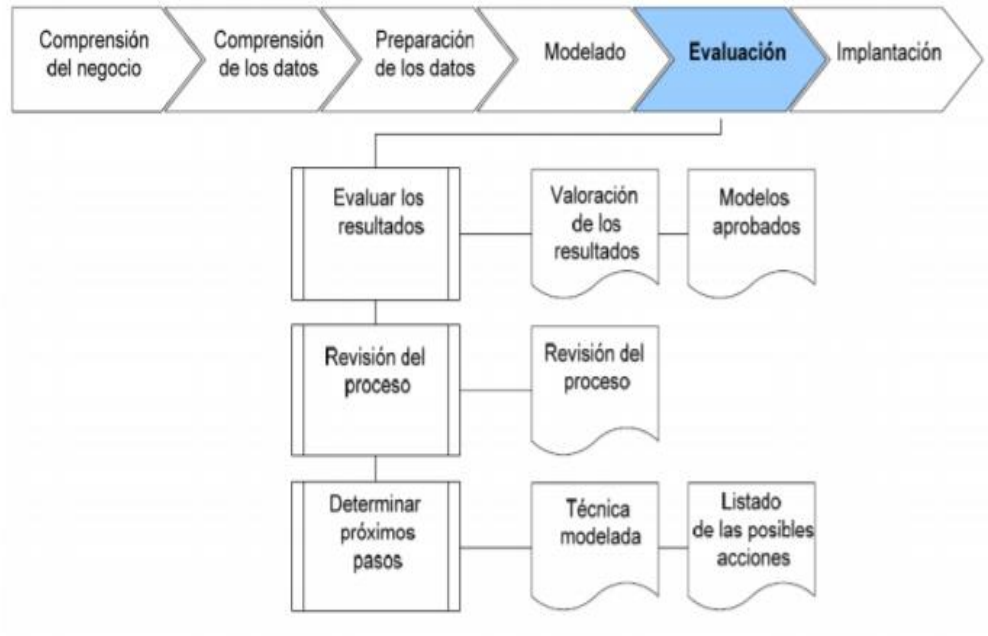


Figura 2.18 **Fase de evaluación.**

Fuente: CRISP-DM (2015).

2.5.6. Fase de implementación

En esta fase una vez que el modelo ha sido construido y validado se comienza la puesta en marcha del proyecto, se instala el modelo resultante. Ya sea como conjunto de datos o como parte del proceso.

Las tareas que se ejecutan en esta fase son las siguientes:

- *Plan de implementación,* para implementar el resultado de MD en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación.
- *Monitorización y Mantenimiento,* si los modelos resultantes del proceso de Minería de Datos son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

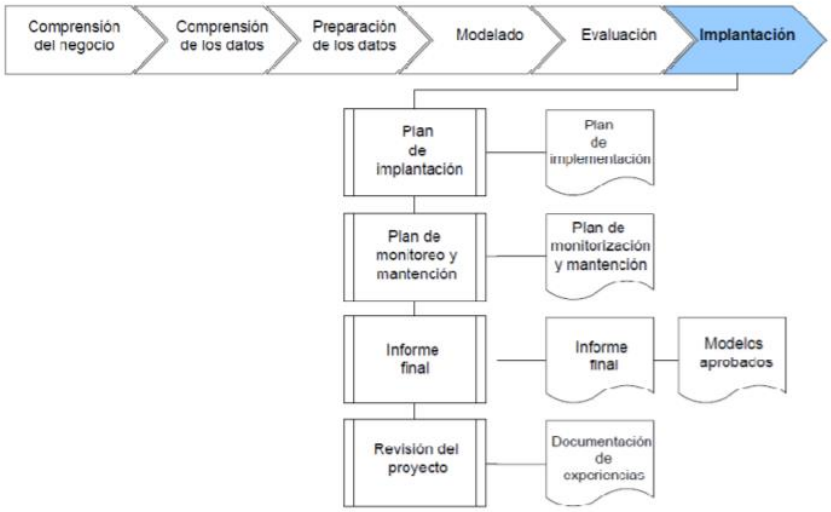


Figura 2.19 **Fase de Implementación.**

Fuente: CRISP-DM (2015).

- *Informe Final,* es la conclusión del proyecto de MD realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.
- *Revisión del proyecto,* en este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

2.6. INGENIERÍA DEL SOFTWARE

"La ingeniería de software es el establecimiento y uso de principios fundamentales de la ingeniería con objeto de desarrollar en forma económica software que sea confiable y que trabaje con eficiencia en máquinas reales" (Bauer, 1969 citado en Pressman, 2010, p.11).

Según Pressman (2010), menciona que la IEEE[8] define a la ingeniería de software como la aplicación de un enfoque sistemático, disciplinado y cuantificable al desarrollo, operación y mantenimiento del software. También asegura que la ingeniería de software está formada por un proceso que tiene métodos y un arreglo de herramientas que buscan desarrollar un software de calidad.

A su vez Cota (1994) se refiere a la ingeniería de software como a la aplicación de los principios de la ciencia de la computación y las matemáticas en busca de soluciones efectivas económicamente.

Zabala (2002) citado por Hidalgo (2014), menciona que la ingeniería del software es la rama de la ingeniería que aplica los principios de la ciencia de la computación y las matemáticas para lograr soluciones costo – efectivas (eficaces

[8] IEEE, Abreviación del Ingles "Institute of Electrical and Electronic Engineers".

en costo o económicas) a los problemas relacionados con el desarrollo del software.

En la actualidad la ingeniera de software es empleada para el desarrollo de sistemas de información, y lograr de esta manera productos que sean eficaces y económicos (Hidalgo, 2014).

La ingeniera de software busca el desarrollo, operación y mantenimiento del software como un producto con calidad, eficiencia y económico, mediante métodos sistemáticos.

2.6.1. Proceso del Software

Jacobson (2000) define al proceso de ingeniería de software "como un conjunto de etapas parcialmente ordenadas con la intención de logra un objetivo, en este caso, la obtención de un producto de software de calidad".

La ingeniería de software, al igual que otras ingenierías, debe trabajar con elementos gerenciales y humanos, además de los elementos técnicos propios. Sin embargo, a diferencia de las otras ingenierías, su producto, el software, es inmaterial. El desarrollo de software no puede, por tanto, ser manejado y controlado como otros procesos para productos físicos. El desarrollo de software es una actividad compleja por naturaleza.

La ingeniería de software, trabaja con elementos humanos y gerenciales. Pero el software no es un producto que se pueda tocas es decir es inmaterial y por lo tanto el desarrollo del software es una activad compleja por naturaleza (Roa, 2017).

2.6.2. Proceso Unificado Racional (RUP)

Es un proceso de ingeniería de software que suministra un enfoque para asignar tareas y responsabilidades al interior de una organización de desarrollo. Su objetivo es asegurar la producción de software de alta calidad que satisfaga la necesidad del usuario final en de un tiempo y presupuesto previsible.

RUP[9] permite a todos los integrantes de un equipo de trabajo, conozcan y compartan el proceso de desarrollo, una base de conocimientos y los distintos modelos de cómo desarrollar el software, utilizado un Lenguaje Unificado de Modelado (UML), se constituye la metodología más utilizada para el análisis, implementación y documentación de sistemas orientados a objetos.

Las fases de esta metodología hacen referencia a un lapso de tiempo el que a su vez corresponde a un determinado flujo de trabajo lineal que resulta en objetivos cumplidos, cada fase utiliza determinados artefactos para alcanzar estos objetivos.

Así tenemos 4 fases bien definidas y disciplinas que se desarrollan en forma de cascada implicando que antes de empezar con la siguiente disciplina se debe de culminar la anterior.

2.6.2.1. Fases del RUP

- **Fase de inicio,** durante esta fase de inicio las iteraciones se centran con mayor énfasis en las actividades de modelado de la empresa y en sus requerimientos
- **Fase de elaboración,** durante esta fase de elaboración, las iteraciones se centran en el desarrollo de la base del diseño, encierran los flujos de trabajo de requerimientos, modelo de la organización, análisis, diseño y una parte de implementación orientada a la base de la construcción
- **Fase de construcción,** durante esta fase de construcción, se lleva a cabo la construcción del producto por medio de una serie de iteraciones las cuales seleccionan algunos casos de uso, referidas a su análisis y diseño además se procede a su implementación y pruebas. En esta fase se realiza una pequeña cascada para cada ciclo, se realizan tantas iteraciones hasta que se termine la nueva implementación del producto.

[9] RUP, por sus siglas en inglés Rational Unified Process.

- **Fase de transición,** esta fase de transición busca garantizar que se tiene un producto preparado para su entrega al usuario.

2.6.3. Lenguaje Unificado de Modelado (UML)

UML es un lenguaje visual para especificar, construir y documentar sistemas, define una notación que se expresa como diagramas que sirven para representar modelos o partes de ellos.

Sus objetivos se pueden resumir de la siguiente manera:

- Modelar todo tipo de sistema.
- Creación de un lenguaje de modelado.
- Acoplamiento: modelo – artefacto.
- Manejar problemas.

Esta modelo consta de diferentes diagramas entre los cuales podemos mencionar algunos y una breve descripción de que se realiza en cada uno de estos diagramas, pero cabe recalcar que se pueden realizar combinación entre estos.

- **Diagrama de casos de uso,** es una descripción de las acciones del sistema, nos muestra los requerimientos del sistema desde el punto de vista del usuario.
- **Diagramas de clases,** el diagrama es utilizado para describir la estructura del sistema que muestra sus clases.
- **Diagrama de objetos,** diagrama de objeto es un gráfico de instancias que incluye objetos y datos, el mismo es una instancia de un diagrama de clases.
- **Diagrama de estado,** se utiliza para especificar el estado en el que se encuentra un objeto y sus cambios, es decir representa situaciones durante la vida del objeto.

- **Diagrama de secuencias,** el diagrama muestra la mecánica de la interacción con base en tiempos, donde las clases y los objetos representan la información.
- **Diagrama de actividades,** el diagrama de actividades muestra la naturaleza dinámica de un sistema mediante el modelado de flujo ocurrente de actividad en actividad. Una actividad representa una operación en alguna clase del sistema que es resultante del cambio en el estado del sistema.
- **Diagrama de colaboraciones,** el diagrama de colaboraciones describe las interacciones entre los objetos en términos de mensajes secuenciados. Son una representación entre el diagramad de clases, de secuencias y casos de uso.
- **Diagrama de componentes,** el diagrama de componentes muestra al individuo con sus respectivas dependencias entre ellos.
- **Diagrama de distribuciones,** el diagrama de distribuciones muestra la arquitectura física de un sistema de informático. Puede representar los equipos y dispositivos, mostrar sus interconexiones y el software que se encontrara en cada máquina.
- **Diagrama de despliegue,** es un diagrama utilizado para modelar la disposición física de los dispositivos de software en nodos. Algunos de estos diagramas muestran el despliegue de modelados de sistemas empotrados, sistemas cliente-servidor y sistemas completamente distribuidos.

2.7. HERRAMIENTAS

Las herramientas de la ingeniería de software proporcionan el soporte automatizado o semi automatizado para el proceso y los métodos. Cuando las herramientas se integran de forma que la información que cree una de ellas pueda usarla otra, se dice que se ha establecido un sistema para el soporte del

desarrollo del software, que con frecuencia se denomina ingeniería del software asistida por computadora (Pressman, 2010).

Las herramientas que se emplean en la Minería de Datos buscan la extracción de conocimiento y se las puede clasificar principalmente en:

- Técnicas de verificación, en las que se busca probar una hipótesis.
- Métodos de descubrimiento, incluyen las de predicción se aplican para la búsqueda de patrones.

Es así que dentro de la Minería de Datos existen varias herramientas dedicadas a la búsqueda de patrones para la predicción entre las cuales se podemos mencionar a SPPS Clemente, Oracle Data Miner y WEKA entre otros. A continuación, se lista las herramientas que se utilizaran para la implementación de la presente propuesta de tesis:

2.7.1. Minería de Datos

- **Weka** (Waikato Environment for Knowledge Analysis, Entorno para Análisis del Conocimiento desarrollado por la Universidad de Waikato), es una plataforma de software para aprendizaje automático y Minería de Datos escrito en Java. Weka es un software libre distribuido bajo licencia GNU-GPL. servirá para realizar la Minería de Datos ya que este cuenta con diferentes herramientas y una gran cantidad de algoritmos que se especializados en el área.

2.7.2. Sistema operativo

- **Windows,** es una edición súper completa diseñado para toda la familia de los productos Microsoft, es un sistema operativo propietario que funciona bajo la arquitectura de x86 bits. y x64 bits. Es uno de los sistemas más comercial en el mundo. Su modelo es de desarrollo privativo, Share Source; tipo de núcleo monolítico, Uno de los aspectos más importantes de Windows 10 es el enfoque en la armonización de

experiencias de usuario y funcionalidad entre diferentes tipos de dispositivos.

2.7.3. Base de datos

- **PostgreSQL,** es un servidor de Bases de Datos relacionales Orientadas a Objetos, de software libre bajo licencia BSD. Se opta por postgreSQL es un magnifico gestor de bases de datos. Tiene prácticamente todo lo que tienen los gestores comerciales, haciendo de él una muy buena alternativa GPL.

2.7.4. Lenguaje de programación

- **Java,** es un lenguaje de programación y una plataforma informática comercializada por primera vez en 1995 por Sun Microsystems. Hay muchas aplicaciones y sitios web que no funcionarán a menos que tenga Java instalado y cada día se crean más. Java es rápido, seguro y fiable. Desde portátiles hasta centros de datos, desde consolas para juegos hasta súper computadoras, desde teléfonos móviles hasta Internet, Java está en todas partes.

2.7.5. Herramienta IDE

- **NetBeans,** es un programa que sirve como IDE (un entorno de desarrollo integrado) que nos permite programar en diversos lenguajes. ofrece herramientas de primera clase para el desarrollo de aplicaciones web, corporativas, de escritorio y móviles con Java. Siempre es el primer IDE en ofrecer soporte para las últimas versiones de JDK, Java EE y JavaFX. Proporciona descripciones generales inteligentes para ayudarle a comprender y gestionar sus aplicaciones, lo que incluye el soporte inmediato para tecnologías populares, como Maven.

  NetBeans contiene tecnologías innovadoras listas para usar y es el estándar en el desarrollo de aplicaciones, gracias a sus características

integrales para el desarrollo de aplicaciones, las constantes mejoras en el editor de Java y el perfeccionamiento del rendimiento y la velocidad.

2.7.6. Herramientas case

- **MagicDraw UML**, es una herramienta CASE desarrollada por No Magic. La herramienta es compatible con el estándar UML 2.3, desarrollo de código para diversos lenguajes de programación (Java, C++ y C#, entre otros) así como para modelar datos.

2.8. OBJETIVO DEL DESCUBRIMIENTO DE CONOCIMIENTO

Según la CRES [10](2008) citado por Lupín y sus colegas (2013), "la educación superior es un bien público social, un derecho humano y universal y un deber del Estado". Conclusión que llegaron alrededor de 3.500 países integrantes de la Conferencia Regional de Educación Superior en América Latina y el Caribe.

Por otra parte, la UNESCO (2019), la educación superior ha sido reconocida como un derecho humano y un bien público social, que "para el 2030 aseguran el acceso en condiciones de igualdad para todos los hombres y las mujeres a una formación técnica, profesional y superior de calidad, incluida la universitaria". Y además que la educación superior es un elemento fundamental para hacer realidad su desarrollo económico y social, reduciendo la pobreza y las desigualdades socioeconómicas en los países. Sin embargo, la situación actual del país no favorece a que dicho derecho sea ejercido de manera plena. El Instituto de Estadística de la UNESCO supervisa las metas mediante el indicador de "la tasa bruta de matrícula en la educación superior".

2.8.1. Deserción Universitaria

La deserción universitaria no se debe definir a todos los abandonos de estudios, ni todos los estos abandonos merecen la intervención institucional. La deserción en la educación superior puede ser interpretada desde varias

[10] CRES siglas de Conferencia Regional de Educación Superior para América Latina y el Caribe

perspectivas y diferentes tipos de abandono, ya sean por el comportamiento individual, metas individuales, institucionales y estatales. Que la deserción es un fenómeno muy complejo y que el investigador o funcionario debe elegir cuidadosamente aquella definición que más se adecue sus interés y metas. (Tinto, 1989).

Es necesario precisar que es la deserción, diferenciarla de otros fenómenos educativos, reconocer sus variables y sus niveles, para determinar los factores de deserción aplicables a estudiantes de educación superior (Páramo & Correa, 1999).

Según RAICES citados por Romero (2016), cita que la deserción es un sinónimo de abandono, de mortalidad escolar el cual define como la suspensión, repetición, cambio de carrera o abandono antes de obtener el título.

“La deserción afecta a todos los niveles educativos, ya sean estos de educación primaria, secundaria y universitaria, por lo que se puede observar la insuficiente capacidad de retención de niños, adolescentes y adultos” (Romero, 2016)

“La deserción es el abandono del sistema educativo sin haber concluido el mismo”. Es el eslabón final en el fracaso escolar, con una probabilidad de haber repetido y alargado su educación debilitando su autoestima, lo cual lleva a un fracaso (Perassi, 2009).

El termino deserción se utiliza para aquellos alumnos que abandonan sus estudios por diferentes razones. Refiriéndose a cualquier tipo de estudio ya sea escolar o universitario (Pérez & Gardey, 2012).

Según Himmel (2002) define a la deserción como el “abandono prematuro de un programa de estudios antes de alcanzar el título o grado”. Se debe tomar en cuenta el tiempo prudente para asegurar la deserción.

La deserción voluntaria puedes entenderse como la renuncia por parte del alumno de la carrera universitaria como un abandono no informado a dicha

institución, y la no voluntaria por una decisión institucional que obligaría al retiro del alumno, fundamentada en los reglamentos universitarios (Himmel, 2002).

Se entiende por deserción estudiantil al abandono definitivo de las aulas de clase, la no continuidad de la formación académica que la sociedad lo requiere y lo espera, por diferentes razones. El ausentismo y el retiro forzoso son fenómenos que van junto a la deserción (Páramo & Correa, 1999).

Dentro de los estudiantes que desertan se los puede clasificar en tres grupos, los que tienen metas más amplias, metas educativas restringidas y estudiantes que trabajan (Tinto, 1989).

Para Bueno (2019), el abandono de los estudios universitarios significa un mal uso del recurso económico público ya que la matricula no cubre el total del costo de educación y además esto produce una frustración personal.

La mayor parte de la deserción está en la población estudiantil de bajos recursos, pero dicha deserción muchas veces no es voluntaria. Fuera de sus motivaciones muchas veces son circunstancias externas, por lo que las autoridades pertinentes deben buscar la igualdad de oportunidades mediante la otorgación de diferentes becas (Ramirez, Diaz, & Salcedo, 2017).

2.8.2. Índices de deserción

Según Romero (2016), menciona que la deserción en la educación superior es “la cantidad de estudiantes que abandonan el sistema de educación superior entre uno y otro periodo académico (semestre o año)”. Dicho índice se lo calcula entre el total de matriculados del primer periodo, menos los egresados del mismo periodo y más los alumnos matriculados del siguiente semestre, esto genera un nuevo estado un estado de alumnos matriculados sin deserción ideal.

Se define la deserción universitaria como el área entre la permanencia inicial y la permanencia real. Tomando en cuenta el registro de curricular por periodos. La deserción se define la permanecía como la cantidad de estudiantes respecto del

total de estudiantes del ano de ingreso (Rodriguez, Espinoza, Ramirez, & Ganga, 2018).

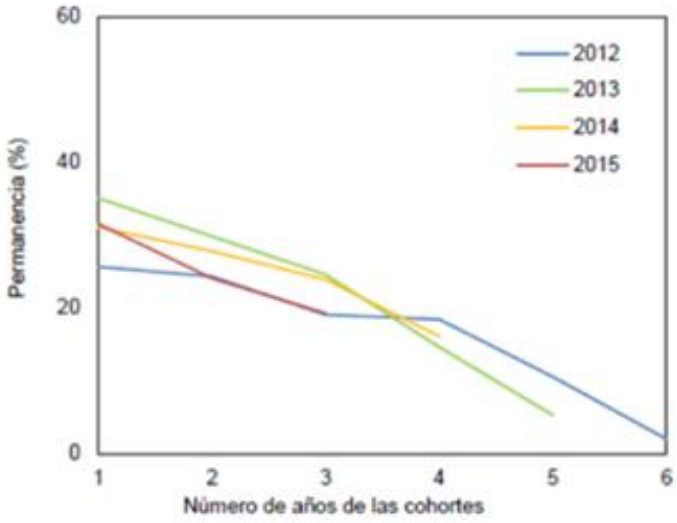


Figura 2.20 Permanecía estudiantil carrera C1.

Fuente: Rodriguez, Espinoza, Ramirez, & Ganga (2018).

2.8.3. Aspectos Académicos

Dentro de los Reglamentos Generales de la U.P.E.A. en capitulo III esta reconoce a dos tipos de estudiantes:

- Estudiante regular, es aquel que, habiéndose matriculado y programado asignaturas, participa y cumple obligatoriamente con todas las actividades, exigencias académicas establecidas en los planes de estudio y programas docentes.
- Estudiante libre, es aquel que por motivos plenamente justificados se ve impedido de asistir y participar regularmente a las clases de cada asignatura, por lo que está sujeto a un régimen especial establecido por

reglamento específico, cumpliendo con el plan de estudios correspondiente.

2.9. MÉTRICA DE CALIDAD

Todo producto debe ser analizado bajo parámetros o métricas estandarizadas que indiquen la calidad que ofrecerá este, es así que el software como producto tiene que ser constantemente sometido a mediciones en el proceso de desarrollo del software y sus productos, para suministrar información relevante a tiempo. Según Pressman (2010), menciona que la calidad es algo que se ve, pero en el caso del software es algo difícil de definir.

Las métricas de calidad del software comprenden un amplio rango diversas, como ser el aseguramiento de y control de calidad, modelos de fiabilidad, modelos y evaluación de ejecución, modelos y medidas de producción.

Para este propósito existen diferentes métricas o nomas entre las que están las los Factores de Calidad McCall, Modelo de FURPS (Funcionality, Usability, Reliability, Performance y Supportability), Normas ISO 9126, MOSCA (Modelo Sistémico de Calidad) y la QSOS (Qualification and Selection of Open Source software) entre otras.

2.9.1. ISO/IEC 9126

La norma ISO/IEC 9126, es una norma internacional para la evaluación de la calidad del software. Permite especificar y evaluar la calidad del software desde diferentes criterios asociados con adquisición, requerimientos, desarrollo, uso, evaluación, soporte, mantenimiento, aseguramiento de la calidad y auditoria de software.

El estándar ISO/IEC 9126 se desarrolló con la intención de identificar los atributos clave del software de cómputo. El estándar identifica seis atributos clave de calidad:

- **Funcionalidad,** La funcionalidad es la capacidad que tiene el software de cumplir y proveer las funciones para satisfacer las necesidades explícitas e implícitas de una organización. Los atributos que toma en cuanta son adecuación, exactitud, interoperabilidad, seguridad y conformidad

Tabla 2.3
Condiciones especificadas

Característica	Ponderación
Adecuación	Proporcionar un conjunto apropiado de funciones para tareas y objetivos especificados
Exactitud	Proporcionar los resultados o efectos correctos o acordados
Interoperabilidad	Interactuar con uno o más sistemas especificados
Cumplimiento funcional	Adherirse a normas, convenciones o regulaciones en leyes y prescripciones

Nota: recuperado de Ingeniería de Software.

- **Confiabilidad,** Es la capacidad que tiene el software que nos permite asegurar un nivel de funcionamiento adecuado a las condiciones específicas, esta se refiere a cuatro criterios, madurez, tolerancia y facilidad de recuperación.

  La confiabilidad se calcula mediante la siguiente formula:

  $$\text{Confiabilidad} = 1 - (\text{n error}/\text{n LDC}) \times 100$$

  Donde:

  N error = número de errores

  LDC = líneas de código

- **Usabilidad,** Es la capacidad del software de ser usado con facilidad de forma atractiva. La usabilidad está determinada por los usuarios finales y los usuarios indirectos del software, dirigidos a todos los ambientes, a la preparación del uso y el resultado obtenido. Los parámetros que toma en cuenta son la facilidad de comprensión, facilidad de aprendizaje y operatividad.
- **Eficiencia,** La eficiencia del software se refiere a la forma del desempeño adecuado según al número de recursos utilizados y las condiciones planteadas, asimismo se debe tomar en cuenta los aspectos como la configuración de hardware, el sistema operativo, entre otros. Toma en cuenta el comportamiento de tiempos, utilización de recurso, conformidad de recursos y conformidad de eficiencia.

  Su fórmula es:

  Eficiencia del software = Eficiencia / LCC

  Donde:

  Eficiencia del software = disponibilidad * Confiabilidad * Mantenibilidad * Capacidad

  LCC = Costo de ciclo de vida

- **Capacidad de mantenimiento**, es la cualidad que tiene el software para ser modificado en ella se incluyen correcciones, mejoras del software, referidos a los cambios en el entorno, y especificaciones de requerimientos funcionales. Comprende los atributos de Facilidad de análisis, Facilidad de cambio y estabilidad.

  Para el cálculo de mantenibilidad se usa la formula:

  $$Mantenibilidad = (Mt – (Fc + Fa + Fd))/Mt$$

  Donde:

  Mt = número de módulos en la versión actual.

Fc = número de módulos en la versión actual que han cambiado.

Fa = número de módulos en la versión actual añadido.

Fd = número de módulos en la versión anterior que se ha borrado.

- **Portabilidad,** Es la capacidad que tiene el software para ser trasladado de un entorno a otro. Se refiere a la facilidad de instalación, facilidad de ajuste y facilidad de adaptación al cambio.
  Su fórmula es:

  Portabilidad = 1 – (ndpm/ndim)

  Donde:

  ndpm = número de días para portar el modelo

  ndim = número de días para implementar el modelo

Pressman (2010), menciona que al igual que otros factores de la calidad de software, los factores ISO 9126 no necesariamente conducen a una medición directa. Sin embargo, son una base útil para hacer mediciones indirectas y una lista de comprobaciones excelente para evaluar la calidad del sistema.

2.10. EVALUACIÓN DE COSTOS

Por otra parte, también deben ser analizados los costos de desarrollo de software ya que esto es muy importante a la hora de selección del software. Del Valle (2014), menciona que el análisis de los costos es el proceso de identificar la calidad y cantidad de los recursos en términos económicos, esfuerzo, capacidad, conocimiento y tiempos que afectaran directamente a la entidad donde se aplicar dicho software. Para este fin existen diferentes herramientas de estimación de costos entre los cuales están COCOMO, CoCots, CoStar, CostModeler, SoftCost, entre otras más.

2.10.1. COCOMO II

El Modelo Constructivo de Costos (COCOMO[11]),desarrollado por Barry Boehm en 1981. Es un modelo jerárquico de estimación de costos de software. Debido a deficiencias encontradas surge COCOMO II un modelo que permite el costo, el esfuerzo y el tiempo cuando se planifica una nueva actividad de desarrollo de software.

Estos cambios incluyen el gasto de tanto esfuerzo en diseñar y gestionar el proceso de desarrollo software como en la creación del producto software, un giro total desde los Main Frame que trabajan con procesos Batch nocturnos hacia los sistemas en tiempo real y un énfasis creciente en la reutilización de software ya existente y en la construcción de nuevos sistemas que utilizan componentes software a medida. (Boehm, 1981)

Estos y otros cambios hicieron que la aplicación del modelo COCOMO original empezara a resultar problemática. La solución al problema era reinventar el modelo para aplicarlo a los 90. Después de muchos años de esfuerzo combinado entre USC-CSE1, IRUS y UC Irvine22 y las Organizaciones Afiliadas al Proyecto COCOMO II, el resultado es COCOMO II, un modelo de estimación de coste que refleja los cambios en la práctica de desarrollo de software profesional que ha surgido a partir de los años70. Este nuevo y mejorado COCOMO resultará de gran ayuda para los estimadores profesionales de coste software.

Por tanto, COCOMO II es un modelo que permite estimar el coste, esfuerzo y tiempo cuando se planifica una nueva actividad de desarrollo software.

El principal cálculo en el modelo COCOMO es el uso de la ecuación del esfuerzo para estimar el número de personas o de meses necesarios para desarrollar el proyecto. El resto de resultados del modelo se derivan de esta medida.

Por un lado, COCOMO define tres modos de desarrollo o tipos de proyectos:

[11] COCOMO, por su significado en inglés COnstructive COst MOdel

- **Orgánico**, proyectos de software pequeños y sencillos, menores de 50 KDLC líneas de código, en los cuales se tiene experiencia de proyectos similares y se encuentran en entornos estables.
- **Semi – acoplado**, proyectos intermedios en complejidad y tamaño (menores de 300 KDLC), donde la experiencia en este tipo de proyectos es variable y las restricciones intermedias.
- **Empotrado**, proyectos bastante complejos, en los que apenas se tiene experiencia y se engloban en un entorno de gran innovación técnica. Además, se trabaja con unos requisitos muy restrictivos y de gran volatilidad.

Y por otro lado existen diferentes modelos que define COCOMO:

- **Modelo básico**, Calcula el esfuerzo (y el costo) del desarrollo en función del tamaño del software, expresado en las líneas estimadas de código (LDC).
- **Modelo intermedio**, Calcula el esfuerzo del desarrollo en función del tamaño del programa y de un conjunto de "conductores de coste".
- **Modelo avanzado**, Incluye todo lo del modelo intermedio además del impacto de cada conductor de coste en las distintas fases de desarrollo.

2.10.1.1. Ecuaciones de COCOMO

- **Estimación del esfuerzo de desarrollo**

  La ecuación básica que utilizan los tres modelos es:

  $$E = a*(KLDC)^b * m(X) \text{ (personas x mes)}$$

  Dónde:

  - E es el esfuerzo estimado de desarrollo, representa las personas-mes necesarios para ejecutar el proyecto.
  - a y b son constantes con valores definidos en cada submodelo.
  - KLDC es la cantidad de líneas de código, en miles.
  - m(X) es un multiplicador que depende de 15 atributos.

- **Estimación del tiempo de desarrollo**

  Para el cálculo del tiempo de desarrollo se utiliza la siguiente fórmula:

  $$T = c * (E)^d \text{ (meses)}$$

  Dónde:

  - T es el tiempo de duración del desarrollo.
  - E es el esfuerzo estimado.
  - c y d son constantes con valores definidos en cada submodelo.

- **Estimación del número de personal promedio**

  Para estimar la cantidad de personas se utiliza la fórmula:

  $$P = E/T \text{ (personas)}$$

  Dónde:

  - P es el número de personas.
  - E es el esfuerzo estimado.
  - T es el tiempo de duración del desarrollo.

- **Estimación de productividad**

  Para estimar la productividad se utiliza la fórmula:

  $$PR = LDC/E \text{ (LDC/persona-mes)}$$

  Dónde:

  - PR es la relación cantidad de líneas de código por persona al mes.
  - KLDC es la cantidad de líneas de código, en miles.
  - E es el esfuerzo estimado.

Estos modos de desarrollo permiten utilizar cuatro valores constantes. Estos valores constantes, codificados aquí como “a”, “b”, “c” y “d”, son propuestos por el modelo COCOMO para complementar las ecuaciones de cálculo usadas en el modelo.

Tabla 2.4
Tabla de estimación de esfuerzo

MODO	A	b	c	d
Orgánico	2.40	1.05	2.50	0.38
Semi acoplado	3.00	1.12	2.50	0.35
Empotrado	3.60	1.20	2.50	0.32

Fuente: Grupo de Investigación de costos (Beltrán, 2008)

Estos valores son para las fórmulas:

- Esfuerzo nominal en personas (E).
- Tiempo de desarrollo del proyecto (T).
- Personas necesarias para realizar el proyecto (NP).
- Costo total del proyecto (CT).

Se puede observar que a medida que aumenta la complejidad del proyecto (modo), las constantes aumentan de 2.4 a 3.6, que corresponde a un incremento del esfuerzo del personal. Hay que utilizar con mucho cuidado el modelo básico puesto que se obvian muchas características del entorno.

Cada uno de estos multiplicadores de esfuerzo, tiene una valoración que se clasifica en una escala de 6 valores desde “muy bajo”, “bajo”, “nominal”, “alto”, “muy alto” y “extraordinariamente alto”. Estos multiplicadores de esfuerzo ajustan el valor real del esfuerzo. Los factores seleccionados se agrupan en cuatro categorías:

- **Atributos del producto de software**
    - RELY: garantía de funcionamiento requerida al software. Indica las posibles consecuencias para el usuario en el caso que existan

defectos en el producto. Va desde la sola inconveniencia de corregir un fallo (muy bajo) hasta la posible pérdida de vidas humanas (extremadamente alto, software de alta criticidad).

  - DATA: tamaño de la base de datos en relación con el tamaño del programa. El valor del modificador se define por la relación: D / K, donde D corresponde al tamaño de la base de datos en bytes y K es el tamaño del programa en cantidad de líneas de código.
  - CPLX: representa la complejidad del producto.

- **Atributos del hardware**
  - TIME: limitaciones en el porcentaje del uso de la CPU.
  - STOR: limitaciones en el porcentaje del uso de la memoria.
  - VIRT: volatilidad de la máquina virtual.
  - TURN: tiempo de respuesta requerido
- **Atributos del personal involucrado en el proyecto**
  - ACAP: calificación de los analistas.
  - AEXP: experiencia del personal en aplicaciones similares.
  - PCAP: calificación de los programadores.
  - VEXP: experiencia del personal en la máquina virtual.
  - LEXP: experiencia en el lenguaje de programación a usar.
- **Atributos propios del proyecto**
  - MODP: uso de prácticas modernas de programación.
  - TOOL: uso de herramientas de desarrollo de software.
  - SCED: limitaciones en el cumplimiento de la planificación.

CAPITULO III

MARCO APLICATIVO

MARCO APLICATIVO

El presente capitulo tiene como propósito el poner en práctica lo mencionado en los capítulos anteriores, se analiza y se desarrolla el modelo de predicción en base a Minería de Datos, de manera que satisfaga las necesidades en el análisis de los índices de deserción de los alumnos de la Universidad Pública de El Alto.

3.1. INTRODUCCIÓN

La Minería de Datos es la búsqueda de patrones en bases de datos, mediante un proceso el cual se aplica diferentes algoritmos a la base de datos. El resultado de estos patrones es visto en un resumen de los datos de entrada.

El modelo debe ser realizado mediante pasos metodológicos que exige el tratamiento de datos confiables y actualizados, para que este proceso sea exitoso la base de datos debe contener gran cantidad de información, ser normalizados y posteriormente analizados.

El presentes trabajo considera la metodología de investigación científica que plantea Hernández, los cuales se consolidan en los cinco capítulos al interior del estudio realizado sobre los índices de deserción universitaria. Estos conceptos se detallan a continuación.

Para el planteamiento de la hipótesis de “Debido a la aplicación de técnicas de Minería de Datos y la ingeniería de Software se tiene el modelo predictivo del índice de deserción en base a factores del alumno, teniendo una eficacia del 90% en la población estudiantil de la Universidad Pública de El Alto” ,se partió por la concepción de la idea principal no sin antes revisar la bibliografía correspondiente y trabajos similar al presente, por lo que se analizó la problemática de la UPEA

respecto al índices de deserción universitaria considerados estos puntos en el capítulo uno.

Siguiendo la metodología de investigación científica, es necesario puntualizar y definir los conceptos necesarios para el presente documento que comprenden desde enfoque, las metodologías y las herramientas, por lo que se consideró a varios autores y diferentes textos, tanto impresos como digitales, ya que, en base a ello se elaboró el capítulo dos correspondiente al marco teórico.

Una vez determinado el marco teórico, se recolectó los datos del Sistema de Información y Estadística de la U.P.E.A., que hace referencia a la población universitaria como objeto de estudio de la presente investigación, de manera estratificada y parametrizada, dando lugar al procesamiento de los datos mediante el modelo de índices de deserción de alumnos en base a Minería de Datos en el capítulo tres.

Los datos, son seleccionados, estandarizados, procesados y analizados, según el método de investigación deben ser medibles y sustentables, que correspondan al aspecto científico, los cuales se procesaron en el capítulo tres y cuatro con la ayuda de métricas de calidad de software como: QSOS, COCOMO II y T-STUDENT.

El reporte final del presente trabajo de investigación sobre el índice de deserción estudiantil, así como las sugerencias se presentan en el capítulo cinco.

3.2. METODOLOGÍA DE LA INVESTIGACIÓN

3.2.1. Tipo investigación

Según Hernández, la investigación correlacional tiene como finalidad conocer la relación o grado de asociación que exista entre dos o más conceptos, categorías o variables. Es decir, intentar predecir el valor aproximado que tendrá un grupo de individuos o casos en una variable, a partir del valor que poseen en las variables relacionadas.

En la presente investigación es de tipo correlacional ya que se desea saber la relación entre el modelo de predicción y los índices de deserción, ya que investigación correlacional busca identificar probables relaciones entre variables medibles, buscando saber cómo se puede comportar un concepto o variable conociendo el comportamiento de la otra variable.

3.2.2. Método investigación

Para el presente trabajo de investigación se utilizó el método científico según en el enfoque de Hernández.

Es el camino planeado o la estrategia que se debe seguir para obtener un resultado; éste opera con conceptos, definiciones, hipótesis, variables e indicadores que son los elementos básicos que proporcionan los recursos e instrumentos intelectuales con los que se ha de trabajar para construir el sistema teórico de la ciencia, y así lograr el objetivo de la investigación.

3.2.3. Enfoque de investigación

La investigación sigue un enfoque cuantitativo, ya que se busca comparar y determinar causalidad que implican los índices de deserción. El enfoque cuantitativo se centra en los números arrojados para cada respuesta generalizada cuando se ha realizado la codificación.

Según Hernández utiliza la recolección de datos para probar la hipótesis con base en la medición numérica y el análisis estadístico, con el fin de establecer pautas de comportamiento y probar teorías.

3.2.4. Muestreo

La presente investigación, tomó una muestra probabilística. Para la cual se define al universo de estudio a los alumnos de educación superior pertenecientes a la Universidad Pública de El Alto.

La unidad de análisis inicial fue recolectada en la unidad de Sistema de Información y Estadística perteneciente a esta casa de estudio, en la gestión

2020. Estos datos históricos son pertenecientes a las gestiones 2016 hasta el momento de la solicitud en dicha unidad.

La muestra del universo o población del cual se recolectan los datos y que debe ser representativo de ésta, será determinada mediante la ecuación citado por Hernández.

3.3. COMPRENSIÓN DE PROBLEMA

Tomando en cuenta que la carrera universitaria comprende cinco años de estudios y que la matriculación es anual, esto implicaría que el alumno debería matricularse de manera continua como mínimo cinco años seguidos para culminar dicha carrera, aunque muchos estudiantes pueden hacer una extensión lo cual no limita estos parámetros anuales fijos, es decir pueden extenderse los cinco años lo cual no se lo considera como deserción.

La Minería de Datos busca nuevo conocimiento a partir de bases de datos, en la presente investigación se busca los índices de deserción de los alumnos de la Universidad Pública de El Alto. Se toma en cuenta la base de datos que existe

en la Universidad sobre los registros de datos sobre los alumnos matriculados, el cual es llenado anualmente por cada universitario en el cual se indican aspectos académicos y socio-económicos.

3.4. COMPRENSIÓN DE LOS DATOS

Los datos de entrada que se utilizan en el presente trabajo son la recopilación de los alumnos registrados en el formulario 01 de matriculación anual que tiene la Universidad Pública de El Alto.

3.4.1. Recolección de datos

Figura 3.1 **Formulario 01 de matriculación.**

Fuente: MAE (2017).

El formulario 01 de inscripción tiene por objetivo permitir a los estudiantes de la U.P.E.A. registrar información real, fidedigna y actualizada en línea, a través del sistema denominado MAE[12].

[12] MAE, del Sistema denominado Matriculacion Academica Estudiantil

La base de datos obtenidos sobre el formulario 01, se obtuvieron en formato hoja de cálculo de Excel.

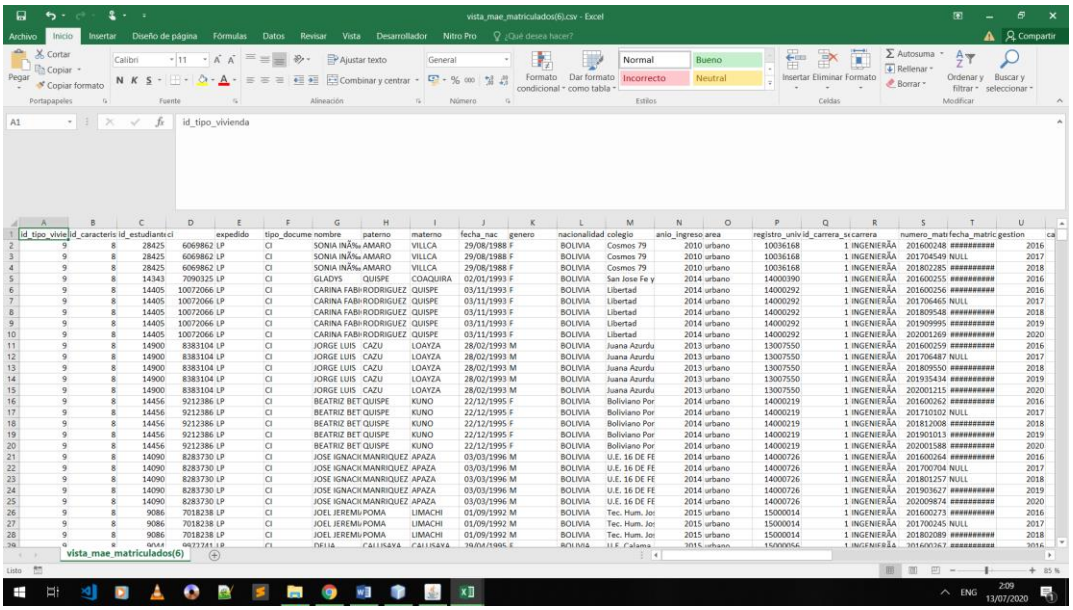


Figura 3.2 **Datos de los alumnos de la U.P.E.A**

Fuente: SIE-U.P.E.A. (2020)

3.4.2. Descripción de los datos

- **Datos Académicos**

  Dentro de los campos que se tiene en esta sección son:

  - *Gestión,* se refiere a la gestión actual de matriculación.
  - *Carrera,* se especifica a la carrera actual que pertenece el alumno.
  - *Sede,* se refiere al lugar o sede donde el alumno pasara las clases.
  - *Modalidad de ingreso,* tipo de ingreso del alumno a la carrera entre las cuales están la modalidad de cursos pre-universitarios, examen de dispensación, excelencia académica, admisión especial, cambio de carrera, carrera paralela, convenio institucional o traspaso de carrera.

- **Datos Personales**

  Se registra información sobre datos personales del alumno como ser:

  - *CI,* número correlativo del cedula de identidad del alumno.
  - *Exp.,* iniciales de acuerdo al lugar donde se expedido el CI del alumno, como ser La Paz (LP), Cochabamba(CBBA), Beni(BN), Oruro(OR), Pando(PND), Potosí(PT), Sucre(SC), Tarija(TJ) y Perú(PERU).
  - *Tipo de Doc.* tipo de documento con el cual se está registrando el alumno como ser Cedula de Identidad (CI), Documento Nacional de Identificación (DNI) o Pasaporte.
  - *Paterno,* apellido paterno del alumno.
  - *Materno,* apellido materno del alumno.
  - *Nombre(s),* es el o los nombres del alumno.
  - *Fecha Nacimiento,* fecha de nacimiento del alumno.
  - *Género,* género del alumno, M si es masculino o F si es femenino.
  - *Nacionalidad,* nacionalidad del alumno, si es de Bolivia, Perú, Argentina u otros países.
  - *Departamento de nacimiento,* departamento en el cual nació el alumno.
  - *Provincia de nacimiento,* provincia de nacimiento del alumno.

- **Dirección Ubicación Actual**

  En esta sección se especifica los datos actuales a la hora de matriculación del alumno como ser:

  - *Departamento*
  - *Provincia*
  - *Ciudad*
  - *Distrito*

- *Zona*
- *Calle*
- *Numero*

- **Datos de Egreso de Secundaria**

  En esta sección se registra los datos referidos al último año de colegio del alumno entre los cuales están:

  - *Colegio,* nombre del colegio del alumno en el último año.
  - *Año de egreso*, año en el cual el alumno curso y aprobó el último curso de secundaria o su equivalente en otro país.
  - *País colegio,* país donde se encuentra el colegio citado anteriormente.
  - *Tipo colegio,* tipo de dependencia del colegio, puede ser pública, privada, de convenio u otra.
  - *Ciudad/Localidad,* ciudad donde se encuentra dicho colegio.
  - *Área,* dentro de que área se encuentra el colegio puede ser urbana o rural.

- **Universidad que Expide el Título de Bachiller**

  Dentro de esta sección se especifica los datos sobre el título de bachiller y que institución lo avala.

  - *Universidad*
  - *Año de título de bachiller*
  - *Nro. De título de bachiller*

- **Datos Socio-Económicos**

  En esta sección se especifica alguna información sobre la situación socio-económica del alumno como ser:

  - *La vivienda que habita,* información sobre la vivienda que habita puede ser propia, propia de padres, alquiler, prestada o anticrético.

- *Característica vivienda,* puede ser vivienda, casa, departamento u otro.
- *Trabaja,* SI en caso de que el alumno trabaje o NO en caso de no trabajar.
- *Jornada laboral,* especificar el tipo de jornada laboral que desempeña puedes ser tiempo completo, medio tiempo, tiempo horario u otro.

- ***Numero de hermanos que estudia en la universidad,*** el alumno debe indicar la cantidad numérica de hermanos que estudian en la universidad sin importar la carrera.

3.5. PREPARACIÓN DE LOS DATOS

En base a los datos obtenidos se debe realizar una preparación de los datos para poder analizarlos con las herramientas de Minería de Datos.

Tabla 3.1
Tabla sobre información de los alumnos.

CAMPO	MUESTRA
id_tipo_vivienda	9
id_caracteristica_vivienda	8
id_estudiante	29****
Ci	68*****
Expedido	LP
tipo_documento	CI
Nombre	IVAN RODRIGO
Paterno	HIDALGO
Materno	MAMANI
fecha_nac	05/03/1989
Genero	M
Nacionalidad	BOLIVIA
Colegio	San Simón de Ayacucho
anio_ingreso_estudiante	2009
Área	Urbano
registro_universitario	90*****
id_carrera_sede	1
Carrera	INGENIERÍA DE SISTEMAS

numero_matricula	201604544
fecha_matriculacion	22/02/2016
Gestión	2016
carrera_id	1
area_id	1
id_sede	1
sede	VILLA ESPERANZA
direccion_departamento	LA PAZ
direccion_ciudad_localidad	EL ALTO
tipo_colegio	PUBLICO
nombre_modalidad	PRUEBA DE SUFICIENCIA ACADÉMICA
anio_ingreso_estudiante	2009
periodo_gestion	2016
anio_egreso_colegio	2007
area	Urbano
numero_hermano_upea	1
trabaja	SI
tipo_jornada_laboral	tiempo completo
colegio	San Simón de Ayacucho
tipo_colegio	PUBLICO
localidad_colegio	La Paz
id_pais_colegio	1
id_universidad	0
numero_titulo_bachiller	24633
anio_titulo_bachiller	2013
fecha_registro_estudiante	22/02/2016
nombre_caracteristica_vivienda	Casa
nombre_vivienda	Propia de padres

Nota. Elaboración propia

3.5.1. Importación a base de datos

Para poder trabajar en el proceso de limpiado mediante consultas, se procedió a subir los datos obtenidos a una base de datos.

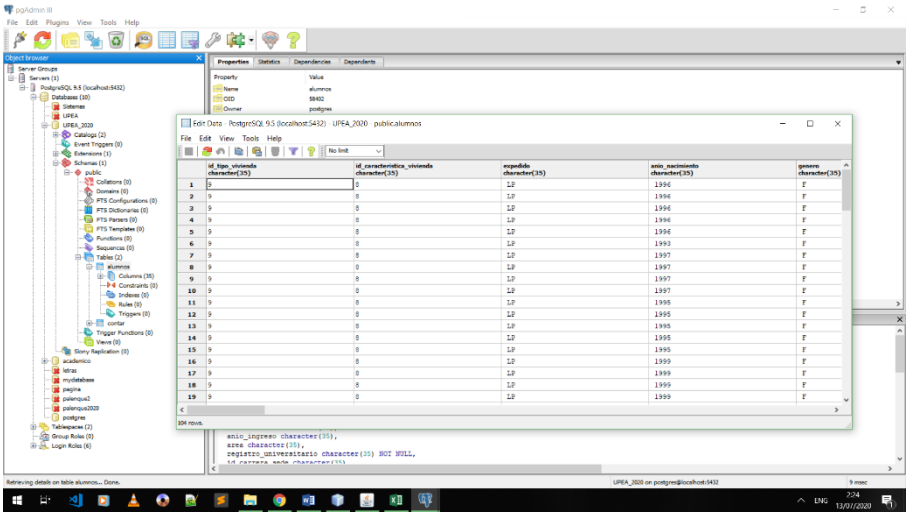


Figura 3.3 **Base de datos en PostgreSQL.**

Fuente: Elaboración propia.

3.5.2. Selección de datos

Para la aplicación de los datos obtenidos es necesario adecuarlos a las necesidades del estudio, por esta razón se procede a realizar la respectiva selección de los datos que usaremos ya que no todos son relevantes.

Tabla 3.2
Selección de campos.

CAMPO	MUESTRA
Registro Universitario	316
Modalidad Ingreso	PRUEBA DE SUFICIENCIA ACADÉMICA
Anio_Ingreso	2016
Genero	M
Nacionalidad	BOLIVIA
Tipo Colegio	PUBLICO
Área Colegio	Urbano
Tipo Vivienda	Casa
Característica Vivienda	Alquilada
Trabaja	NO
Jornada Laboral	No Trabaja
Hermanos	0

Nota. Elaboración propia.

3.5.3. Limpieza de datos

Dentro del proceso de limpieza se realizó el cambio de los caracteres que el software WEKA no los reconocería y que implicaría un problema para la extracción de patrones.

Tabla 3.3
Normalización de datos

CAMPO	MUESTRA
Registro_Universitario	9001277
Gestión	2016
Modalidad_Ingreso	CURSO PRE-UNIVERSITARIO
Anio_Ingreso	2010
Genero	F
Nacionalidad	BOLIVIA
Tipo_Colegio	PUBLICO
Area_Colegio	Urbano
Tipo_Vivienda	Casa
Caracteristica_Vivienda	Propia de padres
Trabaja	TIEMPO COMPLETO
Hermanos	1
Índice de permanencia	Desercion_Leve

Nota. Elaboración propia.

3.5.4. Transformación y estructuración

Luego de ya a ver limpiado los datos fue necesario estructurar y convertir los datos a un formato que se pueda exportar a una base de datos para su mejor tratamiento ya que la cantidad de datos es muy grande. Para trabajar con WEKA, se debe tener en cuenta que este utiliza archivos con formato arff[13].

[13] arff, acrónimo de Attribute-Relation File Format.

```
----------prueba.arff----------
--------
1 % Archivo de prueba para Weka.
2 @relation prueba
3
4 @attribute nombre STRING
5 @attribute ojo_izquierdo {Bien, mal}
6 @attribute dimension NUMERIC
7 @attribute fecha_analisis DATE "dd-MM-yyyy HH:mm"
8
9 @data
10 Antonio, bien,38.43,"12-04-2003 12:23"
11 'Maria José',?,34.53,"14-05-2003 13:45"
12 Juan, bien,43,"01-01-2004 08:04"
13 Maria,?,?,"03-04-2003 11:03"
```

Figura 3.4 **Ejemplo de archivo arff.**

Fuente: Viscaino (2008).

Este formato está compuesto por una estructura claramente diferenciada en tres partes, la cabecera, declaraciones de atributos y sección de datos.

```
@relation 122

@attribute Genero {M,F}
@attribute Nacionalidad {NACIONAL,EXTRANGERO}
@attribute Tipo_Colegio {PUBLICO,PRIVADO,CONVENIO}
@attribute Area_Colegio {urbano,rural}
@attribute Tipo_Vivienda {Casa,Departamento,Habitacion,Otro}
@attribute Caracteristica_Vivienda {Alquilada,'Propia de padres',Prestada,Anticretico,Adjudicada}
@attribute Jornada_Laboral {'No Trabaja','medio tiempo','tiempo completo','tiempo horario'}
@attribute Hermanos numeric
@attribute Ind_Permanencia {Permanencia,Desercion,Desercion_Relativa}

@data
M,NACIONAL,PUBLICO,urbano,Casa,Alquilada,'No Trabaja',0,Permanencia
M,NACIONAL,PRIVADO,urbano,Departamento,'Propia de padres','medio tiempo',0,Permanencia
M,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','No Trabaja',0,Desercion
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',3,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','tiempo completo',0,Permanencia
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','No Trabaja',0,Desercion_Relativa
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Desercion
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','medio tiempo',0,Desercion
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',2,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Permanencia
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',1,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',1,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',1,Permanencia
M,NACIONAL,CONVENIO,urbano,Departamento,Alquilada,'tiempo completo',0,Desercion_Relativa
M,NACIONAL,PRIVADO,urbano,Casa,Alquilada,'medio tiempo',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Habitacion,Alquilada,'No Trabaja',0,Permanencia
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','tiempo completo',0,Permanencia
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',2,Desercion
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo horario',2,Permanencia
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Habitacion,Alquilada,'No Trabaja',2,Permanencia
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',0,Desercion_Relativa
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',1,Desercion
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo completo',2,Desercion_Relativa
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Desercion_Relativa
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','tiempo horario',0,Desercion
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
M,NACIONAL,PRIVADO,urbano,Habitacion,'Propia de padres','medio tiempo',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Habitacion,'Propia de padres','medio tiempo',0,Permanencia
F,NACIONAL,PRIVADO,urbano,Casa,'Propia de padres','No Trabaja',0,Desercion_Relativa
F,NACIONAL,PUBLICO,urbano,Departamento,'Propia de padres','tiempo completo',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Desercion_Relativa
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Permanencia
F,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','No Trabaja',0,Desercion_Relativa
M,NACIONAL,PUBLICO,urbano,Casa,'Propia de padres','medio tiempo',0,Desercion_Relativa
M,NACIONAL,PRIVADO,urbano,Casa,Alquilada,'No Trabaja',0,Desercion_Relativa
F,NACIONAL,PUBLICO,urbano,Casa,Alquilada,'tiempo completo',0,Desercion_Relativa
```

Figura 3.5 **Documento en formato arff.**

Fuente: Elaboracion propia.

3.6. MODELADO

La Minería de Datos es la búsqueda de patrones en bases de datos mediante la aplicación de algoritmos, los datos resultantes son estadísticamente factibles y entendibles para el ser humano.

3.6.1. Técnicas de modelado

En la presente investigación se utilizó algoritmos de clasificación. Un árbol de decisión es un conjunto de reglas organizadas jerárquicamente de tal modo que la decisión final se determina pro el cumplimiento de las condiciones desde la raíz hasta algunas de sus hojas. Las reglas de un árbol de decisión se las describe mediante atributos, las que pueden ser objetivas o situaciones a partir de la cual se devuelven respuestas, que llegan a ser las decisiones.

Los valores de entrada pueden ser discretos o continuos y en los primeros son simples. Un árbol de decisión suele tener un nodo interno, un nodo de probabilidad, un nodo hoja y las ramas que brindan las posibles rutas respecto a as decisión que se toma.

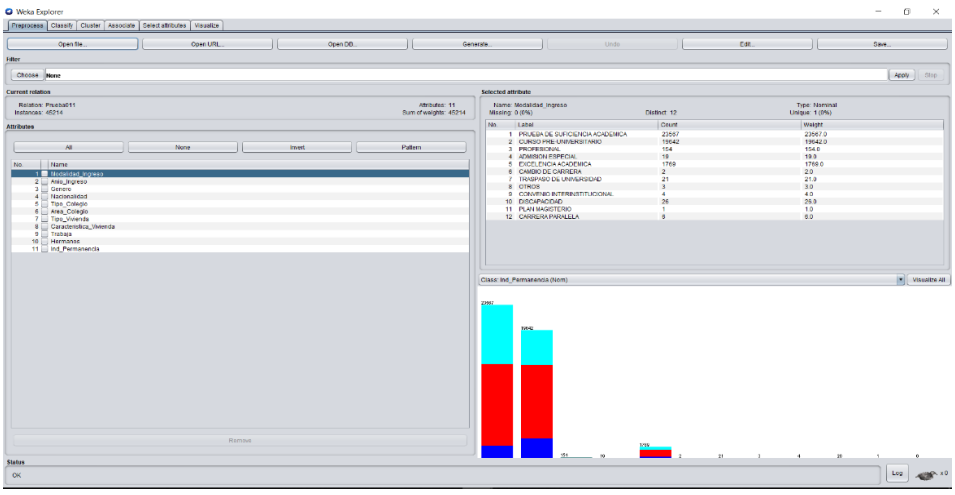


Figura 3.6 **Introducción de datos.**

Fuente: Elaboracion propia.

Los algoritmos basados en Reglas, son una alternativa popular de los árboles de decisión. El antecedente o predicción de una regla es una serie de pruebas como las que se hacen en el nodo en árboles de decisión. El consecuente o conclusión da la clase o clases que aplica a instancias cubiertas por esa regla o tal vez da una probabilidad de distribución acerca de las clases.

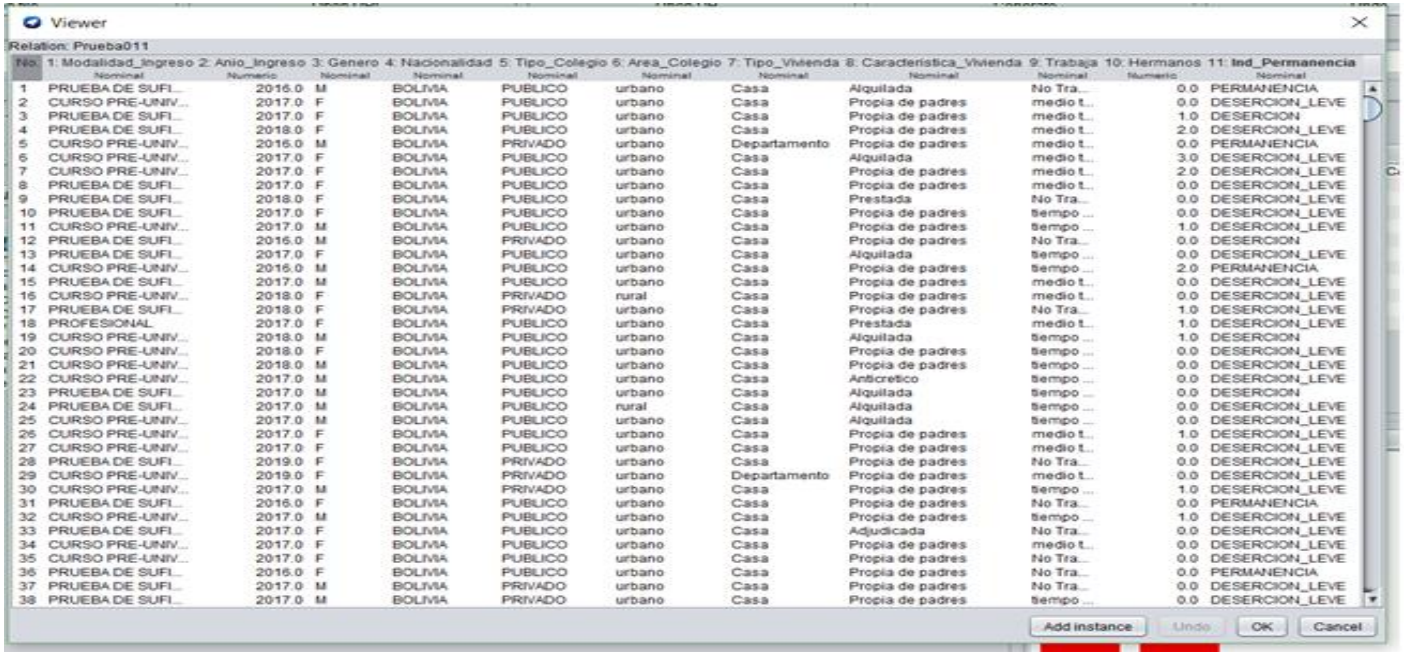


Figura 3.7 **Datos en WEKA.**

Fuente: Elaboración propia.

3.6.2. Pruebas en diferentes algoritmos

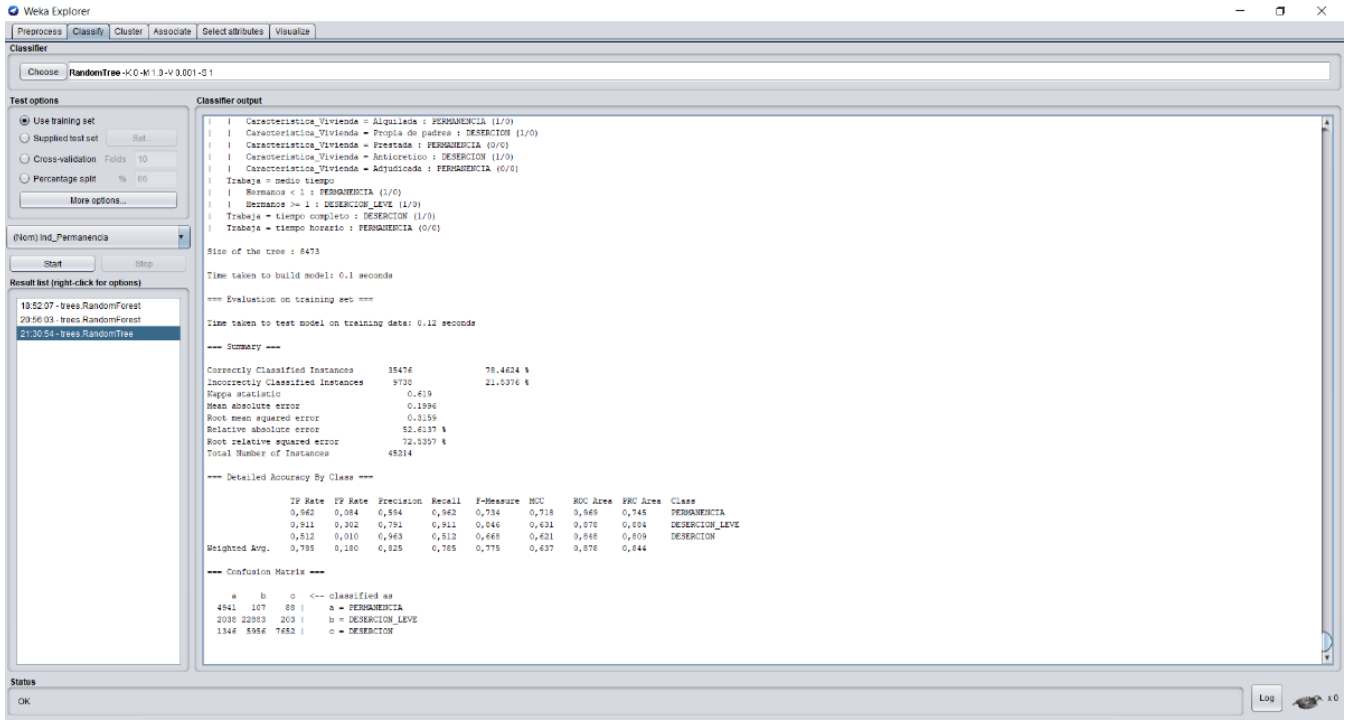


Figura 3.8 **Clasificación mediante algoritmo Random Tree**

Fuente: Elaboracion propia.

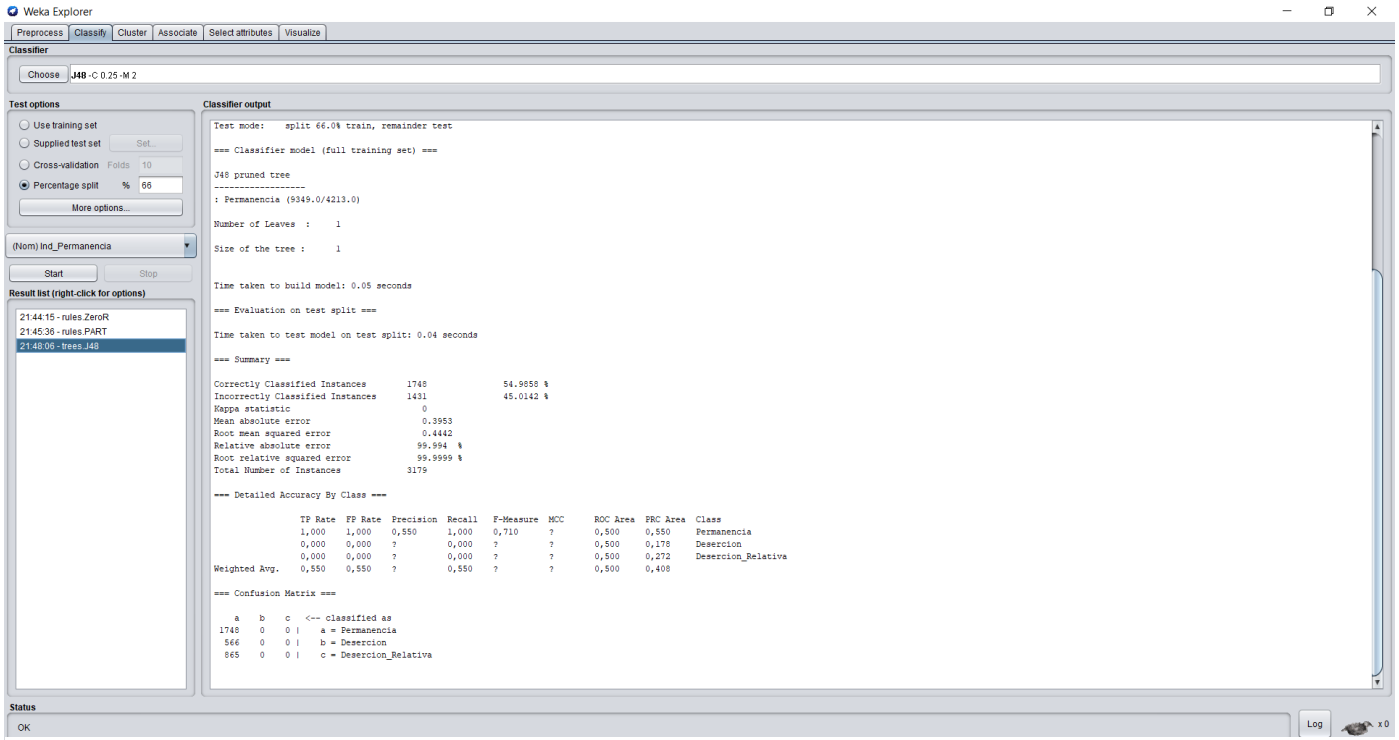


Figura 3.9 **Clasificación mediante algoritmo J48**

Fuente: Elaboracion propia

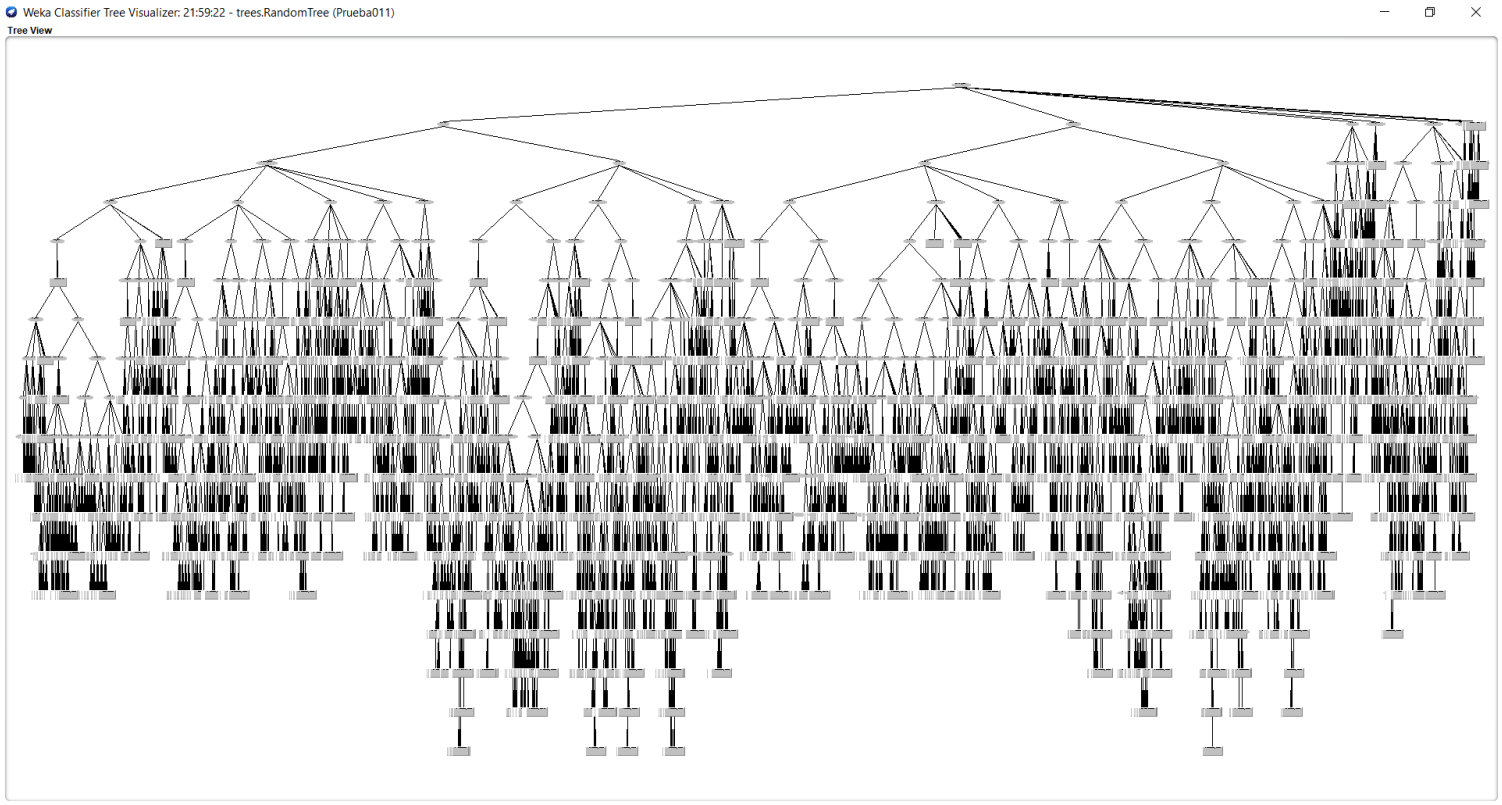


Figura 3.10 **Árbol completo generado por WEKA**

Fuente: Elaboracion propia

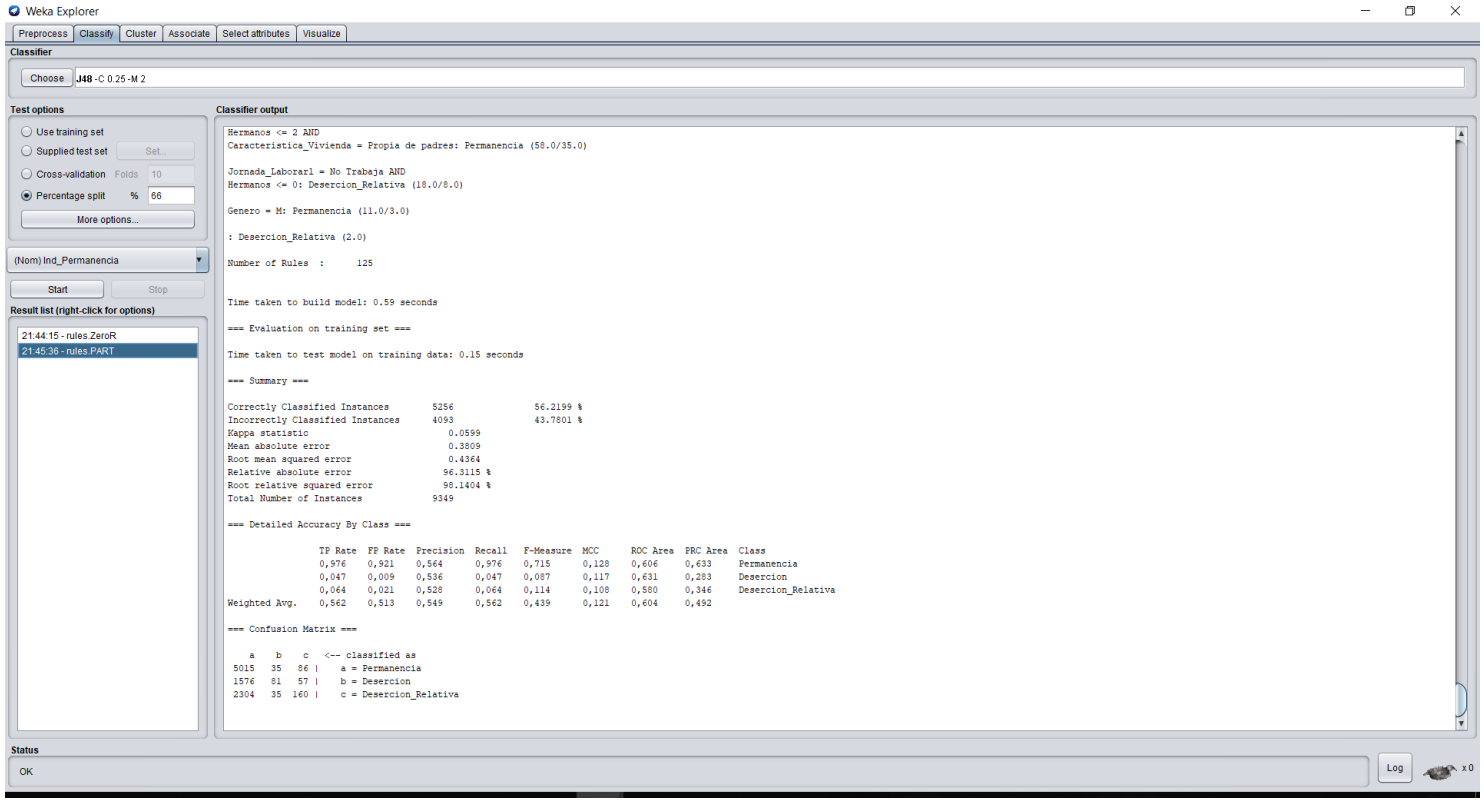


Figura 3.12 **Clasificación mediante algoritmo PART**

Fuente: Elaboracion propia

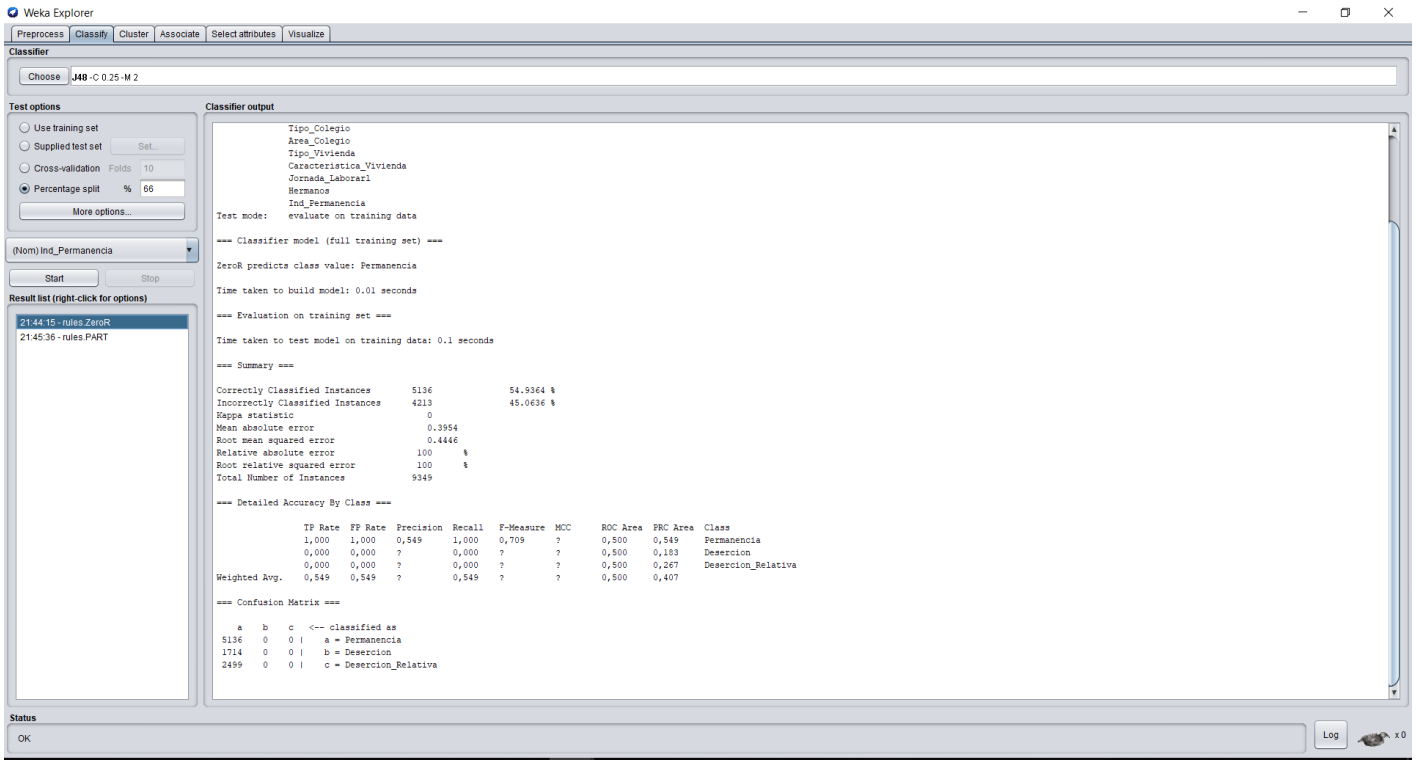


Figura 3.13 **Clasificación mediante algoritmo ZeroR**

Fuente: Elaboracion propia

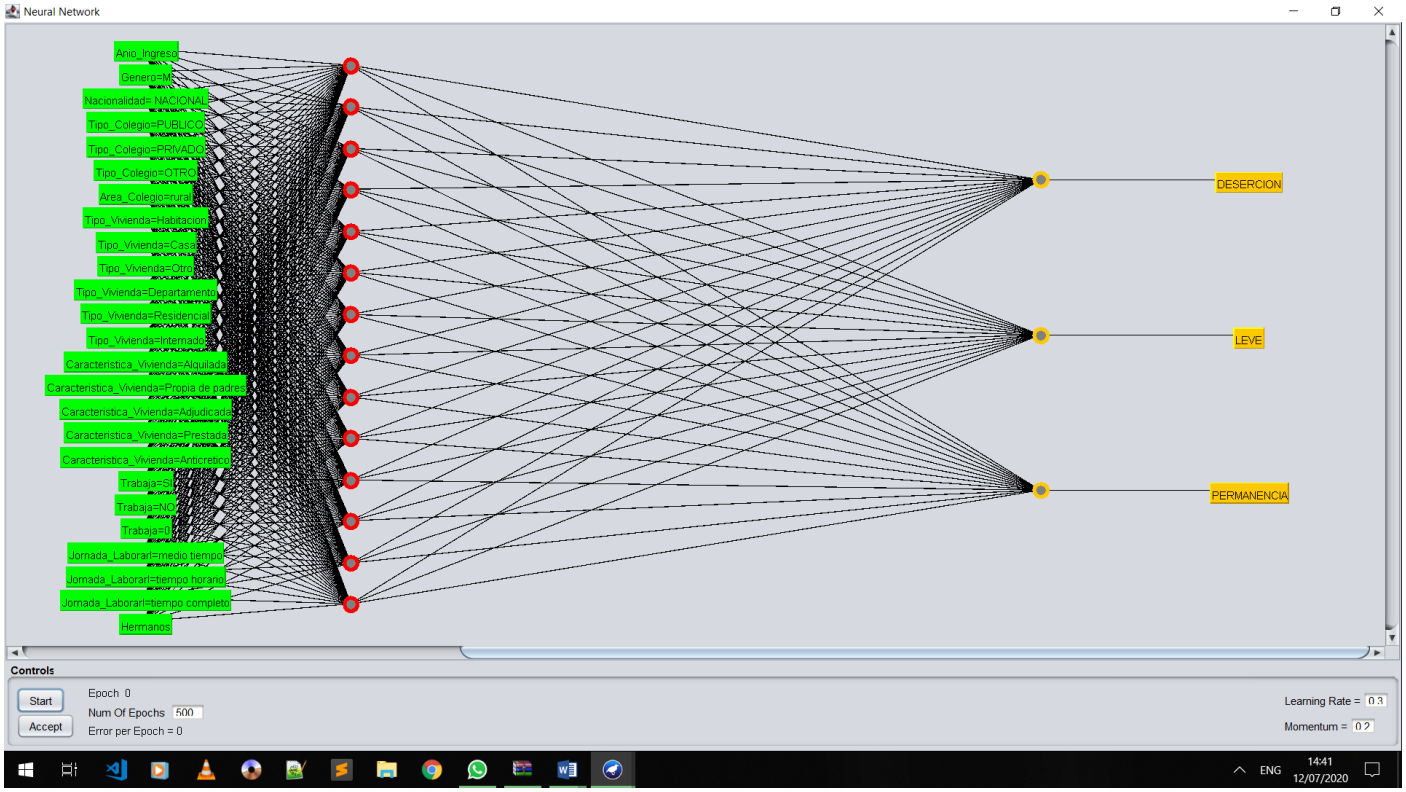


Figura 3.14 **Algoritmo Multiperceptron**

Fuente: Elaboracion propia

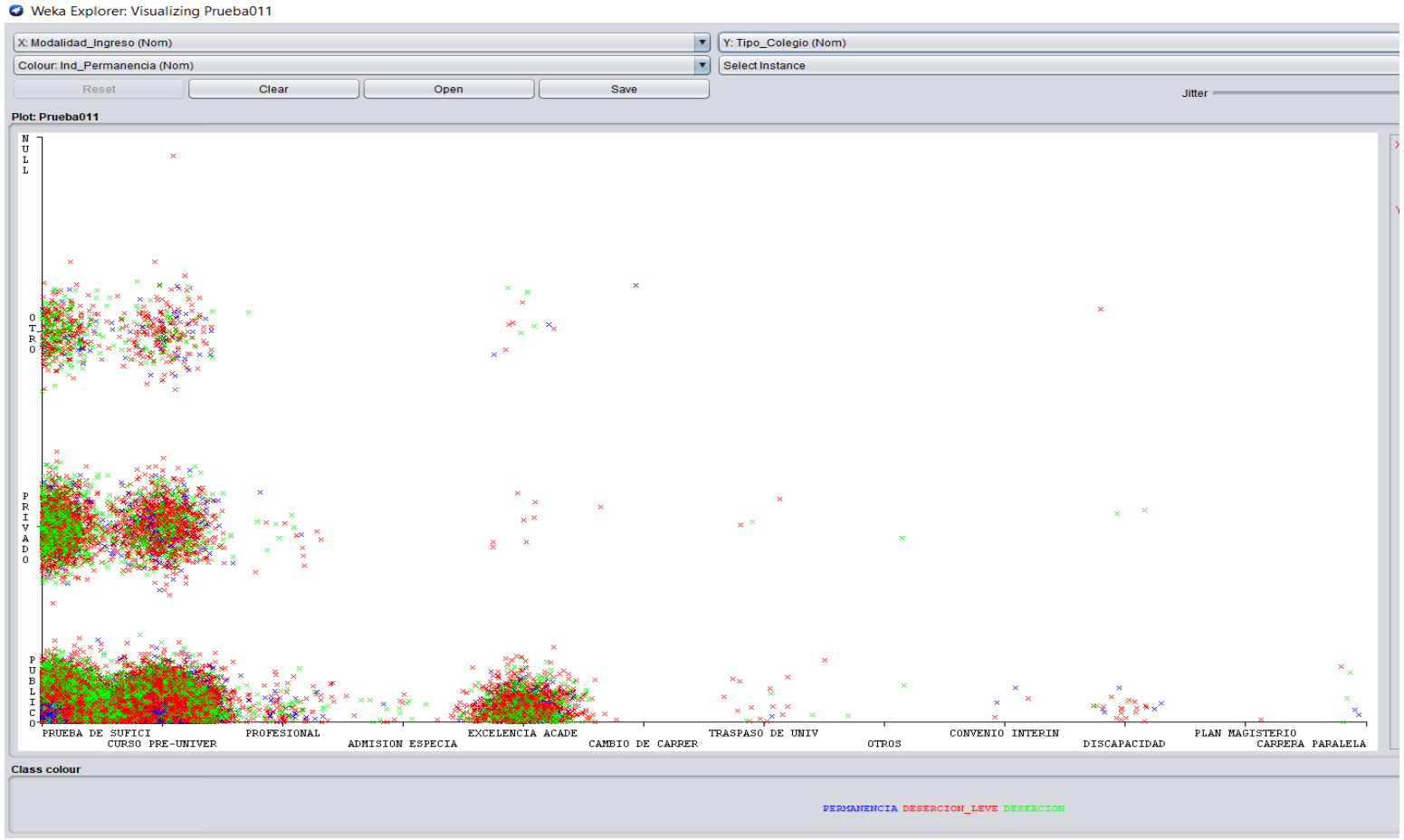


Figura 3.15 **Modalidad ingreso y tipo de colegio**

Fuente: Elaboracion propia

3.7. EVALUACIÓN

Luego de la aplicación de los diferentes algoritmos de árbol clasificación de Minería de Datos se obtuvo los resultados mostrados en la tabla 3.4 en el que se aprecia la comparación de los diferentes algoritmos. Analizando la tabla se puedo determinar que luego de entrenar con diferentes tipos de algoritmos de clasificación que tiene el software Weka, se pudo discriminar como mejor algoritmo para la predicción en base a Minería de Datos aquellos que mostraron el menor porcentaje de error absoluto. Estos algoritmos serán implementados en el modelo de predicción.

Tabla 3.4
Resumen de algoritmo RandomTree

Correctly Instances	Classified	35476
Incorrectly Instances	Classified	9738
Kappa statistic		0.619
Mean absolute error		0.1996
Root mean squared error		0.3159
Relative absolute error		52.6137%
Root relative squared error		72.5357%
Total Number of Instances		45214

Fuente: Elaboración propia

Tabla 3.5
Matriz de confusión de RandomTree

a	b	C	clasificación
4941	107	88	**a=NO DESERCION**
2038	22883	203	**b=DESERCION LEVE**
1346	5956	7652	**c=DESERCION**

Nota: Elaboración propia

Tabla 3.6
Resumen de algoritmo PART

Number of Rules	6
Correctly Classified Instances	46
Incorrectly	7
Kappa statistic	0.8017
Mean absolute error	0.1394
Root mean squared error	0.264
Relative absolute error	31.8988 %
Root relative squared error	56.5045 %
Total Number of Instances	53
Ignored Class Unknown Instances	9297

Nota: Elaboración propia

Tabla 3.7
Matriz de confusión PART

a	b	c	clasificación
13	1	0	**a=NO DESERCION**
3	16	3	**b=DESERCION LEVE**
0	0	17	**c=DESERCION**

Nota: Elaboración propia

Tabla 3.8
Comparación de resultados de algoritmos

ALGORITMO	INSTANCIAS CLASIFICADAS CORRECTAS	INSTANCIAS MAL CLASIFICADAS	ESTADISTICA KAPA	ERROR ABSOLUTO
J48	84,90%	15,09%	0,7664	36,60%
MULTILAYER PERCEPTRON	92,45%	7,54%	0,8852	17,59%
IBK	78,40%	21,53%	0,619	21,35%
PART	86,79%	13,20%	0,8017	31,89%
ZeroR	41,50%	58,49%	0,3793	100%
DECISION TABLE	81,13%	18,86%	0,7203	59,29%
LMT	84,94%	15,09%	0,7664	35,75%
REPTree	76,52%	23,47%	0,5855	52,86%
RANDOM FOREST	92,452%	7,547%	0,8852	25,42%
RANDOM TREE	92,45%	7,54%	0,8862	14,74%

Nota. Elaboración propia

3.8. DESARROLLO DEL MODELADO EN BASE A RUP

Es a partir de este punto que se aplican los fundamentos teóricos, para el desarrollo del prototipo vistos en el capítulo anterior.

3.8.1. Fase de análisis y casos de uso

3.8.1.1. Análisis de requerimiento

- Requerimiento relacionado con el contenido, el modelo de índices de deserción PREDESMIN, debe almacenar los datos referidos a la información del registro de alumnos.
- Sobre los requisitos relacionados a la estructura, el modelo de predicción tiene la estructura básica de tal modo que contiene una breve explicación sobre su aplicación en la página principal.
- Requisitos relacionados con los usuarios, el modelo cuenta con un tipo de orientación hacia las personas con conocimiento básicos en el uso de máquinas de computación.

3.8.1.2. Casos de uso

- **Identificar los actores**

  Se clasifica como actores a:

  - Usuario, que navega y hace uso del módulo para el análisis y visualización de los resultados.

- **Identificación de actividades del actor**

Tabla 3.9
Casos de uso

Caso de Uso	**Llenado de la base de datos**
Actores	Diferentes usuarios
Flujo de eventos	✓ Selecciona archivo ✓ Sube datos ✓ Selecciona ✓ Aplicación de Minería de Datos

Nota: Elaboración propia

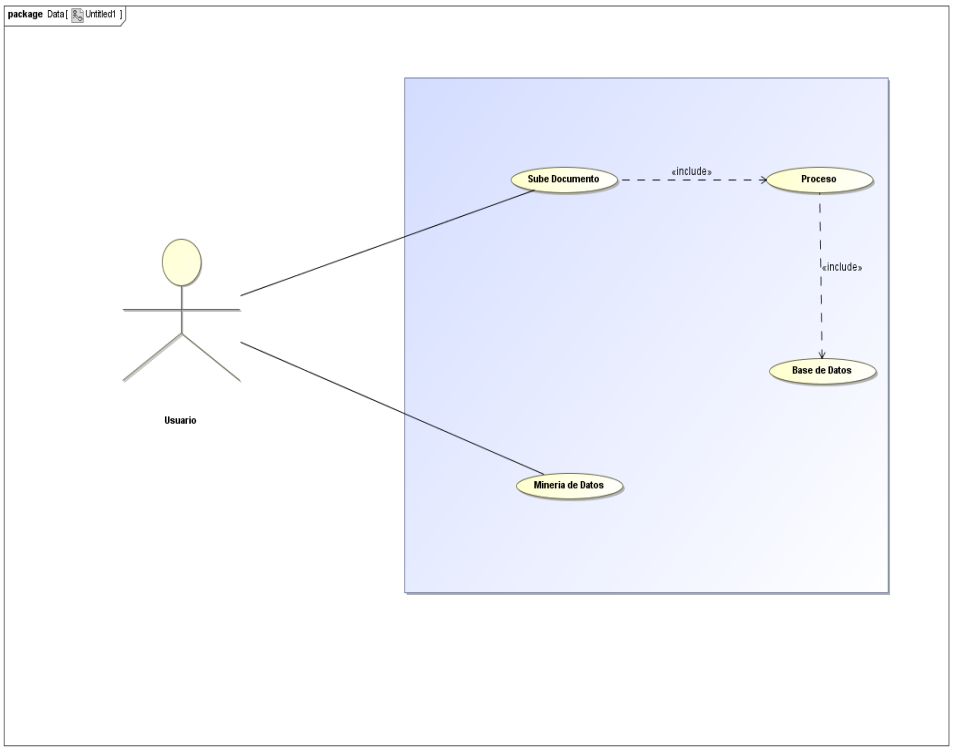


Figura 3.16 **Modelo de casos de uso**

Fuente: Elaboracion propia

Tabla 3.10
Descripción de caso de uso PREDESMIN

Caso de Uso	**Uso del módulo PREDESMIN**
Actores	Diferentes usuarios
Flujo de eventos	✓ Selecciona archivo ✓ Selecciona algoritmo ✓ Aplica Minería de Datos ✓ Obtiene resultados de Minería de Datos

Nota: Elaboración propia

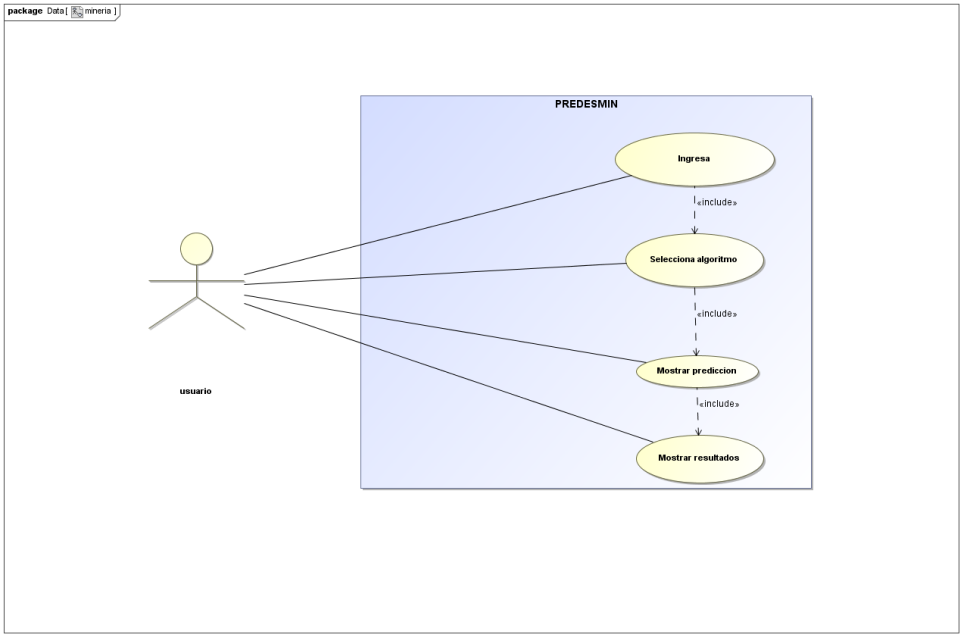


Figura 3.17 **caso de uso minería de datos**

Fuente: Elaboracion propia

3.8.2. Modelo conceptual

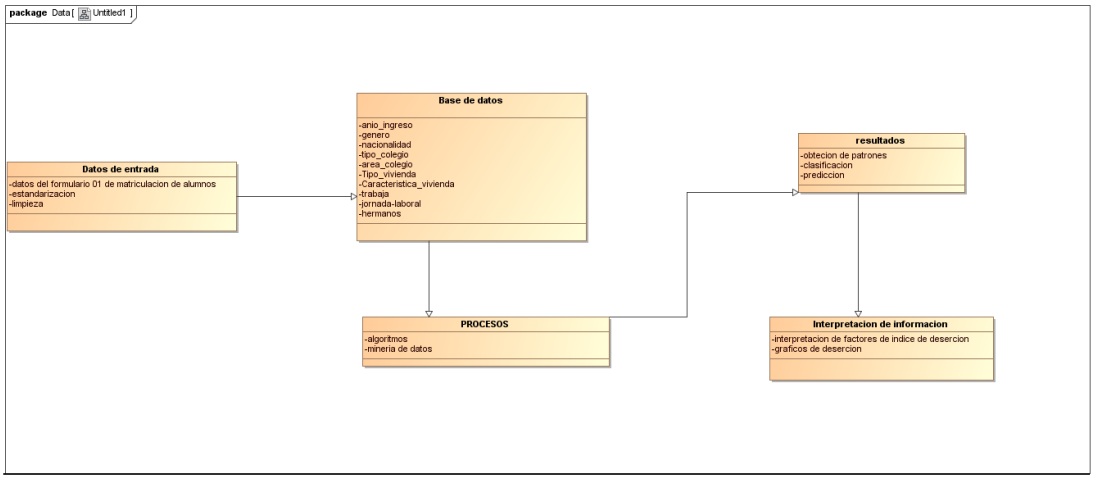


Figura 3.18 **Modelo conceptual**

Fuente: Elaboración propia

.

3.8.3. Modelo de presentación

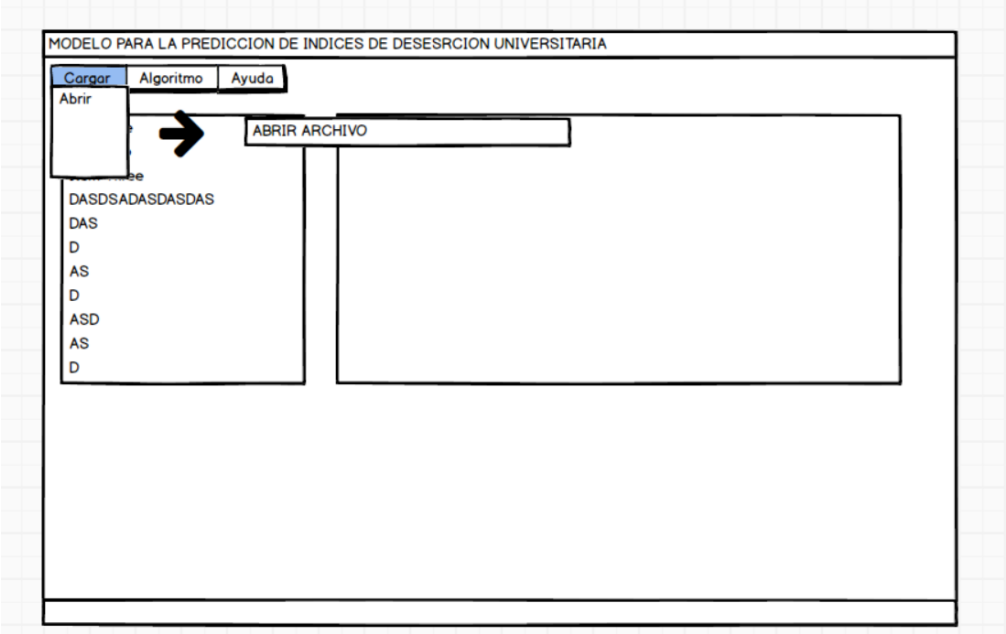


Figura 3.19 **Modelo de presentación, cargar archivo**

Fuente: Elaboracion propia

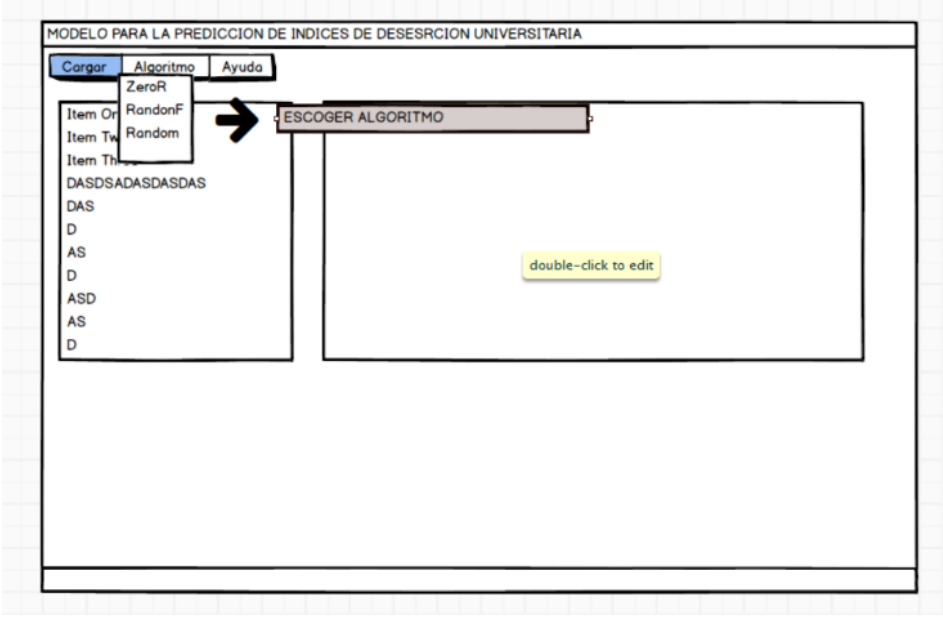


Figura 3.20 **Modelo de presentación, ejecución de búsqueda**

Fuente: Elaboracion propia

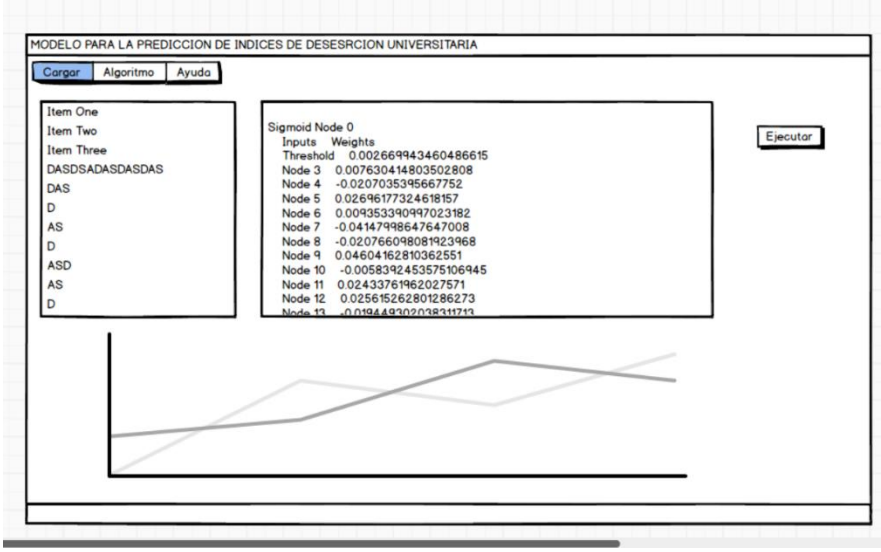


Figura 3.21 **Modelo de presentación, seleccionar algoritmo**

Fuente: Elaboracion propia

3.9. ARQUITECTURA

Tabla 3.11
Especificaciones de Hardware

Especificaciones de Hardware	
Memoria RAM	4GB
Procesador	Intel Core i5
Velocidad	2.1Ghz.
Disco duro	1TB

Fuente: Elaboración propia

Tabla 3.12
Especificaciones de Software

Especificaciones de Software	
Sistema operativo	Windows 10
Minería de Datos	WEKA 3.9.4
Gestor de base de datos	PostgreSQL
Framework java	NeatBeans 7.4
Java jdk	Openjdk-7.4-jdk
Java jre	Openjde-7.4-jde

Fuente: Elaboración propia

3.10. IMPLEMENTACIÓN DEL MODELO

3.10.1. Etapas de funcionamiento del modelo

Las etapas del modelo de funcionamiento son: lectura de archivo, cargar archivo, seleccionar algoritmo y visualización de resultados.

- **Lectura**

  Para la lectura optima del archivo, este debe tener un formato con extensión “arff”, el cual será el archivo de lectura.

- **Cargar archivo**

  El archivo de formato “arff” se carga en el prototipo, este se visualiza en “DATOS”.

- **Selección de algoritmo**

  El modelo cuenta con algoritmos de Minería de Datos, de los cuales uno debe ser seleccionado para la aplicación en los datos cargados previamente.

- **Visualización de resultados**

  El modelo PREDESMIN, nos muestra los resultados a los que arriba, luego de haber sido sometido a los algoritmos que la Minería de Datos nos provee el prototipo.

Figura 3.22 Funcionamiento del Modelo

Fuente: Elaboracion propia

3.10.2. Creación del formulario principal

Para la creación del prototipo se utiliza Neatbeans 7.4

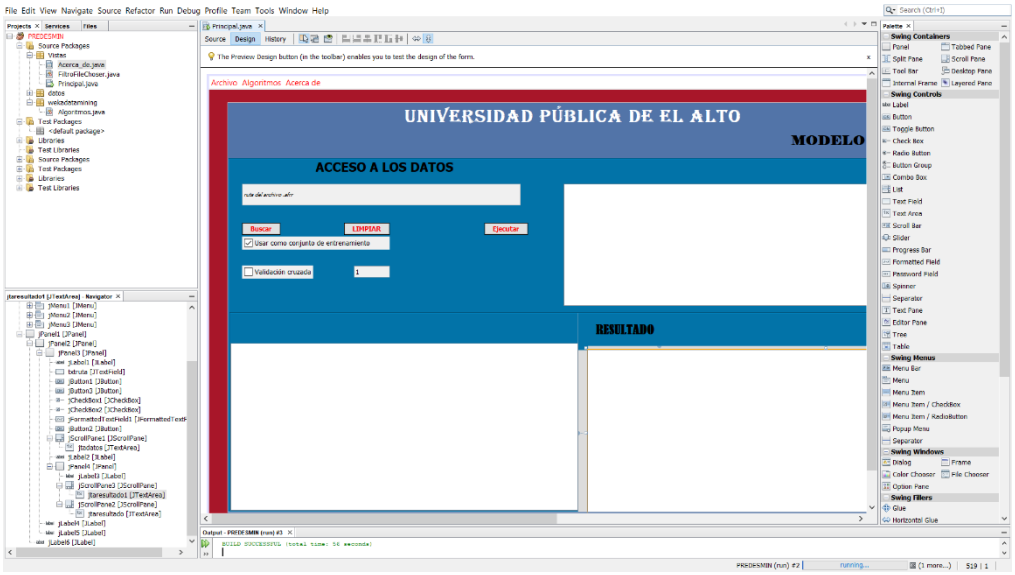


Figura 3.23 **Formulario principal PREDESMIN**

Fuente: Elaboracion propia

3.10.3. Implementación de algoritmos

```
//algoritmo ZeroR
    public String ejecutarZeroR() {
        resultado = "\n";
        ZeroR zeror = new ZeroR();
        try {
            zeror.buildClassifier(coleccion);
            String matrix = matrixConfusion(zeror);
            resultado = resultado + zeror.toString();
            resultado = resultado + matrix;
```

```
        } catch (Exception exception) {
            resultado = exception.toString();
            mostrarMsj();
        }
        return resultado;
    }
//algoritmo Part
    public String ejecutarPart() {
        resultado = "\n";
        PART part = new PART();
        try {
            part.buildClassifier(coleccion);
            String matrix = matrixConfusion(part);
            resultado = resultado + part.toString();
            resultado = resultado + matrix;
        } catch (Exception exception) {
            resultado = exception.toString();
            mostrarMsj();
        }
        return resultado;
    }
```

3.10.4. Compilación

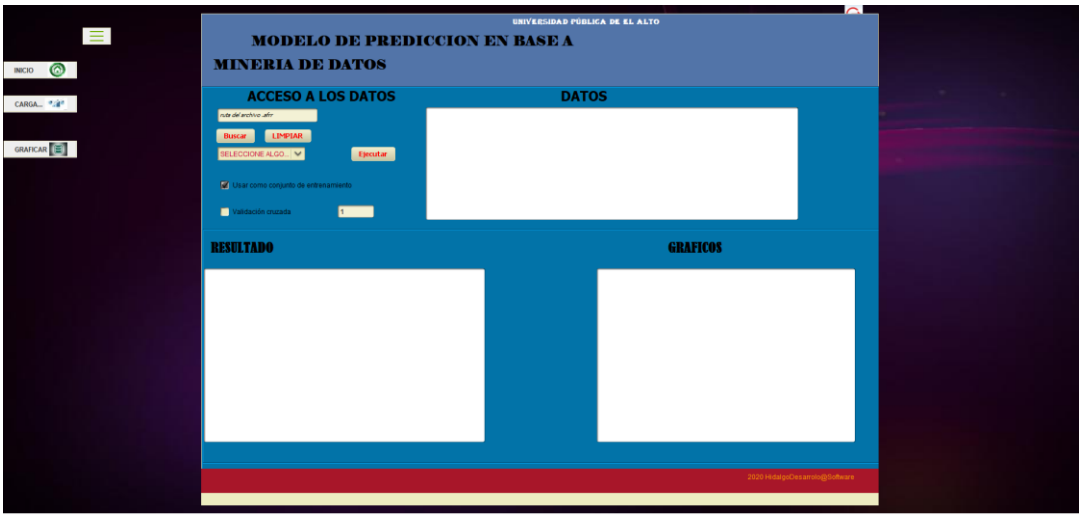


Figura 3.24 **Compilación del proyecto**

Fuente: Elaboracion propia

3.10.5. Resultados

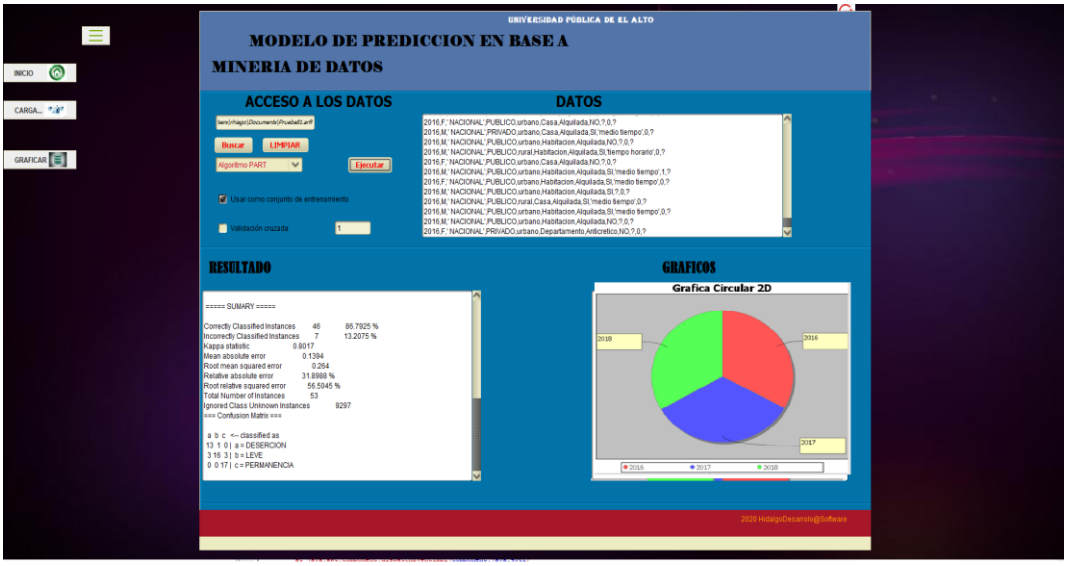


Figura 3.25 **Visualización de los datos**

Fuente: Elaboracion propia

Los resultados mostrados por el algoritmo de Minería de Datos son los que muestran un mayor porcentaje de instancias clasificadas correctamente en contraposición de aquellas mal clasificadas tienden a 0%, el índice Kapa tiene el valor de 1, que indica la concordancia perfecta, asimismo podemos ver que el error absoluto relativo es relativamente bajo. Por otra parte, la matriz de confusión resultante muestra una correcta clasificación entre atributos de los factores de índices de deserción.

3.11. MÉTRICA DE CALIDAD

Con el fin de tener un producto de calidad, la métrica ISO/IEC 9126 de calidad de software, proporciona una clasificación sobre la calidad en un conjunto estructurado.

3.11.1. Funcionalidad

Las funcionalidades son aquellas que satisfacen las necesidades implícitas o explicitas. A continuación, se muestra la ponderación de las características funcionales.

Tabla 3.13
Ponderación de la funcionalidad

Característica	Ponderación
Adecuación	90%
Exactitud	90%
Conformidad	90%
Cumplimiento funcional	90%
Promedio	90%

Fuente: Elaboración propia.

Por tanto, se deduce que el prototipo tiene una funcionalidad del 90%.

3.11.2. Confiabilidad

Los Atributos relacionados con la capacidad del software dado para mantener su usabilidad bajo condiciones establecidas en un tiempo establecido.

$$\text{Confiabilidad} = 1 - (5/4679) \times 100$$

$$\text{Confiabilidad} = 99\ \%$$

El sistema tiene una confiabilidad del 99%.

3.11.3. Usabilidad

Los atributos y su conjunto están relacionados con el esfuerzo necesario para su uso, y en la valoración individual de tal uso.

Tabla 3.14
Ponderación de métricas internas usabilidad

Característica	Métrica interna	Puntaje
Interfaz de usuario amigable	I1: Interfaz de datos amigable	90
	I2: Interfaz de gráficos amigable	90
Comprensión	C1: Comprensión de datos	85
	C2: Comprensión de gráficos	90
Operatividad	O1: Correcta operacionalidad de la interfaz	95
	O2: Correcta operacionalidad de visualización de datos	95
	O3: Correcta operacionalidad de los gráficos	95
Atractividad	A1: Atractividad de la interfaz	90
	A2: Atractividad de los gráficos	90
	A3: Atractividad de la visualización de los datos	85

Fuente: elaboración propia.

Tabla 3.15
Totales de métricas internas usabilidad

Métrica	Puntaje promedio
Interfaz de usuario amigable (I)	90
Comprensión (C)	87.5
Operatividad (O)	95
Atractividad (A)	88.3

Fuente: elaboración propia

Con los datos obtenidos en la tabla 3.15 se aplica en la fórmula:

$$\text{Usabilidad} = \Sigma(xi/n)$$

$$\text{Usabilidad} = 360.8/4$$

$$\text{Usabilidad} = 90\%$$

3.11.4. Eficiencia

Para pode obtener el cálculo de la eficiencia del sistema se consideró ponderar las características esenciales que el sistema desempeña.

Tabla 3.16
Evaluación de desempeño

Característica de desempeño	Ponderación
Rapidez de inicio	4
Rapidez de proceso	5
Proceso rápido de búsqueda	5
Fluidez	5
Disponibilidad	4

Fuente: Elaboración propia.

En base a los datos de la anterior tabla se podría llegar a tener una idea de la eficiencia, para ello se utilizó la siguiente formula:

$$Eficiencia = \Sigma x_i / n * 100/n$$

$$Eficiencia = 23/5 * 100/5$$

$$Eficiencia = 92\%$$

3.11.5. Mantenibilidad

Son el conjunto de atributos los cuales nos da la opción de poder corregir, aumentar o modificar los errores del software, dicho resultado se obtiene mediante la siguiente formula:

$$Mantenibilidad = (Mt – (Fc + Fa + Fd))/Mt$$

Donde:

Mt = número de módulos en la versión actual.

Fc = número de módulos en la versión actual que han cambiado.

Fa = número de módulos en la versión actual añadido.

Fd = número de módulos en la versión anterior que se ha borrado.

Entonces:

$$Mt = 1; Fc = 1; Fa = 0; Fd = 0$$

$$Mantenibilidad = (3 – (0+0+0))/3$$

$$Mantenibilidad = 1$$

$$Mantenibilidad = 100\%$$

3.11.6. Portabilidad

Es la capacidad que tiene el software para ser trasladado de un entorno a otro. Se lo calcula mediante la fórmula:

$$Portabilidad = 1 – (ndpm/ndim)$$

Donde:

ndpm = número de días para portar el modelo.

días. ndim = número de días para implementar el modelo

$$Portabilidad = 1 – (1/6)$$

$$Portabilidad = 0.83*100$$

$$Portabilidad = 83\%$$

3.11.7. Resultados

Calculando de manera independiente cada uno de los factores en cuanto a las características de la norma ISO 9126, estos resultados nos sirven para poder realizar los cálculos.

Tabla 3.17
Análisis global de calidad

N°	Característica	Resultado
1	Funcionabilidad	90%
2	Confiabilidad	99%
3	Usabilidad	90%
4	Eficiencia	92%
5	Mantenibilidad	100%
6	Portabilidad	80%
	Evaluación de la Calidad Global	91%

Fuente: Elaboración propia.

Según Pressman dice que el resultado de la evaluación de una métrica o modelo si supera el 65% es aceptado. Por lo que el 91% encontrado en la medición es aceptable para el modelo.

3.12. EVALUACIÓN DE COSTOS

Con ayuda de COCOMO II, en este punto se estima el costo de producción del software desarrollado.

3.12.1. Puntos de función

La estimación por puntos de función está en la medida de la funcionalidad del sistema de información y un conjunto de factores individuales del sistema. los puntos de función son estimadores que puede ser de utilidad en las etapas iniciales del modelo. La medida de puntos de función está cuantificada en base a diferentes funcionalidades. La tabla siguiente describe los componentes relacionados con su complejidad asignada a cada uno de los factores que se deben considerar para la estimación del modelo.

Tabla 3.18
Puntos de función no ajustado.

Tipo de Parámetros	Cantidad	Factor de ponderación	Total
Entrada	3	5	15
Salida	3	6	18
Archivos	3	5	15
Consultas	3	6	18
Interfaces	3	5	15
Total puntos de función no ajustado			81

Fuente: Elaboración propia.

Según el estimado de interfaces de la tabla anterior, se procede a clasificarlos según su complejidad y luego multiplicar por los pesos establecidos de acuerdo a COCOMO II, para estimar los puntos función ajustados.

En la siguiente tabla se muestran los 14 factores de ajuste donde se pondera con un puntaje que se encuentra entre 0 y 5.

Tabla 3.19
Ponderación de ajuste ce complejidad

Nº de Factor	Factor	Valor 0 - 5
1	Mecanismo de recuperación	3
2	Comunicación de datos	5
3	Rendimiento	5
4	Configuración usada rigurosamente	2
5	Entrada de datos en línea	1
6	Factibilidad operativa	4
7	Actualización en línea	1
8	Interfaces complejas	3
9	Proceso interno complejo	4
10	Reusabilidad de código	4
11	Fácil instalación	5
12	Instalaciones múltiples	3
13	Facilidad de cambios	3
14	Funciones de proceso distribuido	1
	∑ Fi	44

Fuente: Elaboración propia.

Con el promedio encontrado, se reemplaza los datos e la fórmula de punto de función ajustado.

$$PFA = \text{cuenta total} * (0.65 + 0.01 * \sum \boldsymbol{Fi})$$

$$PFA = 81 * (0.65 + 0.01 * 44)$$

$$PFA = 88.29$$

3.12.2. Aplicación de COCOMO II

Para el desarrollo del sistema se deben considerar diversas plataformas que soporten el trabajo, así mismo el lenguaje que será utilizado como también el administrador de base de datos que soportará el sistema.

Para poder calcular las líneas de código, utilizamos el valor del punto de función ajustado, de igual forma utilizaremos el valor de Factor de línea de código del lenguaje de programación utilizada para el desarrollo.

Tabla 3.20
Factor LCD/PF de lenguaje de programación.

Lenguaje	Nivel	Factor LCD/PF
C	2.5	128
ANSI/basic	5	64
Java	6	53
PL/I	4	80
Visual Basic	7	46
ASP	9	36
PHP	11	29
Visual C++	9.5	34

Fuente: Pressman, (2002).

Reemplazamos los datos en la fórmula para calcular las líneas de código:

$$LDC = PFA * Factor\ LDC/PF$$

$$LDC = 88.29 * 53$$

$$LDC = 4679.37$$

$$KLDC = 4679.37/1000$$

$$KLDC = 4.67$$

Tomando en cuenta lo visto en el capítulo dos, el proyecto se enmarca en el modelo básico. De acuerdo a la cantidad de líneas de código el modelo pertenece a un modo orgánico por lo que los valores para a y b serán 2.40 y 1.05 respectivamente según la tabla 2.6.

Las ecuaciones de COCOMO II brinda la siguiente forma:

- **Estimación de esfuerzo de desarrollo**

$$E = a*(KLCD)^b$$

El esfuerzo se estima:

$$E = 2.40*(4.67)^{1.05}$$

$$E = 12.10\ (personas/mes)$$

Es decir que se requiere el esfuerzo de 12 personas trabajando en el desarrollo del sistema.

- **Estimación del tiempo de desarrollo**

Para el cálculo del tiempo se utiliza nuevamente la expresión de COCOMO II para la determinación del tiempo estimado del proyecto.

$$\mathbf{T = c * (E)^d}$$

Donde c y de son constantes que de acuerdo al modo orgánico establecido por COCOMO estos valores son 2.50 y 0.38 respectivamente.

Reemplazando estos valores en la formula se tiene el cálculo del tiempo expresado en meses.

$$T = 2.5 * (12)^{0.38}$$

$$T = 6.42 \text{ (meses)}$$

El tiempo estimado de trabajo es de aproximadamente es 6 meses según los datos obtenidos de la ecuación.

- **Estimación de la productividad**

La productividad que se debe esperar de cada programador está dará por la siguiente expresión:

$$\mathbf{PR = LDC/E}$$

$$PR = 4679/12$$

$$PR = 389.91 \text{ (LDC/persona-mes)}$$

Se espera que un programador genere 390 líneas de código al mes.

- **Cálculo de personal promedio**

Para el cálculo del personal promedio se aplica la fórmula:

$$\mathbf{P = E/T}$$

$$P = 12.10 / 6.42$$

$$P = 1.88 \text{ (personas)}$$

Estos resultados indican que se requiere dos personas trabajando por unos seis meses, desarrollando 390 líneas de código en todo este periodo.

Para el cálculo total del software se considera el sueldo aproximado de un ingeniero de sistemas junior de 3000 Bs. mensuales.

$$CT = 3000 * (P * T)$$

$$CT = 3000 * (2 * 6)$$

$$CT= 36000 \text{ Bs.}$$

Por lo que se concluye que el costo estimado del prototipo es de 36000 Bs., un tiempo de 6 meses y 2 personas trabajando en el mismo.

3.12.3. Costo desarrollo del sistema

Para esta etapa se considera diversos factores, principalmente los relacionados con el desarrollo del prototipo.

Tabla 3.21
Costo de elaboración del prototipo

Detalle	Importe
Análisis y diseño del prototipo	700 Bs.
Material de Escritorio	120 Bs..
Conexión a internet	298 Bs.
Otros	50 Bs.
Total	1.168 Bs

Fuente: Elaboración propia.

3.12.4. Costo total

Para el cálculo del costo total se tomó en cuenta el costo del software calculado anteriormente y el costo de elaboración.

Tabla 3.22
Costo total del prototipo

Detalle	Importe
Costo del software	36.000 Bs.
Costo de elaboración	1.168 Bs.
Total	37.168 Bs.

Fuente: Elaboración propia.

Considerando la tabla anterior se concluye que el costo total del software es de 37.168 Bs.

CAPITULO IV

PRUEBAS Y RESULTADOS

PRUEBAS Y RESULTADOS

En este capítulo se describe las pruebas realizadas al prototipo del modelo PREDESMIN, los resultados obtenidos con los diferentes ensayos, para la prueba de la hipótesis, que es fundamental para el presente trabajo de investigación.

4.1. PRUEBAS AL MODELO

El modelo de predicción está en pleno funcionamiento, en la que se puede visibilizar las ventanas de entrada, de selección del tipo de algoritmo, el número de iteraciones que pueden ser predefinidas por el sistema o de manera manual para posteriormente, cargar los datos y mostrar los resultados de manera gráfica, como se ve en los gráficos.

Tabla 4.1
Fragmento de los datos para la predicción

M	NACIONAL	PUBLICO	urbano	Habitación	Alquilada	SI	medio tiempo	0	?
M	NACIONAL	PUBLICO	urbano	Departamento	Alquilada	SI	tiempo horario	1	?
F	NACIONAL	PUBLICO	urbano	Departamento	Alquilada	SI	medio tiempo	0	?
M	NACIONAL	PUBLICO	rural	Otro	Alquilada	NO		1	?
M	NACIONAL	PUBLICO	urbano	Casa	Alquilada	NO		0	?
M	NACIONAL	PUBLICO	urbano	Casa	Alquilada	SI	tiempo horario	0	?
M	NACIONAL	PUBLICO	urbano	Habitación	Prestada	NO		0	?
M	NACIONAL	PUBLICO	urbano	Casa	Adjudicada	NO		0	?
M	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	SI	tiempo horario	0	?
F	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	NO		0	?

M	NACIONAL	PUBLICO	rural	Casa	Propia de padres	SI	medio tiempo	0	?
M	NACIONAL	PUBLICO	rural	Casa	Propia de padres	SI	tiempo horario	0	?
F	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	SI	medio tiempo	1	?
M	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	NO		0	?
F	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	NO		0	?
F	NACIONAL	PUBLICO	urbano	Casa	Propia de padres	SI	medio tiempo	1	?

Cabe puntualizar, que los datos de la tabla 4.1 fueron en primera instancia estandarizados y entrenados para que posteriormente fueran introducidos en el modelo de predicción y este puede realizar los procesos internos.

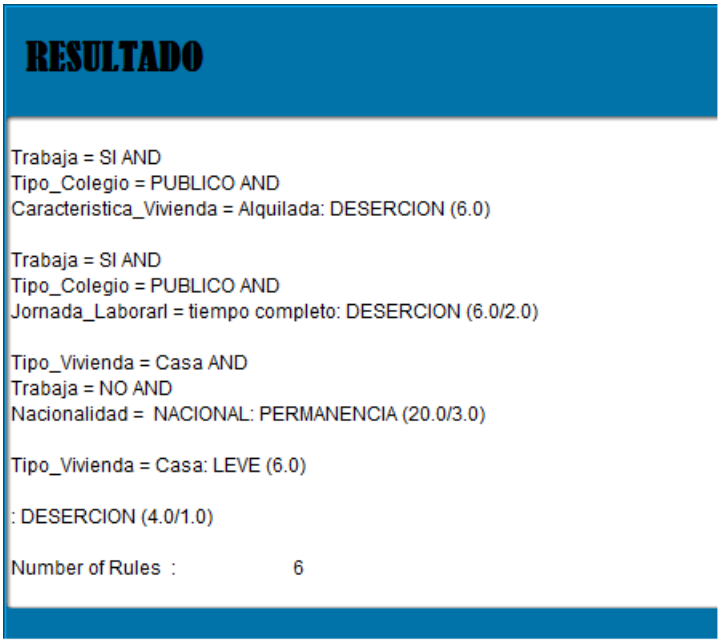


Figura 4.1 resultado del algoritmo PART

Fuente: elaboracion propia

Se aplicó el algoritmo PART, de lo que se obtuvo 6 reglas de clasificación, de las cuales se pudo observar que:

El índice de deserción será afectado por los factores principales de tipo de trabajo, tipo de colegio, característica de vivienda y jornada laboral.

Existiendo una deserción mayor cuando el alumno trabaje, y este trabajo implique un trabajo de tiempo completo, el tipo de colegio del que proviene es de tipo público, y por ultimo si la vivienda es alquilada.

Por otra parte, si el alumno trabaja medio tiempo tiene una casa propia de padres, el índice de deserción disminuye, pero aún hay un porcentaje de deserción considerable.

Pero, si el alumno es nacional, no trabaja y tiene una casa tiene la probabilidad de permanecer en la universidad y no así una de deserción.

Con estos datos podemos también aseverar, que el modelo de predicción del índice de deserción en base a Minería de Datos, está funcionando de manera efectiva.

4.2. PRUEBA DE HIPÓTESIS

Hernández (2014), cita que "Una hipótesis se retiene como un valor aceptable del parámetro, si es consistente con los datos. Si no lo es, se rechaza (pero los datos no se descartan)".

En este punto se realiza la prueba de hipótesis planteada en el capítulo uno, demostrando si la hipótesis tiene una confianza del 90%.

4.2.1. Planteamiento de la hipótesis

Nos planteamos una hipótesis nula (H_0) y la hipótesis de investigación (H_1).

- **Hipótesis nula**

  H_0: "El modelo predictivo del índice de deserción en base a factores del alumno, no tendrá una eficacia del 90% en la población estudiantil de la Universidad Pública de El Alto"

- **Hipótesis de investigación**

  H_1: "El modelo predictivo del índice de deserción en base a factores del alumno, tendrá una eficacia del 90% en la población estudiantil de la Universidad Pública de El Alto".

4.2.2. Tamaño de muestra

En este caso se determinará el tamaño de la muestra a partir de la siguiente fórmula:

$$n = \frac{Z^2 * p * q * N}{e^2(N-1) + Z^2 * p * q}$$

Dónde:

n=nuestra
Z=nivel de confianza
p=probabilidad a favor
q=probabilidad en contra
e=error de muestra
N= población

Entonces tenemos:

n=?
Z=90% = 1.96
p=65% = 0.5
q=35% =0.35
e=5% = 0.05
N= 9344

Reemplazando en la fórmula tenemos:

$$n = \frac{1.96^2 * 0.65 * 0.35 * 9344}{0.05^2(9344-1) + 1.96^2 * 0.65 * 0.35}$$

$\boldsymbol{n}$ **=337**

Determinar el nivel de significancia (rango de aceptación de la hipótesis alternativa)

Según Hernández (2014), la significancia para una tesis de investigación es del 5% en términos de probabilidad 0.05.

Cálculo del valor crítico

En este caso se debe determinar el valor crítico a partir de la tabla de T de Student tabla 4.2.

Tabla 4.2
T – Student para el punto crítico

Grados de libertad	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3137	12.7062	31.8210	63.6559
2	0.8165	1.8856	2.9200	4.3027	6.9645	9.9250
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8408
4	0.7407	1.5332	2.1318	2.7765	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9979	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7970
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3007	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6458
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9852	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2903	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758

Fuente: Distribución-T-Student (2018). Por Flores, K.

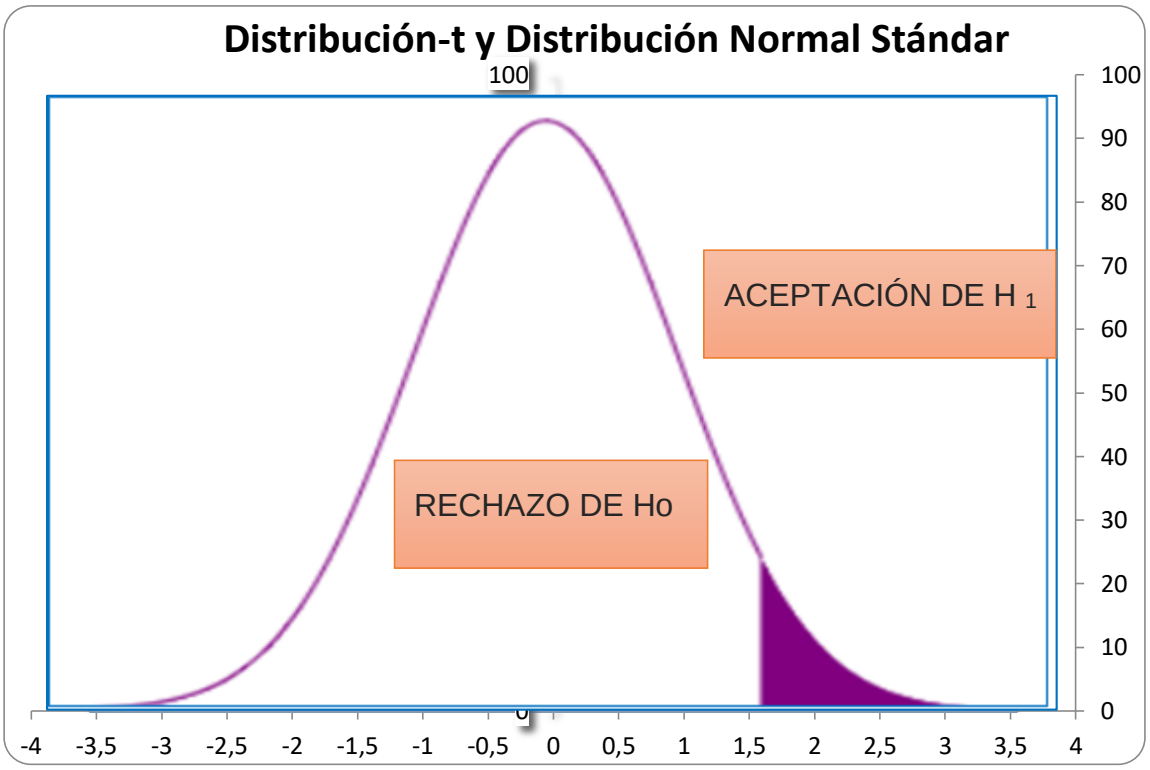


Figura 4.2 Campana de Gauss representando T Student

Fuente: Elaboración propia

De la tabla T-Student se determina que se tiene como valor para el punto crítico:

$t_{crítico}=1.64$

Determinamos el punto de prueba con la siguiente fórmula de:

$$t_{prueba} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Donde:

t=?

X=90.5

μ=90

σ=5.2

n=280

Reemplazando los valores tenemos

$$t = \frac{90.5 - 90}{\frac{5.2}{\sqrt{280}}}$$

t = 1.60

Según la tabla anterior el valor critico es: $t_{critico}$ = 1.64

Para comparar con el valor t obtenido anteriormente: t_{prueba}= 1.60

Si $|t\ critico| > |t\ prueba|$ por lo que se rechaza Ho que es la prueba nula.

Por tanto, la probabilidad de obtener los datos si H1 es cierta según α es de 90%, es aceptada.

Mediante las fórmulas de T - Student de una cola, demostramos que la hipótesis cumple con el modelo de predicción de índice de deserción en base a factores del alumno, tendrá una eficacia del 90%.

CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES Y RECOMENDACIONES

El presente capítulo describe las conclusiones alcanzadas en el desarrollo de la propuesta de investigación, el estado de la hipótesis y las recomendaciones que se da al lector sobre futuros temas como una continuación al presente trabajo de investigación.

El modelo de PREDESMIN, desarrollado en el presente trabajo de investigación, por las características propias de la Minería de Datos, cabe recalcar que el modelo no siempre puede tener resultados exactos al 100%, por lo que teniendo en cuenta que el resultado que arroje el prototipo al modelo de Minería de Datos debe ser tomado como parcial.

5.1. ESTADO DE LOS OBJETIVOS

El objetivo general descrito en el capítulo 1 menciona: “Desarrollar un modelo de predicción en base a la Minería de Datos y los factores socioeconómicos, sobre el índice de deserción de alumnos de la Universidad Pública de El Alto”.

En el Capítulo 3, se presenta el proceso de desarrollo del modelo de predicción PREDESMIN. Este proceso es realizado utilizando Minería de Datos, mostrando la predicción del índice de deserción de alumnos de la Universidad Pública de El Alto.

Por lo anterior descrito se logró alcanzar en su totalidad, ya que se construyó el modelo de predicción en base a la Minería de Datos.

En cuanto a los objetivos específicos se justifica cada uno de ellos en los siguientes incisos:

- “Analizar los algoritmos de minería de datos que sean útiles para el modelo de predicción”, En el capítulo 3 se entrenó con los diferentes algoritmos de Minería de Datos con los que cuenta WEKA, de los cuales se seleccionaron los de mayor relevancia.
- “Plantear un modelo de predicción en base a la Minería de Datos”, El modelo de Minería de Datos expuesto en capítulo 3, tiene como base a los modelos de Minería de Datos descritos como referencia en el capítulo 1 y la base teórica en el capítulo 2, con la función de ambas bases se logra el desarrollo del modelo.
- “Implementar un prototipo en base a algoritmos de Minería de Datos”, en el capítulo 3 se detalla la implementación del prototipo, logrando el objetivo planteado satisfactoriamente.
- “Identificar los factores de deserción en base a la aplicación de algoritmos de minería de datos”, Gracias a la aplicación de los varios algoritmos que tiene WEKA, se logró obtener parámetros sobre factores de los índices de deserción de alumnos, cumpliendo satisfactoriamente con este objetivo.

5.2. ESTADO DE LA HIPÓTESIS

La hipótesis establecida en capítulo 1 es la siguiente: “Debido a la aplicación de técnicas de Minería de Datos y la ingeniería de Software se tiene el modelo predictivo del índice de deserción en base a factores del alumno, teniendo una eficacia del 90% en la población estudiantil de la Universidad Pública de El Alto”. La planificación y estructura de la secuencia de los pasos a seguir a la hora de aplicar el modelo de Minería de Datos, la eficiencia de los algoritmos se sustenta al orden de las tareas que el usuario realiza.

Por tanto, el modelo de predicción de deserción en base a la Minería de Datos, asegura la calidad de los resultados del mismo y de acuerdo a la metodología aplicada ISO se logró el 91% de eficacia al momento de aplicar el modelo.

5.3. CONCLUSIONES

El desarrollo del “Modelo de Predicción en Base a Minería de Datos sobre Índices de Deserción de Alumnos”, llega a las siguientes conclusiones.

La Minería de Datos es una herramienta potente y de mayor alcance, puesto que realiza el tratamiento con cantidades enormes de datos, encontrando un nuevo conocimiento en las bases de datos almacenadas periódicamente en la unidad de registros y admisiones.

El modelo de predicción en base a Minería de Datos, nos muestra los resultados en términos del índice de deserción de los alumnos de la Universidad Pública de El Alto, con base en los factores socioeconómicos descritos como información básica al ingreso y continuidad de la educación superior.

El modelo de minería de datos muestra la aplicación de algoritmos óptimos en la predicción, teniendo en cuenta la identificación de los factores de deserción universitaria.

La evaluación del modelo de predicción en base a la minería de datos está sujeta a la aceptación de los valores de la proporción expuesta en el presente trabajo de investigación.

5.4. RECOMENDACIONES

Considerando que la información que se pudo recabar y generar durante la investigación, se detalla las siguientes recomendaciones:

- ✓ Se recomienda la actualización de la base de datos de la U.P.E.A. para tener resultados más eficientes.

- ✓ Ampliar la información de los estudiantes para generar modelos que permitan explorar con mayor profundidad la relación entre los índices de deserción universitaria y los factores socioeconómicos.
- ✓ Se recomienda recabar datos académicos sobre la asignación de materias, el historial académico para continuar con la presente investigación.
- ✓ Se recomienda poner en practica otros algoritmos de minera de datos existente para verificar los resultados obtenidos en esta investigación

BIBLIOGRAFÍA

BIBLIOGRAFÍA

Agrawal, R., & Shafer, J. C. (1996). *Parallel Mining of Association Rules.* IEEE Transactions on Knowledge and Data Engineering.

Aguilar, R. (2003). *Minería de Datos. Fundamentos, Técnicas y Aplicaciones.* España.

Alonso, C., & Rodriguez, J. (s.f.). *Introduccion a la Mineria de Datos y al Aprendizaje Automatico.* Obtenido de UVA/UBU: https://www.infor.uva.es/~calonso/MUI-TIC/MineriaDatos/01IntroduccionMDyAA.pdf

Apaza Quevedo, B. R. (2009). *MODELOS DE PREDICCIÓN DE PREVALENCIA DE ENFERMEDADES PARA CENTROS DE SALUD BASADO EN MENERIA DE DATOS.* Obtenido de Repositorio Institucional UMSA: https://repositorio.umsa.bo/bitstream/handle/123456789/1540/T.1794.pdf?sequence=3&isAllowed=y

Aruquipa, V. (2015). Modelo Predictivo de las Preferencias Políticas de la Población Votante Boliviana. *(Tesis de Grado).* Universidad Mayor de San Andrés, La Paz.

Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques* (Second ed.). Indianapolis: Wiley Publishing,Inc.

Boehm, B. W. (1981). *Software Engineering Economics.* Prentice-Hall.

Bueno, J. (2019). *¿Qué es el abandono universitario? ¿Por qué, cuándo y dónde se da?* Obtenido de Informacion: https://www.diarioinformacion.com/opinion/2019/05/19/abandono-universitario-da/2150042.html#:~:text=El%20abandono%20del%20grado%20o,producirle%20a%20la%20persona%20afectada.

Carrillo, R., & Gimenez, H. (29 de Enero de 2014). *Prezi*. Obtenido de Modelos De Predicción: https://prezi.com/yifr4gi6p1r1/modelos-de-prediccion/

Chiavenato, I. (2006). *Introduccion a la Teoria General de la Administracion* (Séptima ed.). Mexico: McGram-Hill Interamericana. Obtenido de https://esmirnasite.files.wordpress.com/2017/07/i-admon-chiavenato.pdf

Choque Aspiazu, G. (3 de agosto de 2009). *Ciencia y Computacion*. Obtenido de Mineria de Datos Predictiva: https://www.eldiario.net/computacion/7-090803/supl7_01.html

Cota, A. (1994). *Ingenieria de Software. Soluciones Avanzadas .*

CRIS-DM. (2015). *CRISP-DM cros insustry standard process for data mining*. Obtenido de cris-dm.eu: http://crisp-dm.eu/home/crisp-dm-methodology/

Del Valle, D. (2014). *Estimación de costos de desarrollo de software*. Obtenido de gestiopolis: https://www.gestiopolis.com/estimacion-de-costos-de-desarrollo-de-software/

Fayyad, U., Pieatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases.* Obtenido de Kdnuggets: https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Ferrel, O., Geoffey, H., & Ferrel, L. (2009). *INTRODUCCIÓN A LOS NEGOCIOS EN UN MUNDO CAMBIANTE* (Séptima ed.). Mexico: McGRAW-HILL/INTERAMERICANA.

Flores, M. (2005). *Gestión del conocimiento organizacional en el taylorismo y en la teoría de las relaciones humanas*. Obtenido de revistaespacios.com: https://www.revistaespacios.com/a05v26n02/05260241.html#inicio

García, R. (2005). *Mineria de datos basada en sistemas inteligentes.*

Gonzáles, A. (2006). Desarrolo de técnicas de minería de datos en procesos industriales: Modelización en línea de producción de acero. *(Tesis Doctoral).* Universidad de la Rioja, Logroño.

Heriquez, H. (2018). Desarrollo de una aplicación web para la mejora del control de asistencia de personal en la Escuela Tecnológica Superior de la Universidad Nacional de Piura. *(tesis de licenciatura).* Universidad Inca Garcilaso de la Vega, Lima.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación.* México D.F.: McGraw-Hill.

Hernández, C., & Dueñas, M. (2009). *Hacia una metodología de gestión del conocimiento basada en minería de datos.* Obtenido de repositorio.uigv.edu.pe: http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1&isAllowed=y#:~:text=Seg%C3%BAn%20SAS%20m%C3%A1s%20que%20una,proyecto%20de%20miner%C3%ADa%20de%20datos.

Hidalgo, E. (2014). Aplicación De Minería De Datos Para El Análisis Del Rendimiento Académico En Estudiantes De Secundaria. *(Tesis de Licenciatura).* Universidad Pública de El Alto, El Alto- Bolivia.

Himmel, E. (2002). *Modelos de Analisis de la Desercion Estudiantil en la Educacion Superior.* Obtenido de Revista Calidad en la Educacion: https://www.calidadenlaeducacion.cl/index.php/rce/article/view/409/409

Jacobson, I. (2000). *El lenguaje Unificado de Modelado. Manual de referencia.* Madrid: Pearson Educacion S.A.

Landa, J. (2016). *Tratamiento de los Datos*. Obtenido de fcojlanda.me: http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/

LEXICO. (2020). *Lexico.com*. Obtenido de LEXICO powered by OXFORD: https://www.lexico.com/es/definicion/indice

Lupín , B., Pesciarelli, S., & Agustinelli, S. (2013). *DECLARACIÓN DE LA EDUCACIÓN SUPERIOR COMO BIEN PÚBLICO.* Obtenido de Jornadas Nacionales sobre Pedagogía de la Formación del Profesorado: https://fh.mdp.edu.ar/encuentros/index.php/jnpfp/3jnpfp/paper/viewFile/1167/500

Mamani Padilla, D. I. (2019). *MODELO DE MINERIA DE DATOS BASADOS EN FACTORES ASOCIADOS PARA LA PREDICCION DE DESERCION ESTUDIANTIL UNIVERSITARIA.* Obtenido de Repositorio UNAM: http://repositorio.unam.edu.pe/bitstream/handle/UNAM/94/T095_72389106_T.pdf?sequence=1&isAllowed=y

Marcano, Y., & Rodriguez, R. (2014). *Minería de datos aplicado a la deserción estudiantil.* Obtenido de Revistas UPEL: http://revistas.upel.edu.ve/index.php/educare/article/viewFile/2600/1255

Marquez Granado, E. P. (2006). *DESARROLLO DE UN MODELO DE MINERIA DE DATOS ACADEMICOS.* Obtenido de Repositorio Institucional UMSA: https://repositorio.umsa.bo/bitstream/handle/123456789/306/T-1372.pdf?sequence=3&isAllowed=y

Martinez, A., & Rios, F. (2006). *Los conceptos de conocimiento, epistemología y paradigma, como base diferencial en la orientación metodológica del trabajo de grado*. Obtenido de Revista de Epistemoliga de Ciencias Sociales: https://www.moebio.uchile.cl/25/martinez.html

Mercado, D., Pedraza, L., & Martinez, E. (2015). *Comparación de Redes Neuronales aplicadas a la prediccion de Series de Tiempo .* Obtenido de Scielo: http://www.scielo.org.co/pdf/prosp/v13n2/v13n2a11.pdf

Molina, J., & Garcia, J. (2006). *Técnicas de Análisis de Datos. Aplicaciones Prácticas Utilizando Microsoft Excel y Weka.* Madrid.

Páramo, G., & Correa, C. (1999). Desercion Estudiantil Universitaria. Conceptualizacion. *Revista Universidad Eafit.*, 65-78.

Pautsch, J. (2009). Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. *(Tesis de Grado).* Universidad Nacional de Misiones, Apóstoles.

Perassi, Z. (2009). *¿ES LA EVALUACIÓN CAUSA DEL FRACASO?* Obtenido de Revista Iberoamericana de Educacion: https://rieoei.org/historico/documentos/rie50a03.pdf

Pérez, J., & Gardey, A. (2012). *Definiciones de desercion*. Obtenido de definicion.de: https://definicion.de/desercion/

Pressman, R. (2010). *Ingeniería del Software. Un Enfoque Práctico.* México: Mc Graw Hill.

Quintero Velasco, I. (2016). *ANÁLISIS DE LAS CAUSAS DE DESERCIÓN UNIVERSITARIA .* Obtenido de Repository.Unad.edu.co: https://repository.unad.edu.co/bitstream/10596/6253/1/23783211.pdf

Ramirez, T., Diaz, R., & Salcedo, A. (2017). ¿Abandono o deserción estudiantil? Una necesaria discusion conceptual. *Investigacion y Posgrado*, 63-74.

Roa, L. (2017). *Ingenieria de Software.* Obtenido de cic.puj.edu.co: http://cic.puj.edu.co/wiki/lib/exe/fetch.php?media=materias:is1:01_lectura_ingenieria_software.pdf

Rodriguez Suárez, Y., & Diáz Amador, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas*, 74-80.

Rodriguez, A., Espinoza, J., Ramirez, L., & Ganga, A. (2018). *Deserción Universitaria: Nuevo Análisis Metodológico*. Obtenido de Formación universitaria: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-50062018000600107#t1

Romero, S. (2016). *INVESTIGACIÓN SOBRE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD AUTÓNOMA "JUAN MISAEL SARACHO" DE LA*

CIUDAD DE TARIJA. Obtenido de uajms.edu.bo: http://www.uajms.edu.bo/revistas/wp-content/uploads/2017/10/Inv-des-art1.pdf

Siebes, A. (2020). *Mineria de datos y estadisticas.* Obtenido de SpringerLink: https://link.springer.com/chapter/10.1007/978-3-7091-2588-5_1

Thuraisingham, B. (1999). *Data Mining. Technologies, Technigues Tools and Trends CRC Pres LLC.*

Timarán, S., Hernández, I., Caicedo, S., Hidalgo, A., & Alvarado, J. (2016). *Descubrimiento de Patrones de Desempeño Académico.* Bogotá: Ediciones Universidad Cooperativa de Colombia.

Tinto, V. (1989). *DEFINIR LA DESERCION: UNA CUESTION DE PERSPECTIVA.* Obtenido de Publicaciones ANUIES: http://publicaciones.anuies.mx/pdfs/revista/Revista71_S1A3ES.pdf

Vizcaino, P. (2008). APLICACIÓN DE TÉCNICAS DE INDUCCIÓN DE ÁRBOLES DE DECISIÓN A PROBLEMAS DE CLASIFICACIÓN MEDIANTE EL USO DE WEKA (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS). Bogotá, Colombia.

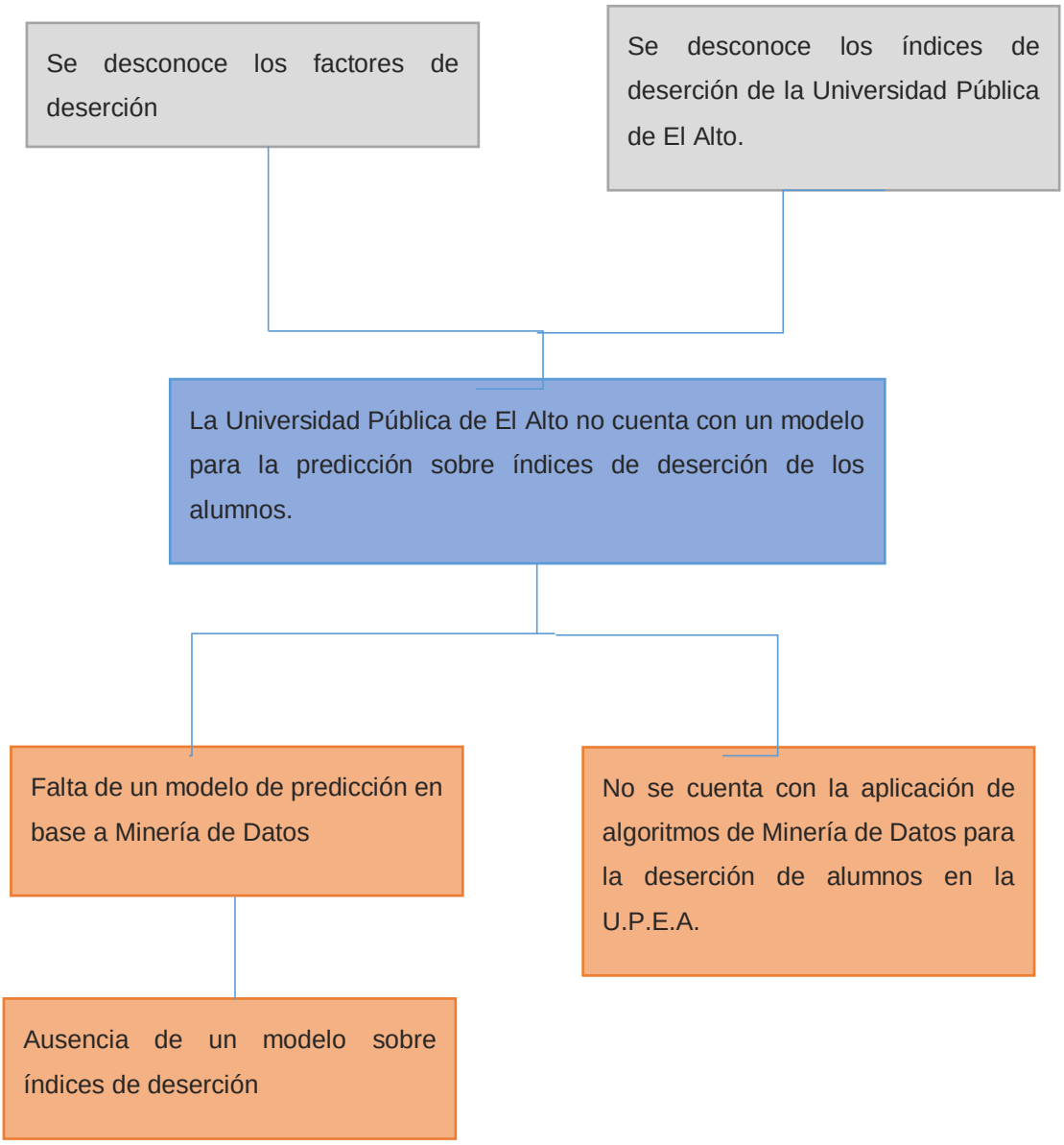
Wang, X. Z. (1999). *Data Mining and Knowledge Discovery for Process Monitoring and Control*. Obtenido de doc.lagout.org: https://doc.lagout.org/Others/Data%20Mining/Data%20Mining%20and%20Knowledge%20Discovery%20for%20Process%20Monitoring%20and%20Control%20%5BWang%201999-09-15%5D.pdf

Witten, I. H., & Frank, E. (2005). *Data Mining. Practical Machine Learnig Tools and Techniques* (Second ed.). San Francisco: Morgan Kaufmann Publishers.

ANEXOS

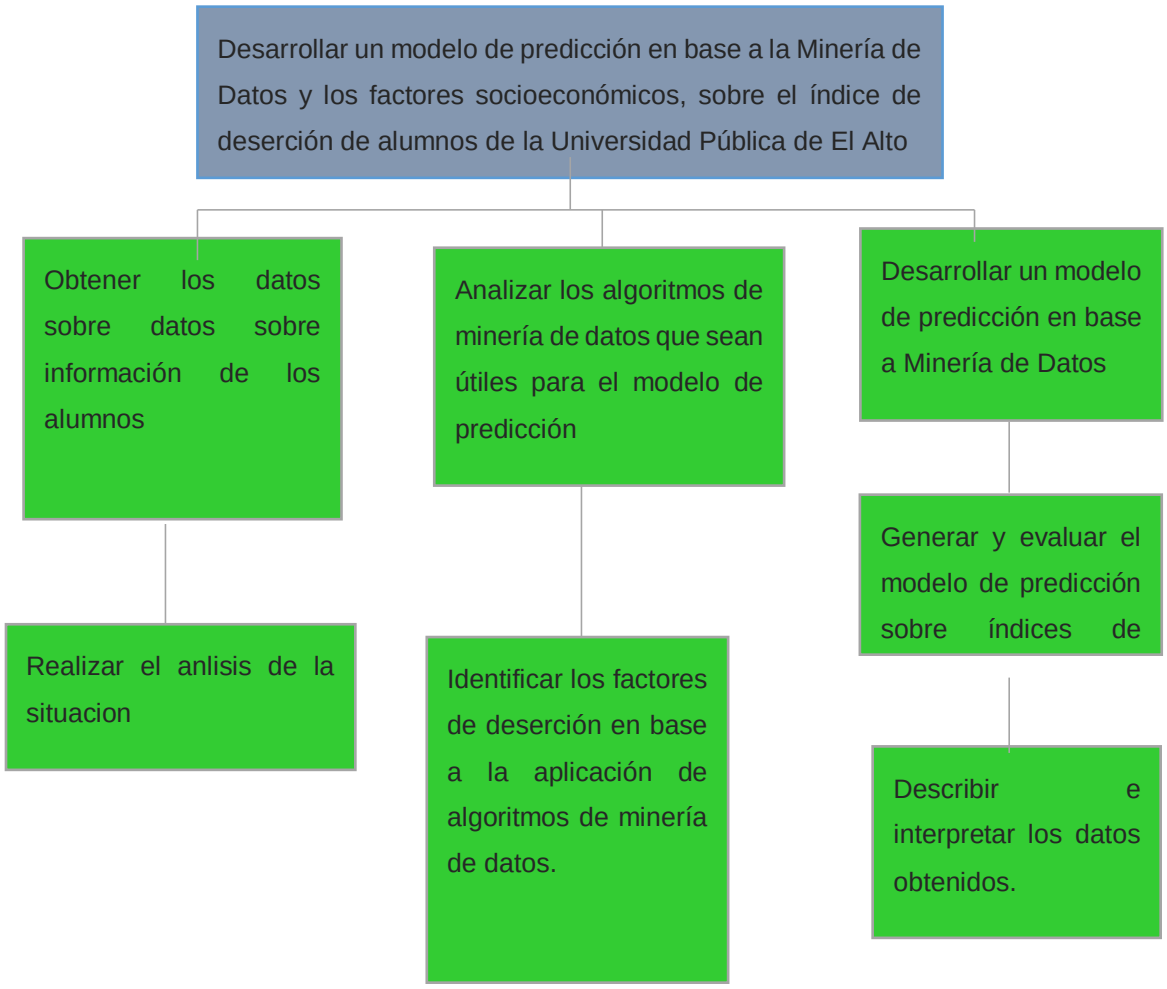
ANEXO A.

ÁRBOL DE PROBLEMAS



ANEXO B.

ÁRBOL DE OBJETIVOS



La Paz - El Alto agosto de 2020

Señor
Ing. David Carlos Mamani Quispe
DIRECTOR DE CARRERA
INGENIERÍA DE SISTEMAS
Presente. –

REF.: AVAL DE CONFORMIDAD

Distinguido Ingeniero,

Mediante la presente tengo a bien comunicarle mi conformidad de la tesis de grado **"MODELO DE PREDICCIÓN BASADO EN MINERÍA DE DATOS SOBRE ÍNDICES DE DESERCIÓN DE ALUMNOS CASO: UNIVERSIDAD PÚBLICA DE EL ALTO".** Que propone el postulante Ivan Rodrigo Hidalgo Mamani, con cedula de identidad 6872515 LP., para su defensa pública, evaluación correspondiente a la materia de Taller de Licenciatura II, de acuerdo al reglamento vigente de la carrera de Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales

Atentamente.

La Paz - El Alto agosto de 2020

Señor
Ing. Enrique Flores Baltazar
TUTOR METODOLÓGICO TALLER ll
Presente. —

REF.: Aval de conformidad

Distinguido Ingeniero,

Mediante la presente tengo a bien comunicarle mi conformidad de la tesis de grado **"MODELO DE PREDICCION BASADO EN MINERIA DE DATOS SOBRE INDICES DE DESERCION DE ALUMNOS CASO: UNIVERSIDAD PÚBLICA DE EL ALTO".** Que propone el postulante Ivan Rodrigo Hidalgo Mamani, con cedula de identidad 6872515 LP., para su defensa pública, evaluación correspondiente a la materia de Taller de Licenciatura II, de acuerdo al reglamento vigente de la carrera de Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales

Atentamente.

Ing. Juan Regis Muñez Sirpa
TUTOR ESPECIALISTA

La Paz - El Alto agosto de 2020

Señor
Ing. Enrique Flores Baltazar
TUTOR METODOLÓGICO TALLER ll
Presente. —

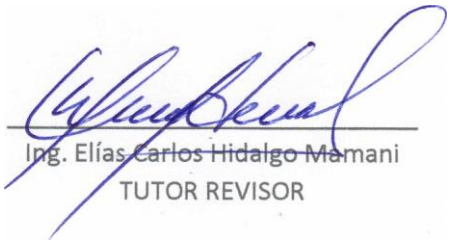
REF.: Aval de conformidad

Distinguido Ingeniero,

Mediante la presente tengo a bien comunicarle mi conformidad de la tesis de grado **"MODELO DE PREDICCION BASADO EN MINERIA DE DATOS SOBRE INDICES DE DESERCION DE ALUMNOS CASO: UNIVERSIDAD PÚBLICA DE EL ALTO".** Que propone el postulante Ivan Rodrigo Hidalgo Mamani, con cedula de identidad 6872515 LP., para su defensa pública, evaluación correspondiente a la materia de Taller de Licenciatura II, de acuerdo al reglamento vigente de la carrera de Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales

Atentamente.



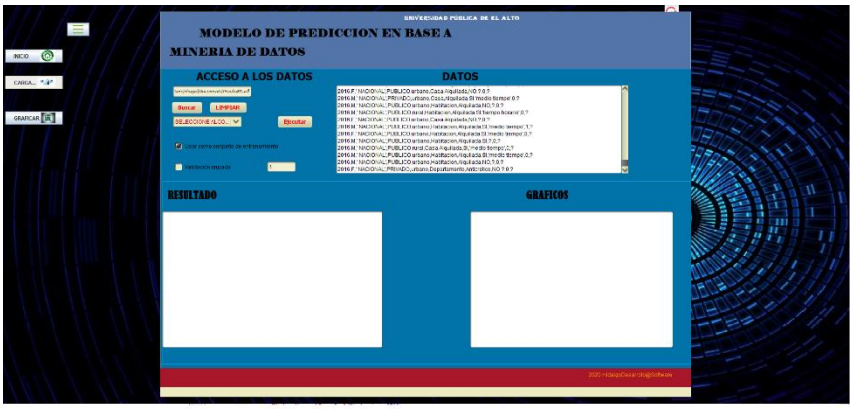
ANEXO C.

MANUAL DE USUARIO

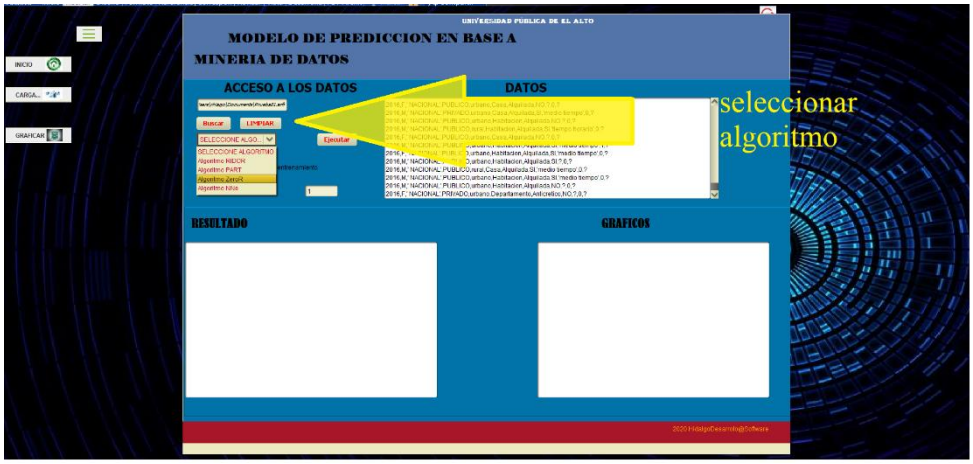
MANUAL DE USUARIO

El modelo de predicción en base a Minería de Datos, es una herramienta destinada a realizar predicciones de factores sobre índices de deserción de alumnos.

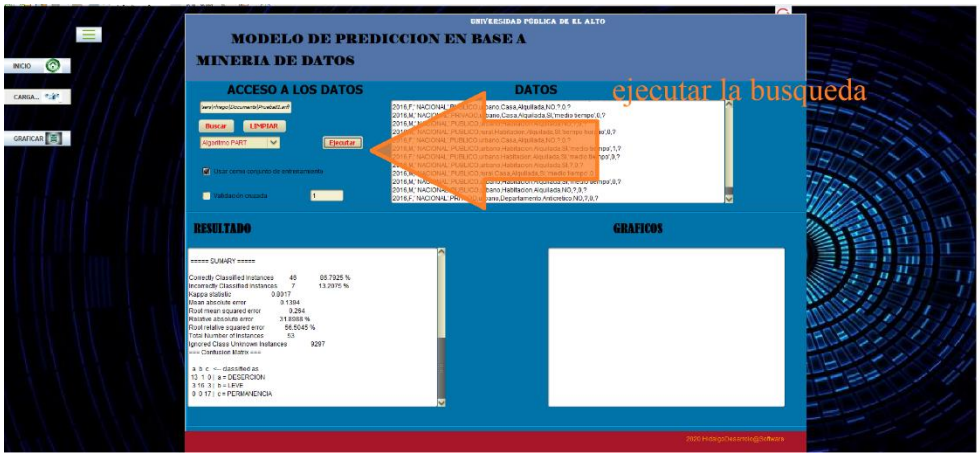
- **Ejecución del modelo de minería**
  - Cargado del archivo arff



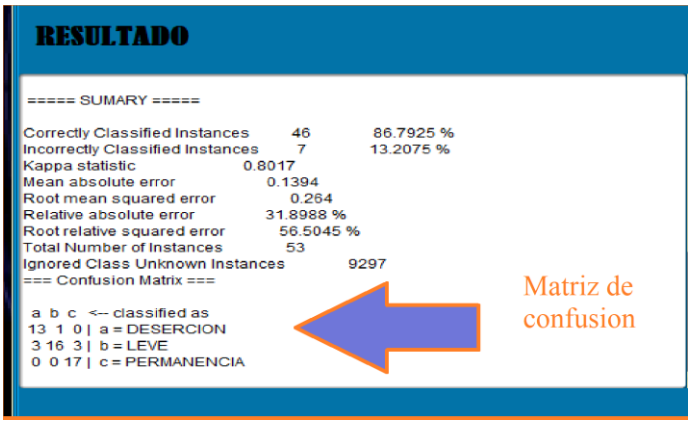
  - Escoger algoritmo a ser utilizado.

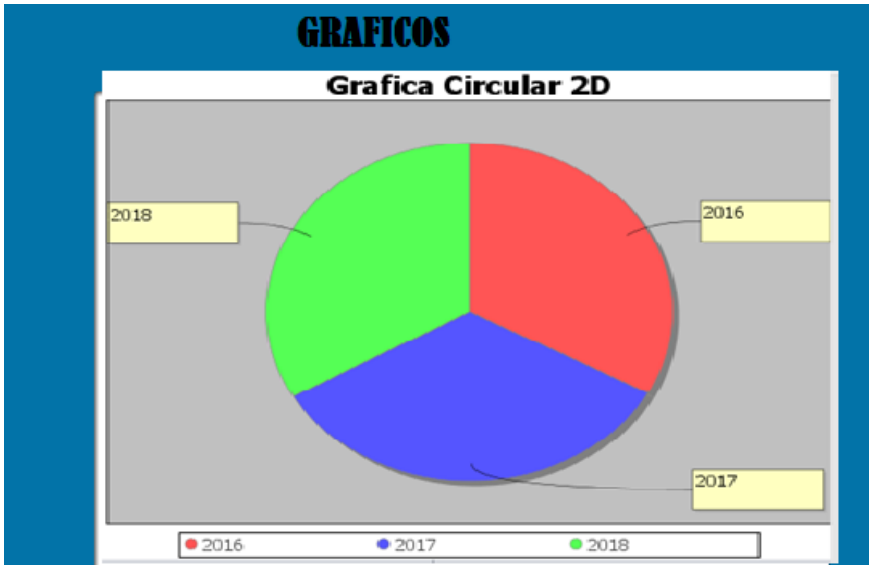


- Ejecutar búsqueda de patrones.



- Visualización de los resultados.


RESULTADO
===== SUMARY =====
Correctly Classified Instances 46 86.7925 %
Incorrectly Classified Instances 7 13.2075 %
Kappa statistic 0.8017
Mean absolute error 0.1394
Root mean squared error 0.264
Relative absolute error 31.8988 %
Root relative squared error 56.5045 %
Total Number of Instances 53
Ignored Class Unknown Instances 9297
=== Confusion Matrix ===
a b c <-- classified as
13 1 0 | a = DESERCION
3 16 3 | b = LEVE
0 0 17 | c = PERMANENCIA
Matriz de confusion


GRAFICOS
Grafica Circular 2D
2018
2016
2017
2016
2017
2018

ANEXO D.

Resultado del analisis de prediccion

Time taken to build model: 0 seconds

=== Predictions on training set ===

inst#	actual	predicted	error	prediction
1	1:DESERCION	1:DESERCION		1
2	2:LEVE	2:LEVE		1
3	1:DESERCION	1:DESERCION		0.75
4	1:DESERCION	1:DESERCION		0.75
5	2:LEVE	2:LEVE		1
6	1:DESERCION	1:DESERCION		0.75
7	2:LEVE	2:LEVE		1
8	2:LEVE	1:DESERCION	+	0.75
9	2:LEVE	2:LEVE		1
10	1:DESERCION	1:DESERCION		0.667
11	1:DESERCION	1:DESERCION		0.667
12	1:DESERCION	1:DESERCION		0.667
13	1:DESERCION	1:DESERCION		0.667
14	2:LEVE	2:LEVE		0.909
15	2:LEVE	2:LEVE		0.909
16	2:LEVE	2:LEVE		0.909
17	2:LEVE	2:LEVE		0.909

18 2:LEVE 2:LEVE 0.909

19 3:NO DESERCION 3:NO DESERCION 0.85

20 3:NO DESERCION 3:NO DESERCION 0.85

21 3:NO DESERCION 3:NO DESERCION 0.85

22 3:NO DESERCION 3:NO DESERCION 0.85

23 3:NO DESERCION 3:NO DESERCION 0.85

24 3:NO DESERCION 3:NO DESERCION 0.85

25 3:NO DESERCION 3:NO DESERCION 0.85

26 2:LEVE 2:LEVE 1

27 2:LEVE 2:LEVE 1

28 2:LEVE 1:DESERCION + 0.667

29 2:LEVE 1:DESERCION + 0.667

86 1:? 2:LEVE 0.909

87 1:? 2:LEVE 0.909

88 1:? 3:NO DESERCION 0.85

89 1:? 3:NO DESERCION 0.85

90 1:? 1:DESERCION 0.667

91 1:? 2:LEVE 0.909

92 1:? 3:NO DESERCION 0.85

93 1:? 1:DESERCION 0.667

94 1:? 2:LEVE 0.909

95 1:? 3:NO DESERCION 0.85

96 1:? 3:NO DESERCION 0.85

97	1:?	2:LEVE	0.909
98	1:?	2:LEVE	0.909
99	1:?	2:LEVE	0.909
100	1:?	3:NO DESERCION	0.85
.			
.			
.			
.			
9348	1:?	1:DESERCION	1
9349	1:?	1:DESERCION	0.75
9350	1:?	1:DESERCION	0.75

=== Evaluation on training set ===

Time taken to test model on training data: 4.08 seconds

=== Summary ===

Correctly Classified Instances	46	86.7925 %
Incorrectly Classified Instances	7	13.2075 %
Kappa statistic	0.8017	
Mean absolute error	0.1394	
Root mean squared error	0.264	

Relative absolute error 31.8988 %

Root relative squared error 56.5045 %

Total Number of Instances 53

Ignored Class Unknown Instances 9297

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,929	0,077	0,813	0,929	0,867	0,818	0,872	0,006	DESERCION
0,727	0,032	0,941	0,727	0,821	0,734	0,745	0,006	LEVE
1,000	0,083	0,850	1,000	0,919	0,883	0,794	0,004	NO DESERCION
Weighted Avg. 0,868	0,060	0,878	0,868	0,864	0,804	0,794	0,005	

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 13  1  0 |  a = DESERCION
  3 16  3 |  b = LEVE
  0  0 17 |  c = NO DESERCION
```

: DESERCION (4.0/1.0)