UNIVERSIDAD NACIONAL DE LOJA

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

CARRERA DE INGENIERÍA EN SISTEMAS

"Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos"

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

Autor:

Darwin Andrés Becerra Encarnación

Director:

Ing. Henry Patricio Paz Arias, Mg. Sc.

LOJA – ECUADOR

2014 - 2015

Certificación de director

Ing. Henry Patricio Paz Arias, Mg. Sc.

OCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS

CERTIFICA:

Que el egresado **Darwin Andrés Becerra Encarnación** autor del presente trabajo de tesis, cuyo tema versa sobre **"ESTUDIO DEL RENDIMIENTO ACADÉMICO APLICANDO TÉCNICAS DE MINERÍA DE DATOS"**, ha sido dirigido, orientado y discutido bajo mi asesoramiento y reúne a satisfacción los requisitos exigidos en una investigación de este nivel por lo cual autorizo su presentación y sustentación.

Loja, 28 de octubre de 2014

Ing. Henry Patricio Paz Arias, Mg. Sc.

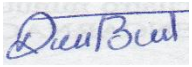
DIRECTOR DEL TRABAJO DE TITULACIÓN

Autoría

Yo **Darwin Andrés Becerra Encarnación** declaro ser autor del presente trabajo de tesis y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales, por el contenido de la misma.

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

Autor: Darwin Andrés Becerra Encarnación.

Firma: 

Cédula: 1104645005

Fecha: 15 de enero de 2015

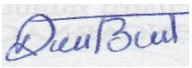
CARTA DE AUTORIZACIÓN DE TESIS POR PARTE DEL AUTOR, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.

Yo **Darwin Andrés Becerra Encarnación** declaro ser autor de la tesis titulada: **"ESTUDIO DEL RENDIMIENTO ACADÉMICO APLICANDO TÉCNICAS DE MINERÍA DE DATOS"**, como requisito para optar al grado de: **Ingeniero en Sistemas**; autorizo al Sistemas Bibliotecario de la Universidad Nacional de Loja para que con fines académicos, muestre al mundo la producción intelectual de la Universidad, a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional.

Los usuarios pueden consultar el contenido de este trabajo en el RDI, en las redes de la información del país y del exterior, con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia de la tesis que realiza un tercero.

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

Firma: 

Autor: Darwin Andrés Becerra Encarnación

Cédula: 1104645005

Dirección: Loja (Tnte. Maximiliano Rodríguez y Perú)

Correo Electrónico: dabecerrae@unl.edu.ec

Teléfono: s/n **Celular:** 0986951899

DATOS COMPLEMENTARIOS

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

Tribunal de Grado: Ing. Ana Lucía Colala Troya, Mg.

Ing. Waldemar Victorino Espinoza Tituana, Mg.

Ing. Franco Hernán Salcedo López, Mg.

Agradecimiento

El presente Trabajo de Titulación primeramente agradezco a Dios por bendecirme para poder llegar a culminar esta meta.

También a la Universidad Nacional de Loja por permitir realizar mis estudios en la carrera de Ingeniería en Sistemas.

Además agradezco a los docentes que aportaron en mi formación profesional y a mi Director del Trabajo de Titulación quien con sus conocimientos, su experiencia y su motivación ha logrado en mí que pueda terminar mi carrera con éxito.

Darwin Andrés Becerra Encarnación

Dedicatoria

Mi Trabajo de Titulación lo dedico con todo mi amor y cariño.

A Dios porque ha estado siempre guiándome para poder llegar a este objetivo.

A mis padres que me dieron la vida, educación, consejos, por ser el pilar fundamental y por hacer todo en la vida para que yo pudiera lograr mis sueños.

A mis hermanos por ser parte importante en mi vida y porque siempre me han incentivado a seguir adelante.

Darwin Andrés Becerra Encarnación

Cesión de derechos

Darwin Andrés Becerra Encarnación autor principal del presente Trabajo de Titulación, autoriza a la Universidad Nacional de Loja, al Área de la Energía, las Industrias y los Recursos Naturales No Renovables y por ende a la Carrera de Ingeniería en Sistemas hacer uso del mismo en lo que estime sea conveniente.

a. Titulo

“Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos”

b. Resumen

La Minería De Datos en el ámbito de la educación investiga, prepara y explora los datos para sacar la información oculta en ellos, básicamente surge para intentar ayudar a comprender el contenido de un repositorio de datos.

Durante los últimos años las universidades han adquirido un gran interés sobre el Rendimiento Académico de los estudiantes y en determinar qué factores influyen, pretendiendo de esta manera evidenciar cuáles son los que más afectan al rendimiento académico de los estudiantes, para ello se analizó los datos académicos, personales, socioeconómicos e institucionales correspondientes a los períodos 2010-2013 de la Universidad Nacional de Loja del Área de la Energía, las Industrias y los Recursos Naturales No Renovables.

Posteriormente se realizó un estudio de las metodologías de Minería de Datos, seleccionando CRISP-DM, debido a que es fundamental en este estudio, porque contiene una guía para el desarrollo del proyecto. Así mismo en la aplicación de técnicas de minería de datos se optó por la Clasificación, que contiene algoritmos como RIDOR, ID3, C4.5, JRIP y PART que fueron utilizados en la herramienta RAPIDMINER que permitió llevar los procesos y obtener resultados del modelo, los mismos que se analizó y evaluó a través de datos reales determinar los factores que influyen en el Rendimiento Académico.

.

Summary

Educational data mining can be apply to research, prepares and explores the data to get the information that is hidden in them; it is used to help to understand the content of a data repository.

During last years, universities have acquired a great interest on the Academic Performance of students and want to determine what factors influence in it, trying to demonstrate which of them are the most affecting academic performance, to do that personal, socioeconomic and institutional data was analyzed from 2010 to 2013 of Loja National University into Energy Area, Industries and Non-Renewable Natural Resources.

After that, a study of data mining methodologies, selecting CRISP-DM, it is essential in this study because it contains a guide to develop the project. Also in the application of data mining techniques we chose the Classification that contains RIDOR, ID3, C4.5, PART and JRIP which were used in the RapidMiner tool, it allowed to develop the processes and get results from the model, finally they were analyzed and assessed using real data to determine factors influencing the Academic Performance.

Índice de Contenidos

Índice General

Índice de Figuras

Índice de Tablas

c. Introducción

Actualmente el rendimiento académico de los estudiantes ha adquirido un gran interés en todas las universidades, debido a que permite mejor el nivel de educación superior en las mismas, una de ellas es la Universidad Nacional de Loja, ya que acoge gran cantidad de estudiantes en distintas áreas y con la información que los mismos proporcionan a los sistemas informáticos de la Universidad, se podrá obtener un modelo sobre el rendimiento académico de los estudiantes y factores que influyen.

Es por ello necesario contar con métodos eficientes y automáticos para explorar las grandes Bases de Datos que almacenan información útil, procesando de forma rápida y fiable la información para encontrar patrones que permitan resolver un problema [1], con esto surgió la Minería de Datos (MD) que ayuda a examinar grandes cantidades de datos para descubrir factores que influyen en el rendimiento académico, donde el objetivo es encontrar modelos a partir de los datos, descubrir patrones que apoye a la toma de decisiones que logren beneficios [2].

En los últimos años han surgido diversas herramientas de MD, como RapidMiner, Weka, Oracle Data Mining, entre otras, basadas en técnicas que facilitan el procesamiento de datos y permiten realizar un análisis de los mismos, con el objetivo de determinar el rendimiento académico [3]. Así mismo se utilizó la metodología CRISP-DM como una guía para desarrollan el presente artículo.

En el presente Trabajo de Titulación el objetivo es determinar el rendimiento académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de MD, donde se propone la utilización de las mismas, para detectar cuáles son los factores (académicos, personales, socioeconómicos e institucionales) que influyen en el rendimiento académico de los estudiantes.

Basado en los lineamientos de la Universidad Nacional de Loja el Trabajo de Titulación está estructurado de la siguiente manera: Resumen donde se presenta una descripción de lo que se desarrolló en el presente Trabajo de Titulación. Introducción que engloba

contenido general referente a este proyecto. La Revisión Literaria en el cual se describe conceptos referentes a la minería de datos como son: las etapas, áreas donde se puede aplicar, técnicas descriptivas-descriptivas para extraer conocimiento y herramientas computacionales que son necesarias para analizar el conocimiento extraído. También se realizó un estudio de las Metodologías de Minería de Datos que sirvió como una guía para llevar a cabo el presente proyecto. La Metodología donde se describe materiales, métodos y técnicas que fueron utilizadas. Discusión presenta un análisis de los objetivos con una descripción de los procesos que se llevaron a cabo para cumplir con los mismos. Conclusiones se redacta las opiniones que se dieron al finalizar el proyecto. Recomendaciones donde se destaca trabajos futuros.

d. Revisión de Literatura

1. CAPITULO I: MINERÍA DE DATOS

La Minería De Datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la Minería de Datos es encontrar modelos a partir de los datos [4].

El objetivo principal de la minería de datos consiste en extraer información y transformarla en una estructura comprensible para su posterior uso, donde para ello, la minería prepara, sondea y explora los datos para sacar la información oculta en ellos, básicamente surge para intentar ayudar a comprender el contenido de un repositorio de datos [5, 6].

Existen algunas diferencias y/o ventajas entre aplicar minería de datos con respecto a sólo utilizar modelos estadísticos [7]:

- La minería de datos es más amplia ya que es un proceso completo formado por varias etapas y que incluye muchas técnicas, entre ellas, las estadísticas. Este proceso de descubrimiento de información está formado por las etapas de pre-procesado, la aplicación de técnicas de minería de datos (una de ellas puede ser estadística) y la evaluación e interpretación de los resultados.
- En las técnicas estadísticas (análisis de datos) se suele utilizar como criterio de calidad la verosimilitud de los datos dado el modelo. En minería de datos suele utilizar un criterio más directo, por ejemplo, utilizando el porcentaje de datos bien clasificados.
- La minería de datos está orientada a trabajar con cantidades muy grandes de datos (millones y billones de datos). En cambio la estadística no suele funcionar tan bien en bases de datos de tan gran tamaño y alta dimensionalidad.

1.1. Etapas de la Minería de Datos

El proceso de minería de datos pasa por las siguientes fases como se puede observar en la siguiente figura 1:

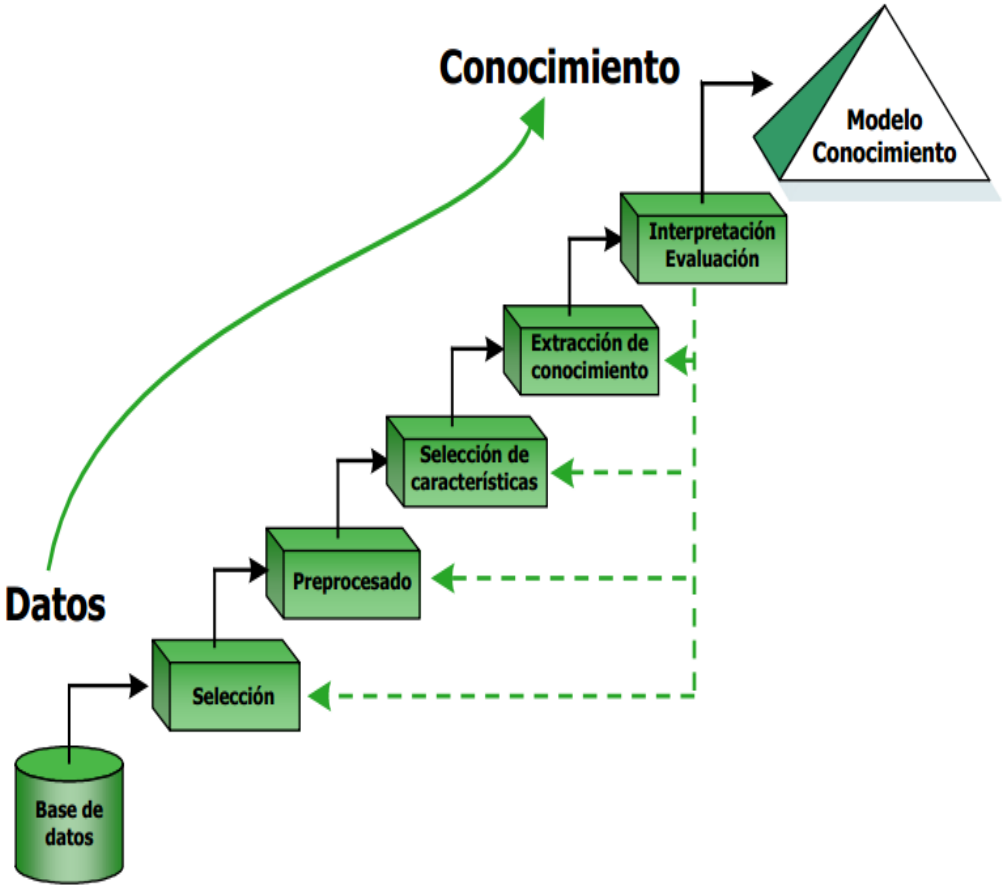


Figura 1: Etapas de la Minería de Datos [8]

A continuación se describen cada una de las etapas observadas en la figura 1:

- **Selección de datos:** En esta etapa se recoge toda la información disponible de los estudiantes. Para ello primero se debe de seleccionar el conjunto de factores que pueden afectar y después se deben de recoger a partir de las diferentes fuentes de datos disponibles. Finalmente toda esta información se debe integrar en un único conjunto de datos [9].
- **Pre-procesado.** En esta etapa se preparan los datos para poder aplicar, posteriormente, las técnicas de minería de datos. Para ello, primero se realizan tareas típicas de pre-procesado como: limpieza de datos, transformación de variables y particionado de datos. Además se han aplicado otras técnicas como

la selección de atributos y el re-balanceado de datos para intentar solucionar los problemas de la alta dimensionalidad y desbalanceo que presentan normalmente este tipo de conjuntos de datos [9].

- **Extracción de Conocimiento:** Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos [10].
- **Interpretación y Evaluación:** Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos [10].

1.2. Áreas de Aplicación de la Minería de Datos [11]:

- **Educación**
  - Selección o captación de estudiantes
  - Detección de abandonos o fracasos
  - Estimación del tiempo de estancia en la institución
- **Financieras:**
  - Detección de uso fraudulento de tarjetas de crédito
  - Predicción del gasto en tarjeta de crédito por grupos
  - Análisis de riesgos en concesión de créditos
  - Identificación de reglas de mercado a partir de datos históricos
  - Reconocimiento de clientes infieles
- **Comercio**
  - Análisis de la cesta de la compra
  - Evaluación de campañas publicitarias
  - Segmentación de clientes
  - Estimación de stocks, de costes, de ventas

- **Seguros**
  - Predicción de qué tipo de clientes contratan nuevas pólizas
  - Identificación de patrones de comportamiento para clientes con riesgo
  - Identificación de comportamiento fraudulento
- **Medicina**
  - Diagnóstico de enfermedades
  - Detección de pacientes con riesgo de sufrir una patología concreta
  - Gestión hospitalaria y asistencial. Predicciones temporales de los centros sanitarios para el mejor uso de recursos
  - Tratamiento de imágenes medicas
- **Bioinformática**
  - Búsqueda de genes
  - Predicción de la estructura secundaria de las proteínas
- **Otras áreas**
  - Telecomunicaciones: detección del fraude
  - Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo spam.
  - Hacienda: detección de fraude fiscal
  - Web: análisis del comportamiento de los usuarios, análisis de los log de un servidor web.
  - Deportes: detección riesgo de lesiones a partir de datos médicos.

1.3. Técnicas de la Minería de Datos

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos con vistas al descubrimiento de información previamente desconocida y que pueda servir de ayuda en el proceso de toma de decisiones [8].

Las técnicas más representativas son:

1.3.1. Técnicas No Supervisadas o de Descubrimiento del Conocimiento

a. **Clustering o Agrupamiento.-** Es el proceso de agrupar los datos en clases o en clústeres, de tal forma que, los datos de un mismo clúster tienen una alta similitud y a su vez, son muy diferentes de los de otro clúster [12].

Al hacer clústeres, se puede identificar regiones densas y regiones dispersas en el espacio de características, y por lo tanto, descubrir distribución de patrones y correlaciones entre los atributos [12].

b. **Reglas de Asociación.-** Es la exploración de los datos con el propósito de identificar relaciones entre los datos, dentro de una fuente o base de datos [13].

   Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos [14].

1.3.2. Técnicas Supervisadas o Predictivas

a. **Predicción.-** Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos [14]. Además comprende el uso de algunas variables o campos de la base de datos para predecir valores futuros o desconocidos, o incluso otras variables de interés. También se centran en encontrar patrones comprensibles para el ser humano que describan la información que tenemos [15].
b. **Regresión.-** Es una técnica utilizada para inter y extrapolar las observaciones, las cuales pueden clasificarse como regresión lineal o no lineal. Hablamos de modelo de regresión cuando la variable de respuesta y las variables explicativas son todas ellas cuantitativas. Si sólo disponemos de una variable explicativa hablamos de regresión simple, mientras que si disponemos de varias variables explicativas se trata de un problema de regresión múltiple [15].
c. **Arboles de Decisión.-** Son ampliamente usados y pueden ser fácilmente explicados basándose en el criterio usado para dividir los datos en las extremidades del árbol. Los árboles de decisión son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos [16].
d. **Clasificación.-** Técnica que permite encontrar modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones. Además empareja o asocia datos a grupos predefinidos [17].
e. **Redes Neuronales.-** Las redes neuronales simulan el cerebro humano mediante el aprendizaje de un conjunto de datos de formación y la aplicación del

aprendizaje para generalizar los patrones para la clasificación y predicción [16]. Las redes neuronales consisten en modelos predecibles, no lineales que aprenden a través del entrenamiento, generalizando los patrones que se encuentran en él, para clasificarlos y hacer pronósticos con ellos. Una vez la red neuronal ha sido entrenada, puede trabajar con gran cantidad de datos en una fracción del tiempo gastado por un humano. Las redes neuronales son ampliamente usadas para detectar actividades fraudulentas [16].

f. **Clasificación Bayesiana.-** Son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular [14].
g. **Lógica Borrosa.-** Surge de la necesidad de modelar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud, es decir permite el tratamiento probabilístico de la categorización de un colectivo. Así, para establecer una serie de grupos, segmentos o clases en los cuales se puedan clasificar a las personas por la edad, lo inmediato sería proponer unas edades límite para establecer tal clasificación de forma disjunta [14].
h. **Métodos estadísticos.-** El objetivo de la modelización estadística consiste en explicar el comportamiento de una variable a partir del conocimiento de otras. Subyacente al concepto de modelización está la idea de que una variable tiene una cierta variabilidad y que esta variabilidad está relacionada con el comportamiento de otras variables [15].

1.3.3. Algoritmos de Clasificación

A continuación se realiza una descripción de los algoritmos de Clasificación que se aplicaron en el presente Trabajo de Titulación.

1.3.3.1. ID3 o Induction Decision Trees.

Algoritmo desarrollado por J. Ross Quinlan en 1983. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos [18].

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

Los elementos son [18]:

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

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

Permite construir un árbol de arriba abajo, de forma directa y no ejecuta vuelta atrás (backtracking) en su búsqueda. Una vez que el algoritmo selecciona un atributo, nunca reconsiderara esta elección. Los dominios de los atributos y de las clases deben ser discretos. Usa el concepto de ganancia de información para seleccionar el atributo más útil en cada paso [19].

Para decidir qué atributo es el más apropiado a usar en cada nodo del árbol se utiliza una propiedad estadísticas llamada ganancia de información, que mide que tan bien clasifica ese atributo a los datos de entrenamiento. Así que elige el nodo del árbol que tenga mayor ganancia de información y luego expande sus ramas utilizando la misma metodología [20].

1.3.3.2. JRIP

Este algoritmo genera un listado de reglas obtenidas básicamente a partir de listas de decisión. Consiste en hacer una lista ordenada de reglas conjuntivas y evaluarlas en orden para encontrar la primera regla que se cumple sobre el ejemplo a clasificar. Una vez encontrada dicha regla se ha encontrado la regla más eficiente para ese ejemplo y es asignado con una etiqueta de valor de salida [21].

En si construye reglas de la forma si-sino empleando operadores de conjunción y disyunción [22].

1.3.3.3. C4.5

El algoritmo C4.5 construye un árbol de decisión a partir de los datos, realizando particiones recursivas de este conjunto y dicha construcción se lleva a cabo mediante la estrategia de primero en profundidad. Se realizan todas las pruebas posibles para dividir al conjunto de datos que se tiene y se selecciona aquella que presenta mayor ganancia de información. El algoritmo C4.5 simplemente es una implementación libre del C4.5 [23].

El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma, se construye un árbol de decisión mediante el algoritmo de construcción "divide

y vencerás", evaluando la información en cada caso utilizando los criterios de Entropía, Ganancia de información, según sea el caso [23].

El algoritmo C4.5 con respecto a ID3 es, que C4.5 puede manejar atributos continuos. Este se basa en la utilización del criterio radio de ganancia. De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección [24].

Antes de realizar la partición de datos, el algoritmo toma en cuenta todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información [25].

Mejoras del algoritmo C4.5 [25]:

- Se manejan atributos continuos.
- Se mejora la eficiencia computacional.
- Se lleva un control de qué tan profundo va a ser el tamaño del árbol de decisión construido.
- Reducir errores en la poda.
- Se evita el sobreajuste (overfitting) de datos, esto es que se aprendan a clasificar demasiado bien los datos de prueba, entonces al momento de mostrar ejemplos desconocidos, este no los clasifique de la misma forma que clasifico los ejemplos de prueba.
- Manejo de atributos con diferentes valores.
- Manejo de datos de entrenamiento con valores desconocidos.

1.3.3.4. RIDOR (Ripple Down Rule)

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 lograr disminuir el error. Luego genera

una expansión similar a un árbol de excepciones. La excepción es un conjunto de reglas que predice clases. Este algoritmo es usado para generar dichas excepciones [26].

1.3.3.5. PART

Evita el paso de optimización global que se usa en las reglas del C4.5, genera una lista de decisión sin restricciones usando el procedimiento de divide y vencerás. Además construye un árbol de decisión parcial para obtener una regla. Para poder podar una rama (una regla) es necesario que todas sus implicaciones sean conocidas. El PART evita la generalización precipitada, y usa los mismos mecanismos que el C4.5 para construir un árbol. La hoja con máxima cobertura se convierte en una regla y los valores ausentes de los atributos se tratan como en el C4.5, es decir, la instancia se divide en piezas [27].

En cuanto al tiempo máximo para generar una regla, es el mismo que para construir un árbol podado, y esto ocurre cuando los datos tienen ruido. En el mejor de los casos el tiempo necesario es el mismo que para generar una regla sencilla, y esto se da cuando los datos no presentan ruido [27].

Para elaborar una regla se construye un árbol de decisión podado para el conjunto de instancias en cuestión, se toma la hoja que cubra el mayor número de instancias y se transforma en regla, y se descarta el árbol [28].

2. CAPITULO II: HERRAMIENTAS DE MINERÍA DE DATOS

Las herramientas de MD sirvieron para extraer conocimientos desde base de datos que contienen grandes cantidades de información

2.1. WEKA

Es una herramienta que permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos [29].

Ventajas [30]:

- Es un software desarrollado bajo licencia GNU-GPL.
- Contiene una extensa colección de técnicas para preprocesamiento y modelado de datos.
- Soporta varias tareas de minería de datos, especialmente preprocesamiento, agrupamiento, clasificación, regresión, visualización y selección.
- Proporciona acceso a Bases de Datos usando SQL, gracias a la conexión “Java Database Connectivity” (JDBC).

WEKA contiene métodos de clasificación, regresión, clustering y reglas de asociación [30].

2.1.1. Entornos de trabajo de WEKA

Define 4 entornes de trabajo:

- **Simple CLI:** la interfaz "Command-Line Interfaz" es simplemente una ventana de comandos java para ejecutar las clases de WEKA [31].
- **Explorer:** permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos [32]. También posee 6 sub-entornos de ejecución [33]:
  - Preprocess: Incluye las herramientas y filtros para cargar y manipular los datos.
  - Classify: Acceso a las técnicas de clasificación y regresión
  - Cluster: Integra varios métodos de agrupamiento
  - Associate: Incluye técnicas de reglas de asociación
  - Select Attributes: Permite aplicar diversas técnicas para la reducción del número de atributos.
  - Visualize: En este apartado podemos estudiar el comportamiento de los datos mediante técnicas de visualización.
- **Experimenter:** Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala [33].
  Es un entorno grafico que permite al usuario crear, ejecutar, modificar y analizar experimentos sobre tareas de clasificación de un modelo ágil y eficaz [34].
- **KnowledgeFlow:** Permite generar proyectos de minería de datos mediante la generación de flujos de información [33].

2.2. RAPID MINER

Es un entorno de código abierto para aprendizaje automático y minería de datos. Permite realizar todos los procesos que intervienen en un proyecto: la adquisición de datos, la transformación de los datos, la selección de datos, la selección de atributos, la transformación de los atributos, el aprendizaje/modelización y la validación [35].

Además permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico [36]. Lo que hace posible aumentar la productividad a través de modelos que solucionan los problemas de predicción, clasificación y segmentación de la información [37].

Ventajas [37]:

- Está desarrollado en Java.
- Es multiplataforma.
- Representación interna de los procesos de análisis de datos en ficheros XML.
- Contiene más de 500 técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva.
- Permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML.
- Puede usarse de diversas maneras:
  - A través de un GUI.
  - En línea de comandos.
  - En batch (lotes)
  - Desde otros programas, a través de llamadas a sus bibliotecas.
- Incluye gráficos y herramientas de visualización de datos.
- Software de código abierto.

2.3. ORACLE DATA MINING

Es una herramienta de software desarrollada por la empresa Oracle para aplicar técnicas de minería de datos a grandes volúmenes de datos [38].

Permite a los analistas de datos para trabajar directamente con los datos dentro de la base de datos, explorar los datos gráficamente, construir y evaluar varios modelos de

minería de datos, aplicar los modelos de minería de datos Oracle con los nuevos datos y desplegar Oracle predicciones y perspectivas de minería de datos [38].

Permite analizar los datos, explorar los datos, construir y evaluar modelos y aplicar estos modelos a nuevos datos, así mismo integra todas las etapas del proceso de la minería de datos [36].

ODM funciona dentro de la base de datos de Oracle, así que no hay necesidad de exportar los archivos a un paquete de software estadístico fuera de la base de datos, lo que reduce los costos y permite mejorar la eficiencia. Con un lenguaje de procedimiento integrado/ lenguaje de consulta estructurado (PL / SQL) permite a los usuarios construir modelos [36].

Es la herramienta más potente para trabajar con bases de datos de Oracle, si bien habrá que pagar una licencia por su uso [36].

2.4. R

R es un entorno de software libre para el cálculo estadístico y gráficos. Se proporciona una amplia variedad de técnicas estadísticas y gráficos [39].

R puede ser extendido fácilmente a través de paquetes. Hay alrededor de 4000 paquetes disponibles en el repositorio de paquetes CRAN [39].

Es orientado a objetos e interpretado, por lo tanto permite al usuario interactuar con la línea de comandos al mismo tiempo que crea gráficos vectoriales de alta calidad [40].

Ventajas [40]:

- Un conjunto integrado de herramientas de análisis de datos.
- Está disponible de manera gratuita para múltiples plataformas.
- Posee funciones estadísticas que son útiles para minería de datos.
- Funciones gráficas para análisis y visualización de los datos.
- Un buen gestor de datos.
- Un conjunto de operadores para cálculos en arrays (vectores de gran tamaño)

2.4.1. Paquetes de R para minería de datos:

- Rattle: El paquete de Rattle, es uno de una serie de herramientas analíticas para el análisis datos [41]. Lleva juntos una multitud de paquetes de R que son

esenciales para la minería de datos, pero a menudo no es fácil para el principiante de usar [41].

Permite cargar de diferentes formatos los datos para trabajar en minería de datos [42].

- FactoMineR: Es un R paquete dedicado al análisis de datos multivariados.

  Las principales características de este paquete son [43]:

  - La posibilidad de tener en cuenta los diferentes tipos de variables (cuantitativas o categóricas).
  - Diferentes tipos de estructura de los datos (una partición de las variables, una jerarquía de las variables, una partición en los individuos) y finalmente, la información complementaria (individuos y las suplementarias).
  - Es un paquete de R dedicada al análisis multivariado exploratorio de datos.

3. CAPITULO III: METODOLOGÍAS PARA MINERÍA DE DATOS

El objetivo principal de las metodologías es tener un proceso o pasos estándares para la resolución de problemas con las herramientas y técnicas de minería de datos. Por ende que surgieron metodologías, las cuales se describen a continuación [44]:

3.1. METODOLOGÍAS

3.1.1. CRISP-DM (Cross-Industry Standard Process for Data Mining)

La metodología CRISP – DM es una guía para el desarrollo de proyectos enfocados a la Minería de Datos. Esta metodología puede trabajar con cualquier herramienta para desarrollar el proyecto, es decir es una metodología equitativa [45], [46].

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas y como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos [47].

3.1.1.1. Fases de CRISP-DM

A continuación se describe cada una de las fases [48-50].

a. **Entendimiento del negocio.-** Esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio, para convertir este conocimiento en un problema de definición de minería de datos y un plan preliminar diseñado para alcanzar los objetivos.
b. **Entendimiento de los datos.-** Esta fase inicia con una colección inicial de datos y procede con actividades para familiarizarse con ellos, identificar problemas de calidad en los mismos, descubrir una primera idea de estos o detectar conjuntos interesantes que permitan formar hipótesis en la búsqueda de información escondida.
c. **Preparación de los datos.-** Cubre todas las actividades para construir la base final de datos. Es preferible que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Estas tareas incluyen tabulación, documentación y selección de atributos, también como transformación y limpieza de datos para las herramientas de modelado.
d. **Modelado.-** Se seleccionan y aplican varias técnicas, y sus parámetros son calculados a los valores óptimos. Por lo general hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por lo tanto será a menudo necesario devolverse a la fase de preparación de datos.
e. **Evaluación.-** Al llegar a esta fase se ha construido un modelo (o modelos) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la entrega final del modelo es importante evaluarlo y revisar los pasos ejecutados para construirlo, de tal forma que este lo más cercano posible de alcanzar los objetivos del negocio.
f. **Despliegue.-** La creación del modelo por lo general no es el final del proyecto. Incluso si el propósito del modelo es incrementar conocimiento sobre los datos, el conocimiento ganado necesitará ser organizado y presentado de una manera que el cliente lo pueda usar. A menudo implica aplicar modelos en vivo dentro del proceso de toma de decisiones de una organización.

3.1.2. SEMMA (Simple, Explore, Modify, Model y Assess)

Es el proceso de selección, exploración y modelado de grandes conjuntos de datos para descubrir patrones desconocidos [48], [51].

3.1.2.1. Fases de SEMMA

Las fases son [49]:

a. **Muestreo.-** en la que se extrae una muestra representativa de la población sobre la cual se aplicará el análisis. La etapa de muestreo es opcional, aconsejable cuando el tamaño del conjunto de datos es demasiado extenso.
b. **Exploración.-** en donde se procede a realizar una exploración de la información para simplificar el problema y así optimizar la eficiencia del modelo. Se pretende determinar cuáles son las variables explicativas que se utilizarán como entradas del modelo.
c. **Modificación.-** en la cual se modifican los datos de la base para que tengan el formato adecuado para la entrada del modelo.
d. **Modelado.-** que permite modelar los datos permitiendo al software la búsqueda automática de una combinación de datos que predicen confiablemente las salidas deseadas. En esta etapa se debe establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Las técnicas utilizadas para el modelado incluyen métodos estadísticos tradicionales, redes neuronales, técnicas adaptativas, lógica difusa, árboles de decisión, reglas de asociación y computación evolutiva.
e. **Valoración.-** que consiste en la valoración de los datos evaluando usabilidad y confiabilidad de lo encontrado en el proceso y estimando que tan bien se comporta haciendo una comparación con otros métodos estadísticos.

3.2. Comparación de Metodologías

- Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Explotación de Datos en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Explotación de Datos en un proceso iterativo e interactivo [52].

- La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Explotación de Datos donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema [52], [53].
- Otra diferencia significativa entre la metodología SEMMA y la metodología CRISP-DM radica en su relación con herramientas comerciales. La metodología SEMMA sólo es abierta en sus aspectos generales ya que está muy ligada a los productos SAS donde se encuentra implementada. Por su parte la metodología CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Explotación de Datos siendo su distribución libre y gratuita [53].
- Una gran diferencia se encuentra en el comienzo del proyecto que será analizado ya que para la metodología SEMMA se centra más en las características técnicas en cada proceso, donde empieza con el muestreo de los datos; en cambio la metodología CRISP -DM realiza la minería de datos, de manera más amplia de acuerdo a los objetivos empresariales y el análisis del problema [54].
- SEMMA está especialmente enfocada al desarrollo del modelo de minería, y quedan fuera de su alcance otros aspectos del proyecto como el conocimiento del problema en estudio o la planificación de la implementación [54].
- Las dos metodologías identifican técnicas de explotación de información utilizables, solo CRISP-DM identifica los distintos problemas de inteligencia de negocio y hace una caracterización parcialmente abstracta de los mismos [49].

3.3. Elección de la Metodología

La metodología a utilizar para el presente Trabajo de Titulación es CRISP-DM ya que es más completa que SEMMA, así mismo es importante mencionar que CRISP-DM contiene una etapa enfocada al entendimiento del negocio y es más flexible ya que se adapta a cualquier herramienta de Explotación de Datos.

e. Materiales y Métodos

Para el presente Trabajo de Titulación denominado “Estudio del Rendimiento Académico aplicando Técnicas de Minería de Datos”, los métodos utilizados están basados en la investigación bibliográfica porque se realizó un análisis de los problemas que afectan el rendimiento académico basándose en fuentes bibliográficas confiables y casos de éxito. Además es un proyecto de desarrollo porque se implementó un modelo que permita estimar en rendimiento académico de los estudiantes.

Para la recolección de la información se utilizó el Método Científico para la formulación de la revisión literaria, en donde se indicó temas como: Casos de Éxito, Rendimiento Académico, Técnicas de Minería de Datos, Herramientas de Minería de Datos que estén enfocados en el proceso del presente proyecto. También el Método Deductivo para ayudar a conocer los problemas de las universidades sobre el rendimiento académico de los estudiantes y a través de esto las universidades tomen decisiones que permitan mejoras. Así mimo el Método Inductivo se lo utilizó para obtener información académica de cada uno de los estudiantes y hacer un análisis de cada uno de los inconvenientes que tienen para poder obtener el problema general de estudio, y así enfocarnos directamente en resolver dicho problema.

Además se utilizó técnicas para recopilar información como: Bibliográfica para la revisión de diferentes fuentes de información confiables enfocados en el tema de Trabajo de Titulación con el fin de detectar problemas y causas que afectan el rendimiento académico. También la técnica de Observación permitió conocer la realidad académica de los estudiantes, así como también permitió seguir obteniendo información necesaria a lo largo del desarrollo del presente proyecto.

También se empleó la metodología donde se determinó una secuencia de pasos ordenados que nos permitan cumplir los objetivos del Trabajo de Titulación con el fin de obtener los resultados esperados; la misma que es Cross-Industry Standard Process for Data Mining, CRISP – DM que es una guía para el desarrollo de proyectos enfocados a la Minería de Datos. Como metodología, incluye descripciones de las fases normales de

un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

f. Resultados

Para el desarrollo del presente Trabajo de Titulación se definieron fases que permitieron cumplir con los objetivos planteados, además cada una de las fases también están basadas en la metodología CRISP-DM; a continuación se describen cada una de ellas y sus actividades:

1. PRIMERA FASE: Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.

1.1. Recopilar información de fuentes académicas, artículos científicos sobre las diversas técnicas de Minería de Datos que permitan determinar el rendimiento académico.

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos con vistas al descubrimiento de información previamente desconocida y que pueda servir de ayuda en el proceso de toma de decisiones [8]. Además, existen investigaciones enfocadas a la utilización de Técnicas de Minería de Datos (ver TABLA I) en diversas áreas y que serán de gran importancia para poder determinar el rendimiento académico. Es por ello en base a estas investigaciones se analizó casos de éxito para poder determinar el rendimiento académico.

TABLA I.

TÉCNICAS DE MINERÍA DE DATOS

TÉCNICAS DE MINERIA DE DATOS	REFERENCIAS
Clustering o Agrupamiento	[3][55-57]
Reglas de Asociación	[3] [58 -59]
Clasificación: Árboles de Decisión	[1] [60 - 63]

Estas Técnicas de Minería de Datos se describen a continuación, donde se realizó una revisión bibliográfica donde se describirá su funcionamiento y además se mencionó casos de éxito donde se aplique las Técnicas de MD.

1.1.1. Agrupamiento o Clustering.- es el conjunto de procedimientos diseñados para encontrar grupos naturales basados en las similitudes presentes en un conjunto de patrones. Para que las técnicas de análisis de clúster sean eficientes debe existir algún tipo de similitud entre los datos [55].

Considerando al agrupamiento de datos como una tarea de clasificación no supervisada, la gran mayoría de métodos pueden ser divididos en: 1) jerárquicos, que producen una serie de particiones anidadas basadas en un criterio para unir o dividir grupos basados en su similitud, y 2) particionales, que identifican la partición que optimiza un criterio agrupamiento [56].

1.1.1.1. Casos de Éxito de Agrupamiento o Clustering

Existen investigaciones donde utilizan Clustering o Agrupamiento para determinar el rendimiento académico (ver TABLA II).

TABLA II.

CASOS DE ÉXITO DE CLUSTERING

Caso de Éxito	Descripción
Exploración de Datos Académicos a través de la aplicación de Técnicas de Minería de Datos en Weka.	La presente investigación expone la aplicación del proceso denominado Descubrimiento de Conocimiento en Bases de Datos (KDD), conocido como Minería de Datos (MD), sobre la información académica de la Universidad Gastón Dachary (UGD). Dicho proceso consiste en una serie de etapas iterativas que incluyen el pre y post procesamiento de datos, hasta obtener conocimiento nuevo. Para ello, se realizaron numerosas selecciones y

	depuraciones de datos, utilización de diferentes criterios de representación y aplicación de diferentes técnicas y algoritmos. La fuente de datos proviene de la información proporcionada al ingreso (personales y antecedentes educativos) y durante el lapso de sus estudios; con la debida protección de datos personales, creando una vista minable con las características de las titulaciones seleccionadas, una colección de individuos sobre los cuales se realizó el estudio para extraer conocimiento útil en lo que se refiere a rendimiento académico, correspondiente un periodo de 10 años (1999-2009). Se ha detectado mediante ciertos algoritmos correspondientes a las técnicas de asociación, clustering, selección de atributos y clasificación, que existen tendencias y relaciones entre los datos pertenecientes a los departamentos evaluados; existiendo particularidades y coincidencias entre estos [3].
Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa.	En el presente artículo se presenta la implementación del algoritmo denominado C-Means para la agrupación de datos en conjuntos difusos, como técnica de minería de datos, esta técnica se implementó en el programa SM2D 1.2 Beta (Software Minería Datos Difusa), y se presenta como ejemplo el análisis del rendimiento académico de la asignatura fisiología vegetal [57].

1.1.2. Reglas de Asociación.- Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos [58].

Las reglas de asociación son parecidas a las reglas de clasificación. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y coocurrencias de eventos [59].

1.1.2.1. Casos de Éxito Reglas de Asociación

TABLA III.
CASOS DE ÉXITO DE REGLAS DE ASOCIACIÓN

Caso de Éxito	Descripción
Exploración de Datos Académicos a través de la aplicación de Técnicas de Minería de Datos en Weka.	La presente investigación expone la aplicación del proceso denominado Descubrimiento de Conocimiento en Bases de Datos (KDD), conocido como Minería de Datos (MD), sobre la información académica de la Universidad Gastón Dachary (UGD). Dicho proceso consiste en una serie de etapas iterativas que incluyen el pre y post procesamiento de datos, hasta obtener conocimiento nuevo. Para ello, se realizaron numerosas selecciones y depuraciones de datos, utilización de diferentes criterios de representación y aplicación de diferentes técnicas y algoritmos. La fuente de datos proviene de la información proporcionada al ingreso (personales y antecedentes educativos) y durante el lapso de sus estudios; con la debida protección de datos personales, creando una vista minable

	con las características de las titulaciones seleccionadas, una colección de individuos sobre los cuales se realizó el estudio para extraer conocimiento útil en lo que se refiere a rendimiento académico, correspondiente un periodo de 10 años (1999-2009). Se ha detectado mediante ciertos algoritmos correspondientes a las técnicas de asociación, clustering, selección de atributos y clasificación, que existen tendencias y relaciones entre los datos pertenecientes a los departamentos evaluados; existiendo particularidades y coincidencias entre estos [3].

1.1.3. Clasificación.- Sirve para identificar las características que indican el grupo al que cada caso pertenezca. Este modelo puede ser usado para comprender los datos existentes y para predecir cómo se comportará algún nuevo caso [60].

La minería de datos crea los modelos de clasificación examinando datos que a su vez se encuentran ya clasificados como casos, e inductivamente encuentra un modelo que predice. Estos casos ya clasificados pueden venir de una base de datos histórica. Estos datos se pueden extraer de una muestra experimental de la base de datos [60].

1.1.3.1. Casos de Éxito Clasificación

Existen investigaciones en diversas áreas donde utilizan Árboles de Decisión para determinar el rendimiento académico (ver TABLA IV).

TABLA IV.

CASOS DE ÉXITO ÁRBOLES DE DECISIÓN

Caso de Éxito	Descripción
Validación de los predictores del rendimiento académico obtenidos mediante minería de datos usando análisis de componentes principales.	A partir de la aplicación de un grupo de técnicas de Minería de Datos como el clustering, los árboles de decisión y algoritmos de aprendizaje inductivo; se pretende clasificar a los estudiantes de acuerdo a su rendimiento académico, para posteriormente encontrar patrones ocultos y reglas que los caractericen; basado en las relaciones que se establecen entre el centro de procedencia de los estudiantes, nivel de escolaridad de los padres y provincia de origen con sus resultados académicos en el primer curso en la universidad. Estos resultados pueden mejorar el proceso de formación académica y elevar la calidad de la educación en la Universidad de las Ciencias Informáticas (UCI) [1].
Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.	En este trabajo presenta un análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia (UPV) a lo largo de tres cursos,

	Este análisis relaciona el rendimiento con las características socioeconómicas y académicas de los alumnos, que se obtienen en el momento de su matrícula, y que se recogen en la base de datos de la universidad. Para el estudio utiliza técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, etc. De entre las técnicas de minería de datos existentes, hemos utilizado dos de ellas para generar los modelos predictivos del rendimiento: los árboles de decisión y la regresión multivariante [61].
Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil.	Este artículo presenta los resultados de la evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008. En este trabajo

	se utiliza: la clasificación como tipo de tarea de minería, el árbol de decisión como tipo de modelo, el J48 (implementación en Weka del algoritmo C4.5) y el FT como algoritmos de minería [62].
Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios.	Este proyecto tiene por objetivo construir modelos predictivos del rendimiento académico de los estudiantes de las diversas carreras de la FACENA de la UNNE. Para la formulación y ajustes de los modelos de predicción, se utilizarán alternativamente, técnicas de minería de datos clásicas y métodos simbólicos o inteligentes, evaluando su desempeño en la predicción del rendimiento académico de los alumnos. Los resultados obtenidos a partir del desarrollo de este proyecto, constituirán un aporte significativo para los procesos de evaluación y acreditación universitarios, los elementos proporcionados por el análisis del rendimiento del alumnado contribuirá a la mejora de la calidad del sistema educativo [63].

1.2. Elaborar un análisis comparativo de las diversas Técnicas de Minería de Datos.

Las diferentes técnicas de Minería de datos mencionadas anteriormente son factibles para determinar el rendimiento académico, pero para la elección de la más idónea se realizó un análisis comparativo de cada una de ellas (ver TABLA V).

TABLA V.

ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS

Técnica de Minería de Datos	Análisis
Clustering o Agrupamiento	Es el conjunto de procedimientos diseñados para encontrar grupos naturales basados en las similitudes presentes en un conjunto de patrones [55]. Los métodos como el Agrupamiento difuso asocia cada patrón con todos los grupos usando una función de pertenencia [56]. Existen algoritmos de clustering como el Simple K-Means que permite definir el número de clusters que se desean obtener. También el algoritmo Cobweb, DBSCAN, entre otros. También los casos de éxito mencionados anteriormente utilizan esta técnica para determinar el rendimiento académico (ver TABLA II).
Reglas de Asociación	Las reglas de asociación se utilizan para descubrir hechos que ocurren en común

	dentro de un determinado conjunto de datos [58]. Esta técnica es más utilizada en soporte para la toma de decisiones, diagnóstico y predicción de alarmas en telecomunicaciones y análisis de información de ventas [59]. Uno de los algoritmos de esta técnica más utilizado es el algoritmo A priori [64] donde se definen reglas de minería de datos para grandes transacciones sobre bases de datos [59, 64]. Existen pocos trabajos realizados a través de esta técnica para determinar el rendimiento académico (Ver TABLA III).
Clasificación: Árboles de Decisión	Sirve para identificar las características que indican el grupo al que cada caso pertenezca. Este modelo puede ser usado para comprender los datos existentes y para predecir cómo se comportará algún nuevo caso [60]. La minería de datos crea los modelos de clasificación examinando datos que a su vez se encuentran ya clasificados como casos, e inductivamente encuentra un modelo que predice. Estos casos ya clasificados pueden venir de una base de datos histórica. Estos datos se pueden

	extraer de una muestra experimental de la base de datos [60]. Así mismo, existen investigaciones que permiten determinar el rendimiento académico a través de la clasificación utilizando la técnica árboles de decisión (Ver TABLA IV). Un árbol de decisión es un conjunto de condiciones o reglas organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas [65]. Además, el árbol de decisión contiene algoritmos como el J48 o C4.5 que permiten determinar el rendimiento académico y es el más utilizado, este algoritmo genera una estructura de reglas y evalúa su bondad usando criterios que miden la precisión en la clasificación de los casos. Algoritmos como ID3, JRIP, RIDOR, PART, C4.5, entre otros.

1.3. Seleccionar la técnica de Minería de Datos que permita identificar de mejor manera el rendimiento académico.

Las técnicas de Minería de Datos analizadas en la TABLA V permiten analizar e identificar factores sobre el rendimiento académico y son una gran ayuda para que autoridades puedan tomar decisiones y mejorar el nivel de educación.

Por ende, en base al análisis realizado en la TABLA V la técnica más idónea es: Clasificación que serán necesarias para determinar el rendimiento académico basados en los datos académicos de los estudiantes. Además existen diferentes casos de éxito donde utilizan estas técnicas y que han sido eficientes para determinar el rendimiento académico (ver TABLA II y IV). Así mismo, esta técnica sirvió de apoyo para realizar el presente proyecto.

1.4. Evaluar la técnica de Minería de Datos para comprobar si se adapta al entorno en el que se va a trabajar.

La elección de la Técnica de Minería de Datos como Clasificación se adaptan al problema a resolver, debido a que son fundamentales ya que generan un conjunto de condiciones o reglas organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones y permitir dar conclusiones [55, 65].

Además en la sección de Resultados se aplican algunos de los algoritmos para poder determinar el rendimiento académico que corresponden a esta técnica (ver Sección 3.4. Implementar el modelo computacional en la herramienta de Minera de Datos seleccionada).

2. SEGUNDA FASE: Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.

2.1. Analizar indicadores que permitan estimar el rendimiento académico.

2.1.1. COMPRESIÓN DEL NEGOCIO

2.1.1.1. Objetivos de negocio.

La Universidad Nacional de Loja (UNL) es una Institución de Educación Superior cuya misión es la formación académica y profesional, con sólidas bases científicas y técnicas, pertinencia social y valores; la generación y aplicación de conocimientos científicos, tecnológicos y técnicos, que aporten al desarrollo integral del entorno y al avance de la ciencia; el fortalecimiento del pensamiento, la promoción, desarrollo y difusión de los saberes y culturas; y, la prestación de servicios especializados.

Es por ende que la idea principal del presente Trabajo de Titulación (TT) es Determinar el Rendimiento Académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos.

La determinación del rendimiento académico es fundamental para poder obtener resultados confiables sobre el desempeño académico de los estudiantes universitarios. La UNL cuenta con valiosa información académica de los estudiantes, cuyo problema radica que dicha información no es analizada para determinar el desempeño académico de los estudiantes.

A partir de esto es necesario aplicar Técnicas de Minería de Datos que permitan realizar un estudio con la finalidad de evaluar el desempeño académico de los estudiantes.

Los objetivos del negocio son:

- ✓ Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.
- ✓ Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.
- ✓ Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.

2.1.1.2. Criterios de Éxito

- ➢ El Análisis las Técnicas de Minería de Datos permitirá seleccionar la que más se adapte a los datos académicos de los estudiantes universitarios.
- ➢ El diseño de un Modelo Computacional se desarrollará en base a los datos académicos de los estudiantes de la Carrera de Ingeniería en Sistemas del Área de la Energía, Las Industrias y los Recursos Naturales No Renovables de la UNL.
- ➢ La implementación del Modelo Computacional ayudará a obtener resultados que permitan determinar cuál es el desempeño académico de los estudiantes.

2.1.1.3. Evaluación de Resultados

Se realiza un análisis de los recursos, requerimientos, supuestos y restricciones que son útiles en la realización del TT los mismos que se detallan a continuación.

2.1.1.4. Recursos

a. Talento Humano

- Director de Tesis: es el encargado de revisar los avances del proyecto y dar tutorías que permitan cumplir con el TT de manera eficiente.
- Investigador: la función es recolección de información, casos de éxito, recolección de datos académicos, análisis de técnicas de Minería de Datos, implementación del modelo computacional y uso de herramientas de Minería de Datos.

b. Recursos Hardware

- Computador: Utilización de computadora para realizar la documentación del TT, para guardar en un repositorio los datos académicos de los estudiantes y además posee las características necesarias para la realización del TT.
- Impresora: obtención de la documentación física de los avances del TT.
- Disco Duro.- guardar la información de cada uno de los avances del proyecto y también realizar un respaldo de los datos académicos de los estudiantes.

c. Recursos Software

- MySQL: Es un sistema gestor de base de datos donde está alojada la Base de Datos de la información académica de los estudiantes.
- Paquete de Ofimática de Microsoft Office 2010 (licencia estudiantes): Serán utilizados para la digitalización de cada una de las fases del TT.

d. Servicios

- Internet: utilizado para consultar información confiable referente al TT.

e. Transporte

- Bus y Taxi: Para trasladarse a la Universidad Nacional de Loja para recibir tutorías y además presentar avances del Trabajo de Titulación.

f. Materiales

- Resma de papel: Para poder imprimir sobre las hojas la documentación de los avances del TT y también para la documentación final.
- Cartuchos de Tinta: Para poder utilizarlos al momento de imprimir la documentación del TT.

- Copias: Copias de la documentación del TT.
- Carpetas: Para poder adjuntar la documentación del TT.
- CD: Para guardar la información del TT.

g. Datos

La recolección de la información académica de los estudiantes se obtuvo a través del Web Service de la UNL donde esta almacenada esta información, donde el Director de la Unidad de Telecomunicaciones e Información (UTI) proporciono los permisos para acceder a los datos académicos de los estudiantes.

2.1.1.5. Riesgos y Contingencia

Se identifica los riesgos que pueden ocurrir en el proyecto describiendo las consecuencias, y se sugieren acciones que pueden ser tomadas para reducir al mínimo tales riesgos. En la siguiente tabla (ver TABLA VI) se describe los riesgos y que acciones tomar frente a ellos.

TABLA VI.

RIESGOS Y CONTINGENCIAS

Riesgos	Contingencias
Pérdida de información	Realizar respaldo a cada uno de los avances del proyecto de tesis, subiéndolo a un repositorio en línea y almacenando en un disco externo.
Pérdida de tiempo en actividades del TT	Considerar el tiempo establecido en cada una de las actividades para poder cumplir con el cronograma establecido, tomando en consideración aquellas que causen retrasos en el mismo.
Datos faltantes	Realizar un análisis de los datos con el fin de comprobar que estén todos los parámetros necesarios.

2.1.1.6. Terminología

Permite describir términos para que sean comprendidos de mejor manera:

a. Términos del Negocio

- Rendimiento Académico: Es el resultado cuantitativo que se obtiene en el proceso de aprendizaje de conocimientos, conforme a las evaluaciones que realiza el docente mediante pruebas objetivas y otras actividades complementarias [66].
- UNL: Universidad Nacional de Loja
- UTI: Unidad de Telecomunicaciones e Información
- TT: Trabajo de Titulación
- Web Service: es un servicio ofrecido por una aplicación que expone su lógica a clientes de cualquier plataforma mediante una interfaz accesible a través de la red utilizando tecnologías (protocolos) estándar de internet [67].

b. Términos de Minería de Datos

- Minería de Datos: es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la Minería de Datos es encontrar modelos inteligibles a partir de los datos [9].
- Base de Datos: es un almacén donde guardamos una colección o conjunto de informaciones (texto, imagen, sonido, video) las cuales se encuentran relacionadas entre sí y pueden ser accesibles y consultadas en cualquier momento. [68].
- Sistemas de Gestión de Base de Datos: es una aplicación que permite a los usuarios definir, crear y mantener la BD y proporciona un acceso controlado a la misma [69].
- Técnicas de Minería de Datos: sirven para la obtención de patrones en los datos y para la extracción de información valiosa [70].

- Modelo Computacional: representaciones que pueden ser procesadas por un computador y así proveer datos simulados, comparables con datos obtenidos de personas reales [71].

2.1.1.7. Presupuesto

A continuación se detalla el presupuesto que involucra el desarrollo del Trabajo de Titulación denominado “Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos”.

a. Talento Humano: Para llevar a cabo el Trabajo de Titulación es indispensable contar con Talento Humano que es quien desarrollará el Trabajo de Titulación y un tutor para que guíe en el proceso del mismo. En la siguiente tabla (ver TABLA VII) se detalla los valores de acuerdo al Ministerio de Relaciones Laborales [72]:

TABLA VII.

TALENTO HUMANO

TALENTO HUMANO			
TALENTO HUMANO	**HORAS**	**PRECIO / H ($)**	**V. TOTAL ($)**
Investigador	1340	5,00	6.700,00
Director de Tesis	120	8,00	960,00
		SUBTOTAL	**7.660,00**

b. Recursos Físicos: La computadora portátil se utilizará para ir desarrollando y documentando el Trabajo de Titulación, los mismos se detallan en la tabla (ver TABLA VIII) a continuación:

TABLA VIII.

RECURSOS FÍSICOS

RECURSOS FÍSICOS					
RUBRO	**CANT.**	**VALOR ($)**	**TIEMPO / M**	**PRECIO / M ($)**	**V. TOTAL ($)**
Computador Portátil Hp dv4-1120us	1	1.100,00	11	20	220,00
Flash Memory 4GB	1	12,00	11	1,00	11,00
Disco Duro Samsung (1TB)	2	100,00	11	8,00	88,00
Impresora	1	50,00	11	2.50	27,50
				SUBTOTAL	**346.50**

c. Los Recursos Software: son necesarios para ir digitalizando el Trabajo de Titulación y la herramienta para implementar el modelo que se desarrollará y de esta manera poder visualizar los resultados (ver TABLA IX).

TABLA IX.

RECURSOS SOFTWARE

RECURSOS SOFTWARE			
RUBRO	**CANT.**	**V. UNITARIO ($)**	**V. TOTAL ($)**
Paquete de Ofimática de Microsoft Office 2013 (licencia estudiantes)	1	140,00	140,00
MySQL	1	0.00	0.00
RAPIDMINER	1	0.00	0.00
		SUBTOTAL	**140,00**

d. Servicios: el internet son útiles para realizar consultas sobre casos de éxito y temas enfocados con el Trabajo de Titulación (ver TABLA X).

TABLA X.

SERVICIOS

SERVICIOS				
RUBRO	**CANT. HORAS POR MES**	**V. UNITARIO/HORA ($)**	**CANT./MES**	**V.TOTAL ($)**
Internet	240 H	0,60	10	144,00
Llamadas telefónicas	1 H	0,15	10	90,00
SUBTOTAL				**234,00**

e. Transporte: son medios necesarios para poder trasladarse a la Universidad para recibir tutorías y además presentar avances del Trabajo de Titulación (ver TABLA XI).

TABLA XI.

TRANSPORTE

TRASPORTE				
RUBRO	**CANT.**	**V. UNITARIO ($)**	**CANT./MES**	**V.TOTAL ($)**
Bus	40 pasajes	0,25	10	100,00
Taxi	5 carreras	1,50	10	75,00
			SUBTOTAL	**175,00**

f. Recursos Materiales: son fundamentales para evidenciar de manera física los avances del Trabajo de Titulación como se puede observar en la tabla (ver TABLA XII).

TABLA XII.

RECURSOS MATERIALES

RECURSOS MATERIALES			
RUBRO	**CANT.**	**V. UNITARIO ($)**	**V.TOTAL ($)**
Resma de papel	6	4,00	24,00
Cartuchos de Tinta	3	15,00	45,00
Perfiles	6	0,45	2,70
Copias	400	0.02	8,00
Carpetas	6	0,30	1,80
CD's	7	0,60	4,20
		SUBTOTAL	**85,70**

g. Datos: La recolección de los datos académica de los estudiantes se obtuvo a través del Web Service de la UNL donde esta almacenada esta información, donde el Director de la Unidad de Telecomunicaciones e Información (UTI) proporciono los permisos para acceder a los datos académicos y personales de los estudiantes (ver TABLA XIII).

TABLA XIII.

RECURSOS DATOS

DATOS	
DATOS	**V.TOTAL ($)**
Datos Académicos	0,00
SUBTOTAL	**0,00**

El Presupuesto Total será financiado por el desarrollador del Trabajo de Titulación, a continuación se resume los gastos que involucra el mismo (observar TABLA XIV).

TABLA XIV.
PRESUPUESTO TOTAL

PRESUPUESTO TOTAL	
Talento Humano	7660,00
Recursos Físicos	346,50
Recursos Software	140,00
Servicios	234,00
Transporte	175,00
Recursos Materiales	85,70
Datos	0,00
TOTAL	8640,20
IMPREVISTOS (10% DEL TOTAL)	864,02
TOTAL PRESUPUESTO + IMPREVISTOS	**9504,22**

2.1.1.8. Plan del Proyecto

Se describe las etapas para ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas y dependencias.

En la siguiente tabla (ver TABLA XV) se describe el Plan del Proyecto:

TABLA XV.

PLAN DE TRABAJO

ID	FASE	TAREAS	RECURSOS	DURACION (SEMANAS)	DEPENDENCIA	ENTRADAS	SALIDAS
1	**Comprensión del Negocio**	Objetivos del Negocio	Investigador	6	0	Información sobre el negocio	Información para entender el negocio
		Evaluación de la Situación					
		Plan del Proyecto					
2	**Comprensión de los Datos**	Recolectar los datos iniciales	Investigador	7	1	Recolección inicial de los datos, analizando cada uno de ellos.	Obtención de una Base de datos para ser administrada
		Descripción de los datos					
		Explotación de los datos					
		Verificar la calidad de los					

		datos					
3	**Preparación de los Datos**	Seleccionar los datos	Investigador	12	2	Realización del pre-procesado de los datos	Exploración de los datos para adaptarlos a las técnicas de Minería de Datos
		Limpiar los datos					
		Estructurar los datos					
		Integrar los datos					
4	**Modelado**	Seleccionar Técnica de Modelado	Investigador	8	3	Información confiable de las Técnicas de Minería de Datos	Selección de la Técnica de Minería de Datos y construcción del Modelo
		Construir el Modelo					
		Evaluar el modelo					
5	**Evaluación**	Evaluar los resultados	Investigador	6	4	Modelo de Minería de Datos	Evaluación e interpretación del Modelo de Minería de Datos

2.1.2. COMPRENSIÓN DE LOS DATOS

2.1.2.1. Recolectar Datos Iniciales:

Se realizó una recolección inicial de los datos relacionados con el problema, también un análisis de los mismos con el fin de identificar las relaciones entre ellos.

Los datos académicos obtenidos corresponden a estudiantes del Área de la Energía, Las Industrias y los Recursos Naturales No Renovables (AEIRNNR) de la UNL, de los periodos 2010 – 2013, además existe información personal, institucional de los estudiantes.

Los datos académicos se obtuvieron del Web Service de la UNL, estos datos fueron almacenados en una Base de Datos y que es administrada mediante MySQL, la misma que consta de 19 tablas. En la siguiente figura (ver Figura 2) se puede observar las tablas que conforman la BD:

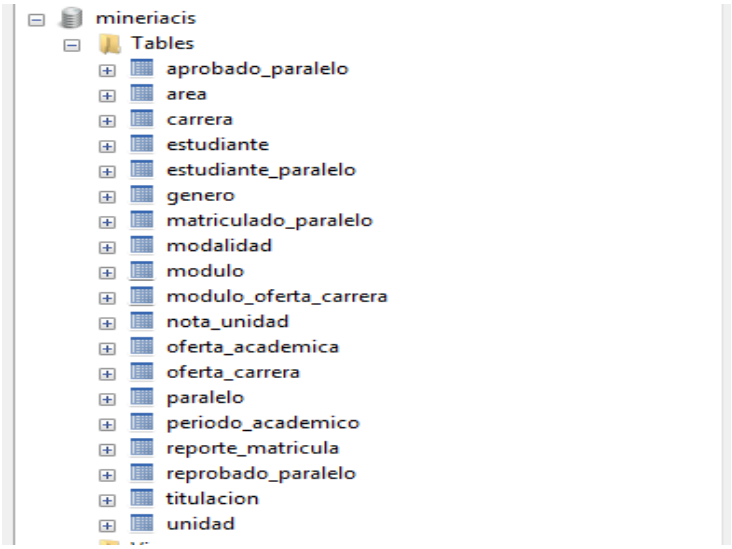


Figura 2. Base de datos

Esta BD contiene información:

- Académica: Información de las Materias, Nota de cada materia, entre otras.
- Personal: Información personal de cada estudiante.
- Institucional: Información de área, carrera, módulos, entre otras.

Además el modelo relacional de la BD integrada por las 19 tablas, conformada por entidades, atributos y las relaciones existentes entre ellas se puede observar en la siguiente figura (ver Figura 3):

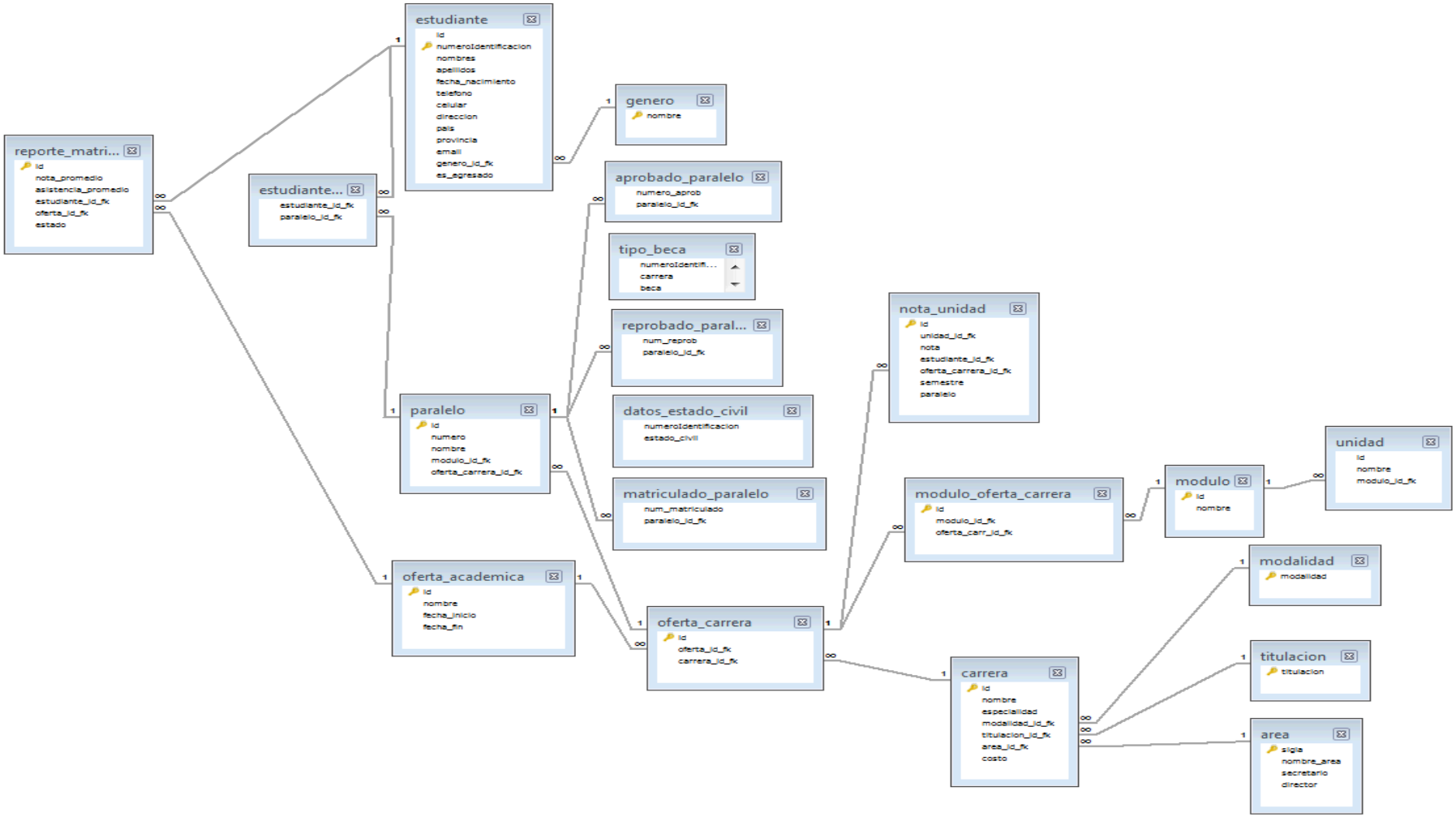


Figura 3: Modelo de la Base De Datos

2.1.2.2. Descripción de los Datos

Los datos inicialmente recopilados se encuentran almacenados en tablas relacionadas en la BD, la cual está integrada por 19 tablas, las mismas que se describen a continuación:

Descripción de la tabla area (ver TABLA XVI):

TABLA XVI.
ATRIBUTOS DE LA TABLA AREA

Atributos	Descripción
sigla	Siglas para identificador de cada una de las áreas.
nombre_area	Nombre del área
secretario	Nombre del secretario del área
director	Nombre del director del área

Descripción de la tabla carrera (ver TABLA XVII):

TABLA XVII.
ATRIBUTOS DE LA TABLA CARRERA

Atributos	Descripción
id	Identificador único de la carrera
Nombre	Nombre de la carrera
Especialidad	Nombre especialidad de la carrera
modalidad_id_fk	Tipo de modalidad (presencial, semipresencial, distancia)
titulación_id_fk	Tipo de titulación (Pregrado, Postgrado, Tecnico_Tecnologico)
area_id_fk	Siglas del área a la que pertenece la carrera.
costo	Costo de la carrera

Descripción de la tabla estudiante (ver TABLA XVIII):

TABLA XVIII.

ATRIBUTOS DE LA TABLA ESTUDIANTE

Atributos	Descripción
id	Identificador único del estudiante
numeroIdentificacion	Número de cedula del estudiante
Nombres	Nombres del estudiante
apellidos	Apellidos del estudiante
fecha_nacimiento	Fecha de nacimiento del estudiante
telefono	Teléfono del estudiante
Celular	Celular del estudiante
direccion	Dirección del estudiante
País	País donde vive el estudiante
Provincia	Provincia donde nació el estudiante
Email	Email del estudiante
genero_id_fk	Tipo de Género del estudiante (Masculino, Femenino)
es_egresado	Para identificar si el estudiante es egresado o no.

Descripción de la tabla aprobado_paralelo (ver TABLA XIX):

TABLA XIX.

ATRIBUTOS DE LA TABLA APROBADO_PARALELO

Atributos	Descripción
numero_aprob	Número de estudiantes aprobados
paralelo_id_fk	Identificador del Paralelo del estudiante.

Descripción de la tabla estudiante_paralelo (ver TABLA XX):

TABLA XX.

ATRIBUTOS DE LA TABLA ESTUDIANTE_PARALELO

Atributos	Descripción
estudiante_id_fk	Identificador del estudiante
paralelo_id_fk	Identificador del Paralelo al que pertenece el estudiante.

Descripción de la tabla genero (ver TABLA XXI):

TABLA XXI.

ATRIBUTOS DE LA TABLA GENERO

Atributos	Descripción
Nombre	Tipo del género del estudiante (Masculino, Femenino)

Descripción de la tabla matriculado_paralelo (ver TABLA XXII):

TABLA XXII.

ATRIBUTOS DE LA TABLA MATRICULADO_PARALELO

Atributos	Descripción
num_matriculado	Número de estudiantes matriculados por paralelo
paralelo_id_fk	Identificador del Paralelo al que pertenece el estudiante.

Descripción de la tabla modalidad (ver TABLA XXIII):

TABLA XXIII.

ATRIBUTOS DE LA TABLA MODALIDAD

Atributos	Descripción
Modalidad	Tipo de modalidad (presencial, semipresencial, distancia)

Descripción de la tabla modulo (ver TABLA XXIV):

TABLA XXIV.

ATRIBUTOS DE LA TABLA MODULO

Atributos	Descripción
id	Identificador del módulo
nombre	Nombre del módulo

Descripción de la tabla modulo_oferta_carrera (ver TABLA XXV):

TABLA XXV.

ATRIBUTOS DE LA TABLA MODULO_OFERTA_CARRERA

Atributos	Descripción
Id	Identificador del módulo
modulo_id_fk	Identificador del módulo
oferta_carr_id_fk	Identificador de la oferta de carrera a la cual

Descripción de la tabla nota_unidad (ver TABLA XXVI):

TABLA XXVI.

ATRIBUTOS DE LA TABLA NOTA_UNIDAD

Atributos	Descripción
unidad_id_fk	Identificador del nombre de la unidad
Nota	Nota del estudiante
estudiante_id_fk	Identificador del estudiante
oferta_carr_id_fk	Identificador de la oferta de la carrera

Descripción de la tabla oferta_academica (ver TABLA XXVII):

TABLA XXVII.

ATRIBUTOS DE LA TABLA OFERTA_ACADEMICA

Atributos	Descripción
id	Identificador de la oferta académica
nombre	Nombre de la oferta académica
fecha_inicio	Fecha de inicio de la oferta académica
fecha_fin	Fecha final de la oferta académica

Descripción de la tabla oferta_carrera (ver TABLA XXVIII):

TABLA XXVIII.

ATRIBUTOS DE LA TABLA OFERTA_CARRERA

Atributos	Descripción
id	Identificador de la oferta de la carrera
oferta_id_fk	Identificador de la oferta académica
carrera_id_fk	Identificador de carrera a la cual pertenece la oferta de la carrera

Descripción de la tabla paralelo (ver TABLA XXIX):

TABLA XXIX.

ATRIBUTOS DE LA TABLA PARALELO

Atributos	Descripción
id	Identificador del paralelo
numero	Número del paralelo
nombre	Nombre del paralelo
modulo_id_fk	Número del módulo del paralelo
oferta_carrera_id_fk	Identificador de la Carrera del paralelo

Descripción de la tabla periodo_academico (ver TABLA XXX):

TABLA XXX.

ATRIBUTOS DE LA TABLA PERIODO_ACADEMICO

Atributos	Descripción
id	Identificador del periodo académico
fecha_periodo	Fecha del periodo académico

Descripción de la tabla reporte_matricula (ver TABLA XXXI):

TABLA XXXI.

ATRIBUTOS DE LA TABLA REPORTE_MATRICULA

Atributos	Descripción
id	Identificador del reporte de la matricula
nota_promedio	Promedio de las notas del estudiante
asistencia_promedio	Promedio de asistencias del estudiante
estudiante_id_fk	Identificador del estudiante
oferta _id_fk	Identificador de la oferta académica
estado	Estado de la matrícula (EstadoMatriculaAprobada, EstadoMatriculaReprobada)

Descripción de la tabla reprobado_paralelo (ver TABLA XXXII):

TABLA XXXII.

ATRIBUTOS DE LA TABLA REPROBADO_PARALELO

Atributos	Descripción
num_reprob	Número de estudiantes reprobados por paralelos
paralelo_id_fk	Identificador del paralelo

Descripción de la tabla titulación (ver TABLA XXXIII):

TABLA XXXIII.

ATRIBUTOS DE LA TABLA TITULACION

Atributos	Descripción
titulación	Tipo de titulación (ACE, POSTGRADO, PREGRADO, TECNICA, ARTESAL Y POPULAR, TECNICO_TECNOLOGICO)

Descripción de la tabla unidad (ver TABLA XXXIV):

TABLA XXXIV.

ATRIBUTOS DE LA TABLA UNIDAD

Atributos	Descripción
nombre	Nombre de la Unidad
modulo_id_fk	Identificador del módulo a la cual pertenece la unidad

2.1.2.3. Exploración de los Datos

Luego de haber realizado una descripción de los datos académicos que conforman la BD de los estudiantes universitarios, es necesario realizar una exploración de los mismos, con el fin de hacer un análisis estadístico y además conocer la distribución que existe en los datos por cada variable.

Para la exploración se incluyeron los datos de todos los estudiantes matriculados en las diferentes carreras del Área de la Energía, Las Industrias y los Recursos Naturales No Renovables (AEIRNNR), tomando en cuenta los periodos académicos: Sept. 2010-Jul. 2011, Sept. 2011-Jul. 2012 y Sept. 2012-Jul. 2013

a. Distribución por período académico: La carrera que presenta el mayor número de estudiantes es Ingeniería en Sistemas (ver TABLA XXXV, ver figura 4).

TABLA XXXV.

NÚMERO DE ESTUDIANTES POR PERÍODO ACADÉMICO.

Carrera	Período Académico			
	Sept. 2010-Jul. 2011	**Sept. 2011-Jul. 2012**	**Sept. 2012-Jul. 2013**	**Subtotal**
Ingeniería en Sistemas	831	741	542	2114
Ingeniería en Electromecánica.	582	540	344	1466
Ingeniería en Geología Ambiental y Ordenamiento Territorial	240	257	248	745
Ingeniería en Electrónica y Telecomunicaciones	305	327	250	882
			Total	5207

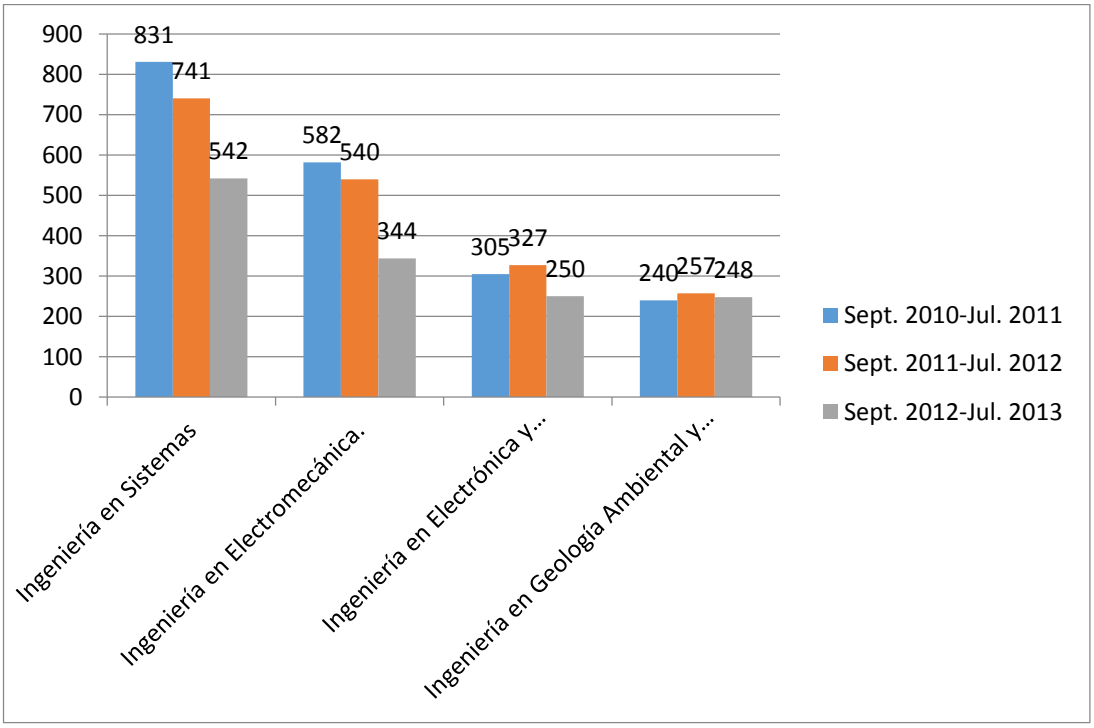


Figura 4. Estudiantes por Periodo Académico

b. Distribución según el Género: En el AEIRNNR la Carrera de Ingeniería en Sistemas posee la mayor cantidad de estudiantes en el género Femenino y la Carrera de Ingeniería en Electromecánica resalta más el género Masculino (ver TABLA XXXVI y Figura 5).

TABLA XXXVI.

ESTUDIANTES POR GÉNERO

Carrera	Número de Estudiantes	
	Masculino	**Femenino**
Ingeniería en Sistemas	1326	788
Ingeniería en Electromecánica.	1413	53
Ingeniería en Geología Ambiental y Ordenamiento Territorial	482	263
Ingeniería en Electrónica y Telecomunicaciones	642	240
Total	**3863**	**1244**

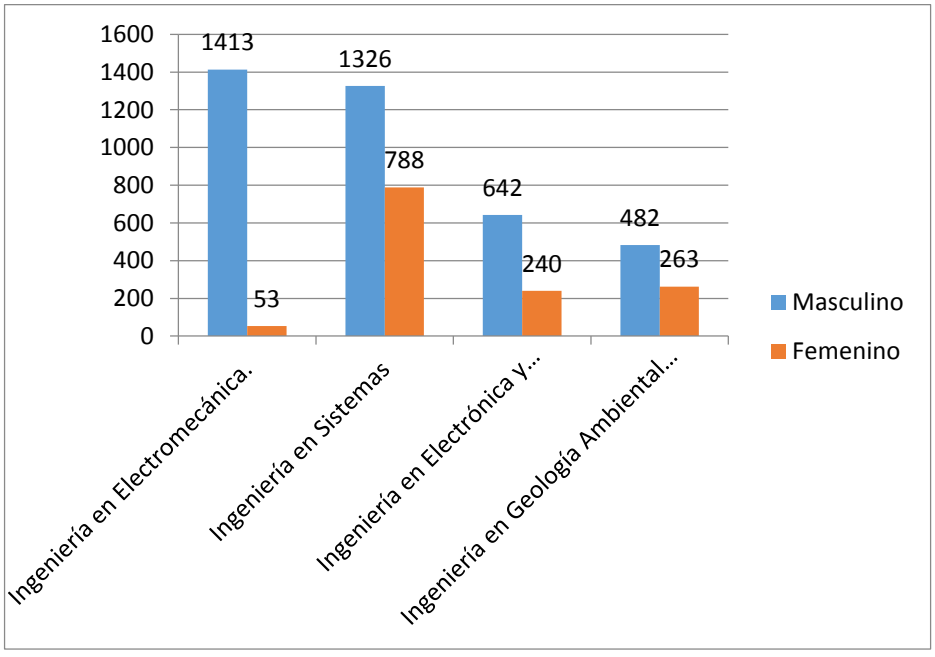


Figura 5. Estudiantes por Género

c. Distribución según el Estado de No Egresados. En el AEIRNNR la carrera de Ingeniería en Sistemas posee el mayor número de estudiantes Egresados y en Formación (ver TABLA XXXVII y Figura 6).

TABLA XXXVII.

ESTUDIANTES EGRESADOS Y EN FORMACIÓN

Carrera	Número de Estudiantes	
	Egresado	**En Formación**
Ingeniería en Sistemas	468	1646
Ingeniería en Electromecánica.	444	1022
Ingeniería en Geología Ambiental y Ordenamiento Territorial	176	569
Ingeniería en Electrónica y Telecomunicaciones	2	880
Total	**1090**	**4117**

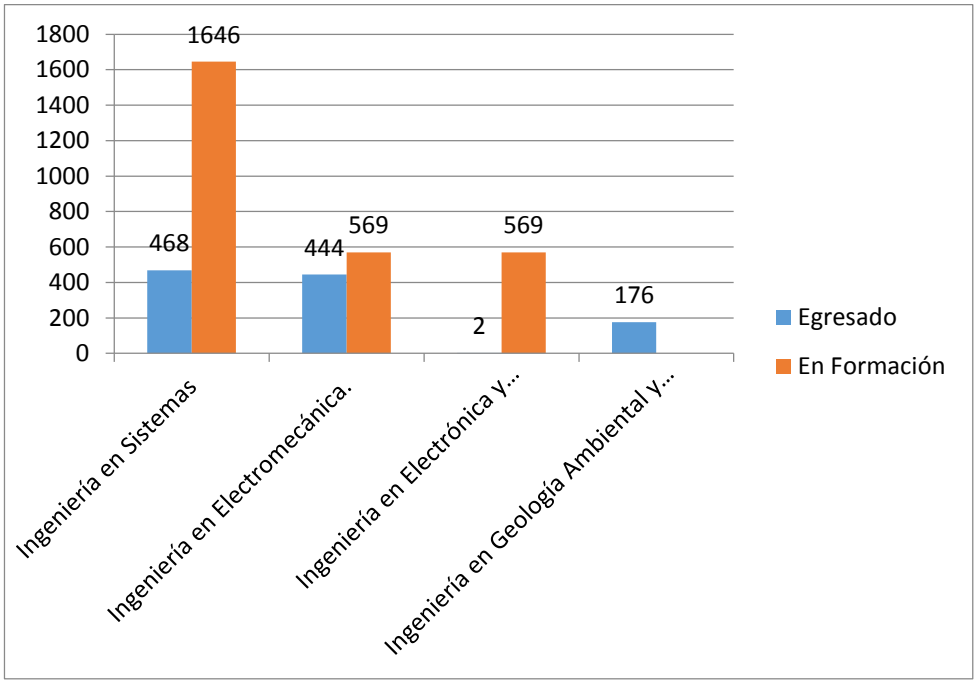


Figura 6. Estudiantes Egresados y En Formación

d. Distribución según la Edad: En el AEIRNNR a edad de los estudiantes con mayor número está en el rango (ver TABLA XXXVIII y Figura 7).

TABLA XXXVIII.

EDAD ESTUDIANTES

Carrera	Edad	
	16-26	**27-37**
Ingeniería en Sistemas	1978	177
Ingeniería en Electromecánica.	1512	187
Ingeniería en Geología Ambiental y Ordenamiento Territorial	602	122
Ingeniería en Electrónica y Telecomunicaciones	888	80
Total	**4180**	**566**

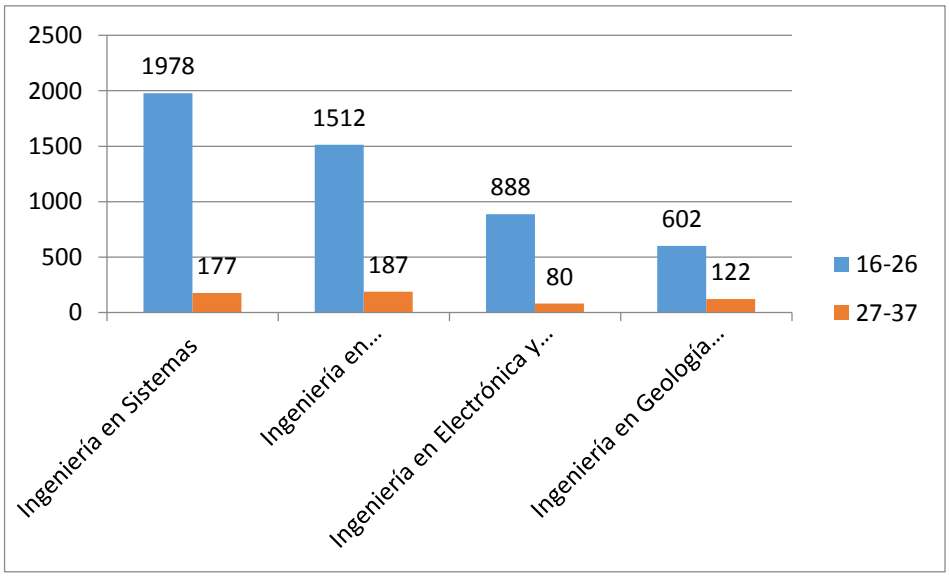


Figura 7: Edad Estudiantes

2.1.2.4. Verificar la Calidad de los Datos

Luego de realizar la exploración de los datos, se encontraron algunos campos que contienen valores con inconvenientes en su calidad, como son:

fecha_nacimiento: este atributo contiene datos de la fecha del estudiante en tipo date.

nota_promedio: este atributo contiene datos con notas de estudiantes creadas por defecto con el valor de 0.00.

2.2. Seleccionar indicadores para construir el modelo computacional que permita estimar el rendimiento académico.

2.2.1. PREPARACIÓN DE LOS DATOS

En la presente fase se construirá la estructura de datos final, en la cual se procederá a ingresar los datos íntegros, sin errores, ni faltantes. Además en esta etapa se transformarán los datos para que puedan ser usados de manera eficiente por técnicas de minería de datos.

2.2.1.1. Selección de datos

La selección de los datos estará enfocada en el tema principal que permita Determinar el Rendimiento Académico. Es por ello que se realizó la eliminación de tablas que no son útiles en la minería de datos, estas tablas son: periodo_academico, área, modalidad, titulación.

- Tabla periodo_academico: Esta tabla es eliminada porque contiene campos que no son necesarios para crear el modelo computacional, también porque no está relacionada con ninguna de las demás tablas.
- Tabla area: Esta tabla es eliminada porque todos los datos que se encuentran en la Base de Datos son del AEIRNNR.
- Tabla modalidad: Esta tabla es eliminada porque todos los datos que se encuentran en la Base de Datos son de la modalidad presencial.
- Tabla titulacion: Esta tabla es eliminada porque contiene campos que no son necesarios para crear el modelo computacional.

2.2.1.2. Limpieza de los datos

En la presente tarea se realizó una limpieza de datos, con la finalidad de dar tratamiento a las inconsistencias encontradas, y así poder generar un modelo de calidad, esto consiste en la eliminación de atributos que contienen datos nulos o tratamiento de valores ausentes, reducción del volumen de datos, eliminar registros duplicados y completar campos vacíos.

Por ende, se eliminaron atributos de varias tablas los mismos que se describen a continuación (ver TABLA XXXIX):

TABLA XXXIX.

ATRIBUTOS ELIMINADOS

Tabla	Columna
carrera	• especialidad • modalidad_id_fk • titulación_id_fk • area_id_fk • costo
estudiante	• telefono • celular • direccion • país • provincia • email

En la base de datos los campos que se han tomado en cuenta para la generación del modelo predictivo, se refiere a datos personales, datos académicos e institucionales.

El campo nota_promedio perteneciente a la tabla reporte_matricula, posee valores que son asignados por defecto, en este caso presenta valores de 0.0, lo que se realizó es eliminar los registros que poseen dicho valor.

2.2.1.3. Construcción de Datos

En la presente actividad se procedió a realizar la construcción de los datos que permitirán construir el modelo computacional, los campos que se han considerado para

establecer en la data set definitivo, para el presente estudio, son los siguientes (ver TABLA XL):

TABLA XL.

ESTRUCTURA DE DATOS PARA DETERMINAR EL RENDIMIENTO ACADÉMICO.

Atributo	Tipo de Datos	Tipo de Contenido	Valores
numeroIdentificacion	Nominal	Continuo	
nota_promedio	Real	Continuo	• malo • bueno • sobresaliente
edad	Nominal	Continuo	• a • b • c
genero	Nominal	Discreto	• 0 • 1
estado_matricula	Nominal	Discreto	• 0 • 1
promedio_asistencia	Nominal	Continuo	• b • m • a
nombre_carrera	Nominal	Discreto	• a • b • c • d
tipo _beca	Nominal	Discreto	• A • B • C • D • N
estado_civil	Nominal	Discreto	• a • b • c • d • e
modulo	Nominal	Discreto	▪ a ▪ b ▪ c

origen_estudiante	Nominal	Discreto	▪ u ▪ r ▪ p ▪ c
numero_hijos_estudiante	Integer	Continuo	▪ S ▪ N
etnia_estudiante	Nominal	Discreto	▪ m ▪ b ▪ i
situación_laboral_estudiante	Nominal	Discreto	▪ S ▪ N
situación_laboral_madre	Nominal	Discreto	▪ S ▪ N
situación_laboral_padre	Nominal	Discreto	▪ S ▪ N
horario	Nominal	Discreto	▪ Matutino ▪ Vespertino
tipo_colegio_estudiante	Nominal	Discreto	▪ 0 ▪ 1

A continuación se describe cada uno de los atributos de la estructura de minería de datos para determinar el Rendimiento Académico de la Tabla XL.

- numeroIdentificacion: Atributo que contiene el número de cedula del estudiante.
- nota_promedio: Atributo que contiene el promedio final del módulo de cada estudiante. En la tabla (ver TABLA XLI) se indica este atributo.

TABLA XLI.

ATRIBUTO NOTA_PROMEDIO.

Siglas	Descripción
malo	Promedio menor a 7.0
bueno	Promedio de 7.00 - 9.00
sobresaliente	Promedio de 9.00-10.00

- edad.- Campo nominal que contiene los rangos de valores de la edad de los estudiantes, se han tomado en cuenta todos los rangos de la edad de los estudiantes. En la tabla (ver TABLA XLII) se indica este atributo.

TABLA XLII.

ATRIBUTO EDAD.

Siglas	Descripción
a	Estudiantes menores a 20 años
b	Estudiantes entre 20 y 21 años
c	Estudiantes mayores a 21 años

- genero.- Campo nominal que contiene los valores de masculino y femenino. En la tabla (ver TABLA XLIII) se indica este atributo.

TABLA XLIII.

ATRIBUTO GENERO

Valor	Descripción
0	Género masculino
1	Género femenino

- estado_matricula: Campo nominal que contiene los valores de Estado Matricula Aprobada y Estado Matricula Reprobada. En la tabla (ver TABLA XLIV) se indica este atributo.

TABLA XLIV:

ATRIBUTO ESTADO_MATRICULA

Valor	Descripción
0	Cuando el Estado Matrícula es Aprobada
1	Cuando el Estado Matrícula es Reprobada

- promedio_asistencia: El promedio por módulo de la asistencia de cada estudiante a clases. En la tabla (ver TABLA XLV) se indica este atributo.

TABLA XLV.

ATRIBUTO PROMEDIO_ASISTENCIA.

Siglas	Descripción
b	Promedio de asistencia de los estudiantes es bajo, menor a 80%
m	Promedio de asistencia de los estudiantes es medio, entre 80% y 90%
a	Promedio de asistencia de los estudiantes es alto, entre 90% y 100%

- nombre_carrera: nombre de la carrera a la que pertenece cada estudiante (ver TABLA XLVI).

TABLA XLVI.

ATRIBUTO NOMBRE_CARRERA.

Valor	Siglas	Descripción
a	CIS	Ingeniería en Sistemas.
b	CIE	Ingeniería en Electromecánica.
c	CIET	Ingeniería en Electrónica y Telecomunicaciones.
d	CGAOT	Ingeniería en Geología Ambiental y Ordenamiento Territorial.

- tipo_beca: El tipo de beca que se ofrece al estudiante. En la tabla (ver TABLA XVLII) se indica este atributo

TABLA XLVII.

ATRIBUTO TIPO_BECA.

Siglas	Descripción
A	Tipo de beca A del estudiante
B	Tipo de beca B del estudiante
C	Tipo de beca C del estudiante
D	Tipo de beca D del estudiante
N	No posee ningún tipo de beca

- estado_civil: Campo que contiene el estado civil que tiene cada estudiante (ver TABLA XLVIII)

TABLA XLVIII.

ATRIBUTO ESTADO_CIVIL.

Siglas	Descripción
a	Soltero
b	Casado
c	Divorciado
d	Unión Libre
e	Viudo

- modulo: El módulo al cual pertenece cada estudiante. En la tabla (ver TABLA XLIX) se indica este atributo.

TABLA XLIX.

ATRIBUTO MODULO.

Siglas	Descripción
a	Módulo de los estudiantes de 1-3
b	Módulo de los estudiantes de 4-6
c	Módulo de los estudiantes de 7-10

origen_estudiante: El lugar de procedencia del estudiante, el cual puede ser del sector urbano (El Sagrario, San Sebastián, Sucre y el Valle), el sector rural (Chantaco, Chuquiribamba, El Cisne, Gualel, Jimbilla, Malacatos, San Lucas, Santiago, San Pedro de Vilcambamba, Quinara, Taquil, Vilcambamba y Yangana), de algún cantón de la provincia de Loja (Loja, Calvas, Catamayo, Chaguarpamba, Celica, Espindola, Gonzanamá, Macará, Paltas, Puyango, Saraguro, Sozoranga, Zapotillo, Pindal, Quilanga y Olmedo) o de otra provincia del Ecuador. En la tabla (ver TABLA L) se indica este atributo.

TABLA L.

ATRIBUTO ORIGEN_ESTUDIANTE.

Siglas	Descripción
U	Si el estudiante pertenece al sector urbano.
R	Si el estudiante pertenece al sector rural.
C	Si el estudiante pertenece a algún cantón de la provincia de Loja.
P	Si el estudiante pertenece a otra provincia del Ecuador.

numero_hijos_estudiante: El número de hijos que tiene el estudiante. En la tabla (ver TABLA LI) se indica este atributo.

TABLA LI.

ATRIBUTO NUMERO_HIJOS_ESTUDIANTE.

Siglas	Descripción
S	Si el estudiante tiene por lo menos un hijo.
N	Si el estudiante no tiene hijos

etnia_estudiante: La etnia del estudiante. En la tabla (ver TABLA LII) se indica este atributo.

TABLA LII.

ATRIBUTO ETNIA_ESTUDIANTE.

Siglas	Descripción
m	Si el estudiante es mestizo
b	Si el estudiante es blanco
i	Si el estudiante es indígena

situación_laboral_estudiante: Describe si el estudiante trabajo o no. En la tabla (ver TABLA LIII) se indica este atributo.

TABLA LIII.

ATRIBUTO SITUACIÓN_LABORAL_ESTUDIANTE.

Siglas	Descripción
S	Si el estudiante trabaja.
N	Si el estudiante no trabaja.

situación_laboral_madre: Describe si la madre del estudiante trabajo o no. En la tabla (ver TABLA LIV) se indica este atributo.

TABLA LIV.

ATRIBUTO SITUACIÓN_LABORAL_MADRE.

Siglas	Descripción
S	Si la madre del estudiante trabaja.
N	Si la madre del estudiante no trabaja.

situación_laboral_padre: Describe si el padre del estudiante trabajo o no. En la tabla (ver TABLA LV) se indica este atributo.

TABLA LV.

ATRIBUTO SITUACIÓN_LABORAL_PADRE.

Siglas	Descripción
S	Si el padre del estudiante trabaja.
N	Si el padre del estudiante no trabaja.

horario: El horario de clases de cada una de las carreras a la que pertenece el estudiante, puede ser matutino o vespertino. En la tabla (ver TABLA LVI) se indica este atributo.

tipo_colegio_estudiante: Describe si el estudiante realizó sus estudios secundarios en un colegio público o particular. En la tabla (ver TABLA LVI) se indica este atributo.

TABLA LVI.

ATRIBUTO TIPO_COLEGIO_ESTUDIANTE.

Siglas	Descripción
0	Si el estudiante realizó sus estudios secundarios en un colegio público.
1	Si el estudiante realizó sus estudios secundarios en un colegio privado.

2.2.1.4. Integración de datos

En esta tarea se integró la base de datos con nuevos datos que fueron otorgados por el Departamento de Bienestar Estudiantil y la UTI.

Las tablas que se le agregó a la base de datos son:

- Tabla tipo_beca que contiene los siguientes atributos (cedula, carrera, tipo_beca).
- Tabla datos_estado_civil que contienen los siguientes atributos (cedula, estado_civil).
- Tabla socioeconómicos que contiene los siguientes atributos (cedula_estudiante, etnia_estudiante, tipo_colegio_estudiante, numero_hijos_estudiante, situación_laboral_estudiante, canton_estudiante, situacion_laboral_padre, situacion_laboral_madre)

Luego de realizar la creación de las tablas se procedió a ingresar los datos desde un archivo CSV que fue proporcionado por el Departamento de Bienestar estudiantil sobre los estudiantes que poseen algún tipo de beca y también los datos personales y socioeconómicos de los estudiantes que fueron proporcionados por la UTI.

2.3. Plantear un modelo computacional mediante la Técnica de Minería de Datos para estimar el rendimiento académico de los estudiantes.

Luego de construir el dataset final se implementará en la herramienta de minería de datos con el fin de construir el modelo computacional y obtener resultados óptimos que permitan determinar el rendimiento académico.

2.3.1. MODELADO:

En esta fase se utilizó el conjunto de datos establecidos, los cuales fueron procesados a través de una Herramienta de Minería de Datos que permitió implementar la o las técnicas necesarias para la construcción del modelo.

2.3.1.1. Selección de técnica de modelado

En esta tarea se seleccionó la técnica de minería de datos que más se adapte al problema con el fin de construir el modelo que permita determinar el rendimiento académico.

Luego de realizar un estudio comparativo de las técnicas (ver Sección Resultados: 1.2. Elaborar un análisis comparativo de las diversas técnicas de Minería de Datos), se seleccionó la Clasificación que contiene diferentes algoritmos los mismos que fueron

aplicados en el presente Trabajo de Titulación, los cuales son: ID3, C4.5, JRIP, RIDOR y PART.

2.3.1.2. Generación del diseño de pruebas

Luego de construir un modelo, hay que crear un mecanismo para probar la calidad y validez del modelo. Por ende, se trabajará con un conjunto de datos para entrenamiento (CE) y también para pruebas (CP) y validación del modelo que permita determinar el rendimiento académico.

El tamaño del CE normalmente es mayor que el del CP (2/3 1/3 4/5 1/5). Los elementos del CE suelen obtenerse mediante muestreo. El CP está formado por los elementos no incluidos en el CE [73].

A continuación, se describe el diseño de pruebas a realizar en los diferentes algoritmos:

- **Algoritmos de Árboles de Decisión: ID3, C4.5**

  El diseño de pruebas que se realizó con estos algoritmos fueron el conjunto de datos de entrenamiento, para el cual se utilizó un 67% del total de datos, además el otro 33% permitió evaluar el modelo creado mediante la validación cruzada.

- **Algoritmos de Reglas RIDOR, PART, JRIP**

  Mediante estos algoritmos también se realizó las mismas pruebas que los algoritmos anteriores, trabajando con el 67% para entrenamiento del total de datos, además el otro 33% permitió evaluar el modelo creado mediante la validación cruzada.

Luego de seleccionar los algoritmos de minería de datos es necesario realizar un estudio comparativo de las herramientas que permitan construir el modelo a través de estos algoritmos (ver Sección Resultados: 3.2. Análisis comparativo de las diferentes herramientas de Minería de Datos que permitan adaptar el modelo computacional realizado).

3. TERCERA FASE: Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.

3.1. Recopilación de información en fuentes académicas, artículos científicos sobre herramientas de Minería de Datos que permitan adaptar el modelo computacional realizado.

Para la implementación del dataset construido se realizó un estudio de algunas herramientas de minería de datos, las mismas que permiten obtener el modelo computacional y resultados que permitan determinar el rendimiento académico de los estudiantes. Las herramientas estudiadas fueron WEKA, RAPID MINER, ORACLE DATA MINING Y R (ver Sección Revisión Literaria: 2. CAPÍTULO II HERRAMIENTAS DE MINERÍA DE DATOS).

3.2. Análisis comparativo de las diferentes herramientas de Minería de Datos que permitan adaptar el modelo computacional realizado.

Las diferentes herramientas de Minería de datos mencionadas anteriormente son factibles para implementar el modelo computacional que permita determinar el rendimiento académico, pero para la elección de la más idónea se realizará un análisis comparativo de cada una de ellas (ver TABLA LVII).

TABLA LVII.

ANÁLISIS COMPARATIVO DE HERRAMIENTAS

HERRAMIENTAS	LICENCIA	S.O.	INTERFAZ	TÉCNICAS
WEKA	GNU GPL	Multiplataforma	Posee una interfaz amigable	• Pre procesamiento de datos. • Técnicas Descriptivas • Técnicas Predictivas • Visualización. • Selección de atributos.
R	GNU GPL	Multiplataforma	Para un principiante no	• Aprendizaje automático • Técnicas Descriptivas

			es fácil de utilizar	• Técnicas Predictivas
RAPIDMINER	GNU GPL	Multiplataforma	Posee una interfaz amigable	• Pre procesado de datos. • Técnicas Descriptivas • Técnicas Predictivas • Múltiples Operadores
ORACLE DATA MINING	PROPIETARIA	Multiplataforma	Para un principiante no es fácil de utilizar	• Pre procesado de datos. • Técnicas Descriptivas • Técnicas Predictivas

3.3. Selección de la mejor herramienta de Minería de Datos que permitan adaptar el modelo computacional realizado.

La herramienta a utilizar para el presente Trabajo de Titulación es RAPIDMINER debido a que posee gran cantidad de operadores que permiten generar de mejor manera un modelo que permita determinar el rendimiento académico. Además, posee una licencia libre, es multiplataforma, posee una interfaz amiga y permite aplicar algoritmos de minería de datos como Clasificación, Clustering, Redes Neuronales, Árboles de Decisión, Algoritmos genéticos, Regresión Lineal, además contiene complementos que permiten implementar los algoritmos que contiene WEKA, para poder generar un modelo de calidad y también permite validar el modelo.

3.4. Implementar el modelo computacional en la herramienta de Minera de Datos seleccionada.

3.4.1. Construcción de Modelo

Luego de seleccionar los algoritmos de minería de datos, estas se ejecutan sobre los datos para generar el o los modelos. Los algoritmos permiten seleccionar los parámetros, el cual es un proceso iterativo que permite obtener mejores resultados.

3.4.1.1. Rendimiento Académico

En esta tarea se realiza la construcción del modelo que permita determinar el rendimiento académico, como se mencionó anteriormente (ver Sección Resultados: 2.3.1.2. Generación del diseño de pruebas) se trabajó con el 67% para entrenamiento

del total de datos, además el otro 33% permitió evaluar el modelo creado mediante la validación cruzada, pesos a cada uno de los atributos, estadística de kappa, entre otros parámetros. Además se estableció el atributo nota_promedio como objetivo que permita determinar el rendimiento académico.

A continuación se describen cada uno de los algoritmos que permitieron crear los modelos.

3.4.1.1.1. ID3

Para poder llevar a cabo este proceso se realizó una serie de pasos que permitan enlazar operadores los cuales están detallados en el ANEXO 1.

En este algoritmo se estableció los siguientes parámetros que permitió obtener los mejores resultados del modelo generado por el mismo basado en un proceso iterativo:

- criterion (criterio) se seleccionó information_gain (*ganancia de información),* aquí se especifica el criterio de selección de atributos y de divisiones numéricas.
- *minimal size for split=4* es el número mínimo de divisiones que se pueden dar por cada nodo.
- *minimal leaf size=2* es el tamaño mínimo de cada hoja.
- *minimal gain=0.3* es la ganancia mínima que debe lograrse con el fin de producirse una división

Para la validación cruzada se fijó el parámetro *number of validations=10* que es el número de subconjuntos que se generan para evaluar el algoritmo.

Luego de la configuración de los parámetros del algoritmo se obtuvo los siguientes resultados (ver TABLA LVIII):

TABLA LVIII.

RESULTADOS DE INSTANCIAS CLASIFICADAS DEL ALGORITMO ID3

ID3	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de kappa
Conjunto de Entrenamiento:	94.84	5.16	0.825
Validación Cruzada:	85.04	14.96	0.529

Así mismo se generó la matriz de confusión durante la fase de entrenamiento la cual se indica a continuación (ver Figura 8).

	true malo	true bueno	true sobresaliente
pred. malo	411	22	1
pred. bueno	27	2438	96
pred. sobresaliente	2	9	35

Figura 8. Matriz confusión fase de entrenamiento algoritmo ID3

Además se generó la matriz de confusión luego de aplicar la validación cruzada la cual se indica a continuación (ver figura 9):

	true malo	true bueno	true sobresaliente
pred. malo	166	105	1
pred. bueno	32	1103	67
pred. sobresaliente	0	19	4

Figura 9. Matriz confusión validación cruzada algoritmo ID3

Las dos matrices de confusión permiten observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente.

También el algoritmo permitió obtener las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo y poder obtener conclusiones a partir de estas reglas. Las reglas generadas se muestran en la siguiente figura (ver Figura 10):

```
|   tipo_beca = A
|   |   horario = Matutino
|   |   |   situacion_laboral_estudiante = N: bueno {malo=0, bueno=48, sobresaliente=0}
|   |   |   situacion_laboral_estudiante = S
|   |   |   |   numero_hijos_estudiante = N
|   |   |   |   |   modulo = c
|   |   |   |   |   |   tipo_colegio_estudiante = 0
|   |   |   |   |   |   |   edad = c
|   |   |   |   |   |   |   |   etnia_estudiante = m
|   |   |   |   |   |   |   |   |   promedio_asistencia = a
|   |   |   |   |   |   |   |   |   |   nombre_carrera = d
|   |   |   |   |   |   |   |   |   |   |   genero = 0: bueno {malo=0, bueno=1, sobresaliente=0}
|   |   |   |   |   |   |   |   |   |   |   genero = 1
|   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_padre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   estado_civil = a
|   |   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_madre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   canton_estudiante = u: sobresaliente
|   |   |   |   numero_hijos_estudiante = S: bueno {malo=0, bueno=5, sobresaliente=0}
```

```
|   |   horario = Vespertino
|   |   |   canton_estudiante = c
|   |   |   |   modulo = a
|   |   |   |   |   tipo_colegio_estudiante = 0
|   |   |   |   |   |   edad = a: bueno {malo=0, bueno=1, sobresaliente=0}
|   |   |   |   |   |   edad = b
|   |   |   |   |   |   |   etnia_estudiante = m
|   |   |   |   |   |   |   |   promedio_asistencia = a
|   |   |   |   |   |   |   |   |   numero_hijos_estudiante = N
|   |   |   |   |   |   |   |   |   |   nombre_carrera = b
|   |   |   |   |   |   |   |   |   |   |   situacion_laboral_estudiante = N
|   |   |   |   |   |   |   |   |   |   |   |   genero = 0
|   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_padre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   |   estado_civil = a
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_madre = S: sobresaliente
|   |   |   |   |   |   |   |   |   |   nombre_carrera = c: bueno {malo=0, bueno=1, sobresaliente=0}
```

```
|   |   |   |   modulo = b
|   |   |   |   |   situacion_laboral_estudiante = N: bueno {malo=0, bueno=4, sobresaliente=0}
|   |   |   |   |   situacion_laboral_estudiante = S
|   |   |   |   |   |   tipo_colegio_estudiante = 0
|   |   |   |   |   |   |   edad = c
|   |   |   |   |   |   |   |   etnia_estudiante = m
|   |   |   |   |   |   |   |   |   promedio_asistencia = a
|   |   |   |   |   |   |   |   |   |   numero_hijos_estudiante = N
|   |   |   |   |   |   |   |   |   |   |   nombre_carrera = c
|   |   |   |   |   |   |   |   |   |   |   |   genero = 0
|   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_padre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   |   estado_civil = a
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_madre = S: bueno
```

```
|   |   horario = Vespertino
|   |   |   modulo = a
|   |   |   |   tipo_colegio_estudiante = 0
|   |   |   |   |   edad = a
|   |   |   |   |   |   etnia_estudiante = b: bueno {malo=0, bueno=1, sobresaliente=0}
```

```
|   promedio_asistencia = b
|   |   tipo_beca = A: bueno {malo=0, bueno=1, sobresaliente=0}
|   |   tipo_beca = D: malo {malo=1, bueno=0, sobresaliente=0}
|   |   tipo_beca = N
|   |   |   horario = Matutino
|   |   |   |   modulo = a: malo {malo=29, bueno=0, sobresaliente=0}
|   |   |   |   modulo = b
|   |   |   |   |   tipo_colegio_estudiante = 0
|   |   |   |   |   |   edad = b: malo {malo=6, bueno=0, sobresaliente=0}
|   |   |   |   |   |   edad = c
|   |   |   |   |   |   |   etnia_estudiante = m
|   |   |   |   |   |   |   |   numero_hijos_estudiante = N
|   |   |   |   |   |   |   |   |   nombre_carrera = a
|   |   |   |   |   |   |   |   |   |   situacion_laboral_estudiante = N
|   |   |   |   |   |   |   |   |   |   |   genero = 0
|   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_padre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   estado_civil = a
|   |   |   |   |   |   |   |   |   |   |   |   |   |   situacion_laboral_madre = S
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   canton_estudiante = c: malo
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   canton_estudiante = p: malo
```

Figura 10: Reglas de Inferencia Algoritmo ID3

Las reglas más importantes generadas por el algoritmo ID3 (ver figura 10) se describen a continuación:

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante no trabaja; entonces el rendimiento académico es bueno.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios

secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y el género es masculino; entonces el rendimiento académico es bueno.

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y pertenece a la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial y el género es femenino y el padre trabaja y el estudiante es soltero y la madre trabaja y el estudiante pertenece al sector urbano; entonces el rendimiento académico es sobresaliente.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y tiene hijos; entonces el rendimiento académico es bueno.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de primero a tercero y realizó sus estudios secundarios en un colegio público y tiene entre 20 y 21 años y es mestizo y no tiene hijos y pertenece a la carrera Ingeniería Electromecánica y no trabaja y el género es masculino y el padre trabaja y el estudiante es soltero y la madre trabaja; entonces el rendimiento académico es sobresaliente.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y no trabaja; entonces el rendimiento académico es bueno.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y trabaja realizó sus estudios secundarios en un colegio público y es mayor a 21 años y es mestizo y el promedio de asistencia es mayor a 90% y no tiene hijos y pertenece a la carrera Ingeniería Electrónica y Telecomunicaciones y el género es masculino y es soltero y el padre trabaja y la madre trabaja; entonces el rendimiento académico es bueno.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a los módulos de primero a tercero realizó sus estudios secundarios en un colegio público y es menor de 20 años y es blanco; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de primero a tercero; entonces el rendimiento académico es malo.

- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto y es mestizo y no tiene hijos y pertenece a la carrea de Ingeniería en Sistemas y no trabaja y es masculino y es soltero y el padre trabaja y la madre trabaja y pertenece a un cantón de la provincia de Loja o a otra provincia; entonces el rendimiento académico es malo.

3.4.1.1.2. C4.5

Para poder llevar a cabo este proceso se realizó una serie de pasos que permitan enlazar operadores los cuales están detallados en el ANEXO 2.

En este algoritmo se estableció los siguientes parámetros que permitió obtener los mejores resultados del modelo generado por el mismo basado en un proceso iterativo:

- U: false es el uso del árbol sin podar
- C: 0.9 es el ajuste umbral de confianza para la poda.
- M: 2.0 es establecer el número mínimo de casos por hoja.
- R: false su uso reduce la poda error.
- N: vacío es establecer el número de pliegues para reducir la poda error.
- B: false Sólo para uso de divisiones binarias.
- S: false no lleva a cabo la cría subárbol.
- L: false para no limpiar después el árbol se ha construido.
- A: false Laplace suavizado de probabilidades predichas.
- Q: 1 Semilla de datos aleatorios arrastrando los pies.

Para la validación cruzada se fijó el parámetro *number of validations=10* que es el número de subconjuntos que se generan para evaluar el algoritmo.

Luego de la configuración de los parámetros del algoritmo se obtuvo los siguientes resultados (ver TABLA LIX):

TABLA LIX.

RESULTADOS DE INSTANCIAS CLASIFICADAS DEL ALGORITMO C4.5

C4.5	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de kappa
Conjunto de Entrenamiento:	92.21	7.79	0.723
Validación Cruzada:	90.98	9.02	0.652

Así mismo se generó la matriz de confusión durante la fase de entrenamiento la cual se indica a continuación (ver Figura 11).

	true malo	true bueno	true sobresaliente
pred. malo	380	45	2
pred. bueno	60	2424	130
pred. sobresaliente	0	0	0

Figura 11. Matriz confusión fase de entrenamiento algoritmo C4.5

Además se generó la matriz de confusión luego de aplicar la validación cruzada la cual se indica a continuación (ver figura 12):

	true malo	true bueno	true sobresaliente
pred. malo	153	18	0
pred. bueno	45	1209	72
pred. sobresaliente	0	0	0

Figura 12. Matriz confusión validación cruzada algoritmo C4.5

Las dos matrices de confusión permiten observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. También el algoritmo permitió obtener las reglas

de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo y poder obtener conclusiones a partir de estas reglas (ver Figura 13):

```
promedio_asistencia = a
|   edad = a
|   |   modulo = a: malo (10.0/3.0)
|   |   modulo = b: bueno (2.0)
|   |   modulo = c: malo (0.0)
|   edad = b: bueno (31.0/5.0)
|   edad = c
|   |   modulo = a
|   |   |   horario = Matutino: bueno (6.0/2.0)
|   |   |   horario = Vespertino: malo (4.0)
|   |   modulo = b: bueno (33.0/9.0)
|   |   modulo = c
|   |   |   nombre_carrera = a: malo (10.0/4.0)
|   |   |   nombre_carrera = c: malo (1.0)
|   |   |   nombre_carrera = b: bueno (3.0)
|   |   |   nombre_carrera = d: malo (0.0)
promedio_asistencia = b: malo (129.0/6.0)
promedio_asistencia = m
|   nombre_carrera = a: malo (22.0/4.0)
|   nombre_carrera = c: malo (4.0)
|   nombre_carrera = b
|   |   edad = a: malo (1.0)
|   |   edad = b: malo (3.0)
|   |   edad = c: bueno (14.0/4.0)
|   nombre_carrera = d: bueno (5.0)
```

Figura 13: Reglas de Inferencia Algoritmo C4.5

Las reglas más importantes generadas por el algoritmo C4.5 (ver figura 13) se describen a continuación:

- Si el promedio de asistencia es mayor a 90% y es menor a 20 años y pertenece a los módulos de cuarto a sexto; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es mayor a 90% y tiene entre 20 y 21 años; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es mayor a 90% y es mayor de 21 años y pertenece a los módulos de primero a tercero y el horario es matutino; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es menor a 80% y pertenece a la carrera de Ingeniería Electrónica y Telecomunicaciones; entonces el rendimiento académico es malo.

- Si el promedio de asistencia es menor a 80% y es menor a 20 años; entonces el rendimiento académico es malo.

3.4.1.1.3. JRIP

Para poder llevar a cabo este proceso se realizó una serie de pasos que permitan enlazar operadores los cuales están detallados en el ANEXO 3.

En este algoritmo se estableció los siguientes parámetros que permitió obtener los mejores resultados del modelo generado por el mismo basado en un proceso iterativo:

- F: 3.0 se utiliza como conjunto de poda.
- N: 2.0 es el peso mínimo de los casos dentro de una fracción.
- O: 2.0 es el número de carreras de optimizaciones.
- D: false Permite definir el modo de depuración.
- S: 1.0 es la semilla de aleatorización.
- E: false Si no comprueba la tasa de error> = 0.5 en la detención de criterios
- P: false permite utilizar la poda

Para la validación cruzada se fijó el parámetro *number of validations=10* que es el número de subconjuntos que se generan para evaluar el algoritmo.

Luego de la configuración de los parámetros del algoritmo se obtuvo los siguientes resultados (ver TABLA LX):

TABLA LX.

RESULTADOS DE INSTANCIAS CLASIFICADAS DEL ALGORITMO JRIP

JRIP	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de kappa
Conjunto de Entrenamiento:	92.27	7.73	0.728
Validación Cruzada:	89.78	10.22	0.621

Así mismo se generó la matriz de confusión durante la fase de entrenamiento la cual se indica a continuación (ver Figura 14).

	true malo	true bueno	true sobresaliente
pred. malo	386	49	2
pred. bueno	54	2420	130
pred. sobresaliente	0	0	0

Figura 14. Matriz confusión fase de entrenamiento algoritmo JRIP

Además se generó la matriz de confusión luego de aplicar la validación cruzada la cual se indica a continuación (ver figura 15):

	true malo	true bueno	true sobresaliente
pred. malo	156	37	0
pred. bueno	41	1188	72
pred. sobresaliente	1	2	0

Figura 15. Matriz confusión validación cruzada algoritmo JRIP

Las dos matrices de confusión permiten observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente.

También el algoritmo permitió obtener las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo y poder obtener conclusiones a partir de estas reglas (ver Figura 16):

```
(estado_matricula = 1) and (promedio_asistencia = b) => nota_promedio=malo (129.0/6.0)
(estado_matricula = 1) and (promedio_asistencia = m) and (nombre_carrera = a) => nota_promedio=malo (22.0/4.0)
(estado_matricula = 1) and (edad = a) and (modulo = a) => nota_promedio=malo (11.0/3.0)
(estado_matricula = 1) and (canton_estudiante = p) and (edad = c) and (promedio_asistencia = a) => nota_promedio=malo (17.0/8.0)
 => nota_promedio=bueno (1318.0/112.0)
```

Figura 16: Reglas de Inferencia Algoritmo JRIP

Las reglas más importantes generadas por el algoritmo JRIP (ver figura 16) se describen a continuación:

- Si estado de matrícula es reprobado y promedio de asistencia es menor a 80%; entonces el rendimiento académico es malo.
- Si estado de matrícula es reprobado y promedio de asistencia es menor a 80% y pertenece a la carrera de Ingeniería en Sistemas; entonces el rendimiento académico es malo.
- Si estado de matrícula es reprobado y es menor a 20 años y pertenece a los módulos de primero a tercero; entonces el rendimiento académico es malo.
- Si estado de matrícula es reprobado y pertenece a otra provincia y es mayor a 21 años y el promedio de asistencia es menor a 80%; entonces el rendimiento académico es malo.

3.4.1.1.4. PART

Para poder llevar a cabo este proceso se realizó una serie de pasos que permitan enlazar operadores los cuales están detallados en el ANEXO 4.

En este algoritmo se estableció los siguientes parámetros que permitió obtener los mejores resultados del modelo generado por el mismo basado en un proceso iterativo:

- C=0.5 el umbral de confianza para la poda.
- M= 2 el número mínimo de objetos por hoja.
- R=false es el uso reducido de poda error.
- N=vacío establece el número de pliegues para la reducción de poda error, a veces se usa como poda.
- B=false se usa para divisiones binarias.
- U=false genera una lista de decisión sin podar.
- Q=6 es la semilla de datos aleatorios.

Para la validación cruzada como en todas las pruebas que se realizó se fijó el parámetro number of validations=10 que es el número de subconjuntos que se generan para evaluar el algoritmo.

Luego de la configuración de los parámetros del algoritmo se obtuvo los siguientes resultados (ver TABLA LXI):

TABLA LXI.

RESULTADOS DE INSTANCIAS CLASIFICADAS DEL ALGORITMO PART

PART	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de kappa (%)
Conjunto de Entrenamiento:	92.83	7.17	0.753
Validación Cruzada:	89.98	10.02	0.621

Así mismo se generó la matriz de confusión durante la fase de entrenamiento la cual se indica a continuación (ver Figura 17).

	true malo	true bueno	true sobresaliente
pred. malo	392	49	2
pred. bueno	48	2416	115
pred. sobresaliente	0	4	15

Figura 17. Matriz confusión fase de entrenamiento algoritmo PART

Además se generó la matriz de confusión luego de aplicar la validación cruzada la cual se indica a continuación (ver figura 18):

	true malo	true bueno	true sobresaliente
pred. malo	152	30	0
pred. bueno	46	1195	72
pred. sobresaliente	0	2	0

Figura 18. Matriz confusión validación cruzada algoritmo PART

Las dos matrices de confusión permiten observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. También el algoritmo permitió obtener las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo y poder obtener conclusiones a partir de estas reglas (ver Figura 19):

```
estado_matricula = 0 AND
promedio_asistencia = a AND
tipo_beca = N AND
etnia_estudiante = m AND
edad = b AND
situacion_laboral_padre = S: bueno (191.0/29.0)

estado_matricula = 0 AND
promedio_asistencia = a AND
tipo_beca = N AND
etnia_estudiante = m AND
edad = c: bueno (137.0/16.0)
```

```
estado_matricula = 0 AND
horario = Matutino AND
etnia_estudiante = m AND
tipo_beca = N AND
modulo = b: bueno (19.0)
```

```
tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
edad = b AND
canton_estudiante = p: bueno (6.0)

tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
numero_hijos_estudiante = N AND
edad = b AND
nombre_carrera = a: bueno (6.0/1.0)
```

```
tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
numero_hijos_estudiante = N AND
modulo = b AND
situacion_laboral_padre = S AND
edad = c: bueno (22.0/6.0)

tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
numero_hijos_estudiante = N AND
nombre_carrera = d: bueno (9.0/2.0)

tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
numero_hijos_estudiante = N AND
nombre_carrera = c AND
situacion_laboral_padre = S: malo (7.0/1.0)
```

```
tipo_beca = N AND
etnia_estudiante = m AND
estado_civil = a AND
numero_hijos_estudiante = N AND
modulo = c: malo (9.0/3.0)

tipo_beca = N AND
etnia_estudiante = m AND
situacion_laboral_estudiante = N AND
estado_civil = a AND
modulo = a AND
edad = a: malo (8.0/2.0)
```

Figura 19: Reglas de Inferencia Algoritmo PART

Las reglas más importantes generadas por el algoritmo PART (ver figura 19) se describen a continuación:

- Si estado de matrícula es aprobado y promedio de asistencia es mayor a 90% y no posee un tipo de beca y es mestizo y tiene entre 20 y 21 años y el padre trabaja; entonces el rendimiento académico es bueno.
- Si estado de matrícula es aprobado y promedio de asistencia es mayor a 90% y no posee un tipo de beca y es mestizo y es mayor a 21 años; entonces el rendimiento académico es bueno.
- Si estado de matrícula es aprobado y horario de clases es matutino y es mestizo y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto; entonces el rendimiento académico es bueno.
- Si no posee un tipo de beca y es mestizo y es soltero y tiene entre 20 y 21 años y pertenece a otra provincia; entonces el rendimiento académico es bueno.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y tiene entre 20 y 21 años y pertenece a la carrera de Ingeniería en Sistemas; entonces el rendimiento académico es bueno.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a los módulos de cuarto a sexto y el padre trabaja y es mayor a 21 años; entonces el rendimiento académico es bueno.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a la carrera de Ingeniería Electrónica y Telecomunicaciones y el padre trabaja; entonces el rendimiento académico es malo.
- Si no posee un tipo de beca y es mestizo y es soltero y no trabaja y no tiene hijos y pertenece a los módulos de séptimo a décimo; entonces el rendimiento académico es malo.
- Si no posee un tipo de beca y es mestizo y es soltero y no trabaja y pertenece a los módulos de primero a tercero y es menor de 20 años; entonces el rendimiento académico es malo.

3.4.1.1.5. RIDOR (Ripple Down Rule)

Para poder llevar a cabo este proceso se realizó una serie de pasos que permitan enlazar operadores los cuales están detallados en el ANEXO 5.

En este algoritmo se estableció los siguientes parámetros que permitió obtener los mejores resultados del modelo generado por el mismo basado en un proceso iterativo:

- F: 5.0 es el ajuste el número de pliegues para IREP.
- S: 2.0 es establecer el número de baraja para cambiar aleatoriamente los datos con el fin de obtener una mejor regla.
- A: false es el conjunto de si utilizar la tasa de error de todos los datos para seleccionar la clase por defecto en cada paso. Si no se establece, el alumno sólo utilizará la tasa de error en la cordillera de los datos de la poda.
- M: false es el conjunto de la bandera de si el uso de la clase mayoritaria como la clase por defecto en cada paso en lugar de elegir la clase por defecto basado en la tasa de error.
- N: 10.0 es establecer los pesos mínimos de los casos dentro de una fracción.

Para la validación cruzada se fijó el parámetro *number of validations=10* que es el número de subconjuntos que se generan para evaluar el algoritmo.

Luego de la configuración de los parámetros del algoritmo se obtuvo los siguientes resultados (ver TABLA LXII):

TABLA LXII.

RESULTADOS DE INSTANCIAS CLASIFICADAS DEL ALGORITMO RIDOR

RIDOR	Instancias correctamente clasificadas (%)	Instancias incorrectamente clasificadas (%)	Índice de kappa
Conjunto de Entrenamiento:	90.66	9.34	0.698
Validación Cruzada:	90.11	9.89	0.634

Así mismo se generó la matriz de confusión durante la fase de entrenamiento la cual se indica a continuación (ver Figura 20).

	true malo	true bueno	true sobresaliente
pred. malo	412	128	2
pred. bueno	28	2338	123
pred. sobresaliente	0	3	7

Figura 20. Matriz confusión fase de entrenamiento algoritmo RIDOR

Además se generó la matriz de confusión luego de aplicar la validación cruzada la cual se indica a continuación (ver figura 21):

	true malo	true bueno	true sobresaliente
pred. malo	160	36	0
pred. bueno	38	1189	72
pred. sobresaliente	0	2	0

Figura 21. Matriz confusión validación cruzada algoritmo RIDOR

Las dos matrices de confusión permiten observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. También el algoritmo permitió obtener las reglas de inferencia que permiten determinar cómo está construido el árbol generado por este algoritmo y poder obtener conclusiones a partir de estas reglas (ver Figura 22):

```
nota_promedio = malo  (1497.0/1299.0)
          Except (estado_matricula = 0) and (promedio_asistencia = a) => nota_promedio = bueno  (677.0/1.0) [343.0/2.0]
          Except (estado_matricula = 0) and (canton_estudiante = p) => nota_promedio = bueno  (31.0/0.0) [15.0/0.0]
          Except (estado_matricula = 0) and (horario = Vespertino) => nota_promedio = bueno  (54.0/1.0) [37.0/1.0]
          Except (estado_matricula = 0) and (modulo = c) => nota_promedio = bueno  (19.0/0.0) [6.0/0.0]
          Except (promedio_asistencia = a) and (edad = b) and (modulo = a) => nota_promedio = sobresaliente  (12.0/1.0) [2.0/0.0]
                   Except (estado_matricula = 1) => nota_promedio = bueno  (8.0/0.0) [5.0/0.0]
                   Except (canton_estudiante = p) => nota_promedio = bueno  (39.0/3.0) [20.0/2.0]
                   Except (nombre_carrera = a) => nota_promedio = bueno  (38.0/5.0) [14.0/1.0]
                   Except (canton_estudiante = c) => nota_promedio = bueno  (24.0/3.0) [9.0/1.0]
                   Except (horario = Vespertino) and (canton_estudiante = r) => nota_promedio = bueno  (3.0/0.0) [2.0/0.0]
                   Except (horario = Vespertino) and (nombre_carrera = c) and (genero = 0) => nota_promedio = bueno  (7.0/1.0) [7.0/1.0]
                   Except (genero = 1) => nota_promedio = bueno  (6.0/1.0) [3.0/1.0]
                   Except (situacion_laboral_padre = S) => nota_promedio = bueno  (30.0/10.0) [15.0/5.0]
          Except (promedio_asistencia = a) and (modulo = b) and (genero = 1) and (canton_estudiante = u) => nota_promedio = bueno  (5.0/0.0)
          Except (estado_matricula = 0) and (canton_estudiante = c) => nota_promedio = bueno  (8.0/0.0) [4.0/0.0]
          Except (promedio_asistencia = a) and (modulo = b) and (numero_hijos_estudiante = N) => nota_promedio = bueno  (27.0/6.0) [11.0/2.0]
          Except (promedio_asistencia = m) and (estado_matricula = 0) and (nombre_carrera = a) and (edad = b) => nota_promedio = bueno  (4.0/
          Except (promedio_asistencia = m) and (nombre_carrera = d) => nota_promedio = bueno  (9.0/2.0) [4.0/1.0]
          Except (promedio_asistencia = m) and (estado_matricula = 0) => nota_promedio = bueno  (8.0/1.0) [4.0/2.0]
```

Figura 22: Reglas de Inferencia del Algoritmo RIDOR

Las reglas más importantes generadas por el algoritmo RIDOR (ver figura 22) se describen a continuación:

El rendimiento académico es malo excepto:

- Si estado de matrícula es aprobado y el promedio de asistencia es mayor a 90%; entonces el rendimiento académico es bueno.
- Si estado de matrícula es aprobado y pertenece a otra provincia; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es mayor a 90% y tiene entre 20 y 21 años y pertenece a los módulos de primero a tercero; entonces el rendimiento académico es sobresaliente.
- Si el promedio de asistencia es mayor a 90% y pertenece a los módulos de cuarto a sexto y el género es femenino y pertenece al sector urbano; entonces el rendimiento académico es bueno.
- Si el promedio de asistencia es mayor a 90% y pertenece a los módulos de cuarto a sexto y no tiene hijos; entonces el rendimiento académico es bueno.

3.5. Evaluar el modelo computacional en un escenario real con datos académicos.

3.5.1. Evaluación de Modelos

En esta tarea se realiza la evaluación del modelo generado por cada uno de los algoritmos, comparando los resultados obtenidos, para lo cual se consideraron diferentes parámetros que permitan evaluar el modelo; algunos de estos se tomaron en cuenta para la evaluación que implica separar los datos de forma que para cada iteración tengamos un solo conjunto de dato de prueba y todo el resto de los datos para entrenamiento [74]. Además el estadístico kappa que mide la coincidencia de la predicción con la clase real de Kappa [75]. Interpretación de la estadística Kappa es [75]: Ajuste pobre = Menor de 0.20; Acuerdo justo = 0.20 a 0.40; Ajuste moderado = 0.40 a 0.60; Buen ajuste = 0.60 a 0.80; Muy de acuerdo = 0.80 a 1.00. Los parámetros de cada algoritmo que permitieron evaluar el modelo se muestran en la siguiente tabla (ver TABLA LXIII):

TABLA LXIII.

RESULTADOS DE LOS ALGORITMOS

Algoritmo	**Conjunto de Datos**	***Instancias correctamente clasificadas (%)***	***Instancias incorrectamente clasificadas (%)***	***Índice de Kappa***	**Error Cuadrático**	**Error Relativo (%)**	**Error Cuadrático Medio**	**Error Cuadrático Relativo**
ID3	C E	94.84	5.16	0.825	0.039	7.55	0.197	1.942
	C P	85.04	14.96	0.529	0.097	12.71	0.310	4.163
C4.5	C E	92.21	7.79	0.723	0.070	13.73	0.265	2.613
	C P	90.98	9.02	0.652	0.078	14.90	0.279	3.332
JRIP	C E	92.27	7.73	0.728	0.073	14.04	0.270	2.663
	C P	89.78	10.22	0.621	0.091	16.54	0.301	3.595
PART	C E	92.83	7.17	0.753	0.062	12.21	0.250	2.467
	C P	89.98	10.02	0.621	0.083	14.82	0.288	3.443
RIDOR	C E	90.66	9.34	0.698	0.093	9.34	0.306	3.017
	C P	90.11	9.89	0.634	0.099	9.89	0.314	3.751

En el Anexo 6 también se tienen resultados de estos algoritmos con márgenes de error muy altos, sin agrupar algunos de los datos tomados en cuenta para obtener el modelo. Así mismo en el Anexo 8 se realizó un estudio del rendimiento académico tomando en cuenta factores académicos, institucionales y personales.

La TABLA LXIII nos proporciona los resultados obtenidos por cada algoritmo donde se puede observar que el algoritmo ID3 proporciona mejor resultados en el conjunto de entrenamiento con un total de 94.84% de instancias clasificadas correctamente y 5.16% clasificadas incorrectamente, además el índice de kappa es alto con el 0.825, que significa que la coincidencia de la predicción con la clase real está muy de acuerdo, así mismo un error cuadrático de 0.039 y un error relativo de 7.55%; y en la validación cruzada el algoritmo C4.5 arroja los mejores resultados con un total de 90.98% de instancias clasificadas correctamente 9.02% clasificadas incorrectamente, índice de kappa 0.652 que significa que la coincidencia de la predicción con la clase real es considerada buen ajuste, así mismo un error cuadrático de 0.078 y error relativo de 14.90%.

En la siguiente figura se indican los resultados de las instancias clasificadas correctamente e incorrectamente luego de evaluar los diferentes algoritmos, donde se puede evidenciar que el algoritmo C4.5 es el que tiene menor margen de error, de 9.02% (ver Figura 23):

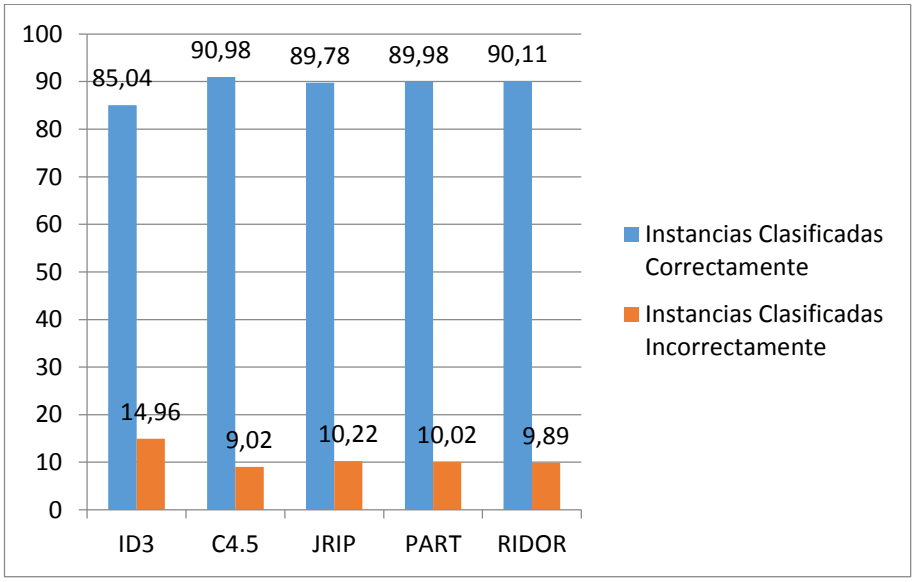


Figura 23: Instancias Clasificadas Correctamente e Incorrectamente

3.6. Demostrar la visualización de los resultados del rendimiento académico de los estudiantes mediante la herramienta de Minería de Datos.

Luego de implementar el dataset en la herramienta de minería de datos se procedió a demostrar los resultados obtenidos por cada uno de los algoritmos, los cuales fueron detallados en la sección anterior (ver Sección Resultados: 3.5. Evaluar el modelo computacional en un escenario real con datos académicos).

3.7. Interpretar los resultados obtenidos por la Herramienta de Minería de Datos acerca del rendimiento académico de los estudiantes.

3.7.1. EVALUACIÓN

En esta fase se evalúa el modelo, con el fin de comprobar que los resultados cumplen los objetivos, así mismo se obtiene el modelo que permite determinar el Rendimiento Académico y los parámetros que se utilizaron para construir el modelo y cuál de estos factores es el que más influye.

3.7.1.1. Evaluación de los Resultados

Esta tarea involucra la evaluación del modelo en relación a los objetivos, para la cual se evaluó el algoritmo que obtuvo mejores resultados que fueron analizados en la fase anterior (ver Sección Resultados: 3.5.1. Evaluación de Modelos) para determinar el rendimiento académico, por ende se utilizaron datos académicos, personales institucionales y socioeconómicos de los estudiantes del AEIRNNR, los mismos que permitieron obtener los siguientes resultados:

3.7.1.1.1. Estudio del Rendimiento Académico aplicando técnicas de minería de datos

Los resultados de este algoritmo clasificaron cada atributo en base a valores establecidos al atributo objetivo que es nota_promedio los cuales obtuvieron un valor de 1 cuando eran clasificadas correctamente en cualquiera de estos casos como bueno, malo y sobresaliente y 0 en caso contrario, es decir sino eran clasificadas correctamente, y aquellas que tienen valores reales, la que más se aproxime a 1, pertenece a dicho estado (ver figura 24).

Row No.	nota_promedio	confidence(malo)	confidence(bueno)	confidence(sobresaliente)	prediction(nota_promedio)
1	bueno	0.011	0.936	0.053	bueno
2	bueno	0.011	0.936	0.053	bueno
3	bueno	0.011	0.936	0.053	bueno
4	bueno	0.011	0.936	0.053	bueno
5	malo	0.686	0.314	0	malo
6	malo	0.686	0.314	0	malo
7	bueno	0.011	0.936	0.053	bueno
8	bueno	0.011	0.936	0.053	bueno
9	bueno	0.011	0.936	0.053	bueno
10	bueno	0.011	0.936	0.053	bueno
11	bueno	0.011	0.936	0.053	bueno
12	bueno	0.011	0.936	0.053	bueno
13	bueno	0.011	0.936	0.053	bueno
14	bueno	0.011	0.936	0.053	bueno

Figura 24. Clasificación de estudiantes mediante algoritmo C4.5

Así mismo se obtuvo el rendimiento académico del AEIRNNR donde este algoritmo obtuvo 380 malo, 2424 bueno y 4 sobresaliente durante la fase de entrenamiento; y 153 malo, 1209 bueno y 12 sobresaliente durante la fase de validación. Estos resultados se obtuvieron de acuerdo a los atributos que fueron tomados en cuenta (ver TABLA XL) para generar el modelo mediante el algoritmo C4.5. Mediante estos resultados se puede determinar que el rendimiento académico es bueno, con promedios de notas de 7.00 – 9.00 (ver Figura 25):

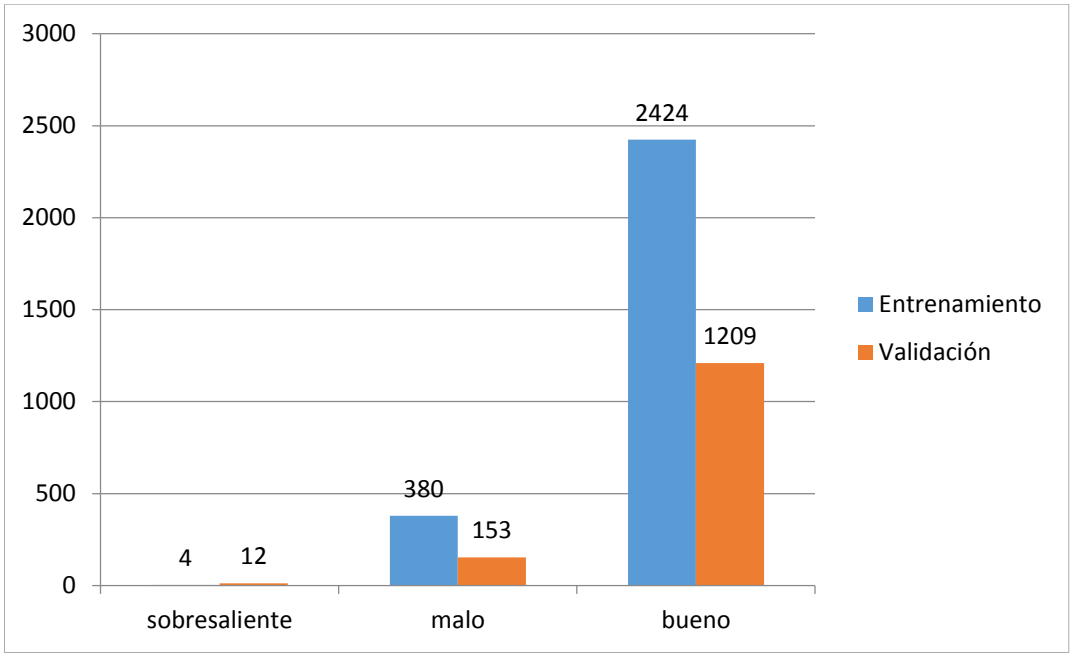


Figura 25: Rendimiento Académico de los estudiantes

Basado en las reglas, el rendimiento académico de los estudiantes es sobresaliente cuando:

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y el género es femenino y el padre trabaja y el estudiante es soltero y la madre trabaja y el estudiante pertenece al sector urbano.
- Si el promedio de asistencia es mayor a 90% y tiene entre 20 y 21 años y pertenece a los módulos de primero a tercero.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de primero a tercero y realizó sus estudios secundarios en un colegio público y no tiene hijos y pertenece a la carrera Ingeniería Electromecánica y no trabaja y el género es masculino y el padre trabaja y el estudiante es soltero y la madre trabaja; entonces el rendimiento académico es sobresaliente.

Basado en las reglas, el rendimiento académico de los estudiantes es bueno cuando:

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante no trabaja.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y pertenece a la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial y el género es masculino.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y tiene hijos.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y no trabaja.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y trabaja realizó sus estudios secundarios en un colegio público y es mayor a 21 años y es mestizo y el promedio de asistencia es mayor a 90% y no tiene hijos y pertenece a la carrera Ingeniería Electrónica y Telecomunicaciones y el género es masculino y es soltero y el padre trabaja y la madre trabaja.

- Si el promedio de asistencia es mayor a 90% y es menor a 20 años y pertenece a los módulos de cuarto a sexto.
- Si estado de matrícula es aprobado y horario de clases es matutino y es mestizo y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto.
- Si no posee un tipo de beca y es mestizo y es soltero y tiene entre 20 y 21 años y pertenece a otra provincia.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a los módulos de cuarto a sexto y el padre trabaja y es mayor a 21 años.
- Si el promedio de asistencia es mayor a 90% y pertenece a los módulos de cuarto a sexto y el género es femenino y pertenece al sector urbano.

Basado en las reglas, el rendimiento académico de los estudiantes es malo o bajo cuando:

- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de primero a tercero.
- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto y es mestizo y no tiene hijos y no trabaja y es masculino y es soltero y el padre trabaja y la madre trabaja y pertenece a un cantón de la provincia de Loja o a otra provincia.
- Si el promedio de asistencia es menor a 80% y es menor a 20 años.
- Si estado de matrícula es reprobado y es menor a 20 años y pertenece a los módulos de primero a tercero.
- Si estado de matrícula es reprobado y pertenece a otra provincia y es mayor a 21 años y el promedio de asistencia es menor a 80%.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a la carrera de Ingeniería Electrónica y Telecomunicaciones y el padre trabaja.
- Si no posee un tipo de beca y es mestizo y es soltero y no trabaja y no tiene hijos y pertenece a los módulos de séptimo a décimo.

3.7.1.1.2. Factores de Rendimiento Académico

El Rendimiento Académico es un indicador del nivel de aprendizaje alcanzado por el estudiante y no siempre es lineal, es decir está relacionada por factores [66].

En el presente Trabajo de Titulación el objetivo es determinar el Rendimiento Académico aplicando Técnicas de Minería de Datos, para el cual se determinaron atributos que permitan obtener este objetivo, los mismos que se encuentran asociados entre sí, los cuales son: datos académicos, personales, institucionales y socioeconómicos del estudiante los cuales se detalla a continuación:

- Académicos: modulo, promedio_asistencia, estado_matricula, tipo_colegio_estudiante.
- Personales: estado_civil, género, edad, etnia_estudiante, numero_hijos_estudiante, procedencia_estudiante.
- Institucionales: nombre_carrera, tipo_beca, horario.
- Socioeconómicos: situacion_laboral_estudiante, situacion_laboral_padre, situacion_laboral_madre.

Luego de clasificar los atributos, a través del algoritmo C4.5 se obtuvieron pesos para cada uno, que permitieron determinar cuáles inciden con mayor relevancia en el Rendimiento Académico (ver TABLA LXIV).

TABLA LXIV.

PESO DE LOS ATRIBUTOS

Factor Influyente	Atributo	Peso	Total	% por Factor
Académico	modulo	0.65	2.77	26,03
	promedio_asistencia	0.62		
	estado_matricula	0.74		
	tipo_colegio_estudiante	0.76		
Personal	estado_civil	0.82	4.12	38,72
	genero	0.67		
	edad	0.81		

	etnia_estudiante	0.45		
	numero_hijos_estudiante	0.69		
	procedencia_estudiante	0.68		
Institucional	tipo_beca	0.84	1.83	17,20
	nombre_carrera	0.56		
	horario	0.43		
Socioeconómicos	situacion_laboral_estudiante	0.83	1.92	18,05
	situacion_laboral_padre	0.58		
	situacion_laboral_madre	0.51		

De acuerdo a estos resultados se puede determinar que el factor que más incide en el Rendimiento Académico es el factor Personal con un porcentaje del 38,72% (ver Figura 26).

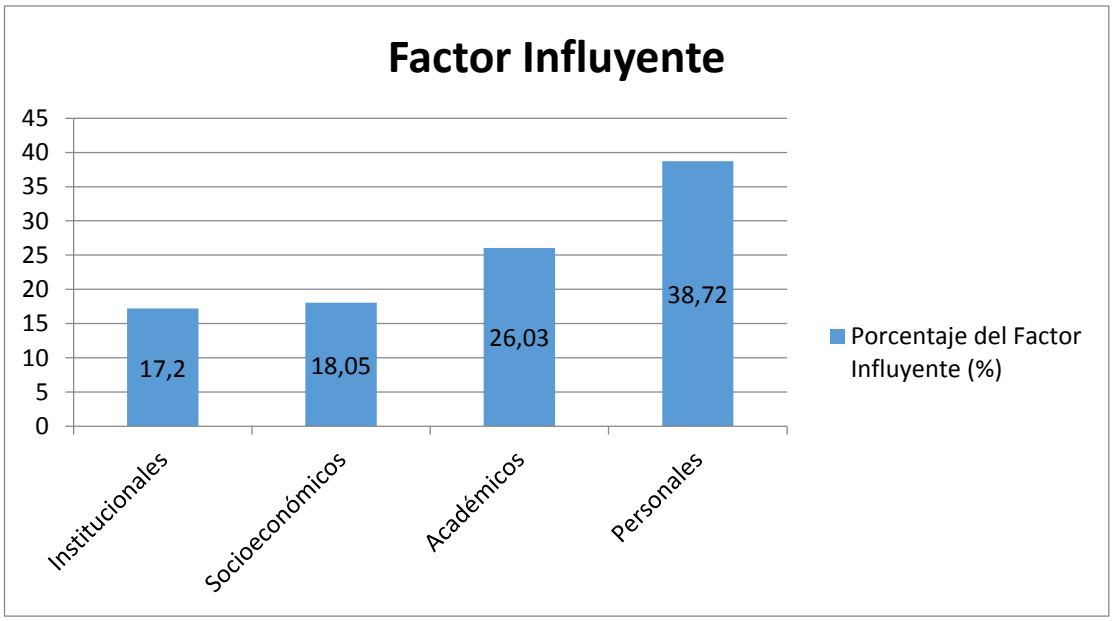


Figura 26. Factor Influyente del Rendimiento Académico.

3.7.1.2. Modelos Aprobados

Luego de validar los modelos obtenidos con cada uno de los algoritmos donde fueron considerados diferentes parámetros para su validación, el modelo que fue aprobado para determinar el Rendimiento Académico es el que se obtuvo a partir del algoritmo C4.5, debido a que permitió obtener los mejores resultados.

g. Discusión

1. Desarrollo de la propuesta alternativa

- **OBJETIVO ESPECÍFICO 1: Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.**

  Durante esta fase se procedió a realizar la búsqueda de información sobre las diferentes técnicas de Minería de Datos y casos de éxito donde se hayan aplicado para determinar el rendimiento académico, así mismo se realizó un análisis comparativo de las mismas con el fin de seleccionar la que más se adapte al presente Trabajo de Titulación.

- **OBJETIVO ESPECÍFICO 2: Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.**

  En esta fase primeramente se realizó la recolección inicial de los datos relacionados con el problema los cuales están almacenados en una base de datos, también se realizó un análisis de los mismos con el fin de identificar las relaciones entre ellos, estos datos corresponden a los estudiantes del AEIRNNR de la UNL. Además en esta etapa se transformarán los datos para que puedan ser usados de manera eficiente por técnicas de minería de datos y se construyó una estructura de datos para obtener el modelo.

- **OBJETIVO ESPECÍFICO 3: Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.**

  En esta fase se procedió a implementar la estructura de datos en una herramienta de Minería de Datos, para la cual se realizó un estudio comparativo de las diferentes herramientas donde se pueda implementar esta estructura y se seleccionó la más idónea, así mismo se aplicaron los algoritmos pertenecientes a la técnica de Clasificación, a través de los cuales se obtuvieron diferentes modelos, los cuales fueron comparados con el fin de obtener el modelo que permita determinar el rendimiento académico.

2. Valoración técnica económica ambiental

En la realización del presente Trabajo de Titulación se analizó la información académica de los estudiantes para medir el rendimiento de los mismos, con el fin que las autoridades de la universidad tomen decisiones que permitan corregir las deficiencias encontradas.

En el ámbito académico permitió adquirir capacidades y habilidades necesarias para poder llevar con éxito el desarrollo del proyecto, además se obtuvo conocimientos y experiencia que permitieron resolver problemas, así mismo permitió en las universidades tener en cuenta los factores que influyen en el rendimiento académico de los estudiantes.

Así mismo para recoger la información académica de los estudiantes se usó de técnicas que permitan recopilar información apropiada. También se utilizó técnica de Minería de Datos adecuada para analizar la información obtenida y extraer conocimiento de la misma que permitió evaluar el rendimiento de los estudiantes, además se utilizó la herramienta de Minería de Datos para implementar el modelo sobre el rendimiento académico.

Igualmente se tuvo el recurso humano y económico necesario para poder realizar el presente proyecto, así como el tiempo necesario que implica el desarrollo del mismo y la guía prestada por el tutor correspondiente.

Además contribuyó a la reducción de impactos negativos al medio ambiente, debido a que los resultados obtenidos serán visualizados en la herramienta de Minería de Datos ahorrando de esta manera recursos tales como: papel y tinta.

Con lo descrito anteriormente se pudo alcanzar los resultados esperados en cuanto al estudio del rendimiento académico aplicando técnicas de Minería de Datos.

h. Conclusiones

Luego de finalizar el Trabajo de Titulación se obtuvieron las siguientes conclusiones:

- ✓ La utilización de minería de datos permitió realizar una comprensión de los datos y descubrir factores que influyen en el rendimiento académico de los estudiantes.

- ✓ El empleo de técnicas de minería de datos ayudan a obtener modelos computacionales, ya que contiene diferentes algoritmos que ayudan a manipular los datos para obtener resultados que permita determinar el rendimiento académico. Sin embargo mediante la técnica de Clasificación a través de los diferentes tipos de algoritmos evaluados, C4.5 permitió obtener los mejores resultados de acuerdo a los datos académicos, institucionales y personales de los estudiantes.

- ✓ La herramienta de minería de datos RAPIDMINER fue fundamental en el desarrollo de este proyecto, ya que contiene gran número de algoritmos y complementos que permite hacer uso de diferentes algoritmos de otras herramientas, además tiene operadores que ayudan a facilitar el desarrollo de los procesos para crear los modelos aplicables para el análisis de los datos.

- ✓ Mediante el modelo de minería de datos obtenido y evaluado a través de datos reales se pudo comprobar que el rendimiento académico de los estudiantes está considerado bueno, con promedio de notas entre 7.00 - 9.00.

i. Recomendaciones

Luego de finalizar el Trabajo de Titulación son importantes las siguientes recomendaciones:

- Tener en cuenta los factores obtenidos que más influyen en el rendimiento académico con el fin de tomar decisiones y proponer estrategias que permitan ayudar a los estudiantes a mejorar el rendimiento académico.

- Aplicar nuevas técnicas de minería de datos como reglas de asociación, agrupamiento o clustering con el fin de obtener diferentes resultados del rendimiento académico a los que se obtuvieron en este Trabajo de Titulación.

- Agrupar los datos que se van a utilizar para la minería de datos, con el propósito de obtener mejores resultados al momento de aplicar algoritmos de minería de datos.

j. Bibliografía

[1]. Corso Cynthia Lorena, Colacioppo Nicolás Leonardo, Descubrimiento de factores relacionados al aprendizaje de resolución de problemas en el contexto de Ingeniería, Universidad Tecnológica Nacional - Facultad Regional Córdoba, En línea: http://conaiisi.frc.utn.edu.ar/PDFsParaPublicar/1/schedConfs/4/17-471-1-DR.pdf

[2]. Eckert Karina Beatriz, Exploración de Datos Académicos a través de la aplicación de Técnicas de Minería de Datos en Weka, Universidad Gastón Dachary - Ingeniería en Informática con Orientación en Sistemas de Información, [En línea]: http://www.42jaiio.org.ar/proceedings/simposios/Trabajos/EST/16.pdf

[3]. Ernesto González Díaz, Zady Pérez Hernández, Ivet Espinosa Conde, Susel Álvarez Reyes, Obtención De Patrones Y Reglas En El Proceso Académico De La Universidad de las Ciencias Informáticas Utilizando Técnicas de Minería de Datos, [En línea]: http://eprints.rclis.org/10937/1/Suzel.pdf

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

[5]. Felipe de Jesús Núñez Cárdenas, Raúl Hernández Palacios, Víctor Tomás Tomás Mariano, Ana María Felipe Redondo, Identificación de Estilos de Aprendizaje en Alumnos Universitarios de Computación de La Huasteca Hidalguense Mediante Técnicas de Minería de Datos, Universidad Autónoma del Estado de Hidalgo, [En línea]: http://www.uaeh.edu.mx/scige/boletin/huejutla/n2/a1.html

[6]. Paola García García, Carlos Azaustre Rodríguez, Minería de Datos aplicada a las Redes Sociales, Universidad Carlos III de Madrid - I.T.T. Telemática, [En línea]: http://www.it.uc3m.es/jvillena/irc/practicas/08-09/08.pdf

[7]. Carlos Márquez Vera, Cristóbal Romero Morales y Sebastián Ventura Soto, Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos, [En línea]: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

[8]. Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda, Minería de Datos, Universidad Carlos III de Madrid – Ingeniería Telemática, [En línea]: http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/07-mineria-de-datos

[9]. José Ignacio González Gómez, Generalidades de la Minería de Datos, Universidad de La Laguna – Departamento de Economía Financiera y Contabilidad, [En línea]: http://www.ecofin.ull.es/users/jggomez/D%20Bdr_Erp/6%20Mineria/Mineria.pdf

[10]. Sofía J. Vallejos, Minería de Datos, Universidad Nacional del Nordeste - Facultad de Ciencias Exactas, Naturales y Agrimensura, [En línea]: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf

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

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

[13]. Corso Cynthia Lorena, Gibellini Fabián, Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios, Universidad Tecnológica Nacional - Departamento de Ingeniería en Sistemas de Información - Laboratorio de Sistemas de Información, [En línea]: http://www.41jaiio.org.ar/sites/default/files/16_JSL_2012.pdf

[14]. José Manuel Molina López y Jesús García Herrero, Técnicas de Análisis de Datos, Instituto Tecnológico Superior de Calkiní en el Estado de Campeche, [En línea]: http://www.itescam.edu.mx/principal/sylabus/fpdb/recursos/r94663.PDF

[15]. Agustín José Calleja Gómez, Minería de Datos con Weka para la Predicción del Precio de Automóviles de Segunda Mano, Universidad Politécnica de Valencia - Escuela Técnica Superior De Informática Aplicada, [En línea]: http://riunet.upv.es/bitstream/handle/10251/10097/PFC_DSIC-80_Agust%C3%ADnCalleja.pdf

[16]. José Antonio García Bermúdez y Ángela María Acevedo Ramírez, Análisis para Predicción de Ventas Utilizando Minería de Datos en Almacenes de Ventas de Grandes Superficies, Universidad Tecnológica de Pereira - Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación - Ingeniería de Sistemas y Computación, [En línea]: http://repositorio.utp.edu.co/dspace/bitstream/11059/1339/1/006312G216.pdf

[17]. Braulio José Solano Rojas, Introducción a la Minería de Datos, Universidad de Costa Rica

[18]. Algoritmo ID3, Centro de Estudios y Aplicaciones Logísticas, Facultad de Ingeniería de la Universidad Nacional de Cuyo, En línea: http://ceal.fing.uncu.edu.ar/data_mining/Algoritmos/algoritmo1.pdf

[19]. Alejandro Guerra Hernández, Arboles de Decisión, Universidad Veracruzana - Departamento de Inteligencia Artificial, En línea: http://www.uv.mx/aguerra/documents/2009-mpi-12.pdf

[20]. Guillermo Roberto Solarte Martínez, José A. Soto Mejía, Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares, Universidad Tecnológica de Pereira, En línea: http://revistas.utp.edu.co/index.php/revistaciencia/article/viewFile/1487/947

[21]. Nora Marcela Aguilar Caro, Aplicación de Métodos de Aprendizaje Automático para la Desambiguación Del PP Attachment en español.

[22]. Vanessa Aguiar Pulido, Resolución de problemas de optimización combinatoria utilizando técnicas de computación evolutiva. Una aplicación a la biomedicina. Universidad de la Coruña - Tecnologías de la Información y las Comunicaciones. En

línea:
http://ruc.udc.es/bitstream/2183/11930/2/AguiarPulido_Vanessa_TD_2014.pdf

[23]. Benjamín Moreno Montiel, Minería Sobre Grandes Cantidades de Datos, Universidad Autónoma Metropolitana - Ciencias y Tecnologías de la Información, En línea:
http://mcyti.izt.uam.mx/archivos/Tesis/Generaci%F3n2007/ICR_Benjam%EDnMoreno.pdf

[24]. Heriberto Cruz Hernández, Minería de Datos, Centro de Investigación y Estudios Avanzados del IPN

[25]. Servente, M. García-Martínez, R. Algoritmos TDIDT Aplicados a la Minería de Datos Inteligente, En línea: http://laboratorios.fi.uba.ar/lsi/R-ITBA-26-datamining.pdf

[26]. Ronald Augusto Velandia Ortega, Fredy Leonardo Hernández Suárez, Evaluación de Algoritmos de Extracción de Reglas de Decisión para el Diagnóstico de Huecos de Tensión, Universidad Industrial de Santander - Facultad de Ingenierías Físico-Mecánicas - Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones, En línea: http://tangara.uis.edu.co/biblioweb/tesis/2010/134742.pdf

[27]. Julio Villena Román, Minería de datos, Universidad Carlos III de Madrid – Departamento de Ingeniería Telemática, En línea: http://www.it.uc3m.es/jvillena/irc/practicas/03-04/18.mem.pdf

[28]. Frank E, Witten IH. Generating Accurate Rule Sets Without Global Optimization. In: Shavlik J, editor. Proceedings of the Fifteenth International Conference on Machine Learning; 1998; San Francisco, CA: Morgan Kaufmann Publishers; 1998.

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

[30]. J.L. Cubero, F. Berzal, F. Herrera, Fundamentos de Minería de Datos, Universidad de Granada - Dpto. Ciencias de la Computación e I.A., [En línea]: http://sci2s.ugr.es/docencia/m1/Preprocesamiento-Weka-MD.pdf

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

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

[33]. José Hernández Orallo y César Ferri Ramírez, Introducción a WEKA, Universidad Politécnica de Valencia - Departamento de Sistemas Informáticos y Computación, [En línea]: http://users.dsic.upv.es/~jorallo/docent/doctorat/weka.pdf

[34]. Abdelmalik Moujahid e Iñaki Inza, Manual de prácticas de minería de datos usando el software WEKA, University of the Basque Country - Department of Computer Science and Artificial Intelligence, [En línea]: https://addi.ehu.es/bitstream/10810/4627/1/tr10-1.pdf

[35]. Práctica de laboratorio de aprendizaje inductivo, Universidad Politécnica de Cataluña – Facultad de Informática, [En línea]: http://www.lsi.upc.edu/~bejar/apren/lab/apind13141q.pdf

[36]. Francisco José García González, Aplicación de Técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Universidad de Granada

[37]. Juan Carlos Díez, Iván Martín y Manuel Aranda, RAPIDMINER, Grupo de Investigación – Departamento de Lenguajes y Sistemas Informáticos [En línea]: http://www.kybele.etsii.urjc.es/docencia/SI_GII_M/2012-2013/Material/[Expo]Presentacion%20RAPID%20MINER.pdf

[38]. ORACLE, Oracle Data Mining, [En línea]: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html

[39]. Yanchang Zhao (2013), R and Data Mining: Examples and Case Studaies, Informáticos [En línea]: http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

[40]. Sergio Hernández, (2012), Minería de Datos de gran escala usando R, Universidad Católica del Maule – Laboratorio de Procesamiento de Información Geoespacial.

[41]. The R User Conference, University of Castilla - La Mancha, [En línea]: http://www.edii.uclm.es/~useR-2013/docs/useR2013_abstract_booklet.pdf

[42]. Daniel Tuttle Ospina, Minería de Datos en el paquete Rattle del Lenguaje R, Universidad Nacional de Colombia – Escuela de Sistemas, [En línea]: http://tecaprendizajeest.wikispaces.com/file/view/Miner%C3%ADa+de+datos+con+eL+paquete+rattle+-+tutorial.pdf

[43]. FACTOMINER, Porque utilizar FactoMineR, [En línea]: http://factominer.free.fr/

[44]. Metodologías, Tareas y Técnicas de Minería de Datos, Escuela de Administración, Finanzas e Instituto Tecnológico, [En línea]: http://bdigital.eafit.edu.co/bachelorThesis/005.74CDR397/marcoTeorico.pdf

[45]. Metodología de Aplicación del Data Mining (DM), Universidad Politécnica Salesiana, [En línea]: http://www.dspace.ups.edu.ec/bitstream/123456789/47/10/Capitulo4.pdf

[46]. Juan Miguel Moine, Ana Silvia Haedo y Silvia Gordillo, Estudio comparativo de metodologías para minería de datos, Universidad Nacional de La Plata - Facultad de Informática, [En línea]:

http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

[47]. Manual CRISP-DM de IBM SPSS Modeler, IBM, [En línea]: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

[48]. Hernando Camargo y Mario Silva, Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP, Universidad el Bosque – Ingeniería en Sistemas, [En línea]: http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

[49]. Juan Ángel Vanrell, Un Modelo de Procesos para Proyectos de Explotación de Información, Universidad Tecnológica Nacional - Ingeniería en Sistemas de Información, [En línea]: http://www.unla.edu.ar/sistemas/gisi/tesis/vanrell-tesisdemagister.pdf

[50]. José Alberto Gallardo Arancibia, Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM), Universidad Politécnica de Madrid - Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software - Facultad de Informática, [En línea]: http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

[51]. Edgar Valencia Romero, Aplicación de las redes neuronales a la minería de datos, Universidad Nacional Autónoma de México – Facultad de Ciencias, [En línea]: http://www.dynamics.unam.edu/DinamicaNoLineal2/docencia/Tesis/Tesisedgar.pdf

[52]. Norka Gómez López, René Iván González Fernández, Alejandro Rosete Suárez, Predicción de complicaciones cardiacas utilizando Minería de Datos: Estado del Arte, Instituto Superior Politécnico "José Antonio Echeverría", [En línea]: http://ccia.cujae.edu.cu/index.php/siia/siia2008/paper/viewFile/1155/234

[53]. Enrique J. Fernández, Asistente para la Gestión de Documentos De Proyectos de Explotación de Datos, Universidad Politécnica de Madrid – Magister en Ingeniería de Software, [En línea]: http://iidia.com.ar/rgm/tesistas/fernandez-tesisdemagister.pdf

[54]. Juan Miguel Moine, Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo, Universidad Nacional de La Plata - Facultad de Informática, [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1

[55]. Manuel Alfredo Pech Palacio, Minería de Datos Espaciales, Universidad de las Américas Puebla [En línea]: http://catarina.udlap.mx/u_dl_a/tales/documentos/msp/pech_p_ma/capitulo2.pdf

[56]. Carlos Eduardo Bedregal Lizárraga, Agrupamiento de Datos utilizando técnicas MAM-SOM, Universidad Católica San Pablo, [En línea]: http://users.dcc.uchile.cl/~cbedrega/publications/Tesis.pdf

[57]. J. Sanabria Garzón, Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa, Universidad de Los Llanos, Red de Revistas Científicas de América Latina, el Caribe, España y Portugal Sistema de Información Científica, [En línea]: http://www.redalyc.org/pdf/896/89680103.pdf

[58]. Braulio José Solano Rojas, Tareas de la minería de datos: reglas de asociación y secuencias, Universidad de Costa Rica, [En línea]: http://bsolano.com/ecci/claroline/backends/download.php/UHJlc2VudGFjaW9uZXMvNy5fVGFyZWFzX2RlX2xhX21pbmVy7WFfZGVfZGF0b3MsX3JlZ2xhc19kZV9hc29jaWFjafNuLnBkZg%3D%3D?cidReset=true&cidReq=CI2352

[59]. Reglas de Asociación, Coordinación de Ciencias Computacionales (CCC) del Instituto Nacional de Astrofísica, Óptica y Electrónica [En línea]: http://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/apriori.pdf

[60]. Neftalí de Jesús Calderón Méndez, Minería de Datos una Herramienta para la Toma de Decisiones, Universidad de San Carlos de Guatemala - Escuela de Ingeniería en Ciencias y Sistemas, [En línea]: http://biblioteca.usac.edu.gt/tesis/08/08_0307_CS.pdf

[61]. R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica, Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos, Universidad Politécnica de Valencia [En línea]: http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf

[62]. Osvaldo M. Spositto, Martín E. Etcheverry, Hugo L. Ryckeboer, Julio Bossero, Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil, Universidad Nacional de La Matanza - Departamento de Ingeniería e Investigaciones Tecnológicas, [En línea]: http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

[63]. Porcel, Eduardo; Dapozo, Gladys; López, María V., Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios, Universidad Nacional del Nordeste - Departamento de Informática, [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/19846/Documento_completo.pdf?sequence=1

[64]. José M. Molina, Jesús García, Capítulo 3. Técnicas de Minería de Datos basadas en Aprendizaje Automático.

[65]. Paula Andrea Vizcaíno Garzón, Aplicación de Técnicas de Inducción de Árboles de Decisión a Problemas de Clasificación Mediante el uso de Weka (Waikato Environment For Knowledge Analysis), Fundación Universitaria Konrad Lorenz - Facultad de Ingeniería de Sistemas, [En línea]: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf

[66]. Rendimiento Académico, Universidad Francisco Gavidia – Facultad de Ingeniería y Arquitectura, Ingeniería en Ciencias de la Computación, [En línea]: http://wwwisis.ufg.edu.sv/wwwisis/documentos/TE/371.262-B634f/371.262-B634f-CAPITULO%20II.pdf

[67]. Marco Besteiro, Miguel Rodríguez, Web Services, En línea: http://www.ehu.es/mrodriguez/archivos/csharppdf/ServiciosWeb/WebServices.pdf

[68]. Marco Antonio Cruz Chávez, Conceptos básicos de bases de datos, Universidad Autónoma del Estado de Morelos - Centro de Investigaciones en Ingeniería y Ciencias Aplicadas, En línea: http://www.gridmorelos.uaem.mx/~mcruz//cursos/miic/bd1.pdf

[69]. Sistemas gestores de bases de datos, Pontificia Universidad Católica del Ecuador - Dirección de Informática, En línea: ftp://ftp.puce.edu.ec/Facultades/Ingenieria/Sistemas/Base%20de%20Datos%20II/Sistemas%20Gestores%20de%20Bases%20de%20Datos%20Capitulo%201.pdf

[70]. María N. Moreno García, Luis A. Miguel Quintales, Francisco J. García Peñalvo y M. José Polo Martín, Aplicación de Técnicas de Minería de Datos en la Construcción y Validación de Modelos Predictivos y Asociativos a Partir de Especificaciones de Requisitos de Software, Universidad de Salamanca - Departamento de Informática y Automática, En línea:, http://ceur-ws.org/Vol-84/paper4.pdf

[71]. Susan Goldman, Richard Golden, Paul van den Broek, ¿Por qué son útiles los modelos computacionales de comprensión de textos?, En línea: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342007000300008

[72]. Ministerio de Relaciones Laborales, Presupuesto General, [En línea]: http://www.relacioneslaborales.gob.ec/

[73]. Introducción al Aprendizaje Automático y a la Minería de Datos con Weka, Herramientas de la Inteligencia Artificial, Universidad Carlos III de Madrid - Ingeniería Informática

[74]. Oldemar Rodríguez, Validación Cruzada (cross-validation) y Remuestreo (bootstrapping), En línea: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci%C3%B3n_-_CV.293124233.pdf

[75]. M. Correa, C. Bielza, J. Pamies-Teixeira, J. R. Alique, Redes Bayesianas vs Redes Neuronales en Modelos para la Predicción del Acabado Superficial, Consejo Superior de Investigaciones Científicas, En línea: http://digital.csic.es/bitstream/10261/13826/1/redes.pdf

k. Anexos

Anexo 1: Proceso para construir el Modelo mediante el algoritmo ID3

Para poder realizar el proceso que permita obtener el modelo se enlazaron diferentes operadores (descripción de operadores ver Sección Anexos: ANEXO 6), el proceso que se obtuvo es el siguiente (ver figura 27):

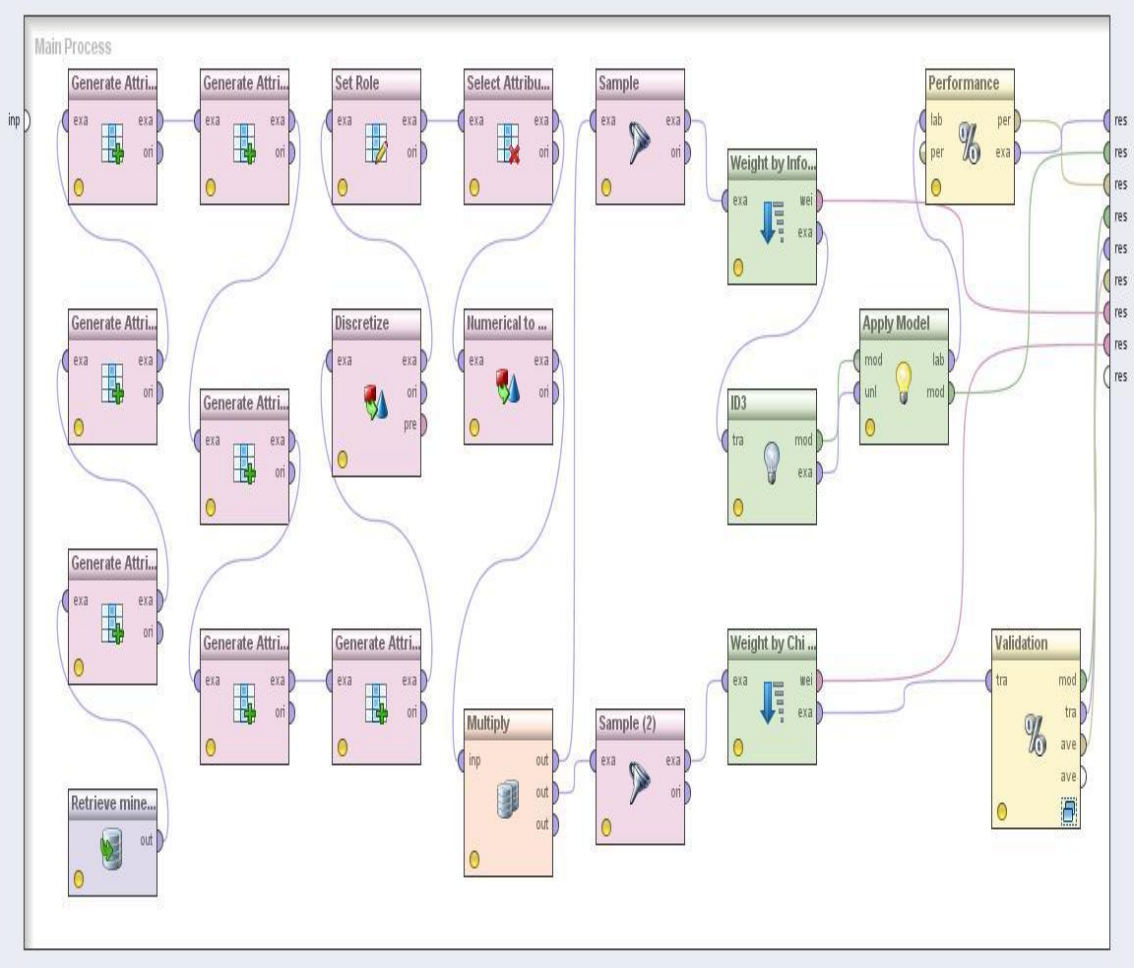


Figura 27: Proceso del algoritmo ID3

Anexo 2: Proceso para construir el Modelo mediante el algoritmo C4.5

Para poder realizar el proceso que permita obtener el modelo se enlazaron diferentes operadores (descripción de operadores ver Sección Anexos: ANEXO 6), el proceso que se obtuvo es el siguiente (ver figura 28):

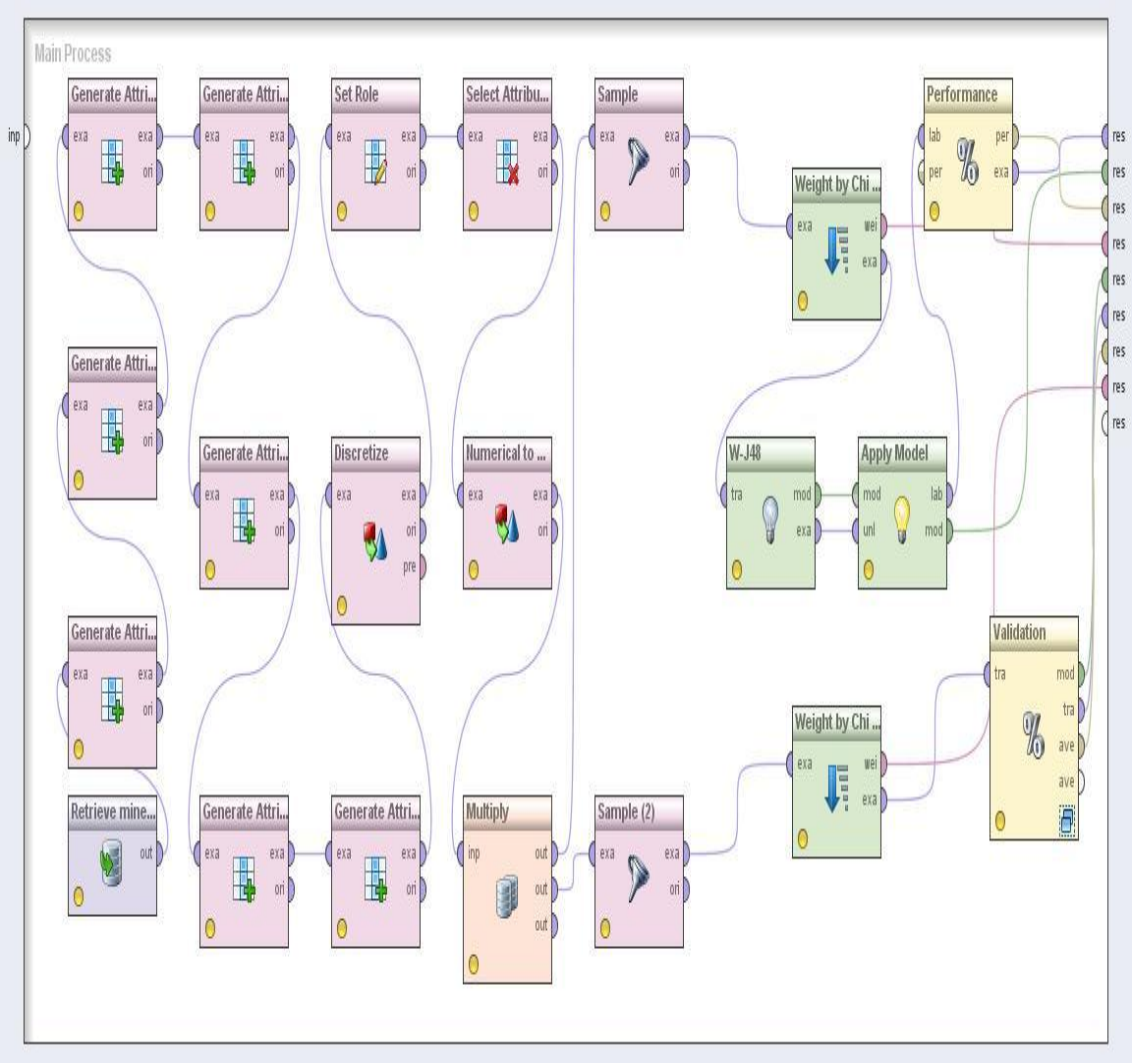


Figura 28: Proceso del algoritmo C4.5

Anexo 3: Proceso para construir el Modelo mediante el algoritmo JRIP

Para poder realizar el proceso que permita obtener el modelo se enlazaron diferentes operadores (descripción de operadores ver Sección Anexos: ANEXO 6), el proceso que se obtuvo es el siguiente (ver figura 29):

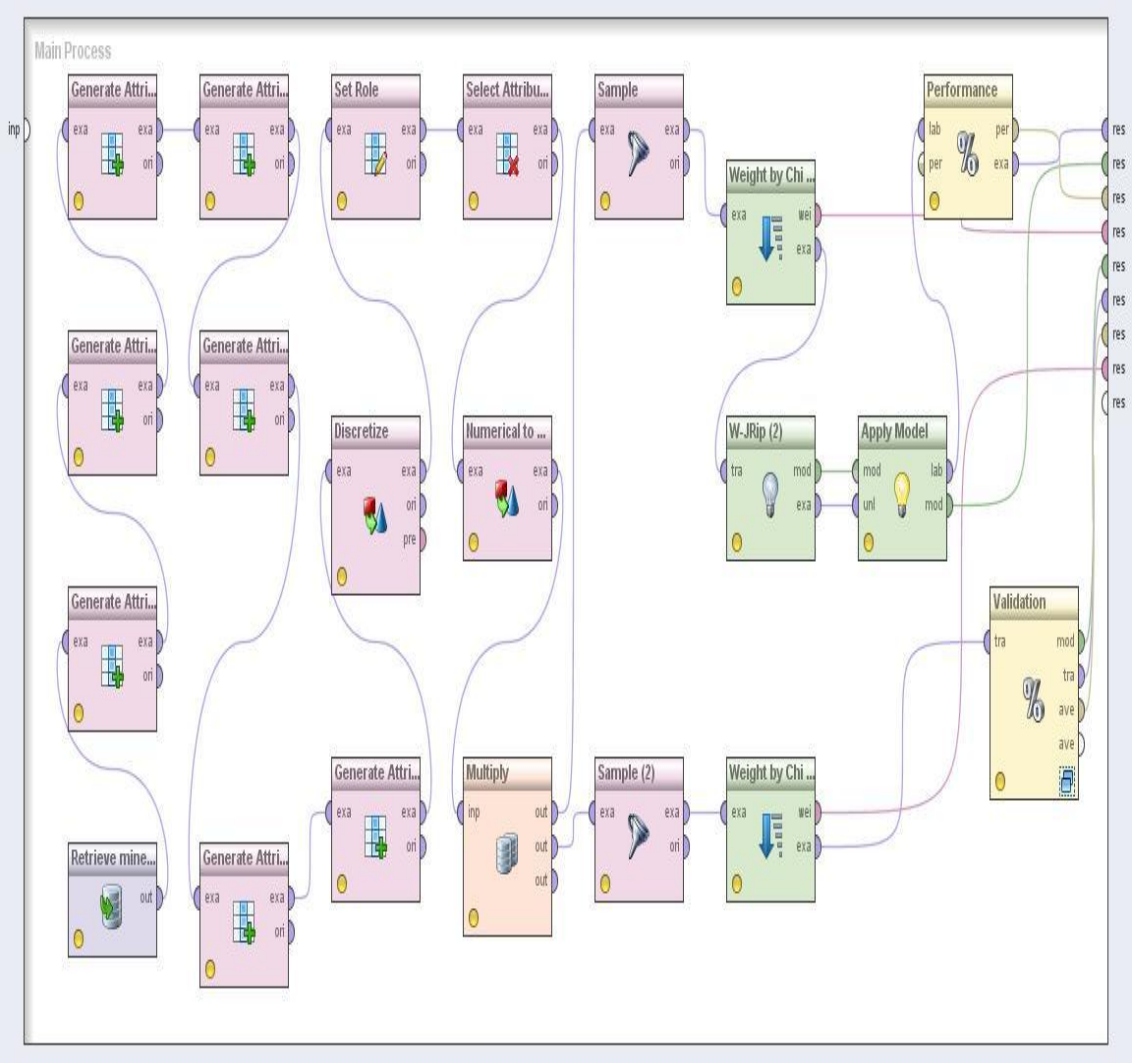


Figura 29: Proceso del algoritmo JRIP

Anexo 4: Proceso para construir el Modelo mediante el algoritmo PART

Para poder realizar el proceso que permita obtener el modelo se enlazaron diferentes operadores (descripción de operadores ver Sección Anexos: ANEXO 6), el proceso que se obtuvo es el siguiente (ver figura 30):

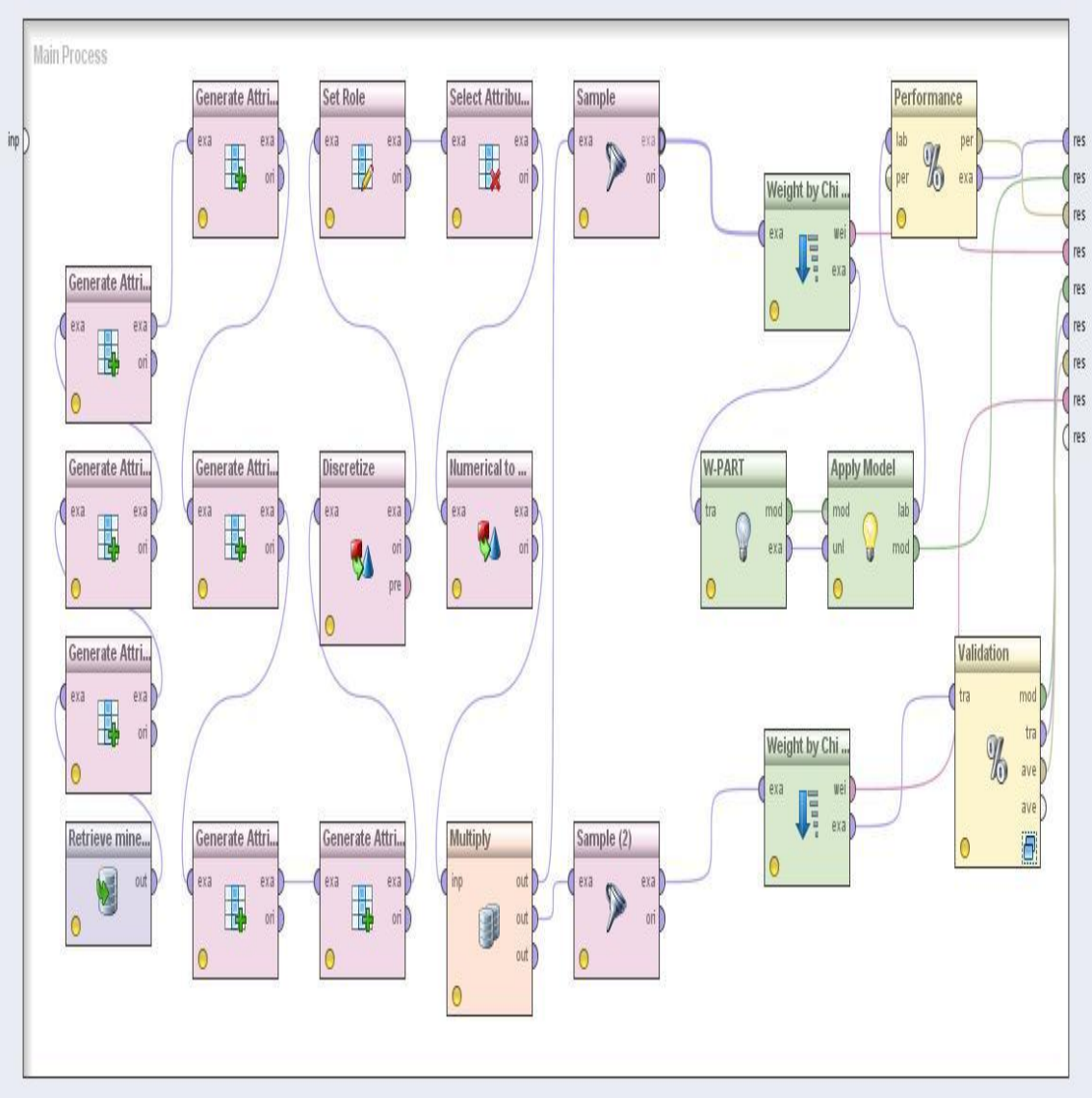


Figura 30: Proceso del algoritmo PART

Anexo 5: Proceso para construir el Modelo mediante el algoritmo RIDOR (Ripple Down Rule)

Para poder realizar el proceso que permita obtener el modelo se enlazaron diferentes operadores (descripción de operadores ver ANEXO 6), el proceso que se obtuvo es el siguiente (ver figura 31):

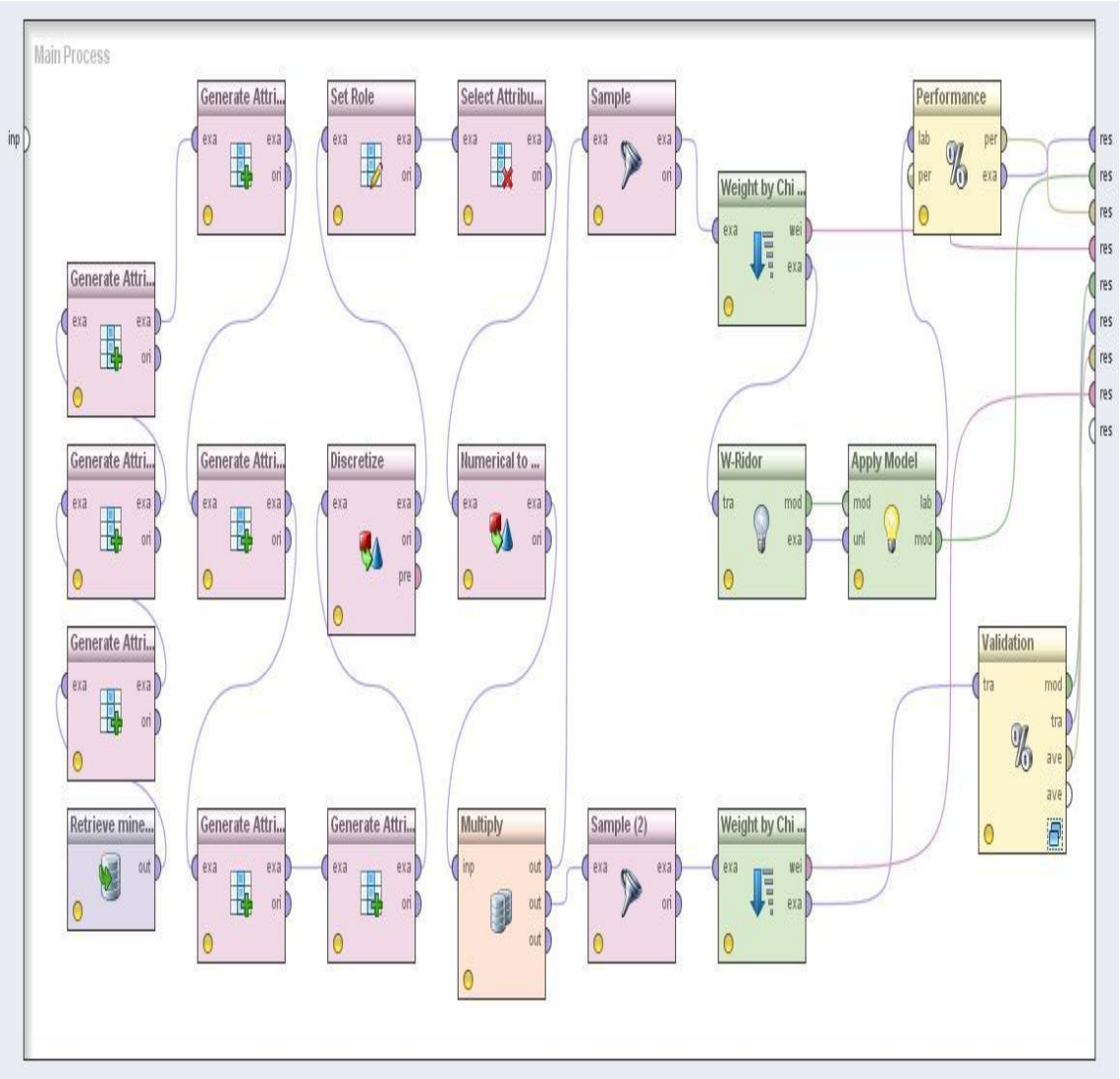


Figura 31: Proceso del algoritmo RIDOR

Anexo 6: Resultados obtenidos sin agrupar los datos

Luego de realizar todo el proceso de datos y sin agrupar algunos de los datos tomados en cuenta para crear el modelo, los resultados obtenidos fueron bajos, con márgenes de error muy altos, los atributos que no se agruparon fueron: edad, modulo y asistencia_promedio.

En la siguiente tabla se detallan los resultados obtenidos (ver TABLA LXV):

TABLA LXV.

RESULTADOS DE LOS ALGORITMOS CON DATOS NO AGRUPADOS

Algoritmo	Conjunto de Datos	*Instancias correctamente clasificadas (%)*	*Instancias incorrectamente clasificadas (%)*	*Índice de Kappa*	Error Cuadrático	Error Relativo (%)	Error Cuadrático Medio	Error Cuadrático Relativo
ID3	C E	97.53	2.47	0.963	0.014	2.65	0.117	0.183
	C P	36.37	63.73	0.146	0.485	48.66	0.695	desconocido
C4.5	C E	84.45	15.55	0.761	0.101	19.12	0.317	0.498
	C P	55.31	44.69	0.293	0.386	50.88	0.621	0.947
JRIP	C E	73.13	26.87	0.581	0.193	35.27	0.439	0.689
	C P	55.84	44.16	0.297	0.344	50.00	0.586	0.906
PART	C E	81.35	18.65	0.714	0.117	22.32	0.341	0.536
	C P	55.71	44.29	0.326	0.374	45.71	0.611	0.932
RIDOR	C E	62.35	37.65	0.387	0.377	37.65	0.618	0.970
	C P	57.98	42.02	0.303	0.420	42.02	0.640	0.975

Anexo 7: Operadores utilizados mediante la Herramienta de Minería de Datos

- **Generate Attributes:** Este operador construye nuevos atributos a partir de los atributos del conjunto de ejemplos de entrada, por ejemplo se lo utilizo en el parámetro estado_matricula para asignar el valor de 0 si el estado de matrícula es Aprobada, de lo contrario se le asignó 1.

- **Discretize:** Este operador permite convertir todos los atributos numéricos del conjunto de datos en atributos nominales, en este proceso permitió cambiar el parámetro numérico nota_promedio y asignar la clase a la que pertenece como valores menores a 7.00 malo y de 7.00-9.00 bueno y 9.00-10.00 sobresaliente.

- **Set Role:** Permite determinar el atributo objetivo para el aprendizaje, el mismo que es nota_promedio.

- **Select Attributes:** Permite seleccionar los atributos para crear el modelo.

- **Numerical to Polynominal:** Permite cambiar el tipo de variable de numérico a polinomial.

- **Multiply:** Permite dividir el conjunto de datos en dos subconjuntos uno para entrenamiento y otro para pruebas.

- **Sample:** Permite obtener una muestra para el conjunto de entrenamiento con un total de 67% y para pruebas del 33% del total de datos.

- **Weight by Chi Squared Statistic:** Permite obtener el peso de cada atributo.

- **Apply Model:** Este operador aplica un modelo a un conjunto de datos. Los modelos contienen información sobre los datos con los han sido entrenados.

- **Performance:** Este operador permite obtener los resultados como es la matriz de confusión donde indica las instancias clasificadas correctamente y las que no se clasificaron correctamente, entre otros.

- **X-Validation:** Permite realizar la validación cruzada, donde se establece el número de validaciones, que es el número de subconjuntos que se crean para evaluar el modelo.

Anexo 8: Rendimiento Académico Aplicando Técnicas de Minería de Datos

El presente estudio se realizó tomando en cuenta los factores personales, académicos e institucionales de los estudiantes, correspondientes a los períodos 2010-2013 de la Universidad Nacional de Loja del Área de la Energía, las Industrias y los Recursos Naturales No Renovables.

Los factores tomados en cuenta para este estudio son los siguientes (ver TABLA LXVI):

TABLA LXVI.
ESTRUCTURA DE DATOS PARA DETERMINAR EL RENDIMIENTO ACADÉMICO

Atributo	Tipo de Datos	Tipo de Contenido	Valores
numeroIdentificacion	Nominal	Continuo	
nota_promedio	Real	Continuo	• malo • bueno • sobresaliente
edad	Nominal	Continuo	• a • b • c
genero	Nominal	Discreto	• 0 • 1
estado_matricula	Nominal	Discreto	• 0 • 1
promedio_asistencia	Nominal	Continuo	• b • m • a
nombre_carrera	Nominal	Discreto	• a • b • c • d
tipo _beca	Nominal	Discreto	• A • B • C • D • N

estado_civil	Nominal	Discreto	• a • b • c • d • e
modulo	Nominal	Discreto	▪ a ▪ b ▪ c

La TABLA LXVII nos proporciona los resultados obtenidos por cada algoritmo donde se puede observar que el algoritmo ID3 proporciona mejor resultados en el conjunto de entrenamiento con un total de 93.00% de instancias clasificadas correctamente y 7.00% clasificadas incorrectamente, además el índice de kappa es alto con el 0.757, que significa que la coincidencia de la predicción con la clase real está muy buen ajuste, así mismo un error cuadrático de 0.057 y un error relativo de 11.18%; y en la validación cruzada el algoritmo C4.5 arroja los mejores resultados con un total de 90.78% de instancias clasificadas correctamente 9.22% clasificadas incorrectamente, índice de kappa 0.679 que significa que la coincidencia de la predicción con la clase real es considerada buen ajuste, así mismo un error cuadrático de 0.080 y error relativo de 14.25%.

TABLA LXVII.

RESULTADOS DE LOS ALGORITMOS AGRUPADOS

Algoritmo	Conjunto de Datos	*Instancias correctamente clasificadas (%)*	*Instancias incorrectamente clasificadas (%)*	*Índice de Kappa*	Error Cuadrático	Error Relativo (%)	Error Cuadrático Medio	Error Cuadrático Relativo
ID3	C E	93.00	7.00	0.757	0.057	11.18	0.239	2.356
	C P	89.11	10.89	0.635	0.087	13.46	0.294	2.296
C4.5	C E	92.57	7.43	0.740	0.064	12.58	0.254	2.505
	C P	90.78	9.22	0.679	0.080	14.25	0.282	2.896
JRIP	C E	92.34	7.66	0.731	0.072	13.93	0.269	2.65
	C P	90.65	9.35	0.679	0.084	15.42	0.289	2.95
PART	C E	92.67	7.33	0.745	0.059	11.63	0.243	2.403
	C P	89.78	10.22	0.655	0.089	14.16	0.297	3.049
RIDOR	C E	90.17	9.83	0.610	0.098	9.83	0.314	3.358
	C P	89.05	10.95	0.619	0.109	10.95	0.329	3.443

En la siguiente figura se indican los resultados de las instancias clasificadas correctamente e incorrectamente luego de evaluar los diferentes algoritmos, donde se puede evidenciar que el algoritmo C4.5 es el que tiene menor margen de error, de 9.22% (ver Figura 32):

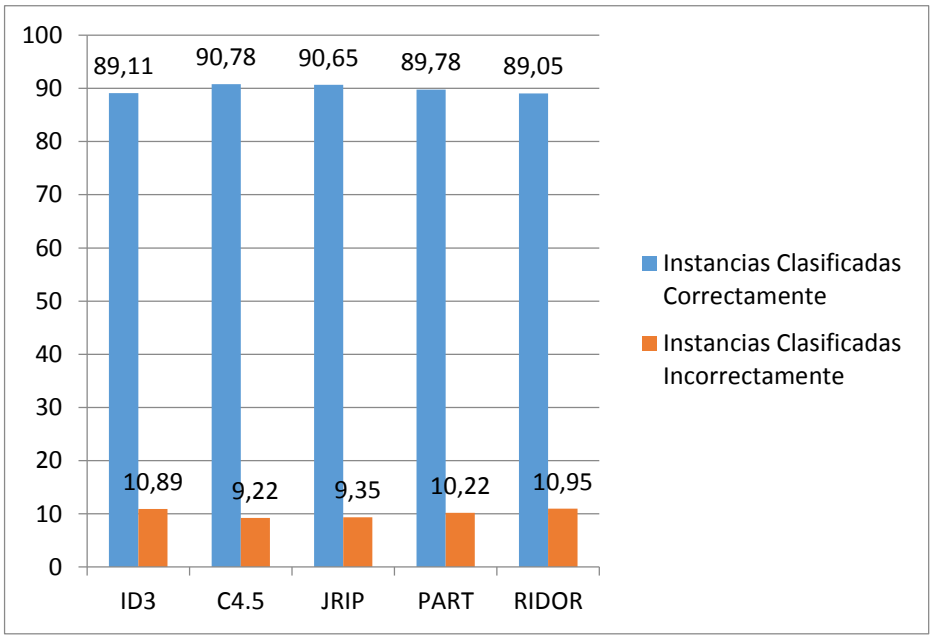


Figura 32: Instancias Clasificadas Correctamente e Incorrectamente

Así mismo se obtuvo el rendimiento académico del AEIRNNR donde este algoritmo obtuvo 389 malo, 2421 bueno y 5 sobresaliente durante la fase de entrenamiento; y 177 malo, 1178 bueno y 4 sobresaliente durante la fase de validación. Estos resultados se obtuvieron de acuerdo a los atributos que fueron tomados en cuenta (ver TABLA LXVI) para generar el modelo mediante el algoritmo C4.5. Mediante estos resultados se puede determinar que el rendimiento académico es bueno, con promedios de notas de 7.00 – 9.00 (ver Figura 33):

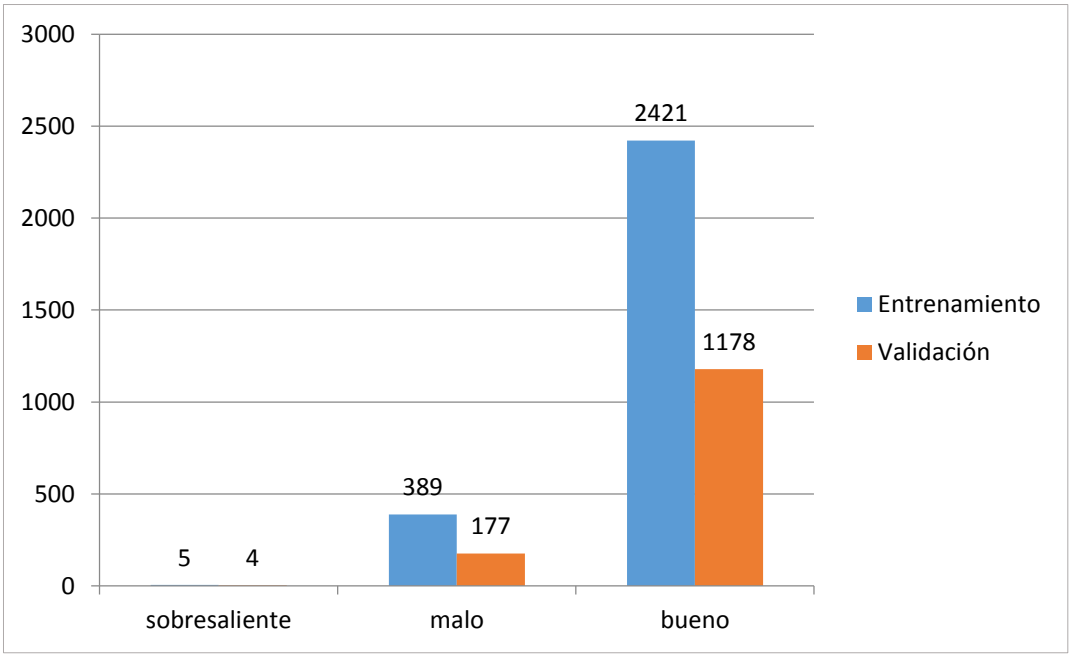


Figura 33: Rendimiento Académico de los estudiantes

Basado en las reglas, el rendimiento académico de los estudiantes es sobresaliente cuando:

- El estudiante tiene un promedio de asistencia mayor a 90% y tiene un tipo de beca A o D y pertenece a la carrera de Ingeniería en Electromecánica o Ingeniería en Sistemas.
- El estudiante pertenece a los módulos de primer a sexto y tienen entre 16 a 28 años y es soltero.

Basado en las reglas, el rendimiento académico de los estudiantes es bueno cuando:

- El estudiante tiene un tipo de beca D, pertenece a los módulos de primero a tercero, tiene entre 16 y 21 años y el promedio de asistencia es alto, es decir entre 90% y 100%.
- El estudiante no tiene un tipo de beca, pertenece a los módulos de primero a tercero, es soltero tiene entre 16 y 21 años.
- El estudiante tiene un tipo de beca A, pertenece a los módulos de primero a tercero, el promedio de asistencia es alto, es decir entre 90% y 100% y tiene entre 16 y 21 años.

- El estudiante tiene un pro medio de asistencia es alto, es decir entre 90% y 100% y pertenece a los módulos de cuarto a décimo.

Basado en las reglas, el rendimiento académico de los estudiantes es malo o bajo cuando:

- El estudiante que no posee un tipo de beca, el promedio de asistencia es bajo, es decir menor a 80%, el género es masculino o femenino y tienen de 16 a 19 años.
- El estudiante que no posee un tipo de beca, el promedio de asistencia es bajo, es decir menor a 80%, pertenece a los módulos de primero a tercero y tienen de 16 a 19 años.
- El estudiante que es divorciado y tiene entre 21 y 23 años, el promedio de asistencia es bajo, es decir menor a 80%, pertenece a los módulos de cuarto a sexto.

Factores de Rendimiento Académico

El Rendimiento Académico es un indicador del nivel de aprendizaje alcanzado por el estudiante y no siempre es lineal, es decir está relacionada por factores [66].

En el presente Trabajo de Titulación el objetivo es determinar el Rendimiento Académico aplicando Técnicas de Minería de Datos, para el cual se determinaron atributos que permitan obtener este objetivo, los mismos que se encuentran asociados entre sí, los cuales son: datos académicos, personales, institucionales y socioeconómicos del estudiante los cuales se detalla a continuación:

- ❖ Académicos: modulo, promedio_asistencia, estado_matricula.
- ❖ Personales: estado_civil, género, edad.
- ❖ Institucionales: nombre_carrera, tipo_beca.

Luego de clasificar los parámetros y mediante el algoritmo C4.5 se obtuvieron pesos en cada una de ellos que permitieron determinar cuáles de estos inciden con mayor relevancia en el Rendimiento Académico (ver TABLA LXVIII).

TABLA LXVIII.

PESO DE LOS ATRIBUTOS

Factor Influyente	Atributo	Peso	Total	% por Factor
Académico	modulo	0.080	2.003	95,56
	promedio_asistencia	0.923		
	estado_matricula	1.0		
Personal	estado_civil	0.012	0.055	2,62
	genero	0.017		
	edad	0.026		
Institucional	tipo_beca	0.014	0.038	1,82
	nombre_carrera	0.024		

De acuerdo a estos resultados obtenidos mediante el algoritmo C4.5 se puede determinar que el factor que más incide en el Rendimiento Académico es el factor Académico, tomando en consideración que el factor Institucional y Personal también afecta, pero no igual que el factor Académico (ver Figura 34).

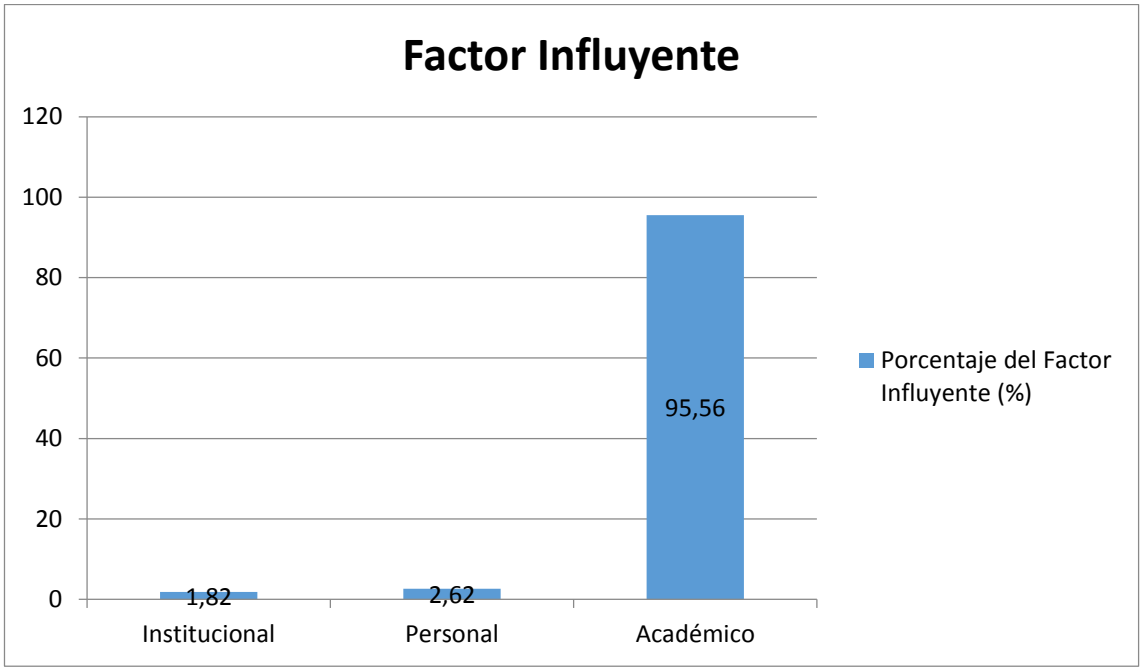


Figura 34. Factor Influyente del Rendimiento Académico.

Anexo 9: Artículo Científico

ESTUDIO DEL RENDIMIENTO ACADÉMICO APLICANDO TÉCNICAS DE MINERÍA DE DATOS

Study of Academic Performance Applying Data Mining Techniques

Darwin Becerra Encarnación[1], Henry Paz Arias[2]

Universidad Nacional de Loja[1], Ecuador[1], dabecerrae@unl.edu.ec[1]

Universidad Nacional de Loja[2], Ecuador[2], hpaz@unl.edu.ec[2]

RESUMEN

Durante los últimos años las universidades han adquirido un gran interés sobre el Rendimiento Académico de los estudiantes y en determinar qué factores influyen, pretendiendo de esta manera evidenciar cuáles son los que más afectan al rendimiento académico, para ello se analizó los datos académicos, personales, institucionales y socioeconómicos correspondientes a los períodos 2010-2013 de la Universidad Nacional de Loja del Área de la Energía, las Industrias y los Recursos Naturales No Renovables. Posteriormente se realizó un estudio de las metodologías de Minería de Datos, seleccionando CRISP-DM, debido a que es fundamental en este estudio, porque contiene una guía para el desarrollo del proyecto. Así mismo en la aplicación de técnicas de minería de datos se optó por la Clasificación, que contiene algoritmos como RIDOR, ID3, C4.5, JRIP y PART que fueron utilizados en la herramienta RAPIDMINER que permitió llevar los procesos y obtener resultados del modelo, los mismos que se analizó y evaluó a través de datos reales para determinar el Rendimiento Académico y mediante el cual se permitirá mejorar el proceso de formación académica.

Palabras Clave:

Rendimiento Académico, Minería de Datos, Técnica, Modelo, Algoritmo.

ABSTRACT

During last years, universities have acquired a great interest on the Academic Performance of students and want to determine what factors influence in it, trying to demonstrate which of them are the most affecting academic performance, to do that personal and institutional data was analyzed from 2010 to 2013 of Loja National University into Energy Area, Industries and Non-Renewable Natural Resources. After that, a study of data mining methodologies, selecting CRISP-DM, it is essential in this study because it contains a guide to develop the project. Also in the application of data mining techniques we chose the Classification that contains RIDOR, ID3, C4.5, PART and JRIP which were used in the RapidMiner tool, it allowed to develop the processes and get results from the model, finally they were analyzed and assessed using real data of Educational Performance and factors that influence in it.

KeyWords:

Academic Performance, Data Mining, Technical, Model, Algorithm.

INTRODUCCIÓN

Actualmente el Rendimiento Académico de los estudiantes ha adquirido un gran interés en todas las universidades, una de ellas es la Universidad Nacional de Loja. Con la información que los estudiantes proporcionan a los sistemas informáticos de la Universidad, se podrá obtener un modelo sobre el rendimiento académico de los estudiantes y además factores que determinen el mismo.

Es por ello necesario contar con métodos eficientes y automáticos para explorar las grandes Bases de Datos, procesando de forma rápida y fiable la información para encontrar patrones de conocimiento para resolver un problema [1]. Para esto la Minería de Datos permite examinar grandes cantidades de datos para descubrir factores que permitan determinar el rendimiento académico, cuyo objetivo es encontrar modelos inteligibles a partir de los datos, descubrir patrones cuya utilización apoye decisiones que reporten beneficios a la organización [2].

En los últimos años han surgido diversas herramientas de MD, como RapidMiner, Weka, Oracle Data Mining, entre otras, basadas en técnicas que facilitan el procesamiento de datos y permiten realizar un análisis de los mismos, con el objetivo de determinar el rendimiento académico [3]. Así mismo se utilizó la metodología CRISP-DM como una guía para desarrollan el presente artículo.

En el presente artículo el objetivo es determinar el rendimiento académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos, donde se propone la utilización de estas técnicas, para detectar cuáles son los factores (académicos, personales, institucionales y socioeconómicos), que permitan determinar el rendimiento académico de los estudiantes e identificar cuáles son los factores que más influyen.

El artículo está estructurado de la siguiente manera: Resumen: contiene la síntesis de la temática del artículo. Introducción: contenido del artículo. Materiales y Métodos: se detalla las técnicas, métodos y la metodología utilizada. Estado del Arte: Minería de Datos, Técnicas de Minería de Datos, Herramienta de Minería de Datos y Metodología de Minería de Datos. Diseño e Implementación: contiene la explicación resumida de todo el trabajo técnico realizado. Resultados: resultados alcanzados. Trabajos Relacionados: aquellos trabajos que se relacionan con el presente artículo. Conclusiones y Trabajos Futuros: Conclusiones que se obtienen a partir del trabajo realizado y trabajos futuros que podrían realizarse a partir de los resultados alcanzados. Agradecimientos. Referencias. Bibliográficas: permitió extraer información relacionado con el presente trabajo.

MATERIALES Y MÉTODOS

En el presente artículo los métodos utilizados están basados en la investigación bibliográfica porque se realizó un análisis de los problemas que afectan el rendimiento académico basándose en fuentes bibliográficas confiables y casos de éxito. Además es un proyecto de desarrollo porque se implementó un modelo que permita estimar en rendimiento académico de los estudiantes.

Para la recolección de la información se utilizó el método científico para la formulación del marco teórico, en donde se indicó temas como: casos de éxito, rendimiento académico, técnicas de minería de datos, herramientas de minería de datos que estén enfocados en el proceso del presente proyecto. También el método deductivo para ayudar a conocer los problemas de las universidades sobre el rendimiento académico de los estudiantes y a través de esto las mismas podrá tomar decisiones que permitan mejoras. Así mimo el método inductivo se lo utilizó para obtener información académica de cada uno de los estudiantes y hacer un análisis de cada uno de los inconvenientes que tienen para poder

obtener el problema general de estudio, y así enfocarnos directamente en resolver dicho problema.

Además se utilizó técnicas de minería de datos para recopilar información como: bibliográfica para la revisión de diferentes fuentes de información confiables enfocados en el tema con el fin de detectar problemas y causas que afectan el rendimiento académico. También la técnica de observación permitió observar la realidad académica de los estudiantes, así como también permitió seguir obteniendo información necesaria a lo largo del desarrollo del presente proyecto.

También se empleó la metodología donde se determinó una secuencia de pasos ordenados que nos permitan cumplir los objetivos; la misma que es Cross-Industry standard Process for Data Mining, CRISP – DM que es una guía para el desarrollo de proyectos enfocados a la minería de datos.

ESTADO DEL ARTE

3.1. Minería de Datos

La Minería De Datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la Minería de Datos es encontrar modelos a partir de los datos [1].

3.2. Técnicas de Minería de Datos

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos con vistas al descubrimiento de información previamente desconocida y que pueda servir de ayuda en el proceso de toma de decisiones [2].

Las técnicas más representativas son:

a. Técnicas No Supervisadas o de Descubrimiento del Conocimiento

- Clustering o Agrupamiento.- Es el proceso de agrupar los datos en clases o en clústeres, de tal forma que los datos de un mismo clúster tienen una alta similitud y a su vez, son muy diferentes de los de otro clúster [3].
- Reglas de Asociación.- Es la exploración de los datos con el propósito de identificar relaciones entre los datos, dentro de una fuente o base de datos [4]. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos [5].

b. Técnicas Supervisadas o Predictivas

- Árboles de Decisión.- Son ampliamente usados y pueden ser fácilmente explicados basándose en el criterio usado para dividir los datos en las extremidades del árbol. Los árboles de decisión son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos [6].
- Clasificación.- Técnica que permite encontrar modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones. Además empareja o asocia datos a grupos predefinidos [7].
- Redes Neuronales.- Las redes neuronales simulan el cerebro humano mediante el aprendizaje de un conjunto de datos de formación y la aplicación del aprendizaje para generalizar los patrones para la clasificación y predicción [6].
- Lógica Borrosa.- Surge de la necesidad de modelar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud, es decir permite el tratamiento probabilístico de la categorización de un colectivo [5].

3.3. Herramienta de Minería de Datos

La herramienta de MD sirvió para extraer conocimientos desde base de datos que contienen grandes cantidades de información.

RapidMiner es un entorno de código abierto para aprendizaje automático y minería de datos. Permite realizar todos los procesos que intervienen en un proyecto: la

adquisición de datos, la transformación de los datos, la selección de datos, la selección de atributos, la transformación de los atributos, el aprendizaje/modelización y la validación [8]. Además permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico [9].

3.4. Metodología de Minería de Datos

CRISP-DM (Cross-Industry Standard Process for Data Mining) es una guía para el desarrollo de proyectos enfocados a la Minería de Datos. Esta metodología puede trabajar con cualquier herramienta para desarrollar el proyecto, es decir es una metodología equitativa [10], [11].

Fases de CRISP-DM [12-14]:

1. Entendimiento del negocio.- Esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio.
2. Entendimiento de los datos.- Esta fase inicia con una colección inicial de datos y procede con actividades para familiarizarse con ellos, identificar problemas de calidad en los mismos, descubrir una primera idea de estos o detectar conjuntos interesantes que permitan formar hipótesis en la búsqueda de información escondida.
3. Preparación de los datos.- Cubre todas las actividades para construir la base final de datos.
4. Modelado.- Se seleccionan y aplican varias técnicas, y sus parámetros son calculados a los valores óptimos.
5. Evaluación.- Al llegar a esta fase se ha construido un modelo (o modelos) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la entrega final del modelo es importante evaluarlo y revisar los pasos ejecutados para construirlo, de tal forma que este lo más cercano posible de alcanzar los objetivos del negocio.
6. Despliegue.- La creación del modelo por lo general no es el final del proyecto. A menudo implica aplicar modelos en vivo dentro del proceso de toma de decisiones de una organización.

DISEÑO E IMPLEMENTACIÓN

El presente artículo describe la aplicación de Técnicas de Minería de Datos para determinar el Rendimiento Académico de los estudiantes, para lo cual fue necesario aplicarlo en un escenario real con datos académicos, institucionales y personales de los estudiantes de la UNL del Área de la Energía, las Industrias y los Recursos Naturales No Renovables. Para ello es necesaria la utilización de la metodología CRISP-DM como una guía que permita desarrollar el proyecto.

4.1. Compresión del Negocio

La idea principal es Determinar el Rendimiento Académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos. La determinación del rendimiento académico es fundamental para poder obtener resultados confiables sobre el desempeño académico de los estudiantes universitarios. La UNL cuenta con valiosa información académica de los estudiantes, cuyo problema radica que dicha información no es analizada para determinar el desempeño académico de cada estudiante.

A partir de esto es necesario aplicar Técnicas de Minería de Datos que permitan realizar un estudio con la finalidad de evaluar el desempeño académico de los estudiantes.

Los objetivos del negocio son:

- ✓ Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.
- ✓ Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.
- ✓ Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.

4.2. Comprensión de los Datos

Se realizó una recolección inicial de los datos relacionados con el problema,

también se realizó un análisis de los mismos con el fin de identificar las relaciones entre ellos.

Los datos académicos obtenidos corresponden a estudiantes del Área de la Energía, Las Industrias y los Recursos Naturales No Renovables (AEIRNNR) de la UNL, de los periodos académicos 2008 hasta el 2013, además existe información personal, institucional de los estudiantes.

Los datos académicos se obtuvieron del Web Service de la UNL, estos datos fueron almacenados en una Base de Datos y que es administrada mediante MySQL, la misma que consta de 19 tablas.

Esta BD contiene información:

- Académica
- Personal
- Institucional
- Socioeconómicos

En la siguiente figura (ver Fig. 1) se puede observar las tablas que conforman la Base de Datos:

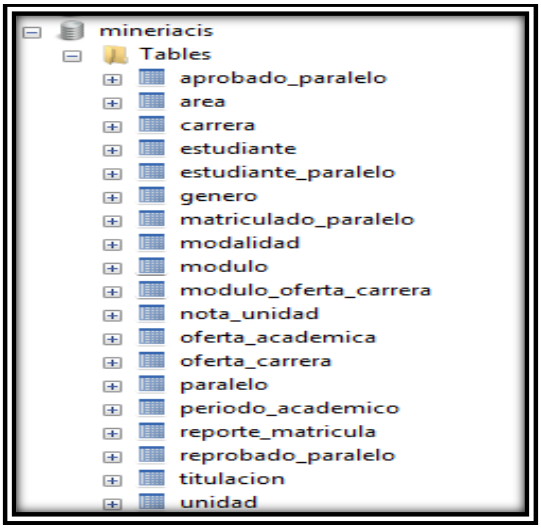


Fig. 1. Base de datos

4.3. Preparación de los Datos

En la presente fase se construirá la estructura de datos final, en la cual se procederá a ingresar los datos íntegros, sin errores, ni faltantes. Además en esta etapa se transformó los datos para que puedan ser usados de manera eficiente por técnicas de minería de datos.

Así mismo se procedió a realizar la construcción de los datos que permitirán construir el modelo computacional, los atributos que se han considerado para establecer en la data set definitivo, para el presente estudio, son los siguientes (ver Tabla I):

TABLA I. Estructura de Datos para determinar el Rendimiento Académico.

Atributo	Valores	Descripción
numeroldentificacion		
nota_promedio	• malo • bueno • sobresaliente	• Promedio de nota bajo • Promedio de nota bueno • Promedio de nota sobresaliente
edad	• a • b • c	• 16-19 años (a) • 19-21 años (b) • 21-30 años (c)
genero	• 0 • 1	• Masculino (0) • Femenino (1)

estado_matricula	• 0 • 1	• Aprobado (0) • Reprobado (1)
promedio_asistencia	• b • m • a	• Baja (b) • Media (m) • Alta (a)
nombre_carrera	• a • b • c • d	• Ingeniería en Sistemas (a) • Ingeniería en Electromecánica (b) • Ingeniería en Electrónica y Telecomunicaciones (c) • Ingeniería en Geología Ambiental y Ordenamiento Territorial (d)
tipo_beca	• A • B • C • D • N	• Tipo de Beca A • Tipo de Beca B • Tipo de Beca C • Tipo de Beca D • No tiene ningún Tipo de Beca N
estado_civil	• a • b • c • d • e	• Soltero (a) • Casado (b) • Divorciado (c) • Unión Libre (d) • Viudo (e)
modulo	▪ a ▪ b ▪ c	▪ Módulo 1-3 (a) ▪ Módulo 4-6 (b) ▪ Módulo 6-10 (c)
origen_estudiante	• u • r • p • c	• Sector Urbano (u) • Sector Rural (e) • Pertenece a otra provincia (p) • Pertenece a otro cantón de Loja (c)

numero_hijos_estudiante	• S • N	• Si tiene hijos • No tiene hijos
etnia_estudiante	• m • b • i	• Mestizo • Blanco • Indígena
situación_laboral_estudiante	• S • N	• Si tiene hijos • No tiene hijos
situación_laboral_madre	• S • N	• Si trabaja • No trabaja
situación_laboral_padre	• S • N	• Si trabaja • No trabaja
horario	• m • v	• Matutino • Vespertino
tipo_colegio_estudiante	• 0 • 1	• Público • Privado

4.4. Modelado

En esta fase se utilizará el conjunto de datos establecidos para procesarlos a través de una Herramienta de Minería de Datos que permita implementar la o las técnicas necesarias para la construcción del modelo.

Es por ello que se utilizó la técnica de Clasificación, donde se evaluaron los algoritmos, los cuales fueron: ID3, C4.5, JRIP, RIDOR y PART, con el fin de construir el modelo que permita determinar el rendimiento académico.

Además para evaluar los modelos generados es necesario crear un mecanismo para probar la calidad y validez del mismo. Por ende, se trabajará con un conjunto de datos para entrenamiento (CE) y también para pruebas (CP) y validación del modelo que permita determinar el rendimiento académico. Diseño de pruebas a realizar en los diferentes algoritmos:

Algoritmos: ID3, C4.5, RIDOR, PART, JRIP

El diseño de pruebas que se realizó con estos algoritmos fueron el conjunto de datos de entrenamiento, para el cual se utilizó un 67% del total de datos, además el otro 33% permitió evaluar el modelo creado mediante la validación cruzada.

Luego de seleccionar los algoritmos de minería de datos es necesario contar con una Herramienta de Minería de Datos, para lo cual se utilizó RAPIDMIER, debido a que posee gran cantidad de operadores que permiten generar de mejor manera un modelo que permita determinar el rendimiento académico, además posee una licencia libre, es multiplataforma, posee una interfaz amiga y permite aplicar técnicas de minería de datos tanto descriptivas y predictivas.

Los resultados obtenidos con cada uno de los algoritmos son (ver Tabla II y Tabla III):

Tabla II: Resultados de Algoritmos

A	CD	*ICC %*	*ICI %*	*IK*
ID3	C E	94.84	5.16	0.82
	C P	85.04	14.96	0.52
C4.5	C E	92.21	7.79	0.72
	C P	90.98	9.02	0.65

JRIP	C E	92.27	7.73	0.72
	C P	89.78	10.22	0.62
PART	C E	92.83	7.17	0.75
	C P	89.98	10.02	0.62
RIDOR	C E	90.66	9.34	0.69
	C P	90.11	9.89	0.63

Tabla III: Resultados de Algoritmos

A	CD	EC	ER (%)	ECM	ECR
ID3	C E	0.039	7.55	0.19	1.94
	C P	0.097	12.7	0.31	4.16
C4.5	C E	0.070	13.7	0.26	2.61
	C P	0.078	14.9	0.27	3.33
JRIP	C E	0.073	14.0	0.27	2.66
	C P	0.091	16.5	0.30	3.59
PART	C E	0.062	12.2	0.25	2.46
	C P	0.083	14.8	0.28	3.44
RIDOR	C E	0.093	9.34	0.30	3.01
	C P	0.099	9.89	0.31	3.75

En la siguiente tabla se indica el significado de las siglas de la Tabla II y III (ver Tabla IV):

Tabla IV: Significado de Siglas

A: Algoritmo	**CP: Conjunto de Pruebas**	**IK: Índice de Kappa**
CD: Conjunto de Datos	ICC: Instancias Clasificadas Correctamente	EC: Error Cuadrático
CE: Conjunto de Entrenamiento	ICI: Instancias Clasificadas Incorrectamente	ER: Error Relativo
ECM: Error Cuadrático Medio	ECR: Error Cuadrático Relativo	

La TABLA II y III nos proporciona los resultados obtenidos por cada algoritmo donde se puede observar que el algoritmo ID3 proporciona mejor resultados en el conjunto de entrenamiento con un total de 93.00% de instancias clasificadas correctamente y 7.00% clasificadas incorrectamente, además el índice de kappa es alto con el 0.76, que significa que la coincidencia de la predicción con la clase real está muy buen ajuste, así mismo un error cuadrático de 0.06 y un error relativo de 11.18%; y en la validación cruzada el algoritmo C4.5 arroja los mejores resultados con un total de 90.78% de instancias clasificadas correctamente 9.22% clasificadas incorrectamente, un índice de kappa 0.68 que significa que la coincidencia de la predicción con la clase real es considerada muy buen ajuste, así mismo un error cuadrático de 0.08 y error relativo de 14.25%.

4.5. Evaluación

En esta fase se evalúa el modelo, con el fin de comprobar que los resultados cumplen los objetivos, además se obtuvo el modelo que permitió determinar el rendimiento académico y los factores que se utilizaron para construir el modelo y cuál de estos es el que más influye en el rendimiento académico.

Así mismo se obtuvo el rendimiento académico del AEIRNNR donde este algoritmo obtuvo 380 malo, 2424 bueno y 4 sobresaliente durante la fase de entrenamiento; y 153 malo, 1209 bueno y 12 sobresaliente durante la fase de validación. Estos resultados se obtuvieron de acuerdo a los atributos que fueron tomados en cuenta (ver Tabla I) para generar el modelo mediante el algoritmo C4.5.

5. RESULTADOS

En la siguiente figura se indican los resultados de las instancias clasificadas correctamente e incorrectamente luego de evaluar los diferentes algoritmos, donde se puede evidenciar que el algoritmo C4.5 es el que tiene menor margen de error, el cual es

9.02%. En la siguiente figura se detallan cada uno de los algoritmos (ver Fig. 2):

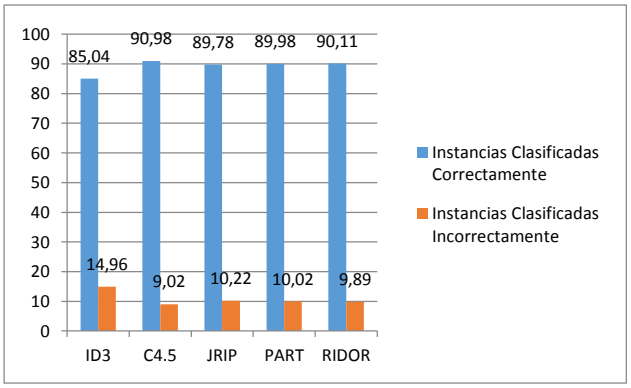


Fig. 2: Instancias Clasificadas Correctamente e Incorrectamente

El modelo obtenido a partir de los diferentes algoritmos antes mencionados permitió determinar el rendimiento académico de los estudiantes, a continuación se presenta el mismo (ver Fig. 3):

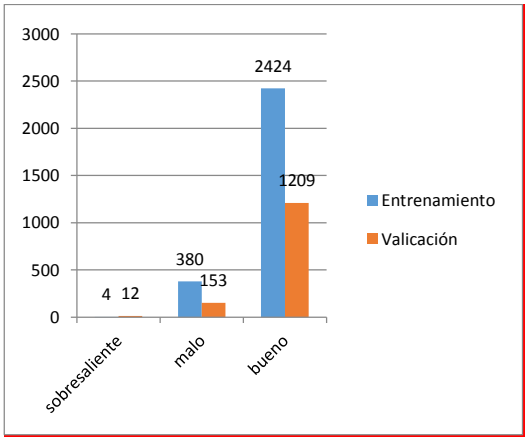


Fig. 3: Rendimiento Académico de los estudiantes

Para verificar lo mencionado, fue necesaria la utilización de los datos académicos, personales e institucionales de los estudiantes del AEIRNNR, los cuales se detalla a continuación:

- Académicos: modulo, promedio_asistencia, estado_matricula, tipo_colegio_estudiante.
- Personales: estado_civil, género, edad, etnia_estudiante, numero_hijos_estudiante, procedencia_estudiante.
- Institucionales: nombre_carrera, tipo_beca, horario.
- Socioeconómicos: situacion_laboral_estudiante, situacion_laboral_padre, situacion_laboral_madre.

En la siguiente tabla se detallan los porcentajes de cada factor del rendimiento académico (ver Tabla V):

Factor Influyente	Atributo	% por Factor
Académico	modulo	26,03
	promedio_asistencia	
	estado_matricula	
	tipo_colegio_estudiante	
Personal	estado_civil	38,72
	genero	
	edad	
	etnia_estudiante	
	numero_hijos_estudiante	
	procedencia_estudiante	
	tipo_beca	17,20

Institucional	nombre_carrera	
	horario	
Socioeconómicos	situacion_laboral_estudiante	18,05
	situacion_laboral_padre	
	situacion_laboral_madre	

Tabla V: Porcentaje de los Factores del Rendimiento Académico

De acuerdo a estos resultados obtenidos mediante el algoritmo C4.5 se puede determinar que el factor que más incide en el Rendimiento Académico es el factor Personal (ver Fig. 4).

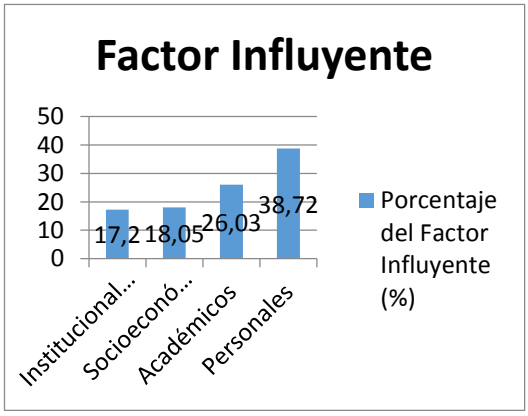


Fig. 4. Factor Influyente del Rendimiento Académico.

Al tener conocimiento del rendimiento académico de esta área, las autoridades podrán tomar medidas y emprender un plan que ayude a mejorar la formación académica de los estudiantes.

6. TRABAJOS RELACIONADOS

- ✓ Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos [15].
- ✓ Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil [16].
- ✓ Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios [17].

7. CONCLUSIONES Y TRABAJO FUTURO

Luego de finalizar el Trabajo de Titulación se obtuvieron las siguientes conclusiones:

- ✓ La utilización de minería de datos permitió realizar una comprensión de los datos y descubrir factores que influyen en el rendimiento académico de los estudiantes.
- ✓ El empleo de técnicas de minería de datos ayudan a obtener modelos computacionales, ya que contiene diferentes algoritmos que ayudan a manipular los datos para obtener resultados que permita determinar el rendimiento académico. Es por ello, que mediante la técnica de Clasificación a través de los diferentes tipos de algoritmos evaluados, C4.5 permitió obtener los mejores resultados de acuerdo a los datos académicos, institucionales y personales de los estudiantes.
- ✓ La herramienta de minería de datos RAPIDMINER fue fundamental en el desarrollo de este proyecto, ya que contiene gran número de algoritmos y complementos que permite hacer uso de diferentes algoritmos de otras herramientas, además tiene operadores que ayudan a facilitar el desarrollo de los procesos para crear los modelos aplicables para el análisis de los datos.
- ✓ Mediante el modelo de minería de datos obtenido y evaluado a través de datos reales se pudo comprobar que el rendimiento académico de los estudiantes está considerado bueno, con promedio de notas entre 7.00 - 9.00.

Los trabajos futuros que se pueden llevar a cabo son:

- ✓ Incorporar nuevos datos al modelo, identificando para ello nuevos factores

los mismos que servirán para determinar el rendimiento académico de los estudiantes.

- ✓ Aplicar nuevas técnicas de minería de datos como reglas de asociación, agrupamiento o clustering con el fin de obtener diferentes resultados a los que se obtuvieron.

8. AGRADECIMIENTOS

El presente Trabajo de Titulación primeramente agradecer a Dios por bendecirme para poder llegar a culminar esta meta.

También a la Universidad Nacional de Loja por permitir realizar mis estudios en la carrera de Ingeniería en Sistemas. Además agradezco a los docentes que aportaron en mi formación profesional y a mi Director del Trabajo de Titulación quien con sus conocimientos, su experiencia y su motivación ha logrado en mí que pueda terminar mi carrera con éxito.

A mis padres por ser el pilar fundamental en mi vida, por sus valores inculcados y el amor que me brindaron día a día, los mismos que permitieron que llegue a esta meta y por ser un ejemplo de vida a seguir. A mis hermanos por ser parte importante en mi vida y porque siempre me han incentivado a seguir adelante.

9. REFERENCIAS BIBLIOGRÁFICAS

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

[2]. Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda, Minería de Datos, Universidad Carlos III de Madrid – Ingeniería Telemática, [En línea]. Disponible en: http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/07-mineria-de-datos

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

[4]. Corso Cynthia Lorena, Gibellini Fabián, Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios, Universidad Tecnológica Nacional - Departamento de Ingeniería en Sistemas de Información - Laboratorio de Sistemas de Información, [En línea]. Disponible en: http://www.41jaiio.org.ar/sites/default/files/16_JSL_2012.pdf

[5]. José Manuel Molina López y Jesús García Herrero, Técnicas de Análisis de Datos, Instituto Tecnológico Superior de Calkiní en el Estado de Campeche, [En línea]. Disponible en: http://www.itescam.edu.mx/principal/sylabus/fpdb/recursos/r94663.PDF

[6]. José Antonio García Bermúdez y Ángela María Acevedo Ramírez, Análisis para Predicción de Ventas Utilizando Minería de Datos en Almacenes de Ventas de Grandes Superficies, Universidad Tecnológica de Pereira - Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación - Ingeniería de Sistemas y Computación, [En línea]. Disponible en: http://repositorio.utp.edu.co/dspace/bitstream/11059/1339/1/006312G216.pdf

[7]. Braulio José Solano Rojas, Introducción a la Minería de Datos, Universidad de Costa Rica.

[8]. Práctica de laboratorio de aprendizaje inductivo, Universidad Politécnica de Cataluña – Facultad de Informática, [En línea]. Disponible en: http://www.lsi.upc.edu/~bejar/apren/lab/apind13141g.pdf

[9]. Francisco José García González, Aplicación de Técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Universidad de Granada

[10]. Metodología de Aplicación del Data Mining (DM), Universidad Politécnica Salesiana, [En línea]. Disponible en: http://www.dspace.ups.edu.ec/bitstream/123456789/47/10/Capitulo4.pdf

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

[12]. Hernando Camargo y Mario Silva, Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP, Universidad el Bosque – Ingeniería en Sistemas, [En línea]. Disponible en: http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

[13]. Juan Ángel Vanrell, Un Modelo de Procesos para Proyectos de Explotación de Información, Universidad Tecnológica Nacional - Ingeniería en Sistemas de Información, [En línea]. Disponible en: http://www.unla.edu.ar/sistemas/gisi/tesis/vanrell-tesisdemagister.pdf

[14]. José Alberto Gallardo Arancibia, Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM), Universidad Politécnica de Madrid - Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software - Facultad de Informática, [En línea]. Disponible en: http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

[15]. Alcover R., Benlloch J., Blesa P., Calduch M., Celma M. Ferri M., Hernández Orallo M., Iniesta L., Más J., Ramírez Quintana M., Robles A., Valiente J., Vicent M., Zúnica L., "Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos, *Universidad Politécnica de Valencia - Dpto. de Estadística e I.O. Aplicadas y Calidad, Dpto. de Informática de Sistemas y Computadores, Dpto. de Sistemas Informáticos y Computación, Dpto. de Física Aplicada*, [En línea]: Disponible en: http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf

[16]. Spositto O., Etcheverry M., Hugo L. Ryckeboer, Julio Bossero, "Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil, *Universidad Nacional de La Matanza - Departamento de Ingeniería e Investigaciones Tecnológicas,* [En línea]. Disponible en: http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

[17]. Porcel E., Dapozo G., López M., "Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios, *Universidad Nacional del Nordeste - Departamento de Informática,* [En línea]. Disponible en: http://sedici.unlp.edu.ar/bitstream/handle/10915/19846/Documento_completo.pdf?sequence=1

Anexo 10: Informe Ejecutivo

a. RESUMEN

Durante los últimos años las universidades han adquirido un gran interés sobre el Rendimiento Académico de los estudiantes y en determinar sus factores que influyen, pretendiendo de esta manera evidenciar cuáles son los que más afectan al rendimiento académico de los estudiantes**.**

Es por ende que este estudio permitió determinar el Rendimiento Académico de los estudiantes del Área de la Energía las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja. Para ello se tomaron en cuenta factores académicos, institucionales, personales y socioeconómicos de los estudiantes de los periodos 2010-2013.

Así mismo los resultados que se obtuvieron permitieron establecer que los estudiantes de esta área se encuentran con un rendimiento académico bueno con promedios de 7.00-9.00 y también se determinaron los factores que más influyen en el rendimiento académico de los estudiantes, el cuál es el factor académico.

b. ANTECEDENTES

En el presente Trabajo de Titulación se analizará la información académica de los estudiantes para medir el rendimiento de los mismos, con el fin que las autoridades de la universidad tomen decisiones que permitan corregir las deficiencias encontradas.

En el ámbito académico permitirá adquirir capacidades y habilidades necesarias para poder llevar con éxito el desarrollo del proyecto, además se obtendrá conocimientos y experiencia que nos permitan resolver problemas que se presenten en la sociedad, así mismo permitirá en las universidades tener en cuenta los factores que influyen en el rendimiento académico de los estudiantes.

Con lo descrito anteriormente el proyecto es viable, por lo tanto se puede justificar que los resultados esperados fueron alcanzados concernientes al análisis del rendimiento académico.

c. OBJETIVO

- ✓ Determinar el Rendimiento Académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos.

d. RESULTADOS

Los resultados obtenidos luego de realizar un estudio a los datos académicos, personales, institucionales y socioeconómicos de los estudiantes de la Universidad Nacional de Loja del Área de la Energía, las Industrias y los Recursos Naturales No Renovables – AEIRNNR de los periodos 2010-2013 con el objetivo de determinar el Rendimiento Académico demuestran que esta área tiene un promedio de buena que corresponde a nota comprendidas en los rangos de 7.00-9.00.

El rendimiento académico del AEIRNNR está conforma de la siguiente manera luego de tomar una muestra de estudiantes de esta área: 533 bajo, 3633 bueno y 16 sobresaliente, los mismos que indican que el rendimiento académico que más sobresale es bueno con 3633 estudiantes. A continuación se presenta el Rendimiento Académico de los estudiantes (ver Figura 1):

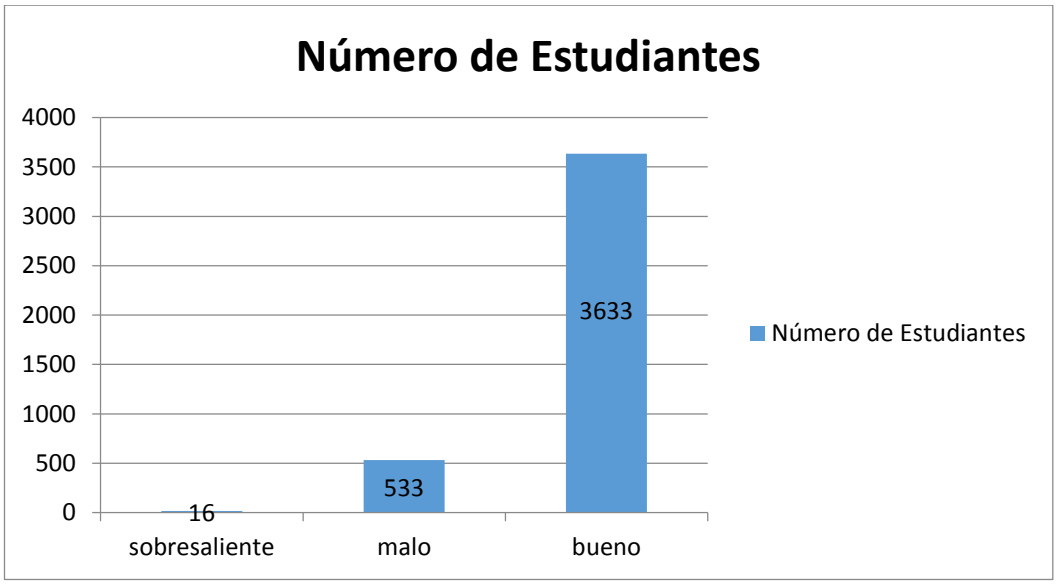


Figura 1: Rendimiento Académico de los estudiantes

Estos resultados se obtuvieron de acuerdo a los atributos que fueron tomados en cuenta, los cuales están categorizados de la siguiente manera: académicos, personales e institucionales del estudiante, los cuales se detalla a continuación: (ver Tabla 1):

Factor Influyente	Atributo	% por Factor
Académico	modulo	26,03
	promedio_asistencia	
	estado_matricula	
	tipo_colegio_estudiante	
Personal	estado_civil	38,72
	genero	
	edad	
	etnia_estudiante	
	numero_hijos_estudiante	
	procedencia_estudiante	
Institucional	tipo_beca	17,20
	nombre_carrera	
	horario	
Socioeconómicos	situacion_laboral_estudiante	18,05
	situacion_laboral_padre	
	situacion_laboral_madre	

Tabla 1: Porcentaje de Factores del Rendimiento Académico

De acuerdo a estos resultados obtenidos (ver Tabla 2) se puede determinar que el factor que más incide en el Rendimiento Académico es el factor Personal. A continuación se presenta el porcentaje de los factores utilizados para determinar el Rendimiento Académico de los estudiantes (ver Figura 2):

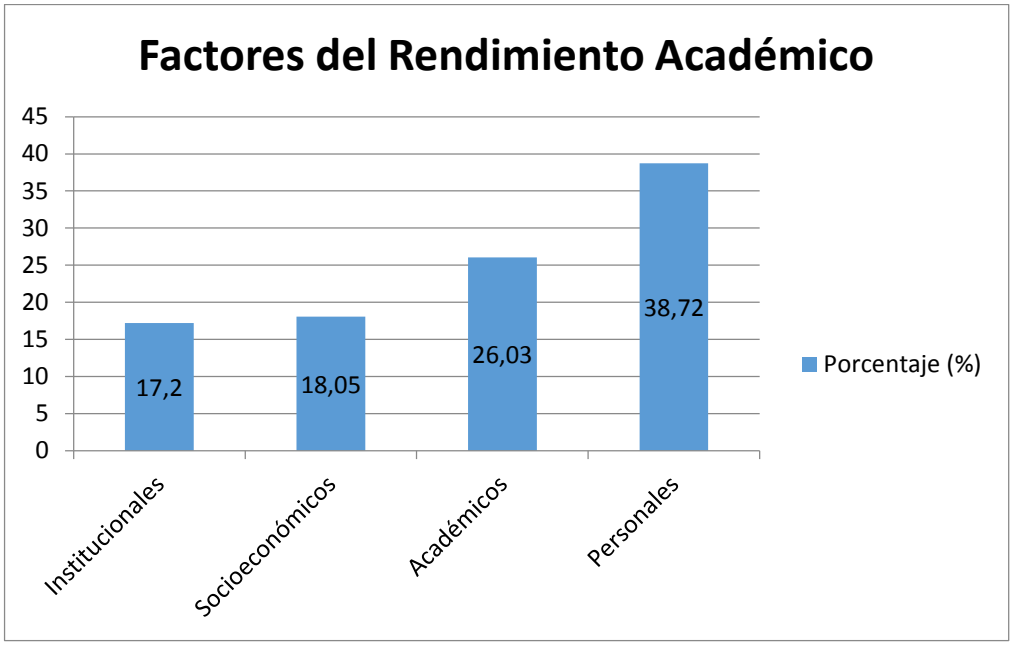


Figura 2: Factores del Rendimiento Académico

Basado en las reglas, el rendimiento académico de los estudiantes es sobresaliente cuando:

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y el género es femenino y el padre trabaja y el estudiante es soltero y la madre trabaja y el estudiante pertenece al sector urbano.
- Si el promedio de asistencia es mayor a 90% y tiene entre 20 y 21 años y pertenece a los módulos de primero a tercero.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de primero a tercero y realizó sus estudios secundarios en un colegio público y no tiene hijos y pertenece a la carrera Ingeniería Electromecánica y no trabaja y el género es masculino y el padre trabaja y el estudiante es soltero y la madre trabaja; entonces el rendimiento académico es sobresaliente.

Basado en las reglas, el rendimiento académico de los estudiantes es bueno cuando:

- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante no trabaja.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y no tiene hijos y está en los módulos de séptimo a décimos y realizó sus estudios secundarios en un colegio público y es mayor de 21 años y es mestizo y el promedio de asistencia a clases es mayor a 90% y pertenece a la carrera Ingeniería en Geología Ambiental y Ordenamiento Territorial y el género es masculino.
- Si posee un tipo de beca A y el horario de clases es matutino y el estudiante trabaja y tiene hijos.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y no trabaja.
- Si posee un tipo de beca A y el horario de clases es vespertino y pertenece a un cantón de la provincia de Loja y está en los módulos de cuarto a sexto y trabaja realizó sus estudios secundarios en un colegio público y es mayor a 21 años y es mestizo y el promedio de asistencia es mayor a 90% y no tiene hijos y pertenece a la carrera Ingeniería Electrónica y Telecomunicaciones y el género es masculino y es soltero y el padre trabaja y la madre trabaja.
- Si el promedio de asistencia es mayor a 90% y es menor a 20 años y pertenece a los módulos de cuarto a sexto.
- Si estado de matrícula es aprobado y horario de clases es matutino y es mestizo y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto.
- Si no posee un tipo de beca y es mestizo y es soltero y tiene entre 20 y 21 años y pertenece a otra provincia.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a los módulos de cuarto a sexto y el padre trabaja y es mayor a 21 años.
- Si el promedio de asistencia es mayor a 90% y pertenece a los módulos de cuarto a sexto y el género es femenino y pertenece al sector urbano.

Basado en las reglas, el rendimiento académico de los estudiantes es malo o bajo cuando:

- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de primero a tercero.

- Si el promedio de asistencia es menor a 80% y no posee un tipo de beca y pertenece a los módulos de cuarto a sexto y es mestizo y no tiene hijos y no trabaja y es masculino y es soltero y el padre trabaja y la madre trabaja y pertenece a un cantón de la provincia de Loja o a otra provincia.
- Si el promedio de asistencia es menor a 80% y es menor a 20 años.
- Si estado de matrícula es reprobado y es menor a 20 años y pertenece a los módulos de primero a tercero.
- Si estado de matrícula es reprobado y pertenece a otra provincia y es mayor a 21 años y el promedio de asistencia es menor a 80%.
- Si no posee un tipo de beca y es mestizo y es soltero y no tiene hijos y pertenece a la carrera de Ingeniería Electrónica y Telecomunicaciones y el padre trabaja.
- Si no posee un tipo de beca y es mestizo y es soltero y no trabaja y no tiene hijos y pertenece a los módulos de séptimo a décimo.

e. CONCLUSIONES

Luego de finalizar el Trabajo de Titulación se obtuvieron las siguientes conclusiones:

- ✓ Los resultados obtenidos permitieron realizar una comprensión de los datos y descubrir factores que influyen en el rendimiento académico de los estudiantes.

- ✓ También se obtuvo el factor que más influye en el rendimiento académico de acuerdo a datos académicos, institucionales y personales de los estudiantes, resultando los datos académicos como el factor que más influye sobre el rendimiento académico.

- ✓ Los resultados fueron evaluado a través de datos reales a través de los cuales se pudo comprobar que el rendimiento académico de los estudiantes está considerado bueno, es decir con promedios de 7.00 – 9.00. Mediante estos resultados las autoridades de la Universidad pueden tomar medidas que les ayude a mejorar la formación académica de los estudiantes.

f. RECOMENDACIONES

Luego de finalizar el Trabajo de Titulación son importantes las siguientes recomendaciones:

- Tener en cuenta los factores obtenidos que más influyen en el rendimiento académico con el fin de tomar decisiones y proponer estrategias que permitan ayudar a los estudiantes a mejorar el rendimiento académico.

- Incorporar nuevos datos, identificando para ello nuevos factores como: socioculturales que abarca el nivel educativo de los padres y también situación laboral del estudiante, las mismas que servirán para determinar e rendimiento académico de los estudiantes.

Anexo 11: Anteproyecto

A. TEMA

"ESTUDIO DEL RENDIMIENTO ACADÉMICO APLICANDO TÉCNICAS DE MINERÍA DE DATOS"

B. PROBLEMÁTICA

1. SITUACIÓN PROBLEMÁTICA

Durante los últimos años las universidades de todo el mundo tienen la preocupación por el Rendimiento Académico (RA) de los estudiantes que se ve afectado por múltiples factores que influyen en él [1], pretendiendo determinar de esta manera cuáles son los que más afectan al rendimiento de los estudiantes [2,3].

Actualmente el proceso de formación educativo ha cambiado debido a que el estudiante es el principal elemento para su aprendizaje, dando como resultado que el estudiante sea el verdadero protagonista de su formación profesional [4,5].

De manera que el RA es fundamental en el ámbito de la educación superior por sus implicaciones en el cumplimiento de la función formativa de las instituciones educativas y el proyecto educativo de los estudiantes debido a que contribuye a la formación profesional en la institución y a nivel individual [6]. Además es un claro indicador del avance exitoso en el transcurso de los estudios de los alumnos, y a su vez también es un pronosticador de la posibilidad de completar exitosamente los estudios [7,8].

Sin embargo el RA se ve afectado por la calidad de vínculo que establece el estudiante con el aprendizaje mismo, teniendo en cuenta que el deseo de saber, la curiosidad, la duda y la pregunta, como elementos de una actitud investigativa, se constituyen en un estilo de vida que caracteriza a los estudiosos y apasionados por la búsqueda del saber [9,10].

Con respecto a la dimensión institucional, algunos autores relacionan de forma directa el rendimiento académico de los estudiantes con el ejercicio de los docentes quienes consideran indispensable el nivel de capacitación y la formación de los mismos, así como su vocación como educadores y calidad humana [11-13].

Por ende la aplicación de Minería de Datos (MD) en el ámbito de la enseñanza, tiene como objetivo obtener una mejor comprensión del proceso de aprendizaje de los

estudiantes y de su participación global en el proceso, orientado a la mejora de la calidad y la eficiencia del sistema educativo [14,15]. A partir de toda la información disponible, las diferentes técnicas de MD pueden ser aplicadas a fin de descubrir conocimiento útil que ayude a mejorar el proceso educativo [7, 16, 17].

Con lo descrito anteriormente se pretende realizar un análisis sobre el rendimiento académico aplicando técnicas de Minería de Datos en base a los sub-problemas que se detallan a continuación:

- Los estudiantes son poco conscientes de la responsabilidad que tienen durante su formación académica por ende proyectan sus dificultades en el sistema educativo como en la utilización de los métodos de enseñanza de los docentes, llevándolos a un fracaso académico [6].

- Si el estudiante no asume una actitud crítica frente a su método de estudio y su compromiso académico, las relaciones en el aula se ven afectadas y el fracaso reiterativo se vuelve en abandono, temor y rechazo hacia el objeto de conocimiento y a quien lo imparte [6, 18].

- El proceso de aprendizaje se basa en la adquisición de nuevos conocimientos por parte de los estudiantes de manera que este se ha visto afectado por múltiples causas como la edad en que ingresan los estudiantes a la universidad, la falta de claridad sobre su identidad profesional, sus aptitudes e intereses provocando el bajo rendimiento académico [6].

- El RA de cada estudiante en la educación superior también se encuentran involucrados los docentes debido a que ellos son la base para la formación de los estudiantes, el RA se encuentra afectado debido a que los docentes no están en constante capacitación y formación continua [9].

- Mediante el Rendimiento Académico se puede determinar el nivel de conocimiento alcanzado por los estudiantes ya que pueden tener el deseo de saber, la curiosidad, la duda y la pregunta, como elementos de una actitud investigativa, el mismo que se ha visto afectado por la calidad de vínculo que establece el estudiante con el aprendizaje [11-13].

2. PROBLEMA DE INVESTIGACIÓN

¿La implementación de un modelo computacional mediante Técnicas de Minería de Datos permitirá determinar el rendimiento académico universitario?

C.JUSTIFICACIÓN

La realización del presente Trabajo de Titulación se analizará la información académica de los estudiantes para medir el rendimiento de los mismos, con el fin que las autoridades de la universidad tomen decisiones que permitan corregir las deficiencias encontradas.

En el ámbito académico permitirá adquirir capacidades y habilidades necesarias para poder llevar con éxito el desarrollo del proyecto, además se obtendrá conocimientos y experiencia que nos permitan resolver problemas que se presenten en la sociedad, así mismo permitirá en las universidades tener en cuenta los factores que influyen en el rendimiento académico de los estudiantes.

Así mismo para recoger la información académica de los estudiantes se hará uso de técnicas que permitan recopilar información apropiada. También se utilizará técnica de Minería de Datos adecuada para analizar la información obtenida y extraer conocimiento de la misma que permita evaluar el rendimiento de los estudiantes, además se utilizará una herramienta de Minería de Datos para implementar el modelo sobre el rendimiento académico.

Igualmente se cuenta con el recurso humano y económico necesario para poder realizar el presente proyecto, así como el tiempo necesario que implica el desarrollo del mismo y la guía continua prestada por el tutor correspondiente.

Además contribuye a la reducción de impactos negativos al medio ambiente, debido a que los resultados obtenidos serán visualizados en la herramienta de Minería de Datos ahorrando de esta manera recursos tales como: papel y tinta.

Con lo descrito anteriormente el proyecto es viable, por lo tanto se puede justificar que se alcanzaran los resultados esperados en cuanto al análisis del rendimiento académico aplicando técnicas de Minería de Datos.

D.OBJETIVOS

OBJETIVO GENERAL

- ✓ Determinar el Rendimiento Académico de los estudiantes mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos.

OBJETIVOS ESPECÍFICOS

- ✓ Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.
- ✓ Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.
- ✓ Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.

E. ALCANCE

El Trabajo de Titulación denominado “Estudio del Rendimiento Académico aplicando Técnicas de Minería de Datos” permitirá medir el rendimiento de los estudiantes en base a datos académicos y de acuerdo a los resultados tomar decisiones a las autoridades de la universidad que les permita mejorar la calidad de la educación.

Además, se trabajará con datos reales de los estudiantes de la Universidad Nacional de Loja del Área de la Energía, Las Industrias y Los Recursos Naturales No Renovables que serán obtenidos por el Sistema de Gestión Académica el mismo que es administrado en el Departamento de Telecomunicaciones e Información (ver Anexo 1) necesarios para el desarrollo del Trabajo de Titulación.

Así mismo el tiempo estimado para realizar el Trabajo de Titulación es de 11 meses a partir de la pertinencia.

Para el presente proyecto se ha estimado en etapas que permiten llevar a cabo el desarrollo, las mismas que se mencionan a continuación:

1. **Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.**

   - ❖ Recopilar información de fuentes académicas, artículos científicos sobre las diversas técnicas de Minería de Datos que permitan determinar el rendimiento académico.
   - ❖ Elaborar un análisis comparativo de las diversas técnicas de Minería de Datos.
   - ❖ Seleccionar la técnica de Minería de Datos que permita identificar de mejor manera el rendimiento académico.
   - ❖ Evaluar la técnica de Minería de Datos para comprobar si se adapta al entorno en el que se va a trabajar.

2. **Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.**

   - ❖ Analizar indicadores que permitan estimar el rendimiento académico.
   - ❖ Seleccionar indicadores para construir el modelo computacional que permita estimar el rendimiento académico.
   - ❖ Plantear un modelo computacional mediante la técnica de Minería de Datos para estimar el rendimiento académico de los estudiantes.

3. **Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.**

   - ❖ Recopilación de información en fuentes académicas, artículos científicos sobre herramientas de Minería de Datos que permitan adaptar el modelo computacional realizado.
   - ❖ Análisis comparativo de las diferentes herramientas de Minería de Datos que permitan adaptar el modelo computacional realizado.

- ❖ Selección de la mejor herramienta de Minería de Datos que permitan adaptar el modelo computacional realizado.
- ❖ Implementar el modelo computacional en la herramienta de Minera de Datos seleccionada.
- ❖ Evaluar el modelo computacional en un escenario real con datos académicos.
- ❖ Demostrar la visualización de los resultados del rendimiento académico de los estudiantes mediante la herramienta de Minería de Datos.
- ❖ Interpretar los resultados obtenidos por la Herramienta de Minería de Datos acerca del rendimiento académico de los estudiantes.

F. MARCO TEÓRICO

CAPITULO 1: CASOS DE ÉXITO, INDICADORES QUE DETERMINAN EL RENDIMIENTO ACADÉMICO EN ESTUDIANTES UNIVERSITARIOS

El rendimiento académico es un claro indicador del avance exitoso en la carrera de estudios y a su vez también es un pronosticador de la posibilidad de completar exitosamente dichos estudios [17, 19, 20].

Además, se debe tenerse en cuenta que se trata de un marco teórico complejo y multidimensional, atravesado y determinado por múltiples factores sociales, económicos, históricos, institucionales e individuales. Por tal motivo el rendimiento académico ha sido representado de diferentes maneras en los diversos estudios que han abordado el tema. En algunos, está representado sólo por el número de materias aprobadas por un alumno en una carrera, en otros por los resultados de test específicamente diseñados o el promedio de notas de las asignaturas cursadas [17, 19, 20].

Por ello existen investigaciones que se orientan en la utilización de técnicas de minería de datos para determinar indicadores que determinan el rendimiento académico en estudiantes universitarios, los mismos que se mencionan a continuación [17, 19, 20].

a. Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos [17].
b. Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de sistemas de la FACENA-UNNE [19].

c. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil [20].

1.1. ANÁLISIS DEL RENDIMIENTO ACADÉMICO EN LOS ESTUDIOS DE INFORMÁTICA DE LA UNIVERSIDAD POLITÉCNICA DE VALENCIA APLICANDO TÉCNICAS DE MINERÍA DE DATOS

1.1.1. Introducción

En este trabajo presentan un análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia, este análisis relaciona el rendimiento con las características socioeconómicas y académicas de los alumnos, que se obtienen en el momento de su matrícula, y que se recogen en la base de datos de la universidad. Han definido un indicador del rendimiento para cada alumno, teniendo en cuenta las calificaciones obtenidas y las convocatorias utilizadas.

Para el estudio utilizamos técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, etc. Esto proporciona una herramienta importante para la acción tutorial, que puede apoyarse en las predicciones de los modelos que se obtienen para encauzar sus recomendaciones y encuadrar las expectativas y el esfuerzo necesario para cada alumno, lógicamente dentro de la cautela habitual a la hora de tratar modelos inferidos a partir de datos.

1.1.2. Factores del Rendimiento Académico

Algunos factores podrían, en gran medida, explicar el éxito o fracaso de un estudiante, como sus características socioeconómicas, edad, estudios previos, entorno al inicio de sus estudios, actividad, o no, laboral durante los estudios, características organizativas y docentes de los centros, planes de estudios, métodos evaluativos, etc. Conocidos estos factores, tanto la universidad como los estamentos responsables de los estudios

preuniversitarios, podrían estudiar acciones que mejoraran el rendimiento de colectivos específicos.

1.1.3. Metodología

La metodología que se ha seguido para la obtención del rendimiento se puede resumir en las siguientes etapas:

1. **Establecer el/los objetivo/s del estudio:** Aplicar técnicas de minería de datos para analizar la influencia de los parámetros (socioeconómicos, características personales, nota de entrada) más relevantes sobre el rendimiento académico de un alumno de primer curso en las titulaciones de informática de la UPV, de forma que nos permita predecir este rendimiento disponiendo únicamente de la información aportada por el alumno en el momento de su matrícula.

2. **Definir la población y la muestra de estudiantes implicada en el estudio:** Está constituida por todos los alumnos de nuevo ingreso en cualquiera de las tres titulaciones de informática de la UPV antes mencionadas. Así, el estudio se ha realizado sobre 569 alumnos de II, 646 alumnos de ITIG y 572 alumnos de ITIS.

3. **Obtención de la vista minable, a partir de la información contenida en la base de datos de la universidad:** Con el fin de crear un almacén de datos y un entorno que facilitara la obtención de datos para realizar el estudio, se decidió integrar los mismos en Oracle. De esta forma se ha podido utilizar como herramienta OLAP el Oracle Discoverer. Con ella se han extraído las vistas minables. Una vista minable puede definirse como una colección de individuos sobre los cuales queremos realizar un determinado estudio, con todas sus características (atributos), que tiene como finalidad poder aplicar el proceso de la minería de datos sobre ella para poder extraer conocimiento útil.

   Así, hemos creado una vista minable por titulación. Cada una de ellas contiene las notas y datos personales de los alumnos de la muestra seleccionada. Para generarla, se ha utilizado el generador de informes del Discoverer, seleccionando, de entre todos los atributos disponibles, aquellos que se consideraron, a priori, que

podrían tener mayor influencia en el rendimiento académico, filtrando el resto. Dichos atributos son:

- Ocupacio P: Ocupación del padre.
- Ocupacio M: Ocupación de la madre.
- Ocupacio A: Ocupación del alumno.
- Ing Nota: Nota con la que el alumno aprueba estudios de acceso.
- Ing Est: Estudios con los que accede a la titulación.

Seguidamente, se procedió a la agrupación de valores de algunos atributos por su elevado número de alternativas, con el fin de reducirlas y hacer más fácilmente interpretables los resultados obtenidos. Tales atributos son:

- D_Altr Estud: Otros estudios universitarios del alumno al ingresar en la titulación.
- D_Estudis P: estudios del padre.
- D_Estudis M: estudios de la madre.
- Dpaises: Derivado del país de nacimiento del alumno, agrupando por zonas geográficas.
- Residencia Alumno: Derivado de la provincia y el código postal donde reside el alumno durante el curso.
- Residencia Familia Alumno: Derivado de la provincia y el código postal donde reside la familia del alumno durante el curso.
- Edad Ingreso: Atributo derivado calculado como la diferencia entre el año de ingreso del alumno y año de nacimiento.

Finalmente, se especificó el tipo de cada atributo como nominal (o categórico) o numérico, siendo todos nominales excepto la nota de acceso a los estudios y la edad del alumno.

4. **Elección del tipo de análisis de datos requerido:** Los modelos de minería de datos que hemos elaborado son de tipo predictivo.

De entre las técnicas de minería de datos existentes, hemos utilizado dos de ellas para generar los modelos predictivos del rendimiento: los árboles de decisión y la regresión multivariante.

- Los árboles de decisión son una serie de decisiones o condiciones organizadas de forma jerárquica, a modo de árbol. Son muy útiles para encontrar estructuras en espacios de alta dimensionalidad y en problemas que mezclan datos categóricos y numéricos. Básicamente, un árbol de decisión es un árbol donde cada nodo representa una condición o test sobre algún atributo y cada rama que parte de ese nodo corresponde a un posible valor para ese atributo. Finalmente, las hojas representan el valor de la variable predicha. Esta técnica se usa en tareas de clasificación, agrupamiento y regresión.

- La regresión multivariante es un método estadístico clásico que permite establecer una relación matemática entre un conjunto de variables independientes X1, X2, ..., Xn y una variable dependiente Y. Se utiliza fundamentalmente en estudios en los que no se puede controlar por diseño los valores de las variables independientes. Los objetivos de un modelo de regresión pueden ser dos: obtener una ecuación que nos permita "predecir" el valor de Y una vez conocidos los valores de las variables independientes, y cuantificar la relación entre las variables independientes y la dependiente con el fin de conocer o explicar mejor la relación. Se trata en este caso de modelos explicativos.

5. **Generación y validación de los modelos:** Para la generación de los modelos se ha utilizado la herramienta SPSS Clementine v.9.0. En concreto, de los árboles de decisión que incorpora el Clementine, hemos utilizado para regresión el árbol C&R, que es un tipo de algoritmo de aprendizaje de árboles que se basa en el algoritmo CART de Leo Breiman. Así mismo, también hemos aplicado el método regresión del Clementine, que implementa una regresión lineal. Dependiendo de la tarea de minería de datos existen diversos criterios que pueden usarse para evaluar los modelos, como, por ejemplo, la precisión predictiva (porcentaje de aciertos) generalmente utilizada en el contexto de la clasificación.

Si la cantidad de datos lo permite, la forma de entrenar y validar un modelo consiste en partir aleatoriamente los datos en dos subconjuntos disjuntos: el de los datos de entrenamiento con el que se genera el modelo, y el de prueba o test (test set), con el que se evalúa el modelo. La Figura 1 muestra el árbol de decisión generado por la herramienta.

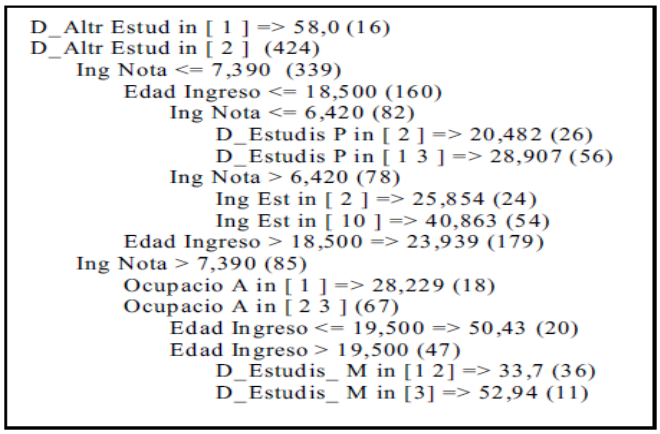


Figura 1. Árbol C&R para la titulación de ITIS

El modelo de regresión lineal multivariante generado para los alumnos de la titulación de ITIS se muestra en la Figura 2. Dicho modelo debe interpretarse de manera que determinados valores de ciertos atributos hacen que el rendimiento de un alumno varíe en un determinado valor, dado por el coeficiente de cada atributo del modelo, y cuyo signo indica si el rendimiento aumenta debido a esa característica, o disminuye.

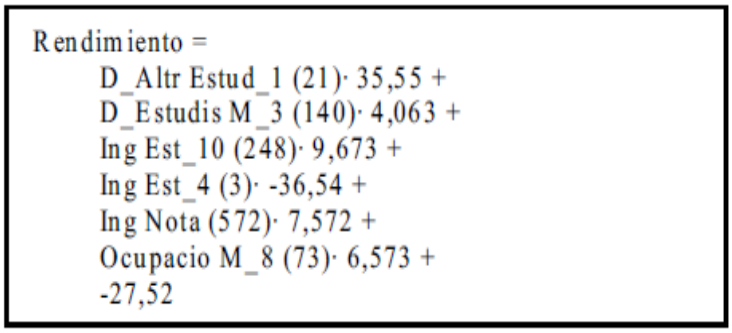


Figura 2. Modelo de regresión para la titulación ITIS

6. **Interpretación de los resultados:** Como puede observarse, los dos modelos mostrados en la sección anterior tienen un error similar. Sin embargo, no tienen en

cuenta los mismos atributos, algo que no parece descartable por el hecho de haber utilizado técnicas diferentes.

a. Análisis del Árbol C&R. Ya que un atributo puede utilizarse a diferentes niveles del árbol (Figura 1) y además repetidamente, hemos calculado para cada atributo el número de ejemplos para los cuales dicho atributo se utiliza. La Tabla 1 incluye esta información para el árbol de la Figura 1.

Atributo	I	IR	IR2
D_Altr Estud	440	1	0,343
Ing Nota	584	1,327	0,455
Edad Ingreso	406	0,922	0,316
D_Estudis P	82	0,186	0,063
D_Estudis M	47	0,106	0,036
Ocupacio A	85	0,193	0,066
Ing Est	78	0,177	0,060
TOTAL	1282	2,913	1

Tabla 1. Análisis de los atributos en el árbol de decisión para ITIS

El análisis ha dado los siguientes resultados:

- Los mejores rendimientos se obtienen para el valor 1 del atributo D_Altr Estud, es decir, los alumnos que ya poseen estudios universitarios, aunque esta condición sólo la cumple un porcentaje relativamente pequeño de alumnos.
- El atributo Ing Nota (nota de ingreso) afecta positivamente.
- El atributo Edad Ingreso afecta negativamente (cuanto mayor es, peor rendimiento).
- El atributo D_Estudis P afecta positivamente para los valores 1 y 3 (padre sin estudios o estudios superiores), y negativamente para el valor 2 (padre con estudios equivalentes a bachillerato).
- El atributo Ocupacio A afecta positivamente para los valores 2 y 3 (alumnos con una ocupación inferior a 15 horas o que no realiza trabajo renumerado), y negativamente para el valor 1 (alumnos con una ocupación mayor o igual a 15 horas semanales).
- El atributo Ing Est afecta positivamente para el valor 10 (alumnos que acceden desde bachillerato LOGSE con PAU), y negativamente para el valor

2 (alumnos que acceden con prueba de acceso pero no provienen de bachillerato LOGSE).

b. Análisis de la regresión lineal. En este modelo aparece en primer lugar el atributo D_Altr Estud con el valor 1 (D_Altr Est_1), indicativo de que poseen ya estudios universitarios. Este factor hace que el rendimiento medio de estos alumnos se incremente en 35,6 puntos, el coeficiente positivo más grande, aunque el número de alumnos que cumple esta condición (21) es reducido.

Los 248 alumnos con el atributo Ing Est_10 (alumnos procedentes de LOGSE con PAU), con un coeficiente de 9,7 positivo tienen rendimiento medio superior a la media. En cambio, aquellos individuos que tienen el atributo Ing Est_4 (titulados universitarios), aparecen con un coeficiente muy negativo, (-36,5), aunque sólo 3 individuos, con edades de ingreso muy superiores a la media (entre 35 y 55 años) aparecen en esta situación.

Probablemente, cargas familiares y de trabajo podrían explicar su rendimiento muy por debajo de la media.

1.2. APLICACIÓN DE MINERÍA DE DATOS CON UNA HERRAMIENTA DE SOFTWARE LIBRE EN LA EVALUACIÓN DEL RENDIMIENTO ACADÉMICO DE LOS ALUMNOS DE LA CARRERA DE SISTEMAS DE LA FACENA-UNNE

En este trabajo se presenta un estudio a través de técnicas de minería de datos que permiten determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste (FACENA-UNNE). Se llevó a cabo un estudio comparativo de diferentes algoritmos clasificadores disponibles en el software Weka, de libre distribución, y se seleccionó el que ofrecía mejores resultados.

La Minería de Datos o Descubrimiento de Conocimiento en Bases de Datos, abarca una variedad de métodos estadísticos y computacionales para investigar la existencia de

relaciones y patrones de comportamiento en almacenamientos electrónicos de datos. Relaciones y patrones emergentes pueden sugerir al investigador explicaciones causales que puedan ser verificadas posteriormente o bien pueden sugerir estrategias de acción para lograr ciertos objetivos de cambio.

El objetivo de este trabajo es presentar un estudio a través de técnicas de minería de datos que permitan determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la FACENA-UNNE.

1.2.1. Materiales y Técnicas a emplear

En este trabajo se utilizó la herramienta Weka (Waikato Environment for Knowledge Analysis) de la Universidad de Waikato, software que se encuentra de manera gratuita en el sitio oficial de esta institución en Internet y contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas.

Los datos utilizados en este análisis fueron obtenidos de un almacén de datos que integra toda la información sistematizada de los alumnos de la Facultad de Ciencias Exactas de la UNNE. El mismo contiene los datos particulares y socio económicos que se registran en el ingreso, los datos de todas las actividades académicas, como asignaturas cursadas y rendidas, trámites de reinscripción y readmisión, reconocimiento de materias y datos del egreso o trámite de graduación.

Así mismo se seleccionaron los alumnos que pertenecen a la carrera Licenciatura en Sistemas de Información que rindieron exámenes finales de las materias que corresponden al primer año en fechas correspondientes al año del ingreso. Con esta información, para cada alumno se calculó: la cantidad de exámenes finales rendidos (número de intentos), la cantidad de exámenes finales aprobados y la cantidad de exámenes finales desaprobados. En función de estos valores, se generaron las tres categorías que identificarán a los alumnos que: 1) en el año de ingreso no rindieron ninguna materia, 2) rindieron pero no aprobaron ninguna y 3) rindieron y aprobaron por lo menos una materia. La consulta resultante se exportó a una planilla de cálculo.

El archivo fue formateado para cumplir con las restricciones del programa Weka que fue utilizado para el procesamiento de los datos, y contiene 2887 registros con las siguientes

variables referidas a los alumnos: año de ingreso (ANIO), sexo (SEXO), estado civil (CIVIL), situación laboral del alumno (SILAAL), grado de instrucción del padre (GRAINSPA), situación laboral del padre (SILAPA), categoría ocupacional del padre (CAOCPA), grado de instrucción de la madre (GRAINSMA), situación laboral de la madre (SILAMA), categoría ocupacional de la madre (CAOCMA), título secundario (TITULO), dependencia del establecimiento secundario (DEPENSEC) y categoría de alumno según la cantidad de materias aprobadas en primer año (CAT_ALUMNO).

La variable o atributo conocido a predecir en este trabajo está representada por CAT_ALUMNO. La misma comprende tres categorías de alumnos, según su rendimiento académico durante el primer año, relacionado con los intentos y resultados de exámenes finales: 1 (no se presentó a rendir nunca), 2 (se presentó a rendir pero no aprobó ninguna materia) y 3 (aprobó una o más materias).

1.2.2. Resultados y Discusión

En la Figura 3 se muestra a través del Explorer de Weka la composición del conjunto de datos y el número de registros por categoría de la variable CAT_ALUMNO.

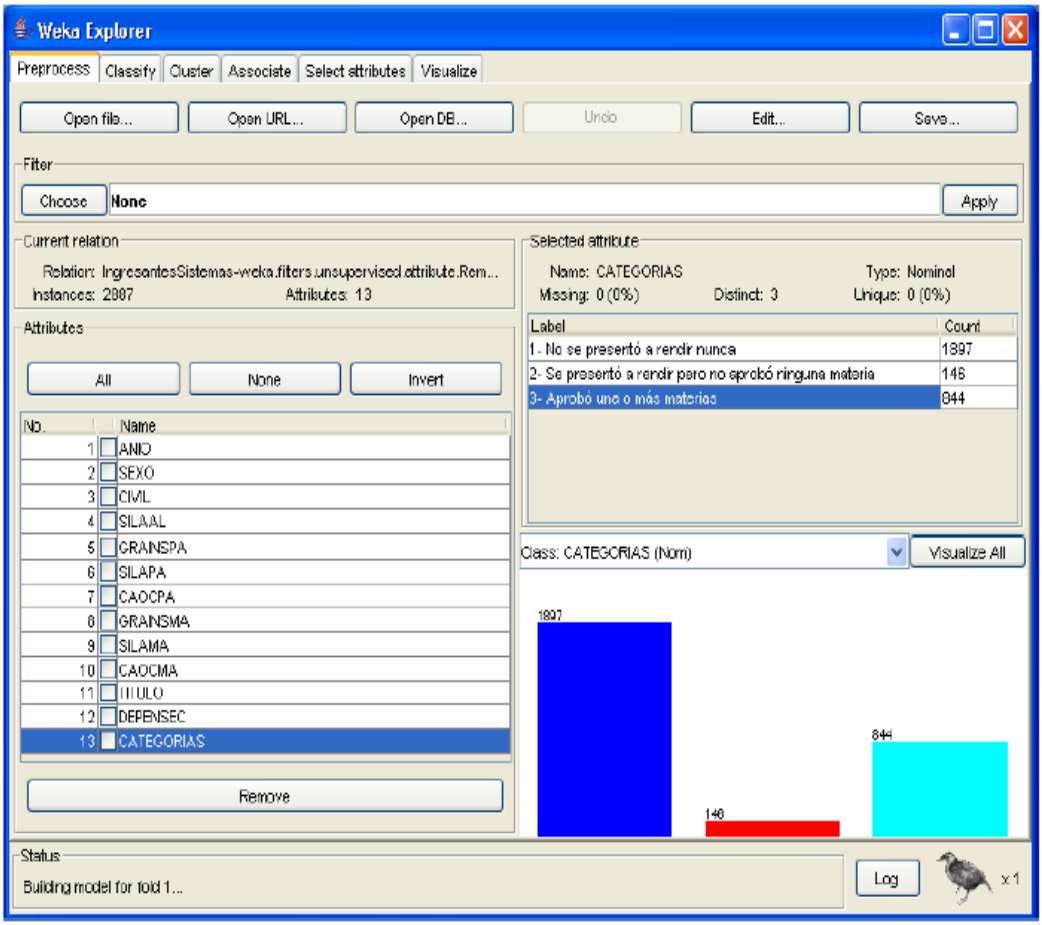


Figura 3. Composición de la primera base de datos estudiada a través de Weka y visualización del número de registros en función de las categorías de CAT_ALUMNO.

Por su parte, en la Figura 4 se visualiza el número de registros por año de ingreso (ANIO), y la proporción de alumnos de categorías 1, 2 y 3 en cada año. Se observa que en los años 2004 y 2005 ha aumentado la proporción de alumnos que no rinden ninguna materia en el primer año y ha disminuido la proporción de alumnos que aprueban al menos una materia durante el primer año.

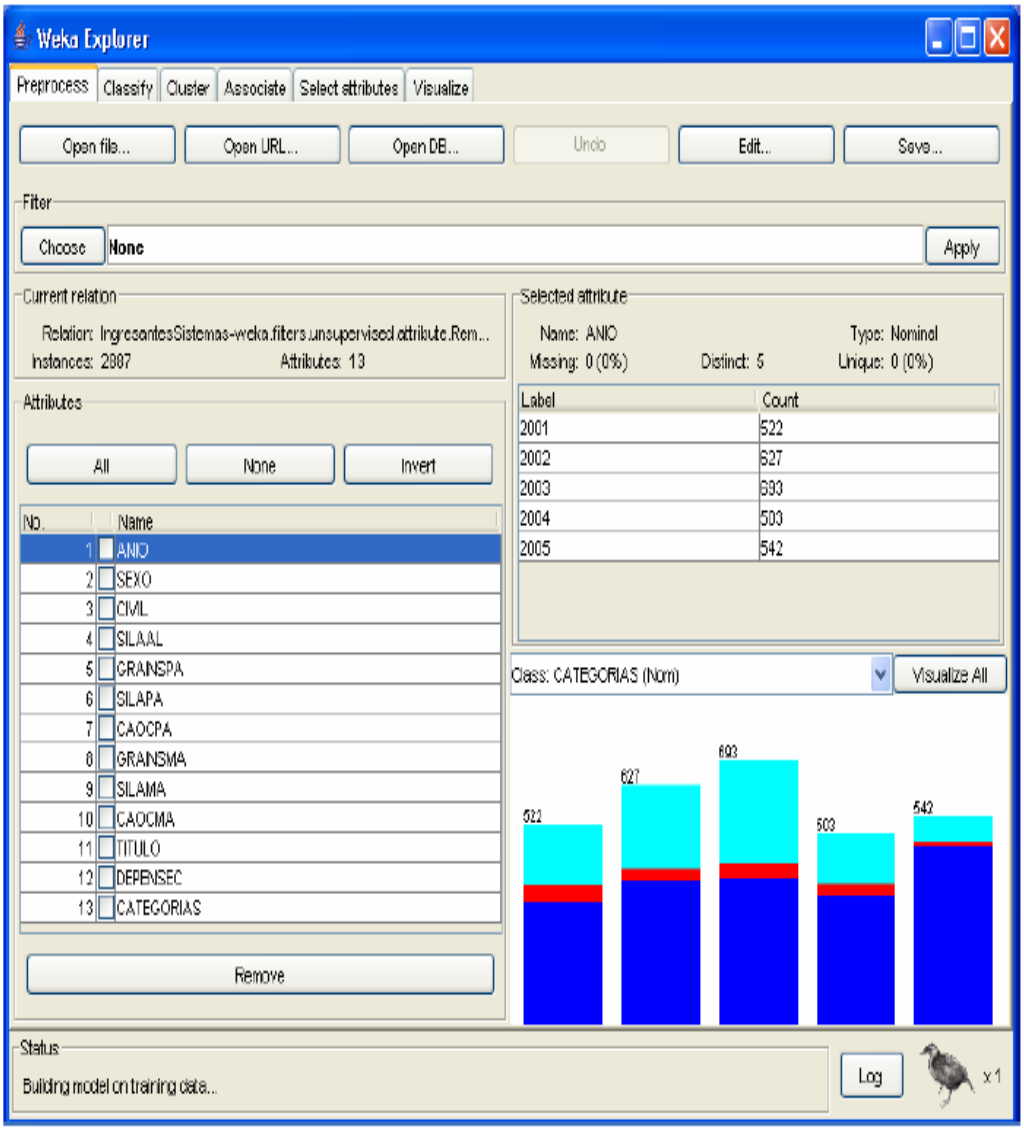


Figura 4. Visualización del número de registros de la variable Año de ingreso (ANIO) en función de la variable CAT_ALUMNO

Finalmente, en la Figura 5 se ilustra el número de registros para las distintas variables en función de las categorías de CAT_ALUMNO.

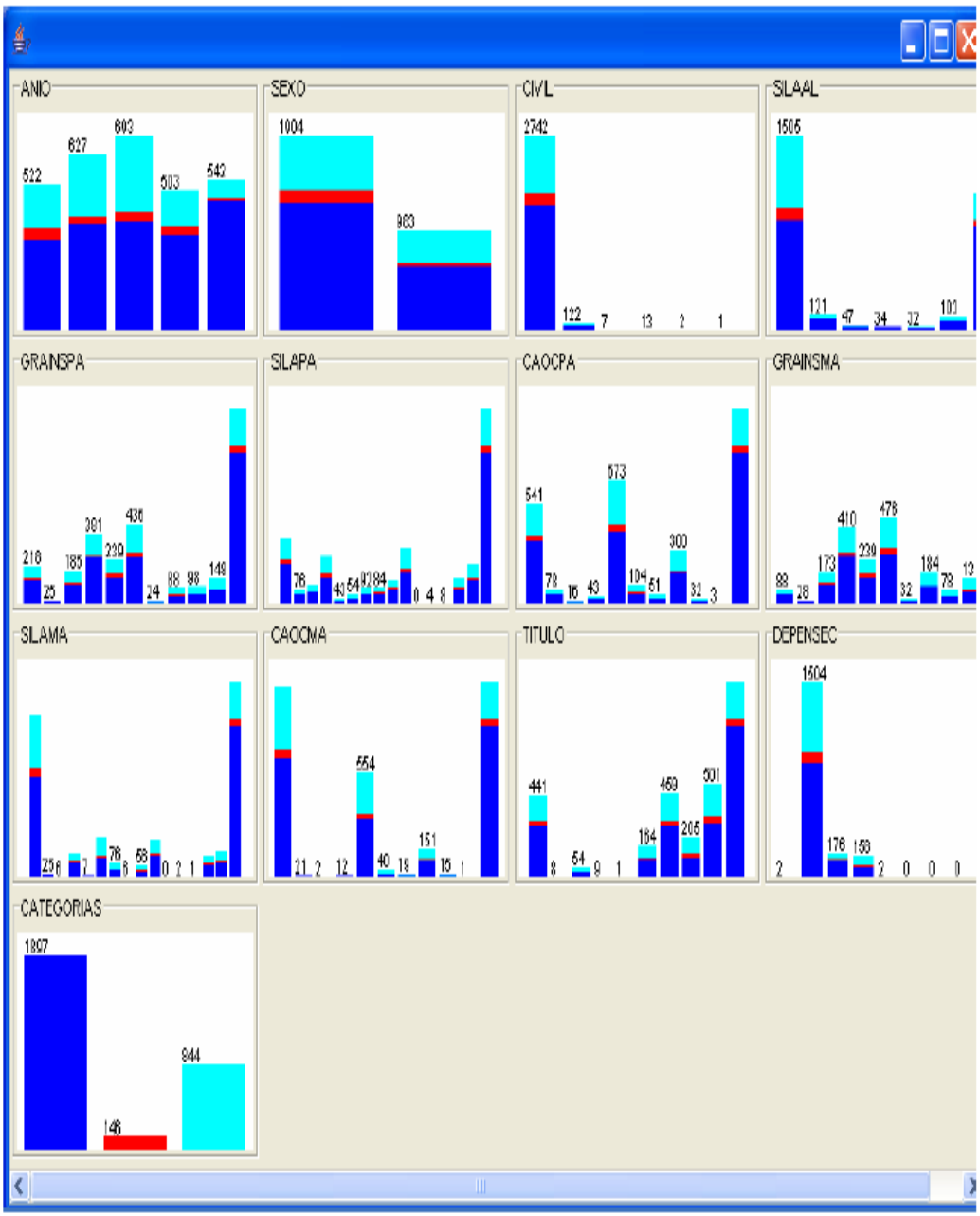


Figura 5. Visualización del número de registros de cada variable en función de la variable CAT_ALUMNO

A continuación, se probaron diferentes algoritmos clasificadores del software Weka, para seleccionar aquél que con un menor error construyese un clasificador para la predicción de la categoría de alumno según su comportamiento durante el primer año (CAT_ALUMNO).

Los mejores resultados fueron obtenidos con el clasificador Logistic (Figura 6), el cual permite estimar y luego emplear modelos de regresión logística múltiple. En el estudio de estos datos se obtuvieron resultados con mediano grado de precisión, ya que el error del clasificador fue de 36,024%, y el porcentaje de instancias clasificadas correctamente fue de 63,97%.

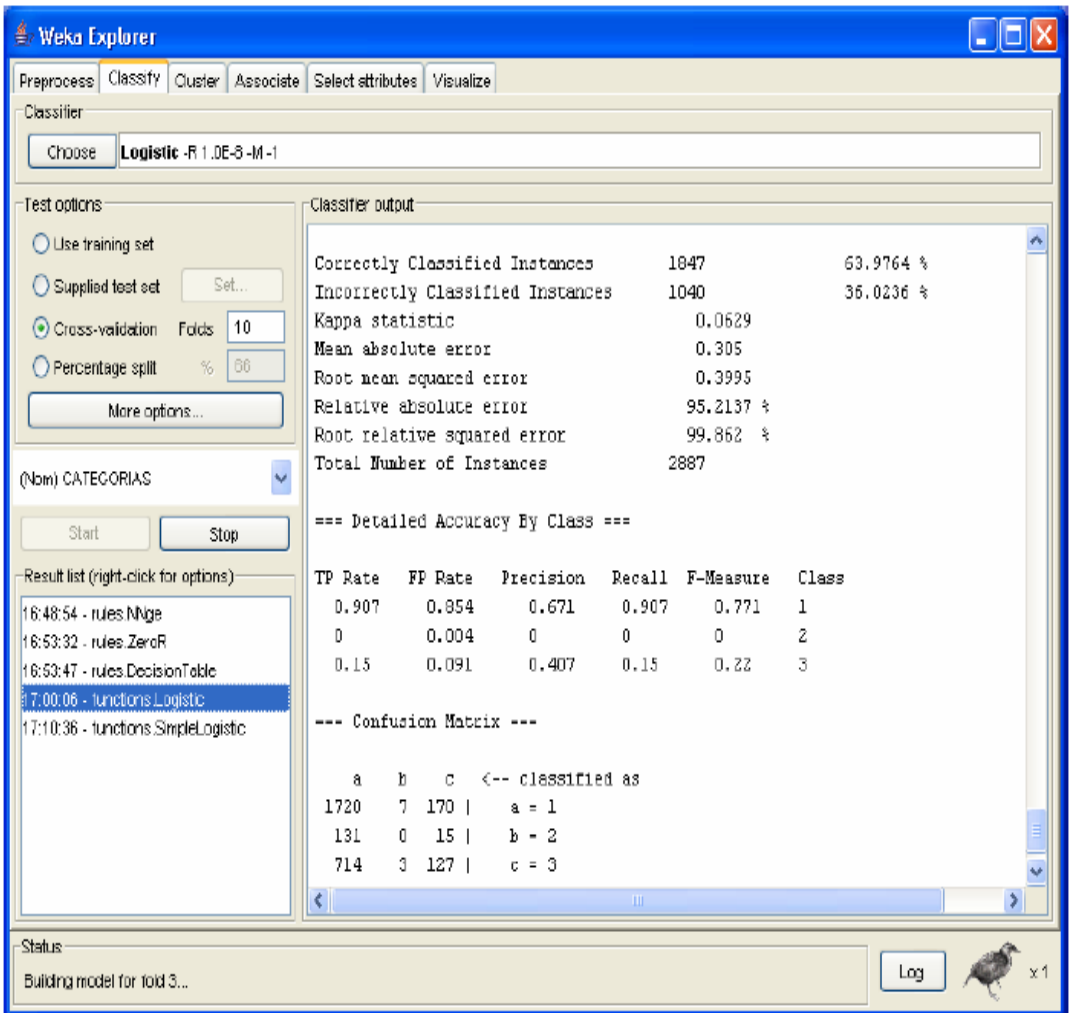


Figura 6. Parte de la salida obtenida mediante el clasificador Logistic de Weka

1.3. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA EVALUACIÓN DEL RENDIMIENTO ACADÉMICO Y LA DESERCIÓN ESTUDIANTIL

Este artículo presenta los resultados de la evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza

(UNLaM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

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

En la Tabla 2 se pueden observar los resultados de un primer análisis cuantitativo del rendimiento de los estudiantes del DIIT durante el año 2008.

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

Tabla 2. Rendimiento de los estudiantes desagregado por carrera.

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

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

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

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

1.3.1. Tecnologías y Herramientas utilizadas

a. Proceso de descubrimiento del conocimiento en base de datos.

El DCDB es un proceso complejo ya que no solo incluye la obtención de los modelos o patrones, sino también la evaluación e interpretación de los mismos. El DCDB es definido en como “el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, compresibles a partir de los datos”. Las principales tareas del proceso de DCDB se pueden resumir en: preprocesar los datos, hacer minería de datos, evaluar los resultados y presentarlos.

En la Figura 7 se puede observar que el proceso de DCDB está organizado en 5 fases:

- **Recopilación e integración:** en esta fase se seleccionan las distintas fuentes de información y se transforman los datos a un formato y unidad de medida comunes generando un almacén de datos.
- **Limpieza, selección y transformación:** en esta fase se eliminan o se corrigen los valores faltantes/erróneos y se seleccionan los atributos más relevantes o se generan nuevos atributos a partir de los existentes para reducir la complejidad de la fase de minería de datos. También se puede reducir la cantidad de instancias.

- **Minería de Datos:** esta es la fase donde se eligen el trabajo a realizar (clasificación, agrupamiento, etc.) y el método a utilizar.
- **Evaluación e interpretación:** en este punto se analizan y evalúan los patrones obtenidos y en caso de ser necesario se retorna a alguna de las fases anteriores.
- **Difusión y uso (presentación):** en esta fase se hace uso de los resultados obtenidos y se difunden entre todos los potenciales usuarios.

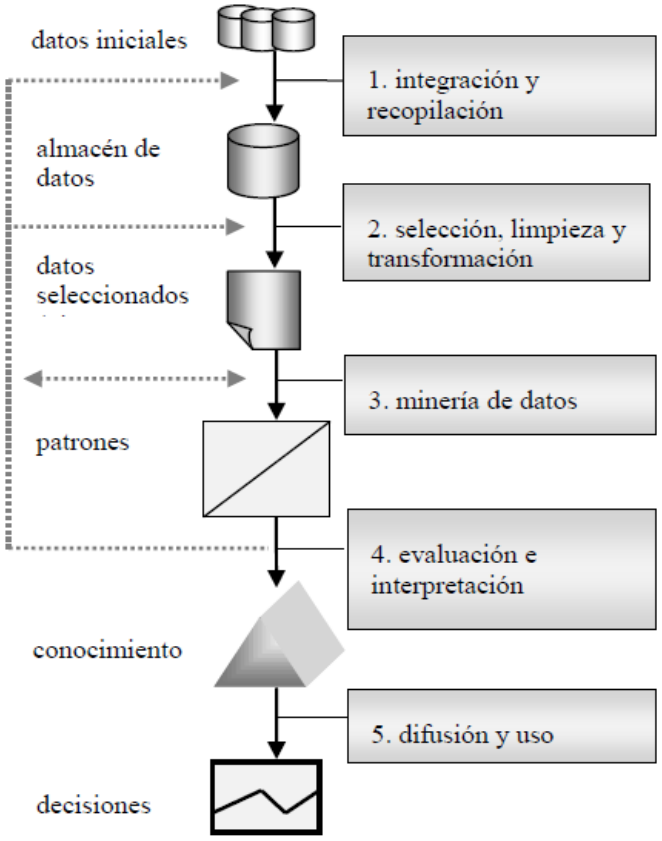


Figura 7. Fases del proceso de descubrimiento de conocimiento en bases de datos, DCDB.

1.3.2. Herramientas de Software.

Las herramientas de software utilizadas en esta investigación fueron:

- **MS SQL Server:** se utilizó para recopilar los datos de las fuentes de información seleccionadas, para realizar una transformación de los datos a partir de la definición de los formatos y las medidas comunes y por último para almacenar los datos transformados (Almacén de Datos).

- **SPSS:** se utilizó para realizar un análisis exploratorio y de correspondencias de los datos. Como resultado del análisis se seleccionaron los atributos más relevantes y se generaron nuevos atributos a partir de los existentes.
- **Weka:** se utilizó para encontrar los patrones que permitan evaluar el rendimiento académico y la deserción estudiantil. Este software contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas.

1.3.3. Resultados del Proceso de Descubrimiento del Conocimiento en Bases de Datos

a. Fase de recopilación e integración: El resultado de esta fase fue la generación de un almacén de datos conformado por 7 tablas cuyas descripciones se pueden ver en la Tabla 3. Para la generación del almacén de datos se tomaron e integraron datos de la base de datos de alumnos de la UNLaM, de la base de datos de encuestas del DIIT y de la bases de datos de colegios de educación secundaria del Ministerio de Educación.

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

Tabla 3. Descripción de las tablas.

En la tarea de integración se transformaron los siguientes atributos:

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

b. **Fase de limpieza, selección y transformación:** Esta fase es la responsable de obtener datos de alta calidad. Para lograr este objetivo se buscó detectar valores anómalos (outliers) y datos faltantes, se realizó una selección de los atributos relevantes y se construyeron nuevos atributos a partir de los existentes.

   Para validar la selección de los atributos relevantes realizada por el Secretario Académico del DIIT y los Coordinadores de las carreras de Informática, Electrónica e Industrial, se realizó un análisis de correspondencias cuyo resultado no modificó los atributos ya seleccionados.

   Los atributos generados, para cada estudiante, fueron:

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

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

Tabla 4. Discretización del atributo índice_materias.

c. Fase de minería de datos:

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

- la clasificación como tipo de tarea de minería.
- el árbol de decisión como tipo de modelo.
- el J48 (implementación en Weka del algoritmo C4.5) [11] y el FT [5] como algoritmos de minería.

En la Tabla 5 se pueden ver los atributos del archivo elaborado para la fase de minería de datos. Este archivo contiene 9545 instancias que representan a los alumnos inactivos, activos y reincorporados. Para entrenar los modelos se utilizó un archivo con 2865 instancias (30% del original), que fueron seleccionadas en forma aleatoria.

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

Tabla 5. Atributos del archivo utilizado en la fase de minería de datos.

Se eligieron como clases los siguientes atributos:

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

❖ **Rendimiento académico:**

El mejor resultado fue obtenido por el algoritmo FT que alcanzó un 78,07% de instancias clasificadas correctamente, mientras que el algoritmo J48 clasificó en forma correcta un 72,53% de las instancias. En la Tabla 6 se puede observar la matriz de confusión generada por el algoritmo FT y en la Tabla 7 la generada por el algoritmo J48.

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

Tabla 6. Matriz de confusión generada por el algoritmo FT.

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

Tabla 7. Matriz de confusión generada por el algoritmo J48.

d. Fase de evaluación e interpretación.

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

Para efectuar la evaluación de los modelos se tomó como medida el porcentaje de aciertos al clasificar una instancia en su respectiva clase.

Por cada algoritmo se realizaron 30 iteraciones y en la Tabla 8 se pueden ver los mejores porcentajes de aciertos.

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

Tabla 8. Porcentaje de aciertos de los algoritmos de clasificación.

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

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

Tabla 9. Porcentaje de aciertos del rendimiento académico malo y de los alumnos inactivos.

Con respecto a la comprensibilidad de los modelos se puede decir:

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

CAPITULO 2: RENDIMIENTO ACADÉMICO

2.1. Definición: Es el resultado cuantitativo que se obtiene en el proceso de aprendizaje de conocimientos, conforme a las evaluaciones que realiza el docente mediante pruebas objetivas y otras actividades complementarias [21].

Por ser cuantificable, el RA determina el nivel de conocimiento alcanzado, y es tomado como único criterio para medir el éxito o fracaso escolar a través de un sistema de calificaciones de 0 a 10 en la mayoría de los centros educativos públicos y privados, para evaluar al estudiante como Deficiente, Bueno, Muy Bueno o Excelente en la comprobación y la evaluación de sus conocimientos y capacidades [21].

2.2. Tipos de Rendimiento Académico

Existen dos tipos de rendimiento académico [21]:

- **Individual.-** Es el que se manifiesta en la adquisición de conocimientos, experiencias, hábitos, destrezas, habilidades, actitudes, aspiraciones, etc.; lo que permitirá al profesor tomar decisiones pedagógicas posteriores.

  - **Rendimiento General.-** Es el que se manifiesta mientras el estudiante va al Centro Educativo, en el aprendizaje de las Líneas de Acción Educativa y hábitos culturales y en la conducta del alumno.

  - **Rendimiento Específico.-** Es el que se da en la resolución de los problemas personales, desarrollo en la vida profesional, familiar y social que se les presenta en el futuro. Se evalúa la vida afectiva del alumno, se considera su conducta parceladamente: sus relaciones con el maestro, consigo mismo, con su modo de vida y con los demás.

- **Social**.- La institución educativa al influir sobre un individuo, no se limita a este sino que a través del mismo ejerce influencia de la sociedad en que se desarrolla.

Se considera factores de influencia social: el campo geográfico de la sociedad donde se sitúa el estudiante, el campo demográfico constituido por el número de personas a las que se extiende la acción educativa.

2.3. Factores que influyen en el proceso de aprendizaje y el rendimiento académico de los estudiantes.

Las características que definen a los jóvenes estudiantes de enseñanza son el resultado de diferentes factores, de manera especial es influyente el contexto educativo inmediato en el que se forman. A continuación se señalan algunos de los factores que influyen en este proceso [21-22]:

- Los métodos utilizados no responden muchas veces a los dinamismos reales de la vida de los jóvenes. La educación sigue siendo considerada por muchos como un proceso de acumulación de conocimientos, por lo que se descuidan otros aspectos importantes de la formación integral como la educación de los sentimientos, el desarrollo de la sensibilidad, entre otros.
- El sistema educativo se mantiene todavía alejado de la realidad y nos prepara para la vida y los compromisos en la sociedad. Al concluir sus estudios, muchos se sienten frustrados pues descubren que no les servirán para conseguir un trabajo, ni para asegurar su futuro.
- La crisis económica ha hecho que los estudiantes cada vez más se vayan vinculando al mundo del trabajo, para aportar económicamente a sus familias o para mantenerse en los estudios. El tiempo limitado para dedicarse al estudio lleva un menor rendimiento académico y a una menor formación.
- El sistema democrático actual ha permitido que los jóvenes busquen espacios para ser protagonistas con respuestas contractivas como grupos de estudio, encuentro, deporte, acción social o participación en el movimiento estudiantil.
- La reforma educativa ha facilitado la apertura al sentido crítico, a la inquietud social y a las primeras experiencias de participación activa.

2.4. Pautas para mejorar el rendimiento académico

El docente puede contribuir a mejorar el rendimiento académico de los alumnos mediante las siguientes actividades [22]:

- Motivar al joven universitario a realizar actividades orientadas al logro y a persistir en ellas.
- Fomentar en los alumnos una alta autoestima.
- Contribuir en la resolución de conflictos personales mediante la orientación y comprensión, de ser necesario recurrir al apoyo psicológico.
- Contar con indicadores fiables del rendimiento académico (notas, informes, revisiones, autoevaluaciones desde diferentes ángulos).
- Distribuir los contenidos teniendo en cuenta las características de los estudiantes.
- Desarrollar talleres de orientación y formación de hábitos de estudio.
- Orientar en cuanto a los métodos, planes y horarios de estudio.

CAPITULO 3: MINERÍA DE DATOS

3.1. Definición.- Se puede definir la Minería de Datos como El proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la Minería de Datos es encontrar modelos inteligibles a partir de los datos [23].

Para que este proceso sea efectivo, debería ser automático o semiautomático y el uso de los patrones descubiertos debería ayudar a la toma de decisiones, y por tanto, un beneficio a la organización [23].

Es el descubrimiento de conocimiento en un conjunto de datos enormemente grande. El conocimiento que se obtiene viene dado en forma de características (patrones) que no son triviales, que son previamente desconocidas y que tienen bastantes posibilidades de ser útiles [24].

3.2. Ventajas [23]:

- A largo plazo, ahorra dinero a la empresa
- Contribuye a la toma de decisiones de forma estratégica
- Mide los resultados en la forma de mejora

- Genera modelos descriptivos, es decir, qué datos influyen en los resultados finales
- Genera modelos predictivos

3.3. Etapas de la Minería de Datos

El proceso de minería de datos pasa por las siguientes fases como se puede observar en la siguiente figura 8:

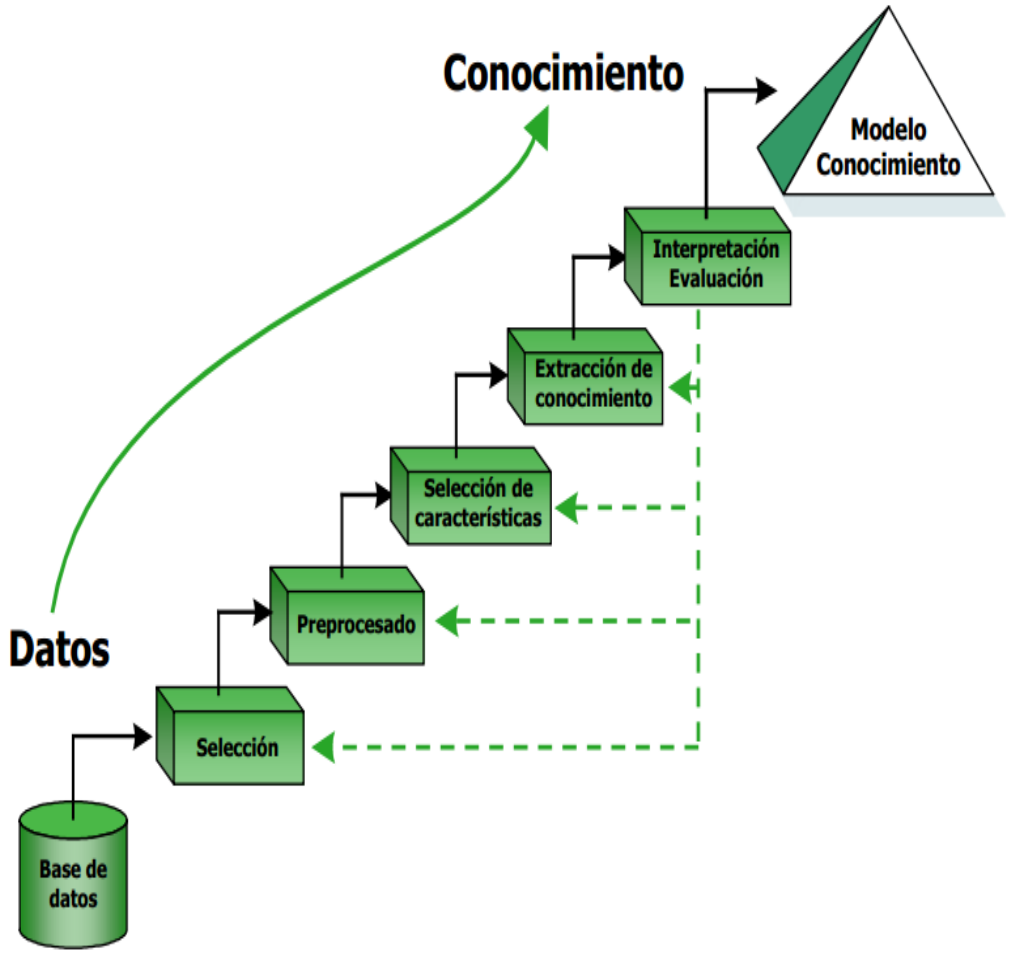


Figura 8: Etapas de la Minería de Datos [25]

- **Preprocesado de datos.-**

  El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto" [26].

Mediante el preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos, según las necesidades y el algoritmo a usar), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering, etc) [26].

- **Selección de Variables**

  Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería [26].

  Los métodos para la selección de características son básicamente dos [26]:

  • Aquellos basados en la elección de los mejores atributos del problema.

  • Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

- **Extracción de Conocimiento**

  Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos [27].

- **Interpretación y Evaluación**

  Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos [27].

3.4. Áreas de Aplicación de la Minería de Datos [28]:

- **Financieras:**
  - Detección de uso fraudulento de tarjetas de crédito
  - Predicción del gasto en tarjeta de crédito por grupos
  - Análisis de riesgos en concesión de créditos
  - Identificación de reglas de mercado a partir de datos históricos
  - Reconocimiento de clientes infieles

- **Comercio**
  - Análisis de la cesta de la compra
  - Evaluación de campañas publicitarias
  - Segmentación de clientes
  - Estimación de stocks, de costes, de ventas

- **Seguros**

  - Determinación de clientes potencialmente caros
  - Predicción de qué tipo de clientes contratan nuevas pólizas
  - Identificación de patrones de comportamiento para clientes con riesgo
  - Identificación de comportamiento fraudulento

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

- **Medicina**
  - Diagnóstico de enfermedades
  - Detección de pacientes con riesgo de sufrir una patología concreta
  - Gestión hospitalaria y asistencial. Predicciones temporales de los centros sanitarios para el mejor uso de recursos
  - Tratamiento de imágenes medicas
- **Otras áreas**

- Telecomunicaciones: detección del fraude
- Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo spam.
- Hacienda: detección de fraude fiscal
- Web: análisis del comportamiento de los usuarios, análisis de los log de un servidor web.
- Deportes: detección riesgo de lesiones a partir de datos médicos.

3.5. Técnicas de la Minería de Datos

La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos con vistas al descubrimiento de información previamente desconocida y que pueda servir de ayuda en el proceso de toma de decisiones [25].

Las técnicas más representativas son:

1. Técnicas No Supervisadas o de Descubrimiento del Conocimiento

a. Clustering o Agrupamiento.- Es el proceso de agrupar los datos en clases o en clústeres, de tal forma que, los datos de un mismo clúster tienen una alta similitud y a su vez, son muy diferentes de los de otro clúster [29].

Al hacer clústeres, se puede identificar regiones densas y regiones dispersas en el espacio de características, y por lo tanto, descubrir distribución de patrones y correlaciones entre los atributos [29].
A diferencia de la clasificación, el Clustering o aprendizaje no supervisado no requiere clases predefinidas (ni conjuntos de entrenamiento) [29].

b. Reglas de Asociación.- Es la exploración de los datos con el propósito de identificar relaciones entre los datos, dentro de una fuente o base de datos [30].

Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas

pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos [31].

2. **Técnicas Supervisadas o Predictivas**
   a. **Predicción.-** Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos [31].

      Además comprende el uso de algunas variables o campos de la base de datos para predecir valores futuros o desconocidos, o incluso otras variables de interés. También se centran en encontrar patrones comprensibles para el ser humano que describan la información que tenemos [32].

      - **Regresión.-** Es una técnica utilizada para inter y extrapolar las observaciones, las cuales pueden clasificarse como regresión lineal o no lineal. Hablamos de modelo de regresión cuando la variable de respuesta y las variables explicativas son todas ellas cuantitativas. Si sólo disponemos de una variable explicativa hablamos de regresión simple, mientras que si disponemos de varias variables explicativas se trata de un problema de regresión múltiple [32].

      - **Arboles de Decisión.-** Son ampliamente usados y pueden ser fácilmente explicados basándose en el criterio usado para dividir los datos en las extremidades del árbol. Los árboles de decisión son estructuras que representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos [33].

   b. **Clasificación.-** Técnica que permite encontrar modelos (funciones) que describen y distinguen clases o conceptos para futuras predicciones. Además empareja o asocia datos a grupos predefinidos [34].

      - **Redes Neuronales.-** Las redes neuronales simulan el cerebro humano mediante el aprendizaje de un conjunto de datos de formación y la aplicación del aprendizaje para generalizar los patrones para la clasificación y predicción [33].

Las redes neuronales consisten en modelos predecibles, no lineales que aprenden a través del entrenamiento, generalizando los patrones que se encuentran en él, para clasificarlos y hacer pronósticos con ellos. Una vez la red neuronal ha sido entrenada, puede trabajar con gran cantidad de datos en una fracción del tiempo gastado por un humano. Las redes neuronales son ampliamente usadas para detectar actividades fraudulentas [33].

- **Clasificación Bayesiana.-** Son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular [31].

- **Lógica Borrosa.-** Surge de la necesidad de modelar la realidad de una forma más exacta evitando precisamente el determinismo o la exactitud, es decir permite el tratamiento probabilístico de la categorización de un colectivo. Así, para establecer una serie de grupos, segmentos o clases en los cuales se puedan clasificar a las personas por la edad, lo inmediato sería proponer unas edades límite para establecer tal clasificación de forma disjunta [31].

- **Métodos estadísticos.-** El objetivo de la modelización estadística consiste en explicar el comportamiento de una variable a partir del conocimiento de otras. Subyacente al concepto de modelización está la idea de que una variable tiene una cierta variabilidad y que esta variabilidad está relacionada con el comportamiento de otras variables [32].

- **Técnicas Genéticas.-** Tienen algo en común con las redes neuronales ya que ésta técnica también tiene su base en la biología. Los algoritmos genéticos aplican los mecanismos de la genética y de la selección natural para buscar conjuntos óptimos de parámetros que describan una función de predicción [33].

3.6. Herramientas de Minería de Datos

3.6.1. WEKA

Se trata de un acrónimo derivado de Waikato Environment for Knowledge Analysis – Entorno para Análisis del Conocimiento de la Universidad de Waikato. Esto es porque fue esta universidad la que inició el desarrollo de Weka en 1993, no obstante, su desarrollo original fue hecho en TCL/TK y C, para en 1997 pasar a reescribirse todo el código fuente del entorno en Java, una plataforma más universal, y añadir las implementaciones de diferentes algoritmos de modelado [32].

Es una herramienta que permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos [35].

WEKA contiene métodos de clasificación, regresión, clustering y reglas de asociación [36]. Como se puede observar en la siguiente figura 9, Weka define 4 entornes de trabajo:

Figura 9: Interfaz de WEKA [32]

- **Simple CLI:** la interfaz "Command-Line Interfaz" es simplemente una ventana de comandos java para ejecutar las clases de WEKA [37]. L a interfaz se puede observar en la figura 10.

  La primera distribución de WEKA no disponía de interfaz gráfica y las clases de sus paquetes se podían ejecutar desde la línea de comandos pasando los argumentos adecuados [37].

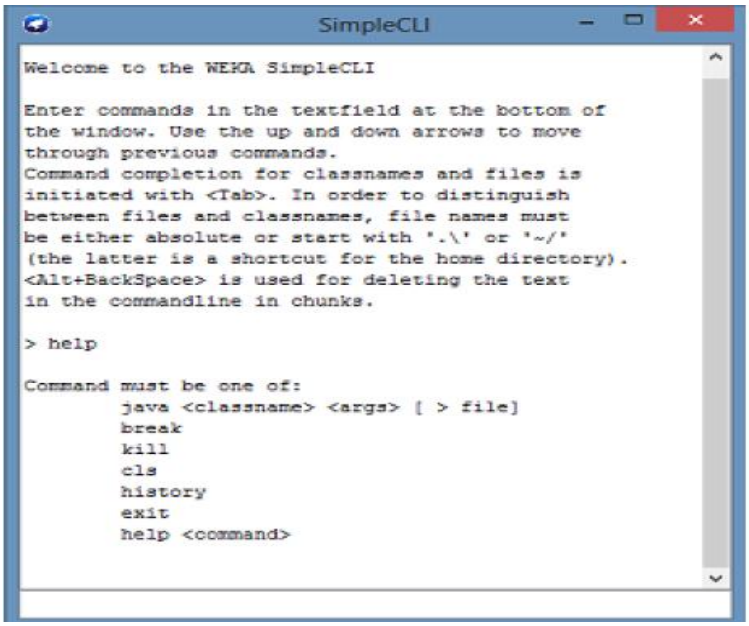


Figura 10: Interfaz SimpleCLI [38]

- **Explorer:** permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos [39].

  Como se puede observar existen 6 sub-entornos de ejecución en la figura 11 [40]:

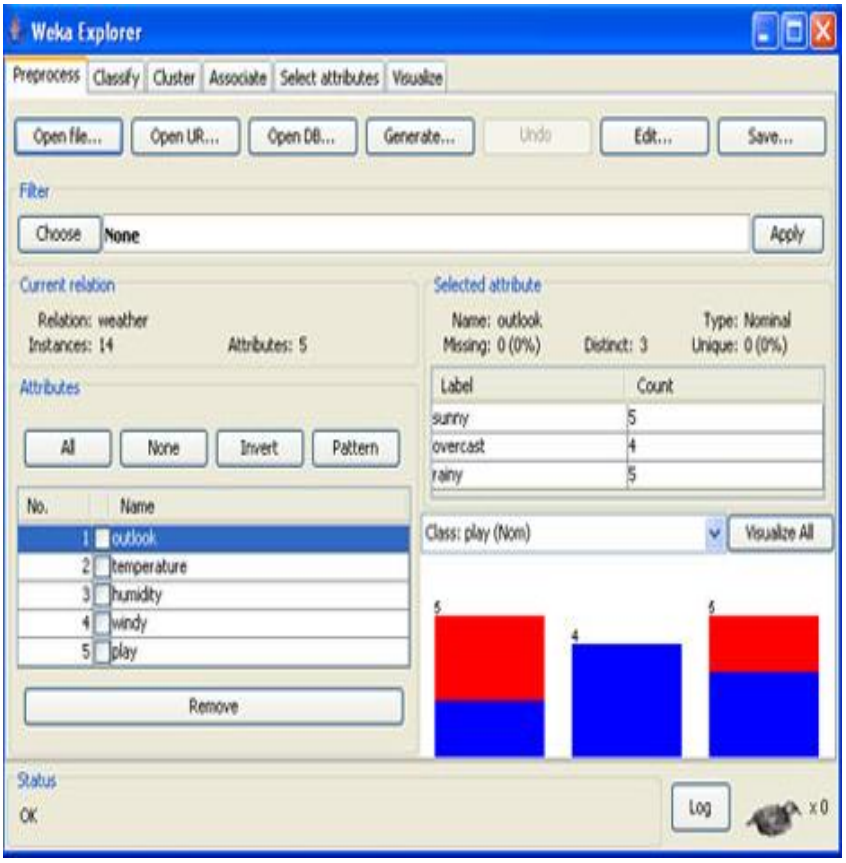


Figura 11: Ventana Weka Explorer [36]

- Preprocess: Incluye las herramientas y filtros para cargar y manipular los datos.
- Classify: Acceso a las técnicas de clasificación y regresión
- Cluster: Integra varios métodos de agrupamiento
- Associate: Incluye técnicas de reglas de asociación
- Select Attributes: Permite aplicar diversas técnicas para la reducción del número de atributos
- Visualize: En este apartado podemos estudiar el comportamiento de los datos mediante técnicas de visualización.

➢ **Experimenter:** Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala [40].

Es un entorno grafico que permite al usuario crear, ejecutar, modificar y analizar experimentos sobre tareas de clasificación de un modelo ágil y eficaz (ver Figura 12) [41].

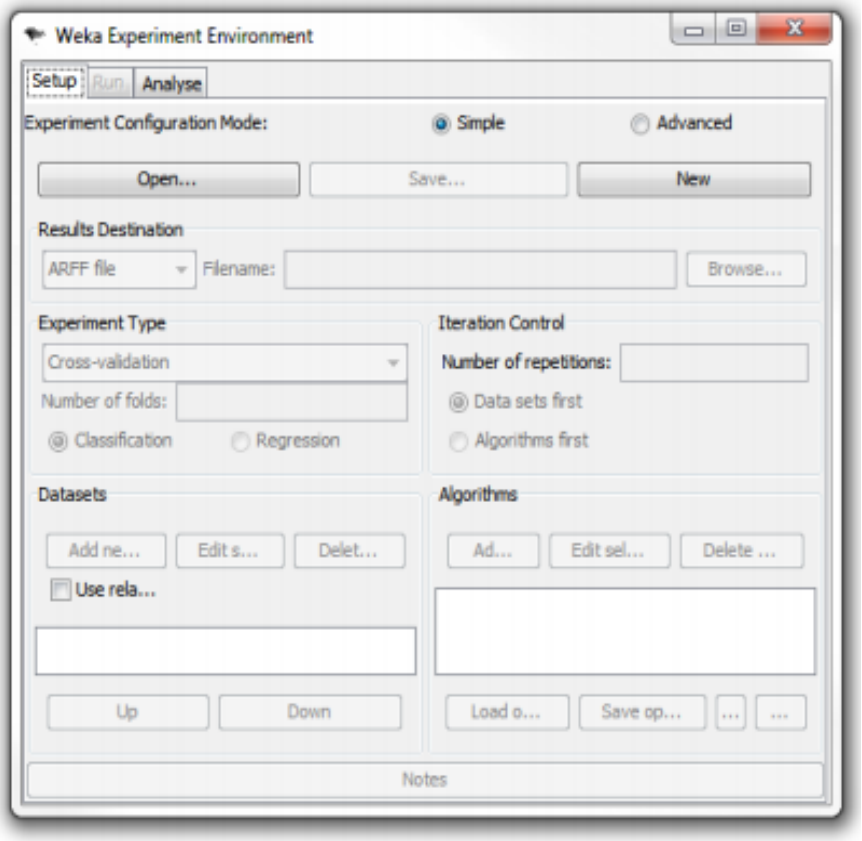


Figura 12: Ventana Experimenter [32]

- **KnowledgeFlow:** Permite generar proyectos de minería de datos mediante la generación de flujos de información [40]. En la Figura 13 se puede observar la ventana:

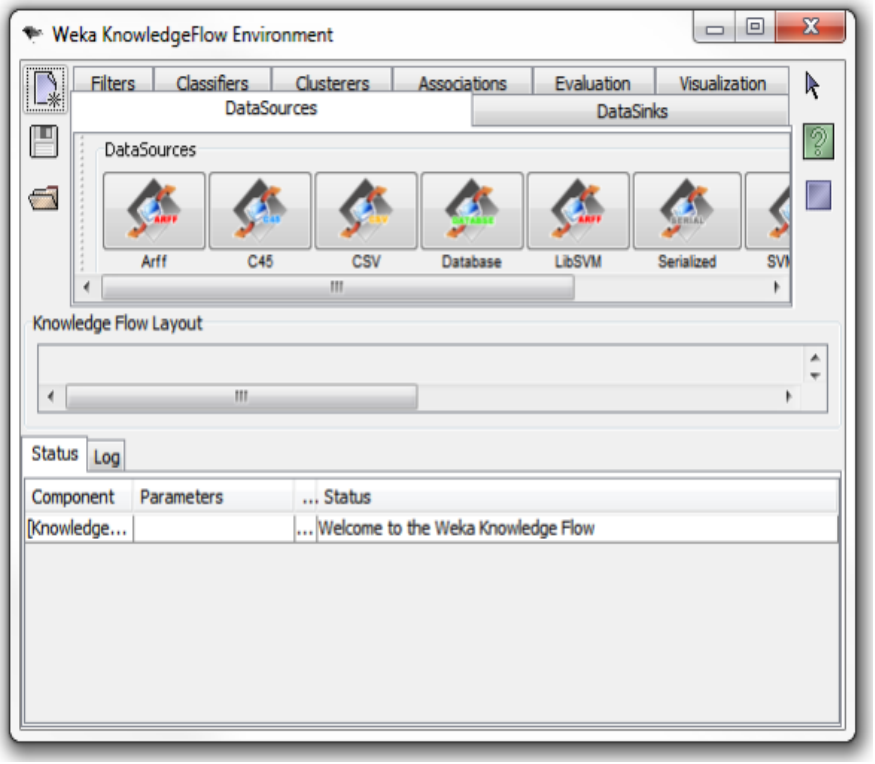


Figura 13: Ventana **KnowledgeFlow [32]**

3.6.2. KNIME

Knime es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de minería de datos [30].

Fue desarrollada originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza (Alemania) bajo la supervisión del profesor Michael Berthold [30].

El diseño de esta herramienta se basa en el diseño de un flujo de ejecución que van reflejando las diferentes etapas de un proyecto de minería de datos [30].

Knime es una plataforma modular de exploración de datos, que permite al usuario crear flujos de datos, o pipelines, de forma visual e intuitiva [42].

Ventajas de Herramienta KNIME [43]:

- Interfaz de usuario amigable
- Alta portabilidad respecto a la fuente de datos
- Diversidad de algoritmos de clasificación.
- Herramientas gráficas adecuadas.

Partes de la Herramienta KNIME:

En la siguiente figura 14 se indica cada una de las partes de esta herramienta [38]:

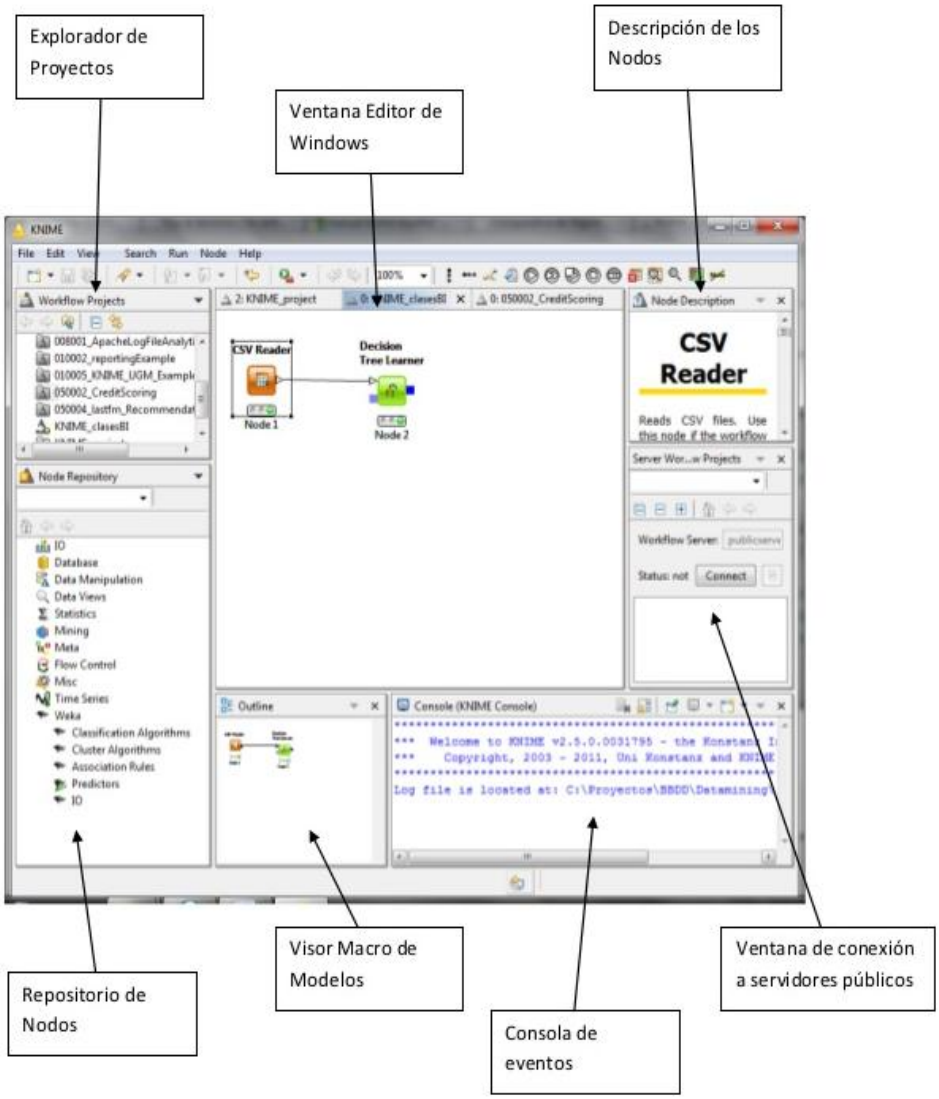


Figura 14: partes de Herramienta Knime [38]

Para ello, *KNIME* proporciona distintos nodos agrupados en fichas, como por ejemplo [38]:

a) Entrada de datos [IO > Read]

b) Salida de datos [IO > Write]

c) Preprocesamiento [Data Manipulation], para filtrar, discretizar, normalizar, filtrar, seleccionar variables, etc.

d) Minería de datos [Mining], para construir modelos (reglas de asociación, clustering, clasificación, MDS, PCA…)

e) Salida de resultados [Data Views] para mostrar resultados en pantalla (ya sea de forma textual o gráfica)

Por otro lado, a través de plugins, los usuarios pueden añadir módulos de texto, imágenes, procesamiento de series de tiempo y la integración de varios proyectos de código abierto, tales como el lenguaje de programación *R*, *WEKA*, el kit de desarrollo de Química y LIBSVM [38].

3.6.3. ORACLE DATA MINING

Oracle Data Mining (ODM) es una herramienta de software desarrollada por la empresa Oracle para aplicar técnicas de minería de datos a grandes volúmenes de datos [44]. La pantalla principal se puede observar en la Figura 15.

Permite a los analistas de datos para trabajar directamente con los datos dentro de la base de datos, explorar los datos gráficamente, construir y evaluar varios modelos de minería de datos, aplicar los modelos de minería de datos Oracle con los nuevos datos y desplegar Oracle predicciones y perspectivas de minería de datos [44].

Figura 15: Ventana Principal de Oracle Data Mining [38]

La herramienta ODM está basada en un esquema de flujo de trabajo, similar a otras herramientas de minería de datos, siendo una extensión del SQLDeveloper, permitiendo analizar los datos, explorar los datos, construir y evaluar modelos y aplicar estos modelos a nuevos datos, así como compartir estos modelos en aplicaciones en línea entregando resultados en tiempo real. La herramienta integra todas las etapas del proceso de la minería de datos y permite integrar los modelos en otras aplicaciones con objetivos similares [38].

ODM funciona dentro de la base de datos de Oracle, así que no hay necesidad de exportar los archivos a un paquete de software estadístico fuera de la base de datos, lo que reduce los costos y mejora la eficiencia. Con un lenguaje de procedimiento integrado/ lenguaje de consulta estructurado (PL / SQL) e interfaces de Java de programación de

aplicaciones (API), Oracle DM permite a los usuarios construir modelos [38].

ODM ofrece dos versiones, una en la que a través de una interfaz gráfica los usuarios podrán aplicar las técnicas de minerías de datos que consideren necesarias y una versión en la que los desarrolladores podrán utilizar la API de SQP para crear aplicaciones a medida [38].

Se trata de la herramienta más potente para trabajar con bases de datos de Oracle, si bien habrá que pagar una licencia por su uso [38].

3.6.4. RAPID MINER

RapidMiner es un entorno de código abierto para aprendizaje automático y minería de datos. Se pueden hacer con RapidMiner todos los procesos que intervienen en un proyecto: la adquisición de datos, la transformación de los datos, la selección de datos, la selección de atributos, la transformación de los atributos, el aprendizaje/modelización y la validación [45].

RapidMiner tiene también disponibles algunos plug-ins desarrollados por la comunidad para el procesamiento de diferentes tipos de datos como datos temporales, procesamiento de texto o la minería de la web [45].

Implementa todos los operadores de data mining, modelos predictivos, modelos descriptivos, transformación de datos, series de tiempo, etc. [38]. RapidMiner permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico [38]. Lo que hace posible aumentar la productividad a través de modelos que solucionan los problemas de predicción, clasificación y segmentación de la información [46].

RapidMiner contiene más de 500 técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, entre otros [46].

Entre las características principales de RapidMiner destacamos que [46]:

- Está desarrollado en Java.
- Es multiplataforma.
- Representación interna de los procesos de análisis de datos en ficheros XML.
- Permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML.
- Permite el desarrollo de programas a través de un lenguaje de script.
- Puede usarse de diversas maneras:
    - A través de un GUI.
    - En línea de comandos.
    - En batch (lotes)
    - Desde otros programas, a través de llamadas a sus bibliotecas.
- Extensible.
- Incluye gráficos y herramientas de visualización de datos.
- Dispone de un módulo de integración con R.
- Software de código abierto.

3.6.5. R

R es un entorno de software libre para el cálculo estadístico y gráficos. Se proporciona una amplia variedad de técnicas estadísticas y gráficos. R puede ser extendido fácilmente a través de paquetes. Hay alrededor de 4000 paquetes disponibles en el repositorio de paquetes CRAN [47].

Es orientado a objetos e interpretado, por lo tanto permite al usuario interactuar con la línea de comandos al mismo tiempo que crea gráficos vectoriales de alta calidad [48].

El entorno de *R* se caracteriza por su flexibilidad e incluye, entre otros [49]:

- Un buen gestor de datos.
- Un conjunto de operadores para cálculos en arrays (vectores de gran tamaño)
- Un conjunto integrado de herramientas de análisis de datos.
- Funciones gráficas para análisis y visualización de los datos.
- Un lenguaje de programación simple que incluye condicionales, bucles, funciones recursivas definidas por el usuario y capacidades de entrada y salida.

¿Por qué usar R para Minería de Datos?

R es un entorno de computación estadística de alta calidad disponible de manera gratuita para múltiples plataformas [48].

Gran parte de las funciones estadísticas y matemáticas que usan los algoritmos de minería de datos forman parte de la distribución base o de R [48].

En relación al proceso de minería de datos, R posee gran cantidad de paquetes estadísticos útiles para realizar este proceso; en especial, destacaremos [49]:

- ***Rattle*:** El paquete de Rattle, es uno de una serie de herramientas analíticas para el análisis datos. Permite a los usuarios analizar datos de múltiples dimensiones o ángulos diferentes, categorizarlo, y resumir la relación identificada. Técnicamente, la minería de datos es el proceso de encontrar correlaciones o patrones entre docenas de campos en grandes bases de datos relacionales. La minería de datos combina conceptuales, herramientas y algoritmos de aprendizaje automático y la estadística para el análisis de grandes conjuntos de datos, con el fin de obtener conocimientos, la comprensión y el conocimiento para la acción [50]. En la siguiente figura se observa la pantalla principal de Rattle.

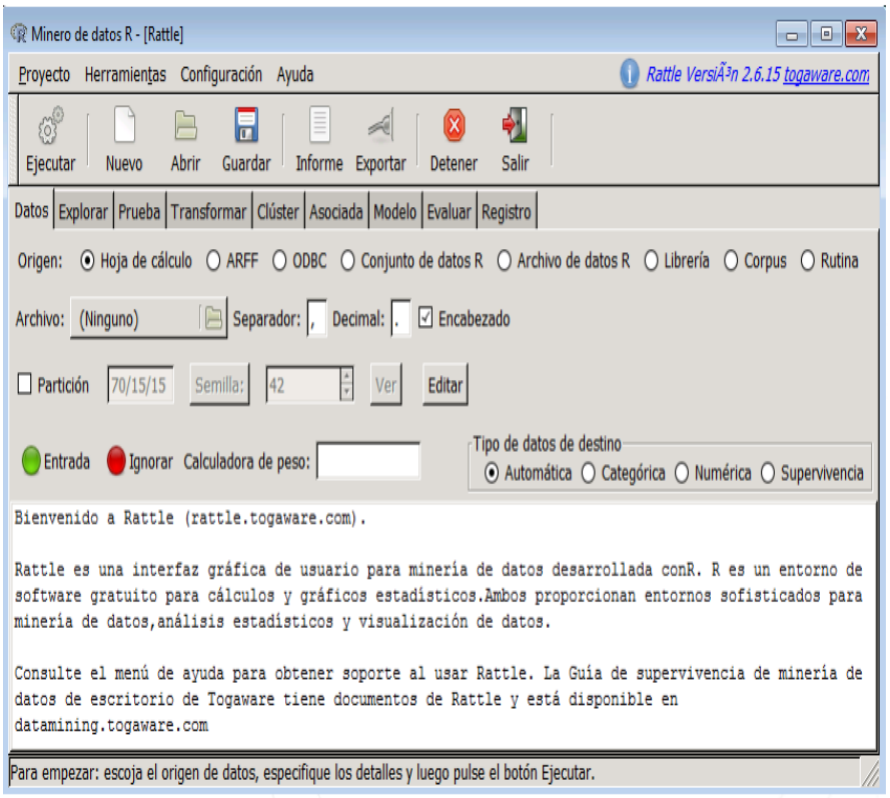


Figura 16: Pantalla Principal de Rattle [50].

Rattle (La herramienta analítica R aprender fácilmente) es una aplicación de minería de datos gráfica y escrita en proporcionar un camino hacia R. Se ha desarrollado específicamente para facilitar la transición, desde básico minería de datos para los datos sofisticados análisis utilizando un lenguaje estadístico de gran alcance. Rattle lleva juntos una multitud de paquetes de R que son esenciales para la minería de datos, pero a menudo no es fácil para el principiante de usar [50].

Dentro de las opciones que ofrece Rattle para cargar los datos se tienen [51]:

- Hoja de cálculo: en la cual se pueden cargar archivos CSV (delimitado por comas). En este tipo de archivo las columnas que corresponden a cada variable se separan por medio de comas y en algunos casos por punto y coma.

- ARFF (Attribute-Relation File Format): es esencialmente un CSV con un encabezado que describe los metadatos. Este tipo de formato fue desarrollado para trabajar con WEKA.
- ODBC (Open Database Connectivity): es un estándar desarrollado para acceder a la información almacenada en una base de datos. R permite conexión con las siguientes bases de datos MS/Excel, MS/Access, SQL Server, Oracle, IBM DB2, Teradata, MySQL, Postgres, y SQLite.
- Conjunto de datos R: esta opción permite cargar datos en Rattle que hayan sido cargados en R.
- Archivo de datos R: Rattle brinda la posibilidad de cargar archivos nativos de R los cuales por lo general tienen la extensión RDATA. Estos archivos pueden contener varios conjuntos de datos.
- Librería: en esta opción se tiene acceso a varios conjuntos de datos disponibles en R.

- **FactoMineR.** Es un R paquete dedicado al análisis de datos multivariados. Las principales características de este paquete es la posibilidad de tener en cuenta los diferentes tipos de variables (cuantitativas o categóricas), diferentes tipos de estructura de los datos (una partición de las variables, una jerarquía de las variables, una partición en los individuos) y finalmente, la información complementaria (individuos y las suplementarias). Además, las dimensiones emitidos desde los diferentes análisis exploratorio de datos se pueden describir de forma automática por variables cuantitativas y / o categóricas. Es un paquete de R dedicada al análisis multivariado exploratorio de datos [52].

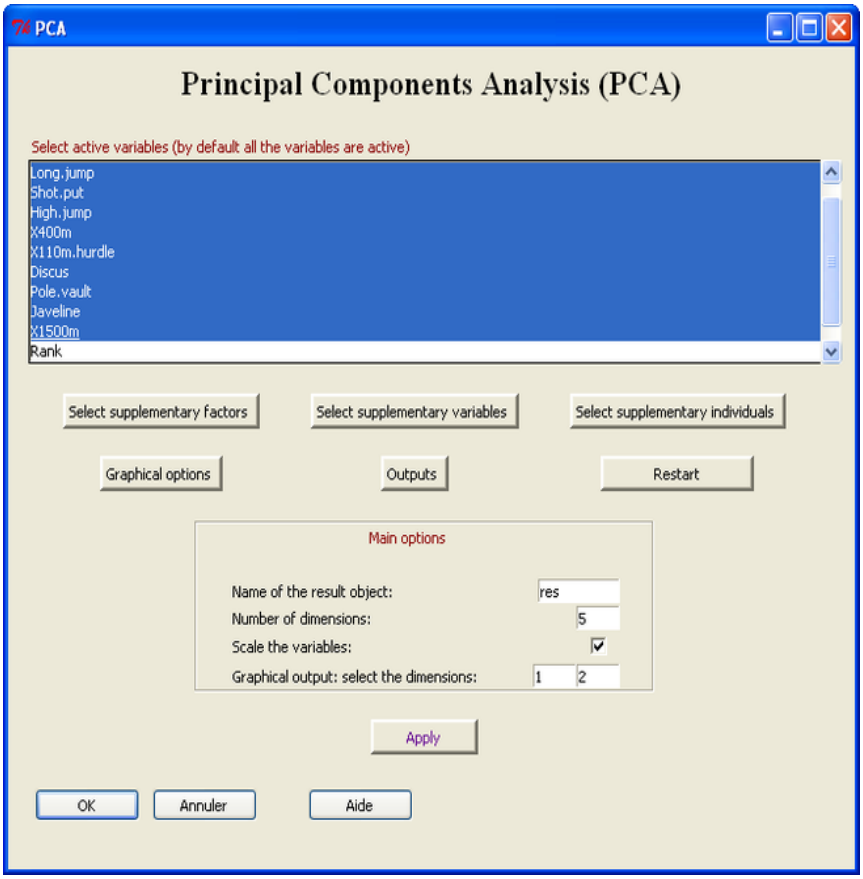


Figura 17: Pantalla Principal de FactoMineR [53]

¿Por qué utilizar FactoMineR? [52]**:**

- Realiza métodos clásicos como el Análisis de Componentes Principales (PCA), análisis de correspondencia (CA), Análisis de Correspondencias Múltiples (ACM), así como los métodos más avanzados.
- Permite añadir informaciones complementarias, como individuos y / o variables suplementarias.
- Proporciona un punto de vista y un montón de salidas gráficas geométrica.
- Ofrece una gran cantidad de ayuda a interpretar (descripción automática de las dimensiones, diversos indicadores, etc.).
- Se puede tener en cuenta una estructura de los datos (estructura de las variables, la jerarquía de las variables, la estructura en los individuos).
- Una interfaz gráfica de usuario está disponible.

G. METODOLOGÍA

Para el desarrollo del presente Trabajo de Titulación denominado “Estudio del Rendimiento Académico aplicando Técnicas de Minería de Datos”, se utilizará una metodología que establezca una secuencia de pasos ordenados que nos permitan lograr cumplir los objetivos del Trabajo de Titulación con el fin de obtener los resultados esperados.

El Trabajo de Titulación es un proyecto basado en la investigación bibliográfica porque se realizó un análisis de los problemas que afectan el rendimiento académico basándose en fuentes bibliográficas confiables y casos de éxito. Además es un proyecto de desarrollo porque se implementará un modelo que permita estimar en rendimiento académico de los estudiantes.

Para la recolección de la información se utilizará el Método Científico para la formulación del marco teórico, en donde se indicará temas como: Casos de Éxito, Rendimiento Académico, Técnicas de Minería de Datos, Herramientas de Minería de Datos que estén enfocados en el proceso del presente proyecto. También el Método Deductivo para ayudar a conocer los problemas de las universidades sobre el rendimiento académico de los estudiantes y a través de esto las mismas podrá tomar decisiones que permitan mejoras. Así mimo el Método Inductivo se lo utilizará para obtener información académica de cada uno de los estudiantes y hacer un análisis de cada uno de los inconvenientes que tienen para poder obtener el problema general de estudio, y así enfocarnos directamente en resolver dicho problema.

Además se utilizará técnicas para recopilar información como: Bibliográfica para la revisión de diferentes fuentes de información confiables enfocados en el tema de Trabajo de Titulación con el fin de detectar problemas y causas que afectan el rendimiento académico. También la técnica de Observación permitirá observar la realidad académica de los estudiantes y docentes, así como también permitirá seguir obteniendo información necesaria a lo largo del desarrollo del presente proyecto.

También se empleará la metodología Cross-Industry Standard Process for Data Mining, CRIP – DM que es una guía para el desarrollo de proyectos enfocados a la Minería de

Datos. Esta metodología puede trabajar con cualquier herramienta para desarrollar el proyecto que esté disponible en el mercado aplicando así una característica adicional que es el de ser una metodología equitativa [54-55].

Es un método probado para orientar sus trabajos de minería de datos. Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas y como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos [56].

A continuación se describe cada una de las fases [57-59].

1. **Entendimiento del negocio.-** Esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio, para convertir este conocimiento en un problema de definición de minería de datos y un plan preliminar diseñado para alcanzar los objetivos.

2. **Entendimiento de los datos.-** Esta fase inicia con una colección inicial de datos y procede con actividades para familiarizarse con ellos, identificar problemas de calidad en los mismos, descubrir una primera idea de estos o detectar conjuntos interesantes que permitan formar hipótesis en la búsqueda de información escondida.

3. **Preparación de los datos.-** Cubre todas las actividades para construir la base final de datos (datos que serán el alimento de las herramientas de modelado) desde una base en bruto. Es preferible que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Estas tareas incluyen tabulación, documentación y selección de atributos, también como transformación y limpieza de datos para las herramientas de modelado.

4. **Modelado.-** Se seleccionan y aplican varias técnicas, y sus parámetros son calibrados a los valores óptimos. Por lo general hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por lo tanto será a menudo necesario devolverse a la fase de preparación de datos.

5. **Evaluación.-** Al llegar a esta fase se ha construido un modelo (o modelos) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la entrega final del modelo es importante evaluarlo más a fondo y revisar los pasos ejecutados para construirlo, de tal forma que este lo más cercano posible de alcanzar los objetivos del negocio. Un objetivo clave es determinar si hay algún evento importante del negocio que no haya sido considerado lo suficiente. Al final de esta fase, se debe tener una decisión sobre el uso de los resultados de minería de datos.

6. **Despliegue.-** La creación del modelo por lo general no es el final del proyecto. Incluso si el propósito del modelo es incrementar conocimiento sobre los datos, el conocimiento ganado necesitará ser organizado y presentado de una manera que el cliente lo pueda usar. A menudo implica aplicar modelos en vivo dentro del proceso de toma de decisiones de una organización, por ejemplo, en la personalización en tiempo real de las páginas web o la puntuación repetida en bases de datos de mercadeo. Sin embargo, dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso repetible de minería de datos a través de la empresa. En muchos casos es el cliente, no el analista de datos, quien realiza los pasos de despliegue. Sin embargo, incluso si el analista no carga con el esfuerzo de despliegue, es importante que el cliente entienda que acciones deben ser llevadas a cabo para hacer uso de los modelos creados.

H. CRONOGRAMA

El cronograma se detalla en la tabla 10 a continuación:

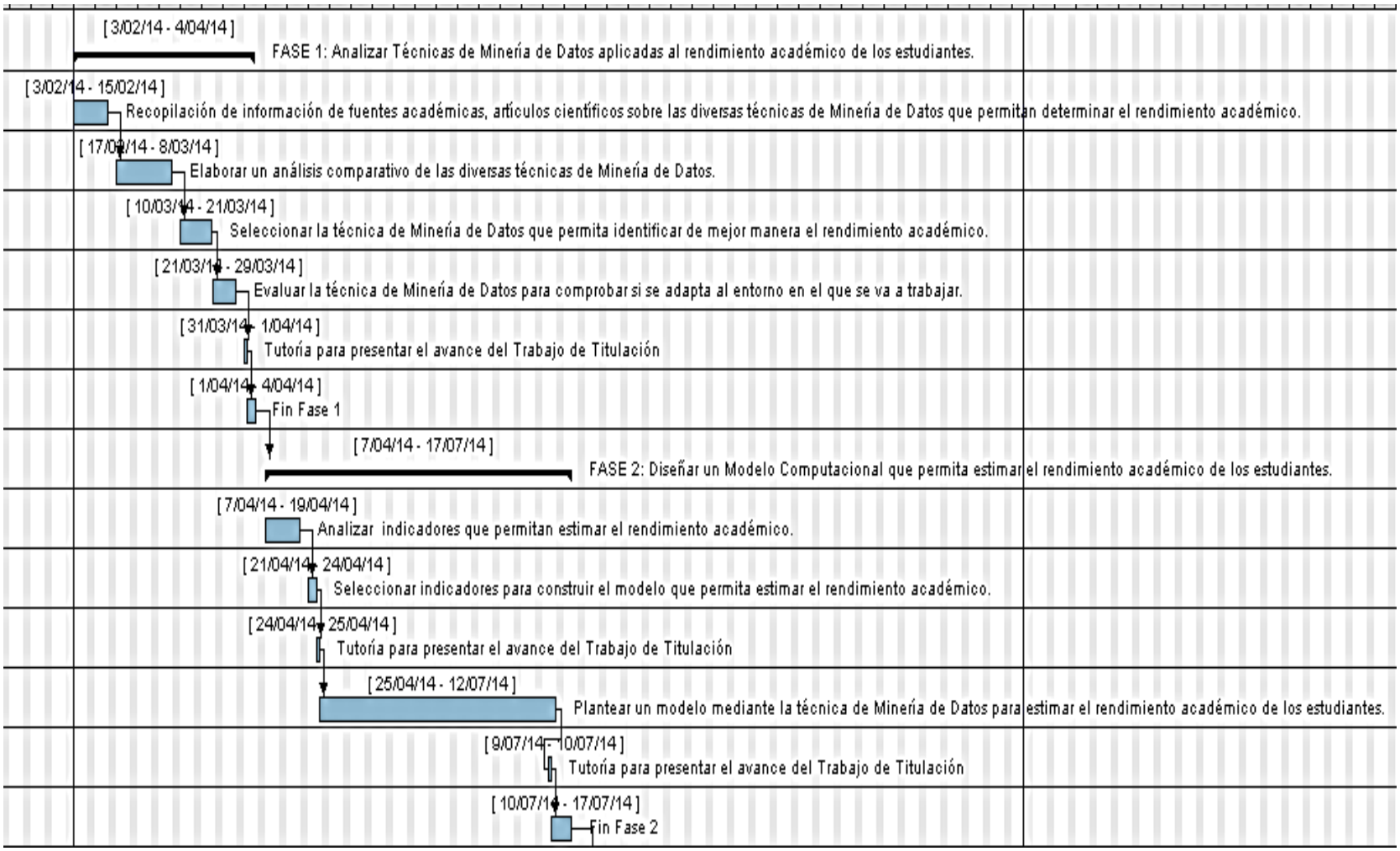
Tabla 10. Cronograma

ID	Nombre	Inicio	Fin
1	**FASE 1: Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.**	**03/02/14**	**04/04/14**
2	Recopilación de información de fuentes académicas, artículos científicos sobre las diversas técnicas de Minería de Datos que permitan determinar el rendimiento académico.	03/02/14	17/02/14
3	Elaborar un análisis comparativo de las diversas técnicas de Minería de Datos.	17/02/14	10/03/14
4	Seleccionar la técnica de Minería de Datos que permita identificar de mejor manera el rendimiento académico.	10/03/14	21/03/14
5	Evaluar la técnica de Minería de Datos para comprobar si se adapta al entorno en el que se va a trabajar.	21/03/14	29/03/14
6	Tutoría para presentar el avance del Trabajo de Titulación	31/03/14	01/04/14
7	Fin Fase 1	01/04/14	04/04/14

Diagrama de Gant de cada una de las fases del Trabajo de Titulación (ver figura 18).

8	**FASE 2: Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.**	**07/04/14**	**22/07/14**
9	Analizar indicadores que permitan estimar el rendimiento académico.	07/04/14	19/04/14
10	Seleccionar indicadores para construir el modelo que permita estimar el rendimiento académico.	21/04/14	24/04/14
11	Tutoría para presentar el avance del Trabajo de Titulación.	24/04/14	25/04/14
12	Plantear un modelo mediante la técnica de Minería de Datos para estimar el rendimiento académico de los estudiantes.	25/04/14	12/07/14
13	Tutoría para presentar el avance del Trabajo de Titulación	14/07/14	15/07/14
14	Fin Fase 2	15/07/14	22/07/14
15	**FASE 3: Implementar el Modelo Computacional sobre el rendimiento académico mediante una herramienta de Minería de Datos.**	**22/07/14**	**13/12/14**
16	Recopilación de información en fuentes académicas, artículos científicos sobre herramientas de Minería de Datos que permitan adaptar el modelo realizado.	22/07/14	06/08/14

17	Análisis comparativo de las diferentes herramientas de Minería de Datos que permitan adaptar el modelo realizado.	06/08/14	22/08/14
18	Selección de la mejor herramienta de Minería de Datos que permitan adaptar el modelo realizado.	22/08/14	03/09/14
19	Tutoría para presentar el avance del Trabajo de Titulación	03/09/14	04/09/14
20	Implementar el modelo en la herramienta de Minera de Datos seleccionada.	04/09/14	03/10/14
21	Evaluar el modelo computacional en un escenario real con datos académicos.	03/10/14	25/10/14
22	Demostrar la visualización de los resultados del rendimiento académico de los estudiantes mediante la herramienta de Minería de Datos.	27/10/14	11/11/14
23	Tutoría para presentar el avance del Trabajo de Titulación	11/11/14	12/11/14
24	Interpretar los resultados obtenidos por la Herramienta de Minería de Datos acerca del rendimiento académico de los estudiantes.	12/11/14	02/12/14
25	Culminación del Trabajo de Titulación	02/12/14	11/12/14
26	Fin Fase 3	11/12/14	12/12/14


[3/02/14 - 4/04/14]
FASE 1: Analizar Técnicas de Minería de Datos aplicadas al rendimiento académico de los estudiantes.
[3/02/14 - 15/02/14]
Recopilación de información de fuentes académicas, artículos científicos sobre las diversas técnicas de Minería de Datos que permitan determinar el rendimiento académico.
[17/02/14 - 8/03/14]
Elaborar un análisis comparativo de las diversas técnicas de Minería de Datos.
[10/03/14 - 21/03/14]
Seleccionar la técnica de Minería de Datos que permita identificar de mejor manera el rendimiento académico.
[21/03/14 - 29/03/14]
Evaluar la técnica de Minería de Datos para comprobar si se adapta al entorno en el que se va a trabajar.
[31/03/14 - 1/04/14]
Tutoría para presentar el avance del Trabajo de Titulación
[1/04/14 - 4/04/14]
Fin Fase 1
[7/04/14 - 17/07/14]
FASE 2: Diseñar un Modelo Computacional que permita estimar el rendimiento académico de los estudiantes.
[7/04/14 - 19/04/14]
Analizar indicadores que permitan estimar el rendimiento académico.
[21/04/14 - 24/04/14]
Seleccionar indicadores para construir el modelo que permita estimar el rendimiento académico.
[24/04/14 - 25/04/14]
Tutoría para presentar el avance del Trabajo de Titulación
[25/04/14 - 12/07/14]
Plantear un modelo mediante la técnica de Minería de Datos para estimar el rendimiento académico de los estudiantes.
[9/07/14 - 10/07/14]
Tutoría para presentar el avance del Trabajo de Titulación
[10/07/14 - 17/07/14]
Fin Fase 2

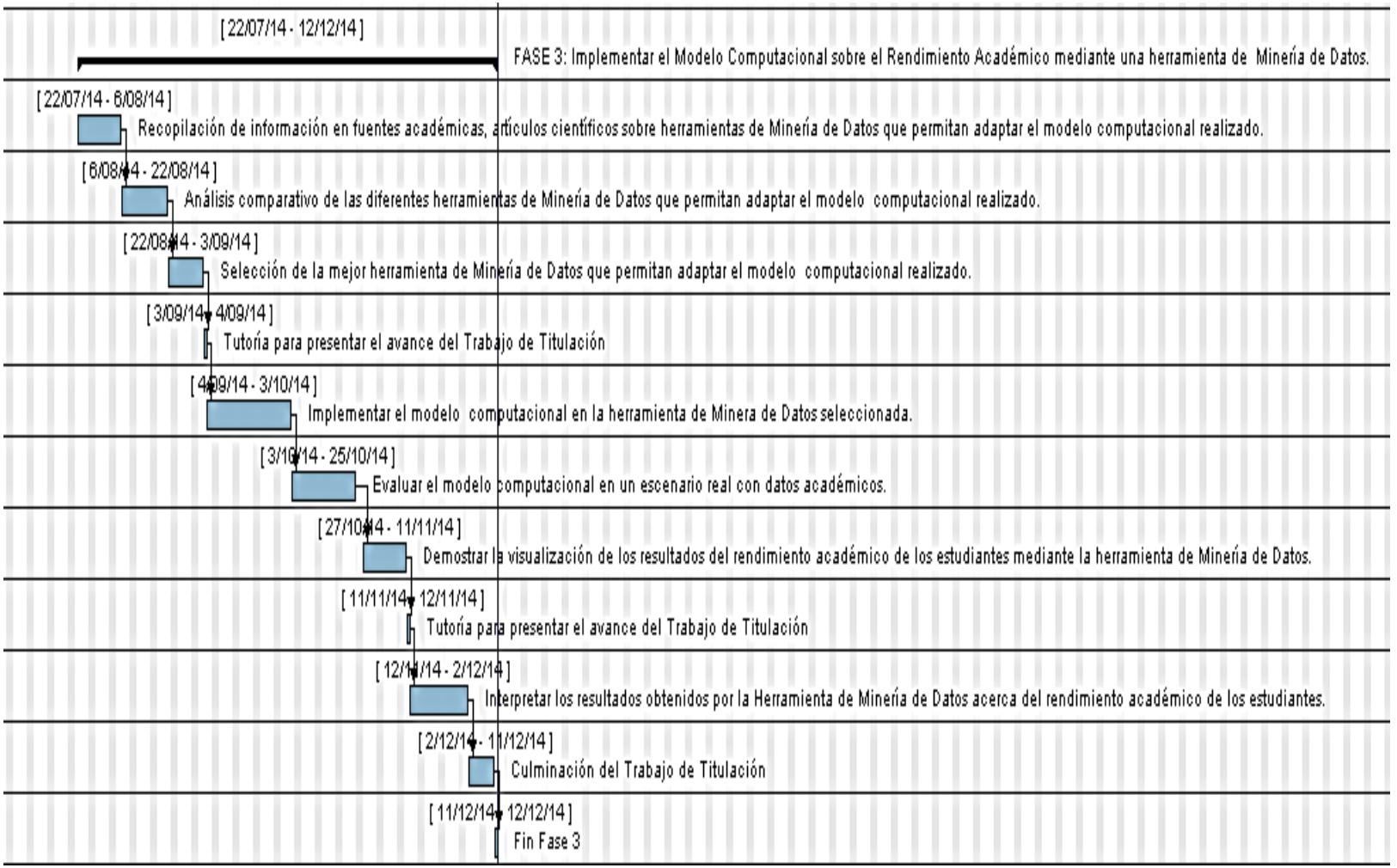


Figura 18: Cronograma

I. PRESUPUESTO Y FINANCIAMIENTO

A continuación se detalla el presupuesto que involucra el desarrollo del Trabajo de Titulación denominado “Estudio del Rendimiento Académico Aplicando Minería de Datos”.

Para llevar a cabo el Trabajo de Titulación es indispensable contar con Talento Humano que es quien desarrollará el Trabajo de Titulación y un tutor para que guíe en el proceso del mismo. En la siguiente tabla 11 se detalla los valores de acuerdo al Ministerio de Relaciones Laborales [60]:

TALENTO HUMANO			
TALENTO HUMANO	**HORAS**	**PRECIO / H ($)**	**V. TOTAL ($)**
Darwin Andrés Becerra Encarnación	1340	5,00	6.700,00
Tutor	120	8,00	960,00
		SUBTOTAL	**7.660,00**

Tabla 11. Talento Humano

Los Recursos Físicos como la computadora portátil se utilizará para ir desarrollando y documentando el Trabajo de Titulación, los mismos se detallan en la tabla 12 a continuación:

RECURSOS FÍSICOS					
RUBRO	**CANT**	**VALOR**	**TIEMPO / M**	**PRECIO / M ($)**	**V. TOTAL ($)**
Portátil Hp dv4-	1	1.100,00	11	20	220,00
Flash Memory 4GB	1	12,00	11	1,00	11,00
Disco Duro Samsung	2	100,00	11	8,00	88,00
Impresora	1	50,00	11	2.50	27,50
				SUBTOTAL	**346.50**

Tabla 12. Recursos Físicos

Los Recursos Software son necesarios para ir digitalizando el Trabajo de Titulación y la herramienta para implementar el modelo que se desarrollará y poder visualizar los resultados (ver Tabla 13).

RECURSOS SOFTWARE			
RUBRO	**CANT.**	**V. UNITARIO ($)**	**V. TOTAL ($)**
Editor de Texto Latex	1	0,00	0,00
Microsoft Proyect 2010	1	80,00	80,00
Paquete de Ofimática de Microsoft Office 2010 (licencia estudiantes)	1	140,00	140,00
Herramientas de Minería de Datos	1	0,00	0,00
		SUBTOTAL	**220,00**

Tabla 13. Recursos Software

Los Servicios como el internet son útiles para realizar consultas sobre casos de éxito y temas enfocados con el Trabajo de Titulación (ver tabla 14).

SERVICIOS				
RUBRO	**CANT. HORAS**	**V.**	**CANT./MES**	**V.TOTAL ($)**
Internet	240 H	0,60	10	1440,00
Llamadas	1 H	0,15	10	90,00
				1530,00

Tabla 14. Servicios

El Transporte son medios necesarios para poder trasladarse a la Universidad para recibir tutorías y además presentar avances del Trabajo de Titulación (ver tabla 15).

TRASPORTE				
RUBRO	**CANT. PASAJES/CARRERAS**	**V. UNITARIO ($)**	**CANT./MES**	**V.TOTAL ($)**
Bus	40 pasajes	0,25	10	100,00
Taxi	5 carreras	1,50	10	75,00
			SUBTOTAL	**175,00**

Tabla 15. Transporte

Los Recursos Materiales son fundamentales para evidenciar de manera física los avances del Trabajo de Titulación como se puede observar en la tabla 16.

RECURSOS MATERIALES			
RUBRO	**CANT.**	**V. UNITARIO ($)**	**V.TOTAL ($)**
Resma de papel	6	4,00	24,00
Cartuchos de Tinta	3	15,00	45,00
Perfiles	6	0,45	2,70
Copias	400	0.02	8,00
Carpetas	6	0,30	1,80
CD’s	7	0,60	4,20
		SUBTOTAL	**85,70**

Tabla 16. Recursos Materiales

El Presupuesto Total será financiado por el desarrollador del Trabajo de Titulación, a continuación se resume los gastos que involucra el mismo (observar tabla 17).

PRESUPUESTO TOTAL	
Talento Humano	**7660,00**
Recursos Físicos	**346,50**
Recursos Software	**140,00**
Servicios	**1530,00**
Transporte	**175,00**
Recursos Materiales	**85,70**
TOTAL	**9937,20**
IMPREVISTOS (10% DEL TOTAL)	**993,72**
TOTAL PRESUPUESTO + IMPREVISTOS	**10930,09**

Tabla 17. Presupuesto Total

J. **BIBLIOGRAFÍA**

[1]. Alvares Aldaco (2009), "Comportamiento de la Deserción y Reprobación en el Colegio de Bachilleres del Estado de Baja California: Caso Plantel Ensenada", X Congreso Nacional de Investigación Educativa. México, 2009.

[2]. F. Araque, C. Roldán, A. Salguero (2009), "Factors Influencing University Drop Out Rates", Computers & Education, vol. 53, pp. 563–574, 2009.

[3]. Carlos Márquez Vera, Cristóbal Romero Morales y Sebastián Ventura Soto, Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos, [En línea]: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf.

[4]. Ángel Cobo Ortega, Rocío Rocha Blanco y Yurlenis Álvarez Díaz (2011), Selección de Atributos predictivos del Rendimiento Académico de estudiante en un modelo de B-Learning, Revista Electrónica de Tecnología Educativa, [En línea]: http://edutec.rediris.es/Revelec2/Revelec37/pdf/Edutec-e_n37_Cobo_Rocha_Alvarez.pdf.

[5]. SALMON, G. (2000): E- moderating: The key to teaching and learning online. London, Kogan Page.

[6]. Isabel Cristina Montes Gutiérrez y Jeannette Lerner Matiz (2011), Rendimiento Académico de los estudiantes de pregrado de la Universidad EAFIT, Universidad

[7]. Porcel, Eduardo; Dapozo, Gladys; López, María V., Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios, Universidad Nacional del Nordeste - Departamento de Informática, [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/19846/Documento_completo.pdf?sequence=1.

[8]. Jairo Elías Gutiérrez, Descubrimiento de Conocimientos en la Base de Datos Académica de la Universidad Autónoma de Manizales Aplicando Redes Neuronales, Universidad Autónoma de Manizales - Facultad de Ingenierías, [En línea]: http://repositorio.autonoma.edu.co/jspui/bitstream/11182/345/1/MSC-JGutierrezInforme(V2).pdf.

[9]. Lerner, J., Vargas, A. et. al. (2004). Los Procesos Pedagógicos y sus Vicisitudes. Reflexiones y aproximaciones. Texto inédito. Biblioteca Luis Echavarría Villegas, Universidad EAFIT, Medellín.

[10]. Bernardo Gargallo López, Cruz Pérez Pérez, Beatriz Serra Carbonell, Francesc Sánchez I Peris e Inmaculada Ros Ros (2007). Actitudes ante el aprendizaje y riesgo académico en estudiantes universitarios. Revista Iberoamericana de Educación,

Universidad de Valencia. [En línea]: http://www.rieoei.org/investigacion/1537Gargallo.pdf.

[11]. Eduardo Vélez, Ernesto Schiefelbein y Jorge Valenzuela (1994). Factores que Afectan el Rendimiento Académico en la Educación Primaria (Revisión de la Literatura de América Latina y el Caribe), [En línea]: http://www.oei.es/calidad2/Velezd.PDF.

[12]. Alejandro Gaviria y Jorge Hugo Barrientos (2001). Determinantes de la calidad de la educación en Colombia. Archivos de economía, DNP, No 159., [En línea]: https://www.dnp.gov.co/portals/0/archivos/documentos/dee/archivos_economia/159.pdf.

[13]. Mella, O., y Ortiz, I. (1999). Rendimiento Escolar. Influencias diferenciales de factores externos. Revista latinoamericana de estudios educativos, Vol. XXIX, Núm. 1, pp. 69-92.

[14]. "Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education". HEINER, CECILY, HEFFERNAN, NEIL y BARNES, TIFFANY. Marina del Rey CA. USA: s.n., 2007.

[15]. Sebastián Ventura Soto. "Minería de Datos en Sistemas Educativos". Universidad de Córdoba - Departamento de Informática y Análisis Numérico, [En línea] http://sci2s.ugr.es/docencia/doctoM6/EducationalDataMining.pdf.

[16]. Karina Eckert y Roberto Suénaga, Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD, Universidad Gastón Dachary, [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/27103/Documento_completo.pdf?sequence=1.

[17]. R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica, Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas

de minería de datos, Universidad Politécnica de Valencia - Dpto. de Estadística e I.O. Aplicadas y Calidad, Dpto. de Informática de Sistemas y Computadores, Dpto. de Sistemas Informáticos y Computación, Dpto. de Física Aplicada, [En línea]: http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf.

[18]. Lerner, J., Vargas, A. et. al. (2004). Los Procesos Pedagógicos y sus Vicisitudes. Reflexiones y aproximaciones. Texto inédito. Biblioteca Luis Echavarría Villegas, Universidad EAFIT, Medellín.

[19]. Dapozo, Gladys; Porcel, Eduardo; López, María V.; Bogado, Verónica; Bargiela Roberto, Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de sistemas de la FACENA-UNNE, Universidad Nacional del Nordeste - Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura, [En línea]: http://sedici.unlp.edu.ar/bitstream/handle/10915/20797/Documento_completo.pdf?sequence=1

[20]. Osvaldo M. Spositto, Martín E. Etcheverry, Hugo L. Ryckeboer, Julio Bossero, Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil, Universidad Nacional de La Matanza - Departamento de Ingeniería e Investigaciones Tecnológicas, [En línea]: http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

[21]. Rendimiento Académico, Universidad Francisco Gavidia – Facultad de Ingeniería y Arquitectura, Ingeniería en Ciencias de la Computación, [En línea]: http://wwwisis.ufg.edu.sv/wwwisis/documentos/TE/371.262-B634f/371.262-B634f-CAPITULO%20II.pdf

[22]. Hábitos de estudio y rendimiento académico, Universidad Francisco Gavidia – Facultad de Ingeniería y Arquitectura, Ingeniería en Ciencias de la Computación, [En línea]: http://wwwisis.ufg.edu.sv/wwwisis/documentos/TE/371.302%2081-G633h/371.302%2081-G633h-Capitulo%20II.pdf

[23]. Definición de Minería de Datos, [En línea]: http://www.oocities.org/es/mineria.datos/definicion_tecnicas_mineria_datos.pdf

[24]. Edgar Acuna, Minería de Datos, Universidad de Puerto Rico-Mayaguez - Departamento de Ciencias Matemáticas, [En línea]: http://academic.uprm.edu/~eacuna/dm1.pdf

[25]. Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda, Minería de Datos, Universidad Carlos III de Madrid – Ingeniería Telemática, [En línea]: http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/07-mineria-de-datos

[26]. José Ignacio González Gómez, Generalidades de la Minería de Datos, Universidad de La Laguna – Departamento de Economía Financiera y Contabilidad, [En línea]: http://www.ecofin.ull.es/users/jggomez/D%20Bdr_Erp/6%20Mineria/Mineria.pdf

[27]. Sofía J. Vallejos, Minería de Datos, Universidad Nacional del Nordeste - Facultad de Ciencias Exactas, Naturales y Agrimensura, [En línea]: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf

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

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

[30]. Corso, Cynthia Lorena, Ing. Gibellini Fabián, Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios, Universidad Tecnológica Nacional - Departamento de Ingeniería en Sistemas de Información - Laboratorio de Sistemas de Información, [En línea]: http://www.41jaiio.org.ar/sites/default/files/16_JSL_2012.pdf

[31]. José Manuel Molina López y Jesús García Herrero, TÉCNICAS DE ANÁLISIS DE DATOS, Instituto Tecnológico Superior de Calkiní en el Estado de Campeche, [En línea]: http://www.itescam.edu.mx/principal/sylabus/fpdb/recursos/r94663.PDF

[32]. Agustín José Calleja Gómez, MINERÍA DE DATOS CON WEKA PARA LA PREDICCIÓN DEL PRECIO DE AUTOMÓVILES DE SEGUNDA MANO, UNIVERSIDAD POLITÉCNICA DE VALENCIA - ESCUELA TÉCNICA SUPERIOR DE INFORMÁTICA APLICADA, [En línea]: http://riunet.upv.es/bitstream/handle/10251/10097/PFC_DSIC-80_Agust%C3%ADnCalleja.pdf

[33]. José Antonio García Bermúdez y Ángela María Acevedo Ramírez, ANÁLISIS PARA PREDICCIÓN DE VENTAS UTILIZANDO MINERÍA DE DATOS EN ALMACENES DE VENTAS DE GRANDES SUPERFICIES, UNIVERSIDAD TECNOLÓGICA DE PEREIRA - FACULTAD DE INGENIERIAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN - INGENIERÍA DE SISTEMAS Y COMPUTACIÓN, [En línea]: http://repositorio.utp.edu.co/dspace/bitstream/11059/1339/1/006312G216.pdf

[34]. Braulio José Solano Rojas, Introducción a la Minería de Datos, Universidad de Costa Rica

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

[36]. J.L. Cubero, F. Berzal, F. Herrera, FUNDAMENTOS DE MINERÍA DE DATOS, Universidad de Granada - Dpto. Ciencias de la Computación e I.A., [En línea]: http://sci2s.ugr.es/docencia/m1/Preprocesamiento-Weka-MD.pdf

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

[38]. Francisco José García González, Aplicación de Técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Universidad de Granada

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

[40]. José Hernández Orallo y César Ferri Ramírez, Introducción a WEKA, Universidad Politécnica de Valencia - Departamento de Sistemas Informáticos y Computación, [En línea]: http://users.dsic.upv.es/~jorallo/docent/doctorat/weka.pdf

[41]. Abdelmalik Moujahid e Iñaki Inza, Manual de prácticas de minería de datos usando el software WEKA, UNIVERSITY OF THE BASQUE COUNTRY - Department of Computer Science and Artificial Intelligence, [En línea]: https://addi.ehu.es/bitstream/10810/4627/1/tr10-1.pdf

[42]. Luis P. Guerra Velasco, Primeros pasos con Knime, Universidad Politécnica de Madrid - Departamento de Arquitectura y Tecnología de Sistemas Informáticos - Faculta de Informática de Madrid, [En línea]: https://laurel.datsi.fi.upm.es/_media/docencia/cursos/inap/ejemplodm.pdf

[43]. Cynthia Corso y Fabián Gibellini, Uso de Herramienta Libre para la generación de reglas de asociación, facilitando la gestión eficiente de inventarios e incidentes, Universidad Tecnológica Nacional – Departamento Ingeniería en Sistemas de Información - Laboratorio de Sistemas de Información, [En línea]: https://sl.linti.unlp.edu.ar/wpcontent/uploads/2012/08/GeneracionReglasAsociacion.pdf

[44]. ORACLE, Oracle Data Mining, [En línea]: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html

[45]. Práctica de laboratorio de aprendizaje inductivo, Universidad Politécnica de Cataluña – Facultad de Informática, [En línea]: http://www.lsi.upc.edu/~bejar/apren/lab/apind13141q.pdf

[46]. Juan Carlos Díez, Iván Martín y Manuel Aranda, RAPIDMINER, Grupo de Investigación – Departamento de Lenguajes y Sistemas Informáticos [En línea]: http://www.kybele.etsii.urjc.es/docencia/SI_GII_M/2012-2013/Material/[Expo]Presentacion%20RAPID%20MINER.pdf

[47]. Yanchang Zhao (2013), R and Data Mining: Examples and Case Studies, Informáticos [En línea]: http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

[48]. Sergio Hernández, (2012), Minería de Datos de gran escala usando R, Universidad Católica del Maule – Laboratorio de Procesamiento de Información Geoespacial.

[49]. Francisco José García González (2013), Aplicación de Técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Universidad de Granada - Master Universitario en Estadística Aplicada.

[50]. The R User Conference, University of Castilla - La Mancha, [En línea]: http://www.edii.uclm.es/~useR-2013/docs/useR2013_abstract_booklet.pdf

[51]. Daniel Tuttle Ospina, Minería de Datos en el paquete Rattle del Lenguaje R, Universidad Nacional de Colombia – Escuela de Sistemas, [En línea]: http://tecaprendizajeest.wikispaces.com/file/view/Miner%C3%ADa+de+datos+con+eL+paquete+rattle+-+tutorial.pdf

[52]. FACTOMINER, Porque utilizar FactoMineR, [En línea]: http://factominer.free.fr/

[53]. FACTOMINER, Descripción de la interfaz gráfica de usuario, [En línea]: http://factominer.free.fr/interface/images/principal-component-analysis-activar.PNG

[54]. Metodología de Aplicación del Data Mining (DM), Universidad Politécnica Salesiana, [En línea]: http://www.dspace.ups.edu.ec/bitstream/123456789/47/10/Capitulo4.pdf

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

[56]. Manual CRISP-DM de IBM SPSS Modeler, IBM, [En línea]: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf

[57]. Hernando Camargo y Mario Silva, Dos caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP, Universidad el Bosque – Ingeniería en Sistemas, [En línea]: http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

[58]. Juan Ángel Vanrell, UN MODELO DE PROCESOS PARA PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN, Universidad Tecnológica Nacional - Ingeniería en Sistemas de Información, [En línea]: http://www.unla.edu.ar/sistemas/gisi/tesis/vanrell-tesisdemagister.pdf

[59]. José Alberto Gallardo Arancibia, Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM), Universidad Politécnica de Madrid - Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software - Facultad de Informática, [En línea]: http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

[60]. Ministerio de Relaciones Laborales, [En línea]: http://www.relacioneslaborales.gob.ec/

K. ANEXOS

Anexo 1: Solicitud para pedir Información académica de los estudiantes del AEIRNNR de la UNL con el fin de poder llevar a cabo el Trabajo de Titulación.

Loja, 11 de Diciembre de 2013

Sr. Ing.
Milton Ricardo Palacios Morocho
DIRECTOR DEL DEPARTAMENTO DE TELECOMUNICACIONES E INFORMACIÓN - UTI

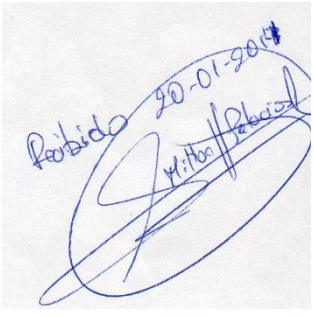
Ciudad.

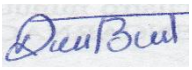
De mis consideraciones:

Yo, Darwin Andrés Becerra Encarnación, portador de la cédula de ciudadanía 1104645005, egresado de la Carrera Ingeniería en Sistemas me dirijo a usted muy comedidamente para solicitarle se me proporcione información académica de los estudiantes del ÁREA DE LA ENERGÍA, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES - AEIRNNR, debido que son de suma importancia para poder desarrollar el Proyecto Fin de Carrera denominado “Estudio del Rendimiento Académico aplicando Técnicas de Minería de Datos”.

Con la certeza de ser atendido favorablemente me suscribo de usted con los más sinceros agradecimientos.

Atentamente:





……..
Darwin Andrés Becerra Encarnación
Egresado
C.I 1104645005

Anexo 12: Certificado de Traducción

Lic.

Carlos Eduardo Zurita Valencia

LICENCIADO EN CIENCIAS DE LA EDUCACIÓN EN LA ESPECIALIDAD DEL IDIOMA INGLÉS

CERTIFICA:

Que la traducción del resumen del Trabajo de Titulación cuyo tema es **"Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos"** es fiel traducción, por lo que su contenido puede ser interpretado de forma correcta.

Atentamente:

…….

Carlos Eduardo Zurita Valencia

Anexo 13: Licencia Creative Commons

"Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos" por Darwin Andrés Becerra Encarnación se distribuye bajo una Licencia Creative Commons Atribución 4.0 Internacional.

Figura 35: Licencia Creative Commons