UNIVERSIDAD JOSÉ CARLOS MARIÁTEGUI

VICERRECTORADO DE INVESTIGACIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

TRABAJO DE SUFICIENCIA PROFESIONAL

MINERÍA DE DATOS AVANZADOS

PRESENTADO POR

BACHILLER JHONATHAN AMERICO QUISPE SUCAPUCA

ASESOR:

ING. JULIAN MANUEL FLORES MANCHEGO

PARA OPTAR TÍTULO PROFESIONAL DE

INGENIERO DE SISTEMAS E INFORMÁTICA

MOQUEGUA – PERÚ

2019

CONTENIDO

CAPÍTULO IV

CONCLUSIONES Y RECOMENDACIONES

CONTENIDO DE TABLAS

Pág.

CONTENIDO DE FIGURAS

Pág.

CONTENIDO DE ECUACIONES

Pág.

RESUMEN

El proyecto buscaba reglas que a través de técnicas de Minería de datos se pueda aplicar a la información de los alumnos de la universidad, se utilizó un proceso de "Knowledge Discovery in Database" a los datos, con el objetivo de encontrar conocimiento útil acerca del desempeño de los estudiantes. A través de distintos factores, se logró determinar los atributos con mayor efecto en el rendimiento de los estudiantes, mediante reglas que de manera precisa indiquen las variables influyentes. Se aplicaron enfoques y técnicas de Minería de Datos que permiten el desarrollo de distintos modelos de predicción. Las etapas desarrolladas en los modelos han sido resueltas por el software llamado "WEKA". Según el modelo, se logró conseguir distintos algoritmos y reglas, dando así la interpretación de los resultados obtenidos. Los resultados son de importancia para influenciar el rendimiento académico de los alumnos y el desempeño del curso, ya que influye en la toma de decisiones acertadas, las cuales serán tomadas en el momento preciso, el conocimiento obtenido puede hacer la diferencia para que un alumno apruebe o desapruebe el curso.

Palabras Clave: Minería de Datos, Conocimiento, Descubrimiento, Base de Datos, Weka, Técnicas.

ABSTRACT

The project sought rules through data mining techniques applied to the information of the university students, using a process of "Knowledge Discovery in Database" to the data, with the objective of finding useful knowledge about the performance of the students. Through different factors, it was possible to determine the attributes with the greatest effect on student performance, by means of rules that precisely indicate the influential variables. Data Mining approaches and techniques were applied that allow the development of different prediction models. The stages developed in the models have been solved by the software called "WEKA". Depending on the model, different algorithms and rules were achieved, thus giving the interpretation of the results obtained. The results are of importance to influence the academic performance of the students and the performance of the course, since it influences the right decisions, which will be taken at the right time, the knowledge obtained can make the difference for a student to pass or fail the course.

Keywords: Data Mining, Knowledge, Discovery, Database, Weka, Techniques.

CAPÍTULO I

INTRODUCCIÓN

Desde hace muchos años, se viene desarrollando la aplicación de avances tecnológicos y sistemas de información, los cuales, se han ido apoderando de las empresas a nivel mundial, por lo que extraer conocimiento a partir de la información ha transformado la industria, siendo ahora el activo más importante de las empresas.

En el presente trabajo, se busca extraer conocimiento y aportar al desempeño de estudiantes y docentes, con la finalidad de mejorar su nivel académico y aprendizaje de estudiantes.

Las diferentes etapas del trabajo explican el proceso que se ha desarrollado, en la primera etapa, va relacionado con el marco teórico y el conocimiento mínimo de los conceptos, para que en los capítulos siguientes se entienda; en la segunda etapa se argumenta la elección del enfoque, técnicas y algoritmos que se van a usar, se documenta todo el proceso de "Knowledge Discovery in Database", explicando la obtención de variables, el proceso que se les ha aplicado, la transformación, la minería de datos y finalmente la interpretación de los resultados, en la tercera etapa se obtiene los resultados de dos experimentos donde se ha aplicado diversas técnicas de minería de datos.

CAPÍTULO II

OBJETIVOS

2.1. Objetivo General

Adquirir conocimiento útil, a partir del comportamiento y calificaciones de los estudiantes que aprueban o desaprueban el curso de informática, aplicando un proceso "Knowledge Discovery in Database, KDD" a la información.

2.2. Objetivos Específicos

a) Definir y aplicar un proceso de "Minería de Datos" a la data recopilada relacionada con el rendimiento académico de los estudiantes.

b) Definir las técnicas y herramientas de "Minería de Datos" más pertinentes a los objetivos establecidos.

c) Establecer variables que determinen el comportamiento, en términos de la calificación final.

d) Interpretar y demostrar la utilidad del conocimiento obtenido mediante experimentos.

CAPÍTULO III
DESARROLLO DEL TEMA

3.1. Marco teórico

3.1.1. El conocimiento en base de datos.

Una expresión vinculada con la minera de datos, es la extracción o descubrimiento del conocimiento en base de dato de gran tamaño, (Knowledge Discovery in Database, KDD). En muchas oportunidades ambas, se han usado del mismo modo, aunque actualmente son muy diferentes. Así Knowledge Discovery in Database, se usa para referirse a un proceso que se compone por etapas, siendo la minería de datos una fase del proceso Knowledge Discovery in Database (Francisco, 2015).

El proceso del descubrimiento del conocimiento en base de datos o Knowledge Discovery in Database, es un método tradicional de transformar datos en conocimiento, desde el año 1989, cuando la necesidad de extraer conocimiento requiere un método de estudio inteligente de datos. Muchos describen la minería de datos de distintas formas, algunos autores lo definen con cinto etapas, otros con menos o muchas etapas más, teniendo en cuenta lo mencionado explicaremos la definición de distintos autores, desde la planificación y comprensión de los datos llegando así a la interpretación y explotación de los resultados.

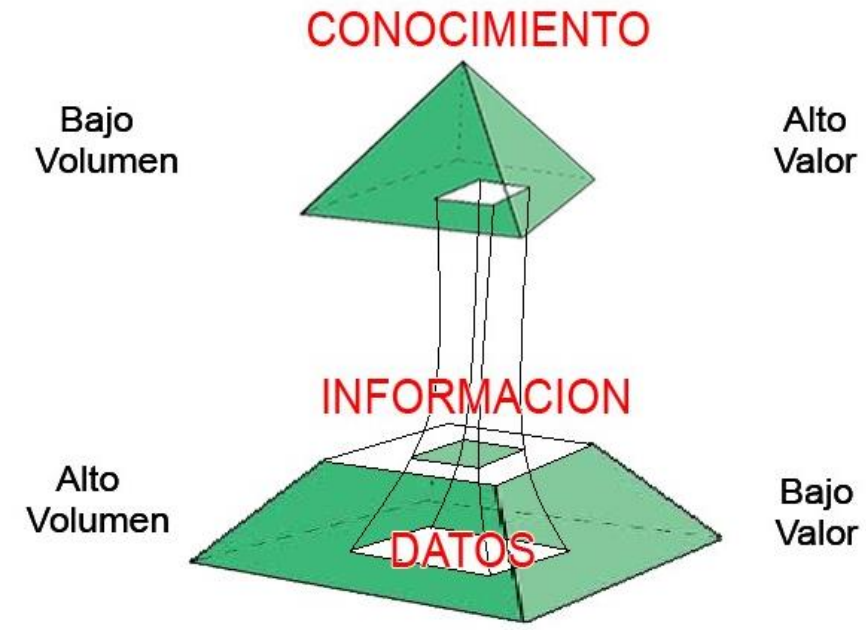

Figura 1. Jerarquía del Conocimiento.

Fuente: Molina, 2002.

3.1.1.1. El proceso KDD según los autores Fayyad, Piatetsky-Shapiro y Smyth.

a. Definición.

Fayyad, Piatetsky-Shapiro, & Smyth, (1996). definen al termino Knowledge Discovery in Database, como "El proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensible en los datos"

b. Propiedades de la definición descrita anteriormente.

Valido nos indica que los patrones deben ser precisos con los nuevos datos, y no solo deben ser de utilidad para los que se han usado en su obtención. Nuevos que nos ayude a encontrar un aporte significativo para el sistema y el usuario. Potencialmente útil los datos deben apuntar a acciones que informen y sirvan de beneficio para el usuario. Comprensible la extracción de patrones no

comprensibles, va a dificultar o imposibilitar su revisión, comprobación, interpretación y su determinación para la toma de decisiones. De hecho, una información que no es comprensible, no proporciona conocimiento útil.

También, estos profesores, definen los conceptos nombrados anteriormente en una forma que puede explicitarse matemáticamente, y que también llevan a una definición efectiva del conocimiento.

Datos un grupo de hechos "x". Patrón una expresión "a", en algún tipo de lenguaje "b" que detalla un subconjunto de los datos "x", el cual tiene que ser original sencilla y simple como la enumeración de los hechos que componen "x". Validez esta propiedad equivale a una función "c (a, x)" que atribuyen una calificación, es decir, un número al patrón. Novedad una función "n (a, x)" que realiza una validación si el patrón no es una recomposición de patrones ya identificados o falsos en caso contrario. Utilidad se representa por una función que evalúa y valida la utilidad "u (a, x)". Los patrones nos permiten realizar o decidir una acción.

Comprensibilidad de Fayyad sugiere como medida cuantitativa la sencillez del patrón, se representa nuevamente con una función "s (a, x)" que asigna un valor. Todo esto lleva como resultado a la fundamental "medida de interés" o "Interestingness Measures", conceptos que son primordiales para el resultado.

Entonces aparece un nuevo término, el cual es definido como una mezcla de Efectividad, Innovación, Provecho y Comprensibilidad que permite valorar y ordenar los patrones. Interestingness se representa por la función "i (a, c, x, n, u, s)"

Al haber intervención humana en el concepto mencionado previamente, se puede decir que la medida de interés es elemental para realizar la definición del conocimiento final. Conocimiento: Un patrón "x" se le atribuye conocimiento si su medida del interés "i" supera hasta cierto punto su origen "u" definido por el mismo usuario.

El conocimiento, lo forman aquellos patrones que se aprenden a detectar y que además se guardan, ya que estos se aplican a los nuevos datos, por tanto, se puede predecir el comportamiento de los sucesos o fenómenos que nos rodean.

Fayyad, Piatetsky-Shapiro & Smyth, (1996) definen. "La mayoría de los trabajos previos en Knowledge Discovery in Database, se centraban en la etapa de Minería de Datos. Sin embargo, los otros pasos son de considerable importancia para el éxito de las aplicaciones de Knowledge Discovery in Database en la práctica". Señalando claramente lo importante de incorporar en esta metodología el pre-procesamiento de la información. Estos autores defines este proceso con cinco etapas o pasos, las cuales son las siguientes: Selección, Pre-procesamiento, Transformación, Minería de Datos e Interpretación/Evaluación.

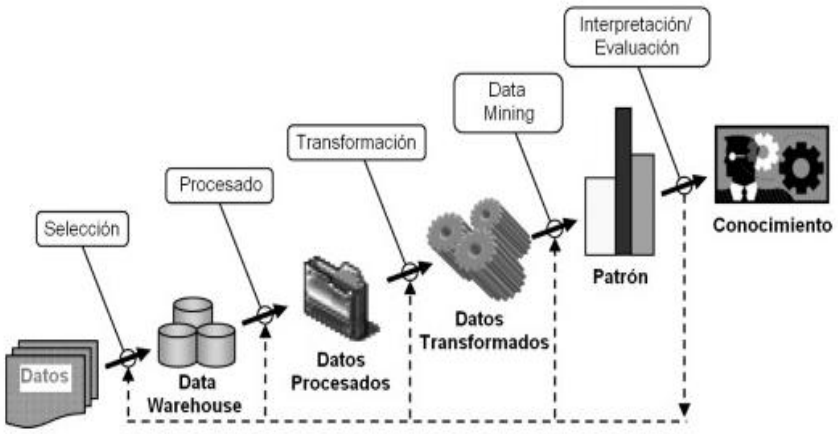

Figura 2. Etapas del Proceso Knowledge Discovery in Database.

Fuente: Fayyad, Piatetsky-Shapiro & Smyth, 1996.

Estos pasos aplicados de una manera repetitiva y participativa, aseguran que el conocimiento que se obtiene sea útil. La finalidad es encontrar patrones que tienen correlación en los datos y puedan ser usados, para realizar predicciones válidas acerca del tema.

c. Etapas del proceso KDD según Fayyad, Piatetsky-Shapiro y Smyth.

A continuación, se explican cada una de las cinco etapas que constituyen el proceso Knowledge Discovery in Database, según estos autores:

Selección de datos: Además de entender el problema y definir de manera exacta los objetivos, se aproxima al paso siguiente, seleccionar los datos, en el cual se elige un grupo de datos los cuales van a ser el objetivo para realizar un análisis y del mismo modo unificar los datos.

Pre-procesamiento de los datos: El objetivo de esta etapa es garantizar la calidad de la información que se va analizar, de esto depende la calidad del conocimiento que luego descubriremos, preparando y limpiando los datos extraídos de diferentes fuentes de datos en una forma transportable. En este paso se incorporan múltiples tareas, tales como, el filtrado de individuos atípicos, la eliminación de ruido, el uso de valores nulos o vacíos, los cuales corresponden a la normalización de los datos. Obteniendo como resultado una estructura de los datos direccionados para su transformación.

En efecto, esta etapa muchas veces es descuidada, pero, sin embargo, es fundamental para la obtención del resultado valido, el cual se da en cantidades grandes de datos que son recopilados por medio de distintos métodos automáticos.

Transformación de los datos: En esta fase consiste en alterar o modificar la estructura de los datos, a una estructura apropiada, con el fin de analizarlos de manera fácil y comprensible. Eso abarca la transformación del esquema inicial de los datos a otros esquemas, la reducción de dimensiones para llevar adelante el trabajo con un número reducido de variables, la eliminación de columnas que varían juntas (agregación), o técnicas más sofisticadas (clustering o análisis de componentes esenciales).

Un buen conocimiento del problema y además una buena intuición, pueden ayudar a determinar un resultado y fracasos ante del descubrimiento del conocimiento. Consolidando los datos de una forma necesaria para la siguiente etapa.

Minería de datos: Es el proceso fundamental donde se aplican distintos métodos especiales, con el fin de obtener ciertos patrones nuevos, válidos, distintos, potencialmente beneficiosos y accesible que están ocultos en los datos. Esta es la etapa de "descubrimiento" dentro del proceso Knowledge Discovery in Database, paso consistente en el uso de determinados algoritmos que producen patrones a partir de la información pre-procesados.

Se ha de seleccionar una técnica de modelado acorde al problema, teniendo en cuenta el objetivo, cabe destacar, que todas las técnicas tienen un conjunto de indicadores que deciden las características del patrón a generar. La selección de los indicadores es un proceso repetitivo y se rige solamente en los resultados obtenidos siempre que siga los pasos adecuados.

Interpretación/Evaluación: Es la última etapa del proceso Knowledge Discovery in Database, se identifican los patrones obtenidos que de verdad son interesantes y luego se estima y analiza el conocimiento extraído en la fase anterior contemplando tres criterios importantes precisión, claridad, e interés.

De todo lo dicho por estos autores, concluyen en que el conocimiento es muy útil e importante dentro de lo que se hace a diario en esta vida, además se disponen de diferentes maneras de aproximarse a él y obtener reglas interesantes para distintos campos.

3.1.1.2. El proceso KDD según los autores Han y Kamber.

a. Definición.

La obtención del conocimiento viene relacionado principalmente con el proceso Knowledge Discovery in Database, que según los autores (Han & Kamber, 2001) se refiere al "Proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información".

Básicamente, para ellos, la extracción del conocimiento de los datos está compuesto por cuatro pasos, que son limpieza e integración (recuperar en la base de datos, los datos relevantes para el análisis), selección y transformación (preparar los datos), minería de datos (elaborar modelos descriptivos/predictivos) y la evaluación del ejemplo (encontrar los modelos descriptivos/predictivos que de mejor manera tengan la solución al problema).

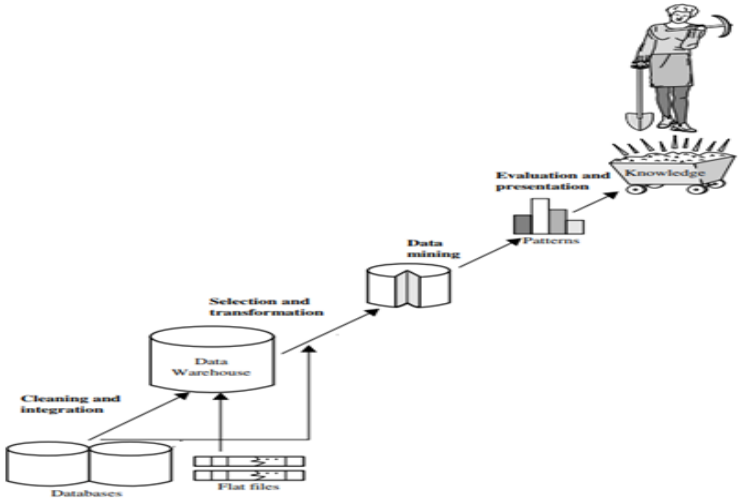

Figura 3. *Fases del Proceso KDD*

Fuente: Han & Kamber, 2001.

b. Etapas del proceso KDD según Han y Kamber.

A continuación, se explican cada una de las cuatro etapas que forman el proceso Knowledge Discovery in Database, según estos autores.

-Limpieza e integración de los datos.

Es el desarrollo que parte del tratamiento de los datos ruidosos, erróneos, faltantes o irrelevantes, y la incorporación de distintas fuentes de datos en una fuente única para la próxima etapa.

Se limpian los valores nulos mediante técnicas, tales como. Ignorar la lista ordenada: se aplica cuando el valor de la etiqueta no existe, ignorando toda una fila o conjunto del registro. Llenar el valor nulo: Se usa cuando la cantidad de datos es pequeña. Usar una constante única para llenar el valor faltante: Acá se reemplaza el valor vacío o nulo por alguna constante. Usar el valor más probable: Se puede conseguir un valor para predecir valores nulos.

Se disminuye el ruido y la inconsistencia en los datos usando distintos métodos: Suavización por media: Cada valor en conjunto de datos es intercambiado por el valor promedio del conjunto de datos. Suavización por mediana: Cada valor en conjunto de datos es intercambiado por la mediana del conjunto de datos. Suavización por límites: Cada valor en conjunto de datos es reemplazado por los valores límites, obviando los valores límites de cada conjunto de datos.

La normalización de datos en particular, se usa para tratar la inconsistencia en los datos, verificando las violaciones en las dependencias funcionales entre atributos y valores ilógicos en los mismos. Se integran los datos de diversas fuentes, combinándose en un almacén único de datos.

- *Selección y transformación de los datos.*

En esta etapa se extraen los datos destacados, seleccionando las variables más fundamentales en el problema. La selección de datos puede ser de forma horizontal y se realiza un muestreo de los datos, habiendo cuatro tipos de muestreo (aleatorio simple, aleatorio estratificado, de grupos y exhaustivo). También puede haber una selección de los datos de forma vertical, seleccionando los atributos más sobresalientes en base a algunos criterios. Otra selección, es la de eliminación de llaves elegidas, las mismas que son variables de códigos de información, nombre y apellidos, teléfonos, etc.

Luego se transforman los datos, consolidando los datos en una forma adecuada para luego ser introducido en el algoritmo de minería de datos. Es una labor de la transformación, es la construcción de nuevos atributos, y en los casos en que los atributos no contribuyan con suficiente poder predictivo.

Dentro de la transformación de los datos, está la discretización de los datos, es decir, convertir valores numéricos en atributos continuos o discretos, todo esto con el fin de hacer cambios en los diferentes tipos de los datos, para ayudar con el uso de técnicas que necesiten diferentes datos específicos.

- *Minería de datos.*

Es la etapa más característica del proceso KDD, y es por eso, que en varias ocasiones se tiende a utilizar esta fase para realizar el nombramiento de todo el proceso. La finalidad es generar conocimiento nuevo que permita utilizar el usuario. Se construye una forma establecida en los datos previamente extraídos. Luego de obtener los patrones se pueden usarse para obtener predicciones.

La etapa de Data Mining se divide, a su vez, en otros tres pasos, estas decisiones se toman antes de iniciar el proceso. Lograr determinar el tipo de trabajo o enfoque de Data Mining es apropiada para el caso. Elegir el tipo de modelo y técnica para realizar la tarea. Elegir el algoritmo correcto de minería de datos para que resuelva la tarea y obtener el tipo de modelo que andamos buscando.

- *Interpretación/Evaluación.*

En la etapa de interpretación y evaluación, se inicia con la validación del modelo obtenido con la finalidad de lograr que sea preciso para el proyecto, verificando que los resultados arrojan sean beneficiosos y válidos. Si el modelo no logra los objetivos esperados, se debe volver a modificar algunas de las actividades anteriores y volver a repetir el proceso o parte de él hasta que logremos el resultante esperado, con lo cual podremos originar un nuevo modelo.

Debido a la existencia de voluminosas bases de datos conteniendo gran cantidad de estos mismo, es que en las organizaciones las decisiones importantes se toman de acuerdo a los que creen y a su experiencia, más que considerando la rica información almacenada, es por esto que mediante el proceso Knowledge Discovery in Database, se intenta solucionar dicha situación con el fin de obtener información útil y valiosa.

3.1.2. Data Mining o Minería de datos.

3.1.2.1. Introducción e Historia.

La idea de Data Mining o minería de datos (en español) no se inicia recién. Debido a que desde los años sesenta, muchos estadísticos manejaban términos como "Data fishing", "Data Mining" o "Data Archeology" con la idea de hallar relaciones o patrones. A comienzos de los años 80, Rakesh, Agrawal, Swami, entre otros, empezaron a utilizar y darle importancia cada vez más a este término, dando la posibilidad a que las empresas se dedicaran a ofrecer servicios utilizando la minería de datos. A finales del año 1980 solo eran un par de empresa las que se beneficiaban con esta tecnología, con el pasar de los años fue creciendo exponencialmente el número de empresas que se dedican a esto, hoy en día son muchas y variadas las organizaciones que ofrecen distintas soluciones, utilizando la minería de datos como herramienta principal. (Agrawal, Imielinski, & Swami, 2013).

Este avance tecnológico, ha logrado ser un buen punto de inicio para el encuentro entre personas del ámbito académico y de los negocios. Intentando presentar su apoyo para lograr comprender el contenido de grandes bases de datos.

La minería de datos, se refiere al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. y que surge por la aparición de nuevas necesidades, se dan ante el incremento del volumen y los diferentes tipos de información que se encuentra en las bases de datos y sobre todo por el reconocimiento de un nuevo potencial. el valor de los datos.

De una forma general, la información pasa de ser un producto a ser considerado una materia prima, que se tiene que pulir y explotar para poder alcanzar el conocimiento. Así a través del conocimiento conseguido, se pueden tomar las mejores decisiones para una organización determinada o para el problema que se tiene a resolver. Todo esto depende de los encargados en la toma de decisiones, los cuales tienen que utilizar el conocimiento obtenido eficientemente, para asegurar un resultado valioso. Claramente, se ve que hallar los patrones no lo es todo, sino que hay que entender, actuar, transformar los datos en conocimiento, y el conocimiento en valor para la organización.

Figura 4. El ciclo de vida de los datos

Fuente: Eyherabide, 2012.

3.1.2.2. Data Mining y su relación con los Data Warehouse.

Antes de mostrar clara y detalladamente el significado de Data Mining y todo lo relacionado, vamos a aclarar un concepto que se relaciona constantemente, aunque no siempre, con la minería de datos, este es el Data Warehouse.

En principio, un Data Warehouse lo podemos asimilar como un término o concepto muy relacionado con la fase de Data Mining y que, en teoría para estudiar mejor las técnicas avanzadas de esta etapa, éstas debiesen estar totalmente incorporadas con el Data Warehouse, así como con herramientas interactivas y flexibles, todo esto nos sirve para cuando tenemos grandes volúmenes de datos o cuando se combinan de formas arbitrarias y no predefinidas.

Es por eso que en esta sección se pretende describir rápidamente lo que es un Almacén de datos o Data Warehouse, para su posterior comprensión, sin profundizar demasiado ni desenfocarnos del tema principal Data Mining.

Las grandes bases de datos de una empresa, usan los computadores como medio para organizar su información o sus datos, de tal forma que sean accesible para los usuarios. A este proceso se le denomina como Data Warehousing.

Según Inmon & Hackathorn, (1992). Define un Data Warehouse es "una colección de datos orientados a temas, integrados, no-volátiles y variantes en el tiempo, organizados para soportar necesidades empresariales".

Estos concentran una gran cantidad de datos de interés para una organización, la cual es distribuida por medio de diferentes herramientas de consulta y de creación de informes, dirigidos para la toma de decisiones, tiene por objetivo principal agrupar los datos con la finalidad de facilitar su siguiente análisis

mediante diversas herramientas, entre ellas, la minería de datos. La etapa de Data Mining extrae de los datos guardados de conocimientos o información de predicción desde el Data Warehouse sin solicitar requerimientos o cuestionarios específicos, a través de sus diferentes técnicas algorítmicas. El almacén de datos ahora pasa a ser un sistema de "información central" en todo este proceso.

Según Francisco, (2015). Define “Actualmente los Data Warehouse y las técnicas OLAP (On-line Analytical Processing), son las formas más adecuadas y efectivas tecnológicamente, se podría decir que son las más avanzadas para integrar, transformar y combinar datos los cuales facilitan al usuario o a otros sistemas el análisis de la información requerida”. Sin embargo, la minería de datos es un conjunto de técnicas de análisis de datos, con una característica muy especial, que permite. Extraer patrones y tendencias para explicar y comprender mejor los datos, predecir conductas futuras.

La gran y principal diferencia entre Data Mining y las demás herramientas que se asocian con Data Warehouse, es que esta etapa no tiene por objetivo transformar y favorecer el acceso a la información, para que el usuario después la analice y sea más comprensible, sino que tiene por propósito analizar los datos obteniendo conocimiento útil para una toma de decisiones de manera adecuada.

Resumiendo, el Data Warehouse facilita la información de gestión entendible, uniforme, actualizada y correcta. Realiza un mejor servicio al cliente con un menor costo para la toma de decisiones, una mayor flexibilidad en el entorno y permite realizar un rediseño de procesos.

Además, las técnicas de Data Mining son usadas significativamente para realizar la explotación de los datos y el análisis de un Data Warehouse, brindando soluciones que están fundamentadas para la interpretación de los datos, por medio de la programación de interfaces de uso general y dar a conocer algoritmos para todos, permitiendo una eficaz exploración y una adecuada organización de los datos.

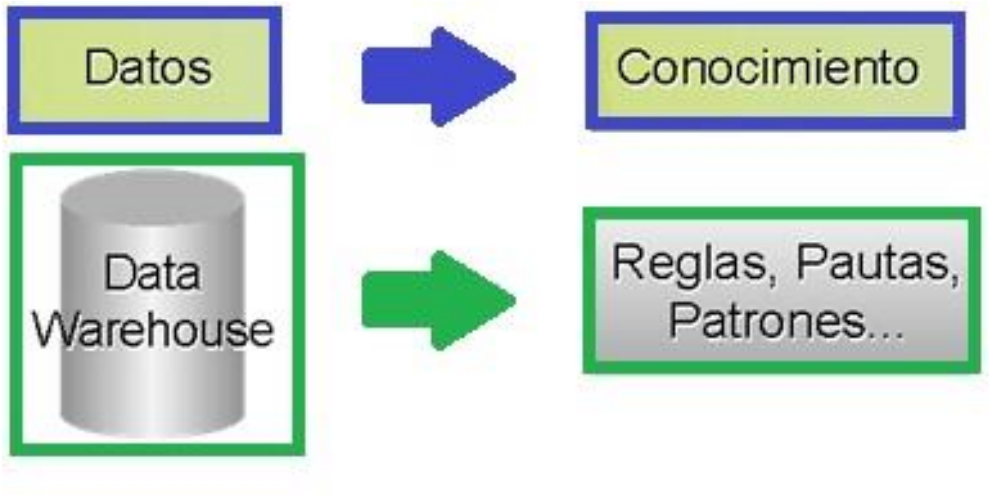

Figura 5. Data Warehouse y su relación con la minería de datos.

Fuente: Francisco, 2015.

3.1.2.3. Minería de datos y sus necesidades.

Los autores Witten & Frank, (2000), definen la minería de datos como "el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos".

Actualmente el análisis de los datos en una base de datos se desarrolla mediante consultas realizadas con lenguajes generales de consultas, como él (Structured Query Language, SQL), estas se efectúan sobre las bases de datos operacionales, es decir, junto al desarrollo transaccional en línea (On-line Transaction Processing, OLTP) de las aplicaciones de gestión. Las consultas en este

procesamiento permiten extraer información resumida. Sin embargo, esto produce información resumida de forma antes establecida, es limitada y poco escalable a grandes volúmenes de datos y la información no es flexible.

Ante las nuevas demandas que surgen debido al crecimiento masivo y la diversidad de fuentes de información, se da un paso más y aparece en la lista otra arquitectura nueva como el almacén de datos (Data Warehouse). Este almacén consiste en un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para favorecer su análisis y dar apoyo a la toma de decisiones, mejorando los procesos de negocios en cualquier organización.

Esta tecnología incluye operaciones de procesamiento analítico en línea (On-line Analytical Processing, OLAP), es decir, técnicas de análisis como pueden ser el resumen, la consolidación o la agregación y también la posibilidad de ver la información desde diferentes enfoques o posición (Francisco, 2015).

A pesar que esta herramienta OLAP, soporta ciertos análisis descriptivos que añaden y permiten transformar de cierta manera los datos en otros datos agregados o combinados, necesitan la habilidad de generar reglas, patrones, pautas, es decir, conocimiento útil que pueda ser aplicado en otros datos.

Existen otras herramientas analíticas, que han sido utilizadas con el fin de analizar los datos y que tienen su origen en la estadística. Aunque estas herramientas son capaces de inferir patrones a partir de los datos (utilizando modelización estadística paramétrica o no paramétrica), el problema que poseen, es que resultan especialmente crípticos para los no estadísticos, es decir, no funcionan bien para bases de datos de grandes volúmenes (con millones de registros, cientos de tablas,

con peso de varios gigabytes y una alta dimensión), que son las que se usan actualmente, además no se incorporan bien con los sistema de información y no so portan algunos tipos de datos frecuentes (atributos nominales con muchos valores, multimedia, datos textuales, etc.)

Es así, como las nuevas necesidades parten de que la importancia no radica en los datos como tales, sino en el conocimiento que se puede extraer a partir de ellos y aún más, en que dicho conocimiento sea utilizable. Estas necesidades y las restricciones de las técnicas existentes ayudan al origen de una nueva generación de herramientas que posibiliten la extracción de conocimiento útil, a través de la información que se disponga, y que, contenidas en una denominación, reciben el nombre de "minería de datos" (Francisco, 2015).

La minería de datos, constituye la fase central del proceso denominado extracción del conocimiento en bases de datos (Knowledge Discovery in Database), y el resultado que genera esta fase son patrones, conjuntos de reglas, ecuaciones, arboles de decisiones, entre otros conceptos.

a. Objetivos del Data Mining.

A continuación, se exponen algunos objetivos importantes dentro de este proceso de minería de datos.

Producir conocimiento nuevo que resulte útil mediante la elaboración de un modelo a partir de la información obtenida. El modelo mencionado es una explicación de los patrones o vinculaciones, que puedan tener los datos y ser utilizados para comprender mejor los datos, explicar situaciones anteriores o realizar predicciones.

Encontrar modelos inteligibles a partir de varios volúmenes de datos encontrados en diversos repositorios de datos. Para que este desarrollo sea positivo tiene que ser asistido o automático.

Modelamiento descriptivo es la construcción de modelos que ayudan a entender los datos que se tiene (estimación de distribución de probabilidades, búsqueda de correlaciones entre variables, etc.)

Modelamiento predictivo es la construcción de un modelo que prediga el valor de una variable determinada, y partir de los patrones descubiertos, se espera poder tomar decisiones confiables que sean de utilidad y beneficio para una organización en particular.

b. Ventajas del Data Mining.

Genera Modelos predictivos: Proporciona relaciones no encontradas e identificadas mediante la ejecución del proceso de minería de datos las cuales sean declaradas como modelo predictivos o reglas del negocio.

Permite a los usuarios dar primacía a las diferentes decisiones y acciones resultando factores que tienen un porcentaje mayor en un objetivo, qué fragmentos de clientela pueden ser desechables y qué partes del negocio son sobrepasados y por qué.

Ayuda a la organización, ya que ahorra dinero y produce oportunidades nuevas de negocio. Admite a los usuarios examinar la base de datos, que sin seleccionar previamente algún subconjunto de variables.

Identifica la llave a la información, desde grandes cantidades de datos producidos por procesos convencionales y extrae los patrones de una forma automatizada, esto contribuye de manera estratégica para la toma de decisiones.

Produce dominio sobre la decisión del negocio a usuarios, que mejor comprenden el entorno y el problema y es capaz de calcular los movimientos y el resultado de mejor forma (Francisco, 2015).

3.1.2.4. El Proceso de Data Mining.

En este punto se utilizan distintos procesos autónomos del procedimiento específico para la obtención del conocimiento usada.

Aunque en minería de datos cada fase concreta y especifica es distinta a la anterior, el proceso normalmente se compone de etapas principales que pueden son cuatro.

a. Determinación de los objetivos.

Esta fase consiste en la delimitación y la finalidad que un cliente desea, el cual bajo la orientación del experto en minería de datos. Un aspecto muy importante en este paso es que la minería de datos no es una meta así misma, es decir, no hay proyecto si el negocio no tiene un objetivo.

b. Pre-procesamiento de los datos.

Mediante el pre-procesado, se depuran los datos (se suprime valores que no son correctos, valores no válidos y desconocidos.), se obtienen ejemplares de los mismos (mayor rapidez a la respuesta del proceso), se enriquecen, el tamaño de los datos es menor, eligiendo las variables que influyen en el problema, sin apenas

sacrificar la calidad del patrón obtenido de la aplicación de minería, además ocurre un cambio de las bases de datos. Esta etapa desgasta alrededor del 75% del tiempo de un proyecto de minería de datos.

c. Determinación del modelo.

Se inicia realizando unos análisis estadísticos de los datos, para tener una primera aproximación se lleva a cabo una percepción gráfica de los mismos. Según formulación de los objetivos y el trabajo que debe llevarse a cabo, se pueda desarrollar algoritmos en diferentes campos de la Inteligencia Artificial.

Mediante la técnica se obtiene un patrón de conocimiento, que es representado por un comportamiento que se observan en los valores de las variables o las conexiones de asociación entre ellas. Los cuales se pueden usar varios procedimientos y a la vez para producir diferentes modelos para luego compararlos.

d. Análisis de los resultados.

Finalmente se procede a su validación, se verifica si los datos obtenidos están acorde y adecuados, luego compararlos con los que se obtuvieron en el análisis estadístico y los de visualización gráfica. El usuario señala si lo obtenido es novedoso y aportan un conocimiento nuevo que permite considerar sus decisiones.

Cabe destacar que, de haber resultado varios modelos a través del uso de diferentes técnicas, es necesario comparar los modelos rastreando aquel que se encaje mejor al problema, viendo su validez y eficacia. Y si ninguno de los modelos planteados logra alcanzar los resultados esperados, es necesario buscar nuevos modelos por lo que es necesario replantear algunos procesos.

3.1.2.5. Tipos de modelos y enfoques algorítmicos de Data Mining.

Según Leo Apostel un modelo es "una representación abstracta, gráfica o visual, física, matemática, de fenómenos, sistemas o procesos a fin de analizar, describir, explicar, simular; en general, explorar, controlar y predecir esos fenómenos o procesos".

Un modelo nos admite decidir un último resultado, a partir de ciertos datos de ingreso en el proceso o sistema. La minería de datos tiene como objetivo obtener conocimiento beneficioso, a partir del análisis de ciertos datos. Estos conocimientos se pueden obtener realizando diferentes reglas inferidas en los datos o patrones, o bien de una presentación más precisa, es decir, una recopilación de los mismos, constituyendo un modelo de datos analizados. Hoy en día, existen muchas formas distintas de interpretar y mostrar los modelos, cada uno define el tipo de técnica a utilizar para poder inferir dichos modelos (Waldo, 2012).

Existen dos tipos de modelos: los predictivos y descriptivos, los cuales tienen distintas tareas, y a la vez estas tienen sus propios requisitos, que obtienen distintos tipos de información.

a. Modelos predictivos.

De las variables de interés (dependientes) se estiman valores desconocidos o futuros, utilizando variables diferentes a los de la base de datos (independientes). En otras palabras, se administra problemas y trabajo en los que hay que interpretar valores para uno o más modelos o explicado de mejor manera, intenta responder o predecir preguntas futuras en base a un trabajo de su comportamiento anterior, además estos son resguardados de una salida (clases, valor numérico o categoría).

Por ejemplo, un modelo predictivo sería el que permita definir las ventas futuras de un determinado producto.

Entre las tareas de predicción encontramos la clasificación y regresión.

- *Clasificación.*

Es una tarea muy utilizada, en este cada ejemplar de la base de datos se relaciona a una clase, la cual se señala por medio del valor de un atributo. Este atributo puede relacionarse con distintos valores discretos correspondientes a una clase determinada, los demás atributos (relevantes), se usan para predecir la clase.

La finalidad es predecir la clase de nuevos registros de las que se desconocen la clase, utilizando las reglas generadas, es decir, maximiza la precisión de clasificación de los registros nuevos que se van agregando, y se calcula de la siguiente manera. Precisión es igual a Predicciones correctas entre el número total de predicciones (correctas o incorrectas).

- *Regresión.*

Se trata de entender una función real en el cual cada registro obtiene un valor real. La misma que es la diferencia esencial con la clasificación, ya que el valor que se va a predecir es real (tipo numérico).

El objetivo de esta tarea lograr el valor real y el error entre el valor predicho sea minimizado, utilizando la función generada. Y dado que la única diferencia con la clasificación es que la regresión predice valores reales, un modelo de regresión podría convertirse fácilmente en un modelo de clasificación.

Se tiene un w0 y w1, donde “x” e “y” son las dimensiones y “n” es la cantidad de ejemplos de la muestra.

$$w1 = \frac{n(\sum xy)(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}$$ ……..….. [Ecuación 1]

w1= valor real uno

x = dimensión uno

y = dimensión dos

n = cantidad de ejemplos

Fuente: Nadinic, 2008.

$$w0 = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2}$$ ……..…..…..… [Ecuación 2]

w0 = valor real dos

x = dimensión uno

y = dimensión dos

n = cantidad de ejemplos

Fuente: Nadinic, 2008.

Obteniendo la siguiente ecuación.

$$y = w0 + w1x$$…….……[Ecuación 3]

y = conclusión real

w0 = valor real dos

w1 = valor real uno

Fuente: Nadinic, 2008.

b. Modelos descriptivos.

Identifican patrones que describen los datos, los cuales sirven para buscar las propiedades de los datos analizados, entregar información entre sus características y la relación de los datos y no para predecir datos nuevos. Dado el análisis de los datos se deduce un modelo representativo que ayude a la solución de la problemática. El objetivo de este modelo no es pronosticar datos nuevos, sino que explicar los existentes.

Por ejemplo, una agencia de transporte analiza los traslados que han realizado sus clientes y supone un modelo representativo, con el fin de identificar conjuntos de personas con similares gustos, para poder realizarles distintas ofertas por cada grupo.

Entre las tareas o enfoques descriptivos encontramos el agrupamiento (clustering), las reglas de asociación, y las correlaciones o análisis correlacional.

- *Agrupamiento o Clustering.*

Es uno de los mejores enfoques dentro del modelo descriptivo y radica en lograr conjuntos desde de los datos, este trabajo no analiza el etiquetado de los datos con una clase, sino que los analiza para producir esta etiqueta, debido a esto no se conocen las clases, ya que el objetivo de este enfoque algorítmico es poder detallar de forma breve el grupo de datos.

Los datos son agrupados fundamentados en el principio de mejorar la semejanza entre las partes de un grupo y minimizar la semejanza entre las partes de grupos diferentes.

En otras palabras, la base de datos se va a segmentar en grupos, donde cada grupo posee objetos que son muy parecidos entre sí, pero que a su vez son muy diferentes a las partes de los otros grupos. Los grupos pueden ser o no disjuntos.

También los resultados obtenidos mediante este enfoque, van a poder ser utilizados como entrada a otros métodos, o para resumir la información de las grandes bases de datos.

- *Asociación.*

Esta tarea busca vinculaciones no expresadas o relación entre los datos de los atributos discretos. El objetivo de la asociación es poder expresar de manera precisa vinculaciones existentes entre los datos de los atributos de un grupo de datos. Es decir, buscar entre los datos algunos patrones que identifiquen ciertas reglas de comportamiento.

Las reglas de asociación están formadas por el antecedente (la primera parte, "si"), ubicado en el lado izquierdo, y el consecuente (la segunda parte, "entonces"), ubicado en el lado derecho, y también proponen la información en forma de declaraciones del tipo "si- entonces" ("Si A, entonces B"). No siempre la efectividad de una agrupación de los atributos, va a implicar la evidencia de una conexión causa-efecto, o sea, para que los datos estén asociados puede o no existir una causa.

Es utilizada, muchas veces, en el análisis de la cesta de compra (Market-Basket Analysis), para identificar productos que son frecuentemente comprados juntos y así obtener valiosas reglas que puedan usarse de mejor manera para ajustar el inventario de un supermercado, por ejemplo.

Existe un caso de las reglas de asociación, las mismas que se denominan “reglas de asociación secuenciales” o “patrones secuenciales”. Y básicamente, lo que se busca con este tipo de reglas es determinar de forma concisa patrones secuenciales en los valores de los datos. Los cuales diferencian de las reglas de asociaciones normales en que las relaciones existentes entre los datos se basan en el tiempo. Por ejemplo, funcionan de la siguiente forma: “Si sucede en el evento X en el preciso momento de tiempo t, entonces sucederá en el evento Y en el momento de tiempo t+n”.

- *Correlación.*

Esta tarea se utiliza para lograr revisar o buscar el grado de igualdad de los datos entre dos atributos numéricos. Posee una fórmula que permite medir la correlación lineal, esta fórmula se le denomina “el coeficiente de correlación r”, que es un valor real y se encuentra entre -1 y 1 ($r \in [-1...1]$). Los atributos tienen un comportamiento similar si r es positivo, es decir, los atributos están perfectamente correlacionados, si r es negativo, los atributos están perfectamente correlacionados negativamente, esto quiere decir, que cuando un atributo se incrementa el otro disminuye, mientras que si r es 0 no hay correlación. Este análisis de correlaciones, es útil cuando se quiere establecer reglas de ítems correlacionados.

3.1.2.6. Técnicas de Minería de Datos.

Las técnicas de minería de datos son la consecuencia de un largo proceso de desarrollo e investigación de productos que derivan de la Inteligencia artificial y estadística, estas técnicas son algoritmos con un grado de complejidad, y que son aplicadas sobre un grupo de valores obteniendo ciertos resultados.

Cada tarea o función puede ser llevada a cabo utilizando distintas técnicas, por ejemplo, para generar modelos predictivos pueden ser inferidos por técnicas como redes neuronales y árboles de decisión, por nombrar algunas técnicas. El tipo de conocimiento, que va a resultar va a marcar claramente la técnica de Minería de Datos a realizar (Hernandez, Ramirez, & Ferri, 2004)

Desde hace mucho tiempo, se aplicaban sobre grandes volúmenes de datos las técnicas de minería de datos. Esto llevo a que muchas empresas públicas y privadas comiencen a crear y alimentan bases de datos, que son principalmente relacionadas y diseñadas para proyectos de minería de datos, en las que centran información fundamentalmente valiosa y útil de todas las áreas de negocio.

a. Redes neuronales artificiales (RNA).

La terminología "red neuronal" (NN, por Neural Network) es usada para expresar modelos matemáticos de la labor del cerebro humano, que expresan las potencialidades del método paralelo masivo y de la muestra distribuida que existe en el cerebro.

Las redes neuronales artificiales (ANN, por Artificial Neural Network), son un paradigma del estudio y de la metodología automática inspirada en la forma en que se desarrolla el sistema nervioso de los seres vivos especialmente de los animales, también las redes neuronales artificiales imitan el real sistema nervioso de forma abstracta y consisten en un sistema relacionado e interconectado por múltiples neuronas que se aportan mutuamente entre ellas, con el fin de generar un estímulo de salida.

Las redes neuronales artificiales evolucionan y una de ellas es el procesamiento de información o datos que se dan en elementos sencillos determinadas como neuronas. Las neuronas o nodos transmiten señales por medio de conexiones estandarizadas. Cada conexión, llamado también como enlace de comunicación, tiene un peso asociado.

Cada neurona aplica un proceso el cual activa la entrada total recibida que son recibidas de las neuronas conectadas, generando un valor de salida que se interpretará como valor de ingreso, los cuales serán transmitidos al resto de la red.

Algunas características de esta técnica son. Están especializadas para resolver los problemas. Los cuales se realizan de manera interactiva y sistemáticamente a ingresos clásicos, cada una con sus respectivas respuestas o salidas. Son usadas para descubrir patrones, clasificaciones de voz e imagen, procesamiento del lenguaje, predicción y optimización. Generan modelos de tipo predictivos, producen un solo modelo de este mismo, se utilizan para resolver problemas de enfoques de Clasificación y de Regresión, aunque también a veces resuelven problemas de Agrupamiento.

La exactitud es generalmente alta. Son una de las técnicas más difícil de comprender, pero poseen ventajas muy significativas como su buen funcionamiento predictivo, la tolerancia a fallos, su auto organización, su flexibilidad, el método de aprendizaje y también pueden ser utilizadas en tiempo real. (Francisco, 2015).

- *Estructura.*

Una red neuronal se fundamenta en el procesamiento distribuido a una red de nodos llamados neuronas o unidades de procesamiento, que son la base de una red

neuronal y tienen la capacidad de trabajar en paralelo. Tanto el paralelismo como el procesamiento distribuido son dos propiedades importantes dentro de la minería de datos, ya que permiten que las redes neuronales puedan procesar cantidades de datos muy elevadas.

Es necesaria mostrar una imagen para facilitar y entender el funcionamiento de una neurona artificial.

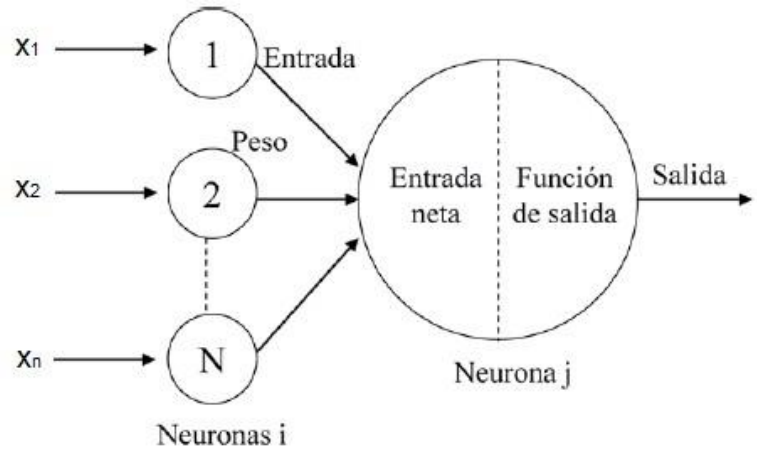

Figura 6. Funcionamiento general de una neurona artificial

Fuente: Palmer & Montaño, 1999.

$y_f = f(\sum_{i=1}^{n} w_{ij} * x_i + \theta_j$ ……..…….. [Ecuación 4]

Para entender de mejor manera la imagen se procede a explicar el funcionamiento de una red neuronal artificial. Estas cuentan con tres funciones.

Función de entrada: Tiene por objetivo combinar los patrones de entrada que llegan a la neurona dentro de una entrada general. Cada uno de los valores de entradas van a multiplicarse por sus correspondientes pesos.

Función de activación: Calcula la activación de una unidad en relacionada con la entrada total y la previa activación. Como ya dijimos cada una de las neuronas aplica un procedimiento de activación a la entrada total que recibe de otras neuronas que están relacionadas, siempre y cuando esta entrada supere un cierto umbral y luego emite una señal hacia las neuronas de la siguiente capa. Si lo obtenido es menor que el valor del umbral, la neurona no envía ninguna señal lo que significa que se encuentra inactiva y algunas funciones de activaciones típicas, no lineales, son.

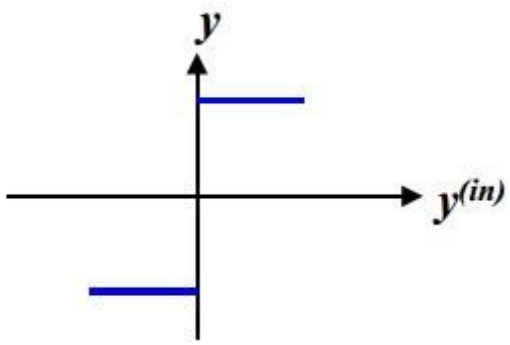

Figura 7. Activación Tipo Escalón

Fuente: Izaurieta & Saavedra, 2000.

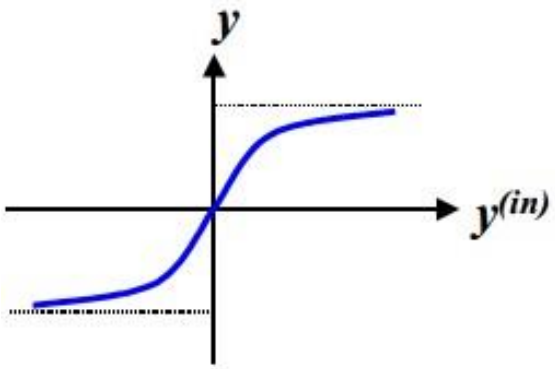

Figura 8. Activación tipo sigmoidea

Fuente: Izaurieta & Saavedra, 2000.

También existen activaciones lineales, cuando ocurre estos se dice que es una "neurona lineal", en caso contrario es una "neurona no lineal". Las neuronas lineales se representan por un cuadrado mientras que las neuronas no lineales se representan por un círculo.

Función de salida. Una vez aplicada la función de activación y esta función está por encima del umbral, es decir, lo supera, envía una señal y el ejercicio final determinando que valor es transportado a las neuronas relacionadas. Cabe mencionar, que, si la función que se encarga de la activación no supera el umbral determinado, no se va a pasar ninguna salida a las otras neuronas y se puede estructurar de diferentes formas.

Las Redes de propagación, como las redes neuronales mono-capas, es la red neuronal más tradicional y fácil, tiene una capa de neuronas que se encargan de proyectar las entradas a una capa de neuronas de salida donde se realizan diversos cálculos. En una red mono-capa, las neuronas de salida pueden ser lineales o no lineales. En este modelo de red, las neuronas se encuentran totalmente conectadas y no existen ciclos.

Perceptrón: Fue el primer modelo de red neuronal artificial supervisada creado. Es capaz de reconocer y aprender patrones sencillos convirtiéndolo en el modelo más simple de las redes neuronales.

Esto gracias a que es de tipo supervisada, por lo tanto, la red tiene que ser entrenada con un grupo de patrones previamente clasificados de manera que, si se los clasifica de forma incorrecta, por medio de una regla de aprendizaje se pueda corregir el error (Rosenblatt, 1958).

La principal limitante del Perceptrón es que tiene reglas que solo tiene como objetivo problemas que sean de dos clases y linealmente separables, es decir, con un híper plano se deben separar los elementos "validos" de los "no validos".

Esto se hace mediante una función que sea discriminante lineal creando una frontera de decisión. Un Perceptrón se puede beneficiar con otros perceptrones u otro tipo de neurona artificial, constituyendo así al desarrollo de redes neuronales más complejas.

Adaline: es una neurona artificial desarrollada por el profesor Bernie Widrow y un estudiante, en la Universidad de Stanford en el año 1960. La misma que es útil para preparar un simple elemento de procesado, el cual es propuesto a realizar una función de transmisión lineal.

La misma que es denominada como “Regla del Mínimo Error Cuadrático Medio” y es fundamental reducir la diferencia que se da entre la salida que se desea y la actual para cada patrón. Cabe destacar, que es necesario calcular la función de error para todo el conjunto de patrones. Se compone de una sola capa de “n” nodos o neuronas generando así “n” salidas, con “m” entradas.

La estructura de esta red es muy similar a la del Perceptrón simple, resuelve problemas linealmente separables, pero es un modelo físico capaz de interpretar el aprendizaje, diferenciándolo del Perceptrón en la forma de usar la salida en la regla de aprendizaje.

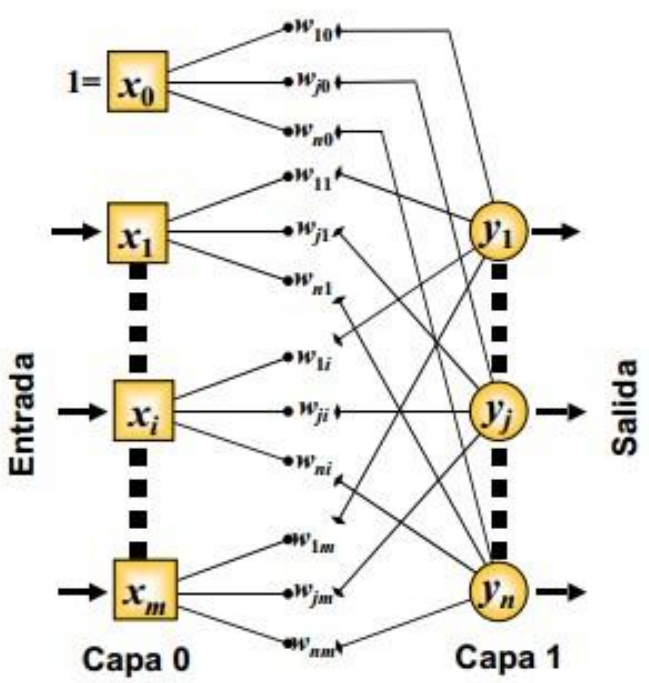

Figura 9. Red mono capa

Fuente: Izaurieta & Saavedra, 2000.

Redes neuronales multicapas: Es una extensión de la red neuronal unicapa o Mono capa, encontrándose un grupo de capas intermedias entre el ingreso y las salidas (capas ocultas). Este tipo de redes puede estar parcialmente o totalmente interconectada y no existen ciclos.

Perceptrón Multicapa: Es una red neuronal en cascada la cual tiene una o más capas ocultas, permitiéndole solucionar problemas que linealmente no se puede separar, resolviendo así la gran limitante del Perceptrón simple.

Su arquitectura se clasifica en tres capas; la "capa de entrada", que está cada neurona está estructurada para que introduzcan patrones de entrada a la red, y propaguen la activación a través de los pesos hasta la "capa oculta", en esta otra capa, se desarrolla un método que son aplicados a las entradas que llegan. Por lo

que la activación se transmite a por medio de los pesos a la capa de salida, los valores dados por las neuronas relacionadas a la capa de salida corresponden con las salidas de la red.

Posee una limitante, esta es efectividad que existe de mínimo locales relacionadas al error, dificultando de forma considerable el entrenamiento a la red, debido a que una vez que se alcance el mínimo, se detiene el entrenamiento, aunque no se logre alcanzar la tasa de convergencia requerida.

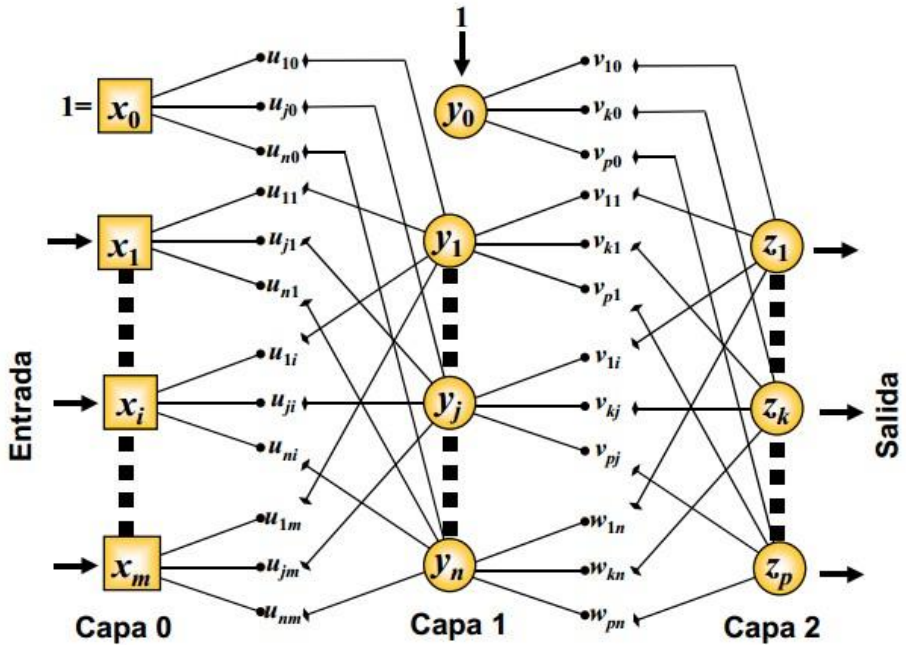

Figura 10. Red multicapa

Fuente: Izaurieta & Saavedra, 2000.

Las redes neuronales recurrentes se caracterizan por la interconexión de vínculos de realimentación, presenta al menos un ciclo cerrado de activación neuronal. Estos vínculos pueden ser entre neuronas de distintas capas, y de la misma capa o, más simple, entre una misma neurona. Esta estructura estudia fundamentalmente el modelo de sistemas no lineales.

Redes de Hopfield: Pertenece al grupo de las redes recursivas o recurrentes, indica que existe una realimentación entre las neuronas. Quiere decir que, al ingresar un patrón de entrada, la información se expande hacia adelante y hacia atrás, creando un modelo. Cuando la evolución esté en algún medio estable, esta se detendrá, o también puede ser el caso que la red no se detenga nunca.

Las redes recurrentes deben obedecer tres objetivos. Dado cualquier estado de inicio, deben converger constantemente a un estado estable. El poder de obtención de cada estado efectivo debe estar delimitado principalmente y de manera precisa debe cumplir algún criterio de métrica (por ejemplo, que el final se encuentre lo más cerca al inicial). Debe poder tener cualquier cantidad de estados estables.

La estructura es de tipo binaria, es decir, sólo tienen dos valores posibles para sus estados. Y los valores posibles pueden ser 1 ó -1, o bien 1 ó 0. Suelen usar como memoria auto-asociativa, esto significa que, si entra un patrón "x" incompleto o ya sea que no sea válido y se encuentre relacionado con el ruido, la red evoluciona hacia el patrón que se encuentre cerca. Esto se produce al entrenamiento que se les enseña, el cual se basa en disminuir los estados que debe recordar la red.

Máquina de Boltzmann: Es una red neuronal recurrente estocástica y consisten en neuronas conectadas entre sí, que pueden estar conectadas bidireccionalmente y que poseen salidas binarias. Útiles para el reconocimiento de patrones, intentando recuperar información no disponible de un estado. Además, pueden tener unidades ocultas.

Se clasifican en dos grupos: las visibles, que son las que conforman la interfaz de la red y las no visibles, que son las que ayudan a mejorar el desempeño de esta.

Una máquina de Boltzmann se puede ver como una red de múltiples unidades y de dos estados (binaria) conectadas de cierta forma, donde el estado prendido = 1 y apagado = 0.

- *Topología.*

La organización más popular de una red neuronal se constituye de tres capas. La "capa de entrada", es donde cada neurona o unión pertenece a una variable independiente lista para examinar, luego unos nodos organizados en una única o varias "capas ocultas", realizan el trabajo de la red y dependen de la dificultad del problema, y una "capa de salida" relacionada con nodos de salida, este nodo corresponde a la variable dependiente. Generalmente es un nodo el que se encuentra en la capa de salida, aunque si hay más de uno, significa que puede realizarse muchas predicciones.

- *Mecanismos de aprendizaje.*

El aprendizaje ofrece una opción eficaz a la programación. Se han propuesto diferentes estrategias de aprendizaje para redes neuronales tales como el aprendizaje supervisado y no-supervisado (Rojas, 1996)

El aprendizaje de una red neuronal más que nada significa que se podrá adaptar a los pesos, es decir, es el proceso por el cual una red neuronal cambia sus pesos para responder a una entrada de información. Dentro de las redes neuronales artificiales existen diversos métodos de aprendizaje siendo dos los más importantes.

Aprendizaje supervisado: Este aprendizaje se realiza debido a que el procedimiento, se realiza por medio de un entrenamiento que es conducido por un agente externo, este va a determinar el resultado que va producir la red desde una entrada determinada. Es el gestor o maestro quien controla la salida de la red, si por algún motivo se encuentren diferencias y no coincida con la esperada, es necesario cambiar los pesos de las conexiones, ajustando la red y aumentando la posibilidad que la salida nueva se aproxime a la anhelada.

Existen tipos básicos de aprendizajes supervisados que son tres. Aprendizaje por corrección de error. Aprendizaje por refuerzo. Aprendizaje estocástico.

Red Backpropagation: Es una red de tipo supervisada, para realizar la aplicación es necesario que se den modelos de redes con dos capas o más de neuronas. Radica en un aprendizaje en grupo predefinido de pares de entradas como también salidas dados previamente, usando un periodo de propagación y adaptación de dos fases.

En principio, se atribuye un patrón de entrada a la red como incentivo para la capa número uno de red, después se va difundir desde la primera capa por medio de las capas con mayor rango de la red, hasta producir una salida. Se revisa y relaciona el resultado que se obtiene en las neuronas de salida con la salida que se desea obtener y se evalúa la cantidad una señal de error para cada neurona de salida.

Luego estas señales de error se propagan y se dirigen hacia atrás desde la capa de salida, con rumbo a todas las neuronas de la capa intermedia que apoyan de manera directa a la salida, recogiendo el porcentaje de error estimado a la

participación de la neurona intermedia en la salida inicial. Este proceso se repite capa por capa de tal forma que en su totalidad de neuronas logren recibir un error que explique su aportación referente al error total.

Teniendo en cuenta el valor recibido, se ajustan los pesos que conectan de cada neurona, de tal modo que la próxima vez que se tenga el mismo patrón, la salida esté más próxima a la que se desea, disminuyendo así el error.

A diferencia de la regla delta en el caso del Perceptrón, esta técnica necesita el uso de neuronas que su función sea de activación continua y se diferencie de las demás. Comúnmente esta función será del tipo "Sigmoidal".

Memoria asociativa bidireccional: Es una red recurrente que implanta una memoria asociativa, se emplea para recordar información al presentarse una cierta información clave. Si la información solo se conoce parcialmente por anticipado, o si tiene ruido, la red es capaz de completar dicha información.

Su arquitectura consta de dos capas de neuronas que-están completamente interconectadas entre capas, las cuales pueden ser de diferentes dimensiones. Las unidades puedes o no tener conexiones de retroalimentación consigo mismas. En esta red ninguna capa puede ser considerada esencialmente solo de salida o, de entrada.

A las capas se les denomina "capa-X" y "capa-Y", respectivamente. Las conexiones de estas capas son bidireccionales, esto quiere decir, que las funciones de la red, de forma iterativa, envían señales hacia atrás y adelante entre las capas hasta que el equilibrio se ha alcanzado.

Aprendizaje no supervisado: Este no requiere de un agente externo para ajustar toda la red. Tampoco percibe información de parte de su entorno que indique si lo generado en respuesta a una entrada definida es correcta o no. De este modo lo único que puede realizar la red es encontrar patrones en los datos de entrada y crear diferentes niveles a partir de los patrones. Así después del entrenamiento cuando ingrese un dato, la red podrá clasificarlo e indicar en que categoría clasificó a dicho dato. Existen dos tipos básicos de aprendizajes supervisados hebbiano, competitivo y comparativo.

Mapas de Kohonen: Las redes auto-organizadas fueron inventadas por Teuvo Kohonen, entre los años 1982 y 1990, Kohonen creó una forma de representar los datos de un campo multidimensional en otro campo de menores dimensiones.

Las redes de Kohonen están compuestas capas son dos: la capa de entrada y la capa de salida de sensores que se encarga de realizar el cálculo. Las neuronas que representan algún parecido entre sus patrones se presentan juntas en el campo de salida. Este campo salido es establecido por el diseñador de la red. La idea está fundamentada en un funcionamiento de aprendizaje biológico por competición, de tal modo que cuando un conjunto de datos de ingreso se proyecta a la red, los pesos de las neuronas se adaptan de modo que la clasificación presente en el campo de entrada sea preservada en la salida.

Máquina de Cauchy: Es un modelo superior a la de Boltzmann, posee la misma arquitectura y funcionamiento, pero se diferencia ya que considera funciones alternativas de probabilidad y de ajuste de temperatura.

Posee una gran una ventaja respecto a la máquina de Boltzmann, esta radica en la rapidez de convergencia. También, se ha demostrado que combinando las funciones de posibilidad y temperatura previa se consigue alcanzar el mínimo global de energía.

Resonancia Adaptativa: Fue desarrollada por Stephen Grossberg y Gail Carpenter para resolver la "Teoría de Resonancia Adaptiva", trata de encontrar un modelo adecuado y que tenga la capacidad de respuesta a problemas de aprendizaje, dado que una red puede aprender patrones nuevos (plasticidad del aprendizaje), y conservar previamente los patrones aprendidos (estabilidad del aprendizaje).

El modelo propone que, ante una determinada información de entrada, se activa solo una de las neuronas de salida, logrando alcanzar su dato de respuesta máximo tras competir con las neuronas que faltan. Para solucionar esto, las redes de resonancia adaptativa proponen añadir un modo de retroalimentación entre neuronas de la capa de salida y entrada, con el objetivo de proveer el aprendizaje de información nueva, pero sin eliminar la que ya se encuentra almacenada.

Hay métodos básicos de entrenar una red de resonancia adaptativa y son dos: lento y rápido. Con el primer método el nivel de entrenamiento de los pesos de la neurona de reconocimiento, y hacia el vector de entrada se determina la cantidad a los valores continuos con ecuaciones diferenciales. En cambio, con el método rápido, se utilizan ecuaciones algebraicas para calcular el grado de ajustes de peso, usándose valores binarios. A medida que la red va a aprendiendo, ésta va creando su propia clasificación, el aprendizaje es de tipo no supervisado, aunque existe una modalidad supervisada.

Durante el desarrollo de aprendizaje, los pesos de las conexiones de la red sufren modificaciones, donde podemos indicar que este proceso ha terminado o que la red ha logrado aprender, cuando los datos resultantes permanecen estables en sus pesos.

El modelo no menor dado el aprendizaje de las redes neuronales es necesario saber cómo se ajusta la red cuando una salida no es como la deseada, o sea, de qué forma se determina lo que se va usar para modificar los valores de los pesos.

En fin, para las redes neuronales artificiales hay una amplia gama de modelos para seleccionar. La mezcla de la topología (el número de neuronas y capas ocultas, y cómo están conectadas), el paradigma de aprendizaje y el algoritmo de aprendizaje definen un modelo de red neuronal artificial (Bigus, 1996).

b. Arboles de decisión (AD).

Un árbol de decisión es un grupo de condiciones que se organizan para realizar una estructura jerárquica, el mismo que contiene cero o más nodos internos y uno o más nodos de hoja. Los nodos internos tienen dos o más nodos secundarios y contienen divisiones, estos demuestran el valor de los atributos expresados. Los arcos de un nodo interno a otro secundario o de menor jerarquía, son enumerados con diferentes salidas de la prueba del nodo interno. Cada nodo hoja posee una etiqueta de clase asociada.

El Árbol de Decisión del mismo modo es un modelo de predicción, que facilita la información para categorizar y representar de forma gráfica una fila de reglas que ocurren de forma posterior, la cual es determinante para tomar una decisión a un valor de salida, resolviendo un determinado problema. La construcción de un Árbol de Decisión se basa en el principio de “divide y vencerás”:

por medio de un algoritmo de aprendizaje supervisado estas realizan separaciones sucesivas del campo de muchas variables para lograr maximizar la distancia entre el conjunto de cada división, es decir, realiza el fraccionamiento que discrimina.

c. Naïve Bayes (NB).

Esta es una de las técnicas más utilizadas y está basada en condiciones probabilísticas, hace uso del Teorema de Bayes propuesto Bayes mediante esto realiza las predicciones. La misma que puede lograr predecir que un determinado caso pertenezca a una determinada clase y es usada principalmente para resolver problemas de clasificación, aunque también genera un modelo predictivo y descriptivo. (Bayes, 1764).

El adjetivo "Naïve" (ingenuo), es considerado de tal forma, ya que supone que el efecto de un valor de atributo sobre una clase dada es independiente de los valores de los otros atributos X. Dada la clase determinara si los atributos son condicionalmente independientes, este va a superar a los clasificadores de redes neuronales artificiales, también a los arboles de decisión. (Han & Kamber, 2001).

3.1.2.7. Herramientas de Data Mining.

a. Introducción.

Las herramientas de Minería de datos predicen futuras comportamientos y tendencias, mediante la exploración de las bases de datos en busca de patrones que se encuentren ocultos, beneficiando a los negocios y ayudando a tomar decisiones conducidas y proactivas por un conocimiento finalizado en el manejo de la información.

También se pueden resolver cuestionarios que demandan mucho tiempo, encontrando datos que ni un profesional experimentado podría encontrar, debido a que se encuentra lejos de sus expectativas. (Francisco, 2015).

Cabe aclarar, que existen múltiples herramientas de minería de datos, pero solo algunas permiten y ayudan a integrar datos históricos con entradas de datos en tiempo real, dando como resultado la posibilidad de ejecutar Data Mining en tiempo real.

Para el desarrollo de modelos de minería de datos existen múltiples herramientas de software, ya sean libres o comerciales.

Según una encuesta hecha en el portal del reconocido sitio "Kdnuggets 20", relacionado con el proceso de extracción de conocimiento respecto a la base de datos, dieron a conocer las herramientas más utilizadas en esta tecnología en los últimos 12 meses (Tomando en cuenta desde el año 2012 hacia atrás).

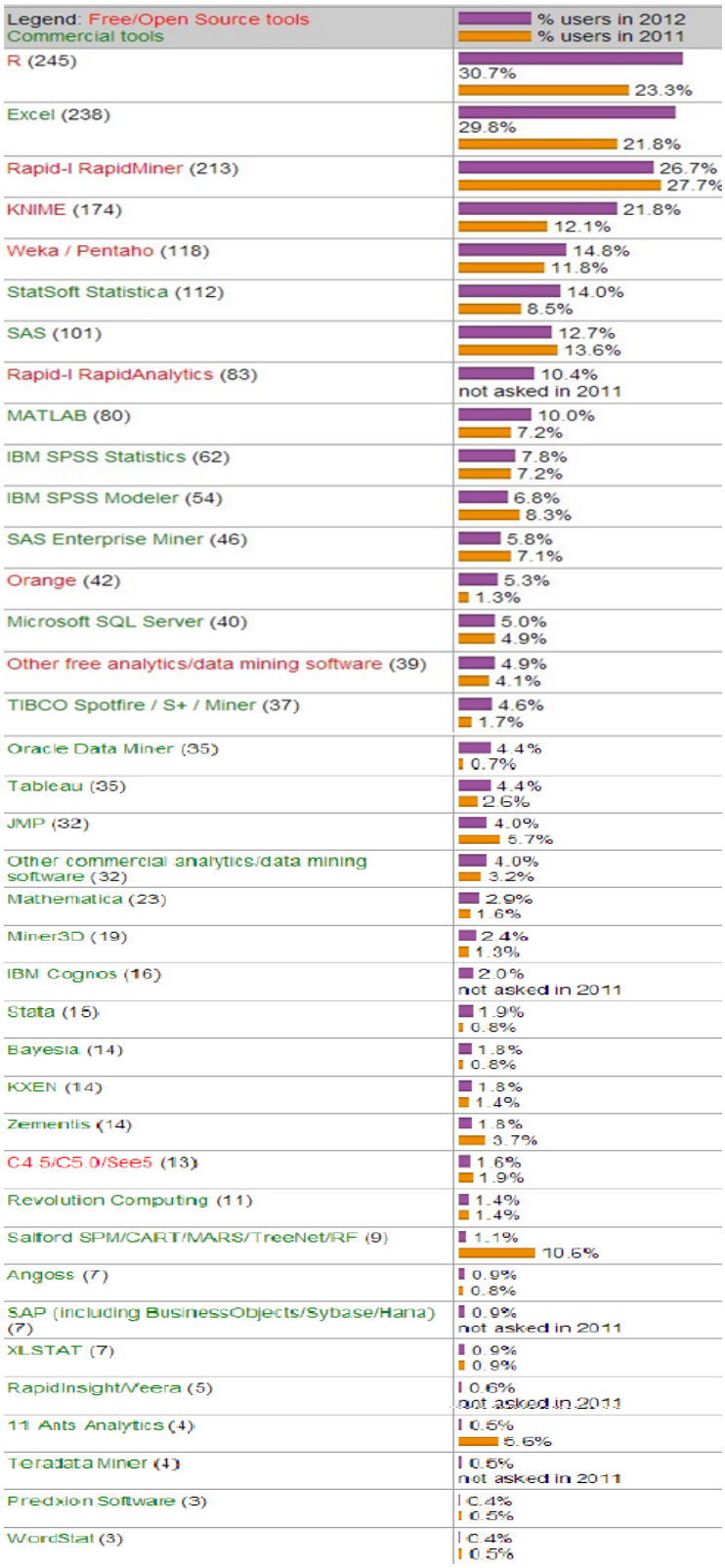

Figura 11. Encuesta de las herramientas utilizadas frecuentemente

Fuente: Kdnuggets, 2012.

De todas estas herramientas, se han seleccionado “WEKA” que, según la encuesta mostrada anteriormente, es una gran herramienta gratuita para trabajar el proceso de minería de datos. (Licencia GNU-GPL) lo que influye fuertemente en la elección y no de otras de tipo comercial, como, por ejemplo: “SPSS CLementine” o “SAS”.

b. WEKA (Waikato Environment for Knowledge Analysis).

La herramienta WEKA, es una plataforma de software que realiza aprendizaje automático y minería de datos desarrollado en Java y elaborado en la Universidad de Waikato (Nueva Zelanda). WEKA es un software libre distribuido bajo licencia GNU-GPL.

En el año 1993 se da inicio al desarrollo de la versión original de WEKA, en los lenguajes de C y TCL/TK, ya en el año 1997 se decidió reescribir el código en Java utilizándose en muchas y muy diferentes áreas, en general con fines docentes y de investigación.

La WEKA (Gallirallus australis) es un ave que no vuela con una naturaleza inquisitiva y se encuentra en las islas de Nueva Zelanda, su nombre se ha convertido en significado de una extensa colección de algoritmos máquinas de conocimiento elaborados por la universidad de Waikato (Nueva Zelanda) implementados en Java. Los algoritmos bien se pueden direccionar a un grupo de datos o llamadas desde su propio código Java.

Figura 12. Ave WEKA

Fuente: Wikipedia, 2018.

Esta herramienta está orientada a la extensibilidad por lo que agregar nuevas funcionalidades es una tarea sencilla.

- *Funcionamiento de WEKA.*

Está constituido por una serie de paquetes y herramientas gráficas y algoritmos para análisis de información y modelado predictivo, relacionados por una interfaz gráfica de usuario la cual es accesible a sus funciones. Estos pueden servir para cualquier proyecto relacionado al análisis de datos y también integrados, asimismo se pueden extender con aportes de parte de los usuarios que produzcan algoritmos nuevos.

La herramienta tiene cuatro interfaces de usuario para de interactuar con ella y así acceder a sus funcionalidades, estas son: Explorer, Experimenter, Knowledge Flow, Workbench, Simple CLI.

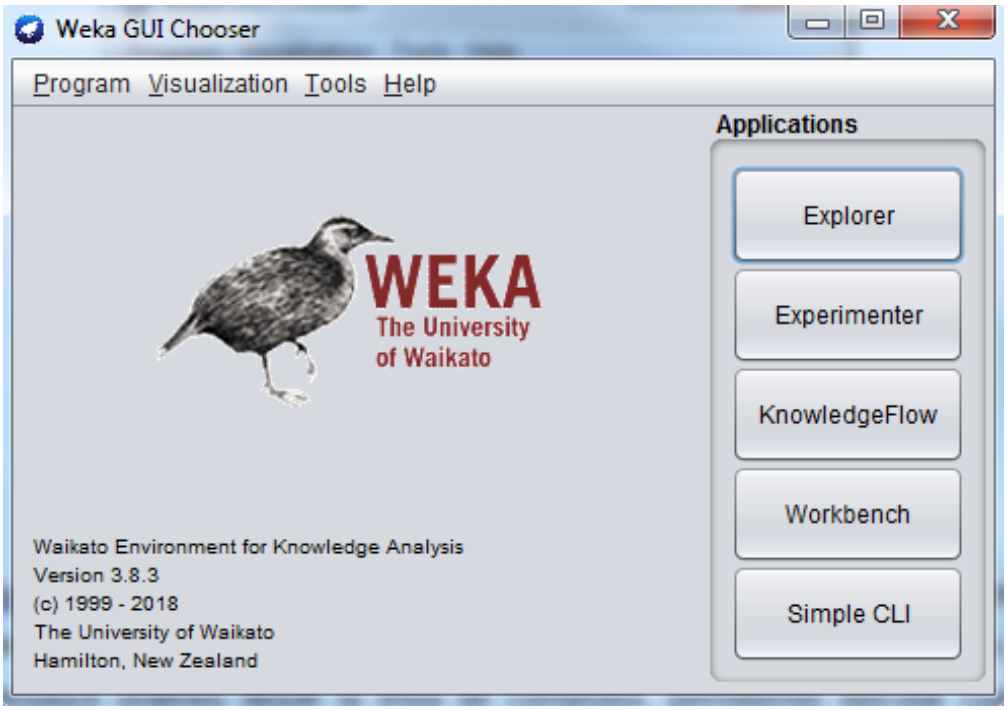

Figura 13. Inicio del Software Weka

El Explorer es el modo descriptivo más usado, posee varios paneles y un único archivo de datos, dando así facilidad para realizar operaciones que permiten ejecutar distintas trabajos, las cuales están insertadas en varios paneles, la cuales son: Pre-procesado de la información (Algoritmos de filtrados son usados para transformar la información y eliminar registros), Clasificación (Algoritmos de clasificación son aplicados para proyectar la exactitud del modelo predictivo resultante), Clustering (Da acceso a las técnicas de agrupamiento para entender una combinación de distribuciones normales), Asociaciones (Las reglas de asociación proponen verificar las interrelaciones fundamentales entre los atributos de los

información), Selección de atributos (Proporciona algoritmos para identificar los atributos más predictivos en un conjunto de datos) y Visualización de datos (Muestra una matriz de puntos dispersos, donde cada punto puede ser analizado en detalle).

Experimenter: Es un modo útil para aplicar varios algoritmos de aprendizaje automático sobre distintos conjuntos, y luego poder realizar contrastes estadísticos entre ellos y obtener otros índices estadísticos, es decir, el Experimenter nos dirá si las diferencias aparentes en porcentajes de aciertos de distintos algoritmos son estadísticamente significativas, o son debidas al azar.

Knowledge Flow: Es la interfaz más cuidada del programa. Su funcionamiento es grafico posee casi las mismas funciones que el Explorer, pero con una interfaz que permite "arrastrar y soltar" facilitando con creces su uso y una ventaja notoria es que ofrece soporte para el aprendizaje incremental.

Workbench: Es una interfaz gráfica de usuario unificada que combina los otros tres (y cualquier complemento que el usuario haya instalado) en una aplicación. El banco de trabajo es altamente configurable, lo que permite al usuario especificar qué aplicaciones y complementos aparecerán, junto con la configuración relacionada con ellos.

Simple CLI: Es una abreviación de Simple Command-Line Interface (Interfaz Simple de Línea de Comandos). Esta interfaz proporciona una consola para poder introducir órdenes desde la línea de comandos, permitiendo ejecutar cualquier operación soportada por WEKA de forma directa. Ya prácticamente no

se usa debido a que se fue recubriendo WEKA con otras interfaces más fáciles de utilizar para el usuario común.

- *Beneficios y problemas.*

Como toda herramienta, WEKA tiene sus pros y su contra, pero que sin duda marcan una dominación las ventajas y razones que favorecen el uso de este software, más que sus carencias.

Beneficios: Software de libre distribución bajo la licencia publica general de GNU. Corre en casi cualquier plataforma. Portabilidad. Contiene una gran colección de técnicas para el pre-procesamiento de datos. Fácil uso para principiantes, debido a uso interfaz gráfica de usuario. Entrega de resultados sencillos de interpretar.

Problemas: Escasa documentación orientada al usuario. Los algoritmos de WEKA no cubren el área de modelado de secuencias.

c. Otras herramientas.

- *De tipo libre.*

KNIME (Konstanz Information Miner): Es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. Está construido bajo la plataforma Eclipse, programado principalmente en Java y distribuido bajo la licencia de GNU-GPL.

Permite una fácil integración de nuevos algoritmos, manipulación de datos y métodos de visualización como modelos. Compatible con WEKA, también incluye métodos estadísticos a través del uso de la herramienta R.

Orange: Es un software informático que opera en múltiples plataformas, sirve para realizar minería de datos y análisis predictivo desarrollado en la facultad de informática de la Universidad de Ljubljana (Eslovenia). Consta de varios componentes desarrollados en C++ que implementan algoritmos de minería de datos y se distribuye bajo la licencia GNU-GPL.

Los componentes pueden ser controlados a través de un entorno grafico o sino también incluye un entorno de secuencias de comandos para la creación de prototipos de nuevos algoritmos y esquemas de prueba utilizando Python.

- *De tipo comercial.*

SPSS Clementine: Es una herramienta estadística integrada de minería de datos, programada en Java, muy usada por la empresa que implementan minería de datos, tiene la capacidad para operar con grandes bases de datos, posee una sencilla interfaz y opera en múltiples plataformas, es capaz de trabajar distintas técnicas como reglas se asociación, regresión, clustering y clasificación, solo por mencionar algunas características. Es de código cerrado (Software propietario).

Además, posee una arquitectura distribuida (cliente/servidor). El programa consiste en un módulo base y módulos anexos, que se han ido actualizando constantemente con nuevos procedimientos estadísticos. Cada uno de estos módulos se compra por separado.

El 2009 Internacional Business Machines Corporación (IBM) compra este software, dándole más funcionalidades y ayudando a la organización que compre el software, a anticipar los cambios de manera que pueda planificar e implementar

estrategias que mejoren los resultados dentro de la empresa, mediante la aplicación del análisis predictivo.

Stadistica es un paquete estadístico usado en el ámbito de la minería de datos, su primera aversión fue lanzada sobre MS-DOS. Dispone de un sistema propio de archivos y es un software no libre. Posee la característica de exportar sus modelos a otros formatos como XML, C/C++, Visual Basic, Java, (Structured Query Language, SQL).

El programa consta de varios módulos. El principal de ellos es la Base, que implementa las técnicas estadísticas más comunes. Y además se complementa con otros módulos más específicos, tales como, “Advanced”, “QC”, “Data Miner”.

Esta última plataforma para Data Mining de stadistica, ofrece el más amplio y eficiente sistema de herramientas intuitivas para todo el proceso de minería de sus datos. Además, dispone de numerosos algoritmos y técnicas de Data Mining.

3.2. Caso práctico

3.2.1. Elección de enfoque, técnicas y algoritmos.

Como ya se ha mencionado anteriormente en este proyecto, para obtener los diversos modelos, que son utilizados para hacer las predicciones, es necesario tomar una serie de decisiones antes de comenzar todo el proceso.

Es obligatorio determinar y elegir estas tres cosas con las que se va a desarrollar el proceso. Elegir el tipo de tarea o enfoque a implementar. Elegir las técnicas para obtener el modelo. Elegir el algoritmo de minería de dato que resuelva la tarea y obtenga el tipo de modelo.

3.2.1.1. Herramienta escogida.

Dentro de las herramientas que se han averiguado y evaluado, hubo una que más destacó por sobre el resto, esta fue "WEKA", se ha podido observar que es una de las herramientas libres mejor catalogadas en el transcurso de los años, esto debido a su potencial funcionamiento y su fácil aprendizaje de uso gracias a su interfaz gráfica de usuario.

Es el mejor software para llevar a cabo el proceso de Data Mining ya que principalmente es un software libre, además contiene muchas herramientas para los datos pre- procesamiento, clasificación, regresión, clustering, reglas de asociación, y visualización y entrega resultados gráficos e interpretables fácilmente.

Se ha obtenido mucha documentación en internet, respecto al uso y manejo de la plataforma de aprendizaje WEKA, esto contrarresta el hecho de que este software no posea "Lenguaje Español". Además, se han observado varios videos-tutoriales, haciendo aún más sencillo su uso, en repositorios web.

Otro fuerte que tiene, es que está disponible libremente bajo la licencia pública general de GNU, también es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma. Incluye una gran colección de técnicas para pre-procesamiento de datos y modelado.

Todas las técnicas que se pueden utilizar, se fundamentan en la aceptación de que los datos están disponibles en un fichero plano (flat file) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (numéricos, nominales u otros).

Otra útil ventaja, que no se ha tenido que manejar, pero que demuestra lo poderosa que es esta herramienta, es que proporciona acceso a bases de datos vía (Structured Query Language, SQL), gracias a la conexión JDBC (Java Database Connectivity) y permite realizar consultas en este lenguaje directamente a la base de datos, en este caso, este ventaja no ha influido en la elección de WEKA, ya que no se ha usado bases de datos en nuestro problema, solo se menciona en este informe, para respaldar que es un software poderoso y multifuncional.

WEKA parece ser una herramienta muy potente, que cuenta con muchas funcionalidades y características provechosas, sus algoritmos cubren casi todas las áreas de la minería de datos, excepto la del modelado de secuencias, pero que en este caso no ha sido requerido este modelado, en consecuencia, cumple con todo lo que se ha requerido aplicar.

3.2.1.2. Enfoques utilizados.

Estas tareas resuelven problemas de tipo predictivo, que es lo que se busca, con el fin de adquirir conocimiento útil sobre los alumnos que están cursando o que van a cursar los cursos de informática.

No se ha utilizado el enfoque de Agrupamiento o Clusters, esto debido a que se han analizado datos que ya están etiquetados con una clase, en nuestro problema ya se conocen las clases, lo que se pretende es predecir valores futuros y saber a qué clase se clasificarán.

3.2.1.3. Técnicas y Algoritmos empleados.

Mediante la herramienta escogida (WEKA), se han utilizado técnicas como reglas de clasificación y árboles de decisión, porque son más fáciles de tomar e interpretar

la salida de los datos que se obtuvieron mediante los determinados algoritmos, puesto que los resultados obtenidos serán expresados como reglas de la forma "Si x ENTONCES y", detectando de esta forma a los alumnos con problemas y tomando una decisión pertinente para esos casos.

Los algoritmos escogidos en este proyecto y aplicados sobre las técnicas de Data Mining, son varios, entre ellos están los algoritmos de inducción de reglas de clasificación: JRip, NNge, OneR, Prism y Ridor. Asimismo, se han usado algoritmos de árboles de decisión, tales como: J48, SimpleCart, NBTree, REPTree, ADTree y RandomTree. Diversos algoritmos que generar distintos modelos, algunos han sido comparado obteniéndose buenas conclusiones.

3.2.2. Documentación Knowledge Discovery in Database.

Para este problema, en particular, se ha utilizado el proceso de extracción de conocimiento en bases de datos propuesto por los autores Fayyad, Piatetsky-Shapiro & Smyth, (1996). El cual lo definen como "El proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensible en los datos". En efecto, constará de las cinco etapas que ellos proponen, para llevar a cabo el proceso de minería de datos.

Se ha escogido por la importancia que estos autores le dan a la etapa de Pre-procesamiento de los datos, esta es una etapa independiente de gran consideración, ya que garantiza la calidad de los datos para el resultado final que se obtenga. A diferencia de los otros autores (Han & Kamber, 2001), que establecen la etapa de pre-proceso de los datos junto con la transformación y selección, restando

importancia a esta etapa. A mi parecer es significativo tomar en cuenta el pre-procesamiento de los datos, como un paso aparte de los demás.

Es por esto que se ha optado por la elección del proceso Knowledge Discovery in Database según los autores Fayyad, Piatetsky-Shapiro, declinando así la definición que dan Han y Kamber.

3.2.2.1. Selección de Datos.

Una vez que se ha establecido el objetivo general y los objetivos específicos, todo esto con el fin de obtener conocimiento útil al final del proyecto. Se ha procedido a recopilar los datos de los alumnos, esto mediante el coordinador de la escuela de Ingeniería de Sistemas e Informática Edson Huertas, quien ha autorizado la realización de la encuesta a los alumnos dentro de la Filial Ilo.

Además, se logró tener precisión de la información que han llenado los alumnos en las encuestas gracias al registro académico que lleva la escuela de los alumnos dando fe que la información sea correcta, y se han facilitado más variables que puedan incidir en el rendimiento académico de los alumnos, esto para que las variables tuvieran un alcance mayor en el proceso Knowledge Discovery in Database, y se puedan inferir reglas aún más abundantes de contenido y que reflejen claramente lo que se pretende alcanzar.

Todas estas variables son importantes para adquirir información útil del comportamiento del alumno y así alertar a tiempo al profesor cuando un estudiante se encuentra en riesgo de reprobar la asignatura de programación.

Al provenir de distintas fuentes los datos, se ha tenido que unificar dichos datos, permitiendo observar de mejor manera lo que se iba a analizar. Esto es

provechoso, ya que para quién trabaja en el proceso o alguien que observe los datos desde afuera, claramente va a tener una idea más evidente de que es lo que se va a examinar posteriormente.

Cabe recalcar, que todos los datos entregados por el coordinador y por el Registro de la escuela profesional han sido resguardados rigurosamente, manteniendo la privacidad de los mismos, a fin de evitar mal uso de ellos y solo ser usados con fines benéficos en este proyecto.

3.2.2.2. Pre-procesamiento de los datos.

En esta etapa de gran trascendencia, que tiene por objetivo garantizar y mejorar la calidad de los datos que se han analizados, para que al final el conocimiento descubierto sea útil y novedoso.

Se han preparado y limpiado los datos de los archivos Excel, los alumnos que eran de distintos ciclos y llevaban cursos de informática entre los años ya mencionados y tenían parte importante de las variables a estudiar.

Algunos problemas encontrados en el pre-procesamiento de los datos son. Casos especiales y aislados de algunos alumnos que no coincidía la información que ha proporcionado, variables que no existen en el registro de la escuela, es decir, faltaban muchos datos y también alumnos que no existen, tampoco contaban con mucha información en los archivos de registro, por ende, se ha optado por la eliminación de estas tuplas y la eliminación de datos inconsistentes.

Es así como también se han limpiado estos valores nulos, mediante el rechazo de las tuplas completamente vacías. Se ha usado la reducción de datos, por medio de un estudio previo, y se ha considerado que variables de todas las que se

tienen, son las que realmente pueden incidir en el rendimiento académico de los alumnos, algunas variables no tienen una gran consecuencia y no fueron tomadas en cuenta al momento en que ha convertido todo en una gran tabla, esto implica que al reducir la dimensión de la tabla final se ha evitado la entropía.

Luego de haber normalizado los datos, se genera el archivo "datafinal.xls" con las 18 variables que determinen el comportamiento, en términos de la calificación final, de los alumnos del primer curso de programación y una variable correlativa.

Tabla 1

Descripción de las variables.

Variable	Descripción
ID	Variable numérica, que funciona como correlativo para saber de qué alumno se trata, no participa de las reglas generadas.
ESCUELA	Variable que indica la escuela del alumno.
NINGRESO	Variable de tipo numérica, indica el puntaje de entrada a la universidad.
PMAT	Variable de tipo numérica, indica el puntaje en matemáticas que obtuvo el alumno.
NOTA	Variable de tipo numérica, que está relacionada con la nota de enseñanza secundaria, que obtuvo el alumno.
COLEGIO	Variable de tipo numérica, que indica el colegio del cual proviene un alumno (particular, estatal, otros.)
TINGRESO	Variable de tipo numérica, que indica el número de años que se demoró en ingresar a la universidad.
ASISTENCIA	Variable de tipo numérica, medida en términos porcentuales, e indica el porcentaje de asistencia de un alumno al curso.
CICLO	Variable de tipo numérica, que indica el ciclo del año en el cual el alumno curso la asignatura.

DOCENTE	Variable que indica el nombre del profesor que impartió el curso.
EXAMEN1	Variable de tipo numérica, que indica la nota del alumno del examen uno.
EXAMEN2	Variable de tipo numérico, que indica la nota del alumno del examen dos.
NFP	Variable de tipo numérico, que indica la nota final de las practicas obtenida por el alumno.
NFT	Variable de tipo numérico , que indica la nota final de los trabajos del ciclo obtenida por el alumno.
NP	Variable de tipo numérico, indica la nota con que el alumno se presenta al final del curso, determina si le corresponde rendir examen o no.
EXAMENFINAL	Variable de tipo numérico, que indica la nota obtenida en el examen, en caso de que el alumno lo haya rendido.
NFINAL	Variable de tipo numérico, que indica la nota final del alumno en el curso, después de haber rendido examen (si APROBO O DESAPROBÓ).
TERMINO	Variable de tipo booleana, que hace alusión a si el alumno termino el ramo en el semestre que lo inscribió, de lo contrario se asume que lo renuncio (V =termino y F=(no termino).
NMATR	Variable de tipo numérico, que indica si es la primera o segunda vez que el alumno está tomando el curso.

3.2.2.3. Transformación de los datos.

Para facilitar el análisis y estudio de los datos se transforma el esquema de la tabla única que se dispone, por un esquema de etiquetado, en donde, los valores numéricos se han cambiado a formato nominal o categórico, cada variable se ha catalogado en una índole diferente, a excepción de todo lo que se relaciona con las notas, estas variables están clasificadas dentro de la misma escala.

A continuación, se dan a conocer las escalas con las que se produjeron las etiquetas correspondientes para cada variable incidente. En la columna de la izquierda aparece el dato con el que aparecían las variables y en la columna de la derecha, aparece la etiqueta por la cual fue sustituida.

a. Variable ESCUELA.

En la siguiente variable la escuela profesional de ingeniería de sistemas e informática fue remplazada por “EPISI” y la otra escuela por “OTRA”, para su posterior análisis.

Tabla 2

Etiqueta de la variable Escuela.

Datos	Etiqueta
Ingeniería de Sistemas e Informática	Episi
Otra Escuela Profesional	Otra

b. Variable NINGRESO.

Esta tabla cataloga los puntajes obtenidos en el examen de admisión realizado, para los alumnos que han rendido la Prueba de Aptitud Académica desde el año 2014.

Tabla 3

Etiqueta de la variable NINGRESO.

Datos	Etiqueta
1-25	C
26-50	B
51-75	A
76-100	AD

c. Variable PMAT.

Esta tabla etiqueta los puntajes de matemática en la secundaria, clasificándolos en “B”, “A”, “AD” de acuerdo al medio de calificación del ministerio de educación del Perú.

Tabla 4

Etiqueta de la Variable PMAT

Datos	Etiqueta
11-13	B
14-17	A
18-20	AD

d. Variable NOTA.

Esta tabla etiqueta a las notas que obtuvieron en la secundaria (con escala de notas del 11 al 20), catalogándolos en “Muy Bajo”, “Bajo”, “Regular”, “Bueno” y “Muy Bueno”.

Tabla 5

Etiqueta de la variable NOTA.

Datos	Etiqueta
11-12	Muy Bajo
13-14	Bajo
15-16	Regular
17-18	Bueno
19-20	Muy Bueno

e. Variable COLEGIO.

Esta tabla etiqueta los datos de colegio, catalogándolos en “Particular”, “Estatal” y “Otros”, esto de acuerdo al colegio del cual los alumnos provenían.

Tabla 6

Etiqueta de la variable COLEGIO

Datos	Etiqueta
1	Particular
2	Estatal
3	Otros

f. Variable TINGRESO.

Esta tabla etiqueta los datos del tiempo que se demoró en entrar a la universidad, catalogándolos en "Normal", "Mucho", "Excesivo".

Tabla 7

Etiqueta de la variable TINGRESO

Datos	Etiqueta
1-2	Normal
3-5	Mucho
6-17	Excesivo

g. Variable ASISTENCIA.

Esta tabla etiqueta los datos de la asistencia del alumno en el curso, catalogándolos en "Muy Baja", "Baja", "Regular", "Buena" y "Muy Buena". Hay casos excepcionales, en donde las asistencias están justificadas previamente con el profesor y otros cursos en donde no se aplicó asistencia, por ende, se deja como etiquetado "Justificado" y "Sin Información" respectivamente.

Tabla 8

Etiqueta de la variable ASISTENCIA

Datos	Etiqueta
0-10	Muy Baja
11-20	Baja
21-30	Regular
31-40	Buena
41-50	Muy Buena
Just	Justificado
Celda vacía	Sin información

h. Variable CICLO.

Esta tabla etiqueta los datos del semestre en que se impartió el curso catalogándolos en "Primer Ciclo" y "Segundo Ciclo". Hay ciertos cursos, donde la información que se me entrego no se especificaba el ciclo al cual pertenecía el curso, estas celdas se las etiqueto como "Sin Información".

Tabla 9

Etiquetado de la variable CICLO

Datos	Etiqueta
1	Primer Ciclo
2	Segundo Ciclo
Vacía	Sin información

i. Variable DOCENTE.

Esta tabla etiqueta los datos del profesor que impartió el determinado curso, catalogándolos en "DOCENTE1" para el docente que se asigne y "DOCENTE2"

en el caso que fuera el profesor que se asigne más docentes impartiendo el mismo curso.

Tabla 10

Etiqueta de la variable DOCENTE

Datos	Etiqueta
Primer Docente	Docente uno
Segundo Docente	Docente dos

j. Variable EXAMEN1.

Esta tabla etiqueta los datos de las notas de los alumnos que obtuvieron en el examen1, catalogándolos en “Muy Mal”, “Mal”, “Regular”, “Bien” “Muy Bien”. Hay ciertos alumnos que no rindieron el examen1 ya sea porque renunciaron la asignatura o por algún caso especial, en estos casos se les aplica el etiquetado de “No Rinde”.

Tabla 11

Etiqueta de la variable EXAMEN1

Datos	Etiqueta
0-4	Mal
5-8	MuyMal
9-13	Regular
14-16	Bien
17-20	MuyBien
Celda Vacía	NoRinde

k. Variable EXAMEN2.

Esta tabla etiqueta los datos de las notas de los alumnos que obtuvieron en el examen2, catalogándolos en “Muy Mal”, “Mal”, “Regular”, “Bien” “Muy Bien”. Hay ciertos alumnos que no rindieron el examen2 ya sea porque renunciaron la asignatura o por algún caso especial, en estos casos se les aplica el etiquetado de “No Rinde”. Se aplica la misma escala que para el examen1.

Tabla 12

Etiquetado de la variable EXAMEN2

Datos	Etiqueta
0-4	Mal
5-8	MuyMal
9-13	Regular
14-16	Bien
17-20	MuyBien
Celda Vacía	NoRinde

l. Variable NFP.

Esta tabla etiqueta los datos de la nota final de la práctica que obtuvieron los alumnos, asignándoles etiquetas como “Muy Mal”, “Mal”, “Regular”, “Bien” “Muy Bien”. Hay ciertos alumnos que no tienen información en esta variable ya que quizás renunciaron a la asignatura, en estos casos se les etiqueta como “No Rinde”.

Tabla 13

Etiqueta de la variable NFP (Nota final de practica).

Datos	Etiqueta
0-4	Mal
5-8	Muy Mal
9-13	Regular
14-16	Bien
17-20	Muy Bien
Celda Vacía	No Rinde

m. Variable NFT.

Esta tabla etiqueta los datos del promedio final que obtuvieron los alumnos en los trabajos en clases, o en el trabajo del ciclo, dependiendo de lo que se haya implementado en dicho curso, catalogándolos en “Muy Mal”, “Mal”, “Regular”, “Bien” “Muy Bien”. Hay ciertos alumnos que no tienen información en esta variable ya que quizás pueden haber renunciado la asignatura, en estos casos se les etiqueta como “No Rinde”.

Tabla 14

Etiqueta de la variable NFT (Nota final del trabajo).

Datos	Etiqueta
0-4	Mal
5-8	Muy Mal
9-13	Regular
14-16	Bien
17-20	Muy Bien
Celda Vacía	No Rinde

n. Variable NP.

Esta tabla etiqueta los datos de la nota de presentación del alumno antes del examen (si es que lo tuviese que rendir), asignándole etiquetas como “Muy Mal”, “Mal”, “Regular”, “Bien” “Muy Bien”.

Hay ciertos alumnos que no tienen información en esta variable ya que quizás pueden haber renunciado la asignatura, en consecuencia, no tendrán nota de presentación, en estos casos se les etiqueta como “Sin Nota”.

También si algún alumno no cumplió con los requisitos mínimos del curso, es decir, falto demasiado, cuando la asistencia para aprobar es el 75%, o si no fue a dar las evaluaciones, sin presentar algún tipo de justificación real, se le considera como NCR y se les etiqueta como “No Cumple Requisitos”.

Tabla 15

Etiqueta de la variable NP (Nota de presentación).

Datos	Etiqueta
0-4	Mal
5-8	MuyMal
9-13	Regular
14-16	Bien
17-20	MuyBien
Celda Vacía	NoRinde
NCR	NoCumpleRequisitos.

ñ. Variable EXAMENFINAL.

Esta tabla etiqueta los datos de la nota de aplazados del alumno, catalogándolos en "Muy Mal", "Mal", "Regular", "Bien" "Muy Bien". Hay ciertos alumnos que no tienen información en esta variable, debido a que simplemente no rindieron el examen o no alcanzaron los requisitos mínimos, por el motivo de que se eximieron o simplemente no fueron a darlo, cuando suceden estos se les etiqueta como "No Rinde".

Tabla 16

Etiquetado de la variable APLAZADOS.

Datos	Etiqueta
0-4	Mal
5-8	MuyMal
9-13	Regular
14-16	Bien
17-20	MuyBien
Celda Vacía	NoRinde

o. Variable NFINAL.

Esta tabla etiqueta los datos de la nota final que obtuvo el alumno, catalogándolos en "Aprobó", "Desaprobó".

Tabla 17

Etiqueta de la variable NFINAL (Nota final)

Datos	Etiqueta
1	Aprobó
2	Desaprobó

p. Variable TERMINO.

Esta tabla etiqueta los datos de la finalización del curso, es decir, si el alumno terminó el curso en el ciclo que lo inscribió. Si es así y el alumno terminó el curso, la etiqueta será "SI", de lo contrario si el alumno no termino el curso, por diferentes motivos, la etiqueta será "NO".

Tabla 18

Etiqueta de la variable TERMINO

Datos	Etiqueta
True	Si
False	No

q. Variable NMATR.

Esta tabla etiqueta los datos de la nota de examen del alumno, catalogándolos en "Primera" y "Segunda", de acuerdo, a si es primera o segunda vez que lleva el curso.

Tabla 19

Etiqueta de la variable NMATR (Número de matrícula).

Datos	Etiqueta
1	Primera
2	Segunda

Cabe destacar, algo no menor dentro de esta etapa, es que después del etiquetado se guarda el archivo con el formato ".CSV (delimitado por comas)", este formato es un tipo de archivo que guarda los datos separados por puntos y comas, estos ya han sido transformados para que más tarde puedan ser reconocidos sin

problema por la herramienta WEKA. Software con el cual trabajaremos todos los datos y aplicaremos las técnicas de minería de datos.

Como se observa se hace por tanto necesario un pre-procesamiento previo a la transformación, en el que se disminuya el tamaño del conjunto almacenado. Así se han consolidado los datos de una forma apropiada y necesaria para la siguiente etapa del proceso Knowledge Discovery in Database.

3.2.3. Minería de Datos.

Una vez que se ha revisado el conjunto de datos y el número de instancias de las clases "Aprobó" y "Desaprobó", se encuentra **balanceado**, es decir, el número de alumnos aprobados es muy similar al de reprobados, tal cual se observa en la siguiente imagen, donde la clase "Aprobó" tiene 242 instancias mientras que la clase "Desaprobó" posee 248 instancias.

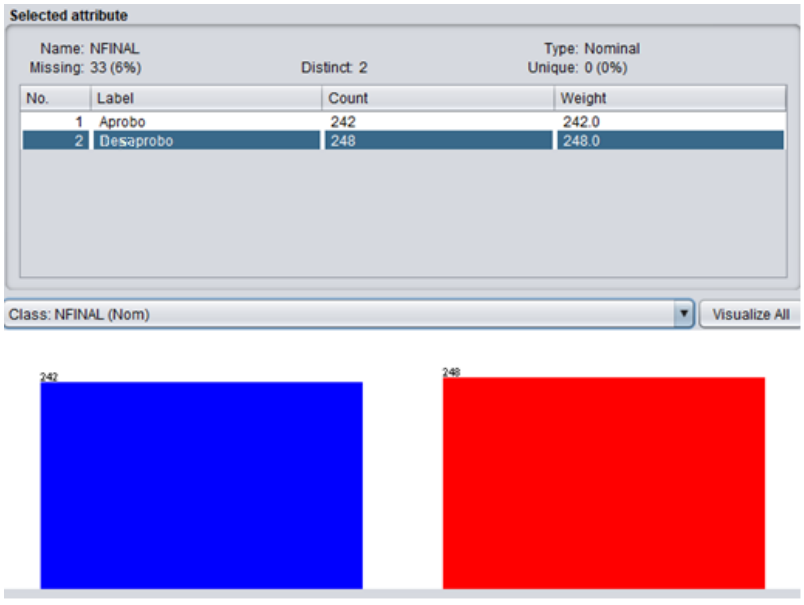

Figura 14. Grafico del atributo NFINAL

Al estar balanceados los datos, nos evitamos que los algoritmos de clasificación tengan en su etapa de entrenamiento una clase mayoritaria, lo que lleva a clasificar en la etapa de prueba con baja sensibilidad a los elementos de la clase minoritaria.

También se ha desechado de utilizar algún filtro o algoritmo que posea WEKA para poder balancear el conjunto de datos. Después de realizar las tareas de pre-proceso y transformación de los datos, se observa que se cuenta con.

Los 10 ficheros de entrenamiento y testeo con los 18 atributos influyentes. Los 10 ficheros de entrenamiento y testeo con los 18 atributos influyentes considerando los costos de clasificación. Los 10 ficheros de entrenamiento y testeo con los atributos que inciden hasta el EXAMEN1. Los 10 ficheros de entrenamiento y testeo con los atributos que inciden hasta el EXAMEN2.

Se han descrito estos cuatro experimentos y las técnicas de minería de datos que han sido utilizadas, para la obtención de los modelos de predicción del rendimiento académico de los estudiantes.

Estos son realizados con el objetivo de obtener la máxima exactitud de clasificación y obtener reglas.

Para el experimento uno, se han tomado 10 algoritmos de clasificación disponibles por la herramienta seleccionada (WEKA), considerando todos los atributos influyentes, mediante la obtención de árboles de decisión y reglas de clasificación, se han conseguido reglas de interés y de fácil comprensión, para que puedan ser leídas por cualquier usuario no experto en el tema.

Las reglas son de manera simple, con un antecedente y un consecuente, son del tipo Si – Entonces, y son fácilmente comprensibles de presentar el

conocimiento. Las reglas determinan una instancia de datos a la clase señalada por el consecuente, si es que los atributos de predicción satisfacen las condiciones declaradas en el antecedente.

También se han obtenido árboles con nodos internos y nodos hojas, este modelo predictivo, en el cual una instancia es clasificada siguiendo un camino de condiciones cumplidas desde la raíz hasta llegar a una hoja, la cual va a corresponder a una clase etiquetada. Hay que mencionar que un árbol de decisión puede convertirse fácilmente a una regla de clasificación.

Los 10 algoritmos utilizados son: Jrip, NNge, OneR, Prism y Ridor, todo estos pertenecientes a la inducción de reglas de clasificación, y ADTree, J48, REPTree, SimpleCart y NBTree, pertenecientes a los árboles de decisión.

El fichero de datos fue particionado en 10 particiones (ficheros de entrenamiento y ficheros de prueba), mediante la opción de "Crossvalidation", que lo que hace es calcular el porcentaje de aciertos esperado haciendo una validación cruzada de "n" hojas (donde "n" es el número de particiones).

Para el segundo experimento, se ha abordado de una forma diferente, ya que se ha realizado una "clasificación sensible al costo", esto quiere decir que ha intentado mejorar el porcentaje de aciertos de la clase "Desaprobó", utilizando un meta-clasificador. Es un meta-clasificador, porque utiliza a otro clasificador base.

La herramienta WEKA permite realizar clasificación teniendo en cuenta el costo, para lo cual se ha utilizado el CostSensitiveClassifier, asociándole una matriz de costo y el clasificador a utilizar, de tal manera que se cosechen mejores resultados de clasificación.

Los primeros dos experimentos, tienen por objetivo principal alcanzar la máxima exactitud de clasificación posible, para que cuando sea añadido una nueva instancia (alumno), se pueda predecir la clase a la que pertenecerá.

Para el tercer experimento, se han considerado solo 11 variables, estas son: Escuela, Ningreso, Nota, PMat, Asistencia, TIngreso, Ciclo, Docente, Colegio, Examen1 y NFinal. Lo que se pretende buscar es mediante las reglas de clasificación y árboles de decisión, obtener reglas que puedan aportar y alertar a tiempo al profesor, es decir, solo se considera hasta el examen1, para en un futuro cuando se haga uso de la aplicación, con solo tener la nota del examen1, se pueda inferir, dentro de que clase está clasificado el alumno. ("Aprobó" y "Desaprobó"). Esto sin duda, ayudará a que el profesor pueda tener en consideración dichos alumnos y tome decisiones pertinentes para mejorar obviamente su rendimiento académico.

Para el cuarto y último experimento, se ha añadido una variable más con respecto al experimento anterior, estas variables son: examen2, esto con el fin de encontrar reglas importantes y ricas en contenido, para cuando el alumno ya haya rendido el examen1 y el examen2, el objetivo de estos dos experimentos (tres y cuatro), es posteriormente analizarlos y determinar cuándo es más pertinente que el profesor tome una acción positiva. Para los experimentos tres y cuatro, se han implementado cuatro algoritmos: SimpleCart, J48, PART y Ridor.

3.2.4. Interpretación y Evaluación de resultados.

En este grupo se muestra y comenta los modelos diversos que se han obtenidos, y han sido generados mediante los algoritmos que mejores resultados de clasificación

han obtenido en la etapa de minería de datos. Los mejores algoritmos que se han experimentado, obteniendo los mejores resultados después del experimento dos, fueron: OneR, Prism y J48.

Todo se ha interpretado ingresando los datos al programa weka se obtuvo valores exactos por ser un software dedicado a la minería de datos.

```
NP:
        Regular -> Aprobo
        Mal     -> Reprobo
        MuyMal  -> Reprobo
        NoRequisitos    -> Reprobo
        SinNota -> Reprobo
        Bien    -> Aprobo
        MuyBien -> Aprobo
(465/490 instances correct)
```

Figura 15. Reglas usando OneR y considerar el costo de la clasificación

El algoritmo OneR descubre pocas reglas y de un solo atributo NP (Nota de Presentación), dando a conocer reglas lógicas de las clases de este atributo, y las clasifica con respecto a las clases del atributo NFINAL (Nota Final).

```
Prism rules
----------
If NP = Regular then Aprobo
If NP = Bien then Aprobo
If NP = MuyBien then Aprobo
If EXAMEN2 = Bien then Aprobo
If EXAMENFINAL = Bien
   and NOTA = MuyBien then Aprobo
If EXAMENFINAL = Regular
   and EXAMEN2 = Mal then Aprobo
If EXAMENFINAL = Regular
   and CICLO = II CICLO then Aprobo
If EXAMENFINAL = Bien
   and ESCUELA = OTRA then Aprobo
If EXAMENFINAL = Regular
   and EXAMEN1 = Mal then Aprobo
If EXAMENFINAL = Regular
   and ESCUELA = OTRA
   and ASISTENCIA = SinInformacion then Aprobo
If EXAMENFINAL = Mal
   and NINGRESO = Bajo then Aprobo
If NOTA = MuyBajo
   and NFP = Regular then Aprobo
If EXAMENFINAL = Mal
   and NFT = NoRinde then Aprobo
If EXAMENFINAL = Mal
   and CICLO = II CICLO
   and NINGRESO = Regular then Aprobo
If EXAMENFINAL = Regular
   and NFP = Regular
   and ESCUELA = EPISI then Aprobo
If EXAMEN1 = MuyBien
   and PMAT = Bueno then Aprobo
If NFP = Bien
   and NFT = MuyBien
   and COLEGIO = Otros then Aprobo
If EXAMENFINAL = Mal
   and NFP = Bien
   and TINGRESO = Mucho then Aprobo
If NFP = Bien
   and NFT = NoRinde
   and COLEGIO = Estatal
   and NOTA = Bien
   and TINGRESO = Normal then Aprobo
If NFP = Regular
   and TINGRESO = Mucho
   and ESCUELA = OTRA then Aprobo
```

Figura 16. Reglas usando Prism y considerar el costo de clasificación aprobó.

En las reglas del algoritmo Prism se observa que este algoritmo descubre una gran cantidad de reglas. También se observa, que genera varias reglas interesantes que hay que tomar en cuenta por los docentes, donde se aplicaran dichas reglas, para mostrar en pantalla a que clase pertenece el alumno.

SI EXAMENFINAL = Regular Y NFP = Regular Y ESCUELA = EPISI **ENTONCES** Aprobó

SI EXAMEN1 = Muy Bien Y PMAT = Bueno **ENTONCES** Aprobó

SI EXAMENFINAL = Regular Y EXAMEN2 = Mal **ENTONCES** Aprobó

La segunda regla dice que, si el alumno tiene habilidades en matemática, o sea, ha obtenido buen puntaje en la PMAT y que además en el EXAMEN1 le fue muy bien, entonces está clasificado dentro de los aprobados. Esta regla fácilmente se puede complementar con las reglas obtenidas en el experimento cuatro.

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

NP = Regular: Aprobo (71.71)
NP = Mal
|   EXAMENFINAL = NoRinde: Reprobo (27.89)
|   EXAMENFINAL = MuyMal: Reprobo (31.21/1.99)
|   EXAMENFINAL = Regular
|   |   TINGRESO = Mucho: Reprobo (2.66)
|   |   TINGRESO = Normal: Aprobo (7.3/1.33)
|   |   TINGRESO = Excesivo: Aprobo (0.0)
|   EXAMENFINAL = Bien: Aprobo (1.99)
|   EXAMENFINAL = Mal
|   |   CICLO = I CICLO: Reprobo (2.66)
|   |   CICLO = SinInformacion: Reprobo (7.97/1.33)
|   |   CICLO = II CICLO: Aprobo (1.99)
|   EXAMENFINAL = MuyBien: Reprobo (3.98)
NP = MuyMal: Reprobo (225.08/3.32)
NP = NoRequisitos: Reprobo (19.92)
NP = SinNota: Reprobo (13.28)
NP = Bien: Aprobo (62.41)
NP = MuyBien: Aprobo (9.96)
```

Figura 17. Árbol obtenido usando J48 y considerando el costo de clasificación.

En el árbol de decisión de la Tabla, se observa que tiene como nodo raíz, el atributo “NP”, siendo sus hojas, las clases del atributo “EXAMENFINAL. También se obtienen buenas reglas que hay que considerar y prestar importancia.

SI NP = Mal Y EXAMENFINAL = Mal Y Ciclo = I CICLO **ENTONCES** Desaprobó

SI NP = Mal Y EXAMENFINAL = Mal Y Ciclo = II CICLO **ENTONCES** Aprobó

SI NP = Mal Y EXAMENFINAL = Regular Y TINGRESO= Normal **ENTONCES** Aprobó

SI NP = Mal Y EXAMENFINAL = Regular Y TINGRESO = Excesivo **ENTONCES** Aprobó

SI NP = Mal Y EXAMENFINAL = Regular Y TINGRESO = Mucho **ENTONCES** Desaprobó**.**

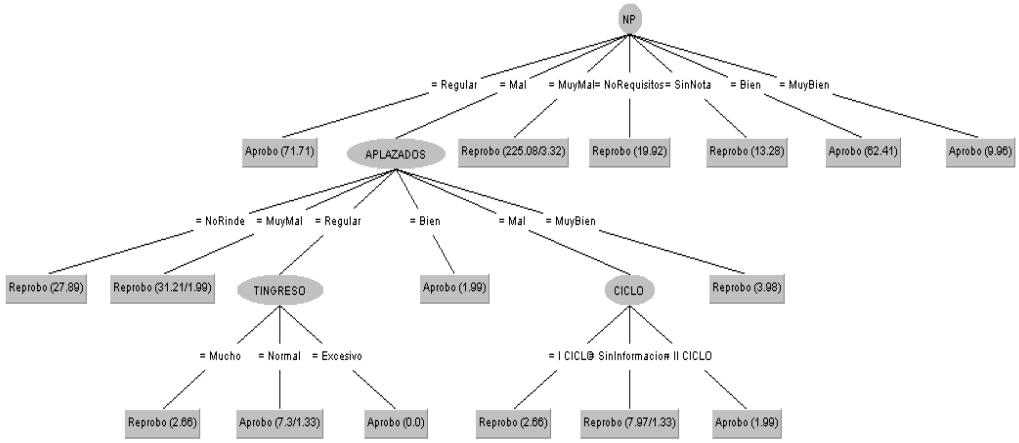

Figura 18. Árbol obtenido por el algoritmo J48.

En el caso del experimento tres, donde se pretende dar una solución después de saber la nota del EXAMEN1, cuando aún hay tiempo de beneficiar el rendimiento del alumno, se generan diversas reglas y conclusiones que benefician a los estudiantes.

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

EXAMEN1 = Mal
|   ESCUELA = OTRA
|   |   ASISTENCIA = SinInformacion
|   |   |   CICLO = I CICLO
|   |   |   |   DOCENTE = DOCENTE1: Aprobo (6.0/1.0)
|   |   |   |   DOCENTE = DOCENTE2: Reprobo (6.0/1.0)
|   |   |   CICLO = SinInformacion: Aprobo (18.0/5.0)
|   |   |   CICLO = II CICLO: Aprobo (0.0)
|   |   ASISTENCIA = MuyBuena: Aprobo (7.0/2.0)
|   |   ASISTENCIA = Buena: Reprobo (2.0)
|   |   ASISTENCIA = Baja: Aprobo (0.0)
|   |   ASISTENCIA = Regular: Aprobo (0.0)
|   |   ASISTENCIA = MuyBaja: Aprobo (0.0)
|   |   ASISTENCIA = Justificado: Aprobo (0.0)
|   ESCUELA = EPISI: Reprobo (45.0/14.0)
EXAMEN1 = Bien: Aprobo (110.0/25.0)
EXAMEN1 = MuyMal: Reprobo (161.0/27.0)
EXAMEN1 = MuyBien: Aprobo (64.0/8.0)
EXAMEN1 = Regular: Aprobo (58.0/22.0)
EXAMEN1 = NoRinde: Reprobo (13.0)
```

Figura 19. Reglas obtenidas usando el algoritmo J48

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = I CICLO Y DOCENTE= DOCENTE1

ENTONCES Aprobó

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = I CICLO Y DOCENTE= DOCENTE2

ENTONCES Desaprobó

Esto se puede deber, a que el docente1 ha impartido más cursos de informática en otras escuelas, por ese motivo tiene más posibilidades de cumplir con la primera regla y que el resultado final sea que el alumno haya "Aprobado".

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = II CICLO **ENTONCES** Aprobó

SI EXAMEN1 = Mal Y ESCUELA = EPISI **ENTONCES** Desaprobó

Como se aprecia, al comenzar con el EXAMEN1 con una mala nota, se tiende a pensar que podría seguir esa tendencia y termina con mala nota final, pero todo depende de los otros atributos, en la primera regla influye el semestre, si le va mal en el primer examen y es el segundo ciclo en el cuál ha cursado el curso, tiene muchas opciones de pasarlo, esto debido a que el primer ciclo, haya renunciado o simplemente lo Desaprobó, entonces vendría siendo la segunda vez que este alumno lleva el curso.

En cambio, sí le fue mal en el examen1 y pertenece a la escuela profesional de ingeniería de sistemas e informática, la regla dice que reprueba el curso, acá todo claramente va a depender a la escuela que pertenezca el alumno (EPISI u OTRA).

En conclusión, se puede decir que, si le fue mal en el primer examen, tiene muchas posibilidades de aprobar aún el curso, no es así en el caso que no rinda o le va muy mal en el EXAMEN1, lo más probable es que termine reprobando el curso.

Son estas reglas las que ya nos permiten ir visualizando a qué clase de la nota final serán clasificadas las instancias, es así como el docente por medio de este conocimiento nuevo adquirido a través de las reglas que nos proporcionó el algoritmo J48, puede tomar medidas favorables en el desempeño del alumno.

```
CART Decision Tree

EXAMEN1=(NoRinde)|(MuyMal)|(Mal)
|  EXAMEN1=(NoRinde)|(MuyMal)|(Bien)|(MuyBien)|(Regular): Reprobo(147.0/27.0)
|  EXAMEN1!=(NoRinde)|(MuyMal)|(Bien)|(MuyBien)|(Regular)
|  |  ESCUELA=(EPISI): Reprobo(31.0/14.0)
|  |  ESCUELA!=(EPISI): Aprobo(24.0/15.0)
EXAMEN1!=(NoRinde)|(MuyMal)|(Mal): Aprobo(177.0/55.0)
```

Figura 20. Reglas generadas por el algoritmo SimpleCart.

Acá se generan reglas en base al EXAMEN1, nuevamente, demostrando que este atributo es importante, en este experimento.

SI EXAMEN1 = Muy Bien O EXAMEN1 = Bien O EXAMEN1 = Regular **ENTONCES** Aprobó

SI EXAMEN1 = Mal Y ESCUELA = EPISI **ENTONCES** Desaprobó

SI EXAMEN1 = Mal Y ESCUELA = OTRA **ENTONCES** Aprobó

Nuevamente el examen1 y la escuela son atributos que hacen la diferencia entre una regla y otra, derivando a que clases pertenecerá el resultado de dichas reglas. Ahora probamos con el algoritmo "Ridor", y se obtienen más reglas interesantes.

```
NFINAL = Aprobo  (490.0/248.0)
           Except (EXAMEN1 = MuyMal)
                and (DOCENTE = DOCENTE1)
                and (ESCUELA = EPISI)
                and (COLEGIO = Estatal) =>
                                    NFINAL = Reprobo  (18.0/0.0) [14.0/3.0]
           Except (EXAMEN1 = MuyMal)
                and (COLEGIO = Otros)
                and (CICLO = I CICLO)
                and (ASISTENCIA = SinInformacion) =>
                                    NFINAL = Reprobo  (15.0/0.0) [6.0/1.0]
           Except (EXAMEN1 = MuyMal)
                and (CICLO = SinInformacion)
                and (ESCUELA = EPISI)
                and (NOTA = Bien) =>
                                    NFINAL = Reprobo  (8.0/0.0) [6.0/0.0]
           Except (EXAMEN1 = MuyMal) =>
                                    NFINAL = Reprobo  (58.0/19.0) [36.0/4.0]
           Except (ESCUELA = EPISI)
                and (EXAMEN1 = Mal)
                and (CICLO = SinInformacion)
                and (PMAT = Bueno) =>
                                    NFINAL = Reprobo  (4.0/0.0) [2.0/0.0]
           Except (EXAMEN1 = Mal)
                and (ESCUELA = EPISI)
                and (PMAT = Regular)
                and (TINGRESO = Normal)
                and (DOCENTE = DOCENTE2)
                and (NOTA = MuyBien) =>
                                    NFINAL = Reprobo  (3.0/0.0) [4.0/1.0]
```

Figura 21. Reglas obtenidas del algoritmo Ridor.

Se han encontrado las siguientes reglas interesantes.

SI EXAMEN1=Muy Mal Y DOCENTE=DOCENTE1 Y ESCUELA=EPISI Y COLEGIO= Estatal **ENTONCES** Desaprobó

SI EXAMEN1=Mal Y Escuela = EPISI Y PMAT = Regular Y TIngreso=Normal Y Docente=Docente2 Y Nota= Muy Bien **ENTONCES** Desaprobó

SI EXAMEN1= Muy Mal Y Colegio = Otros Y Ciclo = I Ciclo Y ASISTENCIA = Sin Información **ENTONCES** Desaprobó

Uno tiende a pensar equívocamente que un alumno proveniente de un establecimiento Estatal, tiene más posibilidades de Aprobar el curso, esto debido al trato y a la modalidad de educación que reciben estos establecimientos en nuestro país, y además porque la gente estigmatiza a los alumnos provenientes de colegios públicos.

Queda evidenciado que esto no siempre es así, ya que como se aprecia en las reglas obtenidas anteriormente, los alumnos que son de colegio estatal al mezclarse con otros factores o atributos, el resultado fue negativo, es decir, han terminado reprobando el curso.

Eso quiere decir que dentro de la universidad todos los factores cumplen un rol importante para aprobar y reprobar el curso de informática.

La escuela ya es un atributo importante junto con el EXAMEN1 y la Escuela Profesional EPISI tiene más posibilidades de reprobar el curso, dependiendo del examen1, esto se comprueba con la última regla que estipula lo siguiente: si él alumno obtuvo muy mala nota en el examen1 y es de primer ciclo el alumno terminará reprobando el curso.

3.3. Representación de resultados

3.3.1. Experimento 1: Validación Cruzada utilizando todos los atributos.

En este primer experimento se han ejecutados los 10 algoritmos (cinco de reglas de clasificación y cinco de árboles de decisión), utilizando todas las 16 variables de los 490 alumnos Luego se ha procedido a aplicar una validación cruzada con 10 particiones, como viene predeterminado en el WEKA.

Cuando se aplica validación cruzada en los diferentes experimentos, lo que se hace es que se realiza el entrenamiento y la prueba 10 veces con las distintas particiones y los resultados que se obtienen son la media de las 10 ejecuciones.

Tabla 20

Validación cruzada de todos los atributos.

Algoritmo	Porcentaje de Acierto "Aprobó"	Porcentaje de Aciertos "Desaprobó"	Porcentaje de Precisión	Promedio Ponderado
JRip	97,1	95,3	96,1	96,15
NNge	95,9	96	95,9	95,93
OneR	89,7	**100**	95,4	95,03
Prism	**96,2**	95,3	95,8	95,77
Ridor	90,5	97,6	94,3	94,13
J48	95,5	97,6	**96,5**	**96,53**
SimpleCart	93,8	97,2	95,6	95,53
ADTree	93,8	95,6	94,7	94,70
REPTree	95	98,4	**96,8**	**96,73**
NBTree	93	96,8	95	94,93

Puede observarse en la tabla anterior que los porcentajes de exactitud o precisión obtenidos para los Reprobados, son más altos que los de Aprobados.

Los algoritmos que consiguen los valores más altos en cada columna están destacados en negrita y son: OneR (para los porcentajes de aciertos en la clase "Desaprobó"), Prism (para los porcentajes de aciertos en la clase "Aprobó") y el algoritmo de árbol de decisión J48 (para los porcentajes de "Precisión" y "Promedio Ponderado").

3.3.2. Experimento 2: Validación cruzada utilizando todos los atributos y considerando el costo de clasificación.

En este experimento se pretende mejorar los porcentajes de aciertos para cada clase. cuando se quiere optimizar la tasa de clasificación sin tomar en cuenta el costo de los errores.

A menudo se pueden llegar a resultados no óptimos, debido al elevado costo que puede causar la mala clasificación de una instancia.

A los 10 algoritmos se le ha aplicado una clasificación sensible al costo, después de varias pruebas en cada algoritmo se ha encontrado una matriz de costos, que permite obtener los mejores resultados de clasificación.

Tabla 21

Validación cruzada de todos los atributos considerando el costo de clasificación.

Algoritmo	Porcentaje de Acierto "Aprobó"	Porcentaje de Aciertos "Desaprobó"	Porcentaje de Precisión	Promedio Ponderado
JRip	90,9	98	94,7	94,5
NNge	93	98	95,6	95,5
OneR	89,7	**100**	95,4	95,0
Prism	**96,6**	95,4	96	96,0
Ridor	89,3	99,6	95	94,6
J48	89,7	99,2	94,9	94,6
SimpleCart	92,1	98,4	95,5	95,3
ADTree	92,6	97,2	95	94,9
REPTree	93,8	98,8	**96,4**	**96,3**
NBTree	91,3	98,8	95,3	95,1

Al comparar los resultados, se ha podido inferir que los resultados obtenidos en la segunda tabla, en la cual se ha considerado el costo de clasificación, han aumentado su porcentaje de acierto en "Desaprobó" en los diversos algoritmos tomados como ejemplo, pero han disminuido en la mayoría el porcentaje de precisión y el porcentaje de acierto de "Aprobó".

Esto es favorable, en el sentido que se sabrá con mayor precisión cuales son los alumnos más vulnerables, dándole a conocer al docente, mediante las reglas y arboles obtenidos por estos algoritmos, cuáles son las variables más influyentes en el rendimiento final del alumno.

3.3.3. Experimento 3: Validación Cruzada usando atributos hasta el Examen1.

El objetivo de este experimento es obtener reglas que ayuden a detectar, hasta que se ha realizado la evaluación del examen1, se encontrara algunos patrones que permitan mostrar los atributos que en conjunto influyen de forma positiva como negativa.

Hay que considerar que solo se toman en cuenta algunos atributos, el principal es el examen1, para luego inferir la nota final a la cual pertenece (Aprobó o Desaprobó).

Al implementar y probar los cuatro algoritmos de clasificación, se infieren algunas reglas interesantes, y que probablemente servirán de base para futuros estudios.

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = I CICLO Y DOCENTE= DOCENTE1 **ENTONCES** Aprobó

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = I CICLO Y DOCENTE= DOCENTE2 **ENTONCES** Desaprobó

SI EXAMEN1 = Mal Y ESCUELA = OTRA Y Ciclo = II CICLO **ENTONCES** Aprobó

SI EXAMEN1 = Mal Y ESCUELA = EPISI **ENTONCES** Desaprobó

SI EXAMEN1 = Muy Bien O EXAMEN1 = Bien O EXAMEN1 = Regular **ENTONCES** Aprobó

SI EXAMEN1 = Mal Y ESCUELA = EPISI **ENTONCES** Desaprobó

SI EXAMEN1 = Mal Y ESCUELA = OTRA **ENTONCES** Aprobó

SI EXAMEN1=Muy Mal Y DOCENTE=DOCENTE1 Y ESCUELA=EPISI Y COLEGIO= Estatal **ENTONCES** Desaprobó

SI EXAMEN1=Mal Y Escuela = EPISI Y PMAT = Regular Y TIngreso=Normal Y Docente=Docente2 Y Nota= Muy Bien **ENTONCES** Desaprobó

SI EXAMEN1= Muy Mal Y Colegio = Otros Y Ciclo = I Ciclo Y ASISTENCIA = Sin Información **ENTONCES** Desaprobó

Al obtener los datos que genera el WEKA en reglas y arboles como las que se han obtenido en este experimento, es mucho más fácil evaluar e influir en el desempeño de los alumnos. Todos estos resultados obtenidos se dan a que la selección de algoritmos fueron todos de tipo “Caja Blanca”, es decir, se obtiene un modelo de salida comprensible para el uso de cualquier tipo de usuario común y corriente.

Al obtener los datos que genera el WEKA en reglas y arboles como las que se han obtenido en este experimento, es mucho más fácil evaluar e influir en el desempeño de los alumnos. Todos estos resultados obtenidos se dan a que la selección de algoritmos fueron todos de tipo “Caja Blanca”, es decir, se obtiene un modelo de salida comprensible para el uso de cualquier tipo de usuario común y corriente.

CAPÍTULO IV

CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

Primera. Para alcanzar el objetivo, se desarrolló un proceso de minería de datos, con la finalidad de establecer diferentes etapas del proceso, que influyan de forma significativa en el rendimiento de los estudiantes. La aplicación de las dos primeras etapas de selección y pre procesamiento de los datos se logró obtener data aplicable para minería de datos.

Segunda. Para lograr los resultados, fue necesario definir enfoques, técnicas y herramientas de minería de datos, para comparar y analizar. Se eligieron enfoques principalmente clasificación y regresión utilizando las técnicas de árboles de decisión y clasificación, los algoritmos empleados mediante estas técnicas son: J48, Prism, OneR, entre los más destacados y precisos.

Tercera. Las variables escuela, ningreso, pmat, nota, colegio, tingreso, asistencia, ciclo, docente, examen1, examen2, aplazados, np, nft, nfp, termino y nmatr las cuales establecieron ciertas reglas que contribuyen a deducir cuáles son las que más influyen en el rendimiento del estudiante, respecto a la nota final y sus clases (“Aprobó” y “Desaprobó”).

Cuarta. Se realizó tres experimentos satisfactorios, analizado los modelos, técnicas y algoritmos que nos ofrece el software weka, se obtuvo buenos resultados y conclusiones satisfactorias, las cuales servirán para que el docente pueda determinar, en el momento que alumnos van a aprobar o desaprobar el curso, y tomar decisiones en el desarrollo de la clase.

4.2. Recomendaciones

Primera. Se recomienda tener más atributos influyentes que pueden salir de lo académico como el aspecto familiar, y el sexo del estudiante los cuales podrían generar nuevas reglas y alcanzar otra etapa del proyecto.

Segunda. Generar un aplicativo que pueda evaluar el rendimiento académico de manera automática, e integrarlo al ERP de la universidad para así poder llevar la educación en la universidad José Carlos Mariátegui a otro nivel.

Tercera. El manejo de minería de datos en el aspecto académico es complejo que, con otros tipos de aplicación, Este contiene distintas variables importantes, pero existen otros aspectos que no siempre se consideran en una encuesta, por ser de índole personal.

REFERENCIAS BIBLIOGRÁFICAS

Agrawal, R., Imielinski & Swami. (2013). *DatabaseMining: Aperformance Perspective, IEEE Transacctions on know ledge and Data Engineering.* Recuperado de Almadem.ibm.

Bayes, T. (1764). *An essay towards solving a problem in the doctrine of chances.* London: Philosophical Transactions of the Royal Society of London.

Bigus, J. (1996). *Data mining with neural networks. solving business problems-from Application Development to decision support.* Nueva York: McGraw-Hill.

Eyherabide, A. (2012). *Introduccion al Data Mining.* Recuperado de Slideshare.

Fayyad, Piatetsky-Shapiro, & Smyth. (1996). *Knowledge Discovery and Data Mining: Towards a Unifying Framework.*

Francisco, R. (2015). *Aplicacion de Data Mining.* Recuperado de Repositorio ubb.

Han, J. & Kamber, M. (2001). *Data Mining Concepts and Techniques.* Asma Stephan.

Hernandez, J., Ramirez, M. J. & Ferri, C. (2004). *Introduccion a la mineria de datos.*

Inmon, B. & Hackathorn, R. (1992). *Los sistemas de información en la sociedad del conocimiento.*

Izaurieta, F. & Saavedra, C. (2000). *Redes Neuronales Artificiales.* Chile: Universidad de Concepcion.

Kdnuggets. (2012). *What Analytics, Data mining, Big Data software you used in the past 12 months for a real project?* Recuperado de http://www.kdnuggets.com

Molina, L. (2002). *Torturando a los datos hasta que confiesen.* FUOC.

Nadinic, M. (2008). *Data Mining y Data Warehousing.* Santiago, Chile.

Palmer, A. & Montaño, J. (1999). *¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adiciones.* España: Universidad de las Islas Baleares.

Rojas, R. (1996). *Neural Networks.*

Rosenblatt, F. (1958). *The perceptron: A probabilistic model for information storage and organization in the brain.* Nueva York.

Waldo, H. (2012). *Extraccion del conocimiento en grandes bases de datos.* La plata.

Wikipedia. (12 de noviembre 2018). *Weka*. Recuperado de Wikipedia: https://www.wikipedia.org/

Witten, I. H. & Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques.* Nueva Zelanda: Morgan.