FACULTAD DE INGENIERIA, ARQUITECTURA Y URBANISMO

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

TÉCNICAS DE MINERÍA DE DATOS PARA PREDICCIÓN DEL DIAGNÓSTICO DE HIPERTENSIÓN ARTERIAL

AUTOR:

DÍAZ AVENDAÑO, ÁNGEL ARNULFO

Pimentel, 06 de noviembre 2016

TÉCNICAS DE MINERÍA DE DATOS PARA PREDICCIÓN DEL DIAGNÓSTICO DE HIPERTENSIÓN ARTERIAL

Aprobación de Tesis

Díaz Avendaño Angel Arnulfo

Autor

Dr. Ramos Moscol Mario

Asesor Especialista/ Vocal del Jurado de tesis

M.Sc.Ing. Chirinos Mundaca Carlos

Presidente del Jurado de tesis

Ing. Cobeñas Sanchez Rosa America

Secretaria del Jurado de tesis

DEDICATORIA

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Dedicatoria

A Dios por ser mi guía incondicional

A mis padres (Arnulfo y Josefa) por su apoyo moral e

Inculcarme a cumplir una meta anhelada.

A mis hermanas (Jacky y Liseth), que son un pilar fundamental de mi vida.

A mi asesor especialista Dr. Mario Fernando Ramos Moscol.

Al M.Sc.Ing. Carlos Alberto Chirinos Mundaca.

A la Ing. Rosa América Cobeñas Sanchez.

Que me apoyaron para lograr este proyecto,

Ya que este logro es fundamental en mi carrera profesional.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

UNIVERSIDAD SEÑOR DE SIPÁN

Índice

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráficos

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tablas

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

RESUMEN

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Resumen

Este Proyecto se incluye dentro de la disciplina de la Extracción Automática de Conocimiento (KDD, Knowledge Discovery in Databases) y más concretamente se centra en la etapa de Minería de Datos (MD). La MD es una de las áreas que más éxito y aplicación ha tenido a la hora de analizar información con el objetivo de extraer nuevo conocimiento. El objetivo de este trabajo fue encontrar patrones y relaciones dentro de los datos permitiendo de la creación de modelos en los que la representación del conocimiento estuvo basada en reglas de asociación y árbol de decisión. Los resultados mostraron que la técnica de regla se asociación es la más acertada para un pre diagnóstico de enfermedad de hipertensión arterial con un nivel de confiabilidad de 98.6 % en sus resultados.

Concretamente, la extracción de reglas de asociación consiste en descubrir relaciones interesantes, y previamente inesperadas, entre los diferentes atributos de un conjunto de datos. Las reglas obtenidas pueden servir de ayuda para poder tomar decisiónes de un pre diagnóstico.

PALABRAS CLAVES: KDD, Técnicas de Minería de Datos, Reglas de Asociación, Árbol de Decisión.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Abstract

This project is included with in the discipline of automatic extraction of knowledge (KDD, Knowledge Discovery in Databases) and specifically focuses on the stage of Data Mining (DM). The MD is one of the most successful areas and implementation has had time to analyze the information in order to extract new knowledge. The objective of this work was to find patterns and relationships in data allowing the creation of models in which knowledge representation was based on association rules and decisión tree. The results showed that the technique of association rule is the right to a pre diagnosis of hypertension disease with a confidence level of 98.6% in its results.

Specifically, the extraction of association rules is to find interesting relationships, and previously unexpected, between the different attributes of a data set. The rules obtained can be helpful to make decisións of a pre diagnosis.

KEYWORDS: KDD, Data Mining Techniques, Association Rules, Decisión Tree.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INTRODUCCIÓN

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INTRODUCCIÓN

Se describe el ámbito en el que se desarrolla la investigación. Se estudia en primer lugar el área de la minería de datos, centrada en el proceso completo de extracción de conocimiento a partir de bases de datos. Se centra en aportar una visión general sobre la minería de datos a modo de introducción, relacionándola con otras disciplinas y estudiando las diferentes etapas que se acontecen en proceso de extracción de conocimiento a partir de bases de datos. Por último, se desarrolla un breve estudio sobre las tareas y aplicaciones de la minería de datos.

Además, se estudian las reglas de asociación y árbol de decisión, de forma que se describe de manera más amplia la parte de la minería de datos que se desarrolla posteriormente en la metodología KDD y que supone foco de estudio de esta investigación.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO I

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO I: PROBLEMA DE INVESTIGACIÓN

En el plan de investigación, se detalla cual es el planteamiento del problema, el marco teórico a considerar en la investigación y cuáles serán las fuentes de datos, la hipótesis, las variables y su operacionalización, además de los métodos de investigación.

1.1 Situación Problemática

Publicó en su investigación Técnicas de Minería de Datos Aplicadas al Diagnóstico de Entidades Clínicas. Por lo cual la Hipertensión Arterial se ha convertido en una de las primeras causas de muertes en el mundo. Según el reporte de la Organización Mundial de la Salud (OMS) del 2012 1 de cada 3 personas en el mundo padece de Hipertensión Arterial; además agrega que 1 de cada 10 personas es diabética. Coinciden que anualmente existen 7.2 millones de muertes por enfermedades del corazón. La hipertensión arterial es la segunda causa de muerte a nivel mundial, se reconoce internacionalmente como "muerte silenciosa" pues en la mayoría de los casos los pacientes tienden a ser asintomáticos. Es por ello que en los hospitales se albergan historias clínicas de pacientes que padecen diversas enfermedades, esta información se aloja en el datamart. Debido al gran volumen de datos existentes, se dificulta la toma de decisiónes de los especialistas para realizar un análisis rápido y efectivo y de esta manera encontrar información útil y valiosa oculta en ellos; por otra parte, la no predicción del comportamiento futuro de algunos problemas de salud presentes en las HCE (Historia Clínica Electrónica) con un alto porcentaje de certeza, basado en el entendimiento del pasado.

(Somoza, 2015) En los últimos años La hipertensión arterial (HTA) es una enfermedad crónica muy extendida a nivel mundial. En España, se estima que el 35% de la población tiene la tensión más alta de lo normal. Con frecuencia, los pacientes no son conscientes de ello, ya que la hipertensión arterial no suele cursar con sintomatología en sus inicios. Sin embargo, unos valores de presión arterial elevados provocan daños en el organismo. Cualquier persona puede contribuir a reducir la presión arterial para prevenir el desarrollo de patologías asociadas. Uno de cada tres adultos tiene presión arterial alta en todo el mundo, afección que ocasiona alrededor de la mitad de todas las muertes por accidente cerebrovascular y enfermedad cardíaca, según el informe estadísticas sanitarias mundiales 2014 de la Organización Mundial de la Salud (OMS).

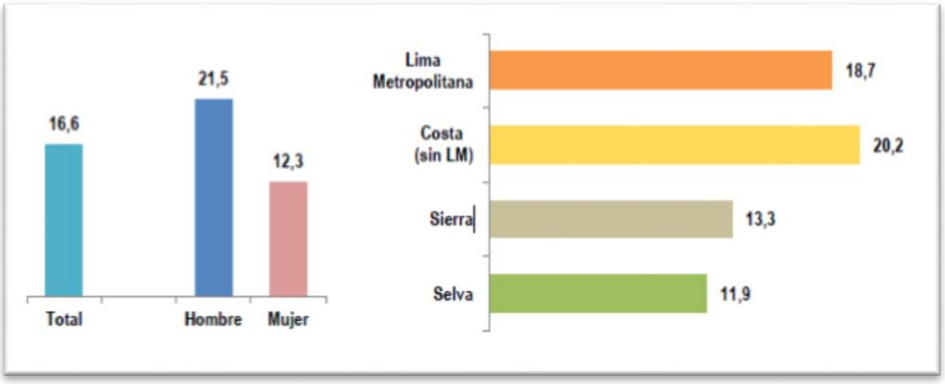
(INEI, 2014). De la población de 15 y más años de edad con presión arterial medida, se encontró un 16,6% con hipertensión arterial; siendo los hombres más afectados (21,5%) que las mujeres (12,3%).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

La prevalencia de hipertensión arterial es mayor en la Costa sin Lima Metropolitana (20,2%) seguido por Lima Metropolitana (18,7%); en tanto, la menor prevalencia se registró en la Selva (11,9%) y en la Sierra (13,3%).

Gráfico 1: PERÚ - Personas de 15 y más años edad con presión arterial alta de acuerdo a medición efectuada, según sexo y región natural.



Fuente: (INEI, Encuesta Demografica y Salud Familiar, 2013)

1.2 Formulación del problema

¿Cómo extraemos información del gran volumen de datos para detallar el modelo de soporte de predicción para el diagnóstico de Hipertensión Arterial?

1.3 Justificación e Importancia

El motivo de desarrollo de esta investigación es que existe un problema real en la capacidad de procesar grandes cantidades de datos, los cuales generan las áreas operativas de cada empresa e institución, el problema que puede ser resuelto con la aplicación de algoritmos de minería de datos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

La presente investigación se justifica entonces por el impacto que representa estudiar los algoritmos de minería de datos en la solución de problemas donde se requiere el uso de grandes repositorios de datos para convertirlos en información útil que genere valor en el ámbito del sector salud, usando para ello técnicas avanzadas de minería de datos; las cuales serán evaluadas para medir su grado de efectividad en el área del sector salud.

Además la presente investigación es importante por su aporte al estudio y conocimiento de las técnicas de minería de datos, tales como árboles de decisión, redes neuronales, clustering, entre otros.

1.4 Limitaciones de la Investigación

Obtención de base de datos a nivel nacional del Hospital Almanzor Aguinaga Asenjo, ya que solo para el estudio se cuenta con la base de datos que se obtuvo en la sede de Chiclayo.

1.5 Objetivo general

Aplicar técnicas de minería de datos para predicción del diagnóstico de hipertensión arterial.

1.6 Objetivos específicos

a. Recopilar información histórica acerca de los pacientes del hospital Almanzor Aguinaga Asenjo como base de estudio en el ámbito local del sector salud.
b. Analizar y evaluar la información sobre algoritmos y técnicas predictivas de minería de datos y determinar los requerimientos del modelo de minería de datos a fin de seleccionar el método más adecuado para el pronóstico de hipertensión arterial

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c. Identificar variables cuantificables y analizar la Intervención de dichas variables con la elaboración de los pronósticos.
d. Mostrar los resultados de las técnicas de minería de datos en una página web de HTML y Java (Netbeans).
e. Realizar pruebas para los modelos de minería de datos.
f. Evaluar los resultados.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO II

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO II: MARCO TEÓRICO

A continuación, se presentan los antecedentes de la investigación, los cuales son usados para estudiar y analizar soluciones previas relacionadas, el marco teórico – científico, donde se detallan los conceptos y técnicas a emplear en el proyecto y la definición de conceptos básicos.

2.1 Antecedentes de la Investigación

(Díaz Pérez, 2012). Esta tesis publica en su investigación **APLICACIÓN DE LA RED DE PROBABILIDAD NEURONAL Y ESCALA DE FRAMINGHAM PARA PREDICCIÓN DE LA HIPERTENSION ARTERIAL**, realizó un estudio con estudiantes de la Facultad de Ciencias de la Salud de la Corporación Universitaria Rafael Núñez con un nivel de confianza del 95% y el error alfa del 5%, de una población de 215 estudiantes de los tres últimos semestres del programa de enfermería y medicina. El muestreo fue aleatorio sistemático con un punto de inicio como intervalo de selección el cual se denominó K que correspondió a dos, de dos en dos hasta completar el tamaño de la muestra de 138 con la escala utilizada fue Framingham Heart Study, para predicción a uno, dos y cuatro años para HTA, teniendo en cuenta factores modificables y no modificables, la cual determina el riesgo cardiovascular (RCV). La tabulación y análisis estadístico se realizó en el programa Excel 2007, la cual se exportó al programa Statgraphics Centurión XVI versión 16.1.15. La información se digitó y se monitoreó para determinar la calidad de los datos incorporados.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Se aplicó Cuadrado para las variables cualitativas y la prueba de análisis multivalente de regresión logística (Pearson) para conocer qué variables formaban parte de la ecuación para el riesgo a hipertensión arterial a uno, dos y cuatros, con un valor de p=0.05. El Clasificador Probabilístico de Red Neuronal (PNN) el estimado se construyó usando una ventana Parzen que pondera las observaciones de cada grupo o variables de acuerdo a su distancia desde la localización especificada; el cual cuenta con una entrada de datos, las cuales son las variables de entrada que son los nombres de (n) variables de entrada, que deben ser factores cuantitativos característicos de las muestras, por lo cual las variables dicotómicas (Variables Dummy) para poder medir la linealidad y relación de los factores de interés El programa (PNN) se entrenó usando Jackknifing, el cual retiene valores del grupo de entrenamiento uno a la vez y determina basándose en el porcentaje de tiempo que el punto retenido es correctamente clasificado. Las gráficas utilizadas fueron el diagrama de red y el gráfico de clasificación, el cual arroja graficas de regiones codificadas por colores que identifican áreas en donde las muestras serían clasificadas en diferentes grupos. Entre los factores modificables tales como: el consumo de café, el manejo del estrés, el consumo de alcohol, el consumo de bebidas negras, fumar; relacionados con los factores no modificables tales como: los antecedentes familiares de hipertensión arterial, el sexo, con la presencia de hipertensión arterial ya diagnosticada, se encontró que todos al final son factores condicionantes para identificar el riesgo: bajo, mediano o moderado y a alto, tomando en cuenta la escala de Framingham como complemento, los cuales precisan la debida atención e intervención inmediata;

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Debido al estilo de vida que lleva la población objeto de estudio y poder reducir el riesgo a hipertensión y transformar o modificar la intensidad de reducción de riesgo con base al riesgo global estimado en la escala y a los niveles de incertidumbre incluidos en el sistema estocástico.

(Salazar Mendiola & Vargas Luna, 2012) Publicó en su investigación **USO DE REDES NEURONALES PARA LA MEDICIÓN AUTOMÁTICA DE PRESIÓN ARTERIAL**, presentó un sistema de monitoreo de PA basado en redes neuronales (RN) que es capaz de dar una medición incluso bajo circunstancias ruidosas. Se desarrollaron el hardware que es capaz de capturar la presión, detectar los sonidos del corazón. Y en lo que corresponde al software se extrae la información más representativa de las señales para alimentar la RN. Las métricas fueron seleccionadas en función del entendimiento fisiológico de la función cardiovascular, el génesis de las señales y su pasividad de determinar las variables de interés. Esta extracción es hecha en segmentos (definidos por los picos de la auscultación). La selección final consta de 16 métricas relacionadas con características del complejo QRS (ECG), pico y valle de la oscilometría (P), amplitud del sonido (HS), dominio de la frecuencia del sonido, obtenido aplicando una transformada Morlet Wavelet (300 escalas), y combinaciones entre estas. Necesarias para implementar el sistema en un estudio clínico.

La correcta optimización de la RN es vital para el buen funcionamiento de todo el sistema.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Por ello, 8 diferentes estructuras de RN fueron probadas para tratar de optimizar su configuración. Todas las redes fueron entrenadas con un set de 1700 métricas, cuyos resultados deseados fueron dados por un humano experto por medio de inspección visual. Todas las redes fueron generadas, probadas y comparadas con MATLAB (Mathworks, Inc., EUA). Las redes se definieron con backpropagation, la función de transferencia de tangente hiperbólica y con 3 capas de diferentes dimensiones. El entrenamiento se hizo utilizando el algoritmo de optimización Levenberg-Marquardt y el desempeño fue medido por medio del error cuadrático medio (MSE). La red con mejor desempeño fue seleccionada e incluida en el algoritmo global. El algoritmo general fue desarrollado en LabView (National Instruments, Inc., EUA) Este concatena la adquisición, el procesamiento, la identificación de sonidos de Korotkoff y finalmente la evaluación de la PA. Como el método propone, el algoritmo implementa la RN para la identificación de sonidos de Korotkoff y cuando estos son detectados, son relacionados con la oscilometría. (Medición de presión) Los valores de PA y niveles de confiabilidad son evaluados por el sistema, obtuvo un resultado durante el estudio piloto se realizó 72 mediciones de PA, de las cuales 1700 sonidos (válidos y no válidos) fueron extraídos para entrenar la red. Estos datos fueron considerados como suficientes para probar la viabilidad del sistema.

(Solarte Martinez & Soto Mejia, 2011). En su investigación **REGLAS DE ASOCIACIÓN EN EL DIAGNÓSTICO DE ENFERMEDADES CARDIOVASCULARES,** presentó una descripción de reglas de asociación y del algoritmo Apriori, para determinar si se debe o no aplicar fármacos a pacientes con enfermedades cardiovasculares, demostrando que es posible diagnosticar la necesidad de administrar fármacos en pacientes con síntomas de enfermedad

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias. El cálculo se efectuó con la herramienta Weka 3.6, mediante la utilización del algoritmo Apriori con un nivel de exactitud del 60%.

2.2 Estado del arte

(Cabrera Hernández, y otros, 2010) Publicó en su investigación **ALGORITMOS PARA EL DIAGNÓSTICO DEL RIESGO DE HTA EN ESCOLARES.** Uso la técnica de Inteligencia Artificial de Algoritmos Genéticos para realizar una estructura que representara soluciones posibles, indicó que una estructura de datos consistía en uno o más cromosomas, que era representado por una cadena de bits, donde cada cromosoma es una concatenación de un número de subcomponentes llamados genes. La posición de un gene en el cromosoma es conocida como el locus del alelo.

En cadena de bits, un gen es un bit, el locus es la posición en la cadena y el alelo es su valor (0 o 1 si es un bit). Para optimizar la estructura de los AGs, una medida de la calidad de cada solución en el espacio de búsqueda es necesaria. La función de adaptabilidad es responsable de esta tarea. En una función de maximización, la función objetiva a menudo actúa como la función de adaptabilidad.

Los AGs usualmente trabajan con funciones de maximización, para los problemas de minimización los valores objetivos de la función puede ser negados y transferido para tomar valores positivos, produciéndose adaptabilidad.

El mecanismo simple de los AGs es el siguiente: Los AGs simples generan aleatoriamente una población de n estructuras (cadenas, cromosomas o individuos). Los operadores de la población actúan transformando la población. Una

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

vez que la aplicación de estos operadores es completada, se puede decir que un ciclo generacional ha concluido.

El operador de selección hace la selección de las cadenas según su adaptabilidad para los siguientes pasos. El operador de cruzamiento realiza la recombinación de material genético a partir de dos cadenas padre. El operador de mutación, al igual que la mutación natural, realiza la mutación de un gen dentro de un cromosoma.

Una probabilidad es asociada a cada uno de estos operadores. El modo de operación de una AG puede ser resumido. El AG se ejecuta para un número fijo de generaciones o hasta que algún criterio de parada es satisfecho.

Los AGs pueden solucionar las dificultades representadas en los problemas reales de la vida que algunas veces no tienen solución por otros métodos. El foco de investigación en los AGs es la robustez: el balance entre la efectividad y la eficiencia necesitada para sobrevivir en muchos ambientes diferentes.

Utilizó una base de casos HTA-children fue usada como aplicación. Esta base es binaria, tiene siete rasgos nominales, 16 rasgos son numéricos y 626 instancias o niños estudiados. Entre las variables nominales se encuentran sexo, color de la piel, edad, y las clasificaciones o diagnósticos de colesterol, triglicéridos y HTA.

Entre las numéricas se encuentran las concentraciones séricas de catalasa, glutatión y súper óxido dismutasa y entre las concentraciones aparece el colesterol, triglicéridos y varios metales. También fueros numéricas las variables peso al nacer y otros. Algunos de estos rasgos presentaban gran cantidad de valores perdidos, por lo que pensamos que los bajos por cientos de clasificación se deban a esto. Según la información contenida en la base y después de aplicar un análisis discriminante resultaron el color de la piel, el sexo y la actividad sérica de la enzima

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

catalasa variables muy importantes. La base se obtuvo como resultado de un estudio aplicado para predecir el riesgo de que un niño sea o no hipertenso. El cálculo de la clasificación individual se efectuó con los clasificadores existentes en la versión 3.7.5 del Weka y los tomados en cuenta en el estudio fueron los siguientes, con sus respectivos niveles de exactitud: Clasificador Exactitud NaivesBayes 62%. Clasificador: Functions. Logistic. 64%, LazyIBK 56%, Trees.J48 63%, Multilayer Perceptron 56%, Trees. ADTree 66%, Functions.SGD 62%, Random Tree 59%, Functions.SMO 61%, Lazy.KStar 57%, Functions. VotedPerceptron 60%

El mejor por ciento de clasificación obtenido por los clasificadores individuales no supera el 67% (0.66), luego fue aplicado el multiclasificador Vote, existente en la versión del Weka, promediando las salidas de los clasificadores base.

El cromosoma resultante corresponde a la combinación de los siguientes clasificadores: weka.classifiers.trees.J48, weka.classifiers.trees.RandomTree, weka.classifiers.lazy.KStar, weka.classifiers.functions.VotedPerceptron.

Este cromosoma provee una combinación de clasificadores que mejora a un 73% la exactitud del sistema multiclasificador, con respecto a los clasificadores individuales, nótese que aún no se logra un buen por ciento de casos correctamente clasificados, pero se mejora la clasificación individual en un 6%.

La investigación que realizó muestra una técnica novedosa que emplea algoritmos genéticos para encontrar un buen conjunto de clasificadores diversos. La función objetivo del Algoritmo Genético involucra la exactitud del sistema multiclasificador y los resultados de la diversidad entre los clasificadores individuales del sistema. Un caso de estudio de la base de HTA es usado para ejemplificar esta contribución.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Once clasificadores base fueron aplicados y sus resultados individuales no superan el 66%. Usando la propuesta del algoritmo genético con medidas de diversidad, se obtiene un multiclasificador que logra mejorar en un 6% la clasificación anterior.

(Cuadrado Rodríguez, y otros, julio 2012), publicó en su investigación **SISTEMA EXPERTO BASADO EN CASOS PARA EL DIAGNÓSTICO DE LA HIPERTENSIÓN ARTERIAL,** describe un sistema experto basado en casos para el diagnóstico de la hipertensión arterial (HTA) en la ciudad de Santa Clara, Cuba, realizado en el marco de un estudio para conocer la incidencia de la enfermedad en esta población. La muestra de la población estudiada estaba formada por 455 hombres y 394 mujeres, entre 18 y 78 años de edad. Los individuos fueron clasificados en normotensos (personas con presión arterial normal), prehipertensos (personas en riesgo de padecer HTA) e hipertensos.

Se realizó un procesamiento estadístico en el que se emplearon técnicas multivariadas como el Análisis Discriminante y la Regresión Logística cuyos resultados, junto a los del Método del Triángulo de Füller, fueron utilizados en el sistema para jerarquizar los factores de riesgo de la HTA y obtener el grado de importancia (peso) de estos. Por medio de la técnica de segmentación CHAID se pudo reducir las comparaciones entre los casos haciendo más eficiente este proceso. La obtención de las funciones de comparación por rasgos para las variables continuas se obtuvo de la aplicación conjunta de un análisis de varianza (ANOVA) y el método TwoStep Cluster Analysis. Todo esto permitió construir la función de semejanza para la comparación entre el nuevo caso a diagnosticar y los casos de la base. La adaptación de la solución de los casos más semejantes se realizó con la aplicación del algoritmo de los k-vecinos (método de clasificación supervisada de aprendizaje, estimación basada en un conjunto de entrenamiento y

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

prototipos que sirve para estimar la función de densidad de las predictoras por cada clase), más cercanos. El sistema experto fue validado finalmente y se comprobó una efectividad en el diagnóstico del 96%.

2.3 Bases teórico científicas

Se presentan los conocimientos o bases teóricas que serán empleadas a lo largo de la investigación.

2.3.1 Data warehouse

Un data warehouse (Cadenillas, 2011) es una base de datos, que constituye el gran almacén de datos que está diseñado fundamentalmente para permitir el acceso en forma fácil a toda la organización, integrar información histórica y consistente, adaptarse a los cambios que se dan en la organización, generar datos dirigido al usuario y presentados en forma consolidada fundamentalmente, para distribución de información y de consultas.

Se puede caracterizar un data warehouse haciendo un contraste de cómo los datos almacenados en un DW, difieren de los datos operacionales usados por las aplicaciones transaccionales u operacionales. El ingreso de datos en el Data Warehouse viene desde el ambiente operacional en casi todos los casos.

El Data Warehouse es siempre un almacén de datos transformados y separados físicamente de la aplicación donde se encontraron los datos del ambiente operacional.

2.3.2 Data Mart

(Alarcón, 2011) Cuando mantenemos una estructura de Data Warehouse, pero adaptada a solo un sector de la organización, se utiliza un Data Mart, que son subconjuntos de Data Warehouse para un área específica de la organización.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Un Data Mart está diseñado para satisfacer las necesidades específicas de grupos comunes de usuarios. Aunque generalmente son subconjuntos del Data Warehouse, también pueden integrar un número de fuentes heterogéneas, e incluso ser más grandes en datos que el almacén central.

2.3.3 Predicción

Según (Bunge, 2001) dice:

El término predicción puede referirse tanto a la «acción y al efecto de predecir como a las palabras que manifiestan aquello que se predice, en este sentido, predecir algo es «anunciar por revelación, ciencia o conjetura algo que ha de suceder.

La predicción constituye una de las esencias claves de la ciencia, de una teoría científica o de un modelo científico. Así, el éxito se mide por el éxito o acierto que tengan sus predicciones. La predicción en el contexto científico es una declaración precisa de lo que ocurrirá en determinadas condiciones especificadas. Se puede expresar a través del silogismo: "Si A es cierto, entonces B también será cierto".

El método científico concluye con la prueba de afirmaciones que son consecuencias lógicas del corpus de las teorías científicas. Generalmente esto se hace a través de experimentos que deben poder repetirse o mediante estudios observacionales rigurosos. Una teoría científica cuyas aseveraciones no son corroboradas por las observaciones, por las pruebas o por experimentos probablemente será rechazada.

Las teorías que generan muchas predicciones que resultan de gran valor (tanto por su interés científico como por sus aplicaciones) se confirman o se falsean fácilmente y, en muchos campos científicos, las más deseables son aquéllas que, con número bajo de principios básicos, predicen un gran número de sucesos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.4 Minería de datos

(Cadenillas, Mineria de Datos, 2011). La minería de datos (data mining) es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos.

Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos. Es el llamado descubrimiento del conocimiento y va direccionando al nivel estratégico directamente. Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos.

2.3.5 Algoritmos de Minería de Datos

(Cadenillas, Mineria de Datos, 2011). Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo, el algoritmo analiza los datos proporcionados en busca de tipos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos.

2.3.5.1 Algoritmo de Clústeres

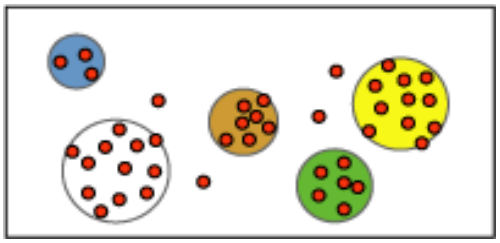
(Algoritmo de Clusteres, 2014)**.** Es un algoritmo que utiliza técnicas interactivas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación en las predicciones.

Los clústeres agrupan los puntos del gráfico e ilustran las relaciones que identifica el algoritmo.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 2: Gráfico de Dispersión de Algoritmo de Clustering



Fuente**:** (Algoritmo de Clusteres, 2014)

El algoritmo de agrupación en clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión, en que no se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de agrupación en clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

Cómo Funciona el Algoritmo

El algoritmo de agrupación en clústeres identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas Un gráfico de dispersión es una forma Útil de representar visualmente el modo en que el algoritmo agrupa los datos, tal como se muestra en el gráfico de dispersión representa todos los casos del conjunto de datos; cada caso es un punto del gráfico.

2.3.5.2 Algoritmo de Asociación

(Microsoft SQL server, 2014). Los modelos de Asociación se generan basándose en conjuntos de datos que contienen identificadores para casos individuales y para los elementos que contienen los casos. Un grupo de elementos de un caso se denomina un conjunto de elementos.

Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen como estos elementos se agrupan dentro de los casos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Las reglas que el algoritmo identifica pueden utilizarse para pre diagnosticar si un paciente puede adquirir enfermedad de HA, basándose con elementos existentes de una Base de Datos Histórica.

Gráfico 3: Reglas Derivadas de un Conjunto de Elementos

Regla
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Fuente: (Microsoft SQL server, 2014)

Como se muestra en la figura, el algoritmo de asociación puede encontrar potencialmente muchas reglas dentro de un conjunto de datos. El algoritmo usa dos parámetros, soporte y probabilidad, para describir los conjuntos de elementos y las reglas que se generan.

Cómo Funciona el Algoritmo

El algoritmo de asociación recorre un conjunto de datos para hallar elementos que aparezcan juntos en un caso. A continuación, agrupa en conjuntos de elementos todos los elementos asociados que aparecen, como mínimo, en el número de casos especificado en el parámetro MINIUM_SUPPORT.

Por ejemplo, un conjunto de elementos puede ser “Mountain 200= Existing, Sport 100=Existing”, y puede tener un soporte de 710. El algoritmo generara reglas se usan para predecir la presencia de un elemento en la base de datos, basándose en la presencia de otros elementos específicos que el algoritmo ha identificado como importantes.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.5.3 Algoritmo de árboles de decisión

(El Algoritmo de Árboles de Decisión, 2014). Es un algoritmo de clasificación y regresión proporcionado para el modelado de predicción de atributos discretos y continuos.

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos.

Utiliza los valores, conocidos como estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

Cómo Funciona el Algoritmo

El algoritmo de árboles de decisión genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si se predice una columna continua o una columna discreta.

El algoritmo de árboles de decisión utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de Analysis Services utilizan la selección de las características para mejorar el rendimiento y la calidad de análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si se utilizan demasiados atributos de predicción o de entrada al diseñar un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la entropía y las redes bayesianas.

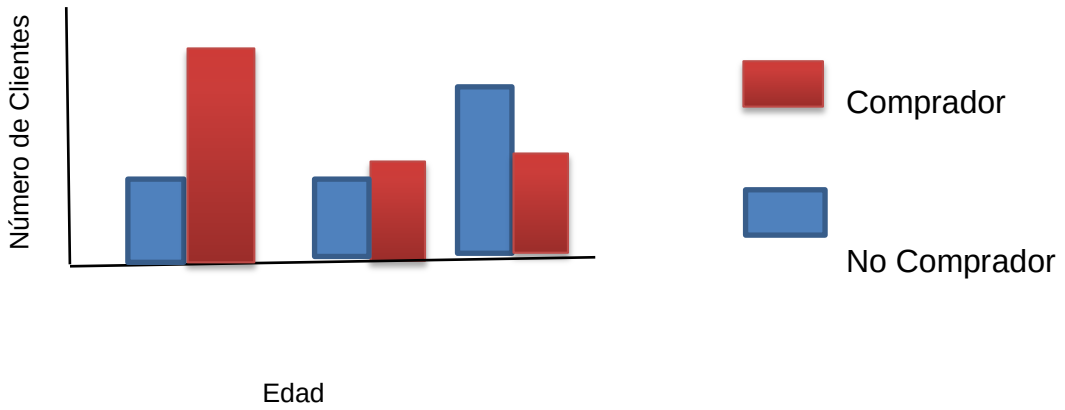
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobre ajustado o sobre entrenado. Un modelo sobre ajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobre ajustar un conjunto de datos determinado, el algoritmo de árboles de decisión utiliza técnicas para controlar el crecimiento del árbol.

Predecir columnas discretas

La forma en que el algoritmo de árboles de decisión genera un árbol para una columna de predicción discreta puede mostrarse mediante un histograma. La Figura muestra un histograma que traza una columna de predicción, comprador, con una columna de entrada, edad. El histograma muestra que la edad de una persona ayuda a distinguir si esa persona comprara una bicicleta.

Gráfico 4: Histograma de una columna de predicción

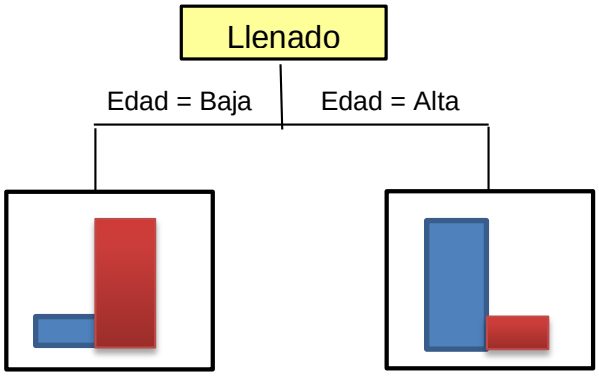


Fuente: (El Algoritmo de Árboles de Decisión, 2014)

La correlación que aparece en la Figura hará que el algoritmo de árboles de decisión cree un nuevo nodo en el modelo.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 5: Llenado de un árbol de decisión



Fuente: (El Algoritmo de Árboles de Decisión, 2014)

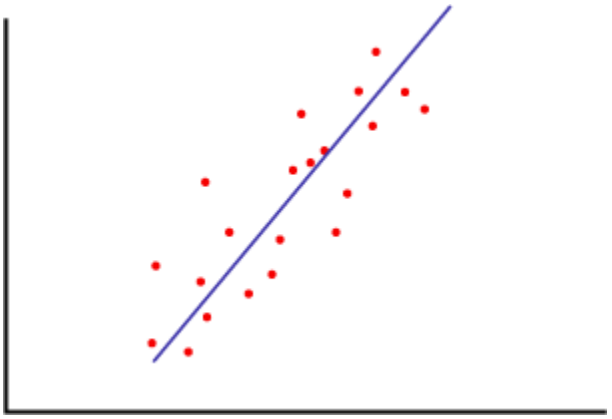
A medida de que el algoritmo agrega nuevos nodos a un modelo. Se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

2.3.5.4 Algoritmo de regresión lineal

(Microsoft SQL server, 2014). Es una variación del algoritmo de árboles de decisión que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación utiliza esa relación para la predicción. La relación toma forma de una ecuación para la línea que mejor represente una serie de datos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 6: Línea de regresión



Fuente: (El Algoritmo de Árboles de Decisión, 2014)

Cada punto de datos del diagrama tiene un error asociado con su distancia con respecto a la línea de regresión.

Hay otros tipos de regresión que utilizan varias variables y también hay métodos no lineales de regresión. Sin embargo, la regresión lineal es un método útil y conocido para modelar una respuesta a un cambio de algún factor subyacente.

Aunque hay muchas maneras de calcular la regresión lineal que no requieren herramientas de minería de datos, la ventaja de utilizar el algoritmo de regresión lineal para esta tarea es que se calculan y se prueban automáticamente todas las posibles relaciones entre las variables.

No tiene que seleccionar un método de cálculo, como por ejemplo para resolver los mínimos cuadrados. Sin embargo, la regresión lineal podría simplificar en exceso las relaciones en escenarios en los que varios factores afectan el resultado

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Cómo funciona el algoritmo

Es una variación del algoritmo de árboles de decisión. Al seleccionar el algoritmo de regresión lineal, se invoca en un caso especial del algoritmo de árboles de decisión, con parámetros que restringen el comportamiento del algoritmo y requieren ciertos tipos de datos de entrada. Además en un modelo de regresión lineal, el conjunto de datos completo se utiliza para calcular las relaciones en el paso inicial, mientras que en un modelo de árboles de decisión estándar los datos se dividen repetidamente en árboles o subconjuntos más pequeños.

2.3.6 Técnicas predictivas de minería de datos

Dado que la Minería de Datos es un campo muy interdisciplinar, existe un conjunto de tareas que cumplen con sus propósitos y que pueden ser utilizadas en áreas de aplicación específicos, diversas técnicas de Minería de Datos se utilizan para llevar a cabo las tareas de la misma, estas técnicas consisten en algoritmos específicos que pueden ser utilizados para cada función.

Dentro de las principales técnicas de Minería de Datos se encuentran:

2.3.6.1 Técnicas de inferencia estadística.

- ✓ Visualización.
- ✓ Razonamiento basado en memoria.
- ✓ Detección de conglomerados.
- ✓ Análisis de vínculos.
- ✓ Árboles de decisión.
- ✓ Redes neuronales.
- ✓ Algoritmos genéticos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.6.2 Técnica de análisis de vínculos

Esta técnica es muy útil para identificar las relaciones entre registros aplicando modelos basados en descubrimiento de patrones presentes en los datos. Dependiendo de los tipos de descubrimiento de conocimiento, las técnicas de análisis de vínculos tienen tres tipos de aplicaciones: descubrimiento de asociaciónes, descubrimiento de patrones secuenciales, y descubrimiento de secuencias de tiempo similares.

A continuación se Analizan brevemente cada una de estas aplicaciones: **Descubrimiento de Asociaciónes:** Las asociaciónes son las afinidades entre los elementos, los algoritmos de descubrimiento de asociaciónes encuentran sistemática y eficientemente combinaciones donde la presencia de un elemento sugiere la presencia de otro. Al aplicar estos algoritmos para las operaciones de compras en un supermercado, se descubren las afinidades entre los productos que pueden ser adquiridos juntos, las reglas de asociación representan tales afinidades entre los datos. Los patrones de soporte y de confiabilidad indican la fuerza de la asociación, las reglas con altos valores de soporte y confiabilidad son más válidas, relevantes y útiles para un grupo u organización.

2.3.7 Metodologías de Desarrollo

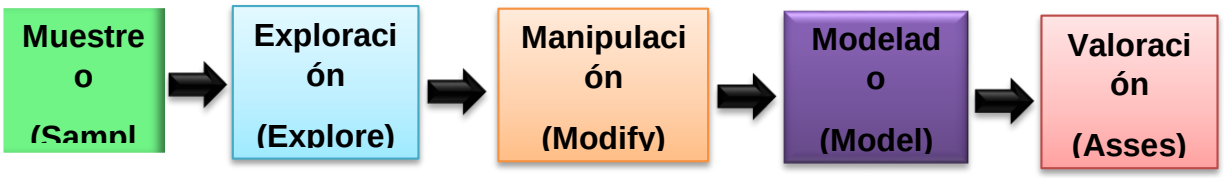
2.3.7.1 Metodología SEMMA (Sample, Explore, Modify, Model y Asses)

(SAS Institute, 2010) Es una empresa con sede en Cary (Carolina del Norte, E.E.U.U). Es una metodología más corta y menos extensa que el CRISP-DM porque se centra más en el desarrollo del proceso de Minería de datos y no se orienta a objetivos empresariales.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Tiene 5 fases cada uno representando a sus siglas SEMMA

Gráfico 7: Fases de metodología SEMMA



Fuente: (SAS Institute, 2010)

Sample: Extracción de una muestra representativa

En esta primera fase de la metodología, se realiza la extracción de un conjunto de datos que sean una buena representación de la población a analizar, esto se hace con el objetivo de facilitar los procesos de minado sobre los datos, reduciendo los tiempos que se necesita para determinar la información valiosa para el negocio.

Explore: Exploración de los datos en la muestra.

En esta fase, se hace un recorrido a través de los datos extraídos en la muestra para detectar, identificar y eliminar datos anómalos, ayudando a refinar los procesos de descubrimiento de información en fases siguientes del proceso. En este punto del proceso, la exploración se puede realizar a través de medios visuales, aunque muchas veces no es suficiente este método, es por eso, que además de la visualización se pueden manejar diferentes técnicas estadísticas como análisis de factores, análisis de correspondencias, entre otros.

Modify: Modificación de los datos.

Esta modificación de los datos se puede realizar creando, seleccionando y transformando las variables en las cuales se va a enfocar el proceso de selección

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

del modelo. Muchas veces se tendrá la necesidad de realizar modificaciones cuando los datos que se están analizando cambien. Esto se debe a que el entorno en el que se trabaja la minería de datos es dinámico e iterativo.

Model: Modelación de los datos

En esta fase, las herramientas de software se encargan de realizar una búsqueda completa de combinaciones de datos que juntos predecirán de una manera confiable los resultados buscados. Es en esta parte donde las técnicas y métodos de minería de datos entran a jugar un papel importante para la solución de los problemas que fueron identificados al iniciar el proyecto de minería de datos.

Assess: Evaluación de los datos obtenidos

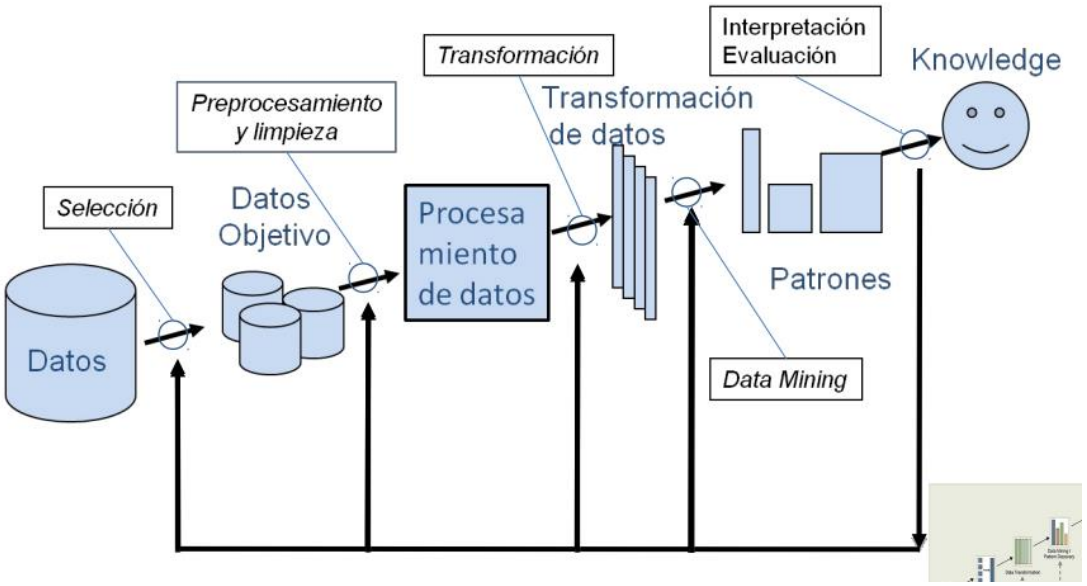
Después de que la fase de modelación presente los resultados obtenidos de la aplicación de los métodos de minería de datos al conjunto de datos. Se deberá realizar un análisis de los resultados para ver si estos fueron exitosos de acuerdo a las entradas que se tuvieron para analizar el problema. Una buena práctica para identificar si los resultados con el modelo creado son los esperados, es aplicar este modelo a una porción de datos diferente. Si el modelo funciona correctamente para esta muestra y para la muestra utilizada para el proceso de creación del modelo, se tiene una buena probabilidad de tener un modelo valido.

2.3.7.2 Metodología KDD

(Metodología KDD, 2010).Es una metodología propuesta por Fayyad en 1996, propone 5 fases: Selección, pre procesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Gráfico 8: Metodología KDD



Fuente: (Metodología KDD, 2010).

Fases

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.
3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
4. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).
5. Seleccionar y aplicar el método de minería de datos apropiado.
6. Evaluación, interpretación, transformación y representación de los patrones extraídos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

7. Difusión y uso del nuevo conocimiento.

2.4 Definición de términos básicos

2.4.1 Jackknifing:

Según (Sanchez Cameron, 2005) dice:

El método Jackknifing es una técnica de muestreo especialmente útil para la varianza y el sesgo de estimación. En el ámbito de la estadística se denomina remuestreo a una variedad de métodos que permiten realizar algunas operaciones. Estimar la precisión de muestra estadísticas. Intercambiar marcadores de puntos de datos al realizar test de significancia.

2.4.2 Almacén de datos

Es una colección de datos orientada a un determinado ámbito (organización, institución, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiónes en la entidad en la que se utiliza. (Kimball, 1998).

2.4.3 Análisis prospectivo de datos

Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos. (Lezcano, 2010)

2.4.4 Árbol de decisión

Estructura en forma de árbol que representa un conjunto de decisiónes. Estas decisiónes generan reglas para la clasificación de un conjunto de datos. (Asencios, 2004).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.4.5 Data Mart

Los Data Mart son subconjuntos de datos de un data warehouse para áreas específicas. Entre las características de data mart destacan: Usuarios limitados, área específica, tiene un propósito específico y tiene una función de apoyo. (Kimball, 1998)

2.4.6 Método

Modo ordenado y sistemático de proceder para lograr un fin / conjunto de reglas. (Getoor & Ben, 2007).

2.4.7 Metodología

Conjunto de métodos que se siguen en una disciplina científica / ciencia del método y de la sistematización científica. (Grudnitsky, 1992).

2.4.8 Minería de datos

(Asencios, 2004) Descubrimiento de relaciones en grandes conjuntos de datos. Conjunto de técnicas aplicadas al proceso de extracción y presentación de conocimiento que yace implícito en grandes conjuntos de datos.

2.4.9 Modelo predictivo

Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos (Lezcano, 2010)

2.4.10 Técnicas

Aplicación práctica de métodos y conocimientos relativos a diversas ciencias. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesado diferente de los datos. (Española).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.4.11 Técnicas de Predicción

Métodos que tienen por finalidad obtener estimaciones o pronósticos de valores futuros de una serie temporal a partir de la información histórica contenida en la serie observada hasta el momento actual. (Getoor & Ben, 2007)

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO III

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO III: MARCO METODOLÓGICO

En el marco metodológico, se detalla la hipótesis de esta investigación, los procedimientos específicos para el desarrollo de la misma, así como sus técnicas e instrumentos a usar en la observación y recolección de datos.

3.1 Tipo y diseño de la investigación

La presente investigación es:

Del tipo Tecnológica: porque utiliza un marco de conocimientos relacionados con la tecnología para aplicarse en el proyecto en estudio.

Su diseño Cuasi-Experimental: porque consiste y se seleccionan los grupos de la muestra en los que se prueba la variable sin ningún tipo de selección aleatoria o proceso de pre selección.

3.2 Población y muestra

3.2.1 Población:

La población está conformada por un total de 1,000.000 pacientes registrados en la base de datos del Hospital Almanzor Aguinaga Asenjo del periodo 2015.

3.2.2 Muestra:

La muestra es de tipo poblacional, es decir los 8,735 registros seleccionados por conveniencia para el estudio.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.3 Hipótesis

La aplicación de técnicas de minería de datos permitirá realizar la predicción de Diagnóstico de Hipertensión Arterial del Sector Salud.

3.4 Variables

3.4.1 Variable dependiente

Predicción de diagnóstico de hipertensión arterial del sector salud.

3.4.2 Variable independiente

Técnicas de minería de datos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.5 Operacionalización de variables

Tabla 1: Operacionalización de variables

Variable	Dimensiones	Indicadores	Fórmula
Predicción de Diagnóstico de Hipertensión Arterial	Cálculo de tiempo de ejecución del sistema	Determina el cálculo de ejecución proyectados por el sistema	**TEM= TS** **TEM=** Tiempo de Ejecución del Modelo. **TS=** Tiempo en Segundos.
	Precisión de pre diagnóstico	Determina la precisión de pre diagnóstico proyectados por el sistema	**PPM1 =** Performance Confidence **PPM2 =** Performance Condicional Operador. **PPM1 =** Precisión de Pre diagnóstico Modelo 1. **PPM2=** Precisión de Pre diagnóstico Modelo 2.
	Error Cuadrático	Margen de Error con respecto a los resultados obtenidos.	E^2 = error cuadrático Error cuadrático que existe entre la validación de prueba y las técnicas de minería de datos.

Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.6 Métodos, técnicas e instrumentos de recolección de datos

3.6.1 La observación

Es una técnica de investigación que consiste en observar personas, fenómenos, hechos, casos, objetos, acciones, situaciones, etc., con el fin de obtener determinada información necesaria para una investigación.

Se realiza la recolección de datos, la cual centraremos en los datos de entrada del sistema en estudio, es decir acerca de los componentes del sistema y las relaciones entre ellas, teniendo en cuenta los datos cuantitativos que son necesarios para el tratamiento de los mismos.

3.6.2 La entrevista

Es una técnica para obtener datos que consisten en un diálogo entre dos personas: El entrevistador “investigador” (el que hace las preguntas) y el entrevistado (el que responde a las preguntas); esta técnica es empleada con la finalidad de obtener toda la información necesaria que pueda ser brindada por la persona entrevistada, la cual, por lo general es una persona entendida en la materia de investigación. (Ver Anexos).

3.7 Procedimientos para la recolección de datos

Se recopila la información haciendo uso de la ficha de la entrevista. El objetivo de estas técnicas de recolección de información es obtener información útil para la validación de la Implementación de Técnicas de Minería de Datos para el Diagnóstico de Enfermedades de Hipertensión Arterial (Ver Anexos).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.8 Análisis Estadístico e Interpretación de los Datos.

La información que será recopilada a través de las diferentes técnicas que se aplicarán en esta investigación, ha sido tratada en un software de medición de datos (rapidminer) y HTML, se realizaron gráficos, tablas estadísticas y tabulaciones tanto en Excel 2010. El software nos permitió evaluar el comportamiento de las variables y permitió demostrar un análisis estadístico que han sido presentados en tablas y gráficos.

Tabla 2: Técnicas e instrumentos.

TECNICAS UTILIZADAS	**INSTRUMENTO DEMOSTRATIVOS**
Levantamiento de la información	Formato de entrevistas. Archivos de bases de datos
Planteamiento de la solución	Metodología KDD
Análisis de resultados	Consultas estadísticas con el software

Fuente: Elaboración propia.

3.9 Principios éticos

El ejercicio de la investigación científica y el uso del conocimiento producido por la ciencia demandan conductas éticas en el investigador. La conducta no ética carece de lugar en la práctica científica. Debe ser señalada y erradicada. Una de las funciones de la ética es la de regular la integridad misma del proceso de la investigación en cuanto a sus valores.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Para el desarrollo de esta investigación se cuenta con los siguientes valores: La base de datos es real; la información recopilada de la institución es veraz y no está manipulada a conveniencia del investigador.

Se está guardando la discreción y confidencialidad que el caso requiere en cuanto a la información confiada por la Institución.

3.9.1 Medio ambiente:

La propuesta de la solución ayudará a la predicción de diagnóstico de Hipertensión Arterial con técnicas de Minería de Datos, como toma de decisiónes en el sector salud.

3.9.2 Confidencialidad

Se asegurará la protección del modelo de predicción en la toma de decisiónes para el sector salud.

3.9.3 Objetividad

El análisis de la situación encontrada se basará en criterios técnicos e imparciales.

3.9.4 Originalidad

Se citarán las fuentes bibliográficas de la información mostrada, a fin de demostrar la inexistencia del plagio intelectual.

3.9.5 Veracidad

La Información mostrada será verdadera, cuidando la confidencialidad de ésta.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.10 Criterios de Rigor Científico.

Tabla 3: Criterios de rigor científico

Criterios	Características científicas del criterio
Confiabilidad	Se realizan limpieza de datos con la información proporcionada por ESSALUD – Chiclayo, para aplicarlos a la investigación para lograr el objetivo de realizar un pre diagnóstico de Hipertensión Arterial con ayuda de las técnicas de minería de datos.
Validación	Se validarán los instrumentos de pruebas de validación y la propuesta de solución a través de juicio de expertos.
Contrastación	Se contrastará la hipótesis a través de métodos estadísticos.

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO IV

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO IV: MARCO ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

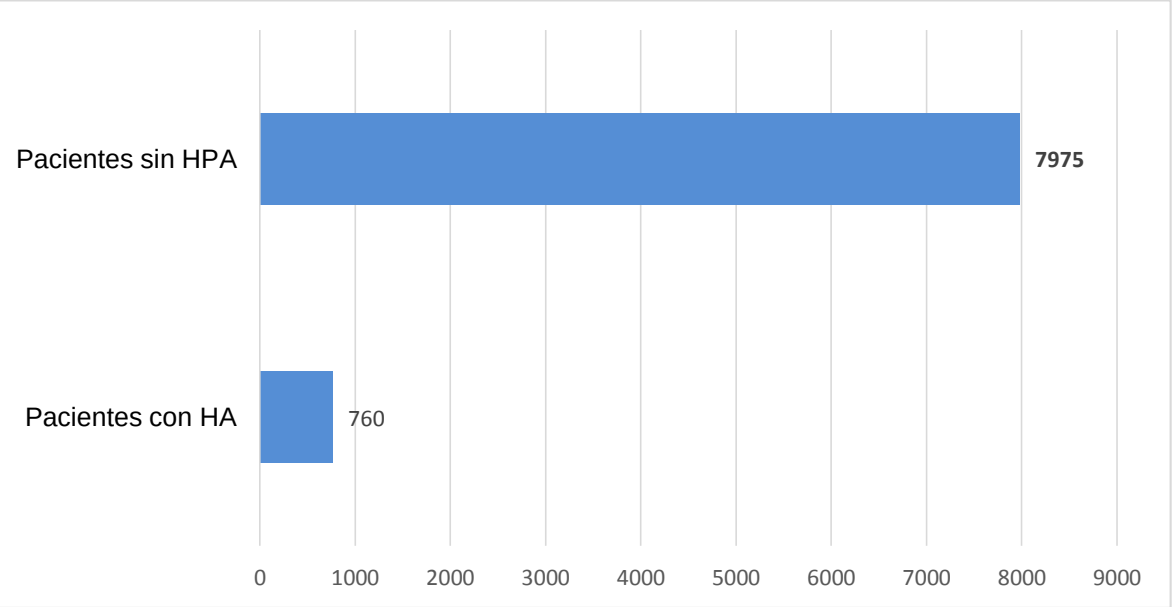
4.1 Resultados en tablas y gráficos

4.1.1 Resultados de las técnicas de minería de datos

4.1.1.1 Reglas de asociación

Registros	*Pacientes con HA*	*Pacientes sin HA*	*Nivel de confianza*
8,735	760	7,975	0.986

Gráfico 9: Resultados con técnica de reglas de asociación



Fuente: Elaboración propia

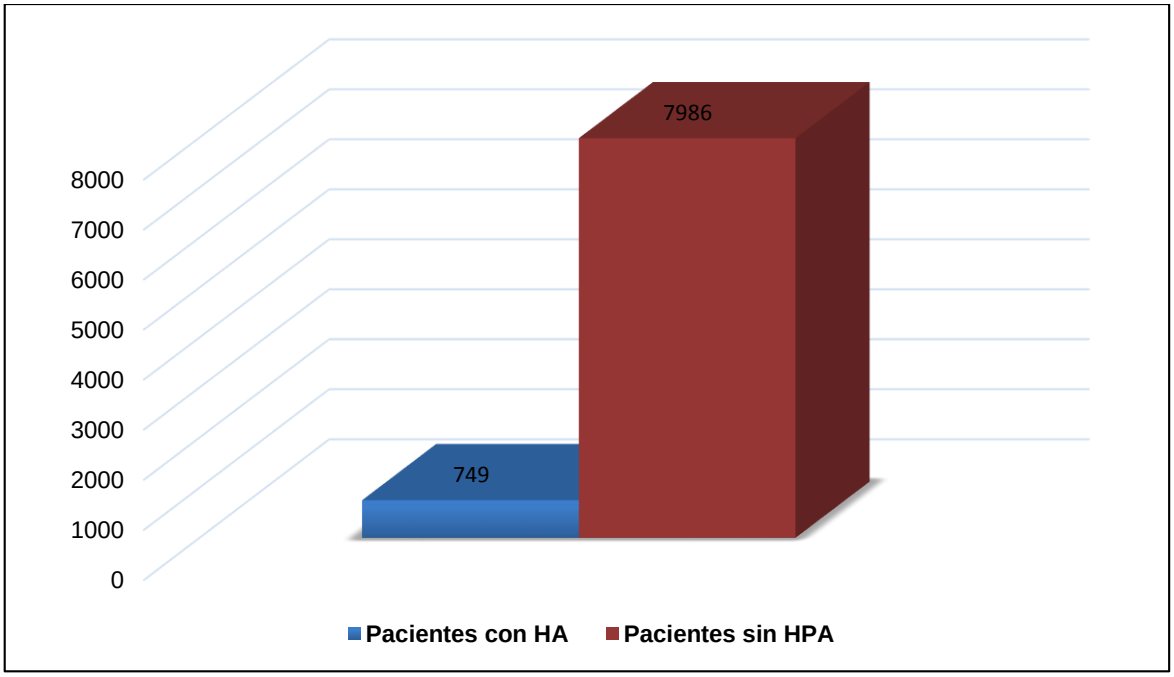
Según el estudio con la técnica de reglas de asociación para la base de datos de pacientes del ESSALUD de 8,735 registros, se llegó a la conclusión que 760 pacientes con un nivel de confianza de 98.6% podrían adquirir la enfermedad de Hipertensión Arterial, este resultado se obtuvo del mismo software.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

4.1.1.2 Técnica de árbol de decisión

Registros	*Pacientes con HA*	*Pacientes sin HA*	*Nivel de confianza*
8,735	749	7986	0.97

Gráfico 10: Resultados con técnica de árbol de decisión



Fuente: Elaboración propia

Según el estudio con la técnica de árbol de decisión para la base de datos de pacientes del ESSALUD de 8, 735 registros, se llegó a la conclusión que 749 pacientes con nivel de confianza de 97 % podrían adquirir la enfermedad de Hipertensión Arterial.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En esta técnica con respecto al nivel de confianza el software no nos brinda este ítem por lo que aplicamos una regla de tres simple para hallar el nivel de confianza y poder evaluar los resultados.

Evaluando nivel de confianza

En la técnica de reglas de asociación obtenemos el resultado de 760 pacientes con un nivel de confianza de 0.986, entonces aplicamos la regla de tres y obtenemos lo siguiente:

$$X = \frac{749 * 0.986}{760}$$

X= 0.97 para árbol de decisión

El resultado de 97% es el nivel de confianza que se calculó para la técnica de árbol de decisión.

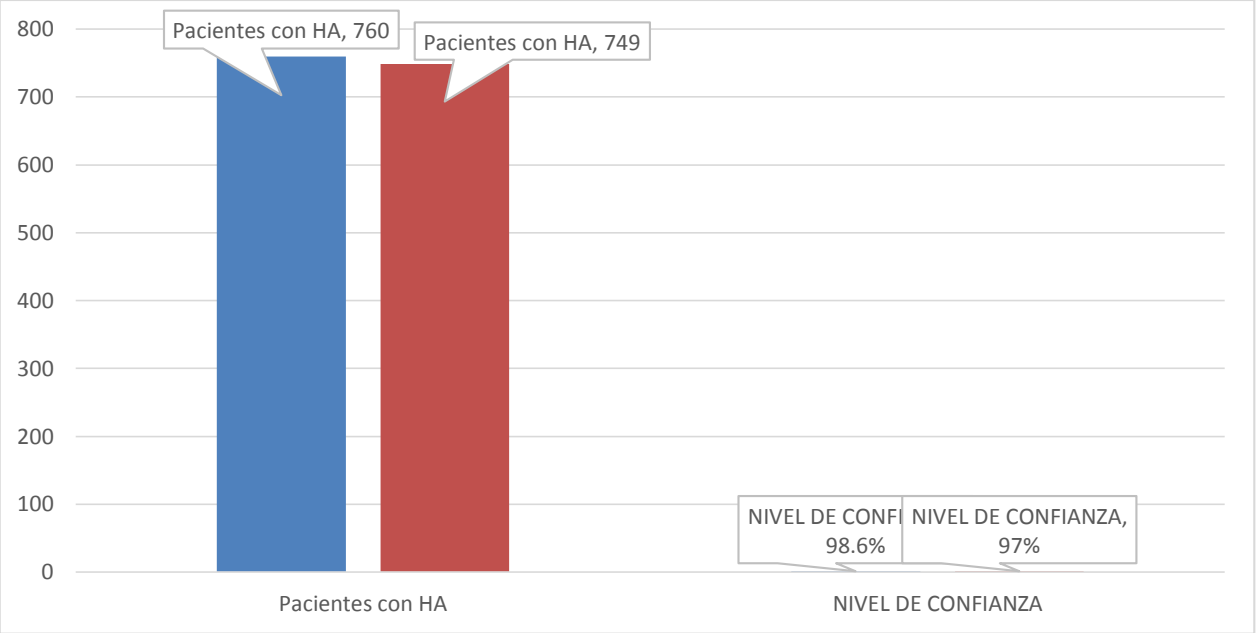
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Evaluando la técnica de regla de asociación y árbol de decisión

Pacientes con HA	*Nivel de confianza*	*Técnica de minería*
760	98.6%	Reglas de asociación
749	97%	Árbol de decisión

Gráfico 11: Técnica de reglas de asociación vs técnica de árbol de decisión



Fuente: Elaboración propia

Por lo tanto la técnica más acertada para la investigación es de técnica de reglas de asociación con un nivel de confianza de 98.6% en los resultados, obteniendo que 760 pacientes podrían adquirir enfermedad de hipertensión arterial.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

4.1.1.3 Tiempo de ejecución del modelo

El tiempo de ejecución del modelo, mide el tiempo que tarda el modelo en generar el pre diagnóstico.

TEM = TS

TEM = Tiempo de Ejecución de los Modelos

TS = Tiempo en Segundos

TEM = 300 seg.

TEM = 5 minutos.

El tiempo que tarda la ejecución de los modelos de minería de datos es de 5 minutos.

Precisión de Pre diagnóstico Modelo 1 = Performance confidence

PPM1 = Performance confidence

La evaluación dentro del software de RapidMiner se realiza a través del operador confidence, que se determina el nivel de confiablidad del modelo de Reglas de Asociación 98.6 % establecido por el software

Precisión de Pre diagnóstico Modelo 2 = Performance condicional operador.

PPM2 = Performance condicional operador

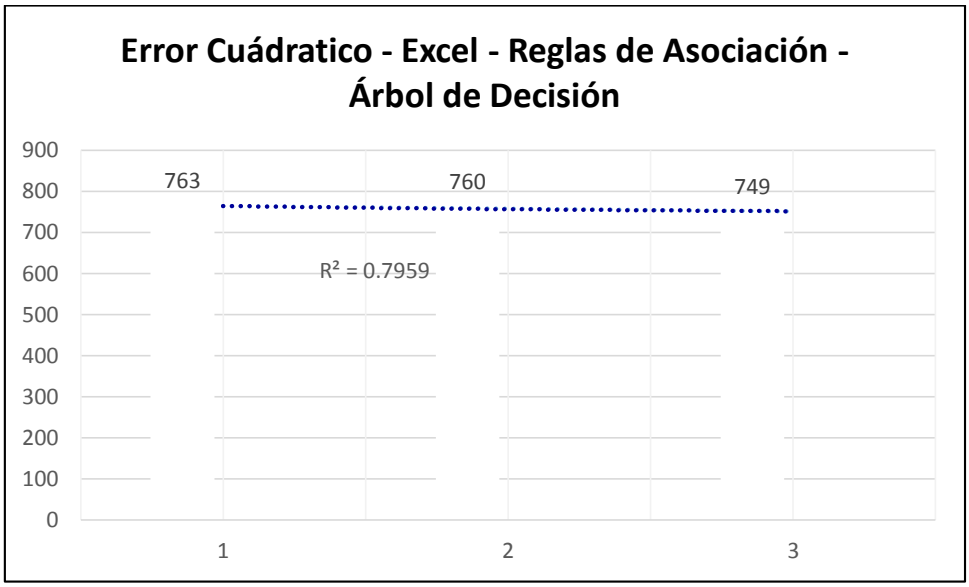
La evaluación dentro del software de RapidMiner se realiza a través del operador condicional, que se determina el nivel de confiablidad del modelo de Árbol de decisión si cumplen las premisas en los rangos establecidos en un 97% establecido por el software.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

4.1.1.4 Error Cuadrático

Pacientes con HA	*Técnica de minería*
763	Excel
760	Reglas de asociación
749	Árbol de Decisión

Gráfico: Error Cuadrático con las Validaciones de Prueba y las Técnicas de Minería de Datos



Fuente: Elaboración Propia

Cuando realizamos la comparación en una hoja de cálculo en Excel con las técnicas de minería de datos (Reglas de Asociación y Árbol de Decisión) y hallamos un error cuadrático, como se aprecia en el gráfico la línea de tendencia entre Excel y técnica de reglas de asociación se mantiene casi en el mismo nivel, en cambio con la técnica de árbol de decisión está lejos de los resultados deseados, obtuvimos un 0.7959 margen de error cuadrático en los resultados.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

4.2 Discusión de Resultados

Los resultados obtenidos representan las pruebas realizadas con el sistema, a través de la aplicación del modelo se ha logrado reducir el porcentaje de inconsistencias.

Los pre diagnóstico arrojados son en base análisis específicos para cada caso (colesterol, colesterol LDL, colesterol HDL y Triglicéridos), se estudió cada comportamiento histórico del mismo, determinado así con ayuda de los modelos de reglas de asociación y árbol de decisión planteado para un pre diagnóstico personalizado, para ello se utilizaron técnicas innovadoras como la minería de datos, orientados al proceso de pre diagnóstico y cálculo de probabilidades, además el uso de la herramienta rapidminer y metodología KDD para tratamiento de las técnicas de reglas de asociación y árbol de decisión, obteniendo como resultado el modelo de técnicas de reglas de asociación la más acertada con el 87% de nivel de confianza en los resultados de pre diagnóstico de enfermedad de hipertensión arterial.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO V

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V: PROPUESTA DE INVESTIGACIÓN

La investigación realizada tiene como función el desarrollo de una solución de pre diagnóstico de enfermedad de Hipertensión Arterial, cuya utilidad adquiere un formato destinado al ámbito de salud. La solución desarrollada debe satisfacer los criterios fundamentales. Para ello se recopiló información histórica de pacientes, que consta de 8,735 registros con ítems (Colesterol Bueno HDL, Colesterol Malo LDL, Triglicéridos) estos registros lo proporcionó el área de Informática de ESSALUD.

Gráfico 12: Datos heurísticas 2015.

1	FECHA	NOMBRE	SEXO	EDAD	COLESTERO	COLESTEROL HD	COLESTEROL LD	TRIGLICERIDO	DGX	DNI
2	02-Jan-15	ABANTO CHUQUIRUNA ITALO	M	60	138	33	65	198		27679956
3	02-Jan-15	CABALLERO ORREGO OSCAR CLODOMI	M	87	123	32	70	106	A41.9	16494412
4	02-Jan-15	CASTRO CABANILLAS JOSE EUGENIO	M	78	153	84	52	87		16622573
5	02-Jan-15	ESPINOZA VIGIL LEYLA NATALIA	F	14	99	0	0	89		73027786
6	02-Jan-15	FERNANDEZ SALAZAR HERNANDO LOR	M	66	143	36	89	87	R06.6	16645652
7	02-Jan-15	GAMARRA VDA DE TAVARA MARIA EU	F	78	204	46	99	297		16577655
8	02-Jan-15	GIL GUTIERREZ SEGUNDO	M	72	180	41	119	99		19202799
9	02-Jan-15	HERRERA ZAVALETA PEDRO ROGER	M	61	221	69	123	142		16625745
10	02-Jan-15	MAYORGA BARCO CARLOS SAMUEL	M	79	109	61	31	84		16522409
11	02-Jan-15	MENA COBOS ROSA TEODORA	F	53	273	55	143	375		17450368
12	02-Jan-15	MU¥OZ OTOLEAS ANTONIO	M	64	147	0	0	148		16581247
13	02-Jan-15	NOVOA ESQUEN LUSMIT DEL SOCORR	F	43	191	0	0	140		17591727
14	02-Jan-15	ODIAGA PEREZ FLORDELIND	F	53	164	37	99	144		33652245
15	02-Jan-15	PEÑA MAZA YUDY	F	39	142	49	64	148		27732535
16	02-Jan-15	SANCHEZ URRELO HILDA CONSUELO	F	91	138	55	58	127		16413761
17	02-Jan-15	SARAVIA RODRIGUEZ ELIZABETH	F	58	131	53	65	64		16781243
18	02-Jan-15	SUYON OLAYA JESUS SALVADOR	M	4	129	38	80	55		73563100
19	02-Jan-15	VELIZ DE HEREDIA IDELSA TEODOL	F	66	234	76	91	335		16643680
20	02-Jan-15	VENEGAS VILLALOBOS MARIA	F	56	242	40	161	202	E14.9	27667864
21	03-Jan-15	ACU¥A AYAY MARIA ELSA	F	85	224	92	114	88		16604870
22	03-Jan-15	BACA DE LINARES VICTORIA OFELI	F	83	193	73	89	151		16429227
23	03-Jan-15	BRIONES COLLANTES DORIS	F	46	150	49	66	176		27365690
24	03-Jan-15	BRUNO NAMUCHE MANUEL	M	69	231	41	138	258		19198882
25	03-Jan-15	BUSTAMANTE VASQUEZ JACOBA	F	55	254	71	147	182		16578564
26	03-Jan-15	CABRERA CORNEJO DANIEL	M	77	203	40	131	161		17434246
27	03-Jan-15	CASTILLO CACHO SEGUNDO PEDRO	M	64	125	40	63	111		16548451
28	03-Jan-15	CASTRO DE SANCHEZ ZENAIDA	F	74	136	66	50	103		16545641
29	03-Jan-15	CHERO GONZALES WILFREDO CRUZ	M	65	157	34	93	150		16553987
30	03-Jan-15	CHERRES PURISACA LUIS GONZAGA	M	48	336	49	242	228		17438559
31	03-Jan-15	CHICOMA CARMONA ALFREDO	M	70	125	42	66	88	I64.X	17427735

Fuente: Área de Informática de ESSALUD

Tabla 4: Resumen de registros proporcionados por el área de informáticos

MES	*REGISTROS*
ENERO	1219
FEBRERO	1143
MARZO	1115
ABRIL	1148
MAYO	1306
JUNIO	507
JULIO	607
AGOSTO	1418
SETIEMBRE	272
TOTAL	**8735**

Información brindada por el área de informática de ESSALUD - 2015.

5.1 Metodologías

Para poder llevar a cabo el desarrollo de la solución de pre diagnóstico, se determinó el uso de una metodología que ofrece las herramientas necesarias para un desarrollo óptimo de la solución antes mencionada, según el enfoque requerido.

Para el desarrollo del modelo de pre diagnóstico, se realizó un análisis entre las metodologías KDD y SEMMA, analizando que cualidades presentan ambas y poder determinar cuál se adecua mejor a las necesidades del ámbito de salud.

Leyenda de evaluación				
1	**2**	**3**	**4**	**5**
Totalmente en desacuerdo	En desacuerdo	Ni en acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.1.1 Evaluación de Metodología SEMMA

Tabla 5: Evaluación de metodología SEMMA

N°	PREGUNTAS	1	2	3	4	5
1	La estructura cuenta con fases, que faciliten la aplicación a la solución de pre diagnóstico.				x	
2	El Hospital cuenta con los recursos necesarios para llevar a cabo el desarrollo de esta metodología.				x	
3	El enfoque que esta metodología emplea da solución al ámbito de salud.			x		
4	El tiempo empleado en la metodología es el requerido para la solución inteligente solicitada por el área de informática del Hospital				x	
5	El modelamiento de datos propuesto favorece las necesidades de la organización de servicios de salud.		x			
6	Las herramientas que la metodología ofrece están acorde con las necesidades de la organización de servicios de salud.			x		
7	Los objetivos de la metodología están orientados a lo que la organización necesita estratégicamente.	x				
	TOTAL	21				

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.1.2 Evaluación de Metodología KDD

Tabla 6: Evaluación de metodología KDD

N°	PREGUNTAS	1	2	3	4	5
1	La estructura cuenta con fases, que faciliten la aplicación a la solución de pre diagnóstico.					x
2	El hospital cuenta con los recursos necesarios para llevar a cabo el desarrollo de esta metodología.				x	
3	El enfoque que esta metodología emplea da solución al ámbito de salud.					x
4	El tiempo empleado en la metodología es el requerido para la solución inteligente solicitada por el área de informática del Hospital					x
5	El modelamiento de datos propuesto favorece las necesidades de la organización de servicios de salud.					x
6	Las herramientas que la metodología ofrece están acorde con las necesidades de la organización de Servicios de Salud.					x
7	Los objetivos de la metodología están orientados a lo que la organización necesita estratégicamente.					x
	TOTAL	**34**				

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.2 Resultados de Evaluación de Metodologías SEMMA VS KDD

De acuerdo al análisis previo realizado

Tabla 7: Resultados de evaluación de metodologías SEMMA VS KDD

Metodología	***Puntaje***
Enfoque SEMMA	21
Enfoque KDD	34

Fuente: Elaboración propia

Por lo tanto se llega a la conclusión que para la construcción de un Modelo de pre diagnóstico, la metodología que más se adecua a las necesidades de la Organización de Servicios de Salud es la que emplea el enfoque de KDD es iterativa, además de tener un enfoque más hacia aspectos técnicos de minería, sin embargo para la solución inteligente se necesita una metodología que se centre más en desarrollo del proceso de minería de datos con fines, adicionando que el tiempo y costo con los que se cuentan son limitados.

En las siguientes secciones se describen los procesos realizados para cada fase del proyecto que garantizan su calidad y cumplimiento.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.3 Comparación de Herramienta Tecnológica

Característica	***RapidMiner***	***Peso***	***Weka***	***Peso***	***SAS***	***Peso***
Licencia	Libre	4	Libre	4	Privativa	1
Entorno de Trabajo	Gráfico	4	Gráfico	2	Gráfico	3
Integración a otros Software	Permite la Integración	4	No permite la integración	1	No permite la integración	2
Cantidad de Registros	Permite el uso de grandes volúmenes de datos	4	Permite el uso de grandes volúmenes de datos	4	Permite el uso de grandes volúmenes de datos	4
Total		16		11		14

Tabla 8: Comparación de herramienta tecnológica

Fuente: Elaboración propia

Se seleccionó la herramienta RapidMiner (16) por ser fácil de uso, ofrece una interfaz amigable, de licencia libre, además de permitir la integración de software para la publicación automática de los resultados con tecnología web y de otros sistemas.

5.3.1 Comparación de Técnicas de Minería de Datos

Tabla Nº 8.1: Leyenda de Evaluación

Nivel de Impacto	**Puntaje**
Cumple	**SI**
No Cumple	NO

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla Nº 8.2: Comparación de Técnicas de Minería de Datos

Criterios	Técnicas de Minería de Datos			
	Clustering	**R. Asociación**	**Regresión Lineal**	**Árbol de Decisión**
Toma decisiónes inteligentes sobre problemas complejos	SI	**SI**	SI	**SI**
Es utilizada para formar relaciones entre datos. Rápida y eficaz, pero insuficiente en espacios multidimensionales donde relacionan más de 2 variables	NO	NO	SI	NO
Dada una base de datos, se construyen diagramas de construcciones lógica	NO	**SI**	NO	**SI**
Sirve para representar y categorizar una serie de condiciones que suceden de forma sucesiva	NO	NO	NO	**SI**
Descubrir hechos que ocurren en común dentro de un conjunto de datos	SI	**SI**	NO	NO
TOTAL	2	**3**	2	**3**

Fuente: Elaboración Propia

Después de realizar el análisis correspondiente, será mediante la técnica de reglas de asociación y árbol de decisión para el caso de investigación de predicción de diagnóstico de hipertensión arterial, ya que son los que más se adecua para cumplir con todos los requisitos que se requieren.

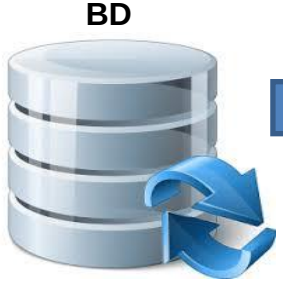
Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

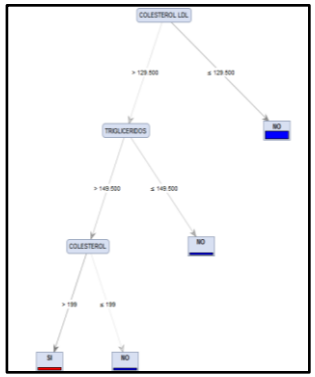
5.4 Arquitectura del proyecto de técnicas de minería de datos para el pre diagnóstico de hipertensión arterial

Gráfico 13: Arquitectura del proyecto de técnicas de minería de datos.

RAPIDMINER

BD **T.M** **REGLAS DE ASOCIACIÓN** **T.M – ÁRBOL DE DECISIÓN**



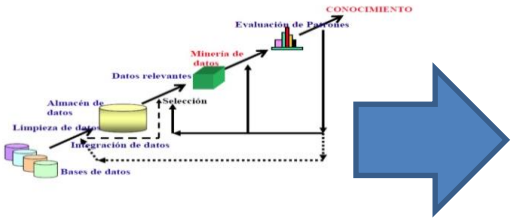


DATOS HEURÍSTICOS DE PACIENTES

GRAFICA

	FECHA	NOMBRE	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	PRESION
2	16-Mar-15	WIQUE VDA DE CAYCAY LAURA ANTO	191	0	0	124	NO
3	31-Jun-15	WIQUE VDA DE CAYCAY LAURA ANTO	230	63	136	155	SI
4	25-Jul-15	WIQUE VDA DE CAYCAY LAURA ANTO	142	26	74	299	NO
5	31-Jan-15	WIQUEN MEDANERO AVELINA	194	0	0	0	NO
6	21-May-15	WIQUEN MILLONES SUSANA	201	41	122	192	NO
7	20-Feb-15	WIQUEN VALLEJOS EMMA EUMELIA	224	0	0	140	NO
8	29-May-15	WIQUEN VALLEJOS EMMA EUMELIA	243	58	157	140	NO
9	05-Sep-15	WIQUEN VALLEJOS EMMA EUMELIA	230	0	0	124	NO
10	31-Jan-15	WIQUEN VALLEJOS LILIANA MARIA	177	0	0	125	NO
11	28-May-15	WIQUEN VALLEJOS LILIANA MARIA	215	38	152	125	NO
12	21-Aug-15	WOPO RICO ROSA GLADYS	254	0	0	63	NO
13	06-Apr-15	WOPO VERA JOSE OMER	190	0	0	110	NO
14	25-Apr-15	ABAD ABAD MARINA	173	0	0	91	NO
15	11-May-15	ABAD BARON CINTHYA MIRHANY	0	44	191	284	NO
16	14-May-15	ABAD CUNYA GABINO	188	59	110	96	NO
17	07-Jan-15	ABAD DE NIQUE JUANA CONSTANCIA	179	32	92	276	NO
18	06-Jun-15	ABAD LITANO HUMBERTO	213	0	0	413	NO
19	11-Jun-15	ABAD SANTIN FANNY M	278	0	0	95	NO
20	03-Sep-15	ABAD SANTIN FANNY M	283	0	0	141	NO
21	03-Mar-15	ABANTO CERDAN SILVERO	160	56	93	56	NO
22	02-Jan-15	ABANTO CHUQUIRUNA ITALO	138	33	65	198	NO
23	09-Jan-15	ABANTO CUBAS JHENY ELIZABET	144	55	79	48	NO
24	20-Aug-15	ABANTO DE OLANO MARIA FIDENCIA	230	58	145	133	NO
25	02-Mar-15	ABANTO LAMAS ALEIDA	220	35	145	199	SI
26	15-May-15	ABANTO LAMAS ALEIDA	206	32	90	419	NO
27	11-May-15	ABANTO VERA VENCESLAO	229	73	129	135	NO
28	18-Aug-15	ABANTO ZELADA ENRIQUE	165	42	88	177	NO
29	29-Apr-15	ACARO MIRANDA ROSA VICTORIA	173	0	0	78	NO
30	02-May-15	ACHA ALBERCA DOSITEO	190	52	109	145	NO
31	29-Apr-15	ACHA CORTEZ AMPARO ESTANI	136	0	0	114	NO
32	10-Jun-15	ACHA CHOPPA AMPARO ESTANI	200	57	123	396	NO

METODOLOGIA KDD



INTERFAZ

Fuente: Elaboración propia

5.5 Aplicación de la metodología KDD

5.5.1 Fase I: Selección de datos

Se selecciona los datos con los que se va a trabajar; con 8735 registros de pacientes; esta información se encuentra en Microsoft Excel 2010, esta fase se utiliza para la técnica de asociación y árbol de decisión.

Gráfico 14: Selección de datos

1	FECHA	NOMBRE	SEXO	EDAD	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	DGX	DNI
2	16-Mar-15	¥IQUE VDA DE CAYCAY LAURA ANTO	F	59	191	0	0	124		16648112
3	01-Jun-15	¥IQUE VDA DE CAYCAY LAURA ANTO	F	59	230	63	136	155		16648112
4	25-Jul-15	¥IQUE VDA DE CAYCAY LAURA ANTO	F	59	142	26	74	206		16648112
5	31-Jan-15	¥IQUEN MEDIANERO AVELINA	F	86	194	0	0	0		16514431
6	21-Mar-15	¥IQUEN MILLONES SUSANA	F	73	201	41	122	192		16514900
7	20-Feb-15	¥IQUEN VALLEJOS EMMA EUMELIA	F	62	224	0	0	140		16519613
8	29-May-15	¥IQUEN VALLEJOS EMMA EUMELIA	F	62	243	58	157	140		16519613
9	05-Sep-15	¥IQUEN VALLEJOS EMMA EUMELIA	F	62	230	0	0	124		16519613
10	31-Jan-15	¥IQUEN VALLEJOS LILIANA MARIA	F	58	177	0	0	125		16419325
11	28-May-15	¥IQUEN VALLEJOS LILIANA MARIA	F	58	215	38	152	125		16419325
12	21-Aug-15	¥OPO RICO ROSA GLADYS	F	50	254	0	0	63		16471299
13	06-Apr-15	¥OPO VERA JOSE OMER	M	69	190	0	0	110		16441709
14	25-Apr-15	ABAD ABAD MARINA	F	48	173	0	0	91		33668262
15	11-May-15	ABAD BARON CINTHYA MIRHANY	F	18	0	44	191	284		74809547
16	14-May-15	ABAD CUNYA GABINO	M	54	188	59	110	96	I06.0	02604724
17	07-Jan-15	ABAD DE NIQUE JUANA CONSTANCIA	F	64	179	32	92	276		17536161
18	08-Jun-15	ABAD LITANO HUMBERTO	M	55	213	0	0	413	A09.X	03835756
19	11-Jun-15	ABAD SANTIN FANNY M	F	49	278	0	0	95		16615267
20	03-Sep-15	ABAD SANTIN FANNY M	F	49	283	0	0	141		16615267
21	03-Mar-15	ABANTO CERDAN SILVERIO	M	83	160	56	93	56		19195864
22	02-Jan-15	ABANTO CHUQUIRUNA ITALO	M	60	138	33	65	198		27679956
23	09-Jan-15	ABANTO CUBAS JHENY ELIZABET	F	16	144	55	79	48		71435299
24	20-Aug-15	ABANTO DE OLANO MARIA FIDENCIA	F	80	230	58	145	133		16443965
25	02-May-15	ABANTO LAMAS ALEIDA	F	46	220	35	145	199		16680042
26	15-May-15	ABANTO LAMAS ALEIDA	F	46	206	32	90	419		16680042
27	11-May-15	ABANTO VERA WENCESLAO	M	76	220	73	120	135		17993181
28	18-Aug-15	ABANTO ZELADA ENRIQUE	M	59	165	42	88	177		26630857
29	29-Apr-15	ACARO MIRANDA ROSA VICTORIA	F	63	173	0	0	78	N39.9	16493764
30	02-May-15	ACHA ALBERCA DOSITEO	M	78	190	52	109	145		33643381
31	29-Apr-15	ACHA CORTEZ AMPARO ESTANI	F	71	136	0	0	114		16632655
32	10-Aug-15	ACHA CORTEZ AMPARO ESTANI	F	71	200	37	122	206		16632655
33	27-Feb-15	ACHA TOGAS MARIIA ISILDA	F	45	251	43	167	209		27714110
34	25-Mar-15	ACOSTA CASTILLO EMERITA	F	74	186	36	91	294		17533777
35	08-Jun-15	ACOSTA COLLAZOS LETTY FLOR	F	33	211	0	0	66		41420640
36	05-Feb-15	ACOSTA DE ENRIQUEZ MARIA LUISA	F	78	153	0	0	236		16528180
37	07-Apr-15	ACOSTA DE ENRIQUEZ MARIA LUISA	F	78	194	0	0	279		16528180
38	04-May-15	ACOSTA DE ENRIQUEZ MARIA LUISA	F	78	221	42	126	265		16528180
39	02-Sep-15	ACOSTA DE ENRIQUEZ MARIA LUISA	F	78	239	0	0	221		16528180

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.5.2 Fase II: Pre procesamiento y limpieza

En esta Fase es donde se realiza un tratamiento de los datos incorrectos y ausentes.

Medición de parámetros

En las tablas se observa los rangos de colesterol y triglicéridos, con el nivel deseable, alto o muy alto.

Tablas 9: rangos de colesterol y triglicéridos, con el nivel deseable, alto o muy alto.

COLESTEROL	
Nivel deseable	Menos de 200 mg/dl (menos de 5,172 mmol/l)
Límite alto	200-240 mg/dl (5,17 – 6,19 mmol/l)
Alto	Más de 240 mg/dl (más de 6,19 mmol/l)

TRIGLICERIDOS	
Deseable:	Menos de 150 mg/dl (menos de 1,69 mmol/l)
Límite alto:	Entre 150 – 400 mg/dl (1,69 – 4,52 mmol/l)
Alto	Entre 400 – 1000 mg/dl (4,52 – 11,29 mmol/l)
Muy alto	Más de 1000 mg/dl (más de 11,29 mmol/l)

Niños y adolescentes de (02 - 19) años de edad.

	COLESTEROL TOTAL	COLESTEROL LDL
Deseable	Menos de 170 mg/dl	Menos de 110 mg/dl
Límite alto	170 a 200 mg/dl	110 a 130 mg/dl
Alto	Más de 200 mg	Más de 130 mg/d

Antes de empezar con el pre procesamiento y limpieza en Microsoft Excel utilizamos fórmulas que para ello se tiene en cuenta las tablas anteriores para poder elaborarlas.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 10: Fórmulas para el procesamiento de datos.

ITEMS	***FORMULAS***
Colesterol	SI(EDAD<=19,SI(COLESTEROL<=170,0,1),SI(EDAD >19,SI(COLESTEROL<200,0,1)))
Trigliceridos	SI(TRIGLICERIDOS<150,0,1)
Colesterol HDL	SI(COLESTEROL HDL<35,0,1)
Colesterol LDL	SI(EDAD<=19,SI(COLESTEROL LDL<=100,0,1),SI(EDAD>19,SI(COLESTEROL LDL<130,0,1)))

Fuente: Elaboración propia

En esta fase para ambas técnicas se emplean las fórmulas antes mencionadas.

Técnica de reglas de asociación

De acuerdo a la edad se aplica una fórmula para el proceso de conversión de los datos a cero o uno según el intervalo de colesterol total, colesterol HDL, colesterol LDL y los triglicéridos.

Gráfico 15: Preprocesamiento de base de datos – reglas de asociación

	ORDEN	FECHA	NOMBRE	EDAD	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	SEXO	DNI
2	1	16-Mar-15	¥IQUE VDA DE CAYCAY LAURA ANTO	59	0	0	0	0	0	16648112
3	2	01-Jun-15	¥IQUE VDA DE CAYCAY LAURA ANTO	59	1	1	1	1	0	16648112
4	3	25-Jul-15	¥IQUE VDA DE CAYCAY LAURA ANTO	59	0	0	0	1	0	16648112
5	4	31-Jan-15	¥IQUEN MEDIANERO AVELINA	86	0	0	0	0	0	16514431
6	5	21-Mar-15	¥IQUEN MILLONES SUSANA	73	1	1	0	1	0	16514900
7	6	20-Feb-15	¥IQUEN VALLEJOS EMMA EUMELIA	62	1	0	0	0	0	16519613
8	7	29-May-15	¥IQUEN VALLEJOS EMMA EUMELIA	62	1	1	1	0	0	16519613
9	8	05-Sep-15	¥IQUEN VALLEJOS EMMA EUMELIA	62	1	0	0	0	0	16519613
10	9	31-Jan-15	¥IQUEN VALLEJOS LILIANA MARIA	58	0	0	0	0	0	16419325
11	10	28-May-15	¥IQUEN VALLEJOS LILIANA MARIA	58	1	1	1	0	0	16419325
12	11	21-Aug-15	¥OPO RICO ROSA GLADYS	50	1	0	0	0	0	16471299
13	12	06-Apr-15	¥OPO VERA JOSE OMER	69	0	0	0	0	1	16441709
14	13	25-Apr-15	ABAD ABAD MARINA	48	0	0	0	0	0	33668262
15	14	11-May-15	ABAD BARON CINTHYA MIRHANY	18	0	1	1	1	0	74809547
16	15	14-May-15	ABAD CUNYA GABINO	54	0	1	0	0	1	02604724
17	16	07-Jan-15	ABAD DE NIQUE JUANA CONSTANCIA	64	0	0	0	1	0	17536161
18	17	08-Jun-15	ABAD LITANO HUMBERTO	55	1	0	0	1	1	03835756
19	18	11-Jun-15	ABAD SANTIN FANNY M	49	1	0	0	0	0	16615267
20	19	03-Sep-15	ABAD SANTIN FANNY M	49	1	0	0	0	0	16615267
21	20	03-Mar-15	ABANTO CERDAN SILVERIO	83	0	1	0	0	1	19195864
22	21	02-Jan-15	ABANTO CHUQUIRUNA ITALO	60	0	0	0	1	1	27679956
23	22	09-Jan-15	ABANTO CUBAS JHENY ELIZABET	16	0	1	0	0	0	71435299
24	23	20-Aug-15	ABANTO DE OLANO MARIA FIDENCIA	80	1	1	1	0	0	16443965
25	24	02-May-15	ABANTO LAMAS ALEIDA	46	1	1	1	1	0	16680042
26	25	15-May-15	ABANTO LAMAS ALEIDA	46	1	0	0	1	0	16680042
27	26	11-May-15	ABANTO VERA WENCESLAO	76	1	1	0	0	1	17993181
28	27	18-Aug-15	ABANTO ZELADA ENRIQUE	59	0	1	0	1	1	26630857
29	28	29-Apr-15	ACARO MIRANDA ROSA VICTORIA	63	0	0	0	0	0	16493764
30	29	02-May-15	ACHA ALBERCA DOSITEO	78	0	1	0	0	1	33643381
31	30	29-Apr-15	ACHA CORTEZ AMPARO ESTANI	71	0	0	0	0	0	16632655
32	31	10-Aug-15	ACHA CORTEZ AMPARO ESTANI	71	1	1	0	1	0	16632655
33	32	27-Feb-15	ACHA TOGAS MARIA ISILDA	45	1	1	1	1	0	27714110
34	33	25-Mar-15	ACOSTA CASTILLO EMERITA	74	0	1	0	1	0	17533777
35	34	08-Jun-15	ACOSTA COLLAZOS LETTY FLOR	33	1	0	0	0	0	41420640
36	35	05-Feb-15	ACOSTA DE ENRIQUEZ MARIA LUISA	78	0	0	0	1	0	16528180
37	36	07-Apr-15	ACOSTA DE ENRIQUEZ MARIA LUISA	78	0	0	0	1	0	16528180
38	37	04-May-15	ACOSTA DE ENRIQUEZ MARIA LUISA	78	1	1	0	1	0	16528180
39	38	02-Sep-15	ACOSTA DE ENRIQUEZ MARIA LUISA	78	1	0	0	1	0	16528180

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

De estos datos se realiza la limpieza de datos, si es menor de 19 años y su colesterol es menor de 170 mg/dl entonces es normal y si tiene más de 170 mg/dl tiene colesterol alto; si es mayor de 19 años y su colesterol es menor que 200 mg/dl entonces es normal, sino tiene colesterol alto.

Si tiene triglicéridos menor a 150 mg/dl entonces es normal sino tiene riesgo.

Si tiene el tolesterol HDL es menor de 35 mg/dl entonces es normal sino tiene riesgo.

Gráfico 16: Datos de limpieza de la base de datos – reglas de asociación

1	ITEM	FECHA	NOMBRE	COLESTEROL	COLESTEROL	COLESTEROL	TRIGLICERIDOS
2	1	16-mar-15	¥IQUE VDA DE CAYCAY LAURA	0	0	0	0
3	2	01-jun-15	¥IQUE VDA DE CAYCAY LAURA	1	1	1	1
4	3	25-jul-15	¥IQUE VDA DE CAYCAY LAURA	0	0	0	1
5	4	31-Jan-15	¥IQUEN MEDIANERO AVELINA	0	0	0	0
6	5	21-mar-15	¥IQUEN MILLONES SUSANA	1	1	0	1
7	6	20-feb-15	¥IQUEN VALLEJOS EMMA EUM	1	0	0	0
8	7	29-may-15	¥IQUEN VALLEJOS EMMA EUM	1	1	1	0
9	8	05-sep-15	¥IQUEN VALLEJOS EMMA EUM	1	0	0	0
10	9	31-Jan-15	¥IQUEN VALLEJOS LILIANA MA	0	0	0	0
11	10	28-may-15	¥IQUEN VALLEJOS LILIANA MA	1	1	1	0
12	11	21-Aug-15	¥OPO RICO ROSA GLADYS	1	0	0	0
13	12	06-Apr-15	¥OPO VERA JOSE OMER	0	0	0	0
14	13	25-Apr-15	ABAD ABAD MARINA	0	0	0	0
15	14	11-may-15	ABAD BARON CINTHYA MIRHA	0	1	1	1
16	15	14-may-15	ABAD CUNYA GABINO	0	1	0	0
17	16	07-Jan-15	ABAD DE NIQUE JUANA CONS	0	0	0	1
18	17	08-jun-15	ABAD LITANO HUMBERTO	1	0	0	1
19	18	11-jun-15	ABAD SANTIN FANNY M	1	0	0	0
20	19	03-sep-15	ABAD SANTIN FANNY M	1	0	0	0
21	20	03-mar-15	ABANTO CERDAN SILVERIO	0	1	0	0
22	21	02-Jan-15	ABANTO CHUQUIRUNA ITALO	0	0	0	1
23	22	09-Jan-15	ABANTO CUBAS JHENY ELIZAB	0	1	0	0
24	23	20-Aug-15	ABANTO DE OLANO MARIA FI	1	1	1	0
25	24	02-may-15	ABANTO LAMAS ALEIDA	1	1	1	1
26	25	15-may-15	ABANTO LAMAS ALEIDA	1	0	0	1
27	26	11-may-15	ABANTO VERA WENCESLAO	1	1	0	0
28	27	18-Aug-15	ABANTO ZELADA ENRIQUE	0	1	0	1
29	28	29-Apr-15	ACARO MIRANDA ROSA VICTO	0	0	0	0

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Si es menor de 19 años y su colesterol LDL es menor de 110 mg/dl entonces es normal y si tiene más de 110 mg/dl tiene colesterol alto; si es mayor de 19 años y su colesterol es menor que 130 mg/dl entonces es normal sino tiene riesgo; quedando de esta manera:

Técnica de Árbol de Decisión

Para está técnica la limpieza de datos se da cuando utilizamos las fórmulas en Microsoft Excel, según los que cumplen los rangos establecidos (colesterol, colesterol HDL, colesterol LDL, triglicéridos), agregamos una columna (Presión) de esta manera se realiza la limpieza de datos obteniendo 0 (NO) y 1 (SI), según sea el caso.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 17: Limpieza de la base de datos – árbol de decisión

	FECHA	NOMBRE	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	PRESION
2	16-Mar-15	¥IQUE VDA DE CAYCAY LAURA ANTO	191	0	0	124	NO
3	01-Jun-15	¥IQUE VDA DE CAYCAY LAURA ANTO	230	63	136	155	SI
4	25-Jul-15	¥IQUE VDA DE CAYCAY LAURA ANTO	142	26	74	206	NO
5	31-Jan-15	¥IQUEN MEDIANERO AVELINA	194	0	0	0	NO
6	21-Mar-15	¥IQUEN MILLONES SUSANA	201	41	122	192	NO
7	20-Feb-15	¥IQUEN VALLEJOS EMMA EUMELIA	224	0	0	140	NO
8	29-May-15	¥IQUEN VALLEJOS EMMA EUMELIA	243	58	157	140	NO
9	05-Sep-15	¥IQUEN VALLEJOS EMMA EUMELIA	230	0	0	124	NO
10	31-Jan-15	¥IQUEN VALLEJOS LILIANA MARIA	177	0	0	125	NO
11	28-May-15	¥IQUEN VALLEJOS LILIANA MARIA	215	38	152	125	NO
12	21-Aug-15	¥OPO RICO ROSA GLADYS	254	0	0	63	NO
13	06-Apr-15	¥OPO VERA JOSE OMER	190	0	0	110	NO
14	25-Apr-15	ABAD ABAD MARINA	173	0	0	91	NO
15	11-May-15	ABAD BARON CINTHYA MIRHANY	0	44	191	284	NO
16	14-May-15	ABAD CUNYA GABINO	188	59	110	96	NO
17	07-Jan-15	ABAD DE NIQUE JUANA CONSTANCIA	179	32	92	276	NO
18	08-Jun-15	ABAD LITANO HUMBERTO	213	0	0	413	NO
19	11-Jun-15	ABAD SANTIN FANNY M	278	0	0	95	NO
20	03-Sep-15	ABAD SANTIN FANNY M	283	0	0	141	NO
21	03-Mar-15	ABANTO CERDAN SILVERIO	160	56	93	56	NO
22	02-Jan-15	ABANTO CHUQUIRUNA ITALO	138	33	65	198	NO
23	09-Jan-15	ABANTO CUBAS JHENY ELIZABET	144	55	79	48	NO
24	20-Aug-15	ABANTO DE OLANO MARIA FIDENCIA	230	58	145	133	NO
25	02-May-15	ABANTO LAMAS ALEIDA	220	35	145	199	SI
26	15-May-15	ABANTO LAMAS ALEIDA	206	32	90	419	NO
27	11-May-15	ABANTO VERA WENCESLAO	220	73	120	135	NO
28	18-Aug-15	ABANTO ZELADA ENRIQUE	165	42	88	177	NO
29	29-Apr-15	ACARO MIRANDA ROSA VICTORIA	173	0	0	78	NO
30	02-May-15	ACHA ALBERCA DOSITEO	190	52	109	145	NO
31	29-Apr-15	ACHA CORTEZ AMPARO ESTANI	136	0	0	114	NO
32	10-Aug-15	ACHA CORTEZ AMPARO ESTANI	200	37	122	206	NO

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

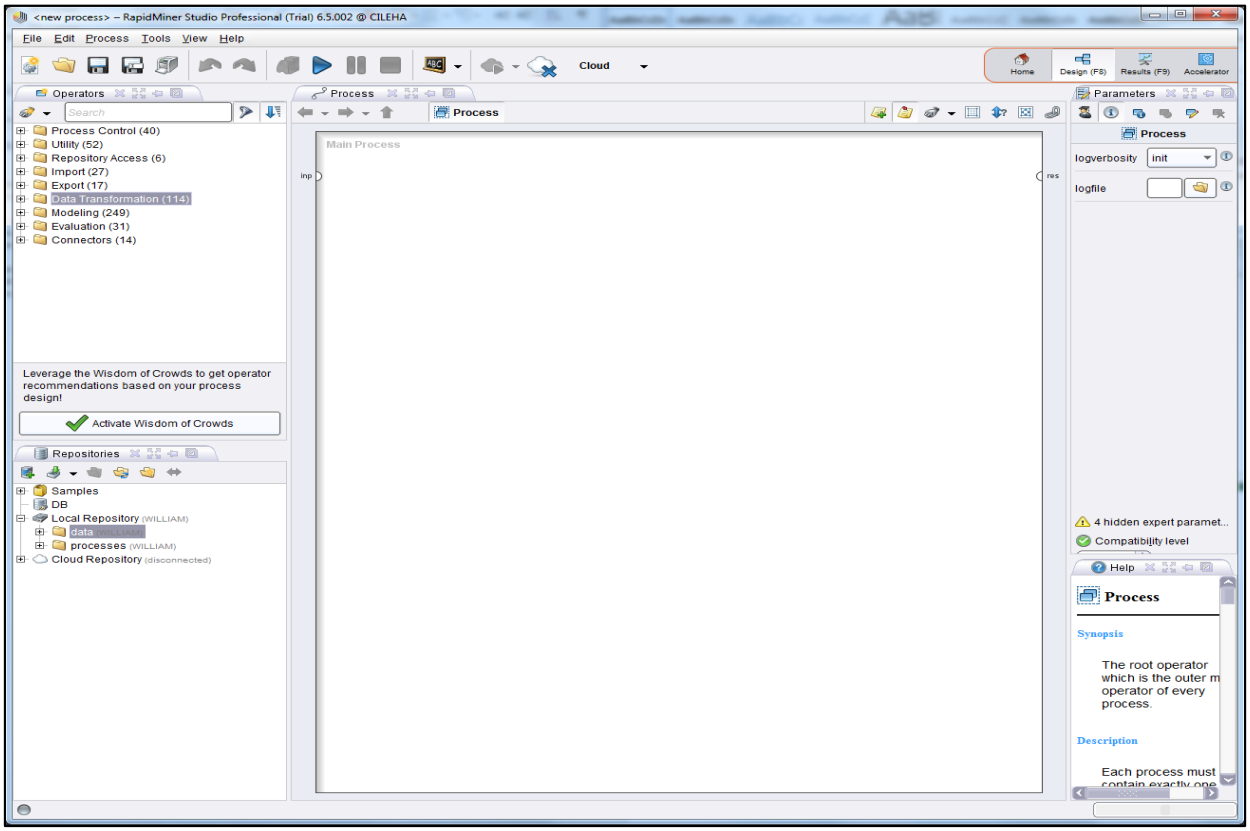
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.5.3 Fase III: Transformación y carga

Transformación de los datos y reducción de la dimensionalidad.

En esta fase se está empleando la herramienta RapidMiner que es de Open Source especial para predicciones con la **técnica de reglas de asociación.**

Gráfico 18: Entorno de RapidMiner



Fuente: RapidMiner

Aquí se cargan los datos para transformarlos y que el software los interprete en RapidMiner.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Obtenemos la importación de la BD de Microsoft Excel a RapidMiner para técnica reglas de asociación.

Gráfico 19: Datos en RapidMiner

16	07-Jan-15	ABAD DE NIQUE JUAN/	0	0	0	1
17	08-jun-15	ABAD LITANO HUMBER	1	0	0	1
18	11-jun-15	ABAD SANTIN FANNY M	1	0	0	0
19	03-sep-15	ABAD SANTIN FANNY M	1	0	0	0
20	03-mar-15	ABANTO CERDAN SILV	0	1	0	0
21	02-Jan-15	ABANTO CHUQUIRUNA	0	0	0	1
22	09-Jan-15	ABANTO CUBAS JHEN'	0	1	0	0
23	20-Aug-15	ABANTO DE OLANO MA	1	1	1	0
24	02-may-15	ABANTO LAMAS ALEIDA	1	1	1	1
25	15-may-15	ABANTO LAMAS ALEIDA	1	0	0	1
26	11-may-15	ABANTO VERA WENCE	1	1	0	0
27	18-Aug-15	ABANTO ZELADA ENRI	0	1	0	1
28	29-Apr-15	ACARO MIRANDA ROSA	0	0	0	0
29	02-may-15	ACHA ALBERCA DOSIT	0	1	0	0
30	29-Apr-15	ACHA CORTEZ AMPARI	0	0	0	0
31	10-Aug-15	ACHA CORTEZ AMPARI	1	1	0	1

Fuente: Elaboración propia

Selección de variables: dependiente e independientes en RapidMiner

Gráfico 20: Variables dependiente e independiente en rapidminer

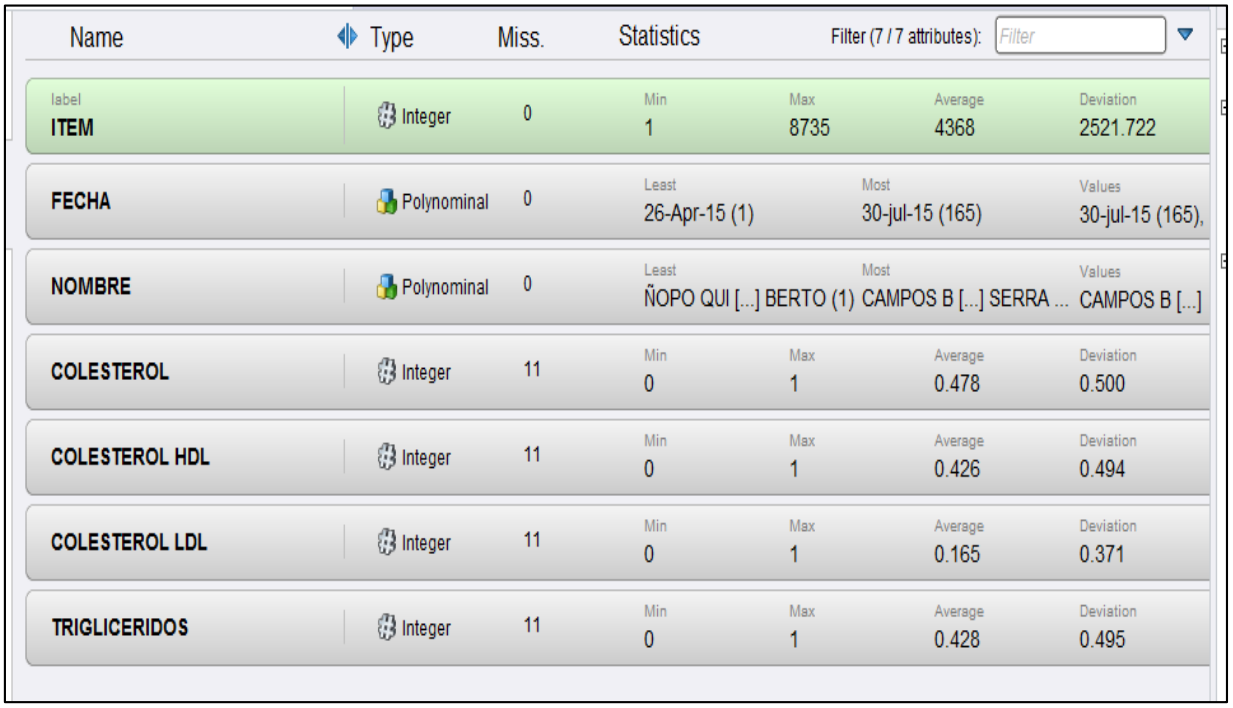
ITEM	FECHA	NOMBRE	COLESTERI	COLESTERI	COLESTERI	TRIGLICERI
integer	polyno...	polyno...	integer	integer	integer	integer
label	attribute	attribute	attribute	attribute	attribute	attribute
1	16-mar-15	¥IQUE VDA I	0	0	0	0
2	01-jun-15	¥IQUE VDA I	1	1	1	1
3	25-jul-15	¥IQUE VDA I	0	0	0	1
4	31-Jan-15	¥IQUEN MEI	0	0	0	0
5	21-mar-15	¥IQUEN MIL	1	1	0	1
6	20-feb-15	¥IQUEN VAL	1	0	0	0
7	29-may-15	¥IQUEN VAL	1	1	1	0
8	05-sep-15	¥IQUEN VAL	1	0	0	0
9	31-Jan-15	¥IQUEN VAL	0	0	0	0
10	28-may-15	¥IQUEN VAL	1	1	1	0
11	21-Aug-15	¥OPO RICO	1	0	0	0
12	06-Apr-15	¥OPO VERA	0	0	0	0
13	25-Apr-15	ABAD ABAD	0	0	0	0

Fuente: RapidMiner

Al procesar tenemos que los tipos de datos son enteros y hay que transformarlos

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 21: Transformación de datos a binomial



Name	Type	Miss.	Statistics				Filter (7 / 7 attributes): Filter
label ITEM	Integer	0	Min 1	Max 8735	Average 4368	Deviation 2521.722	
FECHA	Polynominal	0	Least 26-Apr-15 (1)	Most 30-jul-15 (165)	Values 30-jul-15 (165),		
NOMBRE	Polynominal	0	Least ÑOPO QUI [...] BERTO (1)	Most CAMPOS B [...] SERRA ...	Values CAMPOS B [...]		
COLESTEROL	Integer	11	Min 0	Max 1	Average 0.478	Deviation 0.500	
COLESTEROL HDL	Integer	11	Min 0	Max 1	Average 0.426	Deviation 0.494	
COLESTEROL LDL	Integer	11	Min 0	Max 1	Average 0.165	Deviation 0.371	
TRIGLICERIDOS	Integer	11	Min 0	Max 1	Average 0.428	Deviation 0.495	

Fuente: RapidMiner

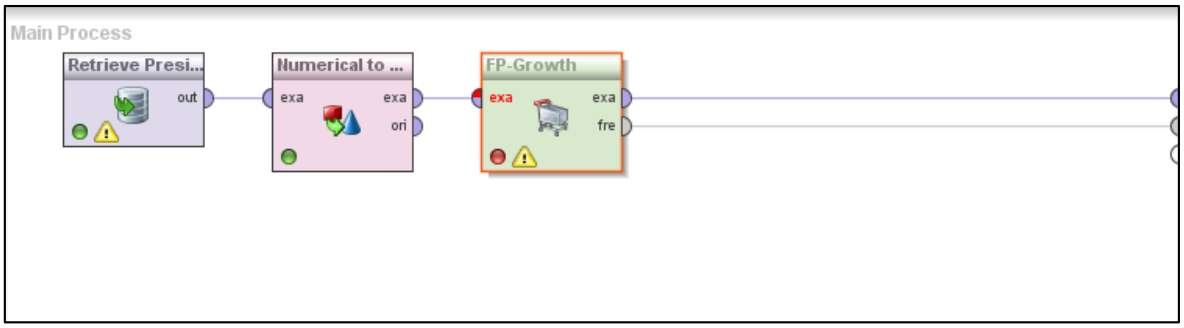
Lo transformamos de numéricos a binomial para procesar los datos, porque son dos resultados o probabilidades que se obtiene como dos categorías (éxito o fracaso). De tal manera que los ceros y unos ahora son falsos y verdaderos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Modelado

Gráfico 22: Lectura de datos – reglas de asosiación



Fuente: RapiMiner

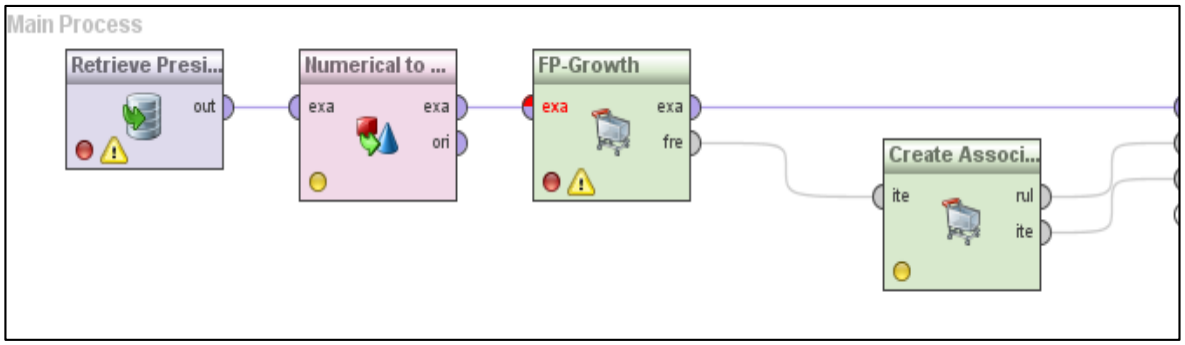
En este proceso se hará una lectura de los datos como se aprecia en la imagen.

En el RapidMiner se importa los datos de Microsoft Excel, para la técnica de regla de asociación.

Entrenamiento

En este proceso se realiza el entrenamiento y validación de los datos, estos deben ser en un periodo de tiempo. En este proceso se añade el operador que permitirá medir el performance del modelo.

Gráfico 23: Entrenamiento y testeo de los datos – reglas de asosiación



Fuente: RapiMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Retrive prezi

Es el operador para leer los datos del repositorio estos datos están en el formato de unos y ceros. Se ha realizado el procedimiento de limpieza para la base de datos.

Numerical to binomial

Transforma los datos unos y ceros en verdaderos y falsos. Es el formato que requiere el siguiente operador binomial. Sea 1(éxito) o 0 (fracaso) para los casos que se de en el estudio.

FP-Growth

Este operador requiere el formato de verdaderos y falsos. Este operador encuentra en la base de datos los conjuntos de asociación para los ítems establecidos por el objeto de estudio.

Create associaton rules

Se encuentran las reglas de asociación que pueden construirse en base a los conjuntos frecuentes que produce el operador anterior.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Reglas de asociación

Asociación se define como:

Sea $I = \{ i_1, i_2, \ldots, i_n \}$ un conjunto de n atributos binarios llamados ítems.

Sea $D = \{ t_1, t, \ldots, t_n \}$ un conjunto de transacciones almacenadas en una base de datos.

Cada transacción en D tiene un ID (identificador) único y contiene un subconjunto de items de I. Una regla se define como una implicación de la forma:

$$X \Rightarrow Y$$

Donde:

$X, Y \subseteq I$ y

$X \cap Y = \emptyset$

Los conjuntos de items X y Y se denominan respectivamente "antecedente" (o parte izquierda) y "consecuente" (o parte derecha) de la regla.

Tabla 11: Lípidos en la sangre.

ID	*Colesterol LDL*	*Colesterol HDL*	*Colesterol*
1	1	1	0
2	0	1	1
3	0	0	0
4	1	1	1

Fuente: Elaboración propia

I ={Colesterol LDL, Colesterol HDL, Colesterol}

A la derecha se muestra una pequeña base de datos que contiene los items, donde el código '1' se interpreta como que el lípido en la sangre (item) correspondiente está

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

presente en la transacción y el código '0' significa que dicho lípido en la sangre no está presente.

Un ejemplo de regla podría ser:

Significaría que si el paciente tiene 'colesterol LDL' y 'colesterol HDL' también existe la posibilidad que tenga 'colesterol' elevado, es decir, según la especificación formal anterior se tendría que:

X ={Colesterol LDL, Colesterol HDL}

Y ={Colesterol}

Soporte y confianza

Una regla necesita un soporte de varios cientos de registros (transacciones) antes de que ésta pueda considerarse significativa desde un punto de vista estadístico. A menudo las bases de datos contienen miles o incluso millones de registros.

Para seleccionar reglas interesantes del conjunto de todas las reglas posibles que se pueden derivar de un conjunto de datos se pueden utilizar restricciones sobre diversas medidas de "**significancia**" e "**interés**". Las restricciones más conocidas son los umbrales mínimos de "**soporte**" y "**confianza**".

El 'soporte' de un conjunto de items X en una base de datos D se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de items:

Formula General de Soporte: $$\mathrm{sop}(X) = \frac{|X|}{|D|}$$

En el ejemplo anterior el conjunto X ={Colesterol LDL, Colesterol HDL}

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tiene un soporte de;

$$\text{sop}(X) = \frac{2}{5} = 0.4$$

Es decir, el soporte es del 40% (2 de cada 5 transacciones).

La 'confianza' de una regla se define como:

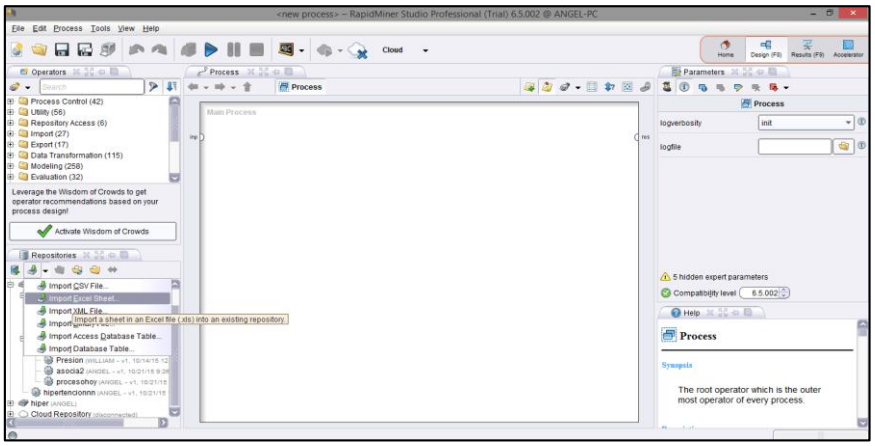
Fórmula General de Confianza

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sop}(X \cup Y)}{\text{sop}(X)} = \frac{|X \cup Y|}{|X|}$$

Técnica de árbol de decisión

Se empleó la herramienta RapidMiner que es de Open Source especial para predicciones con la técnica de árbol de decisión**.**

Gráfico 24: Entorno RapidMiner - árbol de decisión



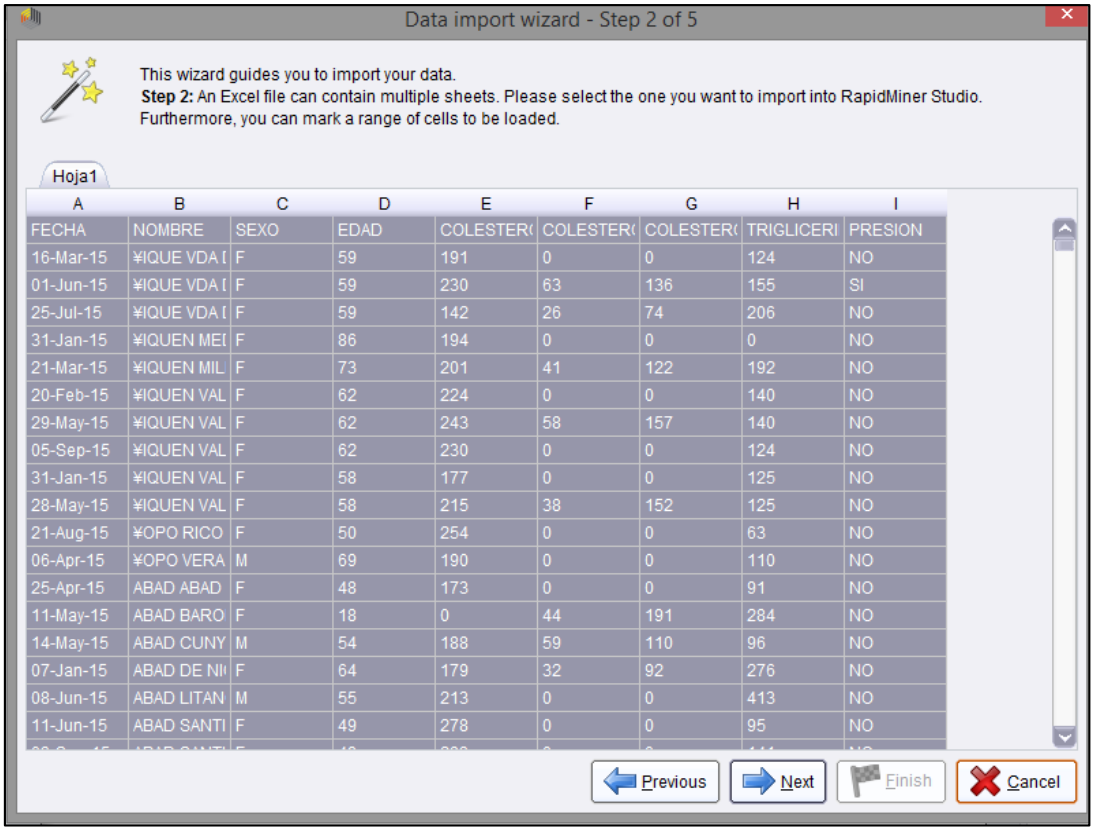
Fuente: RapidMiner

Aquí se cargan los datos para transformarlos y que el software los interprete en RapidMiner.

Obtenemos la importación de la BD de Microsoft Excel a RapidMiner para técnica árbol de decisión.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 25: Datos RapidMiner – árbol de decisión



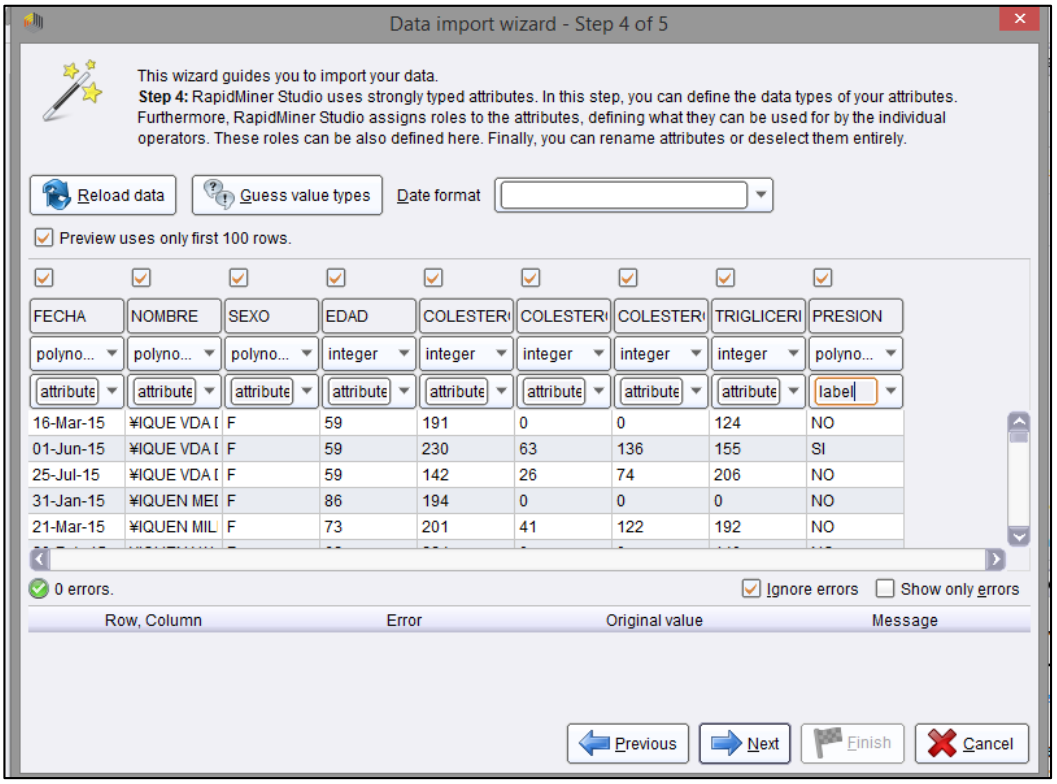
A	B	C	D	E	F	G	H	I
FECHA	NOMBRE	SEXO	EDAD	COLESTER(	COLESTER(	COLESTER(	TRIGLICERI	PRESION
16-Mar-15	¥IQUE VDA [	F	59	191	0	0	124	NO
01-Jun-15	¥IQUE VDA [	F	59	230	63	136	155	SI
25-Jul-15	¥IQUE VDA [	F	59	142	26	74	206	NO
31-Jan-15	¥IQUEN MEI	F	86	194	0	0	0	NO
21-Mar-15	¥IQUEN MIL	F	73	201	41	122	192	NO
20-Feb-15	¥IQUEN VAL	F	62	224	0	0	140	NO
29-May-15	¥IQUEN VAL	F	62	243	58	157	140	NO
05-Sep-15	¥IQUEN VAL	F	62	230	0	0	124	NO
31-Jan-15	¥IQUEN VAL	F	58	177	0	0	125	NO
28-May-15	¥IQUEN VAL	F	58	215	38	152	125	NO
21-Aug-15	¥OPO RICO	F	50	254	0	0	63	NO
06-Apr-15	¥OPO VERA	M	69	190	0	0	110	NO
25-Apr-15	ABAD ABAD	F	48	173	0	0	91	NO
11-May-15	ABAD BARO	F	18	0	44	191	284	NO
14-May-15	ABAD CUNY	M	54	188	59	110	96	NO
07-Jan-15	ABAD DE NI	F	64	179	32	92	276	NO
08-Jun-15	ABAD LITAN	M	55	213	0	0	413	NO
11-Jun-15	ABAD SANTI	F	49	278	0	0	95	NO

Fuente: RapidMiner

Se cambia en tipo de dato entero a texto para que la variable Presión sea dependiente, por consecuente depende de las variables independientes que son trigliceridos, colesterol, etc.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 26: Variable dependiente e independiente – árbol de decisión



Fuente: RapidMiner

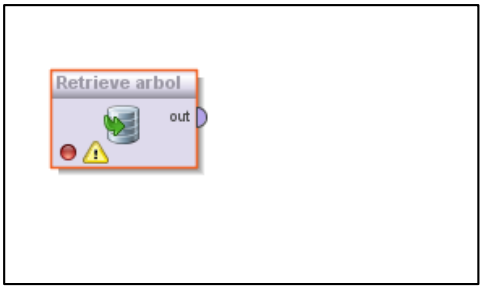
El item presión es la variable dependiente y las variables independientes son: colesterol, colesterol LDL, colesterol HDL y triglicéridos, son los indicadores a evaluar para la variable dependiente.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Modelado

Gráfico 27: Lectura de datos – árbol de decisión



Fuente: RapidMiner

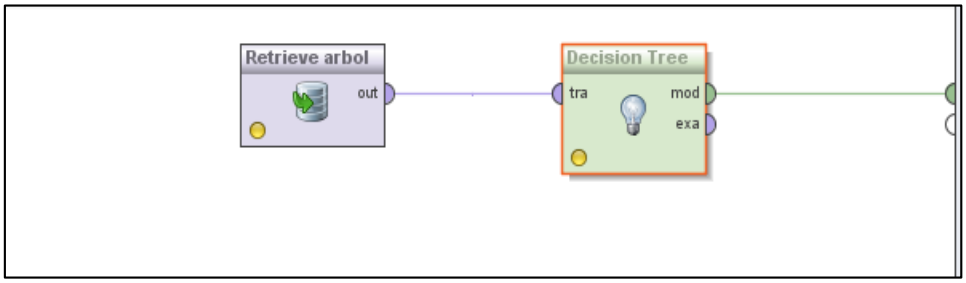
En este proceso se hará una lectura de los datos como se aprecia en la imagen.

En el RapidMiner se importa los datos de Microsoft Excel, para la técnica de árbol de decisión.

Entrenamiento

En este proceso se realiza el entrenamiento y validación de los datos, estos deben ser en un periodo de tiempo. En este proceso se añade el operador que permitirá medir el performance del modelo.

Gráfico 28: Entrenamiento y testeo de los datos – árbol de decisión



Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Retrive Árbol

Es la base de datos (Microsoft Excel) con todas las variables y así como los datos mismos.

Decisión Tree

Es la conexión de la base de datos con decisión tree (árbol de decisiónes).

De esta forma podemos tomar nuestras decisiónes basadas en la probabilidad de un árbol de decisiónes.

Técnica de árbol de decisión

Es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.). De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

Realiza esta labor mediante la construcción de un árbol de decisión.

Los elementos son:

- Nodos: Los cuales contendrán atributos.
- Arcos: Los cuales contienen valores posibles del nodo padre.
- Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Elección del Mejor Atributo

La elección del mejor atributo se establece mediante la entropía. Eligiendo aquel que proporcione una mejor ganancia de información. La función elegida puede variar, pero en su forma más sencilla es como esta:

$$-\left(\frac{|p|}{|d|}\right)\log_2\left(\frac{|p|}{|d|}\right)-\left(\frac{|n|}{|d|}\right)\log_2\left(\frac{|n|}{|d|}\right)$$

Donde *p* es el conjunto de los ejemplos positivos, *n* el de los negativos y *d* el total de ellos. Se debe establecer si el logaritmo es positivo o negativo.

5.5.4 Fase IV: Minería de datos

Donde se obtienen los patrones de interés según la tarea de minería que llevemos a cabo (descriptiva o predictiva).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Reglas de asociación

Gráfico 29: Obtención de patrones para realizar la predicción con reglas de asociación

Row No.	ITEM	COLESTER...	COLESTER...	COLESTER...	TRIGLICER...	FECHA	NOMBRE
1	1	false	false	false	false	16-mar-15	¥IQUE VDA [
2	2	true	true	true	true	01-jun-15	¥IQUE VDA [
3	3	false	false	false	true	25-jul-15	¥IQUE VDA [
4	4	false	false	false	false	31-Jan-15	¥IQUEN MEI
5	5	true	true	false	true	21-mar-15	¥IQUEN MIL
6	6	true	false	false	false	20-feb-15	¥IQUEN VAL
7	7	true	true	true	false	29-may-15	¥IQUEN VAL
8	8	true	false	false	false	05-sep-15	¥IQUEN VAL
9	9	false	false	false	false	31-Jan-15	¥IQUEN VAL
10	10	true	true	true	false	28-may-15	¥IQUEN VAL
11	11	true	false	false	false	21-Aug-15	¥OPO RICO
12	12	false	false	false	false	06-Apr-15	¥OPO VERA
13	13	false	false	false	false	25-Apr-15	ABAD ABAD
14	14	false	true	true	true	11-may-15	ABAD BARO
15	15	false	true	false	false	14-may-15	ABAD CUNY
16	16	false	false	false	true	07-Jan-15	ABAD DE NI
17	17	true	false	false	true	08-jun-15	ABAD LITAN
18	18	true	false	false	false	11-jun-15	ABAD SANTI
19	19	true	false	false	false	03-sep-15	ABAD SANTI
20	20	false	true	false	false	03-mar-15	ABANTO CE
21	21	false	false	false	true	02-Jan-15	ABANTO CH
22	22	false	true	false	false	09-Jan-15	ABANTO CU
23	23	true	true	true	false	20-Aug-15	ABANTO DE
24	24	true	true	true	true	02-may-15	ABANTO LAI
25	25	true	false	false	true	15-may-15	ABANTO LAI
26	26	true	true	false	false	11-may-15	ABANTO VEI
27	27	false	true	false	true	18-Aug-15	ABANTO ZEI
28	28	false	false	false	false	29-Apr-15	ACARO MIRA
29	29	false	true	false	false	02-may-15	ACHA ALBEI
30	30	false	false	false	false	29-Apr-15	ACHA CORT
31	31	true	true	false	true	10-Aug-15	ACHA CORT
32	32	true	true	true	true	27-feb-15	ACHA TOGA
33	33	false	true	false	true	25-mar-15	ACOSTA CA:
34	34	true	false	false	false	08-jun-15	ACOSTA CO

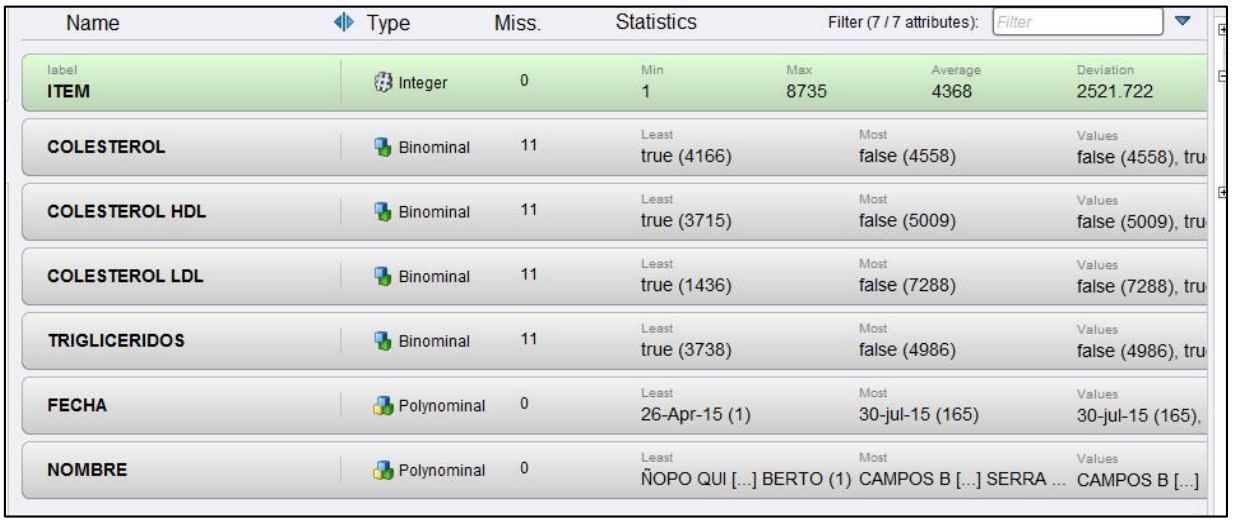
Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Entonces ya empieza a interpretar la información.

Gráfico 30: Interpretación de la información – reglas de asosiación



Name	Type	Miss.	Statistics			
label ITEM	Integer	0	Min 1	Max 8735	Average 4368	Deviation 2521.722
COLESTEROL	Binominal	11	Least true (4166)	Most false (4558)	Values false (4558), tru	
COLESTEROL HDL	Binominal	11	Least true (3715)	Most false (5009)	Values false (5009), tru	
COLESTEROL LDL	Binominal	11	Least true (1436)	Most false (7288)	Values false (7288), tru	
TRIGLICERIDOS	Binominal	11	Least true (3738)	Most false (4986)	Values false (4986), tru	
FECHA	Polynominal	0	Least 26-Apr-15 (1)	Most 30-jul-15 (165)	Values 30-jul-15 (165),	
NOMBRE	Polynominal	0	Least ÑOPO QUI [...] BERTO (1)	Most CAMPOS B [...] SERRA ...	Values CAMPOS B [...]	

Fuente: RapidMiner

Árbol de decisión

Gráfico 31: Obtención de patrones para realizar la predicción en tree

Row No.	PRESION	FECHA	NOMBRE	SEXO	EDAD	COLESTER...	COLESTER...	COLESTER...	TRIGLICER...
1	NO	16-Mar-15	¥IQUE VDA [	F	59	191	0	0	124
2	SI	01-Jun-15	¥IQUE VDA [	F	59	230	63	136	155
3	NO	25-Jul-15	¥IQUE VDA [	F	59	142	26	74	206
4	NO	31-Jan-15	¥IQUEN MEI	F	86	194	0	0	0
5	NO	21-Mar-15	¥IQUEN MIL	F	73	201	41	122	192
6	NO	20-Feb-15	¥IQUEN VAL	F	62	224	0	0	140
7	NO	29-May-15	¥IQUEN VAL	F	62	243	58	157	140
8	NO	05-Sep-15	¥IQUEN VAL	F	62	230	0	0	124
9	NO	31-Jan-15	¥IQUEN VAL	F	58	177	0	0	125
10	NO	28-May-15	¥IQUEN VAL	F	58	215	38	152	125
11	NO	21-Aug-15	¥OPO RICO	F	50	254	0	0	63
12	NO	06-Apr-15	¥OPO VERA	M	69	190	0	0	110
13	NO	25-Apr-15	ABAD ABAD	F	48	173	0	0	91
14	NO	11-May-15	ABAD BARO	F	18	0	44	191	284
15	NO	14-May-15	ABAD CUNY	M	54	188	59	110	96
16	NO	07-Jan-15	ABAD DE NI	F	64	179	32	92	276
17	NO	08-Jun-15	ABAD LITAN	M	55	213	0	0	413
18	NO	11-Jun-15	ABAD SANTI	F	49	278	0	0	95
19	NO	03-Sep-15	ABAD SANTI	F	49	283	0	0	141
20	NO	03-Mar-15	ABANTO CE	M	83	160	56	93	56
21	NO	02-Jan-15	ABANTO CH	M	60	138	33	65	198

Fuente: RapidMiner

Se obtiene los datos de patrones para realizar pre diagnóstico con la técnica de árbol de decisión.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

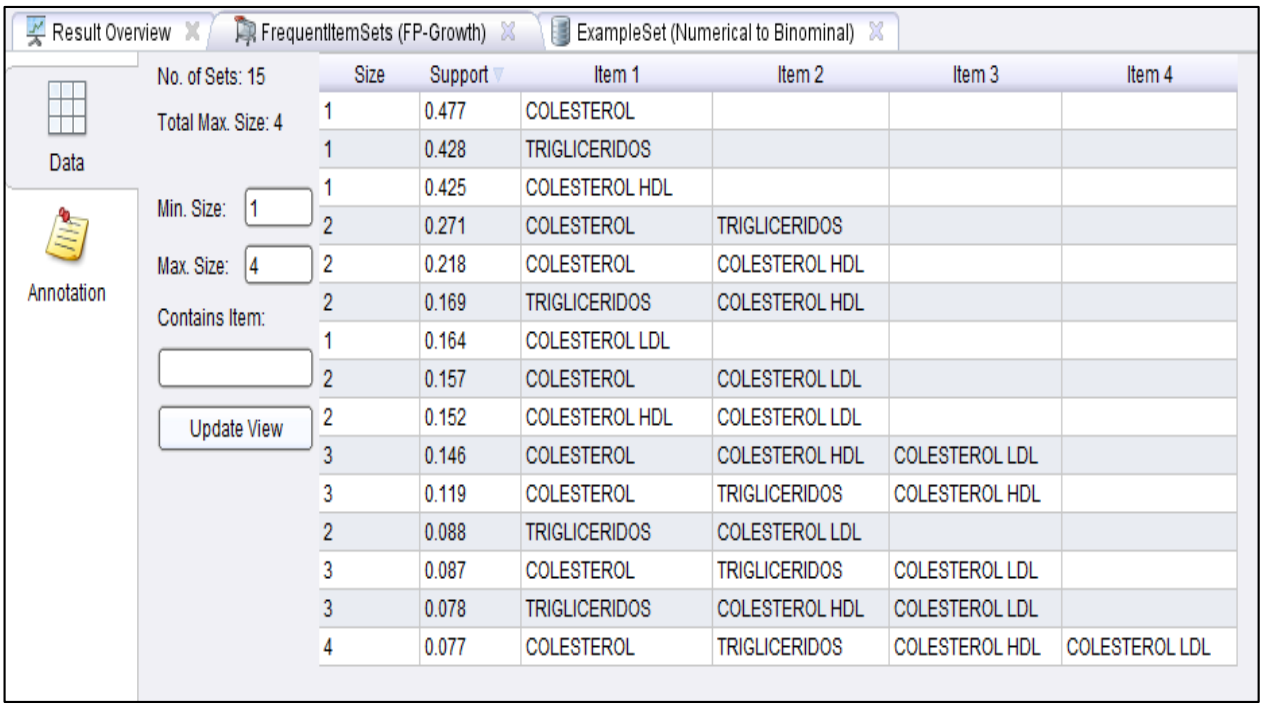
5.5.5 Fase V: Interpretación y Evaluación

Es la interpretación y evaluación del nuevo conocimiento en el dominio de la aplicación. Asociaremos las variables independientes para interpretar los datos y las predicciones. Para esto debemos tener el componente WEKA de RapidMiner y ejecutamos el programa para obtener resultados.

Reglas de Asociación

Esto nos quiere decir que el 47% tiene colesterol alto, el 43% tiene Triglicéridos, el 27% tiene colesterol alto y triglicéridos, etc.

Gráfico 32: Interpretación de los datos con reglas de asociación



Size	Support	Item 1	Item 2	Item 3	Item 4
1	0.477	COLESTEROL			
1	0.428	TRIGLICERIDOS			
1	0.425	COLESTEROL HDL			
2	0.271	COLESTEROL	TRIGLICERIDOS		
2	0.218	COLESTEROL	COLESTEROL HDL		
2	0.169	TRIGLICERIDOS	COLESTEROL HDL		
1	0.164	COLESTEROL LDL			
2	0.157	COLESTEROL	COLESTEROL LDL		
2	0.152	COLESTEROL HDL	COLESTEROL LDL		
3	0.146	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	
3	0.119	COLESTEROL	TRIGLICERIDOS	COLESTEROL HDL	
2	0.088	TRIGLICERIDOS	COLESTEROL LDL		
3	0.087	COLESTEROL	TRIGLICERIDOS	COLESTEROL LDL	
3	0.078	TRIGLICERIDOS	COLESTEROL HDL	COLESTEROL LDL	
4	0.077	COLESTEROL	TRIGLICERIDOS	COLESTEROL HDL	COLESTEROL LDL

Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 33: Obtención de resultados para el pre diagnóstico con reglas de asociación

No.	Premises	Conclusion	Support	Confidence	LaPla...	Gain	p-s	Lift	Convi...
1	TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL, COLESTEROL HDL	0.077	0.876	0.990	-0.099	0.058	4.017	6.289
2	TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL HDL	0.078	0.885	0.991	-0.099	0.041	2.080	4.985
3	COLESTEROL LDL	COLESTEROL, COLESTEROL HDL	0.146	0.887	0.984	-0.183	0.110	4.070	6.932
4	COLESTEROL, TRIGLICERIDOS, COLESTI	COLESTEROL HDL	0.077	0.888	0.991	-0.097	0.040	2.089	5.145
5	COLESTEROL LDL	COLESTEROL HDL	0.152	0.922	0.989	-0.177	0.082	2.168	7.368
6	COLESTEROL, COLESTEROL LDL	COLESTEROL HDL	0.146	0.930	0.990	-0.168	0.079	2.187	8.201
7	COLESTEROL LDL	COLESTEROL	0.157	0.954	0.994	-0.172	0.078	2.000	11.38
8	COLESTEROL HDL, COLESTEROL LDL	COLESTEROL	0.146	0.962	0.995	-0.157	0.074	2.018	13.85
9	TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL	0.087	0.986	0.999	-0.090	0.045	2.067	36.71
10	TRIGLICERIDOS, COLESTEROL HDL, COL	COLESTEROL	0.077	0.990	0.999	-0.079	0.040	2.075	51.03

Fuente: RapidMiner

Luego creamos las Reglas de Asociación para obtener Conclusiones. Toda persona que tiene Triglicéridos y Colesterol LDL también tiene Colesterol elevado en un nivel de confianza de 98.6 % y que en primera vista el 0.986 de los pacientes que representa los 760, tendrá Hipertensión Arterial, ya que tienen Triglicéridos y Colesterol LDL alto.

Gráfico 34: Resultados de las reglas de asociación

AssociationRules

```
Association Rules
[TRIGLICERIDOS, COLESTEROL LDL] --> [COLESTEROL, COLESTEROL HDL] (confidence: 0.876)
[TRIGLICERIDOS, COLESTEROL LDL] --> [COLESTEROL HDL] (confidence: 0.885)
[COLESTEROL LDL] --> [COLESTEROL, COLESTEROL HDL] (confidence: 0.887)
[COLESTEROL, TRIGLICERIDOS, COLESTEROL LDL] --> [COLESTEROL HDL] (confidence: 0.888)
[COLESTEROL LDL] --> [COLESTEROL HDL] (confidence: 0.922)
[COLESTEROL, COLESTEROL LDL] --> [COLESTEROL HDL] (confidence: 0.930)
[COLESTEROL LDL] --> [COLESTEROL] (confidence: 0.954)
[COLESTEROL HDL, COLESTEROL LDL] --> [COLESTEROL] (confidence: 0.962)
[TRIGLICERIDOS, COLESTEROL LDL] --> [COLESTEROL] (confidence: 0.986)
[TRIGLICERIDOS, COLESTEROL HDL, COLESTEROL LDL] --> [COLESTEROL] (confidence: 0.990)
```

Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Técnica de árbol de decisión

Con los 8,735 registros que va a evaluar la técnica de árbol de decisión obtenemos los resultados en la columna presión como indica en la figura SI o NO, son los pacientes que pueden adquirir la enfermedad de HA, sirve un pre diagnóstico ya que por sus colesterol elevado y triglicéridos elevados son indicadores de poder adquirir esta enfermedad, estos datos y rangos de los ítem (colesterol LDL, colesterol HDL y trigleceridos), son desarrollados en el ítem presión.

Gráfico 35: Interpretación de los datos con árbol de decisión

ExampleSet (8735 examples, 1 special attribute, 8 regular attributes) Filter (8,735 / 8,735

Row No.	PRESION	FECHA	NOMBRE	SEXO	EDAD	COLESTER...	COLESTER...	COLESTER...	TRIGLICER...
1	NO	16-Mar-15	¥IQUE VDA [	F	59	191	0	0	124
2	SI	01-Jun-15	¥IQUE VDA [	F	59	230	63	136	155
3	NO	25-Jul-15	¥IQUE VDA [	F	59	142	26	74	206
4	NO	31-Jan-15	¥IQUEN MEI	F	86	194	0	0	0
5	NO	21-Mar-15	¥IQUEN MIL	F	73	201	41	122	192
6	NO	20-Feb-15	¥IQUEN VAL	F	62	224	0	0	140
7	NO	29-May-15	¥IQUEN VAL	F	62	243	58	157	140
8	NO	05-Sep-15	¥IQUEN VAL	F	62	230	0	0	124
9	NO	31-Jan-15	¥IQUEN VAL	F	58	177	0	0	125
10	NO	28-May-15	¥IQUEN VAL	F	58	215	38	152	125
11	NO	21-Aug-15	¥OPO RICO	F	50	254	0	0	63
12	NO	06-Apr-15	¥OPO VERA	M	69	190	0	0	110
13	NO	25-Apr-15	ABAD ABAD	F	48	173	0	0	91
14	NO	11-May-15	ABAD BARO	F	18	0	44	191	284
15	NO	14-May-15	ABAD CUNY	M	54	188	59	110	96
16	NO	07-Jan-15	ABAD DE NI	F	64	179	32	92	276
17	NO	08-Jun-15	ABAD LITAN	M	55	213	0	0	413
18	NO	11-Jun-15	ABAD SANTI	F	49	278	0	0	95
19	NO	03-Sep-15	ABAD SANTI	F	49	283	0	0	141
20	NO	03-Mar-15	ABANTO CE	M	83	160	56	93	56
21	NO	02-Jan-15	ABANTO CH	M	60	138	33	65	198

Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Resultados del tree (árbol de decisión)

Según la técnica del tree (árbol de decisión) empleada en RapidMiner obtenemos el siguiente resultado: Si colesterol ldl > a 129.50 y triglicéridos > 149.5 y colesterol total > =199 entonces 749 pacientes podrían sufrir de hipertensión arterial.

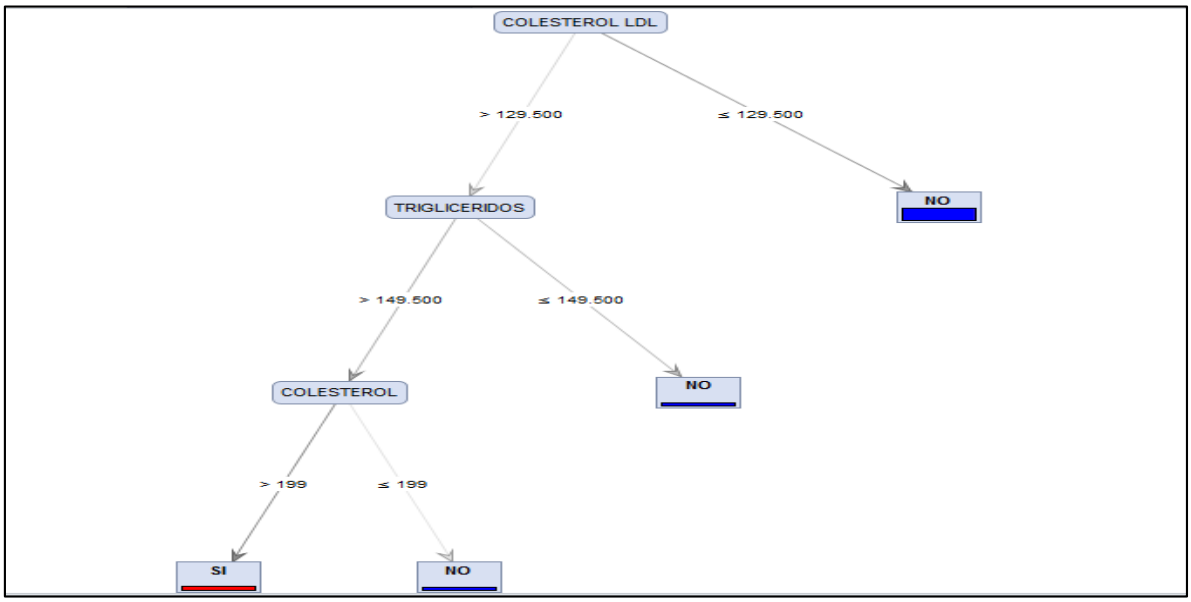
Gráfico 36: Resultados con la técnica árbol de decisión

Tree

```
COLESTEROL LDL > 129.500
|   TRIGLICERIDOS > 149.500
|   |   COLESTEROL > 199: SI {NO=0, SI=749}
|   |   COLESTEROL ≤ 199: NO {NO=10, SI=0}
|   TRIGLICERIDOS ≤ 149.500: NO {NO=625, SI=0}
COLESTEROL LDL ≤ 129.500: NO {NO=7338, SI=13}
```

Fuente: RapidMiner

Gráfico 37: Árbol de decisión



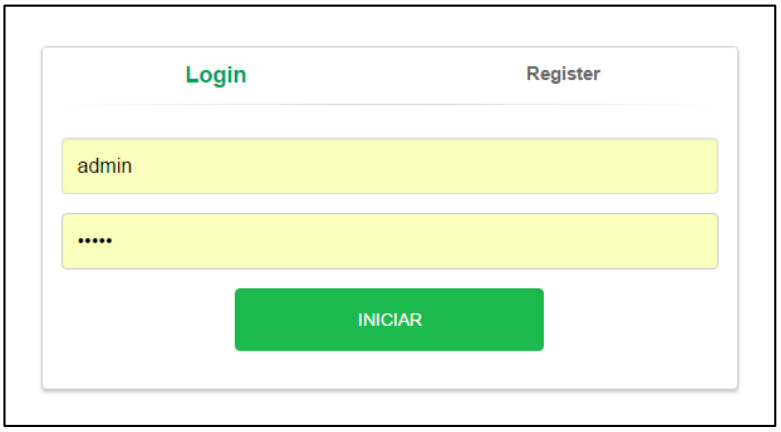
Fuente: RapidMiner

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Lenguaje JAVA – Técnicas de minería de datos

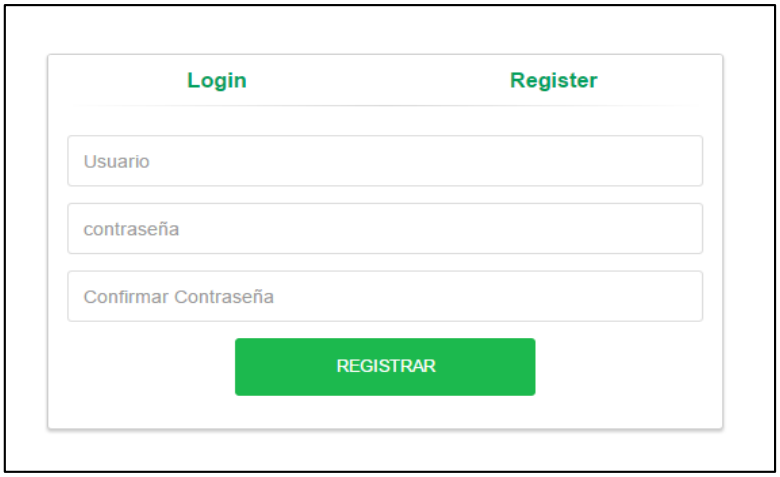
Gráfico 38: Pantalla de acceso al sistema



Fuente: Elaboración propia

Esta Pantalla es la del acceso al sistema, donde se validará su usuario y contraseña, para ingresar a visualizar el pre diagnóstico con Minería de Datos.

Gráfico 39: Pantalla de registro de usuario



Fuente: Elaboración propia

Esta pantalla se visualiza los campos para crear un nuevo usuario, luego se procede a registrar para poder accesar al sistema.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 40: Pantalla de descripción de reglas de asociación – pacientes sin HA

Fuente: Elaboración propia

Esta pantalla muestra una descripción de la técnica de minería de datos, en este caso las reglas de asociación en base al estudio que se realizó con la base de datos proporcionada por el hospital de ESSALUD período 2015, de 8,735 registros

Se concluyó que 760 pacientes padecen de la enfermedad de hipertensión arterial con un nivel de confianza de 98.6 % y 7,975 pacientes no padecen esta enfermedad.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 41: Pantalla de descripción de árbol de decisión – pacientes sin HA

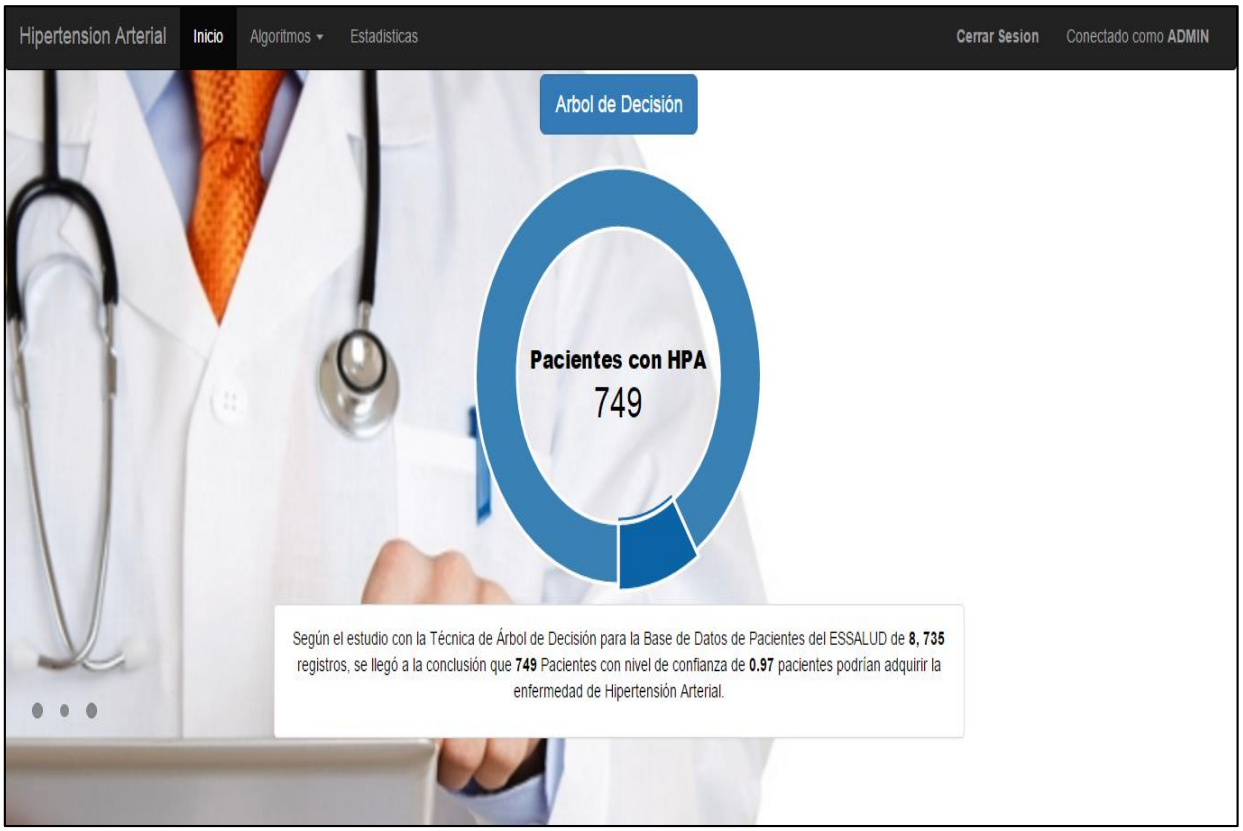
Fuente: Elaboración propia

Esta pantalla muestra una descripción de la técnica de minería de datos, en este caso con el árbol de decisión en base al estudio que se realizó con la base de datos proporcionada por el hospital de ESSALUD período 2015, de 8,735 registros

Se concluyo que 749 pacientes padecen de la enfermedad de hipertensión arterial con un nivel de confianza de 97 % y 7,986 pacientes no padecen esta enfermedad.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 42: Pantalla de descripción de árbol de decisión – pacientes con HA



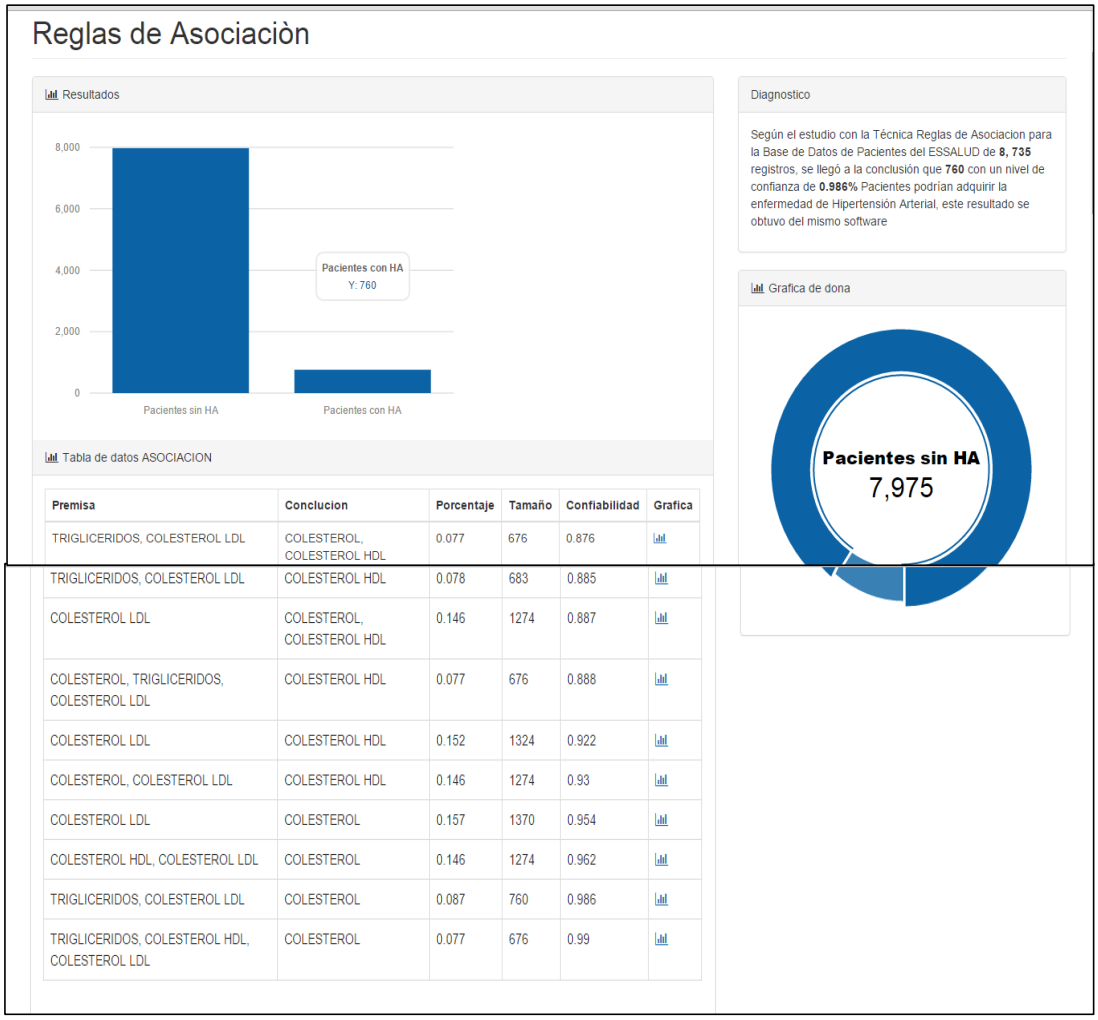
Fuente: Elaboración propia

Esta pantalla muestra una descripción de la técnica de minería de datos, en este caso con árbol de decisión en base al estudio que se realizo con la base de datos proporcionada por el hospital de ESSALUD período 2015, de 8,735 registros

Se concluyó que 749 pacientes padecen de la enfermedad de hipertensión arterial con un nivel de confianza de 97%.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 43: Pantalla de técnica de reglas de asociación



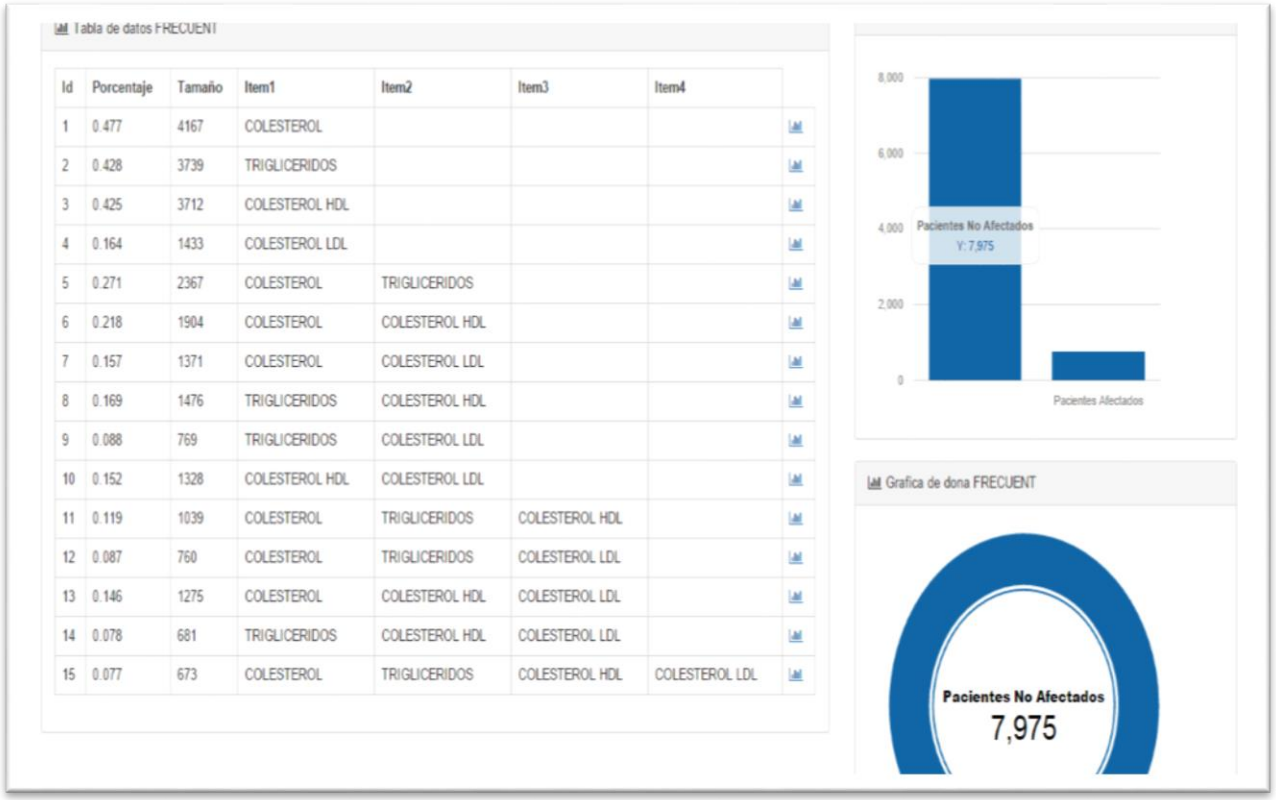
Premisa	Conclucion	Porcentaje	Tamaño	Confiabilidad	Grafica
TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL, COLESTEROL HDL	0.077	676	0.876	
TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL HDL	0.078	683	0.885	
COLESTEROL LDL	COLESTEROL, COLESTEROL HDL	0.146	1274	0.887	
COLESTEROL, TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL HDL	0.077	676	0.888	
COLESTEROL LDL	COLESTEROL HDL	0.152	1324	0.922	
COLESTEROL, COLESTEROL LDL	COLESTEROL HDL	0.146	1274	0.93	
COLESTEROL LDL	COLESTEROL	0.157	1370	0.954	
COLESTEROL HDL, COLESTEROL LDL	COLESTEROL	0.146	1274	0.962	
TRIGLICERIDOS, COLESTEROL LDL	COLESTEROL	0.087	760	0.986	
TRIGLICERIDOS, COLESTEROL HDL, COLESTEROL LDL	COLESTEROL	0.077	676	0.99	

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 44: Pantalla de tabla de frecuencia de reglas de asociación



Fuente: Elaboración propia

En esta pantalla nos muestra una tabla de frecuencia de la técnica de reglas de asociación donde se puede visualizar el tamaño de la población de pacientes que tienen lípidos en la sangre como son: colesterol ldl, colesterol y triglicéridos, en cuanto al colesterol, se observa en la primera fila un tamaño de 4167 pacientes que padecen de colesterol elevado, así podemos seguir visualizando el resto de ítems para cada caso.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 45: Comparación de técnicas de minería de datos

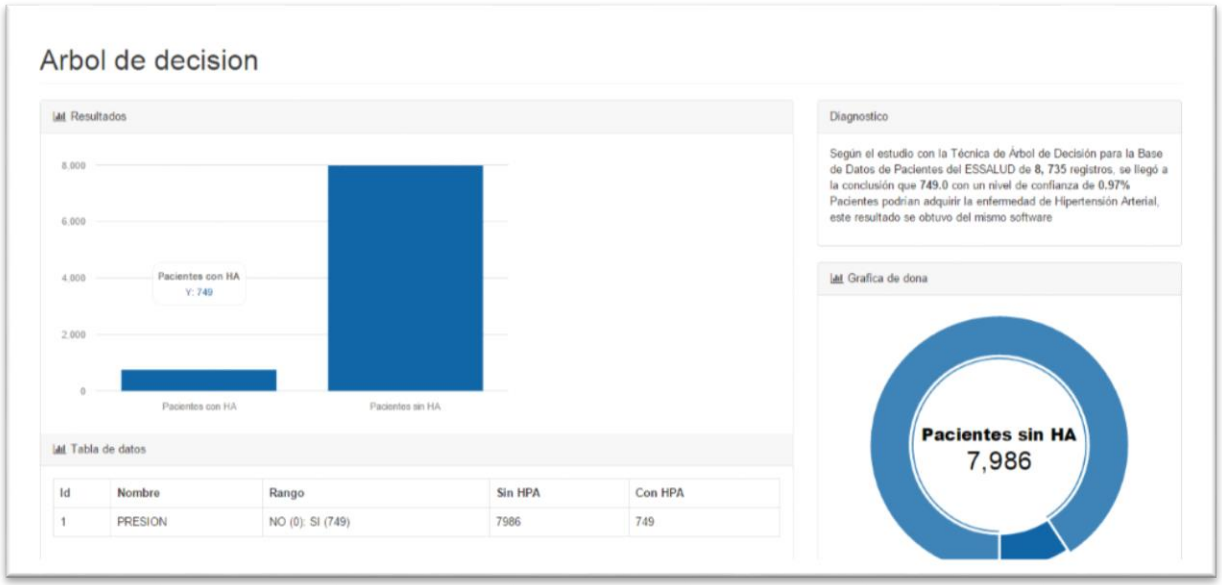
Fuente: Elaboración propia

En esta pantalla se visualiza la comparación de la técnica de reglas de asociación con árbol de decisión para el pre diagnóstico de enfermedad de hipertensión arterial

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 46: Árbol de decisión

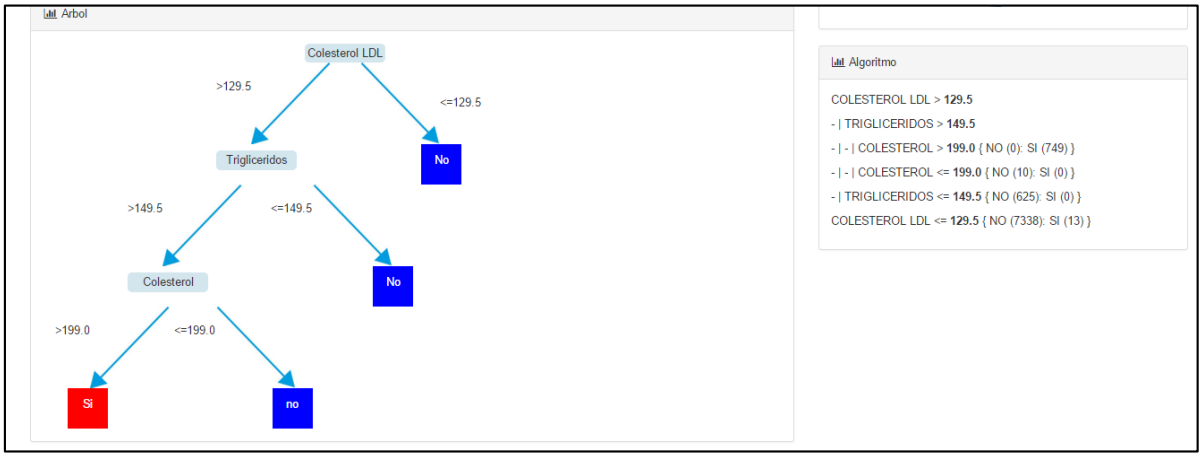


Fuente: Elaboración propia

En esta pantalla nos muestra los datos de la técnica árbol de decisión realizada con la ayuda de la herramienta rapidminer para mostrar los resultados de esta técnica de minería de datos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 47: Gráfico de árbol de decición



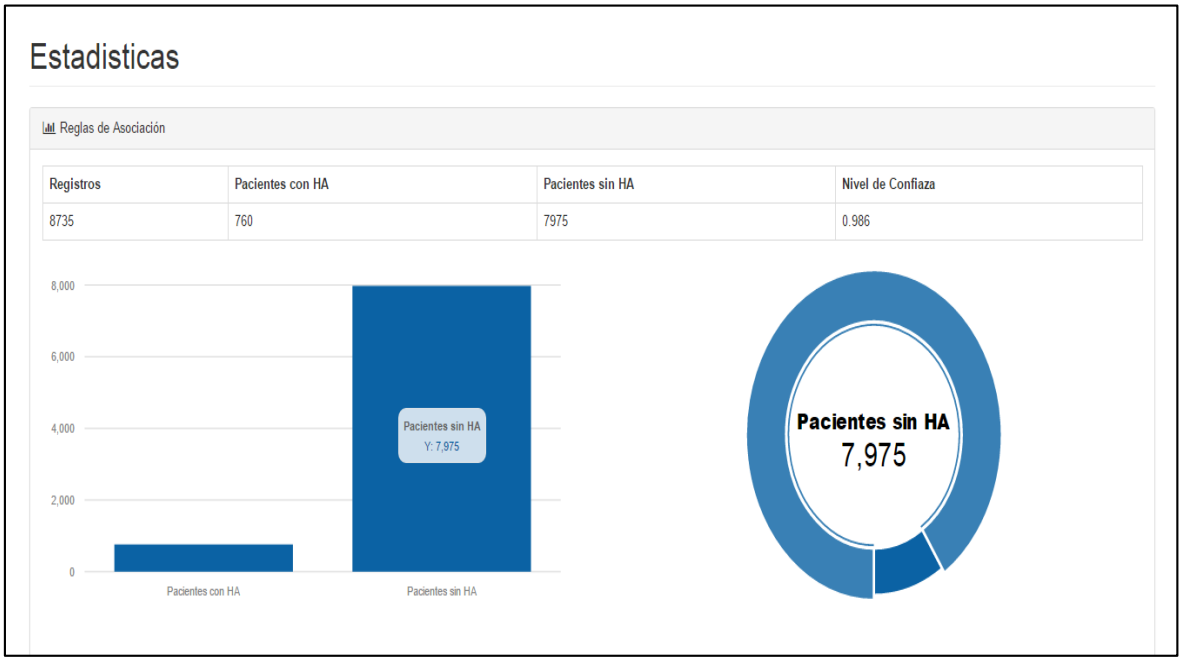
Fuente: Elaboración propia

En esta pantalla se visualiza el gráfico del tree (árbol) con su respectiva leyenda, se muestran los resultados de esta técnica.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 48: Estadisticas de la técnica de reglas de asociación



Fuente: Elaboración propia

En esta pantalla nos muestra los datos de la técnica de reglas de asociación realizada con la ayuda de la herramienta rapidminer para mostrar los resultados de esta técnica de minería de datos.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

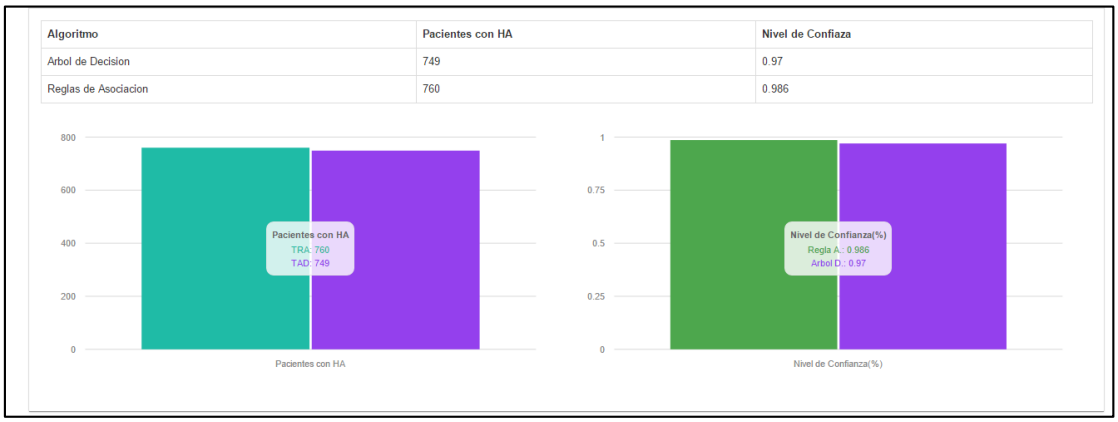
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 49: Estadisticas de la técnica de árbol de decisión

Fuente: Elaboración propia

En esta pantalla nos muestra los datos de la técnica de árbol de decisión realizada con la ayuda de la herramienta rapidminer para mostrar los resultados de esta técnica de minería de datos.

Gráfico 50: Comparación de técnica de reglas de asociación y técnica de árbol de decisión



Fuente: Elaboración propia

En esta pantalla se visualiza la comparación de la técnica de reglas de asociación con árbol de decisión de manera gráfica para el pre diagnóstico de enfermedad de hipertensión arterial.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

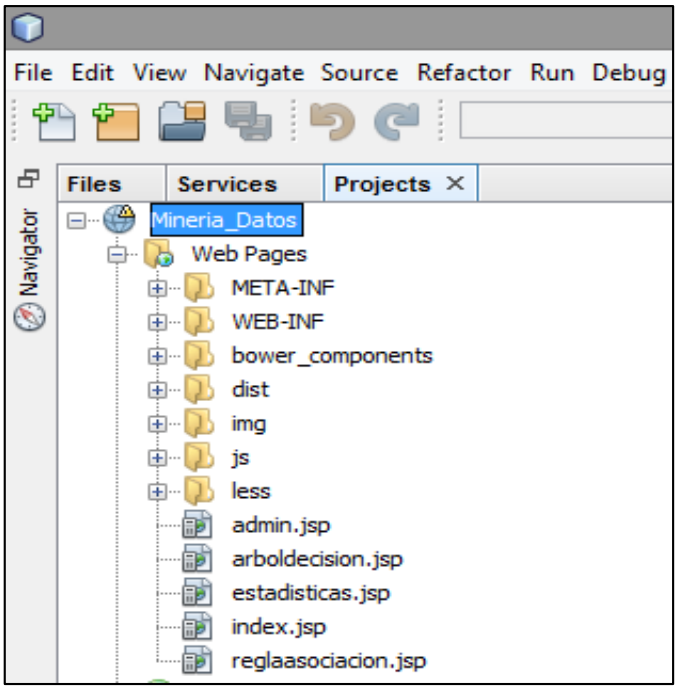
5.6 Publicación web de la investigación

El proyecto consta de 2 parte una la parte web y otra donde está todo el java. En el sitio web esta lo que se necesita para administrar la misma tenemos: las imágenes, java script, el compilador lss es un compilador css, bower_components que son librerías para los gráficos tenemos bootstrap social que son para los iconos y las librerías morrisjs para los gráficos.

A través de la parte web llamamos los resultados y los mostramos.

Ahora a través de la carpeta controlador se conecta el java con la parte web. Y el controladorUsuario.java se enlaza con las demas clases que son: DowUsuario.java., DowReglasAsociación.java, DowArbolDecisión.java, DowMigracion.java. Ver código y su documentación en el Anexo 05

Gráfico 51: Pantalla del proyecto netbeans



Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Pruebas de Validaciones

Para realizar pruebas se recurrió a una hoja de Excel, de este modo se hizo las comparaciones para medir la eficiencia de la técnica de minería de datos.

En primer lugar se tiene los 8,735, se realizó una limpieza de información

Para ganar resultados de conveniencia, para ello se utilizó las fórmulas para cada rango de colesterol HDL, colesterol LDL y triglicéridos.

Ítems	***Fórmula***
Colesterol	SI(EDAD<=19,SI(COLESTEROL<=170,0,1),SI(EDAD >19,SI(COLESTEROL<200,0,1)))
Triglicéridos	SI(TRIGLICERIDOS<150,0,1)
Colesterol HDL	SI(COLESTEROL HDL<35,0,1)
Colesterol LDL	SI(EDAD <=19,SI(COLESTEROL LDL<=100,0,1),SI(EDAD >19,SI(COLESTEROL LDL<130,0,1)))

Cuando se realizó este procedimiento, podemos observar los 0 y 1, es donde se aplicó las formulas antes mencionadas.

ORDEN	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	Fórmula Prueba
1	1	0	1	1	1
2	1	1	1	1	1
3	0	0	0	1	0

0 = indica que el paciente no presenta colesterol, colesterol ldl, colesterol hdl y triglicéridos

1= indica que el paciente presenta colesterol, colesterol ldl, colesterol hdl y triglicéridos

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Fórmula _ Prueba:

SI (COLESTEROL+COLESTERO LDL+TRIGLICERIDOS=3, 1,0)

A	B	C	D	E	F
ORDEN	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	Fórmula_Prueba
1	1	0	1	1	1
2	1	1	1	1	1
3	0	0	0	1	0

Como se puede observar las columnas de las celdas A es de orden (volumen de pacientes, 1 - 8735 registros) B es colesterol, C es colesterol HDL, D colesterol LDL, E triglicéridos, F Verificación Prueba.

En la celda F se aplicó la fórmula de prueba para obtener una aproximación de probabilidades del diagnóstico de Hipertensión Arterial.

A	B	C	D	E	F
ORDEN	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	Fórmula_Prueba
1	1	0	1	1	1
2	1	1	1	1	1
3	0	0	0	1	0
					=CONTAR.SI(763)

Se aplicó una función Contar. Si a los resultados de la celda F (Fórmula_Prueba). Donde se obtuvieron 763 Pacientes podrían adquirir hipertensión arterial, esto se elaboró en una hoja de cálculo de Excel.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En la hoja de cálculo de Excel obtuvimos 763 pacientes que pueden adquirir hipertensión arterial, el modelo que se adecua es el de Reglas de Asociación con 760 pacientes una proximidad muy cerca al caso de estudio de predicción del diagnostico

Pacientes con HA	*Nivel de confianza*	*Técnica de minería*
760	98.60%	Reglas de asociación
749	97%	Árbol de decisión

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO VI

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

- Se recopiló información histórica acerca de los pacientes del hospital, brindo información de un gran volumen de datos de 8,735 registros para el caso de estudio. Ver Gráfico 14.
- Se concluye que las técnicas o algoritmos eficientes para los requerimientos en el estudio de predicción de diagnóstico, se seleccionaron árbol de decisiónes y reglas de asociación. ver tabla 8.1 pág. 72. y ver anexo 03.

- Se concluye que las variables cuantificables para ambos casos son: colesterol, colesterol HDL, colesterol LDL y triglicéridos. para el caso de árbol de decisión El ítem presión es la variable dependiente y las variables independientes son: colesterol, colesterol LDL y triglicéridos. y para el caso de reglas de asociación seria si un paciente tiene triglicéridos y colesterol LDL (son variables independiente) elevado se llega a la conclusión de que podría adquirir un colesterol elevado (variable dependiente). Estas variables son las que se midieron para el estudio de predicción de diagnóstico de H.A.

- Se expone que se ha utilizado una aplicación con soporte en HTML y JAVA WEB, y librerías de Rapidminer.jar para mostrar los resultados obtenidos de Rapidminer (Herramienta de Minería de Datos). ver el ítem 5.6 Publicación web de la investigación.
- Se realizaron las pruebas con una hoja de cálculo de Excel y se procedió a realizar comparación con las técnicas de minería de datos (Reglas de Asociación y Árbol de Decisión) ver anexo 03.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

- Se evaluaron los resultados y se cumplieron las premisas en los rangos establecidos y se logró realizar el diagnóstico proyectado, dando como resultados que 760 pacientes podrían adquirir la enfermedad de hipertensión arterial con un nivel de confianza de 98.6% con la técnica de reglas de asociación y 749 pacientes con un nivel de confianza de 97% con la técnica de árbol de decisión. Ver Gráfico 36.

- Evaluando los resultados se obtiene que la técnica de reglas de asociación es la adecuada para el análisis y soporte de predicción del diagnóstico de hipertensión arterial de los pacientes del Hospital “Almanzor Aguinaga Asenjo – Chiclayo” con el resultado de 760 pacientes del volumen de datos de 8735 datos con un nivel de confianza de 98.6% Ver Gráfico 33.

6.2 Recomendaciones

- Se recomienda ampliar la dase de datos ya que en el Hospital Almanzor Aguinaga Asenjo solo pudieron entregarme la informacion del año 2015 que cuenta con 8735 regitros ya que de los años anteriores no tenian esa informacion porque cada año limpian su base de datos.

- Se recomienda que para utilizar la técnica reglas de asociación se debe realizar la limpieza de los datos comvirtiendolos en ceros y unos para el mejor funcionamiento de dicha técnica.

- Se recomienda utilizar como minimo 8735 registros a más. Ya que conlleva en la incidencia de su nivel de confiabilidad.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

- Debido a que no se cuenta con técnicas para el diagnóstico de hipertensión arterial se sugiere emplear estas técnicas de minería de datos, por su grado de confiabilidad demostrada en la presente investigación.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

BIBLIOGRAFÍA

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

BIBLIOGRAFIA

Zambrano Alarcón. (2011). Data Mart. Análisis, Diseño e Implementación en Data Mart. Tesis PUPC. Lima. Perú.

Algoritmo de Clústeres. (2010). El Algoritmo de Clustering de Microsoft. Obtenido de https://i-msdn.sec.s msft.com/dynimg/IC35369. Revista.

Cabrera Hernández, L., Morales Hernández, A., Casas Cardoso, G., Denoda Pérez, L., Gonzáles Rodríguez, E., & Alfonso Rodríguez, J. (2010). Algoritmos Genéticos con Medidas de Diversidad Para el Diagnostico del Riesgo de HTA en Escolares. Tesis. Cuba.

Chan, d. M. (16 de Mayo de 2012). Directora General de OMS, Organización Mundial de la Salud. Obtenido de www.paho.org/arg. Revista.

Chávez Cárdenas, D., & Denoda Pérez, L. (2012). Casos de Niños en Edad Pediátrica de Modo que se Pueda Contar con un Modelo Capaz de Inferir el Riesgo de HTA. Tesis. Cuba.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Cuadrado Rodríguez, S., González Rodríguez, E. F., Curbelo Hernández, H., Luna Carvajal, Y., Casas Cardoso, G., & Gutiérrez Martínez, I. (2012). Sistema experto basado en casos para el diagnóstico de la hipertensión arterial. Revista Facultad de Ingeniería Universidad de Antioquia, (60).

Dandretta, G. H. (2002). Web mining: implementando técnicas de data mining en un servidor web. Universidad de Belgrano. Buenos Aires. Argentina. Revista.

Díaz Pérez, A. (2012). *Aplicación de la Red de Probabilidad Neuronal y Escala de Framingham para Predicción de la Hipertensión Arterial. Tesis. Cuba.*

Dr. MSc. Guillermo Alberto Pérez Fernández, D. C. (2012). Revista Cubana de Informática Médica. My SciELO.

GROUP, M. (Mayo de 2010). Estudio sobre proyectos de Data Warehouse. Revista. Chile.

INEI. (2013). Encuesta Demográfica y Salud Familiar. Obtenido de *www.inei.gob.pe. Instituto Nacional de Estadística. Informe.*

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INEI. (22 de Mayo de 2014). Enfermedades no Transmisibles. Obtenido de www.inei.gob.pe. Instituto Nacional de Estadística. Informe.

Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. John Wiley & Sons.Libro

Lerman, D. J. (22 de Enero de 2013). Obtenido de http://entremujeres.clarin.com/vida-sana/salud/Hipertension-sintomas-infarto-acv-ataque_cerebral-angina_de_pecho-miocardio-riesgo-sal_0_1334868810.html. Revista.

Lezcano, R. D. (2010). Minería de datos. Grupo de Investigación en Sistemas e Informática. Articulo.

Salazar Mendiola, J. L., & Vargas Luna, J. M. (2012). Uso de Redes Neuronales para la Medición Automática de Presión Arterial. Revista. México.

Salud, E. S. (16 de Mayo de 2012). Obtenido de Organización Mundial de la Salud: www.paho.org/arg. Revista.

Salud, O. M. S (12 de Mayo de 2014). Organización Panamericana de la Salud. Obtenido de www.paho.org/arg. Revista.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Sanitarias, E. (12 de Mayo de 2012). Organización mundial de la salud. Obtenido de www.paho.org/arg. Artículo.

Martínez, G. R. S., & Mejía, J. A. S. (2011). Arboles de decisiónes en el diagnóstico de enfermedades cardiovasculares. Scientia et Technica, 3(49), 104-109. Tesis. Colombia.

Hernández, F. D., & Corales, Y. S. TÉCNICAS DE MINERÍA DE DATOS APLICADAS AL DIAGNÓSTICO DE ENTIDADES CLÍNICAS DATA MINING TECHNIQUES APLIED TO DIAGNOSYS OF CLINICAL ENTITIES. Revista Cubana de Informática Médica, 12(2).

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

ANEXOS

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ANEXOS

Anexo 01: Costos y presupuestos

PRESUPUESTO

Tabla 12: Costos de suministros de oficina y servicios

Partida	Descripción	Cantidad	Costo (S/.)
1	**Bienes**		**142.50**
	Materiales de escritorio		
	Memoria Flash usb	1 unidad	60.00
	Materiales de impresión		
	Fotocopias	2 unidad	40.00
	Impresiones	53 unidad	35.00
	Anillados	3 unidad	7.50
2	**Servicios**		**280.00**
	Servicio básicos		
	Luz		50.00
	Internet		60.00
	Teléfono		30.00
	Pasajes, viáticos, fletes		100.00
	Otros		40.00
Total			**422.50**

Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 13: Costos de equipos

Descripción	**Cantidad**	**Costo (S/.)**
PC Pentium Dual Core - Analista	01	1,300.00
PC Pentium Dual Core - Diseñador	01	1,300.00
PC Pentium Dual Core - Programador	01	1,300.00
Notebook Lenovo Core I5 – Jefe del Proyecto	01	2,700.00
Total		**6,600.00**

Fuente: Elaboración propia

Tabla 14: Costos durante el funcionamiento del proyecto

Ítem	**Descripción**	**Costo**	**Total (S/.)**
Costo de Internet	04 Meses	120.00	480.00
Costo de Software	04 Meses	00.00	00.00
Mantenimiento de Equipos	04 PC	150.00	600.00
Licencia de Windows 8 Profesional	04 PC	540.00	2,160.00
Licencia Original de Antivirus Karpesky	04 PC	80.00	320.00
Total			**3 560.00**

Fuente: Elaboración propia

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 15: Costo de recursos humanos

Descripción	Cantidad	Duración	Costo (S/.)	Costo (S/.)
Analista	01	03 Meses	700.00	2,100.00
Diseñador	01	03 Meses	600.00	1,800.00
Programador	01	03 Meses	800.00	2,400.00
Jefe del Proyecto	01	03 Meses	1000.00	3,000.00
Total				**9,300.00**

Fuente: Elaboración propia

Tabla 16: Costo total de proyecto

Descripción	Total (S/.)
Costo de Suministros de Oficina y Servicios	422.50
Costos de Equipos	6,600.00
Costo Durante el Funcionamiento del Proyecto	3,560.00
Costo de Recursos Humanos	9,300.00
Total	**19,882.50**

Fuente: Elaboración propia

El Financiamiento para esta investigación es de S/. 19,882.50 Soles.

Financiamiento

El Proyecto será financiado en su totalidad por el responsable de la investigación.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Anexo 02: Instrumentos utilizados

Modelo de Entrevista

Nombres y Apellidos:

1. **¿El Hospital Almanzor Aguinaga Asenjo cuenta con una Gran Base De Datos?**

2. **¿De Cuantos Registros se almacena en la Base de Datos del Hospital Almanzor Aguinaga Asenjo?**

3. **¿Cómo se maneja esa información?**

4. **¿Está información sería muy útil para los doctores de diversas especialidades?**

5. **¿Le gustaría a usted que a través de la información adquirida de un paciente determinado se puede predecir un pre diagnóstico?**

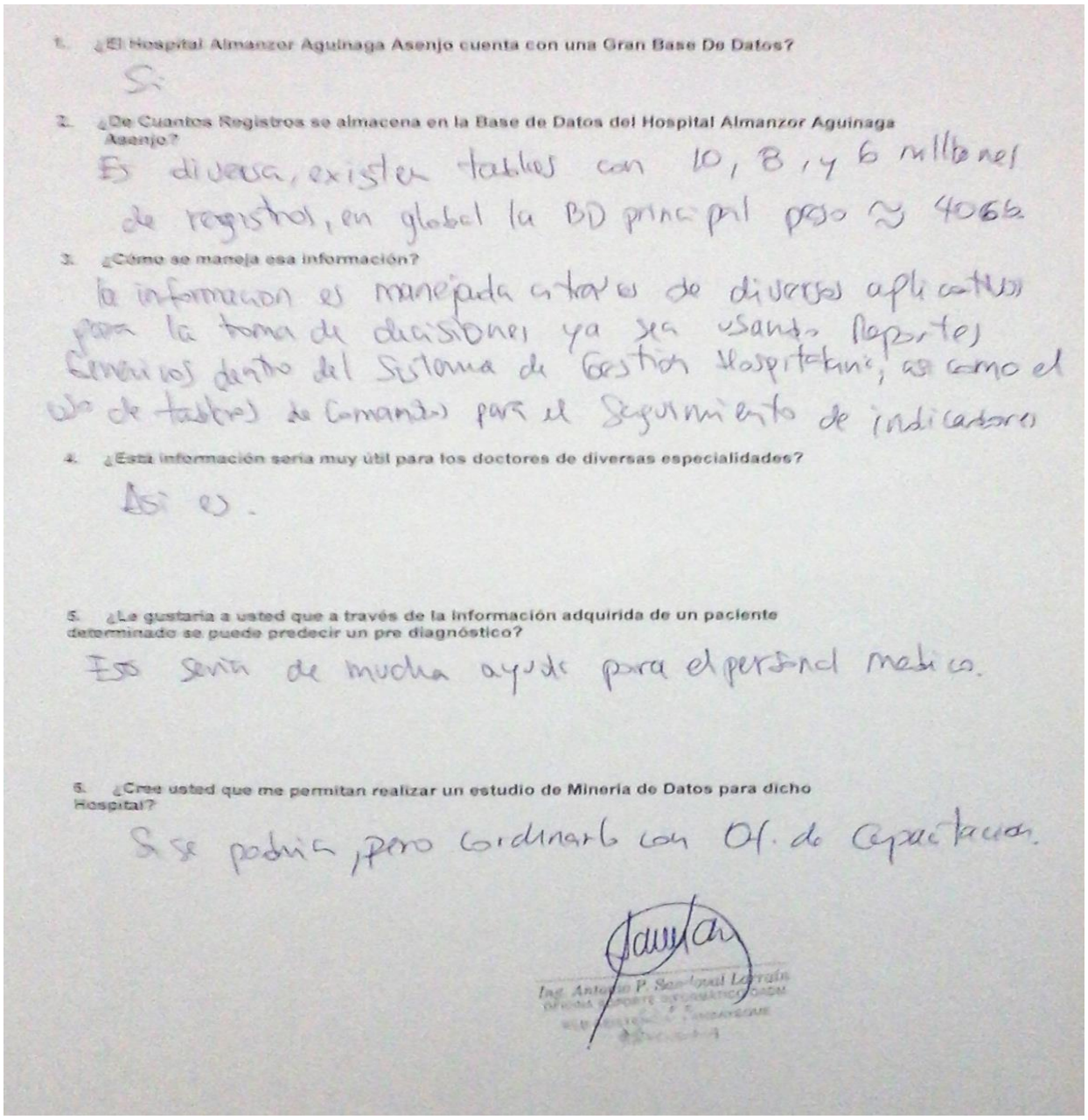
6. **¿Cree usted que me permitan realizar un estudio de Minería de Datos para dicho Hospital?**

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Entrevista – Lugar: Hospital Almanzor Aguinaga Asenjo

Nombres y Apellidos de Entrevistado : Ing. Antonio Sandoval – Área de Informática



1. ¿El Hospital Almanzor Aguinaga Asenjo cuenta con una Gran Base De Datos?

Si

2. ¿De Cuantos Registros se almacena en la Base de Datos del Hospital Almanzor Aguinaga Asenjo?

Es diversa, existen tablas con 10, 8, y 6 millones de registros, en global la BD principal pesa ≈ 40Gb

3. ¿Cómo se maneja esa información?

la información es manejada a través de diversas aplicativos para la toma de decisiones ya sea usando Reportes Gerenciales dentro del Sistema de Gestión Hospitalaria, así como el uso de tableros de Comando para el Seguimiento de indicadores

4. ¿Esta información sería muy útil para los doctores de diversas especialidades?

Así es.

5. ¿Le gustaría a usted que a través de la información adquirida de un paciente determinado se puede predecir un pre diagnóstico?

Esto sería de mucha ayuda para el personal medico.

6. ¿Cree usted que me permitan realizar un estudio de Minería de Datos para dicho Hospital?

Si se podría, pero coordinarlo con Of. de Capacitación.

Ing. Antonio P. Sandoval Lozada

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Anexo 03: Pruebas de resultados obtenidos y visitas en el lugar de investigación

Gráfico 52: Portada de ingreso a EsSalud

Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 53: Área de informática del hospital Almanzor Aguinaga Asenjo

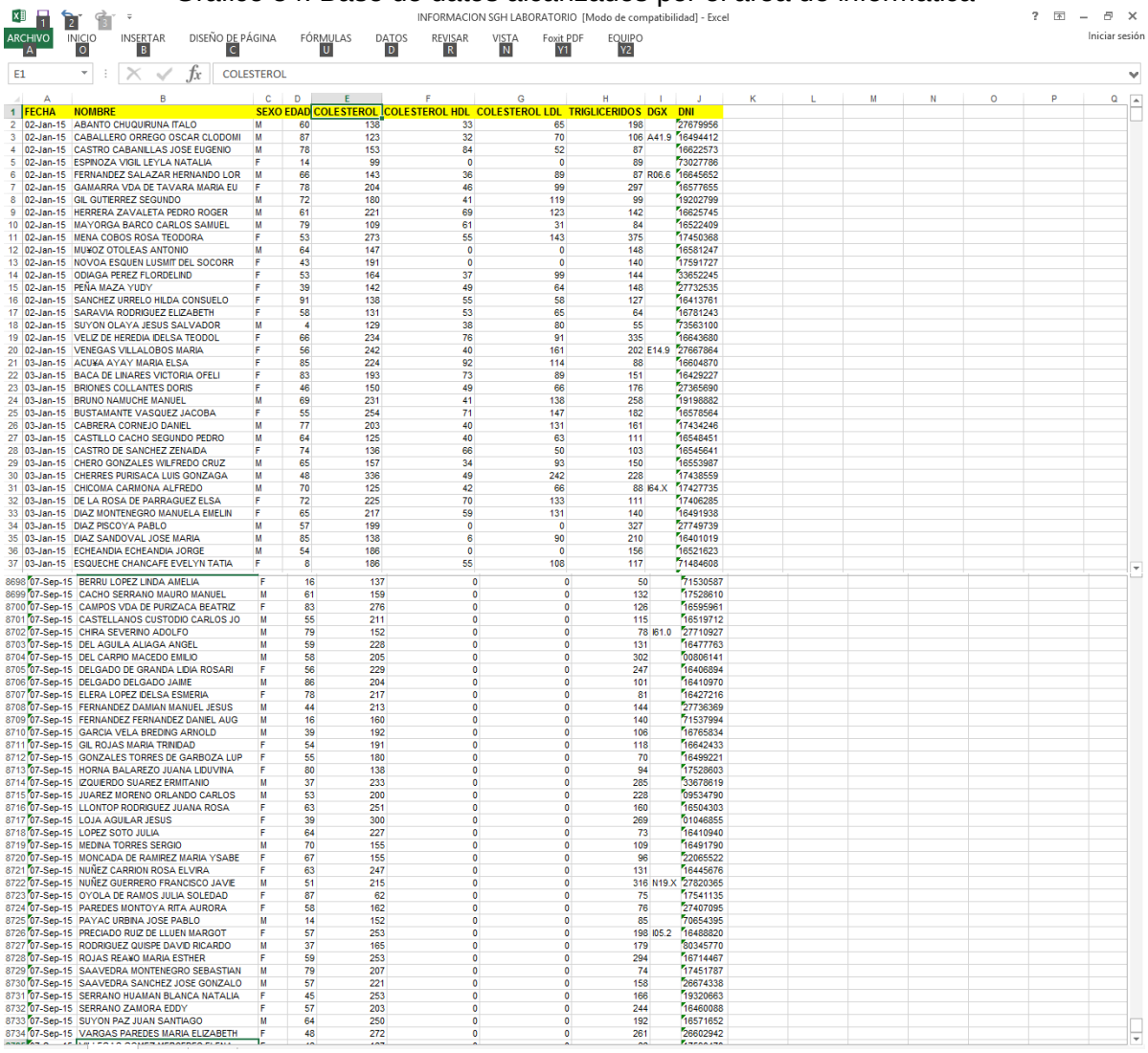
Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Documentos sustentatorios de la base de datos alcanzada por el área de informática con los cuáles realizamos la investigación:

Gráfico 54: Base de datos alcanzados por el área de informática



INFORMACION SGH LABORATORIO [Modo de compatibilidad] - Excel

FECHA	NOMBRE	SEXO	EDAD	COLESTEROL	COLESTEROL HDL	COLESTEROL LDL	TRIGLICERIDOS	DGX	DNI
02-Jan-15	ABANTO CHUQUIRUNA ITALO	M	60	138	33	65	198		27679956
02-Jan-15	CABALLERO ORREGO OSCAR CLODOMI	M	87	123	32	70	106	A41.9	16494412
02-Jan-15	CASTRO CABANILLAS JOSE EUGENIO	M	78	153	84	52	87		16622573
02-Jan-15	ESPINOZA VIGIL LEYLA NATALIA	F	14	99	0	0	89		73027786
02-Jan-15	FERNANDEZ SALAZAR HERNANDO LOR	M	66	143	36	89	87	R06.6	16645652
02-Jan-15	GAMARRA VDA DE TAVARA MARIA EU	F	78	204	46	99	297		16577655
02-Jan-15	GIL GUTIERREZ SEGUNDO	M	72	180	41	119	99		19202799
02-Jan-15	HERRERA ZAVALETA PEDRO ROGER	M	61	221	69	123	142		16625745
02-Jan-15	MAYORGA BARCO CARLOS SAMUEL	M	79	109	61	31	84		16522409
02-Jan-15	MENA COBOS ROSA TEODORA	F	53	273	55	143	375		17450368
02-Jan-15	MUÑOZ OTOLEAS ANTONIO	M	64	147	0	0	148		16581247
02-Jan-15	NOVOA ESQUEN LUSMIT DEL SOCORR	F	43	191	0	0	140		17591727
02-Jan-15	ODIAGA PEREZ FLORDELIND	F	53	164	37	99	144		33652245
02-Jan-15	PEÑA MAZA YUDY	F	39	142	49	64	148		27732535
02-Jan-15	SANCHEZ URRELO HILDA CONSUELO	F	91	138	55	58	127		16413761
02-Jan-15	SARAVIA RODRIGUEZ ELIZABETH	F	58	131	53	65	64		16781243
02-Jan-15	SUYON OLAYA JESUS SALVADOR	M	4	129	38	80	55		73563100
02-Jan-15	VELIZ DE HEREDIA IDELSA TEODOL	F	66	234	76	91	335		16643680
02-Jan-15	VENEGAS VILLALOBOS MARIA	F	56	242	40	161	202	E14.9	27667864
03-Jan-15	ACUNA AYAY MARIA ELSA	F	85	224	92	114	88		16804870
03-Jan-15	BACA DE LINARES VICTORIA OFELI	F	83	193	73	89	151		16429227
03-Jan-15	BRIONES COLLANTES DORIS	F	46	150	49	66	176		27365690
03-Jan-15	BRUNO NAMUCHE MANUEL	M	69	231	41	138	258		19198882
03-Jan-15	BUSTAMANTE VASQUEZ JACOBA	F	55	254	71	147	182		16578564
03-Jan-15	CABRERA CORNEJO DANIEL	M	77	203	40	131	161		17434246
03-Jan-15	CASTILLO CACHO SEGUNDO PEDRO	M	64	125	40	63	111		16548451
03-Jan-15	CASTRO DE SANCHEZ ZENAIDA	F	74	136	66	50	103		16545641
03-Jan-15	CHERO GONZALES WILFREDO CRUZ	M	65	157	34	93	150		16553987
03-Jan-15	CHERRES PURISACA LUIS GONZAGA	M	48	336	49	242	228		17438559
03-Jan-15	CHICOMA CARMONA ALFREDO	M	70	125	42	66	88	I64.X	17427735
03-Jan-15	DE LA ROSA DE PARRAGUEZ ELSA	F	72	225	70	133	111		17406285
03-Jan-15	DIAZ MONTENEGRO MANUELA EMELIN	F	65	217	59	131	140		16491938
03-Jan-15	DIAZ PISCOYA PABLO	M	57	199	0	0	327		27749739
03-Jan-15	DIAZ SANDOVAL JOSE MARIA	M	85	138	6	90	210		16401019
03-Jan-15	ECHEANDIA ECHEANDIA JORGE	M	54	186	0	0	156		16521623
03-Jan-15	ESQUECHE CHANCAFE EVELYN TATIA	F	8	186	55	108	117		71484608
07-Sep-15	BERRU LOPEZ LINDA AMELIA	F	16	137	0	0	50		71530587
07-Sep-15	CACHO SERRANO MAURO MANUEL	M	61	159	0	0	132		17528610
07-Sep-15	CAMPOS VDA DE PURIZACA BEATRIZ	F	83	276	0	0	126		16595961
07-Sep-15	CASTELLANOS CUSTODIO CARLOS JO	M	55	211	0	0	115		16519712
07-Sep-15	CHIRA SEVERINO ADOLFO	M	79	152	0	0	78	I61.0	27710927
07-Sep-15	DEL AGUILA ALIAGA ANGEL	M	59	228	0	0	131		16477763
07-Sep-15	DEL CARPIO MACEDO EMILIO	M	58	205	0	0	302		00806141
07-Sep-15	DELGADO DE GRANDA LIDIA ROSARI	F	56	229	0	0	247		16406894
07-Sep-15	DELGADO DELGADO JAIME	M	86	204	0	0	101		16410970
07-Sep-15	ELERA LOPEZ IDELSA ESMERIA	F	78	217	0	0	81		16427216
07-Sep-15	FERNANDEZ DAMIAN MANUEL JESUS	M	44	213	0	0	144		27736369
07-Sep-15	FERNANDEZ FERNANDEZ DANIEL AUG	M	16	160	0	0	140		71537994
07-Sep-15	GARCIA VELA BREDING ARNOLD	M	39	192	0	0	106		16765834
07-Sep-15	GIL ROJAS MARIA TRINIDAD	F	54	191	0	0	118		16642433
07-Sep-15	GONZALES TORRES DE GARBOZA LUP	F	55	180	0	0	70		16499221
07-Sep-15	HORNA BALAREZO JUANA LIDUVINA	F	80	138	0	0	94		17528603
07-Sep-15	IZQUIERDO SUAREZ ERMITANIO	M	37	233	0	0	285		33678619
07-Sep-15	JUAREZ MORENO ORLANDO CARLOS	M	53	200	0	0	228		09534790
07-Sep-15	LLONTOP RODRIGUEZ JUANA ROSA	F	63	251	0	0	160		16504303
07-Sep-15	LOJA AGUILAR JESUS	F	39	300	0	0	269		01046855
07-Sep-15	LOPEZ SOTO JULIA	F	64	227	0	0	73		16410940
07-Sep-15	MEDINA TORRES SERGIO	M	70	155	0	0	109		16491790
07-Sep-15	MONCADA DE RAMIREZ MARIA YSABE	F	67	155	0	0	96		22065522
07-Sep-15	NUÑEZ CARRION ROSA ELVIRA	F	63	247	0	0	131		16445676
07-Sep-15	NUÑEZ GUERRERO FRANCISCO JAVIE	M	51	215	0	0	316	N19.X	27820365
07-Sep-15	OYOLA DE RAMOS JULIA SOLEDAD	F	87	62	0	0	75		17541135
07-Sep-15	PAREDES MONTOYA RITA AURORA	F	58	162	0	0	76		27407095
07-Sep-15	PAYAC URBINA JOSE PABLO	M	14	152	0	0	85		70654395
07-Sep-15	PRECIADO RUIZ DE LLUEN MARGOT	F	57	253	0	0	198	I05.2	16488820
07-Sep-15	RODRIGUEZ QUISPE DAVID RICARDO	M	37	165	0	0	179		80345770
07-Sep-15	ROJAS REAÑO MARIA ESTHER	F	59	253	0	0	294		16714467
07-Sep-15	SAAVEDRA MONTENEGRO SEBASTIAN	M	79	207	0	0	74		17451787
07-Sep-15	SAAVEDRA SANCHEZ JOSE GONZALO	M	57	221	0	0	158		26674338
07-Sep-15	SERRANO HUAMAN BLANCA NATALIA	F	45	253	0	0	166		19320663
07-Sep-15	SERRANO ZAMORA EDDY	F	57	203	0	0	244		16460088
07-Sep-15	SUYON PAZ JUAN SANTIAGO	M	64	250	0	0	192		16571652
07-Sep-15	VARGAS PAREDES MARIA ELIZABETH	F	48	272	0	0	261		26602942

Fuente: Elaboración propia

Como se muestra fueron un total de 8,735 pacientes para el estudio.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Técnica de Reglas de Asociación

Tabla 17: Resultados de Pruebas – Reglas de Asociación

Premisas	***Conclusión***	***Tamaño***	***Confiabilidad***
Trigliceridos, Colesterol LDL	Colesterol , Colesterol HDL	676	0.876
Trigliceridos, Colesterol LDL	Colesterol HDL	683	0.885
Colesterol LDL	Colesterol , Colesterol HDL	1,274	0.887
Colesterol, Trigliceridos, Colesterol LDL	Colesterol HDL	676	0.888
Colesterol LDL	Colesterol HDL	1,324	0.922
Colesterol, Colesterol LDL	Colesterol HDL	1,274	0.93
Colesterol LDL	Colesterol	1,370	0.954
Colesterol HDL, Colesterol LDL	Colesterol	1,274	0.962
Trigliceridos, Colesterol LDL	Colesterol	760	0.986
trigliceridos, Colesterol HDL, Colesterol LDL	Colesterol	676	0.99

Fuente: Elaboración propia

Según el estudio con la técnica de reglas de asociación para la base de datos de pacientes, de essalud. Que toda persona que tiene triglicéridos y colesterol LDL también tiene colesterol elevado con un nivel de confianza de 98.6% y que en primera vista el 0.986 de los pacientes que representa los 760, tendrá hipertensión arterial, ya tiene triglicéridos y colesterol LDL alto.

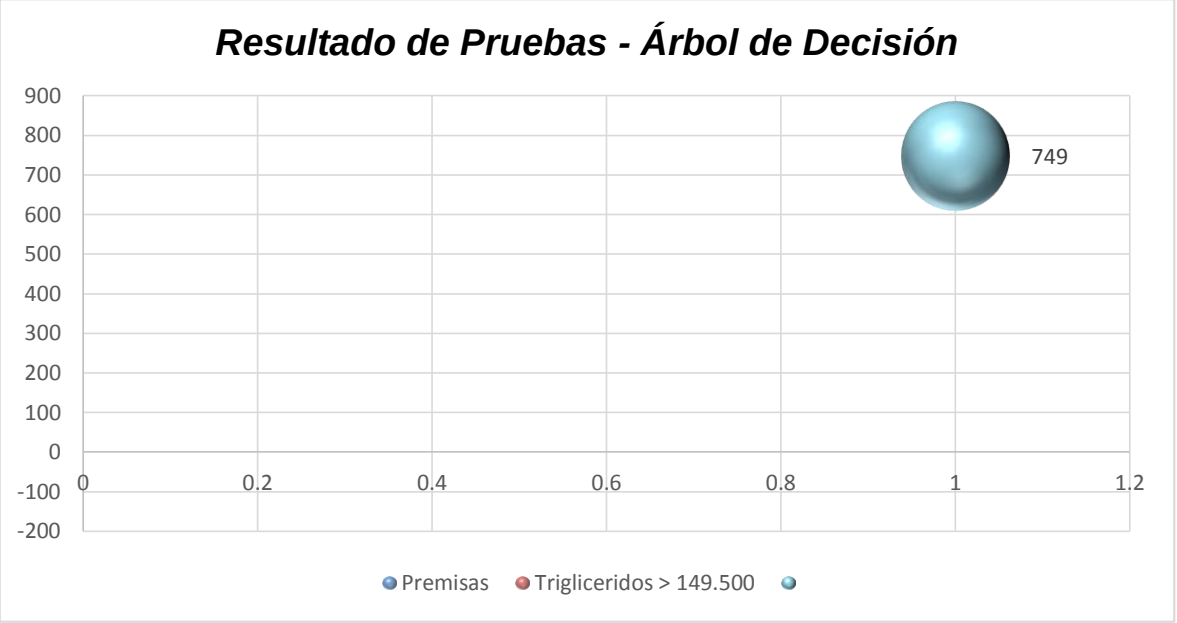
Técnica de Árbol de Decisión

Premisas	***Conclusión***
Colesterol LDL > 129.500	Colesterol LDL alto
Trigliceridos > 149.500	Trigliceridos alto
Colesterol > 199	Colesterol alto

Según el estudio con la técnica de árbol de decisión. Si una persona tiene el colesterol ldl > 129.500, triglicéridos > 149.5, colesterol > 199 se llega a la conclusión que tiene colesterol ldl alto, triglicéridos alto, colesterol alto.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Gráfico 55: *Resultado de pruebas – árbol de decisión*



Fuente: Elaboración propia

749 pacientes podrían adquirir la enfermedad de hipertensión arterial

Reglas de Asociación

Registros	*Pacientes con HA*	*Pacientes sin HA*	*Nivel de Confianza*
8735	760	7975	98.6 %

Según el estudio con la técnica de reglas de asosiación para la base de datos de pacientes del ESSALUD de 8,735 registros, 760 con un nivel de confianza de 98.6% pacientes podrían adquirir la enfermedad de Hipertensión Arterial.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Técnica de Árbol de Decisión

Registros	*Pacientes con HA*	*Pacientes sin HPA*	*Nivel de Confianza*
8735	749	7986	97%

Según el estudio con la técnica de árbol de decisión para la base de datos de pacientes de ESSALUD, de 8,735 registros, 749 pacientes con nivel de confianza de 97 % podrían adquirir la enfermedad de Hipertensión Arterial.

Anexo 04: Modelos matemáticos empleados en el estudio:

Reglas de asociación

Asociación se define como:

Sea $I = \{ i_1, i_2, \dots, i_n \}$ un conjunto de n atributos binarios llamados ítems.

Sea $D = \{ t_1, t, \dots, t_n \}$ un conjunto de transacciones almacenadas en una base de datos.

Cada transacción en D tiene un ID (identificador) único y contiene un subconjunto de items de I. Una regla se define como una implicación de la forma:

$$X \Rightarrow Y$$

Donde:

$X, Y \subseteq I$ y

$X \cap Y = \emptyset$

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Técnica de árbol de decisión

Es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.). De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

Realiza esta labor mediante la construcción de un árbol de decisión.

Los elementos son:

- Nodos: Los cuales contendrán atributos.
- Arcos: Los cuales contienen valores posibles del nodo padre.
- Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

Elección del Mejor Atributo

La elección del mejor atributo se establece mediante la entropía. Eligiendo aquel que proporcione una mejor ganancia de información. La función elegida puede variar, pero en su forma más sencilla es como esta:

$$-\left(\frac{|p|}{|d|}\right)\log_2\left(\frac{|p|}{|d|}\right)-\left(\frac{|n|}{|d|}\right)\log_2\left(\frac{|n|}{|d|}\right)$$

Donde *p* es el conjunto de los ejemplos positivos, *n* el de los negativos y *d* el total de ellos. Se debe establecer si el logaritmo es positivo o negativo.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Anexo 05: Código y documentación del sitio web

ControladorUsuario.java

```
package controlador;

import java.io.IOException;

import java.io.PrintWriter;

import javax.servlet.RequestDispatcher;

import javax.servlet.ServletException;

import javax.servlet.http.HttpServlet;

import javax.servlet.http.HttpServletRequest;

import javax.servlet.http.HttpServletResponse;

import javax.servlet.http.HttpSession;

import modelo.beans.Usuario;

import modelo.down.DowArbolDecisión;

import modelo.down.DowMigracion;

import modelo.down.DowReglasAsociación;

import modelo.down.DowUsuario;


public class ControladorUsuario extends HttpServlet {


  // En este caso yo le pido entrar y lo que realiza aquí es capturar el usuario y el password

  //de las vistas osea del index.jsp y nos lleva a controladorUsuario


  protected void processRequest(HttpServletRequest request, HttpServletResponse response)

      throws ServletException, IOException {

    response.setContentType("text/html;charset=UTF-8");

    try (PrintWriter out = response.getWriter()) {
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
String operacion = request.getParameter("operacion");
//si la operación es iniciar voy a recuperar esos dos campos usuario y password
//y los guardo en un array que se llama parámetros
if (operacion.equals("iniciar")) {

   String[] parametros = new String[2];
   parametros[0] = request.getParameter("usuario");
   parametros[1] = request.getParameter("password");
   // dowusuario tiene metodos que conectan con la base de datos
   DowUsuario oDowUsuario = new DowUsuario();
   //Y  pongo los datos del array en ese método que es
   //oDowUsuarioautenficacion me va a devolver un objeto usuario
   Usuario oUsuario = oDowUsuario.autentficacion(parametros);
   //si hay un usuario y contraseña va a ser diferente a nulo
   //me ejecuta todo y si no me devuelve al index.jsp
   if (oUsuario.getId() != null) {
      // para crear la sesión o inicializarla
      // asignando sesión de usuario.
      // una vez creada o inicalizada le voy a agregar un atributo
      //login y le voy  a dar al usuario
      HttpSession oSession = request.getSession(true);
      oSession.setAttribute("login", oUsuario);

      /**
       * voy a llamar a la clase que tiene los metodos del rapidminer
       * Ejecuntando algoritmos de rapidminer-->
       */
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
// arbol de decisión. llamo a la clase (DowArbolDecisión)que tiene los metodos
//para llamar a rapidminer

DowArbolDecisión oArbolDesicion = new DowArbolDecisión();
oArbolDesicion.generarAlgoritmo();
 // reglas de asociación. llamo a la clase (DowReglasAsociación)que tiene los
// metodos  para llamar a rapidminer
DowReglasAsociación oReglasAsociación = new DowReglasAsociación();
oReglasAsociación.generarAlgoritmo();

/**
 * MIgracion de datos a mysql
 //paso los datos convertidos a mi bd
 // llamo a mi clase Dowmigracion
 */
DowMigracion oMigracion = new DowMigracion();
oMigracion.migracionResultArbolDesicion();
oMigracion.migracionResultReglasAsociación();

// asignando lista de datos importantes.
// asigno valores para mis vistas
// me voy a cargar datos como mis tablas ya estan llenas
// y llamo esos datos con cargardatos y los muestro en mi vista admin.jsp
oSession.setAttribute("datos", oDowUsuario.cargarDatos());
// direccionando a la pagina administrador
RequestDispatcher oDispatcher = request.getRequestDispatcher("admin.jsp");
oDispatcher.forward(request, response);
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
  //en caso contrario me devuelve al index.jsp
    } else {
        response.sendRedirect("index.jsp");
    }
}
  if (operacion.equals("registrar")) {
    String[] parametros = new String[2];
    parametros[0] = request.getParameter("usuario");
    parametros[1] = request.getParameter("password");
    // dowusuario tiene metodos que conectan con la base de datos
    DowUsuario oDowUsuario = new DowUsuario();
    if (oDowUsuario.registrar(parametros)) {

      //autentificando usuario
     //Y  pongo los datos del array en ese método que es
    //oDowUsuarioautenficacion me va a devolver un objeto usuario
        Usuario oUsuario = oDowUsuario.autentficacion(parametros);

      // asignando sesión de usuario.
      // para crear la sesión o inicializarla
      // asignando sesión de usuario.
      // una vez creada o inicalizada le voy a agregar un atributo
      //login y le voy  a dar al usuario
      HttpSession oSession = request.getSession(true);
      oSession.setAttribute("login", oUsuario);

       /**
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

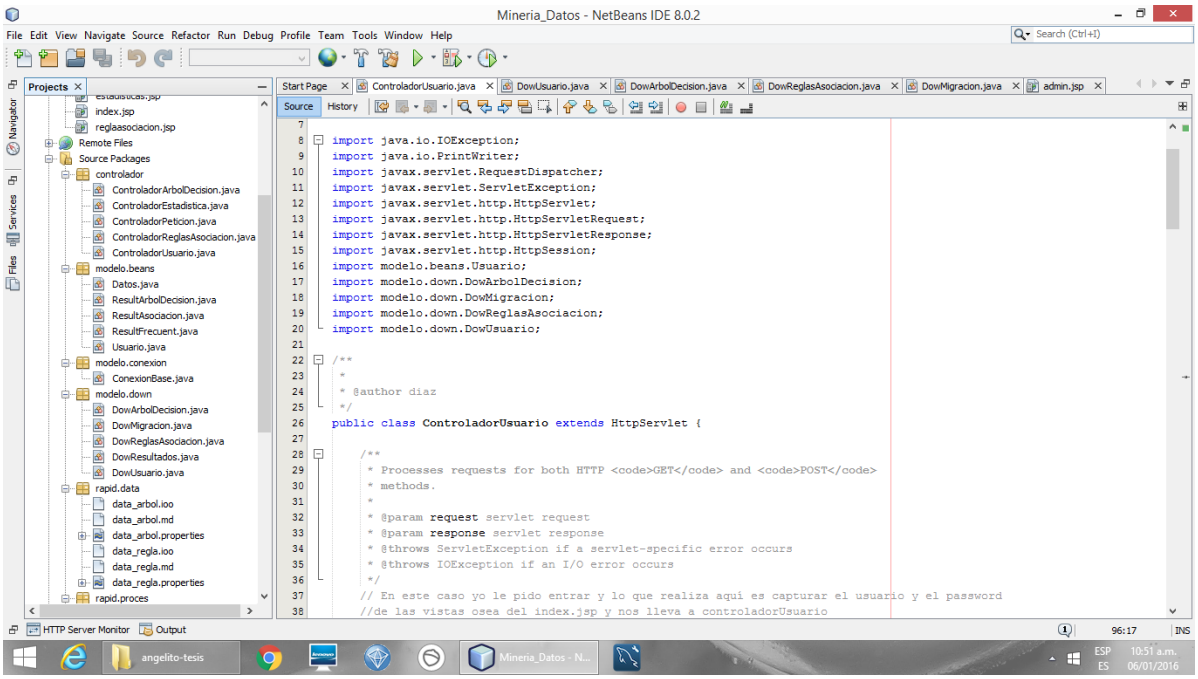
```
 * voy a llamar a la clase que tiene los metodos del rapidminer
 * Ejecuntando algoritmos de rapidminer-->
 */
// arbol de decisión. llamo a la clase (DowArbolDecisión)que tiene los metodos
//para llamar a rapidminer
DowArbolDecisión oArbolDesicion = new DowArbolDecisión();
oArbolDesicion.generarAlgoritmo();
// reglas de asociación
 // reglas de asociación. llamo a la clase (DowReglasAsociación)que tiene los
//metodos   para llamar a rapidminer
DowReglasAsociación oReglasAsociación = new DowReglasAsociación();
oReglasAsociación.generarAlgoritmo();

/**
 * MIgracion de datos a mysql
 //paso los datos convertidos a mi bd
 // llamo a mi clase Dowmigracion
 */

DowMigracion oMigracion = new DowMigracion();
oMigracion.migracionResultArbolDesicion();
oMigracion.migracionResultReglasAsociación();

// asignando lista de datos importantes.
 // asigno valores para mis vistas
 // me voy a cargar datos como mis tablas ya estan llenas
 // y llamo esos datos con cargardatos y los muestro en mi vista admin.jsp
oSession.setAttribute("datos", oDowUsuario.cargarDatos());
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
            // direccionando a la pagina administrador
            // lo que hace requestdispatcher es que todos los atributos que
            //se almacena en esta sesión los envie a admin.jsp
            RequestDispatcher oDispatcher = request.getRequestDispatcher("admin.jsp");
            oDispatcher.forward(request, response);
            //en caso contrario me devuelve al index.jsp
          } else {
            response.sendRedirect("index.jsp");
          }
        }
    }
}
```

Gráfico 56: ControladorUsuario.java



Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DowUsuario.java

```
package modelo.down;


import java.sql.ResultSet;
import java.sql.SQLException;
import java.util.logging.Level;
import java.util.logging.Logger;
import modelo.beans.Datos;
import modelo.beans.Usuario;
import modelo.conexion.ConexionBase;


public class DowUsuario {
   //recibe un string de parametros que es usuario y password
   // y me devuelve un usuario
   public Usuario autentficacion(String[] parametros) {


      try {
         //aqui yo me conecto con la base de datos
         ConexionBase oConexionBase = new ConexionBase();


         Usuario oUsuario = new Usuario();
         // le envío un  metodo call autentificar y dos parametros
         String sql = "{call autentificar(?,?)}";
         // ejecuto que me devuelva los datos envío los parametros y el
         // sql.
         // sql es la sentecia y los parametros  son usuario y paswword
         ResultSet rs = oConexionBase.datosProcedure(parametros, sql);
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

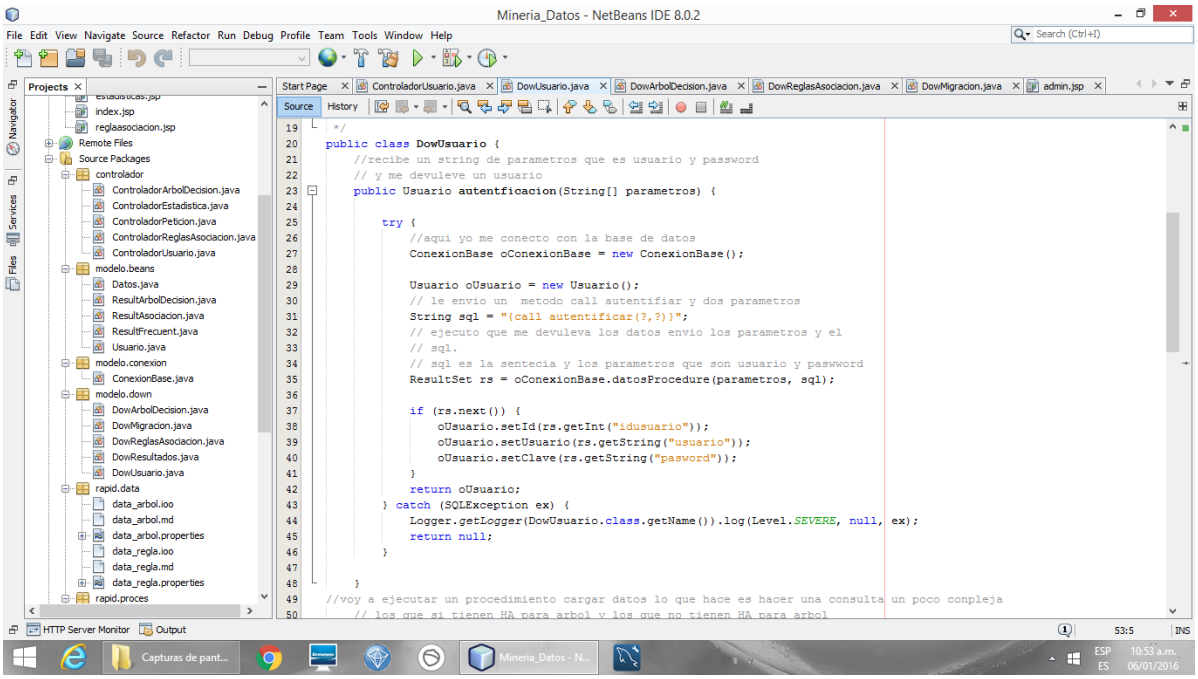
```
            if (rs.next()) {
                oUsuario.setId(rs.getInt("idusuario"));
                oUsuario.setUsuario(rs.getString("usuario"));
                oUsuario.setClave(rs.getString("pasword"));
            }
            return oUsuario;
        } catch (SQLException ex) {
            Logger.getLogger(DowUsuario.class.getName()).log(Level.SEVERE, null, ex);
            return null;
        }
    }
//voy a ejecutar un procedimiento cargar datos lo que hace es  una consulta un poco conpleja
    // los que si tienen HA para arbol y los que no tienen HA para arbol
    // el nivel de  confienza para arbol y el nivel de  confianza para regla

    public Datos cargarDatos() {
        try {
            ConexionBase oConexionBase = new ConexionBase();
            Datos oDatos = new Datos();
            String sql = "{call cargarDatos()}";
            ResultSet rs = oConexionBase.datosProcedure(null, sql);
            // set para asignarle los valores a la bd
            // get para mostrar los datos
            // si consulta tiene algun archivo que me saque ( arbol si tienen HA) a_si
              (arbol si no tiene HA) A_NO  y lo mismo para reglas de asociación b_si y  b_no
 if (rs.next()) {
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
            oDatos.setSi_arbol(rs.getString("a_si"));
            oDatos.setNo_arbol(rs.getString("a_no"));
            oDatos.setSi_regla(rs.getString("b_si"));
            oDatos.setNo_regla(rs.getString("b_no"));
            oDatos.setConfianza_arbol(rs.getString("a_cz"));
            oDatos.setConfianza_regla(rs.getString("b_cz"));
        }
        // una vez que ya esta lleno me lo devuelve en la clase odatos
        return oDatos;
    } catch (SQLException ex) {
        System.out.println("ERROOR--> " + ex.getLocalizedMessage());
        return null;
    }
}
```

Gráfico 57: DowUsuario.java



Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DowArbolDecisión.java

//esta clase o algoritmo es el que llama las librerias del rapidminer

```
package modelo.down;


import com.csvreader.CsvWriter;
//com.rapidminer //process.class esta llamando a la clase de la libreria
import com.rapidminer.RapidMiner;
import com.rapidminer.Process;
import com.rapidminer.example.Attribute;
import com.rapidminer.example.Example;
import com.rapidminer.example.ExampleSet;
import com.rapidminer.operator.IOContainer;
import com.rapidminer.operator.Operator;
import com.rapidminer.operator.OperatorCreationException;
import com.rapidminer.operator.OperatorException;
import com.rapidminer.operator.io.RepositoryStorer;
import com.rapidminer.operator.nio.file.FileObject;
import com.rapidminer.operator.nio.file.SimpleFileObject;
import com.rapidminer.tools.OperatorService;
import com.rapidminer.tools.XMLException;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.net.URL;
import java.util.ArrayList;
import java.util.Iterator;
import java.util.List;
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
import java.util.logging.Level;
import java.util.logging.Logger;


/**
 *
 * @author Iamlu_000
 */
//este algoritmo es el que maneja las librerias de rapidminer para hacer
// esta clase llama a las librerias de rapidminer
public class DowArbolDecisión {


   public boolean generarAlgoritmo() {


      try {
// este metodo crea 3 atributos.
// almacena el resultado (resultset)
//hay dos tipos de conteiner uno para recibir (ioresult) o y otro entregar resultados (ioinput)
         ExampleSet resultSet;
         IOContainer ioResult;
         IOContainer ioInput;
// todo el url es para indicarme en que carpeta se encuentra
         //se dirige a todas las carpetas rapid que son 3 carpetas
         //rapid.data, rapid.proces,rapid.result
         URL url = this.getClass().getResource("/rapid");
         //crea la dirección
         // una vez que tengo la dirección donde estan mis datos principales que
          // es rapid.data(data es la bd) y el rapid.proces
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
String direccion = url.toString().substring(5, url.toString().length());
//esto indica el modo en que se va a ejecutar el rapidminer es en linea de comandos

RapidMiner.setExecutionMode(RapidMiner.ExecutionMode.COMMAND_LINE);
//con esto inicio el rapidminer
//lo que va a ser es jalar el init que es de la clase import com.rapidminer.RapidMiner
//lo que estoy haciendo es importar la libreria
RapidMiner.init();
// la dirección es donde se encuentra todo. Una vez hecho eso me voy a la carpeta proces
//y recupero proceso_arbol.rmp
//el que se dibujo en rapidminer y recupero el proceso Arbol.rmp
String ruta = direccion + "/proces/proceso_arbol.rmp";
//ahora creo un nuevo proceso pero con la ruta es decir crea el proceso con el modelo
//que has creado en rapidminer.
Process oProcess = new Process(new File(ruta));
// ahora falta darle la ruta de la base de datos. los datos
//operatorservice tambien esta en la libreria de rapidminer.jar
// repositorystore se encuentra en la libreria de rapidminer.jar
//lo que hace es crea el operador -- y el repository.class es un inicializador propio del rapimdiner
// y este operador ya esta iniciado.
Operator dataOperator = OperatorService.createOperator(RepositoryStorer.class);
//ya declarado y ya inicializado arriba. ahora con setparameter le voy a indicar como se llama bd
// y le envias  dirección=  rapid + indico en que carpeta esta.
dataOperator.setParameter("data_arbol", direccion + "/data/data_arbol");

// ahora al fileobject le voy a dar la ruta (osea la ruta es donde esta el proceso guardado osea rmp)
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
FileObject file = new SimpleFileObject(new File(ruta));
//ahora el ioinput es para introducir datos.
//y le voy a introducir la ruta ( por intermedio del file)
ioInput = new IOContainer(file);



// aqui colocamos  todo los resultados  en el ioresult
// agarro el proceso oprocess.run y lo ejecuto es decir agarro iounput donde esta la      ruta
  y lo
//envia a un conteiner que esta declarado arriba ioResul.
ioResult = oProcess.run(ioInput);



 // esto coloca los resultados rapid.result
 //Expotar Resultados
 // aqui creo una lista donde todos  los datos que tengo los meto a mi  lista y de mi lista
//y una vez que tengo llena mi lista hay yo recien voy a transformar en un csv para
  pasarlo a mi  bd
 List<Attribute> lista = new ArrayList<>();


if (ioResult.getElementAt(0) instanceof ExampleSet) {
   resultSet = (ExampleSet) ioResult.getElementAt(0);


   for (Example example : resultSet) {
      Iterator<Attribute> allAtts = example.getAttributes().allAttributes();
      while (allAtts.hasNext()) {
         Attribute a = allAtts.next();
         lista.add(a);
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

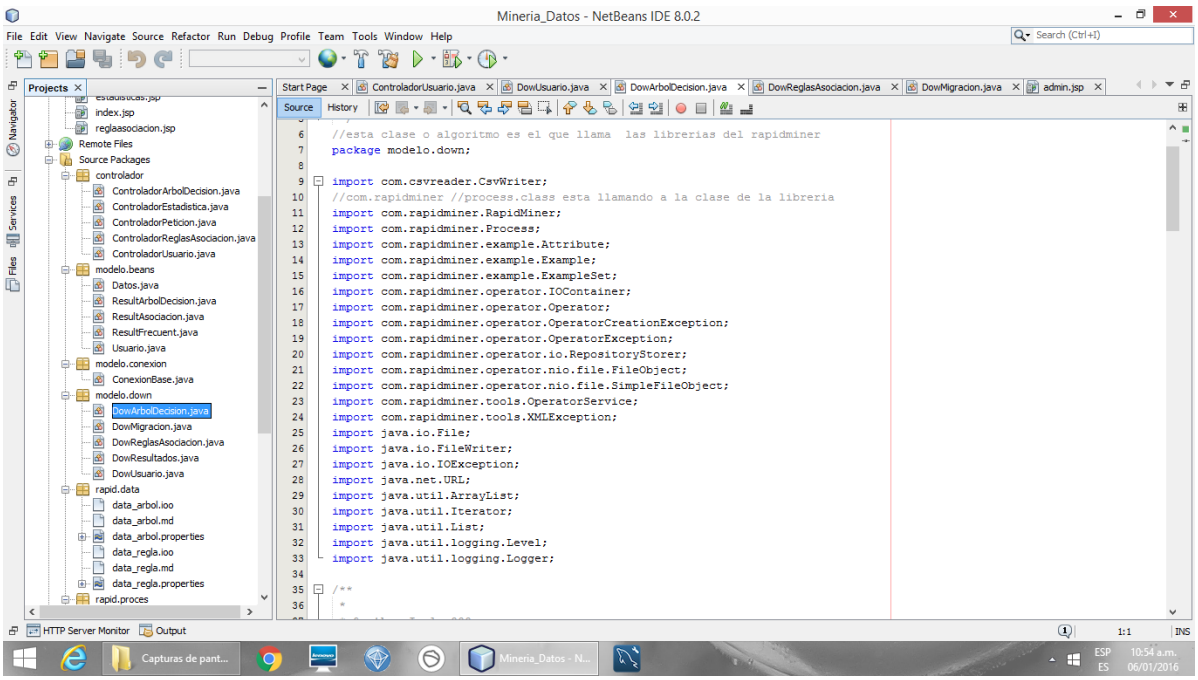
```
        }
    }
}


// pasando datos a archivo csv
 //llamamos a una libreria javacsv esta librreria me ayuda a convertir datos de una lista
  //a csv


for (int i = 0; i < 1; i++) {
    //aqui le indico a donde quiero exportarlo
    String outputFile = direccion + "/result/result_arbol.csv";
    // me devuelve un falso o verdadero si existe o no
    boolean alreadyExists = new File(outputFile).exists();
    // si no  existe el archivo csv
    if (alreadyExists) {
        //lo creo
        File ficheroUsuarios = new File(outputFile);
        ficheroUsuarios.delete();
    }


    try {
          // una vez creado le digo que esten separados x coma
        CsvWriter csvOutput = new CsvWriter(new FileWriter(outputFile, true), ',');
        // una vez hecho eso todos los datos de mi lista los paso al csv
        //hago un for y todo los datos de mi lista lo comienzo a pasar al csv y lo convierto
       //en string
        for (int j = 0; j < lista.size(); j++) {
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
            csvOutput.write(lista.get(i).toString());
            // grabo todo los datos que voy pasando
            csvOutput.endRecord();
        }
        // cierra
        csvOutput.close();

    } catch (IOException e) {
    }
}
```

Gráfico 58: DowArbolDesicion.java



Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DowReglasAsociación.java

```
//esta clase o algoritmo es el que llama  las librerias del rapidminer
package modelo.down;
//com.rapidminer //process.class esta llamando a la clase de la libreria
import com.csvreader.CsvWriter;
import com.rapidminer.RapidMiner;
import com.rapidminer.example.Attribute;
import com.rapidminer.example.Example;
import com.rapidminer.example.ExampleSet;
import com.rapidminer.operator.IOContainer;
import com.rapidminer.operator.Operator;
import com.rapidminer.operator.OperatorCreationException;
import com.rapidminer.operator.OperatorException;
import com.rapidminer.operator.io.RepositoryStorer;
import com.rapidminer.operator.nio.file.FileObject;
import com.rapidminer.operator.nio.file.SimpleFileObject;
import com.rapidminer.tools.OperatorService;
import com.rapidminer.tools.XMLException;
import java.io.File;
import java.io.FileWriter;
import java.io.IOException;
import java.net.URL;
import java.util.ArrayList;
import java.util.Iterator;
import java.util.List;
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
import java.util.logging.Level;
import java.util.logging.Logger;

//este algoritmo es el que maneja las librerias de rapidminer
//esta clase llama a las librerias de rapidminer
public class DowReglasAsociación {

   public boolean generarAlgoritmo() {

      try {
// este metodo crea 3 atributos.
// almacena el resultado (resultset)
//hay dos tipos de conteiner uno para recibir (ioresult)  y otro entregar resultados (ioinput)
         ExampleSet resultSet;
         IOContainer ioResult;
         IOContainer ioInput;
// todo el url es para indicarme en que carpeta se encuentra
// y se dirige a todas las carpetas rapid que son 3 carpetas
//rapid.data, rapid.proces,rapid.result
         URL url = this.getClass().getResource("/rapid");
//crea la dirección
// una vez que tengo la dirección donde estan mis datos principales que
// es rapid.data(data es la bd) y el rapid.proces
         String direccion = url.toString().substring(5, url.toString().length());
//esta indica el modo en que se va a ejecutar el rapidminer es en linea de comandos

         RapidMiner.setExecutionMode(RapidMiner.ExecutionMode.COMMAND_LINE);
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
//con esto inicio el rapidminer
//lo que va a ser es jalar el init que es de la clase import com.rapidminer.RapidMiner
//lo que estoy haciendo es importar la libreria
        RapidMiner.init();
// la dirección es donde se encuentra todo. Una vez hecho eso me voy a la carpeta proces
//y recupero proceso_regla.rmp
//el que se dibujo en rapidminer y recupero el proceso regla.rmp
        String ruta = direccion + "/proces/proceso_regla.rmp";
//ahora creo un nuevo proceso pero con la ruta es decir crea el proceso con el modelo
//que has creado en rapidminer.
        com.rapidminer.Process oProcess = new com.rapidminer.Process(new File(ruta));
// ahora falta darle la ruta de la base de datos los datos
//operatorservice tambien esta en la libreria de rapidminer.jar
// repositorystore se encuentra en la libreria de rapidminer.jar
//lo que hace es crea el operador -- y el repository.class es un inicializador propio del rapimdiner
// y este operador ya esta iniciado.
        Operator dataOperator = OperatorService.createOperator(RepositoryStorer.class);
//ya declarado y ya inicializado arriba. ahora con setparameter le voy a indicar como se llama bd
// y le envias  direccion=  rapid + indico en que carpeta esta.
        dataOperator.setParameter("data_arbol", direccion + "/data/data_regla");
// ahora al fileobject le voy a dar la ruta (osea la ruta es donde esta el proceso guardado osea rmp)
        FileObject file = new SimpleFileObject(new File(ruta));
//ahora el ioinput es para introducir datos.
//y le voy a introducir la ruta ( por intermedio del file)
        ioInput = new IOContainer(file);

// aqui colocamos  todo los resultados  en el ioresult
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
// agarro el proceso oprocess.run y lo ejecuto es decir agarro iounput donde esta la ruta y lo
//envia a un conteiner que esta declarado arriba ioResul.
        ioResult = oProcess.run(ioInput);



         // esto coloca los resultados rapid.result
         //exportar resultados


// aqui creo 2 listas. los datos que tengo los meto a mi  lista y de mi lista
//y una vez que tengo llena mi lista hay yo recien voy a transformar en un csv para pasarlo a mi  bd
        List<Attribute> lista1 = new ArrayList<>();


        List<Attribute> lista2 = new ArrayList<>();


        if (ioResult.getElementAt(0) instanceof ExampleSet) {
           resultSet = (ExampleSet) ioResult.getElementAt(0);


           for (Example example : resultSet) {
              Iterator<Attribute> allAtts = example.getAttributes().allAttributes();
              while (allAtts.hasNext()) {
                 Attribute a = allAtts.next();


                 lista1.add(a);
              }
           }
        }
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

```
if (ioResult.getElementAt(1) instanceof ExampleSet) {
   resultSet = (ExampleSet) ioResult.getElementAt(0);

   for (Example example : resultSet) {
      Iterator<Attribute> allAtts = example.getAttributes().allAttributes();
      while (allAtts.hasNext()) {
         Attribute a = allAtts.next();

         lista2.add(a);
      }
   }
}
/**
 * pasando datos lista 1 a archivo csv
 */
//aqui le indico a donde quiero exportarlo
String outputFile = direccion + "/result/result_association_regla.csv";
// me devuelve un falso o verdadero si existe o no
boolean alreadyExists = new File(outputFile).exists();

if (alreadyExists) {
//lo creo
   File ficheroUsuarios = new File(outputFile);
   ficheroUsuarios.delete();
}

try {
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

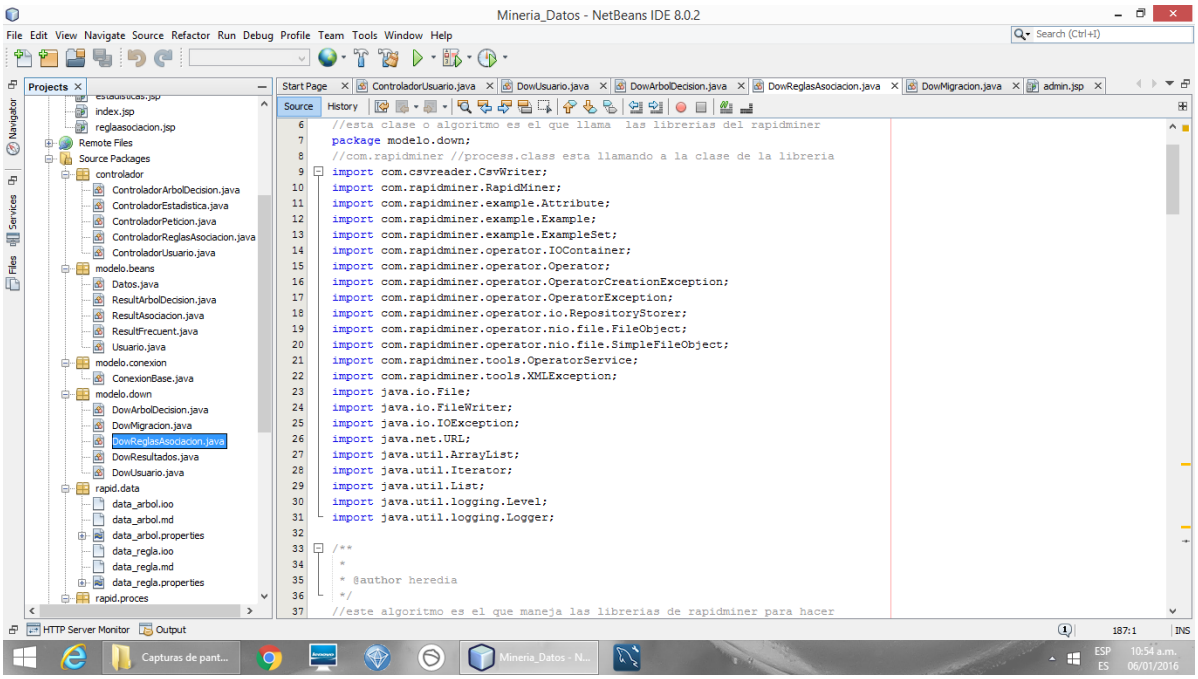
```
// una vez creado le digo que esten separados x coma
        CsvWriter csvOutput = new CsvWriter(new FileWriter(outputFile, true), ',');
// una vez hecho eso todos los datos de mi lista los paso al csv
//hago un for y todo los datos de mi lista lo comienso a pasar al csv y lo convierto en string
        for (int j = 0; j < lista1.size(); j++) {
           csvOutput.write(lista1.get(j).toString());
           // grabo todo los datos que voy pasando
           csvOutput.endRecord();
        }
        // cierra
        csvOutput.close();

    } catch (IOException e) {
    }

    /**
     * pasando datos lista 2 a archivo csv
     */
    //aqui le indico a donde quiero exportarlo
    outputFile = direccion + "/result/result_frecuent_regla.csv";
    alreadyExists = new File(outputFile).exists();

    if (alreadyExists) {
        //lo creo
       File ficheroUsuarios = new File(outputFile);
       ficheroUsuarios.delete();
    }
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
        try {
    //una vez creado le digo que esten separados x coma
        CsvWriter csvOutput = new CsvWriter(new FileWriter(outputFile, true), ',');
// una vez hecho eso todos los datos de mi lista los paso al csv
//hago un for y todos los datos de mi lista lo comienzo a pasar al csv y lo convierto en string
        for (int j = 0; j < lista2.size(); j++) {
          csvOutput.write(lista2.get(j).toString());
          // grabo todo los datos que voy pasando
          csvOutput.endRecord();
        }
        // cierra
        csvOutput.close();
```

Gráfico 59: DowReglaAsociación.java



Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DowMigracion.java

```
package modelo.down;

import java.net.URL;

import modelo.conexion.ConexionBase;


// con esta clase comienzo a migrar los datos a mysql

public class DowMigracion {

//primero migro los datos de arboldesicion.

   public boolean migracionResultArbolDesicion() {


      boolean band;


      try {

         ConexionBase oConexionBase = new ConexionBase();

// primero le indico donde se encuentra csv

         URL url = this.getClass().getResource("/rapid/result/result_arbol.csv");

         String ruta = url.toString().substring(6, url.toString().length());

// si la tabla de mi bd de arbol de decisión estan llena lo borra

         String sql = "DELETE FROM result_arbol_desicion where id<100;";

//ejecuta el sql con la ayuda de la conexion

         oConexionBase.ejecutarExecute(sql);

// con este codigo migro los datos

// primero le doy la ruta a donde esta ubicado

//el INTO TABLE me indica a donde lo voy a importar dentro de la bd hay una tabla que se

//llama result_arbol_desicion

         sql = "LOAD DATA\n"

               + "LOCAL INFILE '" + ruta + "'\n"
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
            + "INTO TABLE result_arbol_desicion\n"
            + "FIELDS TERMINATED BY ',' ENCLOSED BY '\"'\n"
            + "LINES TERMINATED BY '\\n';";
//ejecuta el sql con la ayuda de la conexion
        band = oConexionBase.ejecutarExecute(sql);
//me devuelve una bandera si ejecuto correcta mente me devuelve true si no false
        System.out.println("-->>>" + sql);

      } catch (Exception e) {
         band = false;
      }

      return band;

   }
//primero migro los datos de reglas de asociación.
   public boolean migracionResultReglasAsociación() {

      boolean band;

      try {
         ConexionBase oConexionBase = new ConexionBase();
// primero le indico donde se encuentra csv
         URL url = this.getClass().getResource("/rapid/result/result_frecuent_regla.csv");
         String ruta = url.toString().substring(6, url.toString().length());
// si la tabla de mi bd de result_frecuent estan llena lo borra
         String sql = "DELETE FROM result_frecuent where id<100;";
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

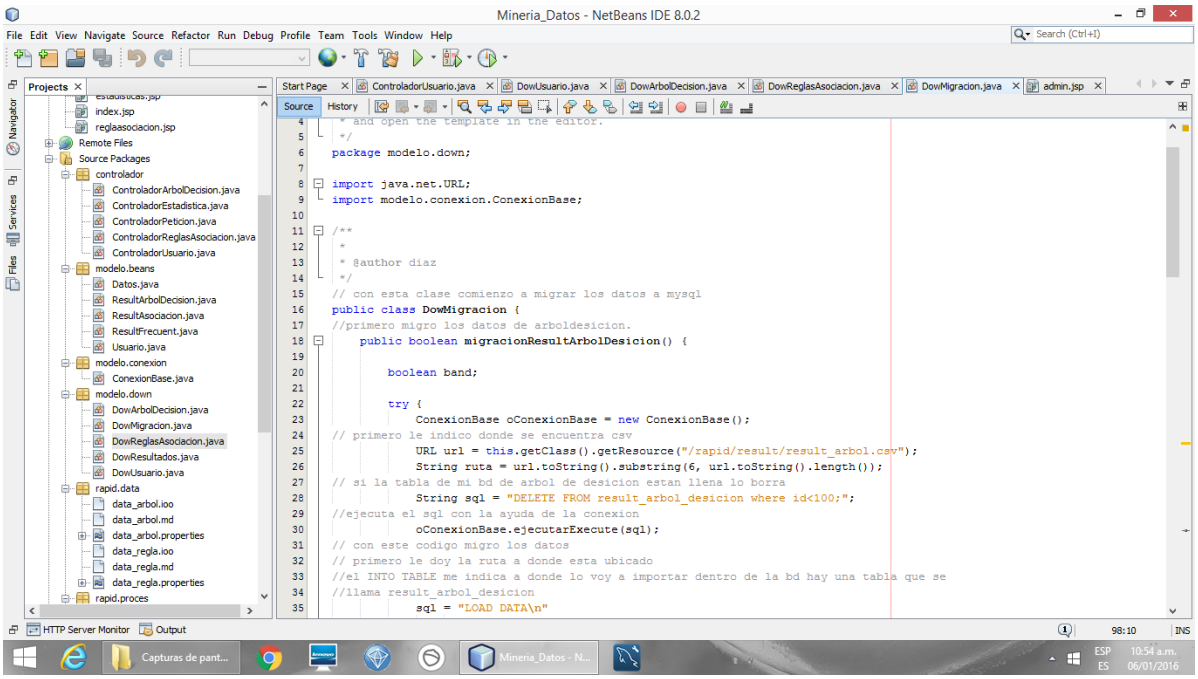
```
//ejecuta el sql con la ayuda de la conexion
        oConexionBase.ejecutarExecute(sql);
// con este codigo migro los datos
// primero le doy la ruta a donde esta ubicado
//el INTO TABLE me indica a donde lo voy a importar dentro de la bd hay una tabla que se
//llama result_frecuent
        sql = "LOAD DATA\n"
            + "LOCAL INFILE '" + ruta + "'\n"
            + "INTO TABLE result_frecuent\n"
            + "FIELDS TERMINATED BY ',' ENCLOSED BY '\"'\n"
            + "LINES TERMINATED BY '\\n';";
//ejecuta el sql con la ayuda de la conexion
        band = oConexionBase.ejecutarExecute(sql);
// primero le indico donde se encuentra csv
        url = this.getClass().getResource("/rapid/result/result_association_regla.csv");
        ruta = url.toString().substring(6, url.toString().length());
// si la tabla de mi bd de result_asociation estan llena lo borra
        sql = "DELETE FROM result_asociation where id<100;";
//ejecuta el sql con la ayuda de la conexion
        oConexionBase.ejecutarExecute(sql);
// con este codigo migro los datos
// primero le doy la ruta a donde esta ubicado
//el INTO TABLE me indica a donde lo voy a importar dentro de la bd hay una tabla que se
//llama result_asociation
        sql = "LOAD DATA\n"
            + "LOCAL INFILE '" + ruta + "'\n"
            + "INTO TABLE result_asociation\n"
```


Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

```
            + "FIELDS TERMINATED BY ',' ENCLOSED BY '\"'\n"
            + "LINES TERMINATED BY '\\n';";
//ejecuta el sql con la ayuda de la conexion
        band = oConexionBase.ejecutarExecute(sql);
//me devuelve una bandera si ejecuto correcta mente me devuelve true si no false
      } catch (Exception e) {
        band = false;
      }

      return band;

   }
}
```

Gráfico 60: DowMigracion.java



Fuente: Elaboración propia.

Tesis publicada con autorización del autor
Algunos Derechos Reservados. No olvide citar esta tesis
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe