IMPLEMENTACIÓN DE LA METODOLOGÍA KDD PARA EL ANÁLISIS DESCRIPTIVO Y PREDICTIVO SOBRE ENFERMEDADES CRÓNICAS (HIPERTENSIÓN ARTERIAL Y DIABETES MELLITUS) PRESENTES EN PACIENTES DE LA ESE HOSPITAL SAN JUAN DE DIOS DE PAMPLONA

Autor

JUAN CAMILO MÉNDEZ FLÓREZ

Universidad de Pamplona

Facultad de Ingenierías y Arquitectura

Ingeniería de Sistemas

Pamplona, Norte De Santander

2021

IMPLEMENTACIÓN DE LA METODOLOGÍA KDD PARA EL ANÁLISIS DESCRIPTIVO Y PREDICTIVO SOBRE ENFERMEDADES CRÓNICAS (HIPERTENSIÓN ARTERIAL Y DIABETES MELLITUS) PRESENTES EN PACIENTES DE LA ESE HOSPITAL SAN JUAN DE DIOS DE PAMPLONA

Autor

JUAN CAMILO MÉNDEZ FLÓREZ

Trabajo presentado para optar por el título de Ingeniero de Sistemas

Director

JOSÉ ORLANDO MALDONADO BAUTISTA

Ph.D. en Ingeniería Informática

Codirector

JAIR CAÑATE CELEDÓN

Ingeniero de Sistemas

Universidad de Pamplona

Facultad de Ingenierías y Arquitectura

Ingeniería de Sistemas

Pamplona, Norte De Santander

2021

Dedicatoria

A Dios por bendecirme la vida y permitirme llegar hasta este momento tan importante en mi formación profesional.

A mi familia, que siempre llevaré en mi corazón y pensamiento; con quien he compartido los mejores momentos de mi vida y que siempre con su esfuerzo y motivación me han ayudado a descubrir el camino para llegar al éxito.

A todos los docentes de la carrera de ingeniería de sistemas de la Universidad de Pamplona, por haber compartido sus conocimientos a lo largo de la preparación de mi profesión, de manera especial al PhD José Orlando Maldonado Bautista, tutor de mi trabajo de grado, igualmente al personal de la ESE Hospital San Juan de Dios de Pamplona por su valioso aporte en este trabajo.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ECUACIONES

Resumen

Este proyecto está basado en la implementación de la metodología KDD cuyo principal objetivo es encontrar conocimiento en un conjunto de datos no procesados.

Por medio de este procedimiento KDD se busca identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles, seleccionando, limpiando, transformando y aplicando técnicas de data Mining a los datos suministrados de enfermedades crónicas (hipertensión arterial y la diabetes mellitus) desde el año 2018 al 2020 de pacientes de la ESE Hospital San Juan de Dios de Pamplona, para un análisis descriptivo y predictivo. Descriptivo permitiendo visualizar el comportamiento de los datos de pacientes a estudiar y un análisis predictivo para hallar información útil y con ella la toma de decisiones pertinentes, siguiendo y cumpliendo las diferentes etapas en el proceso general de esta metodología.

Palabras Clave: KDD, minería de datos, análisis descriptivo, análisis predictivo, enfermedades crónicas.

1. Planteamiento del Problema

1.1 Descripción del Problema y Justificación

En un estudio hecho por McKinsey Global Institute se afirma que el 90% de los datos generados mundialmente en el sector de la salud son comúnmente ignorados, puesto que el hacer un análisis completo de estos datos para convertirlos en información, conlleva mucho tiempo y altos costos.

La ESE Hospital San Juan de Dios de Pamplona es una entidad del estado que ofrece servicios especializados que cubren las necesidades de salud no solo a la población local, sino también a la proveniente de los municipios de Cacota, Chitagá, Cucutilla, Mutiscua, Pamplonita y Silos. Tales municipios se encuentran cercanos a la ciudad de Pamplona; por lo cual cada día ésta institución genera y almacena un gran volumen de datos estructurados y no estructurados, sin llegar a su análisis; por consiguiente, la implementación de la metodología KDD en dicha entidad, mostrará las ventajas que se obtienen al permitir que los datos almacenados tengan una mejor gestión, procesamiento y análisis.

El hospital de Pamplona lleva el registro de cada individuo del programa de pacientes crónicos, en el cual se encuentran diferentes variables como edad, peso, talla, resultados de exámenes de creatinina, hemoglobina glicosilada, colesterol, HDL, LDL, albuminuria, triglicéridos y glicemia, cada uno con sus respectivas fechas. Así, con base en dicha información se puede diagnosticar a un paciente con hipertensión arterial y/o diabetes mellitus.

La implementación de ésta metodología permite la selección, limpieza y transformación de un conjunto de datos para posteriormente, aplicando técnicas o algoritmos de minería de datos, realizar un análisis descriptivo y predictivo. El análisis descriptivo permite sintetizar la información obtenida, analizando su histórico y observando el comportamiento de los datos. El análisis predictivo permite mediante la aplicación de modelos, la detección de patrones y realización de

predicciones. En conjunto dichas herramientas pueden apoyar la toma de decisiones necesarias en patologías crónicas (hipertensión arterial y diabetes mellitus).

1.2 Objetivos

1.2.1 Objetivo General

- Implementar la metodología KDD para el análisis descriptivo y predictivo sobre enfermedades crónicas (hipertensión arterial y diabetes mellitus) presentes en pacientes de la ESE Hospital San Juan de Dios de Pamplona.

1.2.2 Objetivos Específicos

- Obtener la data que se encuentra en las bases de datos del hospital.
- Aplicar la tecnología KDD a los datos obtenidos.
- Emplear técnicas para un análisis descriptivo completo de las enfermedades crónicas hipertensión y diabetes.
- Aplicar técnicas para un análisis predictivo del estado y desarrollo de las enfermedades crónicas (hipertensión arterial y diabetes mellitus).

2. Marco Teórico y Estado del Arte

2.1 Marco Teórico

2.1.1 Minería de Datos

Según (López C. P.), “La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

Igualmente, la minería de datos se podría definir como la construcción de un modelo que ajustado a unos datos proporciona un conocimiento. Por tanto, podemos distinguir dos pasos en una tarea de minería de datos: como primer paso está la elección del modelo, la cual está determinada básicamente por dos condicionantes: el tipo de los datos y el objetivo que se quiera obtener. Y como segundo paso el ajuste final del modelo a los datos, que consiste en realizar una fase de aprendizaje con los datos disponibles.

También se puede determinar que la disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos, orientándose hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining”.

2.1.2 Tipos de Análisis

2.1.2.1 Análisis Descriptivo

“La analítica descriptiva consiste en almacenar y realizar agregaciones de datos históricos”, según el instituto de ingeniería del conocimiento, en este caso se trabaja con datos de pacientes diagnosticados con enfermedades crónicas (diabetes mellitus e hipertensión arterial) desde el año 2018 al 2020 de la E.S.E Hospital San Juan de Dios; donde se aplicarán distintas estrategias de

visualización para resumir el estado de los datos, de forma que, puedan ayudar a la comprensión del estado actual y pasado, permitiendo detectar, visualizar y observar la evolución de los datos.

En general, cualquier tipo de datos puede presentarse de forma intuitiva mediante el uso de técnicas de agregación y visualización de la información. (Instituto de ingenieria del conocimiento)

2.1.2.2 Análisis Predictivo

"La analítica predictiva proporciona herramientas para estimar aquellos datos que son desconocidos o inciertos, o que requieren de un proceso manual o costoso para su obtención. Más allá del puro análisis de la información histórica que realiza la analítica descriptiva las predicciones de datos que realiza la analítica predictiva fortalecen las conclusiones que se van a obtener" según el instituto de ingeniería del conocimiento, en este caso, basadas en estas dos enfermedades a estudiar

"Para estimar la información que no se conoce, la analítica predictiva utiliza una serie de técnicas, entre ellas tenemos:

- Clasificación automática de la información que permite apoyar y facilitar la labor de expertos al analizar mayor volumen de información a menos coste.
- Técnicas de predicción: el fin de esta práctica es facilitar la integración del modelo de predicción en el proceso de negocio, además de las predicciones, el modelo puede generar índices de variabilidad o intervalos de confianza de estas predicciones, informando así no solo del valor más probable sino también de la volatilidad esperada", así lo define el (Instituto de ingenieria del conocimiento)

2.1.3 Aprendizaje Automático

2.1.3.1 Aprendizaje Supervisado

Una de las modalidades que tiene el machine learning es la de aprendizaje supervisado. “Usando esta modalidad, se entrena al algoritmo para que aprenda de los datos introducidos, ingresando manualmente las preguntas, denominadas características, y las respuestas, denominadas etiquetas. Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones.

Existen, a su vez, dos tipos de aprendizaje supervisado:

- Regresión: tiene como resultado un número específico. Si las etiquetas suelen ser un valor numérico, mediante las variables de las características, se pueden obtener dígitos como dato resultante.
- Clasificación: en este tipo, el algoritmo encuentra diferentes patrones y tiene por objetivo clasificar los elementos en diferentes grupos”, según (Zambrano, 2018)

2.1.3.2 Aprendizaje No Supervisado

“A diferencia del aprendizaje supervisado, en el no supervisado solo se le otorgan las características, sin proporcionarle al algoritmo ninguna etiqueta. Su función es la agrupación, por lo que el algoritmo debería catalogarse por similitud y poder crear grupos, sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes del grupo.

Cualquier programa que aplique machine learning puede simplificar trabajos de bases de datos y, por tanto, ahorrarles a muchos empleados centenares de horas de trabajo. Además, se está involucrando activamente la automatización cognitiva, que incluye a imágenes y documentos no estructurados, lo que amplía aún más las capacidades de agrupación, clasificación y regresión”, dicho por (Zambrano, 2018)

2.1.4 Algoritmos

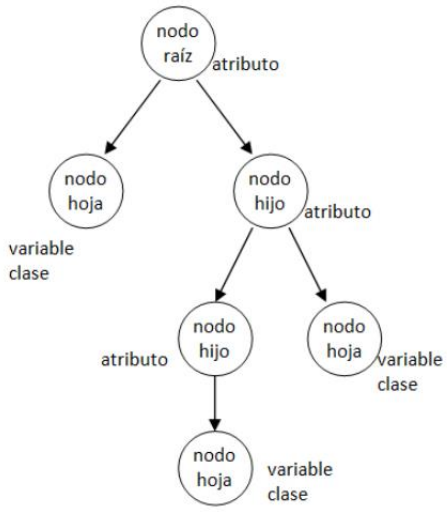
2.1.4.1 Árbol de Decisión

Para (Microsoft, 2018), “El algoritmo de árboles de decisión es un algoritmo de clasificación y regresión 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 estas columnas para predecir los estados de una columna que se designa como elemento de predicción.

Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El árbol se representa por nodos, donde el nodo principal raíz es el atributo a partir del cual se inicia el proceso de clasificación. Los nodos internos o nodos hijos son preguntas acerca del atributo o problema. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver”.

Figura 1 *Representación de la estructura de un árbol de decisión*



Nota: Tomado de (Barrientos Martínez, y otros, 2019)

¿Cómo funciona el algoritmo?

Para (Barrientos Martínez, y otros, 2019), “Un algoritmo de generación de árboles de decisión consta de 2 etapas: la primera corresponde a la inducción del árbol y la segunda a la clasificación.

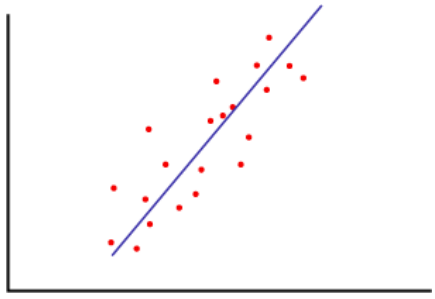
En la primera etapa se construye el árbol de decisión a partir del conjunto de entrenamiento; La construcción del árbol inicia generando su nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de entrenamiento en dos o más subconjuntos; para cada partición se genera un nuevo nodo y así sucesivamente.

En la segunda etapa del algoritmo cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja. El camino a seguir en el árbol lo determinan las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente en él”.

2.1.4.2 Regresión Lineal

Según (Microsoft, 2018), “El algoritmo de regresión lineal 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, utilizar esa relación para la predicción. La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos.

Figura 2 *Regresión Lineal*



Nota: Tomado de (Microsoft, 2018)

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".

¿Cómo funciona el algoritmo?

Según Joaquín Amat Rodrigo, "La regresión lineal simple consiste en generar un modelo de regresión o ecuación de una recta que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente se le identifica como Y y a la variable predictora o independiente como X. El modelo de regresión lineal simple se describe de acuerdo a la ecuación:

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + \epsilon$$

Siendo $\alpha_{0\ldots n}$ los coeficientes, $x_{1\ldots n}$ las variables predictoras y ϵ el error aleatorio. Este último representa la diferencia entre el valor ajustado por la recta y el valor real".

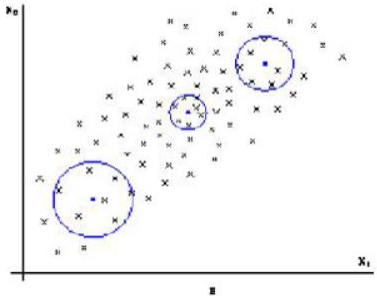
(Rodrigo, 2016)

2.1.4.3 K-NN

Según el artículo escrito por (Rodríguez Rodríguez, Rojas Blanco, & Franco Camacho), "El método de los k-vecinos o k-NN es un método retardado y supervisado (pues su fase de entrenamiento se hace en un tiempo diferente al de la fase de prueba) cuyo argumento principal es

la distancia entre instancias. El método básicamente consiste en comparar la nueva instancia a clasificar con los datos k más cercanos conocidos, y dependiendo del parecido entre los atributos el nuevo caso se ubicará en la clase que más se acerque al valor de sus propios atributos”

Figura 3 *k-NN*



Nota: Tomado de (García Cambronero & Gómez Moreno)

¿Cómo funciona el algoritmo?

(García Cambronero & Gómez Moreno), describe el funcionamiento del algoritmo así: “Se tienen tres conjuntos de datos, el conjunto de entrenamiento que se utiliza para el aprendizaje del clasificador y los conjuntos de validación y de test para comprobar si el clasificador es capaz de generalizar, es decir si presenta buenos resultados.

El proceso de aprendizaje de este clasificador consiste en almacenar en un vector el conjunto de entrenamiento, junto a la clase asociada a cada muestra de este conjunto. En primer lugar, y con motivo del aprendizaje del algoritmo, calcula la distancia euclidiana de cada muestra de entrenamiento, a todas las demás que están almacenadas en el vector y de las que conocemos la clase a la que corresponden, quedando las K muestras más cercanas y clasificando la nueva muestra de entrenamiento en la clase más frecuente a la que pertenecen los K vecinos obtenidos anteriormente. La segunda tarea para diseñar el clasificador, realiza el mismo proceso con los datos

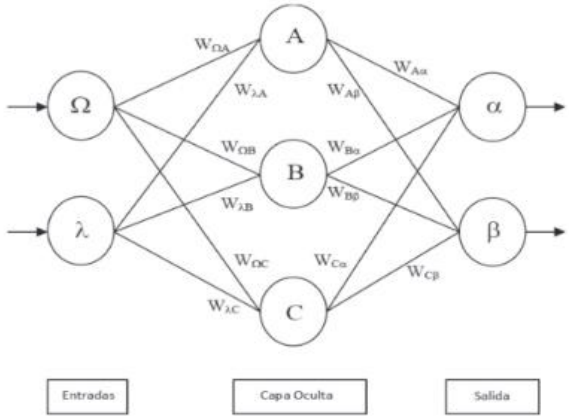
de validación, calcula el porcentaje de clasificación sobre los ejemplos de este conjunto (desconocidos en la tarea de aprendizaje) para conocer su poder de generalización".

2.1.4.4 Redes Neuronales (MLP)

(Mercado Polo, Pedraza Caballero, & Martínez Gómez, 2015) en su master define las redes neuronales de la siguiente manera: "Las Redes Neuronales Artificiales (RNA) son sistemas de procesamiento de la información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. Consisten en un conjunto de elementos simples de procesamiento llamados nodos o neuronas conectadas entre sí por conexiones que tienen un valor numérico modificable llamado peso.

Se distinguen tres tipos de capas: de entrada, de salida y ocultas. Una capa de entrada, está compuesta por neuronas que reciben datos o señales procedentes del entorno. Una capa de salida se compone de neuronas que proporcionan la respuesta de la red neuronal. Una capa oculta no tiene una conexión directa con el entorno. Este tipo de capa oculta proporciona grados de libertad a la red neuronal gracias a los cuales es capaz de representar más fehacientemente determinadas características del entorno que trata de modelar".

Figura 4 *Red Neuronal Perceptrón Multicapa*



Nota: Tomado de (Larrañaga, Inza, & Moujahid)

¿Cómo funciona el algoritmo?

Para (Larrañaga, Inza, & Moujahid) explica el funcionamiento del algoritmo así: “Para conseguir que una red neuronal realice las funciones deseadas, es necesario entrenarla. El entrenamiento de una red neuronal se realiza modificando los pesos de sus neuronas para que consiga extraer los resultados deseados. Para ello lo que se hace es introducir datos de entrenamiento en la red, en función del resultado que se obtenga, se modifican los pesos de las neuronas según el error obtenido y en función de cuanto haya contribuido cada neurona a dicho resultado. Este método es conocido como Backpropagation o propagación hacia atrás. Con este método se consigue que la red aprenda, consiguiendo un modelo capaz de obtener resultados muy acertados incluso con datos muy diferentes a los que han sido utilizados durante su entrenamiento”.

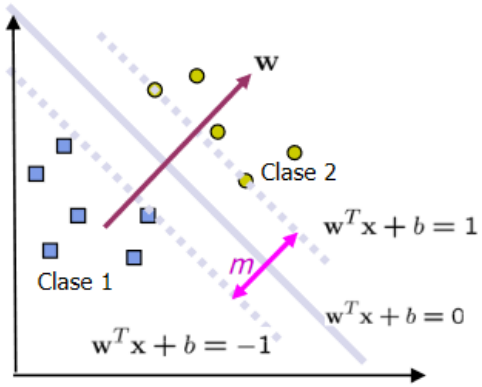
2.1.4.5 Vectores de Soporte (SVM)

Para (Betancurt) en su maestría define a la SVM y su funcionamiento así: “Una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento.

¿Cómo funciona el algoritmo?

Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor y encuentra un hyperplano que los separe y maximice el margen m entre las clases en este espacio como se muestra en la Figura.

Figura 5 *Maquina de soporte Vectorial (SVM)*



Nota: Tomado de (Betancurt)

Maximizar el margen *m* es un problema de programación cuadrática y puede ser resuelto por su problema dual introduciendo multiplicadores de Lagrange. Sin ningún conocimiento del mapeo, la SVM encuentra el hyperplano óptimo utilizando el producto punto con funciones en el espacio de características que son llamadas kernels. La solución del hyperplano óptimo puede ser escrita como la combinación de unos pocos puntos de entrada que son llamados vectores de soporte".

2.1.4.6 Reglas de Asociación

"Son una manera muy popular de expresar patrones de datos en una base de datos. Estos patrones pueden servir para conocer el comportamiento general del problema que genera una base de datos, y de esta manera, se tenga más información que pueda asistir en la toma de decisiones.

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 cómo estos elementos se agrupan dentro de los casos" según (Rueda, 2020).

Figura 6 *Reglas de Asociación*

```
Association Rules
[Colesterol Total] --> [LDL, Trigliceridos] (confidence: 0.423)
[LDL] --> [Colesterol Total, Trigliceridos] (confidence: 0.523)
[LDL] --> [Trigliceridos] (confidence: 0.532)
[Trigliceridos] --> [Colesterol Total, LDL] (confidence: 0.535)
[Trigliceridos] --> [LDL] (confidence: 0.544)
[Colesterol Total, LDL] --> [Trigliceridos] (confidence: 0.558)
[Colesterol Total] --> [Trigliceridos] (confidence: 0.584)
[Colesterol Total, Trigliceridos] --> [LDL] (confidence: 0.725)
[Trigliceridos] --> [Colesterol Total] (confidence: 0.737)
[Colesterol Total] --> [LDL] (confidence: 0.759)
[LDL] --> [Colesterol Total] (confidence: 0.937)
[LDL, Trigliceridos] --> [Colesterol Total] (confidence: 0.983)
```

Nota: Elaborado por el autor

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 genera un gran número de itemsets candidatos y lo hace compactando la base de datos utilizando un árbol FP. La base de datos compactada se divide en bases de datos condicionales y se obtienen de ellas los itemsets frecuentes, sin la generación de candidatos y con solo dos barridos a la base de datos. El primer barrido determina el soporte de cada item para eliminar los no frecuentes. Los items frecuentes mantenidos se ordenan en orden descendiente de soporte en

cada transacción. En un segundo barrido se analiza cada transacción para generar el árbol FP. Al terminar este paso, se generan las reglas de asociación a partir de los itemsets frecuentes obtenidos en el paso anterior". (Martínez, 2020)

2.1.5 Índices de Rendimiento

2.1.5.1 Error Absoluto

RapidMiner lo define como la desviación absoluta promedio de la predicción del valor real, por lo tanto, la desviación media absoluta (MAD) mide la dispersión del error de pronóstico o, dicho de otra forma, la medición del tamaño del error en unidades. Es el valor absoluto de la diferencia entre el valor real y el pronóstico, dividido sobre el número de periodos.

Ecuación 1 *Error Absoluto (MAD)*

$$MAD = \frac{\sum |Real - Pronóstico|}{n}$$

(Betancourt, 2016)

2.1.5.2 Error Relativo

El error relativo promedio es el promedio de la desviación absoluta de la predicción del valor real dividido por el valor real, según RapidMiner. Por lo tanto, Error porcentual medio absoluto (MAPE) entrega la desviación en términos porcentuales y no en unidades como las anteriores medidas. Es el promedio del error absoluto o diferencia entre la demanda real y el pronóstico, expresado como un porcentaje de los valores reales.

Ecuación 2 *Error Relativo (MAPE)*

$$MAPE = \frac{\frac{\sum_{i=1}^{n} |Real - Pronóstico|}{Real_i}}{n} \times 100$$

(Betancourt, 2016)

2.1.5.3 R^2

"El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar. El resultado del coeficiente de determinación oscila entre 0 y 1.

Ecuación 3 *Coeficiente de Determinación (R^2)*

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(Y_t - \bar{y})^2}$$

En la expresión anterior en el numerador de refiere a la expresión de la varianza, donde ŷ es la estimación de un modelo sobre lo que según las variables explicativas vale Y, y el denominador la única diferencia existente respecto a la fórmula original de la varianza es la ausencia de su denominador. Es decir, no dividimos entre N." para (López J. F., 2017)

2.1.5.4 Soporte y Nivel de Confianza

Para (Rodrigo, Reglas de asociación y algoritmo Apriori con R, 2018), "Una regla necesita un soporte de varios cientos de registros 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 en una base de datos se define como el número de transacciones que contienen X dividido entre el total de transacciones":

Ecuación 4 *Soporte (sop(x))*

$$sop(X) = \frac{|X|}{|D|}$$

Y La confianza es la eficiencia de una regla "Si X entonces Y" se define acorde a la ecuación:

Ecuación 5 *Nivel de Confianza (conf (X⇒Y))*

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

2.1.6 Enfermedades Crónicas

"Las enfermedades crónicas son afecciones de larga duración y por lo general de progresión lenta. Toda enfermedad que tenga una duración mayor a seis meses puede considerarse como crónica, éstas son de curso prolongado que necesitan tratamientos continuos para su control y no se resuelven espontáneamente.

Las principales enfermedades crónicas son: enfermedad cardiovascular, cáncer, enfermedades respiratorias crónicas, la diabetes e hipertensión arterial. Los factores de riesgo para estas enfermedades son consumo excesivo de alcohol, aumento en la glucosa sanguínea, aumento en el colesterol sanguíneo, sobrepeso/obesidad, tabaquismo y sedentarismo.

Las consecuencias a nivel individual de padecer estas enfermedades abarcan diferentes ámbitos, éstas limitan la calidad de vida a nivel físico, psicológico, económico y en el ámbito social", según (Guillén, 2019).

En las siguientes tablas se muestran los exámenes que el Hospital San Juan de Dios de Pamplona utiliza para diagnosticar a un paciente con una enfermedad crónica, estos son: creatinina, hemoglobina glucosilada, colesterol total, HDL, LDL, albúmina, triglicéridos, triglicéridos y glicemia catalogados con el nivel normal, alto o muy alto respectivamente.

Tabla 1 *Niveles de Creatinina Sérica*

Creatinina	
Moderadamente bajos	Entre 0 - 0.69 mg/dl
Normal	Entre 0.7 a 1.18 mg/dl
Ligeramente Elevado	Entre 1.19 a 1.75 mg/dl
Moderadamente Elevado	Entre 1.76 a 3.5 mg/dl
Excesivamente elevado	Entre 3.51 a 6.54 mg/dl

Nota: Adaptada de (Salazar GA)

Tabla 2 *Niveles de la Hemoglobina Glicosilada para personas diabéticas y no diabéticas*

Hemoglobina Glicosilada		
	Persona Diabética	**Persona No Diabética**
Normal	Entre 6,5% a 7,0%	Entre 4,0 a 5,6%
Controlado	Entre 7,0% a 7,9%	Entre 6,5 a 7,0%
Prediabetes	-	Entre 5,7 a 6,4%
Diabetes No controlada	Mas de 8,0%	-

Nota: Adaptada de (Diabetes, 2016)

Tabla 3 *Niveles del Colesterol Total*

Colesterol Total	
Aceptable	Menos de 170 mg/dl
Limite Alto	Entre 170 a 200 mg/dl
Alto	Más de 200 mg/dl

Nota: Adaptada de (Sampedro, 2013)

Tabla 4 *Niveles de HDL*

HDL	
Bajo	Menos de 40 mg/dl
Normal	Entre 40 y 60 mg/dl
Beneficioso	Más de 60 mg/dl

Nota: Adaptada de (L., 2016)

Tabla 5 *Niveles de LDL*

LDL	
Normal	Menos de 100 mg/dl
Limite Alto	Entre 100 a 129 mg/dl
Alto	Entre 130 a 190 mg/dl
Muy alto	Más de 190 mg/dl

Nota: Adaptada de (L., 2016)

Tabla 6 *Niveles de Albumina*

Albumina	
Normal	Menos de 20 mg/l
Microalbumina	Entre 20 a 200 mg/l
Macroalbumina	Más de 200 mg/l

Nota: Adaptada de (Ibáñez, 2016)

Tabla 7 *Niveles de los Triglicéridos*

Triglicéridos	
Normal	Menos de 150 mg/dl
Limite alto	Entre 150 a 199 mg/dl
Alto	Entre 200 a 500 mg/dl
Muy Alto	Más de 500 mg/dl

Nota: Adaptada de (L., 2016)

Tabla 8 *Niveles de la Glicemia*

Glicemia	
Hipoglucemia	Menos de 70 mg/dl
Normal	Entre 70 a 120 mg/dl
Alto	Entre 160 a 190 mg/dl
Muy Alto	Más de 215 mg/dl

Nota: Adaptada de (Huizen, 2019)

Tabla 9 *Niveles de la tasa de filtración glomerular (TFG)*

Tasa de filtración glomerular (TFG)	
Estadio I	Mayor de 90 ml/minuto
Estadio II	Entre 59 a 89 ml/minuto
Estadio III	Entre 30 a 59 ml/minuto
Estadio IV	Entre 15 a 29 ml/minuto
Estadio V	Menos de 15 ml/minuto

Nota: Adaptada de (Liendo, 2018)

2.1.7 Hipertensión Arterial

Según (Organización Mundial de la Salud, 2019), "La tensión arterial es la manifestación de un problema circulatorio que se caracteriza por una elevación permanente por la presión de la sangre. La presión de la sangre es la fuerza que hace esta sobre las paredes de las arterias al ser bombeada por el corazón. Cuando una persona sufre de presión alta, su corazón tiene que hacer más fuerza y se debilita. Las complicaciones al no estar controlada esta enfermedad van desde un dolor torácico, infartos, hasta un ritmo cardiaco irregular que puede conllevar a una muerte súbita.

Se estima que en el mundo hay 1130 millones de personas con hipertensión y la mayoría de ellas (cerca de dos tercios) vive en países de ingresos bajos y medianos. En 2015, 1 de cada 4 hombres y 1 de cada 5 mujeres tenían hipertensión.

En Colombia, según la Cuenta de Alto Costo, reportó en el año 2018, 4.048.776 personas diagnosticadas con hipertensión arterial, de estos el 63,3 % pertenecían al régimen contributivo y el 35,8 % al régimen subsidiado. Para ese mismo año, se reportaron más de 317.000 casos nuevos de hipertensión arterial, de los cuales el 61,9 % de los casos se presentó en personas entre 50 y 75 años, y el 3,6 % en menores de 35 años".

2.1.8 Diabetes Mellitus

"La diabetes es una enfermedad crónica que aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce. La insulina es una hormona que regula el azúcar en la sangre" concepto dado por (Organización Mundial de la Salud, 2019). Esta enfermedad puede ser controlada y manejada para evitar complicaciones como ceguera, problemas en el riñón, ataques cardíacos, derrames cerebrales, infecciones, daños en los nervios, problemas en los pies, entre otros. Por lo tanto, es importante aumentar el acceso al diagnóstico, la educación para el autocuidado y la accesibilidad al tratamiento.

"En Latinoamérica y el Caribe las tasas más elevadas de prevalencia de la diabetes corresponden a Belice (12,4%) y México (10,7%). Las ciudades de Managua, Ciudad de Guatemala y Bogotá mantienen tasas de alrededor del 8 al 10%. En América Central y Suramérica la cifra de afectados alcanza los 29,6 millones de personas, lo que vale decir, el 9,4% de la población adulta. Se estima que para el 2040, el número de personas con diabetes se incrementará un 65% en esta Región", estadística dada por (Salud, 2020)

2.2 Estado del Arte

2.2.1 Internacional

- **Perú**: Técnicas de minería de datos para predicción del diagnóstico de hipertensión arterial (Trabajo de Grado). El objetivo de este trabajo fue encontrar patrones y relacione 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 de 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. (Avendaño, 2016)

- **Reino Unido**: Diversos proyectos en curso en el sector salud, tales como caredata.co.uk y 100,000 Genomes Project (el proyecto de los cien mil genomas), están ayudando a entender, diagnosticar y tratar enfermedades, incluyendo el cáncer. También tienen iniciativas de salud incluyen el uso de información para cambiar prácticas, mejorar la prestación de servicios (perspectiva orientada al usuario) y enfocarse en la calidad. (Viviana Cañón, 2017) (Sterckx, 2018)

- **Corea del Sur**: En enero de 2014, el Ministerio de Ciencia, TIC y Planificación del Futuro de Corea del Sur (MSIP) y el NIA lanzaron el programa de Consultoría de Información Médica. “El programa sugiere un servicio de minería de datos que puede ayudar a diagnosticar y personalizar el tratamiento para los pacientes, lo que ayudará a promover la salud pública y agilizar la gestión de las instalaciones médicas. El objetivo final del programa es, naturalmente, reunir los datos médicos recopilados

  MediLatte es una aplicación para el servicio personalizado de información de hospital desarrollado por AD Ventures que comenzó su servicio en octubre de 2012. Utilizando datos abiertos sobre la información del hospital (por ejemplo, ubicación, categoría, número de médicos, tiempo de operación, por ejemplo, revisiones), MediLatte ofrece recomendaciones de servicios médicos personalizados. Su fuente de datos abierta es la información del hospital proporcionada por el Servicio Nacional de Seguros de Salud y los gobiernos locales. En esta plataforma de mercadeo hospitalario, los pacientes reciben

cupones beneficiados y citas rápidas eligiendo hospitales asociados con MediLatte, y los hospitales han reducido los costos de comercialización para atraer a más pacientes a través de las asociaciones de MediLatte.” (Viviana Cañón, 2017)

- **USA**: e-salud: Hay dos soluciones clave en materia de salud: los registros de salud y los registros de las prescripciones. E-records es un sistema que utiliza la tecnología blockchain e integra los datos de los proveedores de salud de Estonia para crear un registro común que todos los pacientes pueden acceder en línea (a través del portal e-Patient). Los médicos pueden acceder a los registros de sus pacientes y leer los resultados de las pruebas. El sistema también recopila datos para las estadísticas nacionales. (Viviana Cañón, 2017)

2.2.2 Nacional

- **Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia (Trabajo de Grado)**: El presente proyecto se basa en la aplicación de minería de datos mediante el algoritmo de clustering K- means que permita la generación de un modelo descriptivo con el análisis de los datos y con el objetivo de identificar posibles comportamientos en enfermedades respiratorias en la ciudad de Bogotá. El conjunto de clústeres generados por la herramienta RapidMiner es la recopilación de datos de un periodo de cinco años de 2012 a 2016, en donde se contemplan el número de casos asociados a 184 diagnósticos de enfermedades respiratorias y la edad de los pacientes corresponde de 0 a 5 años. (Aguilar, 2017)

- **Minería de datos de salud: estudio de los factores personales, familiares y vivienda que influyen en las enfermedades de diabetes e hipertensión a partir de la encuesta de atención primaria en salud del area metropolitana del Valle de Aburrá. (Trabajo de Grado):** La encuesta de Atención Primaria en Salud dada por la OMS se ha convertido en un instrumento vital para conocer los principales factores que influyen en la salud pública y en los determinantes para definir la inversión en salud preventiva. Gracias al Área Metropolitana esta encuesta se recolecta desde mayo de 2015 en dispositivos electrónicos: Tablet, y los datos son almacenados en la nube para su posterior procesamiento y análisis. Este proyecto tiene como fin aplicar técnicas de minería de datos a la encuesta de Atención Primaria en Salud para establecer los factores personales, familiares y habitacionales, que influyen en las enfermedades crónicas: Hipertensión y Diabetes en el Área Metropolitana del Valle de Aburra. (MARLON BORJA PUERTA, 2017)

3. Metodología (KDD)

Para (Fayyad, Piatetsky-Shapiro, & Padhra), “El término descubrimiento de conocimientos en bases de datos o KDD, se refiere al amplio proceso de encontrar conocimiento en los datos y enfatiza la aplicación de alto nivel en métodos particulares de minería de datos. Es de interés para investigadores en aprendizaje automático, reconocimiento de patrones, bases de datos, estadísticas, inteligencia artificial, adquisición de conocimiento para sistemas expertos y visualización de datos.

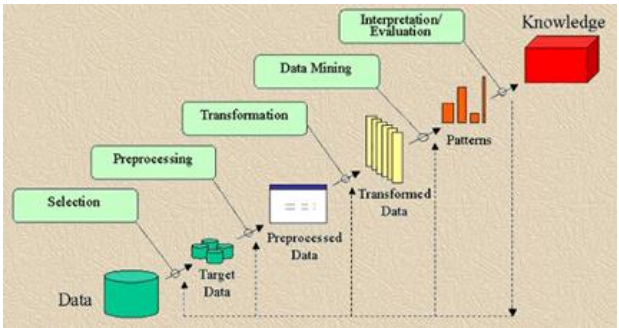
El objetivo unificador del proceso KDD es extraer conocimiento de los datos en el contexto de grandes bases de datos.

Para ello, utiliza métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera conocimiento, de acuerdo con las especificaciones de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformaciones requeridos de esa base de datos”.

3.1 Pasos del Proceso KDD

KDD es un proceso metodológico y además secuencial que se sigue para encontrar conocimiento en un conjunto de datos en bruto. Estos pasos se dividen en 9 que son:

Figura 7 *Ciclo de vida de la metodología KDD*



Nota: Tomado de (Fayyad, Piatetsky-Shapiro, & Padhra)

Según (Landa, 2016) describe cada uno de los pasos de la metodología KDD de la siguiente manera:

1. **Abstracción del escenario**: Es importante conocer las propiedades, limitaciones y reglas del escenario en estudio, para posteriormente definir las metas a alcanzar.
2. **Selección de datos**: Del conjunto de datos recolectados y ya definidos los objetivos por alcanzar, se deben elegir datos disponibles para realizar el estudio e integrarlos en uno solo que puedan favorecer a llegar a alcanzar a los objetivos del análisis.
3. **Limpieza y pre-procesamiento**: Determina la confiabilidad de la información, es decir, realizar tareas que garanticen la utilidad de los datos. Para esto se hace la limpieza de datos (tratamiento de datos perdidos o remover valores atípicos). Esto implica eliminar variables o atributos con datos faltantes o eliminar información no útil para este tipo de tareas.
4. **Transformación de los datos**: En esta etapa se mejora la calidad de los datos con transformaciones que involucran ya sea reducción de dimensionalidad (disminuir la cantidad de variables del conjunto de datos) o bien transformaciones como por ejemplo convertir los valores que son números a categóricos (discretización).
5. **Elección de tareas de Minería de Datos**: Fase en la que se refiere a elegir el paradigma apropiado de Minería de Datos, ya sea la clasificación, regresión o agrupación
6. **Elección del algoritmo**: Posteriormente se procede a seleccionar la técnica o algoritmo, o incluso más de uno para la búsqueda del patrón y obtener conocimiento.
7. **Aplicación del algoritmo**: Una vez seleccionado las técnicas el paso siguiente es aplicarlo a los datos ya seleccionados, limpiados y procesados. Es posible que la ejecución de los algoritmos sean varias intentando ajustar los parámetros que optimicen los resultados.

8. **Evaluación e interpretación**: Una vez aplicado los algoritmos al conjunto de datos, procedemos a evaluar los patrones que se generaron y el rendimiento que se obtuvo para verificar que cumpla con las metas planteadas en las primeras fases.
9. **Entendimiento del conocimiento**: La última etapa es simplemente aplicar el conocimiento encontrado al contexto y comenzar a resolver sus problemáticas.

3.2 Herramientas

Las herramientas que se utilizaran para el desarrollo del proceso de la metodología KDD son:

3.2.1 OpenRefine

OpenRefine (anteriormente Google Refine) es una poderosa herramienta para trabajar con datos desordenados: limpiarlos; transformándolo de un formato a otro; y ampliándolo con servicios web y datos externos. (OpenRefine, s.f.)

Figura 8 *Entorno de OpenRefine*

Nota: Elaborado por el autor

Algunas de sus características principales son:

- OpenRefine funciona con archivos locales o datos de direcciones web en varios formatos de archivo, incluidos CSV, TSV, XLS, XML y otros.

- Tiene la capacidad de filtrar o buscar ciertos elementos que deben cambiarse de alguna manera.
- Puede encontrar entradas duplicadas, celdas vacías, variaciones de entrada, inconsistencias y patrones de errores. (Ham, 2013)

3.2.2 Python

"Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. La sintaxis simple y fácil de aprender de Python enfatiza la legibilidad y, por lo tanto, reduce el costo de mantenimiento del programa. El intérprete de Python y la extensa biblioteca estándar están disponibles en formato fuente o binario sin cargo para todas las plataformas principales y se pueden distribuir libremente; también Python es un lenguaje de programación versátil multiplataforma y multiparadigma que se destaca por su código legible y limpio. Su objetivo es la automatización de procesos para ahorrar tanto complicaciones como tiempo". (Python, s.f.)

Para (Universia, 2020) Python se utiliza en el campo de:

- "Ciencia de los datos. El poder de las bibliotecas Python desarrolladas para el análisis y visualización de datos es asombroso. Con una biblioteca de visualización de datos de Python, puede crearse una amplia variedad de gráficos y representaciones visuales de todo tipo.
- Aprendizaje automático. Python es una herramienta esencial para todos los desarrolladores que quieran sumergirse en el campo del machine learning.
- Desarrollo web. Python se utiliza en el campo del desarrollo web para construir el back-end de aplicaciones web.

- Educación en Ciencias de la Computación. Python se usa ampliamente como herramienta de enseñanza porque es fácil de aprender: su sintaxis es simple y se puede aprender rápidamente".

3.2.3 RapidMiner

"RapidMiner es un entorno para el aprendizaje automático y para procesos de Minería de Datos, bajo el concepto de operador modular permite el diseño de cadenas de operadores complejos anidados para un gran número de problemas de aprendizaje. Hoy en día, RapidMiner es el líder mundial en soluciones de Minería de Datos con código abierto y es ampliamente utilizado por los investigadores y las empresas ya que cuenta con interfaces claras y una especie de lenguaje de script basado en XML lo que la convierte en un entorno de desarrollo integrado para la minería de datos y el aprendizaje automático.

RapidMiner proporciona más de 400 Operadores incluyendo los siguientes:

- Algoritmos de aprendizaje máquina: Cuenta con un gran número de esquemas de aprendizaje para la regresión y las tareas de clasificación.
- Operadores de preprocesamiento de datos: Cuenta con varios operadores como son los de discretización, función de filtrado, la reposición de valores faltantes, normalización, toma de muestras, entre otras.
- De evaluación de la ejecución: Cuenta con la validación cruzada y otros esquemas de evaluación, criterios de rendimiento para clasificaciones y regresiones.
- Visualización: Cuenta con operadores para el registro y la presentación de los resultados. Crear en línea gráficos en 2D y 3D de los datos, modelos aprendidos y los resultados de otros procesos". Según (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006)

4. Obtención de los Datos

Esta es la etapa donde los datos relevantes para el análisis son extraídos de la base de datos propia de la ESE Hospital San Juan de Dios de Pamplona; en este caso se cuenta con los datos recolectados de 3492, de pacientes adscritos a la sección de crónicos de dicho hospital. Estos datos fueron recolectados a partir de exámenes médicos hechos por las EPS correspondientes a cada paciente en los años 2018, 2019 y 2020; esta información se encuentra almacenada en un archivo de Microsoft Excel.

Figura 9 *Base de Datos de la ESE San Juan de Dios de Pamplona sección crónicos*

Diagnóstico de Hipertensión Arterial (HTA)	Fecha Diagnostico de la HTA	Diagnostico Diabetes Mellitus (DM)	Fecha de Diagnostico de la DM	Tipo de Diabetes	Fecha de Ultima Atencion	Tension arterial (Sistolica)	Tension Arterial (Diastolica)	Peso	Talla	IMC	Estado Nutricional	Resultado de Creatinina	Fecha de Ultima Creatinina	Result Hemog Glicos
SI	10/05/2017	SI	10/05/2017	TIPO II NO I	4/10/2017	123	76	51	145	24,26	Normal (18,50	0,44	12/06/2018	
SI	6/08/2012	SI	24/11/2000	TIPO II INSU	30/04/2018	150	70	67	164	24,91	Normal (18,50	1,52	16/01/2018	
SI	22/09/2009	NO			7/05/2018	140	70	71	166	25,77	Preobeso (25	1,62	27/03/2018	
SI	21/10/2017	SI	21/10/2017	TIPO II NO I	21/10/2017	130	85	62	160	24,22	Normal (18,50	0,76	17/08/2017	
SI	29/10/2014	SI	29/09/2014	TIPO II NO I	24/04/2018	120	75	65	150	28,89	Preobeso (25	0,51	24/04/2018	
SI	1/01/2013	SI	1/01/2013	TIPO II NO I	3/09/2018	120	75	95	165	34,89	Obeso Tipo 1 (	0,6	21/02/2018	
NO	17/08/2016	SI	5/05/2011	TIPO II NO I	3/09/2018	190	110	76	168	26,93	Preobeso (25	0,7	21/05/2018	
SI	15/01/2010	SI	15/01/2010	TIPO II NO I	18/10/2017	130	83	66	150	29,33	Preobeso (25	0,9	23/10/2017	
SI	19/03/2010	SI	7/05/2009	TIPO II NO I	16/03/2018	120	75	78	175	25,47	Preobeso (25	0,79	31/01/2018	
SI	26/02/2018	SI	25/02/2018	TIPO II NO I	25/04/2018	150	100	69	158	27,64	Preobeso (25	0,79	2/02/2018	
SI	3/10/2012	SI	30/12/3799	TIPO II NO I	10/05/2018	120	75	82	160	32,03	Obeso Tipo 1 (	1,2	13/02/2017	
SI	12/10/2017	SI	8/01/2014	TIPO II INSU	12/09/2018	120	75	63	152	27,27	Preobeso (25	1,5	22/05/2018	
SI	9/01/2015	SI	16/12/2004	TIPO II INSU	31/05/2018	120	75	83	184	24,52	Normal (18,50	1,7	4/01/2017	
NO	13/12/2016	SI	13/12/2016	TIPO II NO I	3/04/2018	120	70	67	160	26,17	Preobeso (25	1	16/03/2018	
SI	18/01/2018	SI	17/01/2013	TIPO II NO I	17/04/2018	120	75	60	170	20,76	Normal (18,50	0,96	19/01/2018	
SI	24/10/2013	SI	10/05/2007	TIPO II NO I	1/06/2018	120	70	69	153	29,48	Preobeso (25	1,1	1/12/2017	
SI	7/03/2018	SI	25/02/2015	TIPO II NO I	7/06/2018	120	75	85	176	27,44	Preobeso (25	0,9	24/05/2018	
SI	19/02/2016	SI	19/02/2018	TIPO II NO I	8/08/2018	120	60	61	140	31,12	Obeso Tipo 1 (	1,3	10/03/2018	
SI	19/10/2017	SI	12/12/2012	TIPO II NO I	8/06/2018	150	80	84	174	27,74	Preobeso (25	0,9	22/03/2018	
SI	28/06/2012	SI	28/06/2018	TIPO II NO I	13/06/2018	163	80	59	153	25,2	Preobeso (25	0,5	30/01/2018	
NO		SI	16/10/2010	TIPO I INSUI	13/06/2018	120	70	70	172	23,66	Normal (18,50	0,7	29/01/2016	
SI	21/03/2017	SI	21/03/2017	TIPO II NO I	13/06/2018	120	75	70	155	29,14	Preobeso (25	0,7	1/03/2018	
SI	15/06/2011	SI	3/03/2011	TIPO II NO I	14/06/2018	120	70	73	172	24,68	Normal (18,50	1	14/02/2018	

Nota: Elaborado por el autor

En la tabla **Tabla 10** *Diccionario de datos*se muestra el diccionario de las 32 variables de la base de datos de la ESE Hospital San Juan de Dios de la sección de pacientes crónicos junto a su descripción.

Tabla 10 *Diccionario de datos*

Variables	Tipo de Dato	Descripción
Genero	Cadena	Genero del paciente (M, F)
Edad	Entero	Edad del paciente (20 – 90 años)
Diagnóstico de Hipertensión Arterial (HTA)	Cadena	El paciente está diagnosticado con HTA (Si, No)
Fecha del diagnóstico de la HTA	Fecha	Fecha en que se le diagnosticó al paciente HTA
Diagnostico Diabetes Mellitus (DM)	Cadena	El paciente está diagnosticado con DM (Si, No)
Fecha del Diagnóstico de la DM	Fecha	Fecha en que se le diagnosticó al paciente DM
Tipo de Diabetes	Cadena	Tipo de DM que presenta en paciente (Tipo I, Tipo II, Tipo III)
Tensión arterial (Sistólica)	Entero	Sístole del paciente (80 – 210 mmHg)
Tensión Arterial (Diastólica)	Entero	Diástole del paciente (49 – 111 mmHg)
Peso	Entero	Peso del paciente (25-137 kg)
Talla	Entero	Talla del paciente (133 – 184 cm)
IMC	Real	IMC del paciente (12,76 – 53.52 kg/m²)
Estado Nutricional	Cadena	Estado nutricional del paciente (Delgadez, normal, Obeso)
Creatinina	Real	Resultado creatinina del paciente (0,26 – 3,1 mg/dl)
Hemoglobina Glicosilada	Real	Resultado hemoglobina glicosilada del paciente (2,2 – 15,4%)
Colesterol Total	Entero	Resultado colesterol total del paciente (57 – 353 mg/dl)
HDL	Real	Resultado HDL del paciente (24 – 91 mg/dl)
LDL	Real	Resultado LDL del paciente (7,4 – 269 mg/dl)
Albuminuria	Real	Resultado albuminuria del paciente (0,1 – 407 mg/l)
Triglicéridos	Real	Resultado triglicéridos del paciente (30 – 742 mg/dl)

Glicemia	Real	Resultado glicemia del paciente (41 – 435 mg/dl)
Fecha Creatinina	Fecha	Fecha de la última creatinina del paciente
Fecha Hemoglobina Glicosilada	Fecha	Fecha de la última hemoglobina glicosilada del paciente
Fecha Colesterol Total	Fecha	Fecha del último colesterol total del paciente
Fecha HDL	Fecha	Fecha del último HDL del paciente
Fecha LDL	Fecha	Fecha del último LDL del paciente
Fecha Albuminuria	Fecha	Fecha de la última albuminuria del paciente
Fecha Triglicéridos	Fecha	Fecha de los últimos triglicéridos del paciente
Fecha Glicemia	Fecha	Fecha de la última glicemia del paciente
Clasificación del Riesgo Cardiovascular	Cadena	Riesgo cardiovascular presente en el paciente (Bajo, Moderado, Alto)
Estadio HTA	Cadena	Estadio HTA que presenta el paciente (Leve, Controlado, Severa)
Municipio	Cadena	Municipio donde es atendido el paciente

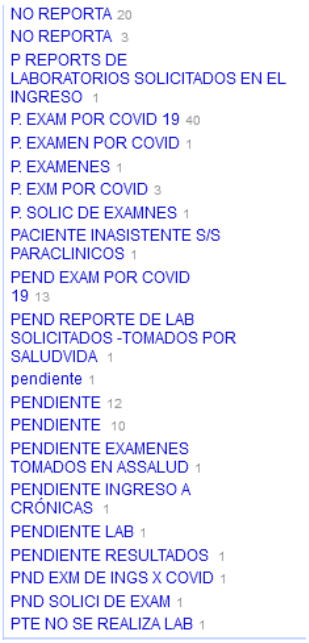
Nota: Elaborado por el autor

5. Limpieza de Datos (Open Refine)

En esta fase se realiza diversas estrategias para manejar datos en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

Como primer paso se eliminó todas las filas que en sus celdas reportaban resultados como: no reporta, pend examen covid, s/s paraclínicos de control, ingreso, reingreso mostradas en la siguiente imagen tomada de registros de la variable creatinina tomada de OpenRefine.

Figura 10 *Registros de la variable creatinina*



Nota: Tomada de OpenRefine

También se identificaron para su posterior tratamiento inconsistencias, dobles espacios, faltas de ortográfica, datos duplicados, espacios vacíos (al principio, final o en medio de los valores), se unificó los tipos de valores por columnas (enteros, reales, cadena, fecha, etc), todo esto utilizando las herramientas que proporciona OpenRefine.

6. Transformación y Carga

En esta fase se lleva a cabo operaciones de agregación o normalización, empleando la herramienta RapidMiner y el lenguaje de programación Python con el fin de una consolidación de los datos para la siguiente fase que es la minería de datos.

Para rectificar los valores de algunas variables, se dio uso al lenguaje de programación Python, como se muestra a continuación:

- IMC: Ya que para la obtención del resultado de la variable IMC se realiza mediante los valores de dos variables diferentes, aplicando la siguiente formula

Ecuación 6 *Índice de masa Corporal (IMC)*

$$IMC = \frac{Peso\ del\ paciente}{Estatura\ \cdot\ Estatura}\ \frac{[Kg]}{[m]}$$

Por consiguiente, se realiza el algoritmo en Python

Figura 11 *Algoritmo trasformación variable IMC*

```
for i in range(len(Peso)):
    Estatura = Talla[i]*0.01
    OperacionIMC = Peso[i]/(Estatura * Estatura)
    n = round(OperacionIMC,2)
    IMCpython.append(n)
```

Nota: Elaborado por el autor

Lo que realiza el anterior algoritmo es sobrescribir los valores originales del vector IMC con el resultado de la operación para cada paciente, obteniendo así los resultados exactos del índice de masa corporal.

- Estado nutricional: esta variable está directamente relacionada con la variable IMC, por lo tanto, se hace el siguiente algoritmo para que cumpla las siguientes reglas:

Tabla 11 *Rango de índice de masa corporal*

Estado Nutricional	Rango de Índice de masa corporal [Kg/m^2]
Delgadez severa	Menor a 16
Delgadez moderada	Entre 16 a 16.99
Delgadez aceptable	Entre 17 a 18.49
Normal	Entre 18.5 a 24.99
Preobeso	Entre 25 a 29.99
Obeso tipo I	Entre 30 a 34.99
Obeso Tipo II	Entre 35 a 39,99
Obeso tipo III	Mayor a 40

Nota: Tomada de (Espinoza, 2016)

Figura 12 *Algoritmo trasformación variable Estado Nutricional*

```
for i in range(len(IMCpython)):

    ## delgadez severa
    if(IMCpython[i] < 16.0):
        ListNE[i] = 'Delgadez Severa (<16)'

    ## delgadez moderada
    if(IMCpython[i]>=16 and IMCpython[i]<17):
        ListNE[i] = 'Delgadez Moderada (16 a 16,99)'

    ## delgadez aceptable
    if(IMCpython[i]>=17 and IMCpython[i]<=18.49):
        ListNE[i] = 'Delgadez Aceptable (17 a 18,49)'

    ## normal
    if(IMCpython[i]>=18.5 and IMCpython[i]<25):
        ListNE[i] ='Normal (18,50 a 24,99)'

    ## preobeso
    if(IMCpython[i]>=25 and IMCpython[i]<30):
        ListNE[i] = 'Preobeso (25 a 29,99)'

    ## obeso tipo 1
    if(IMCpython[i]>=30 and IMCpython[i]<35):
        ListNE[i] = 'Obeso Tipo 1 (30 a 34,99)'

    ## obeso tipo 2
    if(IMCpython[i]>=35 and IMCpython[i]<40):
        ListNE[i] = 'Obeso Tipo 2 (35 a 39,99)'

   # obeso tipo 3
    if(IMCpython[i] >= 40):
       ListNE[i] = 'Obeso Tipo 3 (>=40)'

df['Estado Nutricional 1']=ListNE
```

Nota: Elaborado por el autor

- Estadio HTA: esta variable está directamente relacionada con las variables tensión arterial (Sistólica) y tensión Arterial (Diastólica), por lo tanto el tratamiento para esta variable es similar a la de estado nutricional, por esto motivo se elabora el siguiente algoritmo que cumpla las siguientes reglas:

Tabla 12 *Niveles del Estadio HTA*

Estadio HTA	**Tensión Arterial**	
	Sístole (mmHg)	**Diástole (mmHg)**
Controlado	Menor a 140	Menor a 90
Leve	Entre 140 a 159	Entre 90 a 99
Moderado	Entre 160 179	Entre 100 a 109
Severa	Mayor a 180	Mayor a 110

Nota: Tomado de (Ministerio de Salud, 2017)

Figura 13 *Algoritmo transformación variable estadio HTA*

```
for i in range(len(IMCpython)):

    ## Controlado
    if(sistole[i] < 140 and diastole[i] < 90):
        ListNEstadio[i] = 'Controlado (Sist < 140) (Diast < 90)'

    ## Leve
    if((sistole[i] >= 140 and sistole[i] < 160) and (diastole[i] >= 90 and diastole[i] < 100)):
        ListNE[i] = 'Leve (Sist 140 a 159) (Diast 90 a 99)'

    ## Moderada
    if((sistole[i] >= 160 and sistole[i] < 179) and (diastole[i] >= 100 and diastole[i] < 109)):
        ListNE[i] = 'Moderada (Sist 160 a 179) (Diast 100 a 109)'

    ## Severa
    if(sistole[i] >= 180 and diastole[i] >= 110):
        ListNEstadio[i] = 'Severa (Sist >= 180) (Diast >= 110)'


df['HTA NUEVO']=ListNEstadio
```

Nota: Elaborado por el autor

- Para realizar la técnica de reglas de asociación se crearon nuevas variables aplicando en Excel las siguiente formulas teniendo en cuenta las tablas **Tabla 3** *Niveles del Colesterol Total* a la **Tabla 9** *Niveles de la tasa de filtración glomerular (TFG)*:

Tabla 13 *Formulas para el procesamiento de las variables*

Variables	Formula
Genero	=SI(Genero ="M";1;0)
Colesterol Total	=SI(Colesterol Total >170;1;0)
HDL	=SI(HDL <40;1;0)
LDL	=SI(LDL >100;1;0)
Triglicéridos	=SI(Triglicéridos >150;1;0)

Nota: Elaborado por el autor

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

De estos datos si el género del paciente es masculino toma el valor de 1 y por consiguiente femenino tiene el valor de 0. Si el paciente tiene el colesterol menor a 170 mg/dl es normal tomaría el valor de 0 y si tiene más de 170 mg/dl tiene colesterol alto y tomaría el valor de 1. Si tiene HDL menor a 40 mg/dl es normal y si no tiene riesgo. Si su LDL es mayor a 100 mg/dl tiene riesgo por lo cual tomaría un valor de 1. Si tiene los triglicéridos menores a 150 mg/dl es normal, pero si no tendría los niveles de triglicéridos altos. Quedando de esta manera:

Figura 14 *Preprocesamiento de base de datos para las reglas de asociación*

Genero M=1	Colesterol Total 2	HDL 2	LDL 2	Trigliceridos 2	HTA 2
0	0	0	0	0	1
1	1	0	1	1	1
1	0	0	0	1	1
1	0	0	0	0	1
0	1	0	1	1	1
0	1	0	1	0	1
0	1	0	1	0	0
0	1	1	1	1	1
1	0	0	0	0	1
0	1	0	1	0	1
1	1	1	1	1	1
0	1	0	0	0	1
1	0	0	0	0	1
1	1	0	0	1	0
1	1	0	0	0	1
1	1	0	0	1	1
1	1	0	1	1	1
0	1	0	1	1	1
1	1	0	1	1	1
0	0	1	0	1	1
1	0	0	0	0	0
0	1	0	0	1	1
1	0	0	0	0	1

Nota: Elaborado por el autor

7. Minería de datos (RapidMiner)

En esta etapa es donde se obtienen los patrones de interés según la tarea de minería que llevemos a cabo (descriptiva y predictiva).

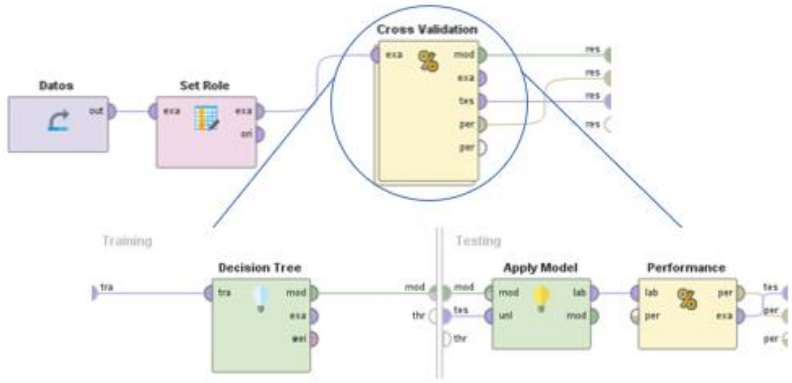
7.1 Algoritmos

Los siguientes algoritmos son modelados en RapidMiner. Para el desarrollo de estos procesos se utiliza los componentes Set Role, Cross Validation y Performance (Regression). Set rol utilizado para seleccionar la variable a predecir quedando así las demás variables como predictoras, Cross Validation utilizado para estimar la precisión con la que un modelo funcionará en la práctica. Teniendo dos subprocesos: un subproceso de entrenamiento y un subproceso de prueba. El subproceso de entrenamiento se utiliza para entrenar un modelo; Luego, el modelo entrenado se aplica en el subproceso de prueba. El rendimiento del modelo se mide mediante el componente Performance (Regression) durante la fase de prueba mediante el error cuadrático medio y R^2.

- **Árbol de decisión**

A continuación, se detalla el modelado para el algoritmo de árbol de decisión aplicado en RapidMiner:

Figura 15 *Proceso para el algoritmo de árbol de decisión*



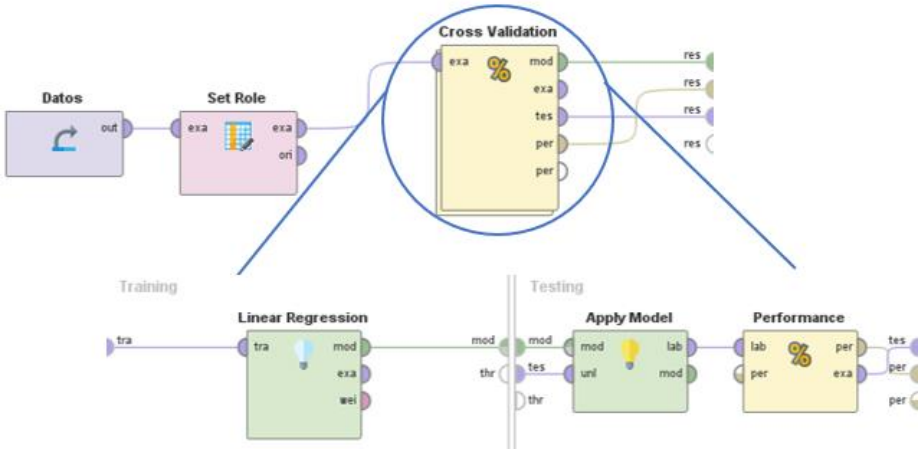
Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizó el componente Decision Tree principalmente. Comenzando el proceso se toma la Data limpia y transformada con su variable objetivo y predictoras ya seleccionadas enviándose subproceso de entrenamiento el cual ejecutará el algoritmo de árbol de decisión bajo los criterios de mínimos cuadrados cuya función es dividir el atributo para que minimiza la distancia al cuadrado entre el promedio de valores en el nodo con respecto al valor verdadero; y un criterio de profundidad máxima igual a 10 que ayuda a restringir la profundidad del árbol de decisión. Después de la ejecución del proceso del algoritmo dentro del subproceso de prueba se envía al proceso de rendimiento.

- **Regresión lineal**

A continuación, se detalla el modelado para el algoritmo de regresión lineal aplicado en RapidMiner:

Figura 16 *Proceso para el algoritmo de Regresión Lineal*



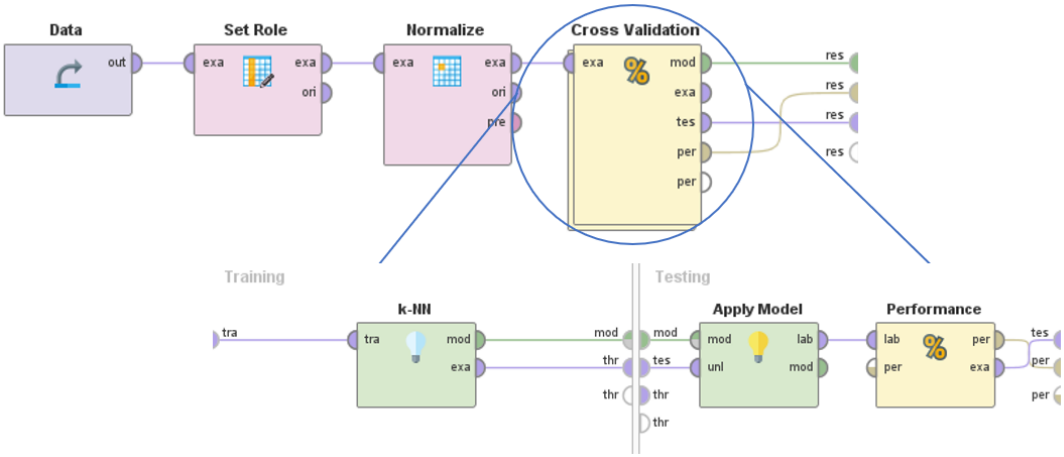
Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizó el componente Linear Regression principalmente. Comenzando el proceso se toma la Data limpia y transformada con su variable objetivo y predictoras ya seleccionadas enviándose subproceso de entrenamiento el cual ejecutará el algoritmo de regresión lineal bajo el criterio de tolerancia mínima igual al 5% utilizado para eliminar características colineales. Después de la ejecución del proceso del algoritmo dentro del subproceso de prueba se envía al proceso de rendimiento.

- **K-NN**

A continuación, se detalla el modelado para el algoritmo de K-NN aplicado en RapidMiner:

Figura 17 *Proceso para el algoritmo de k-NN*



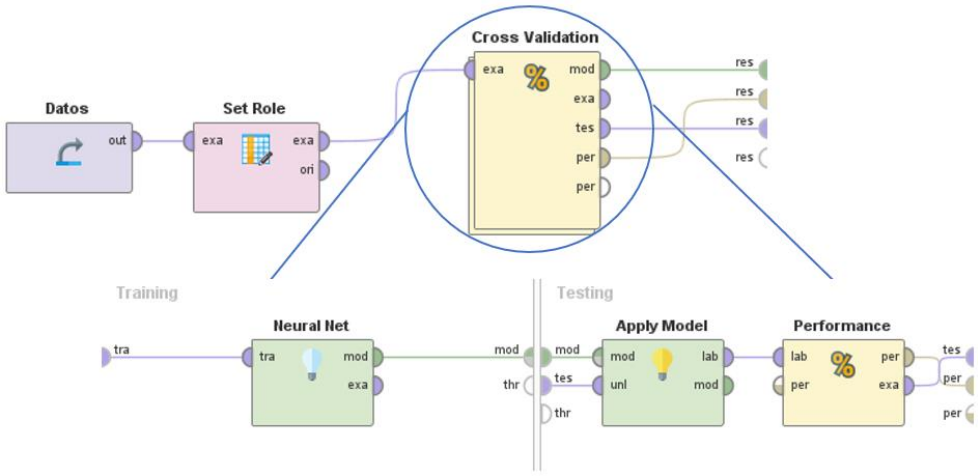
Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizaron componentes de normalización y k-NN principalmente. Comenzando el proceso se toma la Data limpia y transformada con su variable objetivo y predictoras enviándose al proceso de normalización el cual se encargará de ajustar los valores que tienen los datos para que no hallan valores más grandes que otros. El siguiente proceso a realizar es el envío de la tabla normalizada al subproceso de entrenamiento el cual ejecutará el algoritmo de k-NN de acuerdo a los siguientes parámetros: K toma un valor igual a 7 ya que se obtuvo un mejor performance con este valor de K, y el criterio mixedEuclideanDistance para calcular la distancia euclidiana de cada muestra de entrenamiento. Después de la ejecución del proceso del algoritmo dentro del subproceso de prueba se envía al proceso de rendimiento.

- **Redes Neuronales**

A continuación, se detalla el modelado para el algoritmo de redes neuronales aplicado en RapidMiner:

Figura 18 *Proceso para el algoritmo de Redes Neuronales (MLP)*



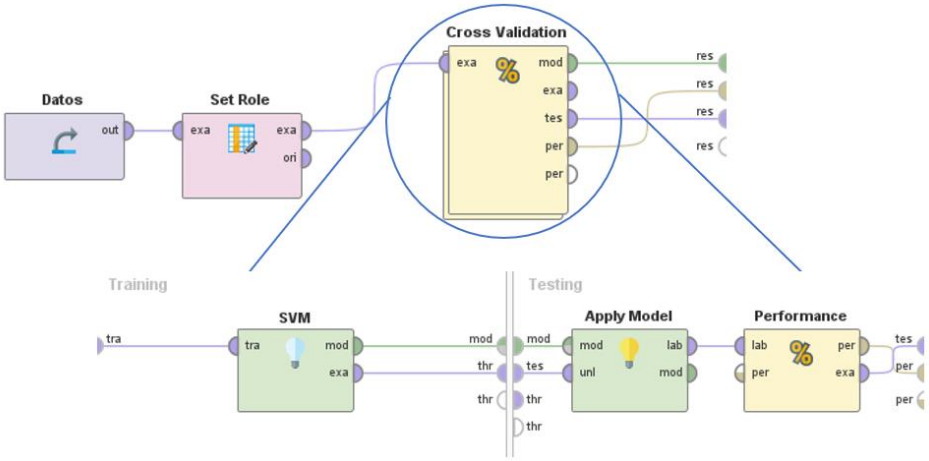
Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizó el componente Neural Net principalmente. Comenzando el proceso se toma la Data limpia y transformada con su variable objetivo y predictoras ya seleccionadas enviándose subproceso de entrenamiento el cual ejecutará el algoritmo de redes neuronales bajo los siguientes parámetros: tomando dos capas ocultas, una con 20 y otra con 10 unidades respectivamente, un total de ciclos de entrenamiento igual a 500. Después de la ejecución del proceso del algoritmo dentro del subproceso de prueba se envía al proceso de rendimiento.

- **SVM**

A continuación, se detalla el modelado para el algoritmo de SVM aplicado en RapidMiner:

Figura 19 *Proceso para el algoritmo de SVM*



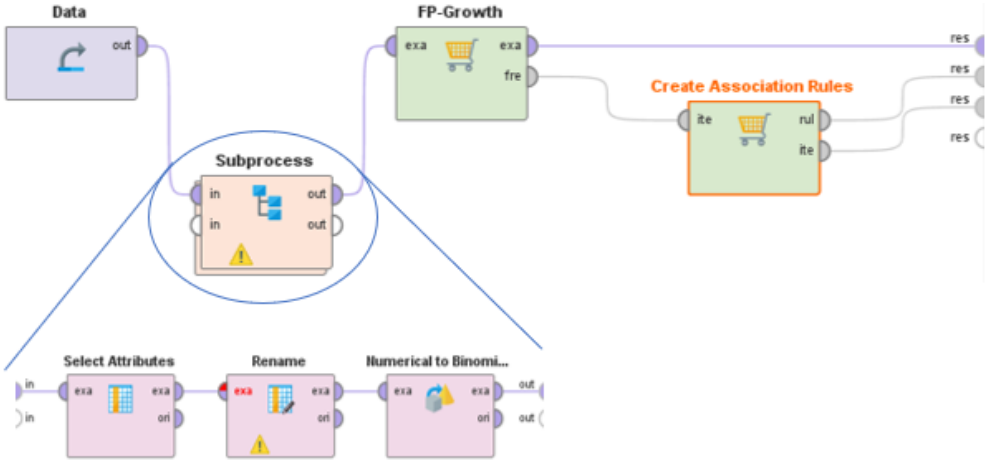
Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizó el componente SVM principalmente. Comenzando el proceso se toma la Data limpia y transformada con su variable objetivo y predictoras ya seleccionadas enviándose subproceso de entrenamiento el cual ejecutará el algoritmo de máquinas de vectores de soporte bajo los siguientes parámetros: Gamma (RBF) de 0.005 y C = 10. Después de la ejecución del proceso del algoritmo dentro del subproceso de prueba se envía al proceso de rendimiento.

- **Reglas de Asociación**

A continuación, se detalla el modelado para el algoritmo de reglas de asociación aplicado en RapidMiner:

Figura 20 *Proceso para el algoritmo de Reglas de asociación*



Nota: Elaborado por el autor en RapidMiner

En este proceso se utilizó los componentes FP-Growth y Create Association Rules principalmente. Comenzando el proceso se toma la Data limpia y transformada enviándose a un subproceso donde se transforma los datos unos y ceros en verdaderos y falsos, para así ejecutar el componente FP-Growth encargado de encontrar en la base de datos los conjuntos más frecuentes para los ítems establecidos por el objeto de estudio, con los parámetros siguientes: un soporte mínimo igual al 20%, para la creación de las reglas un umbral mínimo de confianza de 90% y finalmente, encontrar las reglas de asociación que pueden construirse en base a los conjuntos frecuentes que produce el operador FP-Growth.

7.2 Aplicación de los Algoritmos

Una vez seleccionado las técnicas el paso siguiente es aplicarlo a los datos ya seleccionados, limpiados y procesados.

Para la elección de las variables predictoras junto con su variable dependiente se hizo basadas en el diagrama de correlación de Pearson.

Figura 21 *Matriz de Correlación de Pearson*

	Edad	Sístole	Diástole	Peso	Talla	IMC	Creatinina	Hemoglobina	Colesterol Total	HDL	LDL	Albumina	Trigliceridos	Glicemia	TFG
Edad			-0,114	-0,282	-0,240	-0,165	0,135	-0,121						-0,135	-0,636
Sístole			0,477												
Diástole	-0,114	0,477		0,204	0,115	0,159									0,166
Peso	-0,282		0,204		0,504	0,834				-0,202					0,588
Talla	-0,240		0,115	0,504			0,238			-0,183					0,265
IMC	-0,165		0,159	0,834						-0,116					0,499
Creatinina	0,135				0,238							0,152	0,106		-0,529
Hemoglobina	-0,121											0,115		0,390	
Colesterol Total										0,481	0,805		0,399	0,138	
HDL				-0,202	-0,183	-0,116			0,481		0,391		-0,128		
LDL									0,805	0,391					
Albumina							0,152	0,115					0,149	0,101	
Trigliceridos							0,106		0,399	-0,128	0,149			0,160	
Glicemia	-0,135							0,390	0,138		0,101		0,160		
TFG	-0,636		0,166	0,588	0,265	0,499	-0,529								

Nota: Elaborado por el autor

La interpretación de los valores de la **Figura 21** *Matriz de Correlación de Pearson* es si este coeficiente es igual a 1 o -1 (o cercano a estos valores) significa que una variable es fruto de una transformación lineal de la otra. Teniendo una relación directa al tratarse de 1 (cuando una variable aumenta, la otra también), mientras que existirá una relación inversa al tratarse de -1 (cuando una variable aumenta la otra disminuye).

A continuación, se muestra los índices de correlación de las variables dependientes más importantes para este estudio junto a sus respectivas variables dependientes.

Tabla 14 *Índices de correlación de variables independientes junto a su variable dependiente*

Variable Dependiente	Variables Independientes	Índice de Correlación
Creatinina	Talla	0,238
	Albumina	0,152
	Edad	0,135
	Triglicéridos	0,106
Hemoglobina Glicosilada	Glicemia	0,390
	Albumina	0,115
	Edad	0,121
Colesterol Total	LDL	0,805
	HDL	0,481
	Triglicéridos	0,399
Glicemia	Hemoglobina Glicosilada	0,390
	Triglicéridos	0,160
	Colesterol Total	0,138
	Edad	0,135
TFG	Peso	0,588
	IMC	0,499
	Talla	0,265
	Diástole	0,166

Nota: Elaborado por el autor

Las siguientes tablas (**Tabla 15** a la **Tabla 28**) muestran el rendimiento que tuvo cada uno de los diferentes algoritmos para predecir los valores de cada una de las variables.

Y = {Variables Independientes}; donde Y es la variable a predecir o dependiente.

- Edad = {TFG, Peso, Talla, IMC, Creatinina}

Tabla 15 *Índices de rendimiento para la predicción de la variable Edad*

	Error Absoluto (años)	Error Relativo	R^2
Regresión Lineal	6.350	10.77%	0.542
Árbol de Decisión	8.220	13.79%	0.341
K-NN	8.142	13.90%	0.268
MLP	5.902	10.10%	0.597
SVM	5.319	8.73%	0.661

Nota: Elaborado por el autor

- Sistólica = {Diastólica}

Tabla 16 *Índices de rendimiento para la predicción de la variable Sistólica*

	Error Absoluto (mmHg)	Error Relativo	R^2
Regresión Lineal	9.358	7.15%	0.230
Árbol de Decisión	11.441	8.72%	0.164
K-NN	9.790	7.45%	0.126
MLP	9.445	7.14%	0.244
SVM	8.439	6.44%	0.109

Nota: Elaborado por el autor

- Diastólica = {Peso, IMC, Edad}

Tabla 17 *Índices de rendimiento para la predicción de la variable Diastólica*

	Error Absoluto (mmHg)	Error Relativo	R^2
Regresión Lineal	5.014	6.74%	0.255
Árbol de Decisión	5.695	7.61%	0.171
K-NN	5.218	6.97%	0.173
MLP	5.235	7.05%	0.246
SVM	4.846	6.48%	0.200

Nota: Elaborado por el autor

- Peso = {TFG, Diástole, HDL}

Tabla 18 *Índices de rendimiento para la predicción de la variable Peso*

	Error Absoluto (Kg)	Error Relativo	R^2
Regresión Lineal	0.761	1.17%	0.992
Árbol de Decisión	1.575	2.28%	0.963
K-NN	1.970	2.98%	0.954
MLP	0.588	0.85%	0.996
SVM	0.016	0.02%	1.000

Nota: Elaborado por el autor

- Talla = {Peso, TFG, Creatinina, Diástole}

Tabla 19 *Índices de rendimiento para la predicción de la variable Talla*

	Error Absoluto (cm)	Error Relativo	R^2
Regresión Lineal	0.926	0.59%	0.975
Árbol de Decisión	2.333	1.48%	0.829
K-NN	5.378	3.42%	0.394
MLP	0.796	0.51%	0.988
SVM	1.546	1.00%	0.913

Nota: Elaborado por el autor

- IMC = {Diástole, HDL, Creatinina}

Tabla 20 *Índices de rendimiento para la predicción de la variable IMC*

	Error Absoluto (Kg/m2)	Error Relativo	R^2
Regresión Lineal	0.330	1.23%	0.989
Árbol de Decisión	0.922	3.33%	0.909
K-NN	2.250	8.30%	0.673
MLP	0.362	1.35%	0.992
SVM	0.107	0.38%	0.998

Nota: Elaborado por el autor

- Creatinina = {Talla, Albumina, Edad, Triglicéridos, IMC}

Tabla 21 *Índices de rendimiento para la predicción de la variable Creatinina*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	0.116	14.85%	0.640
Árbol de Decisión	0.137	16.40%	0.624
K-NN	0.160	19.25%	0.367
MLP	0.082	10.36%	0.880
SVM	0.097	10.67%	0.727

Nota: Elaborado por el autor

- Hemoglobina glicosilada = {Edad, Albumina, Glicemia}

Tabla 22 *Índices de rendimiento para la predicción de la variable Hemoglobina Glicosilada*

	Error Absoluto (%)	Error Relativo	R^2
Regresión Lineal	1.286	17.56%	0.200
Árbol de Decisión	1.502	20.58%	0.099
K-NN	1.320	17.95%	0.168
MLP	1.369	18.64%	0.194
SVM	1.286	16.31%	0.200

Nota: Elaborado por el autor

- Colesterol total = {HDL, LDL, triglicéridos}

Tabla 23 *Índices de rendimiento para la predicción de la variable Colesterol Total*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	10.722	6.81%	0.794
Árbol de Decisión	16.709	10.55%	0.658
K-NN	12.206	7.67%	0.792
MLP	10.587	6.63%	0.801
SVM	6.953	5.86%	0.792

Nota: Elaborado por el autor

- HDL = {Colesterol, Peso, Talla, Triglicéridos, IMC}

Tabla 24 *Índices de rendimiento para la predicción de la variable HDL*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	6.703	13.98%	0.388
Árbol de Decisión	8.268	16.95%	0.227
K-NN	7.934	16.65%	0.177
MLP	6.773	13.94%	0.431
SVM	6.484	13.94%	0.425

Nota: Elaborado por el autor

- LDL = {HDL, triglicéridos, glicemia}

Tabla 25 *Índices de rendimiento para la predicción de la variable LDL*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	11.349	17.65%	0.706
Árbol de Decisión	15.879	23.85%	0.537
K-NN	13.876	21.17%	0.681
MLP	12.490	19.13%	0.713
SVM	6.934	7.29%	0.731

Nota: Elaborado por el autor

- Triglicéridos= {Colesterol, Glicemia, Creatinina}

Tabla 26 *Índices de rendimiento para la predicción de la variable Triglicéridos*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	45.400	32.07%	0.387
Árbol de Decisión	59.796	39.31%	0.255
K-NN	58.801	41.98%	0.124
MLP	39.099	28.11%	0.543
SVM	41.473	29.89%	0.481

Nota: Elaborado por el autor

- Glicemia = {Triglicéridos, Colesterol, Edad}

Tabla 27 *Índices de rendimiento para la predicción de la variable Glicemia*

	Error Absoluto (mg/dl)	Error Relativo	R^2
Regresión Lineal	38.675	63.58%	0.164
Árbol de Decisión	55.670	71.09%	0.050
K-NN	42.146	58.36%	0.078
MLP	40.819	67.63%	0.156
SVM	40.544	71.85%	0.137

Nota: Elaborado por el autor

- TFG = {Edad, Creatinina, IMC}

Tabla 28 *Índices de rendimiento para la predicción de la variable TFG*

	Error Absoluto (mL/min)	Error Relativo	R^2
Regresión Lineal	10.312	14.58%	0.830
Árbol de Decisión	13.190	16.61%	0.707
K-NN	14.469	19.33%	0.717
MLP	7.420	9.47%	0.905
SVM	24.804	27.50%	0.461

Nota: Elaborado por el autor

Para la aplicación de las reglas de decisión se utilizaron todas las variables dando como resultado la siguiente tabla de frecuencia

Figura 22 *Diagrama de Frecuencia*

Size	Support	Item 1	Item 2	Item 3
1	0.890	Diagnostico Diabetes Mellitu...		
1	0.864	Diagnóstico de Hipertensión ...		
1	0.671	Tension arterial (Sistolica)		
1	0.639	TFG		
1	0.556	Colesterol Total		
2	0.755	Diagnostico Diabetes Mellitu...	Diagnóstico de Hipertensión ...	
2	0.586	Diagnostico Diabetes Mellitu...	Tension arterial (Sistolica)	
2	0.574	Diagnostico Diabetes Mellitu...	TFG	
2	0.600	Diagnóstico de Hipertensión ...	Tension arterial (Sistolica)	
2	0.568	Diagnóstico de Hipertensión ...	TFG	
3	0.515	Diagnostico Diabetes Mellitu...	Diagnóstico de Hipertensión ...	Tension arterial (Sistolica)
3	0.503	Diagnostico Diabetes Mellitu...	Diagnóstico de Hipertensión ...	TFG

Nota: Elaborado por el autor

Seguidamente, partiendo de la **Figura 22** se obtiene las siguientes reglas de asociación

Figura 23 *Resultado del algoritmo de Reglas de Asociación*

No.	Premises	Conclusion	Support ↓	Confidence
13	Diagnostico Diabetes Mellitus (DM)	Diagnóstico de Hipertensión Arterial (HTA)	0.755	0.847
16	Diagnóstico de Hipertensión Arterial (HTA)	Diagnostico Diabetes Mellitus (DM)	0.755	0.873
10	Diagnóstico de Hipertensión Arterial (HTA)	Tension arterial (Sistolica)	0.600	0.695
21	Tension arterial (Sistolica)	Diagnóstico de Hipertensión Arterial (HTA)	0.600	0.894
7	Diagnostico Diabetes Mellitus (DM)	Tension arterial (Sistolica)	0.586	0.658
15	Tension arterial (Sistolica)	Diagnostico Diabetes Mellitus (DM)	0.586	0.873
5	Diagnostico Diabetes Mellitus (DM)	TFG	0.574	0.645
22	TFG	Diagnostico Diabetes Mellitus (DM)	0.574	0.898
6	Diagnóstico de Hipertensión Arterial (HTA)	TFG	0.568	0.657
20	TFG	Diagnóstico de Hipertensión Arterial (HTA)	0.568	0.889
9	Diagnostico Diabetes Mellitus (DM), Diagnóstico d...	Tension arterial (Sistolica)	0.515	0.683
11	Tension arterial (Sistolica)	Diagnostico Diabetes Mellitus (DM), Diagnóstico d...	0.515	0.767
14	Diagnóstico de Hipertensión Arterial (HTA), Tensi...	Diagnostico Diabetes Mellitus (DM)	0.515	0.858
18	Diagnostico Diabetes Mellitus (DM), Tension arteri...	Diagnóstico de Hipertensión Arterial (HTA)	0.515	0.879
8	Diagnostico Diabetes Mellitus (DM), Diagnóstico d...	TFG	0.503	0.667
12	TFG	Diagnostico Diabetes Mellitus (DM), Diagnóstico d...	0.503	0.787

Nota: Elaborado por el autor

8. Evaluación e Interpretación

8.1 Análisis Descriptivo

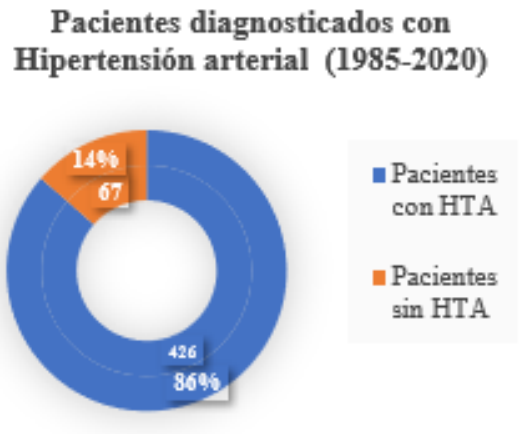
Las siguientes graficas permiten la visualización de datos correspondientes a pacientes con enfermedades crónicas (diabetes e hipertensión) de la ESE Hospital San Juan de Dios de Pamplona, ofreciendo información clara y precisa del comportamiento de estas enfermedades a través de los últimos años.

Se aclara que el programa de crónicos en la ESE Hospital San Juan de Dios de Pamplona, tiene como marco jurídico la Ley 100 de 1993, pero en sí el programa empezó a funcionar desde el año 2000 y se consolidó dicho programa a partir del año 2008.

A raíz de un cambio de contrato con la empresa de tecnología en el hospital la base de datos suministrada y trabajada en este proyecto fue alimentada desde el año 2018 hasta principios del año 2020. Aclarando que los registros desde este momento, época de pandemia por el Covid-19, se encuentran en una base de datos aparte.

- Hipertensión Arterial

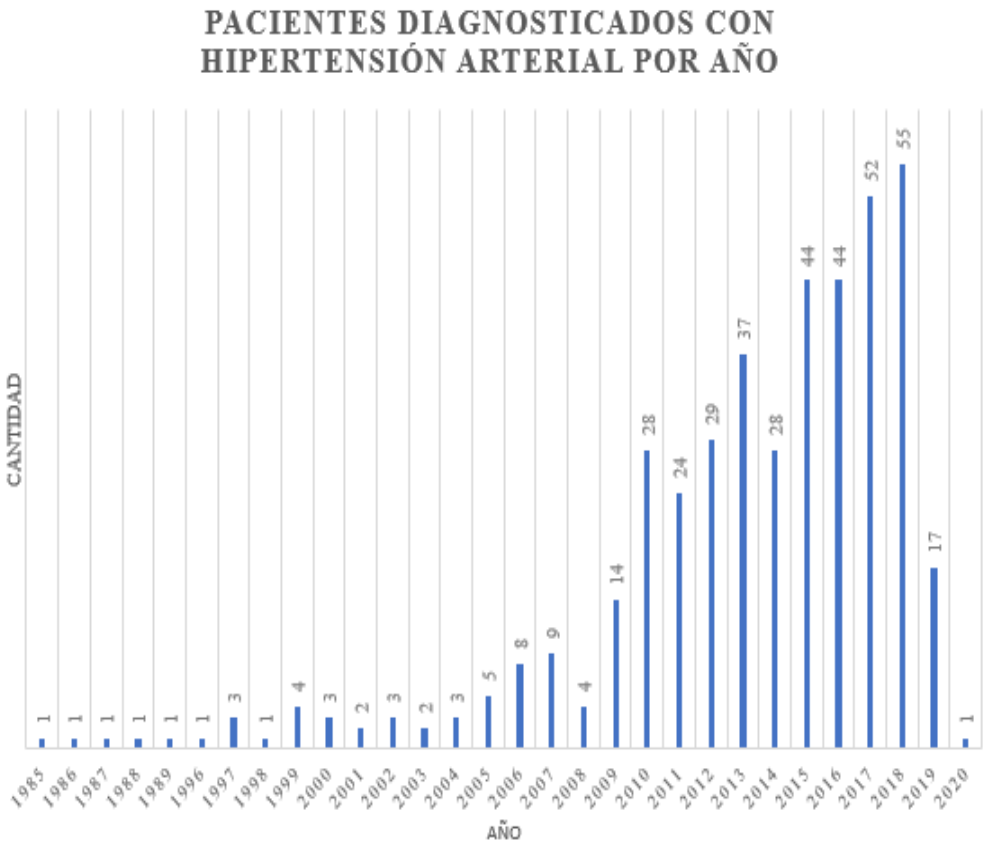
Figura 24 *Distribución de pacientes diagnosticados con HTA*



Nota: Elaborado por el autor

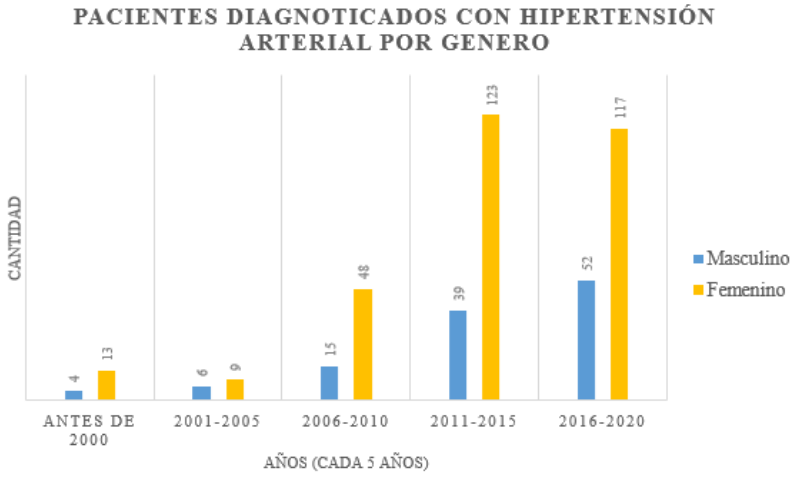
Del total de los pacientes que se encuentran adscritos al programa de crónicos del Hospital San Juan de Dios, 4 de 5 pacientes tienen un diagnóstico de HTA, ubicándolos con riesgo cardiovascular moderado y/o alto. Lo cual permite afirmar que este programa presenta una alta incidencia de hipertensión arterial.

Figura 25 *Número de pacientes diagnosticados con HTA vs Año*



Nota: Elaborado por el autor

Figura 26 *Cantidad de pacientes diagnosticados con HTA por género y por periodo de 5 años*

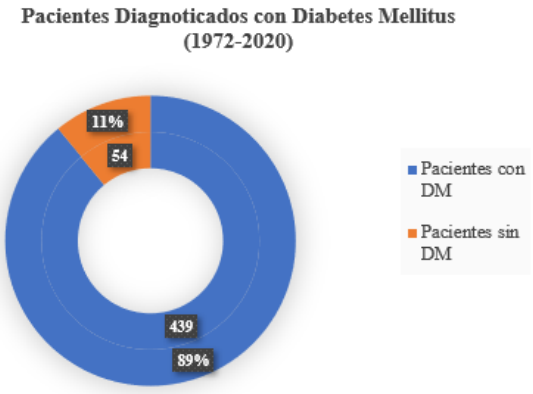


Nota: Elaborado por el autor

Se observa un aumento significativo en la última década en la cifra de diagnósticos de Hipertensión aclarando que en el año 2018 se presentó el mayor número de casos, igualmente se concluye que existe una marcada diferencia de dicha patología entre géneros (mujeres 73% y hombres 27%).

- Diabetes Mellitus

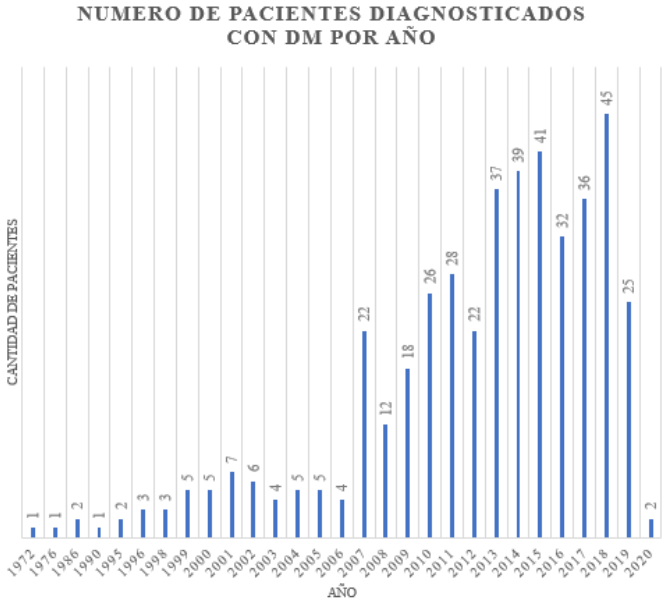
Figura 27 *Distribución de pacientes diagnosticados con DM*



Nota: Elaborado por el autor

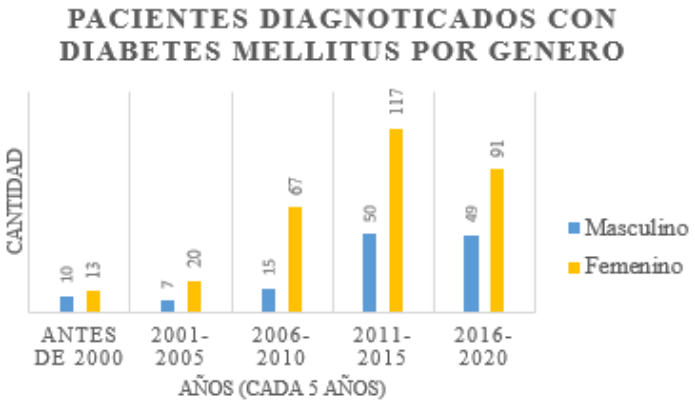
Del total de los pacientes que se encuentran adscritos al programa de crónicos del Hospital San Juan de Dios, 1 de 10 pacientes aproximadamente tienen diabetes mellitus, lo cual permite afirmar que es notable la presencia de esta enfermedad en el programa.

Figura 28 *Número de pacientes diagnosticados con DM vs Año*



Nota: Elaborado por el autor

Figura 29 *Cantidad de pacientes diagnosticados con DM por género y por periodo de 5 años*

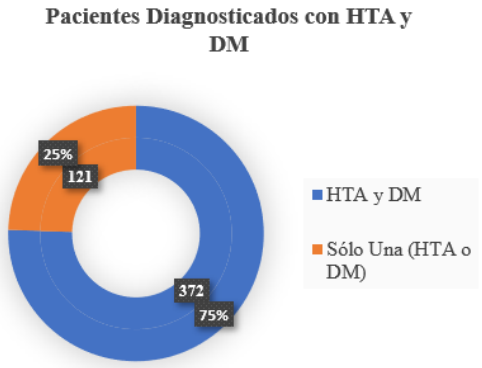


Nota: Elaborado por el autor

Se observa un aumento significativo en la cifra del diagnóstico de Diabetes mellitus desde el año 2007 al 2018, notándose la marcada diferencia de dicha patología entre géneros (mujeres 70% y hombres 30%).

- HTA y DM

Figura 30 *Distribución de pacientes diagnosticados con DM Y HTA*

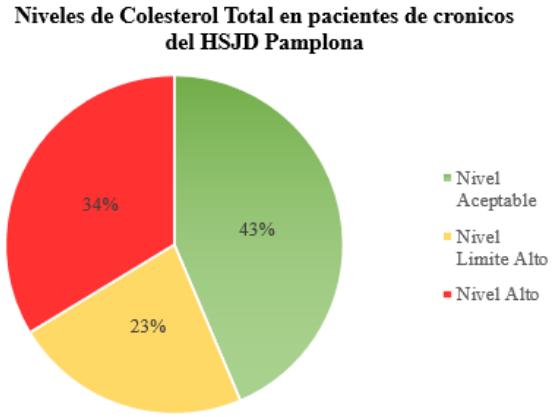


Nota: Elaborado por el autor

Del total de los pacientes que se encuentran adscritos al programa de crónicos del Hospital San Juan de Dios, 6 de 8 pacientes están diagnosticados con las dos enfermedades de hipertensión arterial y diabetes mellitus, y tan solo 2 de cada 8 pacientes está diagnosticado con hipertensión o Diabetes. Las personas con DM tienen mayor probabilidad de tener HTA o viceversa.

- Colesterol Total

Figura 31 *Porcentaje de pacientes basado en los niveles de Colesterol Total*

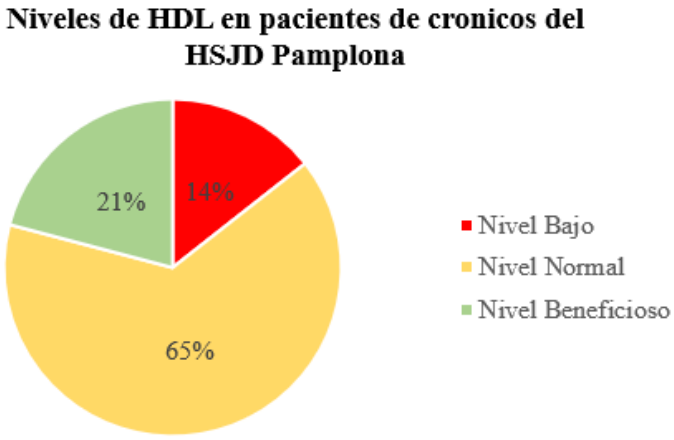


Nota: Elaborado por el autor

Más de la mitad de los pacientes pertenecientes al programa de crónicos presenta colesterol alto o al límite por lo cual tienen tendencia a sufrir de hipercolesterolemia, patología que predispone al desarrollo de enfermedades cardiocerebrovasculares, arterioesclerosis, infarto agudo de miocardio (IAM), secuelas discapacitantes y hasta la muerte.

- HDL

Figura 32 *Porcentaje de pacientes basado en los niveles de HDL*

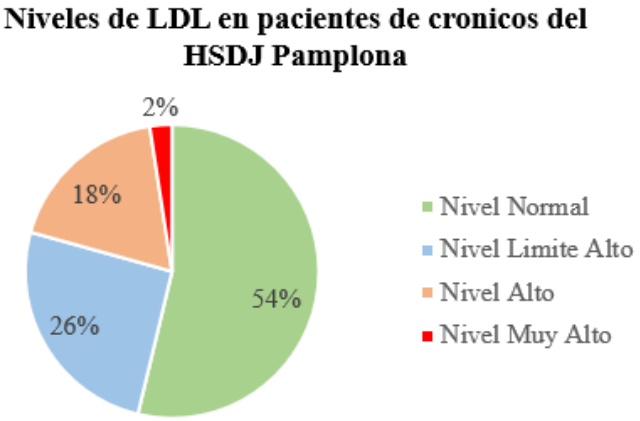


Nota: Elaborado por el autor

Se observa que una buena cantidad de pacientes tienen el HDL normal, el colesterol HDL se puede considerar como el colesterol “bueno” porque un nivel saludable, puede proteger contra los ataques cardíacos y los ataques cerebrales.

- LDL

Figura 33 *Porcentaje de pacientes basado en los niveles de LDL*

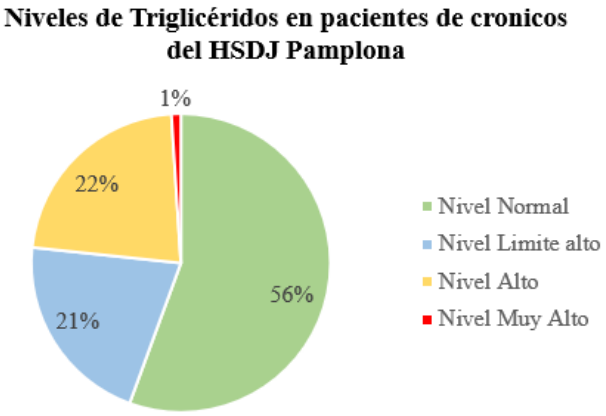


Nota: Elaborado por el autor

Alrededor de la mitad del número de pacientes pertenecientes al programa de crónicas tienen un LDL nivel alto, muy alto y al límite, los cuales tienen tendencia a acumular colesterol malo en las arterias, llevando esto a tener una gran incidencia en las enfermedades de las arterias coronarias.

- Triglicéridos

Figura 34 *Porcentaje de pacientes basado en los niveles de Triglicéridos*

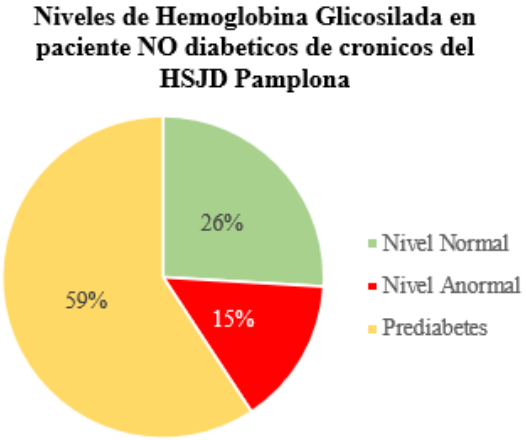


Nota: Elaborado por el autor

Se aprecia que casi la mitad de los pacientes pertenecientes al programa de crónicos presentan triglicéridos altos, muy alto o al límite, lo cual es signo de tener depósitos grasos en las paredes arteriales que posiblemente puede llevar a dichos pacientes a sufrir enfermedades cardiacas o cerebrovasculares.

- Hemoglobina Glicosilada

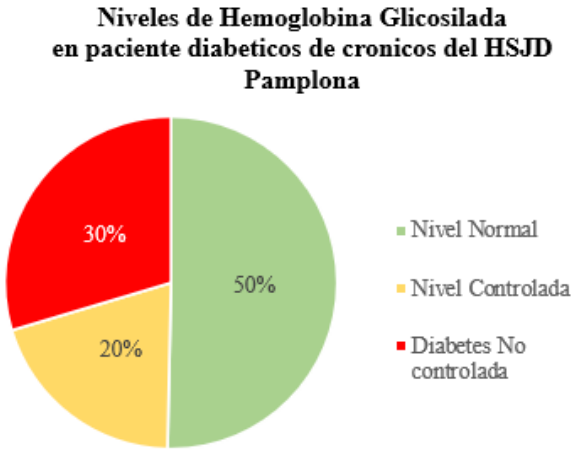
Figura 35 *Porcentaje de pacientes no diabéticos basado en los niveles de Hemoglobina Glicosilada*



Nota: Elaborado por el autor

Se observa en el gráfico que más de la mitad de los pacientes no diabéticos pertenecientes al programa de crónicas tienen prediabetes, la cual es un factor de riesgo para desarrollar diabetes, cifra que permite tomar medidas de prevención teniendo en cuenta los estilos de vida saludables. Las personas con prediabetes pueden necesitar repetir las pruebas cada año.

Figura 36 *Porcentaje de pacientes diabéticos basado en los niveles de Hemoglobina Glicosilada*

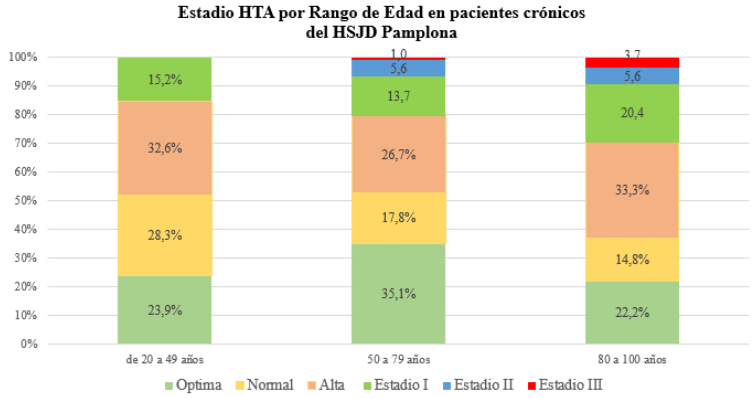


Nota: Elaborado por el autor

Se nota que más de la cuarta parte de los pacientes diabéticos adscritos al programa de crónicos de la E.S.E Hospital San Juan de Dios de Pamplona tienen elevado el nivel de hemoglobina glicosilada, reflejándose una diabetes no controlada; que puede traer como consecuencia un mayor riesgo de sus complicaciones, pudiendo desencadenar en un ataque cardíaco, accidente cardiocerebrovasculares y otros problemas.

- Estadio HTA

Figura 37 *Estadio HTA por rango de edad en pacientes crónicos*



Nota: Elaborado por el autor

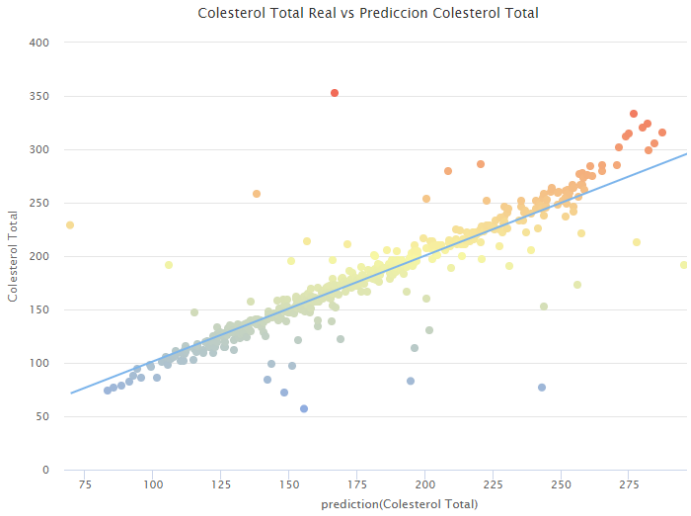
Se evidencia que la hipertensión arterial tiende a ser normal en pacientes jóvenes, pero ésta se va elevando a medida que la persona se envejece después de los 50 años de edad, con factores de riesgo predisponentes.

8.2 Análisis Predictivo

A continuación, se muestra un análisis de los valores predictivos para cada una de las variables

- Colesterol Total

Figura 38 *Colesterol total real vs predicción*



Nota: Elaborado por el autor

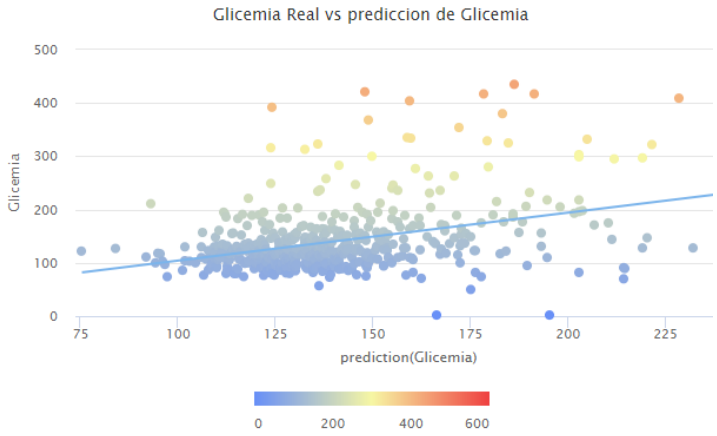
Comparando las desviaciones estándar en los dos conjuntos de datos se puede afirmar que el conjunto de valores predictivos es un conjunto de datos menos disperso al conjunto de datos reales de colesterol total.

De los valores predichos coinciden en el nivel (aceptable, limite alto y alto) 437 de los 493 valores correspondiente al 88,6%.

48 valores cambiaron en 1 nivel, dando como promedio de error absoluto 23,1 y 8 valores cambiaron en 2 niveles, dando como promedio de error absoluto 111,3 mg/dl.

- Glicemia

Figura 39 *Glicemia real vs predicción*



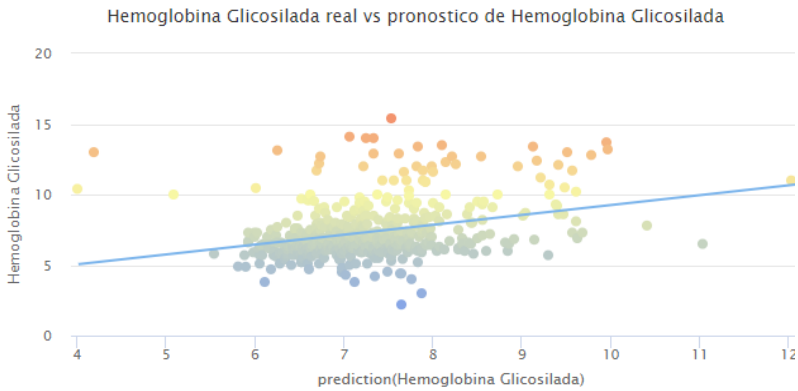
Nota: Elaborado por el autor

De los valores predichos coinciden en el nivel (hipoglucemia, normal, alto y muy alto) 270 de los 493 valores correspondiente al 55%.

220 valores cambiaron en 1 nivel, dando como promedio de error absoluto 43.7 mg/dl y 3 valores cambiaron en 2 niveles, dando como promedio de error absoluto 98,6 mg/dl.

- Hemoglobina Glicosilada

Figura 40 *Hemoglobina Glicosilada real vs predicción*



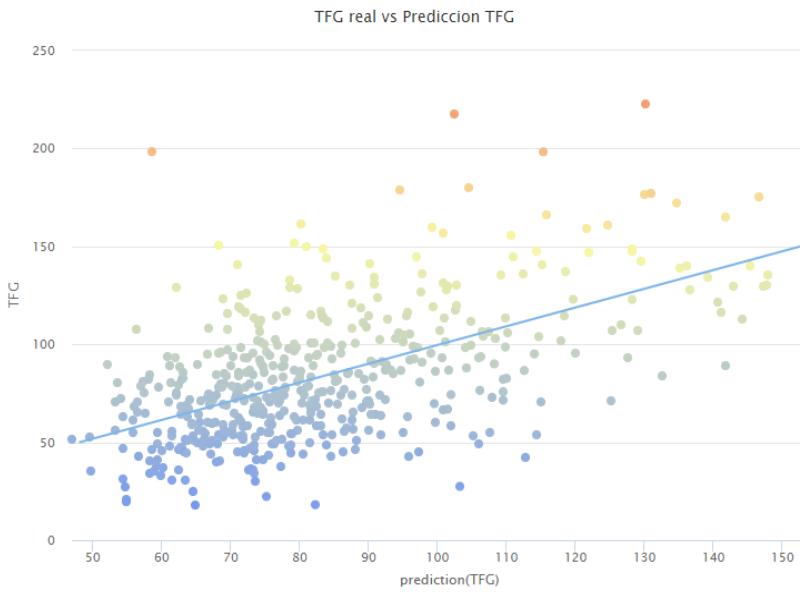
Nota: Elaborado por el autor

Hemoglobina glicosilada para personas diabéticas: De los valores predichos coinciden en el nivel (normal, controlada, diabetes No controlada) 160 de 439 valores. Correspondientes a el 36,4% (si unificamos alto y descontrolada da un 56% de acierto)

Hemoglobina glicosilada para personas No diabéticas: De los valores predichos coinciden en el nivel (normal, prediabetes y anormal) 22 de 54 valores. valores correspondientes al 40% (si unificamos prediabetes y anormal da 56%).

- TFG

Figura 41 *TFG real vs predicción*



Nota: Elaborado por el autor

De los valores predichos coinciden en el nivel (estadio I, II, III, IV, V) 250 de los 493 valores correspondiente al 50,7%.

223 valores cambiaron en 1 nivel, dando como promedio de error absoluto 26,8 ml/minuto; 18 valores cambiaron en 2 niveles, dando como promedio de error absoluto 48,8 ml/minuto y 2 valores cambiaron en 2 niveles, dando como promedio de error absoluto 61,9 ml/minuto.

- De la **Figura 23** *Resultado del algoritmo de Reglas de Asociación* se sacaron las reglas más relevantes en la siguiente tabla:

Tabla 29 *Reglas de Asociación más relevantes*

N°	Premisas	Conclusión	Soporte	Confianza
1	Sistólica, TFG, Colesterol Total	HTA	0.219	0.923
2	Colesterol Total	HTA	0.462	0.832
3	Sistólica	HTA	0.6	0.894
4	Colesterol Total	HTA, DM	0.406	0.730
5	Glicemia	DM	0.227	0.991
6	Colesterol Total	DM	0.499	0.898
7	DM	HTA	0.755	0.847
8	HTA	DM	0.755	0.873
9	Sistólica, Hemoglobina Glicosilada	DM	0.274	0.823
10	LDL, Triglicéridos	Colesterol Total	0.235	0.983
11	Hemoglobina Glicosilada, LDL	Colesterol Total	0.215	0.946
12	TFG, Triglicéridos	Sistólica	0.201	0.707
13	Colesterol Total, LDL	Hemoglobina Glicosilada	0.215	0.51
14	Sistólica, TFG	Hemoglobina Glicosilada	0.215	0.502
15	Colesterol Total, Triglicéridos	TFG	0.213	0.656

Nota: Elaborado por el autor

De la **Tabla 29** se pueden hacer las siguientes conclusiones entre diferentes reglas de asociación:

- Según la regla de asociación 4 y 11, el paciente tiene los niveles de LDL y triglicéridos altos tiende a tener los niveles de colesterol total también altos, a un nivel de confianza del 94%, lo que traería como consecuencia ser diagnosticado con hipertensión arterial y diabetes mellitus.

- Según la regla de asociación 7 y 8, la HTA genera Diabetes en un nivel de confianza de 87%, como también puede ser que la HTA sea consecuencia a las complicaciones de la diabetes en un nivel de confianza de 84%, que en primera vista el 75.5% de los 493 pacientes crónicos del hospital de Pamplona cumplen esta regla.

- Según la regla de asociación 15, la gente con niveles elevados de colesterol total y triglicéridos se sugiere hacer cambios en su dieta para que su función renal no sea afectada.

9. Conclusiones, Recomendaciones y Trabajos futuros

9.1 Conclusiones

- Se obtuvo información de los registros de pacientes crónicos (hipertensión arterial y diabetes mellitus) de la ESE Hospital San Juan de Dios de Pamplona, referente a los años 2018, 2019 y 2020 contando con 3492 registros que fueron recolectados a partir de exámenes médicos hechos por las EPS correspondiente.

- Teniendo como escenario de práctica la ESE Hospital San Juan de Dios Pamplona se facilitó gestionar la solicitud y adquisición de los datos correspondientes al programa de crónicos para su respectivo estudio. Seguidamente se utilizó la herramienta OpenRefine para la limpieza quedando 493 datos de los 3492 suministrados inicialmente; así mismo, se mejoró la calidad de los datos mediante la verificación y cambio de las variables IMC, rango nutricional, estadio HTA, además se asignaron valores binarios a todas las variables para la correcta ejecución del algoritmo de reglas de asociación. Finalmente, se implementó el proceso de minería de datos para obtener el análisis descriptivo y predictivo del comportamiento de las enfermedades crónicas hipertensión y diabetes en la provincia de Pamplona, a partir de los algoritmos de regresión lineal, k-NN, árbol de decisión, redes neuronales, SVM y reglas de decisión, cumpliendo así con el proceso propio de la metodología KDD.

- Una vez analizados los resultados de los algoritmos permitió establecer un análisis descriptivo mostrando la información mediante diagramas de barras y circulares, concluyendo que 4 de 5 pacientes aproximadamente padecen de hipertensión arterial,

igualmente existe una marcada diferencia de dicha patología entre el género, también se establece que 1 de 10 pacientes aproximadamente tienen diabetes mellitus y 6 de 8 pacientes están diagnosticados con las dos enfermedades (HTA y DM).

- Al aplicar las reglas de asociación se generó el soporte de predicción en el cual se observa que los pacientes que tiene los niveles de LDL y triglicéridos altos tienden a tener los niveles de colesterol total también altos, a un nivel de confianza del 98%, lo que traería como consecuencia ser diagnosticado con hipertensión arterial y diabetes mellitus. La HTA genera Diabetes en un nivel de confianza de 87%, como también puede ser que la HTA sea consecuencia a las complicaciones de la diabetes en un nivel de confianza de 84%, que en primera vista el 75.5% de los 493 pacientes crónicos del hospital de Pamplona cumplen esta condición. Es importante destacar que de los valores predichos del colesterol total coinciden en el nivel (aceptable, limite, alto) 437 de 493 valores, correspondiente al 88,6%.

- La ESE Hospital San Juan de Dios de Pamplona me permitió el desarrollo de mi práctica profesional en la dependencia de Informática y Estadística. Aunque fue de forma virtual debido a las circunstancias de la pandemia fue una experiencia enriquecedora, ya que me permitió aplicar mis conocimientos adquiridos en la carrera para proponer soluciones a problemas en un contexto real, aportando una herramienta de análisis en la toma de decisiones de carácter administrativo.

9.2 Recomendaciones

Se recomienda a la ESE Hospital San Juan de Dios de Pamplona, implementar estrategias para garantizar veracidad y confianza en el registro de datos manuales, automatizando la validación de los mismos, con el fin de evitar que los datos digitados sean errados o incompletos, además, ofrecer capacitación de concientización al personal sobre la importancia del buen manejo de los datos y su seguridad.

9.3 Trabajos Futuros

- Desarrollo de un software a partir de los resultados obtenidos en la implementación de la metodología KDD, cuyo fin sería la creación de una herramienta de apoyo para ayudar al profesional en salud a realizar diagnósticos más ágiles precisos y seguros.

- Implementación de la metodología KDD con datos de diferentes áreas tanto de salud como administrativas de la ESE Hospital San Juan de Dios de Pamplona u otras entidades de diferente índole.

10. Referencias Bibliográficas

- Aguilar, J. S. (2017). *Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia. Trabajo de Grado. Universidad Católica de Colombia.* . Bogotá, Colombia: Facultad de Ingeniería. Programa de Ingeniería de Sistemas.
- Avendaño, A. A. (2016). *Tecnicas de minetia de datos para prediccion del diagnostico de hipertension arterial.* Chiclayo, Perú: Universidad Señor de Sipán.
- Barrientos Martínez, R. E., Cruz Ramírez, N., Acosta Mesa, H. G., Rabatte Suárez, I., Gogeascoechea Trejo, M. d., & Blázquez Morales, S. (2019). Árboles de decisión como herramienta en el diagnóstico médico. Universidad Veracruzana.
- Betancourt, D. F. (Marzo de 2016). *Medición del error en pronósticos de demanda.* Obtenido de https://www.ingenioempresa.com/medicion-error-pronostico
- Betancurt, G. A. (s.f.). *LAS MÁQUINAS DE SOPORTE VECTORIAL. Scientia Et Technica.* Obtenido de https://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895
- Diabetes, F. M. (2016). *¿Que es la diabetes?*
- Espinoza, E. N. (2016). *Clasificación del estado nutricional basada en perfiles antropométricos del personal silvoagropecuario femenino de un sector del centro-sur de Chile.*
- Fayyad, U., Piatetsky-Shapiro, G., & Padhra. (s.f.). *Advances in Knowledge Discovery and Data Mining.*
- García Cambronero, C., & Gómez Moreno, I. (s.f.). *ALGORITMOS DE APRENDIZAJE: KNN&KMEANS.* Madrid, España: Universidad Carlos III.
- Guillén, S. I. (5 de Julio de 2019). Enfermedades crónicas no transmisibles como amenaza. Pamplona, España.

- Ham, K. (2013). OpenRefine (version 2.5).Free, open-source tool for cleaning and transforming data. *JMLA*.
- Huizen, J. (2019). *¿Cuáles son los niveles ideales de azúcar en la sangre?*
- Ibáñez, A. S. (2016). *Prevención y progresión de la nefropatía diabética I: epidemiología, patogenia, diagnóstico.*
- *Instituto de ingenieria del conocimiento.* (s.f.). Obtenido de https://www.iic.uam.es/big-data/
- José C. Riquelme, R. R. (2016). *Minería de Datos: Conceptos y Tendencias* . Valencia, España.
- L., M. G. (2016). *Valores Normales de Colesterol y Triglicéridos.*
- Landa, J. (19 de Febrero de 2016). *http://fcojlanda.me/.* Obtenido de http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/
- Larrañaga, P., Inza, I., & Moujahid, A. (s.f.). *Redes Neuronales.* España: Universidad del País Vasco.
- Liendo, M. G. (2018). *Tasa de Filtración Glomerular (TFG renal).*
- López, C. P. (s.f.). *Minería de datos. Técnicas y herramientas.* Thomson.
- López, J. F. (Octubre de 2017). *Coeficiente de determinación (R cuadrado).* Economipedia.com.
- Lozano, A. P. (2017). Cómo introducirse en el universo Big Data | UNIR open class. Madrid, España.
- Marcelo E. Andía, C. A. (2019). *A conceptual guide to use and understand Big Data in clinical research.*

- MARLON BORJA PUERTA, V. C. (2017). • *Minería de datos de salud: estudio de los factores personales, familiares y vivienda que influyen en las enfermedades de diabetes e hipertensión a partir de la encuesta de atención primaria en salud del area metropolitana del valle de aburra.* Medellín: Trabajo de Grado.
- Martínez, C. G. (Febrero de 2020). *REGLAS DE ASOCIACIÓN.* Obtenido de https://rpubs.com/Cristina_Gil/Reglas_Asociacion
- Mercado Polo, D., Pedraza Caballero, L., & Martínez Gómez, E. (2015). *Comparación de Redes Neuronales aplicadas a la predicción de Series de Tiempo (Master en Sistemas informaticos).* Barranquilla, Colombia: Universidad de la Costa.
- Microsoft. (08 de 05 de 2018). *Algoritmo de regresión lineal de Microsoft.* Obtenido de https://docs.microsoft.com/es-es/analysis-services/data-mining/microsoft-linear-regression-algorithm?view=asallproducts-allversions
- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). *YALE: Rapid Prototyping for Complex Data Mining Tasks.* Dortmund.
- Ministerio de Salud, C. (2017). *Dia Mundial de la Hipertension Arterial.*
- OpenRefine. (s.f.). *OpenRefine*. Obtenido de https://openrefine.org/
- *Organización Mundial de la Salud*. (13 de Septiembre de 2019). Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/hypertension
- Python. (s.f.). *Python*. Obtenido de https://www.python.org/doc/essays/blurb/
- Rodrigo, J. A. (2016). *Correlación lineal y Regresión lineal simple.* Obtenido de https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal
- Rodrigo, J. A. (Junio de 2018). *Reglas de asociación y algoritmo Apriori con R*. Obtenido de https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion

- Rodríguez Rodríguez, J. E., Rojas Blanco, E. A., & Franco Camacho, R. O. (s.f.). Clasificación de datos usando el método k-nn. *Revista Universidad Distrital*.
- Rueda, G. G. (2020). *Minado de reglas de asociación aplicado al análisis de riesgo sísmico(Tesis de Grado).* Toluca, México.
- Salazar GA, G. H. (s.f.). *Evaluación de los Niveles de Creatinina Sérica en Pacientes del Hospital de San José de Bogotá.* Bogotá, Colombia.
- Salud, O. M. (8 de Junio de 2020). *https://www.who.int*. Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/diabetes#:~:text=Se%20estima%20que%20en%202016,los%2070%20a%C3%B1os%20de%20edad.
- Sampedro, S. A. (2013). *El colesterol y los triglicéridos.*
- Sterckx, S. D. (2018). *Presuming the promotion of the common good by large-scale health research: the cases of care. data 2.0 and the 100,000 Genomes Project in the UK. In Personalised medicine, individual choice and the common good.* Cambridge University Press.
- Suárez, R. (21 de Agosto de 2018). El 'big data' en salud: presente y futuro de la atención.
- Universia. (Diciembre de 2020). *Para qué sirve Python: qué es y usos.* Obtenido de https://www.universia.net/es/actualidad/orientacion-academica/para-que-sirve-phyton-que-es-y-usos-1154393.html
- Vivas, M. (19 de Mayo de 2020). *https://consultorsalud.com.* Obtenido de https://consultorsalud.com/panorama-de-la-hipertension-arterial-en-colombia/

- Viviana Cañón, A. C. (1 de Noviembre de 2017). *datapopalliance.* Obtenido de http://datapopalliance.org/wp-content/uploads/2018/09/Documento1_VersionFinal_DNP.pdf
- Zambrano, J. (2018). *¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del machine learning y la automatización inteligente.*