INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

"ESTUDIO COMPARATIVO DE ALGORITMOS DE CLASIFICACIÓN DE APRENDIZAJE AUTOMÁTICO EN LA DETECCIÓN DE ENFERMEDADES DEL CORAZÓN."

TESIS

PARA OBTENER EL GRADO DE MAESTRÍA EN SISTEMAS COMPUTACIONALES

P R E S E N T A

L.I. MARCELA LARA CRUZ

DIRECTOR

DR. EDDY SÁNCHEZ DE LA CRUZ

CO-DIRECTOR

DR. RAJESH ROSHAN BISWAL

MISANTLA, VERACRUZ JUNIO, 2020

INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA
DIVISIÓN DE ESTUDIOS PROFESIONALES
AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA

FECHA: 08 de Junio de 2020.

ASUNTO: **AUTORIZACIÓN DE IMPRESIÓN DE TESIS.**

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

MARCELA LARA CRUZ

estudiante de la maestría en SISTEMAS COMPUTACIONALES con No. de Control 172T0801 ha cumplido satisfactoriamente con lo estipulado por el **Lineamiento de Posgrado para la obtención del grado de Maestría** mediante **Tesis.**

Por tal motivo se **Autoriza** la impresión del **Tema** titulado:

ESTUDIO COMPARATIVO DE ALGORITMOS DE CLASIFICACIÓN DE APRENDIZAJE AUTOMÁTICO EN LA DETECCIÓN DE EMFERMEDADES DEL CORAZÓN

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

ATENTAMENTE

DIVISIÓN DE ESTUDIOS PROFESIONALES

DR. EDDY SÁNCHEZ DE LA CRUZ
Presidente

MSC. GALDINO MARTÍNEZ FLORES
Secretario

DR. LUIS CARLOS SANDOVAL HERAZO
Vocal

Archivo.

VER. 01/03/09

F-SA--39

Agradecimientos

A Dios gracias por permitirme llegar hasta este momento, por todos los medios que pones en mi vida para poder realizar mis metas, por mi familia que siempre me motiva y apoya con sus consejos, valores y enseñanza que a base de esfuerzo y dedicación todo es posible.

Gracias al Consejo Nacional de Ciencia y Tecnología (CONACYT), por la beca otorgada para cursar la maestría, y así hacer posible el trabajo presentado en esta tesis. No. CVU 864317.

A mi director de tesis Dr. Eddy Sánchez de la Cruz, mi co-director Dr. Rajesh Roshan Biswal, al maestro Galdino Martínez Flores coordinador del posgrado por su apoyo y estar pendiente en las etapas de este proceso educativo.

También agradezco a los profesores que dedicaron su tiempo y conocimiento en compartirlo en cada una de sus clases.

Muchas gracias a todos los que de alguna manera contribuyeron en esta etapa de mi formación académica.

Dedicatoria

Con inmenso amor a ***mis padres***
Ignacio y Marcela
Por que siempre han sido una de mis mayores motivaciones para seguir superándome, gracias por sus consejos, su amor y apoyo incondicional.
A mis ***hermanos***
Ignacio, Belén, Lisset y Lourdes
por ser parte de mi vida y la alegría de tenerlos me motiva a ser un buen ejemplo para ustedes como su hermana mayor.
A mi ***hijo***
Iraham Jesús
por que desde que llegaste a mi vida eres mi luz y lo que me da fuerzas para seguir adelante en cualquier circunstancia, eres mi motivación para ser una mejor persona en todos los sentidos, intentando darte siempre un buen ejemplo, gracias hijo por darme tanta felicidad e impulso a mi vida.
A mi ***esposo***
Irahan
por ser un excelente compañero de vida, gracias por tu amor, comprensión, apoyo y motivarme siempre para seguir adelante ante cualquier dificultad.
A mi ***mis amigos***
Duanny, Antonio, Iván, Alexis y Caroline por haberse convertido en algo más que compañeros, compartimos momentos muy amenos y otras veces no tanto pero que nos ayudaron a formar una amistad que quedara para siempre.
Gracias de todo corazón a todas aquellas personas que de una u otra manera me apoyaron e impulsaron para poder realizar este logro.

Resumen

La presente tesis de investigación esta enfocada en la utilización de las diferentes técnicas de clasificación que ofrece el aprendizaje automático, aplicadas a problemas de salud relacionadas con enfermedades cardíacas. Al utilizar un conjunto de datos con las características de los síntomas principales que pueden ser causa de alguna enfermedad cardiovascular, se pretende categorizar al paciente como sano o enfermo mediante modelos de clasificación, para ello se utilizaron algoritmos de aprendizaje supervisado que crean funciones a partir de un conjunto de ejemplos de los que conocemos la salida deseada , en este caso se utilizan variables categóricas enumeradas es decir 0 significa no enfermo y 1 enfermo, enfocado a esto la presente investigación trabaja con modelos de clasificación ya que son los que buscan encontrar un sistema capaz de identificar automáticamente para cada objeto la clase a la cual pertenece. Los modelos de clasificación planteados para lograr los objetivos de la investigación son: máquina de soporte vectorial (SVM), regresión logística, random forest y vecino mas cercano (KNN), siendo este ultimo con el que se logro una mayor precisión en la clasificación. El aprendizaje automático es un subcampo de la inteligencia artificial el cual actualmente esta siendo utilizado de manera muy notable en varias áreas que maneja el ser humano, como son la educación, el entretenimiento, investigaciones científicas, vehículos autónomos, robótica y en la salud, por mencionar solo algunas. De esta manera en la vida cotidiana cada día esta mas presente y con mas auge ya que ayuda a simplificar varias actividades del ser humano mediante las herramientas tecnológicas que existen hoy en día al alcance de la mayoría de la población.

Palabras clave: Aprendizaje automático, algoritmos, enfermedad cardíaca, clasificación.

Abstract

This research thesis is focused on the use of different classification techniques offered by machine learning, applied to health problems related to heart disease. By using a data set with the characteristics of the main symptoms that may be the cause of some cardiovascular disease, the objective is to categorize the patient as healthy or ill using classification models, for this supervised learning algorithms are used that create functions from a set of examples of which we know the desired output, in this case enumerated categorical variables are used, that is 0 means not sick and 1 sick, focused on this the present investigation works with classification models since they are those that seek to find a system capable of automatically identifying for each object the class to which it belongs. The proposed classification models to achieve the research objectives are: support vector machine (SVM), logistic regression, random forest and k-nearest neighbors (KNN), being the latter with which greater precision in the classification was achieved.

Machine learning is a subfield of artificial intelligence which is currently being used in a very remarkable way in various areas that human beings manage,such as education, entertainment, scientific research, autonomous vehicles, robotics and health, to name just a few. In this way, in everyday life, each day is more present and with more boom since it helps to simplify various activities of the human being through the technological tools that exist today within the reach of the majority of the population.

Keywords: Machine learning, algorithms, heart disease, classification.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Generalidades

1.1. Introducción

El presente documento de tesis refiere un estudio comparativo de algoritmos de aprendizaje automático para la detección de enfermedades del corazón que constituyen una de las causas más importantes de muerte prematura en todo el mundo. Para mejorar la exactitud en la predicción del riesgo cardiovascular se requiere la evaluación y el tratamiento de múltiples factores presentes en este tipo de padecimientos cardíacos[1]. En la actualidad las enfermedades cardiovasculares (ECV) se han visto como una de las principales causas de decesos a nivel mundial, las estadísticas de la Organización Mundial de la Salud (OMS) muestran que las enfermedades relacionadas con el corazón son las responsables de la perdida de vida de 17,7 millones de personas cada año, esto equivale al 31 % de todas las muertes en el mundo[2] . Las enfermedades cardiovasculares ocurren cuando la grasa y el colesterol se acumulan en las paredes del vaso sanguíneo (arteria) las cuales pueden ser clasificadas en [4]:

- cardiopatía coronaria (infarto de miocardio).
- enfermedad cerebrovascular (apoplejía).
- enfermedad vascular periférica.
- hipertensión arterial(presión alta).
- insuficiencia cardíaca.
- miocardiopatías.
- cardiopatía congénita.
- cardiopatía reumática.

La mayoría de estas enfermedades pueden ser diagnosticadas con anterioridad, no obstante la población no cuenta con una educación de prevención y la mayoría de las veces el área médica no posee la tecnología adecuada para realizar un diagnóstico temprano. Es por ello que este trabajo de investigación, parte de la utilización de las técnicas de aprendizaje automático, el cual es una rama de la inteligencia artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras

aprender por si solas, mediante un conjunto de datos. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada. Como tema de estudio se aborda el estudio y análisis de diferentes algoritmos de clasificación que ofrece el aprendizaje automático, ya que al utilizar estas técnicas se pueden encontrar patrones ocultos en los datos, y a partir de ello poder lograr el desarrollo de modelos que logren ser de utilidad en esta área de salud[5].

El aprendizaje automático es un conjunto de técnicas para analizar datos mediante algoritmos, aprender de ellos y hacer predicciones o clasificar nuevos datos que se dan al programa. La diferencia principal con otros algoritmos de análisis de datos es que en el aprendizaje automático, los algoritmos aprenden por si solos a partir de los datos. Los algoritmos convencionales se programan con una secuencia de condiciones y el trabajo del programa es ver cuando esas condiciones deben realizarse satisfactoriamente. Utilizando el aprendizaje automático se pueden tomar grandes cantidades de datos y el programa los analizará, aprenderá y finalmente los clasificará según sus características principales por si solo[6].

Estas técnicas pueden ser catalogadas como:

- Regresión: Intentan predecir un valor real.
- Clasificación (binaria o multiclase): intentan predecir la clasificación de objetos sobre un conjunto de clases prefijadas.[7].

Modelos desarrollados a partir de estas técnicas serán de gran utilidad para caracterizar el riesgo cardiovascular, permitiendo tomar decisiones eficaces. Los algoritmos de aprendizaje automático están dando buenos resultados en el área de cardiología, gracias a los datos y cómputo. Estos dos elementos permiten que estos algoritmos aprendan conceptos por sí solos. Se trata de ese conjunto de reglas abstractas que por sí solas son construidas, lo que ha traído y permitido que se puedan autoconfigurar.

El aprendizaje automático, en unión con los avances en el área médica da paso a que se automaticen procedimientos que anteriormente se tenían que realizar obligatoriamente de manera manual, hoy en día con el manejo de datos se pueden desarrollar modelos que garanticen una predicción más exacta y de pronto análisis[5].

1.2. Planteamiento del problema

En la actualidad los problemas de salud causados por las enfermedades cardiacas reflejan en las estadísticas un alto porcentaje de decesos a nivel mundial. En México el 20 % del total de muertes fue a causa de estas enfermedades, de acuerdo con la información recopilada por el Instituto Nacional de Estadística y Geografía(INEGI) como se muestra en la figura 1.1[2].

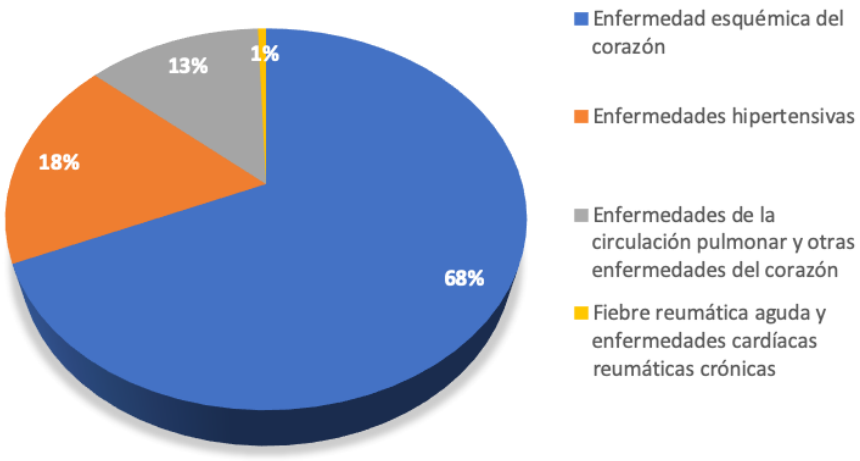

Figura 1.1: Decesos por enfermedades cardiovasculares en 2015 [2].

Dada esta situación es de vital importancia el poder trabajar en medidas preventivas y de detección temprana para minimizar los riesgos relacionados con problemas de las enfermedades cardiacas, las cuales se presentan como una enfermedad progresiva grave, que por lo general se detectan en una etapa avanzada, dejando pocas opciones para dar un tratamiento adecuado.

En la actualidad el área de la salud tiene la facilidad de recopilar grandes volúmenes de datos, pero no cuentan con las técnicas computacionales adecuadas para el procesamiento de esta información, es por ello que se realiza un análisis comparativo de las diferentes técnicas que ofrece el aprendizaje automático.

Una vez planteado el problema a investigar surge la necesidad de realizar un estudio comparativo para la predicción de enfermedades cardíacas aplicando el aprendizaje automático y de esta manera analizar las variables representativas que nos aportaran información relevante para determinar el estado de riesgo del paciente a sufrir alguna enfermedad cardíaca.

1.3. Justificación

Actualmente las nuevas tecnologías, relacionadas con nuestro entorno, están agilizando, automatizando y optimizando diferentes actividades que benefician la salud. Se propone a través del estudio, la exploración y la comparación de las diferentes técnicas que contiene el aprendizaje automático y ser referencia al sector salud con el análisis de bases de datos clínicas, obtenidas de pacientes con padecimientos cardíacos. Al recolectar los datos de cada paciente se observa que existe un problema real en la capacidad de procesar grandes cantidades de datos, los cuales generan en estas áreas médicas, sobre todo en las de especialidades, a este problema se le puede tratar con la utilización de las técnicas del aprendizaje automático para poder contribuir con la determinación de modelos que pueden llevar a encontrar patrones escondidos sobre los datos partiendo de la experiencia de los médicos y las aportaciones de la literatura.

Este proyecto se elabora dado que las enfermedades cardíacas representan un alto índice de fallecimientos, esto abre la oportunidad de poder analizar con los diferentes algoritmos que contiene el aprendizaje automático visualizaciones de datos, para ayudar a predecir tempranamente y así minimizar el riesgo de la salud en los pacientes. El análisis inteligente de datos, brinda la oportunidad de contar con amplias posibilidades para realizar predicciones en las enfermedades cardíacas y así poder contribuir al cuidado de la salud.

Las personas por lo general asisten a un centro de salud cuando presentan alguna enfermedad, no se tiene la cultura de acudir para realizarse exámenes de medicina preventiva. Se necesitan desarrollar acciones rápidas para incrementar y acelerar la

integración del conocimiento en bases de datos sobre los factores que provocan las enfermedades cardiovasculares.

Al obtener una base de datos que nos aporte información útil y adecuada para este tipo de problema que se esta abordando, se procederá a realizar la aplicación de las diferentes técnicas de clasificación que aporta el aprendizaje automático, logrando así un resultado aceptable que sea de apoyo para está área de la salud.

Aplicar inteligencia artificial en el área de la salud cada día se tiene más impacto, ya que se generan grandes cantidades de datos, por ende da como resultado la elaboración de sistemas inteligentes para poder procesar toda esta información y darle un uso adecuado que sea de apoyo para los médicos especialistas.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un estudio comparativo de algoritmos de clasificación de aprendizaje automático para detectar si un paciente esta sano o enfermo de alguna enfermedad cardíaca.

1.4.2. Objetivos específicos

- Comparar la capacidad de los algoritmos de clasificación con aportaciones del estado del arte.
- Visualizar el rendimiento de las diferentes técnicas de predicción de datos.
- Elaborar un análisis experimental de los diferentes algoritmos hacer una clasificación de cuales de estos arrojan mejores resultados respecto a la predicción de enfermedades cardíacas.
- Determinar el mejor algoritmo de acuerdo al porcentaje de clasificación que se visualice, con respecto al dataset de enfermedades del corazón.

1.5. Hipótesis

Utilizando una base de datos y algoritmos de aprendizaje automático es posible encontrar la mejor clasificación para detectar enfermedades cardíacas, y que estos algoritmos sean competitivos en comparación con lo encontrado en el estado del arte.

1.6. Propuesta de solución

Se plantea una solución a través del uso del aprendizaje automático utilizando los algoritmos de clasificación. Ocupando una base de datos que contenga la información necesaria sobre pacientes con padecimientos de enfermedades cardíacas, basándose en esta información se realiza un modelo de predicción implementando la programación en Python y el uso de la herramienta Weka, para extraer las características representativas y poder visualizar los algoritmos más óptimos, con el mejor porcentaje de clasificación ante este tipo de padecimientos cardíacos y así poder realizar un aporte a la investigación en el área de la salud.

Los modelos de clasificación y las técnicas del aprendizaje automático permite obtener mejores predicciones en el área de la salud. Mediante la comparación de algoritmos se pueden medir determinados factores de riesgo en los pacientes, conforme a su edad, sexo, sus de niveles de presión arterial y colesterol, entre otros.

A continuación se describe la elaboración del modelo de clasificación :

1. **Dataset:** El conjunto de datos se obtiene de la base de datos de la fundación clínica de Cleveland. Son de dominio público y están disponibles en línea en el repositorio de aprendizaje automático de UCI (Universidad de California en Irvine) ofrece una colección de bases de datos, teorías de dominio y generadores de datos que son utilizados por la comunidad de aprendizaje automático para el análisis empírico de algoritmos y de aprendizaje automático. El archivo fue creado en 1987 por David Aha y otros estudiantes de posgrado. Desde entonces, ha sido ampliamente utilizado por estudiantes, educadores e investigadores de todo el mundo como fuente principal de conjuntos de datos de aprendizaje

automático. Es uno de los 100 artículos más citados en toda la informática. La versión actual del sitio web fue diseñada en 2007 por Arthur Asunción y David Newman[3]. Está interesado en clasificar a una persona en normal y anormal en relación con problemas de enfermedades cardíacas. Estos datos se dividen en conjunto de entrenamiento y conjunto de prueba, necesarios para la etapa de clasificación.

2. **Preprocesamiento:** Las bases de datos suelen tener datos incompletos, ruido por eso es de gran importancia realizar esta etapa de pre-procesado ya que con este proceso los datos se seleccionan, limpian y transforman, para poder hacer el análisis y que el modelo de clasificación sea de mayor rendimiento.

3. **Selección de características:** Analizar la base de datos para la búsqueda y detección de atributos que sirvan para la clasificación. Ya que al tener un conjunto de datos preparado se ahorra costo computacional y así dar mejor funcionamiento a los clasificadores.

4. **Clasificación:** La clasificación es la manera en que se agrupan los datos en tipos de clases predefinidas de acuerdo a características que poseen en común.

    - Entrenamiento: Son los datos con los que se entrenan los clasificadores para encontrar patrones representativos del conjunto de datos.
    - Prueba: Entrega el error real cometido con el modelo seleccionado. Es decir se evalúan los patrones encontrados por el clasificador y de acuerdo a los resultados que arroje se determina la precisión.

1.7. Alcances y limitaciones

1.7.1. Alcances

Se desarrollara un estudio comparativo de los algoritmos de aprendizaje automático mediante el cual se analizará y clasificará la factibilidad de la detección de enfermedades del corazón con base en las características del conjunto de datos.

1.7.2. Limitaciones

No contar con un dataset basado en información generada en México.

Debido a la complejidad de realizar predicciones de enfermedades cardíacas, con todos los diferentes algoritmos de aprendizaje automático en esta tesis solo se abordaran los algoritmos de clasificación.

1.8. Estructura de la tesis

En este apartado se expresa brevemente como se encuentra estructurado el presente documento de tesis.

Capítulo I se muestra el protocolo de la investigación el cual abarca desde el planteamiento general del proyecto, los objetivos generales y especifico, la justificación, propuesta de solución, los alcances y limitaciones del proyecto.

Capítulo II abarca el marco teórico que es donde se describe las herramientas y los elementos utilizados para llevar a cabo este trabajo.

Capítulo III del documento se muestra una revisión de la literatura relacionada con el tema de investigación.

Capítulo IV muestra la metodología utilizada para desarrollar el presente proyecto.

Capítulo V se dan a conocer los experimentos y resultados obtenidos a través de una serie de pruebas hasta obtener los resultados finales.

Capítulo VI se dan las conclusiones y los trabajos futuros de la investigación realizada.

Capítulo 2

Marco Teórico

2.1. Definiciones y conceptos

2.1.1. Aprendizaje automático

El aprendizaje automático es una área de la inteligencia artificial, el cual va aprendiendo de forma automatizada, lo conforman un conjunto de algoritmos empleados para resolver problemas de toma de decisiones con base a la experiencia que van obteniendo en los casos que han resuelto anteriormente y así mejorar su actuación.

Dentro de este modelo de aprendizaje se encuentra el aprendizaje supervisado, el cual permite buscar patrones en datos históricos relacionando todos los campos con un campo especial, llamado campo o clase objetivo.

Uno de los usos más extendidos del aprendizaje supervisado consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados[10].

2.1.2. Tipos de aprendizaje automático

1. Aprendizaje supervisado

   En el aprendizaje supervisado, la máquina se enseña con el ejemplo. El operador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocidos que incluye las entradas y salidas deseadas, y el algoritmo debe encontrar un método para determinar cómo llegar a esas entradas y salidas. El operador conoce las respuestas correctas al problema, el algoritmo identifica patrones en los datos, aprende de las observaciones, hace predicciones y es corregido por el operador, este proceso sigue hasta que el algoritmo alcanza un alto nivel de precisión y rendimiento.

2. Aprendizaje no supervisado

   El algoritmo de aprendizaje no supervisado estudia los datos para identificar patrones. La máquina determina las correlaciones y las relaciones mediante el análisis de los datos disponibles. En un proceso de aprendizaje no supervisado, el algoritmo de aprendizaje automático es el que interpreta grandes conjuntos de datos y dirige esos datos en consecuencia. El algoritmo intenta organizar

esos datos de alguna manera para describir su estructura, teniendo la necesidad de agrupar los datos en grupos de manera que se vean más organizados [11].

En la siguiente figura 2.1 se muestra los tipos de aprendizaje y la manera en que esta dividido

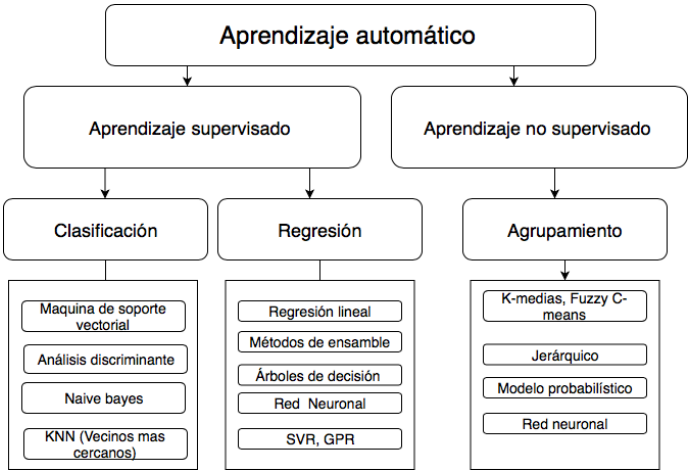

Figura 2.1: Técnicas del aprendizaje automático [9].

3. Aprendizaje por refuerzo

   El aprendizaje por refuerzo enseña a la máquina a través del proceso de ensayo y error. Aprende de experiencias pasadas y comienza a adaptar su enfoque en respuesta a la situación para lograr el mejor resultado posible. Se centra en los procesos de aprendizajes reglamentados, en los que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales. El algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo [12].

Aprendizaje profundo (Deep learning)

El Deep Learning o aprendizaje profundo se define como un algoritmo automático estructurado o jerárquico que emula el aprendizaje humano con el fin de obtener ciertos conocimientos. Destaca porque no requiere de reglas programadas previamen-

te, sino que el propio sistema es capaz de aprender por sí mismo para efectuar una tarea a través de una fase previa de entrenamiento.

También se caracteriza por estar compuesto por redes neuronales artificiales entrelazadas para el procesamiento de información. Se emplea principalmente para la automatización de análisis predictivos.

Los algoritmos que componen un sistema de aprendizaje profundo se encuentra en diferentes capas neuronales compuestas por pesos (números). El sistema está dividido principalmente en 3 capas:

1. Capa de entrada (Intup Layer): Está compuesto por las neuronas que asimilan los datos de entrada, como por ejemplo imagen o una tabla de datos.

2. Capa oculta (Hidden Layer): Es la red que realiza el procesamiento de información y hacen los cálculos intermedios. Cada más neuronas en esta capa haya, más complejos son los cálculos que se efectúan.

3. Capa de salida (Output Layer): Es el último eslabón de la cadena, y es la red que toma la decisión o realiza alguna conclusión aportando datos de salida [13].

Funcionamiento del aprendizaje profundo

Los programas de computadora que utilizan el aprendizaje profundo pasan por el mismo proceso que el niño pequeño que aprende a identificar un objeto. Cada algoritmo en la jerarquía aplica una transformación no lineal a su entrada y usa lo que aprende para crear un modelo estadístico como salida. Las iteraciones continúan hasta que la salida ha alcanzado un nivel aceptable de precisión. La cantidad de capas de procesamiento a través de las cuales deben pasar los datos es por ello que es llamado aprendizaje profundo.

Aplicaciones del aprendizaje profundo

El aprendizaje profundo se utiliza actualmente en las herramientas de:

- Reconocimiento de imágenes
- Procesamiento de lenguaje natural
- Software de reconocimiento de voz

  Estas herramientas están empezando a aparecer en aplicaciones tan diversas como automóviles autónomos y servicios de traducción de idiomas.

Los campos específicos en los que se está utilizando actualmente el aprendizaje profundo incluyen los siguientes:

- Experiencia del cliente.
- Generación de texto.
- Aeroespacial y militar.
- Automatización industrial.
- Investigación médica.
- Visión por computador.

2.1.3. Pasos para realizar aprendizaje automático

1. **Recolectar la información**

   Dependiendo de la problemática que se requiera resolver se debe investigar y obtener los datos que se utilizaran para alimentar la máquina. Es muy importante la calidad y cantidad de información que se obtenga ya que impactará directamente en lo bien o mal que funcione nuestro modelo. Se puede utilizar la información de una base de datos ya existente o bien puede ser creada desde cero, estos datos pueden ser obtenidos de múltiples fuentes y en diferentes formatos.

2. **Preparación de los datos**

   Se debe hacer una selección de características, una vez que se elijan estas impactarán directamente en los tiempos de ejecución y en los resultados, también si es necesario se puede hacer . Se debe tener balanceada la cantidad de datos que se tienen para cada resultado (clase), para que sea representativo, ya que si no, el aprendizaje podrá inclinarse hacia un tipo de respuesta y cuando nuestro modelo intente generalizar el conocimiento fallará. También se debe separar los datos en dos grupos: uno para entrenamiento y otro para evaluación del modelo. Se puede fraccionar aproximadamente en una proporción de 80/20 pero

puede variar según el caso y el volumen de datos con los que se cuente. En esta etapa también podemos preprocesar nuestros datos normalizando, eliminar duplicados y hacer corrección de errores.

3. **Elegir el modelo**

   El tipo de modelo a utilizar se elige de acuerdo al objetivo que se tenga, se pueden utilizar algoritmos de clasificación, predicción, regresión lineal, clustering, Deep Learning , bayesiano, etc y puede haber variantes si lo que se procesa son imágenes, sonido, texto, valores numéricos.

4. **Entrenamiento del modelo**

   Se utiliza el conjunto de datos de entrenamiento para ejecutar la máquina, al realizar esta acción se debe de ver una mejora incremental en el modelo que se halla elegido. Es necesario inicializar los pesos de el modelo aleatoriamente, los pesos son los valores que multiplican o afectan a las relaciones entre las entradas y las salidas, estos se van ajustando automáticamente por el algoritmo seleccionado cuanto más se entrena.

5. **Evaluación del modelo**

   Se debe comprobar el modelo creado contra el conjunto de datos de evaluación que contiene entradas que el modelo desconoce y verificar la precisión del modelo ya entrenado. Si la exactitud es menor o igual al 50 % el modelo no será útil . Si se alcanza un 90 % o más se puede tener una buena confianza en los resultados que nos otorga el modelo [14].

2.2. Métodos de aprendizaje automático

1. Métodos de regresión

   En esta categoría de algoritmos, el programa de aprendizaje automático debe estimar y comprender las relaciones entre las variables. El análisis de regresión es útil para predecir productos que son continuos, es decir se enfoca en una variable dependiente y una serie de otras variables cambiantes, lo que lo hace particularmente útil para la predicción y el pronóstico.

   - Regresión lineal simple

   Es un algoritmo de aprendizaje supervisado que se utiliza en aprendizaje automático y en estadística. Es una aproximación para modelar la relación entre una variable escalar dependiente Y y una o mas variables explicativas nombradas X.

   El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables. Se pretende estudiar y explicar el comportamiento de una variable que notamos y, y que llamaremos variable dependiente o variable de interés, a partir de otra variable, que notamos x, y que llamamos variable explicativa, variable de predicción o variable independiente. Para cumplir dicho objetivo, el primer paso que se debe realizar, es representar las observaciones de ambas variables en un gráfico llamado diagrama de dispersión o nube de puntos. A partir de esta representación el se puede especificar la forma funcional de la función de regresión.

   Empleando la formula siguiente:

$$Y = mX + b \tag{2.1}$$

   donde:

   Y es el resultado

   X es la variable

m la pendiente o coeficiente de la recta

b la constante, conocida como el punto de corte con el eje de Y en la gráfica cuando X es igual a cero.

Este tipo de algoritmo debe minimizar el coste de una función de error cuadrático y los coeficientes que correspondan a la recta óptima.

- Regresión lineal múltiple

En este tipo se manejan múltiples variables independientes que contribuyen a la variable dependiente. Se manejan múltiples coeficientes y a su vez es computacionalmente es más compleja debido a las variables añadidas.

Se trabaja incorporando las n variables independientes con su respectivo coeficiente. Para esto,se construye una matriz de correlación para todas las variables independiente y se incluye la variable dependiente.

A partir de esta matriz, se eligen las variables independientes en orden decreciente de valor de correlación y se ejecuta el modelo de regresión para estimar los coeficientes minimizando la función de error. Se detiene cuando no hay mejora destacada en la función de estimación mediante la inclusión de la siguiente característica independiente. Este método aún puede complicarse cuando hay un gran número de características independientes que tienen una contribución significativa al decidir la variable dependiente. Su formula es la siguiente:

$$yi = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \tag{2.2}$$

En donde:

$$x_1, x_2, ...x_k \tag{2.3}$$

son variables independientes o explicativas.

- Regresión logística

Describe y estima la relación entre una variable binaria dependiente y las variables independientes. Este tipo de regresión uno de los algoritmos de aprendizaje automático más simples y más utilizados para la clasificación de dos clases. Es

fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria. La función logística es también llamada función sigmoide. Esta función es una curva en forma de S que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1. Si la curva va a infinito positivo la predicción se convertirá en 1, y si la curva pasa el infinito negativo, la predicción se convertirá en 0. Si la salida de la función Sigmoide es mayor que 0.5, podemos clasificar el resultado como 1 o SI, y si es menor que 0.5 podemos clasificarlo como 0 o NO.

Regresión logística binaria: la variable objetivo tiene solo dos resultados posibles.

Regresión logística multinomial: la variable objetivo tiene tres o más categorías nominales.

2. Método de agrupamiento

   Se utilizan en el aprendizaje no supervisado, y sirven para categorizar datos no etiquetados. Este método no utilizan la información de salida para la capacitación, sino que permiten que el algoritmo defina la salida. En los métodos de agrupación, solo podemos usar visualizaciones para inspeccionar la calidad de la solución, funciona mediante la búsqueda de grupos dentro de los datos[15].

3. Método de reducción de la dimensionalidad

   Los métodos de reducción de dimensionalidad son algoritmos que mapean el conjunto de los datos a subespacios derivados del espacio original, de menor dimensión, que permiten hacer una descripción de los datos a un menor costo. El método de reducción de dimensionalidad mas común es el análisis de componentes principales (PCA), que reduce la dimensión del espacio de características al encontrar nuevos vectores que maximizan la variación lineal. de los datos [16].

4. Método de redes neuronales y aprendizaje profundo

   Los algoritmos de aprendizaje profundo ejecutan datos a través de varias capas de algoritmos de redes neuronales, las cuales pasan a una representación simplificada de los datos a la siguiente capa [14].

Una red neuronal artificial se entiende por unidades dispuestas en una serie de capas, cada una de las cuales se conecta a las capas anexas. Las redes neuronales artificiales se inspiran en los sistemas biológicos, como el cerebro, y en cómo procesan la información.

Aprenden con el ejemplo y la experiencia, y son muy útiles para modelar relaciones no lineales en datos de alta dimensión.

El aprendizaje profundo es una forma de aprendizaje automático que modela patrones de datos como redes complejas y de múltiples capas. Este método de aprendizaje es una de las maneras más utilizadas de modelar un problema, ya que tiene el potencial de resolver problemas difíciles como la visión por computadora y el procesamiento del lenguaje natural.

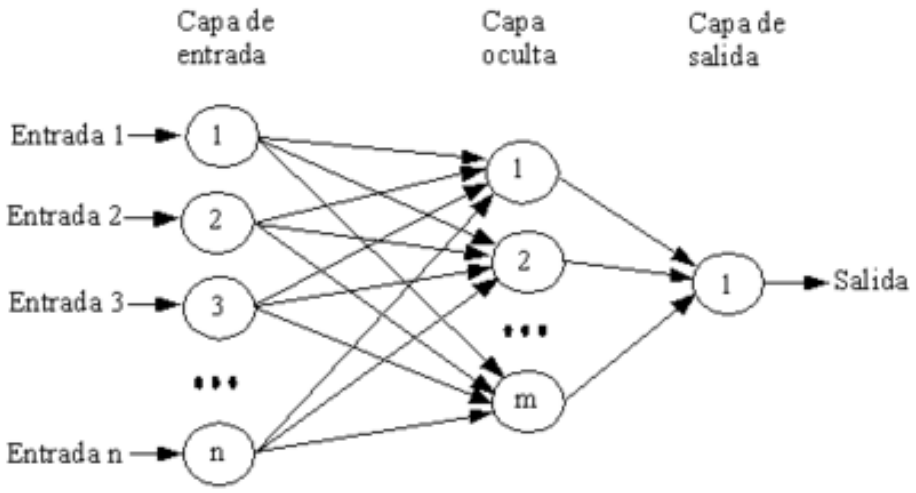

Figura 2.2: Grafo del funcionamiento de una red neuronal[17]

5. Método de clasificación

   Es una subcategoría del aprendizaje supervisado en la que el objetivo es predecir las clases categóricas. La clasificación automática es una colección de algoritmos, ideas y técnicas orientadas a resolver racionalmente el problema general de la clasificación de objetos. La clasificación busca encontrar un sistema capaz de identificar automáticamente para cada objeto la clase a la cual pertenece. Para obtener un modelo de clasificación se necesita que los datos

estén previamente etiquetados, un modelo de clasificación se ajusta usando los datos de entrenamiento.[21].

El método de clasificación predice un resultado de un atributo con valor discreto (a, b, c) dadas unas características (Xo, X1, X2, X3,.... Xn). El método simple de clasificación es el binario, donde se clasifica un registro de variables de entrada en 1 o 0. Un sistema de clasificación predice una categoría, mientras que una regresión predice un número.

Conjunto de datos (dataset)

Es la materia prima del sistema de predicción. Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades.

Instancia, ejemplo o registro

Una instancia es cada uno de los datos de los que se disponen para hacer un análisis

Característica, atributo, factor, propiedad o campo

Son los atributos que describen cada una de las instancias del conjunto de datos[37].

2.3. Algoritmos de Clasificación

Modelos de árbol

Modelos precisos, estables y más sencillos de interpretar básicamente porque construyes unas reglas de decisión que se pueden representar como un árbol. A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas. En estos modelos, destacan los árboles de decisión y los random forest (una media de árboles de decisión). Al ser más precisos y elaborados, obviamente ganamos en capacidad de predicción, pero perdemos en rendimiento.

- Algoritmos de árbol de decisión

Un árbol de decisión es una estructura de árbol similar a un diagrama de flujo que utiliza un método de bifurcación para ilustrar cada resultado posible de una decisión. Cada nodo dentro del árbol representa una prueba en una variable específica, y cada rama es el resultado de esa prueba.

Un árbol de decisiones es el número mínimo de preguntas sí/no que se plantean para evaluar la probabilidad de tomar una decisión correcta, la mayoría del tiempo. Este método permite abordar el problema de una manera estructurada y sistemática para llegar a una conclusión lógica.

Basado en el teorema de bayes su ecuación se presenta con la siguiente fórmula:

$$P(c|x) = P\frac{(x|c)P(c)}{P(x)} \tag{2.4}$$

En dónde:

1. $P(c|x)$ es la probabilidad posterior que deseamos calcular. Está nos indica que la probabilidad de que se tenga una clase c dados los datos en x.

2. P(c) es la probabilidad previa de la clase. Qué tan probable que se obtenga una clase c.

3. $P(x|c)$ es la probabilidad de los datos dada una cierta clase.

4. P(x) es la probabilidad a priori de los datos o predictor [24].

El resultado muestra una apariencia similar a la figura 2.4

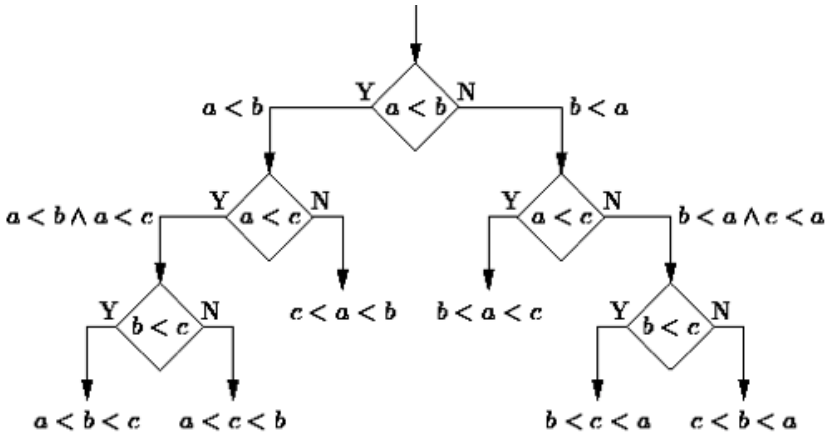

Figura 2.3: Grafo del funcionamiento de un árbol de decisión [25].

Redes neuronales

Las redes neuronales artificiales tratan de replicar el comportamiento del cerebro, donde tenemos millones de neuronas que se interconectan en red para enviarse mensajes unas a otras. Esta réplica del funcionamiento del cerebro humano es uno de los modelos de moda por las habilidades cognitivas de razonamiento que adquieren. El reconocimiento de imágenes o vídeos. Las redes neuronales pueden modificar su comportamiento como respuesta a su entorno. Dado un conjunto de entradas estas se ajustan para producir respuestas consistentes. Una vez entrenada la red neuronal, la respuesta de la red puede ser, hasta un cierto punto, insensible a pequeñas variaciones en las entradas, lo que las hace idóneas para el reconocimiento de patrones, ya que son capaces de abstraer información de un conjunto de entradas. Por ejemplo una red puede ser entrenada en una secuencia de patrones distorsionados de una letra. Una vez que la red sea correctamente entrenada será capaz de producir un resultado correcto ante una entrada distorsionada, lo que significa que ha sido capaz de aprender algo que nunca había visto. [19].

Algoritmo K vecino más cercano (K-NN)

Este tipo de algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecino más cercano de un grupo o de otro. Es así como calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer. Este grupo será, por tanto, el de mayor frecuencia con menores distancias.

Este algoritmo de clasificación es de tipo supervisado ya que el conjunto de datos de entrenamiento ya esta etiquetado, con la clase o resultado esperado.

Basado en Instancia: Es decir que el algoritmo no aprende explícitamente un modelo, lo que realiza es memorizar las instancias de entrenamiento que son usadas como base de conocimiento para ser ocupado en la predicción.

Este tipo de algoritmo es fácil de implementar y funciona muy bien para conjuntos de datos con baja dimensionalidad. El algoritmo K-NN es mucho más rápido que otros algoritmos que requieren entrenamiento tales como la regresión logística o las maquinas de soporte vectorial[22].

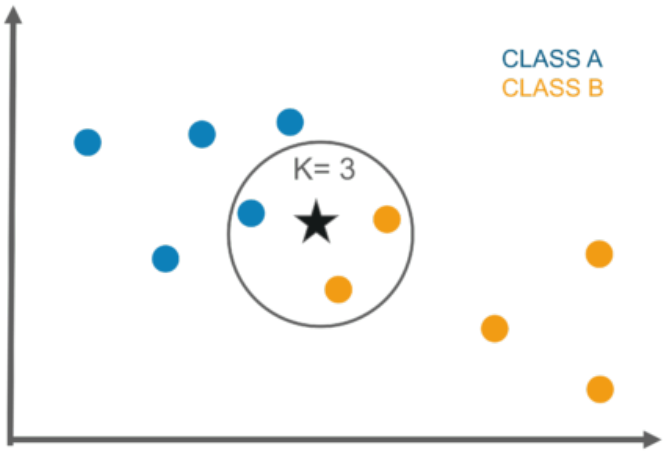

Figura 2.4: Grafo del funcionamiento del algoritmo vecino más cercano(KNN)[18]

Regresión Logística

La regresión logística consiste en obtener una función logística de las variables independientes que permita clasificar a los individuos en una de las dos subpoblaciones o grupos establecidos por los dos valores de la variable dependiente. Se entiende como la probabilidad de que algunos eventos ocurran como una función lineal de un conjunto de variables predictoras.

Para poder calcular este modelo nuestra variable predictora debe ser dicotómica es decir estar entre 0 y 1.

Funcionamiento de la Regresión Logística

Cuando se utiliza regresión logística se debe estimar los parámetros de la ecuación. Su fórmula es la siguiente:

$$(\beta 0, \beta 1, \beta 2, ...\beta k) \tag{2.5}$$

de la función que se pretende evaluar es:

$$Z = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta kXk \tag{2.6}$$

Donde Z es el logaritmo ,el desenlace o el resultado que se está estudiando;

$$\beta 0 \tag{2.7}$$

es la ordenada en el origen de la función de regresión,

$$\beta 1, \beta 2, ...\beta k \tag{2.8}$$

representan los coeficientes de la pendiente de la recta y X1,X2,...Xk son las variables independientes en el caso de las enfermedades del corazón representan el riesgo de padecerla o no. Si nuestros datos se ajustan de manera satisfactoria a este modelo, tendremos la suerte de poder explicar la relación entre las variables independientes y la respuesta de una manera muy sencilla [26].

Maquina de Soporte Vectorial

Las máquinas de soporte vectorial se utilizan para resolver problemas de regresión, agrupamiento y multiclasificación. Su uso abarca varios campos tales como visión artificial, reconocimiento de caracteres, procesamiento de lenguaje natural, análisis de series temporales. Las SVM pertenecen a la categoría de los clasificadores lineales, ya que introducen separadores lineales en el espacio original.

Las máquinas de soporte vectorial(SVM) se pueden usar para:[27].

- Clasificación binaria
- Clasificación multiclase
- Regresión
- Selección de variables
- Identificación de casos anómalos(outliers)
- Agrupamientos (Clustering)

Random forest(Bosque aleatorio)

Este tipo de algoritmo esta construido por una gran cantidad de árboles de decisión ha ensamble. Es un método de aprendizaje para la regresión o clasificación. Es un método de aprendizaje para la regresión o clasificación. El bosque aleatorio recogerá los resultados de la votación de cada árbol de decisión para tomar la decisión al hacer la clasificación. En otro lado, el bosque aleatorio devolverá el valor promedio de los valores de todos los nodos de decisión mientras se hace la regresión [29].

Random forest tiene su base en los arboles de decisión, varios de ellos se juntan hasta formar un bosque aleatorio. Para realizar la clasificación de un nuevo objeto basado en atributos cada árbol realiza una clasificación para finalmente el bosque elige la clasificación con más votos. En la clasificación se usa cuando el resultado deseado es una etiqueta discreta o binaria.

Cada árbol va arrojando un resultado y la respuesta que se repite más veces sera el resultado de la predicción.

De un conjunto de datos se seleccionaron de manera aleatoria varios datos de entrenamiento.

- Con el conjunto de datos seleccionado se desarrolla el algoritmo árbol de decisión para clasificación el cual arrojara un resultado especifico.
- Se selecciona otro conjunto de datos de manera aleatoria del conjunto de entrenamiento y se desarrolla un nuevo árbol de decisión que arrojara un resultado, de manera sucesiva se van creando más arboles formando así lo que es el bosque aleatorio(random forest).
- Teniendo ya todos los arboles de decisión desarrollados se verifica la condición que tuvo más votos, siendo este el resultado final.

En la imagen 2.5 se muestra el método que utiliza el clasificador para generar el modelo random forest.

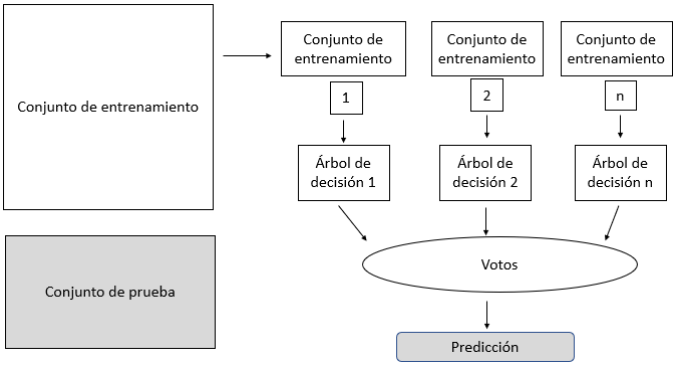

Figura 2.5: Diagrama de clasificación del modelo random forest[30]

2.3.1. Metaclasificadores

Son conjuntos de clasificadores cuyas predicciones individuales son combinadas de alguna manera (típicamente mediante el voto) para clasificar nuevos ejemplos. Los metaclasificadores son empleados para mejorar aún más la precisión en la aplicación de algoritmos en un problema determinado. Esta es una de las áreas mas activas de investigación en aprendizaje supervisado ya que los ensambles son mucho mas precisos que los clasificadores individuales que los componen.

Los ensambles permiten que errores no correlacionados de clasificadores individuales puedan eliminarse por votación mayoritaria.

El objetivo principal es trabajar algoritmos de inducción varias veces y combinar los resultados de alguna forma para obtener un mejor resultado final.

La combinación de varios modelos se puede hacer de diferentes formas, la más común y simple de usar es por voto mayoritario (bagging), realizar un voto pesado (boosting) o usar un nuevo clasificador que decida cómo combinar esos resultados.

El bagging es una de las técnicas de construcción de conjuntos que también se conoce como agregación bootstrap. Dada una muestra de datos, se extraen varias muestras, bootstrapped. Esta selección se realiza de manera aleatoria, es decir, cada variable se puede elegir de la población original, de modo que cada variable es igualmente probable que se seleccione en cada iteración del proceso de arranque.

Boosting es una técnica de aprendizaje secuencial. El algoritmo funciona entrenando un modelo con todo el conjunto de entrenamiento, y los modelos posteriores se construyen ajustando los valores de error residual del modelo inicial. De esta manera, boosting intenta dar mayor peso a aquellas observaciones que el modelo anterior estimó pobremente, sus puntuaciones de precisión y los resultados se combinan para crear una estimación final. Si se realiza una comparación entre estas dos formas de combinación de modelos bagging rara vez obtendrá un mejor sesgo. Sin embargo, boosting podría generar un modelo combinado con errores más bajos, ya que optimiza las ventajas y reduce las dificultades del modelo único.

La desventaja que pueden presentar los metaclasificadores es que los resultados son difíciles de analizar[31].

2.4. Criterios de evaluación de algoritmos

- Validación cruzada

Se parte el conjunto de datos en k subgrupos (k-fold) para después entrenar con k-1 subgrupos de esta manera validar con el restante. Se repite usando cada grupo. Para cada ejemplo es usado el mismo numero para entrenamiento y una vez para pruebas.

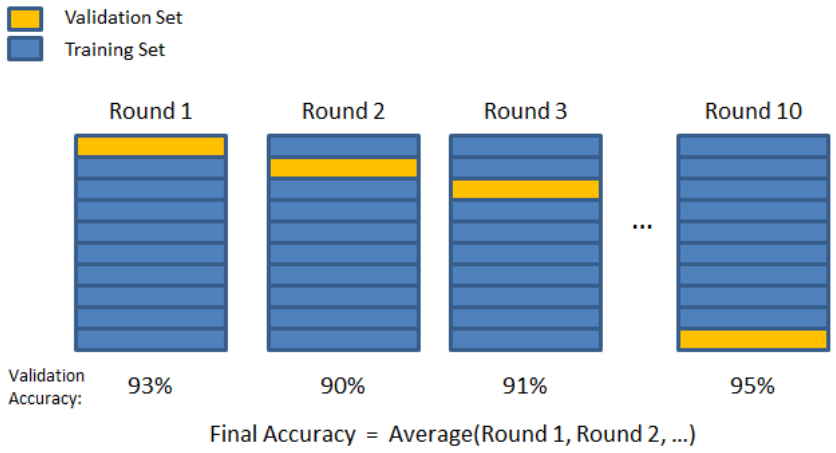

Figura 2.6: Funcionamiento de la validación cruzada con 10 fold [32]

- Criterio percentaje split

Se debe subdividir el conjunto de datos de entrada para entrenamiento en dos: uno para entrenamiento y otro para la validación que el modelo no conocerá de antemano. Esta división se suele hacer del 80 % para entrenar y 20 % para validar. El conjunto de validación deberá tener muestras diversas en lo posible y una cantidad de muestras suficiente para poder comprobar los resultados una vez entrenado el modelo [33].

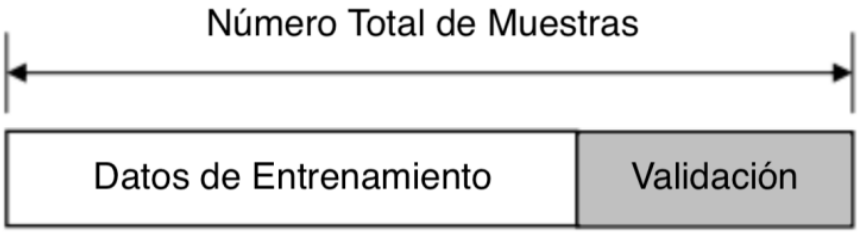

Figura 2.7: Funcionamiento del criterio de evaluación percentage split [33]

2.5. Métricas de evaluación de los clasificadores

2.5.1. Matriz de confusión

Una matriz de confusión es una tabla que a menudo se usa para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba para los que se conocen los valores verdaderos. Permitiendo la visualización del rendimiento de un algoritmo.

El número de predicciones correctas e incorrectas se resume con valores de conteo y se desglosa por clase, esta es la clave de la matriz de confusión. La matriz de confusión muestra las formas en que su modelo de clasificación se confunde cuando hace predicciones. Nos da una idea no solo de los errores que está cometiendo un clasificador, sino más importante aún, de los tipos de errores que se están cometiendo.

	Predicción	Predicción
Clase 1 actual	TP	FN
Clase 2 actual	FP	TN

Tabla 2.1: Matriz de confusión

En donde:

- Clase 1: Positivo
- Clase 2: Negativo

Definición de los términos:

- Positivo (P): la observación es positiva.
- Negativo (N): la observación no es positiva.
- Verdadero positivo (TP): La observación es positiva y se predice que será positiva.
- Falso negativo (FN): La observación es positiva, pero se predice negativa.
- Verdadero negativo (TN): La observación es negativa y se predice que será negativa.
- Falso positivo (FP): La observación es negativa, pero se predice que es positiva [34].

Exactitud

La exactitud de la clasificación es la relación entre las predicciones correctas y el número total de predicciones. Es decir, con qué frecuencia es correcto el clasificador. Es el número de predicciones correctas realizadas por el modelo por el número total de registros.

La exactitud está relacionada con el sesgo de una estimación. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

La exactitud está dada por la relación:

$$Exactitud = \frac{VP+VN}{VP+FP+FN+VN} \tag{2.9}$$

Precisión:

La precisión es la relación entre las predicciones correctas y el número total de predicciones correctas previstas. Esto mide la precisión del clasificador a la hora de predecir casos positivos.

Evalúa los datos por su desempeño de predicciones positivas.

Para obtener el valor de precisión, dividimos el número total de ejemplos positivos correctamente clasificados por el número total de ejemplos positivos predichos.

$$Recall = \frac{VP}{VP+FP} \tag{2.10}$$

Sensibilidad

La sensibilidad también en inglés llamada recall, es la relación entre las predicciones positivas correctas y el número total de predicciones positivas. También se entiende, cuán sensible es el clasificador para detectar instancias positivas. Esto también se conoce como la tasa verdadera positiva.

Indica que la clase se reconoce correctamente.

$$Recall = \frac{VP}{VP+FN} \tag{2.11}$$

Especificidad

Es la tasa negativa verdadera, se calcula como el número de predicciones negativas correctas dividido por el número total de negativos.

$$Especificidad = \frac{VN}{VN+FP} \tag{2.12}$$

Puntaje F1

El puntaje F1 es el promedio ponderado de precisión y sensibilidad. Esta puntuación tiene en cuenta tanto los falsos positivos como los falsos negativos, tomando

como entrada los datos reales de prueba y los datos predichos por el modelo[35].

$$Puntaje \quad F1 = \frac{2*Precisión*Sensibilidad}{Precisión+Sensibilidad} \tag{2.13}$$

2.5.2. Curva ROC

La curva ROC muestra qué tan bien puede distinguir el modelo entre dos cosas, por ejemplo bueno o malo. La curva ROC es necesariamente creciente,muestra la relación que existe entre sensibilidad y especificidad: si se modifica el valor de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad. Si la prueba no permitiera discriminar entre grupos, la curva ROC sería la diagonal que une los vértices inferior izquierdo y superior derecho. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la discriminación fuera perfecta esto se representaría como un 100 % de sensibilidad y 100 % de especificidad, esto depende del entrenamiento de cada modelo.

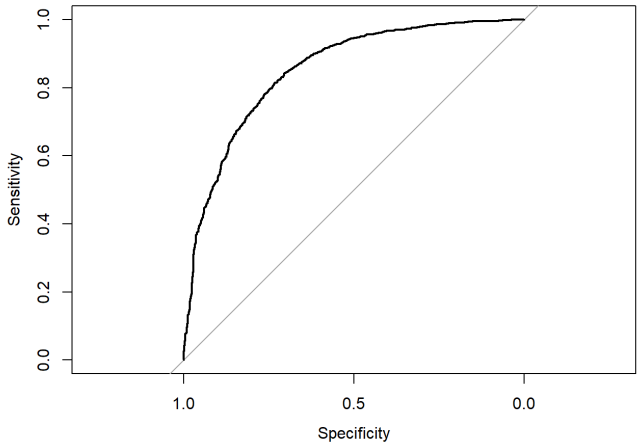

Figura 2.8: Grafo curva ROC [32]

Un parámetro para evaluar la bondad de una prueba diagnóstica que produce resultados continuos es el área bajo la curva llamada AUC. Este área puede interpretarse como la probabilidad de que ante un par de individuos, uno enfermo y el otro sano, la prueba los clasifique correctamente[38].

2.6. Herramientas en las que se trabajaron los algoritmos de clasificación

2.6.1. Weka

Weka es una colección de aprendizaje automático de algoritmos para tareas de minería de datos. los Algoritmos pueden ser aplicados directamente a un conjunto de datos o llamado desde su propio código Java.

Weka contiene herramientas para el procesamiento previo de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización. También es muy adecuado para desarrollar nueva máquina esquemas de aprendizaje[40].
La plataforma Weka aplica varios algoritmos de aprendizaje automático. Para ello se utilizan los siguientes pasos:

1. Abrir el Selector GUI de Weka.
2. Dar clic en el botón Explorador para abrir el explorador de Weka.
3. Abrir un conjunto de datos
4. Hacer clic en Clasificar para abrir la pestaña Clasificar.

La pestaña de clasificación del Explorador es donde puede conocer los diferentes algoritmos y explorar el modelado predictivo.
Al dar clic en el botón Elegir se presenta una lista de algoritmos de aprendizaje automático. Estos a su vez se dividen en varios grupos:

- Bayes : Algoritmos que usan el Teorema de Bayes de alguna manera central, como Naive Bayes.
- Función : Algoritmos que estiman una función, como la regresión lineal.
- perezoso : Algoritmos que usan el aprendizaje perezoso, como k-Vecinos más cercanos.
- meta : Algoritmos que usan o combinan múltiples algoritmos, como Conjuntos.

- misc : implementaciones que no encajan perfectamente en los otros grupos, como ejecutar un modelo guardado.
- Reglas : algoritmos que usan reglas, como las de asociación.
- árboles : algoritmos que usan árboles de decisión, como Random Forest.

Cuando se trabaja con problemas de aprendizaje automático, por lo general no se tiene el algoritmo exacto para utilizar, es por ello que se deben probar con varios algoritmos potentes hasta encontrar el que tenga mejor desempeño.[41]

2.6.2. Phyton

Python es un lenguaje de independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo, lo que ofrece ventajas como la rapidez de desarrollo e inconvenientes como una menor velocidad.

Características del lenguaje Python

- Propósito general

Se pueden crear todo tipo de programas. No es un lenguaje creado específicamente para la web, aunque entre sus posibilidades sí se encuentra el desarrollo de páginas.

- Multiplataforma

Hay versiones disponibles de Python en muchos sistemas informáticos distintos. Originalmente se desarrolló para Unix, aunque cualquier sistema es compatible con el lenguaje siempre y cuando exista un intérprete programado para él.

- Interpretado

Quiere decir que no se debe compilar el código antes de su ejecución. En sí que se realiza una compilación, pero esta se realiza de manera transparente para el programador. En ciertos casos, cuando se ejecuta por primera vez un código, se producen unos bytecodes que se guardan en el sistema y que sirven para acelerar la compilación implícita que realiza el intérprete cada vez que se ejecuta el mismo código.

- Interactivo

Python dispone de un intérprete por línea de comandos en el que se pueden introducir sentencias. Cada sentencia se ejecuta y produce un resultado visible, que puede ayudarnos a entender mejor el lenguaje y probar los resultados de la ejecución de porciones de código rápidamente.

- Orientado a objetos

La programación orientada a objetos está soportada en Python y ofrece en muchos casos una manera sencilla de crear programas con componentes reutilizables.

Funciones y librerías

Dispone de muchas funciones incorporadas en el propio lenguaje, para el tratamiento de strings, números, archivos, etc. Además, existen muchas librerías que podemos importar en los programas para tratar temas específicos como la programación de ventanas o sistemas en red o cosas tan interesantes como crear archivos comprimidos.

Sintaxis clara

Python tiene una sintaxis muy visual, gracias a una notación identada (con márgenes) de obligado cumplimiento. En muchos lenguajes, para separar porciones de código, se utilizan elementos como las llaves o las palabras clave begin y end. Para separar las porciones de código en Python se debe tabular hacia dentro, colocando un margen al código que iría dentro de una función o un bucle. Esto ayuda a que todos los programadores adopten unas mismas notaciones y que los programas de cualquier persona tengan un aspecto muy similar[42].

Principales librerías de python para aprendizaje automático

Las librerías o paquetes que utiliza python son una colección de funciones y métodos que permiten realizar múltiples acciones sin necesidad de escribir demasiado código, ya que contiene módulos integrados que proporcionan diferentes funcionalidades que se pueden usar directamente.

Librerías de datos

- Pandas: Ofrece estructura de datos y herramientas para manipulación y análisis de datos. Es una tabla bidimensional que consiste en etiquetas de columnas y filas llamadas dataframe.

- Numpy Utiliza matrices para sus entradas y sus salidas. En si nos sirve para álgebra lineal.

- SciPy incluye funciones para problemas matemáticos avanzados (integrales, ecuaciones diferenciales, etc)

Librerías de visualización

- Matplotlib: visualizar gráficos.

- Seaborn: Basado en diagramas y series de tiempo.

Librerías de algoritmos

- Scikit-learn: Emplea algoritmos de clasificación , regresión y agrupamiento. Opera de manera simultánea con librerías como NumPy y SciPy.

- TensorFlow: Es recomendable para proyectos grandes y complejos, puesto que aporta mayor control cuando se construyen redes neuronales, ya que es una librería de computación numérica que computa gradientes automáticamente.

- Keras: Diseñada específicamente para hacer experimentos con redes neuronales. Permite crear prototipos rápidamente y de manera fácil.

2.7. Enfermedades Cardíacas

Las enfermedades cardiovasculares (ECV) son un grupo de desórdenes del corazón y de los vasos sanguíneos, entre los que se incluyen[44]:

- Enfermedades cerebrovasculares:

  Enfermedades de los vasos sanguíneos que irrigan el cerebro

- Trombosis venosas profundas y embolias pulmonares:

  Coágulos de sangre (trombos) en las venas de las piernas, que pueden desprenderse (émbolos) y alojarse en los vasos del corazón y los pulmones.

- Arteriopatías periféricas:

  Enfermedades de los vasos sanguíneos que irrigan los miembros superiores e inferiores.

- Cardiopatía reumática: Lesiones del músculo cardiaco y de las válvulas cardíacas debidas a la fiebre reumática, una enfermedad causada por bacterias denominadas estreptococos.

- Cardiopatías congénitas: Malformaciones del corazón presentes desde el nacimiento.

- Cardiopatía coronaria: Enfermedad de los vasos sanguíneos que irrigan el músculo cardiaco.

- Angina de pecho: Es un síndrome clínico que provoca insuficiente suministro de sangre al músculo del corazón. Los síntomas son dolor de garganta, dolor de espalda, náusea, dolor de pecho, ahogo, salivación, sudoración y dolor de opresión. Estos síntomas aparecen con rapidez y se intensifican con el movimiento.

- Infarto cardiaco: Una enfermedad coronaria grave es el infarto con necrosis del miocardio. Primero se desarrolla la estenosis coronaria, queda de repente bloqueado el pericardio , causando fuerte dolor en el corazón. Los síntomas son opresión en el centro del pecho, un dolor que se irradia del pericardio en el pecho, dolores cada vez más frecuentes en el pecho, dolor constante en la

parte superior del estómago, dificultad para respirar, sudoración, sensación de aniquilamiento, desmayo, náusea, vómitos.

- Arritmia: Hablamos del desorden del ritmo cardíaco cuando la coordinación de la actividad eléctrica de los latidos del corazón no funciona correctamente, por lo que nuestro corazón comienza a latir demasiado rápido, demasiado lento o irregularmente. La arritmia puede desarrollarse como consecuencia de la cicatrización del músculo cardiaco después de un infarto cardiaco, enfermedades de válvulas o arterias coronarias del corazón, y espesor anormal de las paredes ventriculares[45].

2.7.1. Factores de riesgo causantes de enfermedad cardíaca

Los factores de riesgo cardiovascular mayormente conocidos son: el tabaco, el colesterol de la sangre, la diabetes, la presión arterial elevada, la obesidad, la falta de ejercicio físico regular (sedentarismo), los antecedentes familiares de enfermedad cardiovascular y el estrés. En la mujer existen factores específicos tales como los ovarios poliquísticos, los anticonceptivos orales y los estrógenos propios. Cuánto mayor sea el nivel de cada factor de riesgo, mayor es el riesgo de tener una enfermedad cardiovascular [46].

- Tabaco

El tabaquismo acelera la frecuencia cardíaca, contrae las arterias principales y puede ocasionar alteraciones en el ritmo de los latidos del corazón. Todo esto hace que el corazón se esfuerce más. Fumar también aumenta la presión arterial, que a su vez aumenta el riesgo de accidentes cerebrovasculares, esto indica que un fumador tiene el doble de riesgo de sufrir infarto de miocardio, que un no fumador [47].

- Colesterol en la sangre

Cuando los niveles colesterol aumentan, normalmente por una alimentación inadecuada alta en grasas saturadas, se transporta demasiado colesterol a los tejidos, de forma que esas moléculas de colesterol se pueden depositar en las arterias, sobre todo las que se encuentran alrededor del corazón y del cerebro, y llegar a taponarlas derivando así graves problemas cardíacos [48].

- Diabetes

Tanto si la producción de insulina es insuficiente como si existe una resistencia a su acción, la glucosa se acumula en la sangre, daña progresivamente los vasos sanguíneos y acelera el proceso de arteriosclerosis aumentando el riesgo de padecer una enfermedad cardiovascular: angina, infarto agudo de miocardio y la muerte cardíaca súbita [49].

- Presión arterial

Es la presión con que la sangre circula por los vasos sanguíneos cuando sale del corazón (tensión arterial sistólica: conocida como presión alta) o cuando el corazón se llena de la sangre que retorna al corazón (tensión arterial diastólica: comúnmente conocida como presión baja). La aparición de hipertensión suele indicar que existe un riesgo cardiovascular [50] .

- Obesidad

Tener sobrepeso o sufrir obesidad incrementa de forma exponencial el riesgo de sufrir una enfermedad cardiovascular. Actualmente el sobrepeso y la obesidad se consideran tan importantes como otros factores de riesgo clásico relacionados con la enfermedad coronaria. El tejido adiposo no sólo actúa como almacén de moléculas grasas, sino que sintetiza y libera a la sangre numerosas hormonas relacionadas con el metabolismo de principios inmediatos y la regulación de la ingesta [51].

- Estrés

El estrés mental induce disfunción endotelial, promueve arritmogénesis, estimula la agregación plaquetaria, aumenta la viscosidad sanguínea por hemoconcentración y estimula factores involucrados en la inflamación. La hiperreactividad cardiovascular de la presión arterial a las pruebas de estrés mental, realizadas en el laboratorio, se ha asociado a un mayor riesgo de desarrollar hipertensión arterial [52] .

- Edad

Con la edad, la actividad del corazón tiende a deteriorarse. Esto puede aumentar el grosor de las paredes del corazón, las arterias pueden endurecerse y perder su flexibilidad, cuando esto sucede, el corazón no puede bombear la sangre tan eficientemente

como antes a los músculos del cuerpo. Las personas mayores tienen un mayor riesgo de sufrir enfermedades del corazón. Aproximadamente 4 de cada 5 muertes debidas a una enfermedad cardíaca se producen en personas mayores de 65 años de edad [53].

- Sexo

Generalmente los hombres tienen un riesgo mayor que las mujeres de sufrir un ataque al corazón, pero esta diferencia es menor cuando las mujeres comienzan la menopausia, ya que investigaciones han demostrado que el estrógeno, una de las hormonas femeninas, ayuda a proteger a las mujeres de las enfermedades del corazón. Pero después de los 65 años de edad, el riesgo cardiovascular es aproximadamente igual en hombres y mujeres [53]

Capítulo 3

Estado del arte

En este apartado del documento muestra la literatura que se encuentran en trabajos previos, relacionados con el tema de las diferentes técnicas del aprendizaje automático utilizadas en la predicción de enfermedades del corazón. Por mencionar algunos de ellos tenemos:

Palaniappan y Awang del Departamento de Tecnología de la Información de la Universidad de Ciencia y Tecnología de Malasia su trabajo muestra un Sistema inteligente de predicción de enfermedades del corazón usando técnicas de minería de datos. El descubrimiento de patrones ocultos y relaciones en los datos a menudo queda sin darles un uso en la tecnología. Las técnicas avanzadas de minería de datos pueden ayudar a remediar esta situación. Esta investigación ha desarrollado un prototipo del Sistema Inteligente de Predicción de Enfermedades Cardíacas (IHDPS, por sus siglas en inglés) utilizando técnicas de minería de datos tales como Árboles de decisión, Naive Bayes y Red neuronal. Los resultados muestran que cada técnica tiene su fuerza única para lograr los objetivos definidos. El uso de perfiles médicos como la edad, el sexo, la presión arterial y el azúcar en la sangre puede predecir la probabilidad de que los pacientes contraigan una enfermedad cardíaca. Permite establecer un conocimiento significativo, por ejemplo, patrones, relaciones entre factores médicos relacionados con la enfermedad cardíaca[54].

Solarte y Soto de la Universidad Tecnológica de Pereira, Pereira, Colombia propone en su trabajo Árboles de decisión en el diagnóstico de enfermedades cardiovasculares. En esta investigación se demuestra empíricamente que es posible diagnosticar la necesidad de administrar fármacos en pacientes con síntomas de enfermedad cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias, mediante la utilización de árboles de decisión con el algoritmo ID3.

Una de las propiedades de esta técnica es que permite una organización eficiente de un conjunto de datos, debido a que los árboles son construidos a partir de la evaluación del primer nodo (raíz) y de acuerdo a su evaluación o valor tomado se va descendiendo en las ramas hasta llegar al final del camino (hojas del árbol), donde las hojas representan clases y el nodo raíz representa todos los patrones de entrenamiento los cuales se han de dividir en clases. La técnica de árbol de decisión conjuntamente con el algoritmo ID3 entrega un conjunto de reglas entendibles que

le permiten al médico tomar la decisión hacerlo de manera rápida[28].

Palacios Pawlovsky Alberto. Este artículo presenta un conjunto basado en distancias para un método KNN(k vecino más cercano) y muestra los resultados de su aplicación en el diagnóstico de enfermedades cardíacas. El conjunto ha sido implementado con dos configuraciones. Una usando tres distancias y otra usando cinco. También agregaron una versión ponderada basada en la precisión promedio que proporciona cada distancia cuando se usa en el método KNN. su conjunto proporcionó una precisión promedio de casi el 85 % para cualquiera de las configuraciones y versiones que fueron probadas con el conjunto de datos UCI Cleveland de la enfermedad cardíaca [55].

Thomas J. y Princy R Theresa. En su trabajo muestran un Sistema de predicción de enfermedades del corazón humano utilizando técnicas de minería de datos Naive Bayes, KNN, algoritmo de árbol de decisión Y red neuronal. En el desarrollo de estas técnicas se observo que la precisión quedo en un 80 % [56].

Boshra Bahrami y Mirsaeid Hosseini Shirvani. El objetivo de esta investigación es evaluar diferentes técnicas de clasificación en el diagnóstico de enfermedades del corazón. Clasificadores como J48 Árbol de decisiones, KNN (vecinos más cercanos),Naive bayes (NB) y SMO se utilizan para clasificar el conjunto de datos. Después de la clasificación, alguna evaluación de desempeño, medidas como precisión, sensibilidad la especificidad, la medida F y el área bajo la curva ROC son evaluados y comparados. Los resultados de comparación mostraron que J48 Árbol de decisión es el mejor clasificador para diagnóstico de enfermedades del corazón en el conjunto de datos existente con una precisión de 83.732 % , K-NN 82.775 %, NB 81.818 % y SMO 82.775 %[57].

Rose y Serna. Estos autores presentan una investigación sobre procesamiento del electrocardiograma para la detección de cardiopatías, presentando como solución un algoritmo para la obtención de los eventos de la señal ECG(electrocardiogramas), para construir un vector de características como entrada y realizar un diagnóstico con una red neuronal. Para este modelo se experimentó con una red neuronal artificial, con retropropagación con gradiente conjugado, escalado de tres capas con 56 neuronas en su capa de entrada, 40 neuronas en su capa oculta y 10 neuronas en su capa de salida. Se realizaron pruebas con más neuronas en la capa oculta pero sin

mejoras significativas, en las pruebas, la precisión aumentó 0.2 % con 100 neuronas en la capa oculta; por tanto, se decidió que el costo beneficio no es significativo[58].

Khateeb Nida et al. En su trabajo del año 2017, trabajo en una investigación de un sistema eficiente de predicciones de enfermedades cardíacas, basado en el algoritmo vecino más cercano(KNN), obteniendo un resultado del 80 % en la precisión de la clasificación [59].

De la Hoz et al. Presentan en su trabajo técnicas de Machine Learning (ml) en medicina cardiovascular. Este estudio está basado en el entrenamiento de un conjunto de datos extraídos del repositorio de Machine Learning con el fin de realizar una comparación de tres técnicas comúnmente utilizadas de Machine Learning, para identificar cuál de ellas es más precisa a la hora de entrenar los datos. Estos datos referentes a enfermedades coronarias, fueron generados por la Universidad de Cleveland en Estados Unidos. Se encuentran en el repositorio de Machine Learning. la regresión logística consigue un resultado desde el punto de vista de precisión del 84,15 %, para la solución de problemas de clasificación de enfermedades cardiovasculares, mientras que las máquinas de soporte vectorial obtienen un nivel de precisión de 82,17 %, lo cual indica que es un método que registra resultados aceptables en el momento de ser utilizado como técnica de clasificación, los árboles de decisión ofrecen un nivel de precisión de 76,56 %, lo que permite deducir que es la técnica menos precisa en el momento del análisis de los datos. La regresión logística ofrece los mayores niveles de precisión, lo cual indica que esta técnica arroja unos resultados de aceptación superior en comparación con las demás técnicas [5].

Ramalingam et al. En este trabajo utilizaron técnicas de aprendizaje automático para ayudar a la industria del cuidado de la salud y los profesionales en el diagnóstico de enfermedades relacionadas con el corazón. Se presenta un estudio de diversos modelos basados en este tipo de algoritmos y técnicas que analicen su rendimiento. Naivebayes, árboles de decisión(DT), random forest (RF) y modelos de conjuntos. Este articulo muestra como los autores hicieron uso de la reducción de la dimensionalidad en los datos realizando una extracción de las características más representativas [60].

Grinenco et al. Estos autores abordan el tema de validación de un modelo de predicción de necesidad de cirugía cardiovascular o cateterismo terapéutico neonatal en fetos con cardiopatías congénitas, para ello plantean validar dicho modelo predictivo realizando un estudio de cohorte de validación, prospectivo y multicéntrico. Para ello realizaron un análisis de regresión logística univariado y multivariado, valoración de calibración del modelo mediante test de Hosmer-Lemeshow, y de discriminación mediante valoración de área bajo la curva ROC (Receiver Operating Characteristic). Los resultados que obtuvieron fueron : En 58 (51,8 %) de 112 pacientes incluidos se requirió TCIN. La adecuación del ajuste del modelo no resultó estadísticamente significativa (p 0,232), y la discriminación fue buena (área bajo la curva ROC 0,833; IC95 %: 0,757-0,909). Para un punto de corte de 0,3 (a partir del cual el riesgo de necesidad de TCIN resultó significativo en el modelo original), hubo sensibilidad de 96,6 %, especificidad de 55,6 %, valor predictivo positivo de 70 % y negativo de 93,8 % [61].

Chaurasia y Pal. El objetivo principal que muestran en su trabajo es informar sobre un proyecto de investigación en el que aprovechan los avances tecnológicos disponibles para desarrollar modelos de predicción para la supervivencia de la enfermedad cardíaca. Utilizaron tres algoritmos (Árbol de clasificación y regresión), ID3 (Iterative Dichotomized 3) y Decision Table (DT) extraídos de un árbol de decisión o clasificador basado en reglas para desarrollar los modelos de predicción utilizando un gran conjunto de datos. También ocuparon métodos de validación cruzada de 10 veces para medir la estimación imparcial [62].

Avellaneda Y Ochoa. En su trabajo presentan como tema de investigación: Implementación de redes neuronales para la detección de la presencia de enfermedades en el corazón, utilizando una base de datos que presenta diferentes características de pacientes, algunos presentan algún tipo de enfermedad del corazón y otros no. Desarrollaron en una primera parte redes neuronales supervisadas, específicamente se implementa un perceptrón multicapa; la segunda parte plantea redes no supervisadas, implementando una red ART2 (Adaptive resonance theory). Épocas, número de neuronas, perceptrón multicapa, redes ART, tasa de aprendizaje, tasa de vigilancia. También se realizó un análisis de PCA (Principal Component Analysis) a la base de datos para reducir el número de entradas a la red neuronal.Los mejores resultados

se obtienen con una tasa de aprendizaje de 0.9, casi independiente del número de neuronas y de la normalización empleada. Para la red ART2, la clasificación de los datos por la red neuronal dependen significativamente de la normalización, llegando a no clasificar si se utilizan valores negativos como entrada.Las redes MLP tiene mejores resultados que las redes ART2, por su porcentaje máximo de clasificación de 70 % se puede decir que este método no es el mejor para la solución del problema presentado. [63].

Kannan y Vasanthi En su documento de investigación tienen como objetivo examinar y comparar la precisión de cuatro algoritmos de aprendizaje automático diferentes para predecir y diagnosticar enfermedades cardíacas mediante los 14 atributos de los conjuntos de datos cardíacos UCI, los algoritmos utilizados son regresión logística 82 %, random forest 80 %, boosting 84 % y maquina de soporte vectorial con un 79 % en su clasificación[64].

Sowmiya y Sumitra del departamento de ciencias de la computación vivekanandha, facultad de artes y ciencias de la mujer. En su trabajo de investigación presentan un estudio analítico de diagnostico de enfermedades del corazón utilizando diferentes técnicas de clasificación. En este trabajo se evaluó el potencial de las nueve técnicas de clasificación de la predicción de la enfermedad cardíaca. arboles de decisión, redes neuronales, naive bayes, SVM, ANN, KNN. El algoritmo que se propuso fue el algoritmo Apriori y SVM (Maquina de Soporte Vectorial) en la predicción de la enfermedad cardíaca. El uso de perfiles médicos tales como la edad, el sexo, la presión arterial, el pecho tipo de dolor, la glucemia en ayunas. Se puede predecir como los pacientes pueden tener una enfermedad cardiaca con base en esto[65].

Muhammad Usman y Nida Khateeb de la universidad tecnológica de auckland presentaron en su investigación un sistema eficiente de predicción de enfermedades cardiacas utilizando la técnica de KNN (vecino más cercanos). En su trabajo emplearon el clasificador KNN (vecino más cercano) para lograr una precisión de aproximadamente el 80 % mediante el uso de 14 atributos. Además, de evaluar varios clasificadores predominantes para resaltar la supremacía del sistema de predicción de la enfermedad cardíaca basado en el clasificador kNN[66].

Campo et al. En su trabajo de investigación presentan un análisis de las enfermedades cardiovasculares se ha convertido en un factor común de investigación, la aplicación de sistemas informáticos inteligentes brindan la posibilidad de identificar de forma anticipada los pacientes que puedan padecer dicha enfermedad, por lo cual se propone en esta investigación la utilización de distintas técnicas de minería de datos como lo son árboles de decisión, las máquinas de soporte vectorial, la regresión logística, el método de NaiveBayes, KNN (vecino más cercanos) y redes neuronales, implementados utilizando un mismo conjunto de datos "Heart Disease Data Set" alojado en el repositorio Machine Learning UCI y bajo un mismo ambiente de prueba, con la finalidad de establecer cuál de las técnicas antes mencionadas logran un mayor porcentaje de precisión a la hora de identificar pacientes que padezcan la enfermedad objeto de estudio; para la realización de las pruebas se utilizó validación cruzada con el fin de seleccionar un porcentaje del conjunto de datos para realizarlas y otro para entrenamiento. Las técnicas que lograron mejores resultados fueron: Regresión Logística y NaiveBayes las cuales alcanzaron un 84 % de precisión, las técnicas de Redes Neuronales, IBK, Máquinas de Soporte Vectorial y Arboles de Decisión obtuvieron porcentajes de precisión inferiores lo cual indica que su desempeño no es el más adecuado para la identificación de este tipo de enfermedad[67].

Capítulo 4

Metodología

En el presente capítulo se aborda la metodología propuesta para dar solución al problema de investigación, de un estudio comparativo de algoritmos de aprendizaje automático para la detección de enfermedades del corazón. Se propone un diseño de algoritmos de clasificación basado en técnicas de aprendizaje automático para analizar y clasificar datos que contienen información de síntomas que pueden causar enfermedades cardiacas.
La metodología que se propone para elaborar este proyecto consta de cuatro fases:

Figura 4.1: Diagrama de la metodología [68]

Fase 1. Análisis de la base de datos original: Los datos iniciales se adquirieron de la base de datos Cleveland del repositorio UCI, contiene 303 instancias y 75 atributos de tipo categórico, entero y real. Para la clasificación se seleccionaron datos que aportaban información más relevante y que sirven mas en el estudio comparativo de algoritmos para la detección de enfermedades del corazón.

Fase 2. Pre-procesamiento de los datos: El realizar esta etapa es fundamental ya que al hacerlo obtenemos calidad en nuestros datos. Para poder llevar a cabo el presente proyecto es de gran importancia hacer una depuración y limpieza de los datos, para trabajar con este dataset se seleccionaron un total de 14 atributos estos son los que representan la información más relevante.
En esta etapa se realizan las siguientes tareas:

- Selección: Elegir las características o atributos que sirvan para diseñar el modelo de clasificación. Al contar con un conjunto de datos que contenga la información de mayor relevancia, es de gran ayuda ya que se ahorra costo computacional al omitir características que son de menor relevancia, y por tanto puede dar un mejor resultado en la ejecución del clasificador.

- Transformación: Los datos originales por lo regular traen ruido, están incompletos o son inconsistentes debido a ello se les debe aplicar algún tratamiento o transformación de datos.

- Limpieza de datos: Consiste en resolver inconsistencias en los datos, completar datos faltantes, eliminar valores atípicos, disminuir el ruido y tratar la redundancia de los datos.

Fase 3. Modelo de clasificación: Esta fase nos permite seleccionar los algoritmos que nos ayuden a obtener mejores resultados de clasificación en la detección de pacientes que pueden presentar algún padecimiento cardiaco. Se agrupan los datos en clases predefinidas de acuerdo a ciertas características o patrones que tienen en común. En esta revisión de algoritmos se pretende utilizar Regresión Logística, Maquinas de Soporte Vectorial, Arboles de Decisión, Random Forest y Multilayer Perceptron, ya que conforme en lo encontrado en el estado del arte son algoritmos que arrojan mejores resultados utilizando estas características o patrones en los datos de síntomas de enfermedades cardiacas.
Para esta fase se llevan a acabo dos tareas de gran importancia para hacer posible la clasificación.

- Entrenamiento: Se toma una parte del conjunto de datos con el que se entrenara al clasificador para que este a su vez vaya aprendiendo y encontrando patrones representativos del conjunto de datos.

- Prueba: El conjunto restante se utiliza para las pruebas, evalúa los patrones encontrados en el entrenamiento los va comparando y así el clasificador va aprendiendo, una vez los resultados obtenidos se puede determinar la precisión de la clasificación.

Fase 4. Evaluación e interpretación. En esta última fase se presenta un análisis sobre los modelos de clasificación utilizados para la detección de pacientes con padecimientos cardiacos, para posteriormente elaborar una argumentación de los resultados obtenidos por el proyecto durante todo el proceso, ofreciendo argumentos y valoraciones que demuestren la factibilidad de las técnicas utilizadas.

Capítulo 5

Experimentos y Resultados

5.1. Experimentos

En esta fase se exponen las pruebas realizadas utilizando de modelos de clasificación para detectar patrones dentro del conjunto de datos cleveland del repositorio machine learning UCI contando con un total de 14 atributos y 303 registros, los atributos mostrados contienen características representativas de síntomas de enfermedades cardíacas. Con base en los datos mostrados en cada uno de los atributos del conjunto se realizaron diferentes pruebas aplicando métodos de clasificación buscando con ello la mejor precisión a la hora de clasificar a un paciente como sano o enfermo.

Las pruebas mostradas a continuación se seleccionaron conforme a los resultados que arrojaron cada uno de los algoritmos de clasificación ya que lo que se busca en la presente investigación es mostrar la comparación y análisis de los algoritmos de clasificación que tengan un mejor comportamiento en el porcentaje de la clasificación.

Para realizar estos experimentos se utilizo, el lenguaje de programación de Python y la herramienta de Weka para tener una observación más amplia del comportamiento de los clasificadores y contar con más opciones para seleccionar los mejores resultados.

5.1.1. Pruebas con modelos de clasificación en python

En la figura 5.1 se puede observar la clase objetivo que se utilizo con las diferentes técnicas de clasificación. La cual es de tipo binaria representando el número 0 a las personas sanas y el número 1 a las personas que tienen algún tipo de enfermedad cardíaca.

Figura 5.1: Clase objetivo

Las dos clases no son exactamente del 50 % en cada una, pero la relación es lo suficientemente buena para elaborar las pruebas sin perder o aumentar los datos.

Los coeficientes de correlación son una forma de medir la relación entre variables. Todos los coeficientes tienen un valor entre -1 y 1, donde -1 muestra una correlación negativa perfecta, a medida que crece la variable A, la variable B tiende a encogerse y 1 muestra una correlación perfecta. Un coeficiente de correlación 0 muestra ausencia completa de relación.

En la figura 5.2 se muestra la correlación que existe entre los datos del conjunto con respecto a la clase objetivo.

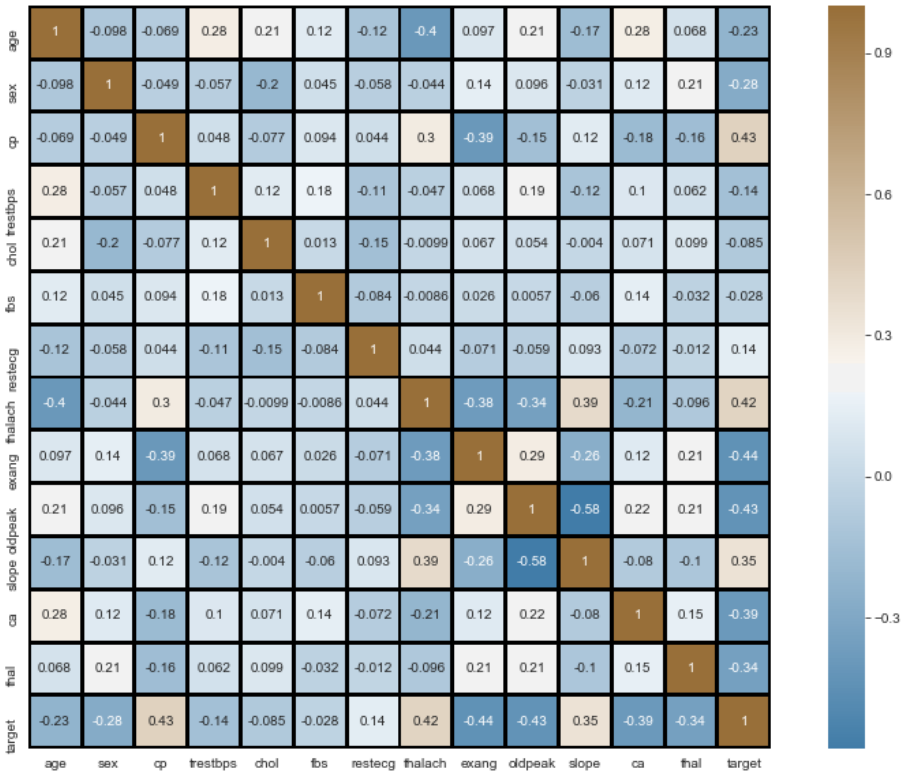

Figura 5.2: Correlación de los datos

Los histogramas que se muestran en la 5.3 se puede ver como cada característica y etiqueta se distribuyen y son útiles para ser aplicadas en las pruebas.

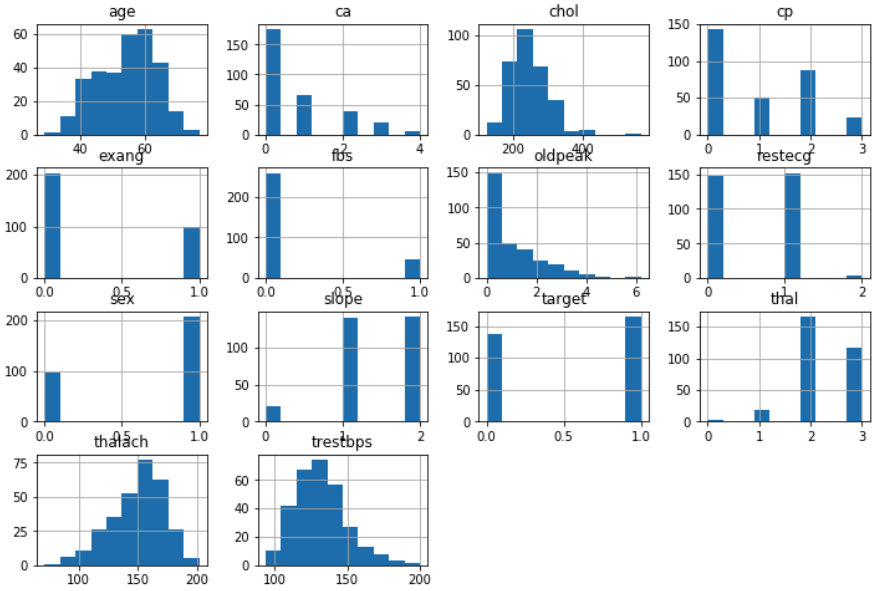

Figura 5.3: Histogramas de los datos

Utilizando aprendizaje automático para predecir si una persona padece o no una enfermedad cardíaca, se trabajo con algoritmos de clasificación los cuales, basado en el estado del arte estos han mostrado resultados muy favorables para este tipo de investigaciones. Después de haber realizado pruebas con los diferentes algoritmos de clasificación que ofrece el aprendizaje automático, se tomaron en consideración los que arrojaron un resultado mayor a un 85 % en la clasificación.

A continuación se muestran los cuatro algoritmos en los cuales se observaron los porcentajes más favorables.

Experimento I Modelo vecino más cercano(KNN)

Este tipo de algoritmo trabaja determinando a que categoría pertenece el nuevo dato que se desea clasificar, seleccionando el número de K vecinos, toma los K vecinos más cercanos al nuevo elemento de acuerdo con la distancia euclidiana, cuenta el numero de elementos entre los K vecinos para seleccionar a que categoría pertenecen y por ultimo asigna el nuevo elemento a la categoría donde se contaron más vecinos.

En esta prueba el modelo de clasificación vecinos más cercanos(KNN) muestra una puntuación la cual varía según los diferentes valores de vecinos que se eligieron. utilizando un numero de 5 k vecinos, aplicando la métrica minkowski, los datos fueron separados para entrenamiento y prueba en un 80/20.

En la tabla 5.1 se muestra la matriz de confusión obtenida a partir del conjunto de datos que utilizo el clasificador KNN, en la diagonal de la matriz se visualiza 26 varaderos positivos, 33 verdaderos negativos, 1 falsos positivos y 1 falsos negativos, esto representa tener buen índice de clasificados correctos.

Predicted true	Sano	Enfermo
Sano	26	1
Enfermo	1	33

Tabla 5.1: Matriz de confusión KNN

De forma mas detallada se observar el reporte de la clasificación en la tabla 5.2 donde se muestra los porcentajes de la capacidad del clasificador para encontrar todas las muestras positivas.

	Precisión	Recall	F1-score	Support
Sano	0.96	0.96	0.96	27
Enfermo	0.97	0.97	0.97	34
prom/total	0.97	0.97	0.97	61

Tabla 5.2: Reporte de clasificación del algoritmo vecino más cercano(KNN)

Experimento II Modelo Random forest (Bosques aleatorios)

El algoritmo random forest calcula la importancia de las características que se utilizan para observar cuanto aumenta la predicción de error cuando los datos son cambiados para esa variable, mientras todos los demás se quedan sin cambios.

En la figura 5.4 se pueden observar todas las características y la importancia que representan dentro del conjunto de datos, es por ello que para realizar este tipo de clasificador se recomienda realizar el análisis de estas clases como se muestra a continuación:

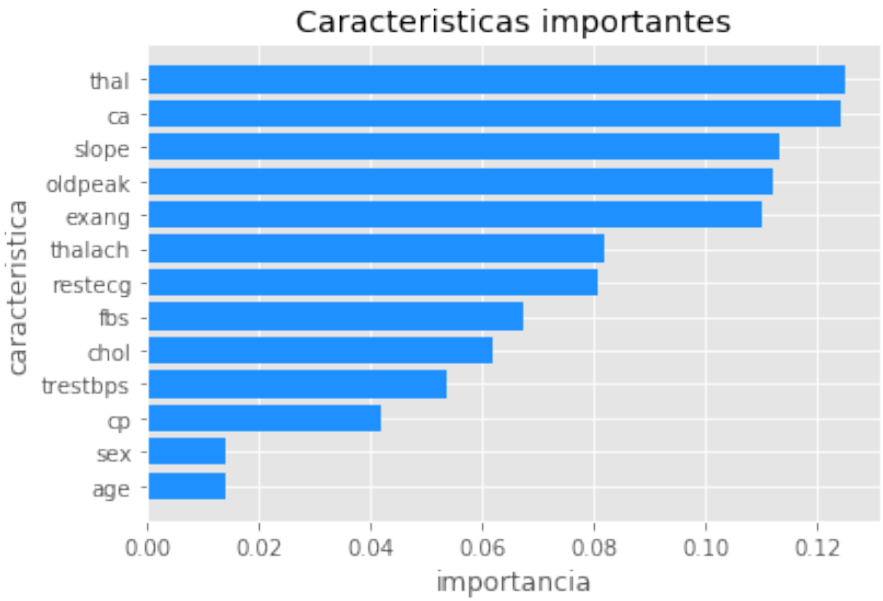

Figura 5.4: Importancia de las características

En la tabla 5.3 se muestra la matriz de confusión la cual fue obtenida a partir del conjunto de datos que se utilizaron con el clasificador bosques aleatorios(random forest), en la diagonal de la matriz se visualiza 27 varaderos positivos, 31 verdaderos negativos, 3 falsos positivos y 4 falsos negativos, esto representa tener buen índice de clasificados correctos, por lo tanto es un resultado favorable en la clasificación.

	Sano	Enfermo
Sano	27	0
Enfermo	3	31

Tabla 5.3: Matriz de confusión del algoritmo random forest

ROC es una curva de probabilidad y AUC representa el grado o la medida de la separabilidad es decir indica la capacidad del modelo para distinguir entre clases. Como se observa en la figura 5.5 el AUC de la curva ROC arroja un 93 %, esto representa la capacidad del modelo para distinguir entre pacientes que se encuentran enfermos o sanos.

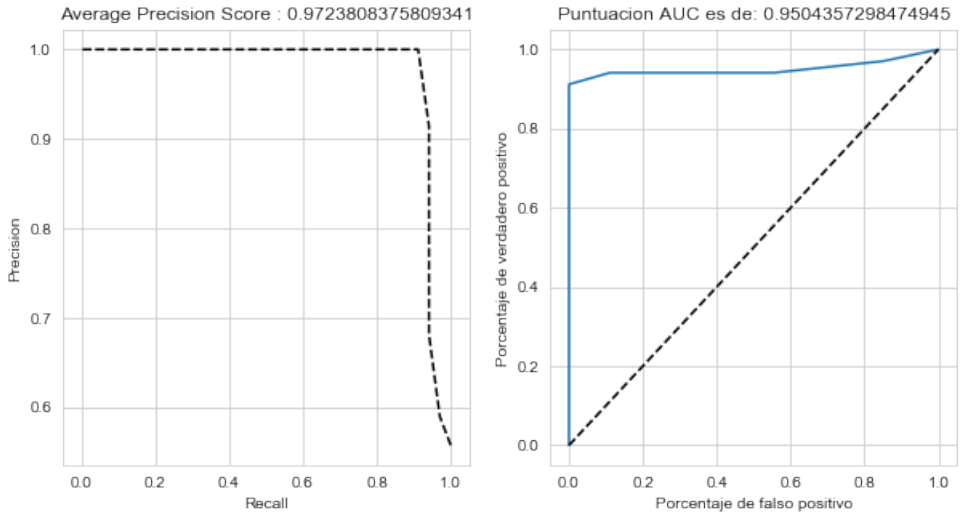

Figura 5.5: Curva AUC del algoritmo random forest

La tabla 5.4 expone la evaluación del desempeño del método random forest de aprendizaje automático obteniendo un 95 %, mostrando la exactitud que se encuentra en la predicción la cual evalúa la eficacia general del algoritmo, entre tanto que la sensibilidad y la especificidad calculan el rendimiento de las clases.

	Precisión	Recall	F1-score	Support
Sano	0.90	1.00	0.95	27
Enfermo	1.00	0.91	0.95	34
prom/total	95	0.95	0.95	61

Tabla 5.4: Algoritmo Random Forest

Experimento III Modelo de regresión logística

Para esta prueba se utilizo la regresión logística binaria, ya que también es uno de los algoritmos que trabaja de forma eficiente con este tipo de conjunto de datos que se esta utilizando.

Este método describe y estima la relación entre la variable objetivo(dependiente) y las variables independientes.

La matriz de confusión 5.5 que arroja este método muestra en su diagonal un resultado de 22 verdaderos positivos y 30 verdaderos negativos dando con esto un porcentaje aceptable para este problema de clasificación, obteniendo un 87 % en la clasificación.

	Sano	Enfermo
Sano	22	5
Enfermo	4	30

Tabla 5.5: Matriz de confusión del algoritmo de regresión logística

La tabla 5.6 se muestra el reporte obtenido de la clasificación:

	Precisión	Sensibilidad	F1-score	soporte
Sano	0.88	0.71	0.79	31
Enfermo	0.75	0.90	0.80	30
prom/total	0.80	0.80	0.80	61

Tabla 5.6: Reporte de clasificación con regresión logística

Experimento IV Máquina de soporte vectorial

Este ultimo experimento se elaboro utilizando el algoritmo máquina de soporte vectorial(SVM) el cual trabaja construyendo un hiperplano en un espacio multidimensional, es uno de los algoritmos de alta precisión el cual se va generando de forma iterativa separando la clase objetivo con un margen que se calcula con la distancia que existe entre la linea hasta los vectores de soporte. Entre mayor sea el margen entre los vectores de soporte en el conjunto de datos dado, este se considera mejor.

Teniendo ya los datos separados en X y Y, se dividen en entrenamiento y prueba el cual quedo en un 80 % para el entrenamiento y un 20 % para prueba, ocupando el parámetro de kernel de función de base radial (RBF).

La matriz de confusión 5.7 muestra que el algoritmo obtuvo un porcentaje aceptable en la clasificación de un 86 % como se observa en la figura 5.6.

Figura 5.6: Porcentaje de clasificación del algoritmo máquina de soporte vectorial

En la matriz de confusión 5.7 se observan 21 varaderos positivos, 32 verdaderos negativos, 6 falsos positivos y 2 falsos negativos.

	Sano	Enfermo
Sano	21	6
Enfermo	2	32

Tabla 5.7: Matriz de confusión del algoritmo maquinas de soporte vectorial(SVM)

En la tabla 5.8 se muestra el reporte de clasificación obtenido del algoritmo maquina de soporte vectorial(SVM)

	Precisión	Sensibilidad	F1-score	soporte
Sano	0.91	0.78	0.84	27
Enfermo	0.84	0.94	0.89	34
prom/total	0.88	0.81	0.81	61

Tabla 5.8: Reporte de clasificación del algoritmo maquinas de soporte vectorial(SVM)

5.1.2. Resultados obtenidos en python

En este apartado se muestra el cuadro 5.9 con los algoritmos y el porcentaje de clasificación que se obtuvo en la fase de pruebas realizadas con la herramienta Jupyter Notebook de el lenguaje de programación interpretado python. Los algoritmos mostrados en la fase anterior utilizan datos de entrenamiento y prueba que permiten a estos modelos aprender de ellos para mejorar de forma gradual el rendimiento predictivo y tomar decisiones basadas en dichos datos.

Como se observa en la tabla 5.9 los algoritmos que dieron un resultado mejor fueron el de vecino más cercano con un 96 % y el de bosques aleatorios con un 95 %.

Modelo	Instancias clasificadas correctas	Instancias clasificadas incorrectas	Precisión
Vecino mas cercano(KNN)	59	2	96 %
Random forest	58	2	95 %
Regresión logística	52	9	87 %
Maquina de soporte vectorial(SVM)	53	8	86 %

Tabla 5.9: Comparación de modelos de clasificación

5.1.3. Resultado con la herramienta weka utilizando meta-clasificadores

Una vez ya realizados los modelos de clasificación y probados con la herramienta de Python, se llevo a cabo una prueba más con el uso de la herramienta weka utilizando validación cruzada, percentaje split y muestra representativa.

En la tabla 5.10 se muestra en la primer columna el tipo de herramienta que se utilizo, en la segunda columna se ubican los clasificadores que se ocuparon para el presente trabajo de investigación, en la tercera, cuarta y quinta columna se observan los criterios de evaluación aplicados en cada uno de los clasificadores.

La tabla 5.10 se puede observar que con la herramienta de Python aplicando validación cruzada el clasificador que arroja una mejor precisión fue la técnica vecino más cercano (KNN) con un 96 % de precisión en la clasificación.

Con el uso de la herramienta de Weka, para probar estos cuatro algoritmos de clasificación, se pudo observar que los resultados arrojados la mejor precisión se obtuvo con los clasificadores multilayer perceptron y J48 la cual fue de 86.1386 % aplicando el criterio de evaluación percentaje split con 66.66 %.

Lenguaje de programación	Algoritmo / ensamblado	Cross-validation 10 folds	Percentage split 66.66 %	Muestra Representativa 170(44)
Python	RandomForest		95 % (80-20)	
Python	K-NN(Vecino mas cercano)		96 % (80-20)	
Python	SVM(Maquina de soporte vectorial)		86 % (80-20)	
Python	Regresión logística		87 % (80-20)	
Weka	RandomForest	80.8581 %	82.1782 %	81.1765 %
Weka	MultilayerPerceptron	77.8878 %	86.1386 %	79.4118 %
Weka	IBK(k-nn) usando 5 k	83.1683 %	78.2178 %	77.6471 %
Weka	SMO(SVM)	83.4983 %	82.1782 %	79.4118 %
Weka	J48 (decisión tree)	78.5479 %	86.1386 %	75.2941 %
Weka	Logistic(Regresión logística)	82.1782 %	83.1683 %	81.7647 %
Weka Meta Classifiers	meta.AdaBoostM1	81.5182 %	81.18.81 %	78.2353 %
Weka Meta Classifiers	meta.AttributeSelectedClassifier	79.538 %	78.2353 %	73.5294 %
Weka Meta Classifiers	meta.Bagging	82.1782 %	79.2079 %	80 %
Weka Meta Classifiers	meta.ClassificationViaRegression	79.538 %	83.1688 %	79.4118 %
Weka Meta Classifiers	meta.CVParameterSelection	54.4554 %	49.505 %	51.7647 %
Weka Meta Classifiers	meta.FilteredClassifier -F	79.538 %	83.1683 %	81.1765 %
Weka Meta Classifiers	meta.IterativeClassifierOptimizer -W	80.5281 %	83.1683 %	70 %
Weka Meta Classifiers	meta.LogitBoost -P	81.1881 %	83.1683 %	80.5882 %
Weka Meta Classifiers	meta.MultiClassClassifier -M	82.1782 %	83.1683 %	81.7647 %
Weka Meta Classifiers	meta.MultiClassClassifierUpdateable -M	83.8284 %	83.1683 %	80.5882 %
Weka Meta Classifiers	meta.MultiScheme -X	54.4554 %	49.505 %	51.7647 %
Weka Meta Classifiers	meta.RandomCommittee -S	79.868 %	81.1881 %	78.8235 %
Weka Meta Classifiers	meta.RandomizableFilteredClassifier -F	67.9868 %	64.3564 %	65.8824 %
Weka Meta Classifiers	meta.RandomSubSpace -P	85.4785 %	85.1485 %	78.8235 %
Weka Meta Classifiers	meta.Stacking -X	54.45545 %	49.505 %	51.7647 %
Weka Meta Classifiers	meta.Vote -S 1 -B	54.4554 %	49.505 %	51.7647 %
Weka Meta Classifiers	meta.WeightedInstancesHandlerWrapper -S 1 -W	54.4554 %	49.505 %	51.7647 %

Tabla 5.10: Comparación de resultados con diferentes clasificadores.

La tabla 5.11 muestra el clasificador multilayer perceptron trabajado en Weka con 500 épocas, se probo con cierto número de épocas para visualizar a partir de que cantidad se obtiene el resultado con mayor precisión. Dando así el mayor porcentaje desde la época 60.

Multilayer perceptron	Iteraciones(epochs)	%
si	10	81.1881 %
si	20	81.1881 %
si	30	81.1881 %
si	40	81.1881 %
si	50	82.1782 %
si	60	86.1386 %
si	70	83.1683 %
si	80	83.1683 %
si	90	84.1584 %
si	100	85.1485 %
si	150	86.1386 %
si	200	85.1485 %
si	250	85.1485 %
si	300	84.1584 %
si	350	85.1485 %
si	400	85.1485 %
si	450	85.1485 %
si	500	86.14 %

Tabla 5.11: Resultados obtenidos con el clasificador multicapa

En la figura 5.7 se exponen el número de épocas o iteraciones del clasificador multilayer perceptron y el porcentaje que arroja en cada una de ellas. Es así como se puede detectar que a partir de la época 60 arroja el mismo resultado que en la época 500, por lo tanto no se es necesario experimentar con más cantidad de épocas.

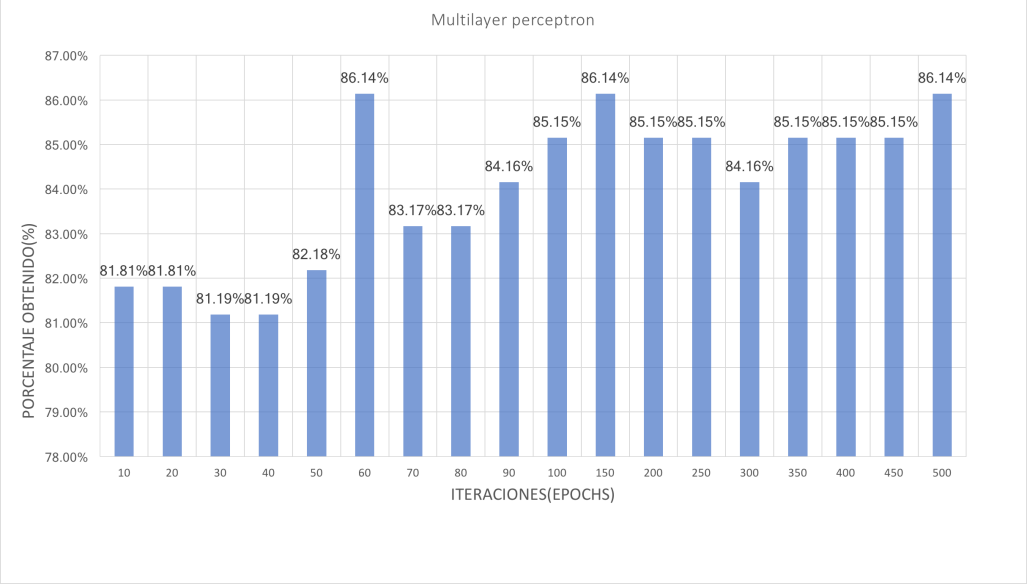

Figura 5.7: Clasificador multilayer comparación de épocas y porcentaje

En la tabla 5.12 se sigue trabajando con la herramienta weka, en este caso se realizo la combinación del clasificador J48 y los metaclasificadores, con la fusión de estos se observa que el porcentaje de precisión llega a 86.1386 % con meta.CVParameterSelection, meta. MultiScheme -X y meta.WeightedInstancesHandlerWrapper.

Lenguaje de programación	Algoritmo / ensamblado	Clasificador	Percentaje split 66.66 %
Weka Meta Classifiers	meta.AdaBoostM1	J48 (decisión tree)	81.1881 %
Weka Meta Classifiers	meta.AttributeSelectedClassifier	J48 (decisión tree)	81.188 %
Weka Meta Classifiers	meta.Bagging	J48 (decisión tree)	84.1584 %
Weka Meta Classifiers	meta.ClassificationViaRegression	J48 (decisión tree)	
Weka Meta Classifiers	meta.CVParameterSelection	J48 (decisión tree)	86.1386 %
Weka Meta Classifiers	meta.FilteredClassifier -F	J48 (decisión tree)	83.16838 %
Weka Meta Classifiers	meta.IterativeClassifierOptimizer -W	J48 (decisión tree)	
Weka Meta Classifiers	meta.LogitBoost -P	J48 (decisión tree)	
Weka Meta Classifiers	eta.MultiClassClassifier -M	J48 (decisión tree)	86.1386 %
Weka Meta Classifiers	meta.MultiClassClassifierUpdateable -M	J48 (decisión tree)	
Weka Meta Classifiers	meta.MultiScheme -X	J48 (decisión tree)	86.1386 %
Weka Meta Classifiers	meta.RandomCommittee -S	J48 (decisión tree)	
Weka Meta Classifiers	meta.RandomizableFilteredClassifier -F	J48 (decisión tree)	64.3564 %
Weka Meta Classifiers	meta.RandomSubSpace -P	J48 (decisión tree)	81.1881 %
Weka Meta Classifiers	meta.Stacking -X	J48 (decisión tree)	49.505 %
Weka Meta Classifiers	meta.Vote -S 1 -B	J48 (decisión tree)	86.1386 %
Weka Meta Classifiers	meta.WeightedInstancesHandlerWrapper -S 1 -W	J48 (decisión tree)	86.1386 %

Tabla 5.12: Resultados obtenidos a partir del uso de diferentes metaclasificadores

5.1.4. Discusión

La propuesta inicial de la investigación fue realizar un análisis de las diferentes técnicas de clasificación para observar y obtener la que arrojara un mejor rendimiento en la clasificación de pacientes con síntomas de enfermedad cardíaca y así determinar si se encuentra sano o enfermo.

Para aplicar estos algoritmos a los modelos de clasificación los experimentos se llevaron a cabo con la herramienta de Python, utilizando la distribución Anaconda en el ambiente Jupyter ya que es un computo interactivo para análisis de datos. Los algoritmos utilizados para el estudio fueron: regresión logística, random forest, vecino más cercano (KNN) y Maquinas de soporte vectorial. Dando como mejor resultado el algoritmo vecino más cercano (KNN) con un 96 % en la precisión.

Con la herramienta de weka se experimento con diferentes metaclasificadores con el objetivo de encontrar los resultados más eficientes, sin embargo el porcentaje mas alto utilizando esta herramienta fue de un 86 % con el metaclasificador multilayer perceptron a partir de 60 épocas.

Partiendo de estos resultados y realizando un análisis los clasificadores más comunes logran tener un mejor desempeño con parámetros adecuados al modelo creando conjuntos óptimos de datos para el entrenamiento, prueba y validación.

Los resultados obtenidos en esta investigación fueron muy favorables en referencia con los del estado del arte como se puede apreciar el mejor resultado obtenido utilizando el conjunto de datos para clasificar si una persona esta sana o enferma de alguna enfermedad cardíaca, logro con el modelo de clasificación vecino más cercano (KNN) con un 96 % de precisión en la clasificación.

A continuación se muestra la tabla 5.13 de la comparación entre el estado del arte y los resultados mostrados en la presente investigación.

Año	Autor	Método	Precisión en la clasificación
2017	Khateeb Nida et al.	Vecino más cercano(KNN)	80 %
2018	Kannan, R., y Vasanthi	Bosque aleatorio	80 %
2020	Lara Cruz Marcela	Vecino más cercano (KNN)	96 %
2020	Lara Cruz Marcela	Bosque aleatorio	95 %

Tabla 5.13: Comparación de resultados de la presente investigación y lo encontrado en la revisión estado del arte

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

En la presente tesis de investigación el objetivo principal fue elaborar un estudio y comparación con las diferentes técnicas de clasificación que se encuentran en el aprendizaje automático, demostrando que este tipo de aprendizaje es aceptable para este tipo de problema médico.

En el área de la salud el aprendizaje automático, específicamente en cardiología puede ser de gran utilidad a la hora de predecir el riesgo cardíaco que pudiera tener un paciente de acuerdo con los síntomas o características que este presente, y poder dar un tratamiento adecuado y de esta manera prolongar la vida del paciente.

En el desarrollo de está investigación se cumplió con el objetivo general el desarrollar un estudio comparativo de algoritmos de clasificación de aprendizaje automático, donde se estudiaron y compararon diferentes técnicas del aprendizaje automático para realizar modelos de clasificación como lo fueron los 4 modelos mostrados: regresión logística, random forest, vecino más cercano (KNN) y maquinas de soporte vectorial. Dando como mejor resultado el algoritmo vecino mas cercano (KNN) con un 96 % en la precisión.

Una vez realizado el objetivo de este estudio se observa como clasificadores sencillos pueden arrojar resultados óptimos para este tipo de problemas de clasificación, dejando así un aporte para la investigación tecnológica que esta enfocada en temas relacionados con el aprendizaje automático aplicado en cuestiones de salud.

Esta investigación permitió visualizar la importancia del aprendizaje automático aplicado en áreas de salud, demostrando que estos modelos de clasificación son herramientas confiables con amplias capacidades de analizar datos de cualquier problema de clasificación ofreciendo resultados eficientes.

Los resultados obtenidos en comparación con el estado del arte fueron muy competitivos y aceptables ya que la mayoría de las técnicas estudiadas y comparadas en esta investigación arrojaron porcentajes desde un 85 % hasta un 96 % en la clasificación.

6.2. Trabajos futuros

El trabajo desarrollado en esta tesis plantea líneas de investigación que pueden ser estudiadas y desarrolladas en un futuro. Entre estas líneas se proponen las siguientes:

1. Ampliar la investigación realizando una evaluación en el comportamiento de las técnicas con un mayor numero de datos.

2. Aplicar las técnicas de clasificación que se encuentran en el aprendizaje automático a otros tipos de enfermedades.

3. Diseñar una aplicación que realice predicciones al generar comunicación entre el paciente y el especialista.

Bibliografía

[1] Vega Abascal, J., Guimará Mosqueda, M., & Vega Abascal, L. (2011). Riesgo cardiovascular, una herramienta útil para la prevención de las enfermedades cardiovasculares. Revista Cubana de Medicina General Integral, 27(1), 91-97.

[2] Consumidor, P. (2018). No rompas más tu corazón. Salud cardiovascular. https://www.gob.mx/profeco/documentos/no-rompas-mas-tu-corazón-salud-cardiovascular?.

[3] Dua, D. y Karra Taniskidou, E. (2017). Repositorio de aprendizaje automático de la UCI [http://archive.ics.uci.edu/ml]. Irvine, CA: Universidad de California, Escuela de Información y Ciencias de la Computación.

[4] Organización Mundial de la Salud. (2018). ¿Qué son las enfermedades cardiovasculares?.

[5] de la Hoz Manotas, A. K., Martínez-Palacio, U. J., y Mendoza-Palechor, F. E. (2013). Técnicas de ML en medicina cardiovascular. Memorias, 11(20), 41-46.

[6] Llodrà Bisellach, G. (2018). Aprendizaje automático para la clasificación de arritmias cardíacas.

[7] Caparrini, F., y Work, W. (2018). Introducción al Aprendizaje Automático - Fernando Sancho Caparrini. Retrieved from http://www.cs.us.es/ fsancho/?e=75

[8] Moreno, A. (1994). Aprendizaje automático.

[9] mathworks. (2018). machine learning. https://la.mathworks.com/discovery/machine-learning.html

[10] Rodríguez, T. (2020). Machine Learning y Deep Learning: cómo entender las claves del presente y futuro de la inteligencia artificial. https://www.xataka.com/robotica-e-ia/machine-learning-y-deep-learning-como-entender-las-claves-del-presente-y-futuro-de-la-inteligencia-artificial

[11] Gonzalo, A. (2020). Tipos de aprendizaje automático. from https://machinelearningparatodos.com/tipos-de-aprendizaje-automatico/

[12] APD España. (2019). ¿Cuáles son los tipos de algoritmos del machine learning?. https://www.apd.es/algoritmos-del-machine-learning/.

[13] SmartPanel. (2019). ¿Qué es el Deep Learning? https://www.smartpanel.com/que-es-deep-learning/.

[14] Bagnato, J. (2019). 7 pasos del Machine Learning para construir tu máquina. Aprende Machine Learning. Available https://www.aprendemachinelearning.com/7-pasos-machine-learning-construir-maquina/.

[15] Roman, V. Aprendizaje No Supervisado En Machine Learning: Agrupación. ¡https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupación-bb8f25813edc¿.

[16] Hernández, J. A. (2016). Métodos de reducción de dimensionalidad: Análisis comparativo de los métodos APC, ACPP y ACPK. Uniciencia, 30(1), 115-122.

[17] Roman, V. Aprendizaje No Supervisado En Machine Learning: Agrupación. ¡https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupación-bb8f25813edc¿.

[18] Edureka. KNN Algorithm Using Python — K Nearest Neighbors Algorithm — Edureka. ¡https://www.edureka.co/blog/k-nearest-neighbors-algorithm/¿.

[19] Rayón, Á. (2018). Guía para comenzar con algoritmos de Machine Learning - Deusto Data. [online] Deusto Data. https://blogs.deusto.es/bigdata/guia-para-comenzar-con-algoritmos-de-machine-learning/

[20] Contreras, F. (2016). INTRODUCCIÓN A MACHINE LEARNING. SUNQU.Contreras, F. (2016). INTRODUCCIÓN A MACHINE LEARNING. SUNQU.

[21] (2019). Retrieved from http://www.stat.rice.edu/ jrojo/PASI/lectures/Costa %20rica/5_Clasif

[22] Flores, D., Ruiz, S., Ruiz, S., Ruiz, S., Peñalver, A., Zabala, A., & Peñalver, A. (2019). El algoritmo K-NN y su importancia en el modelado de datos.Retrieved from https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

[23] Vidaurre, D., Bielza, C., & Larrañaga, P. (2012). Forward stagewise naive Bayes. Progress in Artificial Intelligence, 1(1), 57-69.

[24] nube-programacion. (2017). Clasificador Naive Bayes. 2018, de eenube Sitio web: http://eenube.com/index.php/matematicas/estadistica/99-clasificador-naive-bayes-bayes-ingenuo

[25] http://repositorio.uchile.cl/bitstream/handle/2250/132939/Mejoramiento-de-un-modelo-de-targeting-de-clientes.pdf?sequence=1

[26] Ma. DOLORES FIUZA PÉREZ,J.C.RODRIGUEZ PEREZ. (2000). La regresión logística: una herramienta versátil. 2000, de Revista Nefrología http://www.revistanefrologia.com/es-publicación-nef-articulo-la-regresión-logística-una-herramienta-versátil-X0211699500035664

[27] Betancourt, G. A. (2005). Las máquinas de soporte vectorial (SVMs). Scientia et technica, 1(27).

[28] Solarte Martínez, G. R., y Soto Mejía, J. A. (2011). Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares. Scientia et technica, 16(49).

[29] Chen, MC, Chen, RC, y Zhao, Q. (2018, abril). Combinación de datos de Smartwatch y ambientes para predecir la frecuencia cardíaca. En 2018 Conferencia Internacional IEEE sobre Invención de Sistemas Aplicados (ICASI) (pp. 661-664). IEEE.

[30] Heras, J. Random Forest (Bosque Aleatorio): combinando árboles - IArtificial.net. https://iartificial.net/random-forest-bosque-aleatorio/

[31] Morales, E. (2020). Ensamble de clasificadores. Ccc.inaoep.mx. Available at: https://ccc.inaoep.mx/emorales/Cursos/Aprendizaje2/Acetatos/ensambles.pdf

[32] Ofer, Dan. (2016). Machine Learning for Protein Function.

[33] Bagnato, J. (2019). Qué es overfitting y underfitting y cómo solucionarlo. From https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

[34] Gonzalez, L. (2019). Evaluando el error en los modelos de clasificación - Ligdi González. From http://ligdigonzalez.com/evaluando-el-error-en-los-modelos-de-clasificación-machine-learning/

[35] Gonzalez, L. (2019). Evaluando el error en los modelos de clasificación - Ligdi González.From http://ligdigonzalez.com/evaluando-el-error-en-los-modelos-de-clasificación-machine-learning/

[36] Clavijo Rodríguez, D. L., Bernal Valencia, M., & Silva, J. F. (2006). Sistema inteligente de reconocimiento de enfermedad coronaria (isquemia). Archivos de Medicina (Col), (12).

[37] González, A. and González, A. (2018). Conceptos básicos de Machine Learning Cleverdata.io. https://cleverdata.io/conceptos-basicos-machine-learning/ .

[38] de Ullibarri Galparsoro, L., Fernández, P. (2001). Curvas roc. Atención Primaria en la Red, 5(4), 229-35.

[39] Parra, F. 6 Métodos De Clasificación — Estadística Y Machine Learning Con R. Bookdown.org. ¡https://bookdown.org/content/2274/metodos-de-clasificación.html¿.

[40] Cs.auckland.ac.nz. (2018). https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/I
.

[41] Jason Brownlee. (2016). How to Use Machine Learning Algorithms in Weka. 2016, de Machine Learning Mystery. https://machinelearningmastery.com/use-machine-learning-algorithms-weka/

[42] Alvarez, M. (2018). Qué es Python. DesarrolloWeb.com. https://desarrolloweb.com/articulos/1325.php .

[43] Nicole Chapaval. (2017). Cuatro librerías de Machine Learning: TensorFlow, Scikit-learn, Pytorch y Keras. 2017, de platzi https://platzi.com/blog/librerias-de-machine-learning-tensorflow-scikit-learn-pythorch-y-keras/

[44] World Health Organization. (2018). Enfermedades cardiovasculares. http://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) .

[45] Goldman L.Newby . (2018). enfermedades cardiacas. 2018, de MeddlinePlus https://medlineplus.gov/spanish/ency/patientinstructions/00075

[46] Sans Menéndez, S. (2019). [online] Mscbs.gob.es. http://www.mscbs.gob.es/organizacion/sns/planCalidadSNS/pdf/equidad/07modulo_06.pdf.

[47] Texas Heart Institute. (2019). El tabaquismo y el corazón — Texas Heart Institute. https://www.texasheart.org/heart-health/heart-information-center/topics/el-tabaquismo-y-el-corazon/.

[48] Cachofeiro, V. (2019). [online] Fbbva.es. https://www.fbbva.es/microsites/salud_cardio/mult/

[49] Fundaciondelcorazon.com. (2019). Diabetes y riesgo cardiovascular - Fundación Española del Corazón. https://fundaciondelcorazon.com/prevencion/riesgo-cardiovascular/diabetes.html.

[50] Riojasalud.es. (2019). Qué es la hipertensión arterial (HTA) y qué provoca. https://www.riojasalud.es/ciudadanos/catalogo-multimedia/nefrologia/que-es-la-hipertension-arterial-hta-y-que-provoca?showall=1.

[51] Murillo, A. Z., & Esteban, B. M. (2005). Obesidad como factor de riesgo cardiovascular. Hipertensión y riesgo vascular, 22(1), 32-36.

[52] Armario, P. (2008). Estrés y enfermedad cardiovascular. Hipertensión, 25, 23-34.

[53] Texas Heart Institute. (2019). Factores de riesgo cardiovascular — Texas Heart Institute. [online] https://www.texasheart.org/heart-health/heart-information-center/topics/factores-de-riesgo-cardiovascular/.

[54] Palaniappan, S., y Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

[55] Pawlovsky, AP (2018, enero). Un conjunto basado en distancias para un método kNN para el diagnóstico de enfermedades del corazón. En 2018 Conferencia Internacional sobre Electrónica, Información y Comunicación (ICEIC) (págs. 1-4). IEEE.

[56] Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-5). IEEE.

[57] Bahrami, B., & Shirvani, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2(2), 164-168.

[58] Rose, C., y Serna, M. (2014). Procesamiento del Electrocardiograma para la Detección de Cardiopatías. Memoria del encuentro Nacional de Computación, Ocotlán, Oaxaca.

[59] Khateeb, N., Usman, M. (2017, December). Efficient heart disease prediction system using K-nearest neighbor classification technique. In Proceedings of the International Conference on Big Data and Internet of Thing (pp. 21-26).

[60] V V. Ramalingam,Ayantan Dandapath,M Karthik Raja. (2018). Heart disease prediction using machine learning techniques: A survey. 2018, de researchgate

[61] Grinenco, S., Segovia, M. A., Peña, G., Olmedo, F., Meller, C., Marantz, P., & Izbizky, G. (2016). Validación de un modelo de predicción de necesidad de cirugía cardiovascular o cateterismo terapéutico neonatal en fetos con cardiopatías congénitas. Rev. argent. salud publica, 7(29), 7-13.

[62] Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques.

[63] González, J. A. A., & Rey, C. M. O. (2010). Implementación de Redes Neuronales para la Detección de la presencia de Enfermedades en el Corazón. Redes de Ingeniería, 1(2), 38-46.

[64] Kannan, R., Vasanthi, V. (2019). Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In Soft Computing and Medical Bioinformatics (pp. 63-72). Springer, Singapore.

[65] Sowmiya, C., y Sumitra, P. (2017, marzo). Estudio analítico del diagnóstico de cardiopatías mediante técnicas de clasificación. En Técnicas Inteligentes en Control, Optimización y Procesamiento de Señales (INCOS), Conferencia Internacional de IEEE 2017 (pp. 1-5). IEEE.

[66] Khateeb, N., y Usman, M. (2017, diciembre). Sistema eficiente de predicción de enfermedades del corazón mediante la técnica de clasificación de vecinos K-Nearest. En Actas de la Conferencia Internacional sobre Big Data e Internet of Thing (pp. 21-26). ACM.

[67] Análisis de técnicas de md en diagnóstico de enfermedades cardiovasculares. (2019). Retrieved from https://www.acofipapers.org/index.php/eiei2015/2015/paper/viewFile/1371/478

[68] Diagrama de la metodología. elaboracion propia, Lara Cruz Marcela