Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables

Carrera de Ingeniería en Sistemas

Predicción de enfermedades del corazón utilizando Random Forest y Artificial Neural Networks

Línea de investigación: Sistemas Inteligentes

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS

Autor:

- Bladimir Stanislao Minga Medina

Director y Tutor académico:

- Luis Antonio Chamba Eras

LOJA - ECUADOR

2022

CERTIFICACIÓN DEL DIRECTOR

Ing. Luis Antonio Chamba Eras Mg.SC., PhD.,

DIRECTOR DEL TRABAJO DE TITULACIÓN.

CERTIFICA:

Que el egresado **BLADIMIR STANISLAO MINGA MEDINA**, autor del presente trabajo de titulación, cuyo tema versa sobre "PREDICCIÓN DE ENFERMEDADES DEL CORAZÓN UTILIZANDO RANDOM FOREST Y ARTIFICIAL NEURAL NETWORKS", ha sido dirigido, orientado, discutido bajo mi asesoramiento y ha sido culminado al 100%, reúne a satisfacción los requisitos exigidos en una investigación de este nivel, por lo cual autorizo su presentación y sustentación.

Loja, 14 de marzo del 2022

Ing. Luis Antonio Chamba Eras Mg.SC., PhD.
DIRECTOR DEL TRABAJO DE TITULACIÓN

AUTORÍA

Yo, **Bladimir Stanislao Minga Medina** declaro ser autor del presente Trabajo de Titulación, y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales por el contenido de esta.

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

Autor: Bladimir Stanislao Minga Medina
Correo personal: bladiminga46@gmail.com
Correo institucional: bsmingam@unl.edu.ec

CARTA DE AUTORIZACIÓN

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

Yo, **Bladimir Stanislao Minga Medina** declaro ser autor del Trabajo de Titulación que versa: "**PREDICCIÓN DE ENFERMEDADES DEL CORAZÓN UTILIZANDO RANDOM FOREST Y ARTIFICIAL NEURAL NETWORKS**", como requisito para optar al grado de: **INGENIERO EN SISTEMAS**; autorizo al Sistema Bibliotecario de la Universidad Nacional de Loja para que, con fines académicos, muestre al mundo la producción intelectual de la Universidad a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional:

Los usuarios pueden consultar el contenido de este trabajo en el (RDI), en las redes de información del país y del exterior, con los cuales tenga convenio la Universidad. La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia de la tesis que realice un tercero.

Loja, a 14 de marzo del 2022.

_____

Autor: Bladimir Stanislao Minga Medina
Correo personal: bladiminga46@gmail.com
Correo institucional: bsmingam@unl.edu.ec

DEDICATORIA

El presente trabajo dedico principalmente a mi madre por darme la vida, a mi abuelito y a toda mi maravillosa familia. Dedico también a mi abuelita, que hoy desde el cielo ilumina mi camino.

A mi hijo Matías por ser la fuente de motivación e inspiración, y por ser la luz de mi vida.

A mis amigos, por ser personas extraordinarias.

En fin, quiero dedicar el presente trabajo a cada una de las personas que me acompañaron durante mi trayecto académico y las que estuvieron en las diferentes etapas de mi vida.

Bladimir Minga

AGRADECIMIENTO

En estas líneas quiero agradecer la ayuda que muchas personas me brindaron durante el proceso de investigación y redacción de este trabajo.

Agradezco a mi familia, por acompañarme y apoyarme en todas las etapas de mi vida, a mi tutor, Luis Chamba, por compartir sus conocimientos, por haberme enseñado y orientado en todos los momentos que lo necesité.

A los docentes y colegas que me ayudaron de una manera desinteresada, gracias infinitas por toda su buena voluntad.

A la Universidad Nacional de Loja por todo el conocimiento adquirido en estos años y al estado ecuatoriano, por ser garante del derecho a la educación pública y gratuita, y por las becas de las cuales fui beneficiario.

Bladimir Minga

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. TÍTULO

“PREDICCIÓN DE ENFERMEDADES DEL CORAZÓN UTILIZANDO RANDOM FOREST Y ARTIFICIAL NEURAL NETWORKS”

2. RESUMEN

Las enfermedades cardíacas o enfermedades del corazón son afecciones crónicas que se constituyen como la principal causa de muerte de la población adulta en la mayoría de los países, que según la OMS mueren más personas por este tipo de enfermedades, que por cualquier otra causa. El presente trabajo de titulación tiene como principal objetivo predecir enfermedades del corazón utilizando los algoritmos de machine learning Random Forest y Artificial Neural Networks para clasificación. Los datos que se utilizaron para esta investigación han sido utilizados previamente en diversas investigaciones alrededor del mundo y pertenecen a una recopilación realizada por investigadores del Instituto Húngaro de Cardiología (Hungría), Hospital Universitario de Zúrich (Suiza), Hospital Universitario de Basilea (Suiza) y el Centro Médico VA, Long Beach y Cleveland Clinic Foundation (USA).

En la ejecución de la presente investigación se utilizó el método de revisión bibliográfica para recopilar información que permita fundamentar el desarrollo de la investigación, el proceso de Machine Learning está fundamentado en KDD, se utilizó la técnica de observación activa para la comprensión de los datos y se realizó experimentos con los modelos diseñados. Durante el preprocesamiento de datos se transformaron algunas características utilizando la técnica One Hot Encoding y se utilizó el recall como principal métrica para evaluar el rendimiento de los modelos entrenados. La plataforma Google Colab se utilizó para codificar las instrucciones que permitieron el análisis exploratorio, el preprocesamiento de los datos, el entrenamiento, la validación y el test de los modelos, donde el que mejor rendimiento alcanzó fue Artificial Neural Networks. También se realizó un prototipo de API funcional que se desplegó sobre la plataforma Heroku y para poder realizar predicciones sobre este despliegue se utilizó la aplicación Postman. Entre las principales conclusiones se tiene que los modelos de predicción de enfermedades cardíacas permitieron realizar predicciones prometedoras, ya que con el modelo Artificial Neural Networks se logró alcanzar una accuracy del 88 %, precisión de 85 % y un recall del 93.8 %, frente al 85.3 %, 83% y 90.7% del modelo Random Forest. Los trabajos futuros que se originan luego de haber culminado el trabajo son: Desarrollar un sistema de semaforización automática de pacientes que implemente los modelos entrenados en el presente estudio y Recopilar datos que correspondan exclusivamente a Ecuador y experimentar con otras características para entrenar a los modelos.

ABSTRACT

Cardiac diseases or heart diseases are chronic conditions that are the main cause of death of the adult population in most countries, and according to the WHO, more people die from this type of diseases than from any other cause. The main objective of this degree work is to predict heart diseases using Random Forest and Artificial Neural Networks machine learning algorithms for classification. The data used for this research have been previously used in several investigations around the world and belong to a compilation made by researchers from the Hungarian Institute of Cardiology (Hungary), University Hospital of Zurich (Switzerland), University Hospital of Basel (Switzerland) and the VA Medical Center, Long Beach and Cleveland Clinic Foundation (USA).

In the execution of this research, the literature review method was used to gather information to support the development of the research, the Machine Learning process is based on KDD, the active observation technique was used to understand the data and experiments were carried out with the designed models. During data preprocessing, some features were transformed using the One Hot Encoding technique and recall was used as the main metric to evaluate the performance of the trained models. The Google Colab platform was used to code the instructions that allowed the exploratory analysis, data preprocessing, training, validation and testing of the models, where the one that achieved the best performance was Artificial Neural Networks. A functional API prototype was also made and deployed on the Heroku platform, and the Postman application was used to make predictions on this deployment. Among the main conclusions is that the heart disease prediction models allowed promising predictions to be made, since the Artificial Neural Networks model achieved an accuracy of 88%, precision of 85% and recall of 93.8%, compared to 85.3%, 83% and 90.7% for the Random Forest model. The future works that originate after the completion of the work are: To develop an automatic patient traffic light system that implements the models trained in the present study and to collect data corresponding exclusively to Ecuador and experiment with other characteristics to train the models.

3. INTRODUCCIÓN

Una de las grandes preocupaciones que ha agobiado al hombre es su interés para afrontar las enfermedades, es por ello que la civilización humana desde sus inicios ha intentado identificar y predecir las enfermedades. Actualmente las tecnologías emergentes se muestran muy prometedoras para su aplicación en el ámbito de la salud, el machine learning y organizados en sistemas de Inteligencia Artificial se pueden implementar para fortalecer el sector sanitario. El objetivo principal de la presente investigación es predecir enfermedades del corazón utilizando los algoritmos de machine learning Random Forest y Artificial Neural Networks para clasificación. Se aplica la técnica de clasificación porque en base a ciertas características de los pacientes se va a predecir si desarrollarán o no una enfermedad cardíaca.

El Machine Learning (ML) es de gran utilidad en el diagnóstico precoz de enfermedades, los sistemas que incorporan esta tecnología pueden aprender sobre las condiciones para que un paciente desarrolle una enfermedad. Si se suministran suficientes datos, un sistema de ML no sólo logra detectar una afección con la misma o mayor exactitud que un profesional humano, sino que consigue detectar señales que pueden conducir a una enfermedad y de esta manera el personal clínico puede comenzar a tratar los síntomas de una enfermedad antes de que esta se manifieste [1].

Los avances tecnológicos han permitido que en las últimas décadas sea posible el almacenamiento y procesamiento de grandes cantidades de datos, y actualmente estos datos pueden alimentar complejos algoritmos diseñados específicamente para establecer un sistema de razonamiento que se trata de aproximar lo máximo posible al humano. Estas tecnologías centradas en el análisis de datos están produciendo importantísimos avances en el diagnóstico anticipado de enfermedades, en la mejora de la asistencia primaria y en la administración eficaz de los recursos sanitarios [2].

La identificación de una enfermedad cardíaca en un paciente es compleja y requiere de varios detalles, pruebas de laboratorio y equipos [3]. Por ello la presente investigación no pretende sustituir el enfoque tradicional utilizado para el diagnóstico de enfermedades cardíacas, sino que se intenta apoyar este proceso utilizando tecnologías avanzadas como el Machine Learning. El complejo proceso de diagnóstico de enfermedades del corazón da origen a la necesidad de implementar técnicas de Inteligencia Artificial para apoyar la toma de decisiones y aumentar la

certeza en los diagnósticos médicos, lo que hizo que se plantee la siguiente pregunta de investigación: ¿Es posible predecir enfermedades del corazón utilizando los algoritmos de machine learning Random Forest y Artificial Neural Networks para clasificación?

El Machine Learning es una disciplina científica del ámbito de la IA que intenta extraer toda la información posible de los datos, que no se conforma sólo con la visualización de estos, como podría pasar con las consultas simples; si no que trata de obtener resultados en cuanto a la relación que existe entre los mismos [4]. El ML tiene la finalidad de aprender patrones que generalizan a los datos y no de memorizar los datos o de programar explícitamente las instrucciones. En el presente trabajo se implementó las principales fases de KDD o Descubrimiento de Conocimiento en Bases de Datos, un proceso que se refiere a identificar patrones útiles a partir de los datos. Entre las principales herramientas que se utilizaron está la plataforma Google Colab que se utilizó para codificar las instrucciones de Python para el análisis exploratorio de datos, el preprocesamiento de datos, el entrenamiento y validación de los modelos, esto conjuntamente con la biblioteca scikit-learn. También se utilizó el framework Flask para Python con la finalidad de construir un prototipo de API funcional con el IDE PyCharm.

El dataset utilizado en el proceso de ML es una recopilación de 918 instancias (registros) y 12 características (columnas) conocida universalmente como Cleveland Heart Disease, pero que hace referencia a una selección de datos realizada por investigadores del Instituto Húngaro de Cardiología (Hungría), Hospital Universitario de Zúrich (Suiza), Hospital Universitario de Basilea (Suiza) y el Centro Médico VA, Long Beach y Cleveland Clinic Foundation (USA). Luego de haber realizado el preprocesamiento correspondiente a la data, se procedió a dividir en subconjuntos para el entrenamiento, validación y test, por ende, se dividió primeramente en 80% y 20%, donde este 20% se definió como el conjunto de validación, al 80% restante se lo volvió a dividir nuevamente en 80% y 20%, concretando finalmente a este nuevo 80% como conjunto de entrenamiento y al 20% restante como el conjunto de test. También se utilizó el método de ensayo y error para probar diferentes alternativas para entrenar los modelos modificando sus hiper parámetros.

La presente investigación inicia con una serie de conocimientos básicos acerca del tema de investigación, esto con la finalidad de poner contexto al lector. Luego se aplican los distintos

métodos de investigación utilizados para obtener los resultados para cada objetivo específico. En el primer objetivo se diseñaron los modelos predictivos y para ello primeramente se realizó el análisis exploratorio, el preprocesamiento de los datos y en base a las principales fases del proceso KDD se llevó a cabo todo el proceso de ML. En esta actividad se entrenó y validó los modelos Random Forest y Artificial Neural Networks (ANN), de los cuales el algoritmo que mejores resultados brindó fue ANN, logrando obtener un recall de 93.8 %. En el segundo objetivo se desarrolló un prototipo de API funcional que permite realizar predicciones de enfermedades cardíacas utilizando los modelos entrenados a través de peticiones HTTP, este prototipo se desplegó sobre la plataforma Heroku. En el tercer objetivo se evaluó los modelos a través del prototipo de API mediante tres experimentos planificados. Finalmente se presentan las conclusiones del presente trabajo de titulación, algunas recomendaciones y los trabajos futuros que se originan luego de haber concluido el presente trabajo investigativo.

4. REVISIÓN DE LITERATURA

En este apartado se presenta la base teórica que fundamenta al presente trabajo de titulación. En la sección 4.1 se expone acerca de las enfermedades del corazón, los principales factores que inciden en su desarrollo, causas de estas enfermedades y como se puede predecir este tipo de enfermedades. En la sección 4.2 se presenta la teoría de la Inteligencia Artificial, el machine learning, la importancia en la salud y en la industria 4.0, los algoritmos de machine learning, las métricas que permiten evaluar los modelos y cómo se implementan los modelos de machine learning. Finalmente, en la sección 4.3 se exponen algunos trabajos relacionados considerados importantes para el desarrollo de la presente investigación.

4.1. Enfermedades cardíacas (EC)

Actualmente las enfermedades cardíacas son la causa de muerte más frecuente en casi todos los países del mundo, originándose como consecuencia de diferentes factores que afectan a la salud humana [5]. Este tipo de enfermedades se describen como una gama de afecciones que impactan al corazón de forma negativa y muchas formas de estas enfermedades pueden prevenirse o tratarse con elecciones de un estilo de vida saludable [6]. Existen muchos tipos de enfermedades cardíacas, pero la causa más común es el bloqueo de las arterias coronarias y esta es la razón por lo que ocurren los infartos [7].

4.1.1. Factores de riesgo

Los factores de riesgo son aquellos signos biológicos o hábitos adquiridos que se presentan con mayor frecuencia en los pacientes con una enfermedad concreta. Las EC tienen un origen multifactorial y un factor de riesgo debe ser considerado en el contexto de los otros. Los factores de riesgo cardiovascular tradicionales, se dividen en 2 grandes grupos: no modificables, como: edad, sexo y antecedentes familiares, y los modificables, como: diabetes, hipertensión arterial, dislipidemia, tabaquismo, obesidad y sedentarismo [8].

Entre los factores de riesgo que han recibido el mayor consenso y que han demostrado que solos o combinados se constituyen de mayor riesgo son [9]:

- *La obesidad*: La relación entre el peso y las enfermedades del corazón no son una sorpresa, ya que una persona obesa tendrá más posibilidades de padecer una EC.

- *Actividad física*: El ejercicio constante y un estilo de vida saludable pueden en general repercutir positivamente en el tratamiento de enfermedades, prevenir o retrasar la aparición de la diabetes tipo 2, reducir la presión arterial y ayudar a reducir el riesgo de infarto.
- *Niveles de colesterol*: La acumulación de colesterol es una de las principales causas de la aterosclerosis. Se ha demostrado sistemáticamente que las concentraciones más elevadas de colesterol LDL y de lipoproteínas de no alta densidad a largo plazo se asocian a un mayor riesgo de EC.
- *Glucosa / Diabetes*: La diabetes no es solo una alteración de los niveles de azúcar en sangre, sino que afecta al sistema en general. Los estudios informan de una asociación positiva entre la hipertensión y la resistencia a la insulina.
- *Tabaquismo*: Existe evidencia de que el tabaquismo causa alrededor de 1 de cada 10 muertes por enfermedades cardiovasculares. El humo del tabaco contribuye a las EC, ya que aumenta la placa aterosclerótica y la posibilidad de trombosis.
- *Ingesta de alcohol*: Se ha estudiado que la ingesta de alcohol puede, de hecho, tener un efecto positivo en el metabolismo. Sin embargo, cuando se consumen grandes volúmenes de alcohol, aumenta el riesgo de enfermedades, en particular las EC.

4.1.2. Epidemiología

Cerca del 1% de la población mayor de 40 años presenta insuficiencia cardíaca. La prevalencia de esta enfermedad se dobla con cada década de edad y se sitúa alrededor del 10% en los mayores de 70 años. En la mayoría de países desarrollados la insuficiencia cardíaca es la primera causa de hospitalización en mayores de 65 años, que aproximadamente son el 5% de todas las hospitalizaciones. La insuficiencia cardíaca es un trastorno progresivo y letal, aún con tratamiento adecuado. El control de los factores de riesgo, como la hipertensión y la cardiopatía isquémica, las principales causas de insuficiencia cardíaca, es el único medio para controlar el previsible aumento de esta enfermedad en el futuro [10].

4.1.3. Diagnóstico

Las pruebas que se necesitan para diagnosticar una enfermedad cardíaca dependen del criterio del médico. Independientemente del tipo de enfermedad cardíaca, el médico probablemente, realizará una exploración física y hará preguntas sobre la historia clínica del paciente y de la familia [11].

4.1.4. Síntomas

Generalmente las EC no suelen presentar síntomas previos y su primera manifestación puede ser un ataque al corazón [7]. Pero las características más comunes que se presentan en pacientes con enfermedades cardíacas son los siguientes:

Dolor de pecho experimentado. - El dolor de pecho puede provocar varias sensaciones diferentes.

Presión Arterial. - Es una medida de la fuerza en las paredes de las arterias cuando el corazón bombea sangre a través del cuerpo [12].

Colesterol. – El cuerpo necesita algo de colesterol para funcionar correctamente, pero si tiene demasiado colesterol en la sangre, tiene un mayor riesgo de enfermedades [13].

Azúcar en sangre. - La prueba de glucosa en sangre en ayunas se usa comúnmente para detectar diabetes mellitus [14].

Medición electrocardiográfica. - Es un método de utilidad diagnóstica basado en el registro de la actividad eléctrica cardíaca. El corazón para contraerse y ejercer su función de bomba, necesita ser eléctricamente estimulable, estos estímulos eléctricos producen diferencias de potencial, que pueden registrarse [15].

Frecuencia cardíaca. - La frecuencia cardíaca normal en reposo para adultos oscila entre 60 y 100 latidos por minuto. Si la frecuencia cardíaca en reposo está constantemente por encima de 100 latidos por minuto, es taquicardia. Si no es un atleta entrenado y su frecuencia cardíaca en reposo está por debajo de 60 latidos por minuto es bradicardia [16].

Depresión del ST inducida por el ejercicio en relación con el reposo. - Se refiere a un hallazgo en un electrocardiograma, en el que el trazo en el segmento ST es anormalmente bajo por debajo de la línea base [17].

Angina inducida por ejercicio. - La angina estable generalmente se desencadena por la actividad física. La angina es un dolor en el pecho causado cuando el músculo cardíaco no recibe suficiente sangre rica en oxígeno.

Pendiente del segmento ST de ejercicio pico. - El segmento ST en condiciones normales es plano o isoeléctrico, aunque puede presentar pequeñas variaciones menores de 0.5 mm.

El número de vasos principales. - La fluoroscopia se usa para ayudar al médico a ver el flujo de sangre a través de las arterias coronarias para verificar si hay obstrucciones arteriales [18].

Talasemia. - Es un trastorno sanguíneo transmitido a través de familias (heredadas) en el que el cuerpo produce una forma anormal o una cantidad inadecuada de hemoglobina.

4.1.5. Mortalidad Prematura por enfermedades cardíacas

Los términos para definir la muerte prematura pueden variar de un país a otro, lo que hace difícil en ocasiones comparar directamente los resultados entre los países. Pero la más común es que la muerte se produce antes de la edad promedio de una población determinada, esto basándose en la expectativa de vida de cada lugar.

La mortalidad prematura y los años de vida potencialmente perdidos son los indicadores de la mortandad que podrían reflejar el estado sanitario de un país. En el año 2017 los países de las Naciones Unidas acordaron llevar a cabo un plan estratégico para reducir la mortalidad prematura y se dictó la resolución 71/313 que insta a los gobiernos miembros a trabajar para la reducción de la mortalidad prematura por enfermedades crónicas no transmisibles en un 30% para el año 2030. El análisis de estos indicadores y sus causas constituyen un paso obligado para dictar políticas de salud que lleven a prolongar la vida de las personas con una calidad adecuada [19].

4.1.6. Predicción de enfermedades

Las técnicas de IA aplicadas al diagnóstico de enfermedades han sido utilizadas en estudios de problemas complejos y han alcanzado un aceptado grado de certeza en los resultados obtenidos con respecto a la identificación de enfermedades. Estas aplicaciones son ventajosas debido a que facilitan la construcción y estudio de sistemas capaces de aprender a partir de un conjunto de datos y mejorar procesos de clasificación y predicción.

En las últimas décadas se han realizado varios esfuerzos por aplicar el análisis predictivo en los sistemas de salud, así como lanzar sistemas de aprendizaje automático para facilitar el diagnóstico. En la actualidad la medicina utiliza considerables adelantos que involucran el uso intensivo de la tecnología como la aplicación de técnicas de Inteligencia Artificial, las cuales son factibles cuando aprovechan los datos disponibles y la experiencia clínica. La predicción de enfermedades comienza a utilizarse en la práctica médica con el fin de ayudar a los médicos

en la toma de decisiones y las redes neuronales artificiales han demostrado que producen buenos resultados actuando como una caja negra, en el sentido de que no es posible saber cómo se ha llegado a los resultados obtenidos [20].

4.2. Inteligencia Artificial

La Inteligencia Artificial (IA) es muy antigua como la historia de la humanidad, ya que desde el inicio, con los pueblos primitivos también inicia la historia de la IA [21]. El primer intento por definir a la Inteligencia Artificial lo hizo Alan Turing, el padre de la computación moderna [22], mediante el Test de Turing, el cual intenta demostrar que una máquina es lo suficiente inteligente, si responde como un humano. Aunque parece extraño, no existe un consenso sobre qué es exactamente la IA y, por lo tanto, no hay una definición única, mucho menos exacta.

La popularidad de la IA ha crecido exponencialmente durante los últimos años. El desarrollo e implementación de herramientas para solucionar problemas mediante la IA, han aumentado considerablemente e incluso, su aplicación es muy común dentro de la industria [23]. Las técnicas de la IA son soportes poderosos y valiosos para la toma de decisiones dentro de las actividades corporativas. Estos soportes son capaces de inferir conocimiento de los datos, y a partir de ello generar modelos computacionales que apoyen a la toma de decisiones y de esta manera, resolver de una forma más cómoda los diversos problemas a los que se enfrenta la empresa.

Actualmente, la IA está presente en numerosas actividades que han sido tradicionalmente realizadas por los seres humanos [24], y también se encuentra día a día dentro del sector financiero, la salud, la educación, entre otros. En la Industria 4.0 se utiliza la IA para realizar la transformación digital de los procesos, lo que da origen a una nueva revolución industrial, causada por la evolución de las tecnologías de la información y, específicamente, de la computación y el software [25].

4.2.1. Machine learning

El Aprendizaje Automático o Machine Learning (ML) no es un concepto nuevo, ya en medio siglo XX se definió como una parte de la IA, que usa algoritmos computacionales para proveer a los computadores la capacidad de aprender; es decir, presentar resultados de un problema específico, luego de introducir una suficiente cantidad de datos útiles. Este conocimiento

obtenido no necesita ser programado mediante instrucciones. El ML emplea intensivamente técnicas matemáticas como la estadística y se encuentra íntimamente relacionado con otros campos, como: el modelado, la simulación y la optimización de sistemas [26].

El ML se refiere al subcampo de la computación, especializado en el reconocimiento de patrones ocultos en los conjuntos de datos, y que, a diferencia de la programación clásica, estos algoritmos adquieren de forma autónoma información relevante o el conocimiento implícito, que se halla invisible en un conjunto de datos. Este conocimiento extraído es lo que hace que el programa aprenda [27]. Este aprendizaje aplica inferencias a determinada información para generar una representación adecuada de ciertos aspectos relevantes de algún evento. El ML considera la resolución de problemas como un tipo de aprendizaje que radica en que, una vez conocido un problema, este debe de ser capaz de responder a las nuevas entradas, aplicando la estrategia aprendida [28]. El ML resuelve situaciones por sí solo, a partir de un análisis de datos y mientras más datos existan, mejores resultados se obtiene [23]. El proceso de aprendizaje automático es similar al de la minería de datos, debido a que ambos campos esperan encontrar patrones entre los datos. Sin embargo, en lugar de extraer información para la comprensión humana, como es el caso de la minería de datos; el ML detecta patrones en los datos y ajusta un modelo o programa en consecuencia.

4.2.2. Machine learning en la salud

El machine learning es una ciencia muy prometedora en casi todos los campos, es por ello que desde el origen de la Inteligencia Artificial (IA), la medicina ha sido identificada como uno de los campos más prometedores para su aplicación, ayudando tomar mejores decisiones, a detectar anomalías de forma mucho más eficiente que los humanos, a llevar una mejor atención sanitaria a zonas remotas o a reducir la tensión del personal. Estos son sólo algunos de los campos en los que la IA ha demostrado tener un gran potencial. Los datos clínicos anotados a gran escala, la disponibilidad de paquetes de ML de código abierto y muchos otros factores han impulsado el reciente crecimiento exponencial de la IA [9].

4.2.3. Tipos de Machine Learning

El ML provee a las computadoras la capacidad de aprender a través del reconocimiento de patrones presentes en los datos, sin la necesidad de programar intencionalmente las

instrucciones. El conocimiento que utilizan este tipo de soluciones puede variar conforme se lo alimenta con nuevos datos de entrada.

Los algoritmos del aprendizaje automático se clasifican en diversos tipos, pero los más populares son: el aprendizaje supervisado, aprendizaje no supervisado y el aprendizaje por refuerzo. Donde, los algoritmos de aprendizaje supervisado intentan deducir una función a partir de los datos presentados, los algoritmos no supervisados buscan encontrar asociaciones entre los datos y los algoritmos por refuerzo, intentan determinar el mejor comportamiento de una agente y mejorar su eficiencia realizando cierta tarea basándose en la interacción con su entorno. Cada tipo de aprendizaje contiene diversas técnicas o algoritmos que permiten desarrollar soluciones muy variadas para los problemas encontrados [28].

4.2.4. Algoritmos de Machine Learning supervisado

Los Algoritmos o modelos de aprendizaje supervisado sirven para solucionar problemas de clasificación, donde se predice una clase o para solucionar problemas de regresión, en la que se predice un valor continuo. En el presente trabajo se utiliza la técnica de clasificación porque se infiere la clase a la que corresponde una persona en base a ciertas características, es decir, se predice si el individuo padecerá o no de alguna enfermedad cardíaca.

Los algoritmos que se utilizan en el presente estudio son Random Forest y Artificial Neural Networks, ambos en su versión para clasificación. A continuación, se describe cada uno de ellos.

***Artificial Neural Networks.* –** Es un sistema inspirado en las redes neuronales biológicas, con varios elementos de procesamiento únicos, llamados neuronas. Las neuronas están conectadas entre sí mediante un mecanismo conjunto que consta de un conjunto de pesos asignados [29]. Una red neuronal entrenada, sería capaz de buscar patrones en grupos de datos y mostrar posibles escenarios de comportamiento con base a ciertas variables de entrada [12]. Por lo cual, uno de los requerimientos para el uso adecuado de una red neuronal es disponer de una gran cantidad de datos, para cubrir todos los escenarios posibles [29].

Random Forest. – Este algoritmo ajusta un conjunto de árboles de decisión para aumentar la capacidad de predicción; este tipo de algoritmo se lo puede utilizar tanto para realizar tareas de clasificación y de regresión. Para cada árbol, los datos se dividen de forma recursiva en unidades

más homogéneas, que comúnmente se denominan nodos, con el fin de mejorar la predictibilidad de la variable de respuesta. Los puntos de división se basan en valores de variables predictoras, por lo tanto, las variables utilizadas para dividir los datos se consideran variables explicativas. El valor predicho de una respuesta categórica es la moda de las clases de todos los árboles de decisión ajustados individualmente, y el valor predicho de una respuesta continua es la respuesta ajustada media de todos los árboles individuales que resultaron de cada muestra [30][31].

4.2.5. Knowledge Discovery in Databases (KDD)

El proceso KDD es el proceso de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles que generalmente se considera conocimiento según la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación de la base de datos que se requiera. Principalmente se consideran las siguientes cinco etapas [32]:

1) Selección - Esta etapa consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables o muestras de datos, sobre los que se va a realizar el descubrimiento.
2) Preprocesamiento - Esta etapa consiste en la limpieza y preprocesamiento de los datos objetivo para obtener datos consistentes.
3) Transformación - Esta etapa consiste en la transformación de los datos utilizando métodos de reducción de la dimensionalidad o de transformación.
4) Minería de datos - Esta etapa consiste en la búsqueda de patrones de interés en una forma de representación particular, dependiendo del objetivo (normalmente, la predicción)
5) Interpretación/Evaluación - Esta etapa consiste en la interpretación y evaluación de los patrones extraídos.

El proceso de KDD es interactivo e iterativo, e implica numerosos pasos en los que el usuario toma muchas decisiones. Además, el proceso de KDD debe estar precedido por el desarrollo de una comprensión del dominio de la aplicación, el conocimiento previo relevante y los objetivos del usuario final.

4.2.6. Técnica One Hot Encoding

Es una manera de representar variables categóricas en forma de vectores binarios, es decir, los valores categóricos primero tendrán valores enteros y estos se representan con valores 1 o 0 de

acuerdo a lo que se quiera representar. Esta estrategia consiste en crear una columna para cada valor distinto que exista en la característica que estamos codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0 [33].

4.2.7. Métricas para evaluar los modelos de Machine Learning

Una vez entrenado un modelo, es importante verificar si este tiene un buen rendimiento en relación a los datos nuevos o que no se los haya utilizado durante la fase de entrenamiento del modelo. Se puede poner a prueba un modelo y predecir la respuesta sobre el conjunto de datos de evaluación y luego comprar el resultado predicho, con la observación real.

Existen diferentes métricas que permiten evaluar los modelos de ML, esto con la finalidad de medir la capacidad predictiva del modelo. Se puede encontrar métricas para los modelos de clasificación, métricas para los modelos de regresión y métricas generales, que permiten evaluar a los dos tipos de modelos. La elección de la métrica correspondiente, depende del problema y de la aplicación del ML [34].

En el presente trabajo de clasificación, es importante utilizar algunas métricas para evaluar el rendimiento de los modelos. Como se menciona en [35], las métricas que se utilizan comúnmente para evaluar el rendimiento de los modelos de clasificación, son: Matriz de confusión, Exactitud, Precisión, Sensibilidad, entre otras.

Matriz de confusión. - Confusion matrix en inglés, se utiliza para tener una visión completa al evaluar el rendimiento de un modelo. Esta matriz sirve como base para comprender a las otras métricas y se la define de la siguiente manera:

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos positivos (TP)	Falsos negativos (FN)
	Negativos	Falsos positivos (FP)	Verdaderos negativos (TN)

- Verdaderos positivos (TP): cuando la clase real del punto de datos era Verdadero y la predicha es también Verdadero.

- Verdaderos negativos (TN): cuando la clase real del punto de datos fue Falso y el predicho también es Falso.
- Falsos positivos (FP): cuando la clase real del punto de datos era Falso y el predicho es Verdadero.
- Falsos negativos (FN): Cuando la clase real del punto de datos era Verdadero y el valor predicho es Falso.

Exactitud. **-** También llamada en inglés como accurracy, que en la clasificación es la relación entre las predicciones correctas y el número total de predicciones, o simplemente, con qué frecuencia es correcto el clasificador.

$$Accurracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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.

$$Precisión = \frac{TP}{TP + FP}$$

Recall. **-** También conocida como sensibilidad, es la relación entre las predicciones positivas correctas y el número total de predicciones positivas. O más simplemente, cuán sensible es el clasificador para detectar instancias positivas. Esto también se conoce como la tasa verdadera positiva.

$$Recall = \frac{TP}{TP + FN}$$

4.2.8. Implementar modelos de Machine Learning

El verdadero potencial de las soluciones creadas con ML es evidente cuando éstas pasan a un entorno de producción. Los modelos se pueden implementar a través de una API o mediante contenedores Docker. Una API, es la forma más común de implementar ML, debido a que las APIs pueden conectarse fácilmente a las aplicaciones de producción.

Una API mediante Docker y Flask son las tecnologías que frecuentemente se utilizan como servicio en la mayoría de modelos desarrollados con Python. Los contenedores proporcionan un entorno aislado para que los modelos se ejecuten de forma independiente. Flask es un marco de aplicación web WSGI ligero, diseñado para que la puesta en marcha sea rápida y sencilla, con la capacidad de escalar a aplicaciones complejas.

4.3. Trabajos relacionados

En este apartado se presentan las ideas recopiladas desde 8 artículos que se encuentran relacionados directamente con el tema de estudio. Se habla brevemente de cada uno de ellos, dando a conocer los puntos más relevantes y cómo éstos han servido de apoyo y referencia para desarrollar el presente trabajo de titulación.

El estudio [36] realizado en la Universidad Tecnológica de Gorakhapur en India indica que el aprendizaje automático es la rama de la Inteligencia Artificial que proporciona un apoyo prestigioso en la predicción de cualquier tipo de evento naturales. En esa investigación calcularon la precisión de los algoritmos de aprendizaje automático para predecir las enfermedades del corazón implementando los algoritmos k vecinos más cercanos, árboles de decisión, regresión lineal y la máquina de vectores de soporte. El conjunto de datos utilizado fue Cleveland del repositorio de la UCI con 14 características y la codificación de la programación de Python se realizó en la distribución Anaconda. La principal conclusión fundamenta que los algoritmos k vecinos más cercanos son los mejores algoritmos para predecir las enfermedades cardíacas con un 87% de exactitud.

En el siguiente trabajo de investigación [37] presentan varios atributos relacionados con las enfermedades cardíacas y el modelo basado en algoritmos de aprendizaje supervisado como Naïve Bayes, árboles de decisión, k vecino más cercano y algoritmo de Random Forest. Utilizaron el conjunto de datos existente de la base de datos Cleveland del repositorio de la UCI de pacientes con enfermedades cardíacas, el conjunto estaba conformado por 303 instancias y 76 características, pero consideraron únicamente 14. Los resultados muestran que la puntuación de precisión más alta la consiguieron con K-nearest neighbor con un 90.7% de exactitud.

En la investigación [38] se entrenaron diez clasificadores de aprendizaje automático de diferentes categorías como Bayes, funciones, lazy, meta, reglas y árboles para la predicción

eficiente del riesgo de enfermedades cardíacas utilizando el conjunto completo de atributos del conjunto de datos del corazón de Cleveland con 303 instancias. Como principal resultado mencionan que el clasificador SMO obtuvo un rendimiento notable logrando alcanzar una precisión del 85,148%.

En la siguiente investigación [39] se realizó una revisión sistemática de literatura para identificar las técnicas de Inteligencia Artificial que permitan predecir trastornos cardíacos. Los resultados obtenidos fueron que los árboles de decisión, Naïve Bayes y la red neuronal son técnicas muy importantes para predecir las enfermedades que afectan al corazón, en esa investigación proponen utilizar estas tres técnicas utilizando la métrica precisión para evaluar el desempeño de los modelos.

En el siguiente trabajo proponen una técnica híbrida entre los clasificadores de árboles de decisión y Artificial Neural Networks para mejorar el rendimiento de la predicción de las enfermedades del corazón. El proceso fue realizado mediante el software WEKA utilizando el conjunto de datos de pacientes con enfermedades cardíacas que tomaron del repositorio de la UCI. Para validar el rendimiento de los algoritmos propuestos realizaron una prueba de validación de diez veces, analizando la precisión, recall y especificidad de cada clasificador y de la técnica híbrida. En donde la técnica híbrida (Hybrid-DT) fue la que alcanzó las mejores métricas, un 78.14 % de exactitud y 78 % de recall [40].

En la investigación [41] se evaluó el potencial de seis técnicas de aprendizaje automático para la predicción de enfermedades cardíacas. El recital de estos métodos se evaluó según ocho índices de rendimiento de clasificación diferentes. Además, esos métodos se evaluaron en función de la curva característica operativa del receptor. La precisión de clasificación más alta, del 85 %, se registró con la regresión logística, con una sensibilidad y especificidad del 89 % y el 81 %, respectivamente.

En el artículo [42] se presenta un estudio para mejorar la precisión de la predicción de la insuficiencia cardíaca utilizando el conjunto de datos de enfermedades cardíacas de la UCI. Para ello utilizan múltiples enfoques de aprendizaje automático para comprender los datos y predecir las posibilidades de una EC en una base de datos médica. Los resultados de este estudio comparativo mostraron mejores puntuaciones de precisión en la predicción de enfermedades cardíacas en comparación a los trabajos relacionados. Utilizando RapidMiner lograron obtener

un rendimiento del 93.19 % de exactitud para predecir este tipo de enfermedades con los árboles de decisión.

La investigación [9] tiene como objetivo la implementación de 5 algoritmos clasificadores diferentes y mejorarlos a través de un algoritmo genético que detecta qué combinación de parámetros para cada clasificador da los mejores resultados. Los cinco clasificadores se implementaron utilizando el paquete scikit-learn de Python y los resultados mostraron que algunos algoritmos se adaptan mejor cuando se someten a la evolución, lo que significa que la precisión aumenta a medida que pasan las generaciones y otros algoritmos mostraron una disminución en su rendimiento. El algoritmo que mejor se perfeccionó fué Artificial Neural Networks con una diferencia de precisiones de 9.89 %, pasando de 58.16 % a 68.05 %.

5. MATERIALES Y MÉTODOS

En esta sección se exponen los materiales y métodos que se utilizaron para dar cumplimiento a los objetivos del presente estudio. En el punto 5.1 se manifiesta el contexto en el que se desarrolló la presente investigación de tipo exploratoria, en el 5.2 se presenta el proceso que se efectuó para alcanzar los resultados, en el 5.3 se expone los recursos científicos y técnicos que favorecieron al desarrollo del presente estudio, y finalmente, en el punto 5.4 se muestra a los participantes involucrados en el desarrollo del presente estudio

5.1. Contexto

El presente trabajo de titulación se realizó en la Carrera de Ingeniería en Sistemas de la Facultad de la Energía Las Industrias y los Recursos Naturales no Renovables de la Universidad Nacional de Loja. Para la búsqueda de artículos de investigación se utilizó SCOPUS, IEEE XPLORE y SCIENCE DIRECT como bases de datos académicas. Google Scholar también permitió encontrar literatura no convencional para complementar la investigación. Los datos utilizados en el presente estudio pertenecen a una recopilación realizada por investigadores del Instituto Húngaro de Cardiología (Hungría), Hospital Universitario de Zúrich (Suiza), Hospital Universitario de Basilea (Suiza) y el Centro Médico VA, Long Beach y Cleveland Clinic Foundation (USA). Estos datos se encuentran almacenados en la plataforma Kaggle y han sido utilizados en investigaciones previas como [37], [38], [41], [40] y [36]. Es por ello que adquieren una vital importancia para utilizarlos en la presente investigación.

5.2. Proceso

Para alcanzar el objetivo general del presente Trabajo de Titulación, se efectuó el siguiente proceso:

1) Diseñar los modelos de predicción de enfermedades del corazón.

1.1. Preparación de los datos para el entrenamiento, validación y evaluación de los modelos de ML.

1.2. Desarrollo de los modelos de ML con los dos algoritmos propuestos.

1.3. Evaluación de los modelos mediante conjuntos de datos no utilizados en el entrenamiento.

2) Desarrollar un prototipo de API funcional para interactuar con los modelos.

2.1. Diseño del prototipo de API para la implementación de los modelos.

2.2. Despliegue del prototipo en una plataforma gratuita en la nube.

3) **Evaluar los modelos a través del prototipo de API desarrollado.**

3.1. Planificar y ejecutar escenarios de experimentación.

5.3. Recursos

1) Recursos científicos

1.1. Investigación bibliográfica: esta técnica de tipo expositiva permitió identificar las fases más comunes para llevar a cabo un proceso de machine learning y encontrar información relacionada con el tema de estudio. Además, esta técnica sirvió para elaborar el marco teórico del trabajo de titulación mediante consultas en fuentes bibliográficas científico-académicas.

1.2. KDD: este proceso sirvió como fundamento para realizar el proyecto de Machine Learning, identificando los patrones que se encontraban en los datos y expresándolos como conocimiento a través de modelos de IA.

1.3. Método científico: este método es la base fundamental en toda investigación, por lo tanto, el desarrollo del presente trabajo de titulación se fundamenta en dicho método. Su aplicación inició desde el momento en que se desarrolló la propuesta de TT hasta la culminación del presente trabajo, brindando de esta manera resultados reproducibles para la ingeniería.

1.4. Observación Activa: se utilizó para comprender y manipular el conjunto de datos utilizado para diseñar los modelos. Esta técnica también permitió conocer sobre las enfermedades del corazón y su contexto en general.

1.5. Ensayo y error: es un método heurístico que se utilizó para obtener un conocimiento procedimental y para la alteración conveniente de los hiper parámetros de los algoritmos, esto con la finalidad de encontrar el mejor rendimiento. Es decir, con qué configuración los algoritmos propuestos brindan mejor funcionamiento.

1.6. Método experimental: este método se utilizó para ensayar como los modelos entrenados y así experimentar realizando predicciones con el prototipo de API. Para los experimentos se envió un conjunto de valores de entrada y la respuesta obtenida se comparó con la respuesta real de los pacientes que tenían o no una enfermedad cardíaca.

2) Recursos técnicos

2.1. Python: fue el lenguaje de programación utilizado para codificar las instrucciones necesarias para el tratamiento de los datos y para llevar a cabo todas las tareas del proceso de Machine Learning.

2.2. Scikit-learn: biblioteca de machine learning para el lenguaje de programación Python que se utilizó para diseñar y evaluar los modelos.

2.3. Google Colaboratory: este entorno de Jupyter Notebook gratuito que se ejecuta completamente en la nube permitió escribir y ejecutar el código Python necesario para todo el proceso de diseño de los modelos de predicción de enfermedades cardíacas.

2.4. Google Drive: este servicio se utilizó para alojar los Jupyter Notebook y el conjunto de datos inicial. Lo que permitió poder leer estos recursos desde Google Colab y trabajar con la información contenida en el archivo CSV.

2.5. Heroku: fue la plataforma de computación en la nube que permitió alojar como una API a los modelos entrenados.

2.6. Pycharm: como entorno de desarrollo integrado (IDE) que se utilizó para programar en el lenguaje Python en un entorno local. Con esta herramienta se trabajó principalmente el desarrollo del prototipo de API.

2.7. Postman: es un cliente REST que permitió interactuar con los modelos entrenados, mediante una API desplegada sobre la plataforma Heroku.

2.8. GitHub: este repositorio permitió almacenar el código fuente de los recursos desarrollados durante la ejecución del presente TT.

5.4. Participantes

El presente trabajo es el resultado del esfuerzo y coordinación de los siguientes participantes:

- Bladimir Stanislao Minga Medina, como autor del presente trabajo, que comenzó con la elaboración de la propuesta del proyecto de trabajo de titulación, continuó con el desarrollo de TT, y una vez completados los objetivos planteados, concluyó con la investigación.

- Luis Antonio Chamba Eras, como director para el desarrollo del presente TT, cumpliendo de esta manera con las revisiones correspondientes de los avances académicos y técnicos realizados durante la ejecución del presente trabajo de titulación.

- Francisco Álvarez, como el tutor académico de la asignatura Metodología de Investigación, fundamental para la elaboración de la investigación.

6. RESULTADOS

En el presente trabajo de titulación se planteó tres objetivos específicos, cada uno de ellos compuesto de diferentes fases o actividades que se realizaron durante su ejecución. En el primer objetivo se realizó el diseño de los modelos de machine learning para predecir enfermedades de corazón. En el segundo objetivo se llevó a cabo la construcción y despliegue de un prototipo de API funcional que permite realizar predicciones con los modelos entrenados. Y finalmente, en el tercer objetivo se evaluó los dos modelos de ML a través del prototipo desarrollado.

6.1. OBJETIVO 1: Diseñar los modelos de predicción de enfermedades del corazón.

En este primer objetivo se llevó a cabo utilizando las principales etapas de un proyecto de machine learning basado en el proceso KDD [43]. Las principales etapas que se utilizó son: la recopilación de datos, preparación de datos, entrenamiento del modelo y evaluación del modelo. La imagen 1 representa el flujo que se ejecutó para diseñar dichos modelos.

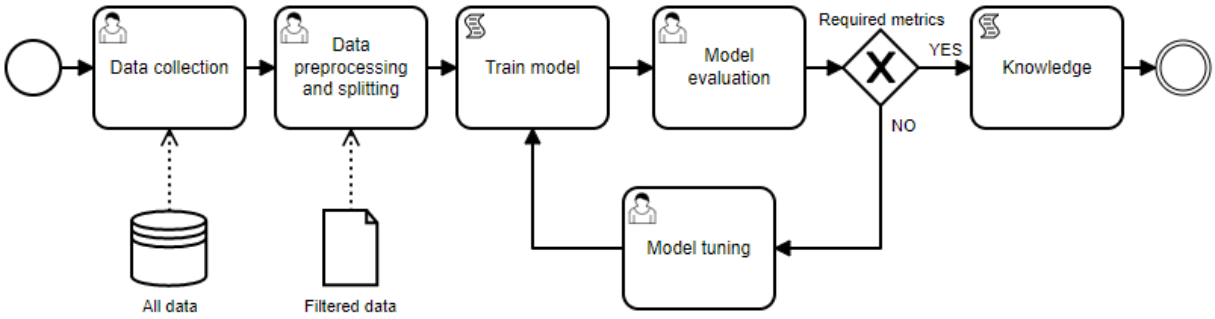


Figura 1 Flujo del proceso de machine learning. Fuente: Elaboración propia

A continuación se explica cada una de las fases que se ejecutaron para obtener los resultados de este objetivo inicial.

6.1.1. Preparación de los datos para el entrenamiento, validación y evaluación de los modelos de ML.

En esta fase se llevó a cabo la recopilación, exploración, preprocesamiento y división del conjunto de datos.

Recopilación de los datos: Se utilizó un conjunto de datos de dominio público conformado por 918 registros y 12 columnas que se encuentra alojado en la plataforma Kaggle y que pertenece a una recopilación realizada por investigadores del Instituto Húngaro de Cardiología (Hungría),

Hospital Universitario de Zúrich (Suiza), Hospital Universitario de Basilea (Suiza) y el Centro Médico VA, Long Beach y Cleveland Clinic Foundation (USA). Estos datos han sido utilizados previamente en diversas investigaciones realizadas en varias partes del mundo y por ello es que ya presentaban un preprocesamiento inicial.

Exploración de los datos: El análisis exploratorio permitió conocer la naturaleza de los datos, los tipos de datos y los valores que contenía el dataset. Este dataset inicialmente contenía 918 registros con 12 columnas o características. En la tabla I se describe las 12 columnas que conformaban el dataset.

TABLA I CARACTERÍSTICAS DEL DATASET INICIAL

CARACTERÍSTICA	TIPO DE DATO	DESCRIPCIÓN
Age	NUMÉRICO	Edad del paciente expresada en años.
Sex	TEXTO	Género del paciente [M: Masculino, F: Femenino]
ChestPainType	TEXTO	Tipo de dolor de pecho [TA: angina típica, ATA: angina atípica, NAP: dolor no anginal, ASY: asintomático]
RestingBP	NUMÉRICO	Presión arterial en reposo [mm Hg]
Cholesterol	NUMÉRICO	Colesterol sérico expresado en mm/dl
FastingBS	NUMÉRICO	Azúcar en sangre en ayunas [1: si BS en ayunas> 120 mg/dl y 0: en caso contrario]
RestingECG	TEXTO	Resultados del electrocardiograma en reposo [Normal: Normal, ST: con anomalía de la onda ST-T (inversiones de la onda T y / o elevación o depresión del ST> 0,05 mv), LVH: que muestra una hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes]
MaxHR	NUMÉRICO	Frecuencia cardíaca máxima alcanzada [Rango entre 60 y 202]
ExerciseAngina	TEXTO	Angina inducida por el ejercicio [Y: Sí, N: No]
Oldpeak	NUMÉRICO	ST [Valor numérico medido en depresión]
ST_Slope	TEXTO	Pendiente del segmento ST del ejercicio pico [Up: ascendente, Flat: plano y Down: descendente]
HeartDisease	BOOLEANO	Enfermedad cardíaca [1: Presencia, 0: Ausencia]

Con el análisis exploratorio se concluyó que las características Cholesterol, RestingBP y MaxHR de tipo de dato numéricos continuos se pueden transformar a tipo de dato enteros discretos que representen categorías. Y que a las características Sex, ChestPainType, RestingECG, ExerciseAngina y ST_Slope se les debe aplicar la técnica One Hot Encoding durante el preprocesamiento de los datos. La actividad completa correspondiente al análisis exploratorio de datos se encuentra en el ***Anexo 1***.

Preprocesamiento de los datos: En esta fase se realizó la transformación necesaria a los datos como: las variables Cholesterol, RestingBP y MaxHR de tipo dato numérico continuo a valores discretos que representan categorías (ver Tabla II). Y para tratar de manera correcta los campos

Sex, ChestPainType, RestingECG, ExerciseAngina y ST_Slope de tipo categóricos se aplicó la técnica One Hot Encoding para convertir cada una de las categorías en nuevas características booleanas en el dataset. Como resultado de este preprocesamiento se obtuvo un dataset de 918 registros y 21 columnas. La tabla II describe el dataset luego de haber sido preprocesado.

TABLA II CARACTERÍSTICAS DEL DATASET PREPROCESADO

CARACTERÍSTICA	TIPO DE DATO	DESCRIPCIÓN
Age	NUMÉRICO	Edad del paciente expresada en años.
RestingBP	NUMÉRICO	Nivel normal: 1 cuando es < 120 En riesgo: 2 cuando es >= 120 y < 229 Peligro: 3 cuando es > 229
Cholesterol	NUMÉRICO	Nivel normal: 1 cuando es < 200 En riesgo: 2 cuando es >= 200 y < 240 Peligro: 3 cuando es > 240
FastingBS	NUMÉRICO	Azúcar en sangre en ayunas [1: si BS en ayunas> 120 mg/dl y 0: en caso contrario]
MaxHR	NUMÉRICO	Nivel normal: 1 cuando es < 100 Hipertensión: 2 cuando es >= 100
Oldpeak	NUMÉRICO	ST [Valor numérico medido en depresión]
Sex_F	BOOLEANO	Femenino: 1, caso contrario 0
Sex_M	BOOLEANO	Masculino: 1, caso contrario 0
ChestPainType_ASY	BOOLEANO	Tipo de dolor de pecho: Asintomático 1, caso contrario 0.
ChestPainType_ATA	BOOLEANO	Tipo de dolor de pecho: Anguina atípica 1, caso contrario 0.
ChestPainType_NAP	BOOLEANO	Tipo de dolor de pecho: Dolor no anguinal 1, caso contrario 0.
ChestPainType_TA	BOOLEANO	Tipo de dolor de pecho: Anguina típica 1, caso contrario 0.
RestingECG_LVH	BOOLEANO	Electrocardiograma en reposo: Hipertrofia ventricular izquierda 1, caso contrario 0.
RestingECG_Normal	BOOLEANO	Electrocardiograma en reposo: Normal 1, caso contrario 0.
RestingECG_ST	BOOLEANO	Electrocardiograma en reposo: Anomalía de la onda 1, caso contrario 0.
ExerciseAngina_N	BOOLEANO	Si: 1, caso contrario 0
ExerciseAngina_Y	BOOLEANO	No: 1, caso contrario 0
ST_Slope_Down	BOOLEANO	Plano: 1, caso contrario 0
ST_Slope_Flat	BOOLEANO	Descendente: 1, caso contrario 0
ST_Slope_Up	BOOLEANO	Ascendente: 1, caso contrario 0
HeartDisease	BOOLEANO	Enfermedad cardíaca [1: Presencia, 0: Ausencia]

Luego del preprocesamiento de datos se seleccionó las siguientes 20 características: *Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak, Sex_F, Sex_M, ChestPainType_ASY, ChestPainType_ATA, ChestPainType_NAP, ChestPainType_TA, RestingECG_LVH, RestingECG_Normal, RestingECG_ST, ExerciseAngina_N, ExerciseAngina_Y, ST_Slope_Down, ST_Slope_Flat, y ST_Slope_Up* como entradas para el entrenamiento de los modelos. Y *HeartDisease* como la variable objetivo (target) que se va a predecir.

División de los datos: El resultado de esta actividad fueron tres subconjuntos de datos que se obtuvieron dividiendo aleatoriamente el dataset preprocesado de la siguiente forma: el dataset inicial se dividió en dos partes, una con el 80% y el otra con el 20%, en donde el del 20% se lo definió como dataset de validación. Seguidamente, el 80% restante, se lo volvió a dividir igualmente en porcentajes de 80% y 20%, obteniendo de esta manera el conjunto de entrenamiento (80%) y el de pruebas (20%). La Figura 2 describe el proceso de división de los datos.

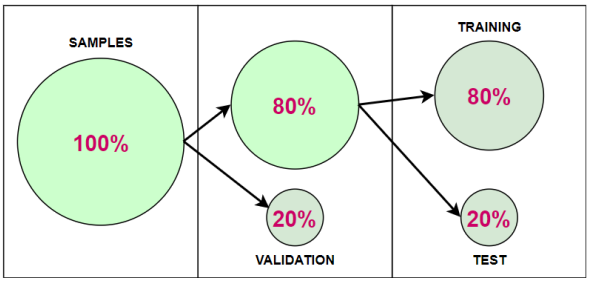


Figura 2 Proceso para la división de los datos. Fuente: Elaboración propia.

Los tres conjuntos resultantes se describen a continuación:

TABLA III DESCRIPCIÓN DE LOS DATASETS

Nombre	Número de registros	Número de columnas
Entrenamiento (training)	587	20
Validación (validation)	184	20
Test	147	20

6.1.2. Desarrollo de los modelos de ML con los dos algoritmos propuestos.

Luego de la preparación de los datos se ejecutó esta fase que consiste en desarrollar los modelos que permiten predecir las enfermedades cardíacas. Para lo cual se llevó a cabo el entrenamiento de los dos algoritmos propuestos utilizando el conjunto de datos correspondiente (training) conformado por 587 registros. Durante el desarrollo de los modelos se experimentó el entrenamiento con varias combinaciones de los principales hiper parámetros de los algoritmos, esto con la finalidad de encontrar las mejores métricas y garantizar una mayor precisión en las predicciones.

6.1.3. Evaluación de los modelos mediante conjuntos de datos no utilizados en el entrenamiento.

En esta última fase del primer objetivo específico se exponen los resultados de la validación de los modelos luego del entrenamiento. Se utilizó el conjunto de datos separado anteriormente para este proceso (conjunto de validación) que contenía 184 registros.

Donde se consiguió los siguientes resultados:

TABLA IV RESULTADOS DE LA EVALUACIÓN DE LOS MODELOS

Nombre	Matriz de confusión	Accurracy	Precisión	Recall
RandomForestClassifier	[[88 9] [18 69]]	0.853	0.830	0.907
MLPClassifier	[[91 6] [16 71]]	0.880	0.850	0.938

6.2. OBJETIVO 2: Desarrollar un prototipo de API funcional para interactuar con los modelos.

Una vez ya diseñados los modelos que permiten predecir las enfermedades del corazón se procedió a desplegarlos a través de un prototipo de API funcional sobre la nube pública Heroku, esto con la finalidad de realizar predicciones desde cualquier cliente web utilizando solicitudes HTTP. El proceso que se llevó a cabo para cumplir con este objetivo se describe a continuación:

6.2.1. Diseño del prototipo de API para la implementación de los modelos.

El diseño general del prototipo de API se muestra en la Figura 3. En la cual, a través del protocolo de comunicación HTTP se propone conectar diversos tipos de clientes hacia la API desplegada sobre una plataforma en la nube.

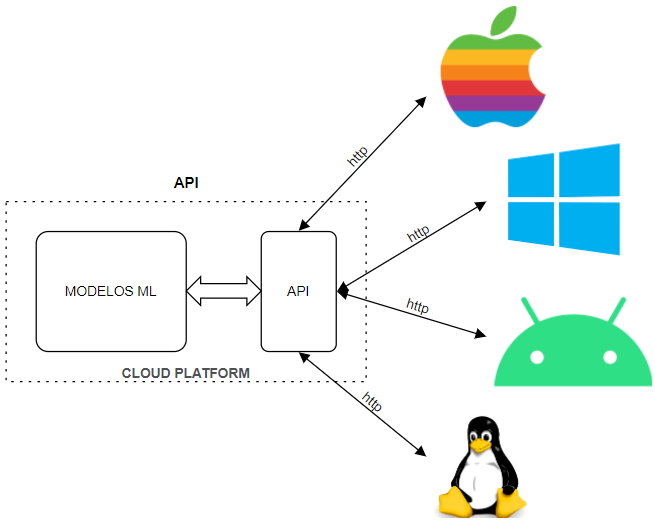


Figura 3 Diseño del prototipo de API. Fuente: Elaboración propia

Se programó las rutas, métodos y clases en el IDE Pycharm, utilizando el framework Flask, en el lenguaje de programación Python. Luego se crearon los archivos necesarios para el despliegue de la aplicación sobre la plataforma Heroku.

6.2.2. Despliegue del prototipo en una plataforma gratuita en la nube.

Para tener una similitud con el funcionamiento real de un sistema de predicciones se procedió a desplegar mediante Git los modelos entrenados en forma de una API a través de un sistema de WebApps para Python. Esta API se desplegó sobre la plataforma Heroku y mediante solicitudes POST se puede intercambiar datos en formato JSON. Las peticiones se pueden realizar para predecir si un paciente o un conjunto de pacientes pueden desarrollar alguna enfermedad cardíaca, esto gracias a que se implementa los modelos de inteligencia artificial entrenados previamente.

Luego del despliegue se concretó el prototipo de API funcional, este nuevo diseño se presenta en la Figura 4. En donde, desde el cliente http (postman) se realizaron peticiones POST hacia la API desplegada sobre la plataforma Heroku. Dentro del prototipo existen clases que contienen los métodos necesarios para realizar las diversas operaciones con los modelos y también para proveer una respuesta de tipo JSON hacia el cliente.

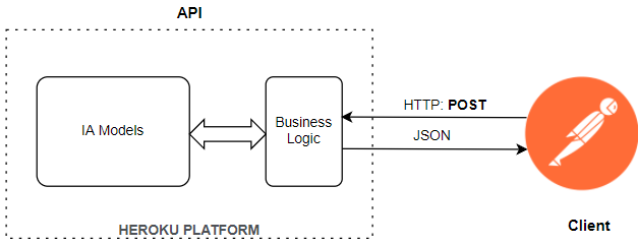


Figura 4 Diseño de prototipo desarrollado. Fuente: Elaboración propia.

La plataforma Heroku proporcionó la siguiente dirección web ***https://predict-heart-diseases-unl.herokuapp.com*** que permite acceder a los servicios web del prototipo de API desplegado.

6.3. OBJETIVO 3: Evaluar los modelos a través del prototipo de API desarrollado.

Este tercer objetivo se llevó a efecto utilizando el prototipo funcional de API que permite realizar predicciones con los dos modelos desarrollados. El conjunto de datos utilizado en esta sesión es el dataset de test que se generó durante la ejecución del primer objetivo específico (Apartado 6.1.1).

6.3.1. Planificar y ejecutar escenarios de experimentación.

Esta actividad se realizó basándose en el proceso experimental presentado en el libro Métodos de investigación en ingeniería del software [27]. A continuación, se describe las tareas que se llevaron a cabo para completar esta experimentación.

6.3.1.1. Alcance

Utilizando el prototipo de API para realizar predicciones de enfermedades cardíacas se pretende evaluar la eficacia de los modelos desarrollados y así conocer la capacidad que tienen dichos modelos para acertar con las predicciones.

6.3.1.2. Planificación

Los datos que se utilizarán se encuentran etiquetados con la clase predictora, es decir se conoce si es que el paciente desarrolló o no una enfermedad cardíaca. Estos datos no han sido utilizados previamente para entrenar ni validar los modelos, por lo tanto, con los valores de entrada se realizará predicciones y estos resultados se comparan con la etiqueta real del conjunto de datos para así conocer la efectividad de los modelos.

El objetivo del presente experimento es conocer el nivel de eficacia para realizar predicciones correctamente con una cantidad limitada y medible de registros. Es por ello que la variable de interés es si un paciente tuvo o no una enfermedad cardíaca.

A través del muestreo por conveniencia se seleccionó para el primer escenario a los 10 primeros registros, para el segundo escenario se utilizó los 10 registros subsecuentes y para el tercer escenario se utilizaron todos los 147 registros disponibles. Cabe recalcar que al momento de generar el dataset para test, los registros ya fueron aleatoriamente seleccionados mediante software, es por ello que en esta actividad simplemente se seleccionan registros consecutivos.

Las predicciones se realizan utilizando el siguiente servicio web *https://predict-heart-diseases-unl.herokuapp.com/predict* que utiliza el método http POST. El cuerpo del JSON que se debe enviar hacia el endpoint es **{"model": string, "data": list}**, donde en *model* se envía el nombre del modelo a utilizar realizar la predicción y en *data* se envía una lista de valores ordenados conforme a la estructura de datos con la cual se entrenó los modelos. Como respuesta del API se obtiene un JSON con la siguiente estructura **{"auto": list, "error": boolean, "message": string, "status": integer}**, donde en *auto* se devuelve una lista con el resultado de las predicciones, una respuesta correspondiente a cada conjunto de valores enviado en la petición, *error* es un valor booleano que indica si la predicción se llevó con éxito, *message* es un valor de tipo cadena que contiene un mensaje de éxito en caso de haber sido una solicitud de predicción exitosa, caso contrario mostrará un mensaje relacionado con el error por el cual no se pudo completar la predicción y *status* es un valor de tipo entero que devuelve 200 cuando la predicción es exitosa, caso contrario devuelve 500 que representa un error en el servidor.

Como los datos están previamente etiquetados y se conoce el resultado real sobre si el paciente desarrolló o no una enfermedad cardíaca, la validación corresponderá a si las predicciones realizadas a través del API coinciden con el resultado real.

6.3.1.3. Operación

Se realizó experimentos planteando tres escenarios y utilizando los mismos datos para realizar predicciones con los dos modelos propuestos, es decir se realizó dos solicitudes por cada escenario manteniendo los mismos datos y solo cambiando el nombre del modelo durante el consumo del servicio web.

Primer escenario: se planteó utilizando 10 registros del dataset de test, en donde se utilizó las filas del 1 al 10. Los resultados de la ejecución de este primer experimento se presentan continuación:

TABLA V PRIMER ESCENARIO DEL EXPERIMENTO CON EL MODELO RANDOM FOREST

N°	Valores de entrada	Real	Predicho
1	40, 2, 2, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
2	64, 2, 3, 0, 2, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	1	1
3	49, 2, 2, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0	1	1
4	36, 1, 3, 0, 2, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0	0	1
5	64, 2, 3, 0, 2, 0.2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	0	1
6	55, 1, 2, 1, 2, 0.4, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1	0	0
7	57, 2, 1, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	1	1
8	54, 2, 3, 0, 2, 3.2, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0	1	1
9	54, 2, 1, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	1	1
10	41, 1, 1, 0, 2, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0

TABLA VI PRIMER ESCENARIO DEL EXPERIMENTO CON EL MODELO ARTIFICIAL NEURAL NETWORKS

N°	Valores de entrada	Real	Predicho
1	40, 2, 2, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
2	64, 2, 3, 0, 2, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1	1	1
3	49, 2, 2, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1	1	1
4	36, 1, 3, 0, 2, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0	0	0
5	64, 2, 3, 0, 2, 0.2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0	0	1
6	55, 1, 2, 1, 2, 0.4, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0	0	0
7	57, 2, 1, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1	1	1
8	54, 2, 3, 0, 2, 3.2, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1	1	1
9	54, 2, 1, 0, 2, 2, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1	1	1
10	41, 1, 1, 0, 2, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0	0	0

Segundo escenario: se planteó utilizando 10 registros del dataset de test, en donde se utilizó las filas del 11 al 20. Los resultados de la ejecución de este segundo experimento se presentan continuación:

TABLA VII SEGUNDO ESCENARIO DEL EXPERIMENTO CON EL MODELO RANDOM FOREST

N°	Valores de entrada	Real	Predicho
1	38, 1, 3, 0, 2, 1.5, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0	1	1
2	51, 2, 3, 0, 2, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	1	1
3	69, 2, 1, 1, 2, 2.5, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0	1	0
4	46, 2, 2, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
5	58, 1, 2, 0, 2, 2.5, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0	1	1
6	49, 2, 3, 0, 2, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
7	69, 2, 2, 1, 2, 0.1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0	0	1
8	63, 2, 2, 0, 2, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0	1	1
9	59, 2, 1, 0, 2, 2, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0	1	1
10	70, 2, 3, 0, 2, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1	0	1

TABLA VIII SEGUNDO ESCENARIO DEL EXPERIMENTO CON EL MODELO ARTIFICIAL NEURAL NETWORKS

N°	Valores de entrada	Real	Predicho
1	38, 1, 3, 0, 2, 1.5, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0	1	1
2	51, 2, 3, 0, 2, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0	1	1
3	69, 2, 1, 1, 2, 2.5, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0	1	1
4	46, 2, 2, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
5	58, 1, 2, 0, 2, 2.5, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0	1	1
6	49, 2, 3, 0, 2, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1	0	0
7	69, 2, 2, 1, 2, 0.1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0	0	1
8	63, 2, 2, 0, 2, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0	1	1
9	59, 2, 1, 0, 2, 2, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0	1	1
10	70, 2, 3, 0, 2, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1	0	0

Tercer escenario: se planteó utilizando todos los 147 registros que conforman el dataset de test. Debido a la cantidad de registros, este resultado no se presenta mediante tablas como los dos experimentos previos, por lo que este experimento se ejecutó utilizando directamente los modelos persistidos como un archivo con formato *pkl*. En la tabla IX se presenta los resultados obtenidos.

TABLA IX RESULTADOS DEL TERCER ESCENARIO DEL EXPERIMENTO

Algoritmo	Muestras	Acierto	Error
RandomForestClassifier	147	119	28
MLPClassifier	147	129	18

6.3.1.4. Análisis

Las figuras que se presentan a continuación permiten interpretar los resultados que se originaron luego de cada uno de los tres escenarios de experimentación.

Primer escenario: Registros del 1 al 10 del conjunto de test.

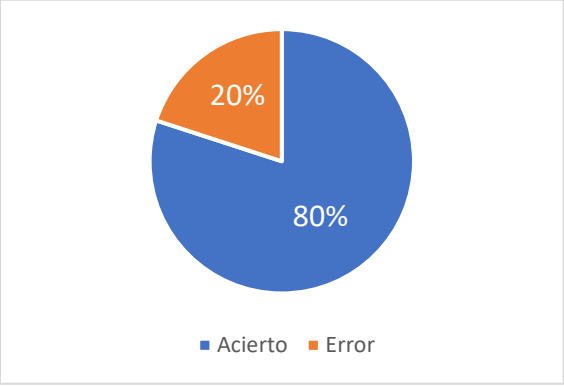


Figura 5 Resultados del primer escenario con el modelo Random Forest

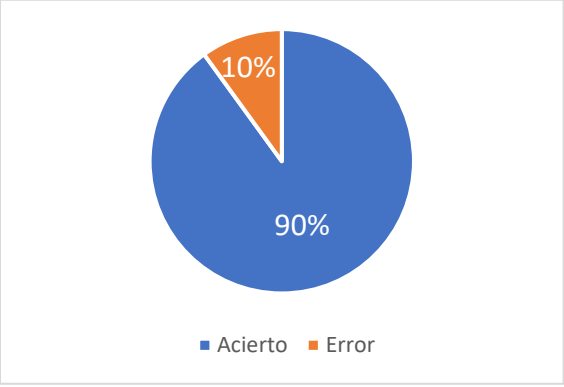


Figura 6 Resultados del primer escenario con el modelo Artificial Neural Networks

Segundo escenario: Registros del 11 al 20 del conjunto de test:

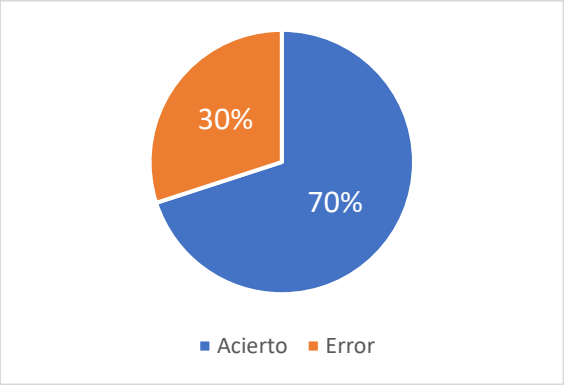


Figura 7 Resultados del segundo escenario con el modelo Random Forest

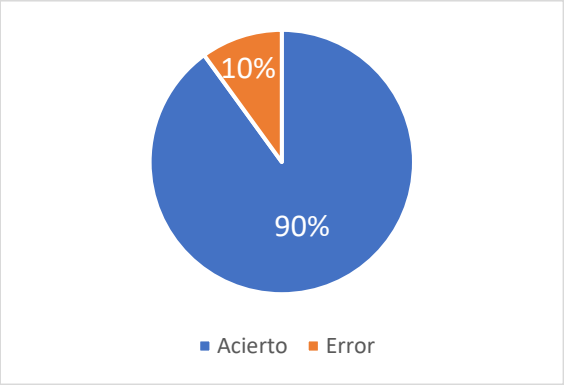


Figura 8 Resultados del segundo escenario con el modelo Artificial Neural Networks

Tercer escenario: Todos los 147 registros de dataset de test.

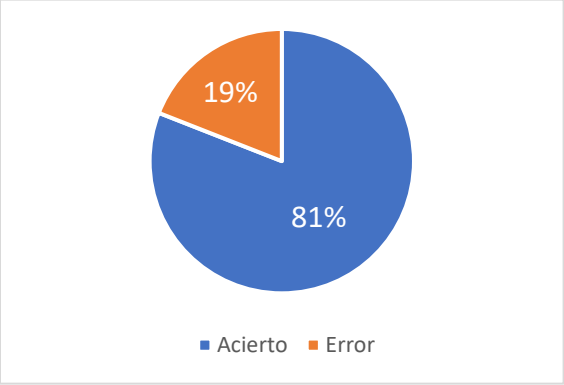


Figura 9 Resultados del tercer escenario con el modelo Random Forest

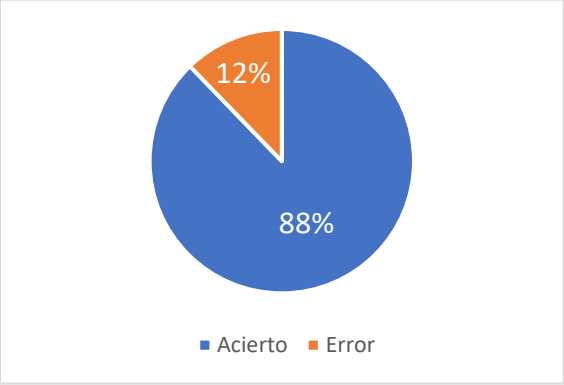


Figura 10 Resultados del tercer escenario con el modelo Artificial Neural Networks

6.3.2. Métricas obtenidas durante el test de los modelos.

En este apartado se presenta un resumen de las métricas obtenidas cuando se evaluaron los dos modelos con todo el conjunto de datos de test.

TABLA X RESULTADOS DEL TEST DE LOS MODELOS ANN Y RFC

Nombre	Matriz de confusión	Accurracy	Precisión	Recall
RandomForestClassifier	[[69 13] [15 50]]	0.810	0.821	0.841
MLPClassifier	[[78 4] [14 51]]	0.878	0.848	0.951

7. DISCUSIÓN

El Machine Learning en la salud brinda enormes ventajas como el diagnóstico precoz de enfermedades y el soporte clínico en el proceso de toma de decisiones. Según los modelos entrenados se puede observar que MLPClassifier (Artificial Neural Networks) tiene mejores ventajas sobre el modelo Random Forest para predecir enfermedades cardíacas.

En el desarrollo del presente trabajo se ejecutaron tres objetivos específicos que permitieron alcanzar el objetivo general. A continuación se detalla la sección correspondiente a la discusión de los resultados por cada objetivo planteado. En el apartado 7.1 se explica la discusión de los resultados contrastándolos con la literatura relacionada con el tema de estudio y sus principales limitaciones y en el apartado 7.2 se presenta la valoración científica, técnica, económica y ambiental del trabajo de titulación.

7.1. Desarrollo de la propuesta alternativa

7.1.1. Diseñar los modelos de predicción de enfermedades del corazón.

El primer objetivo se llevó a cabo basándose en el proceso KDD e implementando sus principales fases. El análisis exploratorio permitió conocer la naturaleza de los datos e introducirse en el contexto de las enfermedades cardíacas, el preprocesamiento de datos fue muy importante porque se manipuló y transformó los datos para garantizar un mejor rendimiento de los modelos. Se aplicó la técnica One Hot Encoding, ya que esta estrategia consiste en crear una columna para cada valor distinto que exista en la característica que se estaba codificando.

Luego de preparar los datos se entrenó y validó los modelos Random Forest (RFC) y Artificial Neural Network (ANN), ambos para clasificación. El diseño de estos modelos se presenta en la sección **6.1**, en donde el algoritmo que mejores métricas brindó es ANN, llegando a obtener una accuracy del 88 %, precisión de 85 % y un recall del 93.8 %, frente al 85.3 %, 83% y 90.7% respectivamente pero que corresponden al modelo RFC. En el estudio de especialización [9] proponen utilizar algoritmos genéticos y en la conferencia internacional sobre ICTCS [40] mencionan a las técnicas híbridas para mejorar el rendimiento de los modelos, pero en el presente trabajo se obtuvo métricas sobresalientes sin la necesidad de recurrir al uso de las estrategias mencionadas.

La mejor exactitud para la predicción de enfermedades cardíacas se obtuvo con Artificial Neural Network, mientras que en las investigaciones relacionadas se obtuvo lo siguiente: por ejemplo en [36] con el algoritmo k vecinos más cercanos lograron obtener un 87 %, en [37] un 90.7 % con K-nearest neighbor, en [38] un 85,148 % con SMO, en [41] 85 % con la regresión logística y en [42] un 93.19 % con árboles de decisión. La literatura previa también permitió conocer que la mayoría de autores evaluaron a los modelos únicamente con la métrica accuracy (exactitud en español) y solo en los estudios [40] y [41] consideran otras métricas como el recall, esta última muy importante cuando se trata de problemas relacionados con la salud. En nuestro estudió se consideró al recall como la métrica más importante para evaluar a los modelos, ya que se necesita conocer la frecuencia para identificar una enfermedad cardíaca. Es decir, el recall indica la proporción de pacientes que se identificaron correctamente por tener una enfermedad cardíaca, sobre el número total de pacientes que realmente presentaban esa condición.

En la presente investigación y en los trabajos de investigación [36], [37], [38], [40] y [42] se utiliza datasets equivalentes, ya que están basados en el repositorio Cleveland, pero los resultados alcanzados en este estudio no se pueden comparar directamente con los resultados de los trabajos relacionados, ya que en estos utilizan modelos de inteligencia artificial diferentes a los propuestos en esta investigación. También es importante recalcar que en investigaciones como [40], [38] y [42] utilizan herramientas que automatizan el proceso de entrenamiento de los modelos como RapidMiner o WEKA, pero en el presente estudio a este proceso se lo realiza de una manera más nativa utilizando el lenguaje de programación Python y aprovechando las ventajas de la computación en la nube, ya que se utilizó Google Colab para realizar la codificación necesaria para todo el proceso de machine learning.

La principal limitación encontrada en el desarrollo de este primer objetivo fue el no poseer un conjunto de datos perteneciente a pacientes exclusivamente del Ecuador, a pesar de que la mayoría de causas de las enfermedades cardíacas son comunes por todo el mundo y por lo tanto los modelos de predicción de enfermedades cardíacas pueden ser considerados genéricos para su aplicación en diferentes países. Otra limitación de este objetivo fue no entrenar otros modelos de machine learning aparte de los propuestos para la investigación e implementar otras estrategias para mejorar el rendimiento de los modelos entrenados.

7.1.2. Desarrollar un prototipo de API funcional para interactuar con los modelos.

El resultado de este segundo objetivo específico permitió evidenciar una forma de cómo se debe implementar los modelos entrenados. A través de una API se puede aprovechar las grandes ventajas que tienen los modelos de machine learning, ya que estas interfaces de programación facilitan el acceso a los servicios a los diferentes clientes web. El prototipo de API diseñado y desplegado sobre la nube demuestra la verdadera utilidad que tiene una solución de machine learning en comparación a la documentación analizada, en donde solo se obtiene resultados de tipo descriptivos y no se expone el beneficio real que proveen este tipo de soluciones basadas en inteligencia artificial.

Actualmente existen diversas plataformas que permiten el despliegue de modelos de machine learning de una forma sencilla sin tener que realizar grandes configuraciones. Un ejemplo de ello es la nube de Heroku, que permitió desplegar el prototipo API como una aplicación web y que puede ser utilizada desde cualquier cliente http (navegador, aplicación móvil, etc.).

La principal limitación de este objetivo fue no haber realizado un proceso completo para el desarrollo de un backend que provea los servicios web sobre los modelos de machine learning entrenados. Las restricciones ocasionadas por usar la versión gratuita de la plataforma Heroku para ejecutar el prototipo de API también fue una limitación que se debe considerar.

7.1.3. Evaluar los modelos a través del prototipo de API desarrollado.

En este tercer objetivo se procedió a evaluar los modelos desplegados a través de medidas que cuantifican el rendimiento y calidad, para de esta manera estimar cómo será el comportamiento de dichos modelos una vez que se los haya puesto en producción.

La evaluación se realizó a través de tres experimentos que permitieron obtener los siguientes resultados basándose en la exactitud para clasificar los casos: en el primer experimento se obtuvo el 80 % de aciertos por parte del modelo RFC, mientras que con ANN se acertó el 90 % de casos. En el segundo experimento se obtuvo el 70 % de aciertos con el modelo RFC y el 90 % con el modelo ANN. Finalmente, en el tercer experimento se obtuvo el 81 % de aciertos con el modelo RFC y el 88 % con el modelo ANN. Estos resultados permitieron conocer que entre RFC y ANN, el mejor modelo para realizar predicciones sobre las enfermedades del corazón es Artificial Neural Network.

Utilizando únicamente a la métrica recall y comparando los resultados alcanzados durante la validación y test de los dos modelos, se obtuvo lo siguiente: el modelo RFC presentó un recall del 90.7 % durante el entrenamiento y en el test bajó 6.6 puntos situándose en un 84.1 %, es decir modelo generó notables cambios cuando se realizó predicciones con un nuevo conjunto de datos. En cambio, el modelo ANN consiguió un recall del 93.8 % durante el entrenamiento y en el test obtuvo un 95.1 %, ganando de esta forma fiabilidad para conseguir la eficacia clínica deseada, ya que las medidas alcanzadas durante la validación y test se mantienen cercanas.

La principal limitación de este último objetivo fue no desarrollar una aplicación web que permita implementar y consumir el API de una manera más amigable para los usuarios finales. Y otra limitación que se evidencia es no haber implementado más servicios web, como para reentrenar los modelos y obtener métricas.

7.2. Valoración técnica, económica y ambiental

Analizado desde tres aspectos, el desarrollo del presente Trabajo de Titulación origina los beneficios que se detallan a continuación.

7.2.1. Valoración Técnica

La valoración técnica del presente trabajo de titulación se da origen a partir de los diversos instrumentos que fueron utilizados para dar cumplimiento al objetivo del presente TT. Los programas informáticos, las plataformas en la nube, el lenguaje de programación python y sus librerías fueron fundamentales para la implementación de los modelos de Machine Learning.

Otras herramientas técnicas que permitieron el desarrollo, monitoreo y finalización del presente trabajo son:

- Translator DeepL: esta herramienta fue utilizada para la traducción al español de los diferentes artículos académicos utilizados en el presente TT.
- Mendeley: fue el gestor bibliográfico utilizado durante el desarrollo del presente trabajo de titulación.
- App Diagrams Net: fue la herramienta en línea gratuita para crear diagramas e ilustraciones utilizados en el presente TT.
- Camunda: aplicación que permitió diseñar los flujos de procesos basándose en el enfoque BPM.

- Repositorio y página web SciHub: este recurso se utilizó para obtener acceso a los diferentes artículos de investigación que se encontraban con acceso restringido.

7.2.2. Valoración Económica

Para el desarrollo del presente trabajo de titulación se necesitó financiamiento para el talento humano y los recursos físicos utilizados directa e indirectamente en el desarrollo del presente trabajo. El software utilizado fue de tipo libre y otros utilizados en su versión gratuita, por lo tanto, no existieron costos relacionados con estas herramientas.

1) Talento Humano

El desarrollo del presente trabajo se llevó a cabo gracias a la colaboración de los siguientes recursos: el estudiante investigador y autor del TT y dos docentes universitarios, uno para el rol de director del TT y otro como tutor académico de la materia de investigación, los dos docentes fueron financiados totalmente por la UNL. En la **Tabla XXVIII**, se muestra el valor económico del talento humano, presente en la investigación.

TABLA XI RECURSOS HUMANOS UTILIZADOS EN EL DESARROLLO DEL TT.

Talento Humano			
Recurso	**Horas**	**Precio**	**Total**
Director de TT	30	$ 14.77	$ 443.1
Tutor académico	30	$ 14.77	$ 443.1
Estudiante	320	$ 5.68	$ 1817.6
Total:			$ 2703.8

2) Recursos físicos

Los recursos físicos necesarios para el desarrollo de la presente propuesta fueron los siguientes:

TABLA XII RECURSOS FÍSICOS UTILIZADOS EN EL DESARROLLO DEL TT.

Recursos físicos	
Recurso	**Precio**
Dispositivos tecnológicos	$ 1300
Materiales de oficina	$ 200
Total:	$ 1500

3) Presupuesto Final

El resultado final del costo del Trabajo de Titulación se describe en la **Tabla XXX.**

TABLA XIII PRESUPUESTO TOTAL DEL TT.

PRESUPUESTO TOTAL	
Recursos Humanos	$ 2703.8
Recursos físicos	$ 1500
Total:	**$ 4203.8**

7.2.3. Valoración Ambiental

Los aportes que realiza el presente trabajo de titulación en el aspecto ambiental son el ahorro de los recursos, ya que las máquinas mediante IA apoyan la toma de decisiones. Otro aporte se produce cuando se utilizan plataformas en la nube, ya que se reduce emisiones de carbono, esto en comparación a los pequeños centros de cómputo. Al ser este un trabajo investigativo que involucra datos digitales, la cantidad de recursos que afectan negativamente al medio ambiente como el papel fue muy mínima.

8. CONCLUSIONES

- Los modelos de predicción de enfermedades cardíacas desarrollados permitieron realizar predicciones muy prometedoras, ya que con el modelo Artificial Neural Networks se logró alcanzar una accuracy del 88 %, precisión de 85 % y un recall del 93.8 %, frente al 85.3 %, 83% y 90.7% respectivamente pero correspondiente al modelo Random Forest. También fue muy importante aplicar estrategias para el preprocesamiento de datos como la técnica One Hot Encoding que permitió mejorar la calidad de los datos y así enriquecer el rendimiento de los modelos.

- La Interfaz de Programación de Aplicaciones (API) desplegada como un prototipo funcional posibilitó la interacción con los modelos persistidos y desplegados sobre la nube de Heroku. Este prototipo de API permitió evidenciar la importancia y funcionalidad que adquieren los modelos de machine learning cuando estos pasan a otros entornos para proveer servicios de predicción.

- Los experimentos permitieron conocer el comportamiento de los modelos entrenados cuando se realizó predicciones con datos desconocidos, lo que permitió identificar que el modelo Artificial Neural Networks mantuvo su rendimiento, mientras que el modelo Random Forest cambió considerablemente su performance en relación a las métricas obtenidas en la validación y test. El proceso de evaluación de los modelos a través del prototipo de API hizo que se determinara que con el conjunto de datos utilizado para entrenar las ANN y RFC, el algoritmo que mejores ventajas brinda fue Artificial Neural Networks.

8.1. Otras Aportaciones

- Se comparte los scripts[1] del análisis exploratorio de datos y de la codificación en Python para el diseño de los modelos, también el repositorio[2] del prototipo de API con la finalidad de que la información y los resultados obtenidos en el presente trabajo sean reproducibles.

[1] https://github.com/bsmm1995/scripts-datasets.git
[2] https://github.com/bsmm1995/tt-ute.git

9. RECOMENDACIONES

- Tal como se aplica en los estudios [9] y [40], se recomienda aplicar técnicas híbridas o implementar algoritmos genéticos con la finalidad de verificar si con el mismo conjunto de datos utilizado en el presente estudio, el rendimiento de los modelos entrenados aumenta.

- Utilizar las plataformas de computación en la nube, ya que para el Machine Learning brindan una gran capacidad de procesamiento, e inclusive permiten el uso gratuito de CPU, GPU y TPU, implementando de esta manera tecnologías que nos direccionan hacia la Industria 4.0 [44].

9.1. Trabajos Futuros

- Como trabajo futuro se recomienda desarrollar un sistema de semaforización automática de pacientes y que implemente los modelos entrenados en el presente estudio, ya que una solución de este tipo permitiría mejorar la toma de decisiones e identificar enfermedades cardíacas de una forma anticipada.

- Recopilar datos que correspondan exclusivamente a Ecuador y experimentar con otras características para entrenar a los modelos. Con el fin de buscar modelos más eficientes y representativos para nuestra región.

10. BIBLIOGRAFÍA

[1] S. VICENTE, «El uso de dispositivos inteligentes y “machine learning” para la predicción de enfermedades.», *Univ. Sevilla*, vol. 1, pp. 1-29, 2019.

[2] I. Rodríguez Rodríguez y J.-V. Rodríguez, «Tecnologías digitales disruptivas aplicadas a la gestión de la pandemia por COVID-19: un análisis a través de la producción científica», *Universidad de Málaga*, Málaga, 2022.

[3] M. M. La Quintana Illanes, «Modelo de Control y Diagnóstico de Enfermedades Cardiovasculares», *Univ. MAYOR SAN ANDRÉS*, p. 73, 2016.

[4] P. I. Dorado-Díaz, J. Sampedro-Gómez, V. Vicente-Palacios, y P. L. Sánchez, «Applications of Artificial Intelligence in Cardiology. The Future is Already Here», *Rev. Esp. Cardiol.*, vol. 72, n.º 12, pp. 1065-1075, dic. 2019, doi: 10.1016/j.recesp.2019.05.016.

[5] E. Martínez Santamaría, M. Lameiras Fernández, M. González Lorenzo, y Y. Rodríguez Castro, «Alteraciones emocionales en personas mayores con enfermedades cardíacas», *Aten. Primaria*, vol. 38, n.º 2, pp. 90-95, jun. 2006, doi: 10.1157/13090430.

[6] MedilinePlus, «Qué es la enfermedad cardiovascular: MedlinePlus enciclopedia médica», *Medlineplus.Gov*, 2020. [En línea]. Disponible en: https://medlineplus.gov/spanish/ency/patientinstructions/000759.htm. [Accedido: 31-oct-2021].

[7] A. V. Ferry *et al.*, «Presenting symptoms in men and women diagnosed with myocardial infarction using sex-specific criteria», *J. Am. Heart Assoc.*, vol. 8, n.º 17, sep. 2019, doi: 10.1161/JAHA.119.012307.

[8] J. Vega Abascal, M. Guimará Mosqueda, y L. Vega Abascal, «Riesgo cardiovascular, una herramienta útil para la prevención de las enfermedades cardiovasculares», *Revista Cubana de Medicina General Integral*, 2011. [En línea]. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21252011000100010. [Accedido: 01-nov-2021].

[9] E. Juliá Martínez, E. L. Rubio, R. María, y M. Quiroga, «Predicción de enfermedades cardiovasculares mediante algoritmos de inteligencia artificial», *Univ. Málaga*, nov. 2020.

[10] F. Rodríguez-Artalejo, J. R. Banegas Banegas, y P. Guallar-Castillón, «Epidemiología de la insuficiencia cardíaca», *Rev. Esp. Cardiol.*, vol. 57, n.º 2, pp. 163-170, feb. 2004, doi: 10.1157/13057268.

[11] P. García-Pavía, M. T. Tomé-Esteban, y C. Rapezzi, «Amyloidosis. Also a Heart Disease», *Rev. Española Cardiol. (English Ed.*, vol. 64, n.º 9, pp. 797-808, sep. 2011, doi: 10.1016/j.rec.2011.05.007.

[12] P. K. Whelton *et al.*, «2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults a report of the American College of Cardiology/American Heart Association Task Force on Clinical pr», *Hypertension*, vol. 71, n.º 6, pp. E13-E115, jun. 2018, doi: 10.1161/HYP.0000000000000065.

[13] O. Medications, «Patient Tools : What You Need to Know What You Need to Know», *NIH Medlin. Plus Mag.*, vol. 7, n.º February, pp. 6-7, 2012.

[14] C. Catellier, «La préparation du diabétique à la chirurgie.», *Laval Med.*, vol. 40, n.º 8, pp. 720-723, oct. 1969.

[15] J. A. Zavala-Villeda, «Descripción del electrocardiograma normal y lectura del electrocardiograma», *Rev. Mex. Anestesiol.*, vol. 40, n.º S1, pp. 210-213, jun. 2017.

[16] L. Vorvick, «Pulso : MedlinePlus enciclopedia médica», *Medline Plus*, 2015. [En línea]. Disponible en: https://www.nlm.nih.gov/medlineplus/spanish/ency/article/003399.htm. [Accedido: 01-nov-2021].

[17] F. Machado, «Elevación del ST inducida por esfuerzo ergométrico en un paciente sin infarto previo: reporte de un caso», *Rev. Uruguaya Cardiol.*, vol. 26, n.º 2, pp. 158-162, 2011.

[18] B. F. Uretsky, R. D. Rifkin, S. C. Sharma, y P. S. Reddy, «Value of fluoroscopy in the detection of coronary stenosis: Influence of age, sex, and number of vessels calcified on diagnostic efficacy», *Am. Heart J.*, vol. 115, n.º 2, pp. 323-333, feb. 1988, doi: 10.1016/0002-8703(88)90478-4.

[19] A. F. D. Herrera, N. B. A. Rojas, y G. H. Vera, «Premature mortality for cardiovascular illnesses in Cuba», *Rev. Cuba. Cardiol. y Cirugía Cardiovasc.*, vol. 24, n.º 4, pp. 1-7, 2018.

[20] N. González Benítez, V. Estrada Sentí, y A. Febles Estrada, «Estudio y selección de las técnicas de Inteligencia Artificial para el diagnóstico de enfermedades», *Rev. Ciencias Médicas Pinar del Río*, vol. 22, n.º 3, pp. 534-544, 2018.

[21] M. Sanz Bobi, «Inteligencia artificial y sistemas expertos», *Rev. Razón y fe Rev. Hispanoam. Cult. Periodo 1, Vol. 233, Número 1169, Página inicial 301, Página Final 311*, 1996.

[22] A. García Serrano, «Inteligencia Artificial. Fundamentos, práctica y aplicaciones», *RC Libros*, 2012.

[23] E. M. Rojas, «Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo», *Rev. Ibérica Sist. e Tecnol. Informação*, pp. 586-599, 2020.

[24] Paola Carranza Bravo, «Introducción a las técnicas de Inteligencia Artificial aplicadas a la gestión financiera empresarial», *Fides et Ratio - Revista de Difusión cultural y científica de la Universidad La Salle en Bolivia*, 2010. [En línea]. Disponible en: http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2071-081X2010000100002. [Accedido: 04-ago-2020].

[25] J. Luis y D. V. Román, «Industria 4.0: la transformación digital de la industria», *Conf. Dir. Y DECANOS Ing. INFORMÁTICA*, 2016.

[26] A. Núñez Reiz, M. A. Armengol de la Hoz, y M. Sánchez García, «Big Data Analysis and Machine Learning in Intensive Care Units», *Medicina Intensiva*, vol. 43, n.º 7. Ediciones Doyma, S.L., pp. 416-426, 01-oct-2019, doi: 10.1016/j.medin.2018.10.007.

[27] C. González-García, «En qué consiste el aprendizaje automático (machine learning) y qué está aportando a la Neurociencia Cognitiva», *Cienc. Cogn.*, vol. 12, n.º 2, pp. 48-50, 2018.

[28] A. Moreno -Eva *et al.*, «Aprendizaje automático», en *Universitat Politècnica de Catalunya*, 1994, UPC., p. 244.

[29] A. H. Moghaddam, M. H. Moghaddam, y M. Esfandyari, «Predicción del índice del mercado bursátil utilizando una red neuronal artificial», *Journal of Economics, Finance and Administrative Science*, 2016. [En línea]. Disponible en: http://www.scielo.org.pe/scielo.php?pid=S2077-18862016000200007&script=sci_arttext&tlng=en. [Accedido: 10-ago-2020].

[30] Y. Everingham, J. Sexton, D. Skocaj, y G. Inman-Bamber, «Accurate prediction of sugarcane yield using a random forest algorithm», *Agron. Sustain. Dev.*, vol. 36, n.º 2, 2016, doi: 10.1007/s13593-016-0364-z.

[31] H. M. Gomes *et al.*, «Adaptive random forests for evolving data stream classification», *Mach. Learn.*, vol. 106, n.º 9-10, pp. 1469-1495, 2017, doi: 10.1007/s10994-017-5642-8.

[32] U. M. Fayyad, «Data mining and knowledge discovery: Making sense out of data», *IEEE Expert. Syst. their Appl.*, vol. 11, n.º 5, pp. 20-25, oct. 1996, doi: 10.1109/64.539013.

[33] P. Rodríguez, M. A. Bautista, J. Gonzàlez, y S. Escalera, «Beyond one-hot encoding: Lower dimensional target embedding», *Image Vis. Comput.*, vol. 75, pp. 21-31, jul. 2018, doi: 10.1016/J.IMAVIS.2018.04.004.

[34] Amazon, «Amazon Machine Learning Guía del desarrollador», 2020.

[35] A. Amidi y S. Amidi, «Hoja de referencia: Consejos y trucos sobre Aprendizaje Automático», *Stanford Univ.*, pp. 1-3, 2018.

[36] A. Singh y R. Kumar, «Heart Disease Prediction Using Machine Learning Algorithms», *Int. Conf. Electr. Electron. Eng. ICE3 2020*, pp. 452-457, feb. 2020, doi: 10.1109/ICE348803.2020.9122958.

[37] V. Sharma, S. Yadav, y M. Gupta, «Heart Disease Prediction using Machine Learning Techniques», *Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020*, vol. 1, n.º 6, pp. 177-181, oct. 2020, doi: 10.1109/ICACCCN51052.2020.9362842.

[38] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, y S. Pranavanand, «Heart disease risk prediction using machine learning classifiers with attribute evaluators», *Appl. Sci.*, vol. 11, n.º 18, sep. 2021, doi: 10.3390/app11188352.

[39] A. Powar, S. Shilvant, V. Pawar, V. Parab, P. Shetgaonkar, y S. Aswale, «Data Mining Artificial Intelligence Techniques for Prediction of Heart Disorders: A Survey», *Proc. - Int. Conf. Vis. Towar. Emerg. Trends Commun. Networking, ViTECoN 2019*, mar. 2019, doi: 10.1109/ViTECoN.2019.8899547.

[40] S. Maji y S. Arora, «Decision Tree Algorithms for Prediction of Heart Disease», *Lect. Notes Networks Syst.*, vol. 40, pp. 447-454, 2019, doi: 10.1007/978-981-13-0586-3_45.

[41] A. K. Dwivedi, «Performance evaluation of different machine learning techniques for prediction of heart disease», *Neural Comput. Appl.*, vol. 29, n.º 10, pp. 685-693, sep. 2018, doi: 10.1007/s00521-016-2604-1.

[42] F. S. Alotaibi, «Implementation of machine learning model to predict heart failure disease», *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, n.º 6, pp. 261-268, 2019, doi: 10.14569/ijacsa.2019.0100637.

[43] U. Fayyad, G. Piatetsky-Shapiro, y P. Smyth, «The KDD Process for Extracting Useful Knowledge from Volumes of Data», *Commun. ACM*, vol. 39, n.º 11, pp. 27-34, nov. 1996, doi:

10.1145/240455.240464.

[44] L. A. Industria *et al.*, «La nube al servicio de las pymes en dirección a la industria 4.0», dic. 2017.

[45] L. González-Támara, «Una introducción a la estadística descriptiva y probabilidad Análisis exploratorio de datos», *Univ. Bogotá Jorge Tadeo Lozano.*, 2017.

11. ANEXOS

11.1. Anexo 1: ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos es una fase crítica en la ciencia de datos y el machine learning, y sin duda, es la que conlleva más tiempo. Utilizando las técnicas propuestas en [45] para resumir mediante gráficas los datos cualitativos y cuantitativos, se procede con el siguiente análisis.

1. Descripción del dataset inicial

El análisis exploratorio se realizó con el lenguaje Python en la plataforma Google Colab, manipulando el conjunto de datos almacenado en Google Drive. En la **Figura 11** se presenta el código implementado para dicho proceso.

```
df = pd.DataFrame()
def _loadCsv():
  global df
  id = '1vPVmPr3-VP7xc8GkC5UvwRJG8Vhtv8BK'
  downloaded = drive.CreateFile({'id':id})
  downloaded.GetContentFile('heart900.csv')
  df = pd.read_csv('heart900.csv')
_loadCsv()
df.shape
```

Figura 11 Script para cargar el conjunto de datos almacenado en Google Drive

Una vez cargado el dataset se empezó con la exploración y comprensión inicial de los datos. La **Figura 12** muestra que el conjunto de datos tiene 918 registros (filas) y 12 características (columnas).

```
print('El archivo tiene el siguiente número de filas y columnas:', df.shape)
print('Las columnas son:', df.columns)
```

```
El archivo tiene el siguiente número de filas y columnas: (918, 12)
Las columnas son: Index(['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS',
       'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope',
       'HeartDisease'],
      dtype='object')
```

Figura 12 Información del dataset inicial

La **Figura 13** presenta los resultados luego de haber aplicado el comando ***describe().*** Este comando se aplicó para mostrar una descripción del dataset, y los resultados indican las propiedades como: la cantidad de registros, la media, la desviación estándar y otras correspondientes a las variables numéricas.

```
1 df.describe()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Figura 13 Resultado del comando describe()

Para continuar con la comprensión del conjunto de datos, se procede a describir cada una de las 12 características, conforme se expresa en la **Tabla XIV**.

TABLA XIV DESCRIPCIÓN DE LAS CARACTERÍSTICAS

CARACTERÍSTICA	TIPO DE DATO	DESCRIPCIÓN
Age	NUMÉRICO	Edad del paciente expresada en años.
Sex	TEXTO	Género del paciente [M: Masculino, F: Femenino]
ChestPainType	TEXTO	Tipo de dolor de pecho [TA: angina típica, ATA: angina atípica, NAP: dolor no anginal, ASY: asintomático]
RestingBP	NUMÉRICO	Presión arterial en reposo [mm Hg]
Cholesterol	NUMÉRICO	Colesterol sérico expresado en mm/dl
FastingBS	NUMÉRICO	Azúcar en sangre en ayunas [1: si BS en ayunas> 120 mg/dl y 0: en caso contrario]
RestingECG	TEXTO	Resultados del electrocardiograma en reposo [Normal: Normal, ST: con anomalía de la onda ST-T (inversiones de la onda T y / o elevación o depresión del ST> 0,05 mv), LVH: que muestra una hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes]
MaxHR	NUMÉRICO	Frecuencia cardíaca máxima alcanzada [Rango entre 60 y 202]
ExerciseAngina	TEXTO	Angina inducida por el ejercicio [Y: Sí, N: No]
Oldpeak	NUMÉRICO	ST [Valor numérico medido en depresión]
ST_Slope	TEXTO	Pendiente del segmento ST del ejercicio pico [Up: ascendente, Flat: plano y Down: descendente]
HeartDisease	BOOLEANO	Enfermedad cardíaca [1: Presencia, 0: Ausencia]

La variable ***HeartDisease*** es la variable que se va a predecir y también es conocida como la variable objetivo (target) que permitirá entrenar los modelos. Por tal motivo se presenta una descripción de sus valores iniciales. Conforme a la **Figura 14** se puede mencionar que el 55.3 % de los pacientes del estudio no tenía una enfermedad cardíaca y que el 44.7 % si tenían algún tipo de esta enfermedad. Esta gráfica evidencia también que los registros originales se encuentran balanceados en relación a la clase predictora.

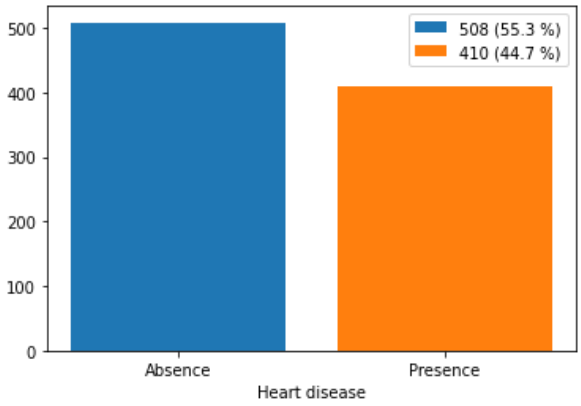
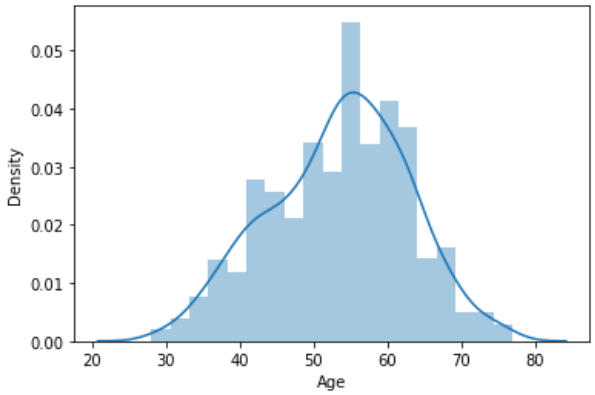


Figura 14 Descripción de la característica HeartDisease

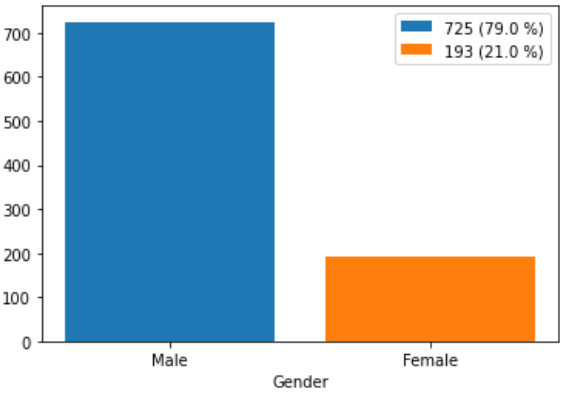
2. Resultados

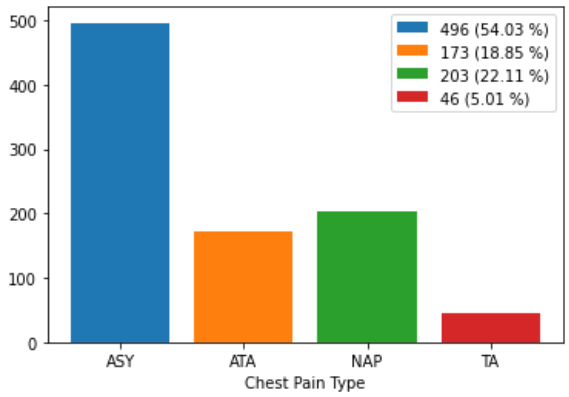
A continuación se presentan las gráficas correspondientes a las 11 características del dataset que servirán como entradas para el entrenamiento de los modelos.



Gráfica que indica la edad de los pacientes. La edad mínima presente en los registros es 28 y la máxima es 77.

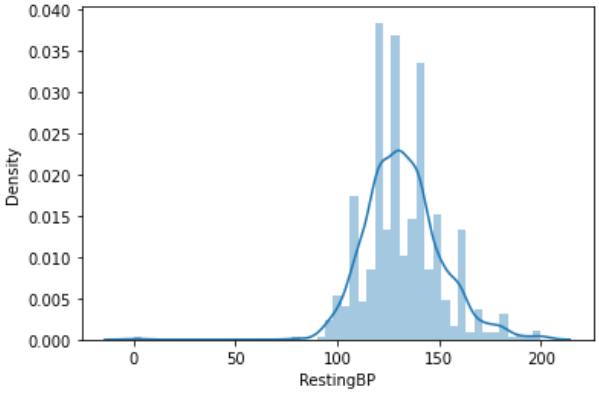
Gráfica que indica el género de los pacientes. El 79 % de registros que conforman el dataset son de género masculino y el 21 % restante son de género femenino.

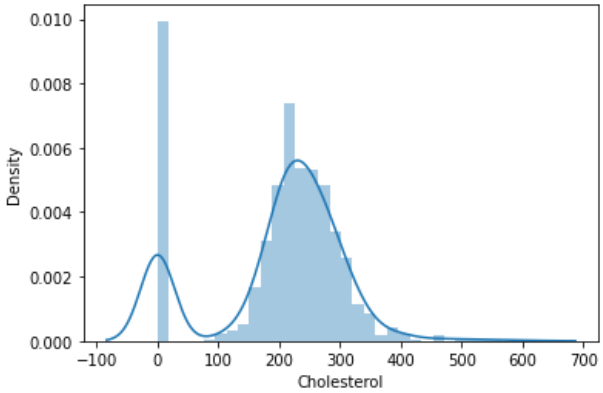




Gráfica que indica el tipo de dolor en el pecho. Donde: ASY: angina típica, ATA: angina atípica, NAP: dolor no anginal y TA: asintomático.

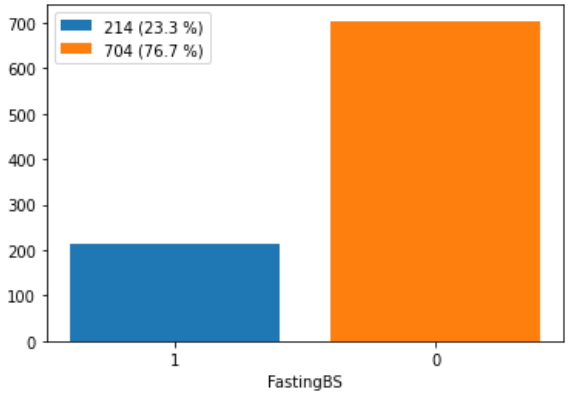
Gráfica que indica los registros de la presión arterial en reposo de los pacientes.

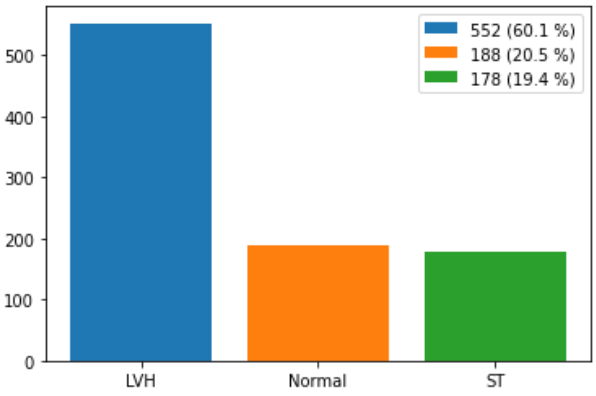




Gráfica que indica el nivel de colesterol en la sangre de los pacientes.

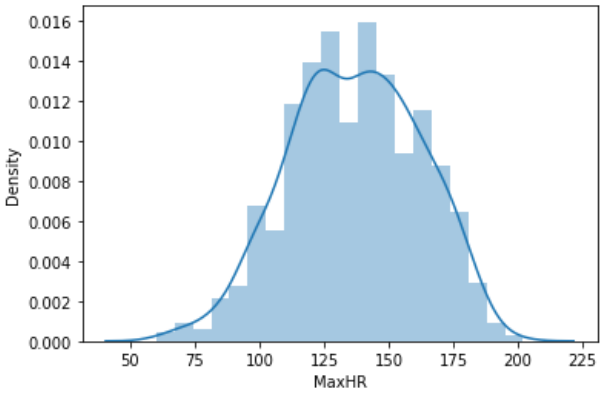
Gráfica que indica el nivel de azúcar en sangre en ayunas. Donde 1: si es mayor a 120 y 0: en caso contrario.

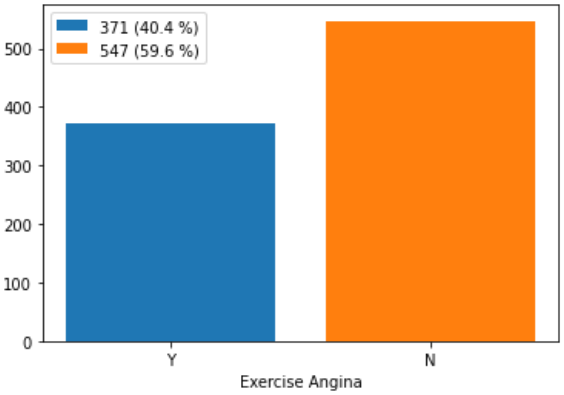




Gráfica que indica los resultados del electrocardiograma en reposo. Donde: Normal: condición normal, ST: anomalía de onda y LVH: hipertrofia ventricular.

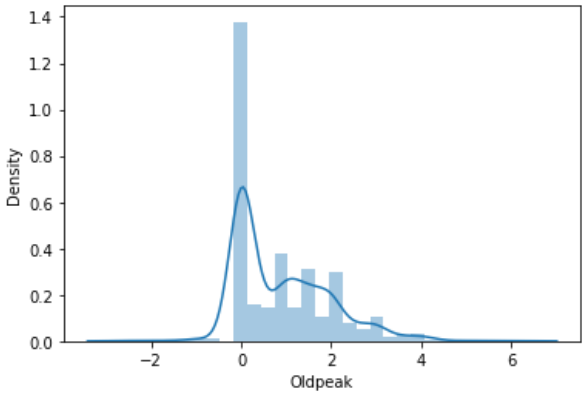
Gráfica que indica la frecuencia cardíaca máxima alcanzada de los pacientes. El rango de estos valores de encuentra entre 60 y 202.

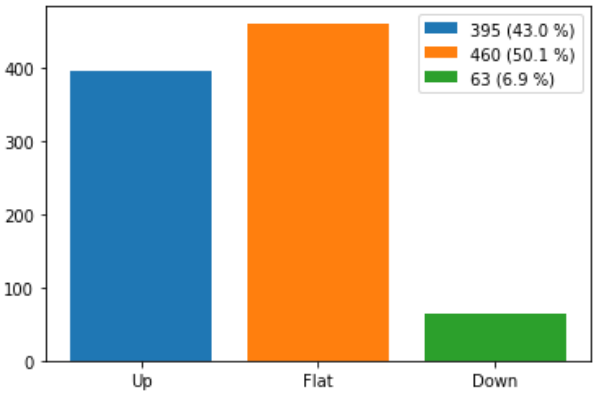




Gráfica que indica la angina inducida por el ejercicio. Donde: Y: si y N: no.

Gráfica que indica la depresión del ST inducida por el ejercicio en relación con el descanso.





Gráfica que indica la pendiente del segmento ST del ejercicio pico. Donde: UP: en pendiente hacia arriba, Flat: plano y Down: en pendiente hacia abajo.

3. Conclusiones

- El análisis exploratorio permitió conocer la naturaleza de los datos, los tipos de datos y los valores que contenía el dataset inicial. Luego de este análisis se pudo concluir que las características Cholesterol, RestingBP y MaxHR de tipo de dato numéricos continuos se pueden transformar a tipos de datos enteros discretos que representen categorías. Y que a las características Sex, ChestPainType, RestingECG, ExerciseAngina y ST_Slope se les debe aplicar la técnica One Hot Encoding durante el preprocesamiento de los datos.

4. Bibliografía

[1] L. González-Támara, “Una introducción a la estadística descriptiva y probabilidad Análisis exploratorio de datos,” Univ. Bogotá Jorge Tadeo Lozano., 2017.