UNIVERSIDAD TECNOLÓGICA DE LOS ANDES

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

Tesis

Algoritmos Machine Learning para mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024

Asesor:

Mag. Olivera Lazo, Maylin Vanesa

Autores:

León Chayña, Karime Erika

Ticona Gamarra, Michael Stevie Antony

Para optar el Título Profesional de:

Ingeniero(a) de Sistemas e Informática

Cusco - Cusco - Perú

2025

UNIVERSIDAD TECNOLÓGICA DE LOS ANDES

FACULTAD DE INGENIERIA

ESCUELA PROFESIONAL DE INGENIERIA DE SISTEMAS E INFORMATICA

Acta N°: 002-2025

ACTA DE SUSTENTACIÓN DE TÍTULO PROFESIONAL

En la ciudad de Cusco, a los 18 días del mes de agosto del 2025, siendo las 11:35 am horas, se reunieron los integrantes del Jurado designado por Resolución Subdirectoral N° 051-2025-UTEA-FI-EPISel- de la Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de INGENIERIA:

Presidente :	Mag. Moreano Córdova Javier
Dictaminante :	Ing. Poccori Umeres Godofredo
Replicante :	Ing. Quispe Salazar Moisés

Para evaluar la sustentación, en la modalidad de:

☒ Tesis ☐ Trabajo de suficiencia profesional

Titulada:

Algoritmos Machine Learning para mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024

Desarrollado por el (los) Bachiller (es):

Br.: León Chayña, Karime Erika
(Apellidos y Nombres)

Br.: Ticona Gamarra, Michael Stevie Antony
(Apellidos y Nombres)

Para optar el Título Profesional de:

Ingeniero(a) de Sistemas e Informática
(Denominación del Título)

Concluido el acto, el Jurado dictaminó que el (la) (los) mencionado(a) (s) bachiller (es) fue (ron) APROBADO (S):

Por: Mayoría
(Unanimidad o Mayoría) (*)

Emitiéndose el calificativo final de:

Bachiller (Apellidos y Nombres)	Calificación (**)
Br. León Chayña Karime Erika	Aprobado Notable
Br. Ticona Gamarra Michael Stevie Antony	Aprobado Notable

Siendo las 12.45 pm horas concluyó la sesión, firmando los integrantes del Jurado.

Presidente: Mag. Moreano Córdova Javier (Firma)

Dictaminante: Ing. Poccori Umeres Godofredo (Firma)

Replicante: Ing. Quispe Salazar Moisés (Firma)

(*): **Mayoría**: Dos integrantes del jurado aprueban o desaprueban; **Unanimidad**: Todos los integrantes del jurado aprueban o desaprueban, Art.18 RGGAT.
(**): 0 a 10: Desaprobado, 11 a 15: Aprobado, 16 a 18: Aprobado Notable, 19 y 20: Aprobado con Distinción, Art. 18 RGGAT.

Algoritmos Machine Learning para mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.docx

INFORME DE ORIGINALIDAD

19%	17%	5%	10%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	1%
2	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	1%
3	repositorio.unap.edu.pe Fuente de Internet	1%
4	Submitted to Universidad Tecnologica de los Andes Trabajo del estudiante	1%
5	repositorio.utp.edu.pe Fuente de Internet	1%
6	www.beaconhealthsystem.org Fuente de Internet	1%
7	ri.unsam.edu.ar Fuente de Internet	1%
8	repositorio.uandina.edu.pe Fuente de Internet	<1%
9	upcommons.upc.edu Fuente de Internet	<1%
10	repositorio.utea.edu.pe Fuente de Internet	<1%
11	repositorio.uss.edu.pe Fuente de Internet	<1%

Metadatos

Datos de los Autores		
Apellidos y Nombres	:	León Chayña, Karime Erika
Tipo de documento de Identidad	:	DNI
Número de Documento de Identidad	:	73745770
URL ORCID	:	https://orcid.org/0009-0007-0377-1221
Apellidos y Nombres	:	Ticona Gamarra, Michael Stevie Antony
Tipo de documento de Identidad	:	DNI
Número de Documento de Identidad	:	48510684
URL ORCID	:	https://orcid.org/0009-0006-6826-7029
Datos del Asesor		
Apellidos y Nombres	:	Mag. Olivera Lazo, Maylin Vanesa
Tipo de Documento de Identidad	:	DNI
Número de Documento de Identidad	:	24005444
URL ORCID	:	https://orcid.org/0000-0002-3491-3428
Datos de la investigación		
Facultad	:	Ingeniería
Escuela Profesional	:	Ingeniería de Sistemas e Informática
Línea de investigación	:	Informática, Sociedad y Gestión de Conocimiento
Rango de años en que se realizó la investigación	:	Abril 2024 – abril 2025
Fuente de financiamiento	:	Autofinanciado
Porcentaje de similitud	:	19%
URL de OCDE	:	https://purl.org/pe-repo/ocde/ford#2.02.04

Dedicatoria

Dedico este trabajo a mi madre, Luzmila.

Ejemplo de perseverancia y fortaleza, por ser fuente constante de inspiración y el pilar más importante en mi vida. Siempre fuiste mi refugio y mi apoyo incondicional, creyendo en mí incluso en los momentos más difíciles de mi vida.

Este logro no es solo mío, también es fruto tuyo, porque esto lo construimos juntas. Te amo infinitamente.

Karime Erika León Chayña

Dedico este trabajo de investigación a mi padre Tonio y mi madre Marieta por su apoyo en mi educación y crecimiento profesional a pesar de todas las dificultades, se logró nuestro objetivo.

A mis familiares, mi gratitud hacia ustedes es imposible de expresar completamente.

A mi hermana Winny J. Ticona Gamarra (QEPD), quien me motivo a seguir adelante y a quien prometí que terminaría mis estudios y titularme.

Michael Stevie A. Ticona Gamarra

Agradecimiento

Agradecemos profundamente a Dios, por iluminar nuestros caminos y brindarnos la fortaleza necesaria para culminar esta etapa académica, guiándonos con sabiduría y dándonos la perseverancia para superar cada desafío, fortaleciendo así nuestro crecimiento personal y profesional.

Expresamos nuestro sincero agradecimiento a nuestras familias, por su apoyo incondicional, amor constante y por ser nuestra principal fuente de motivación a lo largo de este proceso. Su confianza y presencia fueron esenciales para alcanzar esta meta.

De igual manera, extendemos nuestro reconocimiento a nuestra asesora de tesis, Magíster Maylin Olivera, por su dedicación, orientación y valiosas recomendaciones, que contribuyeron significativamente al desarrollo, solidez y culminación de esta investigación.

Asimismo, expresamos nuestra gratitud a la Universidad Tecnológica de los Andes, por brindarnos la oportunidad de formarnos profesionalmente, y a EsSalud, por facilitar el acceso a la información y los recursos necesarios que hicieron posible la realización de este estudio.

Karime Erika León Chayña

Michael Stevie A. Ticona Gamarra

Resumen

En esta investigación tiene como objetivo se evaluar los algoritmos de Machine Learning pueden mejorar la precisión en el diagnóstico de hipertensión arterial, utilizando datos clínicos reales de pacientes atendidos en el Centro Médico Santiago, Cusco 2024.

Para ello, se aplicaron modelos de clasificación supervisada mediante la metodología CRISP-DM, utilizando un conjunto de datos compuesto por 442 registros clínicos. Se desarrollaron y compararon tres algoritmos: Random Forest, Regresión Logística y Red Neuronal. Los resultados mostraron que el algoritmo Random Forest obtuvo el mejor desempeño general con una exactitud del 91.01%, precisión del 93.48%, sensibilidad del 89.58% y especificidad del 92.68%. No obstante, el modelo de Regresión Logística también alcanzó métricas notables, con una exactitud del 88.76% y una precisión del 95.24%, lo que lo convierte en una alternativa robusta y de fácil interpretación para el entorno clínico. Por su parte, la Red Neuronal presentó una exactitud del 78.65%, siendo su sensibilidad (72.92%) su principal limitación, aunque su desempeño fue aceptable.

En conclusión, los algoritmos de Machine Learning permiten mejorar significativamente la precisión del diagnóstico médico en hipertensión arterial. Si bien Random Forest lidera en rendimiento, la Regresión Logística también ofrece buenos resultados, y la Red Neuronal puede ser una opción válida con algunos ajustes adicionales.

Palabras Clave: Algoritmos de Machine Learning, precisión de diagnóstico, hipertensión arterial, métricas de evaluación, tratamiento de datos.

Abstract

This research aims to evaluate Machine Learning algorithms can improve the diagnostic accuracy of hypertension using real clinical data from patients at Santiago Medical Center, Cusco, in 2024.

The study followed the CRISP-DM methodology and used a dataset of 442 clinical records. Three supervised classification algorithms were developed and compared: Random Forest, Logistic Regression, and Neural Network (MLP). Results showed that Random Forest had the best overall performance, achieving 91.01% accuracy, 93.48% precision, 89.58% sensitivity, and 92.68% specificity. However, Logistic Regression also demonstrated strong performance with 88.76% accuracy and 95.24% precision, making it a reliable and interpretable model for clinical use. Meanwhile, the Neural Network achieved 78.65% accuracy, with its main limitation being lower sensitivity (72.92%), although still yielding acceptable results.

In conclusion, Machine Learning algorithms significantly enhance diagnostic precision for hypertension. While Random Forest provides the best results, Logistic Regression offers a solid alternative with lower computational demands, and Neural Networks may be suitable with further optimization.

Keywords: Machine Learning algorithms, diagnostic accuracy, high blood pressure, assessment metrics, data processing.

Índice

Índice de tablas

Índice de figuras

Índice de anexos

I. Introducción

Esta investigación busca mejorar el diagnóstico de la hipertensión arterial, logrando un primer resultado que siga los estándares establecidos para su detección. Para ello, se analizarán distintos algoritmos de Machine Learning y se identificará cuáles permiten una identificación más precisa de esta enfermedad.

La estructura de esta tesis es la siguiente:

Capítulo I: Introducción. Se establece la relevancia del tema y se presenta una preparación general para la comprensión de los capítulos que conforman la investigación.

Capítulo II: Planteamiento del problema. Se describe la situación problemática, identificando el problema general y los problemas específicos. Asimismo, se expone la justificación del estudio y se plantean los objetivos generales y específicos.

Capítulo III: Marco teórico. Se recopilan investigaciones previas a nivel nacional, internacional y local, que respaldan teóricamente el estudio. También se presenta el marco conceptual, que aporta información relevante para la investigación.

Capítulo IV: Metodología. Se detallan los métodos aplicados en la investigación, el tipo y nivel de estudio, así como la descripción de la población, las técnicas y los instrumentos utilizados para la recolección de datos.

Capítulo V: Resultados y discusión. Se exponen los hallazgos obtenidos, acompañados de un análisis detallado y su respectiva discusión.

Capítulo VI: Conclusiones. Se presentan las conclusiones de manera coherente y fundamentada, derivadas del estudio realizado.

Capítulo VII: Recomendaciones. Se formulan recomendaciones prácticas basadas en los resultados obtenidos a lo largo de la investigación.

II. Planteamiento del problema

2.1. Descripción y formulación del problema

A nivel mundial, los datos disponibles indican que en 2021 se estimaba que más de mil millones de personas en todo el mundo sufrían de tensión arterial elevada. Esto equivale a aproximadamente el 25% de la población adulta a nivel global (OMS, 2023).

Es importante tener en cuenta que factores como el envejecimiento de la población y cambios en los hábitos de vida pueden influir en estas estadísticas y provocar variaciones en el futuro. En nuestro país, numerosos individuos tienden a desarrollar enfermedades crónicas de larga evolución y progresión gradual, especialmente a medida que avanzan en la etapa del envejecimiento. Estas dolencias constituyen la principal causa de mortalidad a nivel global, contribuyendo en un 63% al total de defunciones registradas. Entre las afecciones más destacadas se incluyen la artritis, la artrosis, la diabetes y, por supuesto, la hipertensión arterial (Campos et al., 2022).

Esta última merece una atención prioritaria, ya que afecta a un gran número de individuos. Cuando no se controla de manera efectiva, pueden surgir complicaciones más severas que se convierten en factores críticos para provocar diversas afecciones a largo plazo, inclusive poniendo en peligro la vida del enfermo afectado. Por lo tanto, los estudios enfocados en este trastorno resultan de gran importancia, ya que no solo

contribuyen a prevenir complicaciones, sino que también mejoran el diagnóstico, permitiendo su detección en etapas tempranas (Carbo et al, 2022).

Según la OMS (2016), la problemática que estamos tratando se sustenta en una serie de hechos inquietantes. Se trata de una condición insidiosa que afecta a un número considerable de personas, a menudo sin mostrar síntomas evidentes. Esto significa que puede pasar inadvertida durante largos intervalos de tiempo, aumentando así el riesgo de complicaciones graves. Esta enfermedad ejerce un impacto sustancial en la calidad de vida de quienes la padecen, restringiendo su capacidad para llevar una vida activa y satisfactoria, y está asociada con otras enfermedades y el deterioro de órganos vitales, incluso con la posibilidad de resultar mortal. Si bien es más prevalente en individuos de edad avanzada, también puede manifestarse en adultos jóvenes, entre los 25 y 35 años, así como en adultos en general. La hipertensión arterial, comúnmente conocida como tensión arterial alta, se caracteriza por la persistente elevación de la presión en los vasos sanguíneos, lo que puede dar lugar a daños en el organismo. En cada latido del corazón, la sangre es impulsada a través de estos vasos, que tienen la responsabilidad de distribuirla por todo el cuerpo. La presión arterial se produce cuando la sangre ejerce presión contra las paredes de las arterias durante el proceso de bombeo del corazón. Cuando esta presión es elevada, el corazón debe realizar un mayor esfuerzo para propulsar la sangre (OMS, Preguntas y respuestas sobre la Hipertensión, 2016).

Para auxiliar a los profesionales médicos en la evaluación del estado de salud de un paciente o en la determinación de la gravedad de su condición, se utilizan directrices establecidas por organizaciones de renombre, tales como la Organización Mundial de la Salud (OMS), la International Society of Hypertension (ISH), la Sociedad Europea de Hipertensión (SEH) y la Sociedad Europea de Cardiología

(SEC). Estas instituciones desempeñan un papel fundamental en la definición de estándares para la evaluación de la presión arterial y la atención médica relacionada. Este trastorno tiene múltiples orígenes, y en ocasiones, se desarrolla debido a la interacción compleja de diversos factores. Entre estos elementos, se incluye la predisposición genética, ya que la presencia de antecedentes familiares de la enfermedad aumenta el riesgo de padecerla. Además, el estilo de vida juega un rol fundamental, ya que hábitos poco saludables, como la obesidad y el estrés, pueden contribuir a su desarrollo. La edad también se presenta como un factor de riesgo, ya que, con el envejecimiento, se incrementa la probabilidad de sufrir este mal. Otros factores que pueden desencadenar o agravar el problema incluyen la ingesta excesiva de sal, lo que provoca la retención de líquidos y, como consecuencia, un aumento de la presión arterial. Del mismo modo, el consumo excesivo de cafeína puede elevar temporalmente los valores de presión, y algunos medicamentos también pueden tener este efecto (Álvarez et al., 2022).

Si no se controla de manera efectiva, la condición puede acarrear graves consecuencias para la salud. Estas incluyen problemas cardiovasculares que aumentan el riesgo de ataques cardíacos y accidentes cerebrovasculares (conocidos como derrames cerebrales). Además, puede desencadenar enfermedades renales al dañar los vasos sanguíneos que irrigan los riñones, lo que a su vez puede resultar en insuficiencia renal. Existe también la posibilidad de la formación de aneurismas, que pueden provocar hemorragias internas. Por otro lado, se ha observado que la presión alta contribuye al desarrollo de trastornos cerebrales, como la demencia, y afecta negativamente la salud de las arterias (OMS, Preguntas y respuestas sobre la Hipertensión, 2016).

En Perú, la hipertensión arterial emerge como un problema de salud de gran magnitud. Según los datos proporcionados por el Ministerio de Salud del Perú y la Organización Panamericana de la Salud (OPS), alrededor del 23% de la población adulta en el país padece de esta condición. Asimismo, las cifras obtenidas revelan que en Perú existe un total de alrededor de 5.5 millones de individuos mayores de 15 años padecen de hipertensión arterial, lo que corresponde al 22% según los resultados de la Encuesta Demográfica y de Salud Familiar (ENDES). Estos datos subrayan la considerable prevalencia de la hipertensión arterial en el país, agravada por la presencia de diversos factores de riesgo. En consecuencia, resulta imperativo mejorar el acceso a servicios médicos más efectivos como un componente fundamental para hacer frente a esta problemática. (Ayuda Familiar, 2019).

Los métodos tradicionales de diagnóstico tienen limitaciones claras. Se basan en mediciones puntuales de la presión arterial y eso puede generar errores por la variación diaria, el efecto de “bata blanca” o porque algunos pacientes no muestran síntomas. El Machine Learning ofrece una alternativa distinta. Permite analizar gran cantidad de datos clínicos y encontrar patrones que no se ven con los métodos convencionales. Con estos algoritmos se pueden integrar variables como edad, sexo, índice de masa corporal, perímetro abdominal, glucosa, entre otras, lo que ayuda a mejorar la precisión del diagnóstico, detectar casos de manera temprana y predecir mejor el riesgo futuro. Por eso, aplicar modelos de aprendizaje automático resulta una herramienta valiosa para el diagnóstico y control de la hipertensión en el Perú y en particular en Cusco, que es el contexto de esta investigación.

En la Región Cusco, también se ha identificado una importante presencia de personas que padecen hipertensión arterial. De acuerdo con el INEI, en el año 2020, el 14,0 % de la población cusqueña de 15 años a más presentó hipertensión arterial, ya

sea medida por personal capacitado o diagnosticada por un médico. Esta prevalencia evidencia la magnitud del problema en nuestra comunidad y justifica la inclusión de estos casos en nuestra investigación, especialmente los registrados en el Centro Médico Santiago, Cusco, con el fin de comprender con mayor precisión los factores de riesgo, patrones diagnósticos y necesidades de intervención en la región (INEI, 2021).

Frente a esta realidad, nos formulamos la siguiente cuestión: ¿Cómo pueden los algoritmos de Machine Learning mejorar la exactitud en el diagnóstico de la hipertensión arterial en los pacientes del Centro Médico Santiago Cusco 2024? La toma de medidas destinadas a detectar, controlar y tratar esta enfermedad se vuelve esencial para elevar la calidad de vida de los pacientes y minimizar su impacto adverso en la salud. Dada la trascendencia de la hipertensión arterial, se hace imperativo concebir y poner en práctica diversas estrategias que faciliten un seguimiento efectivo en el tratamiento de aquellos que la padecen. Estas estrategias abarcan la concienciación sobre factores conductuales, tales como la importancia de realizar mediciones regulares de la presión arterial, así como la gestión adecuada de la medicación por parte de los profesionales de la salud. Además, resulta esencial aprovechar las tecnologías para la detección y monitoreo de esta afección.

En este escenario, resulta fundamental apostar por la innovación tecnológica como un complemento a los métodos convencionales. La incorporación de nuevas herramientas no solo busca ampliar las posibilidades de diagnóstico, sino también garantizar que los datos obtenidos sean fiables y precisos, lo que repercute directamente en la calidad de la atención que reciben los pacientes con presión arterial elevada. Actualmente, se desarrollan modelos predictivos que permiten una vigilancia más cercana, algunos basados en la lectura de señales eléctricas del corazón y otros en

dispositivos como tensiómetros adaptados que transmiten en tiempo real la información sobre la presión y el ritmo cardíaco a servidores web. Estas innovaciones representan un paso importante en la supervisión y control de la enfermedad, ya que ofrecen a médicos y pacientes recursos más prácticos y exactos para optimizar el tratamiento y prevenir complicaciones.

El propósito de esta investigación radica en el desarrollo de una herramienta que contribuya a perfeccionar el dictamen de la hipertensión arterial. En este contexto, se busca lograr un primer diagnóstico que cumpla con los estándares definidos para la medición y diagnóstico de esta enfermedad. Para alcanzar este propósito, se realizará un análisis y se determinará cuáles son los algoritmos de Machine Learning que pueden respaldarnos en la identificación más precisa de esta enfermedad objeto de estudio.

La formulación del problema es la siguiente:

2.1.1. Problema general

- ¿De qué manera los algoritmos de Machine Learning pueden mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024?

2.1.2. Problemas específicos

- **P.E.1** ¿Qué características son necesarias para construir un dataset eficaz que permita el diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024?
- **P.E.2** ¿Cuáles son los algoritmos de Machine Learning que permitan realizar un diagnóstico de la hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024?

- **P.E.3** ¿Qué nivel de precisión y efectividad alcanzan los diferentes algoritmos de Machine Learning en términos de precisión y efectividad para diagnosticar la hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024?

2.2. Objetivos

2.2.1. Objetivo General

- Evaluar de qué manera los algoritmos de Machine Learning pueden mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

2.2.2. Objetivos Específicos

- **O.E.1** Definir y seleccionar las características clave para la construcción de un dataset que facilite el diagnóstico preciso de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.
- **O.E.2** Determinar y evaluar los algoritmos de Machine Learning que permitan realizar un diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.
- **O.E.3** Comparar la precisión y efectividad de diversos algoritmos de Machine Learning en el diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

2.3. Justificación e importancia

La presente investigación se justifica en la necesidad de implementar instrumentos tecnológicos avanzadas, como el Machine Learning, para perfeccionar el diagnóstico de la hipertensión arterial, una condición crónica que afecta a millones de personas en todo el mundo (OMS, 2023).

Esta investigación no solo busca responder a una problemática de salud pública, sino también contribuir al avance del conocimiento empleando inteligencia artificial en el ámbito médico, especialmente en el Centro Médico Santiago, Cusco.

El incremento en la prevalencia de la hipertensión arterial, impulsado por el envejecimiento poblacional y la adopción de hábitos poco saludables, resalta la urgencia de contar con diagnósticos precisos y oportunos. La implementación de algoritmos de Machine Learning ofrece una oportunidad prometedora para analizar grandes volúmenes de datos, identificar patrones relevantes y realizar predicciones que apoyen la labor de los médicos, reduciendo la probabilidad de errores derivados del cansancio o el desconocimiento (OMS, 2023).

Desde una perspectiva científica, este estudio propone el uso de modelos predictivos que permitan mejorar la precisión diagnóstica, marcando un hito en la utilización de tecnologías avanzadas para abordar desafíos médicos. Además, los resultados obtenidos pueden ser aplicables en el desarrollo de dispositivos médicos inteligentes y en la implementación de estrategias efectivas para otras enfermedades.

En términos económicos, esta investigación se justifica por los beneficios que genera al optimizar los recursos médicos y reducir costos asociados con diagnósticos imprecisos, pruebas adicionales y complicaciones derivadas de un tratamiento tardío o inadecuado. Un diagnóstico temprano y eficaz permite mejorar la calidad de vida de los pacientes, prevenir complicaciones graves y disminuir los gastos en atención médica.

Por último, desde el ámbito social, esta investigación tiene el potencial de fortalecer el sistema de salud al ofrecer herramientas innovadoras para la toma de decisiones médicas, aumentando la accesibilidad y eficiencia en la atención. La investigación también tiene un impacto positivo en la formulación de políticas de salud

pública al identificar grupos de mayor riesgo y facilitar la implementación de estrategias preventivas.

En este contexto, el desarrollo de este estudio no solo beneficia a los pacientes del Centro Médico Santiago, sino que también sienta las bases para futuros avances en la ciencia médica, mejorando la atención sanitaria en el ámbito local y global.

En cuanto a la importancia, la investigación de algoritmos de Machine Learning para mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024 tiene importancia en el ámbito médico y en el ámbito social porque aborda un problema critico en relación con la salud que es la hipertensión arterial la importancia que tiene va a radicar en diferentes aspectos como:

a) El ámbito médico y científico, la investigación impulsa a que se modernice la práctica médica al incluir el uso de Machine Learning para optimizar el diagnóstico, lo que facilitaría la identificación temprana y precisa de los pacientes con hipertensión, que ayudaría a evitar complicaciones graves además de contribuir el avance tecnológico dentro de la medicina local, generando también conocimientos científicos que se puedan aplicar en otro tipo de patologías.
b) Para los pacientes representa la mejora de la calidad de la atención al reducir diagnósticos erróneos previniendo riesgos que se encuentren asociados a los diagnósticos preliminares o diagnósticos posteriores, de la misma forma va a democratizar el acceso a diferentes herramientas tecnológicas avanzadas en zonas descentralizadas como lo es la Región del Cusco.
c) Dentro del sistema de salud la investigación que se presenta contribuye en las eficiencias de los recursos médicos minimizando consultas innecesarias

optimizando tratamientos y fortaleciendo la capacidad del diagnóstico en la atención primaria que reduciría la sobrecarga en los hospitales.

d) Dentro del ámbito social y local la investigación responde a las necesidades específicas de la población esta investigación se adapta a las diferentes particularidades socioeconómicas y culturales.

Por todo lo mencionado esta investigación es esencial para mejorar la precisión diagnóstica, donde se va a poder prevenir complicaciones dentro de la salud además de modernizar el sistema médico y que además va a generar un impacto significativo dentro de la medicina.

2.4. Variables

2.4.1 Variable algoritmos de Machine Learning

El Machine Learning es una disciplina dentro de la inteligencia artificial cuyo objetivo es que los ordenadores tengan la capacidad de aprender a partir del análisis de datos, ya sea mediante ejemplos o experiencias, junto con el uso de algoritmos de aprendizaje, lo hace posible desarrollar modelos de predicción. (Weng, 2020).

2.4.2 Variable hipertensión arterial

Es una afección crónica y no contagiosa, de origen multifactorial, que puede ser controlada, aunque impacta negativamente en la calidad y esperanza de vida. Se vincula con múltiples factores de riesgo que incrementan la posibilidad de desarrollar un infarto agudo de miocardio (IAM), un accidente cerebrovascular (ACV), insuficiencia cardíaca (IC), hipertrofia cardíaca, entre otras complicaciones. (SAC, 2018).

Tabla 1

Operacionalización de las variables

VARIBLE INDEPENDIENTE	DIMENSIONES	INDICADORES
Algoritmos de Machine Learning El aprendizaje automático (Machine Learning, ML) Es un campo dentro de la inteligencia artificial que capacita a los sistemas informáticos para adquirir conocimiento a partir de los datos. Reconocer patrones y realizar predicciones sin necesidad de programación explícita, utilizando algoritmos que analizan situaciones y experiencias para la toma de decisiones (Weng, 2020).	**Redes Neuronales** Las Redes Neuronales imitan el cerebro humano para modelar relaciones no lineales en sistemas complejos. (Tian, 2020).	• Exactitud • Sensibilidad
	Random Forest Random Forest es un algoritmo de aprendizaje supervisado que utiliza múltiples árboles de decisión y muestras aleatorias para mejorar la precisión del modelo y minimizar el sobreajuste. (García, 2018).	• Precisión
	Regresión Logística La Regresión Logística es un método estadístico para clasificación binaria que estima probabilidades. (Weng, 2020).	• Especificidad • F1-score

VARIBLE INDEPENDIENTE	DIMENSIONES	INDICADORES
Hipertensión arterial Es una enfermedad crónica, no transmisible, de etiología múltiples factores, controlable que disminuye la calidad y esperanza de vida. Esta enfermedad se asocia a diversos factores de riesgo que aumentan la probabilidad de padecer un Infarto Agudo de Miocardio (IAM), Accidente Cerebrovascular (ACV), Insuficiencia Cardíaca (IC) e Hipertrofia Cardíaca, entre otros (OMS, 2023).	**Factores sociodemográficos** Los factores sociodemográficos, como la edad y el sexo, influyen significativamente en la prevalencia y control de la hipertensión arterial. (Doryńska et al., 2023)	• Edad • Sexo
	Factores antropométricos Los factores antropométricos mejoran el diagnóstico de hipertensión con Machine Learning. (Nepomuceno de Andrade et al., 2019).	• Peso • Estatura • Índice de masa corporal (IMC) • Perímetro abdominal • Categoría peso
	Factores clínicos y diagnósticos Los factores clínicos de la hipertensión incluyen presión arterial alta, antecedentes de HTA, diabetes y enfermedades renales, diagnosticados con pruebas médicas. (Torres et al., 2021).	• Diagnóstico de HTA • Diagnóstico de diabetes mellitus
	Factores hemodinámicos Los factores hemodinámicos regulan la presión arterial e influyen en la hipertensión, mejorando su diagnóstico con Machine Learning (Torres et al., 2021).	• Presión sistólica • Presión diastólica
	Factores metabólicos La glucosa influye en la hipertensión al afectar la función endotelial y la vasoconstricción, siendo clave en Machine Learning para detección temprana. (Torres et al., 2021).	• Glucosa

III. Marco teórico

3.1. Antecedentes

Barba (2021), en su tesis titulada *Estudio de técnicas de Machine Learning para el diagnóstico del melanoma y otras lesiones cutáneas a partir de imágenes,* desarrollada en la Universidad Oberta de Catalunya, en Barcelona, España, planteó como objetivo implementar un clasificador automático de lesiones de la piel. La metodología usada para este trabajo fue XP donde aplicamos el uso de sprint semanales conteniendo los avances y problemas presentados. Los resultados que se obtuvieron en este estudio fueron que se logró mejorar hasta un 0.75 de f1 – score y 0.86 de accuracy. El principal aporte de esta tesis radica en evidenciar que los algoritmos de Machine Learning contribuyen a incrementar la precisión en los diagnósticos médicos, lo que respalda la posibilidad de aplicar enfoques similares en la detección de la hipertensión arterial. Asimismo, el empleo de la metodología XP, basada en sprints de trabajo, ofrece un modelo flexible para la gestión y desarrollo de modelos predictivos.

Espasa (2022), en el trabajo de fin de grado titulado *Deep Learning en la detección de lesiones cutáneas malignas,* realizado en la Universidad Politécnica de Catalunya, en Barcelona, España, desarrolló un modelo predictivo, el objetivo para este trabajo es crear un modelo predictivo basado en Deep Learning para el diagnóstico del cáncer de piel, a la vez que estudiar su aplicación en un escenario real. La metodología de este estudio se basa en un análisis comparativo entre los resultados

obtenidos al emplear un modelo pre entrenado y un modelo auto entrenado, incluyendo la sistematización de los datos adquiridos. El modelo final alcanzó una precisión del 84.57% y un F1-Score de 76.92%. Uno de los hallazgos más significativos durante la optimización fue la mejora obtenida al utilizar un modelo pre entrenado en comparación con un modelo auto entrenado.

Estos resultados representan un antecedente importante para la presente investigación, porque demuestran que los modelos preentrenados pueden ofrecer un nivel superior de eficacia y optimización en tareas de diagnóstico médico. En este sentido, su aplicación en el análisis de variables clínicas relacionadas con la hipertensión arterial podría contribuir a mejorar la exactitud y eficiencia de los algoritmos predictivos utilizados en este estudio.

Quintanilla et al. (2022), en su artículo *Propuesta de una aplicación de aprendizaje automático con visión artificial como herramienta de apoyo para la detección de melanomas benignos y malignos*, desarrollado en el Departamento de Ciencias de la Computación de la Universidad de las Fuerzas Armadas ESPE, en Sangolquí, Ecuador, plantean como objetivo elaborar una propuesta para la implementación de un aplicativo. La metodología empleada en el estudio sigue fases estructuradas, que incluyen recolección, depuración, implementación, entrenamiento y evaluación. A partir de la propuesta, se concluyó que factores como el color, tamaño y forma de la lesión influyen en el diagnóstico temprano de melanomas, lo que permite realizar un análisis confiable, eficiente y preciso de la imagen. Además, se determinó que el uso y entrenamiento de redes neuronales facilita la detección y desciframiento de patrones, estableciendo correlaciones similares al razonamiento y aprendizaje humano. Como aporte a este proyecto, se destaca la utilidad de una metodología estructurada, que incluye fases como recolección, depuración, implementación,

entrenamiento y evaluación, lo cual es valioso para el desarrollo del modelo de diagnóstico de hipertensión. Asimismo, se resalta la importancia del entrenamiento del algoritmo para detectar patrones y correlaciones, lo cual es crucial para mejorar la precisión del diagnóstico.

Millán y Robles (2020), en su investigación de título *Modelo en Machine Learning para el diagnóstico del cáncer de mama*, elaborada en la Universidad Distrital Francisco José de Caldas, en Bogotá D.C., Colombia. El objetivo principal fue desarrollar un sistema con modelos de Machine Learning para diagnosticar el padecimiento de cáncer de mama. En la metodología, se adopta un enfoque descriptivo. A partir del análisis realizado, se determinó que los modelos de Deep Learning son los más utilizados, con un 28%. No obstante, dado que la investigación se enfoca en modelos de Machine Learning, se seleccionaron aquellos con el mayor porcentaje de uso después de los modelos de Deep Learning. Estos modelos incluyen: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM) y K-nearest Neighbors (KNN), siendo el modelo SVM el más utilizado en Machine Learning, con un 22%. Además, se eligió el modelo Gaussian Naive Bayes para complementar la investigación. El entrenamiento de los modelos seleccionados se realizó utilizando el dataset de cáncer de mama de la Universidad de Wisconsin, que contiene 569 registros, cada uno con 32 características. La librería Scikit-Learn se utilizó para el procesamiento de los datos y el posterior entrenamiento de los modelos. Como aporte, la investigación destaca la importancia de los modelos de Machine Learning en la detección del cáncer de mama, proporcionando bases sólidas sobre cómo aplicar técnicas similares en la detección de hipertensión. Además, la metodología descriptiva empleada permite la selección de modelos más utilizados, como Logistic Regression, Decision Trees, Random Forest, SVM y KNN, lo cual

resulta relevante para el desarrollo de la investigación, ya que orienta sobre los algoritmos más efectivos para el diagnóstico de enfermedades como la hipertensión. También se resalta la importancia de contar con datos de calidad para el entrenamiento del modelo.

Loyola y Chamorro (2021), en la *tesis Implementación de un sistema de diagnóstico clínico aplicando un modelo predictivo de Machine Learning para la detección de neumonía en el Hospital Villa Rebagliati de EsSalud, 2021* en Lima. El objetivo de este proyecto es implementar un sistema de diagnóstico clínico utilizando un modelo predictivo de Machine Learning para facilitar la detección de neumonía en el Hospital Villa Rebagliati de EsSalud. La metodología empleada en este estudio incluye un diseño experimental para medir el efecto de la variable independiente sobre la dependiente, así como un enfoque de investigación aplicada, orientado a resolver un problema práctico mediante la implementación de un modelo de predicción. Además, la investigación sigue un enfoque cuantitativo que permite probar la hipótesis. Los resultados evidenciaron que el sistema de diagnóstico y el modelo de predicción desarrollados constituyen una herramienta valiosa para los médicos, mejorando el tiempo de diagnóstico en un 33.7% en comparación con el sistema anterior utilizado en el hospital. La metodología experimental aplicada en esta tesis proporciona una guía clara sobre cómo estructurar el estudio y evaluar el impacto del modelo predictivo. Asimismo, los hallazgos en cuanto a la reducción del tiempo de diagnóstico demuestran que los modelos predictivos no solo aumentan la precisión, sino también la eficiencia en los procesos médicos. Este estudio destaca cómo un diseño experimental con enfoque cuantitativo permite medir con precisión el impacto de un modelo predictivo en su eficacia diagnóstica, lo que sugiere que la detección de hipertensión podría optimizarse significativamente. Por ello, se plantea la necesidad

de validar los modelos mediante métricas como sensibilidad y especificidad, garantizando su fiabilidad en la práctica clínica.

Laureano (2022), en su tesis presentada *Modelo de Machine Learning usando clasificador de máquinas de soporte vectorial para la detección y clasificación del cáncer de seno usando imágenes mamográficas,* desarrollada en la Universidad Nacional Del Altiplano. El objetivo de este estudio fue desarrollar un modelo de Machine Learning basado en un clasificador de máquinas de soporte vectorial para detectar y clasificar el cáncer de mama a partir de imágenes mamográficas. La metodología aplicada sigue un enfoque cuantitativo, ya que emplea métodos y técnicas de medición, uso de magnitudes y observación. Además, el diseño de investigación es cuasi-experimental, lo que permite evaluar el impacto del modelo predictivo. Los resultados obtenidos mostraron que el modelo alcanzó un 90% de exactitud, con métricas de desempeño detalladas por clase: benigno (precisión: 88.6%, sensibilidad: 83.8%, F1-score: 86.1%), maligno (precisión: 83.9%, sensibilidad: 96.3%, F1-score: 89.7%) y normal (precisión: 94.3%, sensibilidad: 89.2%, F1-score: 91.7%). Al evaluar el modelo con otros conjuntos de datos, se obtuvo un 89% en DDSM, 84% en BCDR y 83% en HRMNB, lo que evidencia su capacidad de generalización. Un aspecto clave de esta investigación es la aplicación de un modelo de Machine Learning validado con distintas métricas de desempeño, resaltando la importancia de utilizar un clasificador robusto y probarlo con diversos conjuntos de datos para medir su eficacia. Asimismo, el uso de un diseño cuasi-experimental con enfoque cuantitativo permite evaluar con mayor precisión el impacto del modelo predictivo, garantizando su aplicabilidad en el diagnóstico clínico.

Calderon et al. (2021), en la tesis de título *Uso de algoritmos de Machine Learning para el diagnóstico de melanomas,* desarrollada en la Universidad de Piura.

El objetivo de este estudio fue desarrollar una herramienta capaz de realizar un diagnóstico rápido y en tiempo real, permitiendo detectar la enfermedad a tiempo e iniciar el tratamiento adecuado. Para mejorar la precisión, se empleó la técnica de data segmentación. Se desarrollaron y compararon modelos basados en Redes Neuronales Convolucionales (RNC) y Random Forest, con el fin de determinar cuál ofrecía mejores métricas de desempeño. La programación se llevó a cabo en el entorno Google Colaboratory. En los modelos de RNC, se evaluaron las arquitecturas DenseNet-121, DenseNet-161, ResNet-34 y ResNet-50, mientras que en Random Forest, se aplicó la técnica de Principal Component Analysis (PCA) y las transformaciones L y P. Los resultados indicaron que el mejor modelo fue ResNet-34, ya que su capacidad para considerar la localidad de los píxeles le permitió superar a Random Forest, el cual no cuenta con esta propiedad debido al aplanamiento de los datos. El aporte que se debe considerar es que es necesario seleccionar el modelo adecuado de acuerdo a la naturaleza de los datos, en la investigación se ha usado la segmentación de los datos para la precisión del diagnóstico por lo que se puede deducir que el proyecto relacionado con la presión arterial sería adecuado que se segmenten los datos considerando variables como edad, genero, etc. Además de tener que evaluar diferentes arquitecturas de RNC, como DenseNet-121, DenseNet-161, ResNet-34 y ResNet-50, que se podría usar en el análisis de diferentes datos, también no se deben de dejar de lado el uso de técnicas para la transformación de datos.

Alarcon y Murga (2020), en su trabajo de grado titulado *Algoritmo para el diagnóstico preliminar de melanoma cutáneo basado en redes neuronales, Naive Bayes y Árboles de Decisión*, desarrollado en la Universidad Nacional De La Amazonía Peruana. El propósito de la investigación fue evaluar la precisión diagnóstica del algoritmo fusionado en la detección preliminar de melanoma cutáneo,

en comparación con los algoritmos de Redes Neuronales, Naive Bayes y Árbol de Decisiones. Se trata de un estudio de tipo aplicado, con un enfoque cuantitativo y un diseño pre-experimental. Se clasificó como investigación aplicada debido a su enfoque en la solución de un problema práctico. Los resultados muestran que los tiempos de ejecución de los algoritmos para el diagnóstico de imágenes dermatoscópicas de melanoma cutáneo fueron: 0.16 ms para el algoritmo fusionado, 0.16 ms para redes neuronales, 0.15 ms para Naive Bayes y 0.16 ms para Árboles de Decisión. El desempeño del algoritmo fusionado fue especialmente satisfactorio, destacando por su sensibilidad, especificidad, precisión y exactitud, lo que confirma su eficacia en el diagnóstico. Un aporte principal es el de la creación y la validación de un algoritmo fusionado donde se integran las redes neuronales, Narve Bayers y los Árboles de Decisión usados para el diagnóstico preliminar; en el caso del documento citado; del melanoma cutáneo, en las que se demuestran tiempos de ejecución competitivos (alrededor de 0.16 ms) y un desempeño sobresaliente en términos de sensibilidad, especificidad, precisión y exactitud. Esta solución evidencia que es viable la combinación de diferentes técnicas de Machine Learning para la mejora de la detección temprana de un diagnóstico clínico optimizando el diagnóstico y el tratamiento oportuno

Valero et al. (2021), en su artículo titulado *Detección de la tuberculosis con algoritmos de Deep Learning en imágenes de radiografías del tórax,* desarrollado en la Universidad Nacional de Moquegua, Ilo - Perú. El objetivo propuesto es mejor precisión para la detección de la tuberculosis. Se han utilizado tres algoritmos de aprendizaje profundo ampliamente reconocidos en el desarrollo de visión computacional: VGG19, MobileNet e InceptionV3, obteniendo resultados prometedores en la detección de la tuberculosis. En particular, MobileNet destacó

sobre los demás, mostrando un desempeño superior en diversas métricas de evaluación. Además, su arquitectura es menos compleja y los pesos generados tras el entrenamiento son significativamente menores en comparación con los otros dos modelos. Se concluye que MobileNet es el algoritmo de Deep Learning más eficiente en comparación con VGG19 e InceptionV3, ya que ofrece una mayor precisión en la detección de la tuberculosis, además de requerir menos recursos computacionales y un menor tiempo de procesamiento. El aporte de esta investigación radica en la selección de modelos de Deep Learning para la detección de enfermedades a través de imágenes médicas. Sin embargo, se considera la evaluación de algoritmos como VGG19, MobileNet e InceptionV3, resaltando la importancia de la precisión del modelo, su eficiencia y su capacidad de implementación en distintos entornos. En particular, MobileNet se distingue por su menor complejidad y mayor optimización en términos de procesamiento y almacenamiento, lo que lo convierte en una opción relevante para la detección de hipertensión arterial a partir de datos médicos. Por ello, es fundamental explorar modelos eficientes y livianos. Por otro lado, la metodología aplicada en la investigación citada sirve como referencia para seleccionar el algoritmo más adecuado y definir la estrategia de optimización en el desarrollo del modelo de Machine Learning, con el objetivo de mejorar la precisión en el diagnóstico de la hipertensión arterial.

Luna y Vargas (2022), en su tesis *Uso de inteligencia artificial para el diagnóstico de covid-19 a través de radiografía de tórax en el Hospital Nacional Adolfo Guevara Velasco, Hospital Regional y Hospital Antonio Lorena, Cusco-Perú, periodo 2020-2021*. El objetivo de la investigación fue: determinar la sensibilidad y la especificidad de la inteligencia artificial para el diagnóstico de COVID 19. En la metodología usada en el tipo de investigación fue experimental transversal,

retrospectivo. El diseño de la investigación es descriptivo. Los resultados muestran que el proyecto permitió evaluar la sensibilidad de la clasificación de imágenes de radiografía de tórax mediante aprendizaje automático, en comparación con el diagnóstico tradicional de COVID-19 a través de pruebas RT-PCR y antigénicas. Se determinó que el modelo de inteligencia artificial alcanzó una sensibilidad del 90.13%, una especificidad del 80.91%, un valor predictivo positivo del 70.24%, un valor predictivo negativo del 94.25% y una precisión del 83.98%, consolidándose como una herramienta eficaz para el diagnóstico de COVID-19. El principal aporte de esta investigación radica en la validación del uso de inteligencia artificial en el diagnóstico médico, demostrando que los modelos de aprendizaje automático pueden lograr altos niveles de sensibilidad y especificidad en la clasificación de imágenes médicas. La metodología utilizada, basada en la comparación con métodos tradicionales, resalta la importancia de validar los modelos con métricas clave como sensibilidad, especificidad, valor predictivo positivo y negativo, así como precisión general.

Rojas (2022), en la tesis presentada con el título *Clasificación de leucocitos en imágenes microscópicas de frotis sanguíneo usando Machine Learning y CNN*. El objetivo planteado es demostrar si la clasificación de leucocitos en imágenes microscópicas de frotis sanguíneo realizado con Machine Learning y CNN (Convolutional Neural Network) cumple un grado de error aceptable. La metodología usada: el tipo de investigación es básica, el nivel de la investigación es correlacional; el método de la investigación es la investigación de acción. El resultado de la investigación muestra que el uso de Machine Learning y de redes neuronales convulsionales para la clasificación de leucocitos en imágenes microscópicas se muestra como una solución viable, con un grado de error comparable o menor al de otras soluciones, también se identificó que los errores en los procedimientos del

laboratorio podrían variar de acuerdo al método de ejecución siendo principal que se consideren referencias como historias clínicas o fichas técnicas El uso de Google Colab hizo posible que el entrenamiento de los modelos se acelere lo que hizo que el tiempo de procesamiento se acelere de horas a minutos. Realizando la comparación con otras investigaciones mostro que las arquitecturas de Redes Neuronales son las que se utilizan más.

Este trabajo aporta de manera importante a la investigación, pues muestra que las Redes Neuronales son eficaces en la clasificación de datos médicos y pueden aplicarse también en la detección de la hipertensión arterial. El uso de Google Colab evidencia que es posible entrenar modelos complejos en menos tiempo, lo que facilita el desarrollo de sistemas predictivos en salud.

3.2. Bases Teóricas

3.2.1. Machine Learning

El Aprendizaje Automático (Machine Learning, ML) es una rama de la inteligencia artificial que posibilita que las computadoras operen sin requerir una programación específica, mediante el uso de algoritmos que analizan y procesan datos para facilitar la toma de decisiones y generar predicciones en contextos reales. Su finalidad es que los sistemas informáticos adquieran conocimientos a partir de datos, experiencias y ejemplos, identificando patrones y construyendo modelos predictivos capaces de anticipar resultados en nuevas situaciones. (Weng, 2020).

3.2.1.1. Algoritmo de Aprendizaje

Estos algoritmos de aprendizaje automático son pedazos de códigos que sirven de ayuda a los usuarios en la exploración y en análisis de manera conjunta con datos que son complejos buscando significados en ellos. Cada uno de los algoritmos es un grupo de instrucciones que no son infinitas y que van de paso en paso sin

equivocaciones para seguir una máquina para obtener objetivos determinados. El objetivo que se tiene en un modelo de aprendizaje automático es el de detectar o establecer patrones para que los usuarios usen para realizar predicciones y/o clasificar la información. En algoritmo de aprendizaje se encuentran patrones en los datos que se le otorga y los clasifica en grupos. Luego compara nuevos datos y los ubica en grupos y es así como puede predecir de que se trata. De este modo, los algoritmos de aprendizaje se fundamentan en códigos de programación y se emplean para llevar a cabo tareas específicas en el momento y contexto adecuados. Su propósito es analizar y detectar patrones similares en los objetos de estudio. Además, tienen la capacidad de realizar predicciones a través del autoaprendizaje, organizando información relacionada o con características similares para generar una salida informativa. Dentro de Machine Learning, los algoritmos se pueden clasificar en tres grandes categorías según la forma en que aprenden a partir de los datos: aprendizaje supervisado, no supervisado y semi-supervisado (MICROSOFT, 2021).

3.2.1.1.1. Categorías de Machine Learning

a) Aprendizaje Supervisado

El método de aprendizaje supervisado, consta del entrenamiento de una función para que este calcule las variables de salidas en función de los datos de entrada (Bhavsar, 2018), de tal manera se define como su objetivo el de incitar un modelo el cual pueda aprender a través de un conjunto de datos donde cada uno de los datos poseen atributos y etiquetas (Bueno, 2018).

A partir de una base de datos utilizada como ejemplos de entrenamiento con una etiqueta de destino específica, de tal forma que $\{(x1, y1), \dots, (xi, yi)\}$ donde xi corresponde al vector de salida del i-ésimo ejemplo y yi es su etiqueta, así mismo, su

objetivo es la creación de un modelo $g : X \rightarrow Y$, donde se pueda predecir la etiqueta Y a partir de futuros datos X (Bell, 2020).

b) Aprendizaje no Supervisado

Los modelos de aprendizaje no supervisado usan un conjunto de datos de entrada $\{x1, x2, \ldots, xn\}$ el cual no contiene una salida determinada, esto es demostrado en la función g; tal que $g: X \rightarrow Y$, para este algoritmo la función g tiene que realizar el mapeo sin una salida Y, este tipo de algoritmos encuentran patrones a partir de los datos que se pueden tomar como ruido no estructurado (Geras, 2021). De tal manera podemos definir que el principal objetivo de este tipo de aprendizaje es descubrir patrones entre las muestras y revelar las clases ocultas detrás de las características (Ortiz, 2019).

c) Aprendizaje Semi – Supervisado

Este tipo de modelos generalmente cuenta con una cantidad de datos de entrenamiento que pueden estar etiquetados y otra gran cantidad de datos que no poseen etiquetas; esto indica que los datos sin etiquetas al utilizarse conjuntamente con una pequeña cantidad de datos etiquetados, logran mejorar sustancialmente el aprendizaje (Jayaram, 2018).

3.2.1.2. Algoritmos de Machine Learning

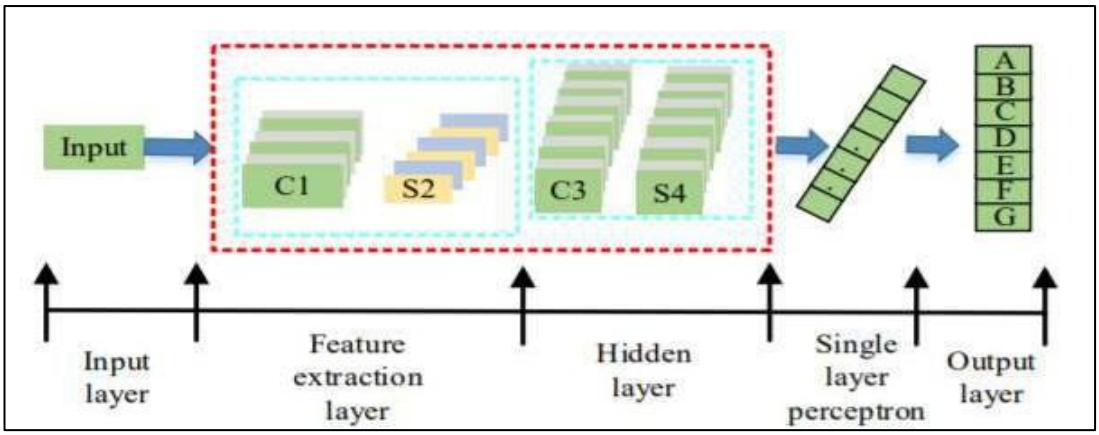
a) Algoritmos de Redes Neuronales

Entre los algoritmos de Machine Learning más destacados se encuentran las Redes Neuronales, las cuales están compuestas por unidades organizadas en capas interconectadas. Cada capa se vincula con las adyacentes, imitando el proceso de análisis y procesamiento de información del cerebro humano.

Estos componentes trabajan de manera conjunta y coordinada para resolver problemas específicos mediante el procesamiento de datos. Las Redes Neuronales suelen emplearse en la modelización de relaciones no lineales o en sistemas donde la interacción entre variables de entrada es altamente compleja y difícil de interpretar

Figura 1

Estructura de una Red Neuronal Convolucional



Nota. En la figura podemos visualizar como es la estructura de una Red Neuronal las cuales se componen por capas que trabajan de manera conjunta en la solución de problemas. Tomado de Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm. 8, 2. (Tian, 2020).

b) Regresión Logística.

Es un método de Regresión Logística y un método estadístico usado para la resolución de problemas en la clasificación binaria, en la que el resultado será de naturaleza dicotómica, lo que quiere decir que solo puede tener dos valores posibles, Es posible utilizarlo para detectar la probabilidad de que el evento se dé (Buitrago, 2020).

c) Random Forest

Es un algoritmo de aprendizaje supervisado que utiliza un enfoque de ensemble Learning para mejorar la precisión y robustez en tareas de clasificación y regresión. Combinan múltiples Árboles de Decisión, que trabajan juntos para que se genere una predicción consolidada. Lo que hace este algoritmo es generar muchos Árboles de

Decisión, pueden ser miles, se usan muestras aleatorias remplazando un conjunto de datos originales, en cada nodo de un árbol, el algoritmo evalúa un subconjunto de manera aleatoria sus características, que va a introducir diversidad entre los árboles. Random Forest es una herramienta poderosa y versátil, ampliamente utilizada en tareas como clasificación, predicción y selección de características, gracias a su capacidad para generalizar bien y manejar grandes volúmenes de datos (García, 2018).

3.2.1.3. Modelos de Machine Learning

a) Modelo de Predicción.

Los modelos predictivos son fundamentales en la minería de datos y por ello siempre es necesario revisar su desempeño y poder de predicción para mejorarlos o inclusive cambiarlos en sus parámetros debido a que los patrones de comportamiento pueden variar según el tiempo o momento en el que fueron desarrollados. (IBM, 2012).

Estos sistemas trabajan de manera conjunta para abordar problemas específicos a través del análisis de datos. Las redes neuronales suelen aplicarse en la identificación de relaciones no lineales o en escenarios donde la interacción entre variables de entrada presenta un alto grado de complejidad, lo que dificulta su interpretación. Será suficiente para entender la información contenida en los datos (Otero, 2018).

b) Modelo de Regresión.

Este de modelo es usado para predecir valores numéricos en función de las variables de entrada, El modelo de regresión es usado para instalar una relación matemática entre un conjunto de variables independiente o características y una variable dependiente continua (Montero, 2018).

El principal objetivo que tiene este modelo es el de comprender la relación existente entre la variable de entrada y predecir un valor numérico que este asociado a

la variable dependiente Por ejemplo, si tienes datos que representan el precio de las casas (variable dependiente) en función de características como el tamaño de la casa, el número de habitaciones, la ubicación, etc. (variables independientes), puedes utilizar un modelo de regresión para predecir el precio de una casa nueva en función de esas características. (Montero, 2018).

c) Modelo de Clasificación

Este modelo se utiliza para asignar etiquetas o categorías a datos nuevos o no etiquetados basándose en patrones identificados en datos de entrenamiento previamente etiquetados. El objetivo principal de un modelo de clasificación es aprender una función que pueda generalizar y asignar correctamente las etiquetas a nuevas instancias de datos. (Gil et al., 2022).

En un problema de clasificación, tienes un conjunto de datos de entrada junto con sus respectivas etiquetas o clases. El modelo de clasificación se entrena utilizando este conjunto de datos para aprender la relación entre las características de entrada y las etiquetas de salida. Luego, el modelo se puede utilizar para predecir la clase de un nuevo conjunto de datos que no ha sido visto durante el entrenamiento. (Gil et al., 2022).

Existen varios tipos de modelos de clasificación en inteligencia artificial, y la elección del modelo depende del problema específico y de las características de los datos. La elección del modelo dependerá de la naturaleza del problema, la cantidad y la calidad de los datos disponibles, así como otros factores. Los modelos de clasificación son fundamentales en una amplia variedad de aplicaciones, como reconocimiento de imágenes, diagnóstico médico, filtrado de spam, entre otros. (Gil el at., 2022).

3.2.1.4. Precisión diagnóstica en el aprendizaje automático (Machine Learning)

a) Precisión

En Machine Learning, la precisión es una métrica utilizada en problemas de clasificación para medir la exactitud de las predicciones positivas realizadas por un modelo. Se define como la proporción de verdaderos positivos (TP) con respecto al total de instancias clasificadas como positivas (TP + FP) (Raschka & Mirjalili, 2019).

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

b) Exactitud (Accuracy)

Es el equilibrio de muestras justamente clasificadas en relación con el total de muestras. Se calcula dividiendo el número de predicciones correctas (verdaderos positivos y verdaderos negativos) entre el número total de muestras (Calderon et al., 2021).

$$Exactitud = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

c) Sensibilidad (Recall)

También conocida como tasa de verdaderos positivos, es la proporción de muestras positivas correctamente identificadas entre el total de muestras positivas. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos (Calderon et al., 2021).

$$TPR = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

d) Especificidad (Specificity)

También conocida como tasa de verdaderos negativos, es la proporción de muestras negativas correctamente identificadas entre el total de muestras negativas. Se

calcula dividiendo el número de verdaderos negativos entre la suma de verdaderos negativos y falsos positivos (Calderon et al., 2021).

$$TNR = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

3.2.1.5. Fases para el tratamiento de dato en Machine Learning

Un proceso fundamental de Machine Learning es el tratamiento de datos, con el cual se puede garantizar la calidad y la eficacia que debe tener el modelo, para lo cual se deben de tener varias fases, una metodología que pueda apoyar para cumplir estas fases es la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) Proceso Estándar Intersectorial para la Minería de Datos, que es usada para estructurar proyectos de data science y Machine Learning (Wirth & Hipp, 2018).

a) Comprensión del negocio

La comprensión del negocio es la primera fase de la metodología CRISP-DM y consiste en entender los objetivos y necesidades de la organización para traducirlos en un problema de minería de datos bien definido. Su propósito es alinear el análisis técnico con las metas del negocio, identificando claramente qué se quiere lograr, qué datos se necesitan y cómo se evaluará el éxito del proyecto. Esta etapa es clave para asegurar que los resultados generen valor real y respondan a una problemática concreta de la empresa o institución (Wirth & Hipp, 2018).

b) Comprensión de datos

La comprensión de los datos es la segunda fase de la metodología CRISP-DM y tiene como objetivo recolectar, explorar y familiarizarse con los datos disponibles para evaluar su calidad y relevancia. En esta etapa se identifican posibles problemas como datos faltantes, inconsistencias o valores atípicos, y se obtienen primeros conocimientos que orientan las siguientes fases del proyecto. Es fundamental para

asegurar que los datos sean adecuados para el análisis y estén alineados con los objetivos del negocio definidos previamente (Wirth & Hipp, 2018).

c) Preparación de datos

La preparación de datos es la tercera fase de la metodología CRISP-DM y consiste en seleccionar, limpiar, transformar e integrar los datos que se utilizarán en el modelado. En esta etapa se eliminan errores, se manejan valores faltantes, se generan nuevas variables relevantes y se ajusta el formato de los datos para que sean compatibles con los algoritmos de análisis. Es una fase crítica, ya que la calidad del modelo depende en gran medida de la calidad de los datos preparados (Wirth & Hipp, 2018).

d) Modelado

El modelado es la cuarta fase de la metodología CRISP-DM y se centra en la aplicación de técnicas estadísticas o de Machine Learning para construir modelos predictivos o descriptivos. En esta etapa se seleccionan los algoritmos apropiados, se ajustan los parámetros y se entrenan los modelos con los datos preparados. También se evalúa el rendimiento inicial de los modelos para identificar cuál ofrece mejores resultados. El éxito del modelado depende tanto de la técnica elegida como de la calidad de los datos utilizados (Wirth & Hipp, 2018).

e) Evaluación del modelado

La evaluación de los modelos es la quinta fase de la metodología CRISP-DM y tiene como objetivo determinar la calidad y utilidad del modelo generado. En esta etapa se analizan métricas de rendimiento (como exactitud, precisión, recall, F1-score, entre otras) para verificar si el modelo cumple con los objetivos del negocio establecidos en la primera fase. También se revisa si existen errores sistemáticos o sesgos, y se comparan diferentes modelos para seleccionar el más adecuado.

Finalmente, se decide si el modelo está listo para ser implementado o si requiere ajustes adicionales (Wirth & Hipp, 2018).

f) Despliegue

En la metodología CRISP-DM es la etapa final en la que se implementa el modelo de minería de datos en un entorno real para su uso práctico. Aunque el objetivo del proyecto puede no ser simplemente construir un modelo, sino obtener conocimiento útil, esta fase implica planificar cómo se utilizarán los resultados, desarrollar sistemas de reporte o aplicaciones que integren el modelo, capacitar a los usuarios y establecer un plan de mantenimiento. El despliegue puede ser tan simple como un informe o tan complejo como una solución automatizada integrada en procesos empresariales (Wirth & Hipp, 2018).

3.2.2. Hipertensión Arterial.

Es una enfermedad crónica, no transmisible, de etiología multifactorial, controlable, que disminuye la calidad y expectativa de vida. Esta enfermedad se asocia a diversos factores de riesgo que aumentan la probabilidad de padecer un Infarto Agudo de Miocardio (IAM), Accidente Cerebrovascular (ACV), Insuficiencia Cardíaca (IC) e Hipertrofia Cardíaca, entre otros (OMS, 2023).

La hipertensión arterial es el principal factor de riesgo que ocasiona daños en los distintos órganos del cuerpo humano generando complicaciones

- Corazón: relajación disminuida del ventrículo izquierdo, dilatación de la aurícula izquierda, arritmias, insuficiencia cardíaca.
- Vasos sanguíneos: presencia de aterosclerosis en arterias carótidas, aumento del grosor carotídeo circundante, aumento de la presión del pulso dependiente de la edad, velocidad en la onda de pulso carotídeo femoral, la rigidez de las grandes arterias y enfermedad arterial periférica de las extremidades inferiores.
- Riñones: función renal reducida con detección de albuminuria.

- Ojos: hemorragias retinianas, micro aneurismas, exudados duros o algodonosos, papiledema.
- Cerebro: daño cerebral, accidente isquémico transitorio, ictus, infartos lacunares, micro sangrados y atrofia cerebral.

3.2.2.1. Síntomas.

La presión arterial alta no tratada aumenta el riesgo de ataque cardíaco, accidente cerebrovascular y otros problemas de salud graves. Es importante controlar la presión arterial al menos cada dos años a partir de los 18 años. Algunas personas necesitan controles con mayor frecuencia. otros (OMS, 2023).

La mayoría de las personas con presión arterial alta no tienen síntomas, incluso si las lecturas de presión arterial alcanzan niveles peligrosamente altos. Se puede tener presión arterial alta durante años sin presentar ningún síntoma. otros (OMS, 2023).

Algunas personas con hipertensión arterial pueden presentar lo siguiente:

- Dolores de cabeza
- Falta de aire
- Sangrados nasales

Sin embargo, estos síntomas no son específicos. No suelen aparecer hasta que la presión arterial alta haya alcanzado un estado grave o pone en riesgo la vida. (OMS, 2023).

3.2.2.2. Tratamiento

Cambiar el estilo de vida puede ayudar a controlar la presión arterial alta. Es posible que el proveedor de atención médica te recomiende hacer cambios en el estilo de vida, que incluyen los siguientes:

- Seguir una dieta saludable para el corazón con menos sal
- Hacer actividad física con regularidad

- Mantener un peso saludable o bajar de peso
- Limitar el consumo de alcohol
- No fumar
- Dormir de 7 a 9 horas diarias

En algunos casos, los cambios en el estilo de vida por sí solos no son suficientes para controlar la presión arterial alta. Si estos no generan los resultados esperados, el profesional de salud puede sugerir el uso de medicamentos para ayudar a reducirla (OMS, 2023).

Los medicamentos que se utilizan para tratar la presión arterial alta incluyen los siguientes:

- Diuréticos
- Inhibidores
- Bloqueadores de los canales de calcio.

Otros fármacos que suelen emplearse en el tratamiento de la presión arterial alta Si las combinaciones de medicamentos previamente mencionadas no logran reducir eficazmente los niveles de presión arterial, el profesional de salud podría considerar otras opciones terapéuticas adicionales.

- Alfabloqueadores.
- Alfabetabloqueadores.
- Betabloqueadores.
- Antagonistas de la aldosterona.
- Inhibidores de la renina
- Vasodilatadores.
- Agentes de acción central.

3.2.2.3. Factores de riesgo asociados a la enfermedad.

Según Thomas (2022), hay muchos factores de riesgo que pueden causar presión arterial alta, como los siguientes:

a) Factores de riesgo sociodemográficos

Edad. El riesgo de tener presión arterial alta aumenta con la edad. Hasta aproximadamente los 64 años. Las mujeres tienen más probabilidades de desarrollar presión arterial alta después de los 65 años (Thomas, 2022).

Se utilizarán los siguientes rangos de edad para el desarrollo de la investigación teniendo en cuenta la presencia de posible hipertensión

- **18-29 años** (Jóvenes adultos) → Hipertensión menos común, posible influencia del estilo de vida.
- **30-39 años** (Adultos jóvenes) → Aumento gradual del riesgo, inicio de enfermedades crónicas.
- **40-49 años** (Adultos de mediana edad) → Mayor prevalencia de hipertensión, factores metabólicos comienzan a influir.
- **50-59 años** (Adultos maduros) → Alto riesgo de hipertensión debido a cambios cardiovasculares y metabólicos.
- **60-69 años** (Adultos mayores) → Prevalencia alta, comorbilidades asociadas.
- **70+ años** (Tercera edad) → Hipertensión frecuente, importancia del monitoreo constante.

Sexo. El sexo es un factor de riesgo sociodemográfico para la hipertensión debido a diferencias biológicas y hormonales. Los hombres son más propensos a desarrollarla a edades tempranas, influenciados por el estilo de vida, mientras que en las mujeres el riesgo aumenta tras la menopausia por la disminución de estrógenos y durante el embarazo por predisposición a hipertensión crónica (Thomas, 2022).

b) Factores antropométricos.

Peso. Se relaciona con el índice de masa corporal porque este es calculado usando el peso en kilogramos (Rodriguez, 2023).

Altura. También tiene un papel importante para el cálculo del índice de masa corporal (Rodriguez, 2023).

Índice de masa corporal. Es la medida usada para poder evaluar si la persona tiene un peso saludable con relación a su altura, se considera un factor importante dentro del contexto de hipertensión arterial porque el sobre peso y la obesidad son factores de riesgo (Rodriguez, 2023).

Perímetro abdominal. Es considerado un factor de riesgo para la hipertensión debido a su asociación con la obesidad central, un componente principal del síndrome metabólico. Este indicador refleja la acumulación de grasa visceral, que tiene un impacto directo en el sistema cardiovascular y el control de la presión arterial. Los varones tienen mas riesgo de obesidad abdominal antes de los 50 años, por otro lado, en las mujeres el riego aumenta después de la menopausia por los cambios hormonales (Thomas, 2022).

Categorías por perímetro abdominal (según OMS y criterios clínicos)

Para hombres*:*

- **Normal:** Menos de 94 cm
- **Riesgo aumentado:** Entre 94 cm y 101 cm
- **Riesgo muy alto:** 102 cm o más

Para mujeres:

- **Normal:** Menos de 80 cm
- **Riesgo aumentado:** Entre 80 cm y 87 cm
- **Riesgo muy alto:** 88 cm o más

Categorías de peso. Es un factor que se asocia de manera importante en el desarrollo de la hipertensión arterial y en su control la clasificación para las categorías del peso está basada en el índice de masa corporal, así como las mediciones del perímetro abdominal los cuales van a ayudar a la identificación del riego de hipertensión (Thomas, 2022).

Estas categorías son:

- Peso bajo = < 18.5 Riesgo reducido
- Peso normal = 18.5 – 24.9 Menor riesgo
- Sobrepeso = 25.0 – 20.9 Riesgo moderado
- Obesidad I = 30.0 – 34.9 Alto riesgo
- Obesidad II = 35.0 – 39.9 Riesgo muy alto
- Obesidad III = >= 40.0 Riesgo extremo

c) Factores hemodinámicos

Presión sistólica. Es la presión máxima que la sangre ejerce contra las paredes de las arterias cuando el corazón se contrae; se conoce como fase sístole, para bombear la sangre al cuerpo, en general se considera que un valor menor a 120 mmHg se considera normal (Thomas, 2022).

Presión diastólica. Es la presión mínima que practica la sangre contra las paredes de la arteria cuando el corazón se encuentra en reposo; conocido como fase diástole; entre latidos, llenándose de sangre, el valor normal para la presión diastólica es menor de 80 mmHg (Thomas, 2022).

d) Factores clínicos y diagnóstico

Diabetes Mellitus. La diabetes mellitus y la hipertensión arterial son enfermedades crónicas que frecuentemente coexisten, aumentando el riesgo de complicaciones cardiovasculares y renales. La diabetes se caracteriza por niveles

elevados de glucosa en sangre debido a la insuficiencia o resistencia a la insulina, mientras que la hipertensión implica una presión arterial elevada. Ambas condiciones comparten factores de riesgo como la obesidad, el sedentarismo y la mala alimentación. Su control requiere una dieta saludable, ejercicio, monitoreo constante y, en muchos casos, tratamiento farmacológico para prevenir complicaciones graves (Torres, et al., 2021).

e) Factores metabólicos

Niveles de glucosa. En el contexto de la Diabetes Mellitus (DM) juega un papel fundamental, ya que esta enfermedad se caracteriza por una alteración en la regulación de los niveles de glucosa en sangre. La glucosa es la principal fuente de energía para las células del cuerpo, y su concentración en la sangre debe mantenerse dentro de ciertos rangos para asegurar un funcionamiento adecuado del organismo. En las personas con diabetes, la glucosa se encuentra en niveles más altos de lo normal debido a un defecto en la producción o el uso de la insulina. La hipertensión y la diabetes generalmente coexisten ya que los niveles de glucosa elevados dañan las arterias y los riñones que van a contribuir a la presencia de la hipertensión (Araya, 2019).

3.2.2.4. Consecuencias de Hipertensión Arterial.

a) Derrame Cerebral. La hipertensión arterial puede provocar la ruptura u obstrucción de los vasos sanguíneos en el cerebro más fácilmente (ERC, 2021).

b) Insuficiencia Cardíaca. La hipertensión arterial puede provocar el agrandamiento del corazón y afectar su capacidad para bombear sangre de manera eficiente. al cuerpo (ERC, 2021).

c) Pérdida de la Vista. La presión arterial alta puede dañar los vasos sanguíneos en los ojos (ERC, 2021).

d) **Ataque Cardíaco.** La presión arterial alta daña las arterias y hace que se estrechen y se endurezcan (ERC, 2021).

e) **Enfermedad/ Insuficiencia Renal.** La presión arterial alta puede dañar las arterias alrededor en los riñones e interferir con su capacidad de filtrar sangre eficazmente (ERC, 2021).

3.3. Definición de términos

- **Pandas**

Es una biblioteca de Python de código abierto utilizada para la manipulación y análisis de datos. Proporciona estructuras de datos flexibles y eficientes, como DataFrame y Series, que permiten trabajar con datos tabulares de manera similar a las hojas de cálculo o bases de datos (McKinney, 2018).

- **Seaborn**

Biblioteca de Python basada en Matplotlib que facilita la creación de gráficos estadísticos atractivos y bien diseñados. Se integra con Pandas para visualizar datos estructurados de manera intuitiva y proporciona herramientas para analizar relaciones, distribuciones y tendencias en conjuntos de datos (Garreta, 2020).

- **matplotlib.pyplot**

Módulo de la biblioteca Matplotlib en Python, utilizado para crear gráficos de forma sencilla y similar a MATLAB. Es una de las herramientas más utilizadas en visualización de datos en ciencia de datos y Machine Learning (Torres, 2020).

- **sklearn.datasets**

El módulo sklearn.datasets en Scikit-Learn proporciona conjuntos de datos predefinidos para pruebas y aprendizaje de algoritmos de Machine Learning. Incluye tanto conjuntos de datos de juguete como aquellos disponibles en bases de datos públicas (Garreta, 2020).

- **sklearn.model_selection**

El módulo sklearn.model_selection en Scikit-Learn proporciona herramientas para dividir conjuntos de datos, evaluar modelos y realizar búsquedas de hiperparámetros. Es fundamental para la validación y optimización de modelos de Machine Learning (Garreta, 2020).

- **sklearn.pipeline**

El módulo sklearn.pipeline en Scikit-Learn permite encadenar múltiples pasos en un flujo de trabajo de Machine Learning. Facilita la integración de preprocesamiento, selección de características y entrenamiento de modelos en un solo objeto, optimizando el código y evitando fugas de datos (data leakage) (Garreta, 2020).

- **sklearn.compose**

El módulo sklearn.compose permite combinar y aplicar múltiples transformaciones a diferentes tipos de datos en un solo paso. Es especialmente útil cuando se tienen datos heterogéneos (por ejemplo, columnas numéricas y categóricas) y se necesita aplicar distintos preprocesamientos antes de entrenar un modelo de Machine Learning (Garreta, 2020).

- **sklearn.preprocessing**

El módulo sklearn.preprocessing proporciona herramientas para transformar datos antes de entrenar modelos de Machine Learning. Estas transformaciones ayudan a mejorar el rendimiento del modelo al normalizar, estandarizar y manejar valores categóricos o ausentes (Garreta, 2020).

- **sklearn.feature_selection**

El módulo sklearn.feature_selection permite seleccionar las características más relevantes de un conjunto de datos, mejorando el rendimiento del modelo y reduciendo el ruido en los datos. Es útil para eliminar variables irrelevantes, reducir el tiempo de entrenamiento y mejorar la interpretabilidad del modelo (Garreta, 2020).

- **sklearn.decomposition**

El módulo sklearn.decomposition se utiliza para la reducción de dimensionalidad, permitiendo representar los datos en un espacio más pequeño sin perder demasiada información. Esto ayuda a mejorar la eficiencia de los modelos de Machine Learning, reducir el ruido y mejorar la interpretabilidad (Garreta, 2020).

- **sklearn.ensemble**

El módulo sklearn.ensemble implementa algoritmos de aprendizaje en conjunto (ensemble Learning), que combinan múltiples modelos (como árboles de decisión) para mejorar la precisión y reducir el sobreajuste en Machine Learning. (Garreta, 2020)

- **sklearn.linear_model**

El módulo sklearn.linear_model proporciona algoritmos de regresión y clasificación lineal, útiles para modelar relaciones entre variables en Machine Learning (Garreta, 2020).

- **sklearn.svm**

El módulo sklearn.svm implementa Máquinas de Soporte Vectorial (SVM), un poderoso conjunto de algoritmos de clasificación y regresión. SVM busca encontrar un hiperplano óptimo que separe los datos con el mayor margen posible (Garreta, 2020).

- **sklearn.neighbors**

El módulo sklearn.neighbors implementa algoritmos basados en vecinos más cercanos (K-Nearest Neighbors, KNN), útiles para clasificación, regresión y reducción de dimensionalidad (Garreta, 2020).

- **sklearn.tree**

El módulo sklearn.tree implementa árboles de decisión para clasificación y regresión, ofreciendo interpretabilidad y flexibilidad en problemas de Machine Learning (Garreta, 2020).

- **sklearn.metrics**

El módulo sklearn.metrics proporciona herramientas para evaluar modelos de clasificación, regresión y clustering en Machine Learning (Garreta, 2020).

- **Google Colaboraty**

Es una plataforma gratuita basada en la nube que permite ejecutar código Python en notebooks de Jupyter, sin necesidad de instalación ni configuración en la computadora. Es ampliamente utilizada en Machine Learning, ciencia de datos y análisis de datos (López, 2022).

- **Sweetviz**

Sweetviz es una biblioteca de Python que genera informes visuales detallados sobre un conjunto de datos en pocos segundos. Es útil para análisis exploratorio de datos (EDA) y comparaciones entre conjuntos de datos (Garreta, 2020).

- **Algoritmo de búsqueda incremental**

El algoritmo de búsqueda incremental es un método utilizado para buscar soluciones paso a paso, explorando el espacio de búsqueda de manera progresiva en lugar de realizar una búsqueda completa desde el inicio. Se usa en problemas donde la solución puede encontrarse de forma iterativa, refinando los resultados en cada paso (Vargas, 2021).

- **PCA**

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad utilizada en Machine Learning y estadística. Su objetivo es transformar un conjunto de datos con muchas variables en un nuevo conjunto de variables (componentes principales) que capturan la mayor cantidad de información con la menor cantidad de dimensiones posibles (Pérez, 2020).

- **Dataset**

Un dataset (conjunto de datos) es una colección estructurada de información organizada en filas y columnas, similar a una tabla en una base de datos o en una hoja

de cálculo. Se utiliza en Machine Learning, análisis de datos y estadística para entrenar modelos, realizar estudios y extraer conclusiones (Pérez, 2020).

- **Predicción**

La predicción es el proceso de estimar valores futuros basándose en datos históricos y patrones identificados en un conjunto de datos. Se utiliza en áreas como Machine Learning, estadística, economía, salud, meteorología y más (Pérez, 2020).

- **Normalización**

La normalización es un proceso de preprocesamiento de datos en el que se ajustan los valores de una variable a una escala común. Se utiliza en Machine Learning, estadística y minería de datos para mejorar la precisión y estabilidad de los modelos (Pérez, 2020).

- **Imputación**

Es el procedimiento mediante el cual se reemplazan los datos faltantes por valores plausibles, con el fin de obtener una base de datos completa y coherente para su análisis (Useche y Mesa, 2006).

- **Sprints**

Un Sprint es el corazón de Scrum; se define como un bloque de tiempo limitado a un mes o menos, durante el cual se crea un incremento de producto terminado, utilizable y potencialmente desplegable. Cada nuevo Sprint comienza inmediatamente después de la finalización del anterior (Schwaber & Sutherland, 2013).

- **Dermatoscopía**

La dermatoscopía es una técnica diagnóstica no invasiva que permite observar estructuras de la piel no visibles a simple vista, mejorando la exactitud diagnóstica en la evaluación de lesiones cutáneas pigmentarias. Su aplicación contribuye a diferenciar lesiones benignas de malignas y a disminuir procedimientos innecesarios de extirpación (Ramírez & Martínez, 2021).

IV. Metodología

4.1. Tipo y nivel de investigación

El presente trabajo de investigación se enmarca dentro de la investigación básica, dado que su objetivo principal es generar conocimientos teóricos que permitan explorar y explicar determinados fenómenos, sin una aplicación práctica inmediata. En este sentido, se orienta a la comprensión profunda de cómo y por qué ocurren dichos fenómenos, contribuyendo al desarrollo de conceptos, principios y teorías que pueden servir de base para futuras investigaciones en la materia. Como señala Borja (2012), la investigación básica se caracteriza por su enfoque en la construcción teórica, lo que la distingue de otros tipos de investigación con fines aplicados.

La investigación desarrollada se clasifica como de tipo básica, dado que su propósito esencial es la generación de conocimientos teóricos respecto al uso de algoritmos de Machine Learning en la mejora de la precisión diagnóstica de la hipertensión arterial, además que los resultados alcanzados podrían ser aprovechados en aplicaciones prácticas futuras, el estudio se centra en analizar y comprender el comportamiento de diversos algoritmos frente a los datos clínicos de los pacientes atendidos en el Centro Médico Santiago.

El nivel de investigación del trabajo es descriptivo, ya que se enfoca en la recopilación, análisis y presentación de información con el fin de detallar las características, propiedades o comportamientos del fenómeno objeto de estudio. Su propósito es brindar una representación clara y sistemática de la realidad investigada,

identificando cómo se manifiestan ciertos aspectos sin ahondar en las causas o relaciones causales entre las variables. Conforme a lo señalado por Hernández, Fernández y Baptista (2014), los estudios descriptivos permiten especificar las propiedades, características y perfiles de los fenómenos analizados, sin centrarse en las relaciones causales entre variables, lo que facilita una comprensión estructurada y objetiva del problema de estudio.

Esta investigación se clasifica como descriptiva porque se enfoca en recoger, analizar y organizar información clínica para mostrar cómo se comportan distintos algoritmos de Machine Learning al diagnosticar hipertensión arterial en los pacientes del Centro Médico Santiago. El objetivo es presentar de manera clara y ordenada los resultados obtenidos, sin buscar explicar las causas del problema, sino describir el rendimiento de los modelos en términos de precisión.

4.2. Ámbito temporal y espacial

4.2.1. Ámbito temporal

El presente estudio se desarrolló en un período de diez meses, comprendido entre abril de 2024 y abril de 2025. Durante este periodo, se llevó a cabo la recolección de datos, garantizando un adecuado registro de la información necesaria para el análisis. Asimismo, para el procesamiento de los datos, se trabajará con la totalidad de la población disponible.

4.2.2. Ámbito espacial

La investigación que se presenta se desarrolló en el Centro Médico de Santiago de la Región Cusco.

4.3. Población y muestra

4.3.1. Población

De acuerdo con Córdova (2003), una población es el conjunto total de elementos, personas u objetos que comparten características, cualitativas o cuantitativas, y que son objeto de estudio en una investigación. Esta población puede estar conformada por un número limitado o ilimitado de elementos, dependiendo de la naturaleza y propósito que se analiza y los objetivos del estudio.

La investigación se fundamenta en una población de 442 historias médicas de pacientes atendidos en el Centro Médico Santiago, Cusco. La base de datos incluye a pacientes diagnosticados con hipertensión, como a los que no tienen hipertensión. Para el análisis, se trabajará con la totalidad de la población disponible.

4.4. Instrumentos

Para el desarrollo de la presente investigación, se empleó una base de datos construida a partir de las historias médicas de los pacientes atendidos en el Centro Médico Santiago, Cusco. Dicha base de datos recopiló información relevante para el análisis y el entrenamiento de los algoritmos de Machine Learning, con el propósito de mejorar la precisión en el diagnóstico de hipertensión arterial.

El instrumento utilizado para la recolección de datos fue una ficha de observación (Anexo 02), diseñada específicamente para extraer variables clave de los registros clínicos.

También se usaron herramientas como Google COLAB y sus librerías para el análisis de datos.

4.5. Procedimientos

El desarrollo de la presente investigación se realizó siguiendo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), reconocida como un

estándar en el ámbito de la minería de datos. Esta metodología se compone de seis fases interrelacionadas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Su estructura flexible y cíclica permite organizar de manera ordenada cada etapa del proceso analítico, asegurando que los resultados obtenidos respondan a los objetivos planteados y que aporten valor a la toma de decisiones (IBM, 2015).

Siguiendo la estructura propuesta por la metodología mencionada, la investigación se desarrolló en fases consecutivas que orientaron desde la definición del problema hasta la evaluación de los modelos, las cuales se explican a continuación.

4.5.1. Fase 1: Comprensión del negocio

Objetivo del proyecto

El objetivo de la investigación es desarrollar modelos de aprendizaje automático que permitan mejorar la precisión del diagnóstico de hipertensión arterial (HTA) en pacientes del Centro Médico Santiago Cusco 2024. Se busca predecir correctamente la presencia de HTA en función de variables clínicas, antropométricas y sociodemográficas.

4.5.2. Fase 2: Comprensión de datos

Recolección de datos

La recopilación de datos se realizó aplicando una ficha de observación (Anexo número 02) en la cual se registraron los siguientes datos en concordancia con personal médico y un coasesor del Centro Médico de Santiago considerando las siguientes características:

- **Edad**: Variable cuantitativa continua que indica el número de años cumplidos por cada paciente al momento de la evaluación.

- **Sexo**: Variable categórica dicotómica que clasifica a los pacientes en masculino o femenino.
- **Peso**: Variable cuantitativa continua que expresa la masa corporal del paciente en kilogramos (kg).
- **Estatura:** Variable cuantitativa continua que representa la altura del paciente en metros (m).
- **Índice de Masa Corporal (IMC):** Variable cuantitativa continúa calculada mediante la fórmula peso (kg) dividido entre la estatura (m) al cuadrado, utilizada para clasificar el estado nutricional del paciente.
- **Perímetro Abdominal**: Variable cuantitativa continua que mide la circunferencia de la cintura en centímetros (cm), considerada un indicador de riesgo cardiovascular.
- **Categoría de Peso:** Variable categórica ordinal que clasifica a los pacientes en diferentes rangos según su IMC, tales como: peso bajo, normal, sobrepeso y obesidad.
- **Diagnóstico de Hipertensión Arterial (HTA):** Variable categórica dicotómica que indica la presencia o ausencia de diagnóstico de hipertensión arterial en el paciente.
- **Diagnóstico de Diabetes Mellitus:** Variable categórica dicotómica que identifica si el paciente ha sido diagnosticado con diabetes mellitus.
- **Presión sistólica:** Variable cuantitativa continua que mide la presión arterial máxima durante la contracción del corazón, expresada en milímetros de mercurio (mmHg).

- **Presión diastólica:** Variable cuantitativa continua que representa la presión arterial mínima cuando el corazón está en reposo entre latidos, expresada en milímetros de mercurio (mmHg).
- **Glucosa**: Variable cuantitativa continua que mide la concentración de glucosa en sangre en miligramos por decilitro (mg/dL), utilizada para evaluar el metabolismo de los carbohidratos y detectar posibles alteraciones como la diabetes mellitus.

Con estas variables fueron seleccionadas con el propósito de analizar su relación con el diagnóstico de hipertensión arterial, se construyó la base de datos en un archivo Excel, la cual es requerida para poder realizar el estudio.

Estos datos se obtuvieron de las historias clínicas de pacientes del Centro Médico Santiago Cusco, estos datos se pueden visualizar en el anexo 03.

Exploración de datos

Para la exploración de datos se siguieron los siguientes pasos:

Figura 2
Vista inicial del dataset

	SEXO	EDAD	PESO	ESTATURA	IMC	PERÍMETRO ABDOMINAL	GLUCOSA	CATEGORÍA PESO	DIABETES MELLITUS	SISTÓLICA	DIASTÓLICA	HTA
0	FEMENINO	34	77.90	1.626	29.464317	99.60	95	SOBREPESO	NO	96	64	NO
1	FEMENINO	24	71.65	1.544	30.055337	94.50	92	OBESIDAD	NO	169	92	NO
2	MASCULINO	68	92.00	1.660	33.386558	1.16	180	OBESIDAD	SÍ	150	113	SÍ
3	FEMENINO	35	66.60	1.523	28.712726	93.30	92	SOBREPESO	NO	126	83	NO
4	FEMENINO	79	61.00	1.430	29.830310	1.08	0	SOBREPESO	NO	129	82	SÍ

El objetivo de la segunda fase de la metodología CRISP-DM es el de familiarizarse con los datos para poder detectar los problemas de calidad que pueden afectar el posterior análisis, por lo que se realizó la inspección inicial del dataset, para identificar las diferentes características estructurales y las primeras inconsistencias.

Los datos que fueron analizados están conformados por variables numéricos como variables categóricas que se relacionan con parámetros biométricos y clínicos de los pacientes

En la figura 2 nos muestra que las variables que se encontraron son: edad, sexo, peso, estatura, índice de masa corporal (IMC), perímetro abdominal, nivel de glucosa, presión arterial sistólica y diastólica, así como los diagnósticos de diabetes mellitus y de hipertensión arterial (HTA).

Para las variables numéricas se tiene que los valores del IMC fueron calculados correctamente a partir de los datos de peso y estatura. Pero también fueron identificados valores atípicos y anómalos como un perímetro abdominal de 1.16 cm y un nivel de glucosa igual a 0 mg/dL, que son resultados fuera del rango definido en las bases teóricas. En este campo se tiene que tener en cuenta que los valores en cero se deben a que los médicos no consignaron estos datos en las historias clínicas, motivo por el cual fueron registrados como 0. Esto sugiere que se realice un tratamiento de los datos en etapas posteriores.

En cuanto a las variables categóricas, las clasificaciones que se asignaron en la columna de categoría peso, presencia de diabetes mellitus y HTA no evidencian irregularidades en los registros, al menos en los cinco primeros pacientes observados.

Con el análisis de exploración realizado se obtuvo una visión generalizada del dataset donde se detectaron aspectos críticos que se ha abordado en la fase de limpieza y en el preprocesamiento. Sin embargo, los datos tienen un potencial muy alto que poder utilizado en modelos predictivos que se encuentran orientados al diagnóstico de enfermedades como la hipertensión arterial.

Figura 3

Información del dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 442 entries, 0 to 441
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   SEXO                 442 non-null    object
 1   EDAD                 442 non-null    int64
 2   PESO                 442 non-null    float64
 3   ESTATURA             442 non-null    float64
 4   IMC                  442 non-null    float64
 5   PERÍMETRO ABDOMINAL  442 non-null    float64
 6   GLUCOSA              442 non-null    int64
 7   CATEGORÍA PESO       442 non-null    object
 8   DIABETES MELLITUS    442 non-null    object
 9   SISTÓLICA            442 non-null    int64
 10  DIASTÓLICA           442 non-null    int64
 11  HTA                  442 non-null    object
dtypes: float64(4), int64(4), object(4)
memory usage: 41.6+ KB
```

Los datos que tenemos en la figura 3 constan de 442 registros y 12 variables que se encuentran relacionadas demográfica y clínicamente como sexo, edad, peso, estatura, índice de masa corporal (IMC), perímetro abdominal, glucosa, presión arterial, y diagnósticos de diabetes mellitus e hipertensión arterial.

Se puede observar que no se tienen valores nulos, significa que es favorable para realizar el análisis porque no será necesario que se apliquen técnicas de imputación o eliminación de datos faltantes.

Tenemos que las variables están tipificadas de manera correcta así tenemos: las numéricas como enteros (int64) o decimales (float64), y las categóricas como objetos (object). Para esta estructura es permisible un tratamiento adecuado para el análisis exploratorio como para el desarrollo de modelos de Machine Learning en las fases siguientes. Esta es una etapa importante en la fase de comprensión de los datos porque permite que se conozcan las estructuras del dataset.

Este tipo de estructura permite un tratamiento adecuado tanto para análisis exploratorios como para el desarrollo de modelos de Machine Learning en fases posteriores.

Estadísticos descriptivos

Figura 4
Estadísticos descriptivos principales de cada variable

	SEXO	EDAD	PESO	ESTATURA	IMC	PERÍMETRO ABDOMINAL	GLUCOSA	CATEGORÍA PESO	DIABETES MELLITUS	SISTÓLICA	DIASTÓLICA	HTA
count	442	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442	442	442.000000	442.000000	442
unique	2	NaN	NaN	NaN	NaN	NaN	NaN	4	2	NaN	NaN	2
top	FEMENINO	NaN	NaN	NaN	NaN	NaN	NaN	PESO SALUDABLE	NO	NaN	NaN	SÍ
freq	252	NaN	NaN	NaN	NaN	NaN	NaN	170	383	NaN	NaN	238
mean	NaN	53.640271	70.332579	1.574975	28.362574	48.337014	67.477376	NaN	NaN	131.764706	86.280543	NaN
std	NaN	17.937015	44.753189	0.090831	17.026911	45.110418	57.391370	NaN	NaN	24.290360	15.530178	NaN
min	NaN	20.000000	44.000000	1.360000	17.553247	0.780000	0.000000	NaN	NaN	90.000000	60.000000	NaN
25%	NaN	39.000000	58.700000	1.510000	23.675165	1.020000	0.000000	NaN	NaN	114.000000	75.000000	NaN
50%	NaN	54.000000	66.500000	1.570000	26.812257	75.750000	92.000000	NaN	NaN	128.000000	83.500000	NaN
75%	NaN	68.000000	76.000000	1.630000	30.239768	88.375000	92.000000	NaN	NaN	149.750000	97.000000	NaN
max	NaN	89.000000	970.000000	1.838000	369.608291	147.000000	362.000000	NaN	NaN	180.000000	120.000000	NaN

La figura 4 muestra los estadísticos descriptivos principales para cada variable de los datos que se están manejando lo que permite que haya una primera aproximación a cómo se comportan y se distribuyen los datos:

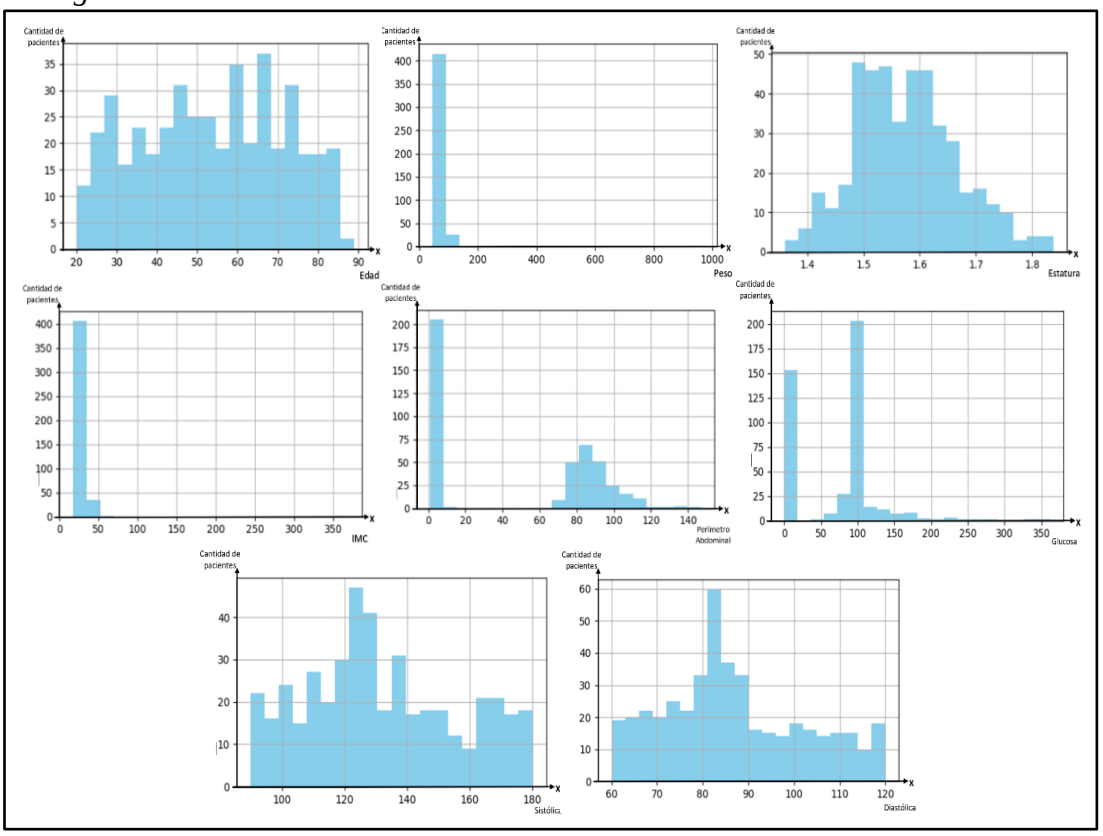
- **Sexo**: El dataset contiene dos categorías, siendo la más frecuente el sexo femenino, con una representación de 252 casos, lo cual sugiere una predominancia de mujeres en la muestra.
- **Edad**: El promedio de edad es de aproximadamente 53.64 años, con un rango que va desde los 20 hasta los 89 años, evidenciando que la muestra está compuesta principalmente por adultos y adultos mayores.
- **Peso**: El peso promedio es de 70.33 kg, con un valor mínimo de 44 kg y un máximo inusualmente alto de 970 kg, lo que indica la posible presencia de valores atípicos o errores de digitación que deberán ser tratados en la etapa de limpieza de datos.

- **Estatura**: La estatura promedio es de 1.57 m, con un mínimo de 1.36 m y un máximo de 1.83 m, valores que se encuentran dentro de rangos plausibles para adultos.
- **IMC (Índice de Masa Corporal)**: El valor promedio del IMC es de 28.36, lo cual indica que, en promedio, los pacientes se ubican dentro del rango de sobrepeso según la clasificación de la OMS. El valor máximo de 369.60 sugiere un error en el cálculo o digitación.
- **Perímetro abdominal**: El promedio es de 48.33 cm, con un valor máximo de 147 cm. Dada su importancia en la evaluación del riesgo cardiovascular, esta variable será fundamental en el análisis posterior.
- **Glucosa**: La media es de 67.48 mg/dL, con una desviación estándar alta de 57.39, indicando una variabilidad significativa en los niveles de glucosa entre los individuos. El valor máximo de 362 mg/dL puede indicar la presencia de casos de hiperglucemia severa.
- **Categoría de peso**: Se identifican 4 categorías distintas, siendo la más frecuente "Peso Saludable" con 170 casos.
- **Diabetes Mellitus**: La mayoría de los pacientes (383 casos) no presentan diagnóstico de diabetes, mientras que 59 sí lo tienen.
- **Presión arterial sistólica y diastólica**: El promedio de la presión sistólica es de 131.76 mmHg y de la diastólica 86.28 mmHg, valores que se sitúan en el límite superior de lo considerado normal, lo que sugiere la posible presencia generalizada de prehipertensión o hipertensión leve en la muestra.
- **HTA (Hipertensión Arterial)**: Hay dos categorías ("Sí" y "No"), con 238 pacientes que presentan hipertensión, lo cual representa aproximadamente el

53.8% del total de la muestra, un dato de gran relevancia para los objetivos del estudio.

Figura 5

Histogramas de las variables numéricas



La figura 5 muestra que para la variable EDAD se tiene una distribución uniforme entre los 20 y 80 años, se muestra que existe una ligera concentración entre las edades de 50 y 70 años, que indica que la muestra la componen en su mayoría los adultos y adultos mayores.

Con la variable PESO los valores que son atípicos en los extremos que van hasta 970 kg sugiere la existencia de errores al momento de ingresar los datos en la fase de limpieza. Por otro lado, la mayoría de valores se encuentran por debajo de los 150 kg.

En la variable ESTATURA, la distribución es aproximadamente normal que se concentra entre 1.5 mt. Y 1.7 mt mostrando consistencia con los valores de la a adulta.

En el IMC (Índice de Masa Corporal) la distribución es sesgada hacia la derecha, con valores que son atípicos y que superan los 300, que demuestra que hay necesidad de depuración de datos. La mayoría se encuentra entre 20 y 40, indicando prevalencia de sobrepeso y obesidad.

En el PERÍMETRO ABDOMINAL la distribución mostrada es unimodal con una ligera asimetría positiva. No obstante, se observa una concentración inusualmente alta de datos en el rango de 0 a 20, lo que indica que existen valores dados en metros (0.9, 1.0, etc.) que no han sido convertidos a centímetros lo que está afectando la correcta interpretación de la variable.

En cuanto a la GLUCOSA se muestra la presencia de dos picos en la distribución, indica una posible separación entre las personas que están sanas y con las que muestran hiperglucemia o diabetes. Demas que se tiene la presencia de valores que son necesarios validar.

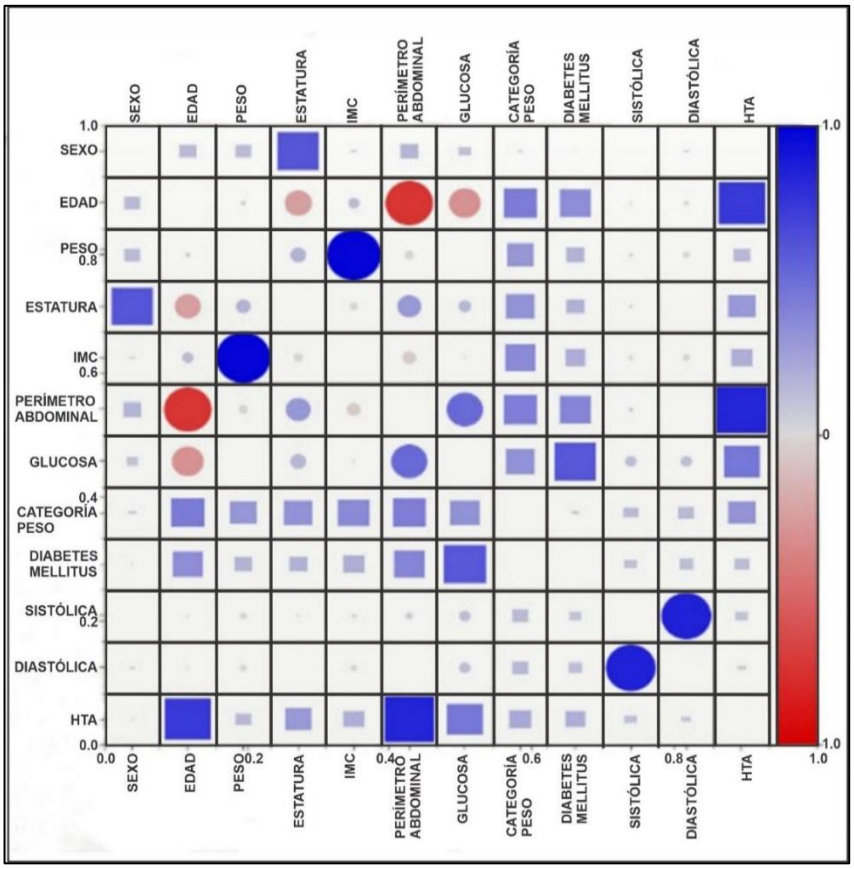
Para PRESIÓN ARTERIAL SISTÓLICA y DIASTÓLICA, en ambos casos la distribución muestra que hay rangos amplios, con una tendencia que es leve hacia la presión sistólica (>140 mmHg). Esto coincide con los diagnósticos clínicos observados en el conjunto de datos.

Los histogramas presentados hacen posible que se identifiquen tendencias y valores que son atípicos para una posterior limpieza y modelado.

Matriz de correlación

Figura 6

Matriz de correlación (Enfoque en Hipertensión Arterial)



En la Figura 6, la matriz de correlación permite identificar la fuerza y dirección de las relaciones entre las variables clínicas y antropométricas analizadas. Se observa una fuerte asociación entre la presión arterial sistólica y diastólica, lo cual es esperable, ya que ambas forman parte del diagnóstico de hipertensión.

Asimismo, el índice de masa corporal (IMC) presenta una alta correlación con el peso, y en menor medida con el perímetro abdominal, lo cual refleja la coherencia entre las medidas antropométricas del conjunto de datos.

También se observa que la relación entre glucosa y perímetro abdominal es baja, por lo que no se evidencia una asociación clara entre estas dos variables en esta muestra.

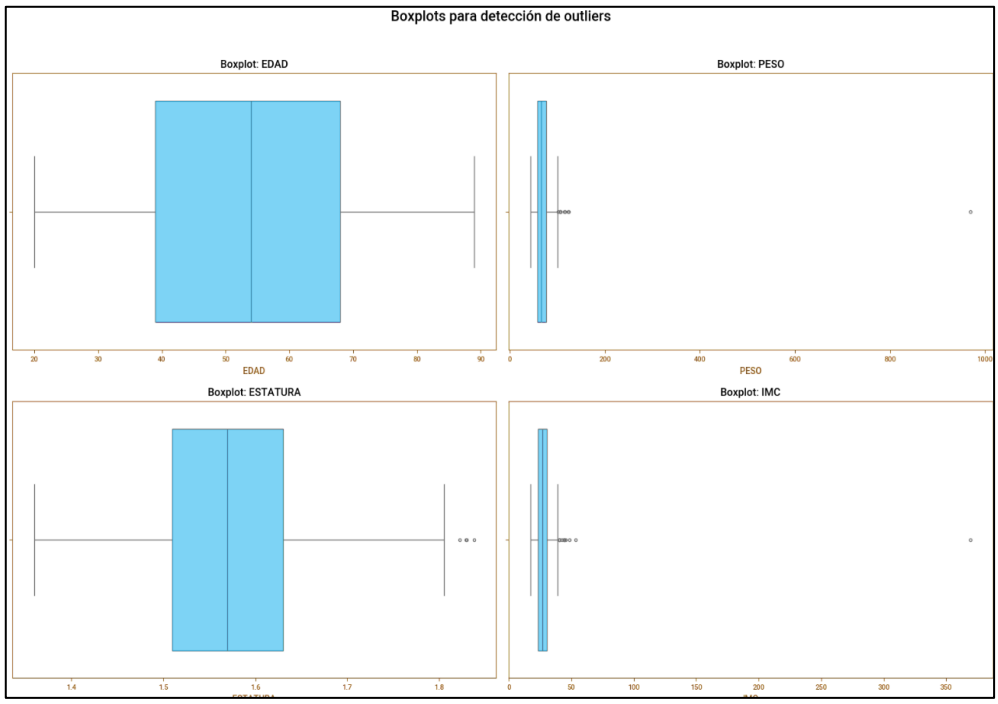
Por otro lado, la edad no muestra correlaciones destacadas con las variables clínicas principales, lo que indica que su efecto puede estar condicionado por otros factores presentes en la población. La variable HTA (hipertensión arterial) presenta relaciones visibles con la presión sistólica, la presión diastólica, el IMC y el perímetro abdominal, lo que refuerza su importancia clínica al momento de identificar factores de riesgo asociados.

En conjunto, esta visualización permite confirmar que las variables de tipo antropométrico y de presión arterial tienen un mayor grado de asociación entre sí y con la presencia de hipertensión, por lo que resultan relevantes para la construcción de modelos predictivos en las siguientes fases.

Outliers (valores atípicos)

Figura 7

Boxplot para detectar outliers edad, peso, estatura, IMC



Con el fin de que se pueda asegurar la presencia de calidad y fiabilidad de los datos que se usaron en la construcción del dataset y del modelado de Machine Learning, se realizó la detección de outliers (valores atípicos) en las variables

numéricas clave: EDAD, PESO, ESTATURA e IMC. El análisis se realizó visualizando diagramas de caja (boxplots) obteniendo los siguientes hallazgos en la figura 7:

- **Boxplot: EDAD**

El diagrama de caja para esta variable presento una distribución de datos que son notablemente simétricos, teniendo a la mayoría de pacientes dentro del rango de edades similares. Tenemos que la mediana se a situado en el centro de la caja aproximadamente que reafirma la simetría. Se destaca que no se tiene identificado puntos individuales que se encuentran fuera de la caja conocidos como bigotes, del boxplot que indica que hay ausencia de outliers significativos confirmando que tiene una alta calidad

- **Boxplot: PESO**

Este boxplot muestra una distribución asimétrica con tendencia a la derecha, donde la mayoría de datos están agrupados, pero se observa un número considerable de puntos individuales fuera del bigote superior, lo que representa outliers extremadamente altos, dentro de estos se incluyen un caso cercano a 1000 kg que sugiere que se presenta un error en el registro de la medición en los datos. La existencia de estos outliers extremos y probablemente erróneos representan una preocupación significativa para la calidad de la variable peso, lo cual requiere ser corregida en las etapas posteriores.

- **Boxplot: ESTATURA**

El diagrama de caja para la estatura exhibió una caja central relativamente compacta, indicando una concentración de la mayoría de los datos en un rango específico de estaturas. La mediana se encontró en el centro de la caja. Se identificaron varios puntos individuales que se extendían más allá del bigote superior, representando

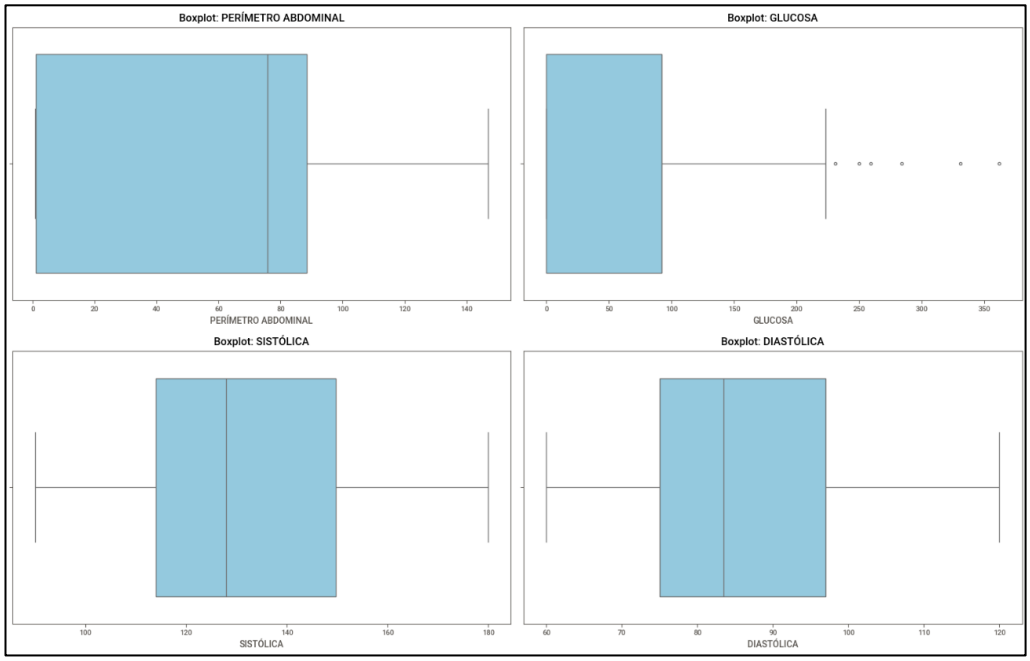
outliers de estatura alta. Aunque estos no son muchos, su presencia sugiere la necesidad de una revisión para determinar si corresponden a errores de registro o a valores reales, pero inusualmente altos en la población.

- **Boxplot: IMC (Índice de Masa Corporal)**

El boxplot del IMC muestra directamente problemas que se identificaron dentro de la variable peso, observándose que se tiene una caja central extremadamente estrecha y compacta en la mayoría de valores de IMC que están agrupados dentro de un rango limitado, en grados altos se identifican puntos individuales que se extienden significativamente a la derecha que supera el bigote superior. Estos representan outliers de IMC extremadamente altos, con un valor más extremo cercano a 350, lo cual es biológicamente imposible y se relaciona directamente con errores previos en la variable PESO. Esto confirma que las inconsistencias en las mediciones de peso impactan en la fiabilidad del IMC, por lo que es indispensable corregirlas antes de realizar el modelado.

Figura 8

Boxplot para detectar outliers en: perímetro abdominal, glucosa, sistólica y diastólica



- **Boxplot: PERÍMETRO ABDOMINAL**

Se observa una distribución fuertemente sesgada hacia la izquierda, con una gran concentración de datos en los valores bajos (alrededor de 0 a 20 cm), lo que evidencia que persisten registros en metros no convertidos a centímetros. La mediana se sitúa por debajo de los valores clínicamente esperados y la caja cubre un rango que no representa adecuadamente los valores típicos de perímetro abdominal en adultos. Este comportamiento confirma la necesidad urgente de corregir la escala de medida antes de continuar con el análisis, ya que afecta la confiabilidad de esta variable clave.

- **Boxplot: GLUCOSA**

La distribución presenta una caja alargada hacia la derecha, con varios puntos individuales que superan los 250 mg/dL, indicando la presencia de valores atípicos elevados. Estos valores, si bien pueden corresponder a casos clínicos reales de hiperglucemia, deben ser contrastados con otros antecedentes para confirmar su veracidad. La mediana se sitúa dentro de un rango clínicamente esperado, lo que

sugiere que, a pesar de los outliers, la mayoría de los datos se encuentra en un rango plausible.

- **Boxplot: PRESIÓN SISTÓLICA**

La variable muestra una distribución relativamente simétrica, sin presencia evidente de outliers. La mediana se ubica cerca de 130 mmHg, dentro del rango limítrofe superior normal, lo que concuerda con los valores promedio previamente descritos. Este comportamiento indica que la variable es confiable y no requiere ajustes adicionales, siendo útil para la evaluación del riesgo hipertensivo.

- **Boxplot: PRESIÓN DIASTÓLICA**

La distribución también es simétrica y no presenta valores atípicos extremos. La mediana se encuentra cercana a 85 mmHg, lo cual es coherente con los valores clínicos normales y con los hallazgos observados en otras representaciones gráficas. Esta variable no presenta distorsiones ni errores de registro, por lo que es adecuada para su análisis sin necesidad de transformaciones adicionales.

4.5.3. Fase 3: Preparación de datos

Limpieza de datos

En esta etapa se aplicaron procesos básicos para mejorar la calidad de los datos y asegurar que el análisis posterior se realice sobre un conjunto limpio, coherente y sin inconsistencias.

Eliminación de duplicados

Se utilizó un comando para verificar y eliminar posibles registros duplicados en el dataset, lo cual es importante para evitar sesgos en el análisis. Para la investigación se confirmó que no existían filas duplicadas.

Figura 9
Eliminación de registros duplicados en el dataset

```
# === Eliminamos duplicados ==
df.drop_duplicates(inplace=True)
```

Se verificó que no existen valores duplicados.

Figura 10
Verificación de valores nulos y valores de glucosa

```
Valores nulos por columna:
 SEXO                    0
EDAD                     0
PESO                     0
ESTATURA                 0
IMC                      0
PERÍMETRO ABDOMINAL      0
GLUCOSA                  0
CATEGORÍA PESO           0
DIABETES MELLITUS        0
SISTÓLICA                0
DIASTÓLICA               0
HTA                      0
dtype: int64
Cantidad de registros con GLUCOSA = 0: 153
```

En la figura 10 se tiene que los resultados verificados de la presencia de valores nulos y de registros que son anómalos en la fase de preparación de los datos, etapa que es fundamental para que se aseguren la calidad y la confiabilidad de los análisis posteriores.

Verificación de valores nulos

Ninguna de las columnas muestra la existencia de valores nulos, que indica que el conjunto de datos se encuentra completo y no es necesaria la imputación de datos que sean faltantes. Estos resultados son importantes ya que al tener datos nulos puede afectar de manera negativa en la precisión de los modelos estadísticos o de Machine Learning.

Validación de valores anómalos en glucosa

Se verificaron de manera específica los registros que contienen valores iguales a 0 en la variable glucosa, los cuales se consideran como atípicos o valores no fisiológicos teniendo en consideración que en condiciones normales los valores en una persona no pueden ser igual a 0. El análisis evidenció que existen 153 registros con este valor atípico, lo que representa una proporción considerable del total de observaciones. Ante este hallazgo se realizan las siguientes acciones:

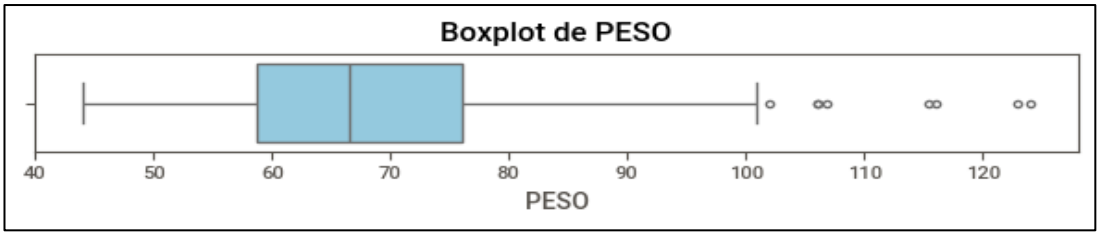
- Se corrige manualmente valores erróneos que se detectaron revisando las historias clínicas.
- Se imputa los datos de la variable GLUCOSA. Los registros con valores igual a cero en la variable glucosa fueron tratados mediante una imputación multivariada, considerando variables clínicas relacionadas como edad, peso, IMC, perímetro abdominal y presión arterial. Este procedimiento permitió estimar valores coherentes, manteniendo la consistencia interna del conjunto de datos y así evitar pérdidas en la data.
- Se corrige valores erróneos que se detectaron revisando las historias clínicas en variables de: PESO, IMC y PERÍMETRO ABDOMINAL.

Transformación y preprocesamiento

Las figuras presentadas a continuación reflejan el estado de las variables numéricas luego de las correcciones en la fase anterior, mostrando una distribución más coherente y adecuada para el análisis exploratorio y el modelado posterior.

Figura 11

Boxplot de peso después de la corrección

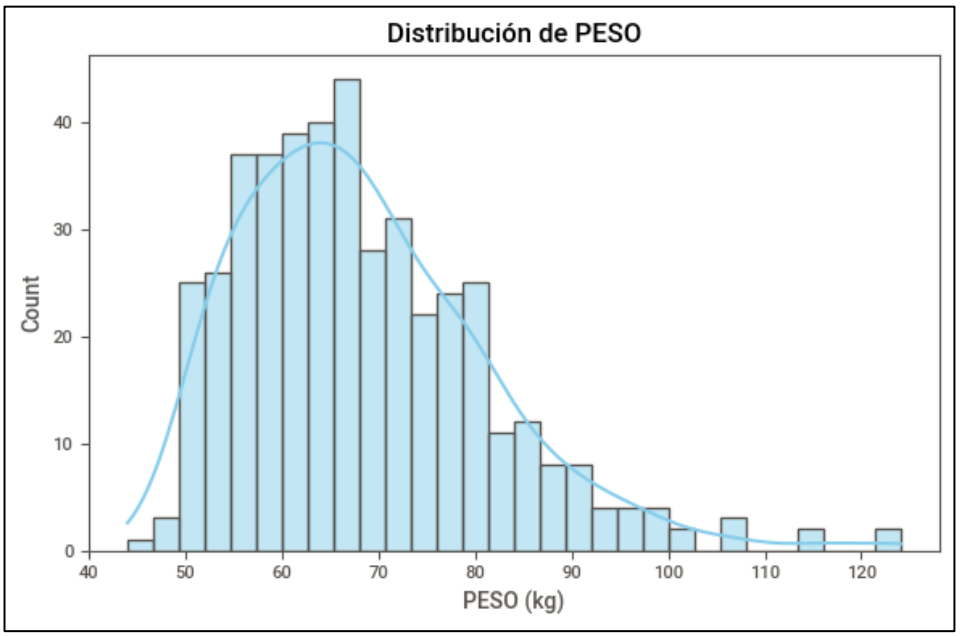


La figura 11 presenta un valor en la mediana de 68 a 70 kg indicando que la mitad de las personas cuentan con un peso menor al valor presentado mientras que la otra mitad es mayor, el rango intercuartílico (IOR) se va extendiendo de forma aproximada desde los 58 kg hasta 76 kg, concentrando al 50% central de los datos. En el Límites de los "bigotes": tenemos Inferior: Cerca de 45 kg. Superior: Hasta aproximadamente 98-100 kg. En Outliers (valores atípicos): A pesar de la corrección, persisten algunos valores atípicos hacia el extremo superior (mayores a 100 kg), lo cual puede deberse a individuos con obesidad severa o condiciones clínicas especiales.

Con los resultados obtenidos se nota que la variable peso se muestra asimétrica hacia a la derecha conocido como un sesgo positivo como se demuestra con la presencia de valores extremos que se encuentran en la parte superior, dentro de los valores que se han identificado no son necesariamente erróneos y teniendo en cuenta la corrección realizada de manera manual de las historias clínicas ya representando casos clínicos reales y además relevantes, como personas con obesidad. Estas observaciones permiten que se validen la mayoría de los datos de peso se encuentran dentro de un rango que se tenía esperado, lo que es importante para garantizar la validez de los modelos predictivos, especialmente si se analiza el riesgo de hipertensión arterial (HTA).

Figura 12

Histograma de distribución de peso



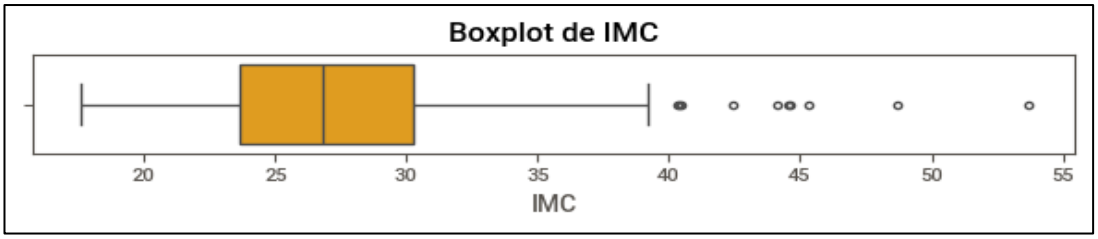
La figura 12 muestra que la distribución del peso corporal en la muestra presenta una asimetría positiva, con la mayoría de los pacientes concentrados entre 55 y 75 kg y un valor modal alrededor de 65-68 kg, pero con una cola extendida hacia la derecha que incluye casos de hasta 125 kg.

Estos valores extremos, validados mediante historias clínicas, representan pacientes con obesidad y son clínicamente relevantes por su asociación con un mayor riesgo de hipertensión arterial (HTA).

La presencia de una proporción considerable de individuos con sobrepeso u obesidad sugiere una posible contribución significativa a la prevalencia de HTA observada en el estudio, y aunque los casos con pesos muy elevados son escasos, su impacto en la salud cardiovascular puede ser desproporcionadamente alto.

Figura 13

Boxplot de IMC después de la corrección



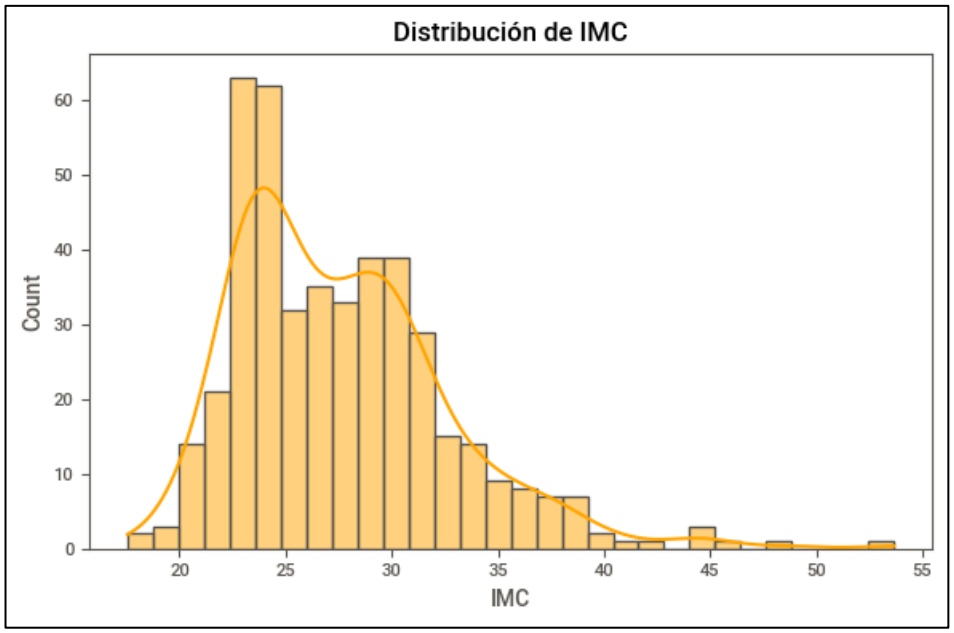
En la figura 13 el boxplot del Índice de Masa Corporal (IMC) muestra que la mayoría de los pacientes tienen un IMC entre 24 y 30, rango que abarca desde sobrepeso hasta obesidad moderada, con una mediana cercana a 27.

Los bigotes indican que los valores habituales del IMC oscilan entre cerca de19 y hasta 38-39, lo cual es coherente con la distribución general, aunque el bigote inferior parece partir desde 19. Mientras que varios outliers por encima de este rango, que alcanzan hasta 54, revelan la presencia de pacientes con obesidad severa.

Estos valores atípicos, aunque poco frecuentes, son clínicamente significativos y deben revisarse para confirmar su validez, ya que pueden tener implicaciones importantes en la evaluación del riesgo de enfermedades cardiovasculares, especialmente hipertensión arterial.

Figura 14

Histograma de distribución de IMC

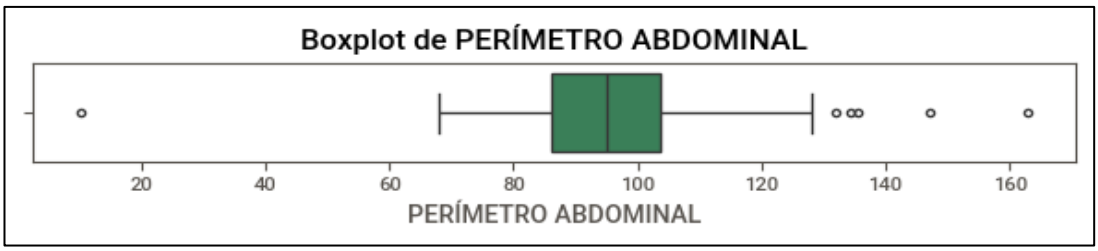


En la figura 14 el histograma del Índice de Masa Corporal (IMC) muestra una distribución asimétrica hacia la derecha con dos picos: uno principal alrededor de 24-25 (peso saludable) y otro secundario entre de 28-31 (sobrepeso), lo que sugiere la presencia de dos grupos predominantes en la población. La mayoría de los pacientes tienen un IMC entre 20 y 32, aunque también se identifican valores extremos cercanos a 54, correspondientes a casos de obesidad severa.

Esta distribución indica que una parte significativa de la muestra presenta sobrepeso u obesidad, factores de riesgo clave para la hipertensión arterial, y resalta la importancia de considerar estos patrones al analizar el perfil de salud cardiovascular de la población.

Figura 15

Boxplot de Perímetro Abdominal después de la corrección

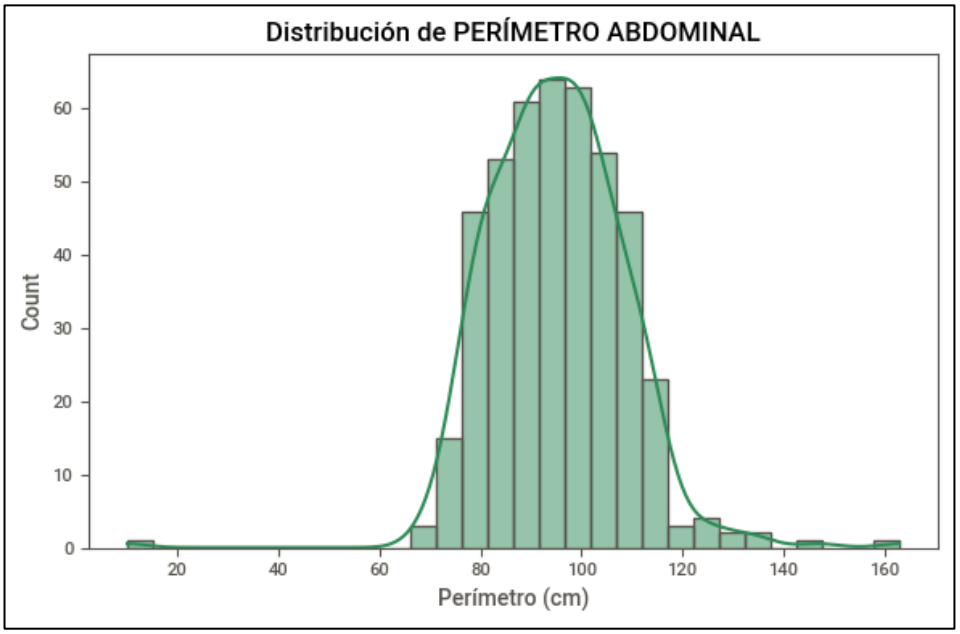


La figura 15 muestra el boxplot del perímetro abdominal revela que la mayoría de los pacientes tienen medidas entre 80 y 105 cm, con una mediana cercana a los 100 cm, lo que indica una alta concentración en ese rango. Sin embargo, se identifican varios outliers: uno extremo inferior con un valor irreal de aproximadamente 10 cm, que probablemente se deba a un error de registro, y varios valores superiores entre 135 y más de 160 cm, que, aunque posibles en casos de obesidad severa, requieren revisión. Además, se observa que los bigotes se extienden hasta aproximadamente 65 y 125 cm.

Esta distribución resalta la importancia de validar y limpiar los datos antes del análisis, ya que la presencia de valores atípicos podría distorsionar las conclusiones sobre la relación entre el perímetro abdominal y el riesgo de hipertensión arterial.

Figura 16

Histograma de distribución de Perimetro Abdominal

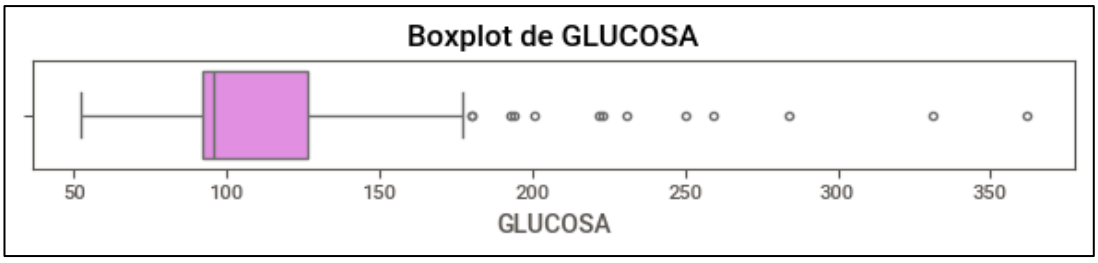


En la figura 16 el histograma del perímetro abdominal muestra una distribución aproximadamente normal con ligera asimetría hacia la derecha, concentrando la mayoría de los valores entre 95 y 105 cm, donde más de 60 pacientes se ubican en ese rango, lo cual es consistente con un riesgo moderado a elevado de hipertensión arterial.

Sin embargo, se identifican valores atípicos significativos: un grupo reducido con medidas superiores a 120 cm, compatibles con obesidad abdominal severa, y un valor extremadamente bajo cercano a 10 cm, que es biológicamente inviable y claramente un error de registro.

Figura 17

Boxplot de Glucosa después de la imputación = 0



En la figura 17 el boxplot de la variable glucosa, generado tras la imputación de los valores igual a cero (GLUCOSA = 0), muestra una distribución más coherente

y clínicamente plausible, eliminando valores biológicamente imposibles como el 0 mg/dL.

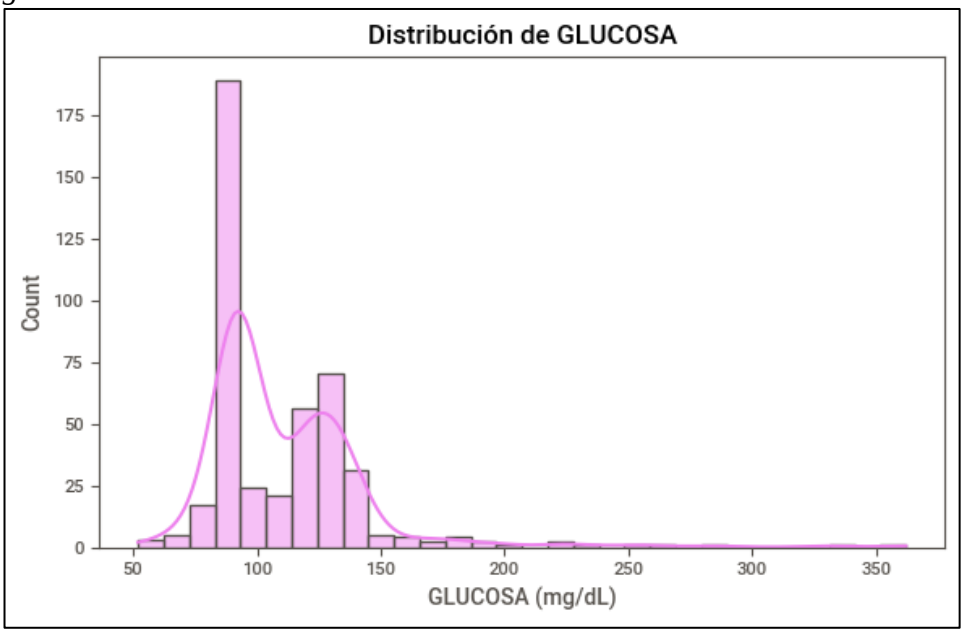
La mayoría de los datos se concentran entre 100 y 125 mg/dL, con una mediana cercana a 100 mg/dL, indicando niveles dentro de rangos normales o limítrofes. Los bigotes se extienden desde aproximadamente 50 hasta 170-175 mg/dL, y no se observan valores extremos en el límite inferior, lo cual confirma el éxito del proceso de imputación.

Sin embargo, se identifican múltiples outliers por encima de los 180 mg/dL, llegando hasta 360 mg/dL, que, aunque estadísticamente atípicos, son biológicamente válidos y clínicamente relevantes, aludiendo a posibles casos de hiperglucemia, prediabetes o diabetes no controlada.

En conjunto, esta imputación ha mejorado la calidad del dataset y permite un análisis más fiable, sin eliminar valores altos que podrían ser indicadores importantes en el diagnóstico de hipertensión arterial.

Figura 18

Histograma de distribución de Glucosa



En la figura 18 el histograma de la variable glucosa, posterior a la imputación de los valores igual a cero (GLUCOSA = 0), revela una distribución multimodal con

al menos dos picos claros: uno principal entre 90 y 100 mg/dL, que representa a la mayoría de los individuos con niveles normales de glucosa, y otro secundario entre 125 y 135 mg/dL, asociado con posibles casos de prediabetes o diabetes incipiente.

La distribución es asimétrica positiva, con una cola extendida hacia la derecha que alcanza hasta 360 mg/dL, lo cual indica la presencia de pacientes con hiperglucemia significativa. El rango general de los valores va de aproximadamente 50 a 360 mg/dL, y la barra más alta supera los 175 casos, confirmando la alta concentración en el rango normal. La ausencia de un pico en 0 mg/dL sugiere que los valores imputados fueron reasignados adecuadamente a intervalos clínicamente plausibles, mejorando la representación real de la población estudiada. Este resultado respalda la validez de la imputación y muestra una distribución que, si bien presenta outliers elevados, es útil para el análisis clínico y el modelado predictivo

Codificación de variables categóricas

Figura 19

Codificación de las variables categóricas

	SEXO	EDAD	PESO	ESTATURA	IMC	PERÍMETRO ABDOMINAL	GLUCOSA	CATEGORÍA PESO	DIABETES MELLITUS	SISTÓLICA	DIASTÓLICA	HTA
0	0	34.0	77.90	1.626	29.464317	99.6	95.00000	3	0	96.0	64.0	0
1	0	24.0	71.65	1.544	30.055337	94.5	92.00000	1	0	169.0	92.0	0
2	1	68.0	92.00	1.660	33.386558	116.0	180.00000	1	1	150.0	113.0	1
3	0	35.0	66.60	1.523	28.712726	93.3	92.00000	3	0	126.0	83.0	0
4	0	79.0	61.00	1.430	29.830310	108.0	137.10089	3	0	129.0	82.0	1

La figura 19 presenta las primeras cinco filas de un conjunto de datos tabular orientado al diagnóstico de hipertensión arterial, donde se observan múltiples variables clínicas y demográficas codificadas y preprocesadas. Las variables incluyen datos categóricos como SEXO (0: femenino, 1: masculino), CATEGORÍA PESO (0: bajo peso, 1: obesidad, 2: peso saludable, 3: sobrepeso), DIABETES MELLITUS y HTA (variable objetivo, 0: no hipertensión, 1: sí hipertensión), así como datos numéricos continuos como EDAD, PESO, ESTATURA, IMC, PERÍMETRO ABDOMINAL,

GLUCOSA, SISTÓLICA y DIASTÓLICA. La diversidad de valores en estas primeras filas refleja variabilidad real entre pacientes, destacando casos con posibles cuadros hipertensivos como el paciente con presión 150/113 mmHg y HTA = 1. Además, la ausencia de ceros en la columna GLUCOSA confirma que ya se ha realizado la imputación de valores erróneos. Según se aprecia en la muestra observada, el conjunto parece correctamente codificado y preprocesado, lo que permite su uso directo en tareas de clasificación supervisada.

Figura 20

Codificación de variables

```
Codificación para 'SEXO':
0: FEMENINO
1: MASCULINO

Codificación para 'CATEGORÍA PESO':
0: BAJO PESO
1: OBESIDAD
2: PESO SALUDABLE
3: SOBREPESO

Codificación para 'DIABETES MELLITUS':
0: NO
1: SÍ

Codificación para 'HTA':
0: NO
1: SÍ
```

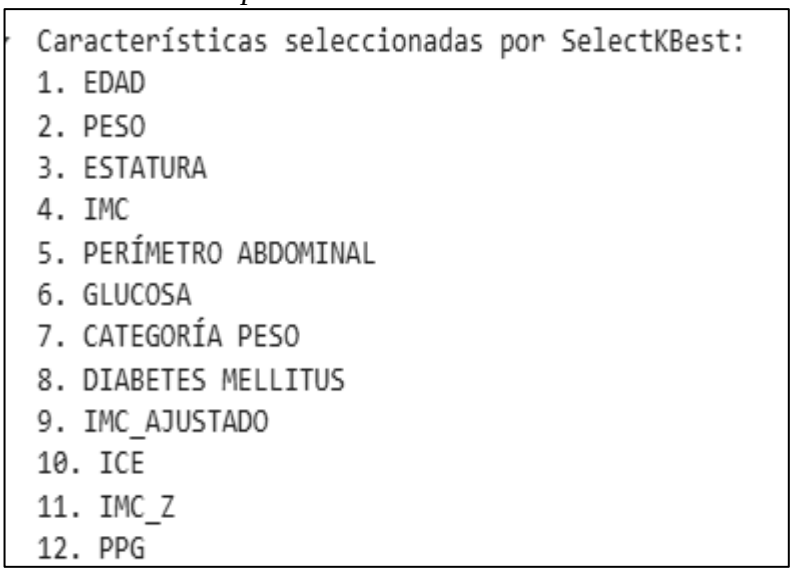
La figura 20 muestra un fragmento de texto que describe el esquema de codificación por etiquetas (Label Encoding) aplicado a cuatro variables categóricas en un conjunto de datos orientado al diagnóstico de hipertensión arterial, con el objetivo de prepararlo para algoritmos de Machine Learning. Las variables codificadas son: SEXO (0: femenino, 1: masculino), CATEGORÍA PESO (0: bajo peso, 1: obesidad, 2: peso saludable, 3: sobrepeso), DIABETES MELLITUS (0: no, 1: sí) y HTA (0: no, 1: sí), siendo esta última la variable objetivo. Si bien las variables binarias como SEXO, DIABETES MELLITUS y HTA han sido codificadas de manera estándar, la

variable CATEGORÍA PESO, con múltiples categorías, ha sido transformada mediante valores numéricos enteros sin un orden cuantitativo explícito. En conjunto, esta codificación permite que los algoritmos de aprendizaje automático interpreten las variables categóricas de manera numérica, optimizando su procesamiento durante el modelado.

Selección de atributos con SelectKBest

Figura 21

Características seleccionadas por SelectKBest



En la figura 21 tenemos el método SelectKBest ha identificado doce variables como las más relevantes para predecir la hipertensión arterial (HTA), incluyendo factores demográficos y clínicos clave como la edad, peso, estatura, IMC y sus variaciones (IMC ajustado e IMC_Z), perímetro abdominal, glucosa, presencia de diabetes mellitus y variables derivadas como categoría de peso, ICE y PPG. Estas variables están correctamente seleccionadas, ya que representan factores de riesgo bien establecidos en la literatura médica. Además, la presencia simultánea de distintos indicadores del IMC aporta información valiosa para el modelo.

La selección de estas variables permite reducir la dimensionalidad del conjunto de datos, facilitando la creación de modelos más simples, interpretables y

potencialmente con mejor desempeño al minimizar ruido y sobreajuste. Aunque SelectKBest se basa en análisis univariantes y no considera interacciones entre variables, constituye un paso fundamental y efectivo para enfocar el modelado en las características más informativas.

Estandarización de variables seleccionadas

Figura 22

Estandarización de variables

Variables estandarizadas:

	EDAD	PESO	ESTATURA	IMC
count	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02
mean	-1.687941e-16	3.657205e-16	8.881784e-16	2.250588e-16
std	1.001133e+00	1.001133e+00	1.001133e+00	1.001133e+00
min	-1.877592e+00	-1.897380e+00	-2.369433e+00	-1.995433e+00
25%	-8.171293e-01	-7.522902e-01	-7.161488e-01	-7.807199e-01
50%	2.007782e-02	-1.446915e-01	-5.483517e-02	-1.582569e-01
75%	8.014711e-01	5.953325e-01	6.064785e-01	5.218313e-01
max	1.973561e+00	4.342191e+00	2.899032e+00	5.168949e+00

	PERÍMETRO ABDOMINAL	GLUCOSA	CATEGORÍA PESO	DIABETES MELLITUS
count	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02
mean	-1.808508e-17	1.165483e-16	-3.727536e-16	5.224579e-17
std	1.001133e+00	1.001133e+00	1.001133e+00	1.001133e+00
min	-1.128239e+00	-1.813086e+00	-2.610064e+00	-3.924882e-01
25%	-1.122920e+00	-6.883136e-01	-1.345099e+00	-3.924882e-01
50%	5.737535e-01	-6.039557e-01	-8.013356e-02	-3.924882e-01
75%	8.352575e-01	7.801553e-01	1.184832e+00	-3.924882e-01
max	2.111752e+00	6.903902e+00	1.184832e+00	2.547847e+00

	IMC_AJUSTADO	ICE	IMC_Z	PPG
count	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02
mean	-1.607563e-16	-5.425524e-17	8.037814e-18	-6.329778e-17
std	1.001133e+00	1.001133e+00	1.001133e+00	1.001133e+00
min	-1.778841e+00	-1.122318e+00	-1.995433e+00	-1.696011e+00
25%	-7.446537e-01	-1.117694e+00	-7.807199e-01	-6.745241e-01
50%	-1.950741e-01	5.322854e-01	-1.582569e-01	-1.423581e-01
75%	5.423643e-01	8.381962e-01	5.218313e-01	4.324372e-01
max	4.962461e+00	2.366546e+00	5.168949e+00	8.026682e+00

La figura 22 muestra las estadísticas descriptivas de las variables luego de aplicar la estandarización tipo z, proceso que consiste en restar la media y dividir por la desviación estándar de cada variable para que todas queden en la misma escala. Esto se realizó en la Fase 3: Preparación de datos, con el objetivo de asegurar la comparabilidad entre variables numéricas y facilitar el funcionamiento de algoritmos de aprendizaje automático sensibles a la magnitud, como SVM, K-NN, redes neuronales o Regresión Logística con regularización. La tabla evidencia que la media de cada variable es cercana a cero (del orden de 10^{-16}), confirmando que el centrado

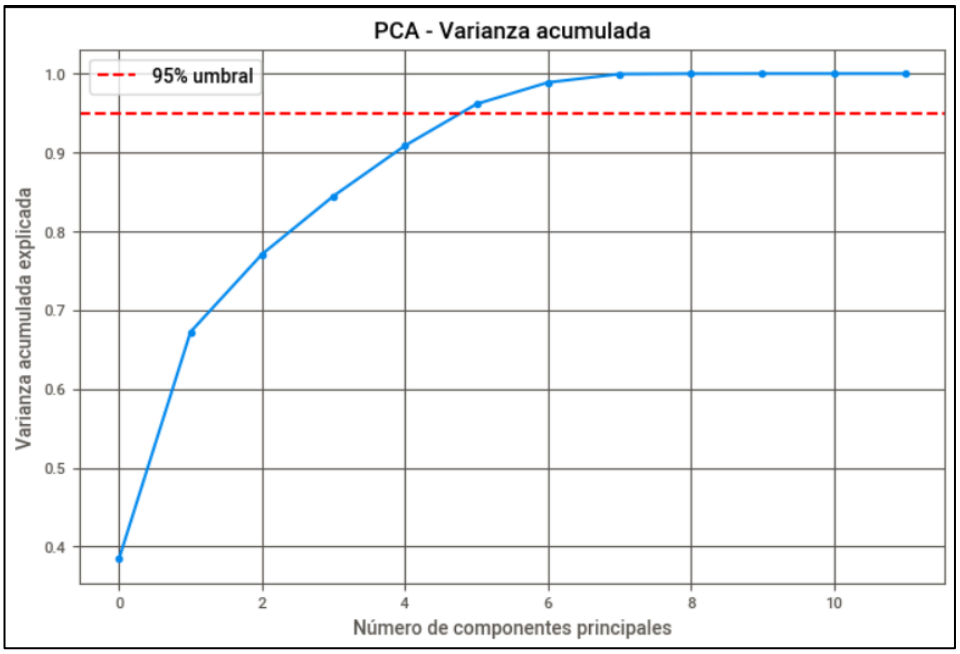
fue correcto, y que la desviación estándar es aproximadamente uno (alrededor de 1.001), lo que indica que la varianza fue ajustada adecuadamente al valor unitario, tal como se espera en un proceso de estandarización z-score. Además, los valores mínimos y máximos transformados en unidades z reflejan que los datos extremos originales, como pesos o niveles de glucosa elevados, se mantienen dentro de un rango estadístico coherente sin alterar la escala global. Este tratamiento tiene importantes implicaciones para el análisis de hipertensión, pues garantiza que variables fundamentales para su predicción (IMC, perímetro abdominal, glucosa, presión arterial, entre otras) sean ponderadas equitativamente en modelos multivariados, evitando que alguna domine por su escala original. También mejora la convergencia y estabilidad numérica en algoritmos basados en gradiente, acelerando el entrenamiento y reduciendo problemas numéricos. Finalmente, facilita la interpretación de los coeficientes en modelos lineales, pues estos representan el cambio esperado en la variable objetivo por cada desviación estándar del predictor, lo que permite comparar la influencia relativa de cada variable en la aparición de hipertensión arterial. En síntesis, la tabla confirma que la estandarización fue exitosa y que el conjunto de datos está adecuadamente preparado para la siguiente fase de modelado con algoritmos que requieren variables escaladas.

Asimismo, se observa que el número de observaciones (count = 442) es constante en todas las variables estandarizadas, lo que indica que no existen valores perdidos en esta etapa del procesamiento de datos.

Análisis de Componentes Principales (PCA)

Figura 23

Análisis de Componentes Principales (PCA)



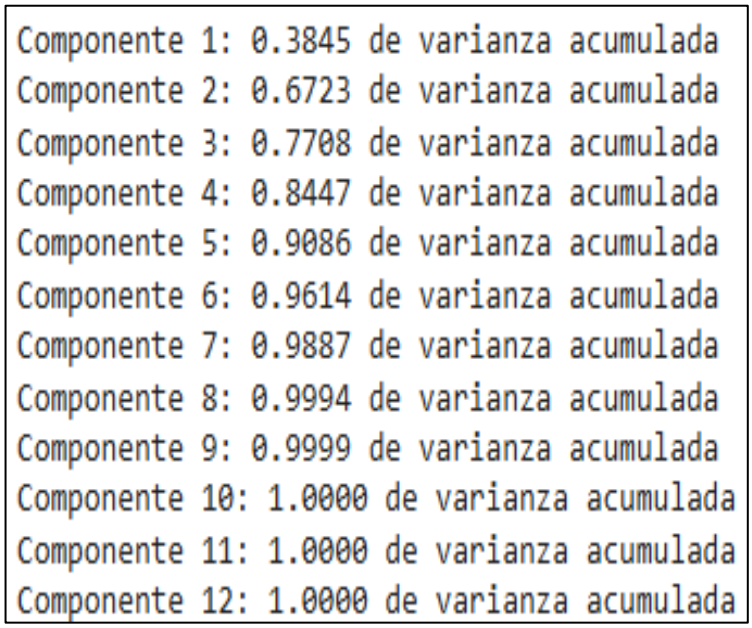
La figura 23 de varianza acumulada del Análisis de Componentes Principales (PCA) muestra cómo la varianza explicada total aumenta al incorporar más componentes principales, con el eje X representando el número de componentes y el eje Y el porcentaje acumulado de varianza explicada. La línea azul indica el incremento progresivo de varianza explicada, mientras que la línea roja discontinua señala el umbral común del 95%, criterio usado para seleccionar componentes significativos. En este caso, las primeras seis componentes superan ese umbral, lo que significa que concentran casi toda la información relevante del conjunto original de variables, permitiendo una reducción efectiva de la dimensionalidad de más de diez variables a solo seis componentes sin pérdida significativa de información.

La curva presenta un "codo" marcado entre los componentes tres y cinco, evidenciando que estos primeros componentes aportan la mayor parte de la varianza, mientras que los componentes posteriores añaden poco valor adicional. Esto implica que, para modelos predictivos, como los dirigidos al diagnóstico de hipertensión

arterial, usar estas seis componentes principales como nuevas variables combinadas puede simplificar el modelo, reducir ruido y complejidad computacional, y mejorar su rendimiento al incorporar la información esencial de variables originales como IMC, peso y glucosa.

Figura 24

Componentes de varianza acumulada



La figura 24 muestra la varianza acumulada explicada por cada uno de los 12 componentes principales generados mediante Análisis de Componentes Principales (PCA), revelando que el primer componente explica el 38.45% de la varianza total, y los dos primeros alcanzan el 67.23%, lo que indica que una gran parte de la información del conjunto de datos puede ser representada con pocos componentes; con el componente 5 se llega al 90.86%, con el componente 6 se llega a 96.14%, que supera el umbral estándar del 95% y con el componente 8 al 99.94%, demostrando que es posible reducir las 12 variables originales a entre 6 y 8 componentes sin pérdida significativa de información, lo cual facilita la simplificación del modelo y mejora su eficiencia computacional manteniendo la mayoría de la variabilidad de los datos.

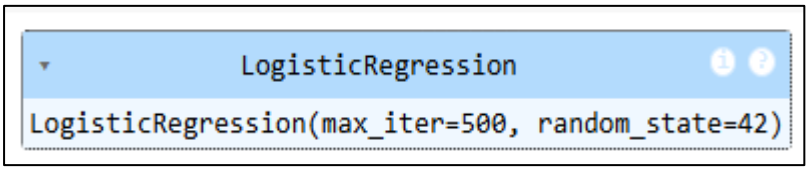
4.5.4. Fase 4: Modelado

Modelado de Regresión Logística

A continuación, se procederá a realizar el modelado mediante Regresión Logística, empleando la librería scikit-learn en Python, con el fin de construir un clasificador capaz de predecir la presencia de hipertensión arterial a partir de las variables seleccionadas.

Figura 25

Modelado en algoritmo de Regresión Logística



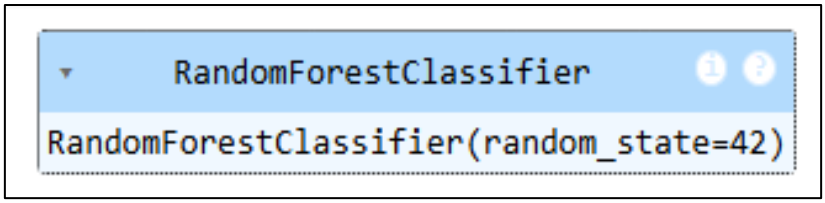
La figura 25 muestra la configuración de un modelo de Regresión Logística (LogisticRegression) en Python mediante la biblioteca scikit-learn, el cual ha sido instanciado con dos parámetros: max_iter=500, que indica que el algoritmo puede realizar hasta 500 iteraciones para converger (lo cual es útil en conjuntos de datos complejos donde la convergencia puede requerir más pasos), y random_state=42, que establece una semilla fija para garantizar la reproducibilidad de los resultados, es decir, que se obtengan los mismos resultados cada vez que se entrene el modelo con los mismos datos. Esta configuración es común en tareas de clasificación binaria, como en el diagnóstico de enfermedades como la hipertensión arterial.

Modelado de Random Forest

Seguidamente se llevará a cabo el modelado mediante Random Forest, utilizando la librería scikit-learn en Python. Este algoritmo, basado en múltiples árboles de decisión, permite mejorar la precisión y la capacidad de generalización del modelo para la detección de hipertensión arterial.

Figura 26

Modelado en algoritmo de Random Forest



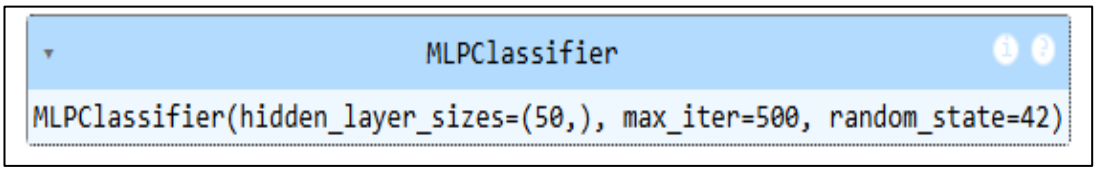
La figura 26 muestra la configuración de un modelo de clasificación llamado RandomForestClassifier del paquete scikit-learn, utilizado comúnmente en tareas de aprendizaje supervisado. En este caso, el clasificador ha sido instanciado con el parámetro random_state=42, lo cual garantiza la reproducibilidad de los resultados al fijar la semilla del generador aleatorio. El modelo Random Forest se basa en un conjunto de árboles de decisión que trabajan de manera conjunta para mejorar la precisión y reducir el sobreajuste, siendo especialmente eficaz en tareas de clasificación como el diagnóstico de enfermedades. Esta configuración básica sugiere que se utilizarán los valores por defecto para el resto de los hiperparámetros del modelo.

Modelado de Red Neuronal

Finalmente, se presenta el modelado mediante una Red Neuronal. Este algoritmo, inspirado en el funcionamiento del cerebro humano, permite identificar relaciones no lineales y complejas entre las variables analizadas, constituyéndose en una alternativa potente para el diagnóstico de hipertensión arterial.

Figura 27

Modelado en algoritmo de Red Neuronal



La figura 27 muestra la inicialización de un modelo MLPClassifier (Clasificador de Perceptrón Multicapa), configurado con una sola capa oculta

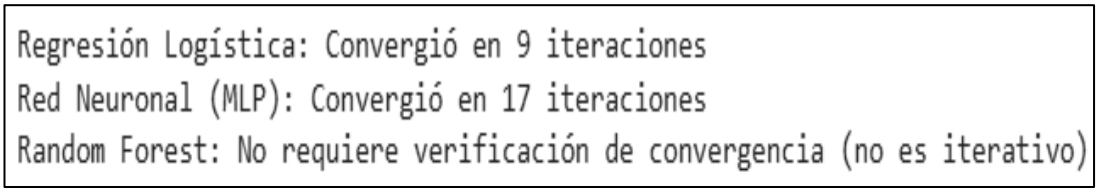
compuesta por 50 neuronas (hidden_layer_sizes=(50,)), lo que le permite aprender patrones complejos de los datos de entrada; se entrenará por un máximo de 500 iteraciones (max_iter=500), lo cual limita la duración del proceso de entrenamiento, aunque podría detenerse antes si el modelo converge, y se ha fijado una semilla aleatoria (random_state=42) para asegurar la reproducibilidad de los resultados, permitiendo obtener siempre los mismos resultados al repetir el entrenamiento bajo las mismas condiciones; en conjunto, esta configuración representa un punto de partida común y razonable para abordar problemas de clasificación como el diagnóstico de hipertensión arterial.

Prueba de convergencia

Para evaluar la estabilidad de los modelos empleados, se analizó el proceso de convergencia durante el entrenamiento, el cual permite determinar si los algoritmos alcanzaron una solución óptima o un punto en el que ya no se producen mejoras significativas en su desempeño.

Figura 28

Datos de la Convergencia de tes modelos de Machine Learning



En la figura 28 los resultados del proceso de convergencia para los modelos de Machine Learning evaluados indican que la Regresión Logística alcanzó una solución óptima en solo 9 iteraciones, lo que refleja que el modelo se ajustó de manera rápida y eficiente, la Red Neuronal, convergió en apenas 17 iteraciones, es notablemente bajo para este tipo de modelo, sugiriendo que los datos eran fácilmente aprendibles, permitiendo una rápida minimización de la función de pérdida. Finalmente, el modelo Random Forest no requiere proceso de convergencia al no ser iterativo, ya que su

funcionamiento se basa en la construcción independiente de múltiples árboles de decisión. En conjunto, estos resultados muestran una alta eficiencia en el entrenamiento de los modelos iterativos, evidenciando que tanto la Regresión Logística como la Red Neuronal alcanzaron un aprendizaje estable en un número mínimo de iteraciones.

Con los resultados obtenidos se tiene que los modelos se encuentran listos para ser evaluados.

4.5.5. Fase 5: Evaluación de los modelos

Modelo de Regresión Logística

Figura 29
Evaluación del algoritmo de Regresión Logística

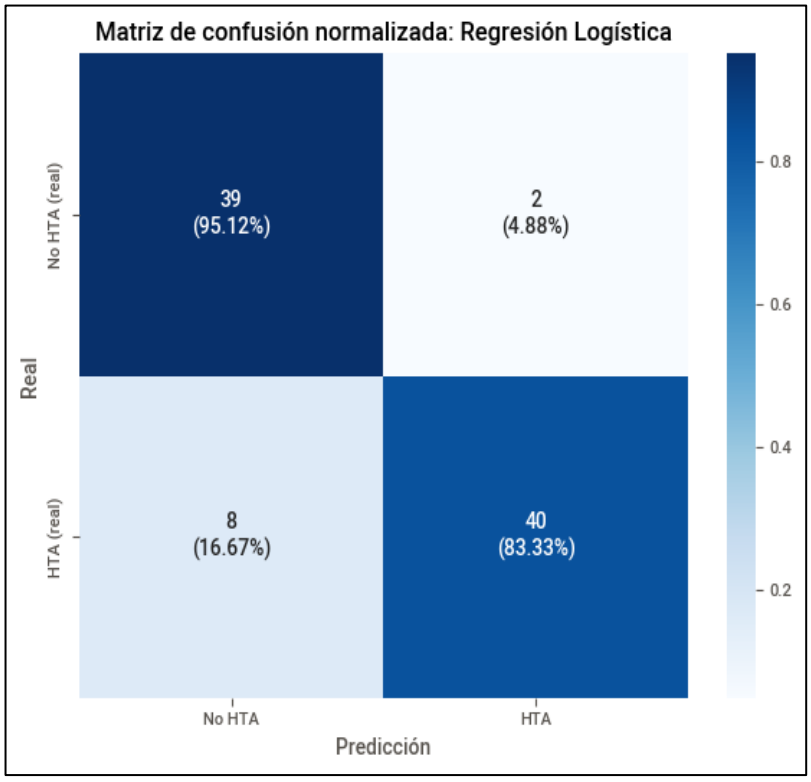
```
Evaluación del algoritmo: Regresión Logística
Métrica global           Valor
Exactitud            0.887640
Precisión global     0.952381
Sensibilidad global  0.833333
Especificidad global 0.951220
F1-score global      0.888889

Métricas por clase:
Clase: No HTA
  Precisión:     0.829787
  Sensibilidad:  0.951220
  F1-score:      0.886364
  Especificidad: 0.833333
Clase: HTA
  Precisión:     0.952381
  Sensibilidad:  0.833333
  F1-score:      0.888889
  Especificidad: 0.951220
```

En la figura 29 la evaluación del modelo de Regresión Logística aplicado a la detección de hipertensión arterial (HTA) muestra un desempeño global sólido, con una exactitud del 88.76%, una precisión del 95.24% y una sensibilidad del 83.33%, lo que indica que el modelo identifica correctamente la mayoría de los casos, especialmente cuando predice la presencia de HTA, aunque aún deja pasar algunos casos reales que son los falsos negativos. También presenta una alta especificidad del 95.12%, lo que demuestra gran eficacia al descartar correctamente a quienes no tienen HTA. A nivel de clases, el modelo se desempeña especialmente bien en la clase "No HTA", con una

sensibilidad de 95.12% y una precisión del 82.98%, mientras que en la clase "HTA" mantiene una excelente precisión (95.24%) y buena sensibilidad (83.33%), aunque esta última es un área que podría mejorarse para reducir el riesgo de no detectar pacientes con hipertensión. En conclusión, la Regresión Logística es un clasificador confiable y preciso, especialmente útil para evitar falsos positivos en el diagnóstico de HTA, aunque sería beneficioso optimizar su sensibilidad para garantizar una mayor detección de los casos reales positivos.

Figura 30

Matriz de confusión para Regresión Logística



En la figura 30 la matriz de confusión generada para el modelo de Regresión Logística aplicado al diagnóstico de hipertensión arterial (HTA) evidenció un desempeño adecuado en la clasificación de los casos.

El modelo alcanzó una sensibilidad del 83.33 %, lo que indica que fue capaz de identificar correctamente a 40 de los 48 pacientes que realmente presentaban hipertensión. Asimismo, obtuvo una especificidad del 95.12 %, clasificando de manera

correcta a 39 de los 41 pacientes que no presentaban HTA. No obstante, se identificaron 8 falsos negativos (16.67 %), es decir, 8 pacientes hipertensos que fueron clasificados erróneamente como no hipertensos, y 2 falsos positivos (4.88 %), correspondientes a pacientes no hipertensos clasificados erróneamente como hipertensos. Estos resultados reflejan que, si bien el modelo posee un buen nivel de precisión global, es necesario seguir optimizando sus parámetros para reducir el margen de error en los diagnósticos, especialmente en la detección de casos positivos, donde una clasificación incorrecta puede tener implicancias clínicas importantes.

Modelo de Random Forest

Figura 31

Evaluación del algoritmo de Random Forest

```
Evaluación del algoritmo: Random Forest
Métrica global           Valor
Exactitud            0.910112
Precisión global     0.934783
Sensibilidad global  0.895833
Especificidad global 0.926829
F1-score global      0.914894

Métricas por clase:
Clase: No HTA
  Precisión:     0.883721
  Sensibilidad:  0.926829
  F1-score:      0.904762
  Especificidad: 0.895833
Clase: HTA
  Precisión:     0.934783
  Sensibilidad:  0.895833
  F1-score:      0.914894
  Especificidad: 0.926829
```

En la figura 31 el modelo Random Forest demostró un rendimiento altamente eficaz para la clasificación de Hipertensión Arterial (HTA), con una exactitud del 0.91, precisión del 0.9347, sensibilidad del 0.8958 y especificidad del 0.9268, lo que evidencia un equilibrio notable entre la detección de casos positivos reales y la minimización de falsos positivos. A nivel de clase, mantuvo métricas sólidas tanto para pacientes con HTA como para aquellos sin la enfermedad, mostrando un mejor balance global respecto a la Regresión Logística, cuya sensibilidad fue de 0.8333, mientras que

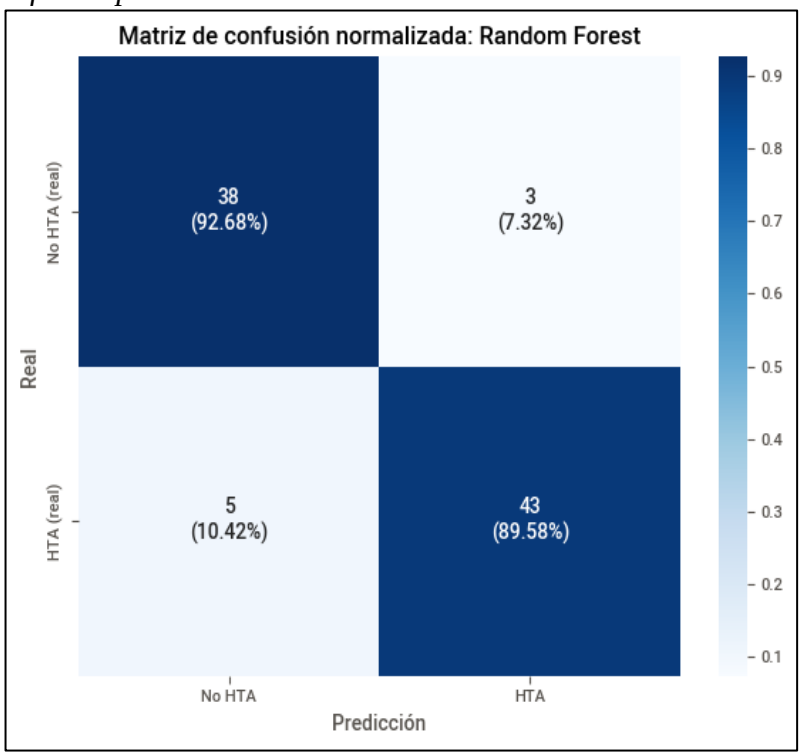
Random Forest alcanzó 0.8958, lo que evidencia su superioridad en la detección de casos positivos, como lo refleja su F1-score de 0.9149.

Para la clase HTA, el modelo alcanzó una precisión de 93.48 %, una sensibilidad de 89.58 % y un F1-score de 91.49 %, evidenciando su eficacia en la detección de pacientes con hipertensión. Para la clase “No HTA”, la precisión fue de 88.37 %, la sensibilidad de 92.68 % y el F1-score de 90.48 %, lo que muestra un equilibrio adecuado en la clasificación de individuos sanos, aunque con una leve menor precisión respecto a la clase HTA.

Estos resultados indican que Random Forest es un modelo robusto y confiable, especialmente útil en contextos clínicos donde es crucial maximizar la detección de pacientes con HTA sin sacrificar la precisión del diagnóstico.

Figura 32

Matriz de confusión para Random Forest



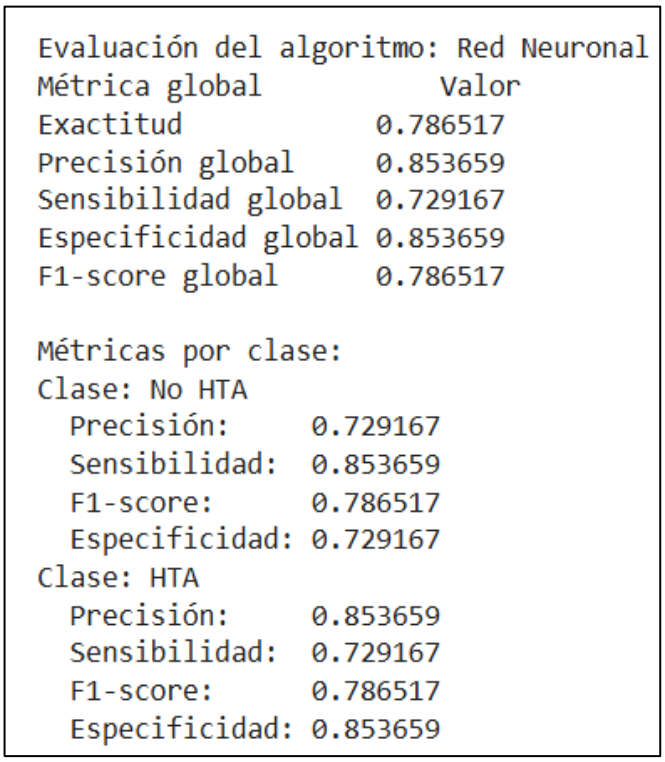
En la figura 32 la matriz de confusión correspondiente al modelo Random Forest evidenciando un desempeño robusto en la clasificación de pacientes con hipertensión arterial (HTA). El modelo logró una sensibilidad del 89.58 %, al

identificar correctamente 43 de los 48 pacientes con diagnóstico real de HTA, y una especificidad del 92.68 %, al clasificar correctamente 38 de los 41 pacientes sin la enfermedad. No obstante, se detectaron 5 falsos negativos (10.42 %), lo cual representa pacientes hipertensos que fueron clasificados como no hipertensos, y 3 falsos positivos (7.32 %), es decir, casos no hipertensos mal diagnosticados como hipertensos. Estos resultados reflejan que el modelo de Random Forest presenta una alta capacidad para detectar patrones clínicos complejos y realizar predicciones precisas, lo que lo convierte en una herramienta confiable para el apoyo en diagnósticos médicos automatizados.

Modelo de Red Neuronal

Figura 33

Evaluación del algoritmo de Red Neuronal

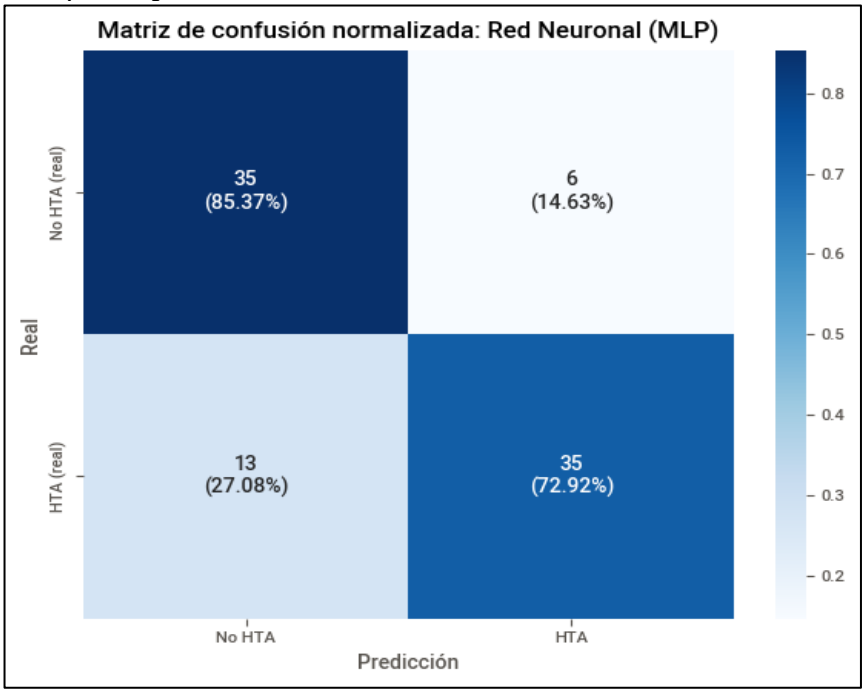


La figura 33 muestra la evaluación del modelo Red Neuronal (MLPClassifier) para la detección de hipertensión arterial (HTA), alcanzando una exactitud del 78.65 %, precisión global del 85.37 % y sensibilidad global del 72.92 %. Estos resultados indican que, aunque el modelo es eficaz al predecir casos positivos, deja sin

detectar aproximadamente un 27 % de los pacientes realmente hipertensos, lo cual es clínicamente relevante. A nivel de clase, la red muestra una precisión del 85.37 % para casos HTA y una especificidad del 85.36 % para casos No HTA, aunque la sensibilidad en la clase HTA (72.92 %) revela una limitación importante, ya que implica una mayor proporción de falsos negativos. En comparación con otros modelos evaluados, como la Regresión Logística (exactitud de 88.76 %, sensibilidad de 83.33 %) y, especialmente, Random Forest (exactitud de 91.01 %, sensibilidad de 89.58 %), el MLPClassifier presenta un rendimiento inferior en detección de casos reales positivos. Por tanto, aunque la Red Neuronal puede ser considerada como un modelo útil, no se perfila como la mejor opción en este estudio para el diagnóstico de HTA, debido a su menor sensibilidad en la clase crítica.

Figura 34

Matriz de confusión para Red Neuronal



La figura 34 presenta la matriz de confusión correspondiente al modelo de Red Neuronal (MLPClassifier) aplicado al diagnóstico de hipertensión arterial (HTA). De los 89 pacientes evaluados, el modelo clasificó correctamente a 35 pacientes sin hipertensión (No HTA), lo que representa una especificidad del 85.37 %, y a 35

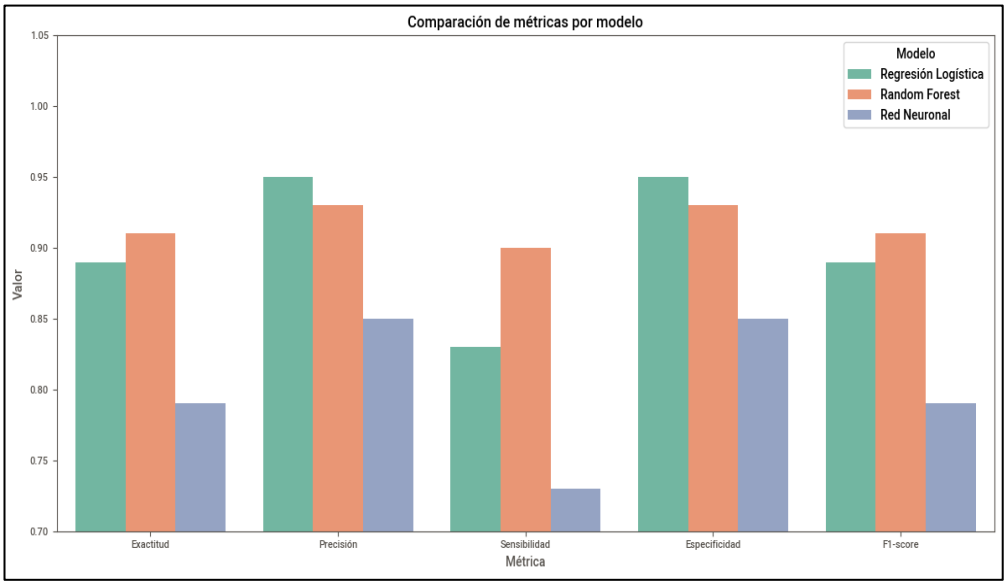
pacientes con hipertensión (HTA), alcanzando una sensibilidad del 72.92 %. No obstante, se identificaron 6 falsos positivos (14.63 %), es decir, pacientes sin HTA mal diagnosticados como hipertensos, y 13 falsos negativos (27.08 %), que representan pacientes hipertensos no detectados por el modelo.

En conjunto, estos resultados reflejan un desempeño moderado del modelo, especialmente en la sensibilidad hacia casos de hipertensión, lo cual representa una limitación crítica en el ámbito clínico. Por tanto, se sugiere considerar mejoras en la arquitectura del modelo o en el proceso de entrenamiento para reducir la tasa de falsos negativos, ya que un error de este tipo podría implicar consecuencias severas en la salud del paciente al no recibir un diagnóstico oportuno.

Comparación de métricas

Figura 35

Comparación de métricas por cada modelo



En la figura 35 se presenta una comparación agrupada de las métricas clave de desempeño (exactitud, precisión, sensibilidad, especificidad y F1-score) para los tres modelos de Machine Learning aplicados al diagnóstico de hipertensión arterial: Regresión Logística, Random Forest y Red Neuronal.

De manera general, se observa que el modelo Random Forest demostró un rendimiento superior en casi todas las métricas analizadas. En términos de exactitud, alcanzó un valor de 0.910, lo que indica que el 91 % de las predicciones realizadas por el modelo coincidieron con los valores reales del conjunto de prueba. Este porcentaje supera tanto a la Regresión Logística (0.887) como a la Red Neuronal (0.786), lo que evidencia su mayor capacidad global de clasificación correcta.

En lo que respecta a la precisión, la Regresión Logística alcanzó el valor más alto con 0.952, ligeramente por encima del Random Forest (0.935). Esto sugiere que el modelo es especialmente eficiente para reducir los falsos positivos, lo cual es clínicamente útil para evitar alarmas innecesarias. La Red Neuronal, obtuvo una precisión menor (0.854), lo que refleja mayor margen de error en esta categoría.

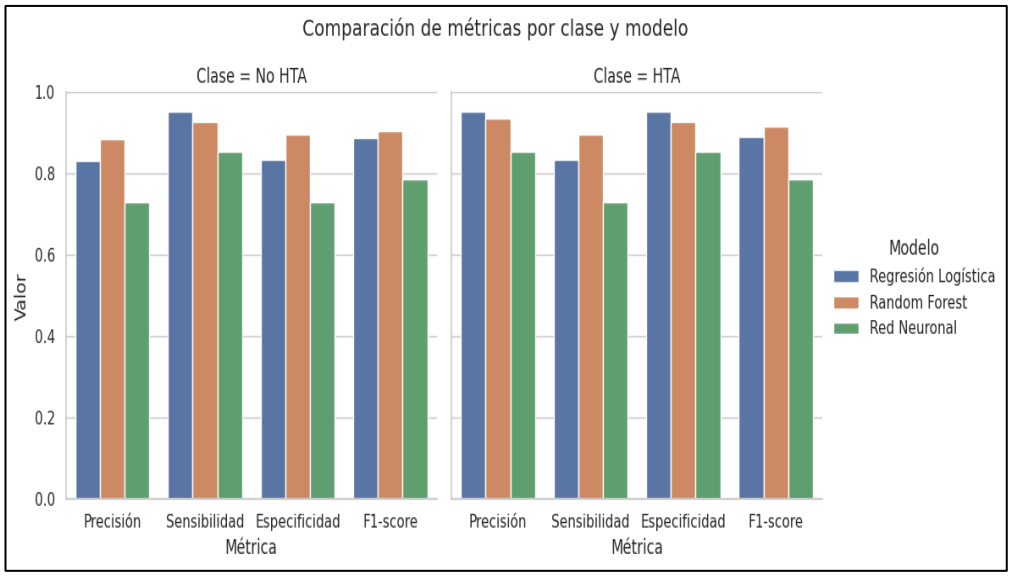
Sin embargo, en la métrica de sensibilidad, que resulta crítica en un entorno médico por su capacidad para identificar correctamente a los pacientes que sí presentan hipertensión, el modelo Random Forest fue el que mejor desempeño mostró, alcanzando un valor de 0.896, siguiendo la Regresión Logística con 0.833, y finalmente la Red Neuronal con 0.729. Este resultado es especialmente relevante porque una sensibilidad baja implica mayor riesgo de falsos negativos, es decir, casos en los que se omite un diagnóstico positivo, lo que puede tener consecuencias clínicas graves.

En cuanto a la especificidad, se observó nuevamente un rendimiento elevado por parte de la Regresión Logística (0.951), seguida por Random Forest (0.927) y la Red Neuronal (0.854). Esto confirma que tanto la Regresión Logística como el Random Forest son modelos eficaces para evitar el sobre diagnóstico de la enfermedad.

Finalmente, el modelo Random Forest se posiciona nuevamente como el mejor, con un valor de 0.915. Este resultado sugiere que Random Forest no solo realiza buenas predicciones positivas y negativas, sino que además mantiene un equilibrio adecuado entre la detección y la confiabilidad diagnóstica. La Regresión Logística obtuvo un F1-score de 0.889, mientras que la Red Neuronal se mantuvo por debajo con 0.787.

Figura 36

Comparación de métricas por clase y modelo



La figura 36 presenta la comparación del desempeño de los modelos Regresión Logística, Random Forest y Red Neuronal según las métricas de precisión, sensibilidad, especificidad y F1-score, diferenciadas por clase ("No HTA" y "HTA"). En la clase No HTA, Random Forest destaca en precisión, especificidad y F1-score, mientras que Regresión Logística lidera en sensibilidad. En contraste, la Red Neuronal muestra el rendimiento más bajo en todas las métricas, evidenciando dificultades para identificar correctamente a los pacientes sin hipertensión.

En la clase HTA, Random Forest mantiene el mejor balance general, con valores altos en precisión y F1-score, siendo eficiente en detectar correctamente a pacientes hipertensos. Regresión Logística resalta por su alta sensibilidad, crucial para

minimizar los falsos negativos. la Red Neuronal, nuevamente, presenta un desempeño inferior, con especial debilidad en sensibilidad.

Figura 37

Métricas de rendimiento de tres modelos diferentes de Machine Learning

	Modelo	Clase	Precisión	Sensibilidad	Especificidad	F1-score
0	Regresión Logística	No HTA	0.829787	0.951220	0.833333	0.886364
1	Regresión Logística	HTA	0.952381	0.833333	0.951220	0.888889
2	Random Forest	No HTA	0.883721	0.926829	0.895833	0.904762
3	Random Forest	HTA	0.934783	0.895833	0.926829	0.914894
4	Red Neuronal	No HTA	0.729167	0.853659	0.729167	0.786517
5	Red Neuronal	HTA	0.853659	0.729167	0.853659	0.786517

<Figure size 1400x600 with 0 Axes>

La figura 37 resume el rendimiento de los modelos evaluados en la clasificación de Hipertensión Arterial (HTA). Haciendo énfasis en las métricas de precisión y sensibilidad que son fundamentales para garantizar diagnósticos correctos en pacientes con la enfermedad. En la clase HTA, Random Forest mostró una combinación sólida de precisión (93.47%) y sensibilidad (89.58%), mientras que Regresión Logística alcanzó una precisión ligeramente superior (95.24%), con una sensibilidad algo menor (83.33%). Estos resultados confirman que ambos modelos son altamente efectivos en la detección de casos reales de HTA, reduciendo tanto falsos positivos como falsos negativos. En cambio, la Red Neuronal presentó un desempeño inferior en ambas métricas, con una sensibilidad del 72.92%, lo que limita su aplicabilidad clínica. En conjunto, Random Forest y Regresión Logística serían las opciones más confiables para el diagnóstico de hipertensión, cada uno con ventajas particulares según la prioridad clínica.

4.6. Análisis de datos

Cada algoritmo fue entrenado bajo las mismas condiciones y evaluado utilizando métricas estándar como la exactitud, precisión, sensibilidad y especificidad. Los resultados fueron los siguientes:

Tabla 2
Resultados de los modelos en valores absolutos

Algoritmo	Verdaderos Positivos (VP)	Verdaderos Negativos (VN)	Falsos Positivos (FP)	Falsos Negativos (FN)	Exactitud	Precisión	Especificidad	F1-score
Regresión Logística	40	39	2	8	79	40	39	44
Random Forest	43	38	3	5	81	43	38	47
Red Neuronal	35	35	6	13	70	35	35	41

Tabla 3
Métricas de rendimiento de los modelos de Machine Learning

Algoritmo	Exactitud (%)	Precisión (%)	Especificidad (%)	F1-score (%)
Regresión Logística	88.76	95.24	95.12	88.89
Random Forest	91.01	93.48	92.68	91.49
Red Neuronal	78.65	85.37	85.37	78.65

En las tablas 2 y 3 se puede observar que el algoritmo Random Forest obtuvo el mejor desempeño general, alcanzando un 91.01 % de exactitud, seguido por la Regresión Logística (88.76 %) y la Red Neuronal (78.65 %). En cuanto al F1-score, que representa el equilibrio entre precisión y sensibilidad, también sobresale el modelo de Random Forest con 91.49 %, lo que confirma su capacidad para clasificar correctamente los casos de hipertensión arterial, minimizando tanto falsos positivos

como falsos negativos. Estas métricas validan la eficacia de los modelos de Machine Learning como herramientas complementarias en la toma de decisiones clínicas, especialmente cuando se utilizan datos clínicos relevantes y bien seleccionados.

4.7. Consideraciones éticas

Garza (2000) menciona que, con el avance constante de la ciencia, el hombre tiende a olvidar sus valores y principios, manifestando falta de empatía y criterio, cabe mencionar que la bioética, nace como respuesta hacia aquellas preguntas que se manifiestan en la medicina y biología, procurando siempre que el avance tecnológico no afecte de forma negativa no sólo al hombre, sino también a la vida vegetal y animal.

En el Código de Ética, capítulo III, artículo 29 del Colegio de Ingenieros del Perú se detalla lo siguiente: “El Ingeniero adquiere un compromiso con la comunidad que debe guiar su actividad profesional a fin de contribuir al estricto cumplimiento de sus obligaciones, a la cabal entrega de sus conocimientos y al proceder honrado donde sea requerido profesionalmente, puesto que se acepta el bienestar y la salud de la sociedad sin un interés lucrativo, se reconoce la seguridad de la vida, salud, bienes y bienestar de la población y de la evolución tecnológica de la nación y se cumplió la elevada misión de guardar y mejorar los recursos naturales y urbanos para una mejor calidad de vida de los habitantes”. (Colegio de Ingenieros del Perú, 2018).

Esta investigación sobre el uso de Machine Learning en el diagnóstico de hipertensión arterial considera principios bioéticos y el Código de Ética del Colegio de Ingenieros del Perú. Se garantiza la confidencialidad y amonificación de los datos médicos para proteger la privacidad de los pacientes. Además, se busca mejorar la precisión diagnóstica sin fines lucrativos, contribuyendo al bienestar de la población. La implementación de los modelos debe evitar sesgos y asegurar resultados justos,

validando rigurosamente las predicciones para garantizar una aplicación ética y segura de la tecnología.

V. Resultados y discusión

5.1. Resultados

Descripción de resultados para cada uno de los objetivos planteados en la investigación

5.1.1. Resultados para el objetivo general

Evaluar de qué manera los algoritmos de Machine Learning pueden mejorar la precisión del diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

Los algoritmos de Machine Learning demostraron una mejora significativa en la capacidad para diagnosticar hipertensión arterial (HTA). Al entrenar modelos con datos clínicos relevantes (como presión arterial, glucosa, IMC, edad, etc.), se logró que los algoritmos aprendieran patrones complejos no lineales que los profesionales podrían pasar por alto.

Entre los modelos aplicados, Random Forest obtuvo el rendimiento global más alto, alcanzando una exactitud del 91.01 %, una precisión del 93.48 %, y un F1-score del 91.49 %, lo que demuestra una clasificación equilibrada y efectiva. Sin embargo, la Regresión Logística también mostró un comportamiento competitivo, con una exactitud del 88.76 %, una precisión del 95.24 %, y una especificidad del 95.12 %, superando incluso al modelo Random Forest en esta última métrica. Esto indica que la diferencia entre ambos modelos no fue marcada, y que ambos constituyen opciones viables para apoyar el diagnóstico clínico de HTA.

Estos resultados respaldan el potencial del uso de algoritmos de Machine Learning como herramientas complementarias en entornos médicos, al ofrecer análisis predictivos confiables basados en múltiples variables clínicas, sin reemplazar el juicio

profesional, y siempre considerando los principios éticos y de validación rigurosa para su aplicación en la práctica.

5.1.2. Resultados para los objetivos específicos

O.E.1 Definir y seleccionar las características y variables clave para la construcción de un dataset que facilite el diagnóstico preciso de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

Tras el proceso de análisis exploratorio, imputación de datos faltantes y codificación de variables categóricas, se procedió a aplicar métodos de selección de características, como SelectKBest, lo cual permitió identificar las variables más influyentes en el diagnóstico de HTA.

Las características clave seleccionadas fueron:

- Presión arterial sistólica y diastólica
- Índice de Masa Corporal (IMC)
- Edad
- Nivel de glucosa
- Perímetro abdominal
- Peso
- Sexo

Estas variables fueron determinantes porque están directamente relacionadas con factores de riesgo de HTA. Al centrar el entrenamiento de los algoritmos en estas variables, se mejoró significativamente la precisión y la eficiencia del modelo.

O.E.2 Determinar y evaluar los algoritmos de Machine Learning que permiten realizar un diagnóstico preciso de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

Se probaron tres modelos distintos: Random Forest, Regresión Logística y Red Neuronal.

Los resultados muestran que el modelo Random Forest fue el que tuvo el mejor desempeño general. Este algoritmo logró una exactitud del 91.01%, lo que significa que acertó en la gran mayoría de los casos. Además, obtuvo una precisión del 93.48% y una sensibilidad del 89.58%, lo que indica que fue muy eficaz tanto en identificar correctamente a los pacientes con hipertensión como en evitar falsos positivos. El modelo de Regresión Logística también presentó buenos resultados, con una exactitud del 88.76%, una precisión del 95.24% y una sensibilidad del 83.33%. Estos números reflejan que es un modelo confiable, además de ser más fácil de interpretar, lo cual puede ser muy útil en el entorno médico. Por otro lado, el modelo de Red Neuronal obtuvo una exactitud más baja (78.65%), con una sensibilidad del 72.92% y una precisión del 85.37%. Aunque su rendimiento fue aceptable, presentó una mayor cantidad de falsos negativos (27.08%), lo que significa que dejó pasar varios casos reales de hipertensión, algo que puede ser riesgoso en el contexto clínico.

El modelo Random Forest se perfila como el más adecuado para apoyar el diagnóstico de hipertensión arterial en este Centro Médico, seguido por la Regresión Logística, que también muestra resultados sólidos. En cambio, el desempeño de la Red Neuronal fue menor, principalmente porque no detectó con suficiente eficacia todos los casos positivos. Estos hallazgos confirman que los algoritmos de Machine Learning pueden ser una herramienta valiosa para mejorar el diagnóstico médico, siempre y cuando se elija el modelo adecuado

O.E.3 Comparar la precisión y efectividad de diversos algoritmos de Machine Learning en el diagnóstico de hipertensión arterial en pacientes del Centro Médico Santiago, Cusco 2024.

La comparación se realizó utilizando gráficos y métricas estadísticas. En general:

Tabla 4

Tabla comparativa del análisis de los algoritmos

Algoritmo	Exactitud (%)	Precisión (%)	Especificidad (%)	F1-score (%)
Regresión Logística	88.76	95.24	95.12	88.89
Random Forest	91.01	93.48	92.68	91.49
Red Neuronal	78.65	85.37	85.37	78.65

En la tabla 4 tenemos la comparación de los algoritmos de Machine Learning utilizados en el diagnóstico de hipertensión arterial (HTA) revela que el modelo Random Forest obtuvo el mejor desempeño general. Este modelo alcanzó una exactitud del 91.01 %, precisión del 93.48 %, especificidad del 92.68 % y un F1-score de 91.49 %, demostrando un excelente equilibrio entre la identificación correcta de casos positivos y negativos, y una capacidad destacada para minimizar tanto falsos positivos como falsos negativos. En segundo lugar, se ubicó el modelo de Regresión Logística, con una exactitud de 88.76 %, precisión de 95.24 %, especificidad de 95.12 % y F1-score de 88.89 %, evidenciando gran eficacia en la correcta clasificación de pacientes no hipertensos, aunque con un desempeño ligeramente inferior en la detección de hipertensos. Finalmente, la Red Neuronal MLP obtuvo los valores más bajos, con una exactitud de 78.65 %, precisión de 85.37 %, especificidad de 85.37 % y F1-score de 78.65 %, lo que sugiere una menor capacidad para distinguir adecuadamente entre pacientes hipertensos y no hipertensos, con una mayor

proporción de falsos negativos. En conjunto, los resultados confirman que los algoritmos de tipo ensamble, como Random Forest, son más robustos y confiables para el apoyo en decisiones clínicas relacionadas con el diagnóstico de hipertensión arterial, especialmente cuando se requiere alta precisión y equilibrio entre sensibilidad y especificidad.

5.2. Discusión de Resultados

Durante el desarrollo de esta investigación, se pudo evidenciar el valor que tienen los algoritmos de Machine Learning como herramientas complementarias en el diagnóstico de hipertensión arterial. A través del análisis de datos clínicos reales del Centro Médico Santiago, se evaluaron tres algoritmos: Random Forest, Regresión Logística y Redes Neuronales, con el objetivo de determinar cuál de ellos ofrecía mayor precisión, sensibilidad, especificidad y efectividad diagnóstica.

A diferencia del estudio de Robles y Millán (2020), en el que los algoritmos de Random Forest y Regresión Logística alcanzaron niveles de exactitud de 90 % y 87 % respectivamente en el diagnóstico de cáncer de mama, los resultados obtenidos en la presente investigación, aplicados al diagnóstico de hipertensión arterial en una muestra de 442 pacientes del Centro Médico Santiago, evidencian un rendimiento ligeramente superior. En nuestro caso, Random Forest logró una exactitud de 91.01 %, mientras que la Regresión Logística alcanzó 88.76 %. Esta diferencia cobra relevancia porque los datos utilizados en este estudio provienen de un contexto clínico local, lo que refuerza la validez externa y la aplicabilidad de los modelos en escenarios reales de salud en la Región Cusco.

Asimismo, al cumplir con el objetivo de comparar la efectividad de diferentes algoritmos, fue evidente que, si bien todos los modelos evaluados mostraron un desempeño aceptable, Random Forest demostró un mejor equilibrio entre precisión,

sensibilidad y especificidad, lo que lo convierte en una alternativa eficaz para el diagnóstico temprano de hipertensión en contextos clínicos. Por su parte, la Regresión Logística, aunque más sencilla y explicativa, mostró limitaciones en cuanto a sensibilidad y especificidad. Las Redes Neuronales presentaron un desempeño competitivo, pero su entrenamiento requiere mayor capacidad computacional, lo cual podría ser un desafío en centros de salud con recursos limitados.

Estos hallazgos se alinean con los resultados obtenidos por Alarcón y Murga (2020), quienes en su estudio sobre diagnóstico de melanoma destacaron el buen desempeño de las Redes Neuronales y Árboles de Decisión, aunque también mencionaron la viabilidad de combinar técnicas para optimizar resultados. En nuestra investigación, aunque no se implementó un algoritmo fusionado, se pudo observar cómo el uso de múltiples modelos permite una visión más integral y comparativa del problema, lo que refuerza la importancia de analizar diferentes enfoques antes de tomar decisiones clínicas basadas en IA.

Además, la experiencia también reafirma lo señalado por Robls (2020) respecto a la necesidad de seleccionar cuidadosamente los algoritmos según la naturaleza del problema y la calidad de los datos disponibles. El uso de modelos bien conocidos como Random Forest y Regresión Logística facilitó el proceso de interpretación y análisis, mientras que la incorporación de Redes Neuronales permitió explorar una alternativa con alto potencial, especialmente para problemas no lineales.

Finalmente, esta investigación no solo cumplió con el objetivo de determinar y evaluar algoritmos adecuados para el diagnóstico de hipertensión, sino que también dejó en evidencia el potencial de las herramientas de Machine Learning para transformar positivamente la práctica médica local, siempre y cuando se usen de forma responsable, con base científica, y pensando en la realidad tecnológica y humana de

cada Centro de Salud. En ese sentido, el desarrollo de modelos predictivos adaptados al contexto de la Región Cusco representa una oportunidad concreta para mejorar la detección temprana, optimizar el tratamiento y, en última instancia, mejorar la calidad de vida de los pacientes hipertensos.

VI. Conclusiones

Se llegaron a las siguientes conclusiones:

1. Los algoritmos de Machine Learning utilizados en esta investigación son: Regresión Logística, Random Forest y Redes Neuronales, demostraron ser herramientas eficaces para mejorar la precisión en el diagnóstico de hipertensión arterial. Mediante el entrenamiento de estos modelos con datos clínicos, se logró identificar patrones complejos y relaciones no evidentes entre las variables, contribuyendo significativamente a la detección de esta condición médica. En particular, Random Forest y Regresión Logística que alcanzaron tasas de precisión superiores al 90%, lo que confirma que el uso de modelos inteligentes puede apoyar de manera confiable la toma de decisiones clínicas en contextos reales.
2. La adecuada selección de características influye directamente en la calidad y el rendimiento de los modelos predictivos. A través del análisis exploratorio de datos se identificaron variables clave para el diagnóstico de hipertensión, como la presión arterial sistólica y diastólica, el índice de masa corporal (IMC), la edad, la glucosa, el perímetro abdominal, el peso y el sexo. La construcción de un dataset sólido con estas variables permitió mejorar la eficiencia del entrenamiento de los modelos y su capacidad de generalización.
3. Los algoritmos evaluados presentaron diferentes niveles de desempeño, siendo Random Forest el más efectivo en términos de precisión, seguido por Regresión Logística y por último Redes Neuronales. Mientras que Regresión Logística destacó por su simplicidad interpretativa, su desempeño fue ligeramente inferior al de los otros modelos más complejos. Las Redes Neuronales mostraron un rendimiento competitivo, aunque su implementación

requiere mayor capacidad computacional y experiencia técnica. Estos resultados evidencian que los algoritmos de Machine Learning pueden integrarse como herramientas complementarias en los sistemas de salud para mejorar la calidad del diagnóstico clínico.

4. La comparación de los modelos basada en métricas como precisión, sensibilidad y F1-score permitió identificar diferencias importantes entre ellos. Random Forest se posicionó como el modelo con el mejor equilibrio entre sensibilidad y precisión, lo que lo convierte en una opción robusta para entornos clínicos donde es crucial detectar con precisión tanto los casos positivos como negativos. Sin embargo, la elección del modelo más adecuado deberá considerar también factores como la interpretabilidad y la facilidad de implementación. En ese sentido, la Regresión Logística continúa siendo una alternativa válida y eficiente en contextos donde se requiere transparencia en la toma de decisiones clínicas.

VII. Recomendaciones

1. Fomentar la adopción progresiva de modelos de Machine Learning en entornos clínicos, como Random Forest y Redes Neuronales, que demostraron ser eficaces en el diagnóstico de hipertensión arterial. Su implementación puede apoyar de forma complementaria a los médicos en la toma de decisiones más precisas y rápidas.
2. Optimizar los procesos de recolección, limpieza y almacenamiento de datos clínicos en los Centros de Salud, asegurando que las variables más relevantes (como presión arterial, IMC, glucosa, edad, perímetro abdominal, etc.) estén completas y correctamente registradas. Esto es esencial para garantizar un entrenamiento eficaz de los modelos predictivos.
3. Establecer programas de formación para el personal de salud en el uso e interpretación básica de los resultados generados por modelos de Machine Learning, especialmente en lo referente a la predicción de enfermedades crónicas como la hipertensión arterial.
4. Incorporar sistemas de evaluación continua del rendimiento de los modelos una vez implementados, para asegurar su actualización periódica con nuevos datos y mantener una alta capacidad de generalización en contextos clínicos reales.

VIII. Referencias

Alarcón Vela, V. M., & Murga Aguilar, D. M. (2020). *Algoritmo para el diagnóstico preliminar de melanoma cutáneo basado en redes neuronales, Naive Bayes y árboles de decisión.* [Tesis de licenciatura, Universidad César Vallejo]

Álvarez-Ocho, R., Torres-Criollo, L. M., Garcés Ortega, J. P., Izquierdo Corone, D. C., Bermejo Cayamcela, D. M., Lliguisupa Peláez, V. D., & Saquicela Salina, A. S. (2022). Factores de riesgo de hipertensión arterial en adultos: una revisión crítica. *Revista Latinoamericana de Hipertensión, 17*(2).

Araya Orozco, M. (2019). *Hipertensión Arterial y Diabetes Mellitus.* Revista Costarricense de Ciencias Médicas, 3 -4. https://www.scielo.sa.cr...

Ayuda Familiar. (2019, 14 de octubre). *12 enfermedades crónicas más comunes en personas mayores.* https://www.ayudafamiliar.es...

Barba Sánchez, A. (2021). *Estudio de técnicas de Machine Learning para el diagnóstico del melanoma y otras lesiones cutáneas a partir de imágenes* [Trabajo final de máster, Universitat Oberta de Catalunya]. http://hdl.handle.net/10609/138407...

Bell, J. (2020). *Machine Learning: Hands-On for Developers and Technical - Professionals.* John Wiley & Sons.

Borja Suárez, M. (2012). *Metodología de la investigación científica para ingenieros.* Editorial Universidad Señor de Sipán.

Bueno, I. (2018, septiembre 7). *Exploring the intersections between Information Visualization and Machine Learning* [Tesis de maestria, Universidad de Sao Paulo]. http://www.teses.usp.br/...

Buitrago, B. (2020, septiembre de 17). *Regresión Logística I: Machine Learning*. Medium. https://medium.com/iwannabedatadriven/...

Calderón Ortiz, J. D., Morales Ticliahuanca, L. F., Roncal Moscol, M. E., & Solórzano Requejo, W. G. (2021). *Uso de algoritmos de Machine Learning para el Diagnóstico de melanomas*. [Tesis de pregrado, Universidad de Piura].

Campos, L., Sán-chez, D. y Abuchar, A. (2019). *Machine Learning y el control de hipertensión arterial*. #ashtag, Revista especializada en ingeniería.

Carbo Coronel, G. M., Berrones Vivar, L. F., & Guallpa González, M. J. (2022). *Riesgos modificables relacionados a la hipertensión arterial. Mas Vita,* 3,4,7.

Colegio de Ingenieros del Perú. (2018). *Código de ética del Colegio de Ingenieros del Perú* (Art.29, Cap. III). https://www.cip.org.pe/...

Córdova Zamora, M. (2003). *Estadística descriptiva e inferencial: Aplicaciones* (5.ª ed.). Moshifera S.R.L.

Doryńska, A., Kęska, R., Wojtyniak, B., Zdrojewski, T., Drygas, W., & Kwaśniewska, M. (2023). Socio-demographic and behavioral factors associated with controlled hypertension after 9 years: Results of a PURE Poland cohort study. Frontiers in Public Health, 11, Article 1167515. https://doi.org/10.3389/fpubh.2023.1167515

ERC. (2021, agosto 4). *Causas y consecuencias de la hipertensión arterial*. El Rincón del Cuidador https://www.elrincondelcuidador.es/...

Espasa Rosell, J. (2022). *Deep Learning en la detección de lesiones cutáneas malignas*. [Trabajado de grado, Escola Técnica Superior d´Enginyeria Industrial de Barcelo].

Familiar, A. (2019, octubre 14). *12 enfermedades crónicas más comunes en personas mayores* https://www.ayudafamiliar.es/...

Garcia Ruiz de León, M. (2018). *Análisis de Sensibilidad Mediante Random Forest*. [Trabajo académico, Escuela Técnica Superior de Ingenieros Industriales - UPM].

Garreta, R. (2020). *Aprendizaje Automático con Python*. Packt Publishing.

Garza G., R. (2000). *Bioética: La toma de decisiones en situaciones difíciles* (p. 13). Trillas.

Geras, K. J. (2021*). Prediction Markets for Machine Learning* - Artificial Intelligence.

Gil Rubio, R., Cruz Perez, E. A., & Perdomo Charry, O. (2022). *Modelos de Machine Learning para clasificar la cartera en un fondo de pensiones*. [Tesis de pregrado, Universidad Santo Tomás].

Google. (2023). *Google Colaboratory*. https://colab.research.google.com/

IBM. (2012, agosto 4). *Minería de textos y el análisis predictivo*. https://developer.ibm.com/es/articles/ap-mineriatextos/

Instituto Nacional de Estadística e Informática. (2021). *Cusco: Enfermedades no transmisibles y transmisibles, 2020*. INEI. https://proyectos.inei.gob.pe/endes/2021/departamentales_en/Endes08/pdf/Cusco.pdf

Jayaram, V. A. (2018). Introduction to special section: Pattern recognition and Machine Learning. Interpretation, 3(4). SAEi-SAEii.

Laureano Yupanqui, C. W. (2022). *Modelo de Machine Learning usando un clasificador de máquinas de soporte vectorial para la detección y*

clasificación del cáncer de seno usando imágenes mamográficas. [Tesis de pregrado, Universidad Nacional del Altiplano].

López, J. &. (2022). *Introducción a Google Colab para el aprendizaje automático*. Académica Española.

Loyola Torres, L. A., & Chamorro Farfán, R. M. (2021). *Implementación de un sistema de diagnóstico clínico aplicando un modelo predictivo de achine Learning para la detección de neumonía en el Hospital Villa Rebagliati de EsSalud, 2021*. [Tesis de pregrado, Universidad Tecnológica del Perú]

Luna Mancilla, P. E., & Vargas Quisca, S. A. (2022). *Uso de la inteligencia artificial para el diagnóstico de COVID -19 a través de radiografía de tórax en hospitales de Cusco, Perú (periodo 2020 -2021)* [Tesis de pregrado, Universidad Andina de Cusco]

McKinney, W. (2018). *Python for Data Analysis*. Q`Reilly Media. http://doi.org/10.5555/2566454

MICROSOFT. (2021, diciembre 4). Ml.NET *Machine Learning*. https://dotnet.microsoft.com/apps/MachineLearningai/ml-dotnet

Millán Gómez, J. A., & Robles Fajardo, J. B. (2020). *Modelo en Machine Learning para el diagnóstico del cáncer de mama* [Tesis de licenciatura, Universidad Autónoma de Ciudad Juárez]. Repositorio Institucional UACJ. https://erecursos.uacj.mx/handle/20.500.11961/17816

Montero Granados, R. (2018). *Modelos de regresión lineal múltiple*. [Tesis, Universidad de Guayaquil].

Nepomuceno de Andrade, G., Matoso, L. F., Miranda, J. W. B., Lima, T. F., Gazzinelli, A., & Vieira, E. W. (2019). *Anthropometric indicators associated with high blood pressure in children living in urban and rural areas*. Revista

Latino-Americana de Enfermagem, 27, e3150. https://doi.org/10.1590/1518-8345.2760-3150.

Organización Mundial de la Salud (OMS). (2023, marzo 16) *Hipertensión arterial*. https://www.who.int/es/news-room/fact-sheets/detail/hypertension.

Organización Mundial de la Salud (OMS). (2016). Preguntas y respuestas sobre la hipertensión. Organización Mundial de la Salud.

Ortiz, R. (2019). *Diomics for diagnosis and assessing brain disease*s: An approach. [Tesis Doctoral, Universidad Politécnica de Valencia].

Otero, L. (2018). *Técnicas conversacionales para la recogida de datos en investigación cualitativa.*

Pérez, J. &. Pérez, R. (2020). *Análisis de Componentes Principales: Fundamentos y Aplicaciones en Ciencia de Datos.* Alfaomega.

Quintanilla, M., Tamayo, D., Sunta-Xi, D., & Quintanilla, J. L. (2022). *Propuesta de una aplicación de aprendizaje automático con visión artificial como herramienta de apoyo para la detección de melanomas benignos y malignos*. [Tesis de pregrado, Universidad de las Fuerzas Armadas ESPE]. https://repositorio.espe.edu.ec/handle/21000/28268.

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3.ª ed.). Pact Publishing.

Ramírez, C., & Martínez, C. (2021). *Dermatoscopía de nevos melanocíticos congénitos acrales: Puesta al día.* Revista Chilena de Dermatología, 37(1), 20–24.

Rodríguez Vera, C. M. (2023). *Relación entre el índice de masa corporal e hipertensión arterial en pacientes adultos* [Tesis de licenciatura, Universidad César Vallejo].2

Rojas Alvarez, G. (2022). *Clasificación de leucocitos en imágenes microscópicas de frotis sanguíneo usando Machine Learning y CNN*. [Tesis de pregrado, Universidad Andina del Cusco].

Schwaber, K., & Sutherland, J. (2013). La guía definitiva de Scrum: Las reglas del juego. Scrum.org & ScrumInc. https://scrumguides.org

Thomas, L. (2022, diciembre 17). *Hipertensión arterial: síntomas y causas*. Mayo Clinic. https://www.mayoclinic.org/es/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410.

Tian, Y. (2020). Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm. 8, 2. IA.

Torres Gonzales, L. C., Mamani Lazo, Y. R., & Choque Quispe, C. A. (2021). *Factores asociados a hipertensión arterial en pacientes atendidos en el centro de salud Ciudad Nueva*. Universidad Nacional Jorge Basadre Grohmann. https://repositorio.unjbg.edu.pe/handle/UNJBG/5003.

Torres, F. J. (2020). *Introducción a la visualización de datos con Python*. Independently published

Useche, L. M., & Mesa, D. M. (2006). *Una introducción a la imputación de valores perdidos*. Terra Nueva Etapa, 22(31), 127–151. http://www.redalyc.org/articulo.oa?id=72103106

Valero Gómez, J. C., Zúñiga Incalla, , A. P., & Clares Perca, C. J. (2021). *Detección de la tuberculosis con algoritmos de deep Learning en imágenes de radiografías del tórax. Revista de Investigación en Salud* - VIVE, pp. 624 - 633.

Vargas, J. &. (2021). *Algoritmos de búsqueda y optimización en inteligencia artificial*. Alfaomega.

Weng, W.-H. (2020). Machine Learning for clinical predictive analytics. En W.-H. Weng (Ed.), *Leveraging data science for global health* (p. 200). Springer.

Wirth, R., & Hipp, J. (2018). *Towards a standard process model for data mining*. *Data Mining and Knowledge Discovery*.

Los anexos, panel fotográfico y otros documentos están resguardados en la oficina de repositorio digital institucional en la Biblioteca Central de la Universidad Tecnológica de los Andes