COMPARACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RIESGO DE INFARTO DE MIOCARDIO

Andersson Camilo Ordoñez Ruiz
Héctor Manuel Muñoz Beltrán
Miguel Felipe Corredor Montejo
William Rene Moreno Romero

Universidad EAN
Especialización en Machine Learning
Docente: MARIE JOSE CHERY LEAL

Bogotá, Colombia
18 de noviembre de 2024

RESUMEN

La presente investigación pretende determinar cuál algoritmo de *Machine Learning* es más efectivo para realizar la predicción de infarto de miocardio usando datos clínicos, con el ánimo de realizar una detección temprana y oportuna para así reducir los índices de morbimortalidad. Para el alcance de dicho objetivo, se brinda un contexto respecto a conceptos como infarto de miocardio, *Machine Learning* y sus diversos tipos de aprendizaje, la definición de algunos algoritmos y métricas de desempeño habituales. El estado del arte resalta investigaciones relevantes en las cuales se ha abordado la aplicación de *Machine Learning* para la predicción de enfermedades cardiovasculares, haciendo énfasis en el infarto de miocardio.

El proceso para la elección de datos clínicos y algoritmos de *Machine Learning*, implementación de algoritmos y análisis de resultados, se realiza en cuatro fases que se encuentran detalladas en la sección Metodología. Mientras que en la sección de análisis y discusión de resultados se realiza una comparación de las métricas de desempeño obtenidas para finalmente concluir que algoritmos ofrecieron un mejor rendimiento.

PALABRAS CLAVE: *Machine Learning*, infarto de miocardio, datos clínicos, modelos de aprendizaje, algoritmos, inteligencia artificial.

1. PROBLEMA DE INVESTIGACIÓN

Las enfermedades cardiovasculares son la principal causa de muerte en todo el mundo, y el infarto de miocardio es una de sus manifestaciones más graves (Organización Mundial de la Salud, 2021). La detección temprana y la predicción precisa del riesgo de infarto de miocardio son cruciales para la intervención oportuna y la reducción de la morbimortalidad (Díaz Delgado, 2022).

Por otro lado, los factores de riesgo son aquellos que se encuentran ligados al desarrollo de una enfermedad mas no son los causantes de esta, entre los factores de riesgo se encuentran los relacionados a la edad, genero, genética y los relacionados al modo de vida como el sedentarismo, alcoholismo, tabaquismo, hábitos alimenticios, etcétera (Cruz Micán et al., 2020).

La prevención de enfermedades cardiovasculares es crucial para reducir la carga de morbilidad y mortalidad, mejorando la calidad de vida de las personas, la Organización Mundial de la Salud (OMS) define la prevención de la enfermedad como:

> La prevención de la enfermedad abarca las medidas destinadas no solamente a prevenir la aparición de la enfermedad, tales como la reducción de los factores de riesgo, sino también a detener su avance y atenuar sus consecuencias una vez establecida. (Organización Mundial de la Salud, 1998, p. 13).

2. PREGUNTA DE INVESTIGACIÓN

¿Cuáles modelos de *Machine Learning* son más efectivos para predecir el riesgo de Infarto de Miocardio utilizando datos clínicos, y cómo comparar su rendimiento?

3. OBJETIVOS

3.1. Objetivo general

Comparar el rendimiento de diferentes algoritmos de *Machine Learning* en la predicción del riesgo de infarto de miocardio utilizando datos clínicos.

3.2. Objetivos específicos

1. Establecer un conjunto de datos clínicos consultando repositorios de uso libre para el entrenamiento de algoritmos de *Machine Learning*.

2. Evaluar el rendimiento de los diferentes algoritmos implementados para la predicción del riesgo de infarto de miocardio, utilizando las métricas de desempeño más relevantes.

3. Determinar la variable con mayor impacto en el riesgo de infarto de miocardio mediante el análisis y la evaluación de las variables más influyentes en su predicción y así establecer cuáles son los determinantes principales de este riesgo cardiovascular.

4. JUSTIFICACIÓN

La predicción temprana de personas con alto riesgo de sufrir un infarto de miocardio es fundamental para mejorar los tratamientos médicos a las personas propensas a sufrir afectaciones cardiacas y reducir las tasas de mortalidad por esta afección. A pesar de que existen métodos tradicionales para evaluar el riesgo cardiovascular, como escalas clínicas y evaluaciones basadas en factores de riesgo, puede ser limitadas e imprecisas, pero con el uso en la medicina de los avances tecnológicos de la inteligencia artificial (IA) y el *Machine Learning* y su capacidad de manejar grandes volúmenes de datos y descubrir patrones complejos que a simple vista pasan desapercibidos, representa una oportunidad para mejorar la predicción temprana de esta afectación y riesgo en la salud pública, previniendo complicaciones más graves y poder salvar vidas.

La utilización de datos clínicos, tales como edad, presión arterial, colesterol, historial médico, entre otros, proporciona una base sólida para la predicción de eventos cardíacos. Sin embargo, la efectividad de los modelos de *Machine Learning* en la predicción de infarto de miocardio depende de varios factores, como el tipo de modelo utilizado, la calidad de los datos y la metodología de evaluación. Por ello, es esencial determinar cuáles modelos de *Machine Learning* son más efectivos en términos de precisión, sensibilidad y otras métricas clave.

Esta investigación es relevante debido a la necesidad de identificar los algoritmos de *Machine Learning* más adecuados para la predicción del riesgo de infarto de miocardio, considerando tanto su rendimiento como su interpretabilidad. Los resultados de este estudio podrían guiar la selección de modelos para aplicaciones clínicas, mejorando la detección temprana y la prevención del infarto de miocardio. La interpretabilidad de los modelos es esencial en el ámbito médico,

ya que permite a los profesionales de la salud comprender las razones detrás de las predicciones y generar confianza en su uso (Krittanawong et al., 2017).

Además, la implementación de modelos predictivos efectivos podría afectar significativamente a la reducción de costos médicos y mejorar la calidad de vida de los pacientes (Riveros et al., 2005), al facilitar intervenciones preventivas oportunas. Dado el creciente volumen de datos clínicos disponibles y el potencial de las tecnologías de *Machine Learning*, este tipo de trabajos son importantes para optimizar la atención médica en pacientes con riesgo cardiovascular, mejorando la eficiencia en la atención médica y prevención.

En el ámbito institucional, este trabajo se enmarca en el campo de investigación “Ciencia, tecnología e innovación”, dentro del grupo de investigación “Ciencias Básicas”, el cual a su vez contiene la línea de investigación “Estadística Aplicada y Ciencia de Datos”.

5. MARCO TEÓRICO

5.1. Enfermedades cardiacas

Este término está directamente relacionado con las afecciones del corazón y los vasos sanguíneos. Estos problemas suceden debido a la acumulación de colesterol en las paredes de los vasos sanguíneos, los cuales con el tiempo se van estrechando como consecuencia de la acumulación de este tipo de placa de colesterol, ocasionando problemas en diferentes partes del cuerpo (Goldman Lee, 2021).

Existen varios tipos de enfermedades cardiovasculares como (Goldman Lee, 2021):

- Cardiopatía coronaria
- Insuficiencia cardíaca
- Arritmias
- Arteriopatía periférica
- Presión arterial alta
- Accidente cerebrovascular
- Cardiopatía congénita

Dentro de estos tipos de enfermedades se encuentra el infarto de miocardio la cual trataremos a continuación.

- **Infarto de Miocardio (Contexto Clínico)**

El infarto agudo de miocardio, conocido también como ataque al corazón, es la necrosis o muerte de una porción del músculo cardíaco que se produce cuando se obstruye completamente el flujo sanguíneo en una de las arterias coronarias (Ranya N. Sweis, 2024).

Sus causas, aunque mayormente asociadas a condiciones médicas de riesgo como hipertensión, el colesterol alto, el tabaquismo, la diabetes, el historial familiar y el estilo de vida sedentario, puede generarse por causas extraordinariamente raras como la generada por trombos (coágulos de sangre) desarrollados en otro lugar del cuerpo diferente al corazón, pero que llegan a las arterias coronaria y generan la obstrucción, conllevando a un infarto de miocardio (Ranya N. Sweis, 2024).

Cualquier persona puede sufrir un infarto de miocardio, haciendo prioritario identificar de forma temprana a las personas con este riesgo. Con ayuda de los datos clínicos recolectados por las diferentes instituciones de salud (Hospital, centros médicos, centros de investigación, etc.) se podría identificar con más eficacia la población con este riesgo, aplicando los tratamientos de manera temprana, factor clave para el éxito en este tipo de afección (Quirón Salud, 2022).

5.1..1. El Infarto de Miocardio como Problema de Salud Pública

El infarto de miocardio se erige como una de las manifestaciones más letales de las enfermedades cardiovasculares, principal causa de mortalidad a nivel global (Organización Mundial de la Salud, 2021). Esta patología, desencadenada por la interrupción del flujo sanguíneo hacia el corazón, conlleva al daño del tejido cardíaco y graves consecuencias para la salud. La detección temprana y la predicción precisa del riesgo de infarto de miocardio resultan esenciales para implementar intervenciones preventivas y terapéuticas oportunas, mitigando así la morbimortalidad asociada (Díaz Delgado, 2022). Esta necesidad ha impulsado la búsqueda de métodos más efectivos que los modelos tradicionales para la identificación de individuos en alto riesgo.

5.1..2. Prevención del Infarto agudo de miocardio

Aunque teóricamente se habla de que cualquier persona puede sufrir un infarto de miocardio, lo cierto es que hay variedad del nivel de riesgo entre unas personas y otras. Factores genéticos, antecedentes familiares y personales, hábitos y estilo de vida hacen más o menos probable que una persona sufra la enfermedad. Es amplio el listado de factores, pero se pueden clasificar en modificables y no modificables (Dattoli García, 2021).

Una clave importante para la prevención es la identificación y control de los factores de riesgo modificables; los cuales pueden llegar a ocupar la mayoría de las causas de los infartos que se pueden identificar y modificar el comportamiento para poder prevenir (Dattoli García, 2021). Algunos de estos factores son:

- Tabaquismo
- Mala alimentación
- Presión arterial alta (hipertensión) no controlada
- Niveles altos de colesterol en sangre
- Diabetes no controlada
- Exceso de peso corporal
- Inactividad física y estrés

5.2. Datos Clínicos / Historia Clínica

La historia clínica es un registro que documenta la relación médico-paciente y contiene información detallada de los eventos médicos en la vida de una persona. Es uno de los elementos más importantes de la práctica médica y está estrechamente regulada por la ley para asegurar su confidencialidad y precisión (Guzmán & Arias, 2012).

Es un documento elaborado exclusivamente por un profesional de la salud, crucial para orientar el tratamiento de un paciente y es la base de investigación científica en medicina. También registra varios hechos de la vida de una persona, incluyendo datos muy íntimos y familiares que deben ser manejados con cuidado (Guzmán & Arias, 2012).

Las variables comunes para predicción de infarto de miocardio son la edad, género, presión arterial, nivel de colesterol, índice de masa corporal, etc. Esta información es extraída de la historia clínica, validando la importancia de una buena elaboración de este instrumento de información.

La calidad de estos datos es importante para procesamiento con técnicas de *Machine Learning*, de faltar datos o presentar inconsistencias, se pueden aplicar técnicas para remplazarlos o corregirlos (media, moda, basada en modelos, vecinos más cercanos, etc.), en predicciones de enfermedades como el infarto de miocardio, la calidad de los datos puede influir en el rendimiento de los modelos de *Machine Learning* encontrados.

Dada la sensibilidad que tienen los datos médicos (incluyen información intima de las personas) están protegidos por la LEY 23 DE 1981, están restringidos la mayor parte de los *dataset* con esta información médica. Sin embargo, existen bases de datos públicas, como las de Framingham o Cleveland (Sun, 2022).

5.3. Inteligencia artificial

La inteligencia artificial tuvo sus origines desde hace varias décadas, pero su auge ha sido exponencial en estos últimos años, dónde se puede indicar que la masificación generada ha sido debido a los diferentes usos que se le ha dado.

Ya no es solamente tema de ciencia ficción, sino que ha traspasado diferentes fronteras para quedarse en las vidas de todos. Es tal su proliferación que ya varios gobiernos han tenido que crear su propia legislación para regular esta industria (Morales Santos et al., 2024).

La combinación de diferentes clases de algoritmos puede llegar incluso a replicar las capacidades humanas. Herramientas usadas por la mayoría de población mundial, como la cámara de equipos celulares para tomar una foto o selfie, ya forma parte de la vida cotidiana o el asistente de voz usado para la búsqueda de cualquier inquietud que se pueda imaginar y diferentes aplicaciones. Plataformas como las de correo electrónico que la mayoría de las personas usan de manera personal o laboral, contienen algoritmos de inteligencia artificial que se encargan de clasificar correos spam, reenviar, etc (Tabima Luque, 2024).

- **Tipos de Inteligencia Artificial**
  Se pueden considerar estos tipos generales:

- Sistemas que piensan como humanos: Se encargan de automatizar la toma de ciertas decisiones llegando en algunos casos a resolver problemas. Son capaces de razonar, aprender, comprender el lenguaje natural y resolver problemas de manera similar a los humanos (Sarker, 2021b) .
- Sistemas que actúan como humanos: este tipo de inteligencia tiene que ver con las herramientas que recrean tareas que normalmente realizaría un humano (Alzubaidi et al., 2021).
- Sistemas que usan la lógica racional: se basan en la lógica formal y el razonamiento deductivo para tomar decisiones. Un ejemplo de este tipo son los sistemas expertos (Sarker, 2021a).

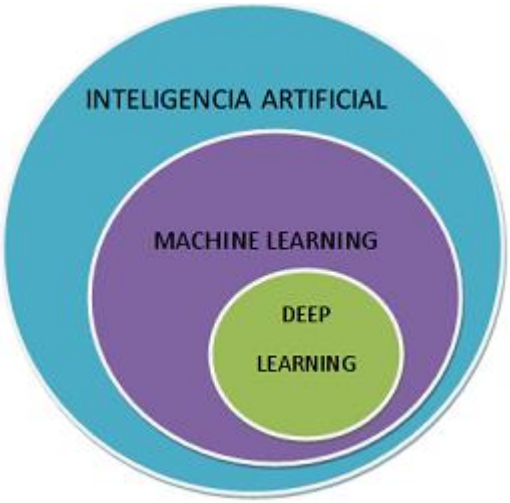
- Sistemas que actúan racionalmente: Para este caso la inteligencia artificial trata de imitar el comportamiento humano, en su aspecto racional basándose en la información disponible y los objetivos deseados (Morishita et al., 2023).

- ***Machine Learning***

Este término se refiere al aprendizaje de maquina y también se le denomina como “inteligencia de aprendizaje automático”. Es una rama o subtipo de la inteligencia artificial tal como se puede apreciar en la Figura 1, que a su vez contiene un subtipo conocido como *Deep Learning* o aprendizaje profundo:

Figura 1

Diagrama de tipos de aprendizaje



Nota: Reproducido de ¿Cuál es la diferencia entre Inteligencia Artificial, *Machine Learning* y *Deep Learning*?, de Mitaritonna Alejandro, 2019, https://www.linkedin.com/pulse/cuál-es-la-diferencia-entre-inteligencia-artificial-y-mitaritonna/, Obra de Dominio Publico

Esta disciplina que hace parte de la inteligencia artificial, a través de algoritmos trata de identificar patrones en grandes volúmenes de datos que de una manera tradicional casi que sería imposible. Mediante esta identificación de patrones se pueden realizar análisis predictivos los cuales posteriormente se pueden implementar en la realización de tareas específicas de forma autónoma (Forero Corba & Bennasar, 2024).

5.3..1. Aplicaciones prácticas del *Machine Learning*

Es muy amplio el espectro de aplicaciones de Machine Learning. A continuación, se listan algunas de los principales tipos de aplicación:

- Sugerencias o recomendaciones: utilizadas por plataformas *e-commerce* para analizar el historial de compras del usuario, así como compras relacionadas de otros usuarios, tendencias y gastos similares (Danilo et al., 2020.).
- Vehículos inteligentes: este tipo de vehículos emplea *Machine Learning* para la configuración interna de temperatura, música, inclinación, luces, frenos (Llopis Sánchez, 2023).
- Redes sociales: Implementan algoritmos para reducir el spam, detección de noticias falsas, contenido no permitido, gustos y tendencias del usuario (Mullah & Zainon, 2021).
- Medicina: Detección temprana de enfermedades como cáncer, análisis de imágenes como ecocardiogramas y radiológicas (Serna-Trejos et al., 2022).
- Búsquedas: Los motores de indexación utilizan estos algoritmos de *Machine Learning* para mostrar resultados de manera eficiente optimizando las búsquedas (Cujar-Rosero et al., 2021).

- Ciberseguridad: Nuevos antivirus emplean *Machine Learning* para la detección de *malware*, escaneo, aceleración de detección entre otras funcionalidades (Pons et al., 2021).

5.3..2. Tipos de aprendizaje en *Machine Learning*

Aprendizaje supervisado

Se denomina así a esta técnica debido a que los algoritmos se entrenan previamente mediante conjuntos de datos de entrenamiento que están etiquetados. Dentro de este tipo de aprendizaje se encuentran los modelos de clasificación (utilizados para variables categóricas) y modelos de regresión (usados con variables continuas) (Valenzuela González, 2022).

Aprendizaje no supervisado

En esta técnica los algoritmos no aprenden a partir de sets de datos de entrenamiento, dispone únicamente de los datos de entrada para identificación de patrones (Arribas Jara, 2018); se asemeja de cierta manera a la forma de aprender cosas nuevas por parte del cerebro humano.

Aprendizaje por refuerzo

Tiene como objetivo aprender políticas o estrategias de control que permiten a un agente interactuar con el ambiente en el que se encuentra contenido, su comportamiento se basa en estímulos. A diferencia de los algoritmos de aprendizaje supervisado, no generaliza situaciones no vistas durante el entrenamiento, tampoco encuentra estructuras escondidas en los datos como lo realizan los algoritmos de aprendizaje no supervisado (Montenegro Meza et al., 2023).

Esta técnica de *Machine Learning* imita el proceso que realizan los seres humanos de ensayo y error para el alcance de objetivos, estos algoritmos descubren por sí mismos las mejores rutas de procesamiento para alcanzar los resultados y se pueden utilizar en entornos complejos con muchas reglas y dependencias, además de que no requieren mucha interacción humana (Torres Jordi, 2021).

5.3..3. Algoritmos de Machine Learning

Diversos algoritmos de *Machine Learning* han sido aplicados en la predicción del riesgo de infarto de miocardio, cada uno con sus propias características y potencial. Algunos de los algoritmos más utilizados incluyen:

Redes Neuronales

Modelos complejos y flexibles capaces de aprender representaciones abstractas de los datos. Pueden manejar relaciones no lineales y capturar patrones sutiles en los datos. Sin embargo, su interpretabilidad es un desafío, lo que puede limitar su aplicación en el ámbito médico (Miotto et al., 2018).

Regresión Lineal

Un modelo lineal ampliamente utilizado para la clasificación binaria. Su principal ventaja es su interpretabilidad, ya que los coeficientes del modelo indican la importancia relativa de cada variable predictora. Sin embargo, puede tener limitaciones en la modelización de relaciones no lineales complejas (Díaz Delgado, 2022).

5.3..4. Algoritmos clasificadores

Esta clase de algoritmos corresponden al conjunto de técnicas utilizadas en *Machine Learning* enfocados en la clasificación de elementos u objetos en diferentes categorías a partir de un entrenamiento previo realizado con un conjunto de datos dado (Sarker et al., 2019). Áreas como la investigación, la academia y por supuesto la industria; los utilizan bastante para sus trabajos que requieren este tipo de herramientas.

Árboles de Decisión

Se considera uno de los métodos más populares para representar clasificadores. Investigadores de diversas disciplinas, como la estadística, el *Machine Learning*, el reconocimiento de patrones y la minería de datos, han abordado el problema de generar un árbol de decisión a partir de los datos disponibles (Rokach & Maimon, 2005)

Random Forests

Un conjunto de árboles de decisión que se entrenan de forma independiente y cuyas predicciones se combinan para obtener un resultado final. Suelen ofrecer un buen rendimiento y son menos propensos al sobreajuste que los árboles de decisión individuales. Sin embargo, su interpretabilidad puede ser más limitada (Rodrigo, 2020).

Máquinas de vectores de soporte (*Support Vector Machines* SVM)

Técnica de algoritmos usada para la clasificación y la regresión. Un hiperplano es usado para separar los datos en distintas clases; suele ser usado para interactuar con datos no lineales (Noble, 2006).

5.3..5. Métricas de desempeño de algoritmos de *Machine Learning*

El desempeño de un algoritmo se puede medir con distintos tipos de métricas. A continuación, se presenta la descripción de las métricas más relevantes:

Matriz de Confusión

La tabla 1 representa la precisión de un modelo de clasificación, muestra el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Las filas de la matriz de confusión corresponden a los valores reales de la variable objetivo, mientras que las columnas corresponden a los valores predichos de la variable objetivo (Kulkarni et al., 2020).

Tabla 1

Matriz de confusión

		Valores Predichos	
		Falso	**Verdadero**
Valores Reales	**Falso**	Verdaderos negativos	Falsos Positivos
	Verdadero	Falsos Negativos	Verdaderos Positivos

Nota: Adaptada de *Foundations of data imbalance and solutions for a data democracy*, de Kulkarni et al., 2020, https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Los términos de tabla 1 se definen a continuación:

- **Verdaderos Positivos y Verdaderos Negativos:** Los valores de la predicción concuerdan con los valores reales de la variable.

- **Falso Positivo:** Corresponde a una predicción errónea, se presenta cuando el valor real de la variable es negativo pero el modelo lo predice como positivo.
- **Falso Negativo:** Es otra predicción errónea, presentada cuando el valor real es positivo pero el modelo lo predice como negativo.

A partir de la matriz de confusión se pueden calcular otras métricas como exactitud, precisión, sensibilidad, especificidad y F1 – *Score*, las cuales se describen a continuación:

Exactitud o *accuracy*

Se define en la ecuación 1 como la relación entre las predicciones correctas realizadas (Verdaderos positivos y verdaderos negativos) y el número total de predicciones (Bernal Vélez et al., 2022).

$$accuracy = \frac{verdaderos\ positivos + verdaderos\ negativos}{Total\ de\ predicciones\ realizadas} \quad \text{(Ecuación 1)}$$

Precisión

Corresponde al porcentaje de verdaderos positivos que fueron predichos, se define en la ecuación 2 como la relación entre verdaderos positivos y la sumatoria entre verdaderos positivos y falsos positivos (Bernal Vélez et al., 2022).

$$precision = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos} \quad \text{(Ecuación 2)}$$

Sensibilidad o *Recall*

Es la tasa de verdaderos positivos e indica la capacidad de predecir los casos verdaderos, se define en la ecuación 3 como una relación entre los verdaderos positivos y la suma de verdaderos positivos con falsos negativos (Bernal Vélez et al., 2022).

$$Recall = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos} \quad \text{(Ecuación 3)}$$

Especificidad o *Specificity*

Corresponde al porcentaje de verdaderos negativos que fueron correctamente predichos, se define en la ecuación 4 como la relación entre verdaderos negativos y la sumatoria entre verdaderos negativos y falsos positivos (Bernal Vélez et al., 2022).

$$Specificity = \frac{verdaderos\ negativos}{verdaderos\ negativos + falsos\ positivos} \quad \text{(Ecuación 4)}$$

F1 – *Score*

Se define como la media armónica de la precisión y el *recall* en la ecuación 5, es una métrica habitualmente usada en modelos de *Machine Learning* aplicados en el campo de la medicina, ya que sirve para medir el desempeño cuando el conjunto de datos se encuentra desbalanceado (Bernal Vélez et al., 2022).

$$F1 - Score = 2 \times \frac{Precisión \times Recall}{Precisión + Recall} \quad \text{(Ecuación 5)}$$

***Receiver Operating Characteristic* (ROC)**

Es una gráfica que enfrenta la tasa de falsos positivos (eje x) contra la tasa de éxito o *Recall* (eje y). Estas tasas se obtienen en función de la variación de un

umbral (valor a partir del cual se decide cual caso es positivo) entre 0 y 1. Es una métrica útil debido a que permite comparar diferentes modelos e identificar cual ofrece mejor rendimiento (Bouza, 2021).

***Area Under Curve* (AUC)**

Corresponde al cálculo del área bajo la curva ROC y se utiliza como resumen de la calidad del modelo, Cuanta más área este contenida dentro de la curva mejor será el clasificador (Bouza, 2021).

5.3..6. El potencial del *Machine Learning* en la predicción del riesgo de infarto de miocardio

El *Machine Learning*, es una rama de la inteligencia artificial, ha emergido como una herramienta prometedora para abordar las limitaciones de los modelos tradicionales. Los algoritmos de *Machine Learning* pueden analizar grandes volúmenes de datos clínicos, identificar patrones complejos y generar modelos predictivos más precisos y personalizados (Rajkomar et al., 2019). Esta capacidad de aprender de los datos y adaptarse a nuevas situaciones convierte al *Machine Learning* en un aliado valioso para la identificación temprana de individuos en alto riesgo de infarto de miocardio.

El *Machine Learning* permite la integración de diversos tipos de datos, incluyendo no solo los factores de riesgo clínicos tradicionales, sino también información sobre biomarcadores, imágenes médicas y otros datos relevantes. Los algoritmos de *Machine Learning* pueden identificar patrones y relaciones no lineales en los datos, lo que podría conducir a una mayor precisión en la predicción y una mejor comprensión de los mecanismos subyacentes al desarrollo del infarto de miocardio.

5.3..7. Predicción del Riesgo de infarto de miocardio con *Machine Learning*

La investigación en la predicción del riesgo de infarto de miocardio mediante *Machine Learning* ha experimentado un crecimiento significativo en los últimos años. Diversos estudios han explorado el uso de diferentes algoritmos y conjuntos de datos, con resultados prometedores en términos de mejora de la precisión predictiva.

Estos estudios, que se incluyen en el item 5.4 , evidencian el potencial del *Machine Learning* para mejorar la predicción del riesgo de infarto de miocardio y otras enfermedades cardiovasculares. Sin embargo, aún persisten desafíos en la selección del algoritmo óptimo, la interpretabilidad de los modelos y la integración de estos modelos en la práctica clínica. La heterogeneidad de los datos clínicos, la presencia de ruido y la necesidad de validar los modelos en poblaciones diversas son algunos de los obstáculos que deben superarse para lograr una implementación efectiva del *Machine Learning* en la predicción del riesgo de infarto de miocardio.

5.4. Estado del arte

El presente estado del arte revisa los avances más relevantes en este campo, destacando los métodos, algoritmos y resultados obtenidos en estudios recientes al usar *Machine Learning* para la predicción de infartos.

- **Diagnóstico de enfermedades cardiacas**

Srinivasan et al., (2023) compara el desempeño de 8 algoritmos de *Machine Learning* para la predicción de enfermedades cardiacas. Realizando la combinación de varias características y algoritmos para desarrollar un modelo de predicción. El

artículo concluye indicando que el modelo de predicción propuesto alcanzó una precisión del 98.07% superando a otros algoritmos analizados como el *Random Forest* y *Decisión Tree* cuya precisión estuvo cerca del 89% y el modelo Naive Bayes el cual obtuvo una precisión del 94.07%.

El estudio de Hajiarbabi, (2024) compara diversos métodos empleados para la predicción de enfermedades cardiacas definiendo 3 categorías de detección: Detección de enfermedad cardiaca basada en información clínica, detección de enfermedad cardiaca basada en electrocardiograma y fonocardiograma, detección de enfermedad cardiaca basada en rayos X. De acuerdo con la clasificación mencionada, el estudio concluye con que los algoritmos *XGBoost (Extreme Gradient Boosting)*, *Random Forest, Ensemble Learning* y métodos de redes neuronales ofrecen el mejor desempeño para la detección de enfermedad cardiaca basada en datos clínicos. El método de redes neurales convolucionales es uno de los mejores métodos para la detección de enfermedad cardiaca basada en señales de electrocardiograma y basada en imágenes de rayos X.

- **Implementación de *Machine Learning* para predicción del riesgo cardiovascular utilizando datos clínicos de rutina y síntomas**

En Weng et al., (2017) demostraron que el *Machine Learning* puede mejorar la predicción del riesgo cardiovascular utilizando datos clínicos de rutina, logrando un aumento significativo en el área bajo la curva ROC (AUC) en comparación con los modelos tradicionales. Motwani et al., (2017) utilizaron *Machine Learning* para predecir la mortalidad por todas las causas en pacientes con sospecha de enfermedad coronaria, logrando una mejora en la discriminación del riesgo en comparación con los modelos clínicos establecidos.

En Nandal et al., (2022) presentan un modelo de *Machine Learning* para la predicción de ataques cardiacos analizando diferentes factores de riesgo como la presión arterial, colesterol alto y diabetes, mediante el uso de modelos como Maquinas de Vectores de Soporte (*Support Vector Machines,* SVM), regresión logística, Naive Bayes y *XGBoost* . Los resultados obtenidos indican que el modelo con mejor predicción fue el de *XGBoost* junto con el de regresión logística, tomando como métrica de desempeño el área bajo la curva.

La investigación de González Cedillo, (2019) compara los modelos Naive Bayesian y Semi-Naive Bayesian, analizando ciertas características medicas como las mencionadas anteriormente (presión arterial, diabetes, sexo). Los algoritmos implementados obtuvieron una precisión del 86%.

El análisis plasmado en el artículo de Mosquera Rojas et al., (2020) señala la importancia de realizar el análisis de electrocardiogramas de manera precisa y oportuna, identificando la necesidad de utilizar técnicas de *Machine Learning* para realizar dichos análisis de una manera automática y confiable. Implementan diversos algoritmos para identificar las señales de electrocardiogramas en 4 categorías: Paciente normal, paciente con fibrilación auricular, paciente con ritmo anormal que puede padecer otra patología y señal ruidosa que no puede ser estudiada. Los algoritmos implementados fueron: Redes Neuronales, *SVM, Random Forest*, Regresión Logística*, XGBoost y MetaCost*. Finalmente, los modelos de clasificación obtuvieron un rendimiento con un *F1 - Score* entre 0.73 y 0.8, siendo el algoritmo de Redes Neuronales el de mejor desempeño.

- **Predicción de arritmias, infartos agudos de miocardio y ataques cardiacos**

El artículo de Patiño et al., (2023) compara y evalúa el aprendizaje y precisión de algoritmos de *Machine Learning* basados en redes neuronales artificiales (

Artificial Neural Networks, ANN), redes neuronales convolucionales (*Convolutional Neural Network,* CNN) y un modelo basado en arboles de decisión *XGBoost*. Para entrenar los algoritmos, los investigadores utilizaron bases de datos que proporcionan electrocardiogramas clasificados con arritmias e infartos agudos de miocardio.

Moreno Sánchez, (2021) presenta un trabajo de grado en el que desarrolla un modelo de *Machine Learning* de aprendizaje no supervisado para la predicción de ataques cardiacos. También experimenta con algoritmos de aprendizaje supervisado con el ánimo de identificar cuales modelos ofrecen un mejor desempeño, teniendo como referencia la métrica *F1-Score*. La investigación concluye que los algoritmos con mejor desempeño respecto a la métrica indicada son el de Regresión Logística y *Random Forest Classifier* alcanzando un F1-*Score* del 85%.

Nandal et al., (2022) analiza diferentes factores de riesgo como la presión arterial, colesterol alto y diabetes, mediante el uso de modelos como *Support Vector Machines*, regresión logística, Naive Bayes y *XGBoost*. Los resultados obtenidos indican que el modelo con mejor predicción de ataque cardiaco fue el de *XGBoost* junto con el de regresión logística, tomando como métrica de desempeño el área bajo la curva, la cual dio resultados del 0.94 y 0.92 respectivamente.

- **Evaluación de Infartos de Miocardio**

Chen et al., (2022) propone modelos de *Machine Learning* para evaluar la gravedad del infarto de miocardio, teniendo en cuenta características fisiológicas, clínicas y paraclínicas. Los investigadores implementan modelos de clasificación para determinar la presencia de infarto y modelos de regresión para cuantificar el porcentaje de miocardio infartado. Los mejores modelos de cuantificación

implementados obtuvieron un error medio de 0.056 y 0.012 correspondiente a los algoritmos *Multilayer Perceptron* y *Support Vector Regression* respectivamente, mientras que el mejor algoritmo de clasificación obtuvo una precisión del 88.67% correspondiente a un algoritmo *Random Forest.*

- **Revisión del Dataset Cleveland**

Ruqiya et al., (2023) analiza estudios previos que han aplicado algoritmos de *Machine Learning* para predecir enfermedades cardíacas, utilizando el *Cleveland Heart Disease Dataset*, que contiene registros clínicos de pacientes con diferentes factores de riesgo cardíaco. Se revisan distintos modelos de *Machine Learning*, como *Random Forest, Support Vector Machines*, y Redes Neuronales, entre otros. Los estudios muestran que algunos algoritmos alcanzan una precisión del 100% en la predicción de enfermedades cardíacas, aunque también se mencionan limitaciones del *dataset*, como la escasez de datos y su distribución desigual por género y edad.

6. METODOLOGÍA

6.1. Primer Nivel

- **Enfoque, alcance y diseño de la investigación**

El enfoque del presente documento se realiza desde una perspectiva cuantitativa, ya que los datos a recolectar son de tipo numérico y/o categórico con el propósito de identificar patrones y relaciones. El diseño no es de tipo experimental debido a que las variables de estudio no son manipuladas, teniendo en cuenta que el conjunto de datos a analizar es extraído de un repositorio de uso libre, la recolección de datos se realizó en un solo momento, por tanto es transversal.

El estudio es de tipo descriptivo y correlacional, debido a que se pretende describir y analizar la relación entre las distintas variables sin establecer causalidad, pero si la relación con el riesgo de infarto de miocardio.

- **Definición de Variables**

Esta fase define las variables que debe contener el *dataset* seleccionado. La elección de estas variables se realizó teniendo en cuenta la revisión bibliográfica que compone el marco teórico y el estado del arte.

6.1..1. Definición conceptual

Las variables de interés deben incluir factores clínicos como la edad, sexo, presión arterial, niveles de colesterol y frecuencia cardiaca. La definición conceptual de cada variable y su relevancia en la predicción del riesgo de infarto de miocardio se presenta a continuación.

Edad: Número de años vividos (Real Academia Española, 2014) desde el nacimiento del paciente hasta el momento de recolección de datos. Suele asociarse a pacientes de edad avanzada con mayor riesgo de sufrir infarto de miocardio.

Sexo: Clasificación biológica de género masculino o femenino (Real Academia Española, 2014). Es una variable relevante para determinar la presencia de infarto de miocardio entre hombres y mujeres.

Presión arterial: Se define como la fuerza que ejerce la sangre contra las paredes de las arterias (Organización Mundial de la Salud, 2023). La presión arterial alta se conoce como hipertensión y es un factor de riesgo importante en la predicción del riesgo de infarto de miocardio.

Nivel de colesterol: Indica la cantidad de colesterol total (lipoproteínas de baja densidad (LDL) y lipoproteínas de alta densidad (HDL)) presente en la sangre (Maldonado Saavedra et al., 2012), un alto nivel de colesterol puede provocar la acumulación de placas en las arterias, incrementando el riesgo de infarto de miocardio.

Frecuencia cardiaca: Esta definida como el número de pulsaciones que realiza el corazón durante un minuto (Vázquez Pérez et al., 2023), es un indicador de capacidad cardiovascular y puede relacionarse con el riesgo de infarto.

6.1..2. Definición operacional

Teniendo en cuenta las definiciones de cada variable, se medirán de la siguiente manera:

Edad: Se mide en años completos, como un número de tipo entero.

Sexo: Se define como una variable categórica binaria, masculino y femenino con la siguiente codificación: [masculino:1, femenino: 0]

Presión arterial: Como unidad de medida se tiene milímetros de mercurio [mmHg]

Nivel de colesterol: Los valores corresponden al colesterol total, como unidad de medida se tiene miligramos por decilitro [mg/dL]

Frecuencia cardiaca: Se mide en número de latidos o pulsaciones por minuto [bpm]

- **Obtención de datos**

Partiendo del hecho que el *dataset* a analizar proviene de un repositorio público y de uso libre, no se realizaran entrevistas ni selección de muestra de población, debido a que este proceso ya fue realizado para la construcción del *dataset.*

6.2. Fases de desarrollo

A continuación, se detallan las actividades de desarrollo de la investigación para dar cumplimiento a los objetivos planteados:

Selección de datos clínicos

- Consulta de repositorios de uso libre para la elección del *dataset*.

Teniendo en cuenta las variables definidas en el item 6.1.2 acorde a la revisión bibliográfica realizada, se elige el *dataset Heart Attack Analysis & Prediction* del repositorio público de Kaggle, para más detalles respecto a la composición del dataset se debe consultar el Apéndice A.

- Definir y procesar las variables de interés del *dataset*

El *dataset* elegido se compone de las siguientes variables:

- **Age**: Edad del individuo (numérica).
- **Sex**: Sexo del individuo (categórica).
  0: Mujer
  1: Hombre
- **Cp**: Tipo de dolor de pecho (Angina) (categórica).
  1: Angina tipica
  2: Angina atipica
  3: Dolor no anginal
  4: Asintomatico

- **Trtbps**: Presión arterial en reposo (numérica).
- **Chol**: Colesterol sérico en mg/dl (numérica).
- **Fbs**: Nivel de azúcar en sangre en ayunas (categórica).
  1: Nivel de azúcar > 120 mg/dl
  0: Nivel de azúcar < 120 mg/dl
- **Restecg**: Resultados del electrocardiograma en reposo (categórica).
  0: Normal
  1: Anormalidad en onda ST-T (inversiones de la onda T y/o elevación o depresión del ST de > 0.05 mV)

2: Hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes

- **Thalachh**: Frecuencia cardíaca máxima alcanzada (numérica).
- **Exng**: Angina inducida por el ejercicio (categórica).
  1 = Si
  0 = no
- **Oldpeak**: Depresión del segmento ST (numérica).
- **Slp**: Pendiente del segmento ST (categórica).
  0: Ascendente
  1: Plana (Horizontal)
  2: Descendente
- **Caa**: Número de vasos principales obstruidos (numérica).
- **Thall**: Defecto reversible del tálamo (categórica).
  0: Normal.
  1: Defecto fijo
  2: Defecto reversible
- **Output**: Presencia de enfermedad cardíaca (categórica).
  0= Menor probabilidad de ataque cardíaco
  1= Mayor probabilidad de ataque cardíaco

Teniendo en cuenta la descripción de las variables del *dataset*, se destaca la presencia de siete variables categóricas, lo que hace necesario implementar el método *get_dummies* para convertir las variables categóricas en variables binarias y también evitar el ordenamiento implícito que conlleva una variable de tipo ordinal, evitando también una interpretación incorrecta de los algoritmos entrenados

Implementación de algoritmos de *Machine Learning*

- Selección de los algoritmos de *Machine Learning* a implementar a partir de la revisión bibliográfica, teniendo en cuenta características como el tipo de algoritmo (regresión, clasificación, redes neuronales) y tipo de aprendizaje.

La variable objetivo del *dataset* escogido es de tipo binario, por ende, se considera adecuado implementar algoritmos de clasificación. Acorde con la revisión bibliográfica realizada en el item 5.4 se opta por escoger los algoritmos *Random Forest*, XGBoost, CNN, SVM, debido a que en los diferentes estudios realizados fueron los algoritmos que obtuvieron mejores resultados, adicionalmente también se valida el comportamiento del algoritmo *K-Nearest Neighbors.* En el Apéndice B se presenta el cuaderno desarrollado para el entrenamiento de los modelos seleccionados.

- Entrenamiento de los algoritmos seleccionados con el conjunto de datos recopilado.

El *dataset* elegido se compone de 303 registros o instancias y 14 columnas, una vez aplicado el método *get_dummies* se genera un total de 27 columnas, obteniendo así una matriz con un total de 8.181 datos. Por buenas prácticas se define que un conjunto correspondiente al 80% de los datos se emplea para realizar el entrenamiento de los algoritmos, mientras que el 20% restante es usado para realizar pruebas de los modelos implementados. Adicionalmente se implementa una estratificación de la variable objetivo, con el fin de asegurar que ambos conjuntos reflejen una distribucion adecuada y proporcional de la variable objetivo *output*

Evaluación del rendimiento de algoritmos

- Evaluación del rendimiento de los algoritmos a partir de las principales métricas de desempeño.

En el item 5.3.2.5 se brinda un contexto de las principales métricas para medir el desempeño de un algoritmo de *Machine Learning,* se definen la matriz de confusión, el área bajo la curva ROC y la exactitud como los principales criterios para la medición del desempeño de los algoritmos. La elección de la matriz de confusión como la métrica principal se fundamenta en la criticidad que representan los falsos negativos, debido a que una predicción de este tipo pone en riesgo a pacientes que no recibirían una atención oportuna.

- Comparación de los resultados obtenidos y generar hallazgos sobre la efectividad de cada algoritmo implementado.

Teniendo en cuenta los resultados obtenidos en la actividad anterior, se realiza una comparación de los criterios definidos para la elección del algoritmo más adecuado en función de las métricas de desempeño.

Análisis de variables influyentes

- Uso de técnicas de análisis de datos para identificar las variables más influyentes en la predicción del riesgo de infarto de miocardio.

Se identifican las variables más influyentes a partir de la matriz de correlación obtenida del *dataset*

- Examen de las variables identificadas para comprender como afectan el riesgo de infarto de miocardio.

- **Técnicas de análisis de datos**

Evaluación del rendimiento de los modelos utilizando métricas como la exactitud, la precisión, el F1-Score y matriz de correlación.

Visualizaciones estadísticas para mostrar la distribución de la población que compone el *dataset*.

6.3. Análisis y Discusión de resultados

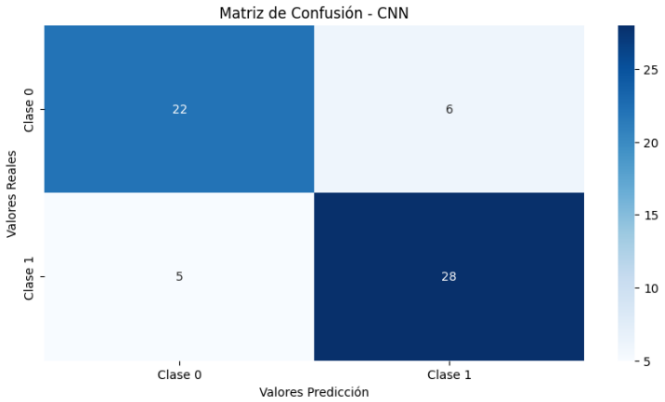
Una vez seleccionados los algoritmos y realizado el respectivo entrenamiento y validación, se realiza el siguiente análisis de resultados a partir de las métricas de desempeño.

- **Análisis matriz de confusión**

Teniendo en cuenta que la matriz de confusión es la principal fuente de datos para el cálculo de otras métricas como la exactitud y el *F1- Score*, en las figuras 2 a 6 se presentan las matrices de confusión de cada uno de los algoritmos implementados.

Figura 2

Matriz de confusión - CNN

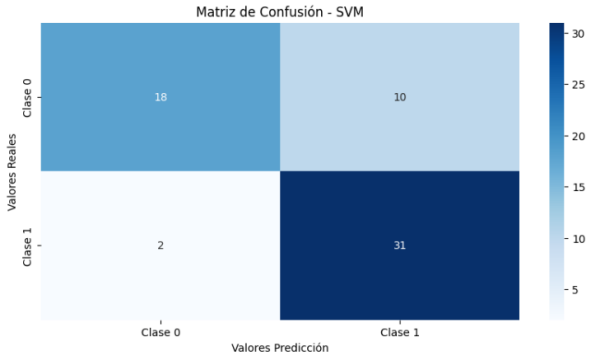


Nota. Fuente: Elaboración propia.

La matriz de la Figura 2, muestra que el modelo CNN tiene buen desempeño a nivel general, con una exactitud del 81% indica que clasifica de manera correcta la mayor parte de las muestras tanto en verdaderos positivos como verdaderos negativos. Se encuentra un leve sesgo con respecto a la clase 1, la cual corresponde a los pacientes que se encuentran en riesgo de sufrir un infarto de miocardio. Para la presente investigación, es recomendable la minimización de los falsos negativos.

Figura 3

Matriz de confusión - SVM

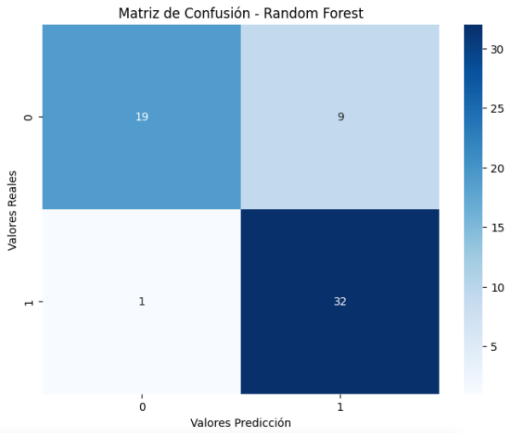


Nota. Fuente: Elaboración propia.

A partir de los resultados de la matriz de la Figura 3, se obtiene una exactitud del 80% indicando que es un modelo fiable en sus predicciones, un bajo número de falsos negativos es un buen indicador de resolución, ya que el no detectar o predecir un paciente en riesgo puede tener consecuencias graves. La métrica de F1 - Score es del 84% indicando que existe un buen equilibrio entre la sensibilidad y exactitud.

Figura 4

Matriz de confusión - Random Forest

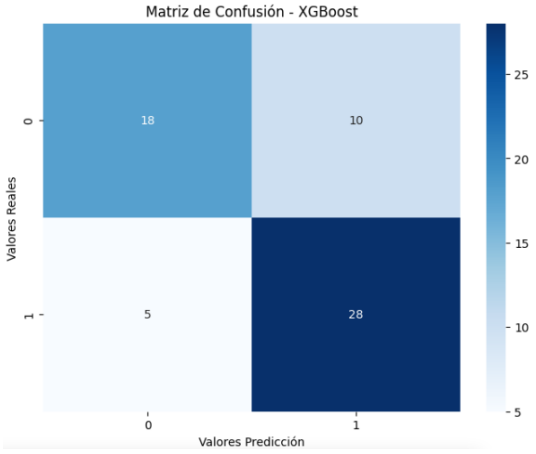


Nota. Fuente: Elaboración propia.

La Figura 4 muestra que el algoritmo *Random Forest* tuvo un mejor desempeño en comparación los demás que se implementaron, ya que la predicción de falsos negativos se redujo a 1. En el contexto de la presente investigación, para las predicciones de falsos positivos tiene menor criticidad, ya que estos resultados pueden servir de apoyo para la generación de alertas y ordenamiento de exámenes especializados.

Figura 5

Matriz de confusión - XGBoost

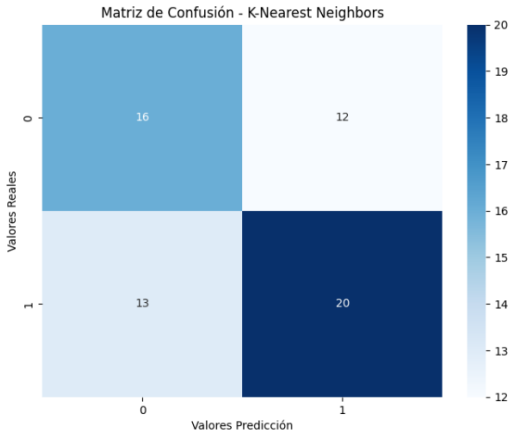


Nota. Fuente: Elaboración propia.

La Figura 5 correspondiente a la matriz de confusión del algoritmo XGBoost, muestra el F1 score de 75%; lo cual es relativamente bueno teniendo en cuenta que se detectaría la mayoría de los casos positivos. Sin embargo, hay un significativo número de pacientes (5) que el algoritmo predijo como falso negativo, lo cual no es tan bueno en el contexto de la investigación.

Figura 6

Matriz de confusión - K Nearest Neighbors



Nota. Fuente: Elaboración propia.

A partir de la matriz de la Figura 6, se puede deducir que este tipo de algoritmo no es tan recomendable para el contexto medico objeto de la investigación, ya que no tuvo unos resultados deseables, llegando a clasificar un alto número de falsos positivos (12) al igual que un significativo número de falsos negativos (13). Se encontró que para el F1 Score obtuvo un 59%. En la práctica esto se traduce en que un alto número de pacientes podría ser clasificados como negativos (44.9%), siendo que pueden ser positivos y un alto número de pacientes negativos (55.1%) serian clasificados como positivos sin cumplir esta condición.

- **Análisis de rendimiento algoritmos**

Se evaluaron cinco algoritmos de aprendizaje supervisado utilizando las métricas de exactitud (*Accuracy*), AUC-ROC y *F1-Score* para determinar su efectividad en el conjunto de datos analizado. En La tabla 2 se realiza la comparación de las métricas de desempeño obtenidas luego de realizar el entrenamiento de los algoritmos.

Tabla 2

Resumen de resultados

Algoritmo	Exactitud	AUC-ROC	F1 - Score
CNN	0.81	0.84	0.82
SVM	0.80	0.88	0.80
Random Forest	0.75	0.89	0.75
XGBoost	0.75	0.84	0.75
K-Nearest Neighbors	0.59	0.65	0.59

Nota. Fuente: Elaboración propia.

- CNN:

Mostró un buen desempeño general, sobresaliendo en *F1-Score*, lo que indica un balance adecuado entre precisión y sensibilidad.

- SVM:

Alcanzó el mayor valor de AUC-ROC (0.88), destacándose como el modelo más efectivo para distinguir entre clases, aunque con una precisión ligeramente menor que la CNN.

- *Random Forest*:

Este modelo tuvo un rendimiento sólido, destacándose en AUC-ROC (0.89), pero su precisión y F1-Score fueron inferiores a los de SVM y CNN.

- XGBoost:

Similar al *Random Forest*, XGBoost mostró un rendimiento equilibrado, pero sin superar a SVM o CNN en métricas clave.

- *K-Nearest Neighbors*:

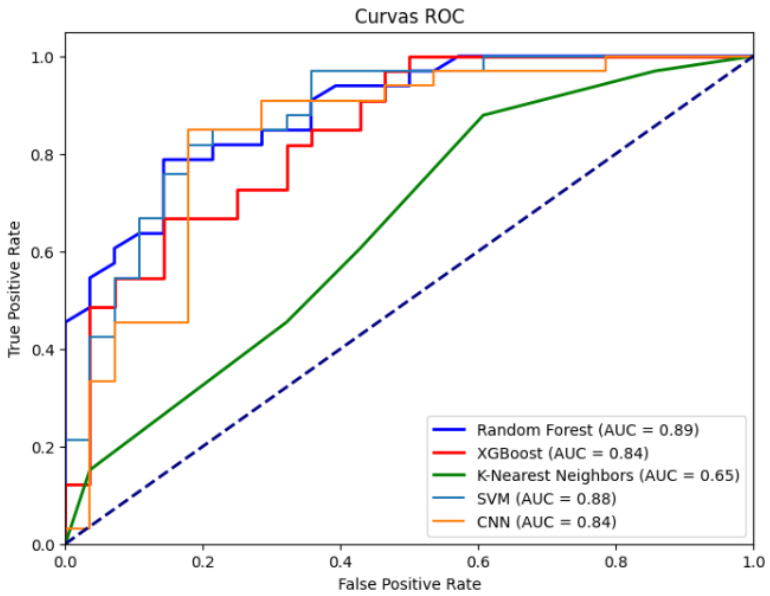
Fue el modelo menos efectivo en todas las métricas, probablemente debido a su sensibilidad a la distribución de los datos y la necesidad de mayor ajuste.

Random Forest mostró el mejor desempeño en términos de balance general entre precisión y F1-Score, aunque SVM destacó en la métrica de AUC-ROC, indicando una alta capacidad de clasificación, el KNN tuvo el desempeño más bajo, sugiriendo que puede no ser adecuado para este conjunto de datos sin optimización adicional.

En la Figura 7 se realiza la comparación de las curvas ROC obtenidas a partir de la implementación de los algoritmos seleccionados. Destaca el algoritmo *Random Forest* ya que abarca un área del 89%, siendo el más efectivo entre los modelos comparados. El algoritmo SVM con un área del 88% tiene un comportamiento similar al de *Random Forest*. Los algoritmos XGBoost y CNN abarcan un área del 84% si bien su capacidad predictiva no es tan alta como las de *Random Forest* y SVM son modelos sólidos y confiables. Finalmente, el modelo *K – Nearest Neighbors* es el que obtuvo menor desempeño entre los algoritmos implementados, lo cual se ve reflejado en su bajo porcentaje de área bajo la curva y los resultados obtenidos a partir de su matriz de confusión.

Figura 7

Comparación de curvas ROC



Nota. Fuente: Elaboración propia.

A partir de la matriz de correlación se obtienen las variables más influyentes en la predicción del riesgo de infarto de miocardio, la tabla 3 organiza los valores de correlación respecto a la variable objetivo de manera descendente. Se destaca que los diagnósticos asociados a un defecto reversible del tálamo (Thall = 2) tienen una fuerte influencia en el riesgo de infarto de miocardio; por tanto, este diagnóstico es crítico al momento de identificar pacientes que se encuentren en riesgo. La angina no inducida por el ejercicio (Exng = 0) es una forma silenciosa pero igualmente peligrosa de enfermedad cardíaca que el paciente no manifiesta bajo estrés físico, pero representa un alto riesgo. La frecuencia cardiaca máxima alcanzada (Thalachh) excesivamente altas puede ser un indicador de condiciones subyacentes peligrosas, tales como taquicardia.

Tabla 3

Correlación de la variable objetivo

Variables	Output
output	1.000.000
thall_2	0.527334
exng_0	0.436757
thalachh	0.421741
slp_2	0.394066
cp_2	0.316742
sex_0	0.280937
cp_1	0.245879
restecg_1	0.175322
cp_3	0.086957
fbs_0	0.028046
thall_0	-0.007293
fbs_1	-0.028046
slp_0	-0.063554
restecg_2	-0.068410
chol	-0.085239
thall_1	-0.106589
trtbps	-0.144931
restecg_0	-0.159775
age	-0.225439
sex_1	-0.280937
slp_1	-0.362053
caa	-0.391724
oldpeak	-0.423572
exng_1	-0.436757
thall_3	-0.486112
cp_0	-0.516015

Nota. Fuente: Elaboración propia.

Entre los cinco algoritmos evaluados, *Random Forest* destacó como el más efectivo, logrando la mayor área bajo la curva ROC del 89% y un buen equilibrio entre precisión y *F1-Score*. SVM también mostró un desempeño competitivo con un AUC-ROC del 88%, mientras que CNN demostró ser confiable con métricas sólidas. Por su parte, XGBoost y KNN obtuvieron un rendimiento inferior, siendo este último el menos adecuado para el conjunto de datos analizado.

Además, se identificaron variables clave que influyen en la predicción del riesgo de infarto de miocardio, como el defecto reversible del tálamo (Thall_2), la

angina no inducida por ejercicio (Exng_0) y la frecuencia cardíaca máxima alcanzada (Thalachh), resaltando su importancia para el diseño de modelos predictivos más precisos.

7. CONCLUSIONES

- El conjunto de datos clínicos seleccionado proporcionó una base sólida para el entrenamiento de los algoritmos de *Machine Learning*, esto permitió un análisis robusto y pertinente en la predicción del riesgo de infarto de miocardio
- Los resultados del entrenamiento de los algoritmos seleccionados evidenciaron que los algoritmos *Random Forest* y SVM presentaron un rendimiento superior de acuerdo con las métricas de desempeño seleccionadas. Mientras que algoritmos como KNN requieren optimización adicional mediante el cambio de hiperparámetros para mejorar su desempeño.

- El resultado obtenido con el criterio bajo la curva ROC concuerda con el análisis realizado a las matrices de confusión de cada algoritmo. Siendo los algoritmos *Random Forest* y SVM los que abarcaron una mayor área bajo la curva y presentaron un menor número de falsos negativos.

- De acuerdo con la tabla 3, se encontró que la variable con mayor influencia en la predicción de pacientes con riesgo cardiovascular es ***thall*** (Defecto reversible Tálamo) con un porcentaje de correlación de 52.73%, seguida por la variable exgn (angina no inducida por el ejercicio) con un 43.67% y la variable taclachh (frecuencia cardiaca máxima alcanzada) con un 42.17%.

- KNN evidenció el rendimiento más bajo en comparación a los otros algoritmos entrenados, por tanto, este algoritmo no es adecuado para este conjunto de datos en su estado actual. Sin una optimización adicional, como el ajuste de hiperparámetros, la selección de características, o el manejo de desequilibrios en los datos, KNN no logra capturar las complejidades necesarias para una predicción precisa en este contexto específico.

- Una cantidad alta de falsos negativos conlleva a que un paciente que se encuentran en riesgo de sufrir infarto de miocardio no sean correctamente identificados y no reciban el tratamiento preventivo adecuado.

8. LISTA DE REFERENCIAS

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*, *8*, 53. https://doi.org/10.1186/s40537-021-00444-8

Arribas Jara, F. (2018). *APRENDIZAJE NO-SUPERVISADO CON MODELOS GENERATIVOS PROFUNDOS* [UNIVERSIDAD AUTONOMA DE MADRID]. https://repositorio.uam.es/bitstream/handle/10486/688035/arribas_jara_fernando_tfg.pdf?form=MG0AV3

Bernal Vélez, M. J., Henao, I. M., Isabel, M., & Arango, I. (2022). *Predicción de enfermedades del corazón usando el algoritmo K-Nearest Neighbors*. https://www.researchgate.net/publication/364476395

Bouza, C. (2021). *LAS CURVAS ROC TEORÍA Y HERRAMIENTAS PARA SU USO* [Universidad de la Habana]. https://www.researchgate.net/publication/351991520_LAS_CURVAS_ROC_TEORIA_Y_HERRAMIENTAS_PARA_SU_USO

Chen, Z., Shi, J., Pommier, T., Cottin, Y., Salomon, M., Decourselle, T., Lalande, A., & Couturier, R. (2022). Prediction of Myocardial Infarction From Patient Features With Machine Learning. *Frontiers in Cardiovascular Medicine*, *9*. https://doi.org/10.3389/fcvm.2022.754609

Cruz Micán, E. O., Poveda Aguja, F. A., & Buitrago Márquez, L. M. (2020). Las TIC en el sector salud, machine learning para el diagnóstico y prevención de enfermedades. *Revista Quantica*, *1*(2), 1–32.

Cujar-Rosero, F., Investigador, E., Pinchao Ortiz, D., Timarán Pereira, R., & Guerrero Restrepo, M. (2021). *Thaqhaña: Un Motor de Búsqueda Inteligente Basado en Recursos Semánticos*. *1*. https://doi.org/10.18687/LACCEI2021.1.1.596

Danilo, M., Liliana, T., Mayra, A., & Gustavo, R. (s/f). *Un enfoque de Machine Learning en el desarrollo de Sistema Recomendadores para Procesos de Investigación*.

Dattoli García, C. A. (2021). Infarto agudo de miocardio: revisión sobre factores de riesgo, etiología, hallazgos angiográficos y desenlaces en pacientes jóvenes. *Archivos de cardiología de México*, *91*(4), 485–492. https://doi.org/10.24875/ACM.20000386

Díaz Delgado, D. (2022). Detección temprana de enfermedades del corazón mediante el aprendizaje automático. *Revista de investigación de Sistemas e Informática*, *15*(1), 33–42. https://doi.org/10.15381/risi.v15i1.23739

Forero Corba, W., & Bennasar, F. N. (2024). Técnicas y aplicaciones del Machine Learning e inteligencia artificial en educación : una revisión sistemática. *RIED. Revista iberoamericana de educación a distancia*, *27*(1), 209–253. https://doi.org/10.5944/RIED.27.1.37491

Goldman Lee. (2021, abril 15). *Goldman-Cecil. Tratado de medicina interna - Edición 26 - Edited by Lee Goldman, MD and Andrew I. Schafer, MD Elsevier*

Inspection Copies. https://inspectioncopy.elsevier.com/book/details/9788491137658

González Cedillo, C. D. (2019). Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian. *Ciencia y Tecnología*, *19*, 117–128.

Guzmán, F., & Arias, C. A. (2012). La historia clínica: elemento fundamental del acto médico. *Revista colombiana de cirugía*, *27*, 15–24.

Hajiarbabi, M. (2024). Heart disease detection using machine learning methods: a comprehensive narrative review. En *Journal of Medical Artificial Intelligence* (Vol. 7). AME Publishing Company. https://doi.org/10.21037/jmai-23-152

Krittanawong, C., Zhang, H. J., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. *Journal of the American College of Cardiology*, *69*(21), 2657–2664. https://doi.org/10.1016/J.JACC.2017.03.571

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. En *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering* (pp. 83–106). Elsevier. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

LEY 23 DE 1981, Pub. L. No. LEY 23 DE 1981 (1981). https://www.mineducacion.gov.co/1621/articles-103905_archivo_pdf.pdf

Llopis Sánchez, A. (2023). *Sistema Inteligente para la planificación y guiado de vehículos*. https://riunet.upv.es/handle/10251/199834

Maldonado Saavedra, O., Ramírez Sánchez, I., García Sánchez, J. R., Ceballos Reyes, G. M., & Méndez Bolaina, E. (2012). Colesterol: Función biológica e implicaciones médicas. *Revista Mexicana de Ciencias Farmacéuticas*, *43*(2), 7–22. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-01952012000200002

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/BIB/BBX044

Mitaritonna Alejandro. (2019, septiembre 26). *¿Cuál es la diferencia entre Inteligencia Artificial, Machine Learning y Deep Learning?* https://es.linkedin.com/pulse/cu%C3%A1l-es-la-diferencia-entre-inteligencia-artificial-y-mitaritonna

Montenegro Meza, M. A., Menchaca Méndez, R., & Menchaca Méndez, R. (2023). Una Introducción amable pero rigurosa al aprendizaje por refuerzo. *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica*, *12*, 1–15. https://www.redalyc.org/articulo.oa?id=512275598001

Morales Santos, Lojo Lendoiro, S., Rovira Cañellas, M., & Valdés Solís, P. (2024). La regulación legal de la inteligencia artificial en la Unión Europea: guía práctica para radiólogos. *Radiología*, *66*(5), 431–446. https://doi.org/10.1016/J.RX.2023.11.008

Moreno Sánchez, J. S. (2021). *Predicción de ataques cardiacos mediante técnicas de Machine Learning*. Universidad Politécnica de Madrid.

Morishita, T., Morio, G., Yamaguchi, A., & Sogawa, Y. (2023). Learning Deductive Reasoning from Synthetic Corpus based on Formal Logic. *Proceedings of the 40th International Conference on Machine Learning*, 25254–25274. https://arxiv.org/abs/2308.07336

Mosquera Rojas, G. E. (2020). *Clasificación de señales ECG para la detección de enfermedades cardíacas : un estudio comparativo*. Universidad de los Andes. http://hdl.handle.net/1992/51547

Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., Andreini, D., Budoff, M. J., Cademartiri, F., Callister, T. Q., Chang, H. J., Chinnaiyan, K., Chow, B. J. W., Cury, R. C., Delago, A., Gomez, M., Gransar, H., Hadamitzky, M., Hausleiter, J., … Slomka, P. J. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *European heart journal*, *38*(7), 500–507. https://doi.org/10.1093/EURHEARTJ/EHW188

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Nandal, N., Goel, L., & TANWAR, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000Research*, *11*, 1126. https://doi.org/10.12688/f1000research.123776.1

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Organización Mundial de la Salud. (1998). *Promoción de la Salud Glosario Organización Mundial de la Salud Ginebra*. https://iris.who.int/bitstream/handle/10665/67246/WHO_HPR_HEP_98.1_spa.pdf

Organización Mundial de la Salud. (2021, junio 11). *Enfermedades cardiovasculares*. OMS. https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Organización Mundial de la Salud. (2023, septiembre 19). *Hipertensión*. https://www.who.int/es/health-topics/hypertension#tab=tab_1

Patiño, D., Medina, J., Silva, R., Guijarro, A., & Rodríguez, J. (2023). Predicción de arritmias e infartos agudos de miocardio usando aprendizaje automático. *Ingenius*, *2023*(29), 79–89. https://doi.org/10.17163/ings.n29.2023.07

Pons, C. F., Recordón, A., & Ruiz Díaz, S. (2021). *Detección y clasificación de zero-day malware a través de data mining y machine learning*. Sociedad Argentina de Informática, SADIO. https://repositorio.uai.edu.ar/handle/123456789/535

Quirón Salud. (2022, octubre 14). *Prevención del Infarto agudo de miocardio*. https://www.quironsalud.com/es/comunicacion/actualidad/infarto-agudo-de-miocardio-ataque-de-corazon-infarto-dolor

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358.

https://doi.org/10.1056/NEJMRA1814259/SUPPL_FILE/NEJMRA1814259_DISCLOSURES.PDF

Ranya N. Sweis. (2024, febrero). *Infarto agudo de miocardio - Trastornos cardiovasculares - Manual MSD versión para profesionales*. https://www.msdmanuals.com/es/professional/trastornos-cardiovasculares/enfermedad-coronaria/infarto-agudo-de-miocardio

Real Academia Española. (2014a). *Diccionario de la lengua española* (Ed. 23.7 en linea). https://dle.rae.es/edad

Real Academia Española. (2014b). *Diccionario de la lengua española* (Ed. 23.7 en linea). https://dle.rae.es/sexo

Riveros, A., Cortazar, C., Alcazar, F., & Sánchez, J. J. (2005). Efectos de una intervención cognitivo-conductual en la calidad de vida, ansiedad, depresión y condición médica de pacientes diabéticos e hipertensos esenciales. *International journal of clinical and health psychology*, *5(3)*, 445–447. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.redalyc.org/pdf/337/33705302.pdf

Rodrigo, J. A. (2020, octubre). *Random Forest python*. https://cienciadedatos.net/documentos/py08_random_forest_python.html

Rokach, L., & Maimon, O. (2005). Decision Trees. *Data Mining and Knowledge Discovery Handbook*, 165–192. https://doi.org/10.1007/0-387-25465-X_9

Ruqiya, M. T., Yaqoob, A. M., Khan, A. A., Shaikh, A. M., & Khan, N. (2023). Review on Cleveland Heart Disease Dataset using Machine Learning. *Quaid-e-Awam University Research Journal of Engineering, Science & Technology*, *21*(1), 87–98. https://doi.org/10.52584/qrj.2101.11

Sarker, I. H. (2021a). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 1–20. https://doi.org/10.1007/S42979-021-00815-1/FIGURES/11

Sarker, I. H. (2021b). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Sarker, I. H., Kayes, A. S. M., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0219-y

Serna-Trejos, J. S., Agudelo-Quintero, E., & Bermúdez-Moyano, S. G. (2022). Machine Learning en ciencias de la salud: Usos y aplicaciones: Machine Learning in health sciences: Uses and applications. *Peruvian Journal of Health Care and Global Health*, *6*(2), 95–96. https://doi.org/10.22258/hgh.2022.62.119

Srinivasan, S., Gunasekaran, S., Kumar Mathivanan, S., Anbu Malar B, B. M., Jayagopal, P., & Teshite Dalu, G. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Scientific Reports* |, *13*, 13588. https://doi.org/10.1038/s41598-023-40717-1

Sun, Y. (2022). Heart disease prediction using enhanced machine learning techniques. *Intelligent Systems and Machine Learning for Industry*, 93–114. https://doi.org/10.1201/9781003286745-5

Tabima Luque, G. A. (2024). *Uso de inteligencia artificial en la detección de estafas por correo electrónico para prevención, protección y apoyo al adulto mayor.* https://repositorio.uniandes.edu.co/entities/publication/059eb73b-ec3a-4e7e-b3cd-018b7cbda680

Torres Jordi. (2021, abril 14). *Introducción al aprendizaje por refuerzo profundo: Teoría y práctica en Python (Spanish Edition): Torres, Jordi: 9798599775416: Amazon.com: Books*. https://www.amazon.com/Introducci%C3%B3n-aprendizaje-por-refuerzo-profundo/dp/B092L5XBQ9/ref=sr_1_1?dib=eyJ2IjoiMSJ9.h5FMKwOtmUESqsxoG68SeA.pYZA9u_vtYN7HcenzIF6GnjT92kuuz0KUYWUstfd1ds&dib_tag=se&keywords=9798599775416&linkCode=qs&qid=1731471686&s=books&sr=1-1

Valenzuela González, G. (2022). *Aprendizaje Supervisado: Métodos, Propiedades y Aplicaciones* [Universidad de Málaga]. https://riuma.uma.es/xmlui/bitstream/handle/10630/25147/TFG_Aprendizaje_Supervisado_GVG.pdf?sequence=4&form=MG0AV3

Vázquez Pérez, J. J., Cervacio Beas, O. N., de Luna Velasco, L. E., & García Ortiz, L. (2023). Frecuencia cardiaca: una revisión sistemática. *Publicación Científica de la Asociación Española en Enfermería en Cardiología*, *XXX*(90), 71–86. https://doi.org/10.59322/90.7186.lr5

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/JOURNAL.PONE.0174944

Apéndice A.

Dataset para el entrenamiento de algoritmos de *Machine Learning*

El *dataset* escogido para el entrenamiento de los algoritmos seleccionados, se encuentra disponible en el siguiente enlace.

https://www.kaggle.com/code/kanncaa1/heart-attack-analysis-prediction/input

Apéndice B.

Implementación de algoritmos de *Machine Learning*

El cuaderno adjunto contiene el código correspondiente a la implementación de los algoritmos seleccionados para la predicción del riesgo de infarto de miocardio.

Apéndice_B_Prediccion_infarto.html