



# Modelo de Red Neuronal para la predicción de Ataques Cardíacos

---

**Integrantes:**

Glendi Michelle Cueto Sunun  
Emiliano Figueroa Monroy  
Raymundo Sánchez Estrada  
Mildred Nadxhielly Vega De Paz

**Profesores:**

Dr. Mauro Delboy Céspedes  
Dr. Rosa María Woo García

## Resumen

Actualmente es esencial la detección temprana de ataques cardíacos ya que permite mejorar la prevención y tratamiento de las enfermedades. En este proyecto, se desarrolla un modelo fundamentado en redes neuronales para la categorización del riesgo de infarto cardíaco a partir de datos clínicos, con un potencial uso en sistemas de cuidado primario en la salud. El procedimiento abarca desde la recopilación y preprocesamiento de datos, normalización de las variables y la puesta en marcha de una arquitectura de aprendizaje profundo. Se entrenó y evaluó la red neuronal a través de técnicas de validación cruzada y métricas de desempeño como la exactitud, precisión, puntuación F, recuperación y la curva ROC.

Con base en la comparación realizada con otros modelos predictivos utilizados que fueron, “Random Forest Classifier”, “Gaussian Naive Bayes” (*GaussianNB*) y Regresión Logística, los hallazgos obtenidos evidencian el gran potencial predictivo del modelo de red neuronal binaria para la detección de pacientes en riesgo, ofreciendo un instrumento valioso para la toma de decisiones médicas y fortaleciendo la atención médica. En conclusión, la implementación y uso de aplicaciones de inteligencia artificial en la predicción de enfermedades cardiovasculares puede contribuir de manera significativa la detección temprana, disminución de la mortalidad y mejora de los recursos preventivos en la salud. Este enfoque define un marco escalable que se puede ajustar a diferentes grupos demográficos y sistemas de registros clínicos electrónicos.

**Palabras clave:** Predicción de ataques cardíacos, redes neuronales, inteligencia artificial, aprendizaje profundo, datos clínicos.

**Abstract**

For effective preventative measures and medical interventions for cardiovascular illnesses, early diagnosis of cardiac events is essential. With potential uses in primary healthcare settings, this study presents a neural network-based model for myocardial infarction risk stratification based on clinical data. Collecting data, preprocessing, variable standardization, and the deployment of a deep learning architecture are all part of the methodology. Cross-validation methods and performance measurements such as accuracy, precision, F-score, recall, and ROC-AUC analysis were used to train and assess the neural network.

The binary neural network shows higher predicting ability for identifying at-risk patients when compared to well-known predictive models like Random Forest Classifier, Gaussian Naive Bayes (GaussianNB), and Logistic Regression. This approach improves proactive healthcare delivery by providing a strong decision-support tool for clinical practice. In general, there is an enormous opportunity for enhancing early diagnosis, lowering death rates, and making the most of preventive healthcare resources by the incorporation of artificial intelligence applications in cardiovascular disease prediction. The recommended structure opens the door for data-driven precision medicine with a scalable solution that can be modified to other demographic groups and electronic health record systems.

**Keywords:** Heart attack prediction, neural networks, artificial intelligence, deep learning, clinical data.

## **Introducción**

Según el informe de la Asociación Estadounidense del Corazón, se deja claro que las enfermedades cardíacas siguen siendo la principal causa de muerte en todo el mundo. “Actualmente hay alrededor de 640 millones de personas que viven con enfermedades cardíacas y circulatorias a lo largo y ancho del planeta, una cifra que se ha duplicado desde 1993” [12].

Lamentablemente hoy en día este número ha ido aumentando debido a los cambios en los estilos de vida, el envejecimiento y el crecimiento de la población mundial. “Y su consecuencia: las enfermedades cardiovasculares son responsables de casi 20 millones de muertes al año (1 de cada 3 defunciones)” [12].

Las afecciones cardiovasculares son una de las principales causas de mortalidad en todo el mundo, hoy en día siendo los ataques al corazón una de las manifestaciones más críticas. El acceso a atención médica enfocada a esta problemática puede ser de acceso limitado en algunas regiones del mundo e incluso de nuestro país, lo que dificulta la identificación temprana de esta afección y la prevención de estos mismos. El aumento de la disponibilidad de datos clínicos, la llegada y evolución de la inteligencia artificial han abierto nuevas oportunidades para implementar nuevas herramientas con enfoque predictivo que faciliten la identificación de personas en riesgo, posibilitando intervenciones a tiempo y reduciendo la carga en los sistemas de salud.

## **Justificación**

El desarrollo de un modelo de predicción de ataques cardíacos basado en redes neuronales es una solución innovadora que puede mejorar significativamente la atención médica. Implementar este tipo de herramientas permite optimizar la detección temprana de pacientes en riesgo, contribuyendo a reducir la tasa de mortalidad y mejorando la eficiencia en la asignación de recursos de salud. Además, la integración de inteligencia artificial en la medicina fortalece la toma de decisiones clínicas y facilita el acceso a diagnósticos más precisos, especialmente en comunidades con escasez de especialistas.

## **Objetivos**

### **Objetivo General**

Desarrollar e implementar un modelo de red neuronal para la predicción temprana de ataques cardíacos, utilizando datos clínicos de pacientes, con el fin de mejorar las estrategias de prevención y tratamiento oportuno de enfermedades cardiovasculares.

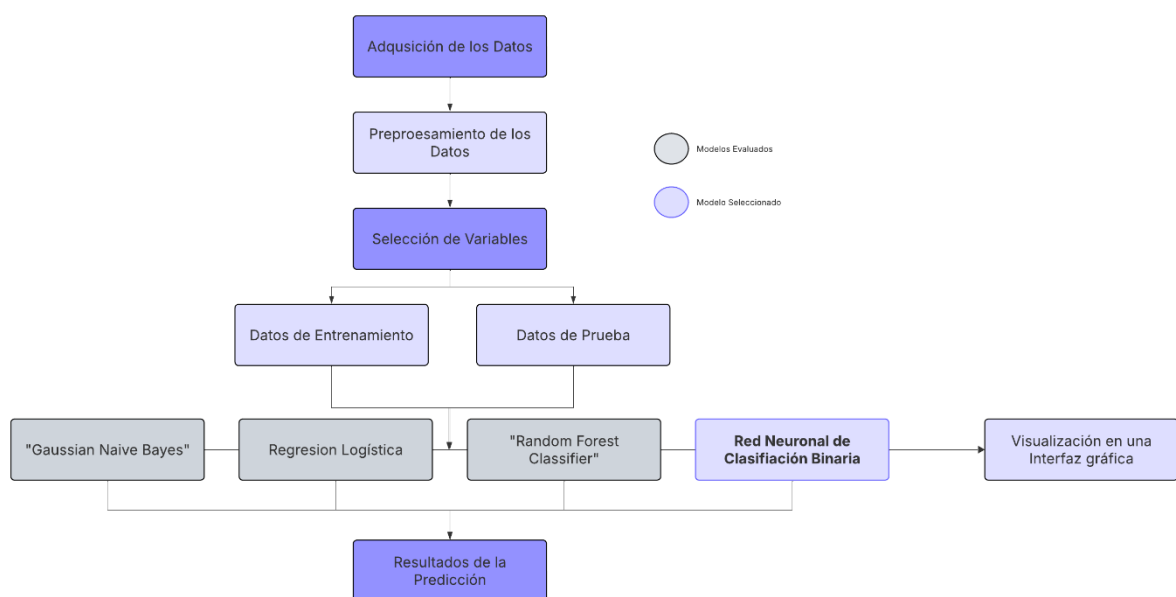
### ***Objetivos Específicos***

1. Recopilar y preprocesar un conjunto de datos clínicos relevantes relacionados con enfermedades cardiovasculares.
2. Diseñar una arquitectura de red neuronal óptima que permita identificar patrones y relaciones complejas entre los distintos factores de riesgo cardiovascular y la ocurrencia de ataques cardíacos.
3. Entrenar y validar el modelo utilizando técnicas de aprendizaje profundo para maximizar su precisión y sensibilidad en la predicción de eventos cardíacos.
4. Implementar un sistema de evaluación comparativa entre el modelo propuesto y

otros algoritmos de predicción desarrollados para cuantificar las mejoras en rendimiento.

5. Desarrollar una interfaz de usuario que facilite la interpretación de los resultados predictivos.

*Figure 1 - Diagrama de Flujo de los Modelos Utilizados*



*Fuente: Elaboración propia.*

En la Figura 1 se analiza el proceso realizado para el proceso de selección y comparación de los modelos utilizados para la predicción de ataques cardíacos y la finalidad de este que es la visualización de sus resultados y uso dentro de una interfaz gráfica amigable para el usuario.

## **Alcance y Limitaciones**

Este proyecto se enfoca en el desarrollo de un modelo de inteligencia artificial capaz de predecir ataques cardíacos en pacientes, utilizando datos clínicos. Se contempla la implementación de técnicas avanzadas de aprendizaje profundo para mejorar la precisión del modelo y su aplicabilidad en entornos médicos.

Sin embargo, existen algunas limitaciones a considerar. La calidad y disponibilidad de datos pueden influir en la efectividad del modelo, y su adopción en el sector salud dependerá de la integración con los sistemas médicos existentes. Además, aunque el modelo puede servir como herramienta de apoyo en la toma de decisiones clínicas, no reemplaza el diagnóstico de un especialista. A futuro, se podrían explorar mejoras en la recopilación de datos, la implementación en hospitales y el desarrollo de interfaces accesibles para profesionales de la salud.

## **Estado del arte**

Las enfermedades cardiovasculares constituyen una causa principal de mortalidad a nivel mundial, siendo el infarto agudo de miocardio una de las presentaciones clínicas más críticas. El uso de modelos de redes neuronales de clasificación binaria para su predicción ha cobrado importancia creciente en la literatura científica contemporánea, dada su capacidad para representar relaciones no lineales complejas en datos clínicos estructurados. A pesar de los avances en modelos predictivos, la validación en contextos clínicos reales y su integración práctica a procesos hospitalarios siguen siendo temas poco abordados de manera sistemática.

Diversos estudios han empleado arquitecturas de redes neuronales profundas como MLP (Multilayer Perceptron), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Units) y modelos híbridos combinando múltiples tipos de capas neuronales. Por ejemplo, Dritsas y

Trigka (2024) comparan seis arquitecturas diferentes —incluyendo MLP, CNN, RNN, LSTM, GRU y un modelo híbrido— aplicadas a un conjunto de datos de predicción de infartos, realizando el desempeño del modelo híbrido y utilizando SHAP (Shapley Additive Explanations) para mejorar la interpretabilidad clínica de los resultados [1].

Por su parte, Nannapaneni et al. (2023) proponen un modelo híbrido basado en CNN, LSTM y GRU que alcanza una precisión del 93,3 % usando el conjunto de datos Cleveland tras aplicar técnicas de normalización y regularización [8]. Estas arquitecturas buscan mejorar el rendimiento en contextos donde los datos clínicos, además de estructurados, presentan relaciones secuenciales o temporales.

El preprocesamiento de datos ha sido una preocupación central, dado que los conjuntos de datos clínicos presentan con frecuencia características como valores faltantes, clases desbalanceadas y ruido estructural. En este sentido, varios autores han aplicado técnicas de balanceo como SMOTE (Synthetic Minority Over-sampling Technique) para abordar la desproporción entre clases que caracteriza los registros clínicos donde los eventos positivos (infartos) suelen ser mucho menos frecuentes que los negativos. Akter et al. (2021) y Krishnan et al. (2021) utilizan esta técnica para mejorar el entrenamiento de modelos como Random Forest y arquitecturas basadas en RNN y GRU, alcanzando altos niveles de precisión (96 % y 98,68 % respectivamente) [3,10].

En cuanto al proceso de validación, la mayoría de los estudios revisados aplican validaciones retrospectivas sobre conjuntos de datos estandarizados, tales como Cleveland, UCI o Kaggle. Estas bases de datos, aunque útiles como benchmarks iniciales, presentan limitaciones sustanciales en cuanto a representación de población y diversidad clínica. Solo una minoría de investigaciones aborda explícitamente la validación en entornos reales. Marqas



et al. (2023) desarrollan un modelo de aprendizaje automático con datos obtenidos de pacientes reales, incluyendo antecedentes médicos, estilo de vida y demografía recopilada de entornos hospitalarios.

Aunque sus resultados son prometedores, los autores reconocen que se requiere una validación externa más amplia para garantizar la generalización del modelo a otras poblaciones [4].

Un aspecto crítico para la integración efectiva de estos sistemas al entorno clínico es su interpretabilidad. En este sentido, los métodos de explicabilidad como SHAP se presentan como herramientas fundamentales para facilitar la comprensión por parte del personal médico e incrementar la confianza en los resultados del modelo. Dritsas y Trigka (2024) emplean SHAP para traducir las salidas del modelo híbrido en variables clínicas comprensibles, como perfiles lipídicos elevados o niveles de presión arterial, mostrando la potencial aplicabilidad del sistema en decisiones médicas [1]. De forma similar, Zhang et al. (2021) integran selección de características embebida mediante métodos de LinearSVC con regularización L1 para reducir dimensionalidad y mejorar tanto la interpretación clínica como la eficiencia computacional del modelo [9].

Sin embargo, la integración operativa de estos modelos en flujos de trabajo reales de la práctica médica se encuentra aún poco desarrollada. Algunos estudios, como los de Almazroi et al. (2023), proponen sistemas de apoyo a decisiones clínicas (CDSS) alimentados por redes neuronales densas, los cuales podrían integrarse a sistemas de historias clínicas electrónicas (EHR). No obstante, la literatura evidencia una carencia importante de intervenciones que consideren tiempos de inferencia, compatibilidad con hardware clínico o cumplimiento con normativas legales como HIPAA o GDPR [6].

Un patrón recurrente preocupante es que muchas publicaciones priorizan métricas como precisión o área bajo la curva ROC (AUC), mientras que métricas de mayor relevancia clínica como la sensibilidad (prevenir falsos negativos) son menos reportadas.

Esto reduce la efectividad práctica de los modelos, dado que en contextos clínicos una alta precisión general puede resultar insuficiente si los casos de infarto no son detectados con la debida sensibilidad. Aunque métricas como precisión, recall y F1-score son frecuentemente utilizadas, su utilización debe ajustarse a los requerimientos de priorización del riesgo clínico.

En conclusión, la literatura reciente demuestra avances importantes en la construcción de modelos de redes neuronales profundas para la predicción binaria de infartos cardíacos. Se han consolidado múltiples arquitecturas y técnicas de preprocesamiento, logrando altos niveles de desempeño en ambientes controlados. No obstante, el campo presenta vacíos importantes en términos de validación externa robusta, integración operativa en sistemas hospitalarios reales y cumplimiento normativo. La adopción de herramientas de explicabilidad como SHAP, combinada con nuevos esfuerzos en validación prospectiva y adaptación institucional, será clave para traducir estos modelos en soluciones clínicas efectivas.

## **Metodología y Diseño del Proyecto**

### **Datos y Preprocesamiento**

#### ***1. Descripción de los Datos***

Los datos analizados son estructurados ya que toda su información encuentra organizada en un mismo formato (numérico), además de que facilitan su comprensión y lectura por la forma en la cual estos se presentan.

*Tabla 1 - Descripción de las variables del DataSet*

No.	Atributos	Tipo de variable	Descripción de las Variables	Valor
1	Edad (Age)	Variable Continua	Rango de Edad de los pacientes.	Múltiples valores en el rango de 29 y 77 (edad en años)
2	Sex	Variable Discreta	Hombre o Mujer representados por su edad.	1: Hombre 0: Mujer
3	CP (Chest Pain) (Tipo de Dolor en el Pecho)	Variable Discreta	Representa el tipo de dolor en el pecho: Angina Típica, Angina Atípica, Dolor No Anginoso, Asintomático.	0: Angina Típica 1: Angina Atípica 2: Dolor No Anginoso 3: Asintomático
4	Presión Arterial en Reposo (Trestbps)	Variable Continua	Representa la Frecuencia Cardíaca en Reposo. (en	Valor continuo múltiple en mm/Hg

			mm/Hg).	
5	Nivel de Colesterol (Chol)	Variable Continua	Nivel de colesterol en mg/dl mediante el sensor de IMC.	Valores continuos múltiples en mg/dl
6	Glucosa en Ayunas (FBS)	Variable Discreta	Representa el nivel de azúcar en la sangre del paciente en ayunas.	0: false (Fbs < 120 mg/dl)  1: true (fbs > 120 mg/dl)
7	Resultados del Electrocardiograma (Restecg)	Variable Discreta	Representa el resultado de ECG, cada número entero representa un nivel de dolor.	0: normal  1: anomalía de onda ST-T (inversión de la onda T y/o elevación o depresión del segmento ST > 0.05 mV)  2: con hipertrofia ventricular izquierda

				probable o confirmada conforme los criterios de “Estes”
8	Frecuencia Cardíaca Máxima alcanzada (Thalach)	Variable Continua	Representa la frecuencia cardíaca máxima del paciente.	Múltiples valores de 71 a 22  Bajo: debajo de los 50 latidos/minuto  Normal: 51 -119 latidos/minuto  Alto: 120 -180 latidos/minuto
9	Angina Inducida por el ejercicio (Exang)	Variable Discreta	Determinar si existe o no angina inducida por ejercicio.	1: Si  0: No
10	Oldpeak	Variable Discreta	Muestra cómo se compara la depresión ST inducida por el	Múltiples valores de número decimales entre 0

			ejercicio con el reposo.	y 6.2
11	Slope	Variable Discreta	Describe el estado del paciente en el momento álgido del ejercicio.	0: pendiente ascendente  1: plano  2: pendiente descendiente
12	Vasos Sanguíneos Principales  (Ca)	Variable Discreta	Cantidad de vasos principales que la fluoroscopia puede colorear.	Numero de vasos principales (0 - 3) coloreados por fluoroscopia → valor de (0,1,2,3)
13	Prueba de Esfuerzo con Talio  (Thal)	Variable Discreta	Pacientes con dolor en el pecho se les hace una prueba de esfuerzo con Talio.  Un trastorno sanguíneo llamado	0: No evaluado  1: Flujo sanguíneo normal  2: Defecto Fijo  3: Defecto Reversible

			talasemia	
14	Variable Objetivo (target)	Variable Discreta	La variable objetivo hace uso de los 13 parámetros del dataset para la generación de resultados de predicción con 2 clases.	0: sin enfermedad  1: enfermedad

*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

## 2. Fuente de Datos

Se hace uso del *Cleveland Heart Disease Dataset* [15], que se encuentra disponible en el repositorio UCI Machine Learning; este dataset contiene registros clínicos de 1025 pacientes 526 con riesgo de ataque cardiaco y 499 sin riesgo de ataque cardiaco.

Se consideran 14 características clínicas: edad, sexo, tipo de dolor torácico, presión arterial en reposo (mm Hg), colesterol sérico (mg/dl), glucemia en ayunas, entre otras, como se logra apreciar en la Tabla 1. [16]

Este conjunto de datos data de 2019 y consta de cuatro bases de datos: Cleveland, Hungría, Suiza y Long Beach V y se cuenta con las licencias necesarias para el uso de sus datos. (*Ilustración 2 y 3*).

### ***3. Procesamiento de los Datos: Técnicas de limpieza, normalización, transformación y análisis exploratorio***

Se hace la lectura del Data Set haciendo uso de la librería de Pandas. (ref. imagen), donde posteriormente se hace el análisis de los valores faltante que se encuentren dentro del dataset.

#### ***Análisis de Características Categóricas***

Durante esta fase del preprocesamiento, se llevó a cabo un análisis exploratorio de las variables categóricas del dataset seleccionado, para comprender su distribución y relación con la variable objetivo que representa la presencia o ausencia de enfermedad cardíaca.

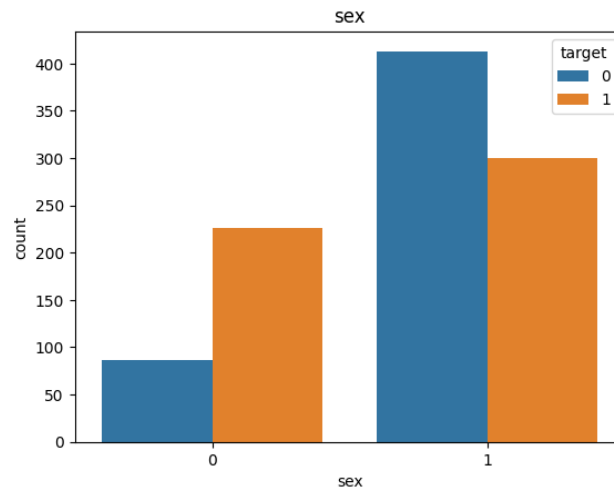
Para la visualización de estas relaciones, se elaboraron gráficos de conteo para cada variable categórica seccionados por la variable objetivo (*target*), donde 0 representa la ausencia de enfermedad cardíaca y 1 indica la presencia de esta. Este análisis se implementó mediante el código (referencia de la *imagen del código*) del apartado de anexos.

Los gráficos obtenidos permitieron la identificación de patrones importantes en varias características categóricas.

La Figura 2 permite la observación del comportamiento de la variable “target” con base en el sexo de los pacientes, donde se aprecia una mayor incidencia de enfermedad cardíaca en los pacientes masculinos (*marcados como 1*), mientras que en el grupo femenino (*marcados como 0*) las dimensiones de casos positivos supera a los negativos.



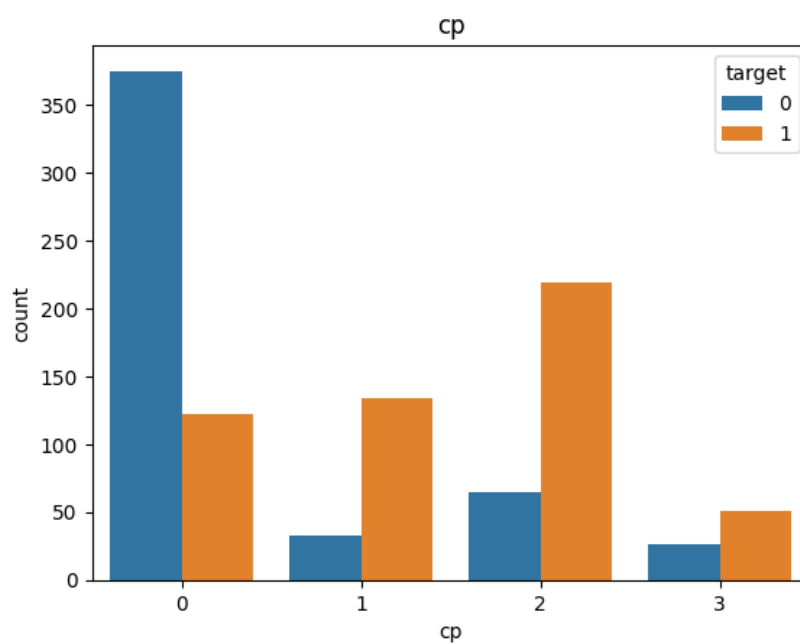
Figure 2- Distribución de la variable "sex" en relación con la variable "target"



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

La Figura 3 representa la distribución de la variable “target” a partir de la variable “cp” (*Tipo de dolor en el pecho*), se observa que la categoría 0 (*Angina Típica*) tiene una relevante asociación con ausencia de enfermedad cardíaca, mientras que las categorías 1 (*Angina Atípica*), 2 (*Dolor no anginoso*) y 3 (*Asintomático*) presentan una mayor proporción de casos positivos, siendo especialmente notoria esta tendencia en la categoría 2.

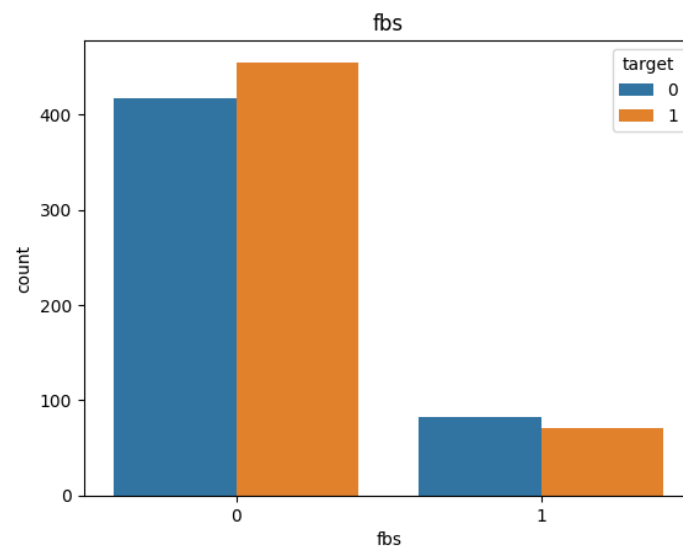
Figure 3 - Distribución de la variable "cp" en relación con la variable "target"



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

La Figura 4 muestra la distribución de la “Glucemia en Ayunas” (fbs), en esta sección el comportamiento de la variable es similar para ambos valores de la variable objetivo, se sugiere que este factor puede poseer menor poder predictivo que otros.

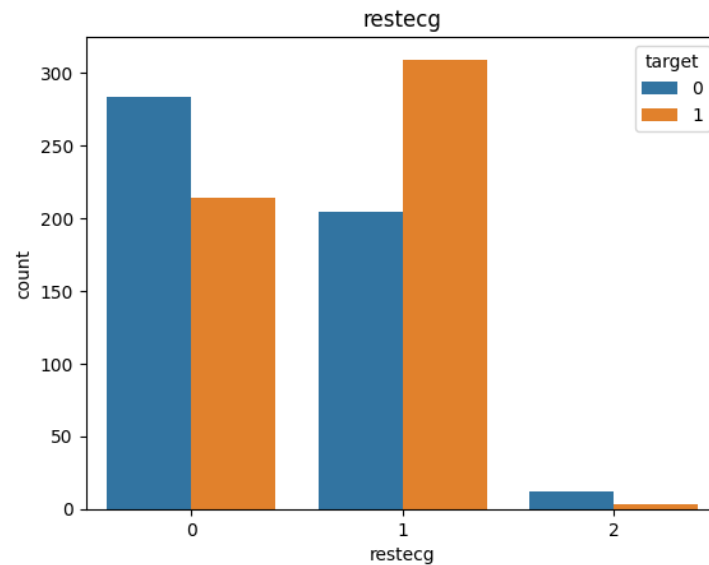
*Figure 4 - Distribución de la variable "fbs" en relación con la variable "target"*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

En la figura 5 se aprecia el comportamiento de la variable “restecg” (Resultados del electrocardiograma), en la categoría 1 se observa una mayor proporción de casos positivos de enfermedad cardíaca, mientras que la categoría 0 representa más casos negativos.

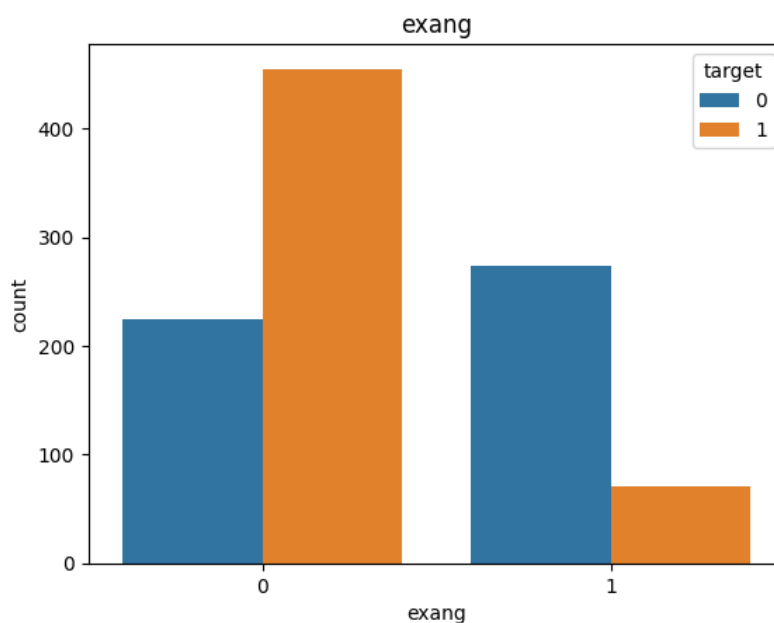
Figure 5 - Distribución de la variable "restecg" en relación con la variable "target"



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

El gráfico de la Figura 6 muestra la distribución de la variable “exang” (Angina inducida por ejercicio), esta presenta una proporción significativamente mayor de enfermedad cardíaca, mientras que aquellos con valor 1 muestran predominantemente ausencia de la enfermedad.

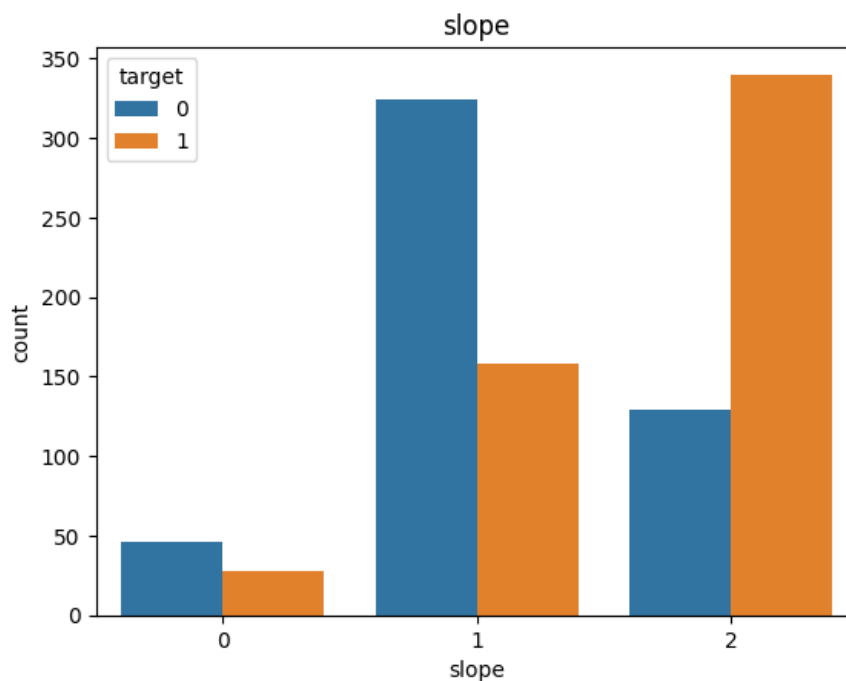
Figure 6 - Distribución de la variable "exang" en relación con la variable "target"



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

La Figura 7 representa como la variable “slope” (Pendiente del segmento ST durante el ejercicio) se encuentra fuertemente ligada con la presencia de enfermedad cardíaca, mientras que la distribución 1 muestra una mayor proporción de casos negativos.

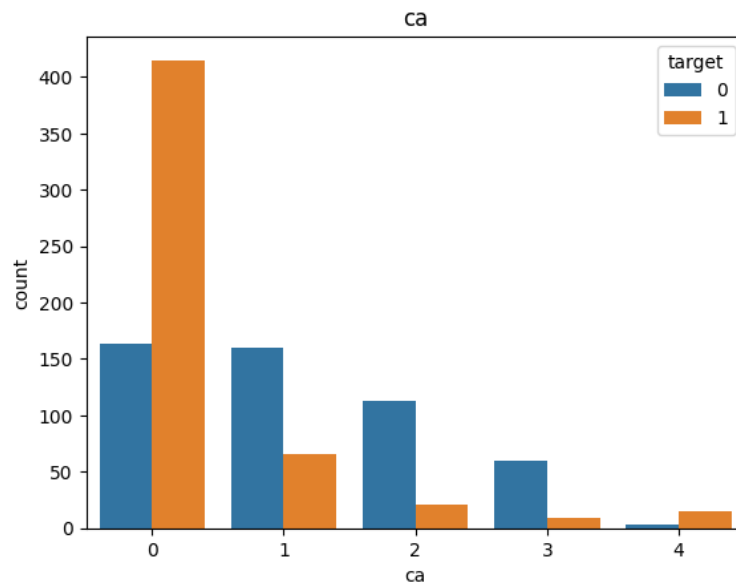
*Figure 7 - Distribución de la variable "slope" en relación con la variable "target"*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

De igual forma en la Figura 8 se observa una clara tendencia donde la variable “ca” (número de vasos afectados por fluoroscopia) (valores 1 - 4), muestra que, a mayor número de vasos afectados, menor es la proporción de casos positivos de enfermedad cardíaca. El resultado 0 muestra una elevada asociación con la presencia de esta enfermedad.

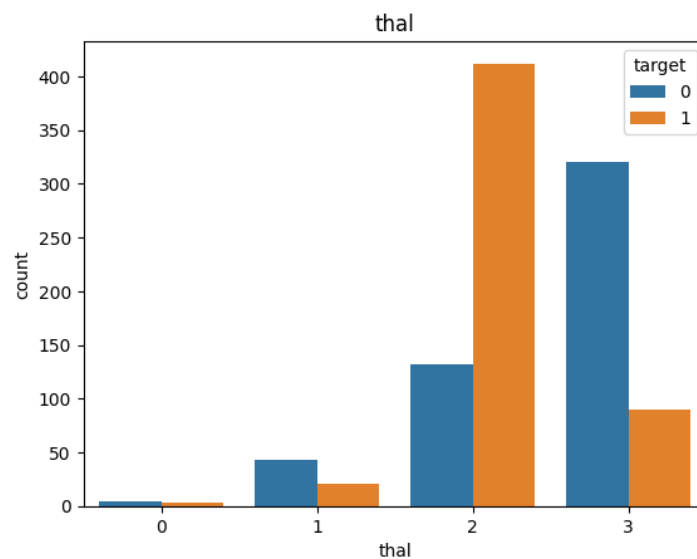
Figure 8 - Distribución de la variable "ca" en relación con la variable "target"



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

En la distribución de la cambiante “thal” (trastorno sanguíneo llamado talasemia) como se logra apreciar en la Figura 9, se observa que la categoría 2 se encuentra ligada con presencia de enfermedad, mientras que la tercera muestra una proporción de casos sin enfermedad.

Figure 9 - Distribución de la variable "thal" en relación con la variable "target"



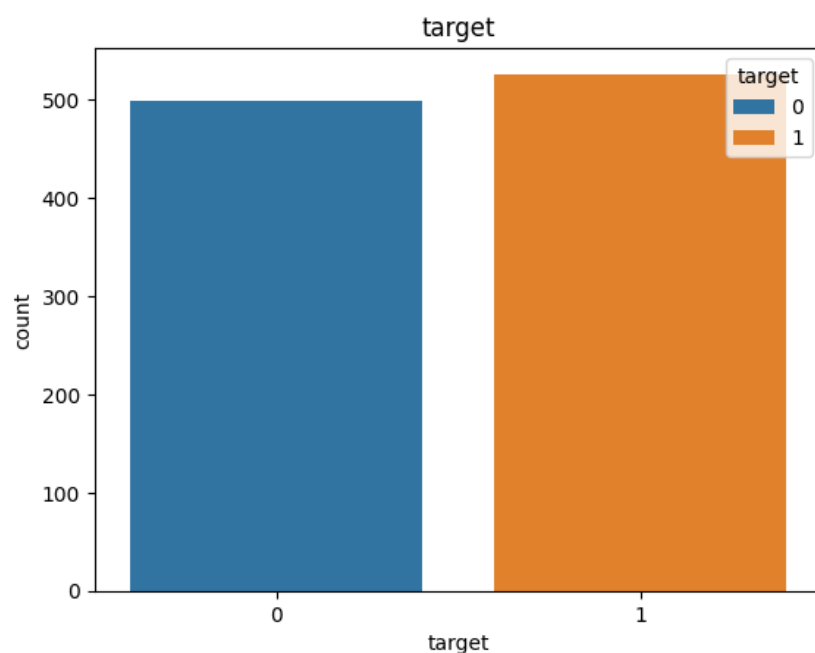
Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

El conjunto de datos muestra un balance adecuado entre casos positivos y negativos de la variable objetivo como se aprecia en la Figura 10, esto es beneficioso para el entrenamiento del modelo de red neuronal, ya que se reduce el riesgo de sesgo hacia una clase predominante.

Este análisis proporciona una comprensión más clara y concisa de la distribución y comportamiento de las variables categóricas, de igual forma apoyó en la identificación de las características con mayor potencial predictivo para el modelo de Red Neuronal.

Variables como “exang”, “ca”, “thal” y “slope” se muestran asociadas particularmente fuertes con la variable objetivo, lo que sugiere una gran importancia como variables predictoras dentro del modelo.

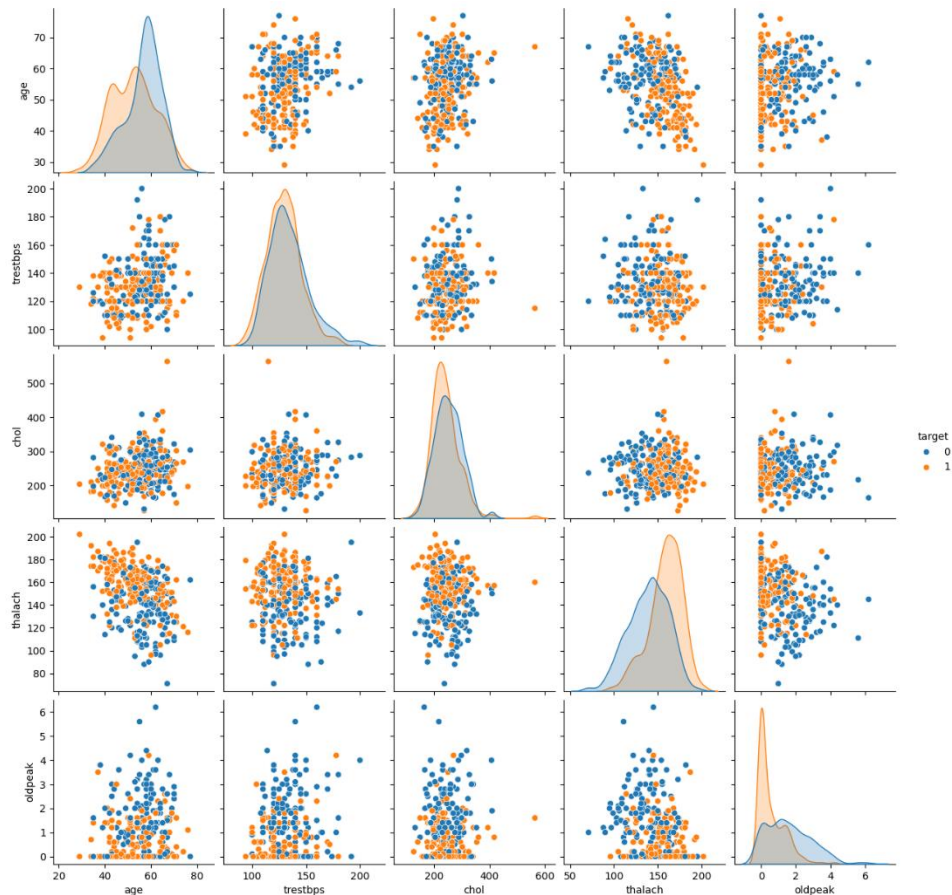
*Figure 10 - - Distribución de la variable Target*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

## Análisis de Características Numéricas.

Figure 11 - Diagrama de dispersión en relación con la variable "target"



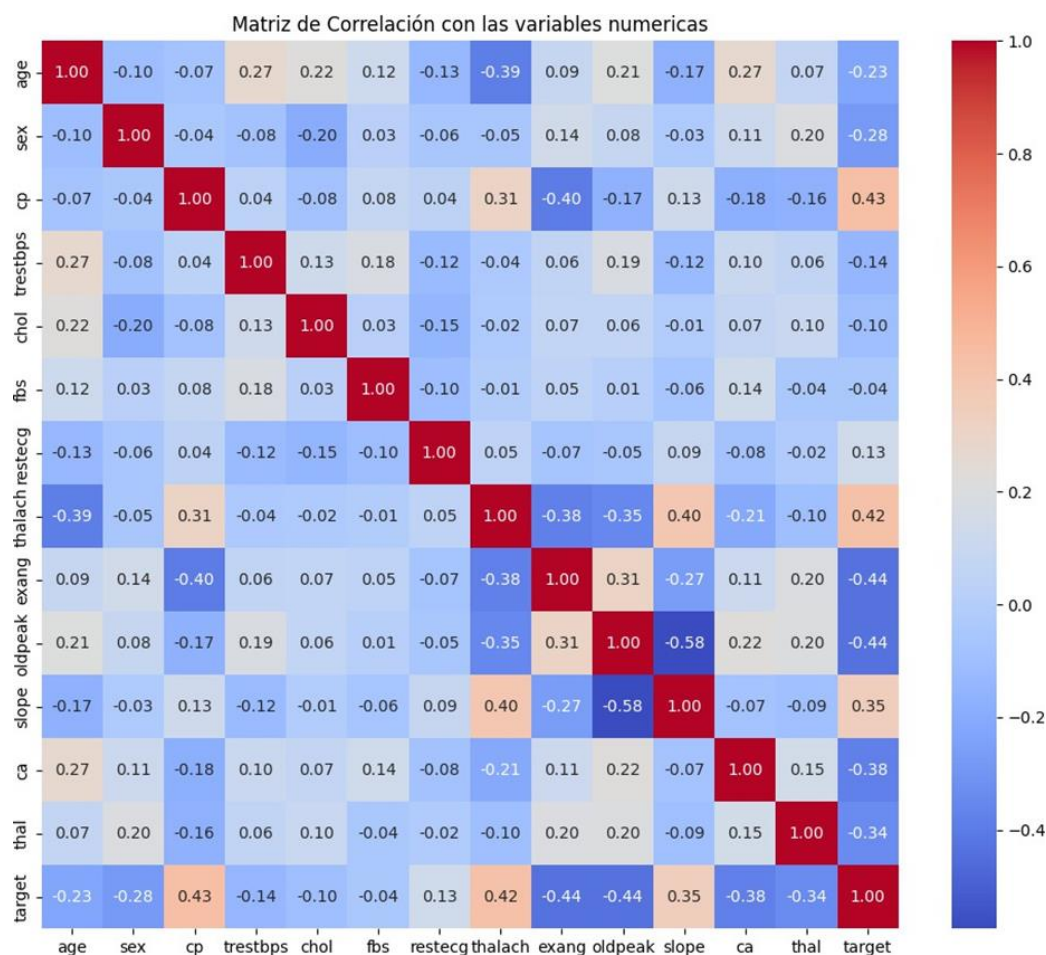
Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

La matriz de gráficos presenta las relaciones entre las 5 variables numéricas del conjunto de datos: edad, presión arterial en reposo, colesterol sérico (*chol*), frecuencia cardíaca máxima (*thalach*) y depresión del segmento ST (*oldpeak*), segmentadas por la presencia (*naranja*) o ausencia (*azul*) de enfermedad cardíaca.

En la diagonal principal de la Figura 11 se aprecia las distribuciones de cada variable numérica, mientras que en el resto de los gráficos se logra visualizar las relaciones entre pares de variables. Las observaciones más relevantes son la frecuencia cardíaca, que se aprecia una clara separación entre grupos, con valores más altos en pacientes con enfermedad cardíaca; de igual manera “*oldpeak*”

presenta valores más bajos en pacientes con enfermedad cardíaca.

*Figure 12 - Matriz de Correlación de las Variables*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

En la Figura 12 se observa que algunas de las variables son críticas para la predicción de los resultados del modelo y, en caso de existir correlaciones débiles que no aporten información al modelo, pueden entrar en consideración de descarte, ya que no aportan información esencial para su desarrollo.



### 1. Variables clave:

- **cp (tipo de dolor de pecho):** +0.43 Correlación positiva moderada
- **thalach (frecuencia cardíaca máxima):** +0.42 Correlación positiva moderada
- **thal (Resultado de Talio):** -0.34 Correlación negativa moderada
- **exang (angina inducida por ejercicio):** -0.44 Correlación negativa moderada
- **oldpeak (depresión del ST):** -0.44 Correlación negativa moderada

Una vez analizados los datos, también se comprobó que no existieran valores nulos o faltantes.

## Preprocesamiento de los Datos

### Variables Numéricas:

Analizando los valores brindados por el mapa de correlación de la Figura 14 y la evaluación de los datos, se decidió llevar a cabo el escalado de los datos numéricos que no fueran categóricos que presentaban valores muy altos.

El escalado de datos asegura que todas las variables contribuyan equitativamente al modelo, evitando que aquellas con rangos naturalmente mayores (age, oldpeak, etc.) tengan una influencia desproporcionada. Esto permite que el algoritmo identifique patrones reales en los datos, en lugar de verse influenciado por diferencias arbitrarias en la escala. [17]

Se escogieron los siguientes valores a escalar: “age”, “trestbps”, “chol”, “thalach”, “oldpeak”.

Estos valores se escalaron usando la función “scaler.fit\_transform” de la librería StandardScaler.

**Variables Categóricas:**

A las variables categóricas encontradas en los datos tales como: "sex", "cp", "fbs", "restecg", "exang", "slope", "ca", "thal".

Al ser estas variables categóricas con asignación de números enteros, se estaría introduciendo de forma implícita un orden o una relación de magnitud entre las categorías que no necesariamente existe. Para este tipo de categorías donde el orden no tiene importancia (sin orden inherente), esto es problemático y puede sesgar el aprendizaje del modelo. La función *get\_dummies* evita este problema al hacer que dichas variables categóricas se transformen en un formato numérico. [18]

**Selección de Datos de Entrenamiento y Prueba**

Para que el modelo pueda empezar a predecir datos, se dividió todos los datos en conjuntos, siendo esta división en dos partes esenciales: entrenamiento y prueba. A su vez, este conjunto de prueba se dividió otra vez en dos, que es el conjunto de validación y el conjunto de prueba. [19]

**Conjunto de entrenamiento:** El conjunto de datos de entrenamiento se utilizó para ajustar los parámetros internos del modelo durante su proceso de aprendizaje; este conjunto de datos permite que el modelo aprenda patrones, relacione las variables predictoras (X) y la variable a predecir (Y).

**Conjunto de Validación:** Estos datos se usaron para evaluar el modelo una vez que fue entrenado; esto prevé el sobreajuste (overfitting), además permite decidir cuándo detener el entrenamiento (early stopping). Estos datos son esenciales, ya que permiten medir el rendimiento del modelo durante su entrenamiento.

***Conjunto de Prueba:*** Estos datos son de uso exclusivo y nos permitieron evaluar el rendimiento final del modelo, ya que contienen métricas nunca vistas durante su entrenamiento o su validación.

Es decir, este conjunto de datos actuó en papel de simular escenarios reales donde el modelo fue puesto a prueba con datos de pacientes que no fueron utilizados en ninguna fase previa.

### **Selección De Los Modelos:**

#### **Regresión Lógica:**

La regresión logística es un algoritmo de clasificación supervisada que permite modelar la relación entre una variable dependiente categórica y una o más variables independientes. Esto para al final poder predecir la probabilidad de que una instancia pertenezca a una de dos categorías (clasificación binaria). [20]

#### ***¿Cómo funciona el modelo de regresión logística?***

La regresión logística estima la probabilidad de que una observación (Y) pertenezca a una clase específica. Para ello, el modelo usó una función de enlace logística (sigmoide), que transformó la combinación lineal de las variables predictoras en un valor entre 0 y 1.

#### **Random Forest Classifier**

El modelo Random Forest Classifier es un algoritmo de aprendizaje automático basado en “ensamble de árboles de decisión”. Este algoritmo combinó múltiples árboles (cada uno de ellos entrenado con una submuestra aleatoria de los datos y un subconjunto aleatorio de características), esto con el fin de producir predicciones más precisas que el de solamente un árbol. [21]

### ***¿Cómo funciona el modelo de Random Forest Classifier?***

Se realizó múltiples subconjuntos de los datos para que cada árbol tuviera uno con que entrenar.

Después de su entrenamiento, al momento de que se realizó la predicción, cada uno de estos árboles vota por una clase (0 o 1) y dicha clase con más votos fue la que se seleccionó al final.

### **Gaussian Naive Bayes**

El Gaussian Naive Bayes es un modelo probabilístico que se basa en el teorema de Bayes; este asume que las características son independientes entre sí dada la clase. La versión gaussiana asume que los valores de las características siguen una distribución normal. [22]

### ***¿Cómo funciona el modelo de Gaussian Naive Bayes?***

Este modelo calculó las probabilidades a priori: Probabilidad de que  $y = 0$  y Probabilidad de que  $y = 1$  a partir de las frecuencias de clases en el conjunto de datos.

Una vez hecho esto, el modelo estimó la media y la desviación estándar de cada característica para cada clase. Apoyado con el teorema de Bayes, calcula la probabilidad a posteriori.

### **Red Neuronal De Clasificación Binaria**

Una Red Neuronal de Clasificación Binaria es un modelo que se inspira en el cerebro humano, compuesto por capas; que, a su vez, estas se componen y se conectan a través de neuronas que transforman los datos de entrada mediante operaciones no lineales.

En la clasificación binaria, la última capa que ayuda a predecir el resultado usa una función de activación sigmoide para generar la predicción. [23]

### ¿Cómo funciona el modelo de Red Neuronal de Clasificación Binaria?

Se creó un modelo con los siguientes parámetros:

**Capa de entrada:** Esta capa se compone de un número de neuronas igual al número de características (variables o features).

**Capas ocultas:** Esta capa está compuesta por neuronas, pero con funciones de activación; en este caso se usó leaky Relu para capturar no linealidades. En la red neuronal utilizada se agregaron 3 capas ocultas de 64, 32 y 16 neuronas, respectivamente.

**Capa de salida:** Esta capa se conformó de solo 1 neurona con activación sigmoide para generar la probabilidad de que Y fuera igual a 1 o 0.

Además, se utilizó un optimizador “Adam” y la función de pérdida “binary cross entropy loss”.

### Obtención De Los Resultados De Cada Modelo

Se utilizaron las métricas como “accuracy”, “precisión”, “recall”, “f1 score” y “auc\_roc” para evaluar el rendimiento y resultados de los modelos; además, se graficó la matriz de confusión de cada uno de ellos. [24]

**Matriz de Confusión:** La matriz de confusión está formada por todas las clases para las que se requiere clasificar y las instancias clasificadas en el proceso de evaluación. Estas instancias se separan en la matriz según si se ha clasificado correctamente o no. [25]

*Tabla 2 - Distribución de la Matriz de Correlación*

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

[26]

- *Accuracy*: Es el número de predicciones correctas entre el número total de predicciones; se calcula como:

$$((TP + TN))/((TP + TN + FP + FN)))$$

- *Precision*: Número de predicciones correctas para una categoría entre el número total de instancias predichas como esa categoría, se calcula como:

$$(TP/(TP+FP)).$$

*Recall*: Proporción de instancias de una categoría que han sido clasificadas correctamente; se calcula como:

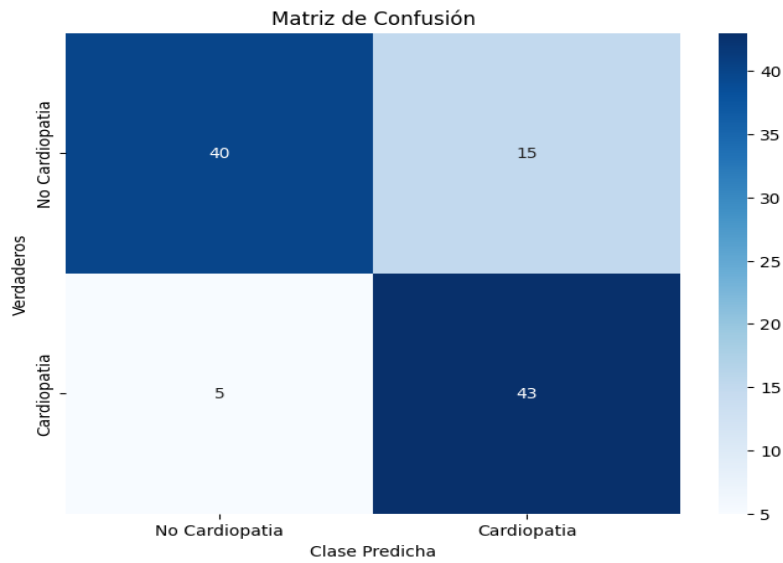
$$(TP/(TP+FN)).$$

F1-score: Número que se obtiene al hacer la operación de:

$$(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}).$$

AUC-ROC: Es un número entre 0.0 y 1.0 que representa la capacidad de un modelo de clasificación binaria para separar las clases positivas de las clases negativas.

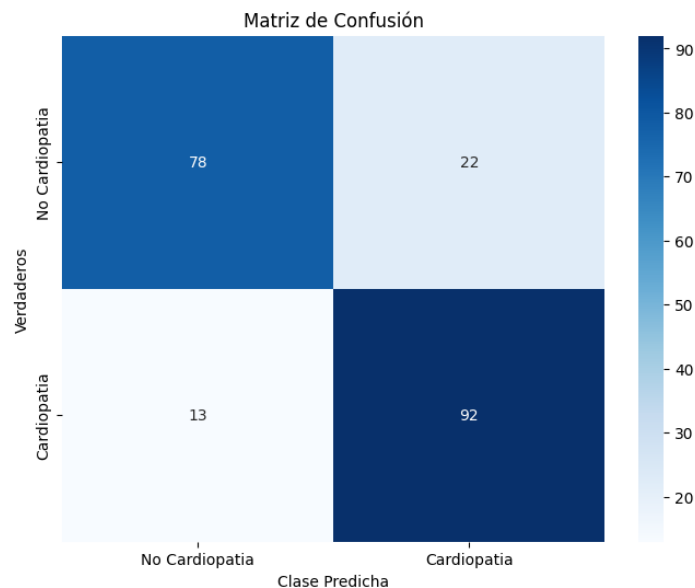
Figure 13 - Matriz de Confusión modelo Regresión Logística



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

Como se observa en la Figura 13, tuvimos 83 resultados correctos (TP+TN) y 20 incorrectos (FP+FN).

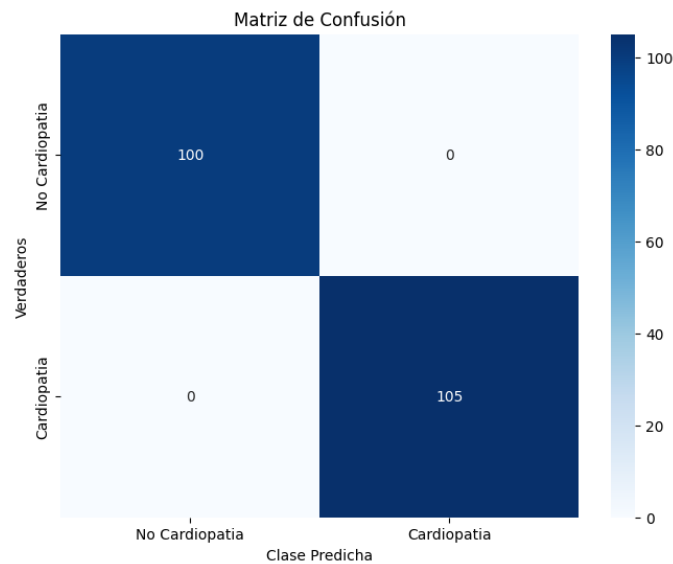
Figure 14 - Matriz de Confusión modelo GaussianNB



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

Como se observa en la Figura 14, tuvimos 170 resultados correctos (TP+TN) y 35 incorrectos (FP+FN).

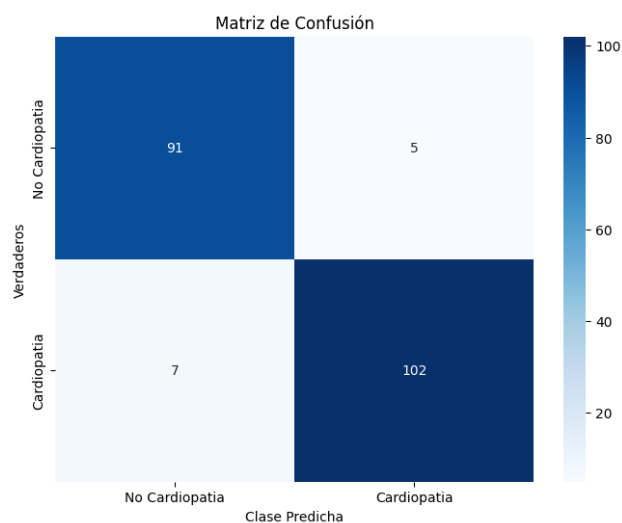
*Figure 15 - Matriz de Confusión modelo RFC*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

Como se observa en la Figura 15, tuvimos 205 resultados correctos (TP+TN) y 0 incorrectos (FP+FN).

*Figure 16 - Matriz de Confusión modelo RNCB*



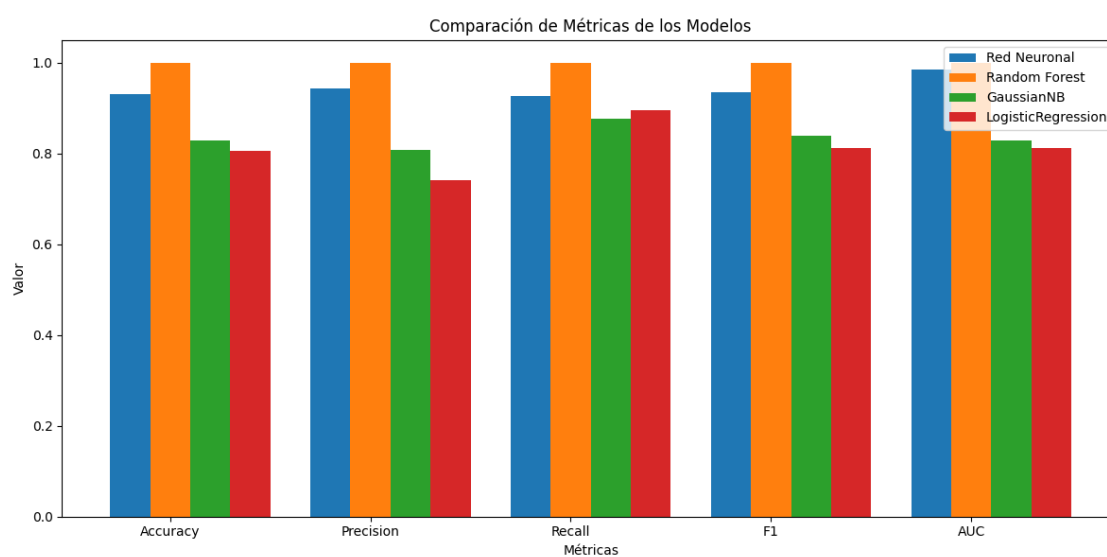
*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*



Como se observa en la Figura 16, tuvimos 193 resultados correctos (TP+TN) y 12 incorrectos (FP+FN).

### Comparación general de todos los modelos seleccionados

Figure 17 - Comparación de las métricas de todos los Modelos



Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.

Nota: Comparación de las métricas resultantes por cada uno de los modelos utilizados.

Tabla 3 - Análisis comparativo de modelos evaluados

Modelo	Accuracy	Precision	Recall	F1 Score	AUC ROC
<b>Red Neuronal CB</b>	<b>0.93</b>	<b>0.94</b>	<b>0.92</b>	<b>0.93</b>	<b>0.98</b>
<b>Random Forest</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>
<b>GaussianNB</b>	<b>0.82</b>	<b>0.8</b>	<b>0.87</b>	<b>0.84</b>	<b>0.82</b>
<b>Regresión Logística</b>	<b>0.8</b>	<b>0.74</b>	<b>0.89</b>	<b>0.81</b>	<b>0.81</b>

Nota: Resultados Obtenidos a partir de los modelos evaluados.

Fuente: Elaboración propia

En la Tabla 3 se observa cómo se evaluaron diferentes modelos de inteligencia artificial para la predicción de ataques cardíacos. Los resultados muestran que el modelo Random Forest alcanzó un rendimiento aparentemente perfecto (100% en todas las métricas), lo que nos sugiere un posible sobreajuste (overfitting) al conjunto de datos de entrenamiento. Esto nos indica que, aunque el modelo clasifica correctamente los datos conocidos, podría tener dificultades para predecir de forma precisa en casos nuevos.

Por otro lado, la Red Neuronal CB demostró un desempeño alto con 93% de exactitud, 94% de precisión y un AUC ROC de 0.98, lo que la convierte en un modelo más equilibrado y potencialmente más adecuado a un entorno clínico. A diferencia de los modelos GaussianNB y Regresión Logística obtuvieron resultados inferiores (82% y 80% de exactitud, respectivamente), lo que nos dice que son buenos modelos, pero la Red Neuronal tiene un mejor desempeño.

### **Conclusiones**

Se seleccionaron estos cuatro modelos de clasificación supervisada con diferentes características y enfoques para comparar su rendimiento: Random Forest, Red Neuronal de clasificación binaria, Gaussian Naive Bayes y Regresión Logística. La elección de estos modelos se basó en su popularidad en modelos de clasificación médica, su diversidad metodológica y su capacidad para capturar distintos tipos de relaciones en los datos.

El modelo Random Forest tuvo un rendimiento perfecto en los datos evaluados, pero su 100% de precisión y exactitud es una clara señal de overfitting, lo que limita funcionamiento en escenarios reales con datos nuevos.

La Red Neuronal es la alternativa más confiable, al combinar alta precisión con un AUC ROC cercano a 1 (0.98), indicando un alto nivel de confiabilidad sin evidencias de overfitting, lo que refuerza su capacidad en escenarios reales.

Los modelos GaussianNB y Regresión Logística presentaron un rendimiento limitado, con una exactitud menor al 85%, lo que los pone en un nivel de efectividad menor comparado a la Red Neuronal. Aunque estos modelos son conocidos por su rapidez y facilidad de interpretación, su desempeño en este caso fue inferior, debido a la complejidad de los datos y a su menor capacidad para capturar interacciones complejas entre variables.

En conclusión, la Red Neuronal de Clasificación Binaria se establece como el modelo más equilibrado entre eficacia y generalización, siendo el más adecuado para su uso en aplicaciones reales. La elección final se fundamenta en su habilidad de adaptarse a patrones complicados sin incurrir en sobreajuste, alcanzando un equilibrio ideal entre exactitud y confianza.

#### **Como trabajo futuro se recomiendan:**

- El desarrollo de una plataforma para médicos que integre el modelo como herramienta de apoyo al diagnóstico.
- Implementar un sistema de monitoreo continuo con dispositivos, como smartwatches u Oura ring, usando datos obtenidos por ellos para predecir eventos cardíacos inminentes y enviar alertas preventivas.

Estas aplicaciones serían de gran utilidad, contribuirían a la Medicina 4.0 y sumarían a la prevención de ataques cardíacos.

### **Referencias Bibliográficas**

- [1] Dritsas, E., & Trigka, M. (2024). Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence. Comput.
- [2] Chitra, S., & Dupuis, L. S. (2024). Heart Attack Prediction Using Neural Networks. Journal of Multidisciplinary Research.
- [3] Akter, S., et al. (2021). Early Diagnosis and Comparative Analysis of Different Machine Learning Algorithms for Myocardial Infarction Prediction. IEEE R10-HTC.
- [4] Marqas, R. B., et al. (2023). A Machine Learning Model for the Prediction of Heart Attack Risk in High-Risk Patients Utilizing Real-World Data. Academic Journal of Nawroz University.
- [5] Prazdnikova, M. (2024). Prediction and Assessment of Myocardial Infarction Risk Based on Medical Report Text Collection. Cybernetics and Computer Technologies.
- [6] Almazroi, A., et al. (2023). A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning. IEEE Access.
- [7] Ali, F., et al. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion.
- [8] Nannapaneni, U. K., et al. (2023). A Hybrid Model for Heart Disease Prediction Using Deep Neural Network. ICCCNT.
- [9] Zhang, D., et al. (2021). Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network. Journal of Healthcare Engineering.

[10] Krishnan, S., et al. (2021). Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. International Journal of Electrical and Computer Engineering.

[11] Rahman, S., et al. (2022). Machine Learning and Deep Neural Network Techniques for Heart Disease Prediction. ICCIT.

[12] Manur, M., et al. (2020). A Prediction Technique for Heart Disease Based on Long Short-Term Memory Recurrent Neural Network. International Journal of Intelligent Engineering and Systems.

[13] Auxilia Anitha Mary, R., & Ramaprabha, T. (2024). Predicting Heart Disease Algorithm Using DNN and MNN in Deep Learning. IRJAEH.

[14] Romero, S. (2025, January 29). Esta es la principal causa de muerte en el mundo en 2025. National Geographic España. [https://www.nationalgeographic.com.es/ciencia/esta-es-principal-causa-muerte-mundo\\_24140](https://www.nationalgeographic.com.es/ciencia/esta-es-principal-causa-muerte-mundo_24140)

[15] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>.

[16] Anderies, Anderies & Tchin, Jalaludin & Putro, Prambudi & Darmawan, Yudha & Gunawan, Alexander. (2022). Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms. Engineering, MAThematics and Computer Science (EMACS) Journal. 4. 87-93. 10.21512/emacsjournal.V4i3.8683.

[17] ¿Por qué y cuándo escalar los datos?, Codificando Bits. (2024, March 5). Codificando Bits. <https://codificandobits.com/tutorial/por-que-cuando-esalar-los-datos/>

- [18] Kaplan, S. (2024, February 22). A Simple Explanation Of Using Get\_dummies In Machine Learning » EML. EML. [https://enjoymachinelearning.com/blog/using-get\\_dummies-in-machine-learning/](https://enjoymachinelearning.com/blog/using-get_dummies-in-machine-learning/)
- [19] Na, & Na. (2020, December 19). Sets de Entrenamiento, Test y Validación, Aprende Machine Learning. Aprende Machine Learning. <https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/>
- [20] Ortega Páez E, Ochoa Sangrador C, Molina Arias M. Regresión logística binaria simple. Evid Pediatr. 2022;18:11.
- [21] Donges, N. (2024, November 26). Random Forest: A complete guide for machine learning. Built In. <https://builtin.com/data-science/random-forest-algorithm>
- [22] Learn Statistics Easily. (2024, July 24). *Qué es: Bayes ingenuo gaussiano* - [https://es.statisticseasily.com/glossario/what-is-gaussian-naive-bayes/#google\\_vignette](https://es.statisticseasily.com/glossario/what-is-gaussian-naive-bayes/#google_vignette)
- [23] Alamilla-Jiménez, E., Bolívar-Cimé, A., & Nájera, E. (2022). REDES NEURONALES Y SU APLICACIÓN EN LA CLASIFICACIÓN DE PATRONES. <https://portal.amelica.org/ameli/journal/115/1153394007/html/>
- [24] Catal, C. (2012). Performance evaluation metrics for software fault prediction studies. Acta Polytechnica Hungarica, 9(4), 193-206.
- [25] Borja-Robalino, R., Monleon-Getino, A., & Rodellar, J. (2020). Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. Revista Ibérica de Sistemas e Tecnologías de Informação, (E30), 184-196.
- [26] Acevedo Charpenel, J. Identificación de estros en perras por medio de deep learning (redes neuronales).

## Anexos

### *Ilustración 1 - Lectura y Análisis de los Datos*

▼ Lectura y Análisis de los Datos

- En esta sección leemos el dataset.

```
# Lectura del dataset
df = pd.read_csv('/content/sample_data/heart.csv')
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

*Nota: Lectura del Dataset,*

### *Ilustración 2 - Licencia del Uso de los Datos*

Deed - Attribution 4.0 International

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable

© <https://creativecommons.org/licenses/by/4.0/>

*Fuente: (Heart Disease - UCI Machine Learning Repository, n.d.)*

*Nota: La data utilizada cuenta con los permisos para su uso y distribución.*

### Ilustración 3 - Permisos para el manejo del Dataset

#### You are free to:

**Share** — copy and redistribute the material in any medium or format for any purpose, even commercially.

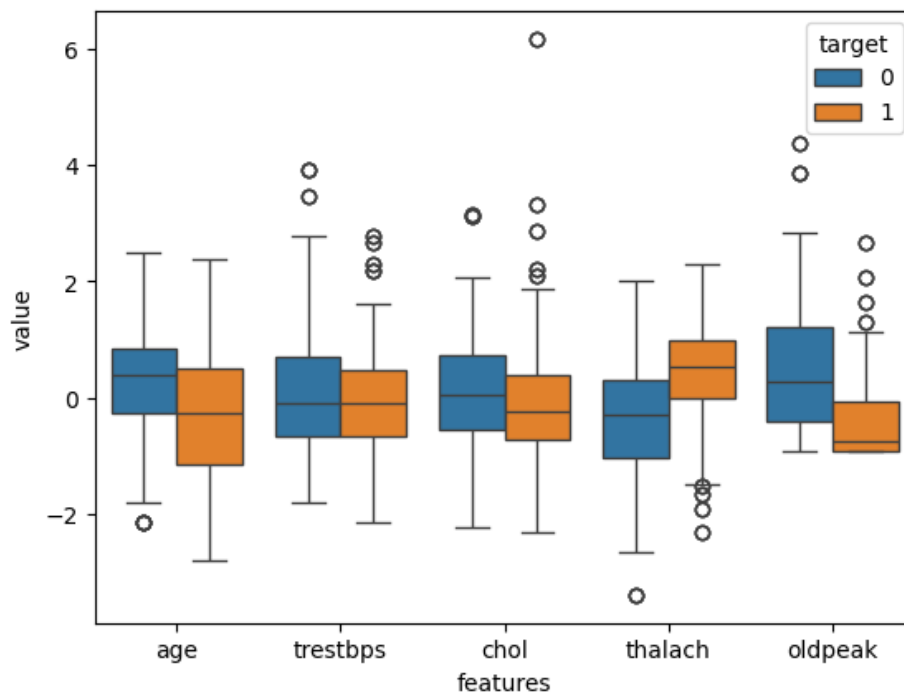
**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

*Fuente: (Deed - Attribution 4.0 International - Creative Commons, n.d.)*

*Nota: La data utilizada cuenta con los permisos para su uso y distribución.*

Figure 18 - Box Plot con las variables escaladas

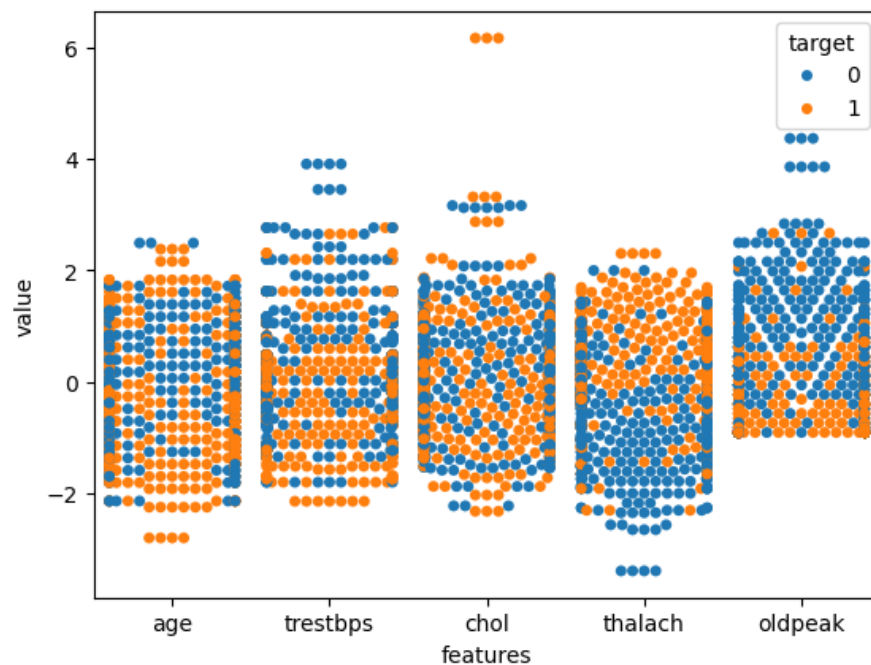


*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*



El diagrama de cajas (Figura 18) plasma la distribución de 5 características clínicas ya estandarizadas (edad, presión arterial en reposo, colesterol, frecuencia cardiaca máxima y depresión del segmento ST) entre pacientes con ataque cardíaco (target = 1, naranja) y los que no poseen ataque cardíaco (target = 0, azul). Se observan diferencias en la frecuencia cardiaca máxima (thalach) y la depresión del segmento ST (olde)

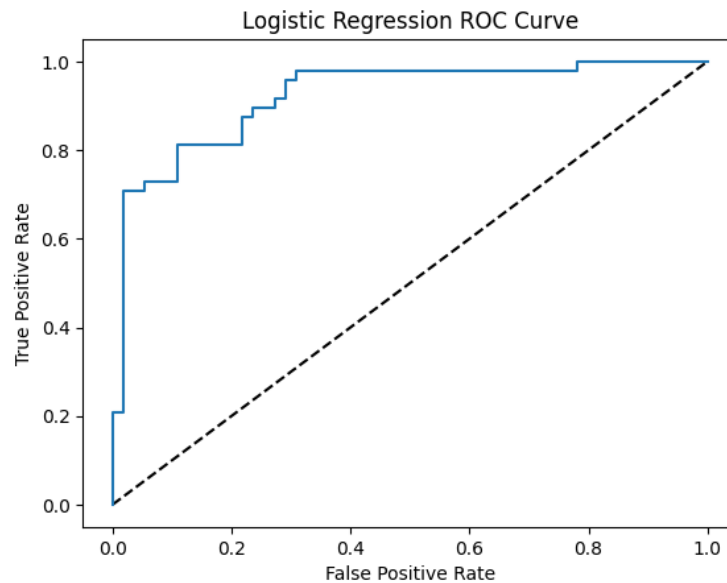
*Figure 19 - Análisis con diagrama de enjambre*



*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

*Figura 19:* Se aprecia un Swarm Plot que comprara la distribución de las características clínicas entre los pacientes con y sin diagnóstico de enfermedad cardiaca. Cada punto representa un individuo, posicionado conforme su valor en el eje Y para la característica indicada en el eje X, este análisis respalda la inclusión de las variables en la red neuronal.

*Figure 20 - Curva ROC de Regresión Logística*

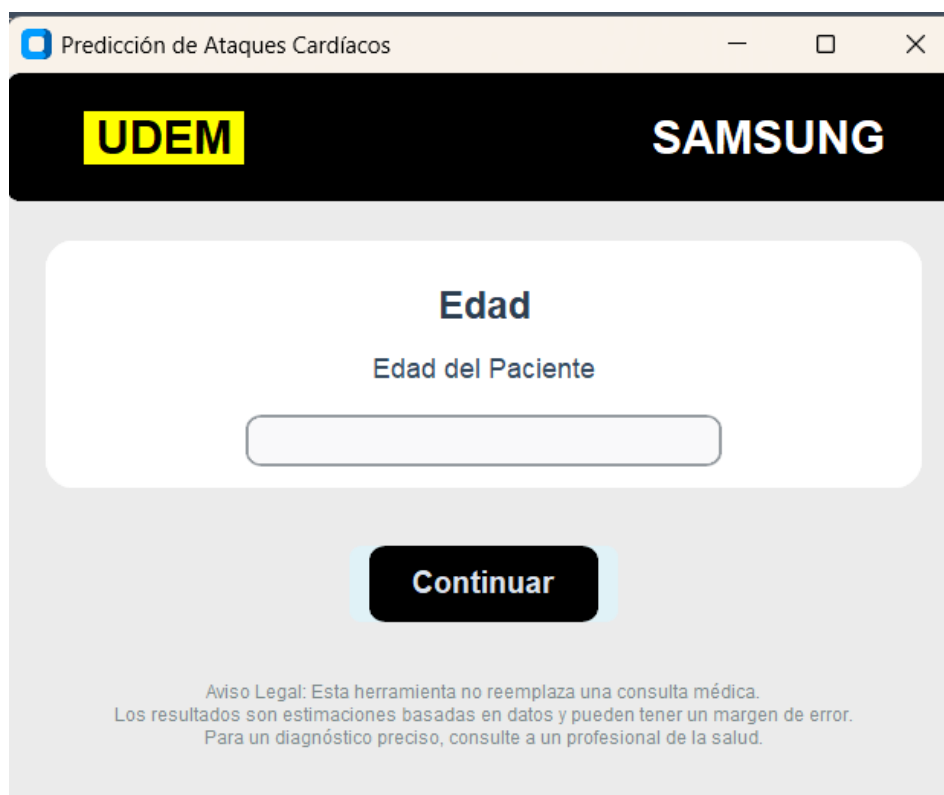


*Fuente: Elaboración propia con base en el dataset Heart Disease UCI, 2019.*

La representación de la Curva ROC de la Figura 19 para el modelo de Regresión Logística en las pruebas de predicción, la línea azul representa el rendimiento del modelo y la línea punteada corresponde a un clasificador aleatorio ( $AUC = 0.5$ ). Se muestra la relación entre los Verdaderos Positivos (TPR) (paciente correctamente identificados) y los Falsos Negativos (FPR) (pacientes sanos incorrectamente detectados como en riesgo).

## Interfaz

*Ilustración 4 - Entrada de datos de Edad*

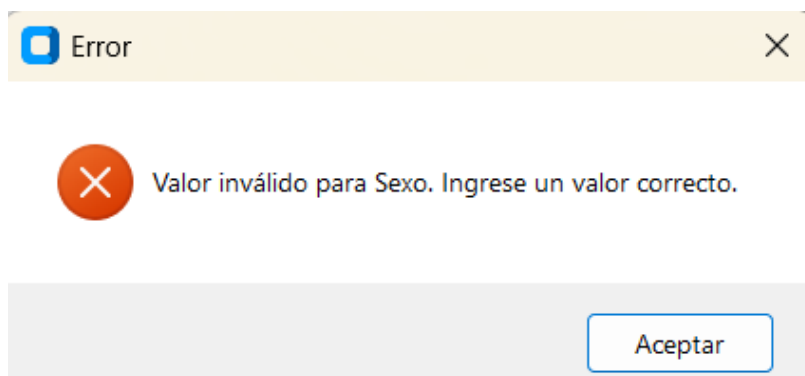


The screenshot shows a web browser window titled 'Predicción de Ataques Cardíacos'. The header is black with a yellow 'UDEM' logo on the left and a white 'SAMSUNG' logo on the right. The main content area is light gray and contains a white rounded rectangle with the title 'Edad' and the subtitle 'Edad del Paciente'. Below this is a text input field. At the bottom of the form is a black button with the text 'Continuar'. Below the button is a legal disclaimer in small text: 'Aviso Legal: Esta herramienta no reemplaza una consulta médica. Los resultados son estimaciones basadas en datos y pueden tener un margen de error. Para un diagnóstico preciso, consulte a un profesional de la salud.'

*Fuente: Elaboración propia.*

*Nota: Se visualiza una ventana de entrada de datos de los pacientes para su predicción.*

*Ilustración 5 - Ventana de validación*



The screenshot shows a yellow error dialog box with a blue square icon containing a white 'X' and the text 'Error'. Below the dialog box is a red circle with a white 'X' and the text 'Valor inválido para Sexo. Ingrese un valor correcto.' At the bottom right of the dialog is a button labeled 'Aceptar'.

*Fuente: Elaboración propia.*

*Nota: Se visualiza una ventana de “warning” para la validación de la entrada de los datos.*

*Ilustración 6 - Entrada de datos del paciente*

The screenshot shows a web browser window titled 'Predicción de Ataques Cardíacos'. The header bar is black with a yellow 'UDEM' logo on the left and a white 'SAMSUNG' logo on the right. The main content area has a light gray background. A white rounded rectangle in the center contains the title 'Tipo de talasemia' in bold. Below it, the text '1: Normal, 2: Defecto fijo, 3: Defecto reversible' is displayed. A text input field is positioned below this text. At the bottom of the white box is a black button with the white text 'Predecir'. Below the white box, there is a small legal disclaimer in gray text: 'Aviso Legal: Esta herramienta no reemplaza una consulta médica. Los resultados son estimaciones basadas en datos y pueden tener un margen de error. Para un diagnóstico preciso, consulte a un profesional de la salud.'

*Fuente: Elaboración propia.*

*Nota: Se muestra una ventana de entrada de datos para la predicción de los resultados.*

*Ilustración 7 - Visualización de Resultados*

The screenshot shows a modal window titled 'Resultado' with a close button (X) in the top right corner. The background is a light gray. In the center, there is a blue circular icon with a white lowercase 'i'. To the right of the icon, the text 'Probabilidad de Ataque Cardíaco: Bajo Riesgo' is displayed. At the bottom right of the window is a white button with a blue border and the text 'Aceptar'.

*Fuente: Elaboración propia.*

*Nota: Se muestra una ventana con los resultados de la predicción.*