Predicción del Riesgo Cardiaco utilizando Técnicas de Aprendizaje Automático

Cristiane de Andrade Coutinho
Universidad Católica de Salta
cristiane.coutinho@sou.inteli.edu.br

30 de enero de 2026

Resumen

Este estudio tiene como objetivo aplicar algoritmos de aprendizaje automático para predecir el riesgo cardíaco a partir de un conjunto de datos clínicos. Se realizó una rigurosa limpieza y transformación de datos, incluida la recodificación de la variable objetivo y la eliminación de atributos con baja correlación. Posteriormente, se implementó una normalización de los datos y un balanceo de clases para garantizar una distribución equitativa en el entrenamiento de los modelos. Se utilizaron algoritmos como Naive Bayes, K-Nearest Neighbors, Gradient Boosting, entre otros, acompañados de validación cruzada mediante el método de *Shuffled Sampling*. La evaluación se llevó a cabo con métricas como exactitud, precisión, sensibilidad (recall) y el área bajo la curva ROC (AUC). Los resultados obtenidos demuestran que, a pesar de una exactitud máxima cercana al 76 %, las técnicas aplicadas garantizan una mayor confiabilidad en la predicción del riesgo cardíaco.

1. Introducción

Las enfermedades cardiovasculares (ECV) representan una de las principales causas de mortalidad en todo el mundo. El corazón, órgano muscular responsable de bombear sangre a través del cuerpo, forma parte esencial del sistema cardiovascular. Este sistema complejo, compuesto por arterias, venas y capilares, controla el flujo sanguíneo y, cuando se ve comprometido, puede causar afecciones graves conocidas como enfermedades cardiovasculares (2).

Según la Organización Mundial de la Salud (OMS), las ECV son responsables de aproximadamente 17,9 millones de muertes cada año, lo que equivale al 31 % de todas las muertes a nivel mundial (2). Estas enfermedades incluyen la cardiopatía coronaria, enfermedad cerebrovascular, hipertensión, enfermedad cardíaca reumática, y otras afecciones del corazón y los vasos sanguíneos. Factores de riesgo como la hipertensión arterial, lípidos sanguíneos anormales, tabaquismo, obesidad, inactividad física, diabetes, edad avanzada, sexo y antecedentes familiares aumentan significativamente la probabilidad de desarrollar una ECV (2).

Ante este escenario preocupante, la detección temprana mediante modelos predictivos se ha vuelto una herramienta crucial en el diagnóstico médico. Los avances en ciencia de datos y aprendizaje automático han permitido el desarrollo de sistemas capaces de procesar grandes volúmenes de datos clínicos, identificar patrones complejos y ofrecer predicciones precisas sobre la presencia o ausencia de enfermedad (2).

1.1. Marco teórico

Con base en estudios recientes, como el publicado por Srinivasan et al. (1) en la revista *Scientific Reports* de *Nature*, se observa un creciente interés en el uso de algoritmos de aprendizaje activo para la predicción de enfermedades cardiovasculares utilizando bases de datos clínicas. Adicionalmente, se consideraron trabajos como el de la revista *International Journal of Engineering Applied Sciences and Technology* (2), que analiza el desempeño de múltiples algoritmos de aprendizaje automático sobre el conocido conjunto de datos UCI para enfermedades cardíacas.

Con base en estas referencias, el presente trabajo propone una arquitectura de *pipeline* de clasificación basada en el uso de algoritmos, complementado con técnicas de preprocesamiento, balanceo de clases, validación cruzada y evaluación mediante métricas como la matriz de confusión, precisión, *recall* y curva ROC.

En este trabajo se aplicaron varios algoritmos de aprendizaje automático para la predicción de enfermedad cardíaca, incluyendo Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN) y Naive Bayes.

Random Forest fue elegido por su capacidad para manejar datos con múltiples variables y su robustez frente al sobreajuste, gracias a la combinación de múltiples árboles de decisión. Gradient Boosting se utilizó debido a su alto rendimiento en tareas de clasificación y su capacidad para corregir errores de modelos débiles a través del aprendizaje secuencial.

El algoritmo K-Nearest Neighbors (KNN) fue incluido por su simplicidad y eficacia en la clasificación basada en la similitud de características entre pacientes. Finalmente, Naive Bayes se seleccionó por su eficiencia computacional y buen desempeño en problemas de clasificación con variables independientes, a pesar de la suposición simplificada de independencia entre variables.

La combinación de estos métodos permite comparar diferentes enfoques y obtener un análisis más completo y robusto para la detección temprana de enfermedad cardíaca.

1.2. Objetivos

El objetivo principal de este trabajo es desarrollar una *pipeline* de procesamiento para un modelo de clasificación binaria, capaz de predecir si un paciente padece o no una enfermedad cardíaca, utilizando técnicas de aprendizaje automático aplicadas sobre un conjunto de datos clínicos extraído de la plataforma Kaggle. La variable objetivo se presenta en formato binario (0 o 1), representando la ausência o presencia de enfermedad respectivamente.

2. Materiales y Métodos

2.1. Datos

El conjunto de datos utilizado en este estudio proviene del repositorio de Kaggle (4) y está basado en un estudio original publicado por Detrano et al. (3).

Este es un conjunto de datos reales que contiene información de pacientes con atributos como edad, sexo, tipo de dolor en el pecho, presión arterial en reposo, colesterol sérico, glucemia en ayunas, resultados del electrocardiograma en reposo, frecuencia cardíaca máxima alcanzada, angina inducida por ejercicio, *oldpeak* (depresión del segmento ST inducida por ejercicio en relación al reposo), pendiente del segmento ST en el pico del ejercicio, número de vasos principales y talasemia.

El estudio original, que contó con 76 atributos, demostró que un subconjunto reducido de estas variables posee mayor aplicabilidad en modelos predictivos (4). Una de las principales tareas con este conjunto de datos es predecir, basándose en los atributos proporcionados de un paciente, si dicha persona tiene o no enfermedad cardíaca. Además, se pueden realizar experimentos para diagnosticar y descubrir distintos conocimientos a partir de estos datos, ayudando a una mejor comprensión del problema.

Los autores del conjunto de datos original son: - Instituto Húngaro de Cardiología, Budapest: Andras Janosi, M.D. - Hospital Universitario, Zurich, Suiza: William Steinbrunn, M.D. - Hospital Universitario, Basel, Suiza: Matthias Pfisterer, M.D. - Centro Médico V.A., Long Beach y Fundación Cleveland Clinic: Robert Detrano, M.D., Ph.D.

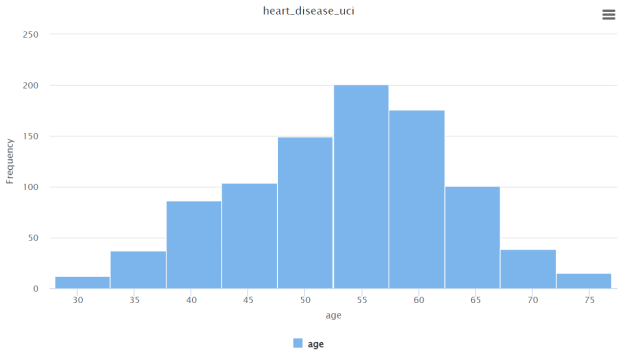
El conjunto de datos cuenta con 920 ejemplos y 16 atributos, donde la variable `num` corresponde al objetivo, es decir, el valor a predecir, y la variable `ID` identifica cada ejemplo.

Examples: 920 Special Attributes: 0 Regular Attributes: 16

2.2. Descripción de los atributos

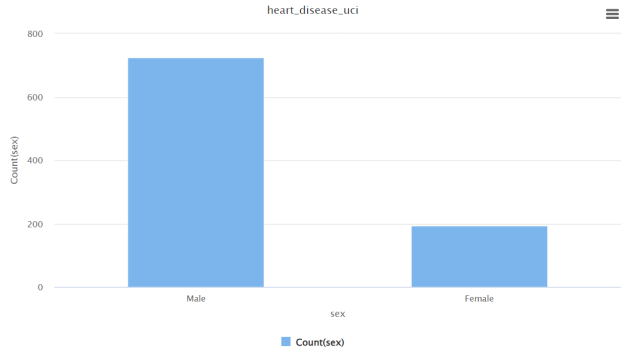
En este estudio no se considerará el atributo `ID`, ya que no aporta información relevante para la construcción del modelo predictivo.

- **Age (Edad):** Representa la edad del paciente. No presenta valores faltantes ni detectamos valores atípicos (*outliers*). La mayoría de los pacientes se concentra en el rango de 50 a 60 años, con un pico alrededor de los 55 años, que corresponde a cerca de 200 casos. Hay pocos pacientes jóvenes (menos de 40 años) y un número considerable de pacientes mayores de 70 años.



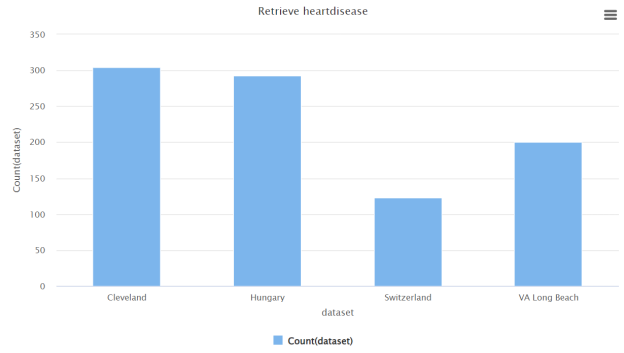
Hipótesis: La edad está altamente correlacionada con la presencia de enfermedad cardíaca, ya que el riesgo cardiovascular aumenta con la edad.

- **Sex (Sexo):** Esta variable indica el sexo del paciente. Se observa un desequilibrio de clases, con mayor cantidad de hombres que de mujeres.

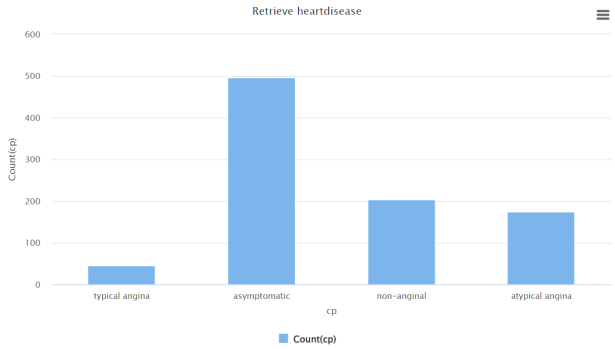


Hipótesis: Aunque el sexo masculino presenta mayor prevalencia de enfermedad cardíaca, debido al sesgo en el conjunto de datos, esta variable será excluida para evitar que el modelo aprenda un sesgo injustificado.

- **Origin (Origen):** Indica la ciudad o lugar de donde provienen los datos. Se descartará esta variable por no ser relevante para el análisis predictivo.

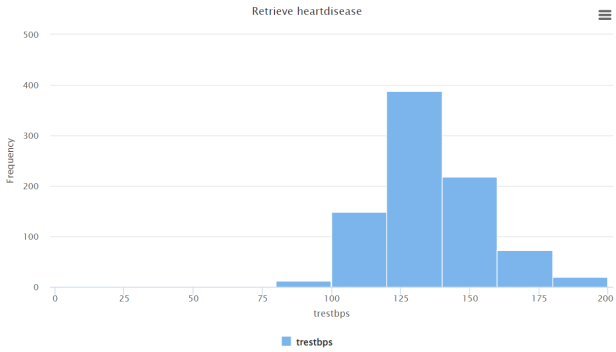


- **CP (Tipo de dolor en el pecho):** Clasificada en cuatro tipos: angina típica, angina atípica, no anginal, y asintomática. La mayoría de los pacientes reportan dolor asintomático.



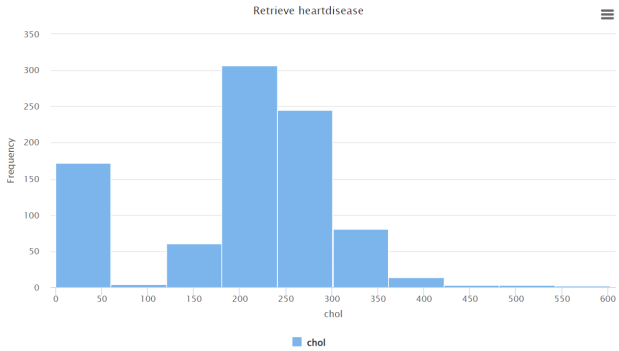
Hipótesis: El tipo de dolor en el pecho puede ser un fuerte indicador para la presencia de enfermedad cardíaca.

- **Trestbps (Presión arterial en reposo):** Medida en mm Hg al ingreso hospitalario, presenta distribución entre 75 y 200 mm Hg.



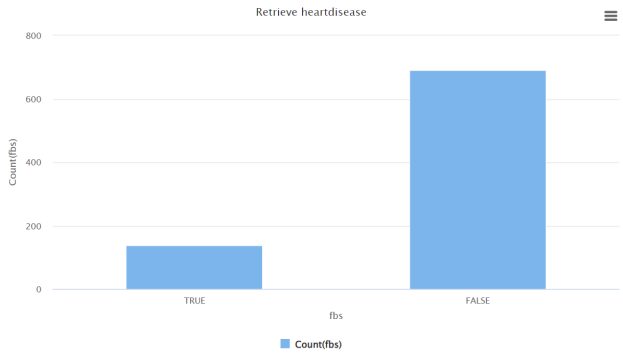
Hipótesis: Valores elevados pueden estar asociados a mayor riesgo cardiovascular.

- **Chol (Colesterol sérico):** Medido en mg/dL, se observan valores extremos bajos (0-50) con pocos casos en el rango 50-120, aumentando nuevamente en valores más altos.



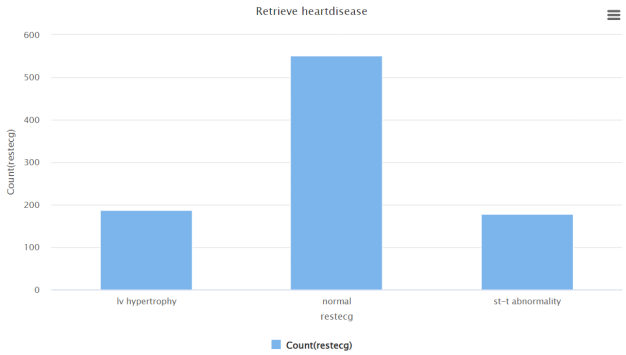
Hipótesis: Niveles elevados de colesterol están relacionados con mayor riesgo de enfermedad cardíaca.x

- **Fbs (Glucemia en ayunas >120 mg/dL):** Variable binaria, la mayoría de los casos son falsos (glucosa normal).



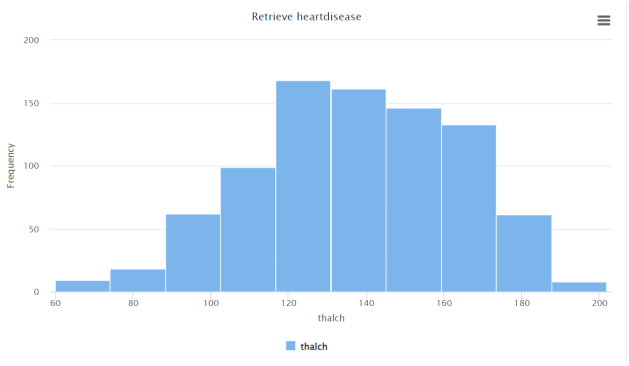
Hipótesis: Glucemia elevada puede estar relacionada con factores de riesgo cardiovascular.

- **Restecg (Resultados del electrocardiograma en reposo):** Categorías: normal, anormalidad del segmento ST-T, hipertrofia ventricular izquierda. La mayoría de los pacientes presentan electrocardiograma normal.



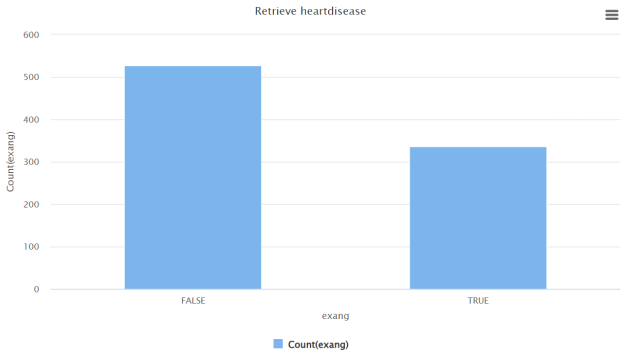
Hipótesis: Anormalidades en ECG podrían estar asociadas con enfermedad cardíaca.

- **Thalach (Frecuencia cardíaca máxima alcanzada):**

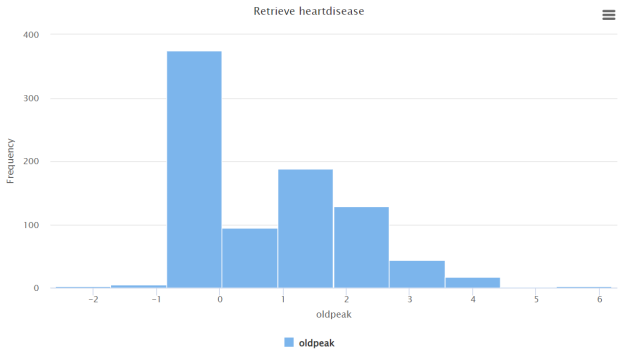


Hipótesis: Una frecuencia cardíaca máxima baja puede ser indicativa de limitación cardiovascular.

- **Exang (Angina inducida por ejercicio):** Variable binaria (verdadero/falso), presenta distribución equilibrada.

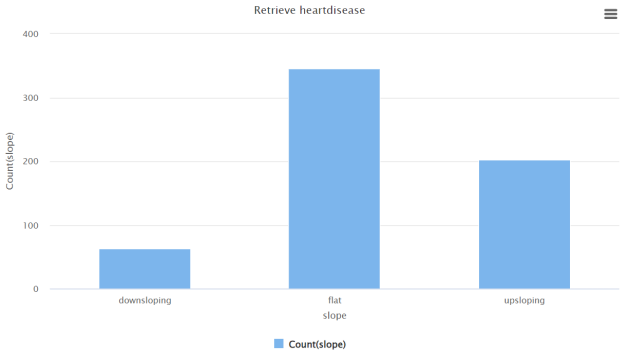


- **Oldpeak (Depresión del segmento ST inducida por ejercicio respecto al reposo):**



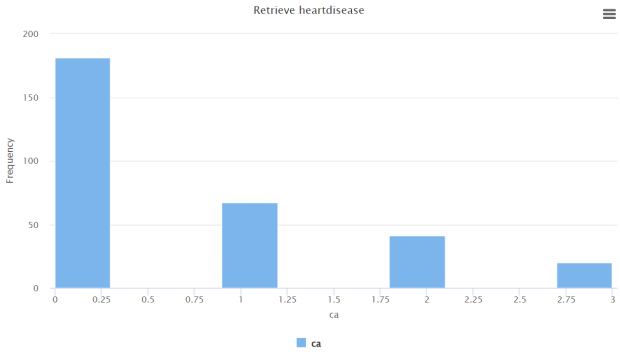
Hipótesis: Una mayor depresión del segmento ST es signo de isquemia miocárdica.

- **Slope (Pendiente del segmento ST en el pico del ejercicio):** Pocos casos con pendiente descendente.



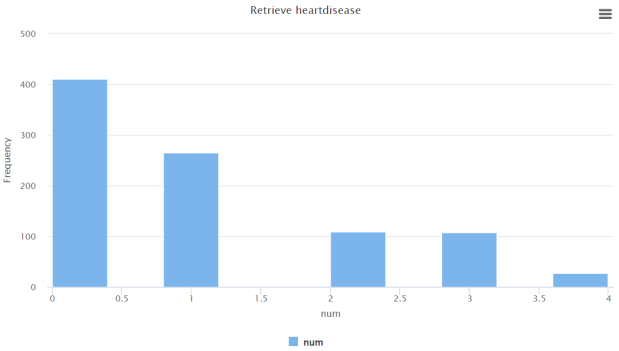
Hipótesis: Pendientes descendentes podrían estar asociadas a mayor riesgo.

- **Ca (Número de vasos principales coloreados por fluoroscopía):** Distribución concentrada en 0 y 0.25.



Hipótesis: Más vasos afectados indican mayor severidad de la enfermedad.

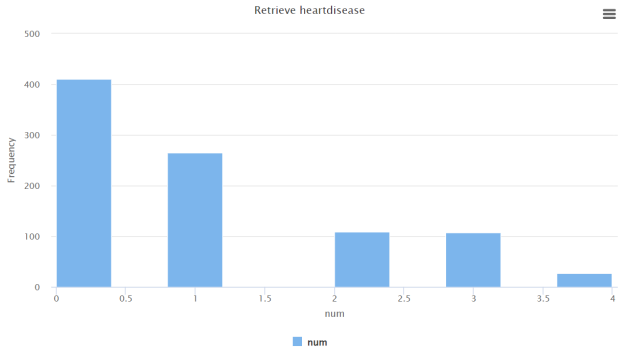
- **Thal (Talassemia):** Categorías: normal, defecto fijo, defecto reversible. Pocos valores corresponden a defecto fijo.



Hipótesis: Anomalías en talasemia pueden ser indicativas de enfermedad cardíaca.

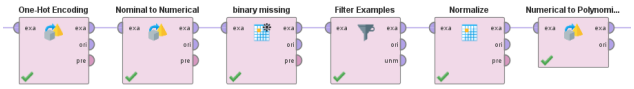
- **Num (Variable objetivo):** Atributo a predecir, con valores entre 0 y 4. 0 indica ausencia de enfermedad cardíaca. 1 a 4 indican presencia con diferente severidad:

- 1: Enfermedad leve
- 2: Enfermedad moderada
- 3: Enfermedad severa
- 4: Enfermedad muy severa



Existe un desequilibrio en las clases, con menor cantidad de casos positivos, por lo que será necesario aplicar técnicas de balanceo para evitar sesgos en el modelo.

2.3. Preprocesamiento y Limpieza de Datos



El preprocesamiento de los datos consistió en diversas etapas fundamentales para garantizar la calidad y la confiabilidad del modelo predictivo. A continuación, se detallan los pasos realizados:

2.3.1. Eliminación de Atributos Irrelevantes

Se eliminaron las variables `ID` y `origin` por no aportar valor predictivo directo al modelo. `ID` actúa únicamente como identificador del paciente, y `origin` indica la procedencia del dato (ciudad/hospital), o seja, una característica que podría introducir sesgos indeseados.

2.3.2. Tratamiento de Datos Faltantes

Algunas variables presentaban valores faltantes. El tratamiento fue realizado conforme a la proporción de valores ausentes y à natureza de la variable:

- Para atributos numéricos con pocos valores faltantes, como `chol`, se optó por la imputación mediante la **media**, tal como propuesto en el estudio de Srinivasan et al. (2023) (1), que destaca la eficiencia de este método en conjuntos de datos biomédicos.

Row No.	mode(fbs)	mode(exang)	median(chol)	median(tres...	median(thal...
1	FALSE	FALSE	223	130	140

- Las variables `fbs` y `exang`, de tipo categórico binario, fueron imputadas usando la **moda**, que en ambos casos correspondía al valor `False`.
- Registros con múltiples valores ausentes en atributos críticos fueron eliminados.

2.3.3. Codificación de Variables Categóricas

Para convertir las variables categóricas en un formato comprensible para los algoritmos de aprendizaje automático, se aplicaron diferentes técnicas:

- **Label Encoding** fue aplicado a variables binarias: `fbs` y `exang`.
- **One-Hot Encoding** se aplicó a las variables categóricas con más de dos clases, como `cp`, `restecg`, `slope` y `thal`, evitando así introducir una falsa relación de orden entre categorías.

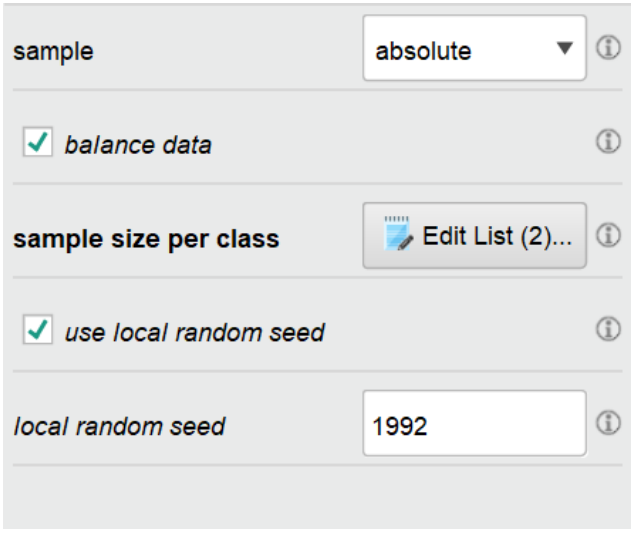
2.3.4. Normalización

Las variables numéricas continuas fueron normalizadas mediante la técnica de **Z-Score**, lo que permite estandarizar su escala:

- `age`
- `chol`
- `thalach`
- `oldpeak`

2.3.5. Balanceo de Clases

El conjunto de datos presentaba un desequilibrio de clases en la variable objetivo (`num`), con predominancia de registros sin enfermedad cardíaca. Se aplicaron técnicas de **balanceo de clases** para evitar el sesgo del modelo hacia la clase mayoritaria.



2.3.6. Transformación de la variable objetivo

Para simplificar el problema de clasificación y facilitar la entrada de datos en los modelos, se realizó una transformación en la variable objetivo `num`. Los valores originales 0 y 1 fueron agrupados y asignados al valor 0, representando la ausencia o presencia leve de enfermedad cardíaca. Los valores 2, 3 y 4 se agruparon y asignaron al valor 1, indicando presencia moderada a severa de la enfermedad. Esta binarización convierte el problema en una clasificación binaria, lo cual es más manejable para los algoritmos utilizados y mejora la interpretación de los resultados.

2.4. Análisis de Correlación

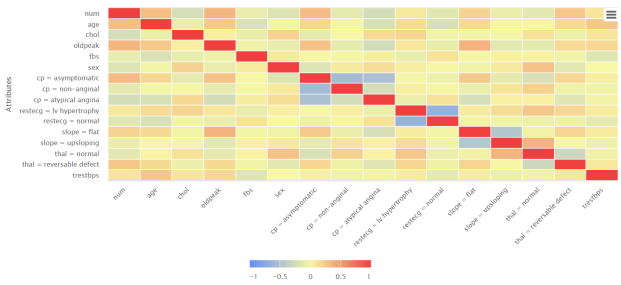
La **análisis de correlación** es una técnica estadística utilizada para evaluar la fuerza y la dirección de la relación lineal entre dos variables. En el contexto de este trabajo, se aplicó para identificar qué atributos del conjunto de datos están más relacionados con la variable objetivo `num`, que representa la presencia o ausencia de enfermedad cardíaca. Esta etapa es fundamental para seleccionar variables que puedan contribuir positivamente al rendimiento de los modelos de clasificación, eliminando atributos redundantes o irrelevantes.

Para ello, se construyó una **matriz de correlación**, que representa en forma tabular los coeficientes de correlación de Pearson entre todas las variables. El coeficiente de Pearson varía entre -1 y 1, donde valores cercanos a 1 indican una fuerte correlación positiva, valores cercanos a -1 indican una fuerte correlación negativa, y valores cercanos a 0 indican poca o ninguna correlación lineal.

La siguiente tabla resume los coeficientes de correlación entre cada atributo y la variable objetivo `num`:

Cuadro 1: Correlación de las variables con la variable objetivo `num`

Variable	**Correlación**
`num`	0.4008
`age`	0.5383
`chol`	0.8354
`thalach`	0.3222
`oldpeak`	0.4344
`fbs`	1.0000
`exang`	0.2852
`ca`	0.4533
`sex`	0.8394
`cp = asymptomatic`	0.0000
`cp = non-anginal`	0.6558
`cp = atypical angina`	0.4436
`restecg = lv hypertrophy`	0.3860
`restecg = normal`	0.6060
`slope = flat`	0.4685
`slope = upsloping`	0.5766
`thal = normal`	0.4770
`thal = reversable defect`	0.6867
`trestbps`	0.9279



Con base en estos coeficientes, se decidió eliminar algunas variables que presentaban baja correlación con el diagnóstico cardíaco, comportamiento poco informativo o problemas de calidad de datos. Las variables eliminadas fueron:

- `thalach` (correlación baja: 0.3222)
- `fbs` (a pesar de alta correlación, contenía ruido e imputaciones excesivas)
- `exang` (correlación muy baja: 0.2852)
- `ca` (alta cantidad de valores faltantes y correlación moderada: 0.4533)
- `cp = asymptomatic` (sin correlación con la variable objetivo: 0.0000)

- `restecg = lv hypertrophy` (correlación baja: 0.3860)

Estas exclusiones ayudaron a refinar el conjunto de datos, evitando el uso de variables com ruído ou redundantes, lo que contribuye a mejorar la capacidad de generalización de los modelos aplicados posteriormente.

2.5. Algoritmos Utilizados

Para la predicción de la variable objetivo `num`, que indica la presencia o ausencia de enfermedades cardíacas, se utilizaron diferentes algoritmos de aprendizaje supervisado. Con el fin de garantizar la robustez y la imparcialidad de los resultados, se aplicaron dos etapas fundamentales de preprocesamiento: el balanceo de clases y la validación cruzada.

2.6. Balanceo de Clases

La variable `num` presenta un desequilibrio significativo, con una mayoría de ejemplos pertenecientes a la clase 0 (ausencia de enfermedad). Esta situación puede hacer que los algoritmos aprendan patrones sesgados, favoreciendo a la clase mayoritaria. Para mitigar este problema, se aplicó *oversampling* a las clases minoritarias, equilibrando así la distribución de muestras entre todas las clases.

2.7. Validación Cruzada

Se utilizó la técnica de *validación cruzada* del tipo *k-fold*, con $k = 10$. Esta técnica consiste en dividir el conjunto de datos en cinco subconjuntos (folds). El modelo se entrena con cuatro de estos subconjuntos y se prueba en el subconjunto restante, repitiendo este proceso cinco veces, asegurando que todas las muestras se usen tanto para entrenamiento como para prueba. Este enfoque reduce la varianza de las estimaciones de rendimiento y ayuda a prevenir el *overfitting*. En particular, se empleó el método de *shuffled sampling*, que consiste en mezclar aleatoriamente los datos antes de dividirlos en conjuntos de entrenamiento y prueba en cada iteración. Esto ayuda a evitar sesgos relacionados con el orden de los datos y garantiza que cada partición sea representativa de la población total.

2.8. Modelos Evaluados

Se aplicaron los siguientes algoritmos:

- **Random Forest**: Un algoritmo de *ensemble* basado en múltiples árboles de decisión. Es robusto al ruido, evita el *overfitting* y permite interpretar la importancia de las variables.
- **Gradient Boosting**: Otro método de *ensemble*, pero que construye los modelos de forma secuencial, corrigiendo los errores del modelo anterior. Ofrece un excelente rendimiento predictivo y es eficaz en la modelización de relaciones no lineales complejas.

- **K-Nearest Neighbors (KNN)**: Un algoritmo basado en la distancia que clasifica una muestra según las etiquetas de los k vecinos más cercanos. Es simple e interpretable, aunque sensible a la elección del parámetro k y a la escala de los datos.
- **Naive Bayes**: Algoritmo probabilístico basado en el Teorema de Bayes. Aunque asume independencia entre las variables (lo que raramente se cumple en la práctica), suele ofrecer buenos resultados en tareas de clasificación, especialmente cuando las variables son categóricas.

Todos los modelos fueron entrenados con los datos balanceados y evaluados mediante validación cruzada, lo que permitió una comparación justa y precisa de su rendimiento.

3. Resultados

En esta sección se presentan los resultados obtenidos para cada uno de los algoritmos implementados, evaluados mediante las métricas de exactitud, precisión, recall, matriz de confusión y curva ROC.

3.1. Métricas de evaluación

Para evaluar el desempeño de los modelos utilizados en la predicción de la enfermedad cardíaca, se emplearon varias métricas que permiten medir distintos aspectos de la calidad de la clasificación:

- **Exactitud (Accuracy):** Indica el porcentaje de predicciones correctas realizadas por el modelo respecto al total de casos evaluados. Es una medida general de desempeño, pero puede ser engañosa si las clases están desbalanceadas.
- **Precisión (Precision):** Representa la proporción de verdaderos positivos entre todos los casos que el modelo clasificó como positivos. Es útil para evaluar la calidad de las predicciones positivas y minimizar falsos positivos.
- **Sensibilidad o Recall:** Mide la proporción de verdaderos positivos identificados entre todos los casos realmente positivos. Es fundamental para detectar correctamente la presencia de la enfermedad, minimizando falsos negativos.
- **Matriz de confusión:** Es una tabla que permite visualizar el desempeño del modelo clasificando los casos en verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, facilitando un análisis detallado de los errores.
- **Curva ROC (Receiver Operating Characteristic):** Representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para distintos umbrales de decisión. El área bajo la curva (AUC) es una medida agregada de desempeño que refleja la capacidad del modelo para distinguir entre clases.

Estas métricas en conjunto ofrecen una visión integral del desempeño del modelo, permitiendo evaluar tanto la precisión como la sensibilidad en la detección de la enfermedad cardíaca.

3.2. Random Forest

- **Exactitud (Accuracy):** 75.96 %
- **Precisión (Precision):** 73.36 %
- **Recall (Sensibilidad):** 82.87 %

Matriz de confusión:

	true 0	true 1	class precision
pred. 0	147	37	79.89%
pred. 1	65	175	72.92%
class recall	69.34%	82.55%	

Figura 1: Random Forest - Matriz

Curva ROC:

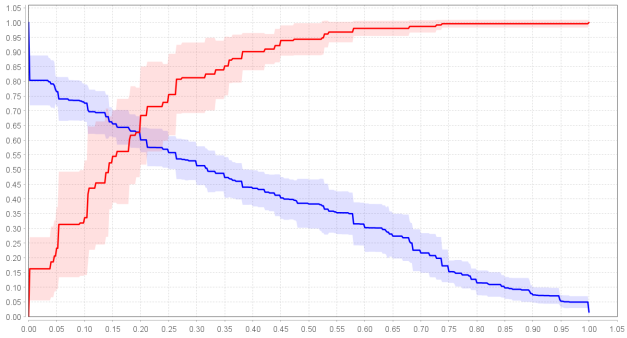


Figura 2: Random Forest - ROC

El modelo evaluado mostró un buen desempeño en la clasificación de pacientes con enfermedad cardíaca. La curva ROC presentó un AUC optimista de 0.822 ± 0.049, lo que indica que el modelo tiene una capacidad del 82.2 % para diferenciar correctamente entre pacientes con y sin la enfermedad cardíaca. Este valor de AUC se considera bueno, ya que se encuentra significativamente por encima del valor de 0.5, que representa un clasificador aleatorio.

En cuanto a las métricas específicas de clasificación, el modelo alcanzó una exactitud (*accuracy*) del 75.96 %, lo que significa que aproximadamente tres cuartas partes de las predicciones fueron correctas. La precisión (*precision*) fue del 73.36 %, indicando que de todos los casos predichos como positivos, el 73.36 % realmente tenían la enfermedad. Además, el *recall* o sensibilidad fue de 82.87 %, reflejando que el modelo identificó correctamente la mayoría de los pacientes enfermos.

En cuanto al Área Bajo la Curva (AUC), el desempeño fue bastante satisfactorio:

- **AUC (optimista):** 0.822 ± 0.049
- **AUC (media):** 0.822
- **AUC (pesimista):** 0.822 ± 0.049

En conjunto, estas métricas muestran que el modelo es capaz de equilibrar la detección de verdaderos positivos sin generar un exceso de falsos positivos, lo cual es crucial para aplicaciones médicas donde la identificación precisa de pacientes con enfermedad cardíaca es fundamental.

3.3. Gradient Boosting

- **Exactitud (Accuracy):** 75.71 %
- **Precisión (Precision):** 75.37 %
- **Recall (Sensibilidad):** 76.92 %

Matriz de confusión:

	true 0	true 1	class precision
pred. 0	158	49	76.33%
pred. 1	54	163	75.12%
class recall	74.53%	76.89%	

Figura 3: Gradient Boosting - matriz

Curva ROC:

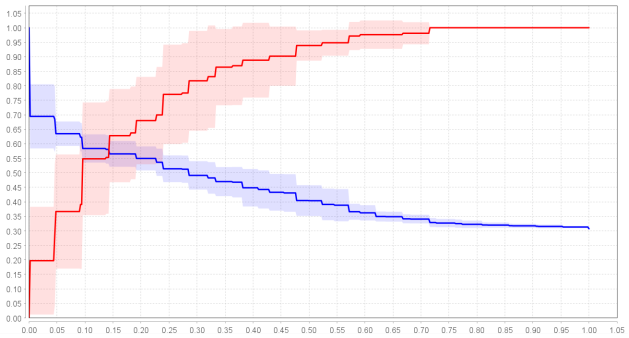


Figura 4: Gradient Boosting - ROC

El modelo de Gradient Boosting también demostró un rendimiento sólido en la clasificación de pacientes con enfermedad cardíaca. La curva ROC presentó un AUC optimista de 0.833 ± 0.069, lo que indica una capacidad del 83.3 % para distinguir entre clases positivas y negativas. El valor de AUC promedio fue de 0.831 y el pesimista de 0.828, mostrando consistencia en la capacidad predictiva del modelo incluso bajo diferentes escenarios de validación cruzada.

La curva ROC generada por este modelo mostró un patrón en forma de .^escaleraçon saltos grandes al inicio, lo cual puede indicar que el modelo identifica correctamente los casos más claros (más fáciles de predecir), mas comete errores conforme se enfrenta a instancias más ambíguas. Ainda assim, como a curva sobe rapidamente no início, esse comportamento é geralmente considerado positivo, pois significa que o modelo tem uma boa taxa de verdadeiros positivos com poucos falsos positivos.

En cuanto al Área Bajo la Curva (AUC), el modelo mostró buenos resultados:

- **AUC (optimista):** 0.833 ± 0.069
- **AUC (media):** 0.831
- **AUC (pesimista):** 0.828 ± 0.072

La precisión y la sensibilidad están equilibradas, lo que sugiere que el modelo tiene una buena capacidad para predecir correctamente los positivos sin generar un número elevado de falsos positivos. En conjunto, los resultados indican que el modelo de Gradient Boosting es una opción eficaz para este tipo de problema de clasificación médica.

3.4. K-Nearest Neighbors (KNN)

- **Exactitud (Accuracy):** 71.18 %
- **Precisión (Precision):** 72.94 %
- **Recall (Sensibilidad):** 71.24 %

Matriz de confusión:

	true 0	true 1	class precision
pred. 0	151	61	71.23%
pred. 1	61	151	71.23%
class recall	71.23%	71.23%	

Figura 5: KNN - Matriz

Curva ROC:

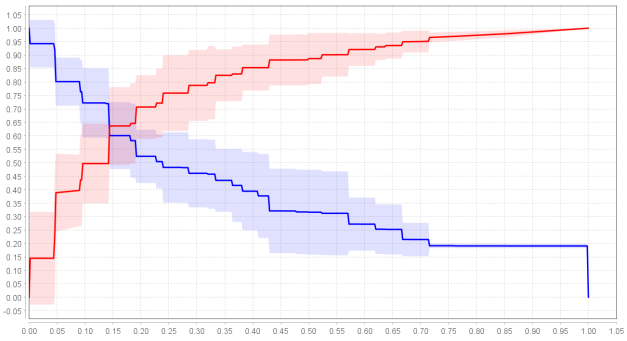


Figura 6: KNN - ROC

El modelo de K-Nearest Neighbors (KNN) presentó un rendimiento aceptable, aunque inferior en comparación con otros algoritmos probados. La exactitud fue del 71.18 %, con una precisión de 72.94 % y una sensibilidad de 71.24 %. Estos valores indican un modelo moderadamente equilibrado, capaz de identificar correctamente una proporción razonable de instancias positivas sin generar un exceso de falsos positivos.

Sin embargo, la curva ROC generada para este modelo presentó un patrón notablemente escalonado tanto en la línea optimista como en la pesimista. Este comportamiento indica

que el modelo produce cambios abruptos en la tasa de verdaderos positivos frente a la de falsos positivos a medida que se ajusta el umbral de clasificación. Las curvas en forma de .[es]calera"son comunes en algoritmos como KNN cuando se trabaja con un conjunto de datos relativamente pequeño o con predicciones menos suavizadas.

En cuanto al Área Bajo la Curva (AUC), el rendimiento también fue aceptable:

- **AUC (optimista):** 0.817 ± 0.063
- **AUC (media):** 0.809
- **AUC (pesimista):** 0.803 ± 0.065

En resumen, aunque el modelo de KNN no alcanza el nivel de precisión y suavidad observado en otros modelos más robustos como Gradient Boosting, aún puede considerarse útil, especialmente si se requiere un enfoque más interpretativo y sencillo para clasificar nuevas instancias.

3.5. Naive Bayes

- **Exactitud (Accuracy):** 76.89 %
- **Precisión (Precision):** 73.22 %
- **Recall (Sensibilidad):** 85.40 %

Matriz de confusión:

	true 0	true 1	class precision
pred. 0	145	31	82.39%
pred. 1	67	181	72.98%
class recall	68.40%	85.38%	

Figura 7: Naive Bayes - Matriz

Curva ROC:

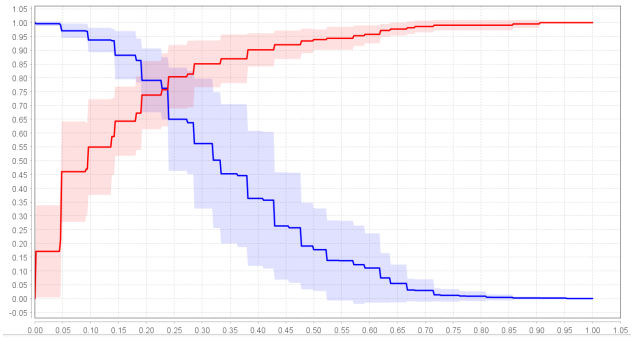


Figura 8: Naive Bayes - ROC

El modelo *Naive Bayes* demostró un rendimiento destacable en términos generales. Presentó una exactitud promedio del 76.89 % con una desviación estándar de 4.89 %, lo que

sugiere una buena capacidad de generalización. Además, la precisión fue del 73.22 %, y el *recall* alcanzó un excelente 85.40 %, lo que indica una fuerte capacidad del modelo para identificar correctamente las instancias positivas (clase 1).

En relación al Área Bajo la Curva ROC (AUC), los resultados fueron muy consistentes:

- **AUC (optimista):** 0.842 ± 0.050
- **AUC (media):** 0.842
- **AUC (pesimista):** 0.842 ± 0.050

A pesar de estos valores altos y estables de AUC, la forma de la curva ROC generada por el modelo presentó un patrón muy escalonado, con saltos pronunciados. Esta falta de suavidad en la curva sugiere que el modelo realiza clasificaciones con probabilidades discretas o poco diferenciadas entre las clases, lo cual es característico de modelos como Naive Bayes, que trabajan con fuertes suposiciones de independencia entre atributos.

En conclusión, aunque la forma de la curva ROC no fue ideal, los valores absolutos del AUC y del *recall* muestran que el modelo es altamente eficaz en la detección de positivos, siendo una opción sólida cuando se busca maximizar la sensibilidad, incluso a costa de una curva menos refinada.

4. Conclusiones

Aunque los modelos no superaron el 76 % en términos de exactitud, los métodos aplicados en el procesamiento y validación de los datos proporcionan una base sólida para confiar en los resultados obtenidos. Se aplicaron técnicas esenciales como la normalización de variables, análisis de correlación para selección de atributos, balanceo de clases y validación cruzada con muestreo aleatorio (*Shuffled Sampling*). Además, se consideró el análisis de la curva ROC para evaluar el rendimiento de los modelos en diferentes umbrales de decisión. Estos enfoques aseguran que, incluso con una precisión moderada, los modelos tengan una capacidad predictiva robusta y coherente dentro del margen del 75 % de exactitud. Los valores de AUC, superiores al 0.80 en la mayoría de los casos, refuerzan esta conclusión, mostrando que los modelos son capaces de distinguir adecuadamente entre las clases positivas y negativas, lo cual es fundamental en contextos médicos como la predicción de enfermedades cardiovasculares.

Referencias

[1] SRINIVASAN, S.; GUNASEKARAN, S.; MATHIVANAN, S. K.; M. B, B. A. M.; JAYAGOPAL, P.; DALU, G. T. *An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database*. Scientific Reports, v. 13, n. 1, p. 13588, 2023.

[2] YAHAYA, Lamido; OYE, Nathaniel David; ADAMU, Abubakar. *Performance Analysis of Some Selected Machine Learning Algorithms on Heart Disease Prediction Using the Noble UCI Datasets*. International Journal of Engineering Applied Sciences and Technology, vol. 5, no. 1, pp. 36–46, 2020.

[3] DETRANO, Robert; JANOSI, Andras; STEINBRUNN, William; PFISTERER, Matthias; SCHMID, Josef J.; SANDHOF, Peter; DUSCHER, Georg; GUZMAN, Clive; HOLMSTROM, Jan; MINGOZZI, Francesco; et al. *International application of a new probability algorithm for the diagnosis of coronary artery disease*. The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.

[4] Redwan Karim Sony. *Heart Disease Dataset*. Kaggle, 2023. Disponible en: `https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data`