Data Science II: Machine Learning para la Ciencia de Datos

Análisis de dataset: Ataques al corazón y su relación con condiciones de riesgo.

Micaela Rios.

- Gráficos para interpretar el dataset

 # Gráficos con Matplotlib

 # Gráfico 2: Colesterol por tipo de dolor de pecho

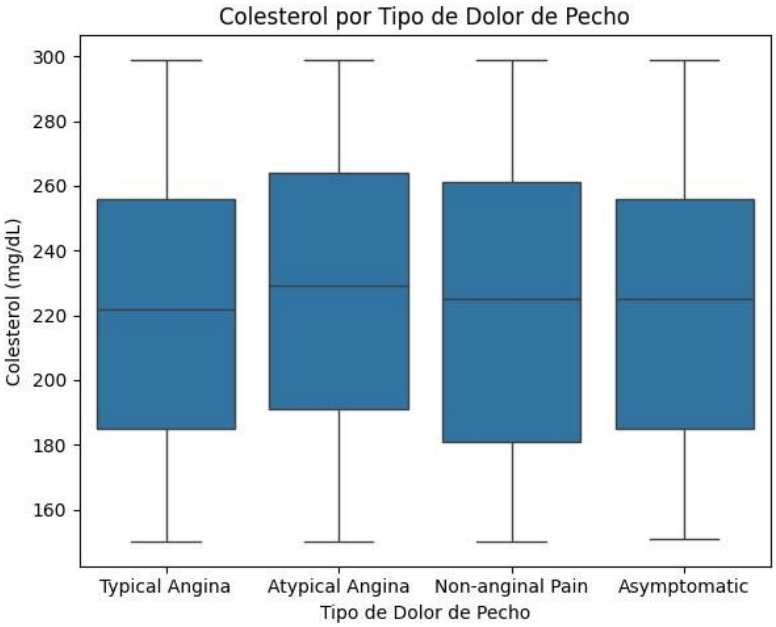
 # Gráfico 3: Edad por estado de diabetes)

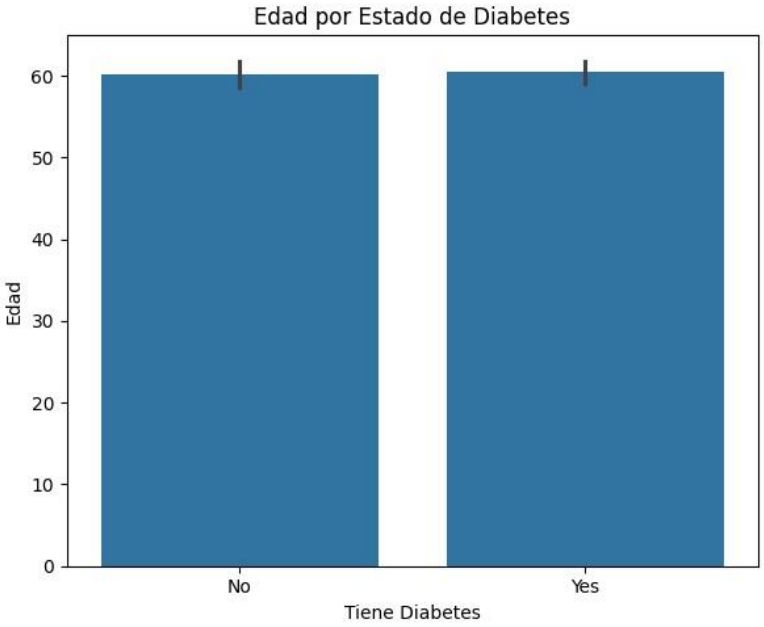
 # Gráficos con Seaborn

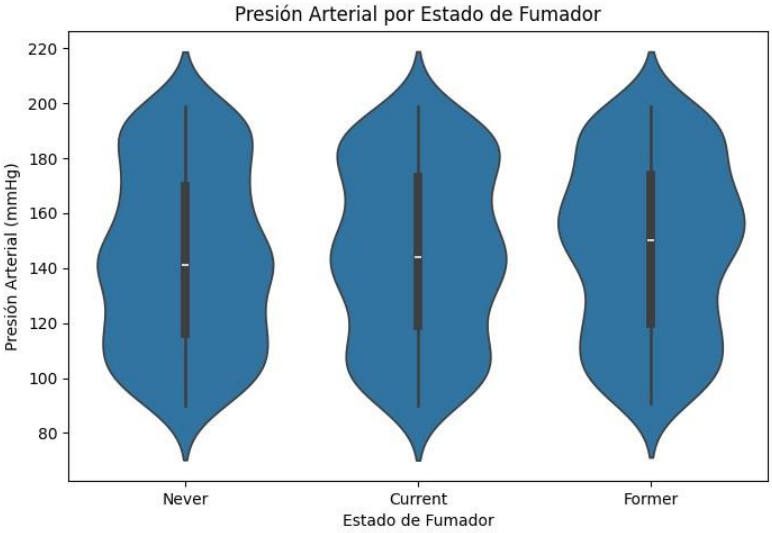
 # Gráfico 4: Presión arterial por estado de fumador

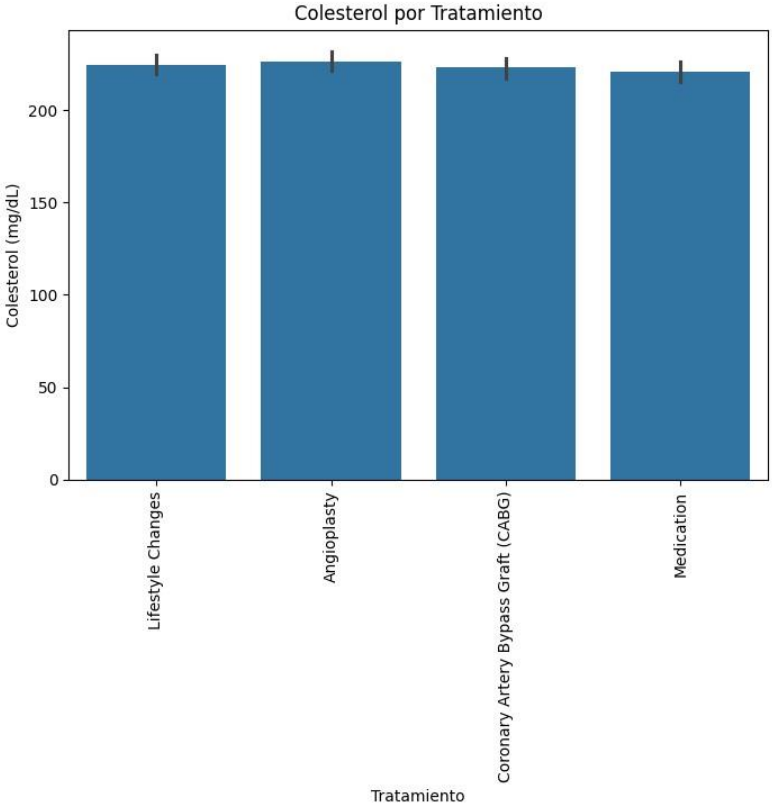
 # Gráfico 5: Colesterol por tratamiento

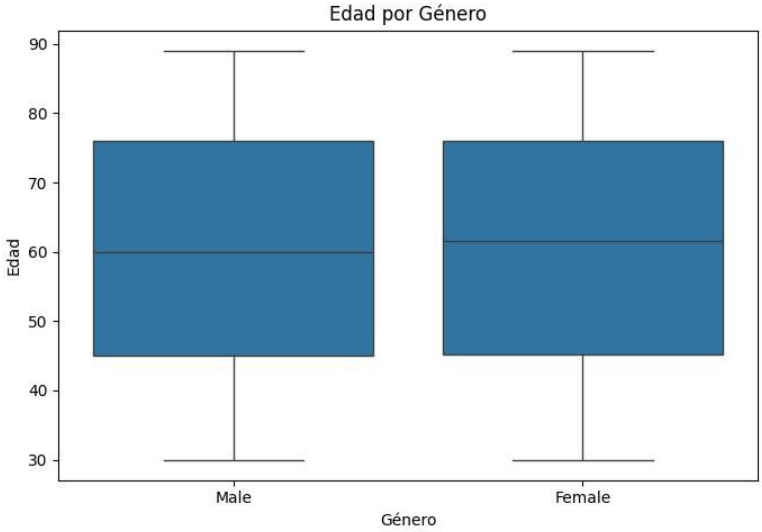
 # Gráfico 6: Edad por género


Colesterol por Tipo de Dolor de Pecho
Colesterol (mg/dL)
300
280
260
240
220
200
180
160
Typical Angina
Atypical Angina
Non-anginal Pain
Asymptomatic
Tipo de Dolor de Pecho


Edad por Estado de Diabetes
Edad
60
50
40
30
20
10
0
No
Yes
Tiene Diabetes


Presión Arterial por Estado de Fumador
Presión Arterial (mmHg)
220
200
180
160
140
120
100
80
Never
Current
Former
Estado de Fumador


Colesterol por Tratamiento
Colesterol (mg/dL)
200
150
100
50
0
Lifestyle Changes
Angioplasty
Coronary Artery Bypass Graft (CABG)
Medication
Tratamiento


Edad por Género
Edad
90
80
70
60
50
40
30
Male
Female
Género

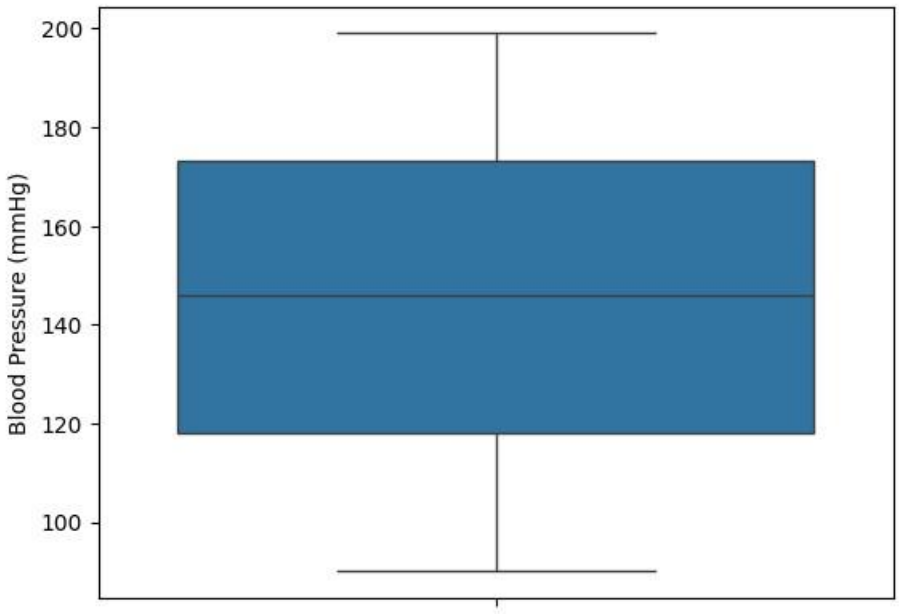
Objetivo del entrenamiento y estudio.

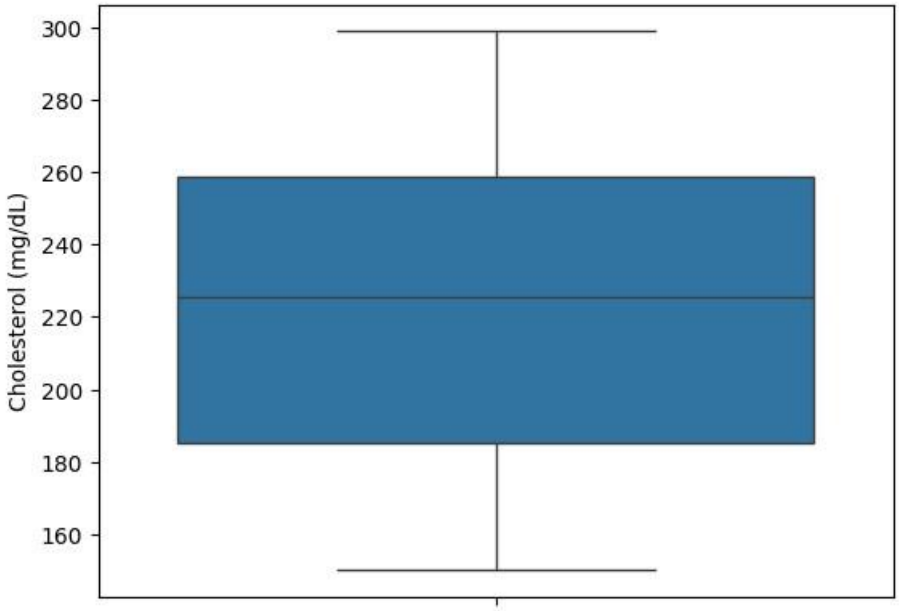
Vamos a analizar e intentar interpretar cómo impactan ciertas condiciones y factores de riesgo en la posibilidad de tener un ataque cardíaco, sobre todo personas de edad más avanzada. De esta forma poder brindar o acercar información a quienes cumplan los mismos requisitos que quienes hayan sido incluidos en la población encuestada para el estudio.

EDA

Codificamos los datos en formatos de texto a números para poder ingresarlos en los modelos para entrenar.

En los dos valores que presentan mayor posibilidad de variabilidad y amplitud no muestran presencia de outliners, por lo que podríamos avanzar de recortar cuantiles.





Vamos a analizar cuáles son los modelos que vamos a probar mediante el entrenamiento que hemos realizado, con el análisis en base a la diabetes como condición para predecir un ataque al corazón.

- Random Forest

  El modelo de Random Forest es muy robusto y preciso porque combina múltiples árboles de decisión. Esto permite manejar datos complejos con relaciones no lineales y características interactivas. Además, al promediar los resultados de varios árboles, reduce el riesgo de sobreajuste (overfitting). Si obtenemos un buen F1 Score con este modelo, significa que está capturando bien las relaciones entre los datos, y la matriz de confusión debería mostrar un buen equilibrio entre las clases positivas y negativas.

- Logistic Regression

  La regresión logística es un modelo sencillo y fácil de interpretar, ideal para análisis exploratorios. Es eficiente computacionalmente y funciona bien cuando las relaciones entre las características y la variable objetivo son aproximadamente lineales. Si el F1 Score es moderado, sugiere que el modelo es útil, pero puede no capturar todas las complejidades de los datos.

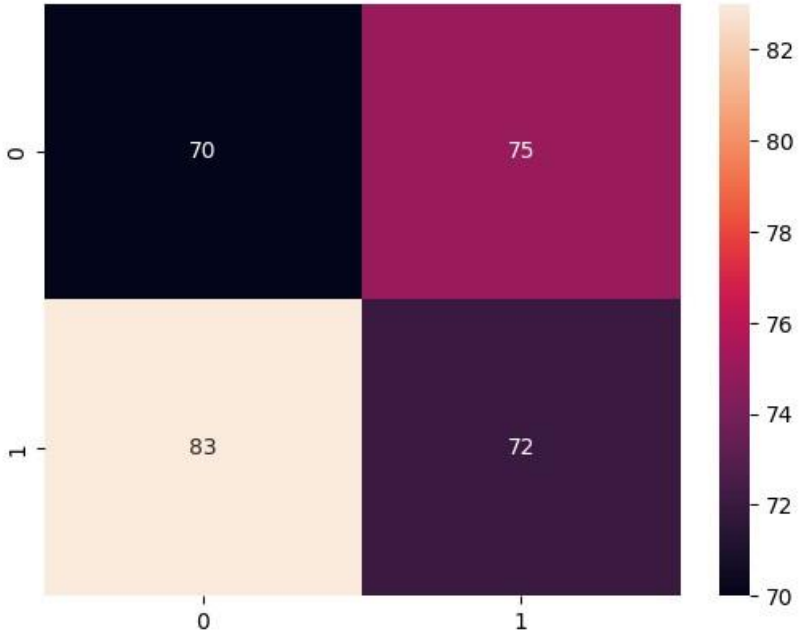
  La matriz de confusión ayudaría a ver si el modelo tiene problemas con clases desbalanceadas.

- XGBoost

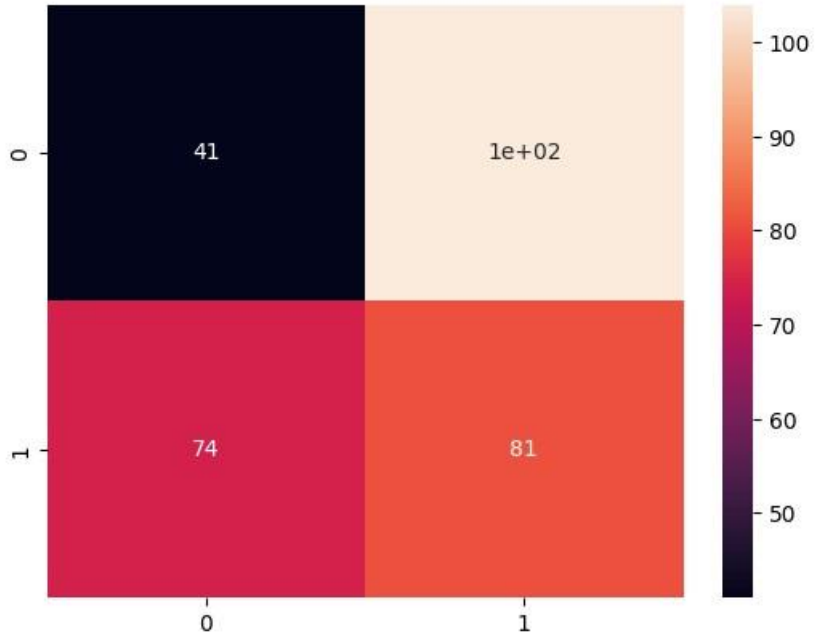
  Es conocido por su alta precisión y rendimiento, especialmente en competiciones de machine learning. Es flexible y permite ajustar muchos hiperparámetros para optimizar su rendimiento. Además, maneja bien los datos desbalanceados. Si

obtenemos un alto F1 Score con XGBoost, indica que el modelo está capturando bien las relaciones complejas en los datos. La matriz de confusión debería mostrar una buena capacidad para predecir ambas clases, incluso si hay desbalanceo.

Random Forest
F1 Score: 0.4768211920529801

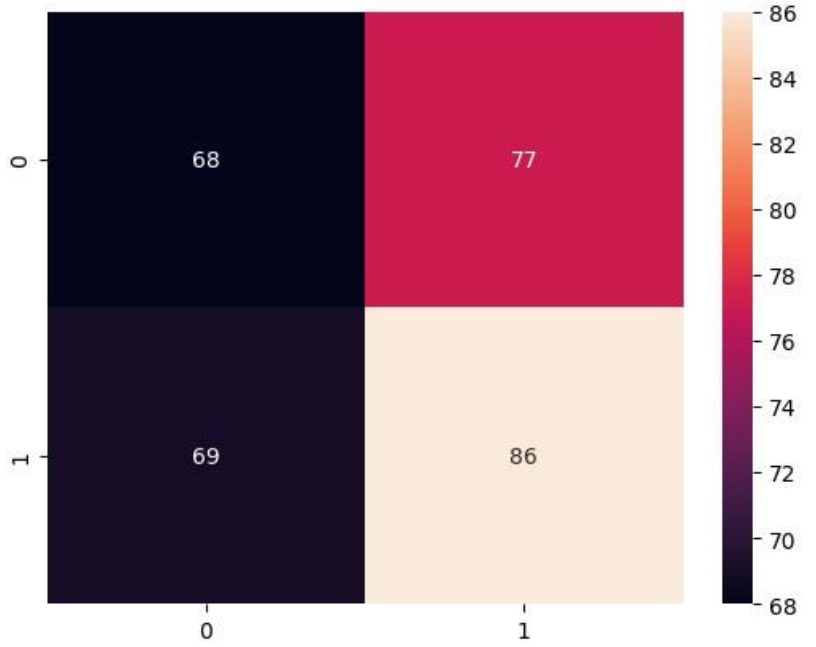


Logistic Regression
F1 Score: 0.4764705882352941

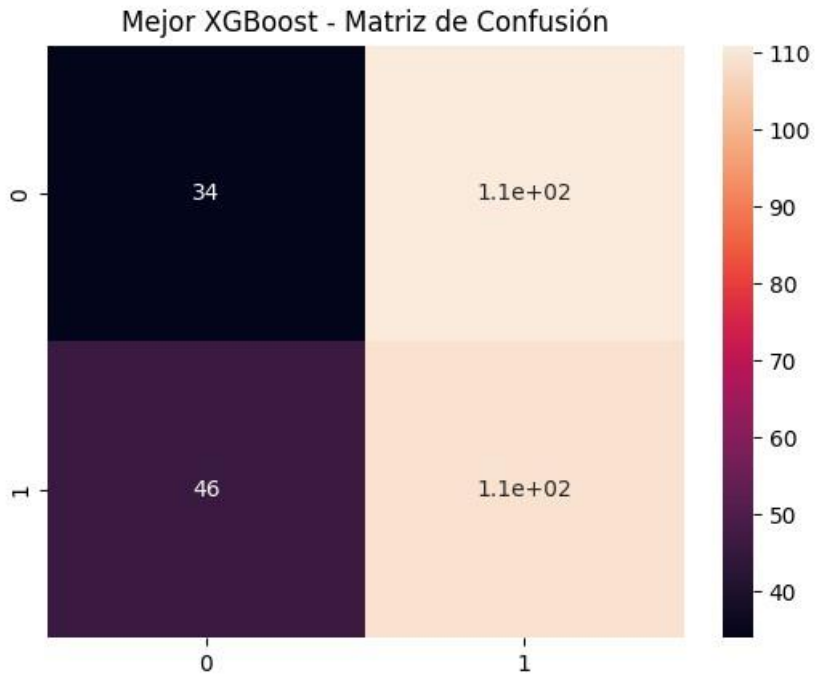


XGBoost

F1 Score: 0.5408805031446541



Debido a la versatilidad de XGboost, si bien dió mejor resultado inicial en la comparación de los 3, para mejorar los resultados obtenidos nos apoyamos en herramientas de inteligencia artificial, para que nos ayude a dar con hiperparámetros que se ejecuten en el modelo, dándonos así un mejor resultado en F1 Score y Matriz de Confusión:



En resumen, estos modelos pueden ser muy efectivos para predecir el riesgo de un ataque al corazón en personas con diabetes, considerando otros factores del dataset. La clave está en entrenar bien los modelos, ajustar los hiperparámetros y evaluar su rendimiento utilizando métricas como el F1 Score y la matriz de confusión. Esto te permitirá identificar el modelo que mejor se adapta a tus datos y proporciona las predicciones más precisas.