TRABAJO

ANÁLISIS INTELIGENTE DE DATOS

Clara Miranda García 100518506

Contenido

Enunciado del Problema 3
Objetivos 3
Análisis Exploratorio de Datos (EDA) 3
Análisis bivariado 5
Ingeniería de Características 11
Aplicación de Modelos 12
Naive Bayes 12
Logistic Regression 13
RandomForest 14
KNN 16
Modelo XGBoost Classifier: 17
ADABOOST 18
Redes Neuronales 19
Resultados 21
Conclusiones 21

Enunciado del Problema

El problema que se pretende abordar en el presente documento es el desarrollo de un modelo predictivo, capaz de identificar correctamente el riesgo de enfermedades cardiovasculares presentes en un set de individuos. El dataset descrito se obtuvo de la página web kaggle https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset.

El trabajo pretende servirse de las variables existentes y otras medidas clínicas para predecir si un individuo tiene un alto riesgo de desarrollar enfermedades cardiovasculares.

Objetivos

Entre los objetivos del trabajo, se intenta realizar un análisis exploratorio completo de los datos para entender sus distribuciones, correlaciones y los outliers que están presentes en las variables, así como descubrir cualquier sesgo o desbalance de datos.

Realizar una ingeniería de características para preparar los datos para el modelado mediante la limpieza, normalización y codificación de las variables necesarias.

Seleccionar la métrica de evaluación adecuada para medir la efectividad de los modelos de machine learning, considerando la precisión, sensibilidad, especificidad y el área bajo la curva ROC.

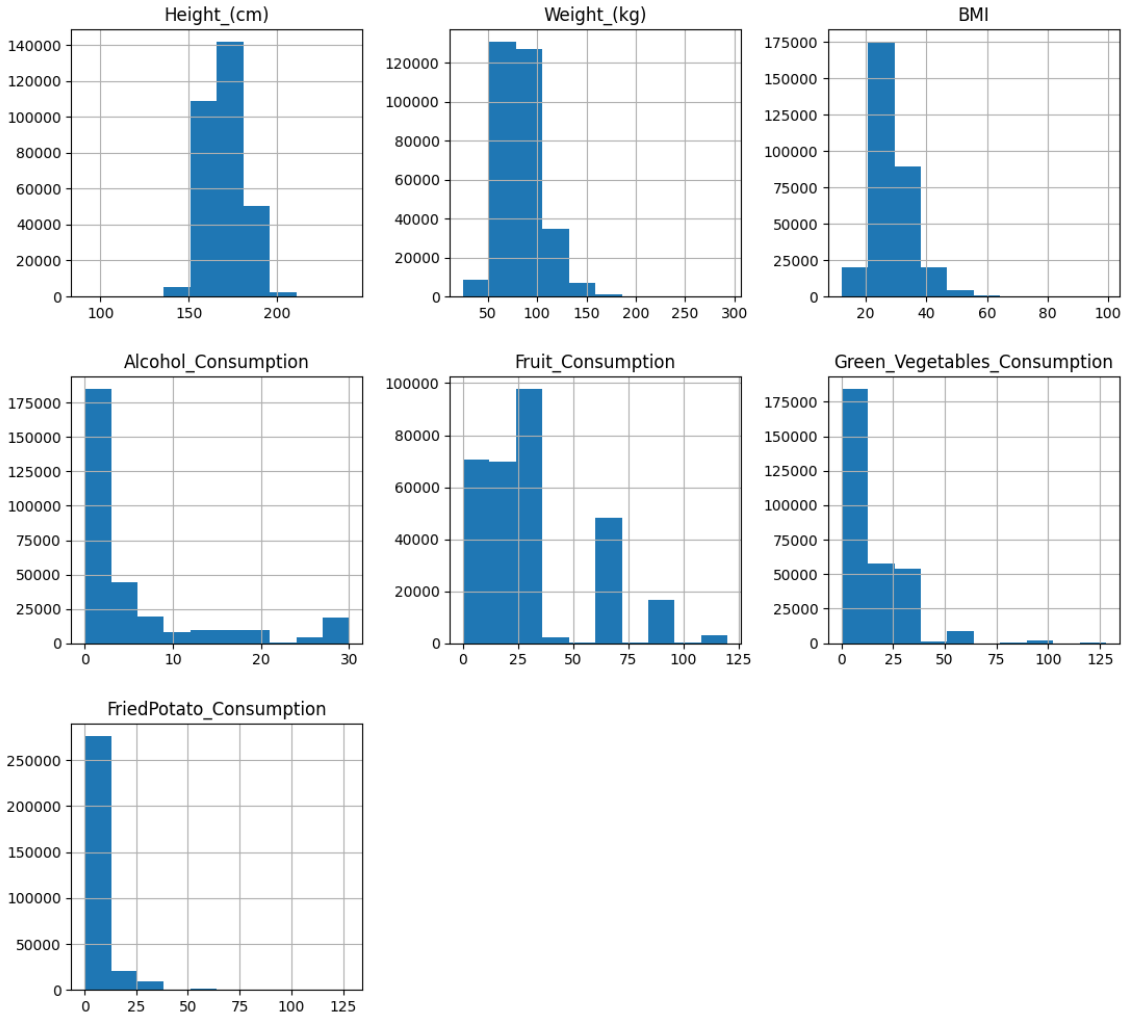
Comparar diferentes modelos de machine learning para identificar el más adecuado en términos de precisión y robustez. Experimentaremos con modelos diferentes: regresión logística, random forest y XGBoost…

Análisis Exploratorio de Datos (EDA)

El “Behavioral Risk Factor Surveillance System” O BRFSS es un dataset estadounidense centrado en distintos aspectos de salud con datos recogidos a través de una encuesta a los residentes de Estados Unidos.

Este dataset está compuesto de diecinueve variables que reflejan datos como el sexo, la edad, la salud general, la frecuencia de revisiones médicas, los hábitos de ejercicio y el historial de si es fumador, etc. El dataset incluye distintas variables representando distintas enfermedades, este estudio se enfocará concretamente en la variable objetivo de si el paciente presenta o no enfermedades del corazón.

Se explora en un primer análisis las primeras columnas del dataset, estadísticas descriptivas y visualizaciones de las distribuciones de las variables. Se lleva a cabo un análisis univariado y bivariado de los datos.



1. Height (cm): La distribución de la altura parece ser aproximadamente normal, con la mayoría de los individuos agrupados alrededor de una altura que podría estar en el rango de 150 a 175 cm. Hay pocos individuos con alturas extremadamente bajas o altas.

2. Weight (kg): La distribución del peso muestra una tendencia similar a la de la altura, con un pico que sugiere que la mayoría de los individuos tienen un peso que podría estar alrededor de 50 a 100 kg. La cola hacia la derecha indica que hay menos individuos con pesos más altos.

3. BMI (Body Mass Index): El Índice de Masa Corporal también muestra una distribución con un pico prominente, probablemente en el rango de 20 a 30, lo que indica un rango de peso normal a sobrepeso para la mayoría de los individuos. La distribución tiene una cola larga hacia la derecha, indicando la presencia de individuos con valores de BMI más altos, posiblemente obesos.

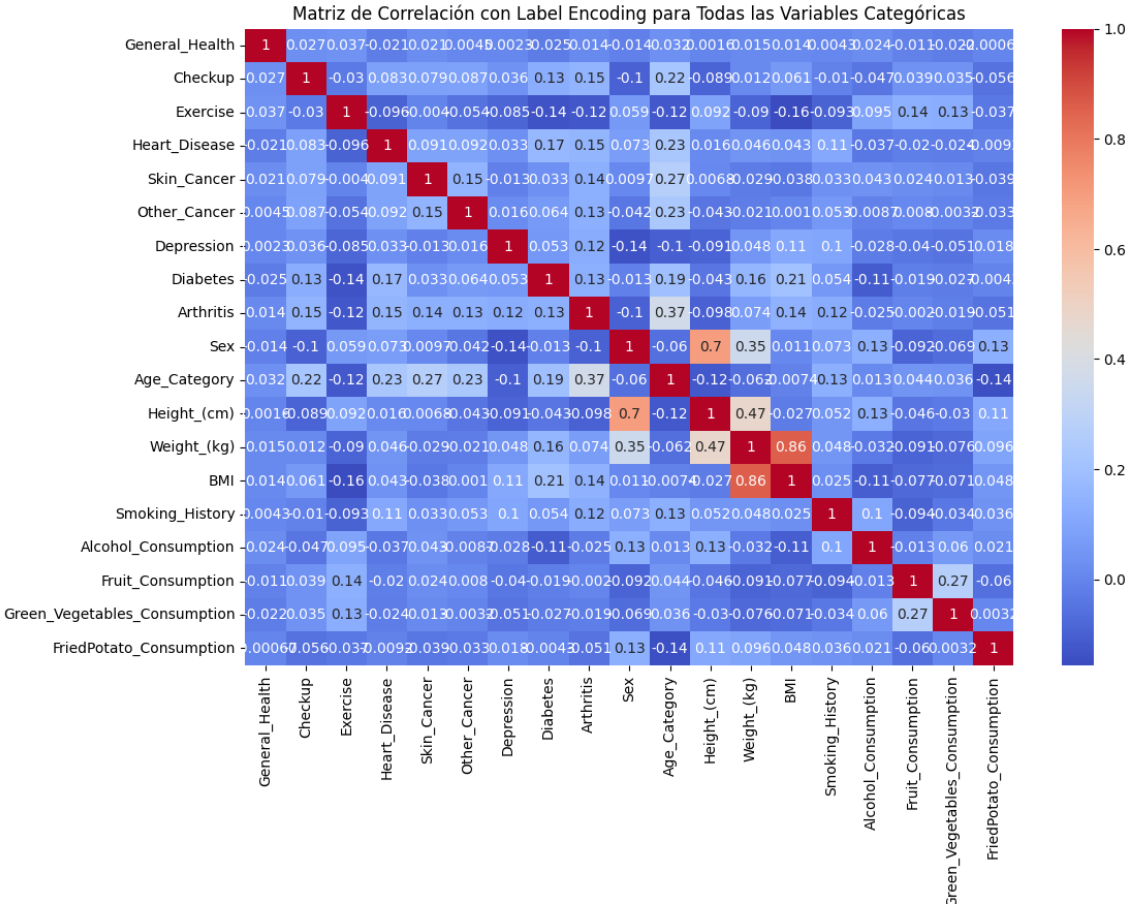
4. Alcohol_Consumption: Esta variable parece estar sesgada a la derecha, con un gran número de los individuos consumiendo poco o ningún alcohol, y muy pocos individuos consumiendo grandes cantidades.

5. Fruit_Consumption: El consumo de frutas varía más que el consumo de alcohol, con picos en valores bajos. Esto podría indicar que algunos individuos consumen frutas con mucha frecuencia o en grandes cantidades.

6. Green_Vegetables_Consumption: El consumo de vegetales verdes tiene un patrón similar al consumo de frutas. Esto sugiere dos grupos de individuos: aquellos que consumen pocas verduras verdes y aquellos que consumen una cantidad considerablemente mayor.

7. FriedPotato_Consumption: La mayoría de los individuos consume una cantidad muy baja de patatas fritas, con un pico muy alto cerca del valor más bajo.

A continuación se procede a enseñar la matriz de correlación para las variables categóricas.

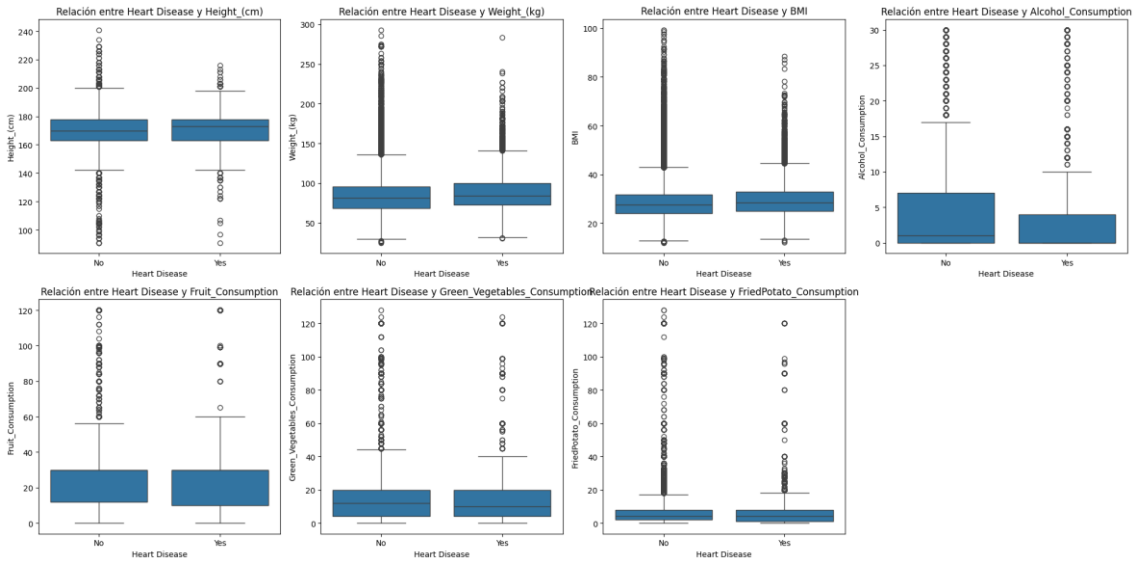


- "Exercise" tiene una fuerte correlación negativa con "Age_Category" (celda azul oscuro), lo que puede sugerir que las categorías de edad más jóvenes tienden a ejercitarse más.
- "BMI" tiene una correlación positiva moderada con "Weight_(kg)" (celda roja), lo cual es esperable ya que el IMC se calcula a partir del peso y la altura.
- Las enfermedades específicas, como "Skin_Cancer", "Other_Cancer", "Depression", "Diabetes", y "Arthritis", parecen tener poca o ninguna correlación fuerte con la mayoría de las otras variables en esta matriz, ya que la mayoría de sus celdas son blancas o de color claro.

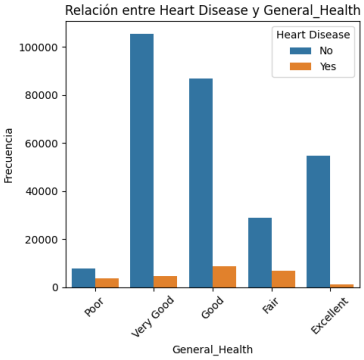
Análisis bivariado

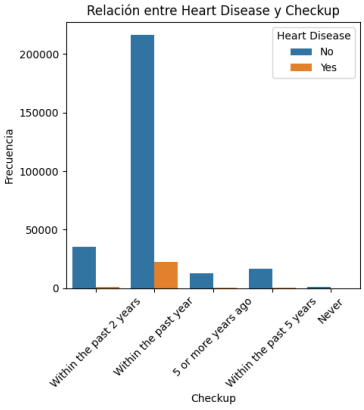
Se quiere realizar un análisis más profundo de las variables, en concreto teniendo en cuenta la relación que estas guardan con la propia variable objetivo, en este caso Heart_Disease. Gracias

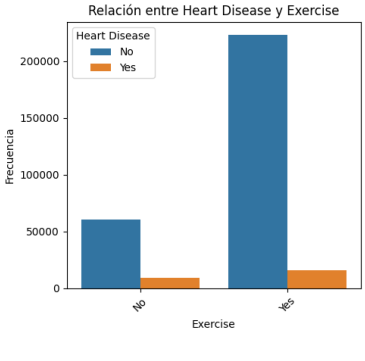
a ello se podrá determinar cómo las variaciones de una variable puede guardar conexión e influencias sobre la variable objetivo. Analizando el análisis bivariado para variables numéricas podemos sacar ciertas conclusiones.

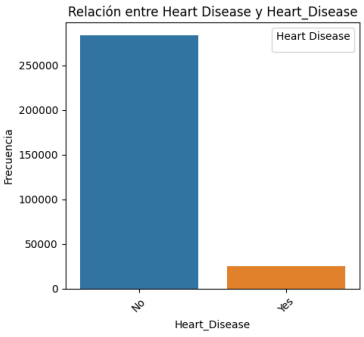


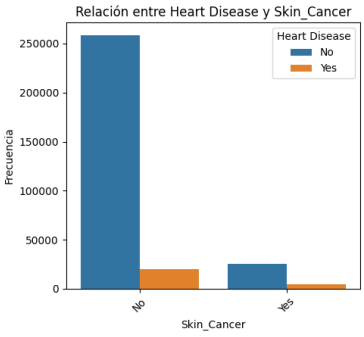
- **Relación entre Heart Disease y Height (cm):** los datos muestran que tanto personas con enfermedad del corazón como aquellas que no presentan esta patología presentan una distribución similar de altura, con medianas aproximadas a los 170 cm. Se extrae de esto que la altura por sí sola no es un predictor significativo de la enfermedad cardíaca. El número de outliers presentes en el grupo sin enfermedad refleja la gran variabilidad y cantidad de datos que hay de este grupo.
- **Relación entre Heart Disease y Weight (kg):** la distribución de peso es similar en ambos grupos presentes. No se observa ninguna diferencia notable entre ellos. En relación a los outliers se vuelve a ver el mayor número de anomalías en el grupo sin enfermedad.
- **Relación entre Heart Disease y BMI:** el índice de masa corporal (BMI) es un factor importante relacionado con las enfermedades cardíacas. De manera muy similar al caso anterior, no se pueden observar muchas diferencias significativas entre ambos grupos y presentan una variabilidad similar. Tanto el tamaño de los boxes como la posición de la mediana y de los "bigotes" están en rangos muy similares.
- **Relación entre Heart Disease y Alcohol_Consumption:** Este gráfico podría mostrar si las personas con enfermedad cardíaca tienden a consumir más o menos alcohol que aquellas sin la enfermedad. Si no hay mucha superposición entre los grupos, el consumo de alcohol podría estar asociado con la enfermedad.
- **Relación entre Heart Disease y Fruit_Consumption:** no se observan diferencias notables en el consumo de frutas, los bigotes y cuartiles son similares. No se cree que sea un factor diferencial para la enfermedad estudiada.
- **Relación entre Heart Disease y Green_Vegetables_Consumption**: al igual que el caso anterior no se observan diferencias importantes entre este consumo entre ambos grupos.
- **Relación entre Heart Disease y FriedPotato_Consumption:** existe menos variabilidad en la distribución y menos outliers en el grupo con enfermedad cardíaca. Quizás se podría atribuir a una posible asociación entre la existencia de la enfermedad cardíaca con un menor consumo de comidas no saludables, como las patatas fritas.

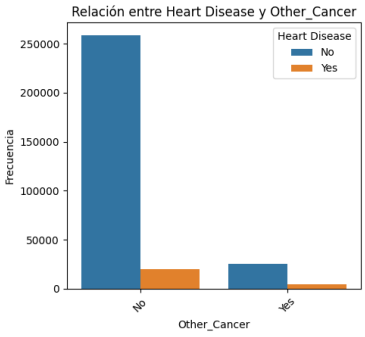

Relación entre Heart Disease y General_Health
Heart Disease
No
Yes
Frecuencia
0
20000
40000
60000
80000
100000
Poor
Very Good
Good
Fair
Excellent
General_Health

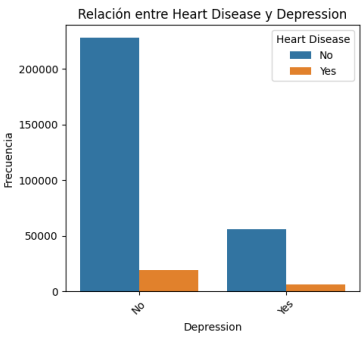

Relación entre Heart Disease y Checkup
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
Within the past 2 years
Within the past year
5 or more years ago
Within the past 5 years
Never
Checkup

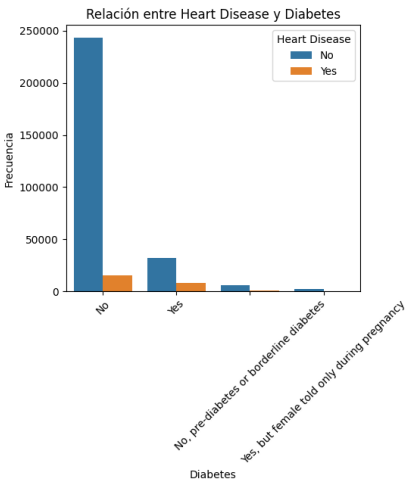

Relación entre Heart Disease y Exercise
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
No
Yes
Exercise

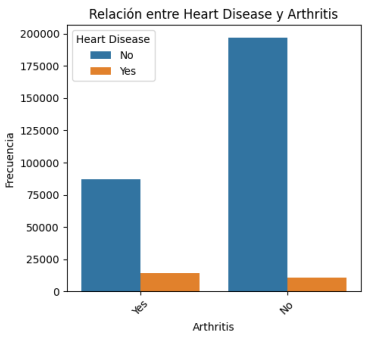

Relación entre Heart Disease y Heart_Disease
Heart Disease
Frecuencia
0
50000
100000
150000
200000
250000
No
Yes
Heart_Disease


Relación entre Heart Disease y Skin_Cancer
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
250000
No
Yes
Skin_Cancer

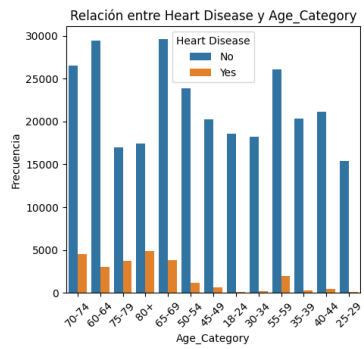

Relación entre Heart Disease y Other_Cancer
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
250000
No
Yes
Other_Cancer

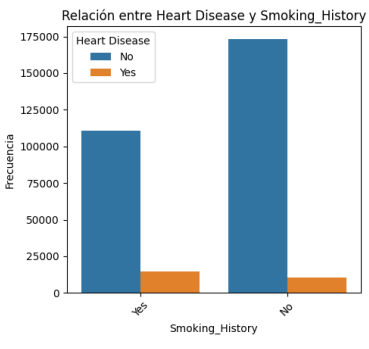

Relación entre Heart Disease y Depression
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
No
Yes
Depression


Relación entre Heart Disease y Diabetes
Heart Disease
No
Yes
Frecuencia
0
50000
100000
150000
200000
250000
No
Yes
No, pre-diabetes or borderline diabetes
Yes, but female told only during pregnancy
Diabetes

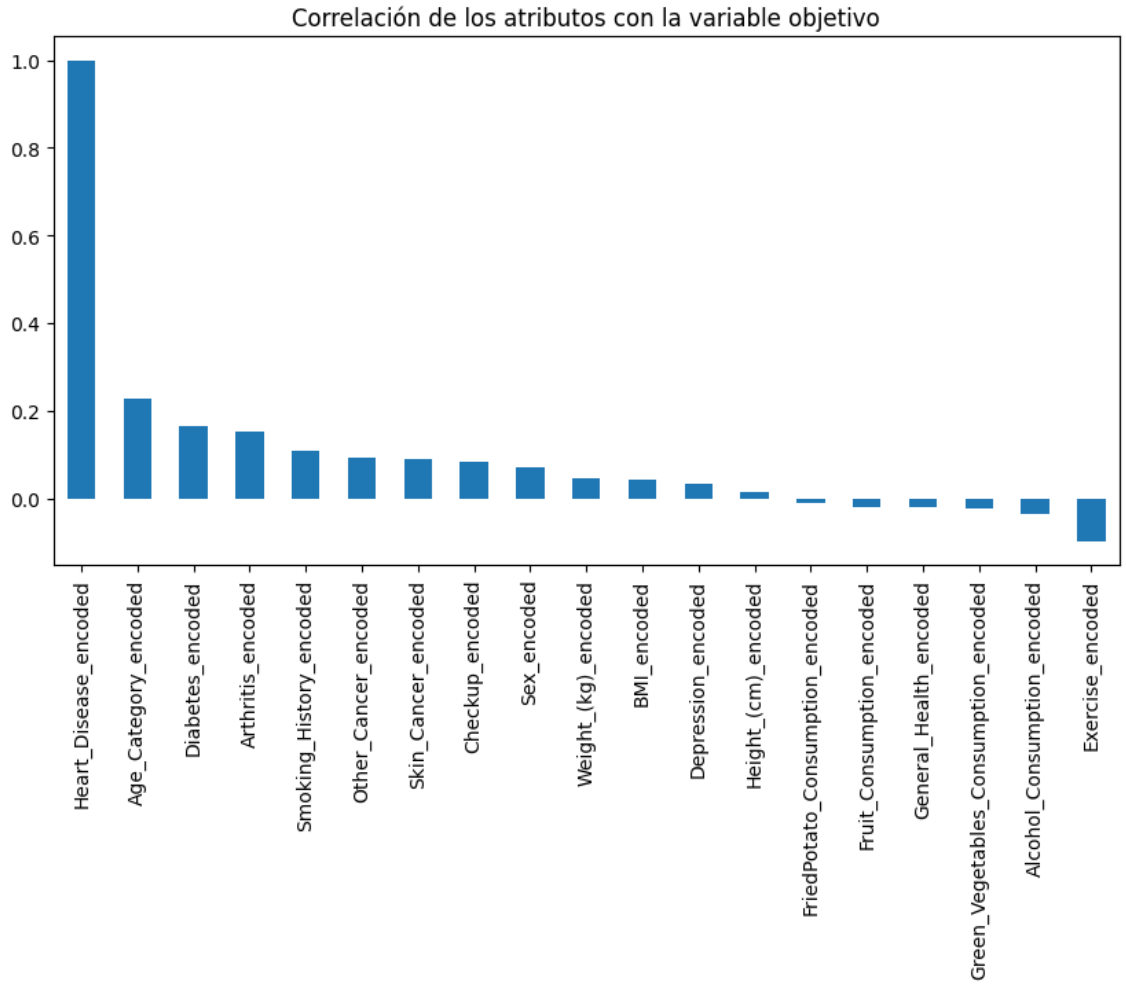

Relación entre Heart Disease y Arthritis
Heart Disease
No
Yes
Frecuencia
0
25000
50000
75000
100000
125000
150000
175000
200000
Yes
No
Arthritis


Relación entre Heart Disease y Sex
Heart Disease
No
Yes
Frecuencia
0
20000
40000
60000
80000
100000
120000
140000
Female
Male
Sex


Relación entre Heart Disease y Age_Category
Heart Disease
No
Yes
Frecuencia
0
5000
10000
15000
20000
25000
30000
70-74
60-64
75-79
80+
65-69
50-54
45-49
18-24
30-34
55-59
35-39
40-44
25-29
Age_Category


Relación entre Heart Disease y Smoking_History
Heart Disease
No
Yes
Frecuencia
0
25000
50000
75000
100000
125000
150000
175000
Yes
No
Smoking_History

- **Relación entre Heart Disease y General_Health**: se observa una clara tendencia en la que las personas que presentan un estado de salud “poor” o malo tienen una mayor tendencia de enfermedad cardíaca. Se podría considerar que existe una relación entre la salud general del individuo y la existencia de las enfermedades cardíacas.
- **Relación entre Heart Disease y Checkup:** la gran mayoría de las personas sin enfermedades cardíacas han tenido alguna clase de revisión en los últimos años, podría sugerir una posible relación entre la regularidad de revisiones y la menor incidencia de enfermedades cardíacas. Sin embargo se cree que es por un sesgo de los datos que existen.
- **Relación entre Heart Disease y Exercise**: la actividad física que realiza el individuo parece estar inversamente relacionada con la enfermedad cardíaca. La mayoría de las personas sin enfermedad del corazón reportan que hacen deporte en comparación con el grupo que sí padece estas enfermedad.
- **Relación entre Heart Disease y Skin_Cancer:** se observa una mayor frecuencia de enfermedades cardíacas en personas sin cáncer de piel.
- **Relación entre Heart Disease y Other_Cancer**: similar al cancer de piel, no se observa que haya una relación directa con la presencia de enfermedad cardíaca, aunque los datos muestren una frecuencia levemente mayor de enfermedad cardíaca en individuos sin otro tipo de cancer.
- **Relación entre Heart Disease y Depression**: la relación parece estar fuertemente asociada con enfermedades cardíacas. A pesar de esto, se quiere remarcar que esto no implica causalidad.
- **Relación entre Heart Disease y Diabetes**: la diabetes muestra una fuerte relación con el grupo de individuos enfermos, con una frecuencia mucho mayor de enfermedades cardíacas entre los individuos que padecen diabetes.
- **Relación entre Heart Disease y Arthritis**: parece existir una relación de ambas variables, ya que la artritis parece estar relacionada con una mayor frecuencia de enfermedad cardíaca.
- **Relación entre Heart Disease y Sex**: hay ligeramente más datos entre el grupo de individuos con sexo “Male” y que presenta algún tipo de patología de corazón.
- **Relación entre Heart Disease y Age_Category:** el número de invididuos con enfermedad cardíaca aumenta con la edad, es consistente con la consideración médica que dicta que el riesgo de enfermedades cardíacas aumenta con la edad.
- **Relación entre Heart Disease y Smoking_History**: los no fumadores tienen una frecuencia menor de enfermedades cardíacas en comparación con las personas fumadoras. Podría existir una fuerte asociación del historial del tabaquismo y las enfermedades cardíacas.



La correlación más alta se observa con Age_Category, lo que refuerza la hipótesis de que la edad es un factor de riesgo significativo para la enfermedad.

Arthritis y Diabetes le siguen, lo que sugiere una fuerte relación con la enfermedad del corazón.

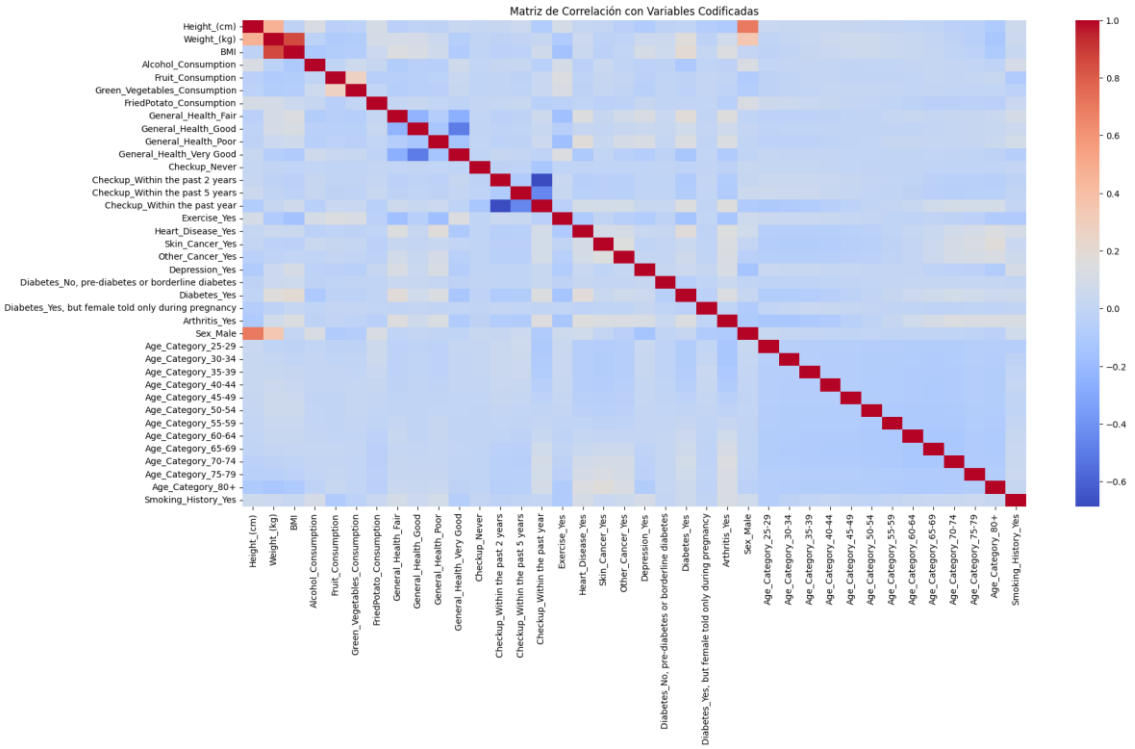
Le siguen variables como Smoking_history, que tiene correlaciones más bajas pero significativa, respaldando la idea de que pueden estar asociados con este tipo de enfermedad.

A partir del resto hay ciertas variables, por ejemplo relacionadas con el peso corporal, dieta y ejercicio que son menos determinantes comparadas con las anteriores.

Llega a haber una correlación negativa con algunas variables, como el Exercise, aunque la correlación es relativamente pequeña, es significativa. Sugiere entonces que a medida que se aumenta el ejercicio disminuye la existencia de la enfermedad del corazón en la población estudiada.

De una manera parecida, “Alcohol_Consumption” parece tener una correlación negativa con la variable objetivo. En este caso, hay que interpretar la correlación negativa de esto, no se cree que esta correlación establezca causalidad. En el marco de personas con enfermedades cardiovasculares es recomendado por expertos de la salud que dichos individuos no consuman grandes cantidades de bebidas alcohólicas. Esta sugerencia podría ser una de las razones por las que el gráfico nos muestra esto.

Finalmente, se codifican las variables y se muestra a continuación la matriz de correlación completa.



- **Edad y enfermedades:** Hay bloques de correlación positiva más oscuros que indican una relación entre categorías de edad avanzada y la presencia de varias condiciones, como enfermedades cardiacas y artritis. Esto es esperado, ya que el riesgo de muchas enfermedades aumenta con la edad.
- **Estado de salud general**: Las variables que representan el auto-reportaje del estado de salud general (por ejemplo, General_Health_Fair, General_Health_Good, etc.) parecen tener una correlación negativa con las condiciones de salud específicas. Esto podría significar que aquellos que reportan un estado de salud mejor tienden a tener menos diagnósticos de enfermedades.
- **IMC y peso**: El Índice de Masa Corporal (BMI) y el peso (Weight_(kg)) están altamente correlacionados, lo que tiene sentido dado que el BMI se calcula a partir del peso y la altura. El BMI podría tener una relación significativa con las condiciones de salud y es probable que sea una variable importante para el modelado predictivo de enfermedades cardiovasculares.
- **Comportamientos de salud:** Hay algunas correlaciones leves entre variables como la historia de fumar y la salud cardíaca, lo que sugiere posibles factores de riesgo o indicadores de enfermedades cardiovasculares.
- **Correlaciones entre enfermedades:** Existe una correlación entre diferentes condiciones de salud (por ejemplo, enfermedades del corazón y otros tipos de problemas de salud como diabetes y cáncer de piel)

Ingeniería de Características

Investigando los datos del dataset se muestra que aproximadamente el 91.9% de las muestras pertenecen a la clase 0 (no enfermo) y solo el 8.1% a la clase 1 (enfermo). Este tipo de desbalance es significativo y puede tener un impacto considerable en el rendimiento de muchos modelos de machine learning, especialmente aquellos que asumen (explícita o implícitamente) que las clases están balanceadas. Por ello se propone un preprocesado de los datos que aborde este reto.

Para la preparación del dataset, abordaremos los siguientes pasos:

1. Gestión de valores faltantes: se pretende sustituir los valores faltantes por estadísticas representativas, o eliminar si existe una cantidad excesiva de datos ausentes en la fila o columna.
2. Codificación de variables categóricas: Usaremos técnicas como Ordinal Encoding.
3. Normalización/Estandarización de variables: las variables numéricas serán estandarizadas para garantizar la existencia de un rango común, importante para algoritmos sensibles a la escala de variables. Ayudará a mejorar la convergencia durante el entrenamiento.

El Ordinal Encoding o el uso de técnicas de embedding pueden ser más adecuadas en ciertas circunstancias. Ordinal Encoding asigna un valor entero único a cada categoría de manera ordenada.

Por otro lado, los métodos de embedding, comúnmente empleados en el aprendizaje profundo, pueden aprender una representación numérica densa de las categorías, lo que permite capturar relaciones potencialmente más complejas entre ellas.

Hay que volver a clarificar que el Ordinal Encoding ordena las categorías numéricamente, lo cual puede ser un problema si el modelo interpreta esta ordenación como una relación de magnitud.

Para el desbalance de clases se pretende implementar la técnica de Synthetic Minority Over-sampling Technique (SMOTE), que crea ejemplos sintéticos de la clase minoritaria. Esta técnica selecciona ejemplos cercanos en el espacio de las características. Además se pretende hacer un submuestreo usando TomekLinks, que elimina los ejemplos de la clase mayoritaria más cercanos a la clase minoritaria, pretendiendo mejorar el límite entre clases.

Para asegurarnos de que las características numéricas contribuyen equitativamente al rendimiento del modelo, se emplea MinMaxScaler. Esta técnica ajusta las variables al rango de 0 a 1.Facilitando el proceso de aprendizaje en modelos sensibles a la magnitud de variables.

Después de aplicar estas técnicas se alcanza un balance, reflejado en 225,784 para la clase (no enfermos) y 227,109 para la clase 1. Al calcular proporciones, vemos que ambas clases presentan valores similares.

Aplicación de Modelos

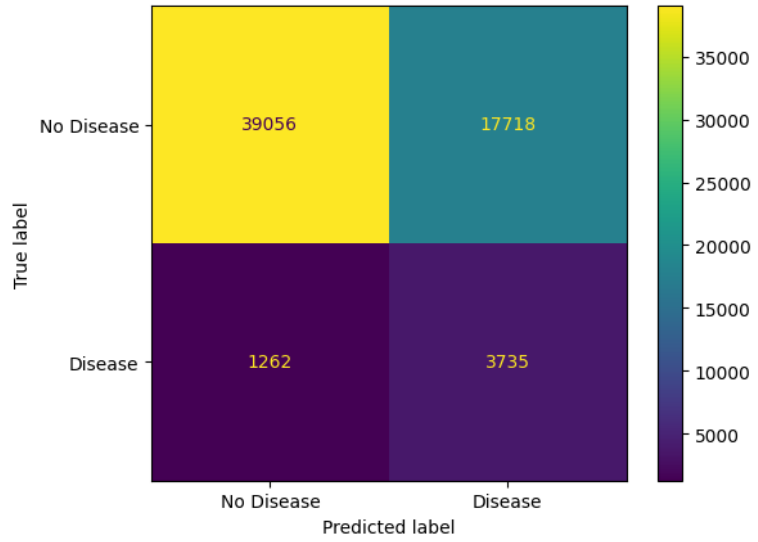
A continuación, definiremos y entrenaremos diferentes modelos, compararemos su rendimiento y seleccionaremos el más adecuado. Después de completar estos pasos iniciales, podemos proceder a la evaluación y mejora de los modelos seleccionados.

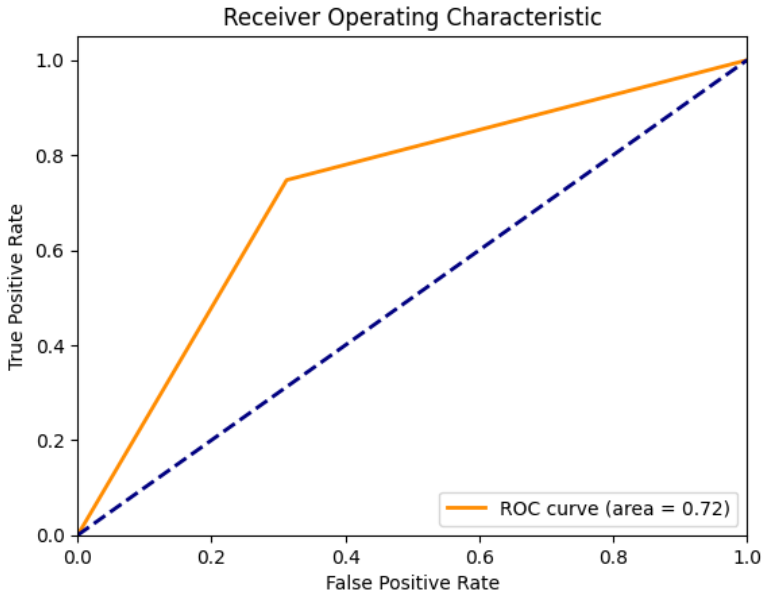
MODELO	HIPERPARÁMETRO
NAIVE BAYES	priors: [None, [0.25, 0.75], [0.5, 0.5], [0.75, 0.25]]
LOGISTIC REGRESSION	C: [0.01, 0.1, 1, 10, 100] solver: ['lbfgs', 'liblinear', 'sag', 'saga']
RANDOM FOREST	n_estimators: [50, 100, 200] max_depth: [10, 20, None] min_samples_split: [2, 5, 10]
K-NEAREST NEIGHBORS	n_neighbors: [3, 5, 7, 9] weights: ['uniform', 'distance'] metric: ['euclidean', 'manhattan']
XGBOOST CLASSIFIER	learning_rate: [0.01, 0.1, 0.2], n_estimators: [100, 200, 300], max_depth: [3, 4, 5]
ADABOOST	n_estimators: [50, 100, 200] learning_rate: [0.1, 0.5, 1.0]}
REDES NEURONALES	hidden_layer_sizes: [(50,), (100,), (50, 50)] activation: ['tanh', 'relu'] solver: ['sgd', 'adam'] alpha: [0.001, 0.01]

Naive Bayes

En cuanto al algoritmo de Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes, asume independencia entre predictores. En este caso se pretende variar los 'priors' que son las probabilidades de cada clase que se pueden ajustar.

	precision	recall	f1-score	support
0.0	0.97	0.69	0.80	56774
1.0	0.17	0.75	0.28	4997
accuracy			0.69	61771
macro avg	0.57	0.72	0.54	61771
weighted avg	0.90	0.69	0.76	61771



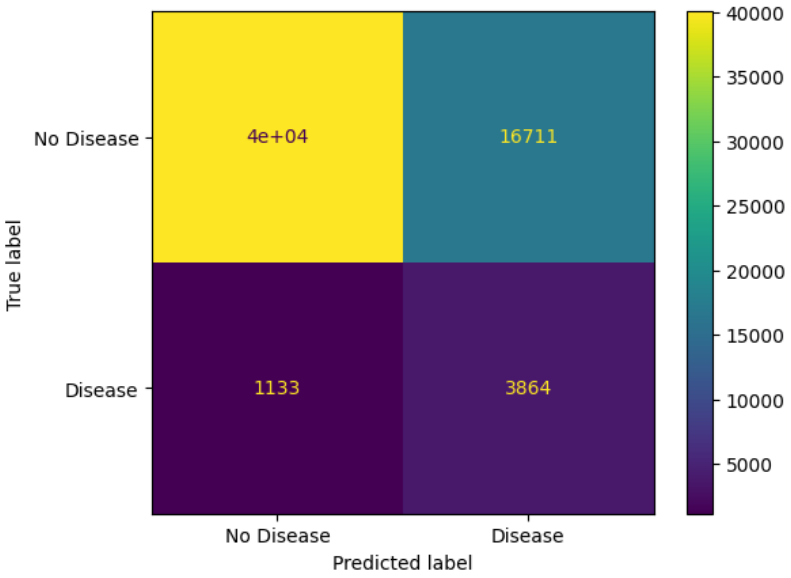


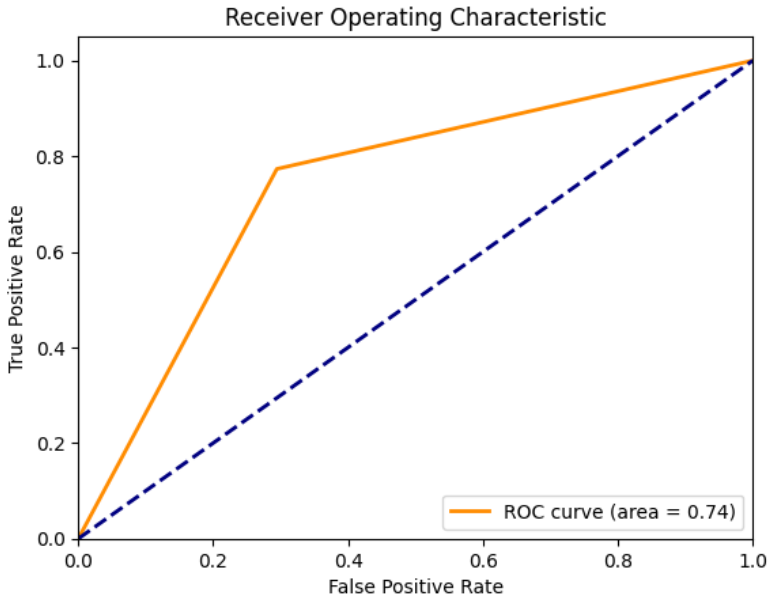
Logistic Regression

Se utilizará la regresión logística, al ser un modelo lineal que permite predecir la probabilidad de que la instancia pertenezca a una clase particular. El parámetro 'C' controla la regularización inversa, ajustarlo puede prevenir el sobreajuste teniendo datos de alta dimensionalidad o en presencia de multicolinealidad. En cuanto a los 'solvers' hacen que el modelo converja de distintas maneras, influyendo en rapidez y precisión.

	precision	recall	f1-score	support
0.0	0.97	0.71	0.82	56774
1.0	0.19	0.77	0.30	4997
accuracy			0.71	61771
macro avg	0.58	0.74	0.56	61771

weighted avg	0.91	0.71	0.78	61771

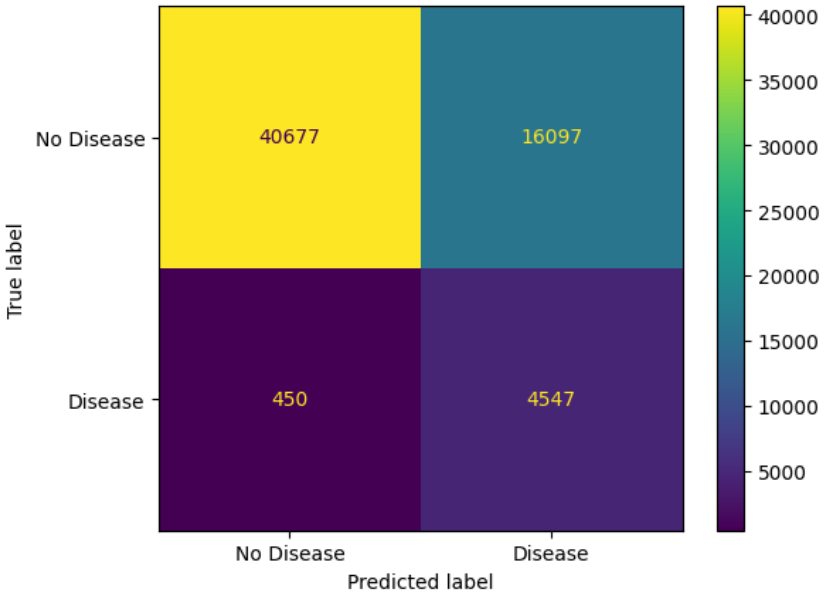


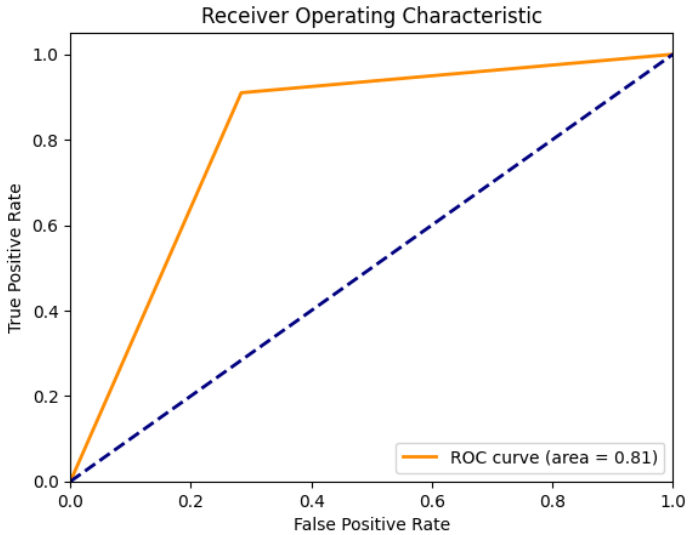


RandomForest

Es un conjunto de árboles de decisión que mejora la generalización mediante la combinación de múltiples árboles. En cuanto a sus parámetros, 'n_estimators' se refiere al número de árboles en el bosque, a mayor número de árboles suele aumentar el rendimiento hasta cierto punto. 'max_depth' controla la profundidad máxima de los árboles, ajudando al sobreajuste, en cuanto a 'min_samples_split' define el número mínimo de muestras necesarias para dividir un nodo interno, a valores más altos, se evita la creación de árboles complejos y demasiado específicos.

	precision	recall	f1-score	support
0.0	0.99	0.72	0.83	56774
1.0	0.22	0.91	0.35	4997
accuracy			0.73	61771
macro avg	0.60	0.81	0.59	61771
weighted avg	0.93	0.73	0.79	61771

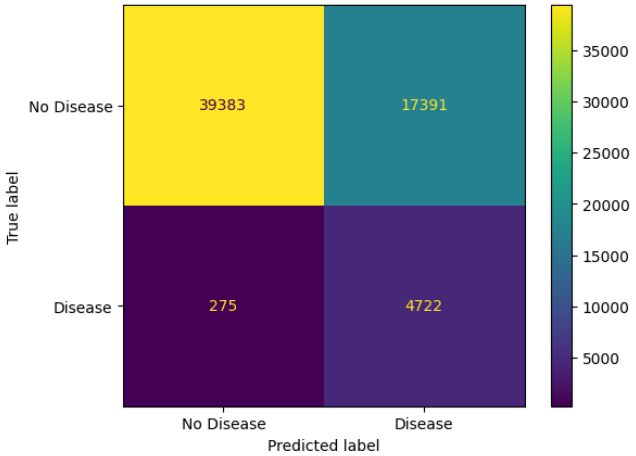


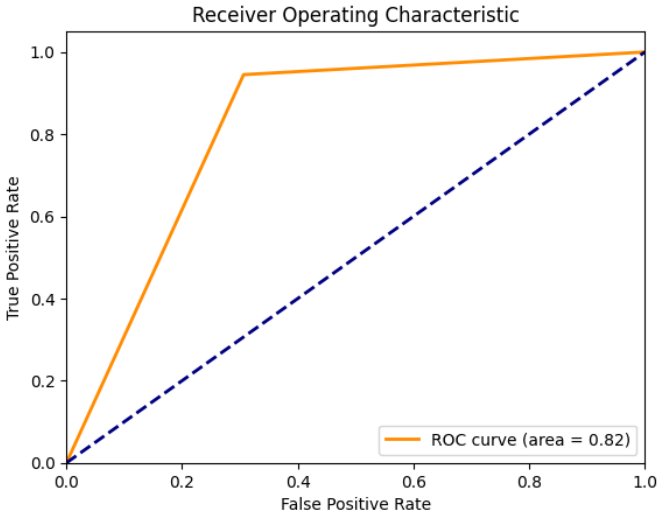


KNN

Algoritmo basado en instancias que clasifica una nueva instancia basándose en la mayoría de etiquetas de clase de los k vecinos más cercanos. Consideramos 'n_neighbors' que sería el número de vecinos a considerar, afectando a la granularidad de las fronteras de decisión. 'weights' puede ajustar la importancia de los vecinos basándose en su distancia, en cuanto a 'metric' define cómo se mide la distancia entre puntos pudiendo influir las fronteras de clasificación. Esto puede ser muy efectivo en conjuntos de datos con muchas clases, como el que se tiene después del remuestreo.

	precision	recall	f1-score	support
0.0	0.99	0.69	0.82	56774
1.0	0.21	0.94	0.35	4997
accuracy			0.71	61771
macro avg	0.60	0.82	0.58	61771
weighted avg	0.93	0.71	0.78	61771

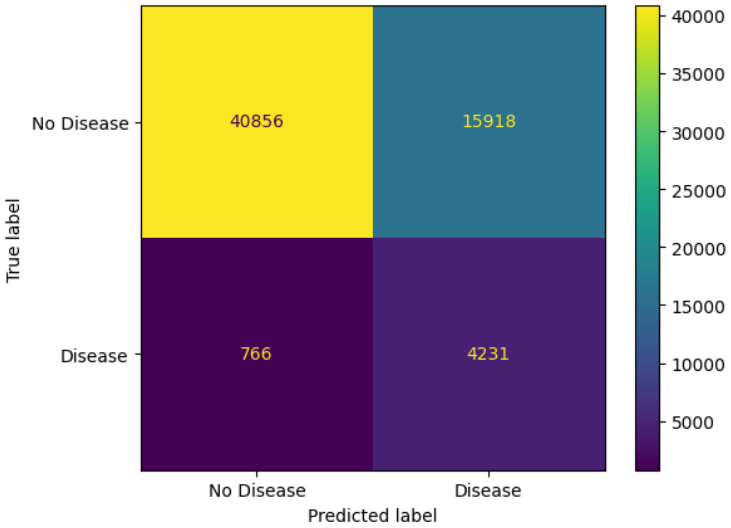


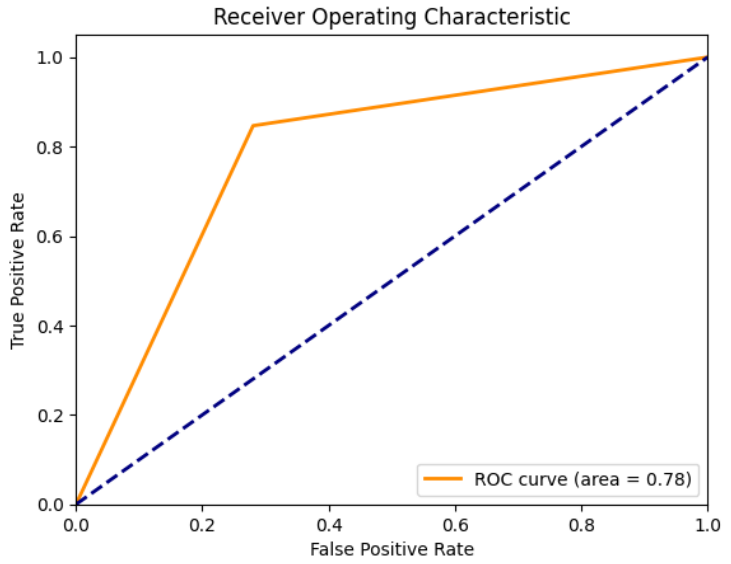


Modelo XGBoost Classifier:

Es un modelo que utiliza árboles de decisión mejorados o árboles de gradient boosting, y es conocido por su rendimiento y velocidad. 'learning_rate' indica el impacto de cada árbol, 'n_estimators' y 'max_depth' al igual que el Random Forest controla el número de árboles y su profundidad.

	precision	recall	f1-score	support
0.0	0.98	0.72	0.83	56774
1.0	0.21	0.85	0.34	4997
accuracy			0.73	61771
macro avg	0.60	0.78	0.58	61771
weighted avg	0.92	0.73	0.79	61771



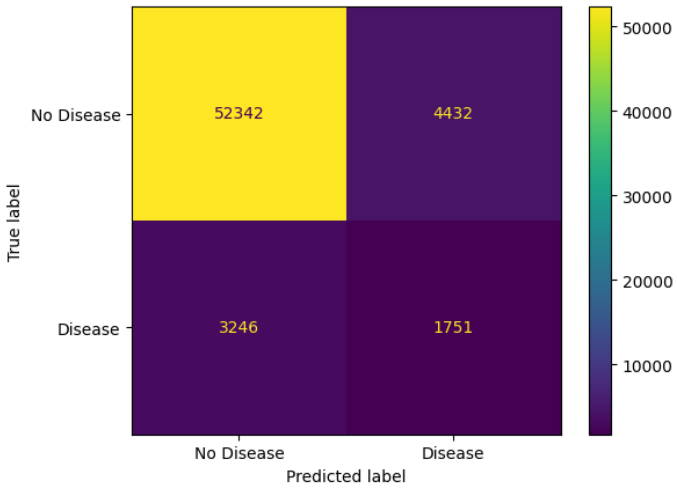


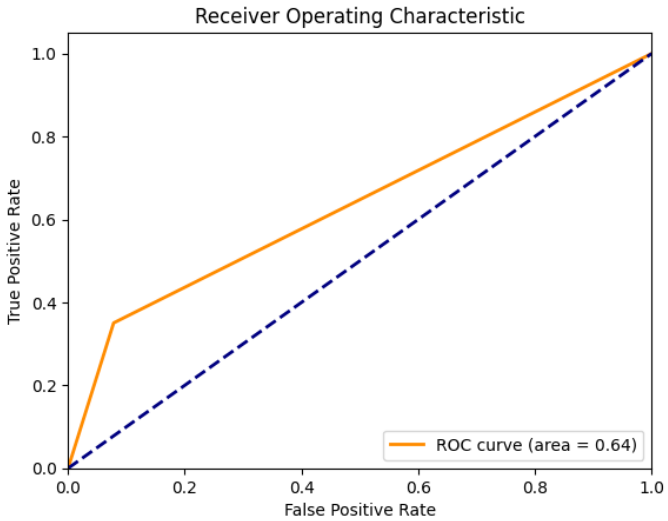
ADABOOST

Algoritmo de ensamblaje que ajusta los pesos de los clasificadores y las instancias en cada interacción para mejorar el modelo donde falló previamente. 'n_estimators' se refiere al número de modelos secuenciales a entrenar y 'learning_rate' afecta cuánto contribuye cada modelo a la versión final.

	precision	recall	f1-score	support
0.0	0.94	0.92	0.93	56774
1.0	0.28	0.35	0.31	4997

accuracy			0.88	61771
macro avg	0.61	0.64	0.62	61771
weighted avg	0.89	0.88	0.88	61771



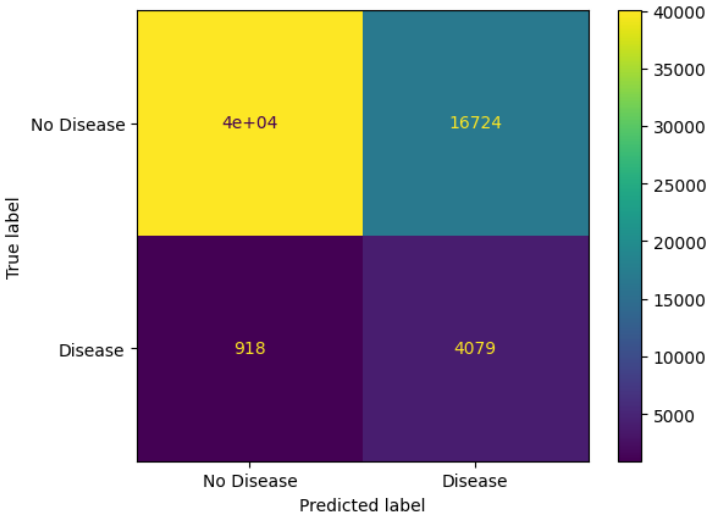


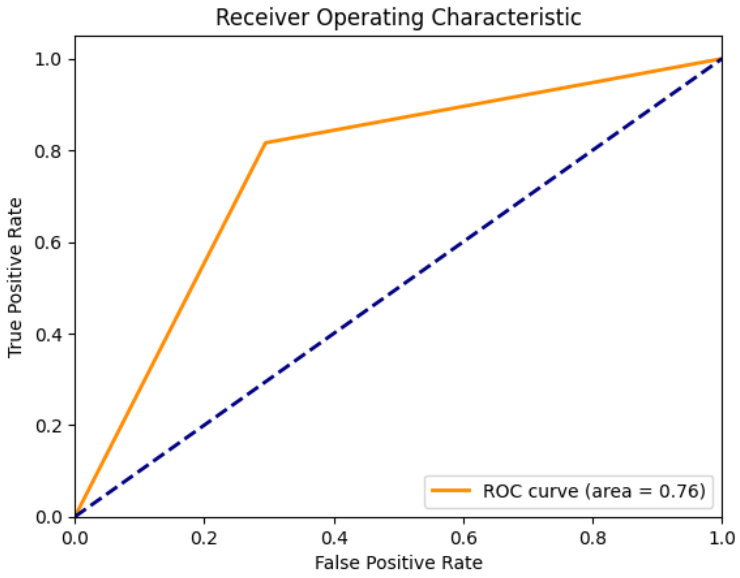
Redes Neuronales

Modelo computacional capaz de capturar relaciones no lineales complejas. En cuanto a sus parámetros, 'hidden_layer_sizes' define la arquitectura de las capas ocultas, 'activation' determina la función de activación para las neuronas, 'solver' es el algoritmo de optimización para el aprendizaje de los pesos y 'alpha' controla la regularización L2.

	precision	recall	f1-score	support

0.0	0.98	0.71	0.82	56774
1.0	0.20	0.82	0.32	4997
accuracy			0.71	61771
macro avg	0.59	0.76	0.57	61771
weighted avg	0.91	0.71	0.78	61771





Resultados

La métrica f1 fue seleccionada como el criterio principal por su capacidad de balanceo entre precisión y recall, aspectos críticos en este contexto. Es crucial identificar los pacientes que tienen una enfermedad sí, pero también asegurar que se detecten la mayoría cantidad posible de casos reales.

Modelo	F1-score (Clase 1.0)	Recall (Clase 1.0)
Naïve Bayes	0.28	0.75
Logistic Regression	0.30	0.77
Random Fores	0.35	0.91
KNN	0.35	0.94
XGBoost Classifier	0.34	0.85
ADABoost	0.31	0.35
Redes Neuronales	0.32	0.82

En cuanto a los mejores resultados obtenidos se considera:

El RandomForest destaca principalmente por su alto recall, indicando que es muy efectivo para detectar correctamente casos de enfermedad cardíaca, capturando el 91% de los casos positivos reales. Su precisión no es la más alta, pero la capacidad del modelo para minimizar los falsos negativos es crucial para llas aplicaciones médicas.

El KNN tiene el recall más elevado de todos los modelos evaluados, lo hace muy valioso para asegurar que una gran cantidad de todos los pacientes sean detectados. Esto viene con el costo de una precisión moderadamente baja, en el contexto médico en el que se haya este trabajo, la importancia del alto recall justifica este compromiso.

En cuanto al XGBoost, ofrece un mejor balance entre recall y precisión. Guarda un equilibrio más conservador entre la detección de casos positivos y mantener un número razonable de falsos positivos.

Conclusiones

La combinación de una preparación de datos cuidadosa con la selección estratégica de las técnicas de modelado ha permitido desarrollar un enfoque para predecir el riesgo de enfermedades cardiovasculares.

Se han aboradod varios aspectos críticos para la predicción correcta de enfermedades cardiovasculares, utilizando técnicas que fueron críticas para la mejora del balance de los datos. Se comenzó con una adecuada preparación de los datos, estudiando distintos tipos de

encoding, resampling y escalado, y seleccionando los que proporcionaban una mejora de la calidad y representatividad de los datos.

Se implementaron la técnica del Ordinal Encoding para conertir las variables categóricas en formatos numéricos que los modelos pudiesen interpretar más eficazmente. Útil sobre todo en datos de las categorías de edad, o el nivel de salud física general. Además, la aplicación del MinMaxScaler fue importante para asegurarse de que todas las características tuvieran el mismo rango. Dado el desbalance de las clases en el dataset, se utilizó técnicas de resampling, SMOTE y Tomek Links para reducir aumentar la clase minoritaria y aumentar la clase mayoritara.

En cuando a los modelos, se evaluaron una selección total de siete modelos distintos, cada uno con sus parámetros optimizados, centrando su evaluación en la métrica de F1-score. Se llevó a cabo un análisis de los resultados y se encontraron los modelos particularmente buenos para este caso de estudio.