Plan de EDA para Nuestro Proyecto

Podemos combinar estos enfoques para desarrollar un plan de EDA exhaustivo:

Análisis univariado y bivariado: Examinar cada variable individualmente y en relación con la variable objetivo para entender las distribuciones y posibles correlaciones.

Visualización de datos: Usar gráficos como histogramas para variables continuas, gráficos de barras para categóricas, y mapas de calor para correlaciones.

Análisis de valores faltantes y outliers: Determinar cómo manejar los datos faltantes y los valores atípicos basados en su impacto en el análisis.

Ingeniería de características: Considerar la creación de nuevas variables basadas en el conocimiento del dominio y la correlación observada entre las variables.

Ejemplos Útiles:

K-Nearest Neighbors (KNN) - Útil para clasificación basada en la similitud de las características del paciente con casos históricos.

Support Vector Machine (SVM) - Efectivo para encontrar el hiperplano óptimo que separa las clases (enfermedad/no enfermedad).

Decision Trees - Proporciona una estructura clara de decisión basada en los atributos más influyentes.

Random Forests - Mejora la predicción de árboles de decisión a través de un enfoque de ensamble, ideal para manejar datasets con muchas características.

Enfoques Sugeridos:

XGBoost: Este algoritmo es excelente para tareas de clasificación y es conocido por su rendimiento superior, especialmente en datasets estructurados como el tuyo.

Logistic Regression: Ideal para problemas de clasificación binaria, como la predicción de la presencia o ausencia de enfermedades cardíacas.

Stacking: Combina múltiples modelos de predicción para mejorar la precisión. Sería útil si decides experimentar con un enfoque de ensamble.

Naive Bayes: Un enfoque rápido y eficaz para la clasificación, especialmente útil cuando las dimensiones del dataset son altas en relación con el tamaño de la muestra.