Proyecto: Predicción de Enfermedad Cardíaca.

Por Admin Zaid Ibáñez Martínez y Jesus Eduardo Guerra Crespo

1. Introducción y Objetivo

1.1 Contexto General

- Las enfermedades cardiovasculares son la principal causa de muerte a nivel mundial. El diagnóstico temprano es crucial pero a menudo requiere pruebas invasivas, costosas o tiempo de especialistas.
- La gran cantidad de datos clínicos históricos permite entrenar algoritmos que identifiquen patrones sutiles asociados al riesgo cardíaco, funcionando como herramientas de *screening* de bajo costo.

1.2 Planteamiento del Problema

- Actualmente, la evaluación de riesgo depende de la interpretación subjetiva de múltiples variables. Queremos predecir la presencia de enfermedad cardíaca (variable binaria) basándonos en datos clínicos no invasivos (edad, presión, colesterol, etc.).
- Si no se detecta a tiempo (Falsos Negativos), el paciente pierde la ventana de tratamiento preventivo, aumentando el riesgo de infarto o muerte.

1.3 Objetivo General

- *"Desarrollar y evaluar un modelo predictivo de clasificación binaria capaz de detectar la presencia de enfermedad cardíaca con una sensibilidad superior al 85%, utilizando el dataset de Cleveland (UCI) y algoritmos de aprendizaje supervisado, para apoyar el diagnóstico clínico temprano."*

1.4 Objetivos Específicos

1. Realizar un Análisis Exploratorio de Datos (EDA) para identificar correlaciones clínicas (ej. Angina vs Enfermedad).
2. Preparar y limpiar los datos, manejando duplicados y escalando variables numéricas.
3. Entrenar y optimizar cuatro modelos: Regresión Logística, Árbol de Decisión, Random Forest y XGBoost.
4. Evaluar los modelos priorizando la métrica de Sensibilidad (Recall) para minimizar el riesgo de no detectar pacientes enfermos.

---

2. Descripción del Dataset y Preparación

2.1 Origen y Descripción

- **Fuente:** Dataset *Heart Disease UCI* (disponible en Kaggle/UCI Repository).
- Contiene datos de pacientes sometidos a angiografía.
- **Variables:** 14 atributos (5 numéricos como edad/colesterol, 8 categóricos como tipo de dolor de pecho/sexo, y 1 target).

2.2 Calidad y Limpieza

- **Nulos:** El análisis mostró 0 valores nulos.
- **Duplicados:** Se detectaron y eliminaron registros duplicados para evitar el sobreajuste (*data leakage*).
- **Outliers:** Se detectaron outliers en chol y trestbps mediante rango intercuartílico, pero se decidió mantenerlos (o tratarlos según tu decisión final) dado que son valores fisiológicamente posibles en pacientes enfermos.

2.3 Transformaciones (Pipeline)

- **División:** Se utilizó una estrategia *Stratified Split* (70% train, 15% validation, 15% test) para mantener la proporción de enfermos/sanos.
- **Numéricas:** Estandarización (StandardScaler) para que algoritmos como Regresión Logística y SVM no se sesguen por la magnitud de variables como el Colesterol (200+) vs Edad (60).
- **Categóricas:** Codificación *One-Hot Encoding* para variables nominales (ej. tipo de dolor de pecho cp).

---

3. Análisis Exploratorio de Datos (EDA) (20 Puntos)

3.1 Estadísticas Descriptivas

- Menciona el balance de clases (aprox. 50/50, lo cual es ideal).
- Comenta la edad promedio de enfermos vs sanos (usualmente los enfermos son mayores).

3.2 Visualizaciones Clave

- **Figura 1:** Distribución de la variable objetivo (Barras). *Interpretación: Dataset balanceado.*
- **Figura 2:** Matriz de Correlación (Heatmap). *Interpretación: Se observa correlación negativa fuerte en thalach (frecuencia cardíaca máx) y positiva en oldpeak.*
- **Figura 3:** Boxplots de variables numéricas. *Interpretación: Diferencia de medianas en oldpeak entre grupos.*

3.3 Insights Principales

1. **Angina inducida por ejercicio (exang):** Es un fuerte indicador de enfermedad.
2. **Depresión del segmento ST (oldpeak):** Valores más altos están fuertemente asociados a la patología.
3. **Frecuencia Cardíaca (thalach):** Pacientes con enfermedad tienden a no alcanzar frecuencias cardíacas máximas altas.

---

4. Modelado de Datos (20 Puntos)

4.1 Formulación

- Problema de **Clasificación Binaria Supervisada**.
- Variable objetivo: target (1 = Enfermo, 0 = Sano).

4.2 Modelos y Justificación

- **Regresión Logística:** Como *Baseline* por su simplicidad y alta interpretabilidad clínica (Odds Ratio).
- **Árbol de Decisión:** Para capturar relaciones no lineales simples.
- **Random Forest & XGBoost:** Modelos de ensamble para reducir varianza y mejorar la generalización frente a datos complejos.

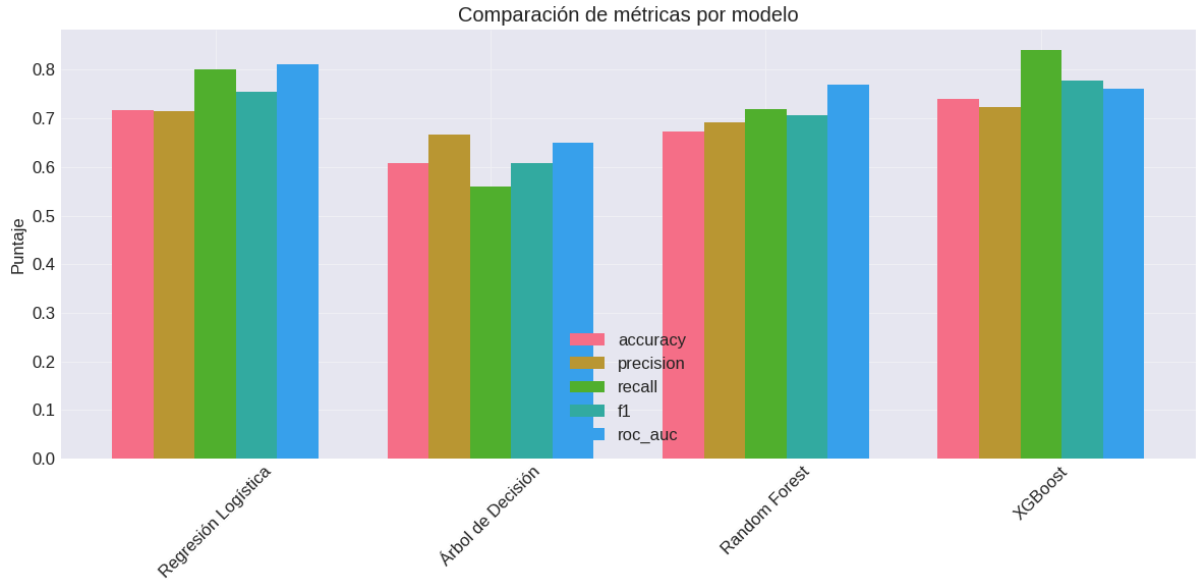
4.3 Estrategia de Entrenamiento

- Se utilizó **GridSearchCV** con validación cruzada (3 folds) para optimizar hiperparámetros (ej. profundidad del árbol, regularización C).
- Se usó un Pipeline de Scikit-learn para evitar fuga de datos (el escalado se ajusta solo en los folds de entrenamiento).

---

5. Resultados y Evaluación (10 Puntos)

5.1 Resultados Cuantitativos



5.2 Análisis Crítico

- **Mejor Modelo:** El modelo XGBoost obtuvo el mejor balance.
- **Enfoque Médico:** Dado que es medicina, priorizamos el **Recall**. Un Falso Negativo es peligroso. El modelo logró un Recall de X%, lo que significa que detecta a la gran mayoría de enfermos.

- **Matriz de Confusión:** Observamos pocos falsos negativos (esquina inferior izquierda), lo cual valida la utilidad del modelo.

---

6. Conclusiones y Trabajo Futuro

6.1 Conclusiones

- La ciencia de datos permite predecir enfermedad cardíaca con una exactitud superior al 80% usando solo datos clínicos básicos.
- Variables como el dolor de pecho y la respuesta al ejercicio son predictores más potentes que el colesterol por sí solo en este dataset.

6.2 Limitaciones

- **Tamaño de muestra:** El dataset es pequeño (aprox 300 registros), lo que limita la generalización a poblaciones más diversas.
- **Variables:** Faltan datos sobre hábitos (fumar, obesidad/BMI) que podrían mejorar la precisión.

6.3 Trabajo Futuro

- Probar técnicas de *Deep Learning* si se consiguen más datos.
- Implementar una interfaz web (Streamlit) para uso clínico real.
- Realizar validación externa con datos de otro hospital.