Proyecto de predición de enfermedades cardíacas

Daniel Arias

30 de marzo de 2025

1. Entendimiento del Problema y de los Datos (EDA)

1.1. Formulación del problema

Formulación del problema

El problema principal consiste en la necesidad de mejorar la eficiencia en la atención médica y reducir los costos hospitalarios mediante la detección temprana de enfermedades cardíacas. El objetivo es construir un modelo predictivo capaz de identificar correctamente la presencia de enfermedad cardíaca en pacientes a partir de variables clínicas como edad, presión arterial, colesterol, tipo de dolor torácico, entre otras. Las principales preguntas que guían este proyecto son: ¿Es posible predecir con precisión la presencia de enfermedad cardíaca a partir de datos clínicos estructurados? ¿Puede un modelo automatizado apoyar decisiones médicas tempranas? Este problema es especialmente relevante en el campo de la ingeniería biomédica y la ciencia de datos aplicada a la salud, ya que el diagnóstico oportuno de enfermedades cardiovasculares es clave para reducir la morbilidad, los costos de tratamiento y la sobrecarga en los sistemas de salud.

Para el desarrollo del proyecto se empleó el set de datos generado por (Janosi, Steinbrunn, Pfisterer, y Detrano, 1988)

2. Exploratory Data Analysis (EDA)

2.1. EDA inicial

Descripción de variables

Dentro del data set a procesar se encuentran las siguientes variables:

- **age**: Edad del paciente en años.
  Variable continua, importante por su relación directa con el riesgo cardiovascular.
- **sex**: Sexo biológico (1 = hombre, 0 = mujer).
  Permite analizar diferencias de riesgo entre géneros.
- **cp**: Tipo de dolor torácico (1 = angina típica, 2 = angina atípica, 3 = dolor no anginoso, 4 = asintomático).
- **trestbps**: Presión arterial en reposo (mm Hg).
  Indicador común de hipertensión.
- **chol**: Colesterol sérico total (mg/dl).
  Nivel elevado es un factor de riesgo cardiovascular.
- **fbs**: Glucosa en ayunas ¿120 mg/dl (1 = sí, 0 = no).
  Relacionado con la presencia de diabetes.
- **restecg**: Resultados del electrocardiograma en reposo (0 = normal, 1 = anormalidad en ST-T, 2 = hipertrofia ventricular izquierda).
- **thalach**: Frecuencia cardíaca máxima alcanzada durante la prueba de esfuerzo.
  Refleja condición funcional del corazón.
- **exang**: Presencia de angina inducida por ejercicio (1 = sí, 0 = no).
- **oldpeak**: Depresión del segmento ST inducida por el ejercicio respecto al reposo.
  Relacionado con isquemia miocárdica.

- **slope**: Pendiente del segmento ST en el pico del ejercicio (1 = ascendente, 2 = plana, 3 = descendente).
- **ca**: Número de vasos coronarios principales coloreados por fluoroscopía (0 a 3).
  Evalúa obstrucción arterial.
- **thal**: Resultados de la prueba de talio (3 = normal, 6 = defecto fijo, 7 = defecto reversible).
- **num**: Diagnóstico de enfermedad cardíaca (0 = ausencia; 1–4 = presencia con distintos grados).
  Esta es la variable objetivo.

Conclusiones generales del EDA inicial

1. El dataset tiene 303 registros con 13 variables importantes para el diagnóstico de posibles enfermedades cardíacas.
2. se encluye una columna adicional **num** que identifica la presencia de una enfermedad cardíaca y la clasifica por grado de severidad.
3. Se encontraron valores nulos en las columnas **ca** y **thal** 4 y 2 casos respectivamente que para el total de todos los datos representan menos del 2 % por lo que se podrían eliminar sin el riesgo de perder mayor información.
4. No se tienen filas ni duplicadas completamente ni duplicados falsos.

2.2. EDA profundizado

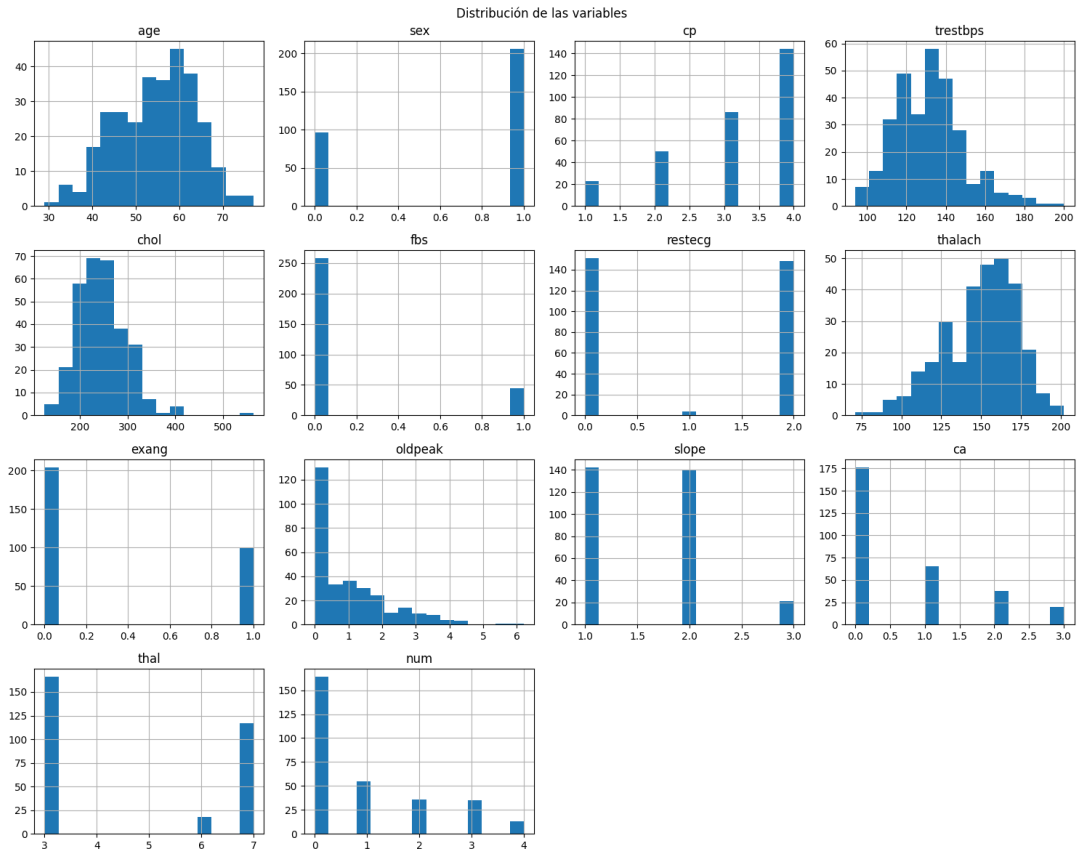

Figura 1: Histograma de cada una de las columnas pertenecientes al data set

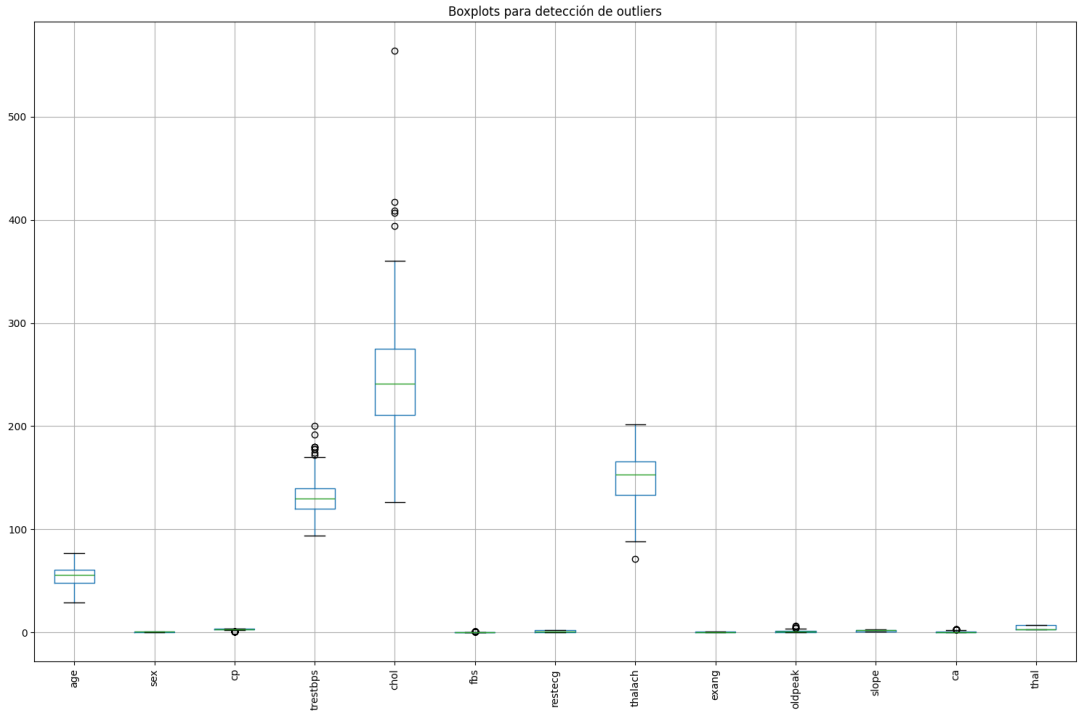

Figura 2: Diagrama de caja de las variables numéricas presentes en el data set

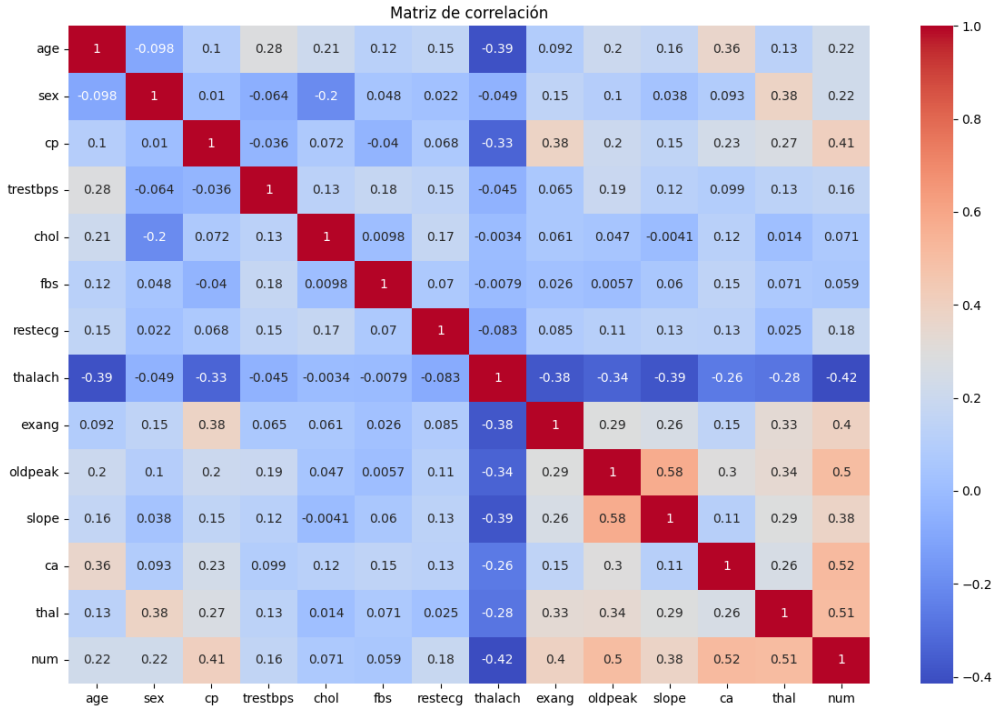

Figura 3: Matriz de correlación entre todas las variables presentes en el data set

- A partir de los histogramas se encuentra que las variables exang, fbs, y sex están altamente desbalanceadas - (mayoría de valores 0).
- Se encontraron outliers en variables como chol, trestbps, y oldpeak, lo cual es clínicamente plausible pero debe validarse médicamente o con expertos.
- La matriz de correlación muestra una correlación negativa entre thalach y exang (esperado: a mayor frecuencia cardíaca, menos probabilidad de angina inducida por ejercicio).

3. Data Wrangling

Dentro del procesamiento de los datos se realizaron las siguientes transformaciones

- **Manejo de valores nulos:** Dado el bajo porcentaje de registros con datos faltantes, se optó por eliminar dichas filas utilizando el método `dropna()`, asegurando así la integridad de los datos sin comprometer el tamaño del conjunto.
- **Verificación del dataset:** Se confirmó que no existían valores nulos tras la limpieza. El conjunto final quedó compuesto por 297 observaciones completas y 13 variables predictoras, además de la variable objetivo `num`.
- **Estandarización de nombres de columnas:** Todos los nombres de columnas fueron convertidos a minúsculas para mantener consistencia y facilitar su manipulación programática.
- **Exportación de datos limpios:** El conjunto de datos procesado fue guardado como `clean_medical_records.csv`, quedando listo para su uso en la etapa de modelado.

Referencias

Janosi, A., Steinbrunn, W., Pfisterer, M., y Detrano, R. (1988). *Heart disease.* UCI Machine Learning Repository. (Disponible en: `https://doi.org/10.24432/C52P4X`)