Tarea I: Heart Disease Dataset

Enlaces de interés

Enlace al repositorio: Link
En este enlace se encuentra toda la información relacionada con la base de datos, incluyendo la descripción detallada de las variables. También se incluye un baseline model performance que puede servir como referencia para evaluar los resultados obtenidos en sus análisis.

Enlace a kaggle: Link
Este enlace los lleva al datacard de Kaggle, donde podéis explorar ideas y enfoques desarrollados por otros usuarios. Es una excelente fuente de inspiración para abordar el dataset desde diferentes perspectivas.

Descripción de la base de datos

La base de datos *Heart Disease* fue obtenida del repositorio de *UCI Machine Learning*. Contiene un total de 303 observaciones y 14 variables, de las cuales 5 son cuantitativas y 9 categóricas. A continuación, se presentan los detalles de estas variables:

- **Edad:** edad del paciente (Cuantitativa discreta)
- **Sexo:** Sexo del paciente (Categórica con dos niveles: Femenino y Masculino)
- **Cp:** Tipo de dolor de pecho (Categórica con 4 tipos)
    - **Tipo 1:** Angina típica
    - **Tipo 2:** Angina Atípica
    - **Tipo 3:** Dolor no anginal
    - **Tipo 4:** Asintomático
- **Trestbps:** Presión arterial en reposo en mm Hg al ingreso al hospital (V. continua)
- **Chol:** Colesterol sérico en mg /dl (V. continua)
- **Fbs:** azúcar en sangre en ayunas > 120 mg/dl (categórica con 2 niveles: verdadero, falso)
- **Restecg:** Resultados electrocardiográficos en reposo (Categórica con 3 niveles)
    - **Nivel 0:** Normal
    - **Nivel 1:** Tener anormalidad de onda ST-T (inversiones de onda T y / o elevación o Depresión de ST de > 005 mV)
    - **Nivel 2:** Muestra hipertropía ventricular izquierda probable o definitiva según los criterios de Estes.
- **Thalach:** Frecuencia cardiaca máxima alcanzada (V. continua)

- **Exang:** Angina inducida por el ejercicio (Categórica con dos niveles: Si, No)
- **Oldpeak:** Depresión del ST inducida por el ejercicio en relación con el descanso (V. continua)
- **Slope:** La pendiente del segmento ST de ejercicio pico (categórica con 3 niveles)
    - **Valor 1:** ascenso
    - **Valor 2:** plano
    - **Valor 3:** descenso
- **Ca:** Numero de vasos principales (0-3) coloreados por fluoroscopia (Categórica con 4 niveles: 1-2-3-4)
- **Thal:** El estado del corazón según la prueba de Thallium (Categórica con 3 niveles)
    - **N =** normal;
    - **DF =** defecto fijo
    - **DR =** defecto reversible
- **Num :** Diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfica) (Categórica con 4 niveles)
    - **Valor 0:** < 50% de estrechamiento del diámetro. No presenta enfermedad
    - **Valor 1:** > 50% de estrechamiento del diámetro. Presenta enfermedad del corazón.

Guía para la realización del ejercicio

Se deberá realizar un preprocesamiento del dataset, prestando especial atención a los ***datos faltantes***, identificando el tipo de variable y calculando el porcentaje de valores ausentes. Además, es fundamental analizar el tipo de datos faltantes para determinar la mejor estrategia de tratamiento.

Asimismo, también se deberá realizar un análisis exploratorio de datos (Exploratory Data Analysis o EDA) para comprender el comportamiento del dataset, las relaciones entre las variables y, en particular, cómo estas se asocian con la variable objetivo, Num.

Finalmente, deberán realizar el modelado de la variable objetivo (Num) utilizando las variables seleccionadas e identificadas como influyentes durante el análisis exploratorio. Será fundamental evaluar el desempeño de los modelos implementados mediante métricas adecuadas para medir el error, como la precisión, el recall, el F1-score o la curva ROC-AUC, dependiendo del enfoque seleccionado.

Se espera que justifiquen la elección del modelo, expliquen los resultados obtenidos y propongan posibles mejoras en caso de que los modelos no alcancen el rendimiento esperado.

A continuación, se presentan algunas ideas para guiar el EDA:

Análisis Univariante:

- ¿Cómo se distribuyen las variables continuas como Trestbps, Chol, Thalach y Oldpeak?
- ¿Existen valores extremos o outliers en estas variables?
- ¿Hay diferencias en la proporción de tipos de dolor de pecho (Cp) entre pacientes con y sin enfermedad?

Análisis Multivariante:

- ¿Cómo varía la frecuencia cardiaca máxima alcanzada (Thalach) según el sexo y la presencia de la enfermedad?
- ¿El tipo de dolor de pecho (Cp) es un indicador relevante para la enfermedad?
- ¿Qué relación existe entre la variable Oldpeak y el diagnóstico (Num)?
- ¿Los resultados electrocardiográficos (Restecg) aportan información importante sobre la variable objetivo?
- ¿Los niveles de colesterol (Chol) están relacionados con la frecuencia de la enfermedad en diferentes grupos de edad?
- ¿Existe una relación entre la edad de los pacientes y el diagnóstico de enfermedad cardíaca (`Num`)?
- ¿Hay diferencias significativas en las tasas de enfermedad cardíaca entre hombres y mujeres dentro del rango de edades predominante?
- ¿Cuál es el umbral de presión arterial que parece estar más asociado con la presencia de enfermedad cardíaca?
- ¿La combinación de colesterol alto y presión arterial elevada aumenta significativamente el riesgo de enfermedad?
- ¿Cómo se distribuyen las categorías de `Restecg` según la variable objetivo?
- ¿Las personas con resultados electrocardiográficos normales presentan menor incidencia de enfermedad en comparación con las otras categorías?

Aspectos Avanzados:

- ¿Es necesario transformar alguna variable continua para mejorar su distribución (por ejemplo, usando logaritmos o estandarización)?

- ¿Cómo afecta la interacción entre variables categóricas (por ejemplo, Sexo y Cp) a la probabilidad de enfermedad?
- ¿Existen patrones claros de datos faltantes en alguna variable? ¿Están relacionados con otras variables del dataset?
- ¿Qué diferencias clave se observan en los pacientes diagnosticados con enfermedad cardíaca en comparación con los que no la tienen?

Preguntas sobre el modelo:

- ¿Qué variables deberían seleccionarse como predictores principales en un modelo basado en el análisis exploratorio?
- ¿Qué transformación de las variables puede ser útil para mejorar la performance del modelo?
- ¿Se justifica el balanceo de los datos si la distribución de la variable objetivo está muy sesgada?

Estas preguntas sirven como guía inicial, pero tienen libertad para explorar otros aspectos del dataset y proponer análisis adicionales que consideren relevantes. ¡El objetivo es profundizar en el conocimiento de los datos y sacar conclusiones significativas!

Guía para la entrega

La entrega deberá realizarse en un **Jupyter Notebook** y debe ser completamente reproducible. El trabajo deberá estructurarse en las siguientes partes:

1. Descripción del problema y objetivos

Incluir una introducción clara que explique el problema a resolver, los objetivos específicos del análisis y las preguntas clave a responder a lo largo del proyecto.

2. Lectura y descripción del dataset

Leer los datos e incluyan una descripción detallada del dataset:

- Dimensiones (número de observaciones y variables).
- Tipos de variables (cuantitativas, categóricas).
- Significado de cada columna y posibles insights iniciales.

3. Limpieza de datos (Data Cleaning)

Realizar el preprocesamiento necesario, incluyendo:

- Identificación de datos faltantes.
- Análisis del porcentaje de valores ausentes en cada variable.

- Clasificación del tipo de datos faltantes (MCAR, MAR, MNAR).
- Estrategias utilizadas para el tratamiento de los valores ausentes y justificación de las mismas.
- Otros tipo de transformaciones y análisis si son necesarios

4. Análisis Exploratorio de Datos (EDA)

Análisis exploratorio de los datos:

- Visualizaciones y estadísticas descriptivas para comprender las distribuciones de las variables.
- Identificación de relaciones clave entre variables y con la variable objetivo (*Num*).
- Respuestas a las preguntas planteadas inicialmente y cualquier insight adicional obtenido.

5. Modelado

Construir modelos predictivos para la variable objetivo (*Num*), basándose en las variables identificadas como relevantes durante el EDA:

- Justificar la selección de las variables y el tipo de modelo.
- Realizar un análisis comparativo de diferentes modelos si es posible.

6. Evaluación

Evaluar el rendimiento de los modelos utilizando métricas adecuadas. Es importante la interpretación de los resultados obtenidos y reflexionar sobre su validez y limitaciones.

7. Conclusiones

Resuman los hallazgos principales, incluyendo:

- Las preguntas iniciales respondidas.
- Las variables clave identificadas.
- Los modelos más efectivos y las métricas de evaluación.
- Posibles áreas de mejora y análisis futuros.

El notebook debe ser claro, bien documentado y contener explicaciones detalladas de cada paso. Se valorará la calidad del código, el uso eficiente de librerías y la claridad en la interpretación de los resultados.