Laboratorio: EDA

Objetivos

Mediante este laboratorio se pretende que aplique los conocimientos adquiridos en los temas trabajando con un conjunto de datos médico.

Descripción

El conjunto de datos con el cual vamos a trabajar se encuentra en el siguiente link:

https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.names

https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv

En esta primera actividad se trata de familiarizarse con los pasos generales a realizar para generar un modelo de aprendizaje automático. Se empieza con el análisis exploratorio de datos (EDA), los pasos que debe desarrollar son:

1. Definición del Problema y Carga de Datos:
    a. Definición del Problema: Entiende claramente el problema que estás tratando de resolver o la pregunta que estás tratando de responder.
    b. Carga de Datos: Importa los datos en tu entorno de análisis. Esto generalmente se hace utilizando bibliotecas como pandas en Python.
2. Inspección Inicial de los Datos:
    a. Revisión de las Primeras Filas: Usa funciones como head() para ver las primeras filas del DataFrame.
    b. Dimensiones del DataFrame: Utiliza shape para conocer la cantidad de filas y columnas.
    c. Información General: Usa info() para obtener información sobre el tipo de datos y valores faltantes.
    d. Descripciones Estadísticas: Aplica describe() para obtener estadísticas descriptivas básicas de las variables numéricas.
3. Limpieza de Datos:
    a. Manejo de Valores Faltantes: Identifica y decide cómo tratar los valores faltantes (eliminarlos, imputarlos, etc.).
    b. Corrección de Tipos de Datos: Asegúrate de que cada columna tenga el tipo de dato adecuado (e.g., convertir cadenas a fechas).

c. Detección y Corrección de Errores: Busca y corrige errores en los datos, como valores fuera de rango o duplicados.

4. Análisis Univariado:
    a. Distribución de Variables: Analiza la distribución de las variables individuales usando histogramas, box plots, etc.
    b. Medidas de Tendencia Central y Dispersión: Calcula medidas como la media, mediana, moda, desviación estándar, y rango.
5. Análisis Bivariado:
    a. Relaciones entre Variables: Examina las relaciones entre pares de variables. Para variables numéricas, usa scatter plots o correlaciones. Para variables categóricas, usa tablas de contingencia o box plots.
6. Análisis Multivariado:
    a. Relaciones entre Múltiples Variables: Usa técnicas como el análisis de componentes principales (PCA) o el análisis de correlación múltiple para entender las relaciones entre múltiples variables.
    b. Visualizaciones Complejas: Utiliza gráficos como heatmaps, pair plots, y gráficos 3D para visualizar relaciones entre más de dos variables.
7. Identificación de Outliers:
    a. Detección de Valores Atípicos: Usa box plots, z-scores, o técnicas de clustering para identificar outliers y decide cómo manejarlos.
8. Visualización de Datos:
    a. Gráficos y Diagramas: Utiliza bibliotecas de visualización (como Matplotlib, Seaborn en Python) para crear gráficos que faciliten la interpretación de los datos.
    b. Dashboards: Considera crear dashboards interactivos si estás trabajando con conjuntos de datos grandes o si necesitas presentar tus hallazgos a otros.

Se pide: Un jupyter notebook explicando cada resultado obtenido, asimismo, cada resultado, diagrama, grafico o tabla generado deberá tener su explicación e interpretación. Si se detecta copia, el trabajo se califica con CERO.

Se presenta el jupyter notebook empaquetado en formato .RAR subido al SIA