Universisdad Técnica Nacional

Minería de datos - ISW 911

Laboratorio # 4

Asignación Grupal integrantes (investigación)

Realice un proceso EDA, Utilice el lenguaje Python para realizar la asignación, Utilice el conjunto de datos que se le proporciona. En su aplicación lea el conjunto de datos desde GIT, el archivo debe ser público para poder realizar la revisión.

Resuelva las siguientes tareas:

1. Carga y comprensión inicial de los datos

- **Cargar el dataset**
- **Verificar el formato**
- **Explorar las primeras filas**
- **Comprobar el tipo de cada columna**
- **Observar la cantidad de filas y columnas**

2. Limpiar y preparar los datos

- **Identificar y tratar valores nulos**
- **Eliminar duplicados**
- **Revisar valores atípicos**
- **Convertir tipos de datos**
- **Cambiar nombre de columnas si procede**
- **Tratar con variables categóricas**

3. Análisis univariado

- **Estadísticas descriptivas**: Obtener una visión general de las distribuciones de las columnas numéricas
- **Histograma y distribución de datos**: Usar histogramas, diagramas de caja (boxplots) y otras visualizaciones para comprender cómo se distribuyen las variables numéricas.
- **Comprobación de asimetría y normalidad**: Usar medidas como el coeficiente de asimetría y la kurtosis para verificar la forma de la distribución.

4. Análisis bivariado

- **Correlación**: Utilizar la matriz de correlación (usando .corr() en pandas) para identificar relaciones entre variables numéricas. Crear un mapa de calor (heatmap) para visualizar las correlaciones.
- **Gráficos de dispersión**: Usar gráficos de dispersión (scatterplot) para identificar relaciones entre dos variables numéricas.
- **Boxplots y análisis por categorías**: Si una variable es categórica, puedes usar boxplots para analizar cómo se distribuyen las variables numéricas dentro de cada categoría.

5. Análisis de distribuciones y sesgos

- **Comprobación de sesgo en los datos**: Utiliza diagramas de caja o distribuciones para detectar la presencia de sesgo o desbalance en las variables.
- **Transformaciones de variables**: Si es necesario, transforma las variables para corregir sesgos (por ejemplo, logaritmos o raíces cuadradas en variables sesgadas).

6. Detección de valores atípicos

- **Boxplots**: Utiliza diagramas de caja para detectar posibles valores atípicos (outliers) en tus datos.
- **Análisis visual de gráficos de dispersión**: Si trabajas con datos numéricos, observa si hay puntos que se alejan significativamente de la tendencia general.
- **Z-scores o IQR**: Aplica la técnica de puntuaciones Z (para distribuciones normales) o el rango intercuartílico (IQR) para detectar y gestionar los outliers.

7. Visualización de los datos

- **Gráficos de barras**: Para datos categóricos, usa gráficos de barras para visualizar la frecuencia o la proporción de cada categoría.
- **Gráficos de líneas**: Si los datos son temporales o secuenciales, usa gráficos de líneas para analizar tendencias a lo largo del tiempo.
- **Diagramas de dispersión**: Para explorar relaciones entre variables continuas.
- **Gráficos de densidad**: Para observar la distribución de los datos en comparación con un histograma.

8. Normalización de datos

- **Normalización, estandarización o escalado:** determine si el conjunto de dato requiere alguna de las técnicas indicadas justifique su respuesta.

9. Resúmenes y conclusiones

- **Resumen de hallazgos**: Documenta los hallazgos clave durante el EDA, como relaciones interesantes entre variables, distribuciones inusuales, la calidad de los datos y las variables más importantes.
- **Preparación para modelado**: Después del EDA, documente las decisiones de qué pasos seguir para el modelado, como la selección de variables, transformación de datos, o el manejo de datos faltantes.