Práctica de Laboratorio - Unidad 2: Introducción a la Limpieza de Datos

Materia: Introducción a la Ciencia de Datos

Unidad 2: Procesamiento y Limpieza de Datos

Título de la Práctica: Limpieza de una Base de Datos Ensuciada

Nombre:

Día y horario de la materia:

Objetivo:

Desarrollar habilidades en el preprocesamiento de datos, incluyendo la identificación y tratamiento de valores faltantes, datos duplicados, y formatos inconsistentes en una base de datos.

Instrucciones:

1. **Recepción de la Base de Datos**
   Descargue el archivo de la base de datos "ensuciada" proporcionado por el profesor. Este archivo ha sido alterado con errores comunes, como valores duplicados, valores faltantes (NaNs), y errores en el formato de algunas columnas.
2. **Análisis Inicial de la Base de Datos**
   Antes de comenzar a limpiar la base de datos, debe realizar un análisis preliminar para comprender la naturaleza y distribución de los errores. Para ello, siga los siguientes pasos:
   - Mostrar un resumen estadístico de los datos.
   - Calcular el porcentaje de valores faltantes por columna.
   - Identificar si hay filas duplicadas.
   - Analizar los tipos de datos de las columnas y si son consistentes con el contenido esperado.
3. **Limpieza de Datos**
   Debe realizar las siguientes tareas de limpieza en la base de datos:
   - **Eliminación o imputación de valores faltantes:** Justificar si decide eliminar filas/columnas con NaNs o utilizar técnicas de imputación (relleno de valores).
   - **Eliminación de duplicados:** Identificar filas duplicadas y eliminarlas.

- **Corrección de tipos de datos:** Asegurarse de que las columnas tengan tipos de datos adecuados (por ejemplo, números como int o float, fechas como datetime, textos como str).
- **Corrección de valores "invalid":** Corregir los valores que fueron etiquetados con cadenas incorrectas como 'bbb'.

4. **Documentación y Reporte**
   Luego de limpiar los datos, redacte un reporte en el cual documente los siguientes aspectos:
   - **Análisis inicial**:
     - Resumen estadístico de la base de datos antes de la limpieza.
     - Tabla que muestre el porcentaje de valores faltantes por columna.
     - Total de filas duplicadas encontradas.
     - Descripción de los tipos de datos originales y los problemas encontrados.
   - **Proceso de limpieza**:
     - Describir qué métodos utilizaron para limpiar la base de datos (eliminación, imputación, etc.).
     - Mostrar antes y después de cada paso clave (por ejemplo, antes y después de eliminar duplicados).
   - **Resultados**:
     - Resumen final de la base de datos después de la limpieza.
     - Confirmar que los tipos de datos son correctos.
     - Tabla que muestre el porcentaje de valores faltantes final por columna.
     - Comprobación de que no hay duplicados ni valores inválidos.
5. **Entrega**
   Subir al aula virtual:
   - El archivo de la base de datos limpio en formato CSV en su onedrive.
   - El reporte en formato PDF con capturas de pantalla de código y las modificaciones del dataframe de las 5 primeras filas.
   - El código que usaron para limpiar la base de datos, en un archivo Jupyter Notebook (.ipynb) en

---

Criterios de Evaluación:

- **Comprensión del problema (20%)**: Se evaluará el análisis inicial del dataset.
- **Calidad de la limpieza (40%)**: Se calificará la corrección efectiva de los problemas en el dataset.
- **Documentación clara y detallada (20%)**: Se evaluará que el reporte sea coherente y que los pasos estén bien explicados.
- **Código funcional (20%)**: El código debe ser claro, eficiente y reproducible.