ACTIVIDAD SESIÓN PROCESAMIENTO Y ESCALAMIENTO DE DATOS

Te han contratado para evaluar un dataset aplicando los conceptos relacionados con el **preprocesamiento de datos**, la **codificación de variables categóricas**, el **escalamiento de datos** y su implementación utilizando la librería **Scikit-Learn**. El trabajo consta de preguntas de reflexión y un caso práctico.

INSTRUCCIONES:

1. Carga de datos (1 punto)

- Descarga el archivo **customer_data.csv** proporcionado en el material complementario.
- Carga el conjunto de datos utilizando **Pandas**.
- Muestra las primeras 5 filas del dataset.

2. Preprocesamiento de datos (3 puntos)

- **Limpieza de datos**:
    - Verifica si hay valores nulos en el dataset y elimina las filas que los contengan.
    - Elimina columnas que no sean relevantes para el análisis (por ejemplo, columnas de identificación).
- **Codificación de variables categóricas**:
    - Aplica **Label Encoding** a la columna Gender (Género).
    - Aplica **One-Hot Encoding** a la columna City (Ciudad).
- **Escalamiento de datos**:
    - Aplica **Min-Max Scaling** a la columna Age (Edad).
    - Aplica **Standard Scaling** a la columna Income (Ingresos).

3. Implementación de técnicas de distancia (3 puntos)

- Calcula la **Distancia Manhattan**, **Distancia Euclidiana** y **Distancia Minkowski** (con p=3p=3) entre los siguientes dos puntos:

- Punto A: [25, 50000] (Edad, Ingresos)
- Punto B: [30, 60000] (Edad, Ingresos)

4. Análisis de resultados (3 puntos)

- **Comparación de técnicas de escalamiento**:
  - Explica las diferencias entre **Min-Max Scaling** y **Standard Scaling**. ¿En qué casos sería recomendable usar cada una?
- **Interpretación de distancias**:
  - Describe las diferencias entre las distancias calculadas (Manhattan, Euclidiana y Minkowski). ¿Qué información adicional proporciona la Distancia Minkowski con p=3p=3?
- **Aplicabilidad**:
  - Explica en qué tipo de problemas de Machine Learning sería útil aplicar **Label Encoding** y en cuáles sería más adecuado usar **One-Hot Encoding**.

INSTRUCCIONES ADICIONALES:

- Puntos totales: 10 puntos.
- Descarga el dataset customer_data.csv.
- Formato de entrega: Comprime el archivo en formato .zip o .rar.
- Archivos a incluir:
  - Un archivo de Python
  - Un documento de texto (.txt o .docx) con las respuestas a las preguntas de análisis.
- Subir el archivo a la plataforma.