1. Introducción

El presente informe desarrolla el análisis de caso correspondiente al módulo 6 del curso Ingeniería de Datos, enfocado en el preprocesamiento y escalamiento de datos. El objetivo principal es aplicar diferentes técnicas de limpieza, transformación, codificación y normalización/estandarización de variables numéricas y categóricas, garantizando así que los datos estén preparados para su uso en modelos de machine learning y análisis avanzado.

2. Metodología aplicada

Para resolver el caso, se realizaron los siguientes pasos:

- 2.1. Imputación de valores faltantes

Se imputaron los valores faltantes en la columna 'Ingresos_USD' utilizando la media, garantizando la consistencia y evitando sesgos en el análisis posterior.

- 2.2. Codificación de variables categóricas

Se aplicaron tres técnicas:

  - Label Encoding: Asigna un número único a cada categoría.}
  - One-Hot Encoding: Genera columnas binarias para cada categoría.
  - Variables Dummy: Similar a One-Hot, pero eliminando la primera categoría para evitar multicolinealidad.
- 2.3. Escalamiento de variables numéricas

Se aplicaron dos métodos:

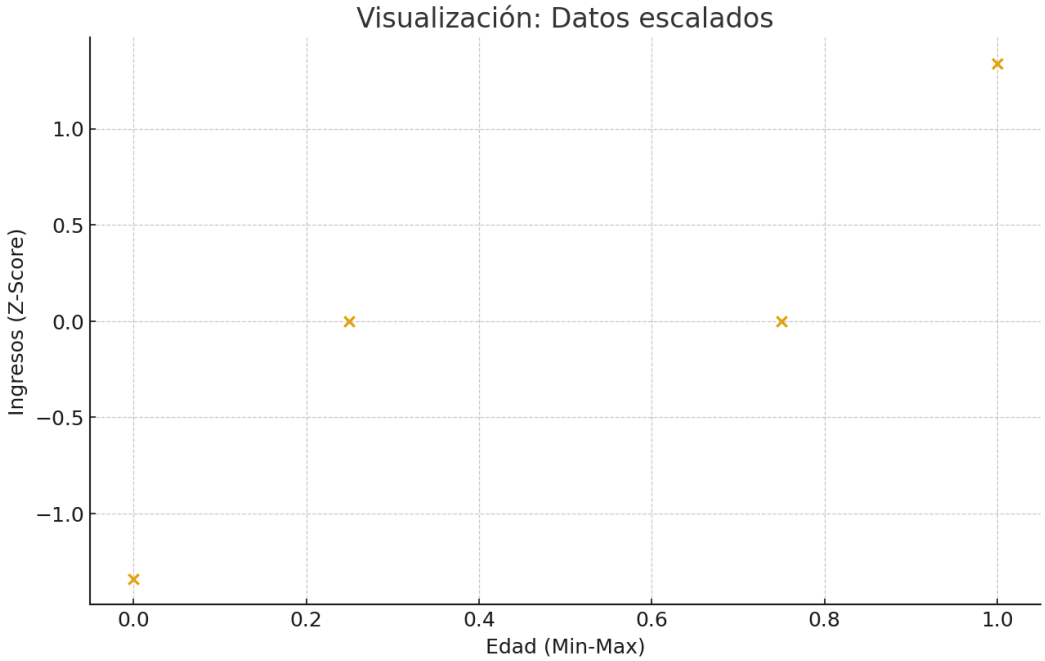
  - Min-Max Scaling: Normaliza los datos en el rango [0,1], conservando las proporciones.
  - Z-Score Scaling (Estandarización): Centra los datos en media 0 y desviación estándar 1, más robusto frente a outliers.

3. Resultados obtenidos

A continuación, se presentan los resultados principales obtenidos tras el preprocesamiento y escalamiento de datos. El dataset original fue transformado incorporando imputaciones, codificaciones y nuevas variables escaladas.

ID	Edad	Ciudad	Ingresos_USD	Ingresos_USD_imputado	Edad_minmax	Ingresos_zscore
1	25	Madrid	30000.0	30000	0.0	-1.34
2	45	Sevilla	50000.0	50000	1.0	1.34
3	30	Madrid	nan	40000	0.25	0.0
4	40	Barcelona	40000.0	40000	0.75	0.0

Gráfico 1. Visualización de datos escalados (Z-Score):



4. Discusión y análisis

¿Por qué es importante realizar estas tareas antes de entrenar un modelo de Machine Learning?

El preprocesamiento de datos es una etapa crítica en cualquier proyecto de ingeniería de datos. Permite garantizar que la información utilizada sea confiable y esté en un formato adecuado para los modelos. Las técnicas aplicadas permitieron:

- Asegurar consistencia mediante imputación.
- Codificar variables categóricas para su uso en algoritmos.
- Escalar variables numéricas para mejorar la eficiencia de los modelos basados en distancia.

¿Qué diferencias observaste entre la normalización y la estandarización?

- Con la normalización, las variables quedaron acotadas entre 0 y 1, útil cuando las magnitudes difieren mucho.
- Con la estandarización, los datos se centraron en 0 con desviación estándar 1, más adecuado cuando hay valores extremos o el modelo asume distribución normal.

5. Conclusiones

La aplicación de técnicas de preprocesamiento y escalamiento de datos mejora significativamente la calidad de los datos y su usabilidad en modelos predictivos. El ejercicio permitió comprender la importancia de preparar correctamente los datos antes de cualquier análisis avanzado o entrenamiento de modelos de machine learning.