

Informe: Preprocesamiento y Escalamiento de Datos

1. Etapas aplicadas

- Imputación de valores nulos: se completó el valor faltante en la columna 'Ingresos' con la media.
- Codificación de variables categóricas:
 - Label Encoding aplicado a la columna 'Ciudad'.
 - One-Hot Encoding para columnas binarias por ciudad.
 - Variables Dummy eliminando una categoría para evitar multicolinealidad.
- Escalamiento de variables numéricas:
 - Min-Max Scaling (valores entre 0 y 1).
 - Z-Score Scaling (media 0 y desviación estándar 1).

2. Respuestas de reflexión

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

Porque los algoritmos de machine learning requieren datos completos, en escalas comparables y sin sesgos provocados por codificaciones incorrectas. Estas tareas permiten mejorar el rendimiento y evitar errores en el modelado.

¿Qué diferencias observaste entre normalización y estandarización?

- Normalización (Min-Max): escala los valores en un rango definido (0 a 1), útil cuando se desconoce la distribución.
- Estandarización (Z-Score): transforma los datos para que tengan media 0 y desviación estándar 1, ideal cuando se supone una distribución normal o se usan algoritmos sensibles a magnitudes absolutas.