Técnicas de Escalado

Bryan Cutipa Carcasi

1. Introducción

En el análisis de datos multivariados, es común que las variables presenten diferentes unidades y escalas. Esto puede afectar negativamente algoritmos de aprendizaje automático, análisis de componentes principales (PCA) o cualquier método sensible a la magnitud de las variables. Las técnicas de **escalado** permiten transformar las variables para que sean comparables.

Este informe presenta la aplicación de cuatro técnicas de escalado a un conjunto de datos simulado con variables socioeconómicas y ambientales.

2. Datos y Métodos

2.1. Conjunto de datos

Se generó un conjunto de 500 observaciones con las siguientes variables:

`ingresos`: distribución log-normal (sesgada positivamente)

`edad`: aproximadamente normal

`pH_agua`: uniforme entre 6.5 y 9.0

`captura_kg`: conteo de captura semanal (Poisson)

2.2. Técnicas aplicadas

1. **Normalización (Min-Max)**: $x' = x - x_{\text{mín}} x_{\text{máx}} - x_{\text{mín}}$, rango [0,1].
2. **Estandarización (z-score)**: $z = x - \mu\sigma$, media 0, desviación 1.
3. **Escalado robusto**: usa mediana e IQR, resistente a *outliers*.
4. **Transformación logarítmica**: útil para datos con sesgo positivo.

3. Resultados

La Figura 1 muestra el efecto de las transformaciones sobre la variable `ingresos`, que originalmente presenta una fuerte asimetría positiva. La transformación logarítmica reduce notablemente el sesgo, mientras que la normalización y la estandarización permiten compararla con otras variables en escalas comunes.

El Cuadro 1 presenta estadísticas resumen de las versiones escaladas.

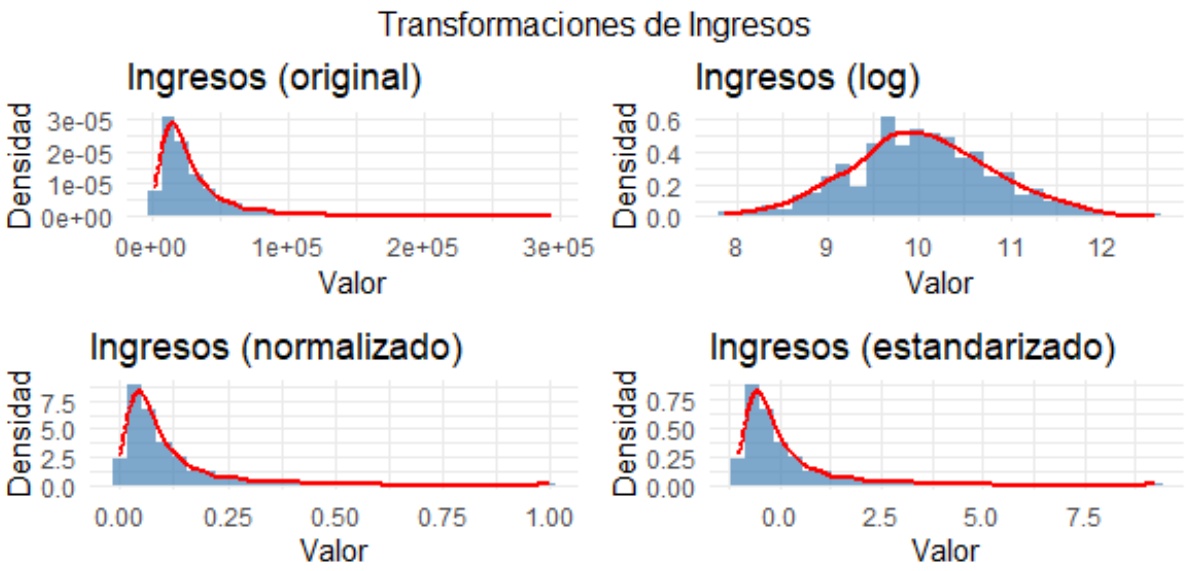

Figura 1: Distribuciones de la variable `ingresos` antes y después de aplicar técnicas de escalado.

Cuadro 1: Estadísticas resumen de transformaciones de `ingresos`

Versión	Media	Desv. Est.	Mediana	IQR
Original	25083	32105	14333	25218
Log	9.98	0.81	9.57	1.10
Normalizado	0.50	0.29	0.34	0.29
Estandarizado	0.00	1.00	-0.33	0.89
Robusto	0.00	1.00	-0.33	1.00

4. Conclusiones

Las técnicas de escalado son esenciales para garantizar la equidad en el tratamiento de variables heterogéneas. La elección del método debe basarse en:

La distribución de los datos (normal vs sesgada),

La presencia de valores atípicos,

Los requisitos del algoritmo posterior (ej. k-NN requiere escalado; árboles no).

En este ejemplo, la transformación logarítmica seguida de estandarización resulta ideal para la variable `ingresos`.