Solucion ejercicio limpieza de datos

1. Eliminación de duplicados

El dataset contiene registros duplicados que no aportan información nueva y por lo tanto hacen que el modelo aprenda patrones incorrectos.

Para limpiar el dataset localizamos las diferentes filas duplicadas, en este caso son la fila 1, 3 y 4. La fila 3 se puede eliminar puesto que es igual que la fila 1, y en la fila 4 nos encontramos con que la edad es distinta, pero tenemos en cuenta que en las filas 1 y 3 la edad es de 20 años por lo tanto asumimos que la fila incorrecta es la 4.

ID_Cliente	Nombre	Edad	Ciudad
101	Ana	20	Madrid
102	Luis	22	Sevilla

2. Tratamiento de valores nulos

Aquí encontramos datos incompletos, estos valores pueden hacer que la calidad del algoritmo disminuya.

En este caso puesto que son dos filas numéricas y hay mas valores que nulos, lo ideal es hacer una media.

La media de las edades es de 27 años y de las medias 7.6 por lo tanto la tabla quedaría.

Alumno	Edad	Nota_examen
Ana	18	8
Luis	27	6
Marta	19	7.6
Juan	45	9

3. Normalización vs Estandarización

En este caso las edades son muy diferentes a las horas de estudio, por lo tanto, el modelo va a tener más en cuenta las edades ya que son números mayores.

Para ello proponemos normalización Min-Max.

Persona	Edad	Horas_estudio
Ana	0.0	0.2
Luis	0.32	1.0
Marta	1.00	0.0

4. Codificación de variables categóricas

El problema de esta tabla es que los niveles académicos de la tabla son letras y el modelo necesita números para funcionar correctamente, para ello proponemos el label encoding asignando valores a cada nivel académico según su importancia.

Primaria = 0, Secundaria = 1, Universidad = 2.

Cliente	Nivel_estudios
Ana	0
Luis	1
Marta	2

5. Outliers: detectar y decidir

Los clientes tienen unos 200kWh de gasto mientras uno tiene 2500kWh esto es un valor que se aleja mucho de la media y hace que el modelo pueda alejarse de la media y el modelo sea de menor calidad.

En este caso, al ser un data set pequeño se puede eliminar ese cliente.

Cliente	Consumo_kWh
A	210
B	230
C	220