MODULO 06 - EJERCICIO 03-A

ALEXIS YURI M.

Trabaja sobre un conjunto de datos crudos con problemas típicos (valores nulos, categorías escalas desiguales) identifica y aplica las transformaciones necesarias para dejarlo listo para el modelado.

1) Información del dataset.

- Archivo: dataset\_vivero.csv

- Tamaño: 102 filas, 6 columnas.

- Filas duplicadas detectados: 2.

- Valores faltantes/nulos (suma total): 31.

2) Detección de problemas.

- Datos numéricos guardados como texto (p. ej., Altura\_cm, Ancho\_cm, Precio), con símbolos/espacios y posible coma decimal.

- Valores nulos en varias columnas.

- Escalas distintas entre variables numéricas.

- Columna identificadora: PlantaID (no debe usarse como variable característica).

3) Transformaciones aplicadas.

- Limpieza de strings: se reemplazó “,” por “.” y se eliminaron símbolos no numéricos (cm, $ …).

- Imputación de faltantes:

Numéricas: media por defecto; mediana si |sesgo| > 1 (más robusto ante outliers).

Categóricas: moda (valor más frecuente).

- Conversión: de tipo objeto a numérico. Para que Altura\_cm, Ancho\_cm y Precio se traten como números y no como categorías.

- Conversión a tipo float cuando ≥80% de los valores eran convertibles.

- Codificación de categóricas con tipo One-Hot convierte categorías nominales en señales binarias sin imponer orden artificial.

- Escalamiento con StandardScaler en variables numéricas (no se incluye PlantaID). Centra en 0 y escala a varianza 1, adecuado para modelos basados en distancia (KNN, SVM) y para estabilizar el entrenamiento.

- Tratamiento de identificadores: PlantaID se dejó fuera de las variables características (es un ID). Se entrega una versión final con el ID preservado como columna auxiliar al inicio del archivo transformado.

4) Justificación de las decisiones tomadas.

StandardScaler vs MinMaxScaler: En datos con unidades heterogéneas y posibles outliers moderados, StandardScaler es más estable. Si fuera a usar redes neuronales o quisiera mantener todas las variables en [0,1], podríamos generar una versión Min-Max sin problema.

Media o mediana: La mediana reduce el impacto de valores extremos cuando el sesgo es alto; en lo demás, la media es adecuada y más informativa.