Glosario

Archivo CSV: Archivo de texto donde los datos están separados por comas.

Archivo .ipynb: Archivo de notebook de Jupyter. Contiene código, texto, gráficos y resultados en un solo documento interactivo.

NumPy: Librería para hacer operaciones matemáticas con arreglos. Es base para muchos cálculos numéricos en ciencia de datos.

Pandas: Librería de Python para manejar datos en forma de tablas (filas y columnas). Permite limpiar, filtrar, agrupar y transformar datasets fácilmente.

DataFrame: Es una tabla de Excel con filas y columnas, sobre la que se hacen análisis y transformaciones.

.shape: Devuelve el tamaño de un DataFrame: cantidad de filas y columnas.

.dtypes: Muestra los tipos de datos que tiene cada columna (texto, número, booleano, etc.).

.drop(): Se usa para eliminar columnas o filas del DataFrame.

.isnull() / .sum(): Detecta y cuenta cuántos valores faltantes hay por columna.

.apply(): Aplica una función a cada elemento de una columna o fila. Es muy útil para transformar datos en masa.

Funciones personalizadas (def): Código definido por el usuario para automatizar tareas, como extraer nombres o calcular valores.

LabelEncoder: Herramienta que convierte datos categóricos (como nombres o géneros) en números para poder usarlos en modelos.

Visualización (Matplotlib / Seaborn): Librerías para crear gráficos como histogramas, boxplots y gráficos de correlación. Seaborn es más visual.

Boxplot: Gráfico que muestra la distribución de una variable y ayuda a detectar valores atípicos (outliers).

Heatmap: Mapa de calor que muestra correlaciones entre variables numéricas con colores.

Histograma (histplot): Gráfico que muestra la frecuencia de valores en una variable numérica.

Countplot: Gráfico de barras para mostrar cuántas veces aparece cada categoría en una variable.

Correlación: Mide cómo se relacionan dos variables numéricas. Va de -1 a 1. Se usa la función corr() para obtener una matriz de correlaciones: cada celda muestra qué tan relacionada está una variable con otra.

Plt.title / plt.xlabel / plt.ylabel: Funciones de Matplotlib para poner título al gráfico y etiquetas a los ejes X e Y.

Z-score: Número que indica cuán lejos está un valor respecto a la media. Se usa para encontrar outliers.

Outliers: Valores que están muy lejos del resto de los datos y que pueden afectar el análisis si no se controlan.

Machine Learning: Rama de la inteligencia artificial que usa datos para crear modelos que predicen o agrupan información.

Aprendizaje supervisado: Se entrena un modelo con datos que tienen una respuesta conocida (etiqueta).

Aprendizaje no supervisado: Se entrena el modelo sin respuestas conocidas. El modelo agrupa o encuentra patrones/etiquetas por sí solo.

Regresión: Técnica supervisada usada para predecir valores numéricos continuos.

Clasificación: Técnica supervisada que predice una categoría (ej: éxito o fracaso). Aunque no se usó un modelo explícito, se prepararon datos para este tipo de análisis.

Clustering (agrupamiento): Técnica no supervisada para agrupar datos similares. Puede aplicarse a actores, géneros, países, etc.

-Modelos supervisados (aprenden con datos que tienen una respuesta/etiqueta):

train_test_split: Divide los datos en dos grupos: entrenamiento y prueba. Se entrena el modelo con uno y se evalúa con el otro.

Regresion Lineal: Modelo que predice un valor numérico a partir de otras variables. Ejemplo: predecir ingresos según presupuesto.

Árbol de desiciones: Modelo basado en reglas. Sirve para clasificar o predecir según características.

Random Forest: Conjunto de muchos árboles de decisión. Mejora la precisión y evita errores de modelos individuales.

cross_val_score: Técnica para evaluar un modelo dividiendo los datos en varias partes y promediando los resultados.

GridSearchCV: Busca automáticamente los mejores parámetros para un modelo, probando combinaciones posibles.

- Modelos no supervisados (no tiene respuestas/etiquetas, encuentran estructura en los datos):

KMeans: Agrupa datos en "k" grupos según su similitud. Ejemplo: agrupar películas por género o perfil de ingresos.

PCA (Análisis de Componentes Principales): Reduce la cantidad de variables conservando la mayor información posible. Ayuda a visualizar y simplificar datos.

DBSCAN: Algoritmo de clustering que detecta grupos en los datos sin necesidad de indicar cuántos clusters queremos. Es útil para detectar grupos de diferentes formas y tamaños, y también puntos que no pertenecen a ningún grupo (ruido).

Standard Scaler: Es otra técnica de procesamiento. Escala los datos para que tengan media 0 y desviación estándar 1. Necesario para modelos que son sensibles a escalas diferentes.

Gaussian Mixture Model: Es un modelo que asume que los datos provienen de una mezcla de varias distribuciones normales multivariadas (Gaussiana). Permite detectar estructuras complejas, modelar clusters con formas elípticas (a diferencia de K-Means que asume clusters esféricos) y estimar la probabilidad de que un punto pertenezca a cada cluster.