📊✨ Importancia de Analizar la Normalidad, Outliers, Correlación y Nulos ✨📊

🔍 ¿Por qué analizar si una variable es normal?

Analizar la normalidad de una variable es esencial porque afecta directamente la elección del modelo y la interpretación.

⭐ ¿Para qué sirve?

• Elegir el modelo adecuado:

  - Si la variable es normal → funcionan bien modelos lineales.

  - Si NO es normal (sesgada) → funcionan mejor árboles o boosting.

• Decidir si la variable debe transformarse (log, raíz, Box-Cox).

• Elegir la métrica correcta: RMSE sufre con outliers, MAE/MAPE no tanto.

• Interpretar el fenómeno: colas largas indican desigualdad o riesgo.

🚨 Importancia de los valores atípicos (Outliers)

Los outliers son valores extremos que pueden distorsionar el análisis.

⭐ Tipos:

• Errores de captura → se corrigen o eliminan.

• Valores reales → a veces deben conservarse.

⭐ Impacto en modelos:

• Modelos lineales → MUY afectados.

• Árboles → casi no les afecta.

• Distancias (KNN, SVM, clustering) → extremadamente sensibles.

🔗 Importancia de analizar la correlación entre variables

La correlación permite entender relaciones entre variables y seleccionar características útiles.

⭐ La correlación sirve para:

• Saber si una variable X aporta algo para predecir Y.

• Detectar multicolinealidad (variables duplicadas).

• Identificar si las relaciones son lineales o no.

⭐ ¿Cómo analizarla?

• Matriz de correlación (Pearson) → numéricas.

• Spearman → relaciones no lineales.

• Visualizaciones: heatmap, scatterplots.

💡 Regla:

• r > 0.7 fuerte

• r 0.3–0.7 moderada

• r < 0.3 débil

🧠 ¿Analizar solo Y o también las variables X?

Siempre debe analizarse **todo el conjunto de variables**.

⭐ Variable objetivo (Y)

• Normalidad, sesgo, outliers, necesidad de transformación.

⭐ Variables predictoras (X):

• Correlación con Y y entre ellas.

• Distribución y outliers.

• Relevancia para el modelo.

⭐ Según el tipo de modelo:

• Lineales → requieren normalidad y pocos outliers.

• Árboles → no requieren normalidad.

• Modelos de distancia → muy sensibles a escala y outliers.

💧 Tratamiento de datos nulos (Missing Values)

Los valores nulos deben analizarse antes de imputarse o eliminarse.

⭐ Tipos:

• Información real ('no aplica')

• Ausencias sin significado

⭐ ¿Cuándo imputar?

• Menos del 60% de nulos

• Métodos: media, mediana, moda, KNNImputer, IterativeImputer, 'desconocido'

⭐¿Cuándo eliminar columnas?

• Más del 70% de nulos

• No aportan información relevante

⭐ ¿Cuándo eliminar filas?

• Pocas filas afectadas (<5%)

• Contienen errores

📘 Resumen General

Analizar normalidad, outliers, correlación y nulos permite:

1️⃣Escoger mejor el modelo.

2️⃣Decidir transformaciones necesarias.

3️⃣Detectar variables útiles.

4️⃣Limpiar adecuadamente los datos.

5️⃣Mejorar precisión y estabilidad del modelo.