Informe de Cierre de la Exploración de Datos

El objetivo de esta fase fue comprender la estructura, calidad y distribución de los datos, así como identificar anomalías que impacten el modelo k-NN.

A. Calidad y Consistencia del Dataset

1. **Instancias y Variables:** El *dataset* final de trabajo consta de **1,788 instancias** y **19 columnas** (18 predictoras y 1 objetivo).
2. **Valores Constantes (Varianza Cero):** La variable 'qs' (Potencia Reactiva del Estator) fue identificada con una desviación estándar próxima a cero (constante) y fue **eliminada**, ya que no aporta capacidad predictiva.
3. **Valores Nulos:** Se confirmó que el *dataset* **no contiene valores nulos (NaN)**, eliminando la necesidad de imputación.

B. Análisis de la Variable Objetivo (Y)

1. **Balance de Clases:** La variable objetivo (CLASE_FALLA) está **perfectamente balanceada**. Cada una de las 12 clases de falla tiene exactamente el mismo número de instancias. Esto es ideal, ya que evita el sesgo del modelo y garantiza que métricas simples como el *Accuracy* sean adecuadas para la evaluación.

C. Análisis de las Variables Predictoras (X)

1. **Escala:** Los gráficos de dispersión (Boxplots) mostraron una **diferencia extrema en las escalas** de las variables.
    - Variables como R, v25x, pg, y qg operan en un rango de cientos o miles.
    - Variables como las corrientes (ixx) y potencias del rotor (pr, qr, ps) operan en un rango de milésimas o centésimas de unidad, haciendo que sus Boxplots sean casi planos (indicando una varianza minúscula).
    - **Implicación para k-NN:** La falta de estandarización haría que las variables con mayor magnitud (ej., R) dominaran la función de distancia, haciendo inútil el cálculo de la similitud.

2. **Outliers:** Se detectó la **presencia de numerosos *outliers*** en la mayoría de las variables. Esto es esperado en datos de fallas eléctricas, donde las anomalías causan picos y lecturas extremas. Estos *outliers* no deben eliminarse sin un análisis exhaustivo, ya que podrían contener información crítica sobre las fallas.

3. **Correlación y Multicolinealidad:** El análisis visual y la naturaleza de los datos confirman la alta multicolinealidad entre las variables de la misma fase (ej., las tres corrientes de red, i25a, i25b, i25c están fuertemente correlacionadas). Esta multicolinealidad es inherente al sistema eléctrico, pero no afecta negativamente al modelo k-NN (a diferencia de modelos basados en coeficientes, como la Regresión Logística).

D. Conclusión y Justificación del Preprocesamiento

La principal conclusión del EDA es la necesidad urgente de tratar la escala de los datos.

Estrategia de Preprocesamiento: Es **obligatorio** aplicar el **Estandarizador (StandardScaler)** para centrar los datos en la media (0) y escalar su varianza (1). Esto garantiza que el algoritmo k-NN calcule distancias justas y que todas las variables predictoras contribuyan de manera equitativa a la clasificación de las fallas.