Módulo 3

Sesión N° 4

ACTIVIDAD:

Manejo de valores perdidos y outliers en un dataset real

Objetivo: Identificar, tratar y documentar el proceso de manejo de valores perdidos y outliers en un conjunto de datos real.

Contexto

Imagina que formas parte del equipo de ingeniería de datos de una empresa industrial. Has recibido un archivo de datos de sensores (“sensores_industriales.csv”) que contiene registros con valores numéricos (por ejemplo: temperatura, presión) y categóricos (por ejemplo: estado del sensor). Este dataset presenta valores perdidos y algunos valores atípicos.

Tu misión:
Aplicar las técnicas aprendidas en la clase para limpiar y preprocesar el dataset, documentar cada paso y justificar las decisiones tomadas.

Entregable:

- Notebook ejecutable (.ipynb) o script (.py) con todo el proceso documentado.
- Archivo CSV final con los datos limpios.
- Informe breve (máx. 1 página) describiendo las decisiones y técnicas utilizadas.
- Formato: grupal.

KIBERNUM
IT Academy

Requerimientos:

1. Carga y exploración inicial
    - Carga el archivo “sensores_industriales.csv” en tu entorno de Python (puedes usar Colab o Jupyter Notebook).
    - Realiza una exploración básica: cantidad de registros, variables, tipos de datos y presencia de valores perdidos.
2. Detección de valores perdidos
    - Identifica las columnas y filas que contienen valores faltantes.
    - Reemplaza los marcadores especiales (como “NA”, “?”, “-999”) por valores nulos reconocidos (NaN).
3. Tratamiento de valores perdidos
    - Elimina filas con más del 50% de sus valores faltantes.
    - Imputa valores faltantes en variables numéricas usando la mediana o KNN, según lo aprendido.
    - Imputa valores faltantes en variables categóricas usando la moda o una categoría especial (“Desconocido”).
4. Detección y tratamiento de outliers
    - Detecta outliers en al menos dos variables numéricas utilizando la regla de 1.5 x IQR.
    - Elimina los registros identificados como outliers extremos o justifica si decides no eliminarlos.
5. Visualización antes y después
    - Genera boxplots y tablas resumen antes y después del tratamiento para evidenciar los cambios.
6. Documentación del pipeline
    - Explica brevemente cada paso realizado y la justificación de las técnicas elegidas (máx. 1 página).
    - Guarda el nuevo dataset limpio y preprocesado en un archivo CSV.