Análisis de datos - Trabajo práctico integrador. Parte 2.

1. Introducción y motivación

Para este trabajo final deberá realizar la preparación del set de datos utilizado en el TP 1 para el entrenamiento del modelo de ML planteado. **El trabajo no requiere el entrenamiento del modelo.**

2. Consignas

El análisis debe abordar los siguientes aspectos:

- Preprocesamiento y limpieza del dataset:
  - Realizar una limpieza general del dataset, eliminando o corrigiendo datos inconsistentes o irrelevantes.
  - Realizar el split del dataset (ej: train y test).
  - Identificar y tratar los valores faltantes en el dataset.
  - Detectar y manejar los outliers utilizando técnicas estadísticas o visuales apropiadas.
  - Escalar y / o normalizar los features.

- Feature engineering:
  - Crear nuevos features en caso de ser necesario. Justificar.
  - Aplicar técnicas de conversión de variables: codificación, discretización.
  - Analizar el balance/desbalance de clases (en el caso que se trate de un problema de clasificación).
  - Proponer y aplicar mecanismos de balance en caso de ser necesario y justificar la selección.

- Reducción de dimensionalidad
  - Evaluar relaciones entre variables y realizar una selección de features con los mecanismos vistos en clase (ej: filtros).
  - Implementar dos técnicas de extracción de features (ej: PCA). Comparar entre sí, y comparar el dataset original con el dataset reducido; evaluar ventajas y desventajas de la reducción.

3. Entrega

Consideraciones:

- La entrega consiste en una o dos notebooks correctamente documentadas y organizadas.
- Las notebooks deben estar compartidas en un repositorio GitHub público (preferentemente el mismo del TP 1).
- Se realizará una única entrega por grupo, a través del campus virtual.
- En la entrega deberán incluir un archivo (txt, pdf) con el link al repositorio GitHub. No es necesario entregar material de apoyo, pdf, slides, zip, etc.; basta con el link. (**NOTA**: el campus obliga a entregar un adjunto. Si no adjuntan nada, la entrega puede fallar).
- La entrega estará habilitada desde el día jueves **14/08/2025 a las 19:00** hasta el día lunes **18/08/2025 a las 23:59** (hora Argentina). Pasada la ventana de tiempo, el campus cierra la posibilidad de entrega de forma automática. Ante cualquier eventualidad, contactar a las docentes.
- **IMPORTANTE!** Al finalizar, asegurarse que la actividad figure "**Entregada**".

4. Evaluación

Se evaluará:

- La comprensión de los temas vistos:
    - Orden de aplicación de las técnicas para prevenir data leakage
    - Manejo de valores faltantes
    - Manejo de datos atípicos (outliers)
    - Conversión de variables
    - Reducción de dimensionalidad
    - etc.
- La calidad de las explicaciones, conclusiones y justificaciones de los pasos realizados.
- Bonus! creatividad (ej: descubrimiento de patrones o datos curiosos, indagar en influencias externas: eventos, condiciones climáticas, época del año, uso de visualizaciones o técnicas no vistas en clase y su explicación).

ChatGPT y demás LLMs: usarlos con responsabilidad