Análisis de datos - Trabajo práctico integrador.

1. Introducción y motivación

Para este trabajo práctico deberán elegir un set de datos (ver 1.1), plantear un posible problema de ML supervisado a partir de los datos elegidos, hacer el EDA completo y realizar la preparación del set de datos para el entrenamiento del modelo de ML planteado.

Deberán trabajar en grupo.

El trabajo consiste en analizar y preparar los datos; no requiere el entrenamiento del modelo.

1.1 Datasets disponibles

En la siguiente tabla tienen los datasets propuestos. Se espera que indaguen en las posibilidades de análisis y visualización según el dataset elegido.

Dataset	Tema
AirBnB Buenos Aires	Alojamiento
Encuesta mundial de salud escolar, Argentina 2018 (EMSE 2018)	Educación
Estadísticas sobre Ejecución de la Pena (SNEEP) (Elegir un año).	Judicial
Recorridos en Ecobicis (Elegir un año).	Tránsito / vehículos
Crímenes reportados en Chicago (Elegir un año).	Policial
Denuncias a la policía de NY (para lo que va del 2025)	Policial
Uso de taxis Yellow Cab en USA (Elegir un año).	Tránsito / vehículos
Eventos de violencia organizada (UCDP Georeferenced Event Dataset, GED)	Conflictos bélicos
Full TMDB Movies Dataset 2024	Ocio
Pacientes con diabetes en 130 hospitales de US	Medicina / salud
Precios Claros - Base SEPA: Elegir alguna de las cadenas grandes de supermercado (Carrefour, Disco, etc.)	Comercio

Uso de High-Volume For-Hire Services (HVFHS) en USA (Elegir un año).	Tránsito / vehículos
Reviews de Yelp (*) Contactar a las profesoras si eligen este (les daremos algunas indicaciones)	Reviews de comercios

1.2 Definición de grupos y elección de datasets

- Los grupos deben ser de 4/5 personas.
- Un mismo dataset no puede ser elegido por más de dos grupos.

Completar esta planilla con los datos del grupo de trabajo y el dataset elegido antes de la clase 2: **30/10/2025**.

2. Consignas

El análisis debe abordar los siguientes aspectos:

- Exploración y comprensión de los datos:
  - Cargar el dataset proporcionado y realizar un análisis exploratorio de los datos.
  - Describir las características principales del dataset, incluyendo el número de observaciones, número de variables y tipos de datos.
  - Identificar patrones generales y distribuciones.
  - Identificar errores, outliers (anomalías), valores faltantes y su tipo (MCAR, MAR, MNAR).

- Aplicación de técnicas de visualización:
  - Utilizar técnicas de visualización adecuadas para ilustrar las principales características del dataset.
  - Asegurarse de que las visualizaciones sean claras, concisas y efectivas para comunicar la información.
  - Interpretar los resultados obtenidos a partir de las visualizaciones.

- Plantear un posible problema de ML supervisado a partir de los datos elegidos.
  - Describir el problema de clasificación o de regresión.
  - Definir la variable target.

- 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 técnicas de extracción de features (ej: PCA). Evaluar ventajas y desventajas de la reducción.

3. Entrega

- La entrega consiste en una presentación de Google Slides (no se requiere entregar código ni repositorios).
- Antes de la clase 8 les compartiremos una presentación de Google Slides para que peguen la información resumida en **5 diapositivas como máximo**.
- La elección de la información a presentar queda a criterio de cada grupo. La idea es que refleje todo el análisis realizado, los desafíos encontrados, los criterios aplicados y las conclusiones obtenidas.

4. Evaluación

- El TP se defenderá de forma oral. La exposición se desglosará en dos sesiones: 11/12/2025 y 15/12/2025.
- La presentación se realizará con diapositivas.
- Cada grupo tendrá asignado un total de 15 minutos (10’ para la exposición y 5’ para devolución de las docentes).
- Es importante que todos los miembros del grupo expongan de igual manera en términos de tiempo. Es altamente recomendable que, durante la presentación, todos los miembros del grupo tengan la cámara prendida.
- Se evaluará:
    - Entendimiento del dataset (características, desafíos, etc.).
    - La elección y aplicación de conceptos de visualización.
    - Justificación de las técnicas elegidas para preparar los datos.

- Profundidad de las conclusiones.
- Calidad de la presentación (diapositivas y exposición).
- El entrenamiento de modelos no es requerido y por lo tanto agregarlo no impacta en la nota.