Trabajo Práctico 1

Introducción

En este trabajo práctico se propone que cada grupo de alumnos se enfrente a diferentes problemas de ciencia de datos y que pueda resolverlos aplicando los contenidos que se ven en la materia.
El objetivo principal del trabajo será aplicar técnicas de análisis exploratorio, preprocesamiento de datos, agrupamiento, clasificación y regresión. En la sección enunciado se detallan los objetivos particulares de cada ejercicio.

Modalidad de entrega

Repositorio

Cada grupo deberá crear su propio repositorio en github: TA047R-2C2025-GRUPOXX
En dicho repositorio deberá estar disponible todo el contenido obligatorio de la entrega (notebooks, modelos entrenados, *datasets*, reportes, presentación) y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Notebooks

El trabajo debe ser realizado en notebooks de python, se espera que las mismas contengan **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. Se debe entregar una notebook para cada ejercicio, con la siguiente nomenclatura :

TA047R_TP1_GRUPOXX_ENTREGA_EJX

Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de *markdown*. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.

Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (*papers*, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, y cualquier referencia que se considere necesaria. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, por lo tanto, deben ser comprensibles por quien los vaya a leer.

Conjuntos de datos (*datasets*)

A partir de las tareas de preprocesamiento, y de las diferentes estrategias que se planteen, es posible que se generen nuevos *datasets* sobre los cuales se entrenarán los modelos. Todo conjunto de datos creado debe ser almacenado y debe estar disponible en la entrega para ser utilizado por el equipo docente.

Modelos

Todos los modelos entrenados tanto para clasificación como para regresión deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizados por el equipo docente.

Reportes

Se deberá confeccionar un reporte (en formato pdf) a modo de resumen de los puntos desarrollados en cada ejercicio. El documento deberá tener la siguiente nomenclatura TA047R_TP1_GRUPOXX_REPORTE y **deberá seguir el template proporcionado por la cátedra**.

Exposición del TP

Cada grupo deberá realizar una exposición de los resultados obtenidos en cada ejercicio del trabajo práctico. La misma no deberá superar los 20 minutos, podrán exponer las notebooks o confeccionar una presentación tipo PowerPoint o similar. **Deberán seguir las pautas mínimas proporcionadas por la cátedra.**

Fechas de entrega

Entrega 14/10/2025 : Deberán informar la entrega por el canal de Slack de su grupo, enviando un *link* al repositorio de **GitHub**. Allí deberá estar disponible la notebook, los modelos entrenados y el reporte con el resumen del trabajo. Para el reporte pueden tomar como referencia el modelo de reporte. **Esta fecha es obligatoria.**

Exposición 21/10/2024 - 23/10/2025: cada grupo realizará una exposición oral de los principales resultados de su trabajo práctico. El equipo docente informará la fecha y la hora en que debe exponer cada grupo. Las pautas mínimas se explican aquí. **Esta fecha es obligatoria para todos los alumnos.**

Reentrega 02/12/2025: aquellos grupos que deban realizar correcciones contarán con esta fecha para volver a entregar el trabajo práctico. Esta es la última oportunidad para aprobarlo.

Enunciado

EJERCICIO 1 - *Análisis Exploratorio de Datos*

El ejercicio tiene como propósito **formular y responder preguntas de investigación** a partir de un subconjunto (3 meses) de datos de viajes en taxis Yellow Cab en USA. El período de análisis será asignado aquí por el equipo docente.
Para responder esas preguntas se deberá realizar un **análisis exploratorio y de preprocesamiento**, que incluya al menos:

- Exploración de variables (cuantitativas y cualitativas), medidas de resumen, correlaciones y gráficos.
- Revisión de datos faltantes y decisión de tratamiento.
- Identificación y análisis de valores atípicos.
- Posible creación de nuevas variables que enriquezcan el análisis.
- Visualizaciones que permitan contrastar hipótesis y comunicar hallazgos.

Los ítems listados son mínimos requeridos, pero cada grupo puede sumar técnicas o enfoques adicionales.

Lo más importante: no se trata solo de aplicar técnicas, sino de **plantear preguntas significativas** que los datos permitan responder (ej.: caracterización de zonas frecuentes, patrones de viajes en tiempo/distancia según horario o día, etc.).

EJERCICIO 2 - Modelos de Clasificación Binaria

Se trabajará con datos de estaciones meteorológicas de Australia, con el objetivo de **predecir si lloverá al día siguiente** (*RainTomorrow*) a partir de los datos del día actual. Cada equipo usará las *Locations* asignadas aquí por el equipo docente.

Etapas del trabajo:

1. **Análisis exploratorio y preprocesamiento:** limpieza de datos, generación de nuevas variables y preparación para los modelos (encoding, normalización, etc.).

2. **Entrenamiento y predicción (80% train / 20% test):**

    - **Modelo 1 – Árbol de decisión:** entrenar, optimizar hiperparámetros con *k-fold CV*, graficar, interpretar reglas y evaluar performance (métricas + matriz de confusión).
    - **Modelo 2 – Random Forest:** entrenar y optimizar con *k-fold CV*, analizar importancia de atributos, mostrar un árbol representativo, evaluar performance (métricas + matriz de confusión).
    - **Modelo 3 – A elección:** entrenar con *cross-validation*, justificar elección, evaluar métricas y matriz de confusión.

3. **Comparación de resultados:** presentar un cuadro comparativo y responder: *¿Qué modelo elegirían para predecir la lluvia y por qué?*

EJERCICIO 3 - Modelos de Regresión

Se trabajará con datos de alojamientos de la plataforma AirBnB. El objetivo es predecir el precio de alquiler (price) en función de la información publicada en los avisos. Cada equipo utilizará el archivo “Detailed Listings data” de una ciudad asignada aquí por el equipo docente.

Etapas del trabajo:

1. **Análisis exploratorio y preprocesamiento**: limpieza de datos, generación de nuevas variables y preparación de features (encoding, normalización, etc.).

2. **Entrenamiento y predicción (80% train / 20% test):**

- **Modelo 1 – Regresión Lineal:** entrenar, identificar variables más influyentes y evaluar performance en train/test (métricas).
- **Modelo 2 – XGBoost:** entrenar con k-fold CV, optimizar hiperparámetros, evaluar performance en train/test (métricas).
- **Modelo 3 – A elección:** entrenar con cross-validation, justificar elección, evaluar métricas en train/test.

3. **Comparación de resultados:** elaborar un cuadro comparativo y responder: *¿Qué modelo elegirían para predecir el precio de alquiler en la ciudad seleccionada?*

EJERCICIO 4 - Agrupamiento (Clustering)

En este punto se busca analizar si es posible agrupar los datos en función de algún criterio. Vamos a utilizar un conjunto de datos que contiene información sobre algunos *tracks* (canciones) de Spotify. Para conocer en detalle cada atributo del dataset pueden consultar el siguiente link. Para esta tarea se propone utilizar el algoritmo **K-Means** y se deberán realizar los siguientes puntos:

a. Analizar la tendencia al *clustering* del dataset.
b. Estimar la cantidad apropiada de grupos que se deben formar.
c. Evaluar la calidad de los grupos formados realizando un análisis de *Silhouette*.
d. Realizar un análisis de cada grupo intentando entender en función de qué características fueron formados.