Unidad 1 - Trabajo Práctico
Preprocesamiento y análisis preliminar de datos

Enunciado

El objetivo de este trabajo práctico es, tomando una de las bases de datos provistas por el curso (https://github.com/ignaciorlando/duia-ml-datasets) o una base propia, aplicar el pipeline completo de preprocesamiento y análisis preliminar de datos que vimos en clase. En particular:

- Determinar el problema que se buscará resolver (por ejemplo, “estimar el costo de una vivienda a partir de sus características”) a partir de los datos.
- Describir los datos en términos de sus características y etiquetas (si existen).
- Identificar variables que puedan introducir sesgos y establecer acciones correctivas, según corresponda.
- Dividir los datos en conjuntos de entrenamiento, validación y test y estudiar si las distribuciones resultantes son similares entre sí.
- Identificar potenciales relaciones entre características y las etiquetas (si existen las etiquetas), o en su defecto entre algunos pares de características. Discutir las relaciones (¿Tienen sentido? ¿A qué pueden deberse?)
- Estandarizar los datos y representar gráficamente un par de características.

Entrega

Un Notebook de Python (Jupyter o Colab) que documente el análisis de los datos y permita ejecutar una a una las diferentes etapas del curado de los datos.

Condiciones de entrega y aprobación

- El trabajo práctico puede hacerse en grupos de hasta 2 personas.
- No puede utilizarse el mismo data set que se usó durante la clase.
- La entrega se realiza a través de Classroom. Incluyan por favor el nombre y apellido completo de los miembros del grupo en el Colab que hagan.
- Si en lugar de usar Colab usan Python puro (no recomendable!), póngannos un link a un repositorio de Github con la entrega correspondiente y un archivo README.md con las instrucciones de ejecución y los nombres de los miembros del grupo.
- Asegurarse de que el código incluya soporte para descargar los datos, o incluir los datos en el repositorio.
- En todo el proceso, evitar potenciales eventos de *data leakage*. Si ocurrieran, hay que realizar una reentrega!