SEMANA 1: METODOLOGÍA CRISP-DM. PROCESANDO DATOS CON PYTHON. FORMATEANDO DATOS. NORMALIZANDO DATOS. VARIABLES CATEGÓRICAS Y NUMÉRICAS

LABORATORIO CALIFICADO N°1

Indicaciones:

El trabajo en su totalidad debe ser hecho en un notebook de 'Jupyter'.

Cree una cabecera que lleve por título el tema de la semana y debajo de él coloque un 'Markdown' con su nombre y apellido.

Ejecute su notebook, grábelo, descárguelo como HTML, conviértalo a PDF y suba ambos archivos en el apartado que el docente indique junto con el archivo CSV que se obtuvo.

Con base a los datos del archivo 'titanic' y tomando como referencia la información en el enlace https://www.kaggle.com/c/titanic/data, realice lo siguiente:

1. Cree un data frame, elimine las columnas 'name', 'ticket', 'cabin', 'boat', 'body' y 'home.dest', contabilice la cantidad de valores únicos por columna e impute a la columna 'age' mediante la mediana, a la columna 'embarked' mediante la moda y elimine las filas que contenga valores nulos. Además, convierta a las columnas 'survived' y 'pclass' a tipo object y, cambie las categorías de 'pclass' de 1, 2 y 3 a '1st', '2nd', '3rd' y las categorías de 'survived' de 0 y 1 a 'no' y 'yes'.

2. Cree una columna con fechas aleatorias que vayan desde las 00:00 horas del 1 de enero de 2025 hasta las 23:59 horas del 31 de diciembre de 2025. Luego, cree una columna con el formato para fecha **18 July, 2025 06:12** y otra columna con el formato **18/07/2025 06:12 AM**, además, extraiga en columnas diferentes el día, día de semana (en inglés), mes y año.

3. Realice una normalización de las columnas 'sibsp' y 'parch' mediante un **bucle for**, asignándoles los nombres 'sibsp_norm' y 'parch_norm', respectivamente, y eliminando las columnas originales.

4. Realice una estandarización de la columna 'age' mediante un **bucle for**, asignándoles el nombre 'age_std' y eliminando la columna original. Además, realice una dummización de las columnas 'pclass' y 'embarked' mediante un **bucle for**, anteponiendo los nombres 'pclass_type' y 'embarked_type' a cada categoría, respectivamente, y eliminando las columnas originales. Por último, descargue el último data frame obtenido como un archivo CSV.

No olvide convertir su HTML a PDF antes de subirlo. También puede revisar el siguiente enlace para poder hacer esto: https://mljar.com/blog/jupyter-notebook-pdf/.