**IT ACADEMY – DATA SCIENCE**

PLANIFICACIÓN PRUEBA INDIVIDUAL:

1.- TIPOS DE ANÁLISIS A REALIZAR.

De acuerdo con las características del dataset con el que vamos a trabajar, compuesto por datos sobre variables meteorológicas de la ciudad de Barcelona, todas ellas numéricas excepto la fecha (día a día) desde marzo de 1983 hasta julio de 2021; planteo un anàlisis de regresión (regresión lineal) y otro de agrupación (clustering).

Una vez tengamos los datos descargados en el formato json en que se encuentran en el sitio web de AEMET (<https://opendata.aemet.es/centrodedescargas/inicio>), generaremos un Data Frame en el Jupyter Notebook con dicha información.

A partir de ahí, iniciamos un proceso de limpieza, formateo y transformación de datos, que nos permita empezar a testear los algoritmos con los que esperamos obtener resultados adecuados a nuestro propósito.

En esta fase de testeo de los algoritmos, haremos distintas pruebas de “feature engineering”, por tal de optimizar el rendimiento de los algoritmos a utilizar.

Hasta conseguir los mejores resultados posibles, tanto en algoritmo supervisado de regresión (regresión lineal) y el no supervisado de agrupación (custering con KMeans).

2.- PRESENTACIÓN DE RESULTADOS.

2.1.- Para el modelo de regresión lineal, obtendremos métricas de evaluación del algoritmo como :

a) MAE : mean absolute error. La más intuitiva. Representa el error medio entre valor real y el valor de predicción del algoritmo.

b) MSE : mean squared error. Similar a MAE pero el ruido es exagerado y los errores

más grandes son “castigados”. Es más difícil de interpretar que MAE ya que no está en unidades originales del data set, sin embargo, generalmente es más popular.

c) RMSE : root mean squared error. La métrica más popular, similar a MSE, sin embargo, el resultado es de raíz cuadrada para que sea más interpretable ya que está en unidades originales del data set. Se recomienda a menudo, que RMSE se utilice como métrica principal para evaluar la bondad del modelo.

2.2.- Para el caso del modelo de agrupación (clustering con KMeans), obtendremos la métrica fundamenal que es el “Silhouette score”.

Hace referencia a un método de interpretación y validación de la coherencia dentro de grupos de datos. (Cómo de bien o mal están definidos los distintos clusters resultantes del modelo).

Igualmente identificaremos las observaciones más cercanas a los centroides de cada cluster, para identificar qué variables identifican a los elementos de cada cluster.