Aprendizaje Automático 22/23

# Práctica 1
# Predicción de la producción de energía solar


***Alejandro de la Vega Ruiz***

*NIA: 100408788*

***Gaston Tomás Ruggeri***

*NIA: 100408777*




***Universidad Carlos III de Madrid***

*Escuela Politécnica Superior*

*Ingeniería Informática y Administración de Empresas*

*Madrid*

## ÍNDICE

# 1. Introducción

El objetivo de la práctica consiste en adquirir conocimientos y destreza respecto a los diferentes tipos de modelos que hay. Para ello seguiremos todos los pasos necesarios.

Primero deberemos determinar qué modelo seleccionaremos como el ajuste de sus hiperparametros posteriormente tendremos que estimar el funcionamiento mediante una evaluación del modelo y finalizaremos con la construcción del modelo final.

El conjunto de datos que utilizaremos consiste en las predicciones meteorológicas del Global Forecast System (GFS) y la finalidad es conseguir un modelo de predicción lo más preciso posible sobre la cantidad de energía generada por las plantas solares en Oklahoma.

# 2. Análisis Exploratorio de Datos (EDA)

La principal finalidad de un EDA es la comprensión de los datos de la mejor forma posible.

Por ello identificamos las principales características, disponemos 75 variables de entrada. En realidad son 15 las variables, pero hacemos una distinción entre 5 franjas horarias a lo largo del día resultado así en 15 variables * 5 franjas = 75 variables a estudiar.

Respecto al número de filas disponemos de una entrada de 4380 filas, cada una de ellas corresponde a un día del año teniendo 12 en total. 365 * 12 = 4380.

| | apcp_sf1_1 | apcp_sf2_1 | apcp_sf3_1 | apcp_sf4_1 | apcp_sf5_1 | dlwrf_s1_1 |
|---|---|---|---|---|---|---|
| **V1** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 279.583582 |
| **V2** | 0.000000 | 0.000000 | 0.010000 | 0.056364 | 0.332727 | 241.907687 |
| **V3** | 0.372727 | 0.021818 | 0.044545 | 0.010000 | 0.007273 | 266.370911 |
| **V4** | 0.002727 | 0.004545 | 0.000000 | 0.000000 | 0.000000 | 246.863048 |
| **V5** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 225.253657 |

5 rows × 76 columns

Analizando las 5 primeras filas podemos comprobar que sigue una estructura de var_t_1 Donde t corresponde con la franja horaria.

Observamos también que los datos parecen estar representados en formato float, mediante la función “disp_df.dtypes” obtenemos:

```
disp_df.dtypes
```

```
apcp_sf1_1     float64
apcp_sf2_1     float64
apcp_sf3_1     float64
apcp_sf4_1     float64
apcp_sf5_1     float64
                ...
uswrf_s2_1     float64
uswrf_s3_1     float64
uswrf_s4_1     float64
uswrf_s5_1     float64
salida           int64
Length: 76, dtype: object
```

Ahora podemos confirmar que las variables de entrada son de tipo Float64 y que la variable de salida es de tipo int64.

Ahora debemos comprobar que el dataset no tenga valores nulos, que en ninguna de las columnas falten datos. Para ello se utiliza el método count y podemos comprobar que efectivamente no hay.

También debemos determinar que de esas 4380 iteraciones con 75 variables de entrada no se encuentren valores duplicados, para ello utilizaremos *disp_df.drop_duplicates()* y luego obtendremos el tamaño mediante shape.

```
disp_df_unicos = disp_df.drop_duplicates()
disp_df_unicos.shape
```

```
(4380, 76)
```

Apreciamos como todos los valores son únicos al no cambiar el tamaño después de hacer el drop.

# 3. Exploración

# 4. Conclusión