0.- INTRODUCCIÓ

L'objectiu inicial del projecte era arribar a reproduir al màxim el temari de la formació de la IT Academy anomenat: "Machine Learning i elaboració d'un projecte final" que havia de fer servir dades de l'AEMET (Agencia Estatal de Meteorologia); però que al final es va impartir amb un temari diferent.
Un cop plantejat el projecte amb la mentora, va recomanar escurçar l'abast del projecte ja que el temari era massa ampli i així ho vaig fer, incloent finalment un model de regressió i un altre de classificació.

1.- PRESENTACIÓ DEL CONJUNT DE DADES ESCOLLIT

Les dades del models s'han obtingut de la web del AEMET (Agencia Estatal de Meteorologia) i s'han pres des de l'estació meteorològica que està a l'aeroport del Prat del Llobregat al període que va des del 1950 fins a finals del 2022.
A l'hora d'extreure les dades de la web, es pot fer directament amb codi Python, però la web té certes limitacions per evitar la saturació en cas de rebre moltes consultes, entre d'altres coses el límit temporal de cada consulta és de períodes com a molt de cinc anys i d'altra banda l'api_key de cada consulta dura només 5 minuts. Tot plegat va obligar a descarregar els fitxers json amb la informació necessària i a partir d'aquí poder treballar des del Python.

El conjunt de dades inicial consta de 26620 registres i 20 camps, al quals s'ha afegit tres camps més:

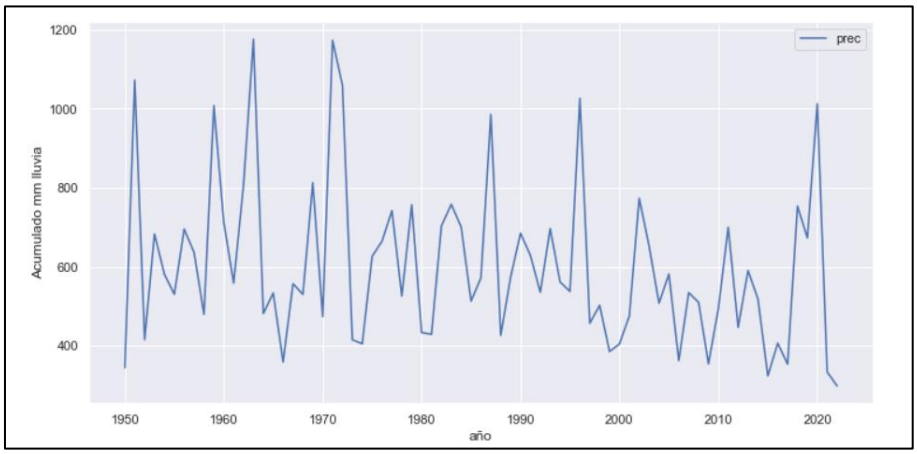
- Lustres: per agrupar les dades en blocs de 5 anys corresponents a cascun dels fitxer json.
- Mes: mes de recollida del registre
- Any: any de recollida del registre

2.- CARACTERÍSTIQUES GENERALS.

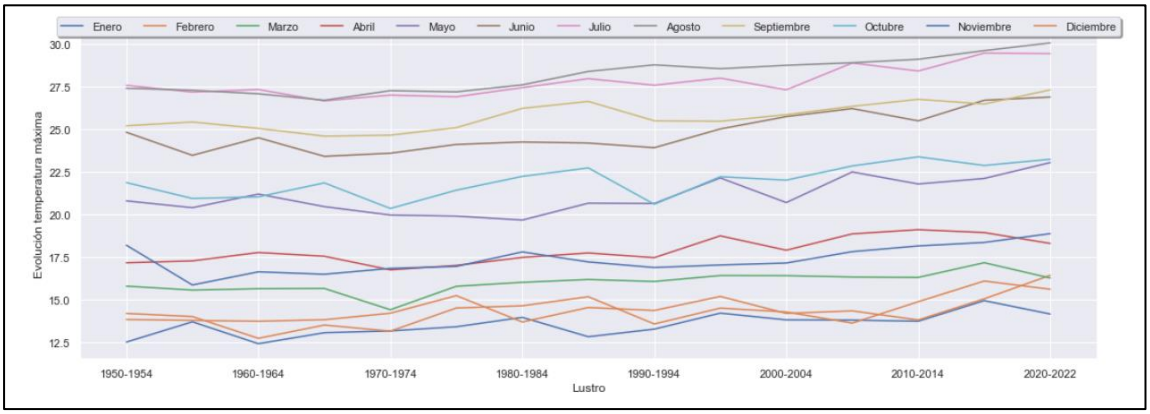
Les dades descarregades corresponen a les preses diàries de paràmetres relacionats amb els fenòmens meteorològics. A nivell general, són dades que inclouen informació de la ubicació de l'estació, temperatura, característiques del vent, pressió atmosfèrica i precipitacions.

Analitzant les dades podem fer varies representacions gràfiques, que ens ajudaran a interpretar millor les dades de data set i trobar determinats comportaments de les dades.

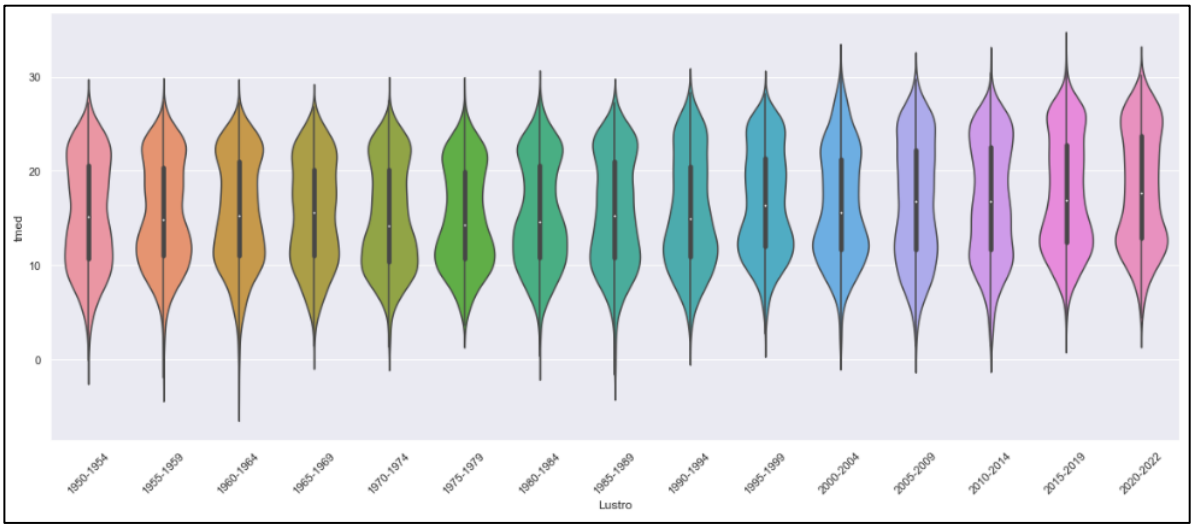
- Pluja anual acumulada durant tot el període on veiem una tendència a la baixa amb el pas dels anys:



- Representació de l'evolució de la mitjana de la temperatura màxima mensual per lustre. Podem veure que la tendència és a pujar la temperatura, independentment del mes que estem observant.



- Finalment, fem una representació de les temperatures: mínima, mitjana i màxima i la seva evolució per lustres per que sigui més fàcil de representar. Adjunto gràfic de la temperatura mitjana, on es veu també una tendència a la pujada durant tot el període.



Tot plegat, justifica de forma molt bàsica el que ja està demostrat sobre l'escalfament global del planeta i l'augment de la sequera en certes regions com la nostra.

3.- DEFINICIÓ DE LES VARIABLES

L'explicació de les variables està inclosa a les metadades que es poden descarregar des de la web d'AEMET i són aquestes:

Metadades genèriques:

unidad_generadora:	Servicio del Banco Nacional de Datos Climatológicos
periodicidad:	1 vez al día, con un retardo de 4 días
descripcion:	Climatologías diarias
formato:	application/json
copyright:	© AEMET. Autorizado el uso de la información y su reproducción citando a AEMET como autora de la misma.
notaLegal:	http://www.aemet.es/es/nota_legal

Metadades específiques:

id	descripción	tipo_datos	unidad	requerido
fecha	fecha del dia (AAAA-MM-DD)	string		true
indicativo	indicativo climatológico	string		true
nombre	nombre (ubicación) de la estación	string		true
provincia	provincia de la estación	string		true
altitud	altitud de la estación en m sobre el nivel del mar	float	M	true
tmed	Temperatura media diaria	float	°C	false
prec	Precipitación diaria de 07 a 07	float	mm (Ip = inferior a 0,1 mm) (Acum = Precipitación acumulada)	false
tmin	Temperatura Mínima del día	float	°C	false
horatmin	Hora y minuto de la temperatura mínima	string	UTC	false
tmax	Temperatura Máxima del día	float	°C	false
horatmax	Hora y minuto de la temperatura máxima	string	UTC	false
dir	Dirección de la racha máxima	float	decenas de grado (99 = dirección variable)(88 = sin dato)	false
velmedia	Velocidad media del viento	float	m/s	false
racha	Racha máxima del viento	float	m/s	false
horaracha	Hora y minuto de la racha máxima	string	UTC	false
sol	Insolación	float	horas	false
presmax	Presión máxima al nivel de referencia de la estación	float	hPa	false
horapresmax	Hora de la presión máxima (redondeada a la hora entera más próxima)	string	UTC	false
presmin	Presión mínima al nivel de referencia de la estación	float	hPa	false
horapresmin	Hora de la presión mínima (redondeada a la hora entera más próxima)	string	UTC	false

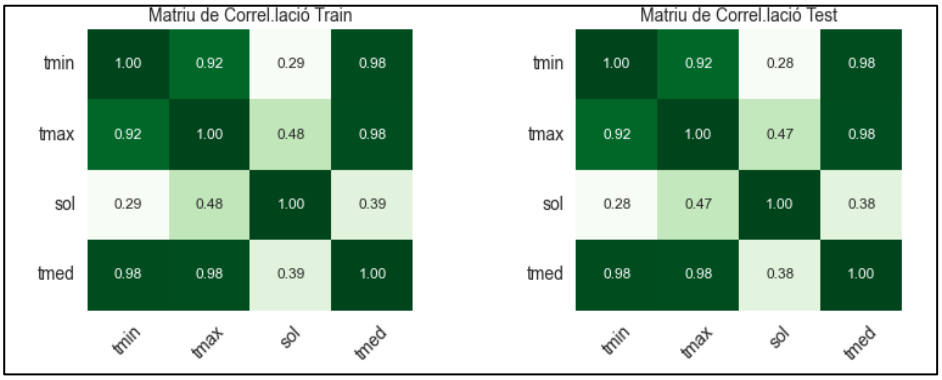
Igualment estarien les 3 variables creades explicades a l'apartat 1.

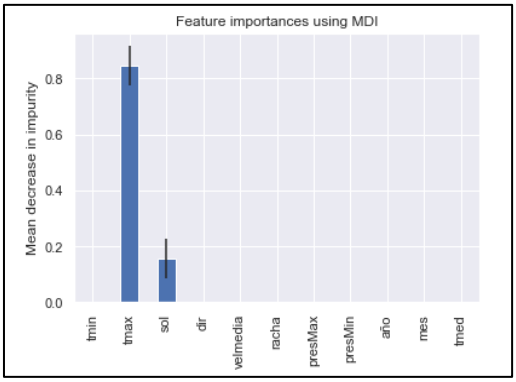
4.- PRESENTACIÓ DELS OBJECTIUS.

Els objectius del projectes són dos: la creació d'un model de regressió per poder predir la temperatura mitjana d'un dia i la creació d'un model de classificació per predir si un dia plourà o no.

4.1.- Model de Regressió – Predicció Temperatura Mitjana

En base a l'anàlisi previ de les dades es determina que els millors camps per fer el model de Regressió són: temperatura màxima, temperatura mínima i sol, com es pot veure a les matrius de correlació i las caracterísitques més importants segons el Random ForestRegressor.



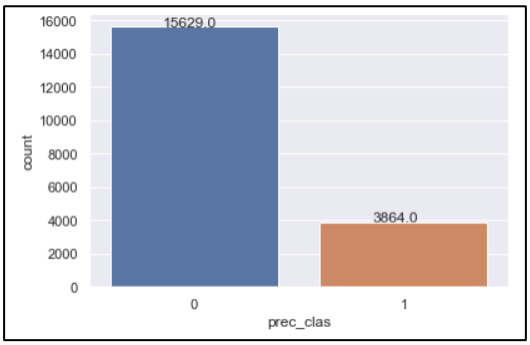


Un cop dividit el data set en els blocs de train/tes i fent l'escalat de les dades apliquem 4 models de regressió amb GridSearch per triar aquell amb un millor comportament de les seves mètriques

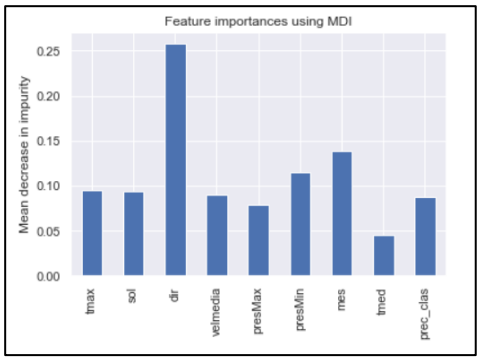
** MSE (Mean Squared Error) **:	** R2 (Coefficient of determination) **:
Multiple Linear Regression GS: 0.0007	Multiple Linear Regression GS: -780.4781
Polynomial Regression GS: 0.0007	Polynomial Regression GS: 1.0
Decission Tree GS: 0.1116	Decission Tree GS: -5.1768
Random Forest GS: 0.0022	Random Forest GS: -6.6699
Neural Network: 0.0007	Neural Network: -780.4847

El model de Polynomial Regression és el que té les millors mètriques de tots ells, especialment pel que fa al valor de R2 que és igual a 1 davant la resta de models que treuen valor negatiu

4.2.- Model de Classificació – Predicció de dies amb pluja



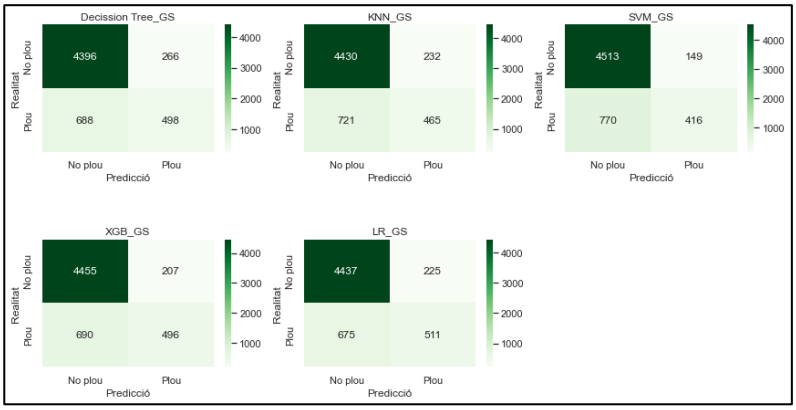
Per dir si un dia ha tingut pluja el seu valor de precipitació ha de ser més gran que el valor d'Ip que prèviament havíem resignat com 0.05 mm de pluja. Fent la gràfica del total de dades trobem que el 80% dels registres corresponen a dies on no ha plogut.



Segons el Random ForestClassifier aquesta es la gràfica de la importància del camps per fer el model classificiació, on destaquen les ratxes màximes de vent, la pressió Minima i el mes, encara que la majoria de variables graficades tènen un pes notable.

En aquest cas treballem amb 5 models de classificació amb GridSearch per triar aquell amb un millor comportament de les seves mètriques i dels resultats de les seves matrius de confussió.

	Accuracy	f1_Score (weighted)	Precission (weighted)	Recall (weighted)	f1_Score (macro)	Precission (macro)	Recall (macro)	f1_Score (micro)	Precission (micro)	Recall (micro)
Decission Tree_GS	0.8369	0.8227	0.8215	0.8369	0.7064	0.7583	0.6814	0.8369	0.8369	0.8369
KNN_GS	0.8370	0.8199	0.8209	0.8370	0.6984	0.7636	0.6712	0.8370	0.8370	0.8370
SVM_GS	0.8429	0.8199	0.8303	0.8429	0.6914	0.7953	0.6594	0.8429	0.8429	0.8429
XGB_GS	0.8466	0.8308	0.8334	0.8466	0.7168	0.7857	0.6869	0.8466	0.8466	0.8466
LR_GS	0.8461	0.8316	0.8327	0.8461	0.7198	0.7811	0.6913	0.8461	0.8461	0.8461



El comportament dels models és bastant similar però podem preseleccionar els models de XGBoost i de Regressió Logística, per què en conjunt són els que millors resultats presenten.