

PRUEBA TÉCNICA

IT ACADEMY 2021

Itinerario:

DATA SCIENCE

Estudiante : Manuel Forcales

Supervisor: Kevin Mamaqi

Septiembre 2021

Sumario

Introducción.....	3
Plan de Análisis.....	4
Fase 1-Extracción de los datos.....	4
Fase 2-Limpieza y preparación de los datos.....	4
Fase 3-Visualización grafica.....	4
Fase 4-Predicción de precipitaciones.....	4
Resultados.....	5
Visualizaciones.....	5
Predicciones.....	10
Aprendizaje Supervisado.....	10
Aprendizaje No Supervisado.....	13
Conclusiones.....	14

Introducción

El objetivo de la prueba técnica final es hacer un análisis y recomendación sobre las precipitaciones de la ciudad de Barcelona basados en sus datos históricos:

Los datos hay que obtenerlos de la siguiente pagina web:

<https://datosclima.es/Aemethistorico/Precipisolstad.php>

y estructurarlos en formato SQL.

Realizar un plan sobre el tipo de análisis a realizar y cómo se presentarán los resultados.

Utilizar PANDAS, NUMPY, MATPLOTLIB (u otras librerías) junto a un cuaderno JUPYTER para la presentación del trabajo realizado.

Subir el resultado final a un repositorio privado de Github y compartirlo una vez finalizado con el tutor.

Tiempo estimado: 72 horas.

Debe tener:

El plan sobre el análisis a realizar.

Incluir un análisis supervisado y uno no supervisado.

Plan de Análisis

Vamos a dividir el trabajo en diferentes fases:

Fase 1-Extracción de los datos

Realizar web scrapping de la web (<https://datosclima.es/>) mediante el uso de la librería Selenium para analizar los valores históricos de las precipitaciones en Barcelona así como otros valores que puedan resultar de interés para la predicción de lluvia (Temperatura, presión atmosférica, Horas de sol,...). Incorporar los Datos en una Base de Datos estructurados

Fase 2-Limpieza y preparación de los datos

Limpiar datos, imputar o eliminar valores nulos, agregar nuevos datos si fuera necesario, eliminar "outliers" si fuera necesario,

Fase 3-Visualización grafica

Análisis Exploratorio (EDA) del histórico de las precipitaciones.

En primer lugar, ver la tendencia de las lluvias a lo largo de estos años. Analizar valores máximos en función del año o el mes. Así mismo analizar el numero de días que llueve anualmente en función de los años y finalmente intentar elucidar si llueve menos actualmente que en el pasado debido a un posible cambio climático.

En segundo lugar, analizar el resto de datos del dataset (presión, días de sol, temperatura,etc..) e intentar relacionar los datos con las precipitaciones.

Fase 4-Predicción de precipitaciones

Aplicar técnicas de aprendizaje supervisado y no supervisado.

Primero analizaremos mediante aprendizaje supervisado si podemos predecir lluvia o no (problema de clasificación) en función del resto de variables. Como ejemplo se propone utilizar el modelo "decision tree" o "Nearest Neighbours".

Finalmente, se aplicara técnicas de "clusterización" para agrupar los datos en función de sus similitudes.

El código de cada apartado se puede encontrar en el GITHUB

link: https://github.com/ManelForcales/prueba_tecnica

Resultados

Visualizaciones

He guardado datos de la estación Meteorológica de Barcelona del Observatorio Fabra (localizado a una altitud de 408 metros) desde el 1 de Enero del 1930 hasta el 31 de Diciembre del 2021.

Las Precipitaciones diarias están registradas en litros/metro cuadrado (l/m²).

La figura 1 muestra el histórico de precipitaciones diarias

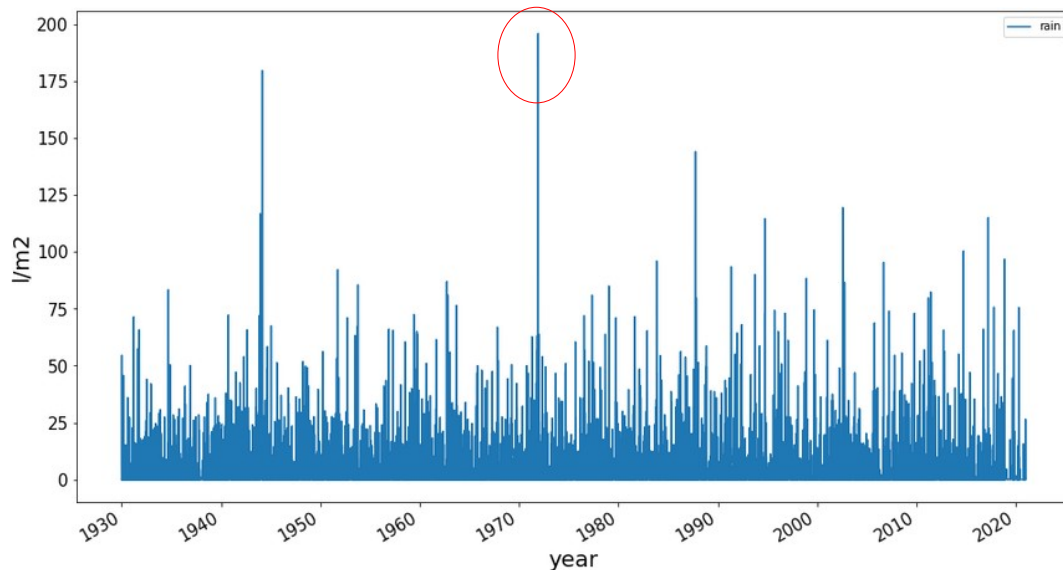


Figura 1

El máximo histórico de lluvias ocurrió el 5 de Diciembre de 1971 con 196 litros/m².

Ha llovido un 22.7 % de tiempo durante los últimos 90 años.

A nivel mensual como se observa en la figura 2, la estación de Otoño (21 Sept- 21 Dic) es la más lluviosa respecto a los días que no llueve con un promedio de 2.5 l/m² en Octubre.

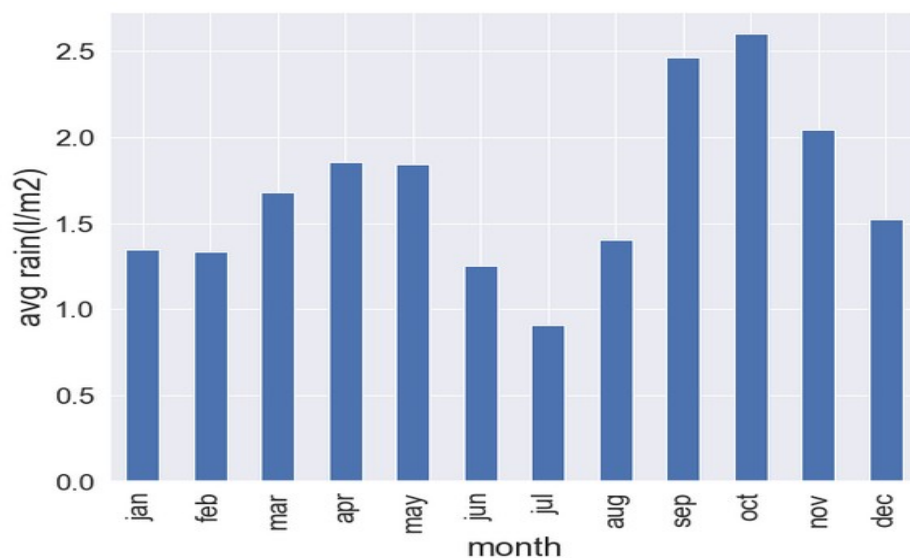


Figura 2

Si miramos el **promedio de las precipitaciones** anuales analizando solamente los días de lluvia observamos en la figura 3 una media anual de **7.5 litros/m²**. No se aprecia ninguna tendencia a lo largo de los años.

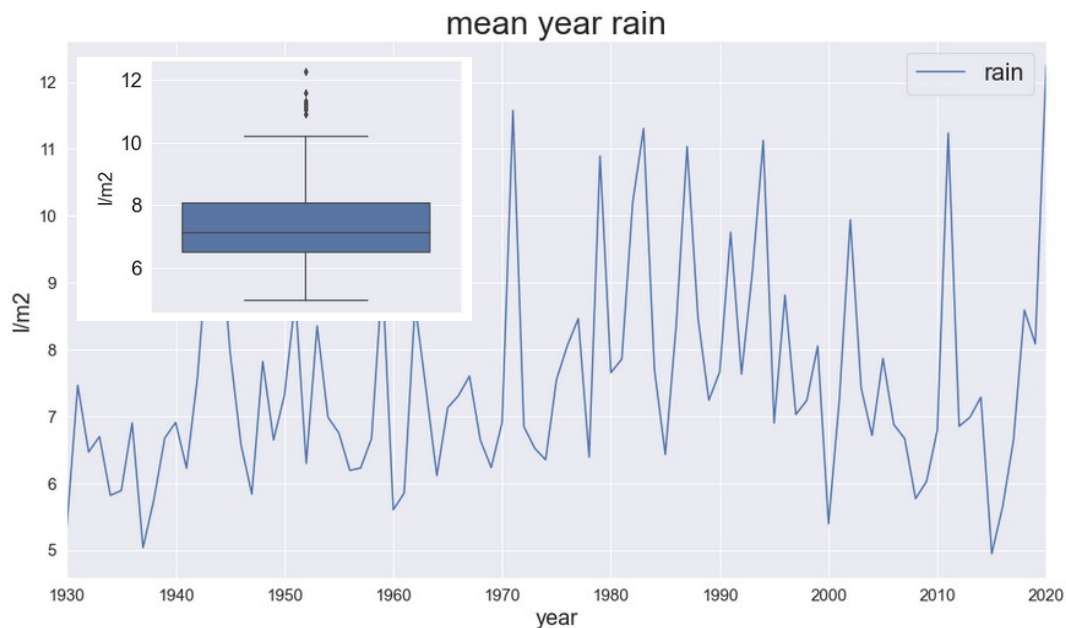


Figura 3

En Otoño además tal y como se aprecia en la figura 4 las precipitaciones son más elevadas.

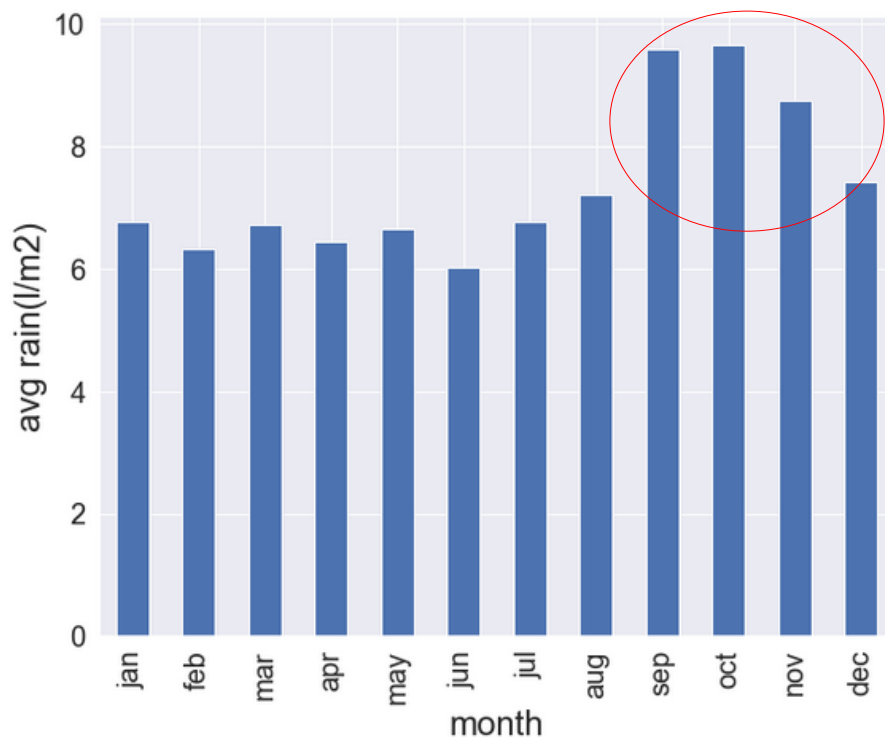


Figura 4

Podemos afirmar que actualmente llueve menos que en el pasado debido al cambio climático?

La respuesta la encontramos en la figura 5. Se han analizado 2 distribuciones de precipitaciones: la primera corresponde a precipitaciones anteriores al año 2000 y la 2ª para precipitaciones posteriores.

El análisis estadístico nos indica que el **P-value es > 5%** con lo cual **no podemos rechazar la hipótesis nula** (H_0 =Hay el mismo volumen de precipitaciones entre el presente y el pasado).

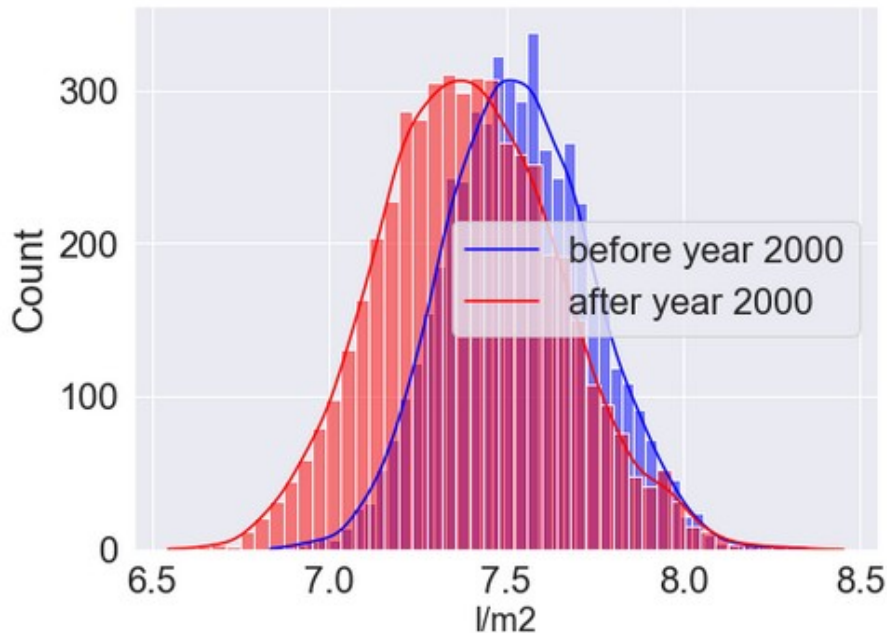


Figura 5

Por el contrario, si analizamos el nº total de días que llueve al año como muestra la figura 6, se aprecia una leve tendencia a la baja.

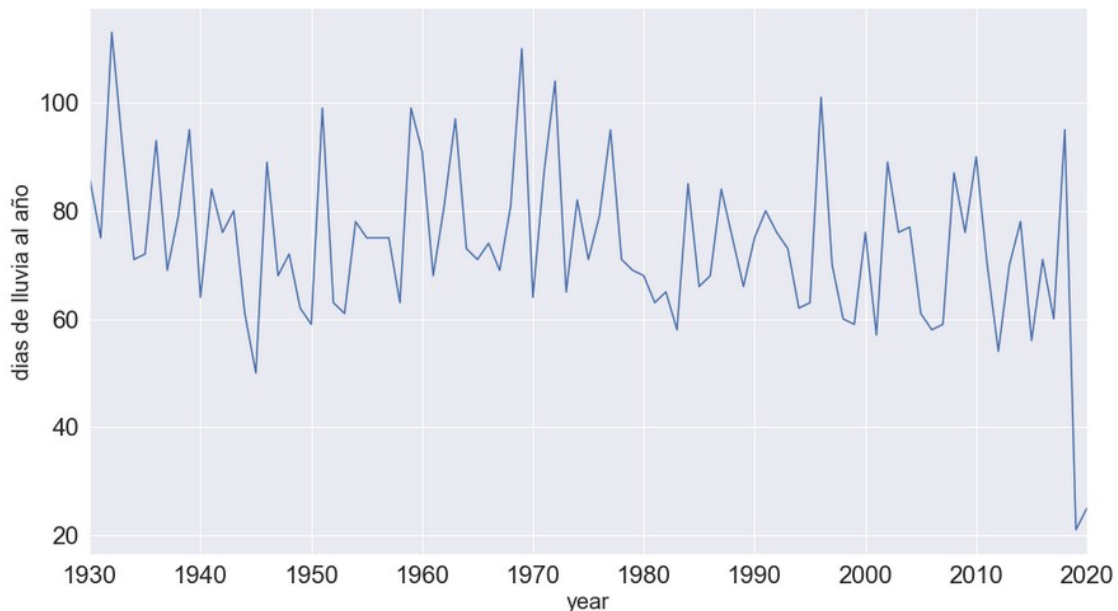


Figura 6

El análisis estadístico de estas 2 distribuciones se puede visualizar en la figura 7 y confirma lo siguiente: **en el pasado llovía más a menudo con una cierta cantidad y**

actualmente llueve menos veces pero cuando lo hace, llueve con más intensidad ya que el volumen de precipitaciones (l/m²) al final del año es similar.

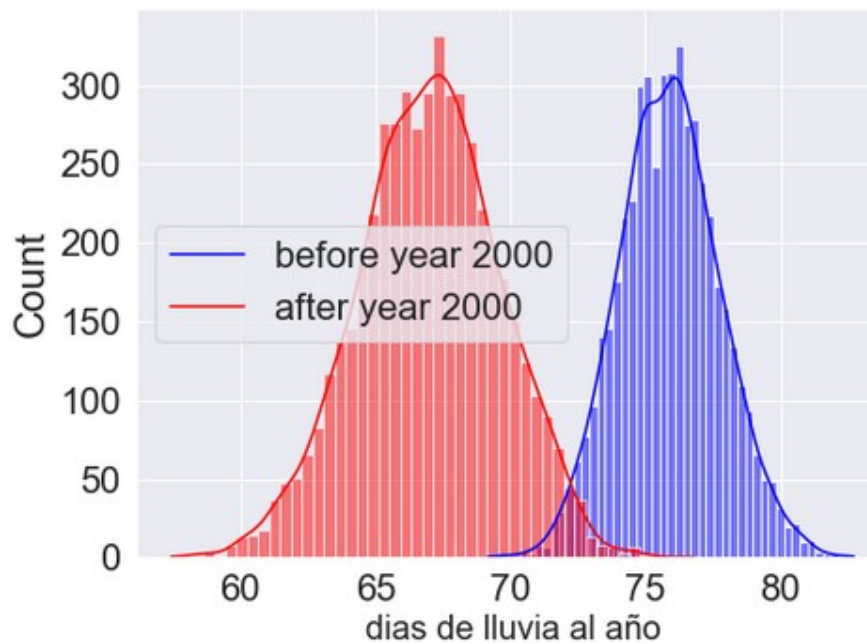


Figura 7

Estas diferencias también se aprecian en las precipitaciones caídas según la época del año o del mes como muestra la figura 8

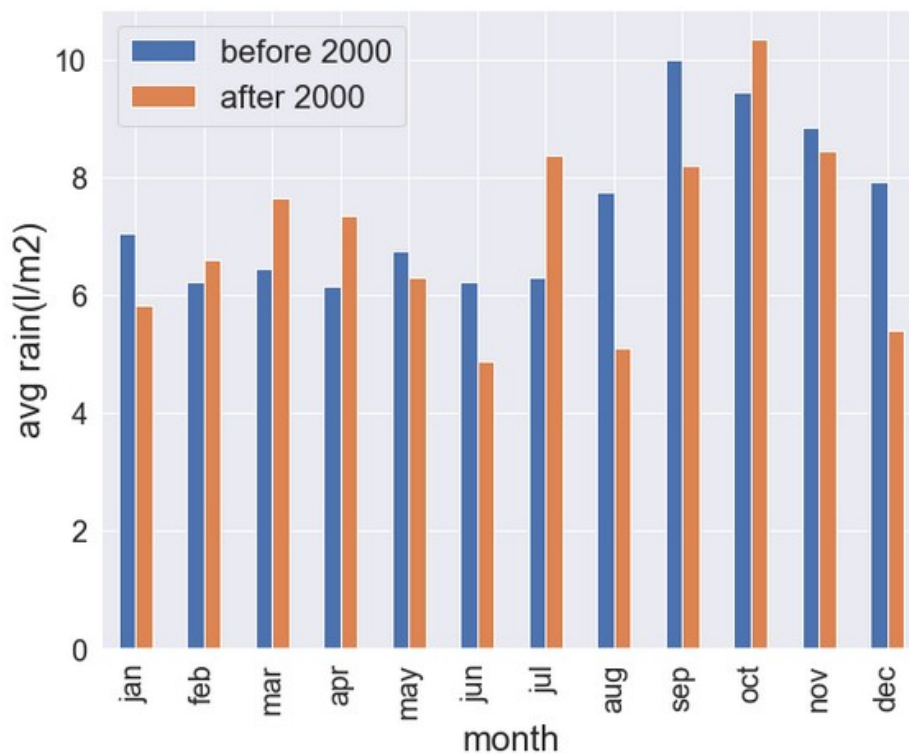


Figura 8

Vamos a analizar el resto de variables del "dataset" y ver que relación tienen:

En la figura 9 se observa por ejemplo que la temperatura media ($T_{\max}+T_{\min}/2$) ha subido más de un grado respecto a 1980. El resultado es significativo indicando que las temperaturas medias actuales son más altas que en el pasado presuntamente debido al cambio climático.

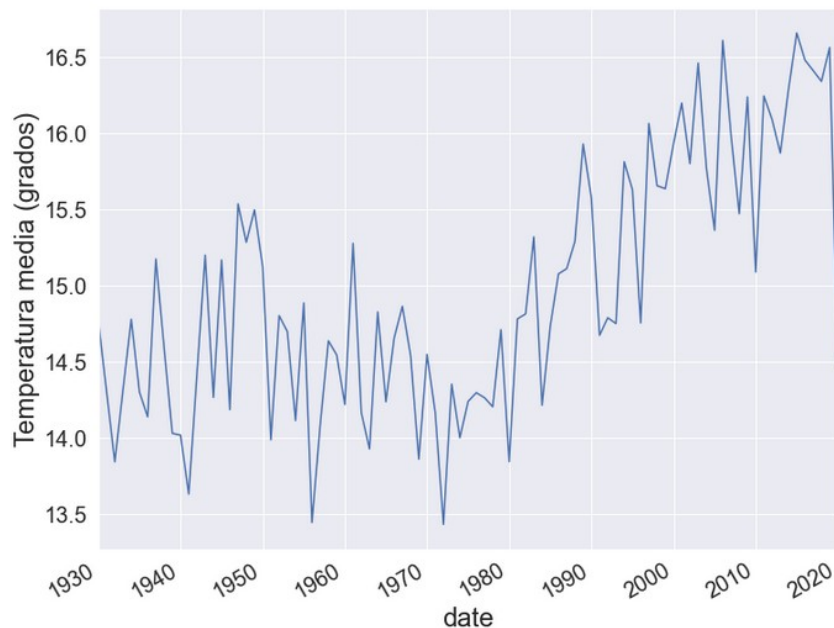


Figura 9

La tabla de correlaciones (tabla 1) muestra que las variables más correlacionadas son:

- Horas de sol con temperaturas media (0.43)
- Precipitaciones con horas de sol (-0.32)
- Temperatura media con mes (0.25)
- Presión media con las Horas de sol (0.22)

	precipitaciones	horas_sol	vel_media	year	month	temp_media	pres_media	lluvia
precipitaciones	1	-0.32	0.11	-0.01	0.03	-0.09	-0.18	0.47
horas_sol	-0.32	1	-0.01	0.12	-0.05	0.43	0.22	-0.5
vel_media	0.11	-0.01	1	-0.01	0	-0.17	-0.15	0.05
year	-0.01	0.12	-0.01	1	0	0.07	-0.01	0.02
month	0.03	-0.05	0	0	1	0.25	0.01	0
temp_media	-0.09	0.43	-0.17	0.07	0.25	1	0.11	-0.14
pres_media	-0.18	0.22	-0.15	-0.01	0.01	0.11	1	-0.3
lluvia	0.47	-0.5	0.05	0.02	0	-0.14	-0.3	1

Tabla 1

Estos datos se pueden resumir en la última fila: se observa una correlación entre lluvia (si/no o 1/0) y un menor numero de horas de sol (día nublado puede ser indicador de lluvia) y por tener una menor presión atmosférica (baja presión puede ser indicador de lluvia). En la figura 10 se observa lo comentado arriba

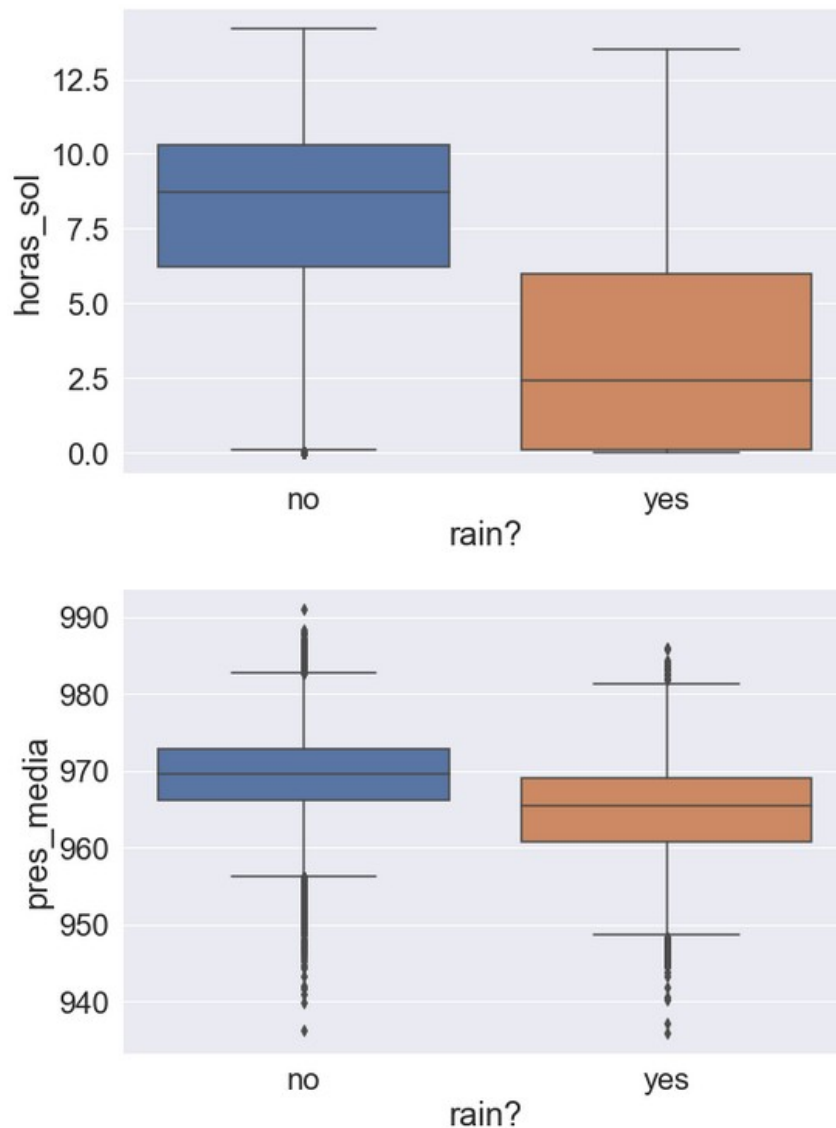


Figura 10

Predicciones

Aprendizaje Supervisado

Para predecir si va a llover o no, estudiaremos el “dataset” a partir del año 1984 en adelante. Esto es debido a que para fechas anteriores, muchos valores están incompletos y no es posible inferir su valor.

La nube de puntos de la figura 11 (horas_sol vs temp_media) nos indica que no hay una clara separación entre los días que ha llovido y los que no. Parece que será un reto obtener un buen clasificador utilizando estas variables.

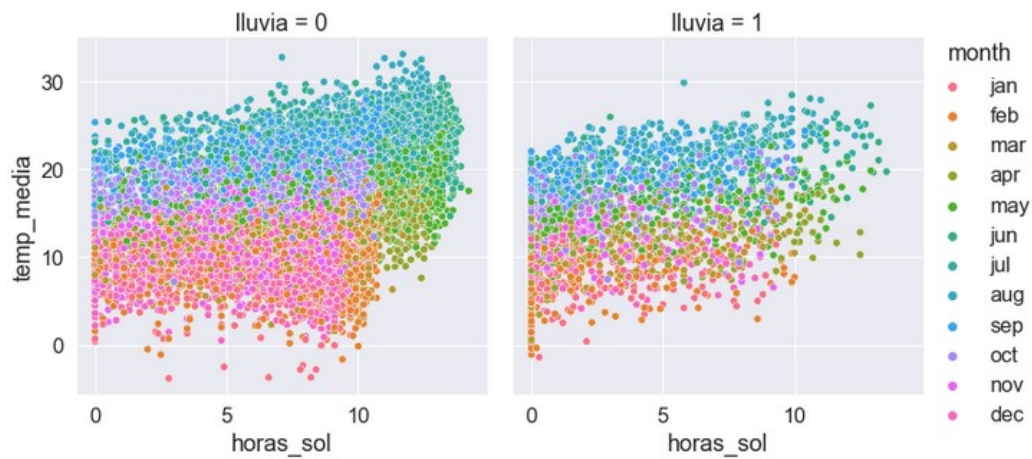


Figura 11

Otro ejemplo lo podemos ver en la figura 12 cuando utilizamos la presión_media vs temperatura_media como variables para predecir lluvia. No existe una clara separación entre los puntos



Figura 12

El Dataset tiene más de 10000 datos (80% del total) donde no existe lluvia y alrededor de 2600 datos (20% del total) donde existe lluvia como se aprecia en la figura 13.

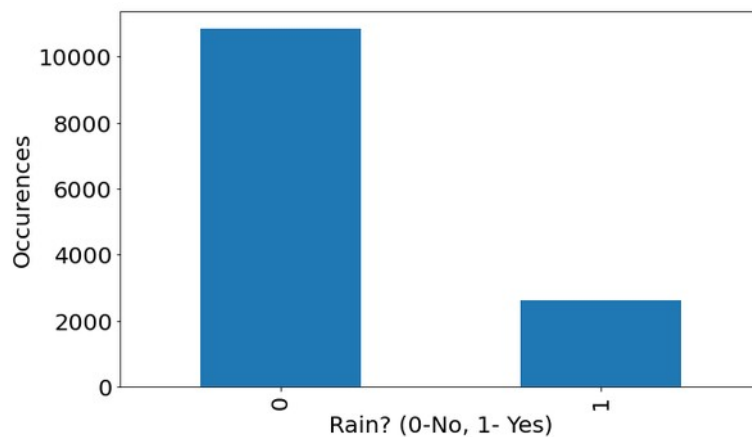


Figura 13

Por tanto tenemos un dataset no balanceado (aka Imbalance). Nuestro modelo debe predecir por encima del 80% sino no es mejor que un modelo puramente aleatorio.

Aplicando “decision tree” obtenemos un puntuación de 86%, la matriz de confianza de la figura 14, muestra que no podemos predecir bien los días de lluvia

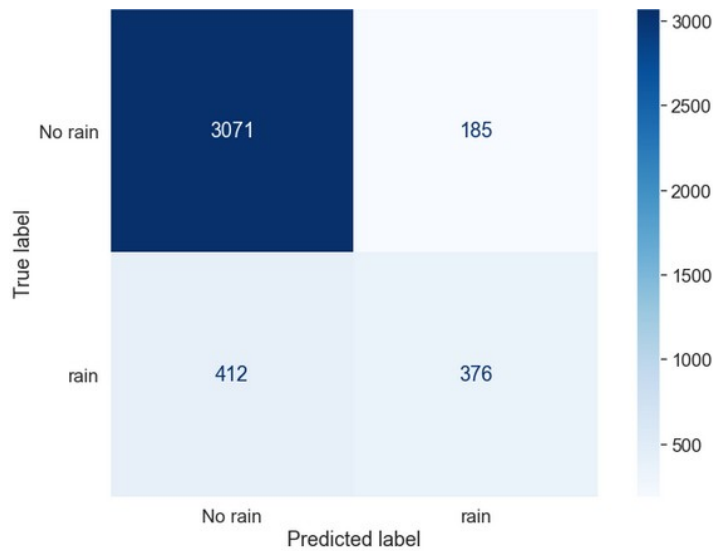


Figura 14

si aplicamos Nearest Neighbors Classification, la matriz de confianza mejora

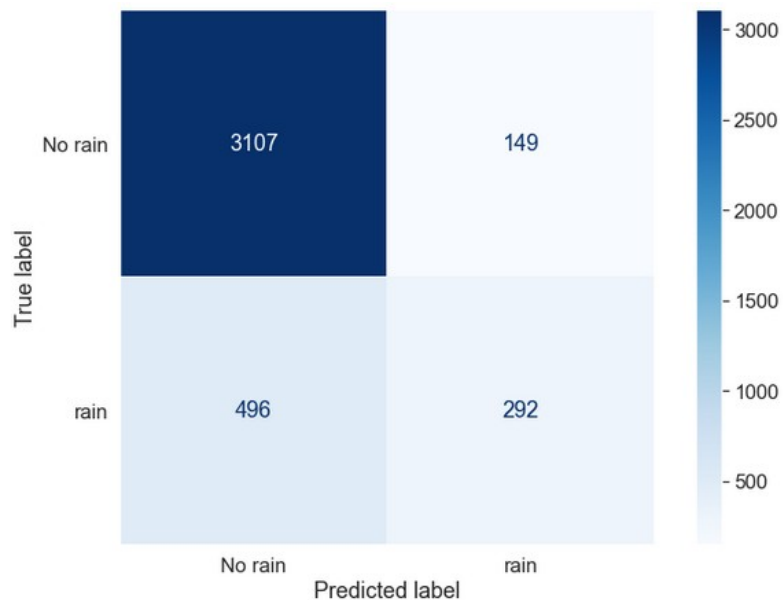


Figura 15

He aplicado técnicas de estandarización/normalización de variables, pero aun así los resultados no mejoran mucho respecto al 86% inicial.

Aprendizaje No Supervisado

Para el proceso de clusterización he quitado las variables referentes a precipitaciones y lluvias así como el mes o año para hacerlo mas sencillo. La idea es determinar el numero de clúster donde estas variables tienden a agruparse debidos a punto en común. He usado la librería Kmeans ya que es una de las mas comunes y bastante directa de implementar.

Primero determino el numero de los diferentes clúster y veo en la figura 16 que la solución más optima es alrededor de 5- 7 clúster con el método del "elbow"

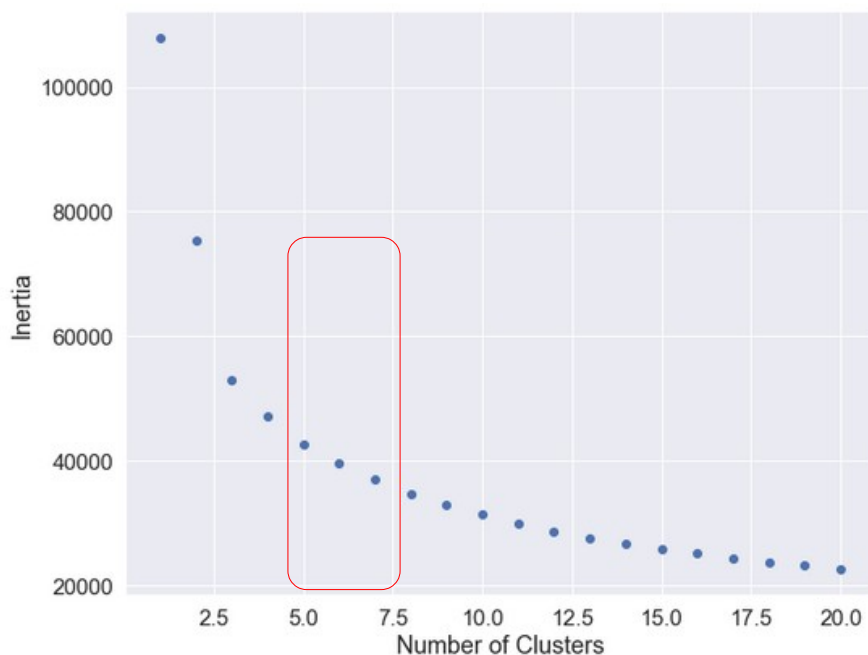


Figura 16

Mientras que usando el método de Silhouette de la figura 17, el valor más optimo son 3 clúster

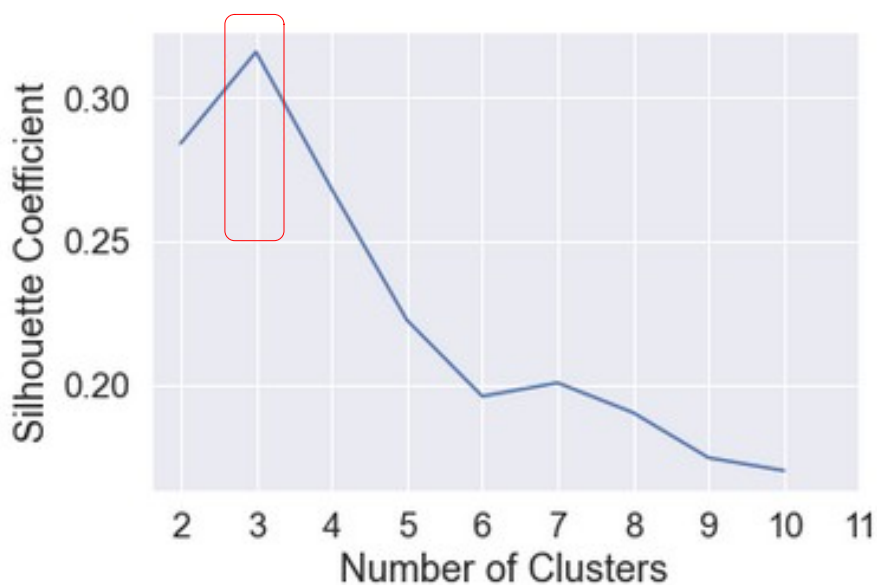


Figura 17

Elijo 3 clúster y evalúo los resultados mirando que valores tienen asignados cada clúster (0,1,2)

En la figura 18 grafica se muestra un ejemplo de los datos iniciales (sabiendo cuales son los que tienen precipitaciones y los que no) y los resultados derivados de la clusterización. Básicamente los puntos del clúster (label 1) són los que aglutinan la mayor parte de datos que contienen las precipitaciones pero como no hay una clara separación entre puntos muchos de ellos quedan mis-clasificados.



Figura 18

Conclusiones

Con las variables actuales es difícil predecir si lloverá o no en Barcelona con datos de temperatura, viento o presión porque no hay una fuerte correlación con las medidas de precipitaciones. Todo parece indicar que a nivel histórico cada vez llueve menos pero cuando lo hace, caen lluvias con mas intensidad.