Instituto Tecnológico Telefónica

**Programa superior en Big Data**

# Efectos del confinamiento en la calidad del aire de la Comunidad de Madrid

**Grupo 7**

**Susana Vara**

**Carmen Reina**

**Alfonso Gallardo**

**Raúl Hervás**

**Walter Ronceros**

Indice

Efectos del confinamiento en la calidad del aire de la Comunidad de Madrid 1

1.- Introducción 3

2.- Objetivos y alcance de proyecto 3

2.1. Motivación del proyecto 3

2.2. Objetivos finales 3

2.3. Limitaciones 3

3.- Desarrollo 3

3.1. Exploración y elección de los datasets 3

3.2. Análisis y unificación de los datasets 3

3.3. Carga de fuentes de datos 3

3.4. Comprensión y análisis descriptivo de los datos 3

3.5. Pre-procesado de datos 3

3.6. Preparación del modelo 4

3.7. Evaluación de algoritmos 4

3.8. Mejora de precisión del modelo 4

4.- Conclusión 4

5.- Anexos 4

6.- Bibliografía 4

# 1.- Introducción

Durante la paralización del país debido al COVID’19 en el mes de marzo es evidente que, sin actividad humana, la naturaleza se abre camino. Durante estas largas semanas de confinamiento se han visto noticias donde animales nunca antes vistos en ciudades se aventuraban a hacer turismo, el olor en la ciudad cambiaba radicalmente pudiéndose percibir olor a vegetación incluso la falta de ruido desorientaba a ciertos animales de ciudad acostumbrados al estrés y contaminación acústica diaria.

Con este estudio queremos ver qué puntos de inflexión pueden verse en la calidad del aire de la Comunidad de Madrid (de ahora en adelante CAM )en las fechas de confinamiento para tratar de obtener el tiempo que tarda la naturaleza en llegar a la máxima calidad de aire y el ser humano en devolverla a sus niveles “habituales”.

# 2.- Objetivos y alcance de proyecto

## 2.1. Motivación del proyecto

Gracias a toda la información recogida de agentes contaminantes por las estaciones situadas por toda la CAM y estando al alcance de todos, así como información metereológica como velocidad y dirección del viento, precipitaciones o radiación solar entre otros queremos contrastarla para tratar de encontrar alguna relación directa entre ellas que pueda ayudar a predecir los niveles de calidad del aire a futuro.

## 2.2. Objetivos finales

El objetivo del presente proyecto consiste en obtener un programa informático que apoyándose en las APIs donde se publica la información de la CAM ( incluido el ayuntamiento de Madrid ),desde enero 2020 hasta final del estado de alarma, sea capaz de predecir o estimar la calidad del aire una vez vuelta la actividad normal.

## 2.3. Limitaciones

Para este estudio damos por hecho ciertas limitaciones insalvables como son :

* No todas las estaciones contienen medición de todos los agentes contaminantes.
* Las mediciones meteorológicas no se pueden prever
* El índice ICA (Índice Calidad del Aire) se obtiene de la medición más adversa de 5 agentes contaminantes por lo que no puede utilizarse para buscar correlaciones directas

Siendo así consideramos que esto va a ser un churro.

# 3.- Desarrollo

La coordinación dentro del equipo la organizamos en Trello, compartiendo ahí tanto la información de negocio que vamos recopilando como los scripts o notebook que vamos generando.

Para el desarrollo del proyecto se empleará el lenguaje de programación Python para la conexión a las APIs y tratamiento de los datos para obtener los datases a estudiar. Se utilizarán varias librerías para el tratamiento y modelado de los datos. Estas son: Pandas, matplotlib, scikit-learn (incluyendo varias sublibrerías), seaborn, Prophet,xgboost.

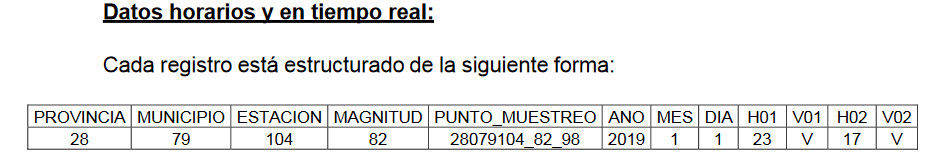
Compartimos un notebook en google colab, así todos los miembros del equipo podemos ver el desarrollo del proyecto:

<https://colab.research.google.com/drive/1hAkG64bXv-BuhfkH0_3qH2lx-pIUcLTK?usp=sharing#scrollTo=B8BFeqAzXKHV>

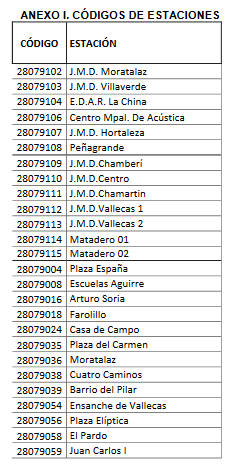
## 3.1. Exploración y elección de los datasets

Elegimos los datos del ayuntamiento y la comunidad de Madrid tanto de calidad del aire como metereológicos . Los datos origen vienen todos con el mismo formato representando para cada hora en cada estación de medida la magnitud a medir.

Las estaciones son los lugares dentro del municipio donde se miden las magnitudes de los contaminantes de calidad del aire o de las magnitudes metereológicas.

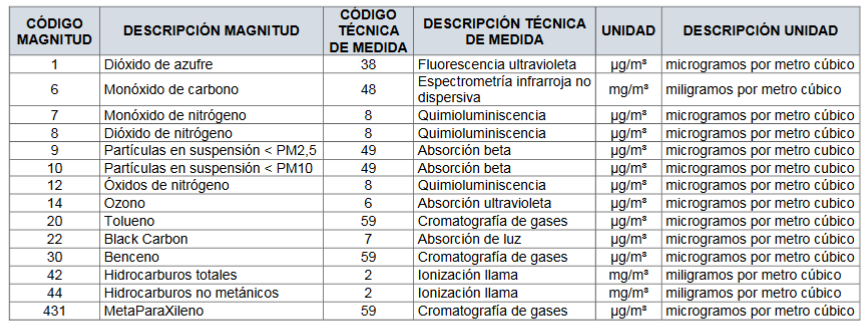


* Las estaciones dentro del municipio de Madrid y en el resto de municipios de la comunidad de Madrid:

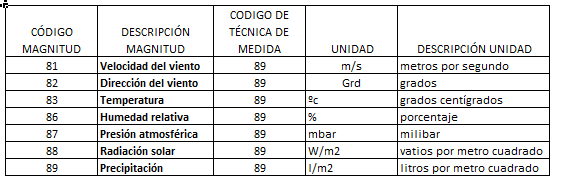




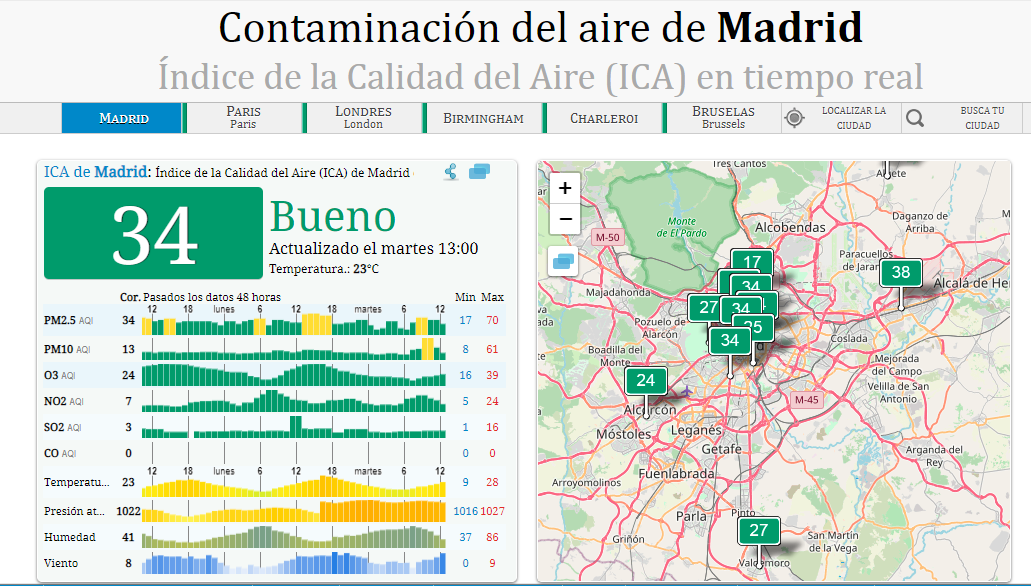
* Las magnitudes de calidad del aire y las técnicas de medida:



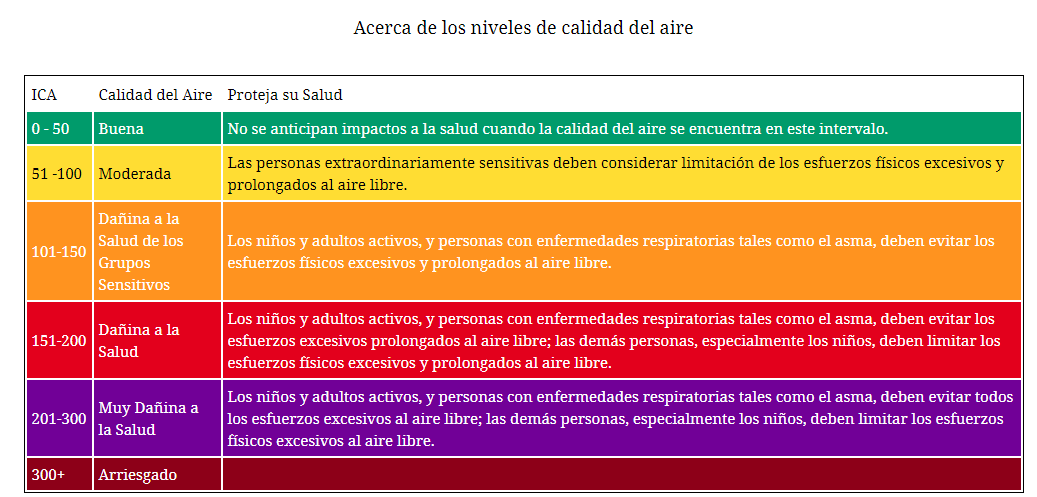
* Las magnitudes metereológicas en las estaciones :



Con toda esta información comenzamos a investigar cómo se mide la calidad del aire, en el sentido de qué marca ese valor que nos comunican en distintos medios de información . Encontramos que para medir la calidad del aire a nivel mundial, se utiliza un índice que llama ICA ( Índice de Calidad del aire) cuyo valor se corresponde con el mayor valor de los contaminantes que se miden en una estación ( dichos valores multiplicados por un factor , para que sean comparables). Si el municipio tiene más de una estación , el valor del ICA total , será el máximo de los ICA de sus estaciones.



(Fuente: <http://aqicn.org/city/madrid/es/>)



Unas definiciones a tener en cuenta para comprender los datasets :

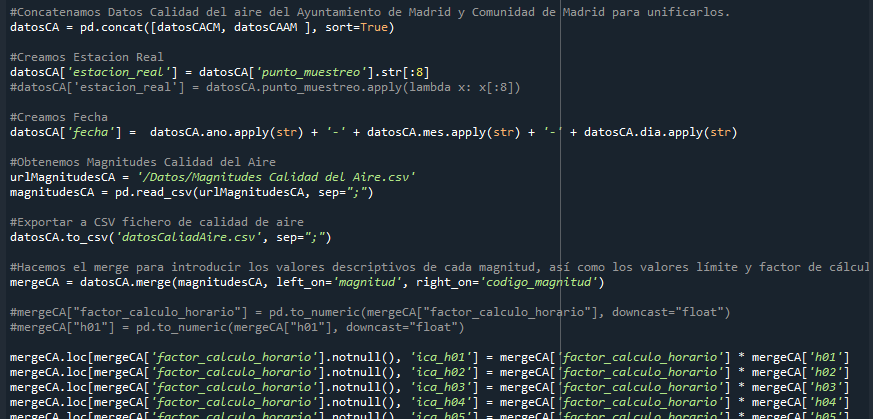
* ICA: Índice de calidad del aire.
* VL: Valor límite. Es un nivel fijado basándose en conocimientos científicos, con el fin de evitar, prevenir o reducir los efectos nocivos para la salud humana, para el medio ambiente en su conjunto y demás bienes de cualquier naturaleza que debe alcanzarse en un periodo determinado y no superarse una vez alcanzado.
* Índice global: Se corresponderá con el mayor de los índices parciales obtenidos para cada contaminante.
* Índice parcial diario: Es el índice obtenido para cada contaminante, asociando el valor 100 con la concentración que representa el valor límite, calculado con las concentraciones diarias u octohorarias de cada contaminante, en función de cómo esté expresado su valor límite, y con la peor de las concentraciones horarias del día en cuestión.
* Índice parcial horario: Es el índice obtenido para cada contaminante, asociando el valorlímite con la concentración que representa el valor límite horario de cada contaminante, obtenido teniendo en cuenta la más elevada de las concentraciones horarias de cada uno delos contaminantes.

## 3.2. Análisis y unificación de los datasets

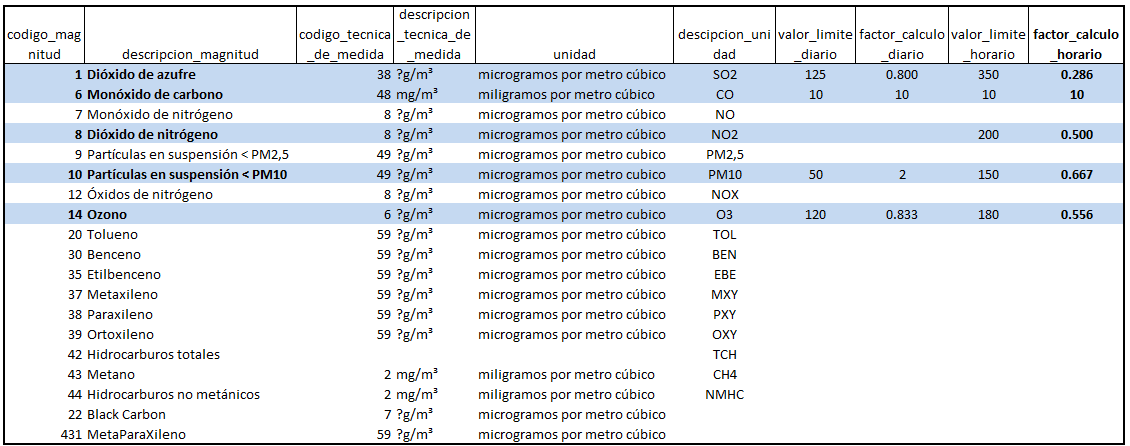
Comenzamos con el análisis de los datasets y lo primero que planteamos es unir por un lado en un dataset los datos de calidad del aire y por otro lo datos metereologicos.

Creamos el campo fecha en un formato válido para el análisis de series temporales.

Para el cálculo del ICA realizamos la multiplicación de los valores de las magnitudes por cada factor :



Estos son los factores de cálculo de los 5 contaminantes que se utilizan para calcular el ICA:



## 3.3. Carga de fuentes de datos

La carga inicial le hemos realizado bajando los csv de los portales de datos abiertos de la comunidad del Madrid y del ayuntamiento de Madrid, tanto para la calidad del aire como para la información metereológica.

<https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

<http://datos.comunidad.madrid/catalogo/dataset/calidad_aire_datos_historico/resource/f525015e-5484-4839-a99e-1588566e93eb>

<https://datos.madrid.es/sites/v/index.jsp?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>

<http://datos.comunidad.madrid/catalogo/dataset/calidad_aire_datos_meteo_historico/resource/f8699435-ba20-45df-a37d-b3c44832d140>

Más adelante desarrollaremos la descarga desde la API de forma automática.

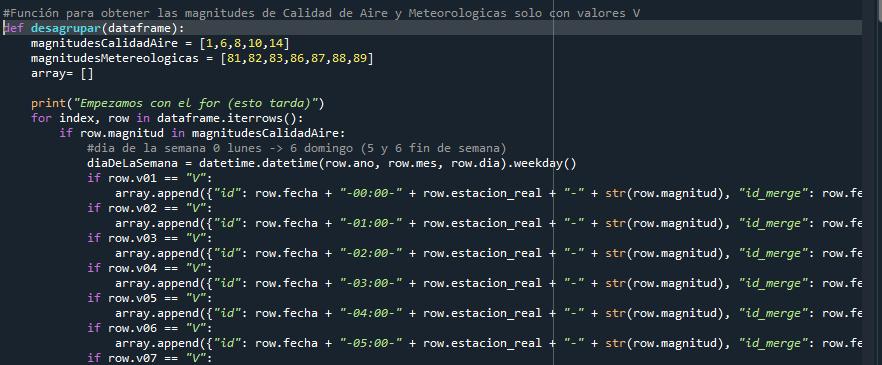
## 3.4. Comprensión y análisis descriptivo de los datos

Una vez cargados los datos en el formato que viene, con los datos de las magnitudes en 24 columnas una para cada hora, pensamos que es mejor para estudiar series temporales desagrupar esa información y poner las magnitudes de calidad del aire en filas. De tal forma en nuestro dataset tendremos para cada dia, hora, estación , magnitud de calidad del aire , su valor , su ICA, y los datos metereológicos en ese momento.

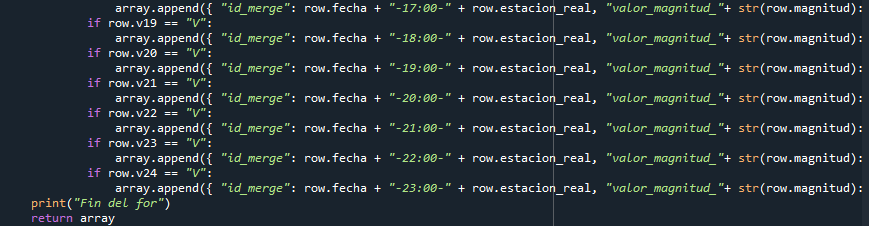
Para esto creamos una función en phyton para transformar el dataset. Convertimos el dataset en un array donde cada fila tendrá los siguientes valores, siempre y cuando la medición en esa hora de esta magnitud en esa estación sea válida ( V):

if row.v01 == "V":

**array**.append({"id": row.fecha + "-00:00-" + row.estacion\_real + "-" + str(row.magnitud), "id\_merge": row.fecha + "-00:00-" + row.estacion\_real, "fechahora": row.fecha + " 00:00", "fecha": row.fecha, "hora": "1", "estacion\_real": row.estacion\_real, "magnitud": row.magnitud, "descripcion\_magnitud": row.descripcion\_magnitud, "factor\_calculo\_horario": row.factor\_calculo\_horario, "ica\_parcial": row.ica\_h01, "valor\_magnitud":row.h01, "provincia": row.provincia, "municipio": row.municipio, "dia\_de\_la\_semana": diaDeLaSemana})

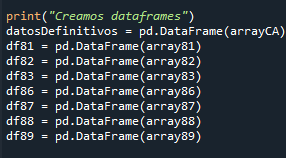


………………………………. Así para todas las horas



A este array con los datos de calidad lo volvemos a convertir en un dataframe y a este le añadimos las columnas de las variables metereologicas.





Este dataset llamado datosDefinitivos es el que empezaremos a procesar.

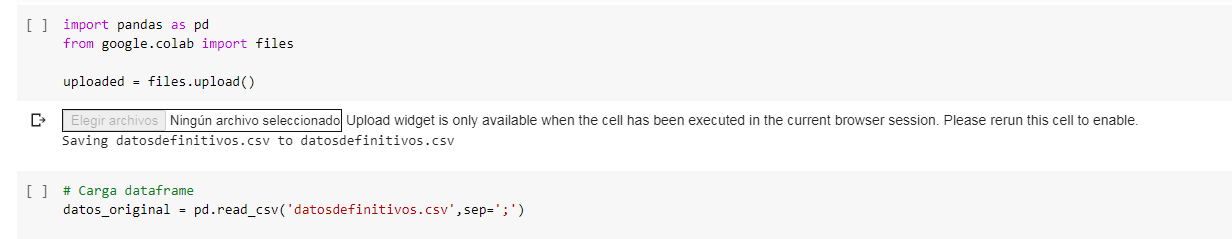
## 3.5. Pre-procesado de datos

El pre-procesamiento y análisis de los datos lo hacemos en google colab con la funcionalidad que tiene de trabajar con notebook.

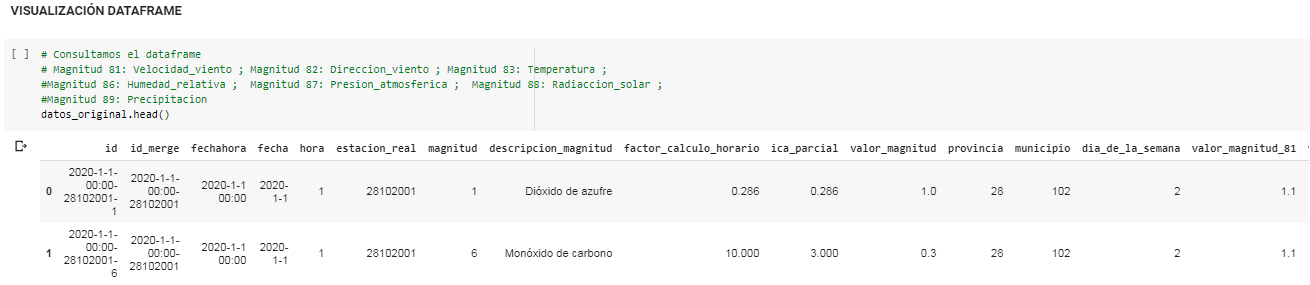
**Parte 1: Carga de datos :**



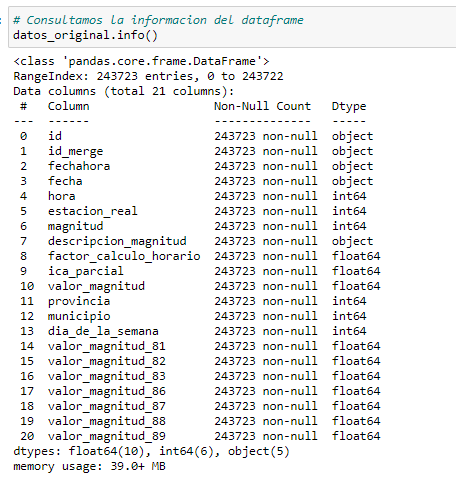
Importamos los datos con google.colab:

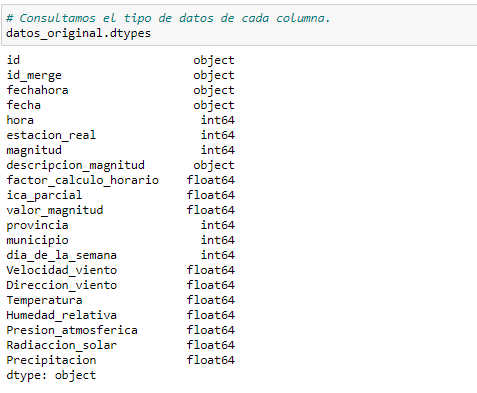


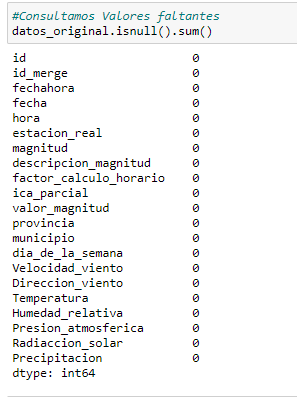
Visualizamos el dataframe :

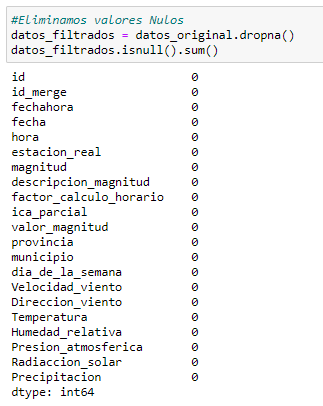


Consultamos la información del dataframe:

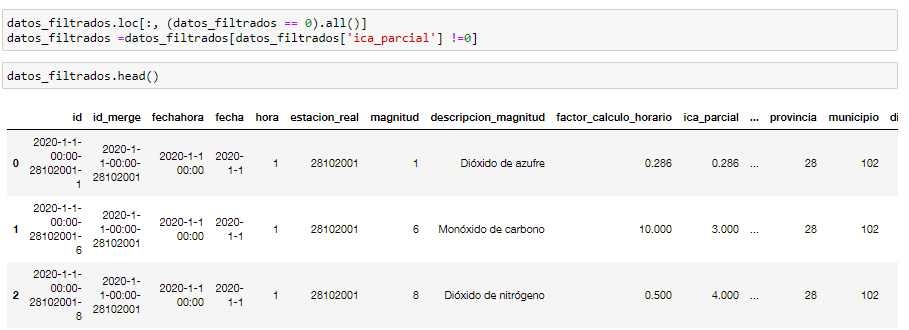


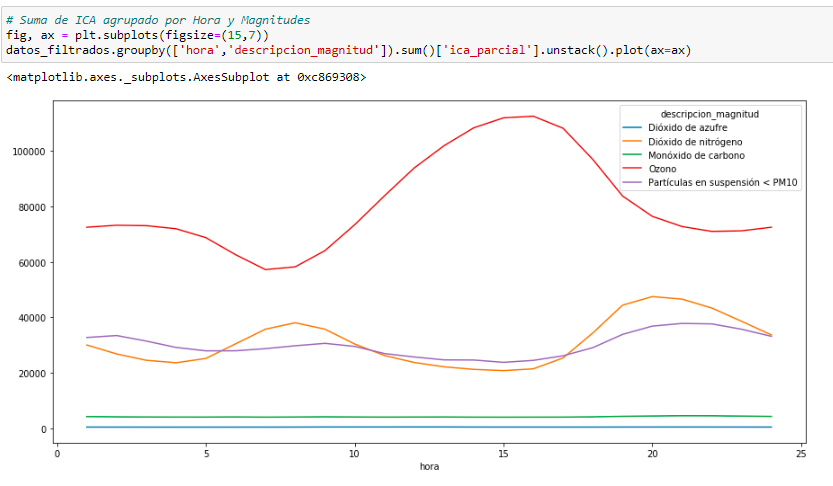


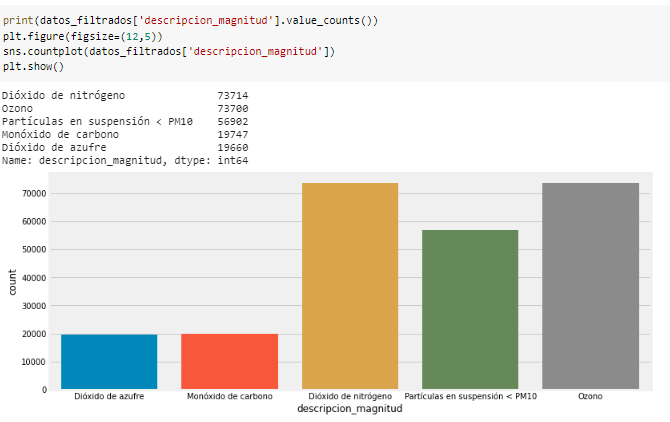


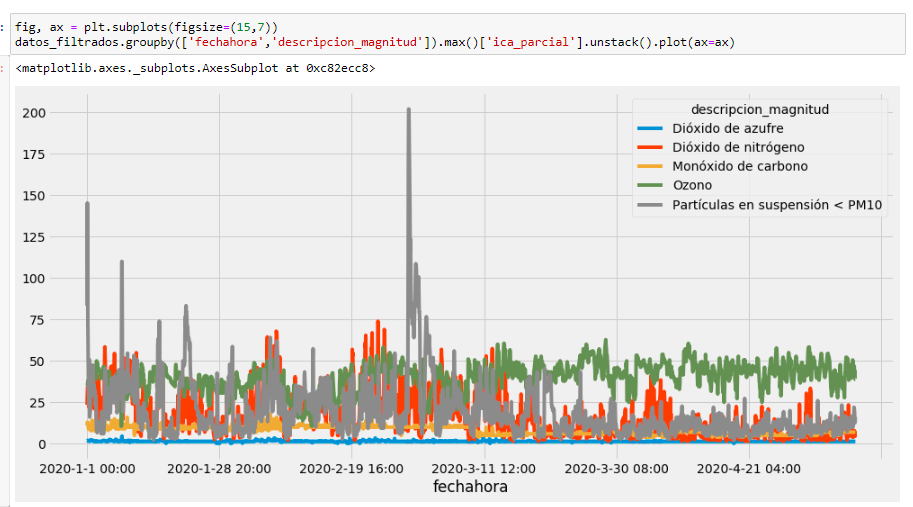


Hacemos un filtrado quedándonos sólo con los que tienen un valor ICA distinto a 0

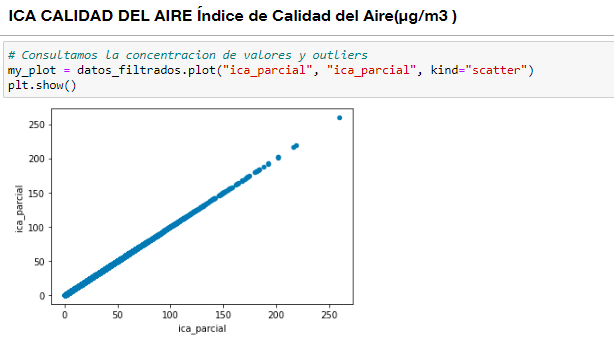


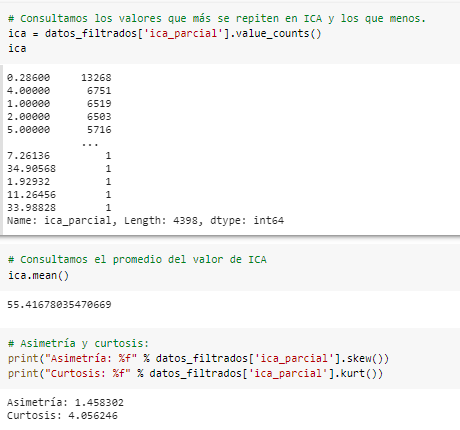


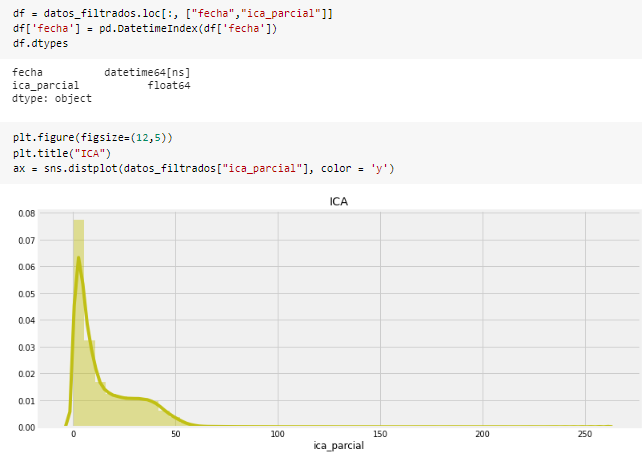


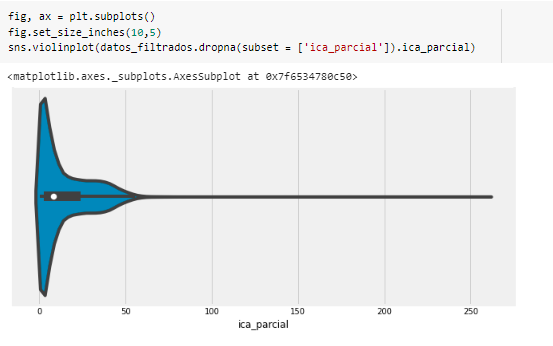


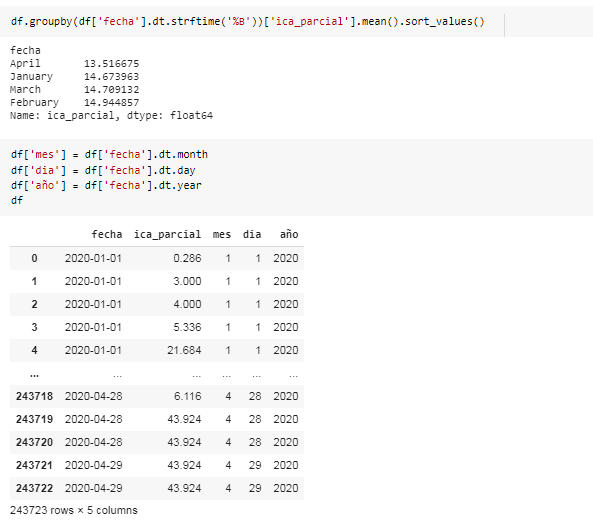
**Parte 2: Compresión de los datos:**

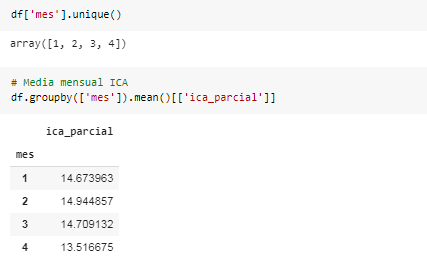


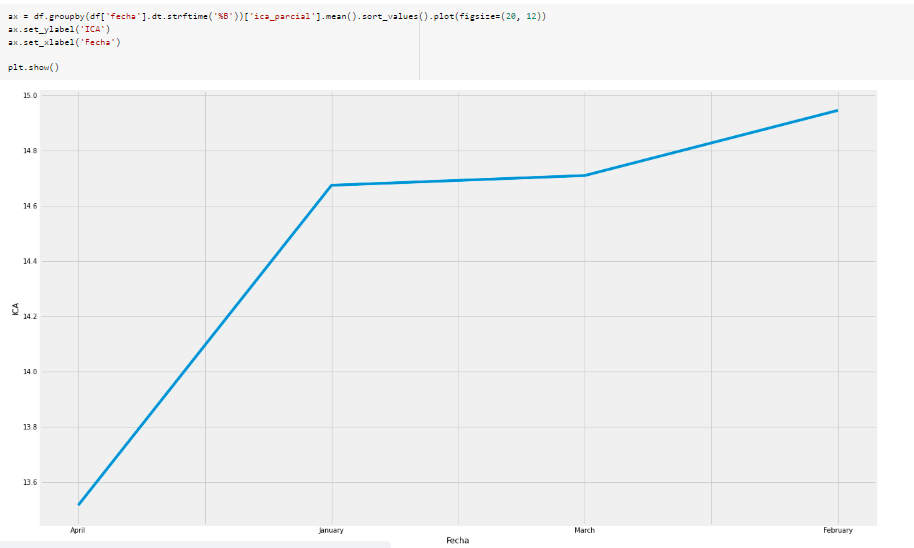


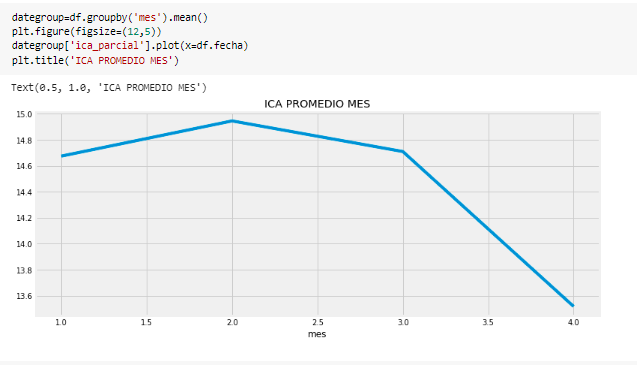


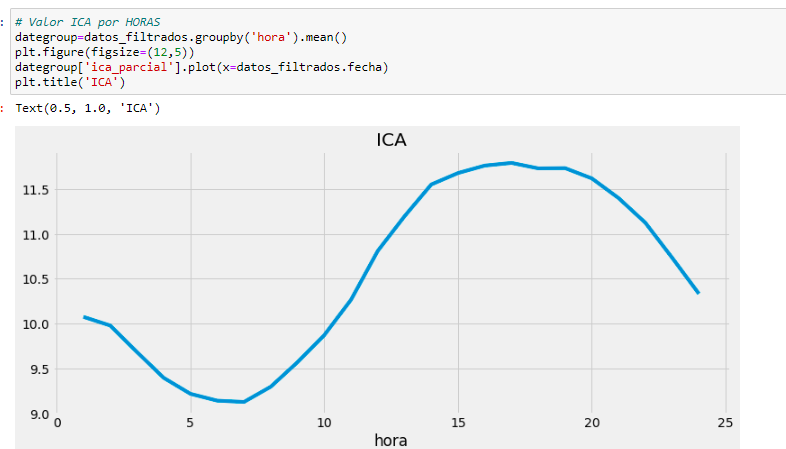






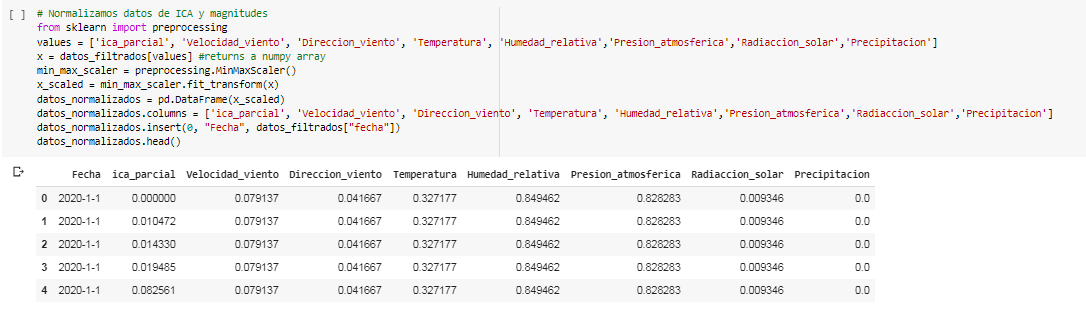


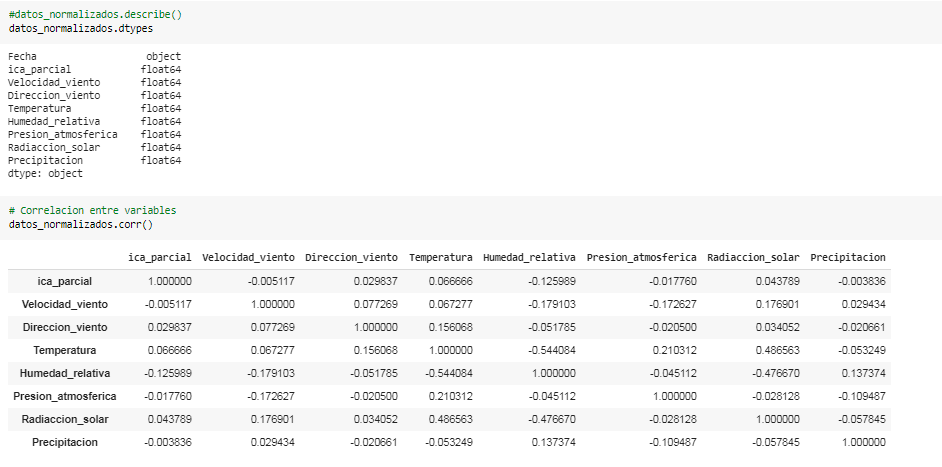


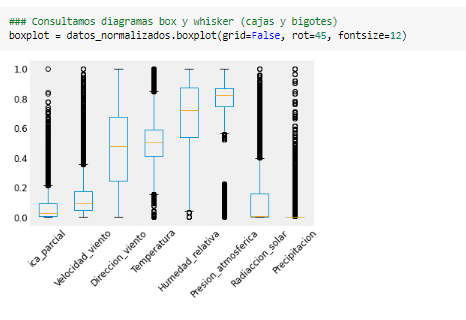


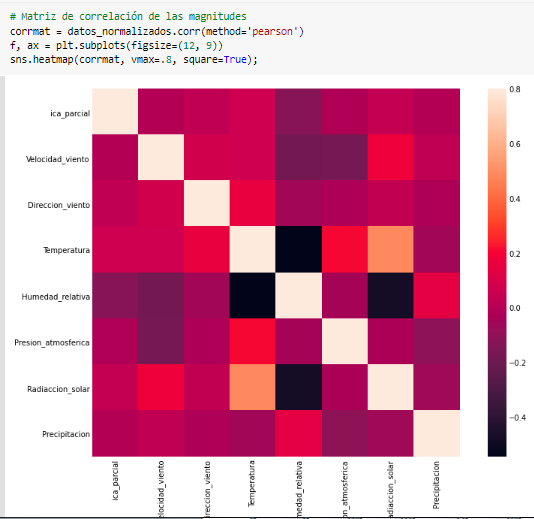
## 3.6. Preparación del modelo

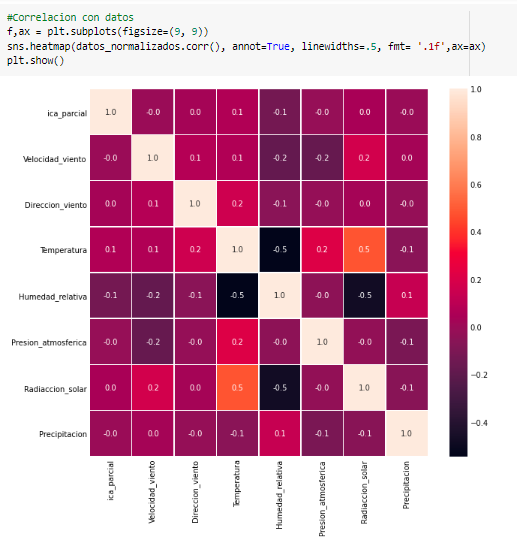
Para preparar los datos para el modelo comenzamos con una visualización de los datos para lo cual normalizando los datos :





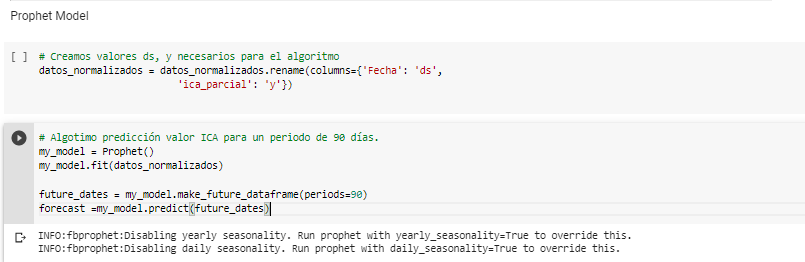


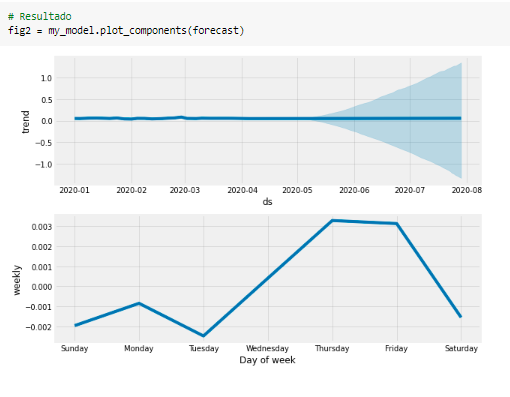




## 3.7. Evaluación de algoritmos

Comenzamos con la modelización , para lo cual el primer modelo que probamos es Prophet Model:





## 3.8. Mejora de precisión del modelo

# 4.- Conclusión

# 5.- Anexos

# 6.- Bibliografía