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](#_Toc43461052)

[1.- Introducción 3](#_Toc43461053)

[1.1 Acerca del Dióxido de Nitrógeno (NO2) 4](#_Toc43461054)

[1.2 Adaptación a la Directiva 2008/50/CE en España 6](#_Toc43461055)

[1.3 Protocolo de Contaminación en Madrid 6](#_Toc43461056)

[2.- Objetivos y alcance de proyecto 8](#_Toc43461057)

[2.1. Motivación del proyecto 8](#_Toc43461058)

[2.2. Objetivos finales 8](#_Toc43461059)

[2.3. Limitaciones 9](#_Toc43461060)

[3.- Desarrollo 9](#_Toc43461061)

[3.1. Exploración y elección de los datasets 10](#_Toc43461062)

[3.2. Análisis y unificación de los datasets 14](#_Toc43461063)

[3.3. Carga de fuentes de datos 15](#_Toc43461064)

[3.4. Comprensión y análisis descriptivo de los datos 17](#_Toc43461065)

[3.5. Pre-procesado de datos 19](#_Toc43461066)

[3.6. Preparación del modelo 27](#_Toc43461067)

[3.7. Evaluación de algoritmos 30](#_Toc43461068)

[3.8. Mejora de precisión del modelo 31](#_Toc43461069)

[4.- Conclusión 31](#_Toc43461070)

[5.- Anexos 31](#_Toc43461071)

[6.- Bibliografía 32](#_Toc43461072)

# 1.- Introducción

La contaminación del aire impacta significativamente en la salud de la población europea, especialmente en las zonas urbanas. Asimismo, tiene un considerable impacto económico, puesto que aumenta los costes sanitarios derivados de las enfermedades que provoca. Esto ha llevado a la calidad del aire a convertirse en una de las principales cuestiones políticas de las últimas décadas en el ámbito europeo.

La importancia de este tema se ve reflejada en las múltiples Directivas Europeas que tratan de establecer unos mínimos de calidad de aire para proteger a la población de excesivos niveles de contaminación.

La progresiva regulación que se ha ido aprobando desde finales de los 70 se ha consolidado en la Directiva 2008/50/CE relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa. Haciendo una muy breve puntualización jurídica, una Directiva Europea es una norma de ámbito supranacional en la que se fija una serie de objetivos a alcanzar, pero deja libertad a los Estados miembros para elegir los medios que consideren más convenientes para alcanzarlos. Dicha norma comunitaria, en su artículo primero señala cuáles son sus objetivos:

* Definir y establecer objetivos de calidad del aire ambiente para evitar, prevenir o reducir los efectos nocivos para la salud humana y el medio ambiente en su conjunto.
* Evaluar la calidad del aire ambiente en los Estados miembros basándose en métodos y criterios comunes.
* Obtener información sobre la calidad del aire ambiente con el fin de ayudar a combatir la contaminación atmosférica y otros perjuicios y controlar la evolución a largo plazo y las mejoras resultantes de las medidas nacionales y comunitarias.
* Asegurar que esa información sobre calidad del aire ambiente se halla a disposición de los ciudadanos.
* Mantener la calidad del aire, cuando sea buena, y mejorarla en los demás casos.
* Fomentar el incremento de la cooperación entre los Estados miembros para reducir la contaminación atmosférica.

Los contaminantes más serios en términos de deterioro para la salud son las partículas en suspensión PM2.5 y los óxidos de Nitrógeno NOx.

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”.

## 1.1 Acerca del Dióxido de Nitrógeno (NO2)

El NO2 se forma a nivel de suelo a partir de emisiones relacionadas con la quema de combustibles fósiles de vehículos, centrales de energía, industrias, etc.

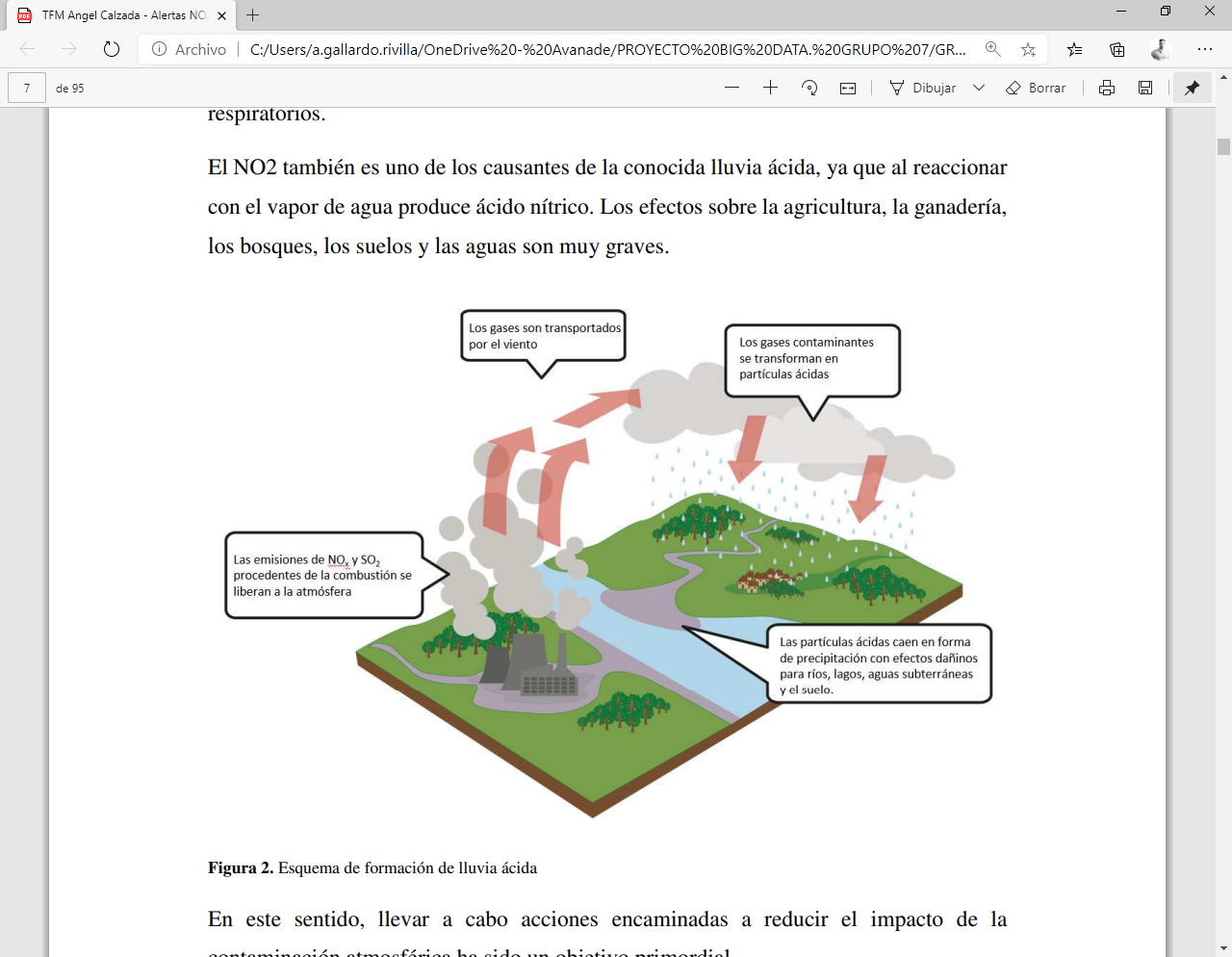


**Figura 1.** Evolución emisiones dióxidos de nitrógeno UE28. (Fuente: Agencia Europea del Medio Ambiente)

Además de contribuir en la formación de ozono, se relaciona al NO2 con efectos nocivos sobre el sistema respiratorio. Los dióxidos de nitrógeno reaccionan con el amoniaco, con la humedad y otros compuestos para formar pequeñas partículas. Estas pequeñas partículas pueden penetrar profundamente en las partes sensibles de los pulmones.

La evidencia científica vincula exposiciones cortas a NO2 (desde 30 minutos hasta 24 horas) con efectos adversos respiratorios, incluida la inflamación de las vías respiratorias en personas sanas y el aumento de los síntomas en personas que padecen de asma. Los estudios también muestran que existe conexión entre la exposición a corto plazo a este contaminante y el aumento de visitas a las emergencias hospitalarias por problemas respiratorios.

El NO2 también es uno de los causantes de la conocida lluvia ácida, ya que al reaccionar con el vapor de agua produce ácido nítrico. Los efectos sobre la agricultura, la ganadería, los bosques, los suelos y las aguas son muy graves.



**Figura 2.** Esquema de formación de lluvia ácida

En este sentido, llevar a cabo acciones encaminadas a reducir el impacto de la contaminación atmosférica ha sido un objetivo primordial.

Los distintos estados miembros de la UE han ido incorporando progresivamente a su legislación distintas medidas para alcanzar los objetivos planteados en la Directiva 2008/50/CE. Algunas de estas medidas consisten en protocolos de actuación en materia de tráfico ante episodios de alta contaminación ya que, como se muestra en la Figura 1, el tráfico por carretera es uno de los principales orígenes del NO2. Madrid, por ejemplo, ha puesto en marcha un protocolo de actuación en episodios de alta contaminación. Esta ha sido una de las motivaciones para realizar el presente trabajo: tratar de estudiar si, a través de la minería de datos podemos predecir si un día se dará o no un episodio de alta contaminación.

## 1.2 Adaptación a la Directiva 2008/50/CE en España

Centrándonos en España, el Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire, establece umbrales de alerta para algunos agentes contaminantes, entre ellos el dióxido de nitrógeno. Se define el umbral de alerta como “el nivel a partir del cual una exposición de breve duración supone un riesgo para la salud humana, que afecta al conjunto de la población y que requiere la adopción de medidas inmediatas.”

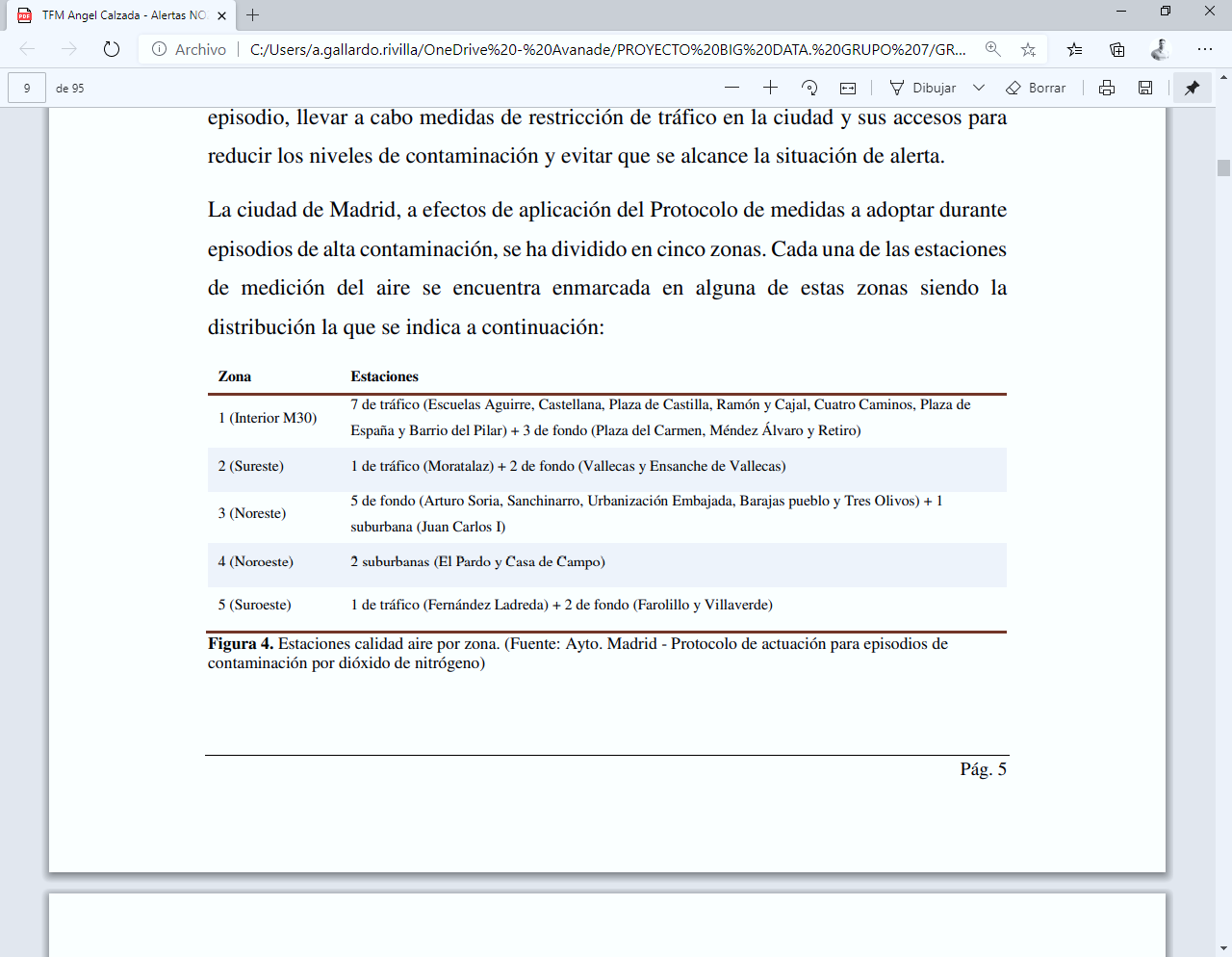
El valor del umbral de alerta para el dióxido de nitrógeno está establecido en 400 microgramos/m3 durante tres horas consecutivas en lugares representativos de la calidad del aire, en un área de al menos 100 km2 o en una zona o aglomeración entera, si esta última superficie es menor. El citado Real Decreto establece asimismo un valor límite horario para la protección de la salud de dióxido de nitrógeno de 200 microgramos/m3 (nivel de aviso) que no debe superarse más de 18 horas al año en ninguna de las estaciones de la red.

## 1.3 Protocolo de Contaminación en Madrid

El Ayuntamiento de Madrid, para llevar a cabo el control de la calidad del aire de la ciudad, dispone del Sistema de Vigilancia, Predicción e Información de la Calidad del Aire que permite conocer, de forma continua y en tiempo real, las concentraciones de contaminantes, con el principal objetivo de proteger la salud de la población y reducir al máximo las situaciones de riesgo.

Las elevadas concentraciones son originadas fundamentalmente por las emisiones del tráfico, y tienen lugar en situaciones con condiciones meteorológicas especialmente adversas, que requieren la ejecución de medidas para reducir los niveles de contaminación y la duración de los episodios, y evitar que llegue a superarse el valor límite horario y que se llegue a alcanzar el umbral de alerta.

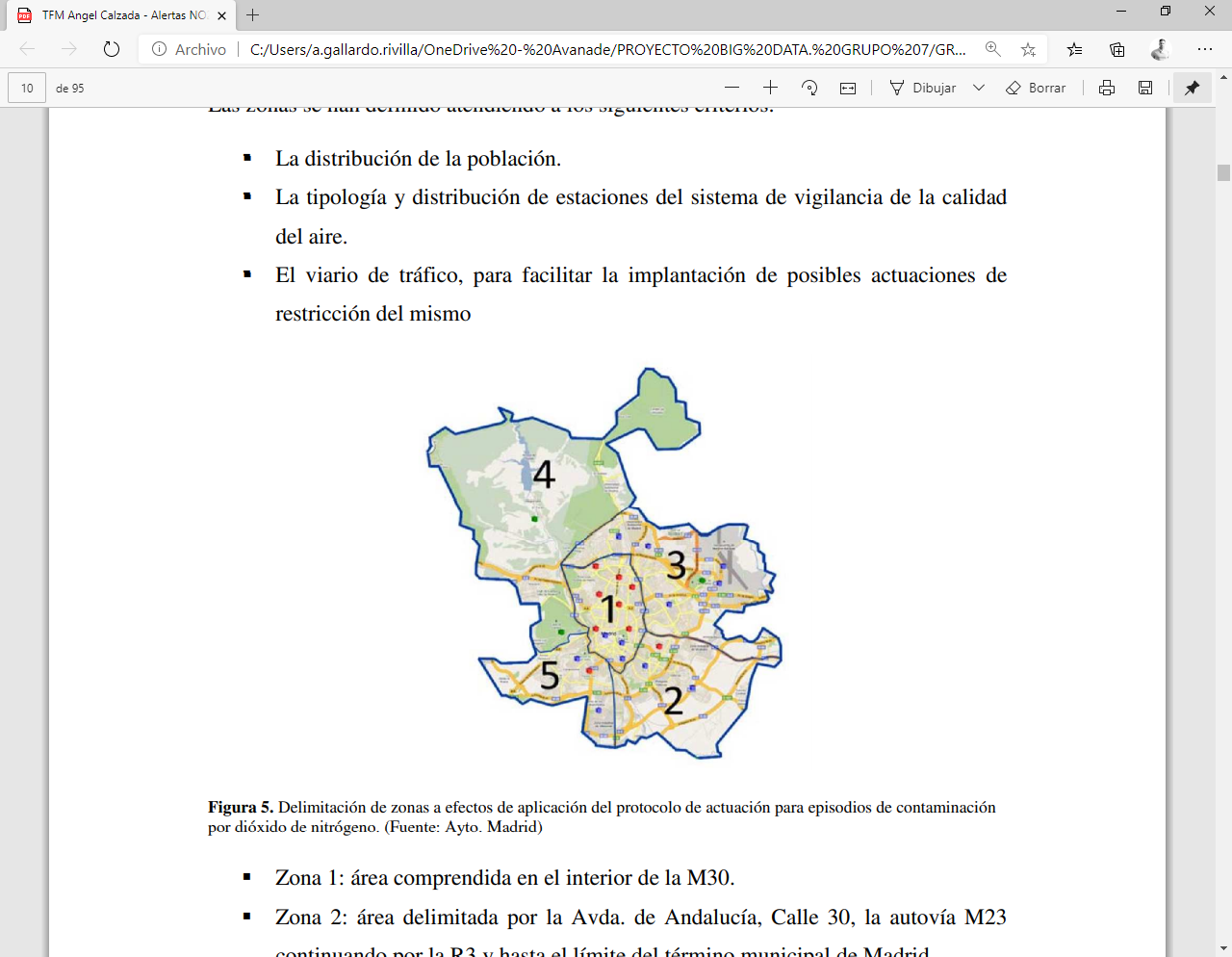
El Ayuntamiento de Madrid ha establecido una división en zonas de tal manera que las situaciones de alerta puedan declararse en áreas más reducidas con alta densidad de población. Igualmente se definen unos niveles de aviso que permitan, en el caso de registrarse concentraciones elevadas de dióxido de nitrógeno, la puesta en marcha de mecanismos de información adicionales, que sirvan tanto para proteger la salud de los ciudadanos como para sensibilizar a la opinión pública, recabar su colaboración para la reducción de la contaminación y, en función de los niveles alcanzados y la duración del episodio, llevar a cabo medidas de restricción de tráfico en la ciudad y sus accesos para reducir los niveles de contaminación y evitar que se alcance la situación de alerta. La ciudad de Madrid, a efectos de aplicación del Protocolo de medidas a adoptar durante episodios de alta contaminación, se ha dividido en cinco zonas. Cada una de las estaciones de medición del aire se encuentra enmarcada en alguna de estas zonas siendo la distribución la que se indica a continuación:



**Figura 3.** Estaciones calidad aire por zona. (Fuente: Ayto. Madrid - Protocolo de actuación para episodios de contaminación por dióxido de nitrógeno)

Las zonas se han definido atendiendo a los siguientes criterios:

* La distribución de la población.
* La tipología y distribución de estaciones del sistema de vigilancia de la calidad del aire.
* El viario de tráfico, para facilitar la implantación de posibles actuaciones de restricción de este.



**Figura 4.** Delimitación de zonas a efectos de aplicación del protocolo de actuación para episodios de contaminación por dióxido de nitrógeno. (Fuente: Ayto. Madrid)

* **Zona 1**: área comprendida en el interior de la M30.
* **Zona 2:** área delimitada por la Avda. de Andalucía, Calle 30, la autovía M23 continuando por la R3 y hasta el límite del término municipal de Madrid.
* **Zona 3:** área delimitada por la autovía M23 y la continuación de la R3, Calle 30 hasta la M40 en la zona oeste y desde allí limita al norte con la M40 hasta el límite del término municipal de Madrid. Esta zona incluye parte del Aeropuerto de Barajas.
* **Zona 4:** área delimitada por el contorno del límite del municipio de Madrid por el norte, la M40 norte, Calle 30 hasta la A5 y el límite del municipio.
* **Zona 5:** área delimitada por el contorno sur de la Casa de Campo, Calle 30, Avda. de Andalucía y el término municipal de Madrid. Se establecen tres niveles de actuación en función de las concentraciones de dióxido de nitrógeno que se registren en las zonas que se han definido:
* Nivel de preaviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 180 microgramos/m3 durante dos horas consecutivas.
* Nivel de aviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 200 microgramos/m3 durante dos horas consecutivas.
* Nivel de alerta: cuando en tres estaciones cualesquiera de una misma zona (o dos si se trata de la zona 4) se superan los 400 microgramos/m3 durante tres horas consecutivas.

# 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 meteorológica 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.

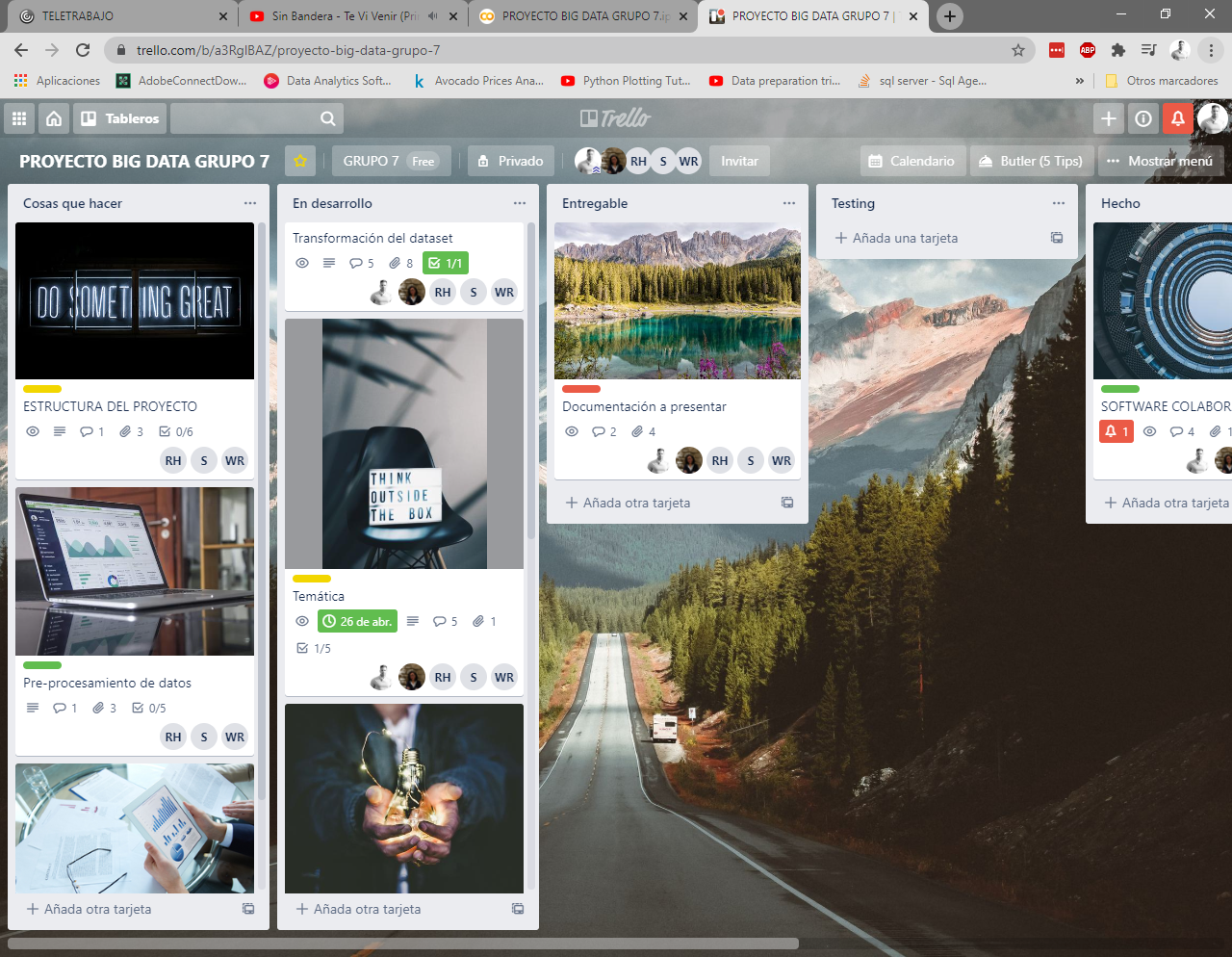
## 2.3. Limitaciones

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.

# 3.- Desarrollo

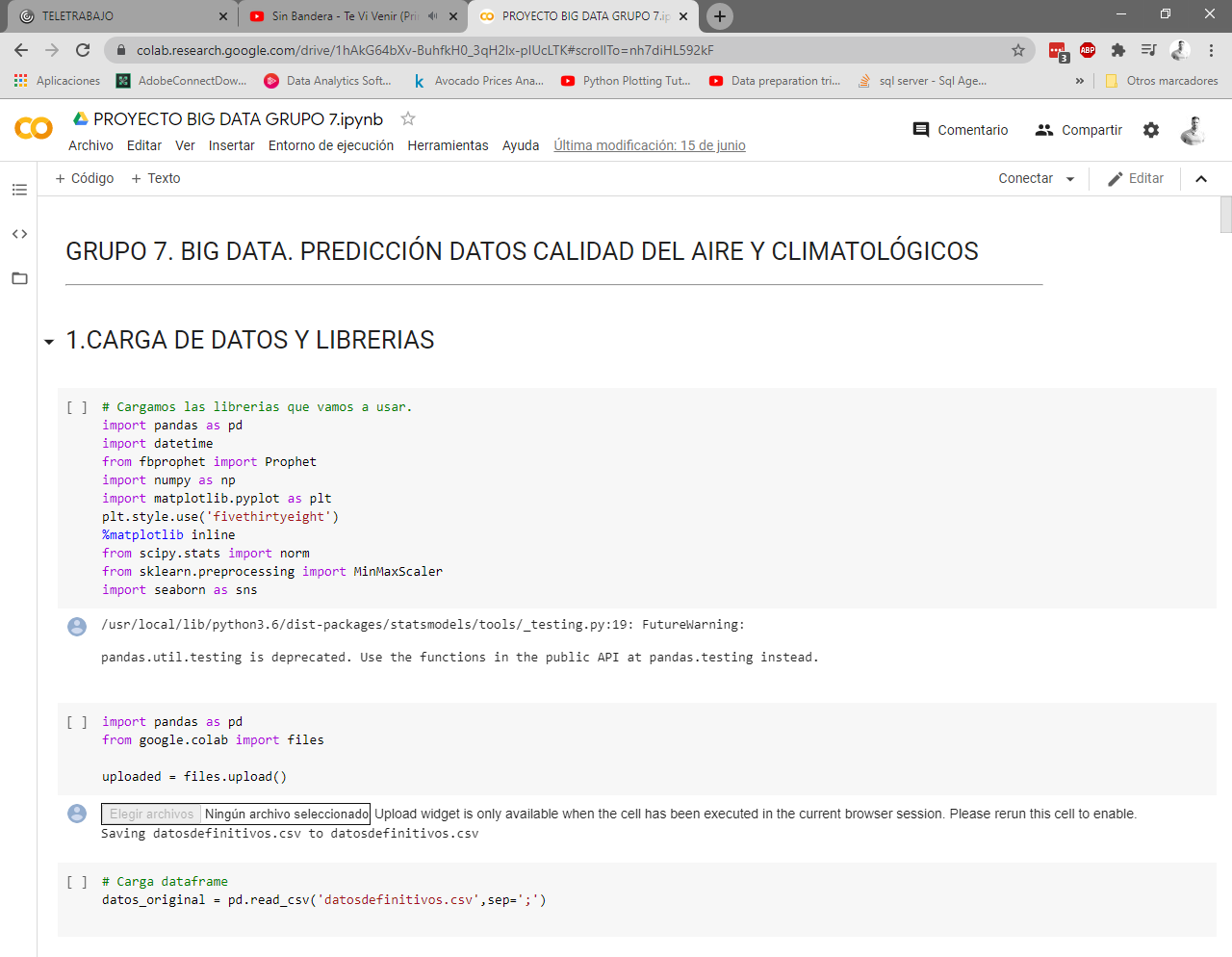
La coordinación dentro del equipo la organizamos en **Trello**, compartiendo ahí tanto la información de negocio que vamos recopilando como los scripts o notebook que vamos generando.



**Figura 5.** Entorno de trabajo en Trello

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:

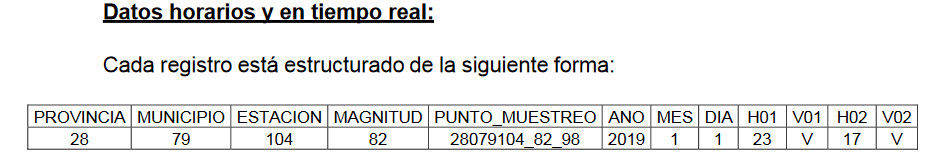


**Figura 6.** Desarrollando en Google Colab.

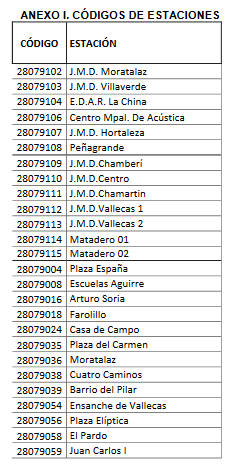
## 3.1. Exploración y elección de los datasets

Elegimos los datos del ayuntamiento y la comunidad de Madrid tanto de calidad del aire como meteorológicos. 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 meteorológicas.

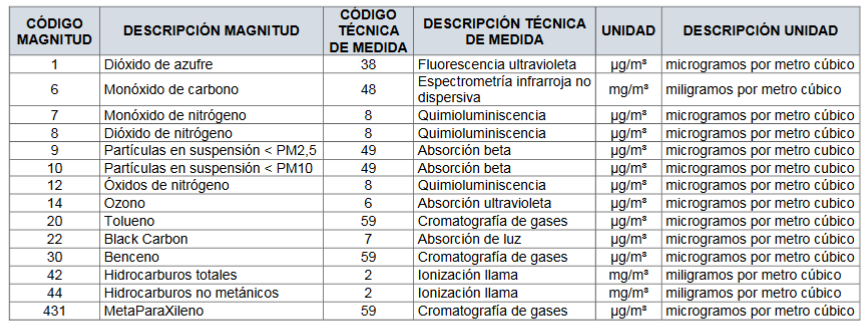


* Las estaciones dentro del municipio de Madrid y en el resto de los municipios de la comunidad de Madrid:

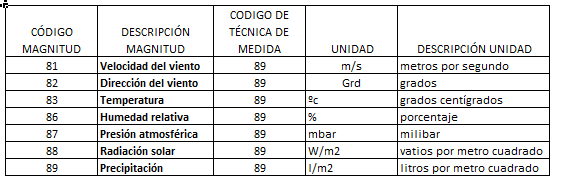




* Las magnitudes de calidad del aire y las técnicas de medida:

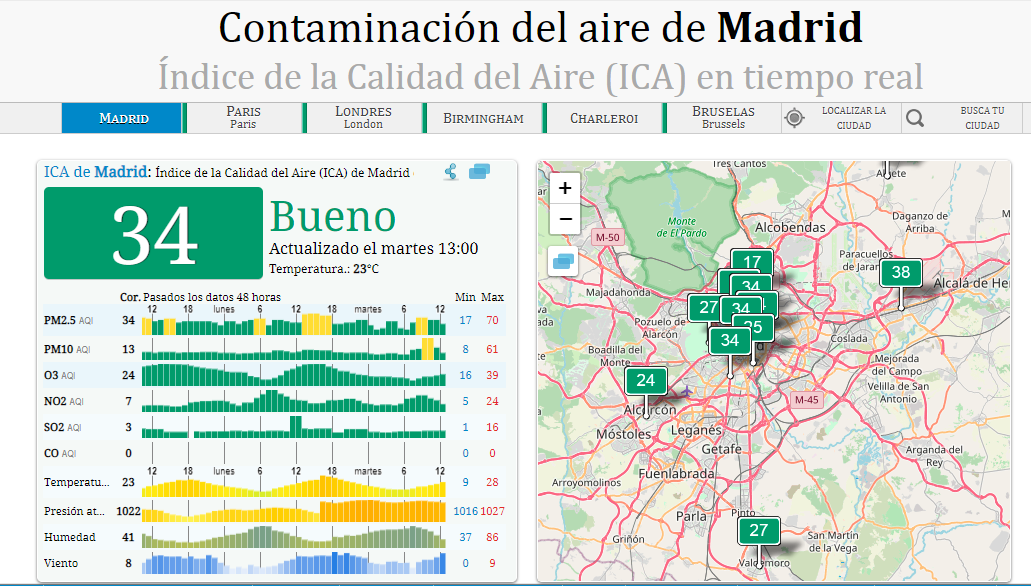


* Las magnitudes meteorológicas en las estaciones:

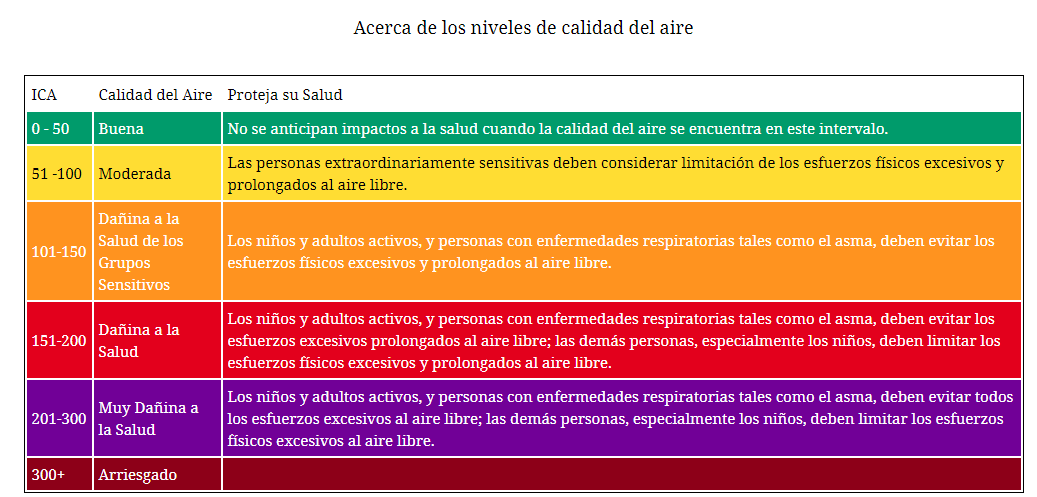


Con toda esta información comenzamos a investigar cómo se mide la calidad del aire, en el sentido de qué marca ese valor que nos comunican en distintos medios de información.

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.



**Figura 6.** Fuente: <http://aqicn.org/city/madrid/es/>



**Figura 7.** Niveles de calidad del aire.

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 valor límite 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 de los 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 datasets los datos de calidad del aire y por otro lo datos meteorológicos.

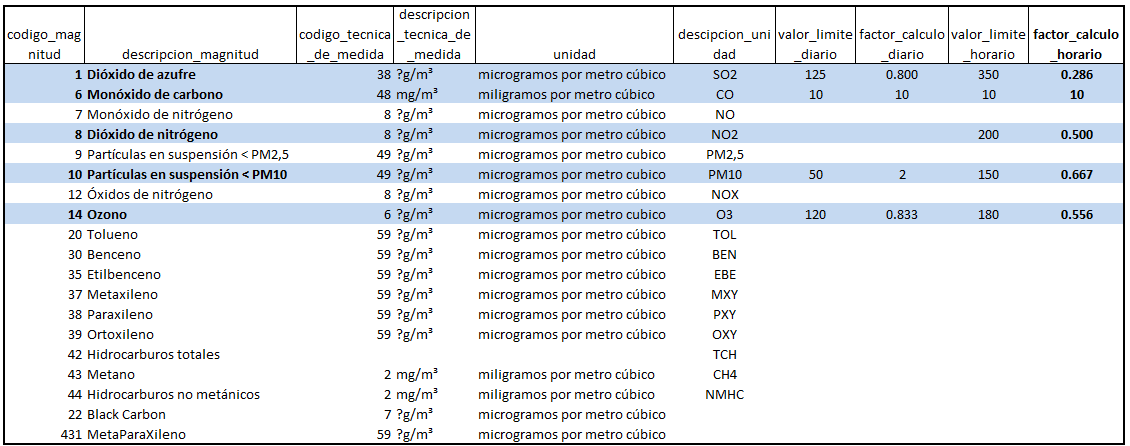
Creamos el campo fecha en un formato válido para el análisis de series temporales.

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



**Figura 8.** Ejemplo de código para el cálculo de los factores.

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



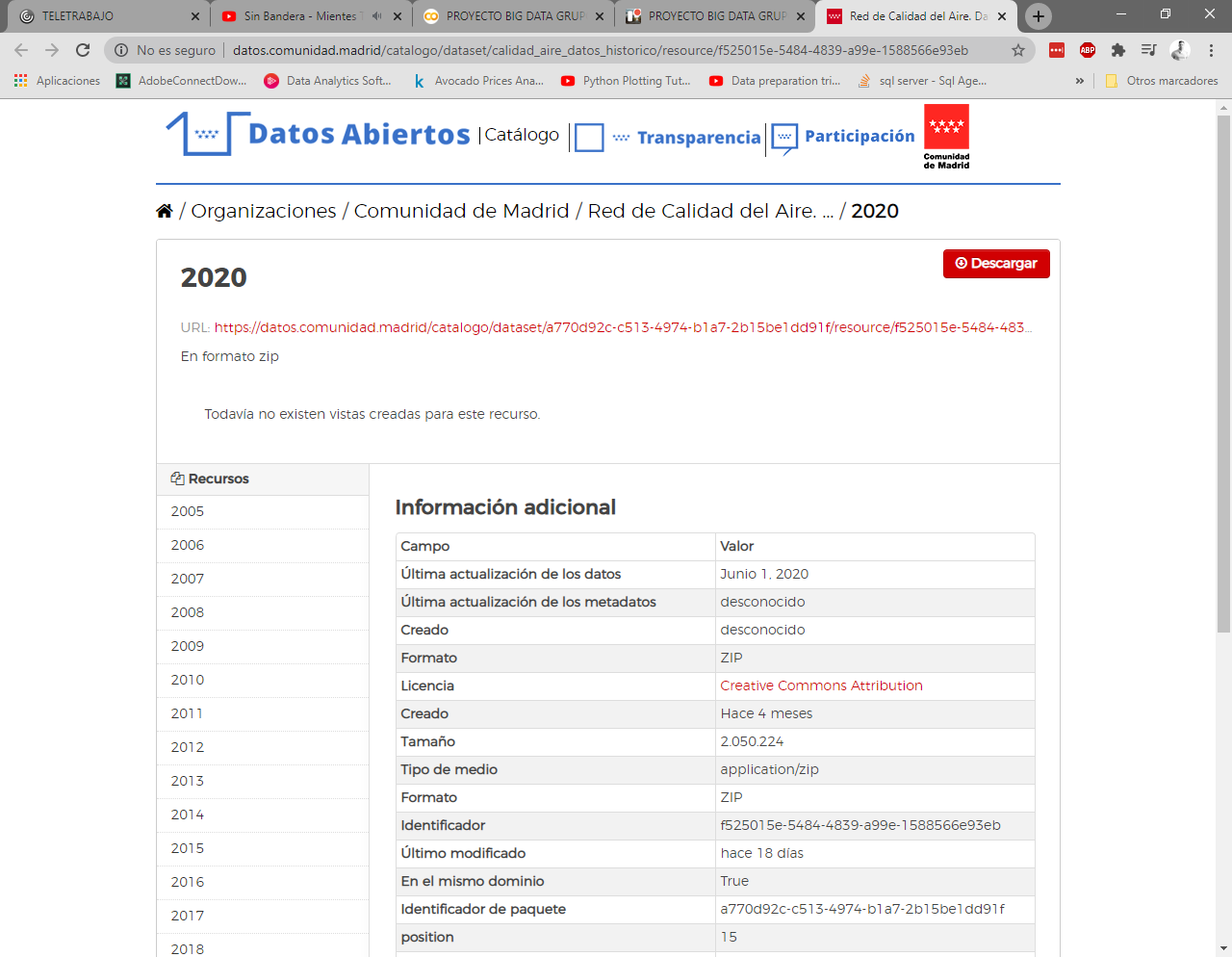
**Figura 9.** Contaminantes para el cálculo del ICA.

## 3.3. Carga de fuentes de datos

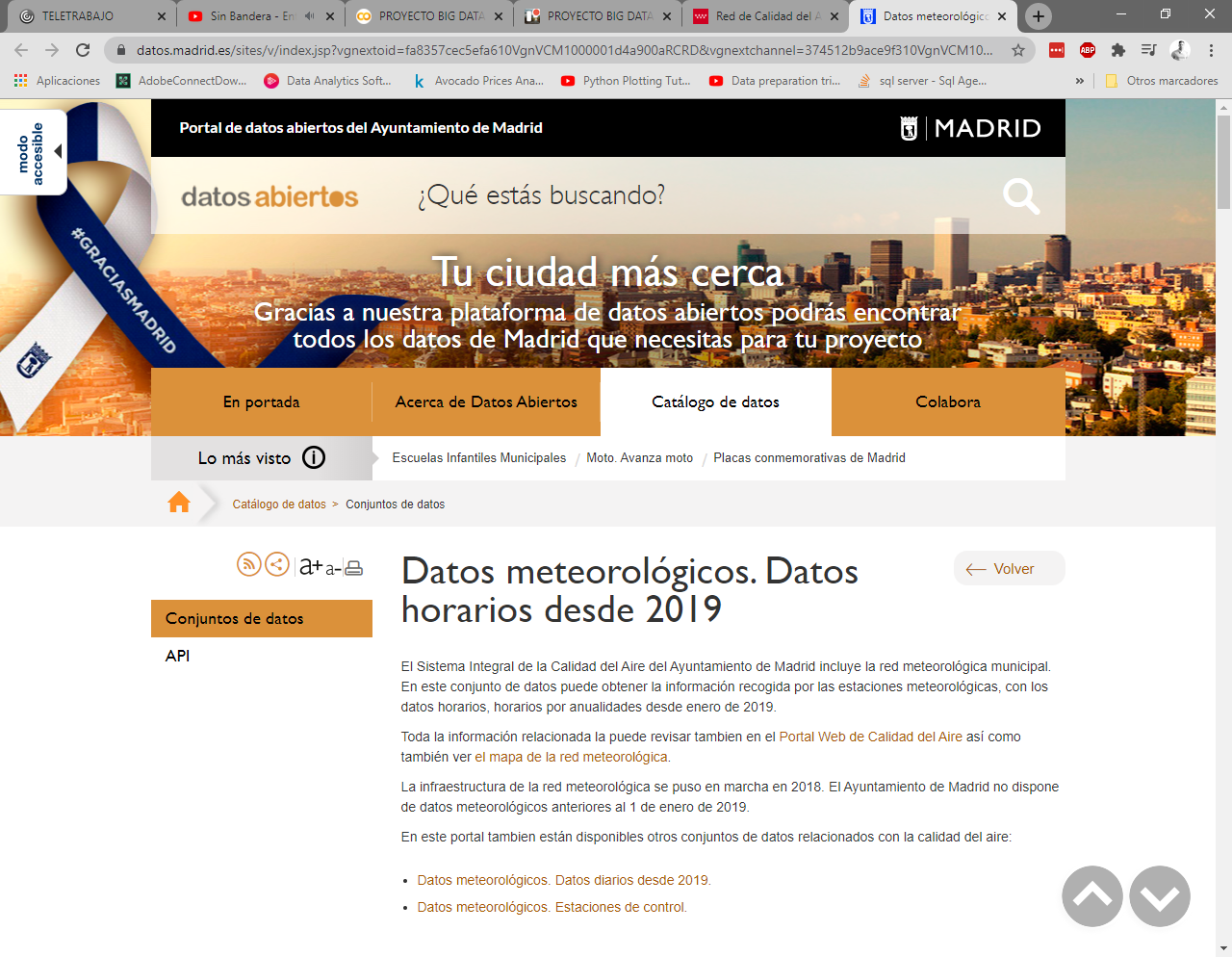
La carga inicial le hemos realizado bajando los csv de los portales de datos abiertos de la comunidad del Madrid y del ayuntamiento de Madrid, tanto para la calidad del aire como para la información meteorológica.



**Figura 10.** Histórico calidad del aire. Consultar [link](https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default).



**Figura 11.** Descripción de los campos del datasets. Consultar [link](http://datos.comunidad.madrid/catalogo/dataset/calidad_aire_datos_historico/resource/f525015e-5484-4839-a99e-1588566e93eb).



**Figura 12.** Datos meteorológicos. Consultar [link](https://datos.madrid.es/sites/v/index.jsp?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD).

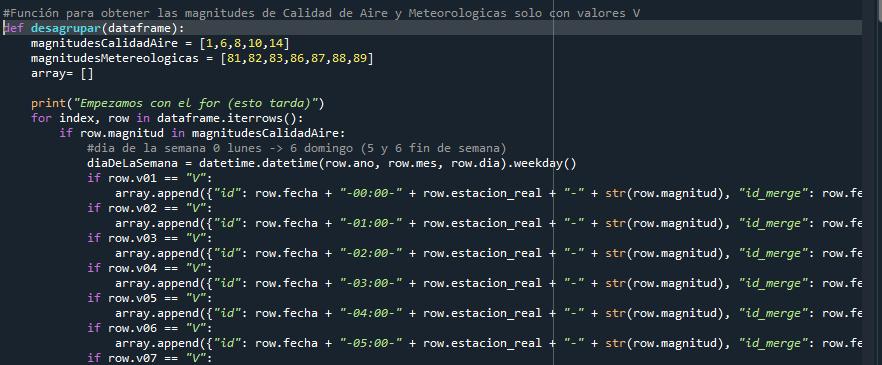
## 3.4. Comprensión y análisis descriptivo de los datos

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 datasets tendremos para cada día, hora, estación, magnitud de calidad del aire, su valor, su ICA, y los datos meteorológicos en ese momento.

Para esto creamos una función en Python para transformar el datasets. Convertimos el datasets 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})

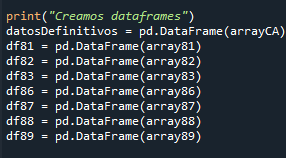


**Figura 13.** Ejemplo array horas.

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 meteorológicas.



**Figura 14.** Unión con las variables meteorológicas.



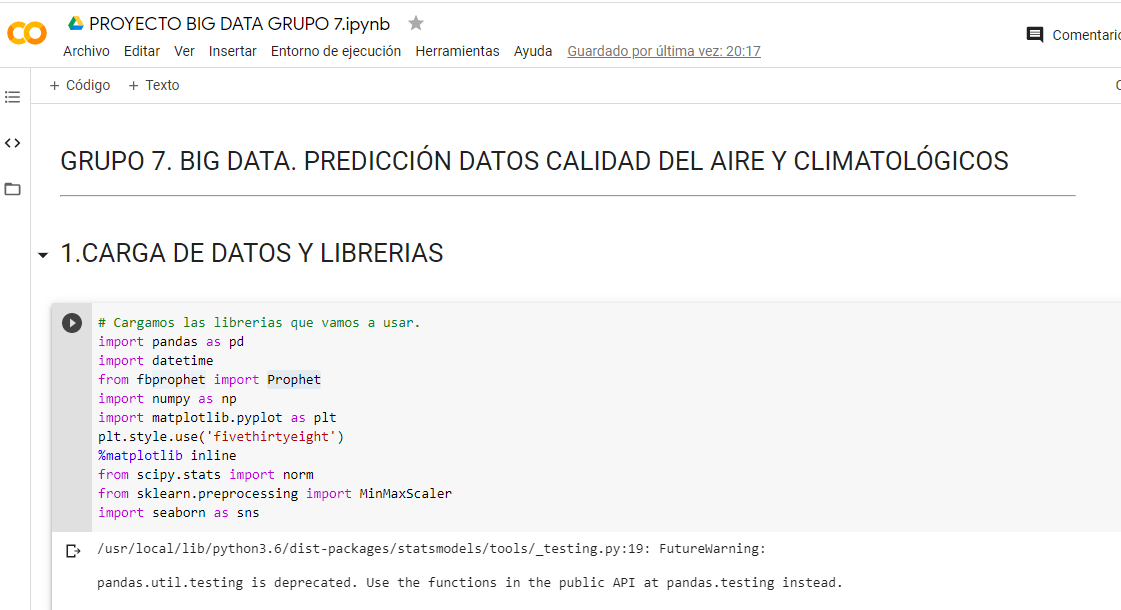
**Figura 15.** Creamos los dataframe.

Este datasets llamado **datosDefinitivos** es el que empezaremos a procesar.

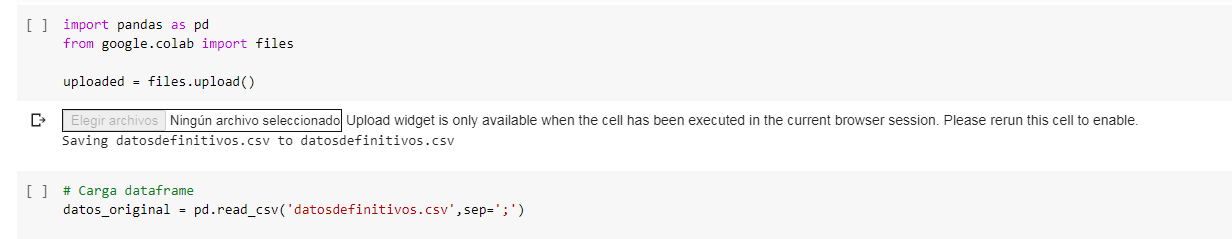
## 3.5. Pre-procesado de datos

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

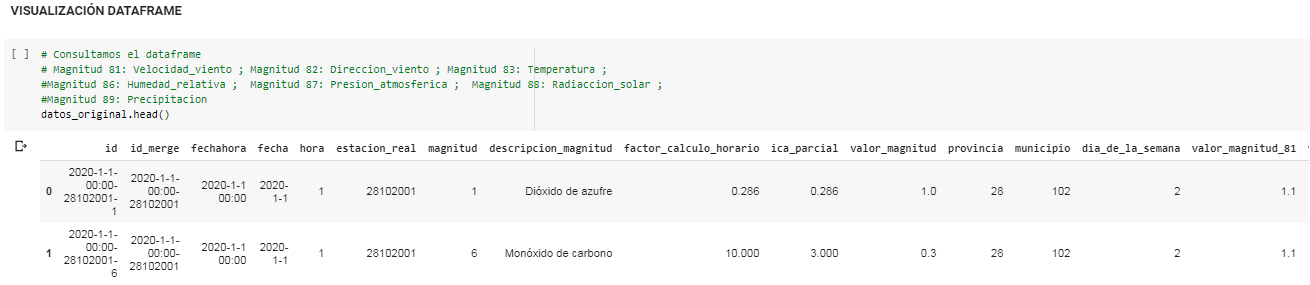
**Parte 1: Carga de datos:**



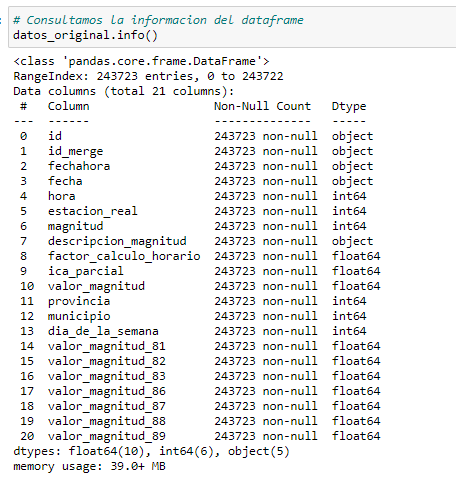
Importamos los datos con google.colab:

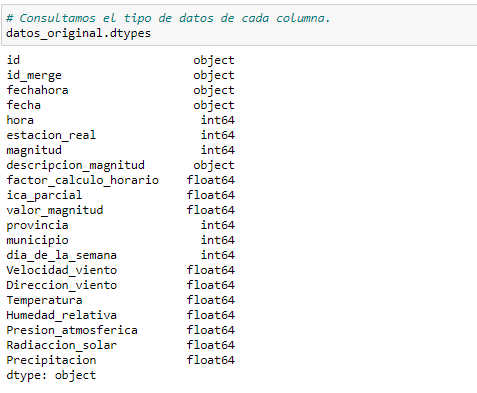


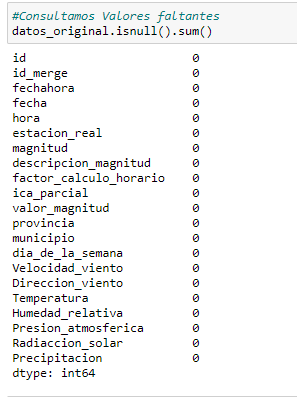
Visualizamos el dataframe :

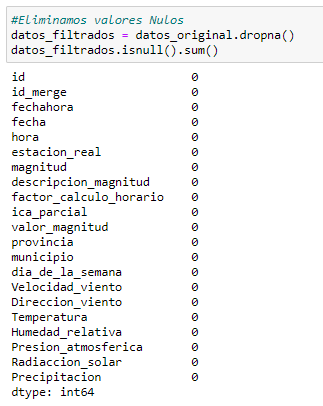


Consultamos la información del dataframe:

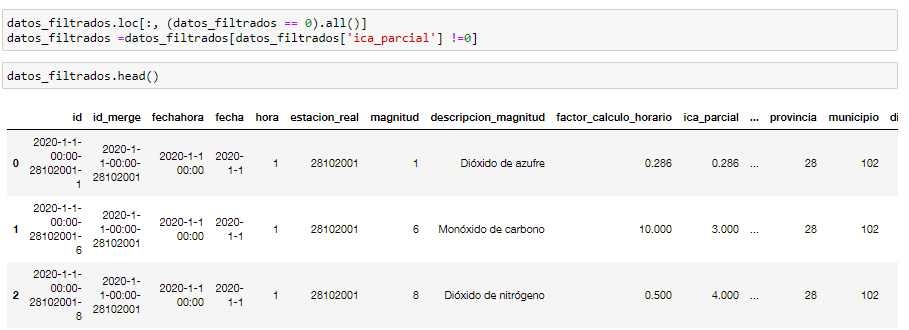


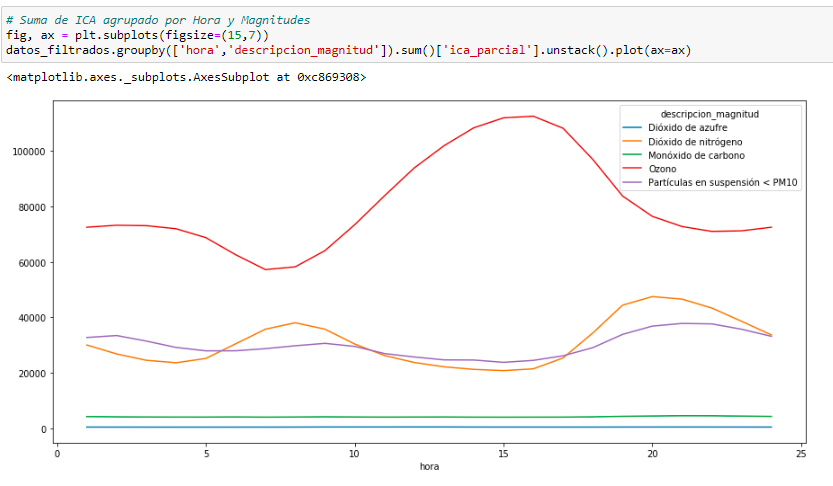


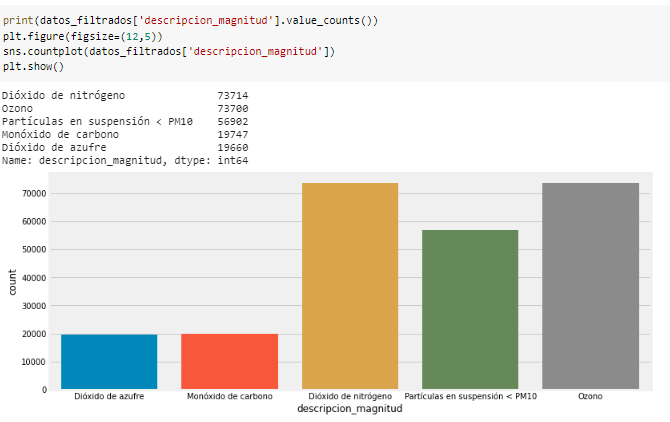


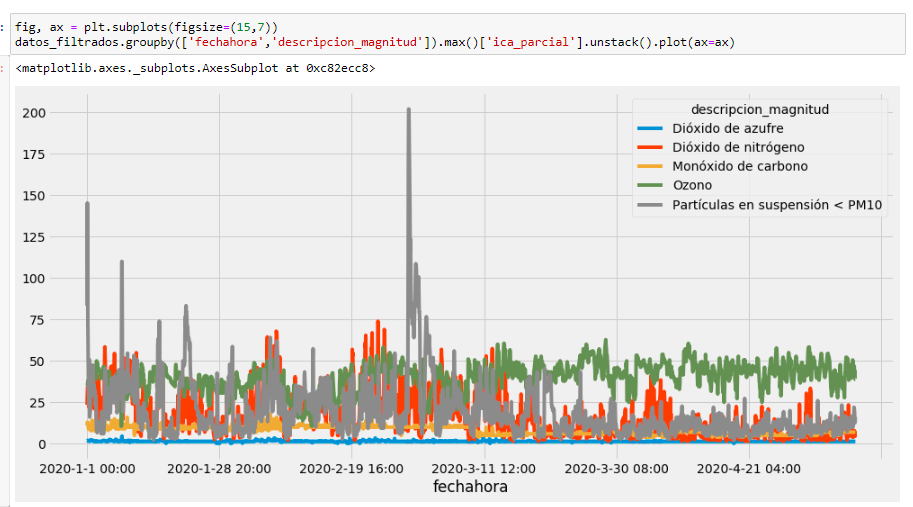


Hacemos un filtrado quedándonos sólo con los que tienen un valor ICA distinto a 0

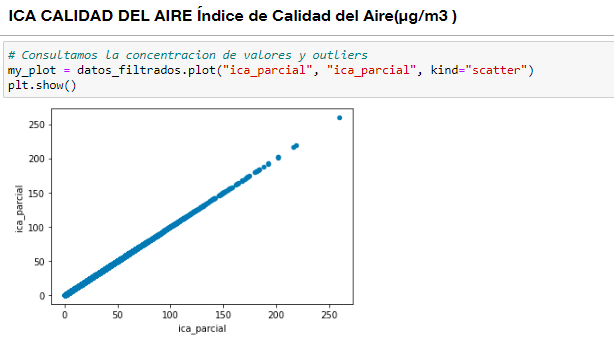


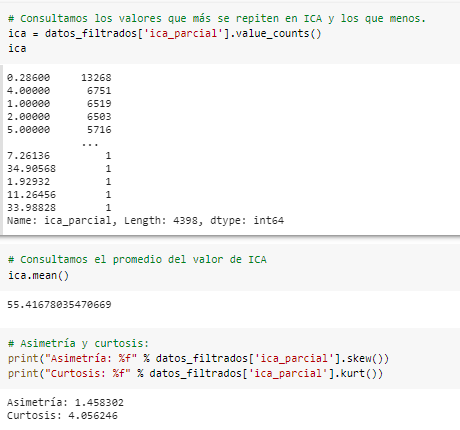


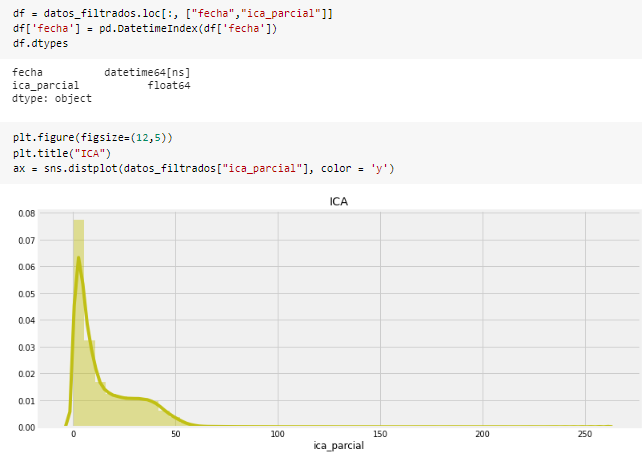


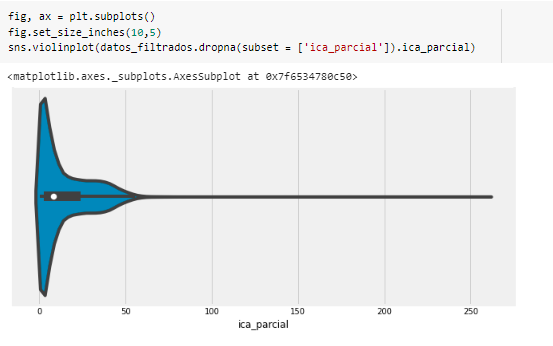


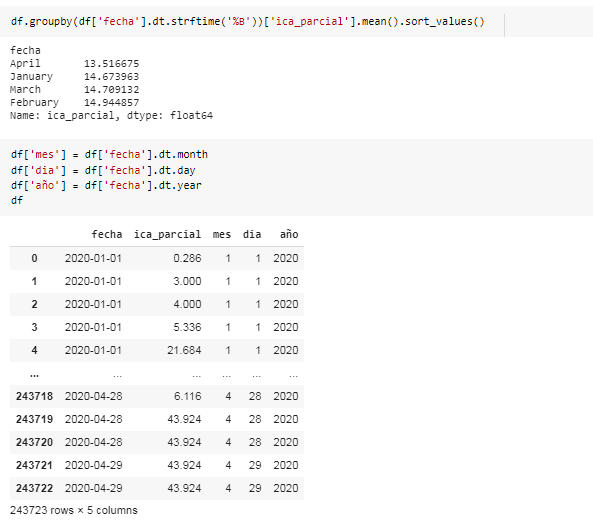
**Parte 2: Compresión de los datos:**

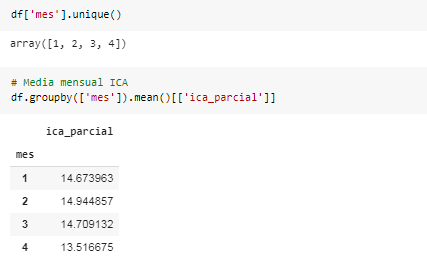


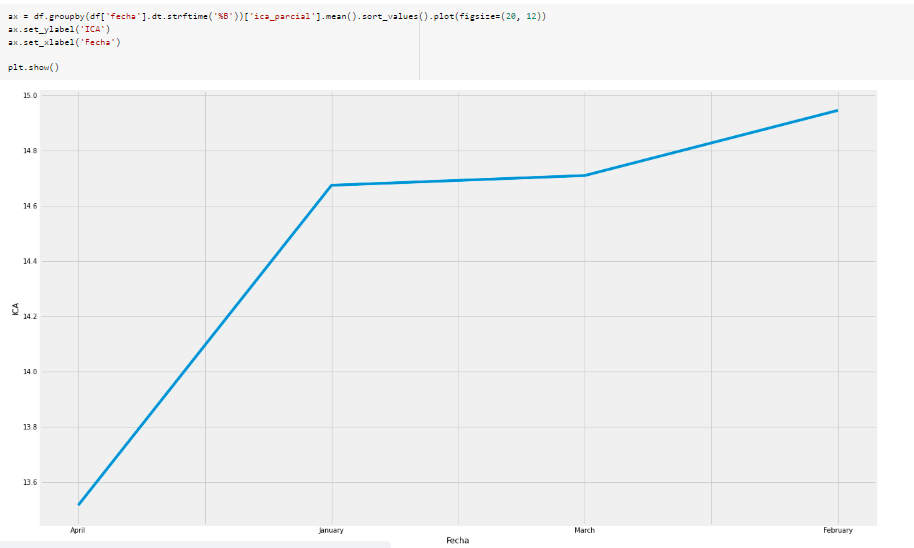


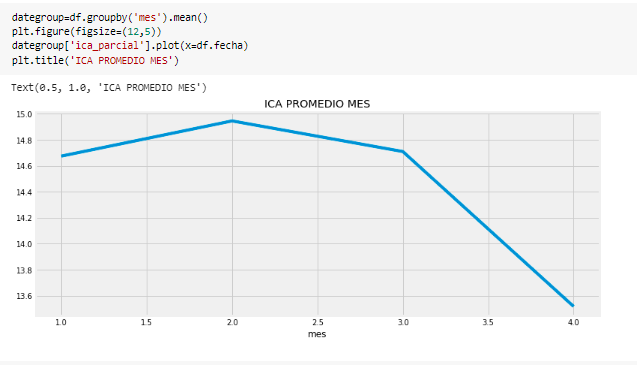


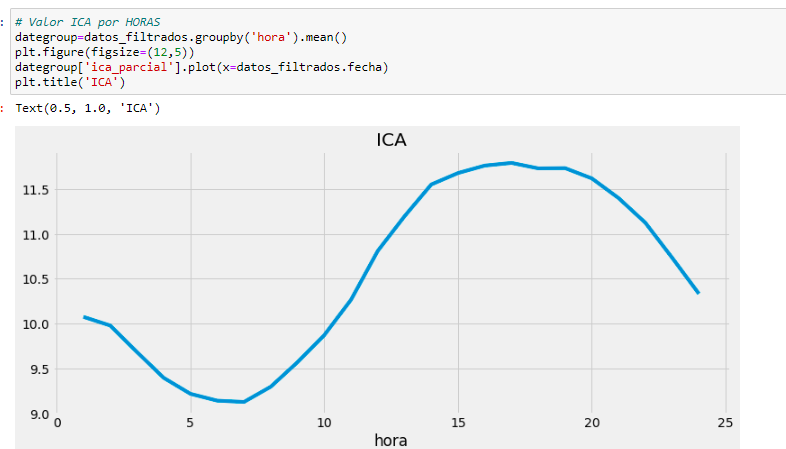






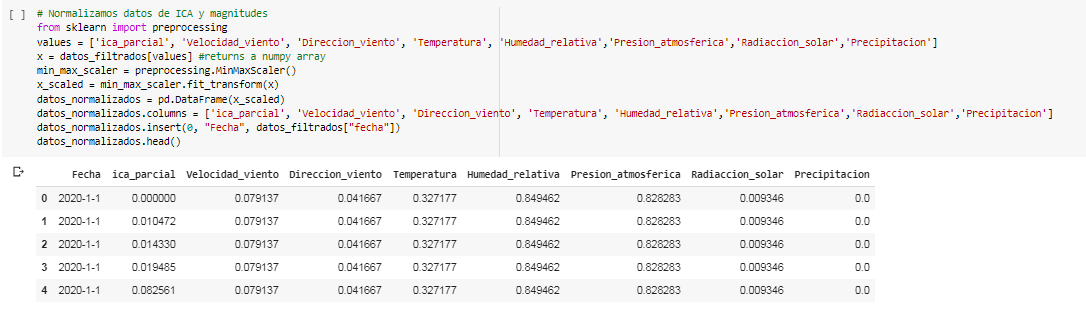


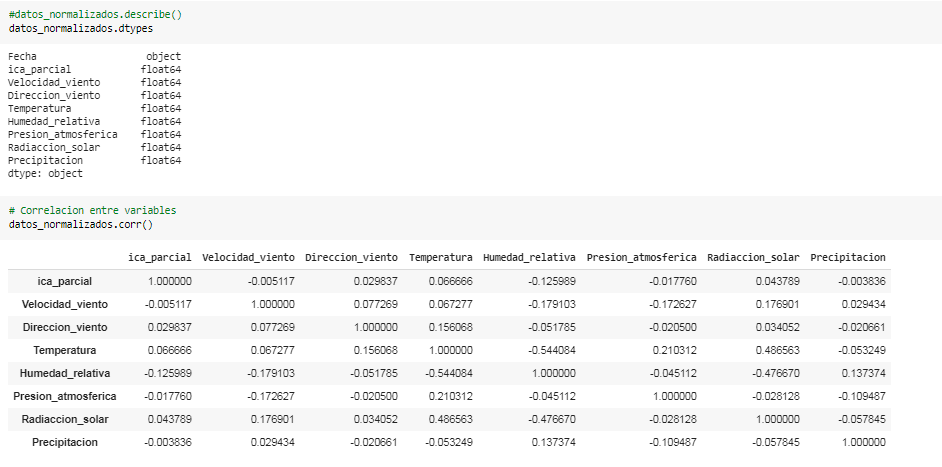


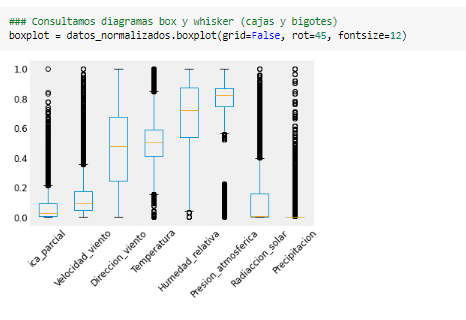


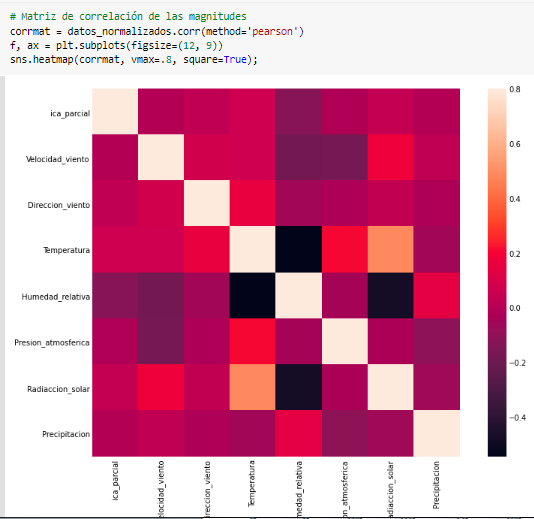
## 3.6. Preparación del modelo

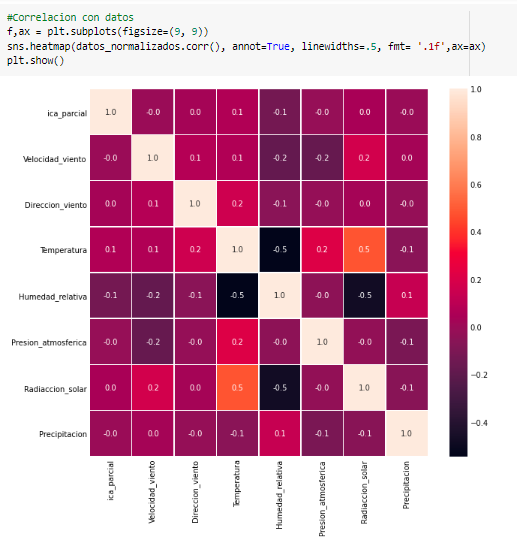
Para preparar los datos para el modelo comenzamos con una visualización de los datos para lo cual normalizando los datos :





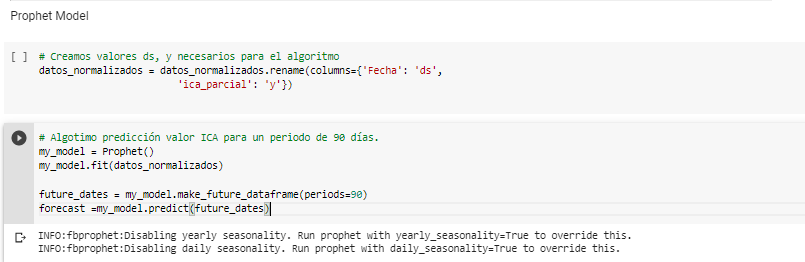


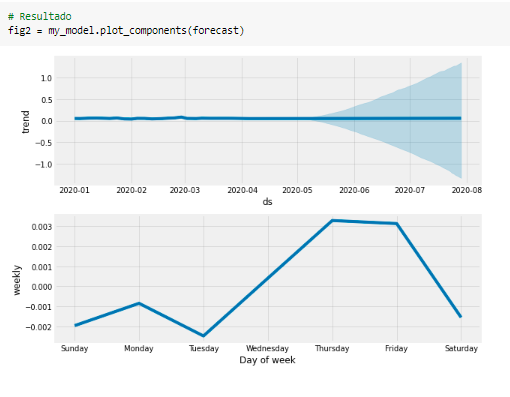




## 3.7. Evaluación de algoritmos

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