

# *Impacto de las partículas contaminantes en la temperatura en el Valle de Aburrá, Antioquia-Colombia.*

**ALUMNO 1:** Laura Cristina López Bedoya

**ALUMNO 2:** Andrés Fernando Morales González

**PROGRAMA:**

Postgrado en Data Science y Machine Learning

**NOMBRE DEL PROYECTO:**

*Impacto de las partículas contaminantes en la temperatura en el Valle de Aburrá, Antioquia-Colombia.*

## Contenido

<b>RESUMEN</b>	<b>4</b>
<b>INTRODUCCIÓN</b>	<b>5</b>
<b>ESTADO DEL ARTE</b>	<b>7</b>
<b>OBJETIVOS</b>	<b>12</b>
Objetivo general	12
Objetivos específicos	12
<b>SOLUCIÓN PLANTEADA</b>	<b>12</b>
Metodología y desarrollo de cada etapa	12
<i>Entendimiento del problema</i>	13
<i>Selección del conjunto de datos</i>	13
<i>Construcción del ambiente de desarrollo</i>	14
<i>Entendimiento de los datos</i>	15
Material particulado	15
Estaciones	18
Temperatura	19
Selección de partículas	20
Análisis descriptivo de los datos	23
<i>Proceso ELT</i>	25
Extracción	25
Carga	27
Transformación	27
<i>Análisis y construcción de los modelos</i>	30
Definición de datos de entrenamiento y prueba	31
Definición de los modelos a construir	32
<b>EVALUACIÓN</b>	<b>33</b>
Evaluación de los modelos con temperatura continua	33
Gráficas de resultado de prueba vs real, QQ y distribución de residuos	33
Árbol de decisiones	33
Random Forest	34
SVR	35

Regresión Lineal Multivariante	35
ANN	36
Evaluación con datos aleatorios	37
Evaluación con dato real	37
Evaluación de los modelos con temperatura categórica	38
Precisión	38
Pruebas aleatorias	38
<b>RESULTADOS</b>	<b>39</b>
<b>CONCLUSIONES Y TRABAJOS FUTUROS</b>	<b>41</b>
Conclusiones	41
Trabajos futuros	41
<b>REFERENCIAS</b>	<b>42</b>

## RESUMEN

El calentamiento global y la calidad del aire han sido motivo de preocupación en los últimos años. En los países y ciudades se han implementado medidas que ayuden a mitigar el daño ambiental y disminuir la contaminación del aire la cual afecta la calidad de vida de las personas y el medio ambiente. Actualmente en el Valle de Aburrá y Medellín (Antioquia-Colombia) se han detectado aumentos en la contaminación del aire debido al uso de medios de transporte y actividad industrial, por lo que se han tomado medidas que ayuden a disminuir el riesgo en la zona. Adicionalmente, se ha observado un cambio de temperatura en la ciudad, razón que nos lleva a preguntarnos si hay una relación directa entre las partículas contaminantes y la temperatura.

En este proyecto se presenta como solución un modelo de predicción de temperatura teniendo como datos de entrada las partículas contaminantes PM2.5, NO, NO2 y Ozono, elegidas luego de un análisis de datos donde se encontraron mayores registros para esta partículas y una correlación poco significativa ( $< 0.7$ ) con la temperatura. Se decide realizar la implementación de un modelo de machine learning para predicción dado que no hay una ecuación que relacione estas partículas contaminantes con la temperatura, y el uso de modelos de machine learning y redes neuronales nos permite predecir la variable objetivo con mayor facilidad y precisión.

Se realizó un análisis de datos y visualización para elegir los datos de entrada de los modelos, asimismo se usaron modelos supervisados como SVR, Árboles de decisión, Random Forest y Regresión lineal multivariante, así como Redes Neuronales, que nos permitieron predecir la temperatura de acuerdo a las medidas de las partículas contaminantes. Al realizar la implementación de los modelos en la evaluación se determina que el mejor modelo de predicción es Random Forest, el cual nos da una mayor precisión en la predicción de la temperatura con uso tanto de datos numéricos como datos categóricos.

## INTRODUCCIÓN

El Valle de Aburrá está conformado por Medellín y 9 municipios cercanos del departamento de Antioquia – Colombia, que son: Barbosa, Girardota, Copacabana, Bello, Envigado, Itagüí, Sabaneta, La Estrella y Caldas. Ésta área se ha caracterizado por ser un foco principal generador de empleo en el departamento, en esta se encuentran centros tecnológicos, fábricas de manufactura y universidades, entidades que cuentan con una gran aglomeración de personas, lo que convierte al Valle de Aburrá en una zona deseada por migrantes del departamento en búsqueda de oportunidades académicas y/o laborales.

El aumento de partículas contaminantes en las ciudades se debe especialmente al incremento de habitantes en los últimos años, como consecuencia se ha aumentado la presencia de medios de transporte y trabajos en fábricas, principales generadores de partículas contaminantes debido al consumo de combustibles fósiles. La calidad del aire puede traer consecuencias leves o graves a la salud, como problemas respiratorios y cardíacos, lo que afecta directamente la calidad de vida de las personas.

Adicionalmente, el crecimiento de las ciudades ha llevado a una notable reducción de espacios verdes y agrícolas, lo que contribuye a la expansión de partículas contaminantes en los espacios abiertos presentando un mayor riesgo a los habitantes. La Organización Mundial de la Salud (OMS) en 2018 publicó dos artículos importantes donde destaca que más del 90% de los niños a nivel mundial respiran aire tóxico cada día, además, mencionan que 9 de cada 10 personas en el mundo respiran aire contaminado. Los indicadores demuestran que cerca de 7 millones de personas mueren cada año por problemas respiratorios causados por la contaminación del aire. [1][18]

«La contaminación del aire representa una amenaza para todos, si bien las personas más pobres y marginadas se llevan la peor parte», dice el Dr. Tedros Adhanom Ghebreyesus, Director General de la OMS. «Es inadmisibles que más de 3000 millones de personas, en su mayoría mujeres y niños, sigan respirando todos los días el humo letal emitido por cocinas y combustibles contaminantes en sus hogares. Si no adoptamos medidas urgentes contra la contaminación del aire, el desarrollo sostenible será una simple quimera.» [1]

Como se ha mencionado las partículas contaminantes que se encuentran en el aire pueden causar graves problemas de salud: Las partículas como el dióxido de azufre (SO<sub>2</sub>) generan aumentos en mortalidad y morbilidad, afectando principalmente las funciones respiratorias. El dióxido de nitrógeno (NO<sub>2</sub>) afecta el sistema respiratorio y puede causar inflamaciones, afectar los pulmones e incrementar resistencia al paso de aire por las vías respiratorias. Los oxidantes como el ozono pueden causar irritación en los ojos y dificultad al respirar cuando se hace actividad física fuerte. El monóxido de carbono (CO) puede disminuir la presencia de oxígeno en la sangre y provocar efectos negativos en el sistema cardiovascular y nervioso. El plomo puede perjudicar las funciones del hígado y los riñones, incluso causar daños neurológicos.

La notable problemática de calidad del aire en el Valle de Aburrá reflejada en el aumento de personas registradas que consultan debido a problemas respiratorios y la contaminación en el ambiente (como nubes de humo) ha generado preocupación en las entidades gubernamentales y la salud pública. Debido a la creciente presencia de partículas contaminantes, se han tomado medidas como pico y placa para restringir la cantidad de

vehículos que transitan por la zona, buscando disminuir la generación de estas partículas y su distribución en el ambiente.

Con este proyecto se busca realizar un análisis estadístico de los datos para llegar a un modelo de machine learning predictivo que permita identificar si hay relación entre el material particulado en el aire del Valle de Aburrá (Antioquia - Colombia) y la temperatura, con el fin de tomar decisiones basadas en los datos y resultados de los modelos realizados para determinar medidas que permitan prevenir la exposición a ambientes con altos niveles de contaminación y altas temperaturas, y definir medidas de prevención y cuidado en la región.

Usaremos la fuente de datos abiertos del Valle de Aburrá e información entregada por SIATA (Sistema de Alertas Tempranas de Medellín y el Valle de Aburrá), con datos recolectados desde 2017/01 hasta 2020/06 con el fin de realizar análisis estadísticos, exploración de datos y crear modelos de Machine Learning con el fin de medir las partículas contaminantes como *dióxido de azufre (SO<sub>2</sub>)*, *dióxido de nitrógeno (NO<sub>2</sub>)*, *monóxido de carbono (NO)* y *ozono*, y su impacto en la temperatura.

Dado que no existe una ecuación que relacione las partículas contaminantes con la temperatura, como solución se plantea el uso de modelos de machine learning y redes neuronales, dado que no existe una relación directa entre las variables de entrada (partículas contaminantes) y la variable objetivo (temperatura). Esos modelos nos permitirán predecir con los valores de temperatura teniendo en cuenta la hora y los niveles presentados de las partículas contaminantes. Como métodos de evaluación de modelos se usará el error cuadrático medio (RMSE), coeficiente de determinación R<sup>2</sup>, Cross-validation para datos numéricos y accuracy para datos categóricos. Al finalizar, basados en estas medidas se elegirá cual es el mejor modelo.

Se usará como metodología base CRISP-DM, donde en una etapa inicial se analiza el problema planteado y los datos necesarios para su desarrollo, luego se pasa a una etapa de análisis descriptivo de los datos donde se define si los datos obtenidos son válidos y si la cantidad es adecuada para entrenar los modelos. Posterior a esto se realiza una limpieza de datos donde se reemplazan los datos no válidos por la media para no afectar las medidas estadísticas. Una vez finalizado el análisis y limpieza de datos se procede con la implementación de modelos de machine learning como SVR, Regresión lineal multivariante, árboles de decisión y random forest, así como redes neuronales. Los modelos mencionados se evaluarán con las métricas mencionadas anteriormente y se definirá cuál es el mejor modelo.

Al realizar la evaluación de los modelos con datos continuos se define como mejor modelo Random Forest, sin embargo, todos presentan valores muy bajos en las métricas usadas para para la evaluación de los resultados, por lo cual se decide realizar otra implementación de modelos con datos categóricos, donde se obtienen mejores resultados dado que los valores de la variable objetivo se reducen a tres categorías con rangos de temperatura, alto, medio y bajo. Con estos los resultados obtenidos presentan una mejoría y de nuevo el mejor modelo de predicción es el Random Forest.

En este artículo se presenta un resumen del problema abordado, la solución planteada, su justificación y evaluación del proceso realizado con un análisis de los resultados obtenidos. El objetivo principal es aplicar un modelo de Machine Learning que permita identificar la relación de las partículas contaminantes con la temperatura, tomando como población de estudio el

Valle de Aburrá en el departamento de Antioquia, Colombia. Al finalizar, se mencionan algunas conclusiones de los resultados obtenidos y trabajos futuros.

## ESTADO DEL ARTE

La reducción de la contaminación atmosférica contribuye a una mejora en la salud de las personas y por ende impacta su calidad de vida. En los últimos años la calidad del aire ha sido uno de los temas principales en el mundo, los niveles de contaminación han aumentado debido al crecimiento de las ciudades y se deben tomar medidas preventivas para evitar enfermedades respiratorias graves.

A continuación, se presentan algunos artículos y bibliografías de proyectos realizados que con el uso de Machine Learning y análisis de datos buscan dar solución a la problemática presentada de contaminación para hallar soluciones que permitan mitigar los riesgos.

### ***Air Quality Prediction: Big Data and Machine Learning Approaches (2018)***

Este artículo presenta una evaluación de la calidad del aire usando métodos de inteligencia artificial, árboles de decisión, redes neuronales, entre otros.

Con el avance de las infraestructuras de IoT, las tecnologías de big data y las técnicas de aprendizaje automático, el monitoreo y la evaluación de la calidad del aire en tiempo real es deseable para las futuras ciudades inteligentes. Este documento informa sobre el estudio de literatura reciente, revisa y compara el trabajo de investigación actual sobre la evaluación de la calidad del aire basado en análisis de big data, modelos y técnicas de aprendizaje automático. Por último, destaca algunas observaciones sobre problemas, desafíos y necesidades de investigación futura. [2]

### ***Air Quality prediction using Machine Learning ALgorithms (2019)***

Este artículo se centra en el uso de técnicas de Machine learning para predecir la concentración de SO<sub>2</sub> en el ambiente. Se utilizan series de tiempo para predecir la cantidad de SO<sub>2</sub> en los próximos meses.

Se realiza un análisis y visualización de datos donde se identifican ciudades con mayores niveles de contaminación y que requieren medidas urgentes, otras ciudades donde los niveles están aumentando y se deben tomar medidas preventivas. Se identificaron otras partículas que son importantes y pueden causar afectaciones mayores en la salud por lo que se sugiere tomar medidas para análisis futuros. [3]

### ***Applicability of machine learning in modeling of atmospheric particule pollution in Bangladesh (2020)***

En este artículo se realizan predicciones de concentración de partículas PM<sub>2.5</sub> y PM<sub>10</sub> con el fin de detectar niveles altos que puedan provocar efectos crónicos en la salud. En la investigación se usaron modelos de Machine learning como SVM y redes neuronales ANN y un modelo de series de tiempo llamado PROPHET. Se usaron

variables como SO<sub>2</sub>, CO, NO<sub>x</sub> y O<sub>3</sub>, que son partículas atmosféricas junto con variables meteorológicas. Los resultados mostraron que el algoritmo con mayor rendimiento fue GPR para ciertos lugares y para otros las redes neuronales obtuvieron resultados más acertados, finalmente se recomienda el uso de modelos combinados para obtener mejores resultados para la predicción de los niveles de concentración de partículas contaminantes en Bangladesh. [4]

***Forecasting Air Pollution Particulate Matter (PM<sub>2.5</sub>) using machine learning regression models (2020)***

En este documento se usan modelos de machine learning para predecir la concentración de partículas en el aire, se toma como conjunto de datos para monitorear la calidad del aire de Taiwán entre 2012 y 2017. Los resultados de los modelos fueron evaluados con medidas estadísticas como el error cuadrático medio (RMSE), error absoluto medio (MAE), error cuadrático medio (MSE) y el coeficiente de determinación(R<sup>2</sup>). [5]

***Air pollution and its health impacts in Malaysia: a review (2020)***

Malasia es un país en desarrollo con enfoque industrializado, se presentan neblinas y contaminación atmosférica debido al tránsito masivo de automóviles. La exposición a gases contaminantes como el ozono y demás partículas causan problemas de salud que se asocian con el aumento de hospitalizaciones y mortalidad.

En este artículo se discute el sistema de monitoreo de la calidad de aire en Malasia y se realiza una comparación con los estándares globales. Se discute acerca de la responsabilidad del gobierno, los impactos en la salud y las oportunidades de investigación con un enfoque de ingeniería, modelos de Machine learning y la falta de proyectos de investigación en este campo. [6]

***Análisis del estado de la calidad del aire en Bogotá (2007)***

Este documento se centra en la realización de un análisis de los registros contenidos en la Red de Monitoreo de la Calidad del Aire de Bogotá. Se construyó una base de datos donde se realizó un proceso de ETL y calidad de datos para ser usada en la evaluación de forma cuantitativa del estado de la calidad del aire de la ciudad. Los resultados sugieren que para contaminantes como óxidos de azufre y de nitrógeno, así como para monóxido de carbono, Bogotá no presenta en la actualidad un problema significativo de contaminación. Sin embargo, las concentraciones atmosféricas de material particulado en la ciudad tienden a estar en niveles elevados a comparación de los niveles sugeridos por las normas de calidad del aire.

En la ciudad no se presenta un problema de contaminación del aire para SO<sub>2</sub>, NO<sub>2</sub>, CO y O<sub>3</sub>. Para todos estos contaminantes las concentraciones registradas por la red de monitoreo de la ciudad suelen ser inferiores a los límites establecidos por la regulación



ambiental local, sin embargo, superan los límites establecidos en particular para zonas industriales. [7]

***Calidad del aire en el Valle de Aburrá Antioquia – Colombia (2009)***

En este artículo se evalúa la contaminación del aire y la presencia de partículas contaminantes en el Valle de Aburrá entre los años 2001 y 2007. Se realiza un análisis estadístico usando series de tiempo para graficar las partículas contaminantes, se logra identificar una variación de los niveles de partículas contaminantes durante el día, los niveles más altos se detectan en horas pico, que tienen gran flujo vehicular. Algunas de las partículas medidas presentan valores superiores a los niveles de precaución para la salud definidos por la OMS, por lo que se sugiere tomar medidas que ayuden a mitigar la propagación de estas partículas con el fin de reducir el riesgo de enfermedades respiratorias y cardiovasculares en el Valle de Aburrá. [8]

***Fortalecimiento de la Red de monitoreo de Calidad del Aire en el Valle de Aburrá con Medidores Pasivos (2008)***

Este documento se centra en el objetivo de fortalecer la monitorización de la calidad del aire del Valle de Aburrá mediante muestreos pasivos de partículas como dióxido de azufre, dióxido de nitrógeno y ozono. Como resultado no se obtuvo un valor superior al establecido por la Norma Colombiana Anual, sin embargo, en algunas estaciones se superó el nivel guía de la OMS para el dióxido de nitrógeno.

El uso de medidores pasivos permite identificar zonas de alto riesgo y evaluar las tendencias de contaminación atmosférica que permitan identificar focos de contaminación que pueden ser perjudiciales para la salud. Finalmente se propone la aprobación de esta técnica de medición en Colombia para tener información sobre la calidad del aire y tomar medidas preventivas a tiempo. [9]

***Big data aplicado al transporte y en las ciudades: adaptación a la ciudad de Sevilla (2018)***

En este documento se analizan los resultados obtenidos de la medición de partículas contaminantes en diferentes barrios de Sevilla, asimismo, se da una posible solución a los efectos causados por la emisión de gases contaminantes. Se recomienda la realización de campañas de conciencia que incentiven el uso de transporte público y la limitación del tránsito de vehículos en la ciudad, además teniendo en cuenta las afectaciones a la salud y a la integridad de la vida vegetal, se sugiere limitar la actividad industrial en diferentes zonas mitigando la emisión de gases mediante el uso de energías renovables. [10]

***Predicción de la calidad del aire de Madrid mediante modelos supervisados (2019)***

En este documento se presenta la aplicación de diferentes modelos de Machine learning para predecir la calidad del aire de Madrid, para ello se usan datos abiertos y se implementan modelos como SVM, redes neuronales convolucionales CNN, LSTM y MLP. A partir de los modelos usados se determina si es posible realizar una predicción con precisión de la calidad del aire y como prevenir con medidas niveles altos que se presenten a futuro.

Se realiza todo el proceso desde ETL, análisis de correlación, modelado y aplicación de modelos de Machine learning, al final se presenta un análisis de los resultados obtenidos al aplicar cada modelo. El mejor resultado se ha presentado al aplicar el modelo SVM, sin embargo, los demás modelos también presentan buenos resultados y el error obtenido es bajo. [11]

***Introducción al estudio de contaminantes atmosféricos mediante minería de datos (2016)***

En este trabajo se realiza una predicción de contaminación atmosférica mediante el uso de Weka, una herramienta de software libre elaborada por la universidad de Waikato. Se usan fuentes de datos obtenidas en bases de datos abiertas y mediante Weka se aplican modelos como regresión lineal con el fin de obtener el valor de un contaminante. Como resultado es necesario una cantidad de datos mayor para el estudio y con ellos realizar una predicción más precisa. [11]

***A systematic review of data mining and machine learning for air pollution epidemiology (2017)***

En el artículo científico se analizan 400 artículos de los cuales al realizar un filtrado se tomaron 47 separados en 3 áreas, fuentes de distribución, predicción de contaminación/calidad del aire y generación de hipótesis. Se aplicaron redes neuronales, árboles de decisión, SVM, algoritmo APRIORI y clasificación con k-means.

La mayoría de las investigaciones fueron realizadas en Europa, China y Estados Unidos, donde se observa que la aplicación de modelos de Machine Learning está tomando fuerza en análisis de salud ambiental. Finalmente, recomiendan el uso de redes neuronales para aplicaciones de epidemiología a causa de la contaminación del aire, teniendo en cuenta que este modelo ha presentado buenos resultados y se ve un buen potencial para futuras investigaciones, además, resaltan la importancia de la calidad de los datos recolectados que se usarán en los modelos de predicción. [12]

***A Deep learning model for air quality prediction in Smart cities (2017)***

En este artículo se resalta la importancia de IoT (Internet of Things) en temas de investigación de diferentes áreas. Las ciudades inteligentes utilizan este tipo de dispositivos para crear vida urbana sostenible. Con el uso de dispositivos de IoT la cantidad de datos ha ido en aumento, los cuales son utilizados por organizaciones

gubernamentales y partes interesadas para realizar predicciones y procesamiento de datos con el fin de garantizar un desarrollo sostenible en las ciudades.

En la predicción se han usado técnicas de aprendizaje profundo para tratar problemas de predicción con big data. En el artículo proponen un modelo basado en redes neuronales LSTM (Long Short Term Memory) para predecir valores de la calidad del aire en una ciudad inteligente, los resultados obtenidos son satisfactorios y pueden ser usados en otras áreas de ciudades inteligentes. [13]

#### ***Deep learning architecture for air quality predictions (2016)***

En este artículo científico parten de la premisa que los modelos actuales de predicción de la calidad del aire son poco profundos, sin embargo, estos métodos arrojan resultados insatisfactorios, lo que los llevó a investigar métodos para predecir la calidad del aire basados en modelos de arquitectura profunda.

Proponen un nuevo método de aprendizaje profundo espacio-temporal (STDL). Se utiliza un modelo codificador automático apilado (SAE) para extraer las características inherentes de la calidad del aire y se entrena el modelo por capas. Este modelo puede predecir la calidad del aire de todas las estaciones en simultáneo y muestra estabilidad temporal en todas las estaciones. Este modelo presenta mejores resultados que otros como por ejemplo el SVR (Support Vector Regression). [14]

#### ***Red neuronal Backpropagation para la predicción de datos de contaminación y prevención de ataques a personas con padecimientos de rinitis alérgica y asma (2017)***

Este artículo trata principalmente de la prevención de enfermedades respiratorias y alérgicas mediante la implementación de modelos de predicción como una red neuronal backpropagation. Los datos se obtienen de sensores distribuidos en varios lugares de la ciudad donde se miden cinco partículas principales. Al entrenar y evaluar el modelo, se verifica en qué punto el error es mínimo, con esto el modelo aplicado tendrá mayor precisión de predicción y será aplicado a una población más grande.

Al final se concluye que el modelo aplicado presenta buenos resultados, cercanos a los datos reales recolectados. [15]

Al analizar los artículos encontrados logramos identificar cuales son los modelos de Machine learning más usados para la predicción de calidad del aire en las ciudades. Los dos principales son SVM y redes neuronales, ambos presentaron resultados prometedores en las investigaciones realizadas. Asimismo, se concluye que los modelos de Machine learning son la mejor alternativa para predicción de calidad del aire y partículas contaminantes, estos presentan resultados acertados y en poco tiempo, lo que permite tomar medidas de precaución en menor tiempo.

## OBJETIVOS

### Objetivo general

Aplicar un modelo de machine learning que permita identificar la relación de las partículas contaminantes con la temperatura en el área metropolitana del Valle de Aburrá, Antioquia - Colombia.

### Objetivos específicos

- Explorar los datos obtenidos de la página web y el API REST de SIATA (Sistema de Alertas Tempranas de Medellín y el Valle de Aburrá).
- Analizar y preparar los datos para usarlos en un modelo de Machine learning.
- Explorar modelos de Machine Learning para identificar el impacto de las partículas contaminantes en la temperatura.
- Analizar la relación de los datos de partículas contaminantes y temperatura.
- Analizar los resultados obtenidos y concluir según lo identificado.

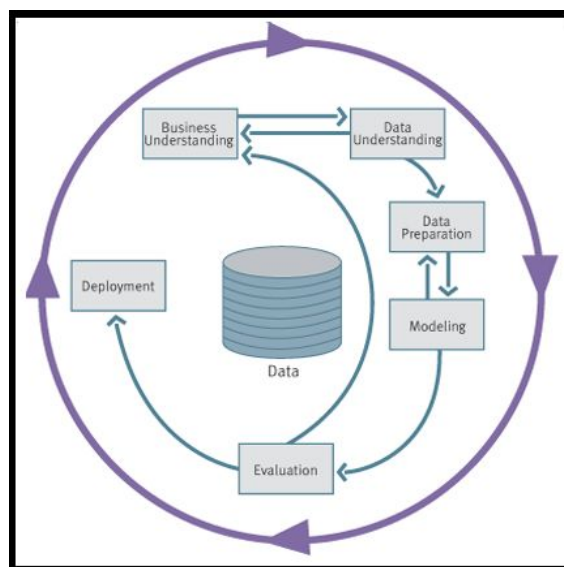
## SOLUCIÓN PLANTEADA

### Metodología y desarrollo de cada etapa

Nuestra metodología fue la siguiente:

1. Entendimiento del problema.
2. Selección del conjunto de datos.
3. Construcción del ambiente de desarrollo.
4. Entendimiento de los datos.
5. Proceso ELT.
6. Análisis y construcción de los modelos.

Nos basamos en la metodología CRISP DM, la cual consta de primero entender el negocio para luego entender los datos, preparar los datos, crear los modelos y evaluarlos.



Img 1. Modelo CRISP DM

## 1. Entendimiento del problema

A nivel global se viene sufriendo los efectos del cambio climático y el calentamiento global. En Medellín y el Valle de Aburrá (Antioquia, Colombia) se vive en una constante alerta sobre la calidad del aire y cómo ésta impacta en la salud de las personas. Debido a que la calidad del aire ya es monitoreada por el SIATA (Sistemas de Alertas Tempranas del Valle de Aburrá y Medellín) decidimos analizar el impacto que tienen las mismas partículas contaminantes del aire en la temperatura, puesto que el principal causante del calentamiento global es la actividad humana y la gran mayoría de las partículas contaminantes en el aire son generadas por actividad industrial y uso de los medios de transporte.

Con lo anterior, nuestro objetivo es poder encontrar una relación entre las partículas contaminantes en el aire y la temperatura mediante un modelo de predicción supervisado.

## 2. Selección del conjunto de datos

Teniendo en cuenta nuestro objetivo y luego de haber definido la ubicación geográfica: Medellín, Colombia, realizamos búsquedas de datos abiertos que nos brindaran la información necesaria para el desarrollo del problema. En un principio se encontraron datos interesantes en la página de datos abiertos del área metropolitana [16], dentro de ellos Control de Emisiones en fuentes móviles, sin embargo, los datos eran muy pocos y solo se encuentran registros del 2017 lo que brinda información poco relevante para el análisis.

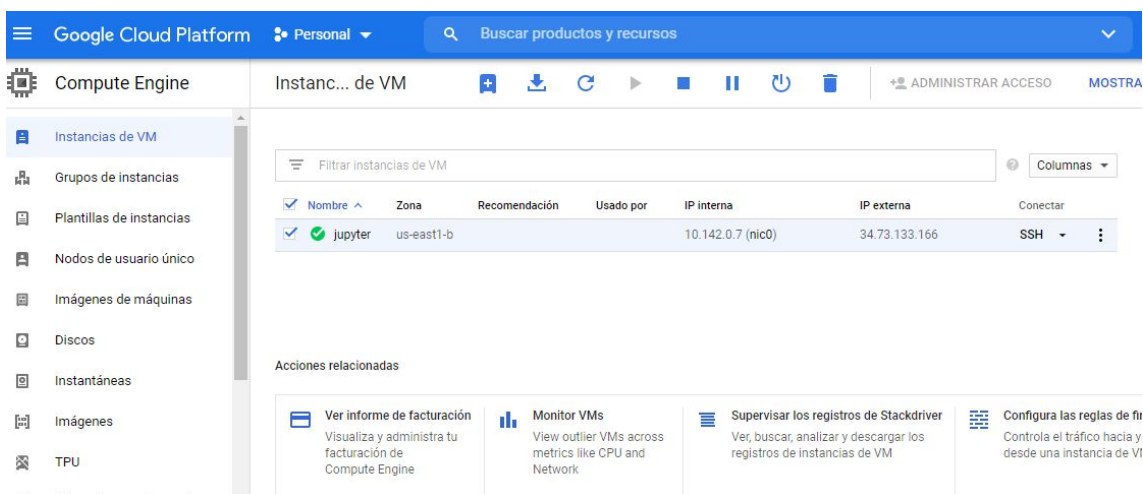
Dentro de la página mencionada se encontraron datos del SIATA. Ya con el conocimiento de que el SIATA recolectaba todos estos datos se decide contactar con la entidad y solicitar la documentación necesaria para extraer los datos.

El SIATA [17] brinda información: meteorológica, hidrológica, partículas contaminantes (calidad del aire), entre otras con la posibilidad de descargar el historial desde el 2017, esto significa: Volumen, Variedad y Veracidad (3 Vs para Big Data), razón por la cual se decide trabajar con estos datos.

## 3. Construcción del ambiente de desarrollo

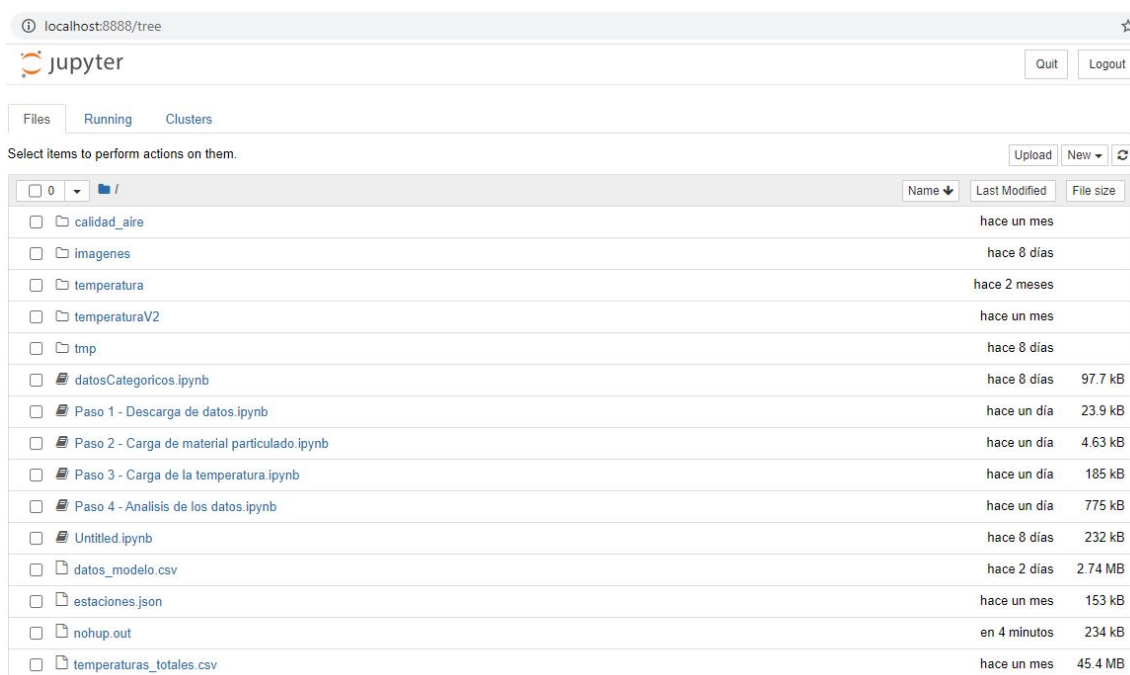
Para realizar este trabajo se eligió como lenguaje de programación Python específicamente Notebooks en Jupyter, base de datos MySQL y Data Studio. **Se usó Google Cloud Platform para crear una máquina virtual donde fue instalado Jupyter y SQL Cloud para la base de datos MySQL.**

Se decidió usar GCP porque se tenía el conocimiento previo visto en el postgrado y porque se contaba con la prueba gratuita.



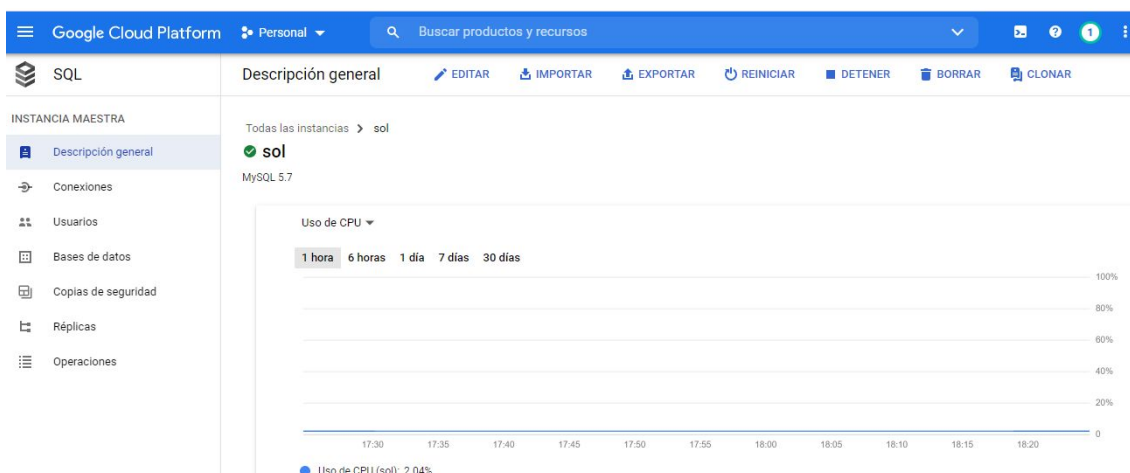
Img 2. GCP.

Se presentaron dificultades al momento de conectarse con Jupyter porque este no permite una conexión directa. Se debe de usar **Tunneling**, por esta razón la dirección en el navegador es localhost:



Img 3. Jupyter.

Esta es la base de datos:



Img 4. SQL Cloud.

## 4. Entendimiento de los datos

### Material particulado

El material particulado se encuentra en formato CSV (archivo separado por comas) con los siguientes campos:

Campo	Descripción	Formato/Medida
Fecha_Hora	Fecha	YYYY-MM-DD HH:mm:ss
codigoSerial	Código de la estación que tomó la medición	número
pm25	Material particulado menor a 2.5 micras	µg/m3
calidad_pm25	Calidad del dato pm25	[1,4]
pm10	Material particulado menor a 10 micras	µg/m3
calidad_pm10	Calidad del dato pm10	[1,4]
pm1	Material particulado menor a 1 micra	µg/m3

calidad_pm1	Calidad del dato pm1	[1,4]
no	Monoxido de Nitrogeno	ppb
calidad_no	Calidad del dato no	[1,4]
no2	Dioxido de Nitrogeno	ppb
calidad_no2	Calidad del dato no2	[1,4]
nox	Oxidos de Nitrogeno	ppb
calidad_nox	Calidad del dato nox	[1,4]
ozono	Ozono	ppb
calidad_ozono	Calidad del dato ozono	[1,4]
co	Monóxido de Carbono	ppm
calidad_co	Calidad del dato co2	[1,4]
so2	Dióxido de Azufre	ppb
calidad_so2	Calidad del dato so2	[1,4]
pst	N/A	N/A
calidad_pst	N/A	N/A
dviento_ssr	N/A	N/A
calidad_dviento_ssr	N/A	N/A
haire10_ssr	N/A	N/A
calidad_haire10_ssr	N/A	N/A



p_ssr	N/A	N/A
calidad_p_ssr	N/A	N/A
pliquida_ssr	N/A	N/A
calidad_pliquida_ssr	N/A	N/A
rglobal_ssr	N/A	N/A
calidad_rglobal_ssr	N/A	N/A
taire10_ssr	N/A	N/A
calidad_taire10_ssr	N/A	N/A
vviento_ssr	N/A	N/A
calidad_vviento_ssr	N/A	N/A

**Tabla 1. Datos de partículas contaminantes.**

Las medidas de las partículas son las siguientes:

Contaminantes	Nomenclatura	Unidades
Material Particulado menor a 1 micra	pm1	$\mu\text{g}/\text{m}^3$
Material Particulado menor a 2.5 micras	pm25	$\mu\text{g}/\text{m}^3$
Material Particulado menor a 10 micras	pm10	$\mu\text{g}/\text{m}^3$
Ozono	ozono	ppb
Monóxido de Carbono	co	ppm
Monóxido de Nitrógeno	no	ppb
Dióxido de Nitrógeno	no2	ppb
Óxidos de Nitrógeno	nox	ppb
Dióxido de Azufre	so2	ppb

**Img. 5 Mediciones de los datos.**

Cada dato tomado tiene una calidad que se interpreta de la siguiente forma:

Valor Flag	Calidad del dato
1	Dato válido
-1	Dato válido por el operado anterior
1.8 – 2.5	Dato dudoso
2.6 – 3.9	Dato malo
>= 4.0	Dato faltante
Dato -9999 y calidad (flag) 1	Equipo fuera de operación

**Img. 6 Rangos de la calidad de los datos.**

## Estaciones

Las estaciones o nubes son los sensores colocados por todo el área metropolitana que capturan los datos. La información de estas estaciones se encuentra en formato JSON y corresponden al **codigoSerial** de los datos del material particulado, Ejemplo:

```
{
  "PM2_5_CC_ICA": 80.36793320234932,
  "PM2_5_mean": 27.923704831397917,
  "PM2_5_last": 8.31825226575,
  "fecha_hora": "2020-08-05T14:00:00",
  "humedad_relativa": 34.8211666667,
  "latitude": 6.26792,
  "longitude": -75.54586359999996,
  "temperatura": 31.6323333333,
  "online": "Y",
  "altitud": 1775,
  "barrio": "El Raizal",
  "vereda": "Zona Urbana",
  "ciudad": "Medellin",
  "estado": "A",
  "nombre": "1",
  "codigo": 1
}
```

Este objeto contiene la información de la estación, donde los datos relevantes para el análisis son el código, la longitud, latitud y altura.

## Temperatura

La temperatura también se encuentra en formato JSON. Ejemplo de un día tomado por uno de los sensores:

```
{
  "date": "2019/6/4",
  "data": [
    21.9128813559,
    21.1713333333,
    21.2175,
  ]
}
```

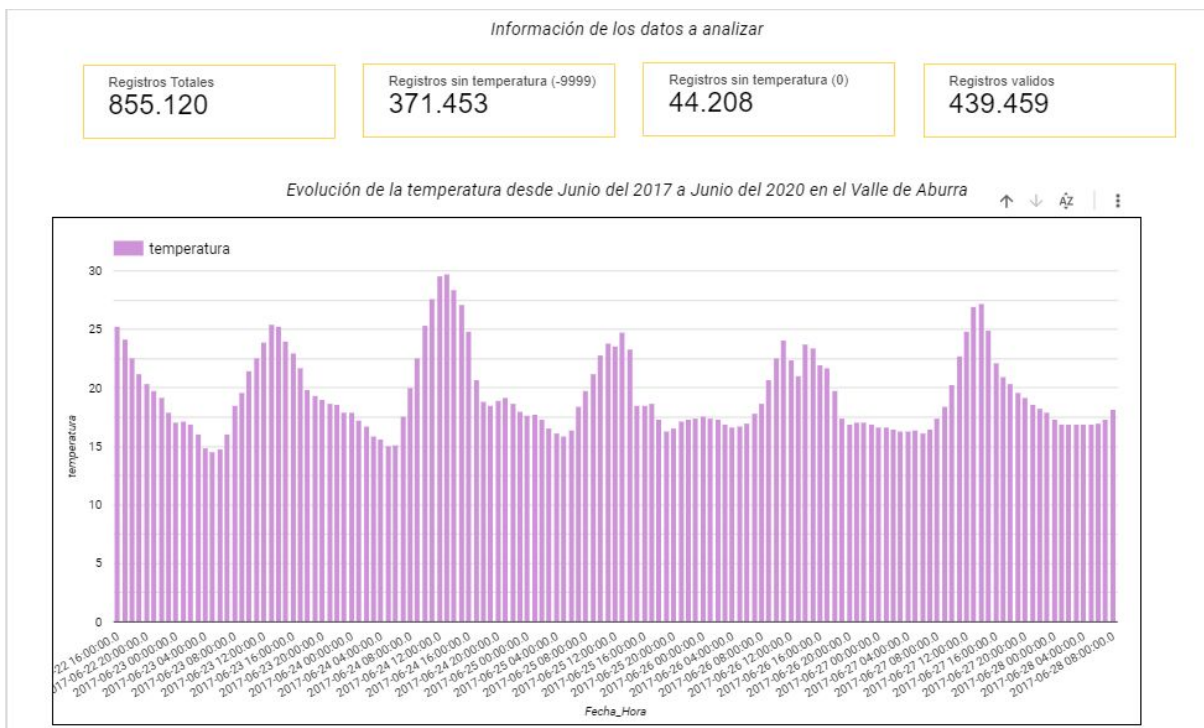
```
20.9398333333,  
20.3101666667,  
19.7935,  
19.6323333333,  
20.8713333333,  
25.3555,  
26.9653333333,  
28.3596666667,  
29.4618333333,  
30.4608333333,  
31.079,  
31.5801666667,  
31.5701666667,  
31.4405,  
29.9893333333,  
28.0673333333,  
26.735,  
25.1461666667,  
24.7545,  
24.5996666667,  
23.5653448276  
]  
}
```

La temperatura provee en el campo **data** una lista con 24 posiciones con la temperatura tomada en todo ese día, lo que corresponde a un dato por hora.

Es importante tener en cuenta que el valor -9999 significa que el dato no fue tomado.

Para lograr un mejor entendimiento de los datos se realiza un análisis en Data Studio que se puede acceder por medio de esta URL:

<https://datastudio.google.com/u/1/reporting/00117590-5ebb-4cef-96de-11658f627921/page/MC0cB>



**Img. 7 Visualización de datos en Data Studio.**

Como se observa, se tiene un total de 855.120 registros, donde los datos válidos que cuentan con el valor de la temperatura son 439.459.

### Selección de partículas

Teniendo en cuenta que no todas las estaciones cuentan con la medición de todas las partículas, se eligen las partículas que más afectan en la calidad del aire basándose en las referencias encontradas en el estado del arte, las cuales son:

- Partículas menor a 10 micrómetros
- Partículas menor a 2,5 micrómetros
- Ozono troposférico (O3)
- Dioxido de nitrógeno
- Dióxido de Azufre

Para ver como es el impacto de todas las partículas en la temperatura se debe escoger la combinación entre ellas que de mayor cantidad de datos para que el modelo se pueda entrenar bien. Se desea incluir en el modelo la mayor cantidad de partículas que se pueda, para ello se realizan diferentes combinaciones con el fin de determinar las partículas que serán incluidas:

PM2.5, PM10, Ozono, NO2, SO2 y CO
0

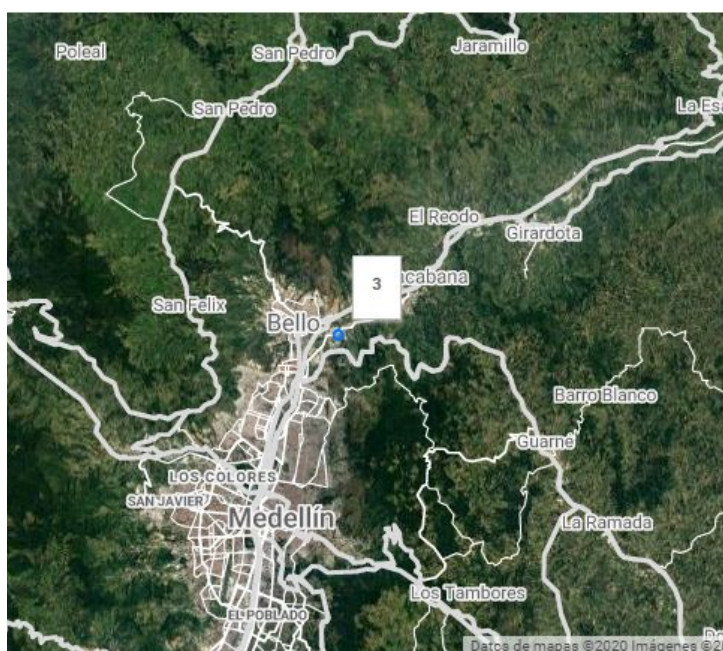
PM2.5, Ozono, NO2, SO2 y CO
8.829

PM2.5, Ozono, NO2 y CO
9.375

**Img. 8 Combinaciones de partículas para determinar la cantidad de datos.**

Solo hay 9375 datos que contienen el registro de partículas menores a 2.5 micrómetros, medición de Ozono, NO2 y CO. Pero estos registros solo se encuentran en la estación 3 en Bello:



**Img. 9 Ubicación de la estación 3.**

Debido a que la combinación de variables PM2.5, Ozono, NO2, SO2 Y CO solo se presentan en una estación se analiza la combinación de otras variables:



**Img. 10 Combinación PM2.5, Ozono, NO2, NO y NOX.**



**Img. 11 Ubicación de estaciones.**

Esta es la combinación de partículas tomadas con el mayor número de estaciones posible, las estaciones son 100, 3, 25 y 12. Con los datos de estas 4 estaciones se realizará la limpieza e implementación de modelos.



## Análisis descriptivo de los datos

	Fecha_Hora	pm25	calidad_pm25	no2	calidad_no2	ozono	calidad_ozono	no	calidad_no	nox	calidad_nox	temperatura
29530	2020-04-10 05:00:00	11.0	1.0	4.3857	1.0	10.70000	1.0	0.4816	1.0	4.8673	2.1	21.0595
22849	2019-04-09 06:00:00	30.0	1.0	19.6766	1.0	4.40000	1.0	21.8299	1.0	41.5065	1.0	21.2763
15415	2018-10-03 09:00:00	30.0	1.0	15.7238	1.0	10.30000	1.0	38.8490	1.0	54.5728	1.0	23.1980
7994	2018-04-01 04:00:00	17.0	1.0	22.5143	1.0	1.02667	1.0	12.1907	1.0	34.7043	1.0	20.2193
425	2017-08-18 23:00:00	23.0	1.0	11.0885	1.0	8.40000	1.0	6.1314	1.0	17.2199	1.0	18.5843
15092	2018-09-23 10:00:00	12.0	1.0	3.6080	2.1	29.52000	1.0	5.0282	1.0	8.6362	2.1	25.9057
16572	2018-10-30 11:00:00	0.0	3.0	7.4008	1.0	18.50000	1.0	8.8936	1.0	16.2944	1.0	22.2642
15268	2018-09-28 01:00:00	10.0	1.0	12.5953	1.0	3.90000	1.0	22.6229	1.0	35.2182	1.0	20.7080
28341	2020-03-07 21:00:00	38.0	1.0	10.0060	1.9	27.90000	2.1	1.4373	1.9	11.4433	1.9	21.7983
7719	2018-03-26 00:00:00	34.0	1.0	17.8440	1.0	2.40000	1.0	12.7272	1.0	30.5711	1.0	19.2698

img. 12 Ejemplo de los datos.

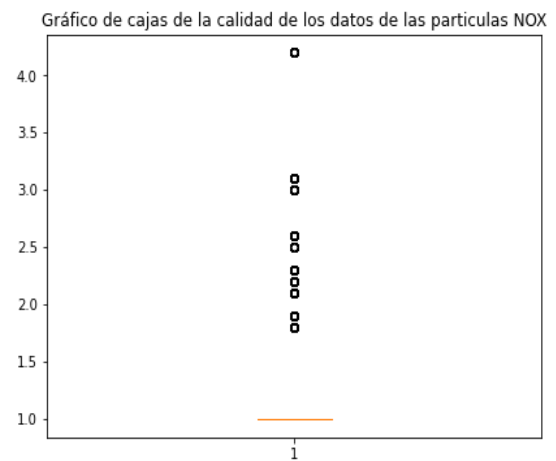
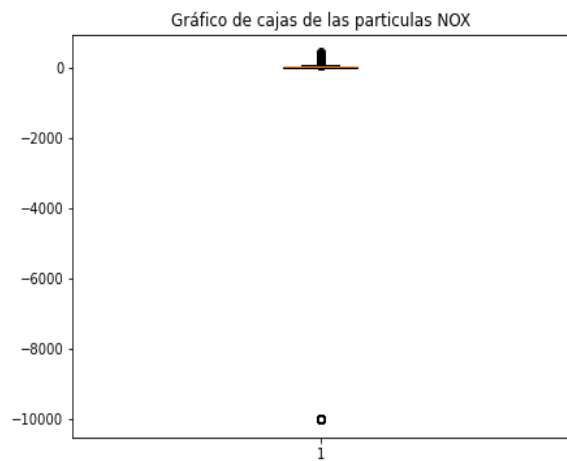
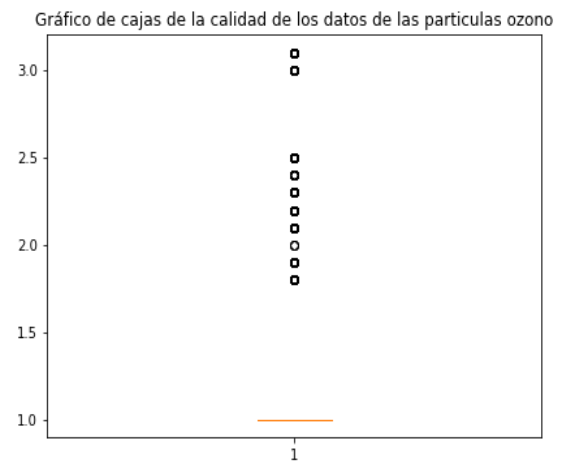
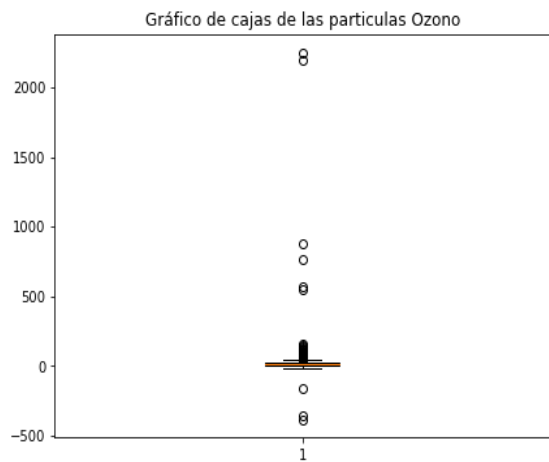
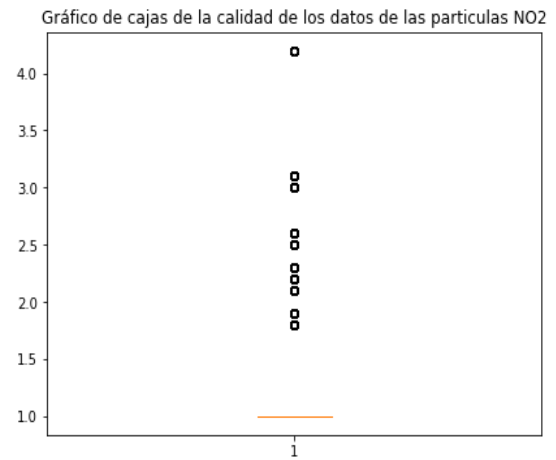
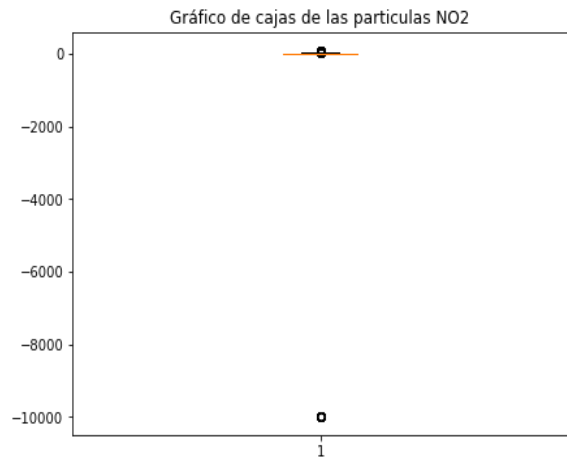
Los tipos de datos mostrados en el ejemplo son de tipo float excepto la fecha. El tamaño del dataset con el que se trabajará el modelo es **32098**.

	pm25	calidad_pm25	no2	calidad_no2	ozono	calidad_ozono	no	calidad_no	nox	calidad_nox	temperatura
count	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000	32098.000000
mean	26.935619	1.169904	-4.163540	1.179355	13.803919	1.202069	-1.527193	1.229013	12.973423	1.199139	23.252272
std	80.116356	0.518216	432.375034	0.463721	23.135484	0.480851	433.036382	0.514107	433.978397	0.478794	3.270600
min	-15.000000	1.000000	-9993.000000	1.000000	-383.300000	1.000000	-9993.000000	1.000000	-9993.000000	1.000000	14.175500
25%	11.000000	1.000000	8.017950	1.000000	3.493375	1.000000	3.388875	1.000000	12.701800	1.000000	20.741200
50%	18.000000	1.000000	12.684750	1.000000	9.600000	1.000000	8.189305	1.000000	22.872400	1.000000	22.492000
75%	27.000000	1.000000	18.863225	1.000000	21.100000	1.000000	20.890075	1.000000	40.181450	1.000000	25.491625
max	995.000000	4.299500	94.375700	4.199200	2246.100000	3.100000	442.075000	4.199200	434.499000	4.199200	35.508700

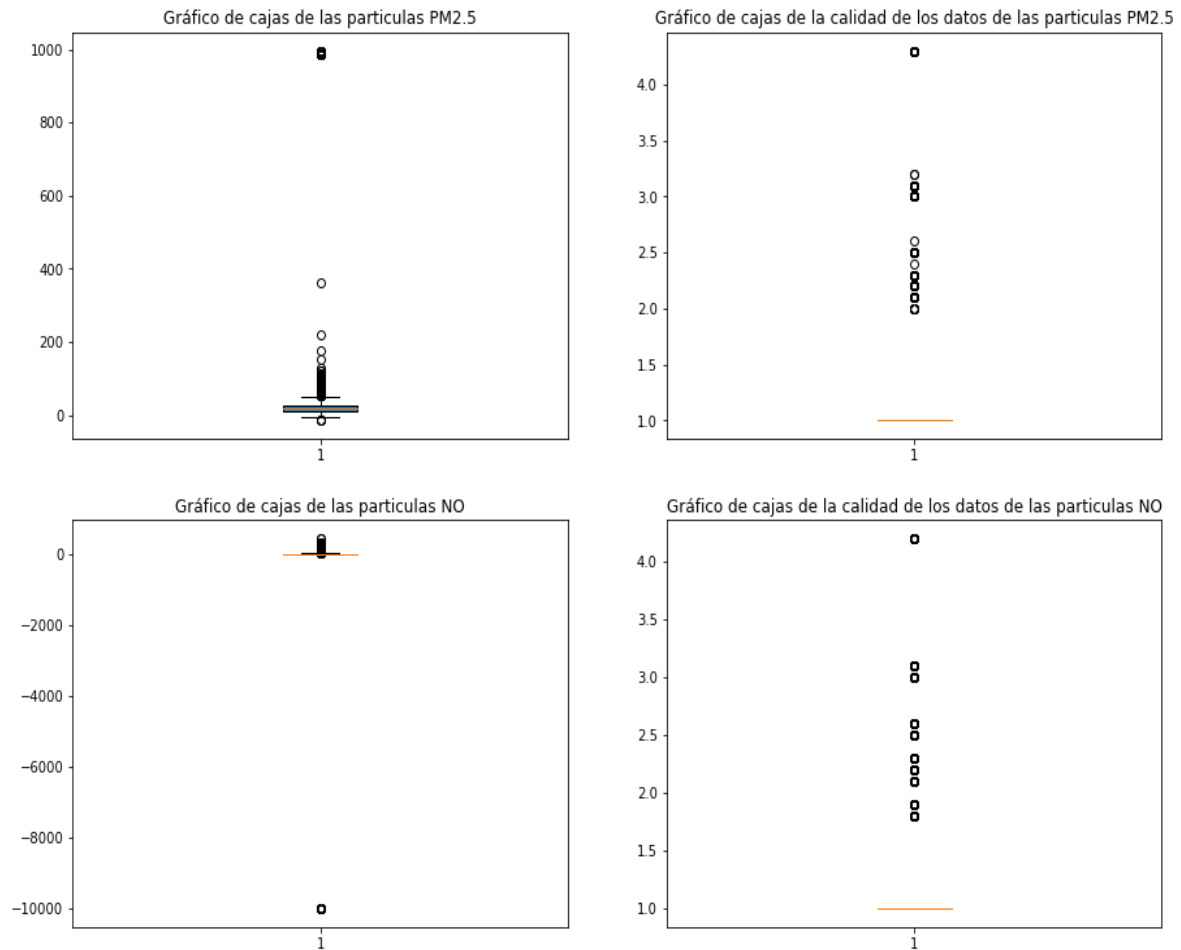
img. 13 Describe de los datos.

Se observa valores menores a cero en los datos de las mediciones, estos pueden corresponder a una medida mal tomada, con calidad mala o datos no existentes. En los valores máximos no se observan datos extraños y los cuartiles muestran que las mediciones rondan en valores acordes (no negativos), por lo que se puede decir que los datos no tomados o malos son pocos. No sobra decir que la media y la desviación estándar están siendo afectadas por los valores extremos.

En los siguientes diagramas de cajas se observa la dispersión de los datos:







**Img. 14 Diagramas de caja de las variables.**





























Gracias a estos diagramas se logra identificar una gran dispersión por datos outliers. Sin embargo, hay una gran concentración en los valores cercanos a 0.

## 5. Proceso ELT

Usando este proceso se tienen los datos cargados que permiten hacer diferentes transformaciones para lo que se llegue a necesitar. En un principio solo se tomaron datos entre el 2019 y 2020, luego de realizar la limpieza de datos y el análisis se identifica que la cantidad de datos disminuye y se requieren más datos para entrenar los modelos. **Gracias al proceso de ELT se puede adicionar datos del 2017 y 2018 sin demasiado esfuerzo.**

### Extracción


















Se descargaron los archivos .csv de la página de SIATA. Extraídos los datos de las partículas lucen de la siguiente forma:

Nombre	Fecha de modificación	Tipo	Tamaño
 estacion_data_calidadaire_3_20170101_20170131.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	158 KB
 estacion_data_calidadaire_3_20170201_20170228.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	143 KB
 estacion_data_calidadaire_3_20170301_20170331.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	158 KB
 estacion_data_calidadaire_3_20170401_20170430.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	154 KB
 estacion_data_calidadaire_3_20170501_20170531.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	158 KB
 estacion_data_calidadaire_3_20170601_20170630.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	154 KB
 estacion_data_calidadaire_3_20170701_20170731.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	158 KB
 estacion_data_calidadaire_3_20170801_20170831 (1).csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	156 KB
 estacion_data_calidadaire_3_20170801_20170831.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	156 KB
 estacion_data_calidadaire_3_20170901_20170930 (1).csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	154 KB
 estacion_data_calidadaire_3_20170901_20170930.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	154 KB
 estacion_data_calidadaire_3_20171001_20171031.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	160 KB
 estacion_data_calidadaire_3_20171101_20171130.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	155 KB
 estacion_data_calidadaire_3_20171201_20171231.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	162 KB
 estacion_data_calidadaire_3_20180101_20180131.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	162 KB
 estacion_data_calidadaire_3_20180201_20180228.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	148 KB
 estacion_data_calidadaire_3_20180301_20180331.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	162 KB
 estacion_data_calidadaire_3_20180401_20180430.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	159 KB
 estacion_data_calidadaire_3_20180501_20180531.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	166 KB
 estacion_data_calidadaire_3_20180601_20180630.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	157 KB
 estacion_data_calidadaire_3_20180701_20180731.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	163 KB
 estacion_data_calidadaire_3_20180801_20180831.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	164 KB
 estacion_data_calidadaire_3_20180901_20180930.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	157 KB
 estacion_data_calidadaire_3_20181001_20181031.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	163 KB
 estacion_data_calidadaire_3_20181101_20181130.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	158 KB
 estacion_data_calidadaire_3_20181201_20181231.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	162 KB
 estacion_data_calidadaire_3_20190101_20190131.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	163 KB
 estacion_data_calidadaire_3_20190201_20190228.csv	22/08/2020 2:50 p. m.	Hoja de cálculo d...	147 KB

**Img. 15 Datos descargados de las partículas contaminantes.**

Los archivos CSV se dividen en una carpeta por cada estación por cada mes.

Para descargar los datos de la temperatura se puede acceder con el uso de un **API REST**, para esto se realizó un script en Python para descargar la temperatura de cada estación existente. Se descargaron datos de todos los días desde el 1ro de Enero del 2017 hasta el 31 de Julio del 2020.

Nombre	Fecha de modificación	Tipo
 1	5/08/2020 10:53 a. m.	Carpeta de archivos
 2	5/08/2020 10:59 a. m.	Carpeta de archivos
 3	5/08/2020 11:03 a. m.	Carpeta de archivos
 4	5/08/2020 11:07 a. m.	Carpeta de archivos
 5	5/08/2020 11:12 a. m.	Carpeta de archivos
 9	5/08/2020 11:15 a. m.	Carpeta de archivos
 10	5/08/2020 11:19 a. m.	Carpeta de archivos
 11	5/08/2020 11:21 a. m.	Carpeta de archivos
 12	5/08/2020 11:24 a. m.	Carpeta de archivos
 13	5/08/2020 11:28 a. m.	Carpeta de archivos
 14	5/08/2020 11:31 a. m.	Carpeta de archivos
 15	5/08/2020 11:34 a. m.	Carpeta de archivos
 16	5/08/2020 11:37 a. m.	Carpeta de archivos
 18	5/08/2020 11:42 a. m.	Carpeta de archivos
 19	5/08/2020 11:46 a. m.	Carpeta de archivos
 20	5/08/2020 11:48 a. m.	Carpeta de archivos
 22	5/08/2020 11:52 a. m.	Carpeta de archivos

**Img. 16 Datos de la temperatura por estación.**

temperatura_2019_6_1.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_2.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_3.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_4.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_5.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_6.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_7.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_8.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_9.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_10.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_11.json	5/08/2020 11:31 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_12.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_13.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_14.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_15.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_16.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_17.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_18.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_19.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_20.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB
temperatura_2019_6_21.json	5/08/2020 11:32 a. m.	Archivo de origen ...	1 KB

Img. 17 Datos de la temperatura de una estación.

## Carga

En la carga se procedió a cargar los datos CSV a una base de datos MySQL, seguida de la temperatura. Esta información fue almacenada en una sola tabla. En esta también se cargaron los datos de la altura, longitud y latitud correspondiente a cada estación para poderlas ubicar geográficamente.

	Fecha_Hora	codigoSerial	pm25	calidad_pm25	pm10	calidad_pm10	pm1	calidad_pm1	no	calidad_no	no2	calidad_no2	nox	calidad_nox	ozc
▶	2017-01-01 00:00:00	3	58	1	-9999	1	-9999	1	12.8549	1	12.3501	1	25.2029	1	0.7
	2017-01-01 00:00:00	6	-9999	1	11	1	11	1	35.4668	1	8.96708	1	44.4347	1	-99
	2017-01-01 00:00:00	11	-9999	1	100	2.3	-9999	1	-9999	1	-9999	1	-9999	1	-99
	2017-01-01 00:00:00	12	65	1	97	1	-9999	1	60.9923	1	13.7697	1	74.762	1	2.5
	2017-01-01 00:00:00	25	56	1	-9999	1	-9999	1	66.5462	1	12.3508	1	78.9	1	0.1
	2017-01-01 00:00:00	28	139	1	-9999	1	-9999	1	31.0268	1	13.406	1	44.4343	1	-99
	2017-01-01 00:00:00	31	54	1	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	1.7
	2017-01-01 00:00:00	37	-9999	1	37	1	-9999	1	0.836667	1	5.97583	1	6.81306	1	1.0
	2017-01-01 00:00:00	38	40	1	61	1	-9999	1	-9999	1	-9999	1	-9999	1	1.5
	2017-01-01 00:00:00	40	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	1.1
	2017-01-01 00:00:00	41	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	0.9
	2017-01-01 00:00:00	43	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	0.7
	2017-01-01 00:00:00	44	11	1	-9999	1	-9999	1	-9999	1	-9999	1	-9999	1	2.4

Img. 18 Base de datos MySQL con los datos.

## Transformación

Con los diagramas de caja realizados en el entendimiento de los datos se puede observar que los datos están con bastantes outliers, los cuales si nos son tratados correctamente dañan el modelo. De estas gráficas se puede decir que:

- El 50% de los datos rondan en valores cercanos a 0 (-100 y 100) en  $\mu\text{g}/\text{m}^3$  para el PM2.5 y ppb para las demás partículas

- Según el IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales) los rango de las partículas son las siguientes:

Rangos ICA	Clasificación	O <sub>3</sub> 8h (ppm)	O <sub>3</sub> 1h (ppm) <sup>1</sup>	PM <sub>10</sub> 24h (µg/m <sup>3</sup> )	PM <sub>2.5</sub> 24h (µg/m <sup>3</sup> )	CO 8h (ppm)	SO <sub>2</sub> 24h (ppm)	NO <sub>2</sub> 1h (ppm)
$0 \leq ICA \leq 50$	Verde	0,000 0,059	-	0 54	0,0 15,4	0,0 4,4	0,000 0,034	(2)
$51 \leq ICA \leq 100$	Amarillo	0,060 0,075	-	55 154	15,5 40,4	4,5 9,4	0,035 0,144	(2)
$101 \leq ICA \leq 150$	Anaranjado	0,076 0,095	0,125 0,164	155 254	40,5 65,4	9,5 12,4	0,145 0,224	(2)
$151 \leq ICA \leq 200$	Rojo	0,096 0,115	0,165 0,204	255 354	65,5 150,4	12,5 15,4	0,225 0,304	(2)
$201 \leq ICA \leq 300$	Morado	0,116 0,374 (0,155 0,404) (4)	0,205 0,404	355 424	150,5 250,4	15,5 30,4	0,305 0,604	0,65 1,24
$301 \leq ICA \leq 400$	Marrón	(3)	0,405 0,504	425 504	250,5 350,4	30,5 40,4	0,605 0,804	1,25 1,64
$401 \leq ICA \leq 500$	Marrón	(3)	0,505 0,604	505 604	350,5 500,4	40,5 50,4	0,805 1,004	1,65 2,04

Fuente: Manual de Operación de Sistemas de Vigilancia de Calidad de Aire del Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire (MAVDT, 2010 Pág. 134).

**Img. 19 Puntos de corte para el cálculo del ICA.**

Con el factor de conversión del 1ppm = 1000ppb se tienen las siguientes medidas:

- 0 µg/m<sup>3</sup> <= PM<sub>2.5</sub> < 500.4 µg/m<sup>3</sup>
- 0 ppb <= NO<sub>2</sub> < 2040 ppb
- 0 ppb <= NO < 2040 ppb
- 0 ppb <= NO<sub>x</sub> < 2040 ppb
- 0 ppb <= Ozono < 604 ppb

Se tomó como base la calidad del dato para hacer la transformación, la cual consiste en reemplazar los datos que tienen una calidad mayor a 2,6 por la media de cada uno:

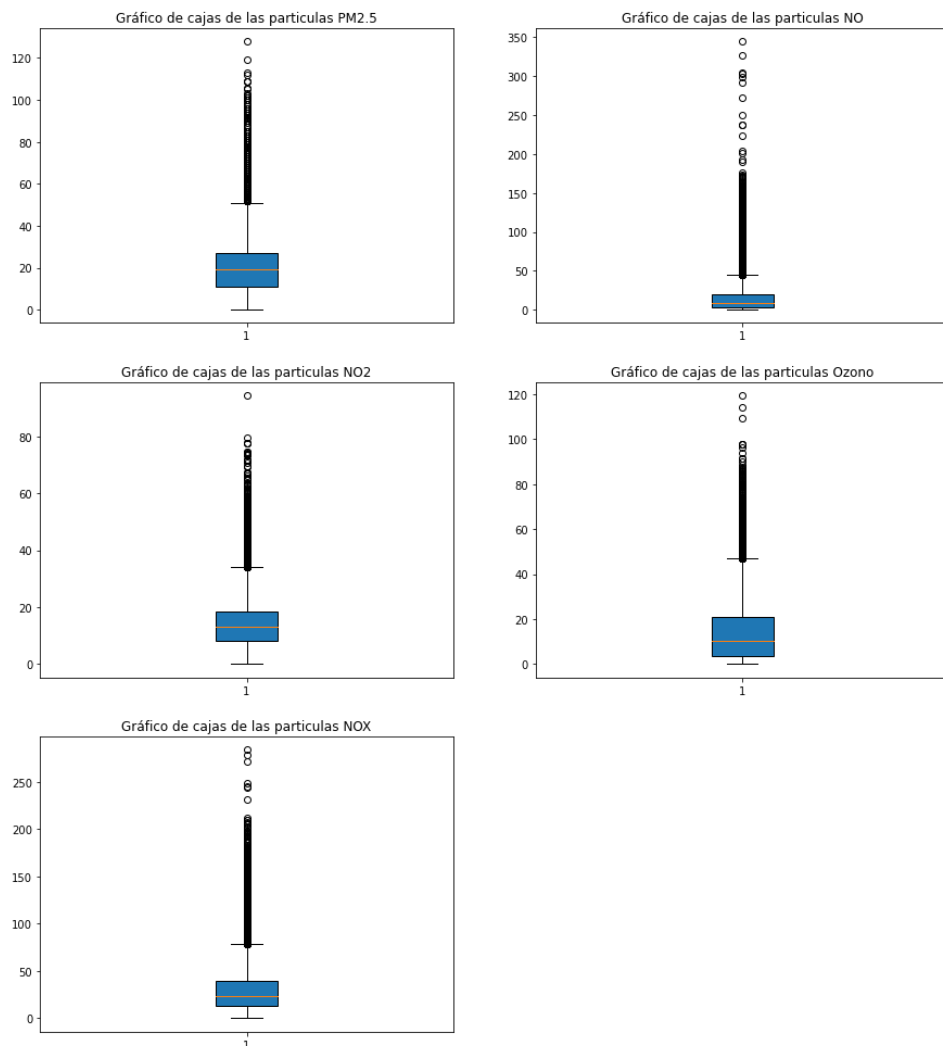
- El 3.54% de PM<sub>2.5</sub> va a ser reemplazado por 20.82 µg/m<sup>3</sup>
- El 3.07% de NO<sub>2</sub> va a ser reemplazado por 14.53 ppb
- El 2.42% de Ozono va a ser reemplazado por 13.74 ppb
- El 3.00% de NO va a ser reemplazado por 17.19 ppb
- El 3.07% de NO<sub>x</sub> va a ser reemplazado por 31.65 ppb

Como dato adicional, se sabe que la temperatura varía durante el día, por lo tanto se adiciona la variable hora. Luego de esto, los datos tienen los siguientes estadísticos:



	count	mean	std	min	25%	50%	75%	max
pm25	32098.0	20.818019	13.070837	0.0000	11.000000	19.000000	27.000000	128.0000
calidad_pm25	32098.0	1.169904	0.518216	1.0000	1.000000	1.000000	1.000000	4.2995
no2	32098.0	14.525961	8.881023	0.0000	8.283775	13.087100	18.623750	94.3757
calidad_no2	32098.0	1.179355	0.463721	1.0000	1.000000	1.000000	1.000000	4.1992
ozono	32098.0	13.736053	12.422501	0.0000	3.618112	10.372500	21.000000	119.4030
calidad_ozono	32098.0	1.202069	0.480851	1.0000	1.000000	1.000000	1.000000	3.1000
no	32098.0	17.192831	22.761418	0.0000	3.638825	8.762465	20.364775	345.0050
calidad_no	32098.0	1.229013	0.514107	1.0000	1.000000	1.000000	1.000000	4.1992
nox	32098.0	31.647010	27.223627	0.0000	13.262075	23.972400	39.499600	283.8110
calidad_nox	32098.0	1.199139	0.478794	1.0000	1.000000	1.000000	1.000000	4.1992
temperatura	32098.0	23.252272	3.270600	14.1755	20.741200	22.492000	25.491625	35.5087

Img. 20 Descripción de los datos.

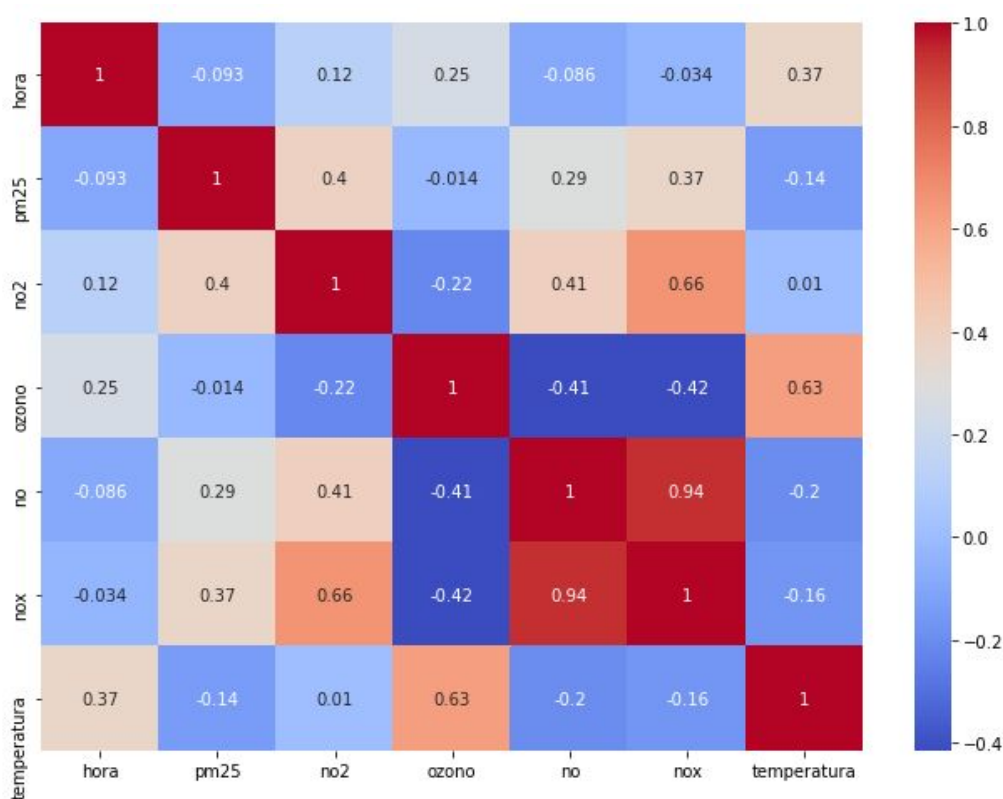


Img. 21 Diagrama de caja de las partículas elegidas.

Todos los estadísticos cambiaron debido a que ya no hay datos extremos y no hay valores negativos. Realizando diagramas de cajas se puede observar el cambio: los datos tienden a ser más estables según su dispersión, sin embargo se muestran algunos como outliers pero no lo son debido a que están dentro de los límites permitidos.

## 6. Análisis y construcción de los modelos

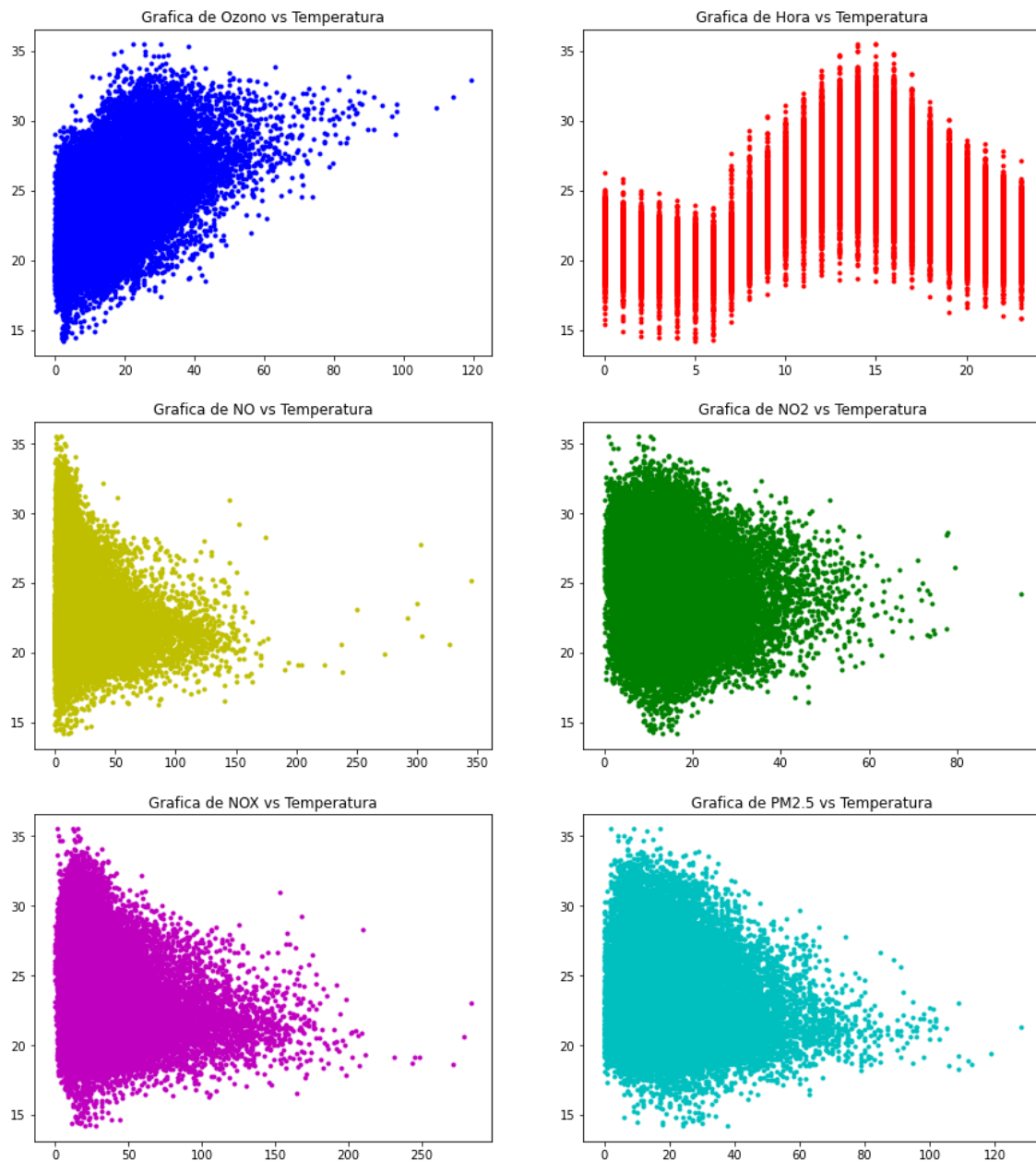
Una vez finalizada la transformación de los datos, limpieza y análisis se procede a construir los modelos. Primero se realizó una matriz de correlación para ver las variables que más impactan en la temperatura y así descartar variables ya sea con una correlación muy alta entre sí o con muy poca influencia en la temperatura:



Img. 22 Matriz de correlación.

Dentro de las transformaciones se realizó la exclusión de la partícula NOX teniendo en cuenta una correlación de 0.94 con la partícula NO y su baja correlación con la variable objetivo (temperatura), además se excluye la calidad de los datos puesto que no brinda información adicional al análisis.

A continuación se muestran 6 gráficas para observar el comportamiento de cada partícula y la hora con la temperatura:



**Img. 23 Gráfica de las variables de entrada Vs variable objetivo.**

El Ozono presenta una correlación casi lineal con la temperatura y es mayor que la correlación entre la temperatura y la hora. Basados en la visualización en Data Studio se intuye que la correlación entre la hora y la temperatura sería mayor, sin embargo, de acuerdo con este análisis no lo es, posiblemente se debe a que hay otros factores que afectan la temperatura mayor que la hora como la estación del año y la humedad atmosférica.

### **Definición de datos de entrenamiento y prueba**

Para los modelos se dividen los datos de entrenamiento y prueba en un 75% y un 25% respectivamente.

### Definición de los modelos a construir

Los modelos que se van a construir son: Árbol de decisiones, Random Forest, Soporte de Regresión Vectorial (SVR), Regresión Lineal Multivariante y Redes neuronales artificiales (ANN).

¿Por qué estos modelos? Se eligió SVR y ANN debido a que en trabajos pasados encontrados en el estado del arte se comentó que estos dos dieron buenos resultados para el análisis de la calidad del aire, aunque en este caso es la temperatura. Sin embargo, el análisis que se llevó a cabo tiene cierta igualdad con la planteada en este proyecto puesto que se construyeron los modelos usando las partículas contaminantes como datos de entrada.

Se eligieron también Árboles de decisiones y Random Forest porque son modelos que dan buenos resultados cuando se tienen variables predictoras y MLR se pensó como un modelo adecuado para este ejercicio, aunque las variables no tenían una relación lineal con la temperatura.

Luego de realizar los primeros modelos se decidió implementar una categorización de la temperatura debido a que los resultados de los modelos para la variable objetivo continua no fueron los esperados.

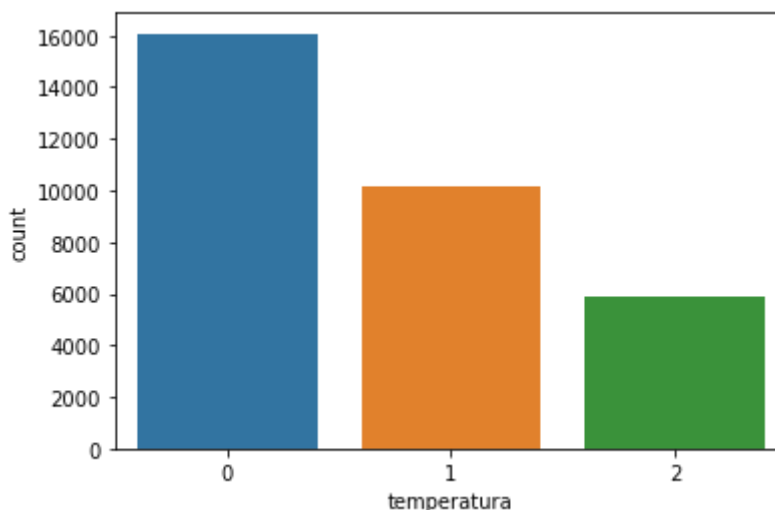
La categorización se hizo de la siguiente forma:



**Img. 24 Rangos de temperatura.**

Teniendo en cuenta la gráfica anterior se definieron los siguientes rangos:

1. Frío, temperatura menor a 22 °C
2. Medio, temperatura entre 22°C y 26°C
3. Alto, temperatura mayor a 26°C



**Img. 25 Variables categóricas.**



Adicional a esto, se implementa un Oversampling pero el resultado de los modelos no mejoró, por que se decide no usarlo.

**IMPORTANTE:** Para ver la construcción de los modelos y el desarrollo del proyecto por favor descargar los Notebooks desde el siguiente enlace:  
[https://drive.google.com/file/d/1WviW2EKZ9bgCGe\\_U09XNAtQvNK6AF029/view?usp=sharing](https://drive.google.com/file/d/1WviW2EKZ9bgCGe_U09XNAtQvNK6AF029/view?usp=sharing)

## EVALUACIÓN

### Evaluación de los modelos con temperatura continua

Para evaluar de los modelos con variable a predecir continua se usaron los siguientes métodos:

1. Error cuadrático medio
2. Coeficiente de determinación (R2 Score)
3. Distribución de los residuos y gráficas QQ
4. Pruebas aleatorias y una prueba real

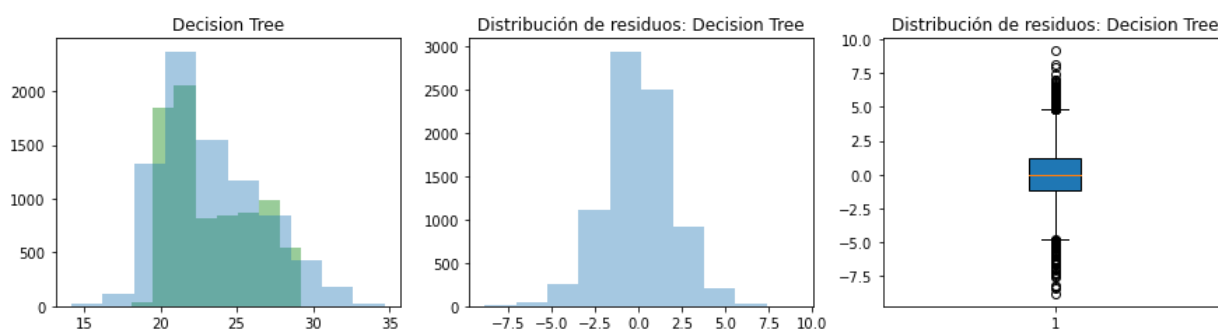
Los resultados de cada modelo son:

Modelo	Error cuadrático medio	R2 Score	Cross-Validation
Árbol de decisiones	1,95	0,64	0,8
Random Forest	1,83	0,68	0,83
SVR	1,99	0,62	0,79
MLR	2,35	0,47	0,7
ANN	5,55	0,48	N/A

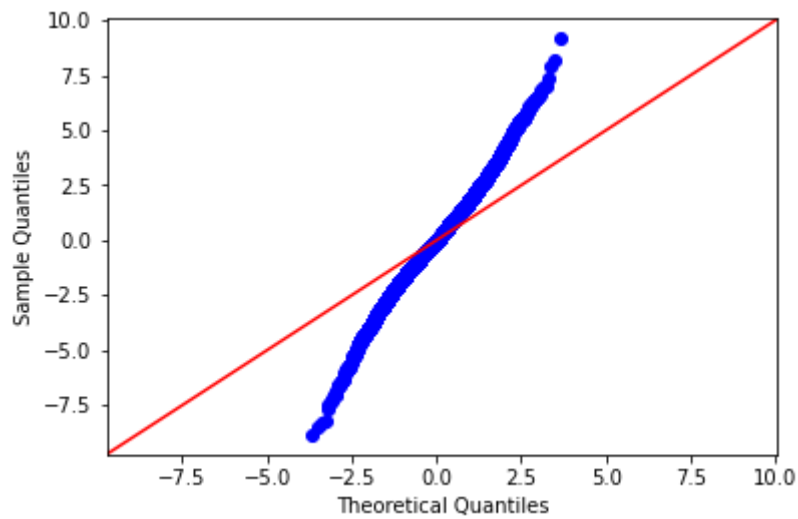
Tabla 2. Resultados de los modelos

### Gráficas de resultado de prueba vs real, QQ y distribución de residuos

#### Árbol de decisiones

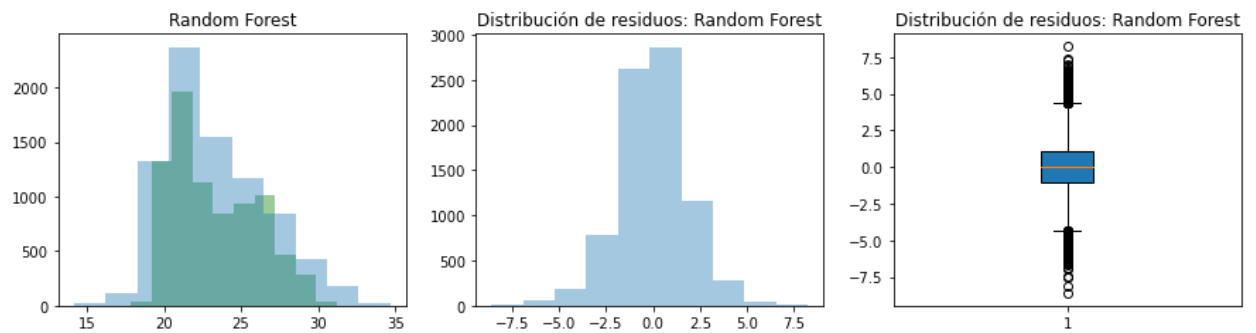


Img. 26 Resultados de los Árboles de decisión.

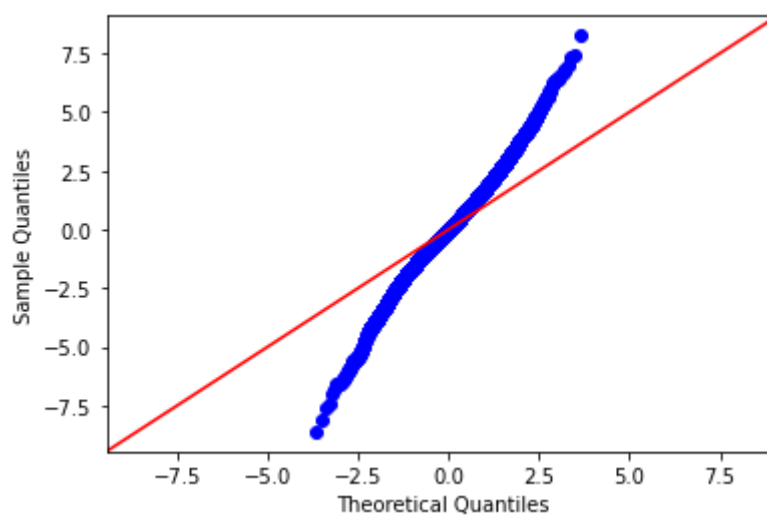


Img. 27 Gráfica QQ Árboles de decisión.

### Random Forest

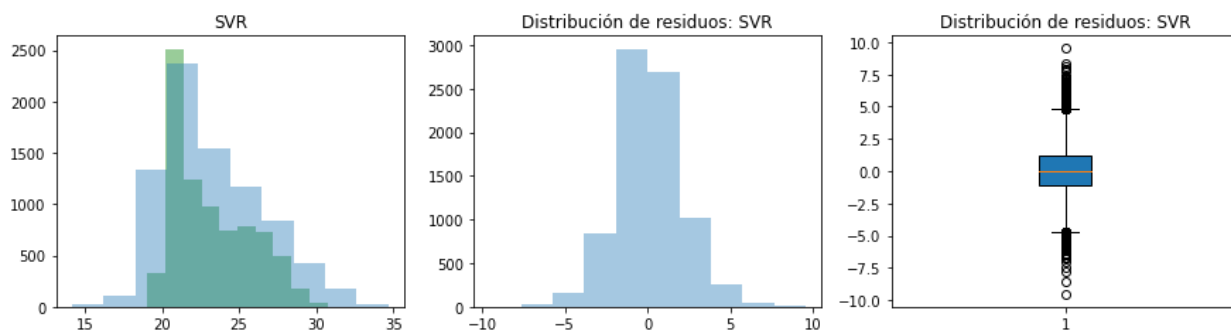


Img. 28 Resultados del Random Forest.

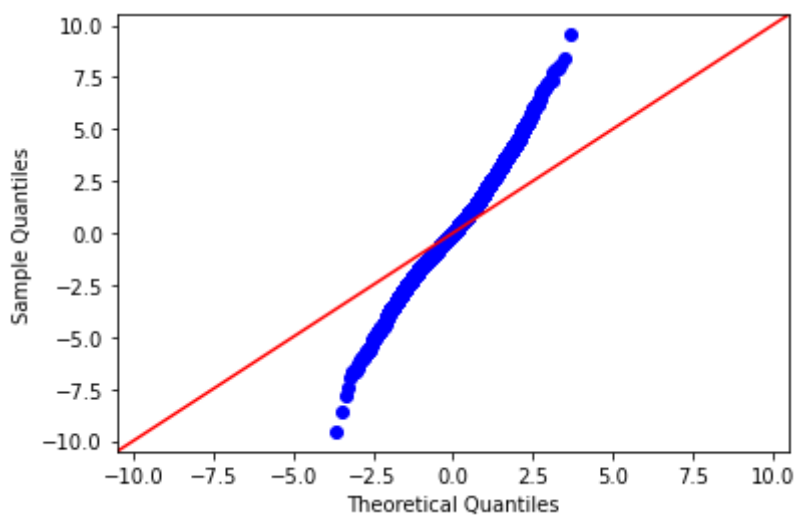


Img. 29 Gráfica QQ del Random Forest.

## SVR

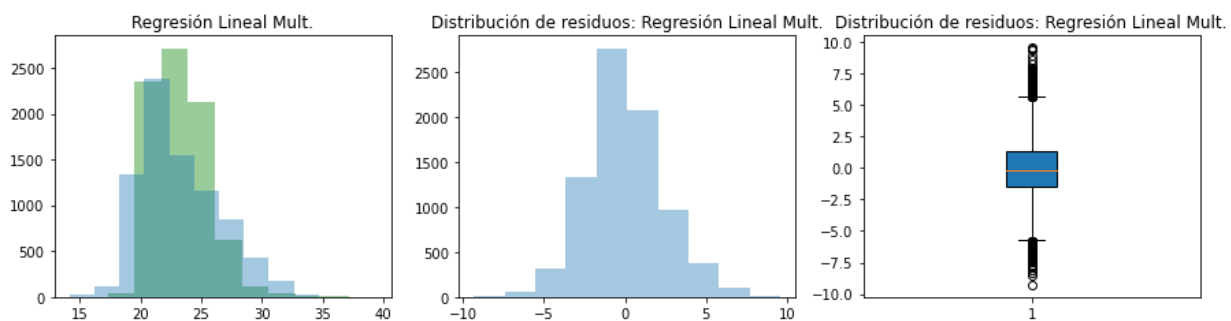


Img. 30 Resultados del SVR.

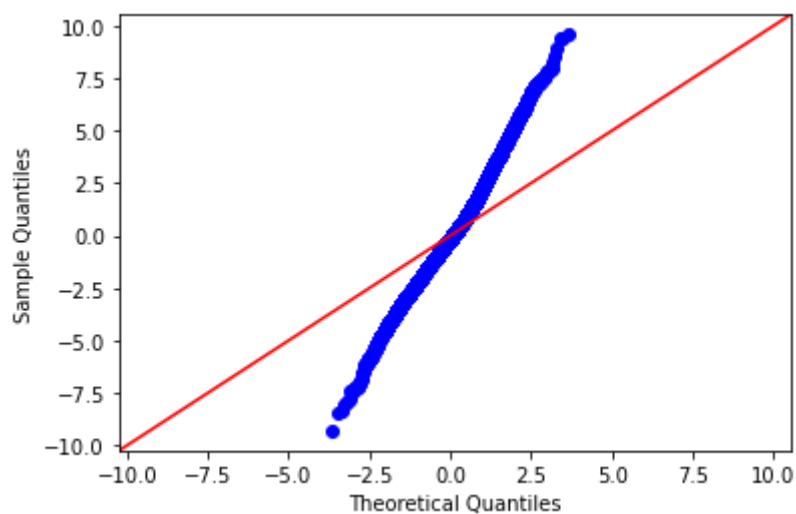


Img. 31 Gráfica QQ del SVR.

## Regresión Lineal Multivariante

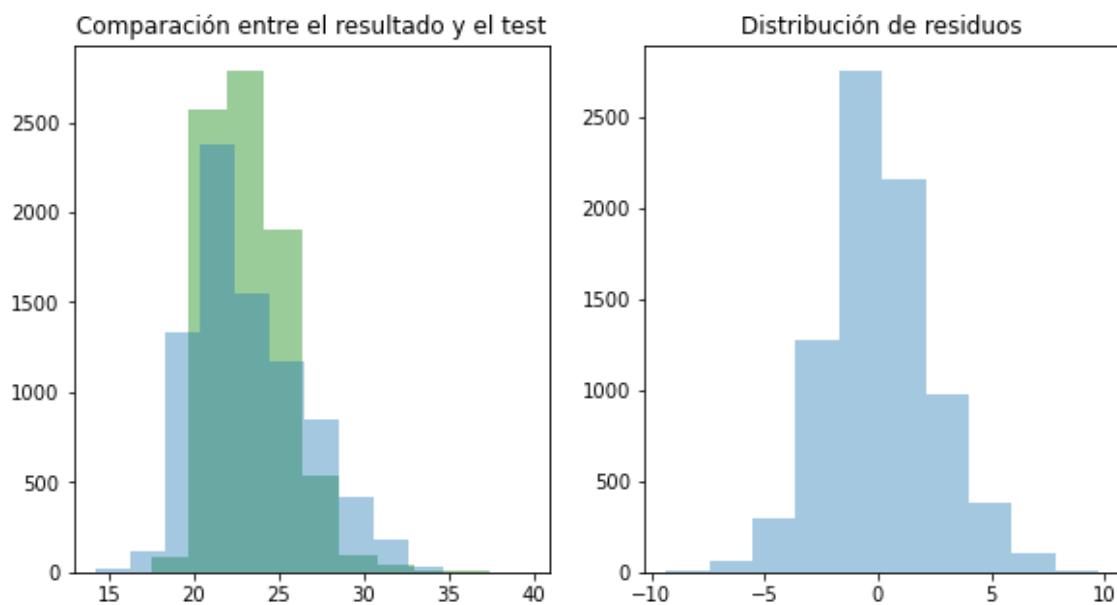


Img. 32 Resultados del MLR.

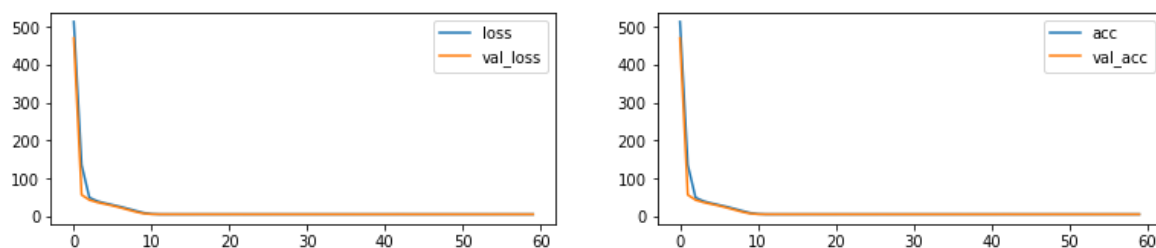


Img. 33 Gráfica QQ del MLR.

## ANN



Img. 34 Resultados de ANN.



Img. 35 Entrenamiento de ANN.

## Evaluación con datos aleatorios

Se realiza la siguiente prueba con el dato aleatorio:

pm25 19.0000

no2 18.6532

ozono 0.7000

no 25.5631

hora 2.0000

Temperatura real: **17.44**

Modelo	temperatura predecida
<i>Árbol de decisiones</i>	20,32
<i>Random Forest</i>	19,57
<i>SVR</i>	20,35
<i>MLR</i>	20,71

**Tabla 3. Resultados de modelos**

Para la ANN se usaron los siguientes:

pm25 17.0000

no2 19.2241

ozono 5.2000

no 22.3033

hora 20.0000

Temperatura real: **20.10**

Temperatura predecida: **22.85**

## Evaluación con dato real

Se realiza una consulta a un dato que no está incluido dentro del dataset del día 30 de Junio del 2020 a las 13:00. La información es la siguiente:

pm25 18.0

no2 5.3097

ozono 33.5

nox 7.6488

hora 13

Temperatura real: **28.63**

Los resultados arrojados por los modelos fueron los siguientes:

Modelo	temperatura predecida
<i>Árbol de decisiones</i>	29,23
<i>Random Forest</i>	29,32
<i>SVR</i>	28,24
<i>MLR</i>	26,16
<i>ANN</i>	26,2

**Tabla 4. Resultados de modelos**

### Evaluación de los modelos con temperatura categórica

Para la evaluación de estos modelos se usó:

1. La precisión de los modelos
2. Matriz de confusión
3. distribución de residuos (al ser enteros 0, 1 y 2)

### Precisión

Modelo	Precisión
<i>Regresión logística</i>	0,66
<i>SVM</i>	0,71
<i>Árbol de decisiones</i>	0,67
<i>Random Forest</i>	0,73
<i>ANN</i>	0,73

**Tabla 5. Resultados de modelos**

### Pruebas aleatorias

Se realizó la prueba para los modelos con los siguientes datos de entrada:

pm25 20.0000

no2 19.5965

ozono 7.1000

no 6.4641

hora 21.0000

Temperatura: **1 (MEDIA entre 26 °C y 28°C)**

Modelo	Temperatura
<i>Regresión logística</i>	0
<i>SVM</i>	0
<i>Árbol de decisiones</i>	1
<i>Random Forest</i>	1

**Tabla 6. Resultados de modelos**

Para la ANN se realizó la prueba con los siguientes datos:

pm25 13.0000

no2 7.5740

ozono 13.9000

no 10.6007

hora 9.0000

Rango de temperatura real: 1

Rango de temperatura predecida 1

## RESULTADOS

Para determinar si el modelo es bueno o no se debe tener en cuenta los valores de las métricas usadas para evaluación de modelos:

- Para temperatura con valores numéricos:

Métrica	Valor objetivo
Error cuadrático medio (RMSE)	0
Coeficiente de determinación (R2)	1
Cross-Validation	1
Distribución de residuos	Distribución normal
Gráfica QQ	Lineal

**Tabla 7. Valores de las métricas.**

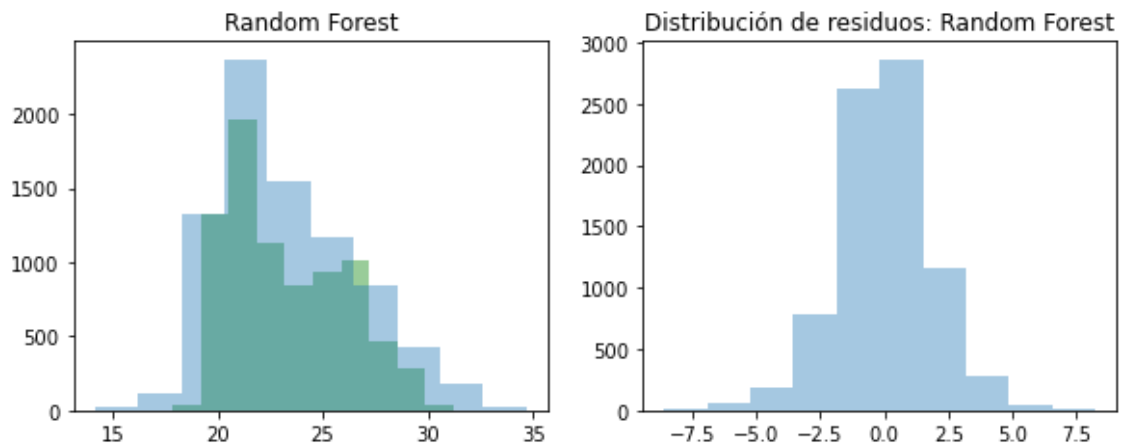
- Para temperatura con valores categóricos:

Métrica	Valor objetivo
Accuracy	1
Matriz de confusión	Diagonal con todos los valores
Distribución de residuos	Distribución normal

**Tabla 8. Valores de las métricas.**

Teniendo en cuenta las métricas usadas para la evaluación de los modelos se obtienen los siguientes resultados:

Se observa que el modelo que presenta mejores resultados para datos de entrada continuos es Random Forest. Este modelo con un R2 de 0.68, error cuadrático medio de 1,83 y cross-validation de 0.83 permite una predicción con mayor precisión de la temperatura.



**Img. 36 Evaluación gráfica del modelo**

En la gráfica de la izquierda se observa una similitud entre los datos reales de salida y los datos que arroja el modelo. La diferencia se debe a las variaciones decimales que se pueden presentar al momento de realizar la predicción, tal como se observa en la gráfica de la derecha donde se muestra la distribución de los residuos la cual presenta una distribución normal, donde la variación de la temperatura predecida puede ser de 8.75°C aproximadamente.

Los resultados obtenidos de los modelos muestran una precisión de predicción del modelo baja aunque mayor al 50% y un error cuadrático medio mayor a 1, esto puede atribuirse a las diferencias de decimales entre los datos predecidos y su valor real. Por esta razón se decide realizar una clasificación de los datos de la variable objetivo, dividiendo los valores de temperatura en tres rangos y con esto realizar nuevamente la evaluación de los modelos.

Se obtiene como mejor modelo con uso de valores categóricos para la variable objetivo el modelo Random Forest, con un accuracy de 0.73 y las redes neuronales con un accuracy de 0.73, ambos modelos con igual precisión.

Adicionalmente, en el análisis de datos y visualización se observa un incremento en los valores de la temperatura desde el año 2017 y un comportamiento relacionado con la hora del día donde la temperatura es baja en horas de la mañana y aumenta entre mediodía y tarde.

En los modelos para variables continuas se observa en las gráficas QQ que los residuos no siguen la diagonal, por lo tanto los modelos no son tan precisos.

Con el cambio de de variable continua a categoría de la temperatura se observa un incremento en el R2 debido a que se aumenta la probabilidad de un resultado esperado, sin embargo, no se tiene la temperatura exacta sino un rango.



## CONCLUSIONES Y TRABAJOS FUTUROS

### Conclusiones

- Se debe de tener precauciones al momento de trabajar con datos de sensores (IoT) porque se pueden presentar demasiados valores erróneos o no tomados debido a que en ocasiones los sensores pueden fallar o apagarse. Para contrarrestar esta desventaja se debe de usar una mayor cantidad de datos para poder entrenar bien los modelos.
- La temperatura es afectada por otros factores climáticos que no fueron tenidos en cuenta, lo que hizo que los modelos fueran menos precisos. Debido a estos mismos factores la temperatura puede variar sin tener en cuenta las variables que se incluyeron en el modelo. Por ejemplo, la temperatura en las mañanas es baja pero en el transcurso del día comienza a subir y en la noche a bajar nuevamente, además dependiendo de la temporada puede cambiar el valor de la temperatura durante el día.
- La correlación entre la temperatura, las partículas y la hora es baja pero por la precisión que se logró con los modelos en conjunto si tienen una relación. El Ozono presenta una correlación alta con la temperatura, donde de manera gráfica se observa una linealidad entre estas dos variables. El Ozono es un gas incoloro e irritante que es producido en su mayor parte por la quema de combustibles, vapores de gasolina y solventes químicos lo que hace más evidente una relación entre la actividad humana y el calentamiento global.
- Los modelos que presentaron mejor precisión a lo largo del trabajo tanto en temperatura continua y categórica fueron el Árbol de decisiones y Random Forest, esto se debe a que ambos generan ramas definiendo un rango dependiendo de los valores de entrada y generando resultados a partir de estas ramas donde no interfieren ecuaciones directamente.
- Se obtiene mayor precisión en la predicción de la temperatura cuando se usan datos categóricos para esta variable, dado que los valores a predecir se reducen a tres categorías y se aumenta la probabilidad de obtener el resultado esperado.

### Trabajos futuros

- Seguir con el proyecto agregando más datos para mejorar la precisión de los modelos. Se podría llegar a usar este modelo para predecir cómo sería la temperatura variando las partículas contaminantes y de esta forma identificar hasta qué nivel se puede limitar para reducir los efectos del calentamiento global y se podría hacer lo mismo para la calidad del aire.
- Poder predecir la temperatura para fechas futuras sin depender de las partículas contaminantes, es decir, únicamente observando el comportamiento de la temperatura a lo largo del tiempo.

- Crear un modelo que permita predecir cuántas personas se enfermaron por la mala calidad del aire y también predecir qué tipo de enfermedades padecerán para poder tomar medidas preventivas.

## REFERENCIAS

- [1] (2018). Nueve de cada diez personas de todo el mundo respiran aire contaminado. Organización mundial de la salud. Recuperado de: <https://www.who.int/es/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breath-e-polluted-air-but-more-countries-are-taking-action>
- [2] Kaur, G., Zeyu, J., Chiao, S., Lu, S. y Xie, G. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development 9(1):8-16. Recuperado de: [https://www.researchgate.net/publication/323012727\\_Air\\_Quality\\_Prediction\\_Big\\_Data\\_and\\_Machine\\_Learning\\_Approaches](https://www.researchgate.net/publication/323012727_Air_Quality_Prediction_Big_Data_and_Machine_Learning_Approaches)
- [3] Bhalgat, P., Bhoite, S., y Pitare, S. (2019). Air Quality Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656. Recuperado de: [https://www.researchgate.net/publication/335911816\\_Air\\_Quality\\_Prediction\\_using\\_Machine\\_Learning\\_Algorithms](https://www.researchgate.net/publication/335911816_Air_Quality_Prediction_using_Machine_Learning_Algorithms)
- [4] Shahriar, S., Kayes, I., Hasan, K., Abdus, M., y Chowdhury, S. (2020). Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh. Recuperado de: [https://www.researchgate.net/publication/343093608\\_Applicability\\_of\\_machine\\_learning\\_in\\_modeling\\_of\\_atmospheric\\_particle\\_pollution\\_in\\_Bangladesh](https://www.researchgate.net/publication/343093608_Applicability_of_machine_learning_in_modeling_of_atmospheric_particle_pollution_in_Bangladesh)
- [5] Doreswamy, H., Harishkumar, K.S., Km, Y., y Gad, I. (2020). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Recuperado de: [https://www.researchgate.net/publication/341908293\\_Forecasting\\_Air\\_Pollution\\_Part particulate\\_Matter\\_PM25\\_Using\\_Machine\\_Learning\\_Regression\\_Models](https://www.researchgate.net/publication/341908293_Forecasting_Air_Pollution_Part particulate_Matter_PM25_Using_Machine_Learning_Regression_Models)
- [6] Afgun, R., Saeed, A., Mahmoud, A., Ramiah, T., Zaman, N. y Abaker, I. (2020). Air pollution and its health impacts in Malaysia: a review. Recuperado de : [https://www.researchgate.net/publication/343142206\\_Air\\_pollution\\_and\\_its\\_health\\_impacts\\_in\\_Malaysia\\_a\\_review](https://www.researchgate.net/publication/343142206_Air_pollution_and_its_health_impacts_in_Malaysia_a_review)
- [7] Gaitán, M., Cancino, J., y Behrentz, E. (2007). Análisis del estado de la calidad del aire en Bogotá. Revista de Ingeniería, núm. 26, noviembre, 2007, pp. 81-92. Recuperado de: <https://www.redalyc.org/pdf/1210/121015050011.pdf>
- [8] Bedoya, J. y Martínez, E. (2009). CALIDAD DEL AIRE EN EL VALLE DE ABURRÁ ANTIOQUIA -COLOMBIA. Dyna, vol. 76, núm. 158, junio, 2009, pp. 7-15. Recuperado de : <https://www.redalyc.org/pdf/496/49612069002.pdf>
- [9] Zapata, C., Quijano, R., Molina, E., Rubiano, C., y Londoño, G. (2008). Fortalecimiento de la Red de Monitoreo de Calidad de Aire en el Valle de Aburrá con Medidores Pasivos. Gestión y Ambiente, vol. 11, núm. 1, mayo, 2008, pp. 67-84. Recuperado de: <https://www.redalyc.org/pdf/1694/169414452004.pdf>

- [10] Casado, A. (2018). BIG DATA APLICADO AL TRANSPORTE Y EN LAS CIUDADES: ADAPTACIÓN A LA CIUDAD DE SEVILLA. Recuperado de: <https://idus.us.es/bitstream/handle/11441/79499/Casado%20Reinaldos%20Alejandro%20TFG%20%282%29.pdf?sequence=1&isAllowed=y>
- [11] Villalba, G. (2019). Predicción de la calidad del aire de Madrid mediante modelos supervisados. Recuperado de: <http://openaccess.uoc.edu/webapps/o2/bitstream/10609/99446/7/gabvilpiTFM0619memoria.pdf>
- [12] Bellinger, C., Shazan, M., Zaïane, O., y Osornio, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. Recuperado de: <https://pubmed.ncbi.nlm.nih.gov/29179711/>
- [13] Kok, I., Ulvi, M., y Ozdemir, S. (2017). A deep learning model for air quality prediction in smart cities. Recuperado de: [https://www.researchgate.net/publication/322512343\\_A\\_deep\\_learning\\_model\\_for\\_air\\_quality\\_prediction\\_in\\_smart\\_cities](https://www.researchgate.net/publication/322512343_A_deep_learning_model_for_air_quality_prediction_in_smart_cities)
- [14] Li, X., Peng, L., Hu, Y., Shao, J., y Chi, T. (2016). Deep learning architecture for air quality predictions. Recuperado de: [https://www.researchgate.net/publication/309096826\\_Deep\\_learning\\_architecture\\_for\\_air\\_quality\\_predictions](https://www.researchgate.net/publication/309096826_Deep_learning_architecture_for_air_quality_predictions)
- [15] Cerna, D., Ramírez, C., Diaz, A., Mosiño, J., Casillas, M., Baltazar, R., y Mendez, G. (2017). Red neuronal Backpropagation para la predicción de datos de contaminación y prevención de ataques a personas con padecimientos de rinitis alérgica y asma. Recuperado de: [https://www.researchgate.net/publication/339204231\\_Red\\_neuronal\\_Backpropagation\\_para\\_la\\_prediccion\\_de\\_datos\\_de\\_contaminacion\\_y\\_prevencion\\_de\\_ataques\\_a\\_personas\\_con\\_padecimientos\\_de\\_rinitis\\_alergica\\_y\\_asma](https://www.researchgate.net/publication/339204231_Red_neuronal_Backpropagation_para_la_prediccion_de_datos_de_contaminacion_y_prevencion_de_ataques_a_personas_con_padecimientos_de_rinitis_alergica_y_asma)
- [16] Datos abiertos del Área Metropolitana. Recuperado de: <https://datosabiertos.metropol.gov.co/>
- [17] SIATA (Sistema de alerta temprana del Valle de Aburrá y Medellín). Recuperado de: [https://siata.gov.co/siata\\_nuevo/](https://siata.gov.co/siata_nuevo/)
- [18] (2018). Más del 90% de los niños del mundo respiran aire tóxico a diario. Organización mundial de la salud. Recuperado de: <https://www.who.int/es/news-room/detail/29-10-2018-more-than-90-of-the-world's-children-breathe-toxic-air-every-day>
- [19] (2019). Índice de calidad del aire: cómo medir la calidad del aire residencial. El Blog de la ventilación eficiente. Recuperado de: <https://www.solerpalau.com/es-es/blog/indice-calidad-aire/>
- [20] ¿Qué es el ICA? Índice de calidad del aire. Área Metropolitana del Valle de Aburrá. Recuperado de: <https://www.metropol.gov.co/ambiental/calidad-del-aire/Paginas/Generalidades/ICA.aspx>

[21] Ozono troposférico. El mosaico de América del Norte: panorama de los problemas ambientales más relevantes. Recuperado de: <http://www3.cec.org/islandora/es/item/986-north-american-mosaic-overview-key-environmental-issues-es.pdf>