Modelos de aprendizaje profundo para la identificación del impacto del COVID 19 en la calidad del aire en Colombia

Karen Lisbeth Núñez Silva y Silvia Juliana Flórez Guerrero

Trabajo de Grado para optar por el título de Ingeniero Industrial

Director

Henry Lamos Díaz

PhD. en Física-Matemática

Universidad Industrial de Santander

Facultad de Ingeniería Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga, Santander

2023

Dedicatorias

A mis padres Melisa y Jorge quienes con su paciencia y apoyo hicieron posible la realización de mis estudios. A mi hermano Santiago por todas las risas y momentos compartidos. A mis abuelos maternos y paternos quienes con sus oraciones me ayudan a encontrar tranquilidad. A mi compañera Silvia, por elegirme para la realización de este proyecto y acompañarnos en cada momento vivido. A mi familia quienes aplazaron muchas veces las ganas de celebrar mi grado. A mis amigas y amigos, quienes fueron testigo de este proceso y me dieron las palabras de aliento necesarias cuando no se veía la luz. A todas las personas que hacen parte de mi vida y también a quienes he encontrado en el camino, pues me llevo lo más bonito de todos.

-Karen Lisbeth Núñez Silva.

A la vida por darme la oportunidad de crecer profesionalmente

A mi madre Martha por ser ejemplo de perseverancia y luchar incansablemente para que sus hijas tuviéramos un mejor futuro

A mis abuelos y demás familiares por acompañarme en cada momento y brindarme palabras de aliento

A mis compañeros, quienes siempre fueron apoyo incondicional en los buenos y malos momentos y quienes siempre me sacaron una sonrisa

A mi compañera Karen Núñez con la que compartimos lágrimas y risas durante el desarrollo de este proyecto y quien fue mi soporte para culminar con éxito mi carrera profesional.

-Silvia Juliana Flórez Guerrero

Agradecimientos

Agradecemos al profesor Henry por su paciencia y apoyo en la realización de este proyecto de grado. También, a nuestros docentes que demostraron amor por su vocación.

Contenido

Introducción 15

1. Generalidades del proyecto 17

1.1. Planteamiento del problema 17

1.2. Justificación del problema 18

2. Objetivos 21

2.1. Objetivo general 21

2.2. Objetivos específicos 21

3. Revisión de literatura 21

3.1. Análisis bibliométrico 21

3.2. Análisis preliminar de la literatura 25

4. Marco de referencia 28

4.1. Marco de antecedentes 28

4.2. Marco teórico 29

4.2.1. COVID-19 29

4.2.2. Índice de la calidad del aire 30

4.2.3. Variables meteorológicas 32

4.2.4. Contaminantes atmosféricos 33

4.2.5. Big Data 35

4.2.6. Aprendizaje Profundo 36

4.2.7. Redes Neuronales Artificiales 36

4.2.8. Sequence to Sequence (Seq2Seq) 39

5. Metodología 40

5.1. Etapa de selección 41

5.2. Etapa de preprocesamiento 63

5.2.1. Dirección del viento. 69

5.2.2. Humedad Relativa del aire. 70

5.2.3. Ozono Troposférico 72

5.2.4. Material Particulado 2.5. 73

5.2.5. Temperatura del Aire. 75

5.2.6. Velocidad del Viento. 76

5.2.7. Dióxido de Nitrógeno 77

5.3. Etapa de procesamiento 78

5.3.1. Modelo MLP 79

5.3.2. Modelo LSTM 80

5.3.3. Modelo Seq2Seq 81

5.4. Etapa de minería de datos 83

5.4.1. Modelo MLP 83

5.4.2. Modelo LSTM 89

5.4.3. Modelo Seq2Seq 95

5.5. Etapa de interpretación 101

6. Resultados y análisis 103

7. Divulgación de conocimiento 114

8. Conclusiones 115

9. Recomendaciones 116

Referencias Bibliográficas 118

Lista de Figuras

Figura 1 Ecuación de búsqueda. 21

Figura 2 Mapeo de palabras clave. 22

Figura 3 Publicaciones por año. 23

Figura 4 Publicaciones de revistas por año. 24

Figura 5 Artículos por autor. 24

Figura 6 Artículos publicados por país. 25

Figura 7 Red Neuronal Artificial con múltiples capas de neuronas. 37

Figura 8 Bucle cerrado de una RNN. 38

Figura 9 Metodología KDD. 40

Figura 10 Estaciones para la Ciudad de Bucaramanga. 43

Figura 11 Distribución de las variables para la ciudad de Bucaramanga. 43

Figura 12 Comportamiento Dirección del Viento en Bucaramanga. 45

Figura 13 Comportamiento Humedad del Aire en Bucaramanga. 46

Figura 14 Comportamiento Ozono Troposférico en Bucaramanga. 47

Figura 15 Comportamiento Material Particulado 2.5 en Bucaramanga. 48

Figura 16 Comportamiento Temperatura del Aire en Bucaramanga. 48

Figura 17 Comportamiento Velocidad del Viento en Bucaramanga. 49

Figura 18 Estaciones para la ciudad de Bogotá. 50

Figura 19 Distribución de las variables para la ciudad de Bogotá. 50

Figura 20 Comportamiento Dirección del Viento en Bogotá. 51

Figura 21 Comportamiento Humedad del Aire en Bogotá. 52

Figura 22 Comportamiento Dióxido de Nitrógeno en Bogotá. 53

Figura 23 Comportamiento Ozono Troposférico en Bogotá. 53

Figura 24 Comportamiento Presión en Bogotá. 54

Figura 25 Comportamiento Material Particulado 10 en Bogotá. 55

Figura 26 Comportamiento Material Particulado 2.5 en Bogotá. 55

Figura 27 Comportamiento Temperatura del Aire en Bogotá. 56

Figura 28 Comportamiento Velocidad del Viento en Bogotá. 57

Figura 29 Estaciones para la Ciudad de Cali. 57

Figura 30 Distribución de las variables para la ciudad de Cali. 58

Figura 31 Comportamiento Dirección del Viento en Cali. 59

Figura 32 Comportamiento Humedad del Aire en Cali. 59

Figura 33 Comportamiento Dióxido de Nitrógeno en Cali. 60

Figura 34 Comportamiento Ozono Troposférico en Cali. 61

Figura 35 Comportamiento Material Particulado 2.5 en Cali. 61

Figura 36 Comportamiento Temperatura del Aire en Cali. 62

Figura 37 Comportamiento Velocidad del Viento en Cali. 63

Figura 38 Diagrama de Cajas y Bigotes para DV. 69

Figura 39 Comportamiento DV para las 3 ciudades. 70

Figura 40 Diagrama de cajas y bigotes para HR. 71

Figura 41 Comportamiento HR para las 3 ciudades. 72

Figura 42 Diagrama de cajas y bigotes para O3. 73

Figura 43 Comportamiento O3 para las 3 ciudades. 73

Figura 44 Diagrama de cajas y bigotes para PM2.5. 74

Figura 45 Comportamiento PM2.5 para las 3 ciudades. 75

Figura 46 Diagrama de cajas y bigotes para TA. 75

Figura 47 Comportamiento TA para las 3 ciudades 76

Figura 48 Diagrama de cajas y bigotes para VV 76

Figura 49 Comportamiento VV para las 3 ciudades. 77

Figura 50 Diagrama de cajas y bigotes para NO2. 78

Figura 51 Comportamiento NO2 para Bogotá. 78

Figura 52 Arquitectura del modelo MLP. 80

Figura 53 Arquitectura del modelo LSTM. 81

Figura 54 Arquitectura del modelo Seq2Seq. 82

Figura 55 Función de pérdida modelo MLP para Bogotá. 83

Figura 56 Predicción de los contaminantes modelo MLP para Bogotá. 84

Figura 57 Función de precisión modelo MLP para Bogotá. 85

Figura 58 Función de pérdida modelo MLP para Bucaramanga. 85

Figura 59 Predicción de los contaminantes modelo MLP para Bucaramanga. 86

Figura 60 Función de precisión modelo MLP para Bucaramanga. 87

Figura 61 Función de pérdida modelo MLP para Cali. 87

Figura 62 Predicción de los contaminantes modelo MLP para Cali. 88

Figura 63 Función de precisión modelo MLP para Cali. 89

Figura 64 Función de pérdida modelo LSTM para Bogotá. 90

Figura 65 Predicción de los contaminantes modelo LSTM para Bogotá. 90

Figura 66 Función de precisión modelo LSTM para Bogotá. 91

Figura 67 Función de pérdida modelo LSTM para Bucaramanga. 92

Figura 68 Predicción de los contaminantes modelo LSTM para Bucaramanga. 92

Figura 69 Función de precisión modelo LSTM para Bucaramanga. 93

Figura 70 Función de pérdida modelo LSTM para Cali. 94

Figura 71 Predicción de los contaminantes modelo LSTM para Cali. 94

Figura 72 Función de precisión modelo LSTM para Cali. 95

Figura 73 Función de pérdida modelo Seq2Seq para Bogotá. 96

Figura 74 Predicción de los contaminantes modelo Seq2Seq para Bogotá. 96

Figura 75 Función de precisión modelo Seq2Seq para Bogotá. 97

Figura 76 Función de pérdida modelo Seq2Seq para Bucaramanga. 98

Figura 77 Predicción de los contaminantes modelo Seq2Seq para Bucaramanga. 98

Figura 78 Función de precisión modelo Seq2Seq para Bucaramanga. 99

Figura 79 Función de pérdida modelo Seq2Seq para Cali 100

Figura 80 Predicción de los contaminantes modelo Seq2Seq para Cali. 100

Figura 81 Función de precisión modelo Seq2Seq para Cali. 101

Figura 82 Relación de métricas de validación. 102

Figura 83 Boxplot escenarios de los contaminantes para la ciudad de Bogotá. 106

Figura 84 Boxplot escenarios de los contaminantes para la ciudad de Bucaramanga. 110

Figura 85 Boxplot escenarios de los contaminantes para la ciudad de Cali. 112

Lista de Tablas

Tabla 1 Cumplimiento de objetivos……..16

Tabla 2 Recursos girados a las instituciones de salud por el COVID-19…….20

Tabla 3 Síntesis de la revisión de literatura……..26

Tabla 4 Descripción general del ICA……....31

Tabla 5 Puntos de corte del ICA……...32

Tabla 6 Nomenclatura de las variables utilizadas…….42

Tabla 7 Cantidad de datos encontrados en la base……64

Tabla 8 Cantidad de datos eliminados……..65

Tabla 9 Datos no válidos……...66

Tabla 10 Datos reemplazados por ciudad…….67

Tabla 11 Estrategias de imputación implementadas…….67

Tabla 12 Estadísticas obtenidas para cada variable en cada ciudad……..68

Tabla 13 Relaciones métricas de validación……...103

Tabla 14 Valores máximos permitidos según la Resolución……..104

Tabla 15 Valores promedios y máximos de los datos preprocesados…….104

Tabla 16 Pruebas estadísticas para el contaminante NO2 en Bogotá…….108

Tabla 17 Pruebas post-hoc para los contaminantes de la ciudad de Bogotá……...108

Tabla 18 Pruebas estadísticas para los contaminantes en la ciudad de Bucaramanga……111

Tabla 19 Prueba Dunn-Bonferroni para las contaminantes de la ciudad de Bucaramanga111

Tabla 20 Pruebas estadísticas para los contaminantes en la ciudad de Cali……...113

Tabla 21 Prueba Dunn-Bonferroni para el contaminante PM2.5 en la ciudad de Cali…...114

Tabla 22 Resultados ANOVA y Kruskal-Wallis ……...114

Lista de apéndices

(Ver apéndice adjunto en la carpeta compartida)

Apéndice A. Artículo de carácter publicable.

RESUMEN

TITULO: Modelos de Aprendizaje Profundo para la identificación del impacto del Covid 19 en la Calidad del Aire en Colombia*

AUTORES:

Núñez Silva, Karen Lisbeth**

Flórez Guerrero, Silvia Juliana**

PALABRAS CLAVE: Calidad del Aire, Covid 19, Deep Learning, Redes Neuronales.

DESCRIPCIÓN:

La contaminación del aire y su impacto en la salud ha sido una problemática mundial, exponerse a los distintos contaminantes atmosféricos ha causado y causa millones de muertes en todo el planeta, con la llegada del COVID-19 y las medidas de confinamiento que trajo consigo, se creería que la calidad del aire mejoró notablemente en la mayor parte de las regiones. Este proyecto de grado propone analizar los índices de Material Particulado (PM2.5), Dióxido de Nitrógeno (NO2) y Ozono Troposférico (O3) antes y durante la pandemia para las ciudades de Cali, Bucaramanga y Bogotá, utilizando modelos de Aprendizaje Profundo para predecir su comportamiento y que con los resultados obtenidos, las autoridades puedan diseñar planes de acción para disminuir estos contaminantes atmosféricos y mejorar la calidad de vida, se eligieron estos modelos Deep Learning ya que tienen la capacidad de procesar grandes volúmenes de datos y modelar patrones no lineales, lo que los hace más precisos que otros métodos estadísticos.

---

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director PhD Henry Lamos Díaz.

ABSTRACT

TITLE: Deep Learning models for identifying the impact of Covid 19 on Air Quality in Colombia*

AUTHORS:

Núñez Silva, Karen Lisbeth**

Flórez Guerrero, Silvia Juliana**

KEYWORDS: Covid 19, Air Quality, Deep Learning, Neural Networks.

DESCRIPTION:

Air pollution and its impact on health has been a worldwide problem, exposure to different atmospheric pollutants has caused and causes millions of deaths around the planet, with the arrival of COVID-19 and the confinement measures it brought with it, it would be believed that air quality improved significantly in most regions. This degree project proposes to analyze the indices of Particulate Matter (PM2. 5), Nitrogen Dioxide (NO2) and Tropospheric Ozone (O3) before and during the pandemic for the cities of Cali, Bucaramanga and Bogota, using Deep Learning models to predict their behavior and with the results obtained, the authorities can design action plans to reduce these air pollutants and improve the quality of life, these Deep Learning models were chosen because they have the ability to process large volumes of data and model nonlinear patterns, which makes them more accurate than other statistical methods.

* Degree Project

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director PhD Henry Lamos Díaz.

Introducción

Según la Organización Mundial de la Salud (OMS, 2016), aproximadamente 3 millones de muertes al año, a nivel mundial, son causadas por la exposición a la contaminación del aire; en 2012 se registraron 6,5 millones de muertes relacionadas a esta problemática.

En Colombia, Bogotá es una de las ciudades con mayores problemas debido a la calidad del aire, causado principalmente por la contaminación por partículas (generado por el sector industrial y de transporte como el hollín, el polvo y el humo), siendo el suroccidente una de las zonas más contaminadas de todo el país por este factor.

Con la llegada de la enfermedad infecciosa del virus SARS-CoV-2 más conocida como el COVID-19, tomó mayor importancia la salud, ya que este factor se vio seriamente vulnerado con los contagios masivos, ocasionando miles de muertes alrededor del mundo. Con el fin de mitigar el impacto de propagación de este virus se propusieron medidas de confinamiento que pausaron las actividades de los seres humanos, como el transporte, las compras, el trabajo, entre otras. Se ha estimado que estas medidas han causado una mejora de la calidad del aire, evidenciando en algunos países la disminución notoria de ciertos contaminantes atmosféricos durante los tiempos de cierres.

Considerando lo anterior, en este proyecto de grado se analizarán antes y durante la pandemia los índices de los siguientes contaminantes atmosféricos con el propósito de determinar su comportamiento: Material Particulado (PM2.5) que se define como una mezcla de partículas sólidas y líquidas encontradas en el aire; Dióxido de Nitrógeno (NO2) formado como subproducto en los procesos de combustión a altas temperaturas como en los vehículos motorizados y las plantas eléctricas y, Ozono Troposférico u ozono superficial (O3)

contaminante secundario originado a partir de reacciones que se activan por la luz solar entre contaminantes primarios como los óxidos de nitrógeno. De igual manera, se pretende aplicar 3 modelos de Aprendizaje Profundo para analizar el comportamiento de las variables escogiendo el de mayor rendimiento, con el fin de apoyar a las autoridades de salud y ambientales a diseñar políticas que contribuyan con la disminución de estos contaminantes atmosféricos, y de esta manera, mejorar la calidad de vida. El modelo de predicción a elegir se basa en los modelos de Aprendizaje Profundo, específicamente en las redes neuronales MLP, LSTM y Seq2Seq, las cuales tienen la capacidad de procesar grandes volúmenes de datos.

Tabla de cumplimiento de objetivos

Tabla 1

Cumplimiento de objetivos

Objetivo	Cumplimiento
Realizar una revisión de literatura orientada al problema de la calidad del aire y el uso de modelos de Aprendizaje Profundo como método de análisis de datos.	Capítulo 3
Aplicar modelos de Aprendizaje Profundo para la predicción de la calidad del aire.	Capítulo 5
Comparar el comportamiento de la calidad del aire antes y durante la pandemia provocada por el COVID 19.	Capítulo 6
Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.	Apéndice A

1. Generalidades del proyecto

1.1. Planteamiento del problema

Existe amplia relación entre la calidad del aire y los efectos graves en la salud, a nivel mundial, 7 millones de muertes prematuras fueron atribuibles a esta problemática en 2016 principalmente en países de bajos y medios ingresos, y se estima que más de 150 millones de personas en América latina viven en ciudades que exceden los índices de calidad del aire propuestos por la Organización Mundial de la Salud (OPS, 2018). Este hecho lo evidencia Michael Brauer en su estudio "How Much, How Long, What, and Where Air Pollution Exposure Assessment for Epidemiologic Studies of Respiratory Disease" (Brauer, 2010) aludiendo que la contaminación del aire es un factor de riesgo en las enfermedades respiratorias. Dicho esto, el cuidado medioambiental ha cobrado mayor importancia, y es por ello, que el gobierno colombiano ha implementado diferentes alternativas para disminuir la contaminación del aire como la disminución de la movilidad de vehículos particulares utilizando y motivando el uso del transporte público, caminar y montar en bicicleta; avanzar en la transición energética, reducir la deforestación, sembrar 180 millones de árboles, conservar la Amazonia y los páramos, entre otras (Grupo ENEL, 2021).

A finales del 2019, en Wuhan, una ciudad de China se empezó a propagar muy rápidamente la enfermedad producida por el coronavirus SARS-CoV-2, la cual ha impactado al mundo por su alto índice de transmisión, afectando en gran medida las vías respiratorias. Esto se puede observar en las altas cifras de contagio a nivel mundial reportadas a 6 de enero de 2022 en países como Estados Unidos (58.473.822 casos confirmados), India (35.226.386 casos confirmados) y Brasil (22.328.252 casos confirmados), los cuales presentan la mayor afectación. Por su parte Colombia, se encuentra ubicado en décimo tercer lugar a nivel

mundial en términos de contagio con 5.242.672 de casos positivos (Our World in Data, 2021). Así, desde su llegada al país, en marzo del 2020, se aceleraron las restricciones de movilidad mencionadas anteriormente con el fin de evitar las aglomeraciones en espacios frecuentados, ya que principalmente se da por el contacto de gotículas respiratorias, causando así la disminución en la actividad de personas y de transporte en las calles. A raíz de esto, la calidad del aire se vio beneficiada al reducirse el dióxido de carbono (CO2), el monóxido de carbono (CO), el óxido de nitrógeno (NOX), entre otras sustancias contaminantes que la afectan (Osso, 2020).

De esta manera, para el desarrollo de este trabajo de investigación se analizará información de la calidad del aire suministrada por las bases de datos del IDEAM de tres de las principales ciudades del país: Bucaramanga (para efectos del presente trabajo de grado se hará referencia al Área Metropolitana de Bucaramanga), Bogotá y Cali, mediante modelos de Aprendizaje Profundo, con el fin de comprobar la hipótesis de que las restricciones a causa del COVID-19 disminuyeron significativamente los niveles de contaminación en Colombia.

Como se profundizará más adelante en el marco de antecedentes, se ha encontrado que este tipo de investigación se ha realizado en Colombia, tomando como muestra Bucaramanga y enfocado al área de logística urbana, utilizando los modelos de Aprendizaje Profundo LSTM, Seq2Seq y SVM, sin embargo, se resalta que en Colombia hay muy poca información actualizada sobre este tema y lo que se ha hallado difiere con los objetivos y el alcance de este proyecto.

1.2. Justificación del problema

“La transformación digital apoyada por Big Data requiere un cambio en la mentalidad de los gerentes, y debe contar con más profesionales enfocados en la innovación y en la solución de problemas que se apoyen en paquetes de software de diversos tipos: simulación,

analítica y estadísticos. Necesitamos adaptar a ese ingeniero tradicional al mundo de hoy", (Unisabana, 2020) dice Gonzalo Mejía, profesor de la facultad de ingeniería de la Universidad de la Sabana.

Desde la ingeniería industrial se utilizan conocimientos avanzados en estadística enfocados en el análisis de datos por medio de herramientas tecnológicas, siendo la ingeniería donde se busca optimizar o manejar los recursos adecuadamente, disminuyendo costos.

En este proyecto de investigación se plasma la problemática de la contaminación del aire en Colombia, la cual, por su baja calidad hace que anualmente el país asuma costos por enfermedades y servicios que ascienden a más de $12 billones y ocasionando aproximadamente 8.000 muertes al año según el informe "Calidad del Aire: Una Prioridad de Política Pública en Colombia" (Departamento Nacional de Planeación, 2018).

Adicionalmente, se enfoca en una pandemia que provocó una emergencia sanitaria, ocasionando dificultades para manejar y distribuir los recursos en la población, como no existían estrategias para manejar eficientemente dichos recursos, los costos en salud aumentaron significativamente.

Según la Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES, 2022), en el año 2021 el Gobierno Nacional giró los recursos mostrados en la Tabla 2 asociados al COVID 19 que inició en marzo del 2020.

Tabla 2

Recursos girados a las instituciones de salud por el COVID-19.

Descripción	Cifra en billones de pesos
Pago del seguro de salud regímenes contributivo y subsidiado	$ 54,4400
Servicios médicos que no están en el PBS	$ 6,0300
Bonificación al personal de salud	$ 0,4096
Canastas COVID 19	$ 1,8200
Pruebas COVID	$ 1,6300
Sostenimiento y mantenimiento camas UCI	$ 0,5711
Apoyo hogares subsidiados contagiados por COVID	$ 0,0767
Residentes médicos	$ 0,1753
TOTAL	$ 65,1527

Por tal razón, se espera que los resultados de este proyecto de investigación sean de gran ayuda para que las entidades pertinentes establezcan estrategias para la mejora de la calidad del aire con el fin de minimizar las enfermedades causadas por esta problemática, ocasionando a su vez una disminución de los costos asociados.

2. Objetivos

2.1. Objetivo general

Desarrollar modelos de aprendizaje profundo para la identificación del impacto del COVID 19 en la calidad del aire en Colombia.

2.2. Objetivos específicos

Realizar una revisión de literatura orientada al problema de la calidad del aire y el uso de modelos de Aprendizaje Profundo como método de análisis de datos.

Aplicar modelos de Aprendizaje Profundo para la predicción de la calidad del aire.

Comparar el comportamiento de la calidad del aire antes y durante la pandemia provocada por el COVID 19.

Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.

3. Revisión de literatura

3.1. Análisis bibliométrico

Con el fin de analizar el impacto en la calidad del aire a causa de las restricciones debidas a la pandemia provocada por el COVID 19 desde el 2020 al 2021, se ha desarrollado la siguiente ecuación de búsqueda Figura 1.

Figura 1

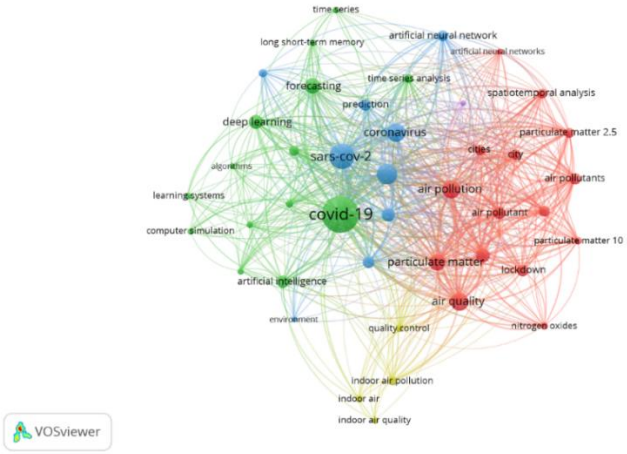
Ecuación de búsqueda.

Scopus: (“Deep Learning" OR "neural network”) AND ("air quality" OR "air pollution" OR "environmental quality") AND (covid$19 OR sars$cov$ OR coronavirus)

Se utilizan las bases de datos disponibles en la plataforma de la Universidad Industrial de Santander (UIS). Dentro de ellas se elige la base de datos Scopus, que permite hacer una búsqueda multidisciplinaria, analizando los problemas que han sido solucionados mediante la optimización de hiperparámetros en otro tipo de investigaciones. Inicialmente, se obtuvieron 322 artículos, cuyos datos fueron insertados en el software VosViewer mediante el cual se hizo un análisis de palabras clave con ocurrencia mínima de 5 menciones, seleccionando las de mayor concordancia con la ecuación de búsqueda como se muestra en la Figura 2 .

Figura 2

Mapeo de palabras clave.



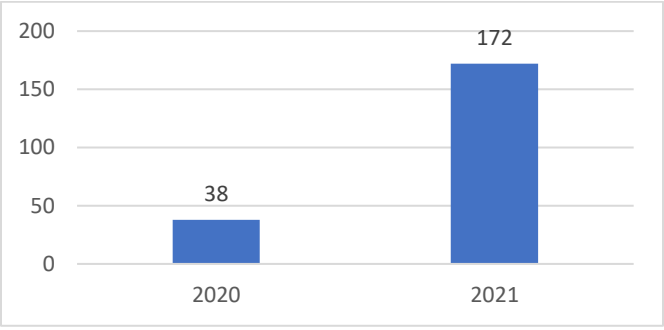
Teniendo en cuenta el análisis hecho anteriormente se procede a filtrar la ecuación de búsqueda ejecutada en Scopus por las palabras claves: "COVID-19", "SARS-CoV-2", "Coronavirus Disease 2019", "Air Quality", "Coronavirus", "Air Pollution", "Particulate Matter", "Forecasting", "Atmospheric Pollution", "Artificial Intelligence", "SARS Coronavirus", "Deep Learning", "Controlled Study", "Air Pollutant", "Prediction", "Lockdown", "Air Pollutants", "Artificial Neural Network", "Cities", "City", "Nitrogen

Dioxide", "Indoor Air Pollution", "Quality Control", "Particulate Matter 2.5", "Covid-19", "Statistical Model", "Nitrogen Oxides", "Predictive Analytics", "Indoor Air Quality", "Learning Systems", "Mathematical Model", "Neural Networks", "Quarantine", "Time Series Analysis", "Long Short-term Memory", "Particulate Matter 10", "SARS", "Simulation", "Time Series", "Artificial Neural Networks", "Carbon Dioxide", "Environment", "pm2.5", "Indoor Air", "air pollution control", "computer simulation", "spatio$temporal analysis", "algorithms", "Indoor Environment" y limitado también por "artículos", emitiendo 210 documentos académicos.

De estos 210 artículos, en la Figura 3, se observan las publicaciones realizadas durante el 2020 y 2021, presentando un 18,1% y un 81,9% respectivamente, siendo el 2021 el año en que más artículos respecto a la temática se publicaron.

Figura 3

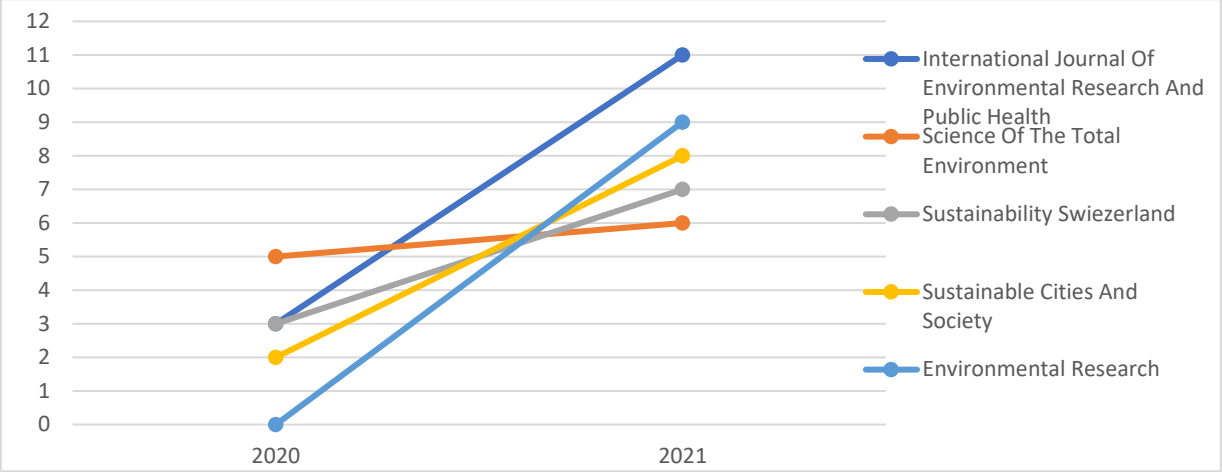
Publicaciones por año.



Considerando las revistas académicas que más tuvieron relevancia en estos dos años para el tema tratado, en la Figura 4 se muestran las cinco con más documentos publicados, resaltando la revista "*International Journal Of Environmental Research And Public Health*" con 14 publicaciones en total, seguida de la revista "*Science Of The Total Environment*" con 11 artículos.

Figura 4

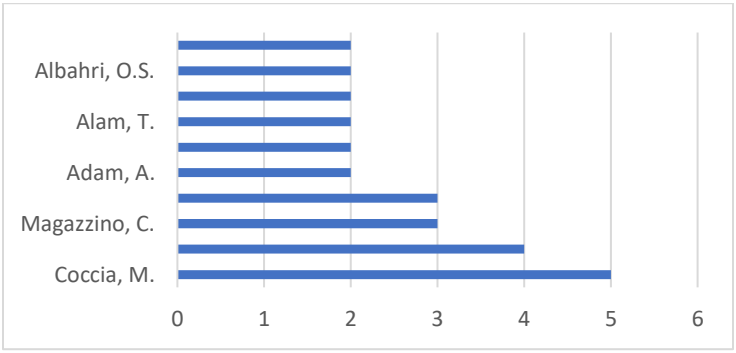
Publicaciones de revistas por año.



Ahora, analizando los 10 autores principales con sus producciones científicas, se puede evidenciar en la Figura 5 que el mayor número de artículos realizados fue por Coccia, M., con un total de 5, seguido por Cao, Magazzino y Mele con 4, 3 y 3 respectivamente. Asimismo, se observa que el mínimo de publicaciones es de 2, siendo este la moda en los 10 autores.

Figura 5

Artículos por autor.

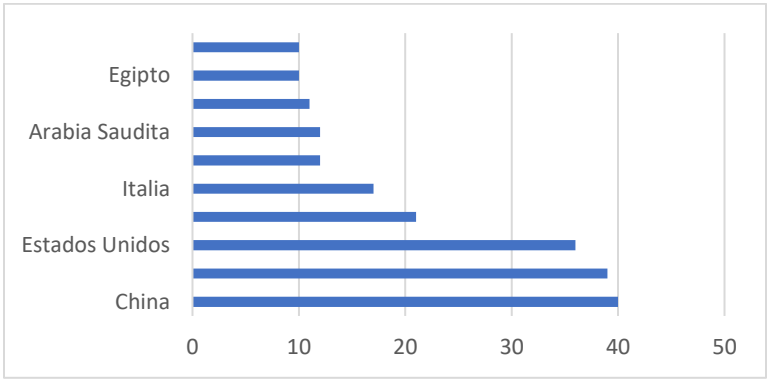


Adicionalmente, con el fin de conocer los países que han tenido mayor participación

en la investigación de este tema, se realiza la Figura 6, en la cual se observa que los países donde más se efectuaron publicaciones son China, India, Estados Unidos, Reino Unido e Italia con 19%, 18,5%, 17,1%, 10% y 8% de artículos publicados respectivamente. Es importante resaltar que en Colombia no se encuentra documentación acerca de este tema o la poca información que se puede recolectar es de más de 5 años atrás, perdiendo relevancia en esta investigación.

Figura 6

Artículos publicados por país.



Finalmente, realizando una depuración a los 210 documentos encontrados, se redujo la cantidad por medio de la lectura de los títulos dando como resultado 30 de estos, de los cuales, con base en sus resúmenes se eligieron finalmente 12 artículos para la investigación.

3.2. Análisis preliminar de la literatura

El análisis de la literatura permite afirmar que los modelos más utilizados en los artículos son aquellos que implican redes neuronales artificiales como MLP, LSTM, ELM, ESN, RBF ya que son capaces de proporcionar mayor precisión en la predicción de los parámetros de calidad del aire. En la Tabla 3 se muestra la síntesis de la revisión de literatura donde se destacan las características de los modelos utilizados en cada artículo.

Tabla 3

Síntesis de la revisión de literatura.

Citación	Modelo	Observaciones
(Zhao, y otros, 2021)	Método CVAE (Codificador automático variacional condicional)	El método se basa en Deep Learning y tiene como objetivo identificar y evaluar de manera más realista los cambios anómalos y abruptos en las PM2.5, así proporciona un nuevo enfoque en la distribución de las fuentes contaminantes, además de detectar el impacto de las condiciones meteorológicas y actividades humanas en las anomalías de los contaminantes del aire y gases de efecto invernadero.
(Ekinci, Omurca, & Ozbay, 2021)	Modelos LSTM (Long-Short Term Memory), BILSTM (Bidirectional Long-Short Term Memory), STACKED LSTM, CNN LSTM Y CONV LSTM	Los modelos Deep Learning proporcionan una precisión adecuada en la resolución de problemas complejos, por lo tanto, usar un modelo de estos puede considerarse como una herramienta prometedora para predecir las concentraciones de O3 a nivel del suelo. El modelo Stacked LSTM tiene la capacidad de modelado más potente para clasificar datos.
(Etchie, Etchie, Jauro, Pinker, & Swaminathan, 2021)	MLPNN	Se utilizan redes neuronales para derivar el promedio mensual a nivel del suelo de los aerosoles.
(Tadano, y otros, 2020)	ELM (Extreme Learning Machine), ESN (Echo State Network), MLP (Multilayer Perceptron), RBF (The Radial Basis Function	Las redes neuronales artificiales demostraron ser herramientas de predicción robustas para estimar el mejor equilibrio entre los

Continuación Tabla 3

Síntesis de la revisión de literatura.

	Networks)	**casos de COVID-19, el porcentaje de cierre y el nivel de contaminantes atmosféricos.**
(Shatnawi & Abu-Qdais, 2021)	RNA	Una RNA adecuadamente entrenada y estructurada puede ser una herramienta útil para predecir los parámetros de calidad del aire con una precisión adecuada.

De igual manera, con la información recopilada se hizo evidente que la escogencia de las variables es de vital importancia para desarrollar correctamente cualquier modelo, por ello, se fraccionan usualmente en dos conjuntos: entradas como las variables meteorológicas (Temperatura máxima, humedad relativa, presión atmosférica, velocidad y dirección del viento) y salidas como las concentraciones diarias de cada contaminante atmosférico (CO [ppm], O3 [mg/m3], NO2 [mg/m3], NO [mg/m3], PM2.5 [mg/m3], y PM10 [mg/m3]).

Otro aspecto clave para mejorar la precisión de cada modelo (Tadano, y otros, 2020) es segmentar el conjunto de datos en tres momentos:

- Entrenamiento: para ajustar los modelos.
- Validación: con el fin de verificar el sobreentrenamiento y definir el número de neuronas en la capa intermedia.
- Prueba: se utiliza para evaluar el rendimiento de los modelos mediante el error cuadrático medio (MSE por sus siglas en inglés).

El presente trabajo de investigación se enfoca en utilizar los modelos de redes neuronales artificiales MLP, LSTM y Seq2Seq, para la comparación y predicción de la

calidad del aire analizando los contaminantes NO2, PM2.5 y O3.

4. Marco de referencia

4.1. Marco de antecedentes

Juan Carlos Álvaro Huanaco y Maira Cecilia Valdivia Valencia en su trabajo de pregrado, "Revisión sistemática de metodologías de análisis de índices de calidad del aire, para determinar el grado de contaminación 2020", hacen alusión a las complicaciones que representa la exposición al material particulado ya que puede ser el causante de afecciones en la salud, empezando por el daño en el ADN. En este trabajo no se encuentra una aplicación de modelos Deep Learning o alguna referencia a las consecuencias causadas por el COVID-19 en el medio ambiente, pero, se destaca el análisis de los índices de calidad del aire (ICA) respecto a las metodologías para medirlos y establecer los niveles adecuados en los que se deben encontrar para evitar riesgos en la salud, aunque no menciona o hace uso de un modelo Deep Learning también se enfatiza su búsqueda de actividades que más alteran estos parámetros, dando como resultado la industrialización, el crecimiento urbanístico, condiciones meteorológicas y fuentes contaminantes.

Alejandro Aurelio Rodríguez Miranda en su tesis doctoral titulada: "Modelización y análisis de la calidad del aire en la ciudad de Oviedo, mediante los enfoques PSO-SVM, Red Neuronal MLP y Árbol de regresión M5", manifiesta la necesidad del desarrollo de técnicas alternativas de diagnóstico como las mencionadas en el título, ya que algunas técnicas tradicionales que se usan actualmente como la monitorización de contaminantes a través de estaciones automáticas, son costosas. Además, al implementar el modelo híbrido PSO-SVM en las ciudades (teniendo en cuenta las especificaciones de cada lugar) ocasionaría que se

mitiguen ciertos problemas de contaminación y a la vez ayudaría a entender el comportamiento de los contaminantes. Aunque no está relacionada con la problemática del COVID-19 o enfocada a la realidad de Colombia, es muy útil para este trabajo de investigación pues en dicha tesis doctoral se puede evidenciar una solución práctica con datos reales sobre la calidad del aire, lo que lo convierte en una buena base al poner en marcha el modelo MLP, el cual se decide utilizar en el presente trabajo de investigación.

Paula Andrea Abril Ortiz y Edgar Leonardo Porras Ojeda en su trabajo de pregrado, "Modelos Deep Learning en Logística Urbana para la predicción de la calidad del aire en la ciudad de Bucaramanga", en su revisión de literatura exponen que los algoritmos más utilizados para la predicción de la calidad del aire son las redes neuronales y sus variaciones, más específicamente, las redes neuronales recurrentes ya que presentan mayor rendimiento en el modelado de estructuras temporales. Por ello, en el presente trabajo de investigación se consideró la implementación de redes neuronales LSTM y seq2seq conforme a lo presentado por los autores.

4.2. Marco teórico

4.2.1. COVID-19

La Organización Mundial de la Salud (OMS , 2020) afirma que el COVID-19 es la enfermedad causada por el nuevo coronavirus conocido como SARS-CoV-2. El 31 de diciembre del 2019 se dio a conocer por primera vez un caso de neumonía vírica en Wuhan, China, cuya propagación avanzó de manera rápida provocando que esta enfermedad se convirtiera en una pandemia, la cual trajo consigo innumerables aspectos negativos que afectan a la población en temas de salubridad, política, medio ambiente, educación, alimentación y economía. El 9 de diciembre del 2021, la base de datos *Our World in Data* (Our World in Data, 2021) presenta 34,010.86 millones de casos positivos con COVID-19 y

671.48 millones de muertes provocadas por esta enfermedad alrededor del mundo.

Los síntomas que presenta esta enfermedad comúnmente son fiebre, tos seca y cansancio y algunos menos frecuentes son pérdida del olfato, congestión nasal, dolor de garganta, dolor de cabeza, entre otros. Según la OMS (2020), las personas que corren mayor riesgo de presentar un cuadro grave son aquellas mayores de 60 años y las que tienen comorbilidades como hipertensión arterial, obesidad o cáncer y problemas cardíacos o pulmonares.

De acuerdo con lo anterior, esta enfermedad es de alta propagación debido a sus síntomas y al no tener los cuidados necesarios como el no uso de mascarilla, lavado de manos, distanciamiento, entre otros, los gobernantes de cada país decidieron declarar estados de alerta proclamando aislamientos y cuarentenas desde marzo del 2020 hasta mediados del 2021 para controlar los contagios.

4.2.2. Índice de la calidad del aire

IDEAM (2021) afirma que:

El índice de la calidad del aire (ICA) es un valor adimensional asociado a un código de colores para reportar el estado de la calidad del aire al que están asociados unos efectos generales que deben ser tenidos en cuenta para reducir la exposición a altas concentraciones por parte de la población (p.2).

Este es calculado para los contaminantes O3, PM10, PM2.5, CO, SO2 y NO2 entre tiempos de 1 y 24 horas de acuerdo con los puntos de corte establecidos en la Resolución 2254 de 2017.

En la Tabla 4 se muestran los valores para los estados de la calidad del aire junto con los efectos que cada uno de estos puede producir.

Tabla 4

Descripción general del ICA.

Rango	Color	Estado	Efectos
0-50	Verde	Buena	La contaminación atmosférica supone un riesgo bajo para la salud.
51-100	Amarillo	Aceptable	Posibles síntomas respiratorios en grupos poblacionales sensibles.
101-150	Naranja	Dañina a la salud de grupos sensibles	Los grupos poblacionales sensibles pueden presentar efectos sobre la salud. 1) Ozono Troposférico: las personas con enfermedades pulmonares, niños, adultos mayores y las que constantemente realizan actividad física al aire libre, debe reducir su exposición a los contaminantes del aire. 2) Material particulado: Las personas con enfermedad cardiaca o pulmonar, los adultos mayores y los niños se consideran sensibles y por lo tanto en mayor riesgo.
151-200	Rojo	Dañina para la salud	Todos los individuos pueden comenzar a experimentar efectos sobre la salud. Los grupos sensibles pueden experimentar efectos más graves para la salud.
201-300	Púrpura	Muy dañina para la salud	Estado de alerta que significa que todos pueden experimentar efectos más graves para la salud.
301-500	Marrón	Peligrosa	Advertencia sanitaria. Toda la población puede presentar efectos adversos graves en la salud humana y están propensos a verse afectados por graves efectos sobre la salud.

Nota. Información tomada de (IDEAM, 2021).

En la Tabla 5 se describen los rangos, las categorías y los puntos de corte de cada contaminante, con el fin de informar a la comunidad lectora los rangos adecuados para no causar daños a la salud.

Tabla 5

Puntos de corte del ICA.

Rango ICA	Categoría	PM10 µg/m3 24 horas	PM2.5 µg/m3 24 horas	CO µg/m3 8 horas	SO2 µg/ m3 1 hora	NO2 µg/m3 1 hora	O3 µg/m3 8 horas	O3 µg/m 3 1 hora
0-50	Buena	0-54	0-12	0-5094	0-93	0-100	0-116	---
51-100	Aceptable	55-154	13-37	5095-10819	94-197	101-189	107-138	---
101-150	Dañina a la salud de grupos sensibles	155-254	38-55	10820-14254	198-486	190-677	139-167	245-323
151-200	Dañina a la salud	255-354	56-150	14255-17688	487-797	678-1221	168-207	324-401
201-300	Muy dañina a la salud	355-424	151-250	17689-34862	798-1583	1222-2349	208-393	402-794
301-500	Peligrosa	425-604	251-500	34863-57703	1584-2629	2350-3853	394	795-1185

Nota. Información tomada de (IDEAM, 2021).

4.2.3. Variables meteorológicas

A continuación, se definen las variables meteorológicas presentes en este estudio.

4.2.3.1. Humedad. Según la Red de Monitoreo de Calidad del Aire de Bogotá **(RMCAB, 2020)** la humedad es la cantidad de vapor de agua que se encuentra en el aire. Esta variable se puede expresar de dos formas: de manera absoluta y se encuentra como humedad absoluta o de forma relativa como humedad relativa/grado de humedad. Para este trabajo de investigación, se hablará de Humedad Relativa.

4.2.3.2. Temperatura ambiente. Es una medida del grado de calor o frío presente en el aire en un momento y lugar determinados (IDEAM, 2018).

4.2.3.3. Dirección del viento. Esta variable se define como la dirección desde la cual sopla el viento, puede ser expresada en grados a partir del norte geográfico (RMCAB, 2020).

4.2.3.4. Velocidad del viento. Es la distancia que recorre una partícula de aire en la unidad de tiempo, se expresa en metros por segundo (m/s), kilómetros por hora (km/h) o nudos. Cuando la velocidad del viento es inferior a 0.5 m/s se dice que el viento está en calma (IDEAM, 2018).

4.2.3.5. Presión atmosférica. Es la presión ejercida por el aire en cualquier punto de la atmósfera. Por lo general se refiere a la presión atmosférica terrestre, pero también se puede extender a la atmósfera de cualquier planeta o satélite (RMCAB, 2020).

4.2.4. Contaminantes atmosféricos

En este proyecto de investigación se analizan los siguientes contaminantes atmosféricos.

4.2.4.1. Material particulado (PM). Es una mezcla de partículas sólidas o líquidas de polvo, humo, cenizas, hollín de Diesel, partículas provenientes de procesos productivos, cemento o polen presentes en la atmósfera, producidas por fuentes naturales o antropogénicas[1] y contienen un amplio rango de propiedades morfológicas, físicas, químicas y termodinámicas. Existen 2 tipos: PM10 que son partículas inhalables con diámetro menor o igual a 10 micrómetros y PM2.5 que son partículas finas inhalables con un diámetro menor o igual a 2.5 micrómetros (EPA, 2021). Los efectos en la salud tras la exposición prolongada o repetitiva de este contaminante pueden ser altamente nocivos, ocasionando agudización de enfermedades cardiovasculares y del asma, cáncer pulmonar, síntomas respiratorios severos e irritación de ojos y nariz. Asimismo, influye de manera global en el medio ambiente con el cambio climático y de manera local, reduciendo la visibilidad en las ciudades (Santamaría, 2008).

4.2.4.2. Dióxido de Nitrógeno (NO2). Compuesto químico formado por Nitrógeno y Oxígeno, forma parte de un grupo de contaminantes gaseosos producto de algunos procesos como incendios forestales y combustión de motores de vehículos (Green Facts, 2006). Sus efectos en la salud son complicaciones en las vías respiratorias, también se le relaciona a enfermedades como autismo, ictus, fallos en el sistema cardiovascular, enfermedades renales y cáncer (Instituto para la salud Geoambiental, s.f.).

[1] Perteneciente o relativo a lo que precede de los seres humanos que, en particular, tiene efectos sobre la naturaleza.

4.2.4.3. Ozono troposférico (O3). Es un contaminante secundario compuesto de hidrocarburos, Dióxido de Nitrógeno (NO2), calor y luz solar. Es el compuesto más representativo de los oxidantes fotoquímicos e ingrediente primordial del smog[2] urbano. Las principales fuentes de emisión de este contaminante son los automóviles y la industria, provocando mayor polución durante el día a causa de que la luz solar desempeña un papel primordial en su formación. Los efectos adversos de la elevada concentración del O3 a la salud humana son negativos y muy significativos, variando de irritación de garganta y ojos, ataques de tos, jadeo, padecimiento de asma, dolores de pecho, hasta daños permanentes a los pulmones y alteraciones del sistema inmunológico. De igual manera, afecta al medio ambiente deteriorando las hojas de los árboles y plantas, reduciendo el rendimiento de cultivos y crecimiento de bosques. Por último, un aspecto positivo de encontrar O3 de forma natural, es que contribuye a remover gases producidos por actividades humanas como el metano (CH4), el monóxido de carbono (CO) y óxidos de nitrógeno (IDEAM, 2002).

4.2.5. Big Data[3]

García y otros (2018) afirman que:

Es el conjunto de arquitecturas y herramientas informáticas destinadas a la manipulación, gestión y análisis de grandes volúmenes de datos desde todo tipo de fuentes, diseñadas para extraer valor y beneficio de estos, con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y sobre todo con un costo reducido (p.15).

[2] Acrónimo de smoke (humo) + flog (niebla) para designar niebla tóxica.
[3] En español se le conoce como datos masivos.

4.2.6. Aprendizaje Profundo

El aprendizaje profundo o Deep Learning es un derivado del Machine Learning [4] (Aprendizaje Automático) que consiste en configurar parámetros con respecto a los datos y entrenar una máquina con el fin de aprender por sí misma, teniendo la capacidad de razonar y sacar sus propias conclusiones como lo haría el ser humano. Entre sus aplicaciones más comunes están el reconocimiento facial y de voz, procesamiento de imágenes o realización de predicciones. Para ello, el aprendizaje profundo utiliza una estructura multicapa para extraer las características inherentes de los datos capa por capa del nivel más bajo al más alto llamada Red Neuronal Artificial (Li, Peng, Hu, Shao, & Chi, 2016), las cuales se asemejan a una red neuronal biológica del cerebro humano, lo que hace que sea una tecnología más robusta y tenga mayor capacidad de aprendizaje.

4.2.7. Redes Neuronales Artificiales

Como se trató al final del punto anterior, las redes neuronales artificiales (RNA) emulan el funcionamiento del cerebro humano, al ser una técnica que aprende de sí misma es capaz de alcanzar una eficacia que se compara a la del ser humano al resolver problemas complejos. “El objetivo de las técnicas basadas en RNA es intentar expresar la solución de problemas complejos como el resultado de combinar pequeñas contribuciones realizadas por un gran número de elementos simples de procesamiento que se hallan interconectados entre sí” (Berzal, 2018, pág. 125).

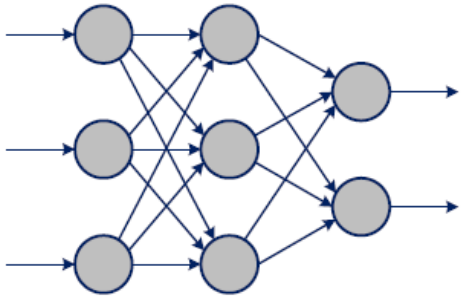
La estructura común de estas redes neuronales se muestra en la Figura 7, donde la capa de entrada (a la izquierda) recibe datos externos para ser traducidos en una respuesta en

[4] Es un subconjunto de la inteligencia artificial que dota a los ordenadores de la capacidad de identificar patrones en datos masivos y realizar predicciones.

la capa de salida (a la derecha) y entre ellas puede haber una o más capas, denominadas capas ocultas.

Figura 7

Red Neuronal Artificial con múltiples capas de neuronas.



Nota. Tomado de (Berzal, 2018).

4.2.7.1. Redes Neuronales Convolucionales (CNN).

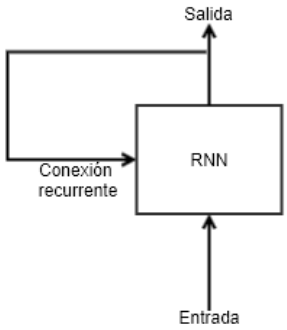
(Ghatak, 2019) afirma: "Las redes convolucionales se componen de capas convolucionales que actúan como extractores de características jerárquicas. Se utilizan principalmente para la clasificación de imágenes o texto, detección de objetos y segmentación de imágenes" (p.43).

4.2.7.2. Redes Neuronales Recurrentes.

Las Redes Neuronales Recurrentes (RNN, por sus siglas en inglés) tienen la capacidad de procesar y obtener información de datos secuenciales, es decir, utilizan información pasada a través de la red neuronal conformando un ciclo, como se muestra en la Figura 8. Esta cualidad las provee de memoria y las hace idóneas para modelar series temporales. Por ende, al analizar videos o música, subtitular imágenes o procesar el lenguaje, este tipo de redes neuronales dependen de su capacidad (Arana, 2021).

Figura 8

Bucle cerrado de una RNN.



Nota. Tomado de (Martín Gutiérrez, 2019).

4.2.7.3. Feedforward Neural Network.

Las Redes Neuronales Feedforward (FNN, por sus siglas en inglés) se componen de la capa de entrada, la capa de salida y las capas intermedias que son capas ocultas. Reciben dicho nombre porque carecen de realimentación y se pueden utilizar para la clasificación y el aprendizaje de funciones no supervisado. Estas redes deducen una cadena de modificaciones entre su entrada y su salida, reflejando una nueva representación de la entrada en cada capa sucesiva mediante una transformación no lineal de las capas inferiores (Ghatak, 2019). A partir de su arquitectura, existen dos derivados que se muestran a continuación:

4.2.7.3.1. Autoencoders o Autocodificador. Las Redes de Autocodificador son FNN que pueden tener más de una capa oculta, permitiendo trabajar con un gran volumen de información, logrando comprimirla y filtrarla adecuadamente. Estas redes neuronales intentan reconstruir los datos de entrada en la capa de salida, dejando ambas capas del mismo tamaño y se entrenan mediante un método de descenso de gradiente, como la propagación hacia atrás *(Tan & Eswaran, 2008)*.

4.2.7.3.2. ***Multilayer Perceptron (MLP).*** Es un complemento de FNN y consta de 3 tipos de capas: entrada, salida y oculta. Las neuronas del MLP se entrenan con el algoritmo de aprendizaje de retropropagación. Estas redes están diseñadas para aproximarse a cualquier función continua y son capaces de resolver problemas que no son separables linealmente. Se usan principalmente en clasificación, reconocimiento, predicción y aproximación de patrones (Abirami & Chitra, 2019).

4.2.7.4. Long Short-Term Memory (LSTM)

Las redes neuronales de memoria a corto y largo plazo son un tipo especial de RNN capaces de aprender dependencias a largo plazo, su diseño se basa en recordar información durante largos períodos de tiempo. Estas redes están compuestas por puertas de olvido, de entrada y de salida (Mañas, 2019).

4.2.8. Sequence to Sequence (Seq2Seq)

Se compone de dos RNN llamadas codificador y decodificador. El objetivo de esta red es tomar una secuencia de elementos (palabras, letras, características de una imagen, etc.), el codificador procesa cada uno de estos elementos, compila la información y la guarda en un vector (llamado contexto), luego, el codificador envía ese contexto al decodificador el

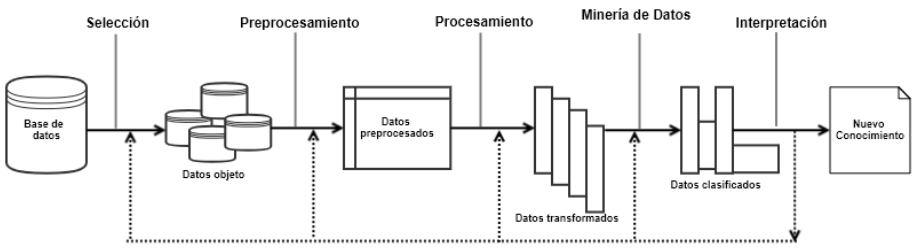
cual produce la secuencia de salida elemento a elemento. Este modelo ha tenido éxito en traducción automática, resumen de texto y subtítulos de imágenes (Alammar, 2018).

5. Metodología

Para el desarrollo de este trabajo de investigación se adoptó la metodología Descubrimiento del Conocimiento en Bases de Datos o "*Knowledge Discovery in Database*" (KDD), la cual se presenta en la Figura 9. Esta metodología fue seleccionada por la gran utilidad en trabajos de investigación que se relacionen con la minería de datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), consta de 5 etapas: selección, donde se busca escoger la base de datos que se ajuste con la temática; preprocesamiento, etapa en la cuál se realiza la limpieza de los datos para eliminar ruido, falta de datos y outliers, para introducirlos en las siguientes etapas; procesamiento, allí se realiza la exploración de los hiperparámetros para ajustar los modelos al conjunto de datos preprocesados; minería de datos, se observa el rendimiento de los modelos con las funciones de precisión y pérdida. Por último, se realiza la interpretación de los resultados obtenidos en cuanto al mejor modelo.

Figura 9

Metodología KDD.



Nota. Adaptado de (García González, Sánchez Sánchez, Orozco, & Obredor, 2019).

5.1. Etapa de selección

En esta primera etapa se realiza una revisión de literatura donde se hace un descubrimiento del conocimiento previo por medio de la base de datos de SCOPUS y se fija la meta que se quiere cumplir con esta metodología. Posteriormente, se hace una búsqueda de los datos que apoyan el desarrollo del proyecto por medio de una fuente de información pública, en este caso la del IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales).

De acuerdo con lo anterior, se selecciona como conjunto de datos las tres ciudades: Bogotá, Cali y Bucaramanga, estableciendo una ventana de tiempo desde el 01 de enero de 2019 hasta el 31 de julio de 2021, con el objetivo de comparar 3 escenarios: antes, durante y después del confinamiento obligatorio, el cual ocurre en las fechas del 12 de marzo al 31 de agosto de 2020, encontrándose en total 116.467 registros, distribuidos así: 8.501 Bucaramanga, 83.236 Bogotá y 24.730 Cali. Las variables meteorológicas seleccionadas como variables de entrada son: DV (Dirección del viento), VV (Velocidad del Viento), HA (Humedad del Aire), P (Presión atmosférica), TA (Temperatura del aire), las cuales se eligieron según el artículo de Shatnawi & Abu-Qdais (2021) y fueron encontradas en dicho rango de tiempo; y los contaminantes seleccionados como variables de salida son: PM2,5 (Material Particulado 2,5), O3 (Ozono Troposférico), NO2 (Dióxido de Nitrógeno) y PM10 (Material Particulado 10), utilizados para comparar y predecir la calidad del aire en el presente trabajo de grado, coincidiendo con los resultados de la revisión de literatura previamente realizada.

Para efectos de este trabajo de investigación, las variables contendrán la nomenclatura descrita en la Tabla 6 junto a sus unidades de medida.

Tabla 6

Nomenclatura de las variables utilizadas.

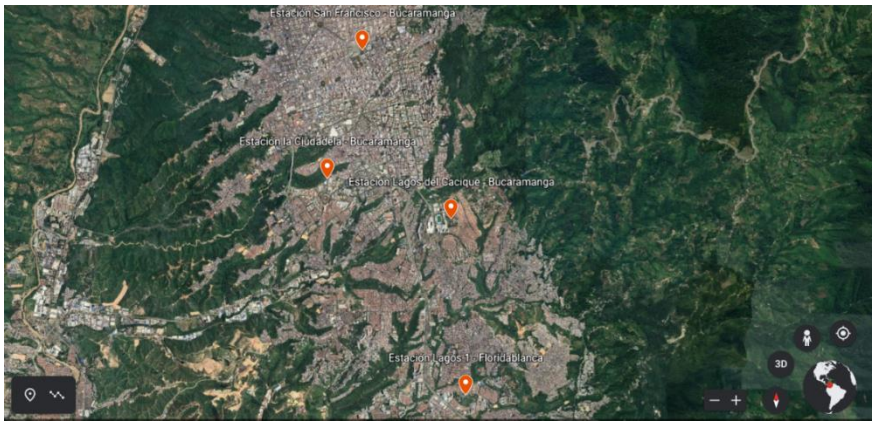
Variable	Unidad de medida	Nomenclatura
Material Particulado 2.5	$\mu g/m^3$	PM2.5
Ozono Troposférico	$\frac{\mu g}{m^3}$	O3
Dióxido de Nitrógeno	$\frac{\mu g}{m^3}$	NO2
Dirección del Viento	$°\ (Grados)$	DV
Velocidad del Viento	m/s	VV
Humedad Relativa	[%]	HR
Presión atmosférica	hPa	P
Material Particulado 10	$\mu g/m^3$	PM10
Temperatura del Aire	℃	TA

Es importante resaltar que las estaciones se encuentran ubicadas en los puntos que se muestran en la Figura 10, Figura 18 y Figura 29, según los criterios de densidad de la población, distribución de fuentes de emisión, meteorología y topografía.

Como se muestra en la Figura 10, el Área Metropolitana de Bucaramanga tiene 4 estaciones denominadas “Estación San Francisco”, “Estación Ciudadela”, “Estación Lagos del Cacique” y “Estación Lagos 1 - Floridablanca” de las cuáles sus registros respectivamente son de: 1.650, 2.294, 1.592 y 2.965.

Figura 10

Estaciones para la Ciudad de Bucaramanga.

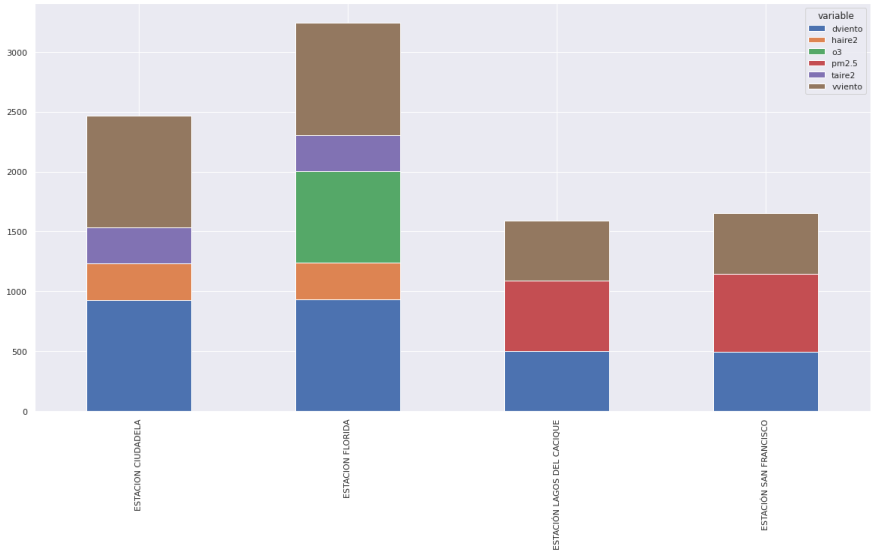


Nota. Gráfico tomado de Google Maps.

Con el fin de conocer la distribución de cada variable por ciudad, se presenta la Figura 11, donde se muestra la cantidad de datos por día que registra cada variable en cada estación para la ciudad de Bucaramanga y su Área Metropolitana.

Figura 11

Distribución de las variables para la ciudad de Bucaramanga.

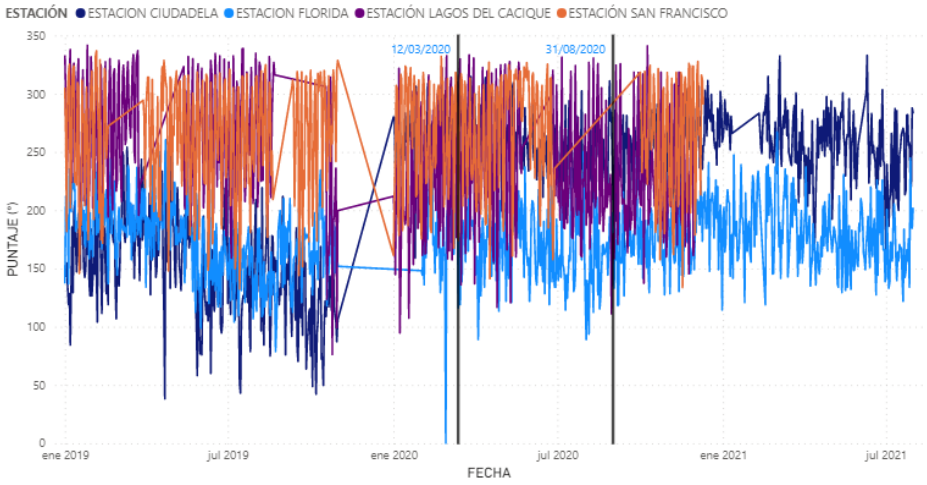


Como se puede observar, hay una gran pérdida de datos para las variables TA, O3 y HA en las estaciones de “Lagos del Cacique” y “San Francisco”, resaltando que, aunque hay valores para las otras variables siguen siendo poca cantidad para el estudio. De igual manera, para la estación “Ciudadela” no se encuentran las variables O3 y PM2.5, las cuales son de vital importancia para la posterior predicción. Por último, para la estación “Florida” la variable faltante es el PM2.5. Cabe mencionar que la situación que se está presentando puede ser debida a que cada estación no mide todas las variables que se utilizan en este proyecto, lo que ocasiona que se tenga la información dispersa entre las estaciones.

Por otra parte, en la Figura 12 se muestra el comportamiento de la Dirección del Viento para las 4 estaciones de la ciudad de Bucaramanga y su Área Metropolitana, aquí se evidencia que para todas las estaciones existen rangos de tiempo que no presentan registros, aproximadamente entre las fechas del 31 de octubre de 2019 hasta el 01 de enero de 2020. De igual manera, la “Estación Ciudadela” registra valores con dirección del viento desde el Noreste (38°) al Suroeste (260°), la “Estación Florida” tiene direcciones del Norte (0°) al Suroeste (267°), la “Estación Lagos del Cacique” registra direcciones desde el Noreste (76°) hacia el Noroeste (341°) y la “Estación San Francisco” registra desde el Sureste (132°) hasta el Noroeste (336°), de lo anterior se puede concluir que las estaciones registran valores de todas las direcciones, resaltando que la “Estación Lagos del Cacique” es la que más variación tiene.

Figura 12

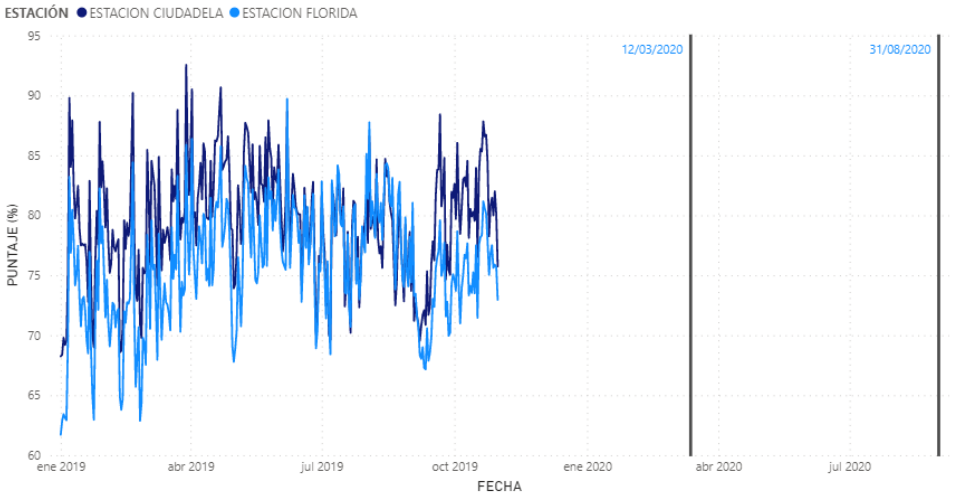
Comportamiento Dirección del Viento en Bucaramanga.



Para la variable HR en la Figura 13 se visualizan datos en las estaciones “Ciudadela” y “Florida” únicamente desde el 01 de enero de 2019 hasta el 31 de octubre de 2019. En ambas estaciones, dicha variable se comporta de manera similar presentando aumento de valores de enero de 2019 hasta abril de 2019, evidenciando una disminución hasta septiembre y nuevamente un aumento hasta finales de octubre del mismo año.

Figura 13

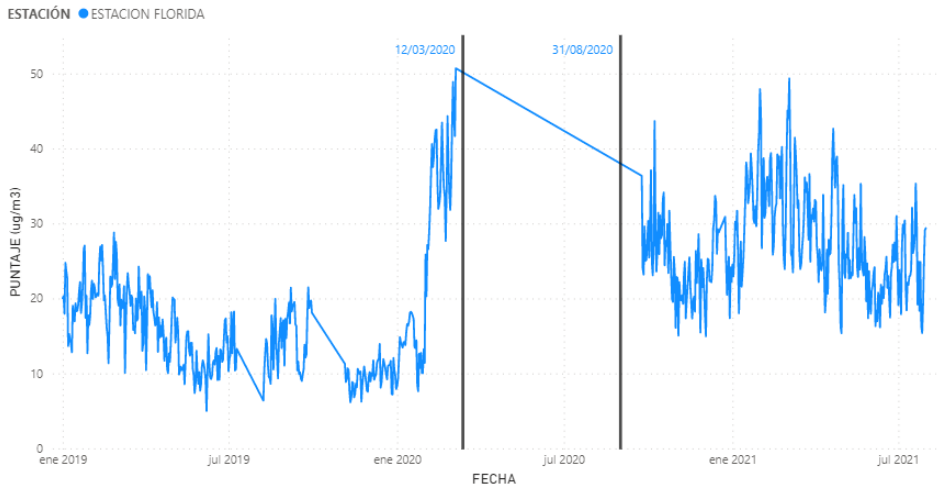
Comportamiento Humedad del Aire en Bucaramanga.



El comportamiento de la variable O3 se encuentra representado en la Figura 14, donde se refleja una gran pérdida de datos para las fechas del 07 de marzo al 23 de septiembre de 2020, coincidiendo con la duración del confinamiento. Además, esta variable sólo presenta registros para la “Estación Florida”. Se evidencia un aumento notorio de valores de febrero a marzo de este año, llegando hasta un máximo puntaje de 50,63 µg⁄m^3, sin embargo, este pico se encuentra en el rango de categoría “Buena” según la Figura 3.

Figura 14

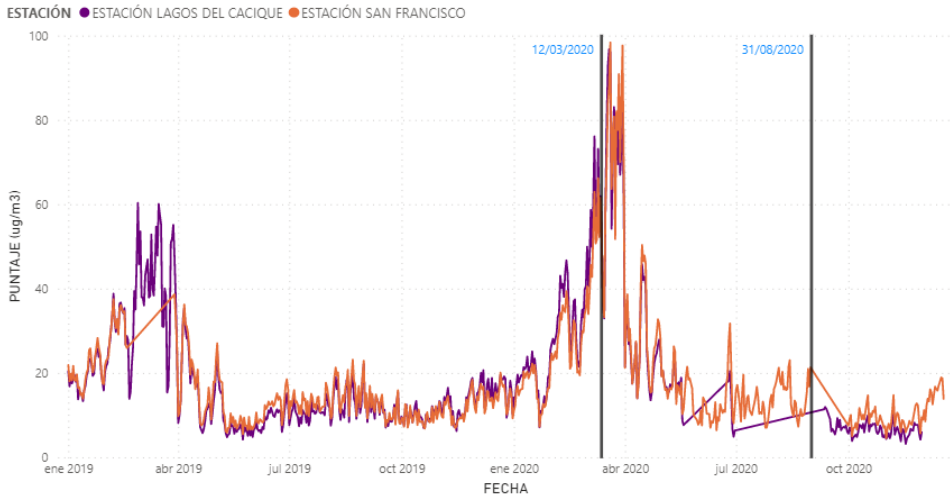
Comportamiento Ozono Troposférico en Bucaramanga.



En la Figura 15, se observa la variable PM2.5, la cual registra datos únicamente para las estaciones “Lagos del Cacique” y “San Francisco”, es posible notar que, durante rangos muy pequeños de tiempo, esta gráfica presenta pérdida de datos. Las dos estaciones se muestran muy similares entre sí, visualizando aumentos significativos para el mes de marzo de 2020, los cuales tratan de nivelarse entre el 01 de abril y el 15 de diciembre de 2020.

Figura 15

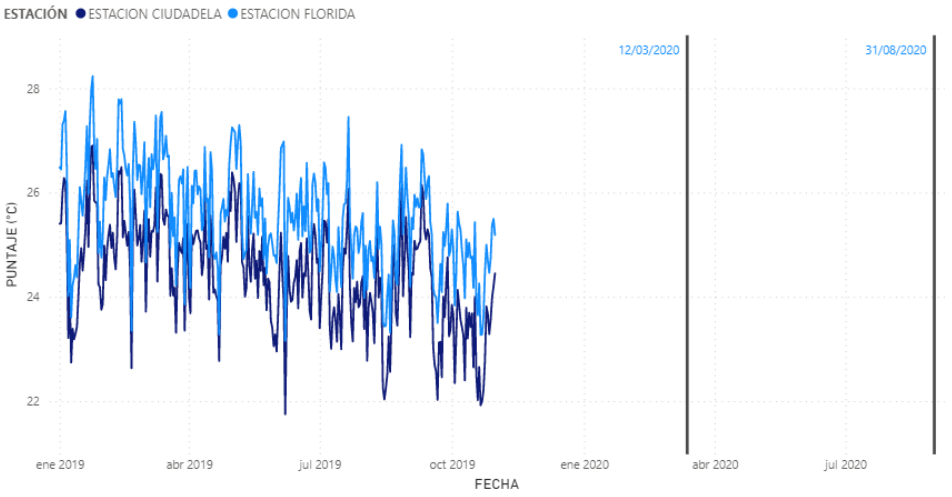
Comportamiento Material Particulado 2.5 en Bucaramanga.



Para la TA se encuentran registros de datos durante los rangos de enero de 2019 hasta octubre del mismo año, situación que se percibe en la Figura 16. Es importante resaltar que ambas estaciones presentan gran similitud en el comportamiento de sus valores.

Figura 16

Comportamiento Temperatura del Aire en Bucaramanga.

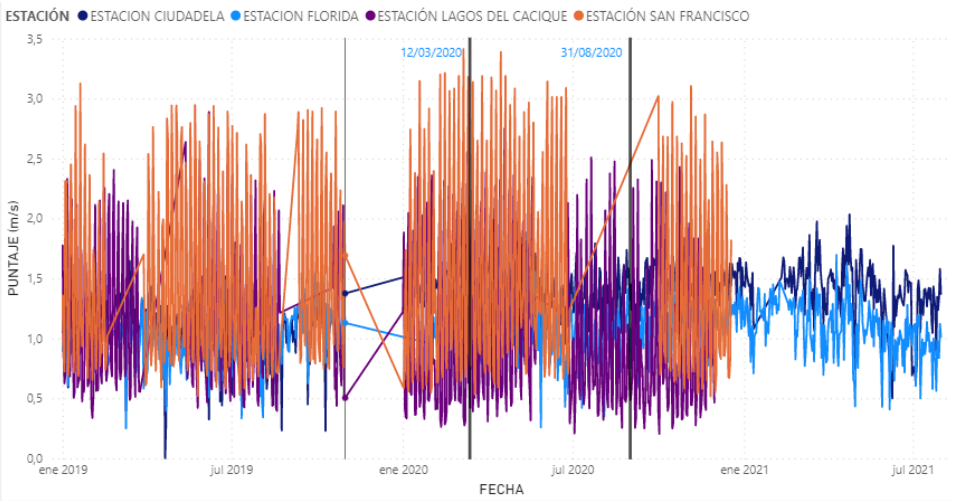


En la Figura 17, se observa el comportamiento de los datos para la variable VV, el

cual es similar a la DV con datos faltantes en los mismos rangos de fechas e incluso en otras fechas adicionales. En las estaciones “Lagos del Cacique” y “San Francisco” estos datos presentan mayor variación, mientras que las estaciones “Ciudadela” y “Florida”, tienen menos variación y son muy similares entre sí.

Figura 17

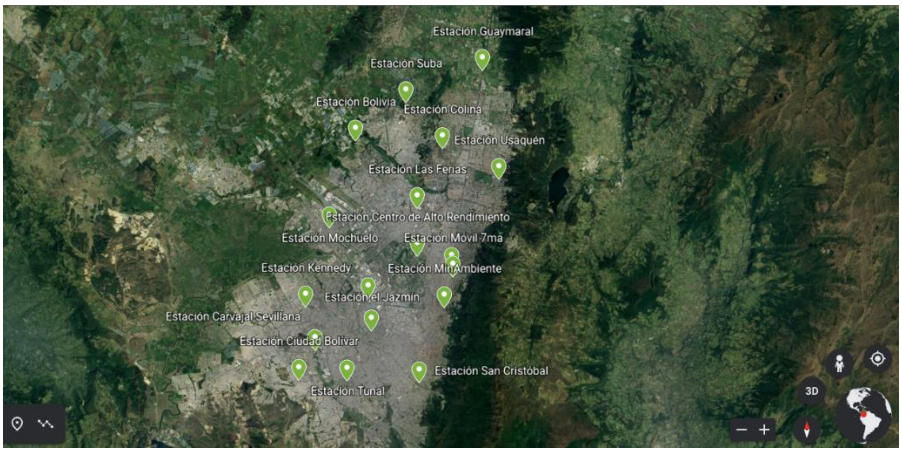
Comportamiento Velocidad del Viento en Bucaramanga.



Para la ciudad de Bogotá se encuentra mayor cantidad de estaciones debido a su área y población, encontrándose 18, denominadas así con su respectiva cantidad de datos: “Estación Bolivia” (1.341), “Estación Carvajal” (5.447), “Estación Centro de alto rendimiento” (3.727), “Estación Ciudad Bolívar” (2.509), “Estación Colina” (1.538), “Estación El Jazmín” (2.722), “Estación Guaymaral” (7.593), “Estación Kennedy” (7.393), “Estación Las Ferias” (8.307), “Estación MinAmbiente” (4.694), “Estación Mochuelo” (1.646), “Estación Móvil” (4.482), “Estación Puente Aranda” (5.251), “Estación San Cristóbal” (6.583), “Estación Suba” (4.938), “Estación Tunal” (8.067), “Estación Usaquén” (4.412) y “Estación Usme” (2.586) como se muestra en la Figura 18.

Figura 18

Estaciones para la ciudad de Bogotá.

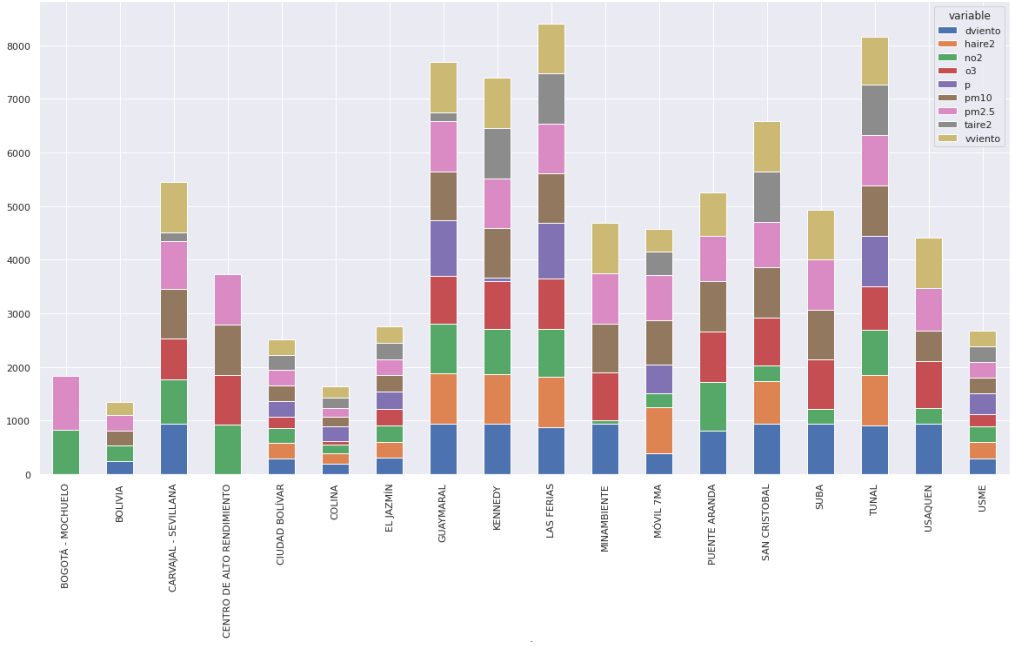


Nota. Gráfico tomado de Google Maps.

Para la ciudad de Bogotá, la distribución de las variables se ve representada en la Figura 19.

Figura 19

Distribución de las variables para la ciudad de Bogotá.



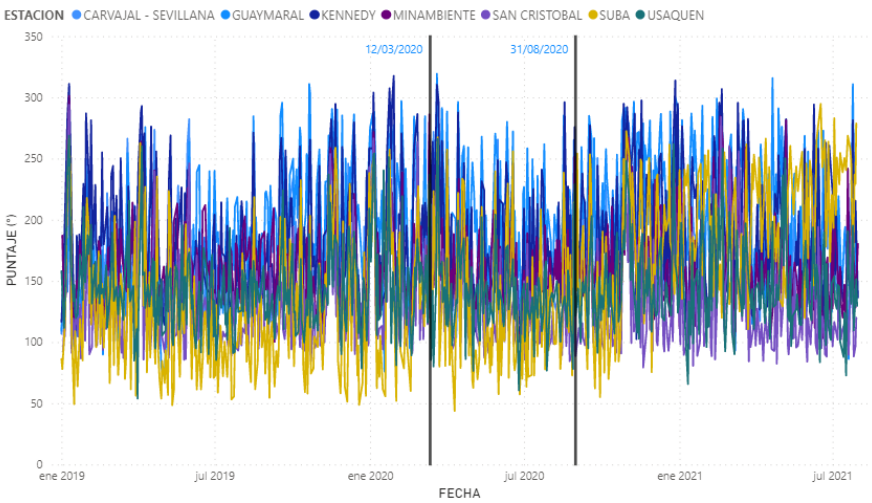
Como se observa, la ciudad de Bogotá tiene mayor cantidad de registros de todas las variables para las 18 estaciones, sin embargo, hay estaciones como “Mochuelo”, “Bolivia”, “Carvajal”, “Centro de Alto Rendimiento”, “Móvil 7ma”, “MinAmbiente”, “Puente Aranda”, “San Cristobal”, “Suba” y “Usaquén” a las cuales les hace falta una o más variables, ocasionado posiblemente por la misma situación de Bucaramanga. Son estaciones que por su ubicación o capacidad sólo pueden medir ciertas variables. Adicionalmente, se percibe que las estaciones con más datos son las de “Guaymaral”, “Kennedy”, “Las Ferias” y el “Tunal”.

Con el fin de visualizar el comportamiento de las variables para las estaciones de la ciudad Bogotá, se decidió mostrar las estaciones más significativas para cada una de las variables, ya que Bogotá contiene 18 estaciones, las cuales al graficarlas todas saturarían las figuras.

En la Figura 20, se observa la variación de DV, la cual en todas estaciones presenta gran similitud, oscilando entre valores de 50° (noreste) a 300° (noroeste).

Figura 20

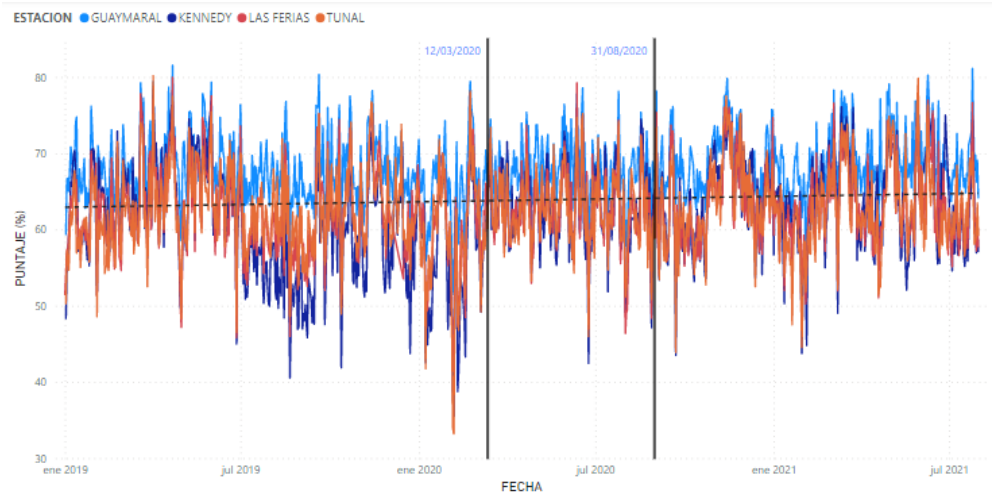
Comportamiento Dirección del Viento en Bogotá.



La variable HR se encuentra representada en las estaciones que registran mayor cantidad de datos en la ventana de tiempo escogida, las cuales son “Guaymaral”, “Kennedy”, “Las Ferias” y “Tunal”. Asimismo, se puede observar que durante este tiempo presenta un patrón de triángulo simétrico, el cual refleja que la tendencia de esta variable es continua, así como se muestra en la Figura 21.

Figura 21

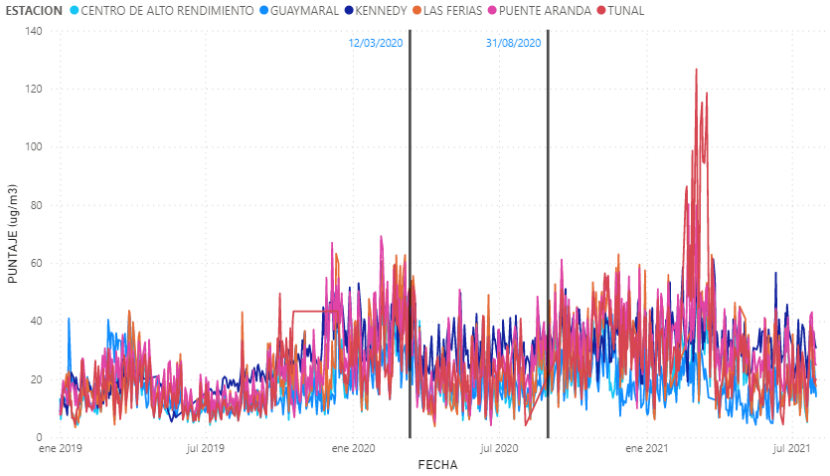
Comportamiento Humedad del Aire en Bogotá.



Para el caso del contaminante NO2 en la Figura 22, se puede observar que durante el confinamiento midió valores con poca variación, manteniéndose ligeramente constante. Aunque se representan las estaciones que más cantidad de datos tienen, se nota la pérdida de estos por rangos de tiempo mínimos. De igual manera, se presentan aumentos abruptos para la estación “Tunal”, pertenecientes a 4 días del mes de marzo de 2021, y, al no tener información de un acontecimiento importante en Bogotá, se puede deducir que son irregularidades en la toma de las mediciones.

Figura 22

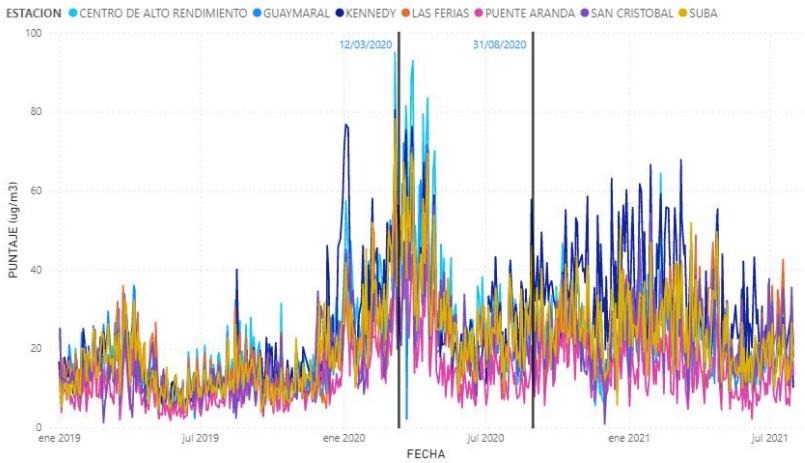
Comportamiento Dióxido de Nitrógeno en Bogotá.



En la Figura 23, se evidencia el contaminante O3, el cual toma valores mínimos dentro del rango de tiempo establecido. Es importante resaltar que durante el inicio del confinamiento esta variable está presentando picos muy altos los cuales van disminuyendo al pasar este tiempo y se mantienen con una tendencia similar con la que se terminó dicho periodo. Sin embargo, los máximos valores se encuentran en la categoría “Buena”.

Figura 23

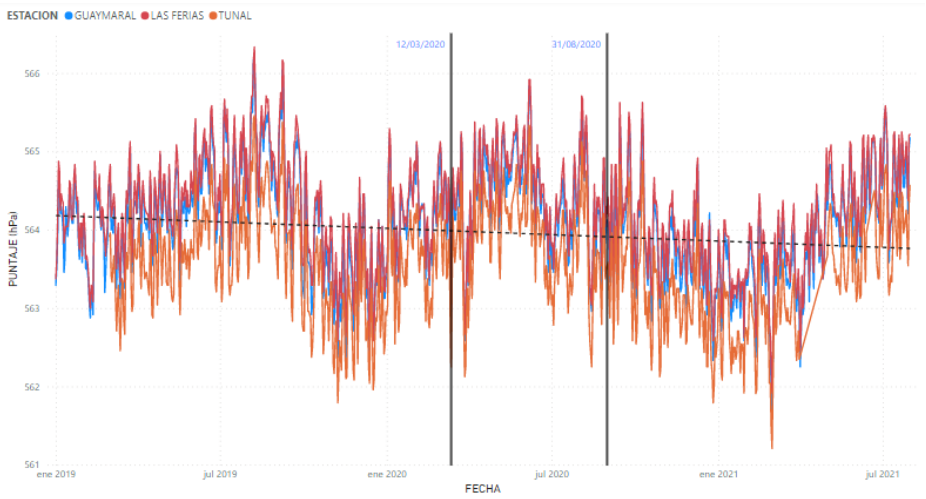
Comportamiento Ozono Troposférico en Bogotá.



En la Figura 24, se demuestra que la variable P presenta una notoria tendencia descendente. Cabe mencionar que esta variable se encontró solamente para esta ciudad en la base de datos inicial.

Figura 24

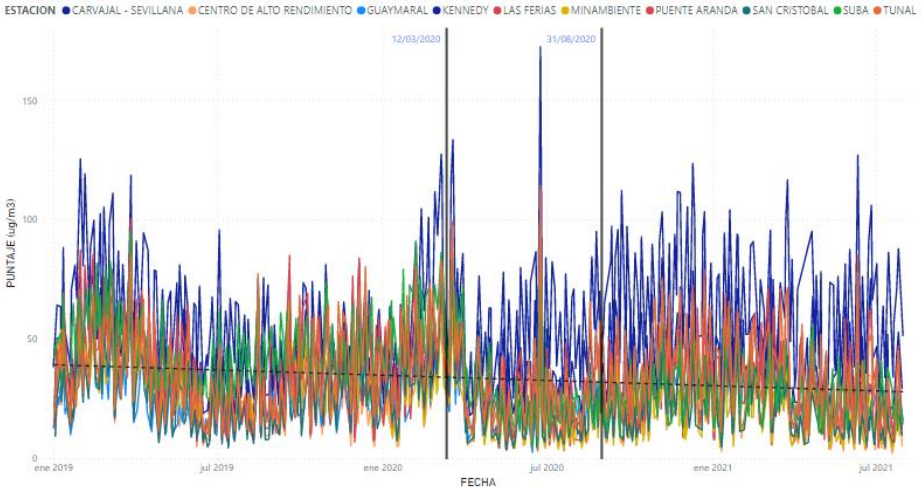
*Comportamiento Presión en Bogotá***.**



Haciendo referencia al contaminante PM10, la base de datos solo presenta valores para la ciudad de Bogotá, así como ocurre con la variable P. En la Figura 25, es posible visualizar que la estación “Carvajal” está registrando altos puntajes respecto a las demás estaciones con un pico muy significativo para el día 24 de junio del 2020, durante el confinamiento. Es notorio que todas las estaciones aumentaron este día, pero no se puede afirmar qué sucedió pues no aparecen reportes de alto impacto que puedan explicar este suceso. Por último, a pesar de que la estación “Carvajal” varía en cantidades más altas, la tendencia es descendente.

Figura 25

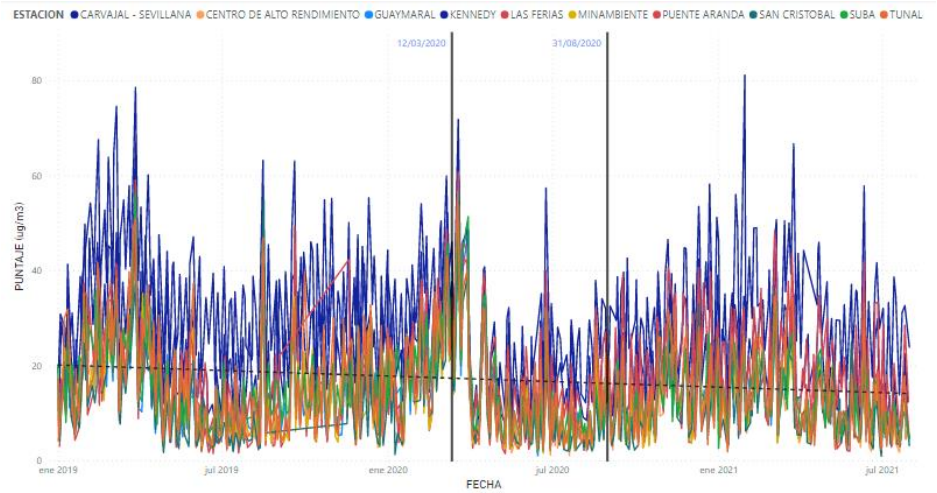
Comportamiento Material Particulado 10 en Bogotá.



Para el contaminante PM2.5, se observa durante el confinamiento se ve una notoria disminución en su comportamiento, sin embargo, después de este vuelven a su estado inicial como se muestra en la Figura 26.

Figura 26

Comportamiento Material Particulado 2.5 en Bogotá.

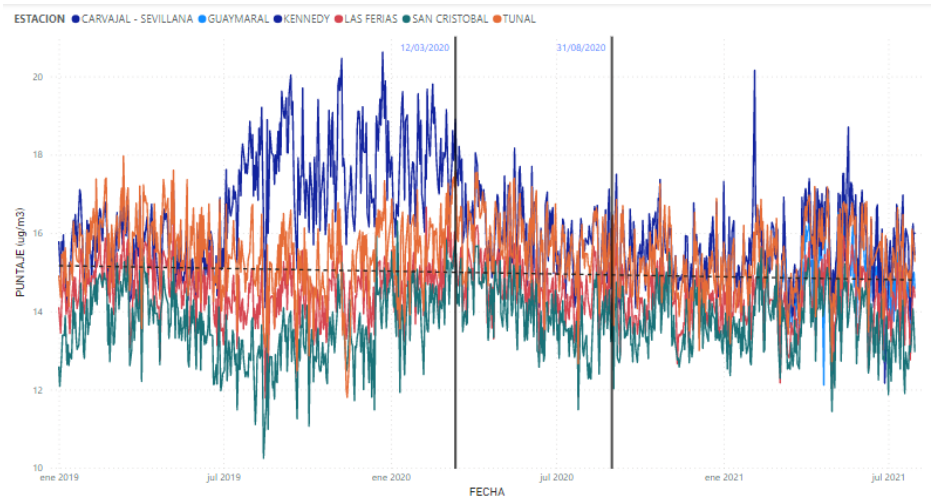


En la Figura 27, se identifica que la estación “Kennedy” tiene valores sesgados de

julio de 2019 a marzo de 2020 presentando un aumento en el puntaje respecto a las demás estaciones, las cuales simultáneamente disminuyen en este rango de tiempo. Así mismo, a pesar de que hay un pico significativo después del confinamiento, la variación tiende a mantenerse continua.

Figura 27

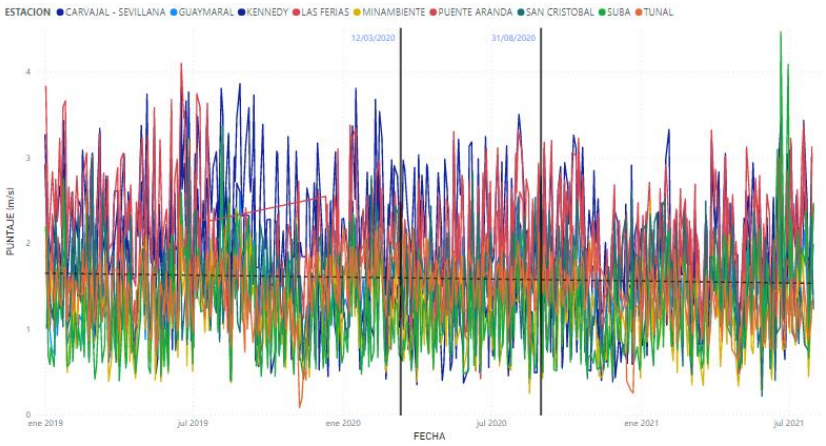
Comportamiento Temperatura del Aire en Bogotá.



Finalizando para la ciudad de Bogotá, se tiene la variable VV, cuyo comportamiento es similar a la variable DV. En esta, las estaciones a pesar de tener gran cantidad de datos presentan pérdidas de estos especialmente para la estación “Puente Aranda”, esta variable se muestra en la Figura 28.

Figura 28

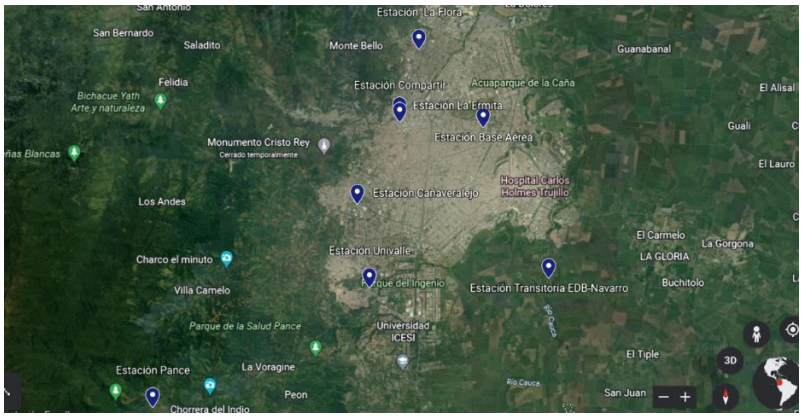
Comportamiento Velocidad del Viento en Bogotá.



Por último, en la ciudad de Cali, en la Figura 29 se encuentran las siguientes 8 estaciones con sus respectivas cantidades: “Estación Base aérea” (1.846), “Estación Cañaveralejo” (3.882), “Estación Compartir” (4.923), “Estación La Ermita” (1.440), “Estación La Flora” (2.978), “Estación Pance” (4.816), “Estación Transitoria Navarro” (1.248) y “Estación Univalle” (3.597).

Figura 29

Estaciones para la Ciudad de Cali.

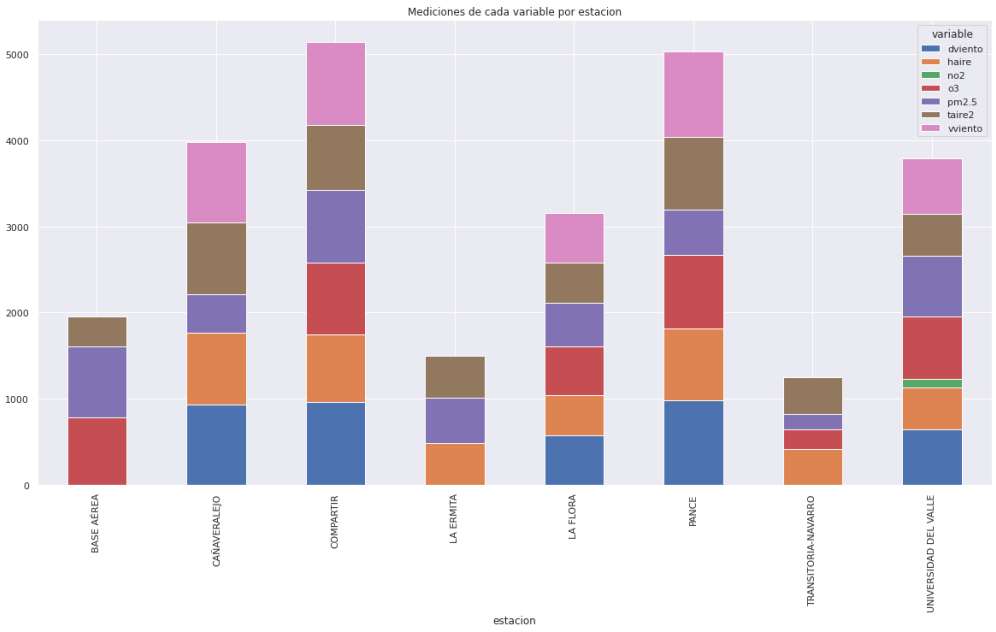


Nota. Gráfico tomado de Google Maps.

Respecto a la ciudad de Cali, en la Figura 30 se muestra la distribución de las variables por estación. Se observa que solo una estación está registrando valores para NO2 debido a que el equipo está por fuera de servicio desde el año 2019 y por motivos presupuestales no se ha podido reparar, según lo menciona el contratista profesional en la calidad del aire de la Alcaldía de Santiago de Cali del Departamento Administrativo de Gestión del Medio Ambiente (DAGMA), el señor Diego Andrés Arias Arana (Alcaldía de Santiago de Cali, 2022). Adicionalmente, se aprecia que las estaciones con mayor número de datos son las de "Compartir" y "Pance".

Figura 30

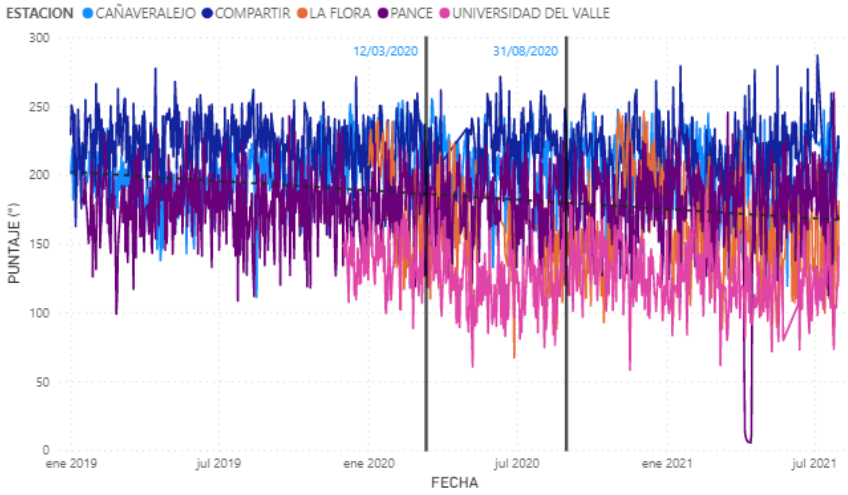
Distribución de las variables para la ciudad de Cali.



A pesar de que se escogen las 5 estaciones más representativas para la variable DV, en la Figura 31 se aprecia una considerable pérdida de datos para algunas de estas sobre todo para el año 2019 e incluso existe un notorio pico en la estación "Pance". Por otro lado, la variación sigue una tendencia decreciente.

Figura 31

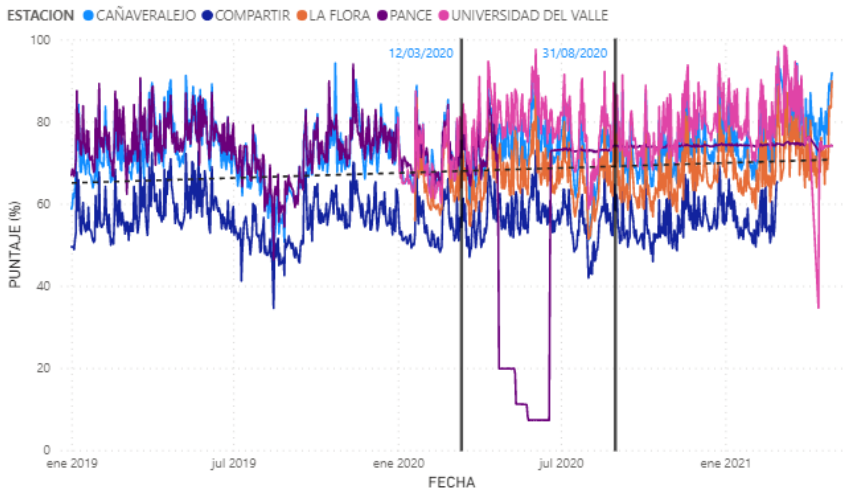
Comportamiento Dirección del Viento en Cali.



La pérdida de datos en Cali sigue estando presente en la variable HR como se puede observar en la Figura 32. De igual manera, la estación “Pance” sigue registrando datos que se salen del comportamiento promedio de las demás estaciones, especialmente durante y después del confinamiento.

Figura 32

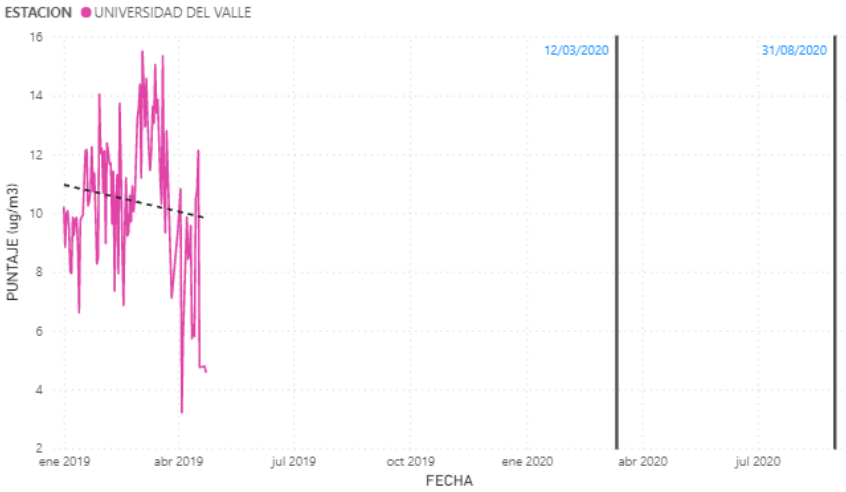
Comportamiento Humedad del Aire en Cali.



La medición de NO2 para Cali es un caso especial, ya que como se mencionó en párrafos anteriores, el equipo que mide este contaminante se encuentra fuera de servicio desde el año 2019, tal como se evidencia en la Figura 33, registrando así muy poca cantidad de datos que sean útiles para el trabajo de investigación.

Figura 33

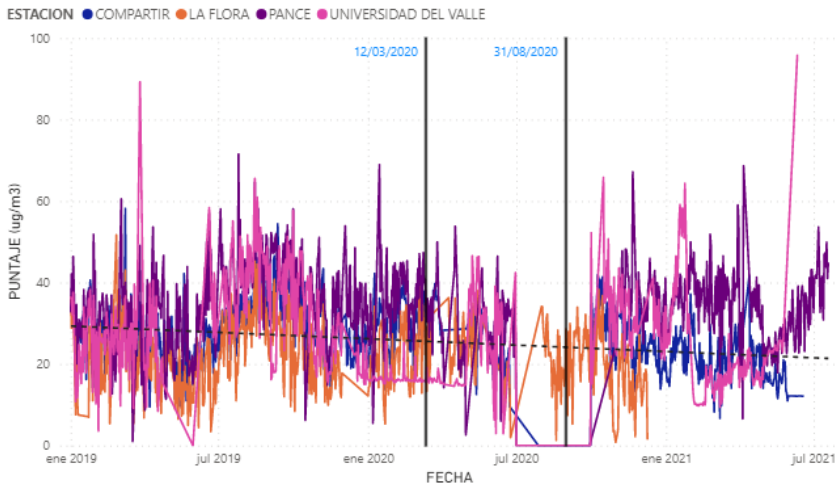
Comportamiento Dióxido de Nitrógeno en Cali.



En la Figura 34 se refleja el contaminante O3 el cual evidencia irregularidades en la toma de sus mediciones, especialmente para las estaciones “Universidad del Valle” y “Compartir”, ya que presentan picos abruptos y pérdida de datos durante el confinamiento y después de este.

Figura 34

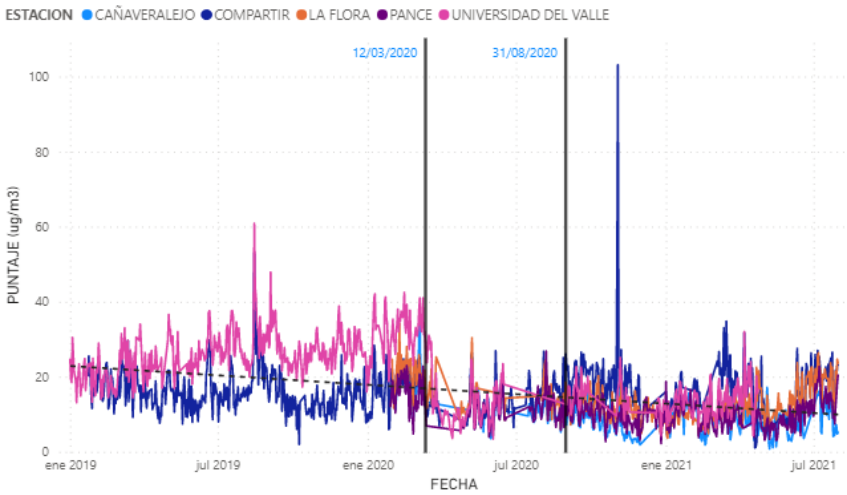
Comportamiento Ozono Troposférico en Cali.



De igual manera que las variables anteriormente mencionadas, en la Figura 35 se muestra el contaminante PM2.5 el cual contiene valores extremos que sesgan el comportamiento general de las estaciones. Así mismo, antes y durante el confinamiento se evidencia faltantes de datos y presenta una tendencia descendente.

Figura 35

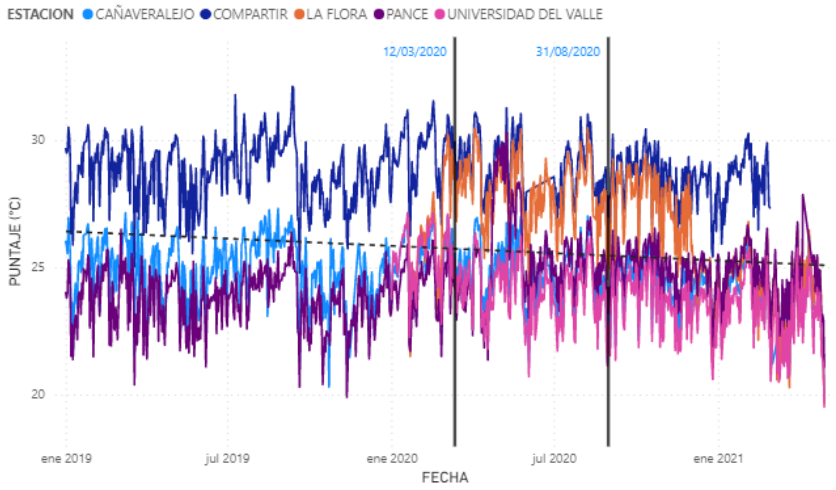
Comportamiento Material Particulado 2.5 en Cali.



Para la variable TA en la Figura 36 se observa que las estaciones “Compartir” y “La flora” miden Temperaturas más altas que las demás estaciones. Se resalta que para todas las estaciones en el período antes del confinamiento muestran patrones de comportamiento por rangos de tiempo y de igual manera, es posible notar que después del confinamiento hay una leve disminución en los valores.

Figura 36

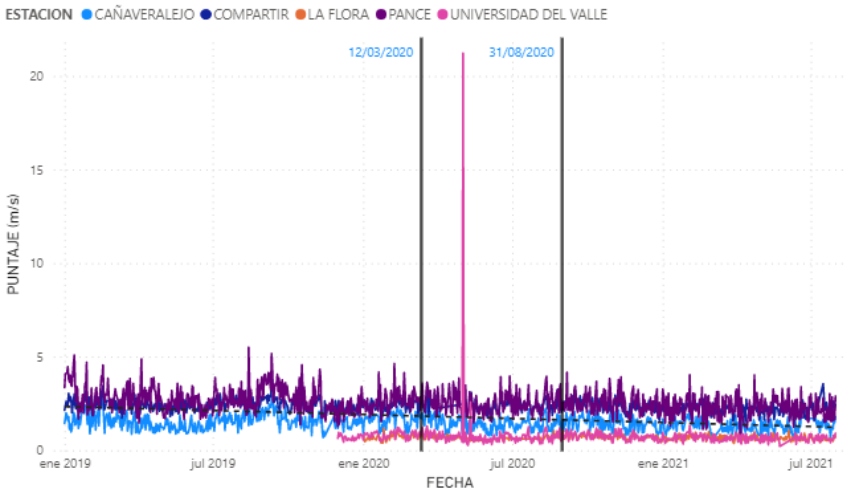
Comportamiento Temperatura del Aire en Cali.



Finalmente, para la ciudad de Cali en la Figura 37 se representa la variable VV, donde se aprecia una toma errónea, pues es un valor fuera del común que afecta las mediciones de la ventana de tiempo establecida.

Figura 37

Comportamiento Velocidad del Viento en Cali.



5.2. Etapa de preprocesamiento

Como segunda etapa se tiene el objetivo de mejorar la calidad de los datos y evitar que en la cuarta etapa, estos sean erróneos o poco confiables, para ello, a los datos objeto se le aplican operaciones para eliminar datos con ruido; se seleccionan estrategias para manejar los datos desconocidos, faltantes y duplicados, los cuales se ignoran o se reemplazan usando técnicas estadísticas como la media, la moda, el mínimo y el máximo, dando como resultado los datos preprocesados (Timarán Pereira, Hernández Arteaga, Caicedo Zambrano, Hidalgo Troya, & Alvarado Pérez, 2016).

Para obtener los datos preprocesados se procede a realizar una limpieza de datos, la cual consiste inicialmente en unificar los nombres de las columnas y las filas, dejar los formatos adecuados de las fechas, quitar los caracteres especiales, cambiar signos decimales, y convertir las variables en columnas.

De acuerdo con el análisis de los datos iniciales que se trató en el numeral anterior, se decide eliminar estaciones que por su baja calidad y cantidad de datos afectarían

directamente al proceso de procesamiento y, por ende, de predicción que se llevará a cabo en el presente proyecto.

Para la ciudad de Bogotá se definió que no se van a tener en cuenta las estaciones “Mochuelo”, “Colina”, “Bolivia”, “Ciudad Bolívar”, “El Jazmín”, “Móvil 7ma”, “Usaquén” y “Usme”, las cuales representan 16.962 datos. Para la ciudad de Bucaramanga no se va a eliminar ninguna estación ya que entre las 4 se complementan y ofrecen datos de buena calidad. Para la ciudad de Cali se eliminan las estaciones “Base aérea”, “La Ermita” y “Transitoria” con un total de 4.534 datos. Estos datos eliminados representan un 18,457% respecto a la totalidad. Los cálculos hechos anteriormente se presentan a continuación.

Tabla 7

Cantidad de datos encontrados en la base.

	Datos iniciales	Datos eliminados de las estaciones	Datos sin estaciones	% Datos eliminados
Bogotá	83.236	16.962	66.274	20,378%
Bucaramanga	8.501	0	8.501	0%
Cali	24.730	4.534	20.196	18,334%
Total	116.467	21.496	94.971	18,457%

Posteriormente, se realiza la eliminación de datos que se comportan de manera anormal, es decir, que se encuentran por fuera de la tendencia del conjunto de datos. Por ello, se hizo uso del método IQR, que consiste en calcular los valores atípicos que se encuentren

por debajo de (Q1-1,5*IQR)[5] y por encima de (Q3+1,5*IQR). Para el presente proyecto, se tomó la decisión de cambiar el valor de 1,5 por 2,5 en la ecuación, valor que determina el tamaño de los bigotes, los cuales indican el máximo y el mínimo valor permitido dentro del comportamiento de los datos. Esta decisión se tomó ya que dentro de los valores que salían atípicos al colocar el 1,5 se encontraban datos que son útiles para la investigación. Al realizar esta acción, en la Tabla 8 se presenta la cantidad de datos eliminados que representa un 0,36% de los datos sin estaciones, quedando con una totalidad de 94.629 considerados datos sin valores atípicos.

Tabla 8

Cantidad de datos eliminados.

	DV	NO2	O3	P	PM10	PM2.5	TA	VV	HA	Total
Bogotá	0	26	63	13	26	18	0	13	5	164
Bucaramanga	0	0	0	0	0	62	0	46	0	108
Cali	5	0	2	0	0	5	0	1	57	70
Total	5	26	65	13	26	85	0	60	62	342

Asimismo, se eliminan 16.456 datos no válidos correspondientes a un 14,129% del total de datos encontrados, de estos, 16.356 son datos pertenecientes a las variables PM10 y P que se encontraron en la ciudad de Bogotá y 100 registros del contaminante NO2 correspondientes a la ciudad de Cali, ya que, al no encontrar suficiente información para la ventana de tiempo establecida, no tienen incidencia sobre los datos finales, esta información se presenta en la **Tabla 9**.

[5] Q1 = Cuartil uno = 25% de los datos. Q3 = Cuartil tres = 75% de los datos. IQR = Rango intercuartil = Q3 – Q1.

Tabla 9

Datos no válidos.

	Datos sin valores atípicos	Datos ruidosos	Datos totales	% Datos ruidosos
Bogotá	66.110	16.356	49.754	24,74%
Bucaramanga	8.393	0	8.393	0%
Cali	20.126	100	20.026	0,497%
Total	94.629	16.456	78.173	17,39%

Realizada la limpieza de datos y como se pudo observar anteriormente, existen estaciones con datos faltantes para las fechas establecidas en la ventana de tiempo, estos datos estarían afectando la serie de tiempo, razón por la cual se decide no trabajar por estaciones sino por ciudad, promediando los valores registrados, y, así obtener datos más organizados y de mayor calidad a la hora de hacer el procesamiento. Adicionalmente, para la ciudad de Bucaramanga en las variables TA y HR, al observar su comportamiento se notó una tendencia constante, por lo cual, se decidió rellenar los datos faltantes con los datos del año anterior.

Seguidamente, se procede a realizar la imputación de datos utilizando el método k-means o k-vecino más próximo. En este, se itera sobre un valor mínimo de 3 y un valor máximo de la décima parte del total de datos, eligiendo el mejor k próximo. Teniendo el caso de que, si el mejor k próximo es mayor o igual a 300, se establece un valor máximo de 300 datos vecinos, así:

```
def get_regressor(x, y):
  min_k, max_k, step = 3, int(x.shape[0]/10), 3
  if max_k > 300:
```

max_k = 300

De acuerdo con lo anterior, se obtiene la **Tabla 10**, donde se observan los datos reemplazados por ciudad, dando como resultado 17.917 datos totales.

Tabla 10

Datos reemplazados por ciudad.

Ciudad	Datos promediados por ciudad	Datos reemplazados	Datos totales
Bogotá	6.599	2	6.601
Bucaramanga	5.052	606	5.658
Cali	5.541	117	5.658
Total	17.192	725	17.917

Por último, debido a que la calidad arrojada por el método de k-vecino no tuvo una buena representación del conjunto de datos para lapsus de tiempo mayores a 30 días, se implementaron estrategias de imputación como se muestra en la Tabla 11 para Bucaramanga y Cali, obteniendo datos más ajustados al problema.

Tabla 11

Estrategias de imputación implementadas.

Ciudad	Variable	Fechas	Decisión
Bucaramanga	PM2,5	18 Dic 2020 – 31 Jul 2021	Se reemplazó por el promedio de Bogotá y Cali.
	O3	4 Mar 2020 – 24 Sep 2020	
	DV	4 Nov 2019 – 29 Dic 2019	Se tomaron datos del 2020 de la misma ciudad.
	VV	30 Oct 2019 – 31 Dic 2019	
Cali	HA	29 Abr 2021 – 31 Jul 2021	Se tomaron datos del 2020 de la misma ciudad.
	TA	24 Abr 2021- 31 Jul 2021	
	O3	29 Jun 2020 – 1 de Oct 2020	Se tomaron datos del 2019 de la misma ciudad.

Teniendo en cuenta todo el procedimiento anterior, se obtuvieron finalmente 17.917

datos preprocesados, los cuales se harán uso en las etapas de procesamiento y minería de datos.

Ahora, se realiza un análisis descriptivo de la media, la desviación estándar, mínimos, máximos y gráficas de las variables, con el fin de conocer y estudiar su comportamiento.

En la Tabla 12 se encuentran tabuladas las estadísticas para los datos obtenidos de las variables meteorológicas y contaminantes para cada ciudad, como se muestra a continuación:

Tabla 12

Estadísticas obtenidas para cada variable en cada ciudad.

	Bogotá						
	DV	HR	NO2	O3	PM2.5	TA	VV
Cantidad de Datos	943	943	943	943	943	943	943
Promedio	172,502	65,022	24,540	48,904	33,214	15,138	1,549
Desviación estándar	33,466	5,553	8,685	12,265	10,684	0,843	0,347
Valor mínimo	103,891	44,333	6,078	18,034	12,496	12,407	0,627
25%	146,909	61,656	17,662	40,408	25,148	14,531	1,296
50%	165,951	64,792	23,725	48,499	32,098	15,099	1,543
75%	194,425	68,653	30,754	57,188	39,440	15,738	1,791
Valor máximo	280,427	79,714	49,960	88,527	84,639	17,428	2,517
	Bucaramanga						
	DV	HR	NO2	O3	PM2.5	TA	VV
Cantidad de Datos	943	943	0	943	943	943	943
Promedio	218,201	77,454	0	22,315	15,985	25,012	1,192
Desviación estándar	29,351	4,482	0	8,992	8,550	0,95	0,253
Valor mínimo	112,969	64,979	0	4,996	4,435	22,454	0,408
25%	198,490	74,937	0	15,132	9,983	24,293	1,012
50%	219,969	77,353	0	22,071	13,721	25,008	1,168
75%	239,612	80,531	0	28,743	18,953	25,666	1,358
Valor máximo	284,731	89,208	0	49,342	47,836	27,567	1,913
	Cali						
	DV	HR	NO2	O3	PM2.5	TA	VV
Cantidad de Datos	943	943	0	943	943	943	943
Promedio	185,449	69,153	0	28,804	16,128	25,726	1,799

Desviación estándar	16,298	6,138	0	7,910	6,099	1,318	0,478
Valor mínimo	124,794	43,294	0	9,988	3,588	20,957	0,905
25%	173,818	65,507	0	22,882	11,487	24,901	1,456
50%	185,384	68,739	0	28,045	15,739	25,849	1,704
75%	196,976	73,107	0	34,239	20,302	26,704	2,112
Valor máximo	233,752	89,018	0	58,400	40,460	28,861	3,677

Posteriormente, se encuentra el análisis de cada variable.

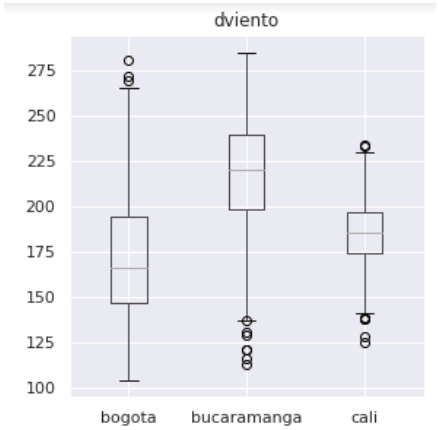
5.2.1. Dirección del viento.

De acuerdo con la Tabla 12, se puede observar que esta variable presenta valores que oscilan entre 103,9 y 284,7 grados, interpretándose como vientos del este y del sur principalmente.

Figura 38), se visualiza la dispersión de estos datos para las 3 ciudades, para Bogotá se encuentra una asimetría positiva y se identifican datos atípicos sobre el límite superior del diagrama; para Bucaramanga se encuentra una asimetría negativa y datos atípicos debajo del límite inferior, mientras que para Cali se presenta una distribución simétrica, pero con datos atípicos por encima y por debajo de los límites. Estos valores atípicos que se presentan en las ciudades se deben a los cambios de temperatura que ocurren durante cada día.

Figura 38

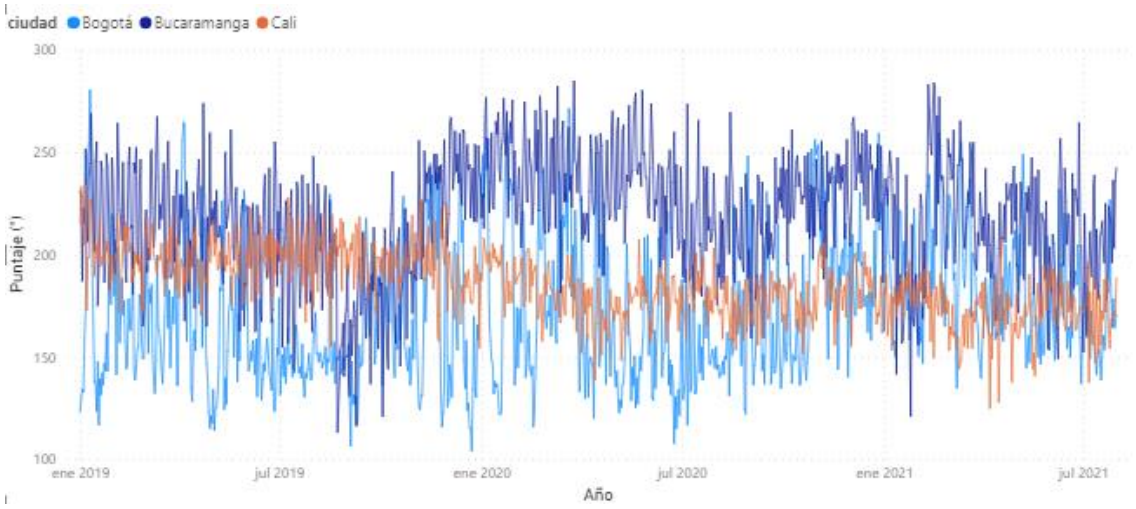
Diagrama de Cajas y Bigotes para DV.



También se puede observar en la Figura 39, que dentro de la ventana de tiempo escogida se observan valores más dispersos para Bogotá.

Figura 39

Comportamiento DV para las 3 ciudades.



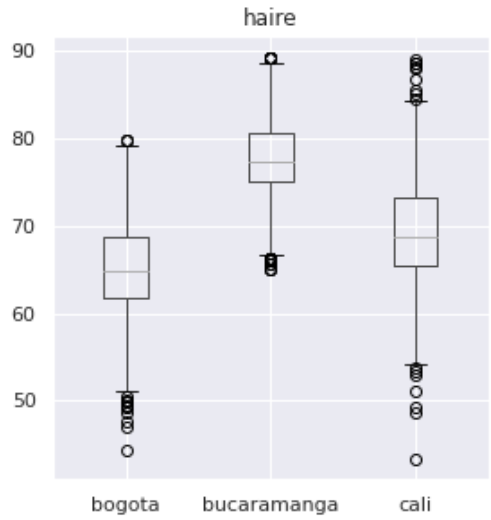
5.2.2. Humedad Relativa del aire.

Esta variable presenta valores en un rango de 43,3 a 89,2%, en la Figura 40 se observa que los 3 diagramas están presentando una distribución asimétrica positiva, también es notorio que existen datos atípicos por encima y por debajo de los bordes de los bigotes en cada diagrama.

Es importante resaltar que los altos niveles de la Humedad Relativa contribuyen con la transmisión y alarga la vida de los virus de la gripe, generando complicaciones en la salud de las personas que padecen COVID 19.

Figura 40

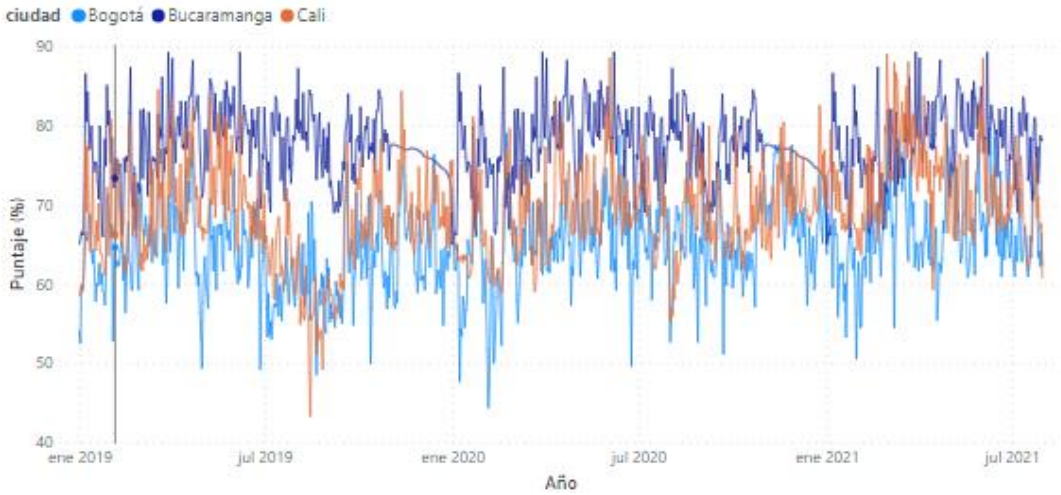
Diagrama de cajas y bigotes para HR.



Para Bucaramanga se visualiza en la Figura 41, el rellenado de datos replicando los del año 2019, adicionalmente se observa que Cali y Bogotá tienen comportamientos similares, este último presentando valores más bajos.

Figura 41

Comportamiento HR para las 3 ciudades.



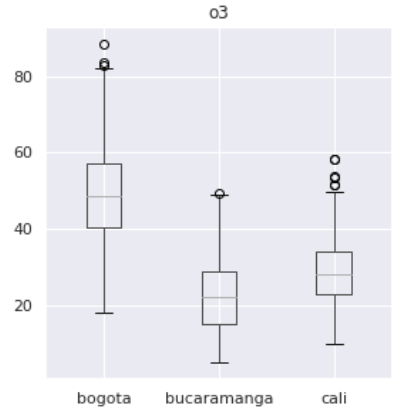
5.2.3. Ozono Troposférico.

En la Figura 42 se puede observar que para esta variable se encuentran pocos valores atípicos sobre el límite superior para las 3 ciudades, reportando así los máximos valores: Cali con 58,4 $\mu g/m^3$, Bogotá 88,5$\mu g/m^3$ y Bucaramanga 49,3 $\mu g/m^3$, estos puntajes se encuentran dentro el rango “Bueno” para la salud según la Tabla 4. Asimismo, se observa que los diagramas presentan distribución normal.

Altos niveles de esta variable contaminante pueden ocasionar insuficiencia pulmonar, empeorando los síntomas del COVID 19. Es decir, Bogotá tiene mayor incidencia sobre esta enfermedad en comparación con Bucaramanga y Cali.

Figura 42

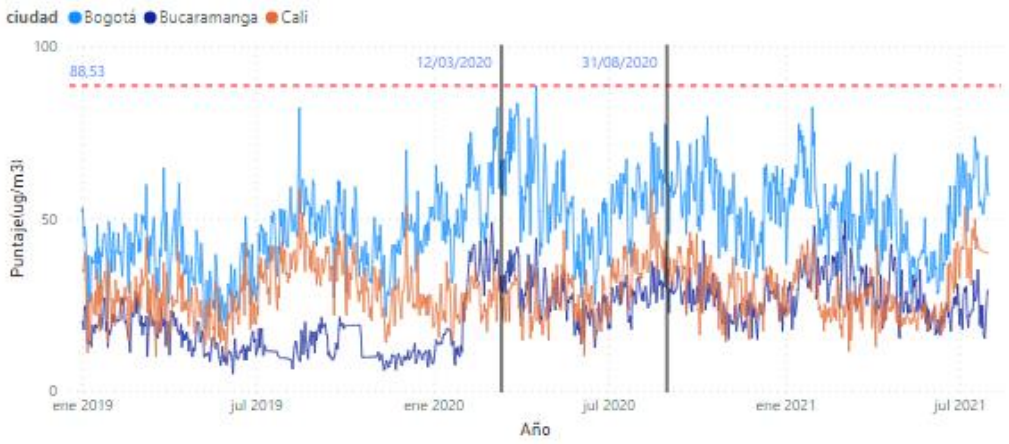
Diagrama de cajas y bigotes para O3.



De este contaminante se puede afirmar que aumentó significativamente en las 3 ciudades a partir de marzo del 2020, siendo Bogotá la ciudad que más presenta altos niveles de contaminación, como se observa en la Figura 43.

Figura 43

Comportamiento O3 para las 3 ciudades.



5.2.4. Material Particulado 2.5.

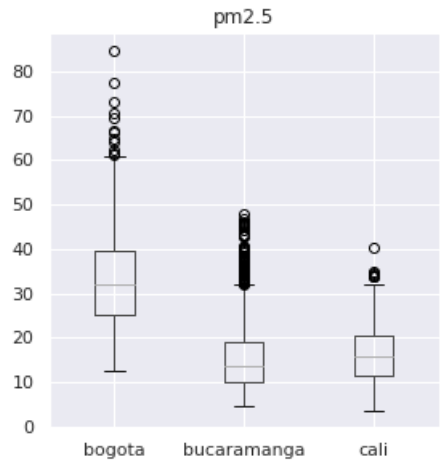
El comportamiento de esta variable para las 3 estaciones presenta valores atípicos y,

además, presenta una asimetría positiva como se observa en la Figura 44 para Bogotá y Bucaramanga, es decir, los datos se concentran en la parte inferior de cada distribución. Así mismo, se puede observar que Bogotá presenta valores más altos comparados con Bucaramanga y Cali, estos alcanzarían una clasificación “Dañina para la salud”, mientras que los valores altos de Bucaramanga y Cali estarían distribuidos en una clasificación entre “Buena” y “Dañina para la salud de los grupos sensibles”.

Con relación a los altos valores dañinos, estos pueden ser producidos por la alta congestión vehicular y el retorno de las empresas después de los confinamientos.

Figura 44

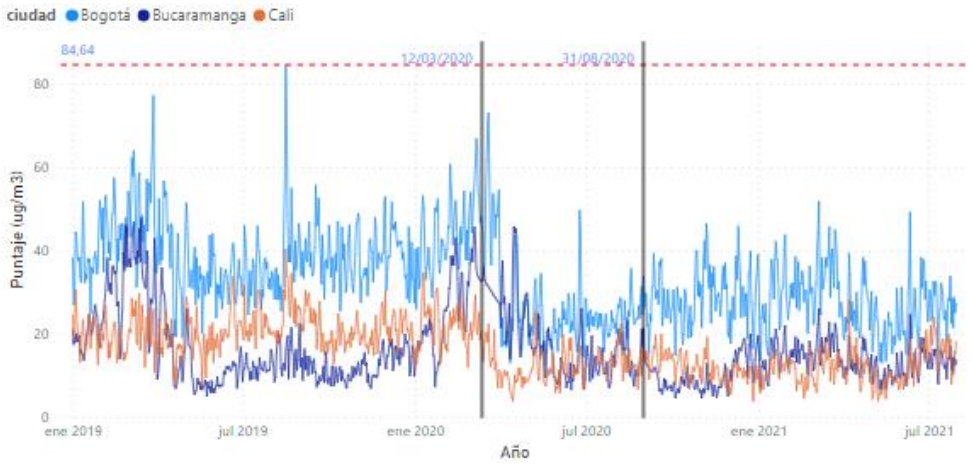
Diagrama de cajas y bigotes para PM2.5.



En la Figura 45 se observa que durante el confinamiento se presentaron picos significativos para las 3 ciudades, sin embargo, a partir de abril del 2020, los valores decrecieron tomando una tendencia de poca variación, a diferencia del 2019 donde estos eran más altos en la ventana de tiempo escogida.

Figura 45

Comportamiento PM2.5 para las 3 ciudades.

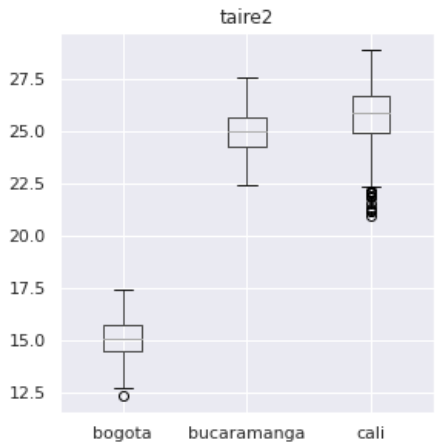


5.2.5. Temperatura del Aire.

Con la Figura 46 se puede observar que el rango de valores para la Temperatura es similar en Bucaramanga y Cali, debido claramente a la ubicación geográfica de estas. Con respecto a la simetría, Bogotá y Bucaramanga presentan una distribución simétrica y Cali, está sesgada ligeramente a la izquierda.

Figura 46

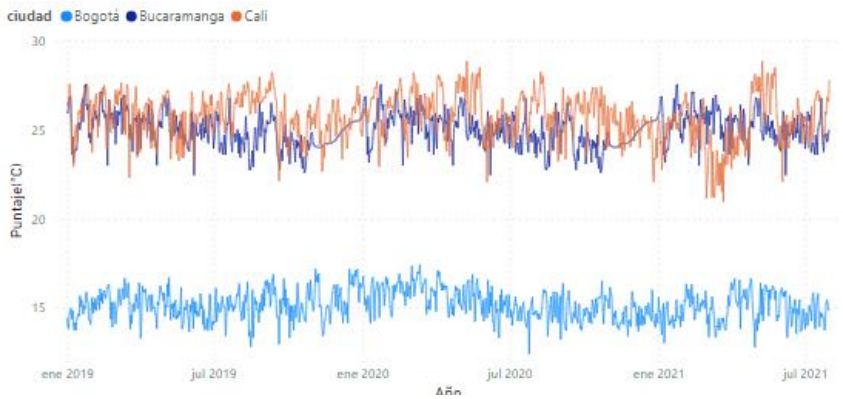
Diagrama de cajas y bigotes para TA.



Para Bucaramanga en la Figura 47 se aprecia la decisión que se tomó en la imputación de datos al duplicar el rango de valores desde el 01 de enero de 2019 hasta el 01 de noviembre de 2019.

Figura 47

Comportamiento TA para las 3 ciudades

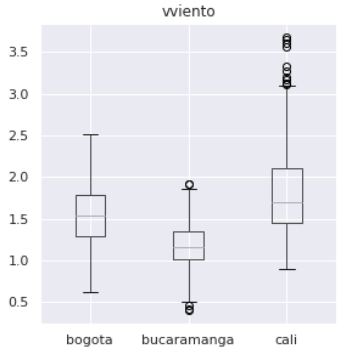


5.2.6. Velocidad del Viento.

Se puede observar en la Figura 48 que Bucaramanga y Cali presentan una distribución asimétrica positiva y datos atípicos, mientras que Bogotá muestra simetría y sin valores atípicos.

Figura 48

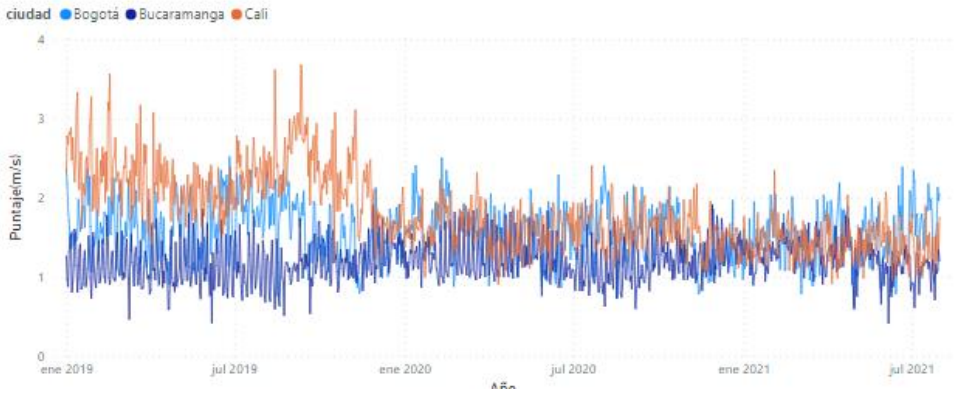
Diagrama de cajas y bigotes para VV



Los valores no presentan una alta variación y son similares para las 3 ciudades, a partir de enero de 2020, situación representada en la Figura 49.

Figura 49

Comportamiento VV para las 3 ciudades.



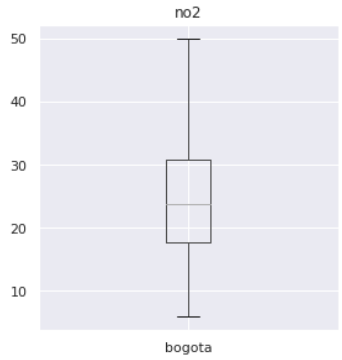
5.2.7. Dióxido de Nitrógeno

En la Figura 50, se encuentra la distribución de NO2 para la ciudad de Bogotá, allí se puede observar que presenta una distribución sesgada ligeramente a la derecha, sin ningún dato atípico y que varía entre 6,1 y 50 $\mu g/m^3$, categorizándose como “Bueno” según la Tabla 4.

El contaminante NO2 es un irritante que afecta a la mucosa de los ojos, nariz, garganta y vías respiratorias. Dosis altas pueden provocar edema y lesiones pulmonares, y una exposición continua a estas dosis contribuyen al desarrollo de bronquitis aguda o crónica. Para los pacientes contagiados con COVID-19 el estar expuestos a este contaminante aumenta el riesgo a una sintomatología más grave (EPA, s.f.) .

Figura 50

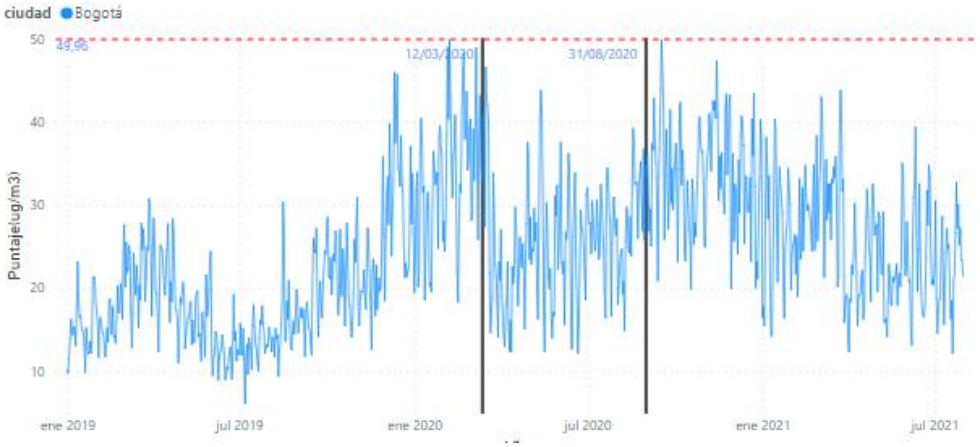
Diagrama de cajas y bigotes para NO2.



En la Figura 51 para la ciudad de Bogotá, se observa un notorio aumento en los valores de este contaminante, el cual se mantuvo en rangos superiores a partir del confinamiento, en comparación con el año 2019.

Figura 51

Comportamiento NO2 para Bogotá.



5.3. Etapa de procesamiento

Una vez hecha la limpieza e imputación de datos, llega la etapa de procesamiento en la cual se procede a ajustar los modelos de Deep Learning. Para el desarrollo de este proyecto

se aplicó para cada ciudad un modelo MLP, LSTM y Seq2Seq como se muestra a continuación.

5.3.1. Modelo MLP

Una vez obtenidos los datos preprocesados, estos se utilizan para el ajuste de los modelos, en esta red cada neurona en una capa está conectada a todas las neuronas de la capa anterior y de la capa siguiente, formando una red de conexiones ponderadas. Durante el entrenamiento, el modelo ajusta los pesos de estas conexiones para minimizar una función de pérdida, lo que permite que el modelo aprenda a mapear las entradas a las salidas deseadas siendo muy popular y versátil en el uso exitoso de una amplia variedad de aplicaciones, como reconocimiento de voz, reconocimiento de imágenes, análisis de sentimientos y predicción de series temporales, entre otros.

Para el presente proyecto, como se observa en la Figura 52, la red neuronal consta de varias capas densas, cada una con 64 unidades. La profundidad de la red se estableció en 8 capas para capturar mejor las relaciones no lineales entre las características de entrada y la salida, y se determinó una cantidad de 40 Epochs. Para la selección de la función de activación se encontró en la literatura que la más utilizada para las redes neuronales modernas como la MLP y CNN es la función ReLU (Goodfellow et al., 2016), sin embargo, después de realizar la exploración de hiperparametros, para este trabajo de investigación la función que más se ajustó al conjunto de datos fue tanh.

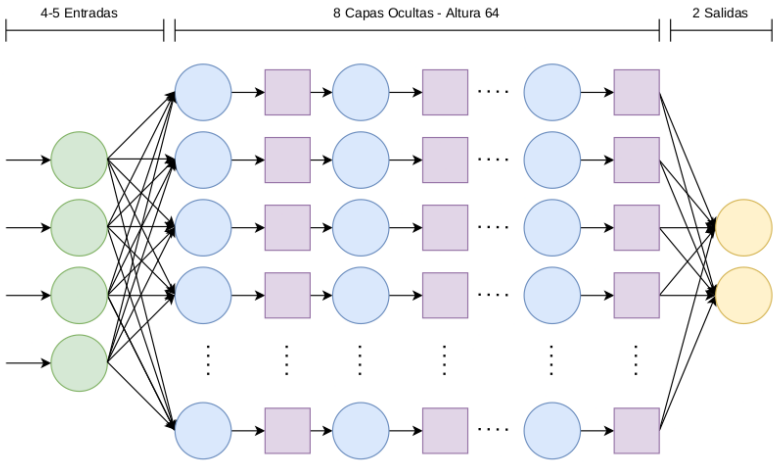
Para prevenir el sobreajuste, se aplicó la técnica de regularización Dropout, que aleatoriamente desactiva un porcentaje de las unidades en la capa anterior durante el entrenamiento. La normalización de lotes no se utilizó en esta implementación, ya que no mejoró significativamente los resultados en los experimentos.

Se estableció un tamaño de lote de 8, lo que significa que se utilizan 8 ejemplos de

entrenamiento en cada iteración del algoritmo. La tasa de aprendizaje se estableció en 0.01 para controlar la cantidad en que los pesos de la red neuronal se ajustan durante el entrenamiento. La función de pérdida seleccionada fue la función de error cuadrático medio (MSE), adecuada para problemas de regresión en los que la salida esperada es un valor numérico y comúnmente utilizada en la literatura (Martinez H, 2020). Con esta, se mide la discrepancia entre las predicciones del modelo y los datos reales.

Figura 52

Arquitectura del modelo MLP.



5.3.2. Modelo LSTM

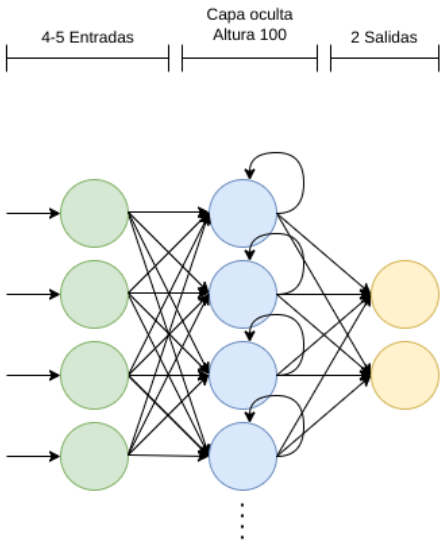
Para predecir datos a partir de una secuencia temporal de datos pasados por una capa oculta de 100 unidades, una función de activación Tanh, sin Dropout, sin Batch Normalization, una tasa de aprendizaje de 0.04, una función de pérdida MSE, 120 Epochs y un tamaño de lote de 8. Estos parámetros se seleccionaron después de realizar pruebas en diferentes combinaciones y seleccionar los que proporcionaron los mejores resultados. La arquitectura de esta red se presenta en la Figura 53.

Cabe destacar que una red LSTM puede procesar secuencias de datos de longitud

variable, esto significa que puede procesar datos en orden y detectar patrones y relaciones en los datos a lo largo del tiempo. Del mismo modo, tiene conexiones recurrentes entre las neuronas de la capa oculta, lo que permite que la información anterior se tenga en cuenta en el procesamiento de la información presente. Esto es especialmente útil en la regresión de datos temporales, donde las observaciones anteriores son válidas para predecir las observaciones futuras.

Figura 53

Arquitectura del modelo LSTM.



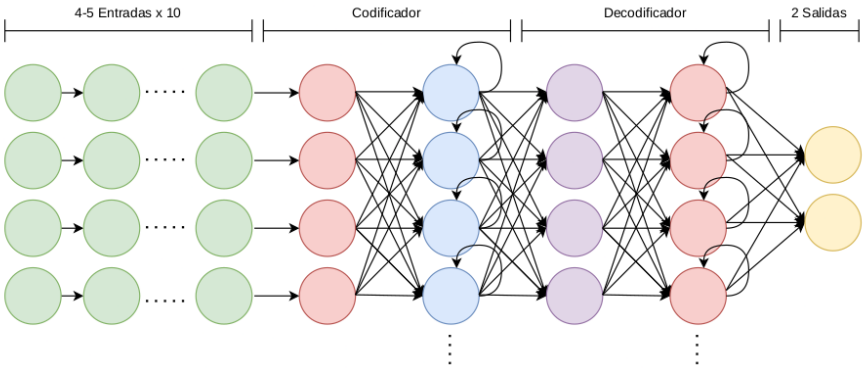
5.3.3. Modelo Seq2Seq

Finalmente, se ajusta un modelo Seq2Seq para predecir datos a partir de una secuencia temporal de 10 datos pasados por una capa oculta codificadora de 64 unidades y otra capa decodificadora de las mismas características, una función de activación tanh, sin Dropout, sin Batch Normalization, una tasa de aprendizaje de 0.04, una función de pérdida MSE, 12 Epochs y un tamaño de lote de 8. En la Figura 54 se muestra la arquitectura

seleccionada.

Figura 54

Arquitectura del modelo Seq2Seq.



Las partes principales de esta red son: un codificador y un decodificador. La ventaja clave se presenta cuando el codificador toma una secuencia de entrada de longitud variable que encapsula toda la información relevante de la secuencia y la convierte en un vector de características fijo, el decodificador toma ese vector de características y lo utiliza para generar una secuencia de salida de longitud variable que corresponde a la secuencia de entrada.

De igual manera, los sistemas de codificador y decodificador pueden capturar relaciones más complejas entre las secuencias de entrada y salida, trabajan juntos para aprender una representación de la secuencia de entrada que es útil para generar la secuencia de salida correspondiente. Esto permite al modelo capturar más información sobre la relación entre los datos, lo que puede mejorar la precisión de las predicciones.

En conclusión, para esta etapa se obtienen los modelos ajustados al conjunto de datos seleccionado gracias a la exploración de hiperparámetros, con el fin de obtener un buen rendimiento en la siguiente etapa.

5.4. Etapa de minería de datos

Con las configuraciones establecidas en la etapa de procesamiento, se procede a iniciar el entrenamiento de cada red para posteriormente predecir los contaminantes en las ciudades de Bogotá, Bucaramanga y Cali y así evaluar el mejor modelo de predicción.

5.4.1. Modelo MLP

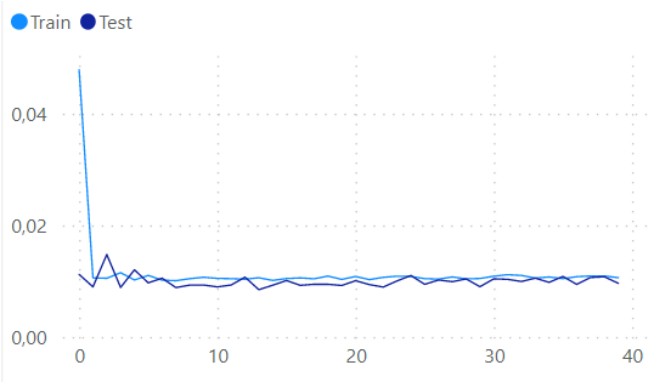
Para el entrenamiento se tomó un 90% del total de datos de cada ciudad mezclados aleatoriamente con el objetivo de poder realizar pruebas con datos que la red neuronal no ha visto antes y permitir la identificación de casos de sobreajuste. Se obtienen las gráficas para identificar el comportamiento de la función de costo para los datos de prueba (val_loss) y para los datos de entrenamiento (loss).

5.4.1.1.Bogotá

La cantidad de datos utilizada fue de 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 55 se puede observar que ocurre un ajuste de los datos a partir de la séptima iteración, presentando un error promedio de 0,01154 para entrenamiento y 0,00996 para la prueba.

Figura 55

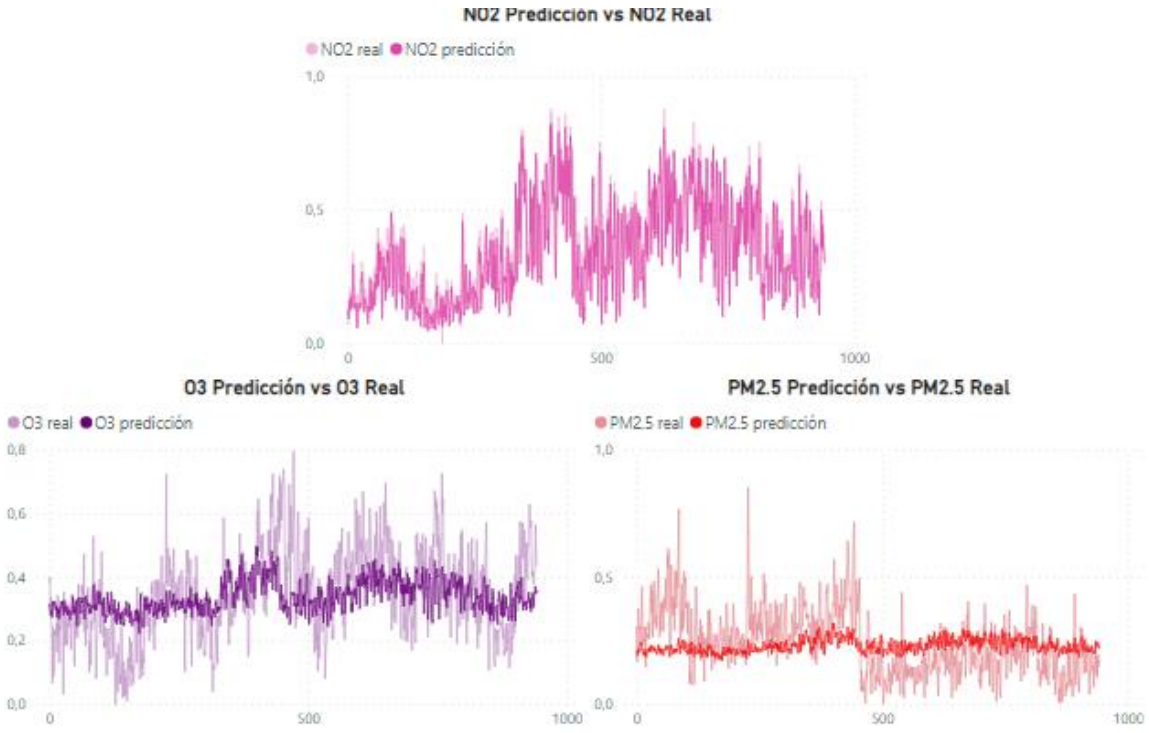
Función de pérdida modelo MLP para Bogotá.



En la Figura 56 se observan las predicciones para los 943 datos con cada uno de los contaminantes (NO2, PM2.5, O3), allí se puede evidenciar que el contaminante NO2 tuvo un buen ajuste a los datos reales, sin embargo, la predicción de PM2.5 y O3 presenta un resultado centrado en el promedio de los datos reales, es decir, el modelo no logra predecir adecuadamente estos contaminantes.

Figura 56

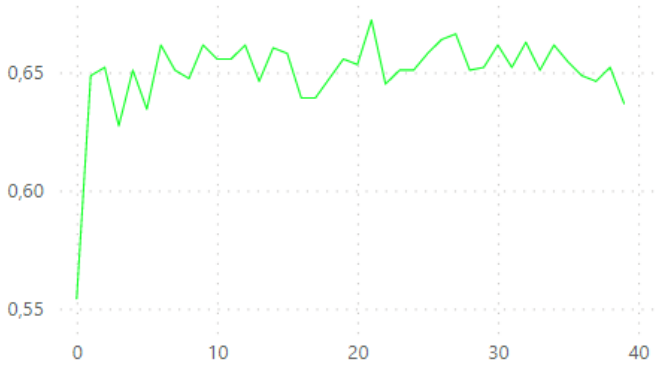
Predicción de los contaminantes modelo MLP para Bogotá.



De acuerdo con lo anterior, el grado de concordancia entre la predicción y los datos reales se representa con la función de precisión mostrada en la Figura 57, la cual evidencia que existe en promedio una precisión de 0,65 que puede ser debida a PM2.5 y O3 por presentar una baja calidad en la predicción.

Figura 57

Función de precisión modelo MLP para Bogotá.

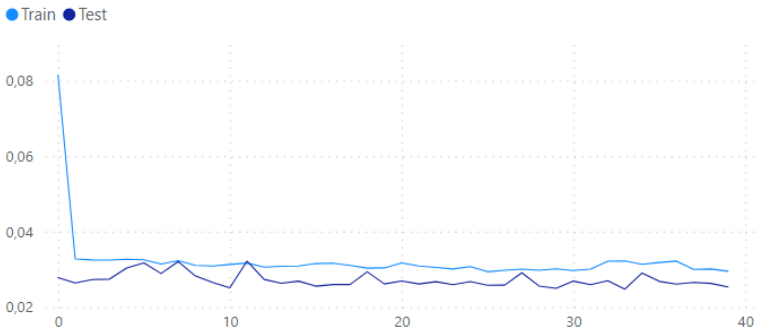


5.4.1.2.Bucaramanga

Para el entrenamiento se utilizaron 5.092 datos y para la validación 566, en esta ocasión el ajuste de los datos ocurre a partir de la iteración 12 como se representa en la Figura 58. El error cuadrático medio tuvo un promedio de 0,03226 para entrenamiento y 0,02709 para la prueba.

Figura 58

Función de pérdida modelo MLP para Bucaramanga.

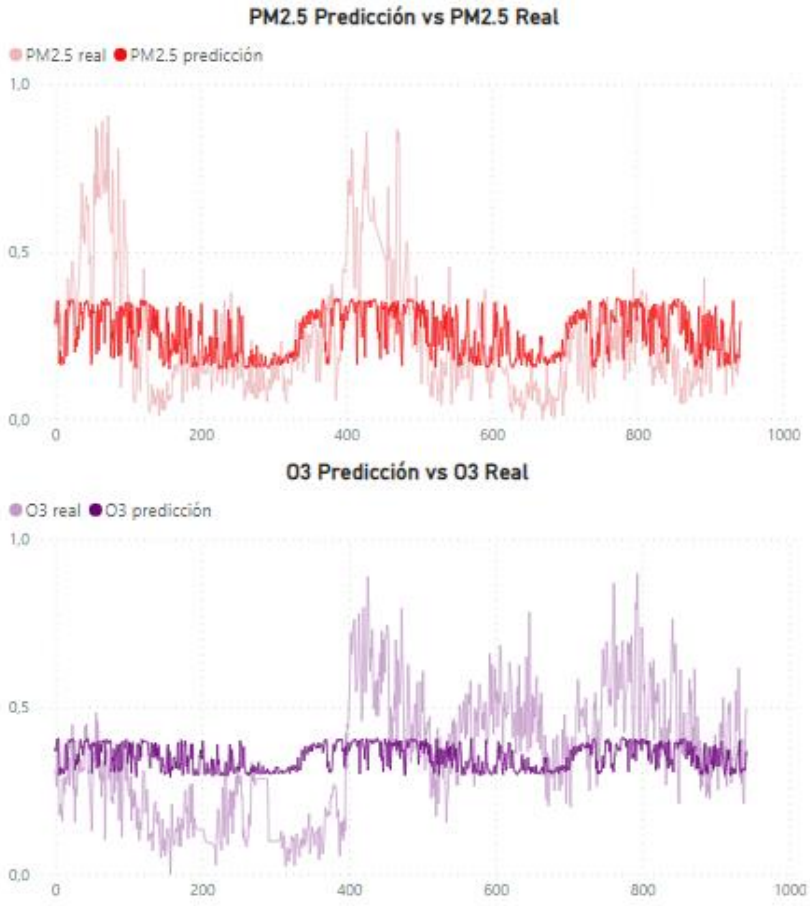


En la Figura 59 se observan las predicciones para la ciudad de Bucaramanga con los

contaminantes PM2.5 y O3, en los cuales se presenta un bajo rendimiento del modelo del mismo modo que ocurre en la ciudad de Bogotá, la predicción se centra en el promedio de los datos reales, pero no es capaz de predecir su variación.

Figura 59

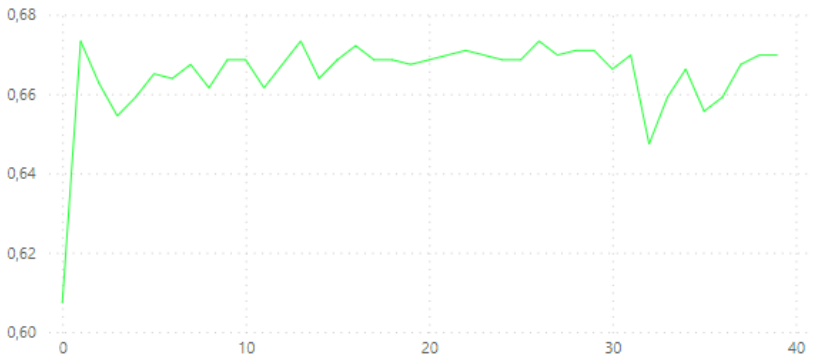
Predicción de los contaminantes modelo MLP para Bucaramanga.



La calidad de esta predicción se ve reflejada en la Figura 60, en donde la precisión del modelo tiene como promedio 0,66, siendo aún muy pobre para considerar que este modelo sea el óptimo.

Figura 60

Función de precisión modelo MLP para Bucaramanga.

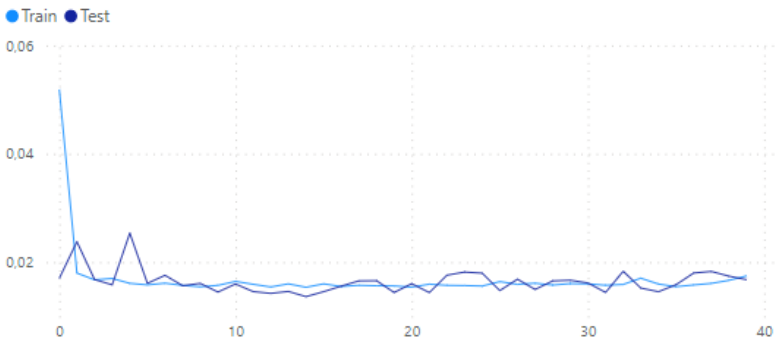


5.4.1.3. Cali

Para el entrenamiento se utilizaron 5.092 datos y para la validación 566, en la Figura 61 se evidencia un ajuste de los datos a partir de la décima iteración. El error cuadrático medio tuvo un promedio de 0,016925 para entrenamiento y 0,016462 para la prueba. A pesar de que este es bastante bajo, se observa que los datos de prueba están presentando variaciones en comparación con los de entrenamiento.

Figura 61

Función de pérdida modelo MLP para Cali.

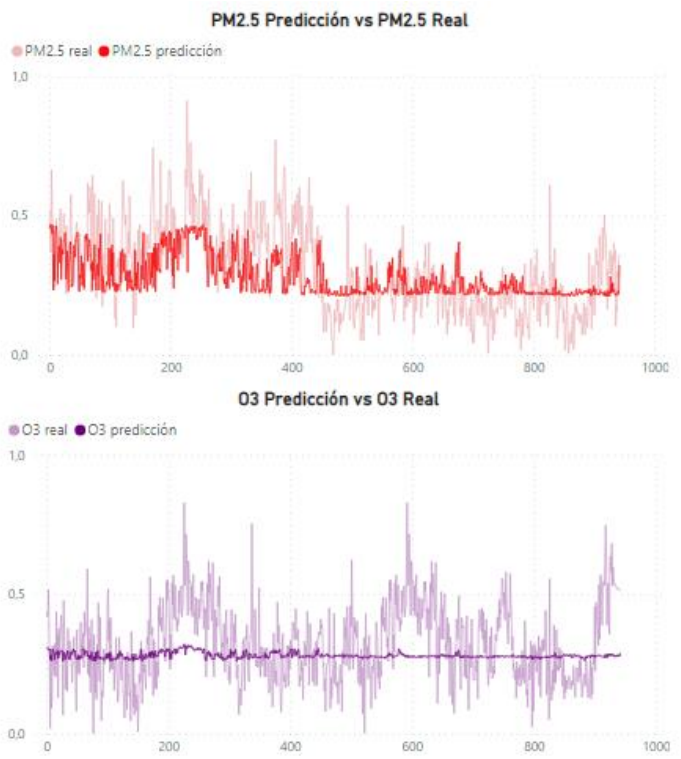


Del mismo modo que ocurre con las ciudades de Bogotá y Bucaramanga, en la

predicción de los contaminantes PM2.5 y O3 para Cali se observa en la Figura 62 que el comportamiento de los datos predichos se ubica hacia el promedio de los datos reales, lo cual, se considera que el modelo no se ajusta al tipo de problema que se está abarcando.

Figura 62

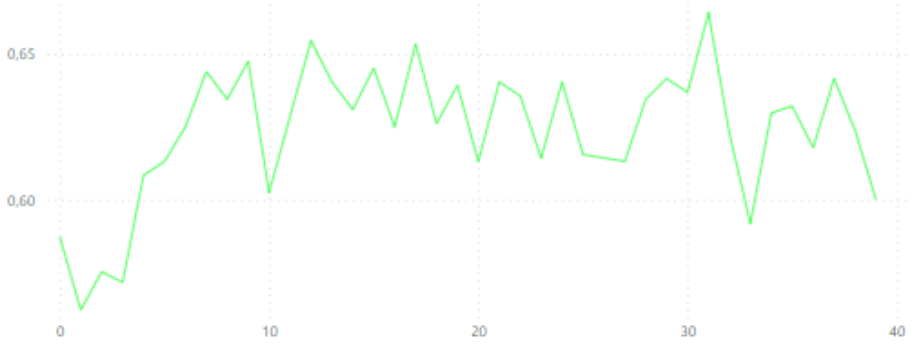
Predicción de los contaminantes modelo MLP para Cali.



Por último, en la Figura 63 se muestra la función de precisión para esta ciudad, donde se corrobora que la precisión del modelo es pobre con un promedio de 0,62.

Figura 63

Función de precisión modelo MLP para Cali.



5.4.2. Modelo LSTM

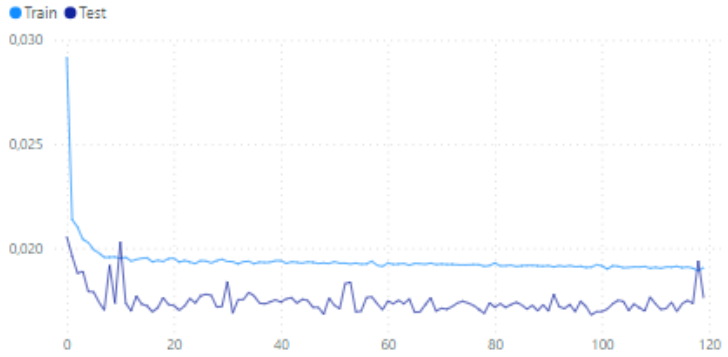
Para el entrenamiento se seleccionó un 90% del total de datos de cada ciudad mezclados aleatoriamente. De igual manera que el modelo anterior, se obtienen gráficas de predicción, pérdida y precisión que permiten analizar el comportamiento de esta arquitectura.

5.4.2.1. Bogotá

La cantidad de datos utilizada fue de 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 64 se puede observar que ocurre un ajuste de los datos a partir de la décima tercera iteración, presentando un error promedio de 0,01938 para entrenamiento y 0,01744 para la prueba. Se evidencia que en la iteración 118 la validación de los datos aumenta de nuevo, traduciéndose en un posible sobreajuste por una cantidad grande de Epochs.

Figura 64

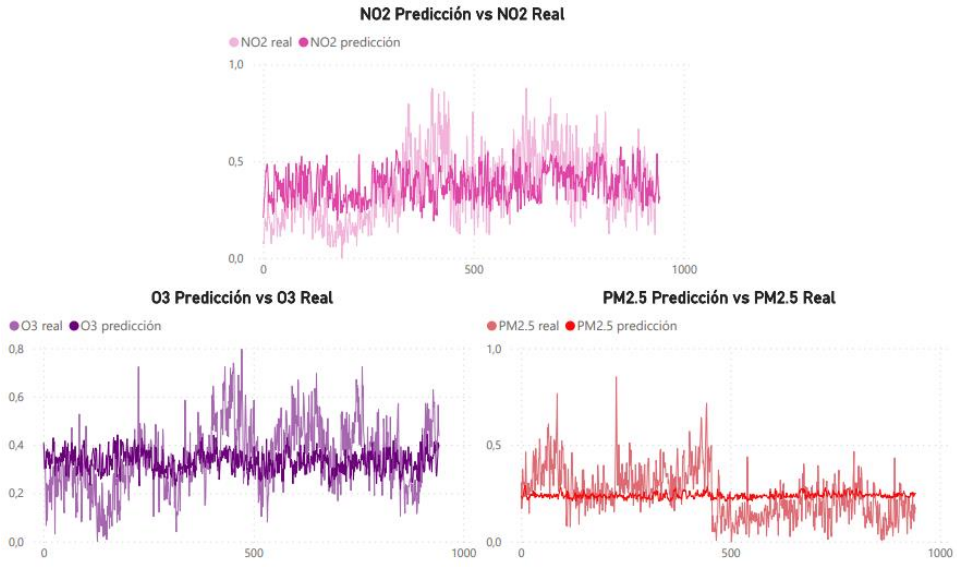
Función de pérdida modelo LSTM para Bogotá.



En la predicción presentada en la Figura 65, se analiza que los contaminantes NO2 y O3 intentan ajustarse más a la variación de los datos reales, sin embargo, no logran acercarse a los límites superiores e inferiores. Por otra parte, el contaminante PM2.5 continúa presentando un comportamiento poco eficiente debido a la naturaleza de sus datos.

Figura 65

Predicción de los contaminantes modelo LSTM para Bogotá.



Teniendo en cuenta lo anterior, la función de precisión para esta ciudad se presenta

en la Figura 66, de la cual se puede decir que no se observa gran variación y tiende a estabilizarse entre un 57 y 60 % a partir del Epoch 5. Se deduce que no se obtiene un valor más alto a causa del contaminante PM2.5.

Figura 66

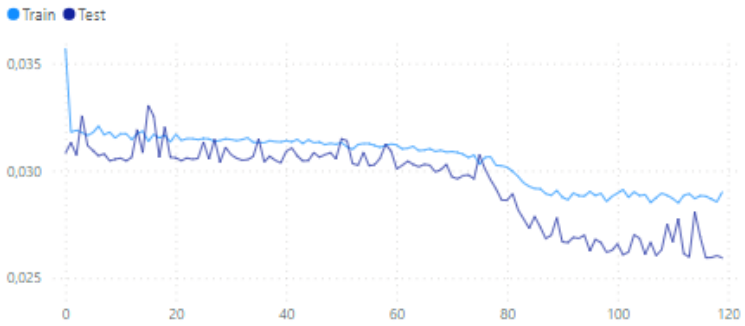
Función de precisión modelo LSTM para Bogotá.

5.4.2.2. Bucaramanga

Para la ciudad de Bucaramanga, se utilizaron 5.092 datos para el entrenamiento y para la prueba 566, en esta ocasión el ajuste de los datos ocurre a partir de la iteración 20 como se representa en la Figura 67, no obstante, se evidencia que a partir de la iteración 75 disminuye la pérdida tanto para el entrenamiento como para la validación en un valor aproximado a 0,002, el cual no es significativo. El error cuadrático medio tuvo un promedio de 0,03054 para entrenamiento y 0,02938 para la validación.

Figura 67

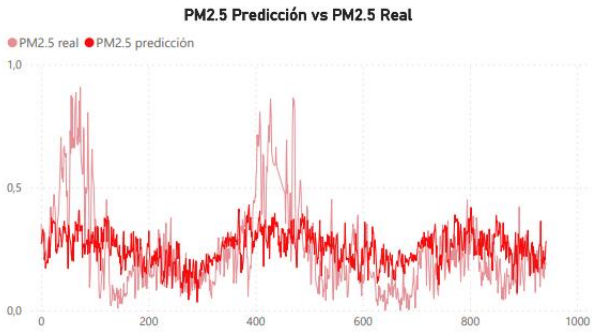
Función de pérdida modelo LSTM para Bucaramanga.

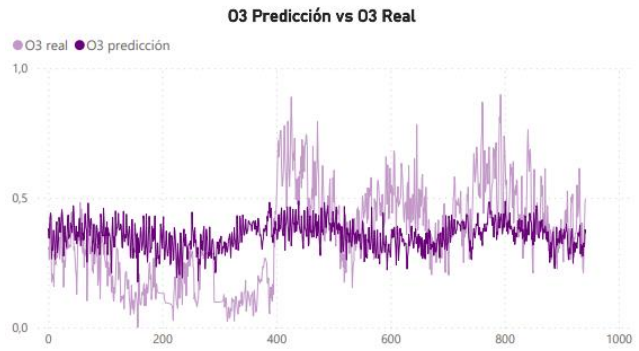


En la Figura 68, se muestra la predicción correspondiente a los contaminantes PM2.5 y O3, en la cual se evidencia que los datos predichos no se ajustan a los datos reales, es decir, no alcanzan a ajustarse a la variación real siendo más notorio este caso en O3. Adicionalmente, en PM2.5 se observa que la predicción intenta imitar el comportamiento de los datos reales contrario a lo sucedido en O3.

Figura 68

Predicción de los contaminantes modelo LSTM para Bucaramanga.

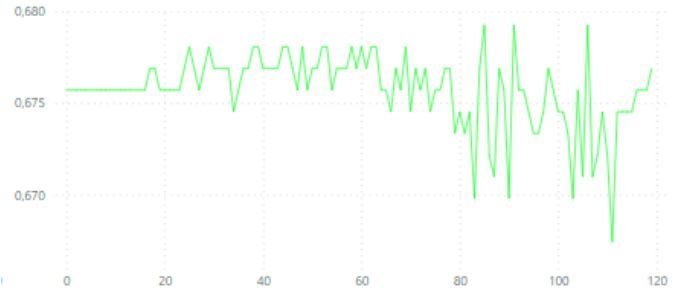




El análisis anterior se ve respaldado con la Figura 69, donde la precisión de la predicción presenta un valor de 0,68 en promedio, notándose una variación insignificante entre 0,67 y 0,68 a partir de la iteración 75, misma incidencia con la función de pérdida presentada en la Figura 68.

Figura 69

Función de precisión modelo LSTM para Bucaramanga.

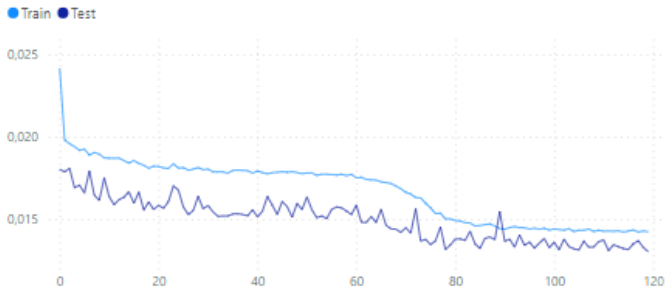


5.4.2.3. Cali

En Cali, 5.092 datos fueron utilizados para entrenar al modelo y 566 para validarlo, en la Figura 70 se evidencia un mayor ajuste de los datos y una disminución en la pérdida a partir del Epoch 80. El error cuadrático medio tuvo un promedio de 0,01670 para entrenamiento y 0,01490 para la prueba.

Figura 70

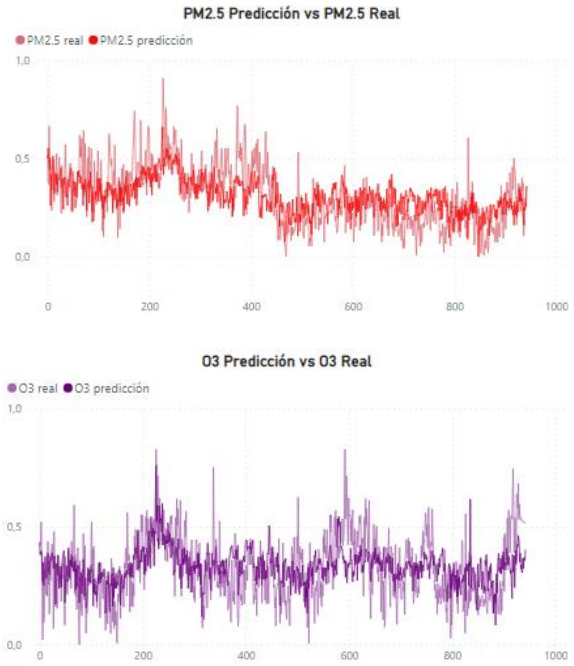
Función de pérdida modelo LSTM para Cali.



En este sentido, en la Figura 71 se evidencia que los datos predichos a comparación con las ciudades anteriores se ajustan un poco más a los datos reales para ambos contaminantes, siguiendo su comportamiento, aunque no logran con éxito una predicción óptima.

Figura 71

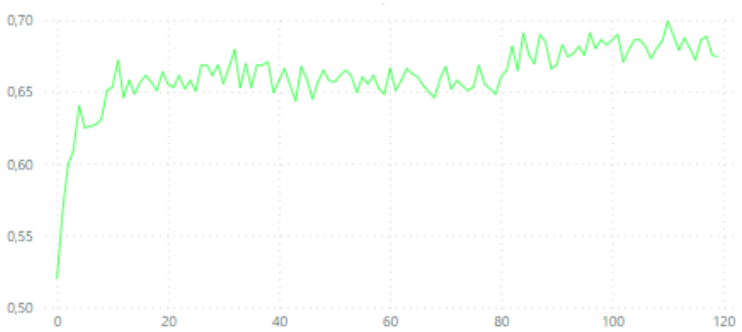
Predicción de los contaminantes modelo LSTM para Cali.



Lo anterior se ve reflejado también en la Figura 72, donde la precisión toma valores mayores a comparación de las otras ciudades analizadas, aunque no son altamente significativos, oscilando entre 0,65 y 0,7.

Figura 72

Función de precisión modelo LSTM para Cali.



5.4.3. Modelo Seq2Seq

Para el entrenamiento se seleccionó un 90% del total de datos de cada ciudad agrupados en secuencias de 10. Se obtienen gráficas de predicción, pérdida y precisión que permiten analizar el comportamiento de este modelo. Adicionalmente, la configuración de la cantidad de los Epochs se decidió en 12, ya que, al aumentar el número de iteraciones, el modelo se sobreajustaba observándose en la función de pérdida, mientras el comportamiento de la validación de los datos aumentaba a medida que aumentan los Epochs, el comportamiento de la prueba disminuía.

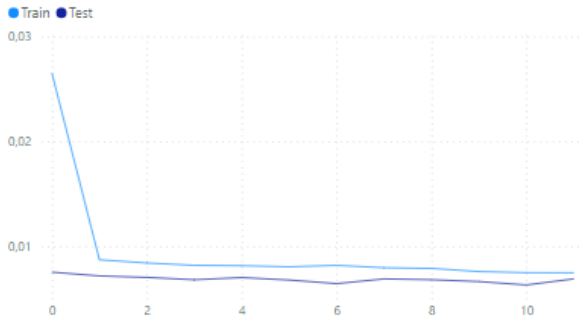
5.4.3.1. Bogotá

En la ciudad de Bogotá, se utilizaron 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 73 se muestra que ocurre un ajuste de los datos a partir de la primera iteración, presentando un error promedio de 0,00887 para entrenamiento y

0,00685 para la prueba.

Figura 73

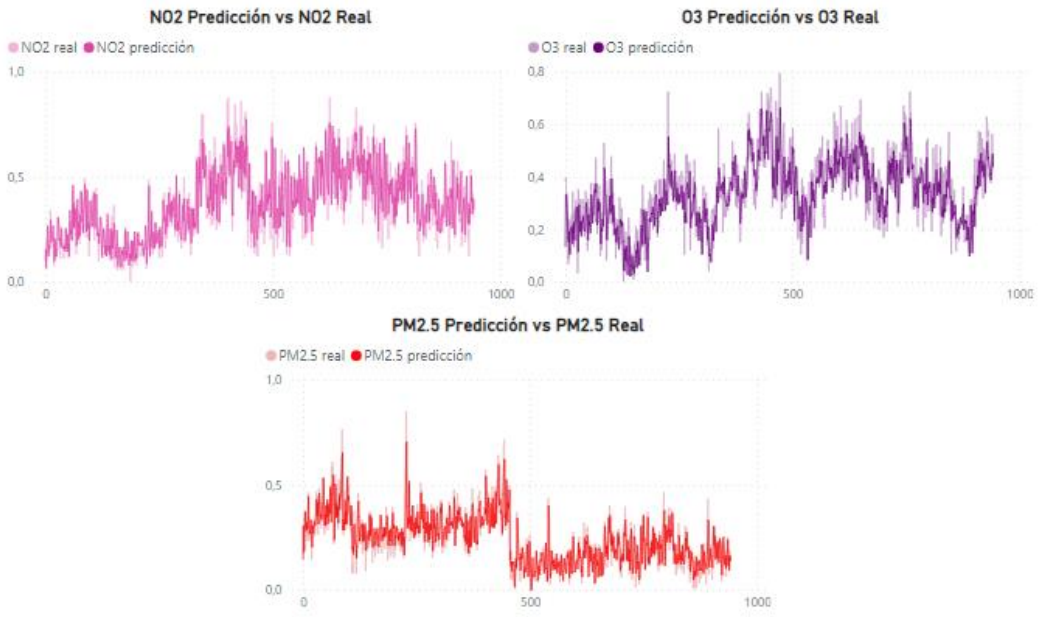
Función de pérdida modelo Seq2Seq para Bogotá.



En la Figura 74, se evidencia que el modelo está realizando una predicción óptima para los 3 contaminantes, ya que se ajusta a los datos reales y sigue el comportamiento de la naturaleza de estos, a diferencia de lo ocurrido en los modelos anteriores.

Figura 74

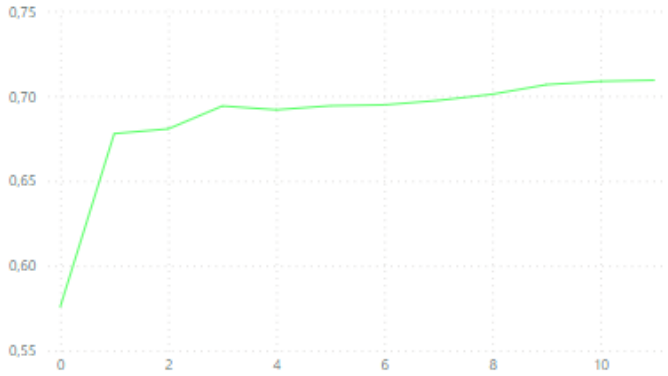
Predicción de los contaminantes modelo Seq2Seq para Bogotá.



A continuación, en la Figura 75, se observa que la precisión de este modelo para Bogotá se encuentra entre 0,68 y 0,71 a partir de la primera iteración, respaldando lo analizado anteriormente.

Figura 75

Función de precisión modelo Seq2Seq para Bogotá.

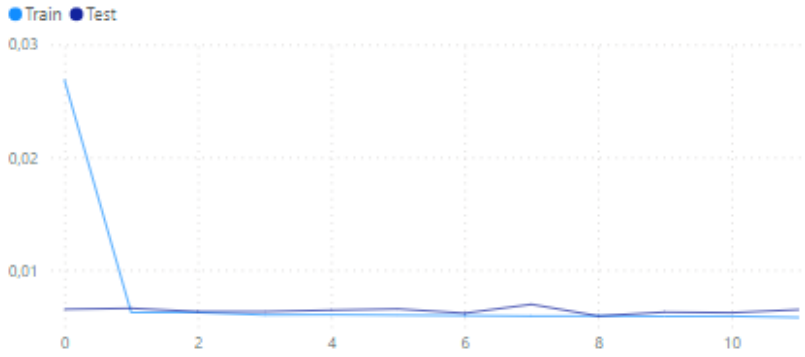


5.4.3.2. Bucaramanga

Los datos utilizados en esta ocasión fueron 5.092 datos para el entrenamiento y 566 para la validación. En la Figura 76 se refleja que el ajuste de los datos ocurre a partir de la primera iteración, siendo el promedio del error cuadrático medio de 0,007723 para entrenamiento y 0,006386 para la prueba. Estos son los más bajos para esta ciudad en comparación con los anteriores modelos analizados.

Figura 76

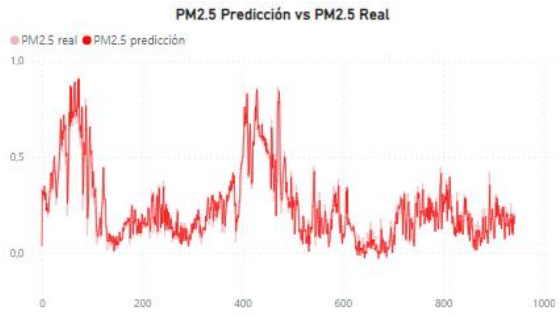
Función de pérdida modelo Seq2Seq para Bucaramanga.

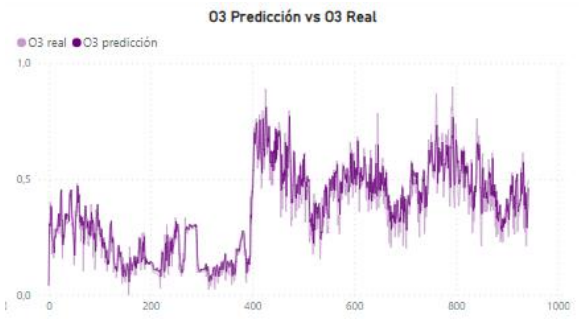


En cuanto a la visualización de la predicción presentada en la Figura 77, su comportamiento es consecuente al análisis realizado anteriormente. Se destaca el contaminante PM2.5 al ajustarse adecuadamente a los datos reales en comparación con O3, el cual, no toma por completo los límites inferiores.

Figura 77

Predicción de los contaminantes modelo Seq2Seq para Bucaramanga.

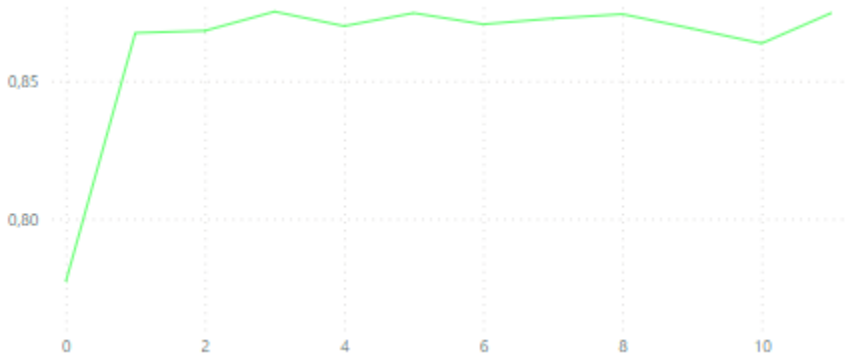




La precisión para esta ciudad toma valores aproximados a 0,86, siendo altos en contraste con los demás modelos. Esto se representa en la Figura 78.

Figura 78

Función de precisión modelo Seq2Seq para Bucaramanga.

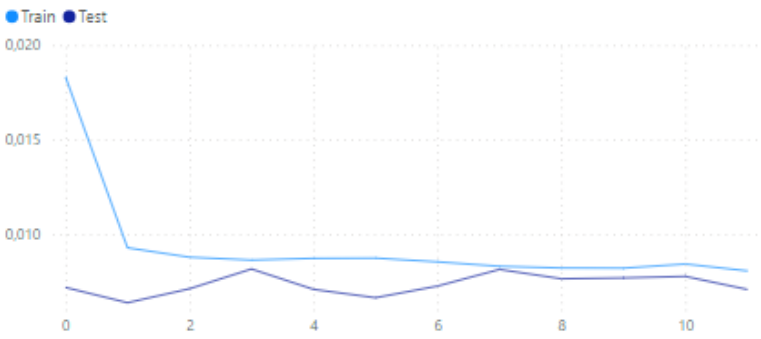


5.4.3.3. Cali

5.092 datos fueron utilizados para entrenar al modelo y 566 para validarlo en la ciudad de Cali, en la Figura 79 se evidencia un mayor ajuste de los datos a partir del Epoch 7. El error cuadrático medio tuvo un promedio de 0,009351 para entrenamiento y 0,007342 para la prueba.

Figura 79

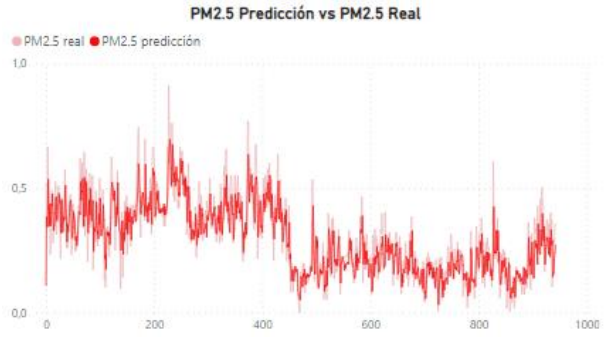
Función de pérdida modelo Seq2Seq para Cali

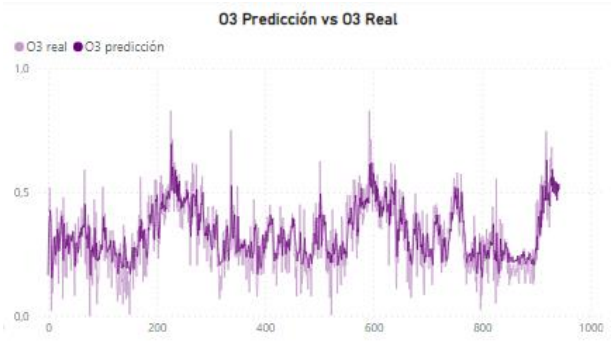


En la predicción de los contaminantes se demuestra que a pesar de que no es óptima, ya que, no predice acertadamente los límites, es mejor que los modelos LSTM y MLP al entender el comportamiento de los datos reales y su función de pérdida es mínima como se muestra en la Figura 80.

Figura 80

Predicción de los contaminantes modelo Seq2Seq para Cali.

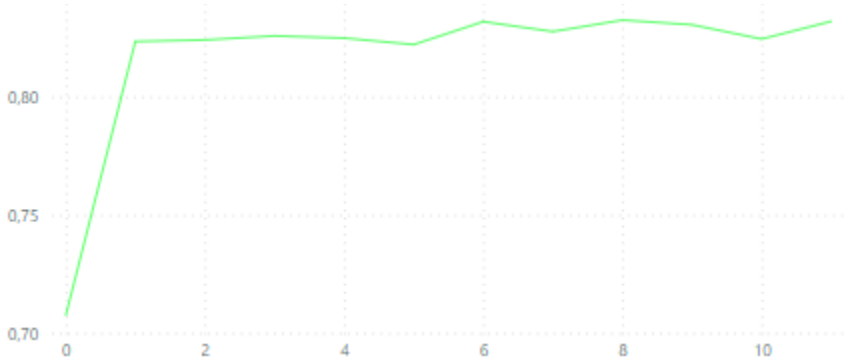




Para corroborar lo dicho anteriormente, en la Figura 81 se observa que la función de precisión se encuentra en 0,82 aproximadamente, siento este valor el más alto para la función de precisión en comparación con el resultado de los anteriores modelos.

Figura 81

Función de precisión modelo Seq2Seq para Cali.

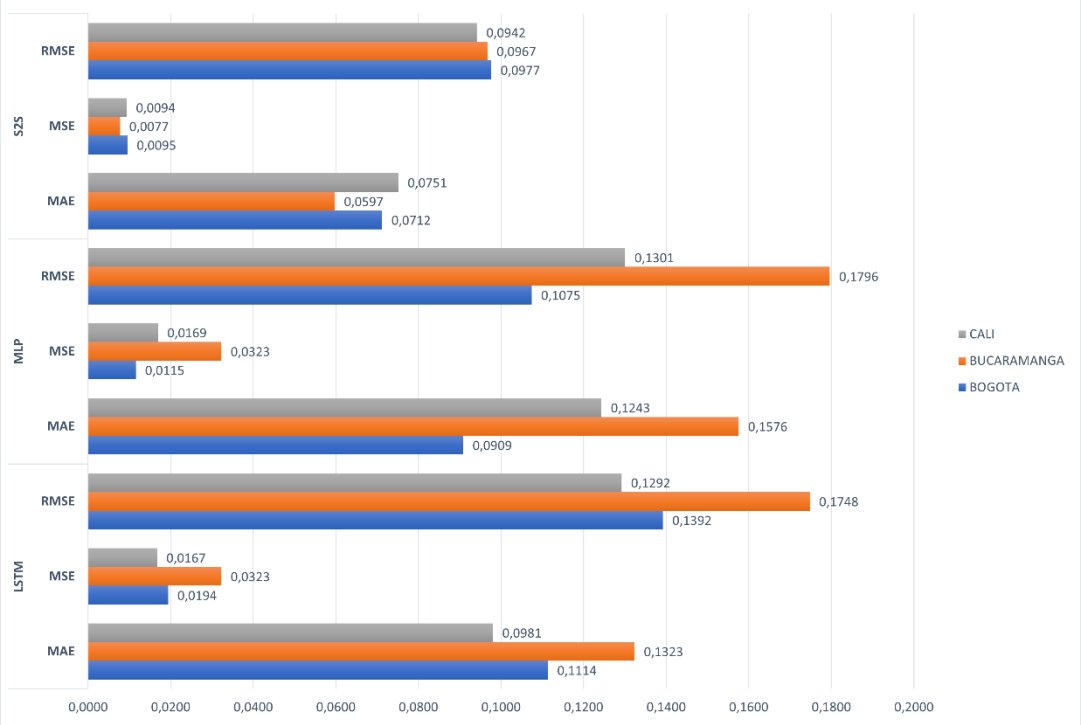


5.5. Etapa de interpretación

Como paso final, se realiza el proceso de verificación y validación de los modelos tratados anteriormente utilizando 3 métricas conocidas como: el error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz cuadrada del promedio de las diferencias cuadradas (RMSE), como se muestra en la Figura 82, las cuales fueron seleccionadas debido a su uso común en el análisis de regresión para comprender el error de predicción de los modelo, como se evidencia en los artículos aquí mencionados.

Figura 82

Relación de métricas de validación.



Para los modelos MLP y LSTM, en general la ciudad con valores de error más altos es Bucaramanga, sin embargo, en el modelo SEQ2SEQ es la que presenta mayor valor de precisión y menor MSE en promedio como se observó en la Figura 82, de igual manera, se destaca que las demás ciudades no presentan una diferencia significativa como sucede en los demás modelos. Por otra parte, la ciudad de Bogotá presentó el menor valor para las tres métricas de validación en el modelo MLP y en la ciudad de Cali se evidenciaron valores más bajos en las métricas para el modelo LSTM respecto a las demás ciudades.

Adicionalmente, el modelo de mayor error es el LSTM para las métricas de MSE y RMSE y, en general, el modelo SEQ2SEQ presenta en promedio los valores más bajos de las métricas de validación, destacándose como el mejor. El anterior análisis se presenta en la Tabla 13.

Tabla 13

Relaciones métricas de validación.

	LSTM			SEQ2SEQ			MLP		
Métrica de validación	MSE	MAE	RMSE	MSE	MAE	RMSE	MSE	MAE	RMSE
Bogotá	0,0193	0,1113	0,1392	0,0095	0,0712	0,0976	0,0115	0,0908	0,1074
Bucaramanga	0,0305	0,1323	0,1747	0,0077	0,0597	0,0967	0,0322	0,1575	0,1796
Cali	0,0167	0,0981	0,1292	0,0093	0,0751	0,0941	0,0169	0,1243	0,1301
Promedio	0,0222	0,1139	0,1477	0,0088	0,0686	0,0961	0,0202	0,1242	0,1390

6. Resultados y análisis

En cuanto a los modelos utilizados, se identificó que el modelo MLP falló en la identificación de patrones en las series de tiempo, lo cual se tradujo en una baja precisión en la predicción de los datos. Por lo tanto, se puede concluir que el MLP no se adaptó para el conjunto de datos utilizado en este proyecto.

Por otra parte, el modelo LSTM, aunque superó el problema del MLP y logró obtener resultados aceptables en la predicción de los datos, no son los mejores ya que, no consiguió interpretar adecuadamente su variación. Sin embargo, una posible variación de este modelo como es el LSTM apilado, podría modelar con mayor capacidad la clasificación de los datos como lo mencionan Ekinci, Omurca, & Ozbay (2021).

Finalmente, gracias al desarrollo de los modelos, se obtuvo que el Seq2Seq tuvo la capacidad de predecir los contaminantes con mejor desempeño que los otros modelos sin presentar sobreajuste, resultado confirmado por el trabajo de grado descrito en el marco de antecedentes realizado por Andrea Abril y Edgar Porras. Esto sugiere que el modelo Seq2Seq es altamente efectivo para la predicción de series de tiempo a pesar de que tiene un costo computacional más alto.

Por otra parte, se realiza un análisis del conjunto de datos preprocesados para determinar el comportamiento del confinamiento en la ventana de tiempo establecida.

Para tener un diagnóstico de dichos datos, se presentan los valores máximos permitidos a partir del 2018 según la resolución 2254 del 01 de noviembre de 2017 (Ministerio de Ambiente y Desarrollo Sostenible, 2017) en la Tabla 14.

Tabla 14

Valores máximos permitidos según la Resolución.

Contaminante	Nivel máximo permisible	Tiempo de exposición
PM2.5	37	24 horas
NO2	200	1 hora
O3	100	8 horas

Adicionalmente, en la Tabla 15, se reportan los valores promedios y máximos presentados en los datos preprocesados con el fin de identificar si alguno de estos supera o se aproxima a los niveles máximos permitidos por las autoridades ambientales.

Tabla 15

Valores promedios y máximos de los datos preprocesados.

			2019			
	NO2	Máximo reportado	O3	Máximo reportado	PM2.5	Máximo reportado
Bogotá	18,7808	45,9985	41,5860	82,0855	38,9908	84,6395
Bucaramanga	N/A	N/A	22,3153	28,7833	15,9850	47,8375
Cali	N/A	N/A	28,9454	58,3996	20,6371	40,4605
			2020			
	NO2	Máximo reportado	O3	Máximo reportado	PM2.5	Máximo reportado
Bogotá	29,8500	49,9599	55,2354	88,5274	30,7479	73,1236
Bucaramanga	N/A	N/A	26,7381	48,8792	16,8541	45,7792
Cali	N/A	N/A	29,0908	58,3996	14,0181	34,7292

Dicho lo anterior, se evidencia que los contaminantes NO2 y O3 para las tres ciudades no superan el valor máximo permitido por las autoridades, por ende, se encuentran siempre en un rango de "Bueno". Por otra parte, el contaminante PM2.5 al menos un día para las distintas ciudades presenta un valor que supera el máximo permitido en el 2019 y para el 2020, solo Bogotá y Bucaramanga superan al menos un día los límites. Asimismo, concluyen Shatnawi & Abu-Qdais (2021) en su artículo, donde las concentraciones de los contaminantes durante el confinamiento no superan los límites impuestos por su gobierno.

Ahora, con el fin de comprobar la hipótesis del planteamiento del problema, la cual describe que la contaminación atmosférica (PM2.5, NO2 y O3) disminuyó después de la etapa del confinamiento provocado por el COVID 19, se realiza inicialmente una gráfica de boxplot donde se comparan visualmente los tres momentos: Antes, Durante y Después del cierre. Posteriormente, para verificar si existen diferencias significativas en el valor medio de los tres escenarios, se aplica un ANOVA[6] estableciendo un nivel de significancia de 1% donde primero se corroboran los supuestos de normalidad con la prueba Kolmogorov-Smirnov[7]; de homocedasticidad[8] con la prueba Levene[9] y se asume que los datos son independientes entre sí. Si efectivamente las medias son diferentes, se realiza una prueba Tukey para determinar cuáles de los escenarios difieren.

Si los supuestos descritos anteriormente no se cumplen, se debe implementar una prueba no paramétrica como Kruskal-Wallis[10] para demostrar si la media de los grupos es

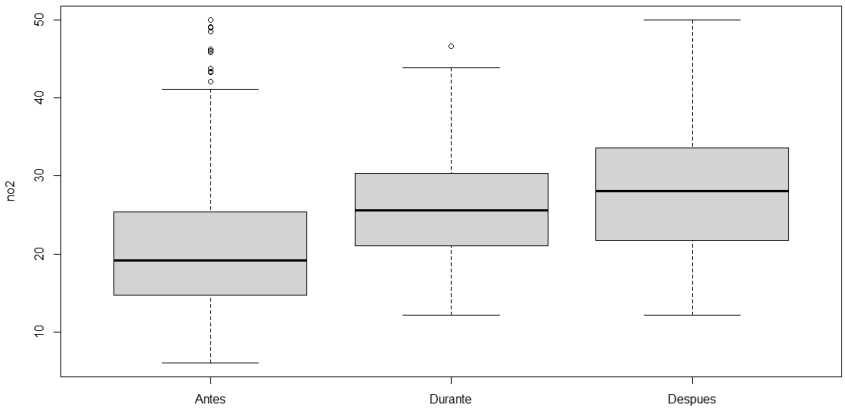
[6] Análisis de la Varianza es una fórmula estadística que se utiliza para comparar las varianzas entre las medias de diferentes grupos.
[7] La prueba Kolmogorov-Smirnov se utiliza para contrastar si un conjunto de datos se ajusta o no a una distribución normal.
[8] El supuesto de Homocedasticidad considera que la varianza es constante (no varía) en los diferentes niveles de los grupos.
[9] La prueba de Levene comprueba si varios grupos tienen la misma varianza en la población.
[10] La prueba de Kruskal-Wallis es una prueba de hipótesis no paramétrica para muestras múltiples independientes que se utiliza cuando no se cumplen los supuestos de un ANOVA.

igual y, ya que esta no proporciona una respuesta a la pregunta de cuáles de los grupos difieren, se requiere hacer una prueba Dunn-Bonferroni[11].

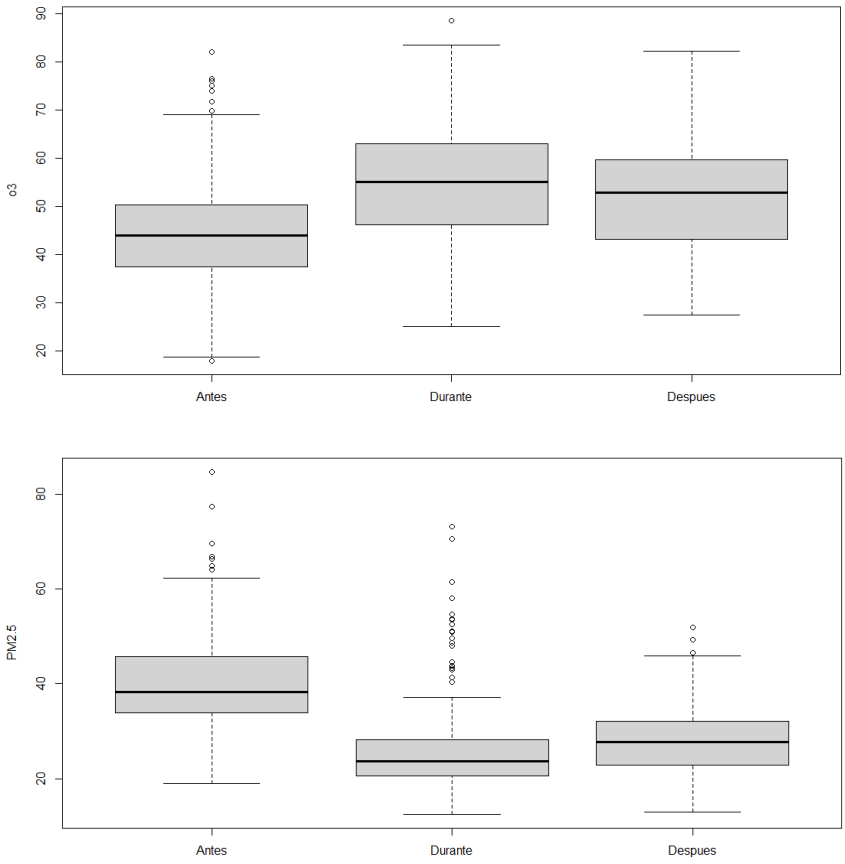
En la **Figura 83** *Boxplot escenarios de los contaminantes para la ciudad de Bogotá.***Figura 83** se muestra la comparación del antes, durante y después del confinamiento para los contaminantes NO2, O3 y PM2.5 en la ciudad de Bogotá, allí se comprueba que en efecto, para el contaminante NO2 después del cierre tuvo mayor concentración, en contraste con lo expuesto por Talbot, y otros (2021) donde refiere que en la mayoría de estudios se reducen estos niveles durante el cierre debido a la disminución en las emisiones de combustión. Para el contaminante O3 se presenta un aumento significativo durante el confinamiento coincidiendo con los resultados obtenidos en el artículo de Ekinci, Omurca, & Ozbay (2021), en el cual, menciona que los niveles de O3 han exhibido una tendencia creciente en muchos países durante este tiempo. Por otra parte, el PM2.5 arroja coincidencias con todos los artículos donde analizan este contaminante durante los bloqueos por el COVID 19, al disminuir su concentración.

Figura 83

Boxplot escenarios de los contaminantes para la ciudad de Bogotá.



[11] La prueba Dunn-Bonferroni compara por pares entre cada grupo independiente e indica qué grupos son diferentes.



En la Tabla 16 se presentan los valores de p[12] para las distintas pruebas realizadas en los contaminantes junto con el resultado arrojado por la hipótesis nula de cada prueba. Se puede concluir que los supuestos no se cumplieron en su totalidad para el contaminante NO2, por ende, se aplica la prueba de Kruskal-Wallis, dando como resultado que los valores medios de los tres escenarios son diferentes. Por el contrario, los contaminantes PM2.5 y O3 presentaron el cumplimiento de los supuestos, por ende, se aplica ANOVA.

[12] El valor p es la probabilidad de que un valor estadístico calculado sea posible dada una Hipótesis nula cierta.

Tabla 16

Pruebas estadísticas para el contaminante NO2 en Bogotá.

Contaminante NO2			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba Kruskal-Wallis
Valor p	0,1921	0,003013	$2{,}2x10^{-16}$
Análisis de la Hipótesis nula	Presenta homocedasticidad	No presenta distribución normal	Las medias de los escenarios son diferentes
Contaminante O3			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba ANOVA
Valor p	0,1936	0,7542	$2x10^{-16}$
Análisis de la Hipótesis nula	Presenta homocedasticidad	Presenta distribución normal	Las medias de los escenarios son diferentes
Contaminante PM2.5			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba ANOVA
Valor p	0,01759	0,03521	$2x10^{-16}$
Análisis de la Hipótesis nula	Presenta homocedasticidad	Presenta distribución normal	Las medias de los escenarios son diferentes

Ahora, dependiendo del resultado de las pruebas realizadas anteriormente, se aplica para cada caso las pruebas post-hoc[13] correspondientes en la Tabla 17.

Tabla 17

Pruebas post-hoc para los contaminantes de la ciudad de Bogotá.

Contaminante	Prueba	Par de escenarios	Valor p
NO2	Dunn-Bonferroni	Antes – Después	$1{,}25x10^{-33}$
		Antes – Durante	$5{,}14x10^{-11}$
		Después – Durante	$2{,}42x10^{-3}$
O3	Tukey	Después – Antes	0

[13] Las pruebas post-hoc son utilizadas para determinar qué medias difieren entre sí, una vez que se ha determinado que existen diferencias entre las medias.

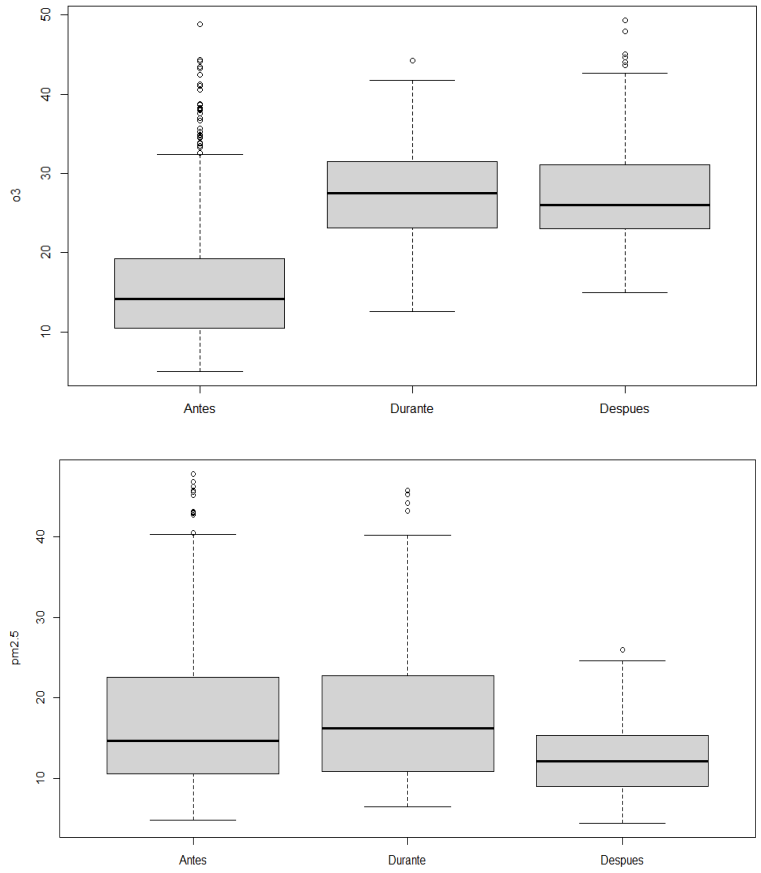
		Durante – Antes	0
		Durante – Después	0,0081
PM2.5	Tukey	Antes – Después	0
		Antes – Durante	0
		Después – Durante	0,1616

Se puede observar que para los contaminantes NO2 y O3, todos los escenarios tienen un valor $p < 0,01$, es decir, hay evidencia significativa de que estos pares de escenarios son aquellos que difieren entre sí. Por otra parte, el contaminante PM2.5 presenta diferencias entre los escenarios “Antes – Durante” y “Antes – Después”, es decir, no se produjo una diferencia significativa con respecto a la etapa “Durante – Después”.

En la Figura 84 se muestra la comparación del antes, durante y después del confinamiento para los contaminantes PM2.5 y O3 en la ciudad de Bucaramanga, se puede observar que las medias son similares en los momentos Durante y Después del confinamiento del contaminante O3 y Antes y Durante del contaminante PM2.5, es decir, ambos presentaron mayor concentración en estos escenarios. En el escenario “Después” se presenta una disminución en el contaminante PM2.5 y un aumento en O3, misma situación presentada por Velders, y otros (2021) donde afirman que las concentraciones observadas se reducen durante el período de confinamiento para el contaminante PM2.5, mientras que aumentó los niveles de NO2 y O3, aludiendo que fue resultado de las estaciones climáticas y del confinamiento.

Figura 84

Boxplot escenarios de los contaminantes para la ciudad de Bucaramanga.



En la Tabla 18 se presentan los valores de p para las distintas pruebas estadísticas realizadas, se concluye que los supuestos no se cumplieron en su totalidad para ambos contaminantes, por lo tanto, la prueba de Kruskal-Wallis fue aplicada para estos, dando como resultado que hay evidencia significativa de que las medias en los tres escenarios son diferentes.

Tabla 18

Pruebas estadísticas para los contaminantes en la ciudad de Bucaramanga.

Contaminante PM2.5			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba Kruskal-Wallis
Valor p	$2{,}2x10^{-16}$	$2{,}2x10^{-16}$	$2{,}2x10^{-16}$
Análisis de la Hipótesis nula	No presenta homocedasticidad	No presenta distribución normal	Las medias de los escenarios son diferentes
Contaminante O3			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba Kruskal-Wallis
Valor p	0,05989	0,00379	$2{,}2x10^{-16}$
Análisis de la Hipótesis nula	Presenta homocedasticidad	No presenta distribución normal	Las medias de los escenarios son diferentes

Posteriormente, se aplica para ambos contaminantes la prueba post-hoc Dunn-Bonferroni como se muestra en la Tabla 19. De esta se puede concluir que los pares de escenarios del contaminante PM2.5 “Antes - Después” y “Después - Durante” son aquellos que tienen diferencias entre sí. En cuanto al contaminante O3 se deduce que los grupos que presentan diferencias en sus valores medios son “Después – Antes” y “Durante – Antes”.

Tabla 19

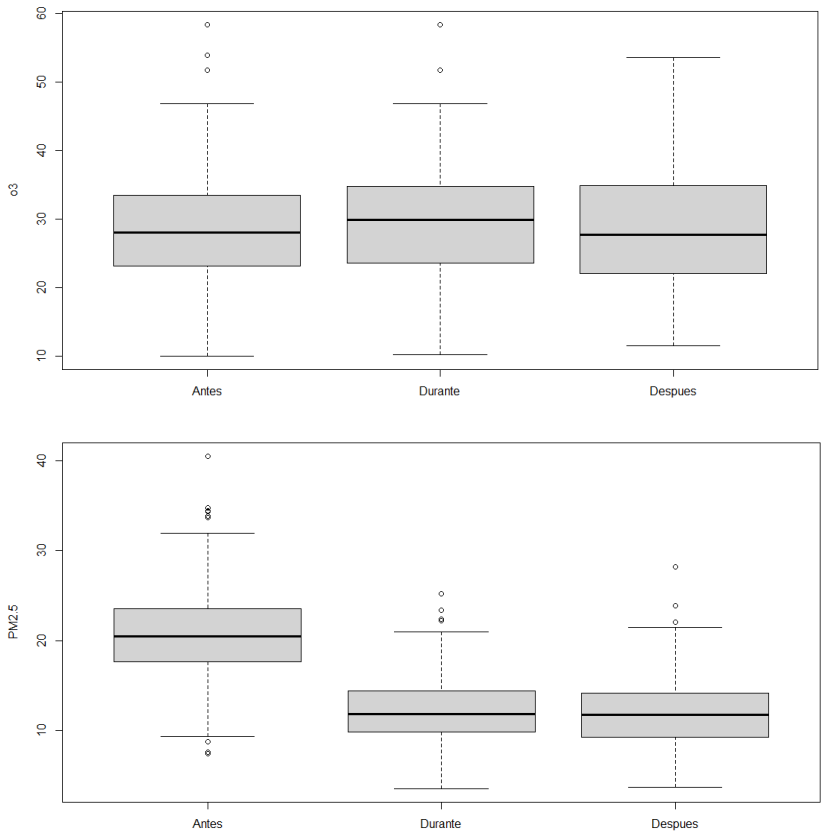
Prueba Dunn-Bonferroni para las contaminantes de la ciudad de Bucaramanga.

Contaminante	Prueba	Par de escenarios	Valor p
PM2.5	Dunn-Bonferroni	Antes – Después	$1{,}71x10^{-13}$
		Antes – Durante	$2{,}23x10^{-1}$
		Después – Durante	$5{,}58x10^{-12}$
O3	Dunn-Bonferroni	Después – Antes	$4{,}46x10^{-70}$
		Durante – Antes	$2{,}45x10^{-50}$
		Durante – Después	$5{,}47x10^{-1}$

Por último, para los contaminantes de la ciudad de Cali se presenta la Figura 85 donde se puede evidenciar que para el contaminante O3 no hay diferencias significativas en los valores medios de cada escenario, mientras que para el PM2.5 las medias de los momentos Durante y Después son similares. La disminución notoria a partir del confinamiento en PM2.5 se ve respaldada por Al-qaness, Fan, Ewees, Yousri, & Elaziz (2021), que aseguran que el confinamiento mejoró las concentraciones de PM2.5.

Figura 85

Boxplot escenarios de los contaminantes para la ciudad de Cali.



Los valores de p para las pruebas estadísticas seleccionadas son tabulados en la Tabla 20, se concluye de igual manera lo que ocurre en la ciudad de Bucaramanga, que los supuestos no se cumplen para ambos contaminantes, por lo tanto, la prueba de Kruskal-Wallis

se aplica y como resultado, se comprueba lo visto en la figura anterior para el contaminante O3, pues no hay evidencia significativa para afirmar que los valores medios de los escenarios son diferentes, por ello, no es necesario aplicar la prueba post-hoc para este. Por otra parte, el PM2.5 presenta diferencias en las medias de sus grupos.

Tabla 20

Pruebas estadísticas para los contaminantes en la ciudad de Cali.

Contaminante PM2.5			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba Kruskal-Wallis
Valor p	0,001494	0,002982	$2{,}2x10^{-16}$
Análisis de la Hipótesis nula	No presenta homocedasticidad	No presenta distribución normal	Las medias de los escenarios son diferentes
Contaminante O3			
Tests	Prueba Levene	Prueba Kolmogorov-Smirnov	Prueba Kruskal-Wallis
Valor p	0,05523	0,003572	0,3528
Análisis de la Hipótesis nula	Presenta homocedasticidad	No presenta distribución normal	Las medias de los escenarios son iguales

Considerando las pruebas anteriores, se aplica la prueba Dunn-Bonferroni únicamente para el contaminante PM2.5 como se observa en la Tabla 21. De esta se puede concluir que los escenarios que difieren entre sí son “Antes – Después” y “Antes – Durante” como también se evidencia en Figura 85.

Tabla 21

Prueba Dunn-Bonferroni para el contaminante PM2.5 en la ciudad de Cali.

Contaminante	Prueba	Par de escenarios	Valor p
PM2.5	Dunn-Bonferroni	Antes – Después	$4{,}84x10^{-96}$
		Antes – Durante	$3{,}56x10^{-59}$
		Después – Durante	$5{,}85x10^{-1}$

"Los resultados del análisis de datos de concentración mensual promedio de contaminantes del aire muestran que las restricciones en movilidad y suspensión de actividades durante el bloqueo de COVID-19 son una causa importante en la disminución de los contaminantes en comparación con los mismos períodos en 2019" (Skirienė & Stasiškienė, 2021), así como se evidencia en los resultados arrojados en esta sección para el PM2.5, el cual se vio impactado positivamente después de los cierres; en contraste con esto, Etchie, Etchie, Jauro, Pinker, & Swaminathan (2021) refieren que en los países tropicales la disminución de los contaminantes no se debe al confinamiento sino a los fuertes efectos de los cambios climáticos estacionales ocurridos en el mismo periodo. En la Tabla 22 se presenta la síntesis del análisis realizado en este capítulo.

Tabla 22

Resultados ANOVA y Kruskal-Wallis

Ciudad	PM2.5	Escenarios que difieren	NO2	Escenarios que difieren	O3	Escenarios que difieren	Resultado
Bogotá	ANOVA	Ant-Des(-) Ant-Dur(-)	Kruskal	Ant-Des(-) Ant-Dur(-) Dur-Des(-)	ANOVA	Ant-Des(+) Ant-Dur(+) Dur-Des(+)	PM2.5 impacto + NO2 y O3 impacto -
Bucaramanga	Kruskal	Ant-Des(-) Dur-Des(-)	N/A	N/A	Kruskal	Ant-Des(+) Ant-Dur(+)	PM2.5 impacto + O3 impacto -
Cali	Kruskal	Ant-Des(-) Ant-Dur(-)	N/A	N/A	Kruskal	No hay diferencias	PM2.5 impacto + O3 no hubo impacto

7. Divulgación de conocimiento

Para la divulgación de los resultados del proyecto se elabora un artículo Apéndice A, el cual contiene los aspectos relevantes del proyecto como la revisión de literatura, la justificación del problema, así como el análisis de los resultados a partir del desarrollo de la investigación.

8. Conclusiones

De acuerdo con la metodología implementada en este proyecto de investigación, se evidencia que los modelos de Aprendizaje Profundo son buenos predictores de series temporales, ya que reconocen patrones presentes en el conjunto de variables de entrada que son diferentes entre sí y son capaces de predecir acertadamente el comportamiento de los contaminantes.

Realizada la revisión de literatura, se encontraron finalmente 12 artículos que exponen la aplicación de diferentes modelos para analizar la problemática de la calidad del aire a partir del COVID 19, donde se evidencia la preocupación mundial de esta temática incentivando las distintas investigaciones para ser apoyo en el control de los contaminantes atmosféricos. Los modelos encontrados para el análisis de los datos en su mayoría pertenecen al Aprendizaje Automático (Random Forest Algorithmic) y a modelos estadísticos simples (ARIMA, prueba t de Welch, Prueba Mann Kendall), sin embargo, el presente trabajo se enfoca en los modelos Aprendizaje Profundo más utilizados como los LSTM y MLP y sus variaciones.

De la aplicación de los modelos de Aprendizaje Profundo LSTM, MLP y Seq2seq se concluye que, para el conjunto de datos utilizados en este proyecto, el modelo óptimo es el Seq2Seq, a pesar de que en la literatura no es utilizado comúnmente. Este arrojó predicciones

ajustadas a los datos reales con porcentajes bajos de las métricas de validación RMSE, MAE y MSE con valores de 9.62, 0.88 y 6.86 para todas las ciudades en cuestión.

Al realizar la comparación de los tiempos “Antes”, “Durante” y “Después” del confinamiento, se infiere que el contaminante PM2.5 disminuyó su concentración para todas las ciudades en comparación con el “Antes”, es decir, hubo un impacto positivo. Por otra parte, para el contaminante O3 se concluye que para las ciudades Bogotá y Bucaramanga tuvo un aumento significativo después del cierre, mientras que inesperadamente en Cali no hubo diferencia en su valor medio para los tres momentos, lo que quiere decir que tuvo un impacto negativo para las primeras ciudades mencionadas. Finalmente, el contaminante NO2 incrementó su concentración después del confinamiento para la ciudad de Bogotá, por lo tanto, hubo un impacto negativo ocasionado por los cierres debido al COVID 19.

En conclusión, a pesar de que el contaminante PM2.5 se vio afectado positivamente gracias a que depende de la contaminación generada principalmente por el tráfico vehicular y de la industria, realmente no hubo una mejora significativa en la calidad del aire, teniendo en cuenta el resultado de los otros dos contaminantes.

9. Recomendaciones

Con lo recopilado en el presente trabajo de investigación, se recomienda para próximas investigaciones el uso de modelos de redes neuronales especializados en la predicción de series de tiempo, como el LSTM o el Seq2Seq, en lugar de modelos generales como el MLP. Así mismo, se sugiere la exploración de técnicas adicionales para mejorar aún más la precisión de la predicción de los datos de las series temporales.

Por otra parte, debido a que se actualizó la plataforma del IDEAM a partir de abril de

2023, donde se agrupan todos los datos de las estaciones de las diferentes ciudades, se recomienda utilizar mayor cantidad de datos para tener mejores predicciones y escoger mayor cantidad de variables de entrada como la radiación solar, precipitación, presión atmosférica, nubosidad, entre otras.

Por último, se considera necesario continuar también con las medidas preventivas que se exponen en el Artículo 15 de la Resolución 2254 del 01 de noviembre del 2017, para evitar la exposición de altos niveles de los contaminantes y así contribuir a la mejora de la calidad de vida de la población. Adicionalmente, se proponen las siguientes:

- Incentivar el uso de bicicletas y transporte público cuando el nivel de los contaminantes de aproximen al límite establecido por el ICA, reduciendo sus tarifas, estableciendo convenios con empresas comprometidas con el medio ambiente para la generación de bonos de descuento o implementando alternativas de cambio de reciclaje por pasajes.
- Realizar talleres de formación y divulgación de educación ambiental a la comunidad.
- Regular estrictamente el parque automotor aplicando sanciones a los que no cumplen con las características ambientales para su movilización, impulsando el uso de transporte con energías renovables como el biocarburante.
- Mejorar la red de vigilancia de la calidad del aire y la detección de incendios forestales para mayor control y prevención.
- Establecer jornadas de plantación de especies arbóreas y arbustivas en los espacios urbanos.
- Fomentar del compostaje colectivo de bioresiduos y el reciclaje de los productos que no han terminado su ciclo de vida.

Referencias Bibliográficas

Abirami, S., & Chitra, P. (2019). *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases.* Science Direct. Obtenido de https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron

ADRES. (09 de febrero de 2022). *El sistema de salud recibió $65,19 billones en el 2021 para financiar la salud de los colombianos*. ADRES. Obtenido de https://www.adres.gov.co/sala-de-prensa/noticias/Paginas/El-sistema-de-salud-recibi%C3%B3-$65,19-billones-en-el-2021.aspx

Aguilar, L. J. (2013). *Big Data: Análisis de grandes volúmenes de datos en las organizaciones.* México D.F.: Alfaomega.

Alammar, J. (09 de mayo de 2018). *Visualizing a Neural Machine Translation Model (Mechanics of Seq2Seq Models With Attention).* Github.io. Obtenido de https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Alcaldía de Santiago de Cali (19 de abril de 2022). Respuesta a radicado No. 202241730100539612. *Respuesta a solicitud de datos de la variable NO2.* https://www.cali.gov.co/participacion/publicaciones/46368/consulte_el_estado_de_su_solicitud/

Al-qaness, M.A.A., Fan, H., Ewees, A.A., Yousri, D., Elaziz, M.A. (2021). Improved ANFIS model for forecasting Wuhan City Air Quality and analysis COVID-19 lockdown impacts on air quality. *Environmental Research*.

Arana, C. (Junio de 2021). Redes Neuronales Recurrentes: Análisis de los modelos especializados en datos secuenciales. Obtenido de https://ucema.edu.ar/publicaciones/download/documentos/797.pdf

Berzal, F. (2018). *Redes neuronales y Deep Learning.* Granada: Editorial Universidad de Granada.

Brauer, M. (2010). How Much, How Long, What, and Where Air Pollution Exposure Assesment for Epidemiologic Studies of Respiratory Disease. *ATSJOURNALS*, 5.

Departamento Nacional de Planeación (2018). *Calidad del Aire: Una Prioridad de Política Pública en Colombia*. Obtenido de https://colaboracion.dnp.gov.co/CDT/Prensa/Presentaci%C3%B3n%20Calidad%20del%20Aire%2015_02_2018.pdf

Ekinci, E., Omurca, S., & Ozbay, B. (2021). Comparative assessment of modeling deep learning networks for modeling ground level ozone concentrations of pandemic lock down period. *Ecological Modelling*, 11.

EPA. (26 de Mayo de 2021). *Particulate Matter (PM) Basics*. EPA. Obtenido de https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM

EPA. (s.f.). *El impacto del Dióxido de Nitrógeno en la calidad del aire interior*. EPA. Obtenido de https://espanol.epa.gov/cai/el-impacto-del-dioxido-de-nitrogeno-en-la-calidad-del-aire-interior#:~:text=El%20NO2%20act%C3%BAa%20principalmente%20como,y%20una%20lesi%C3%B3n%20pulmonar%20difusa.

Etchie, T., Etchie, A., Jauro, A., Pinker, R., & Swaminathan, N. (2021). Season, not lockdown, improved air quality during COVID-19 State of emergency in Nigeria. *Science of the Total Environment*, 11.

Fayyad Usama, Gregory Piatetsky-Shapiro, Smyth Padhraic. (1996). *From data mining to knowledge discovery in databases*. IA Magazine, 37-54.

García González, J. R., Sánchez Sánchez, P. A., Orozco, M., & Obredor, S. (01 de febrero de 2019). *Extracción de conocimiento para la predicción y análisis de los resultados de la prueba de calidad de la educación superior en Colombia*. SCIELO. Obtenido de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062019000400055&lng=en&nrm=iso&tlng=en

García, J., Molina, J. M., Berlanga, A., Patricio, M. A., Bustamante, Á. L., & Padilla, W. R. (2018). *Ciencia de Datos: Técnicas analíticas y aprendizaje estadístico en un enfoque práctico.* Bogotá: Alfaomega.

García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, Á., & Padilla, W. (2018). *Ciencia de Datos: Técnicas analíticas y aprendizaje estadístico en un enfoque práctico.* Bogotá: Alfaomega colombiana S.A.

Ghatak, A. (2019). *Deep Learning with R.* Kolkata: Springer.

Green Facts. (15 de agosto de 2006). *Contaminación del aire Dióxido de Nitrógeno*.Green Facts. Obtenido de http://www.greenfacts.org/es/dioxido-nitrogeno-no2/

Grupo ENEL. (18 de agosto de 2021). *¿Cómo reducir la contaminación del aire en Colombia?*. ENEL.Obtenido de https://www.enel.com.co/es/historias/a202108-disminuye-la-contaminacion-en-el-aire.html

IDEAM. (2002). *Ozono Troposférico*. IDEAM. Obtenido de http://www.ideam.gov.co/web/tiempo-y-clima/ozono-troposferico

IDEAM. (18 de Abril de 2018). *Ficha metodológica operación estadística variables meteorológicas.* IDEAM. Obtenido de http://www.ideam.gov.co/documents/11769/72085840/Ficha+metodologica+variables+meteorologicas.pdf/d5915289-f08c-45c4-ad62-62efe957a1a3#:~:text=Dichas%20variables%20son%3A%20temperatura%20y,y%20velocidad)%20y%20brillo%20solar.

IDEAM. (22 de 01 de 2021). *Proporción de datos del Índice de la Calidad del Aire por autoridad ambiental.* MinAmbiente y Desarrollo. Obtenido de http://www.ideam.gov.co/documents/11769/641368/2.01+HM+Indice+calidad+aire.pdf/5130ffb3-a1bf-4d23-a663-b4c51327cc05#:~:text=El%20%C3%8Dndice%20de%20calidad%20del,por%20parte%20de%20la%20poblaci%C3%B3n.

Instituto para la salud Geoambiental. (s.f.). *Dióxido de Nitrógeno NO2*. Instituto para la Salud Geoambiental. Obtenido de https://www.saludgeoambiental.org/dioxido-nitrogeno-no2?gclid=Cj0KCQiA2ZCOBhDiARIsAMRfv9Kr7ViD1ZF4StBzV1reSUb9B3mZh5EksbZt17ey-BM0onUqOa3_ItIaAu0fEALw_wcB

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 22408-22417.

Mañas, A. M. (16 de junio de 2019). *Notas sobre pronóstico del flujo de tráfico en la ciudad de Madrid*. Bookdown.org. Obtenido de

https://bookdown.org/amanas/traficomadrid/m%C3%A9todos-basados-en-deep-learning.html#lstm-univariado

Martín Gutiérrez, E. (Septiembre de 2019). *Aplicación de modelos de Redes Neuronales Recurrentes a la predicción de emisiones contaminantes de autobuses urbanos.* Universidad Politécnica de Madrid. Obtenido de https://oa.upm.es/66442/7/TFM_ESTRELLA_MARTIN_GUTIRREZ.pdf

Ministerio de Ambiente y Desarrollo Sostenible. (2017). *Resolución 2254 del 1 de noviembre de 2017.* Bogotá.

OMS . (10 de Noviembre de 2020). *Información básica sobre la COVID-19.* Organización Mundial de la Salud. Obtenido de https://www.who.int/es/news-room/questions-and-answers/item/coronavirus-disease-covid-19

OMS. (27 de septiembre de 2016). *La OMS publica estimaciones nacionales sobre la exposición a la contaminación del aire y sus repercusiones para la salud*. Organización Mundial de la Salud. Obtenido de https://www.who.int/es/news/item/27-09-2016-who-releases-country-estimates-on-air-pollution-exposure-and-health-impact

OPS. (2018). *Calidad del aire*. Organización Panamericana de la Salud. Obtenido de https://www.paho.org/es/temas/calidad-aire

Osso, J. D. (20 de Octubre de 2020). *La calidad del aire durante las cuarentenas ocasionadas por el COVID-19.* Departamento de Derecho del Medio Ambiente. Obtenido de https://medioambiente.uexternado.edu.co/la-calidad-del-aire-durante-las-cuarentenas-ocasionadas-por-el-covid-19/

Our World in Data. (9 de Diciembre de 2021). *Coronavirus Pandemic (COVID-19).* Our World in Data. Obtenido de https://ourworldindata.org/coronavirus

RMCAB. (10 de 09 de 2020). *Red de Monitoreo de Calidad del Aire de Bogotá*. Obtenido de http://rmcab.ambientebogota.gov.co/home/map

Russell, R. (2018). *Deep Learning: Fundamentos de aprendizaje profundo para principantes.* Createspace.

Santamaría, J. M. (10 de Abril de 2008). *Efectos del material particulado en la salud*. Zonahospitalaria.com. Obtenido de https://zonahospitalaria.com/efectos-del-material-particulado-en-la-salud/

Shatnawi, N., & Abu-Qdais, H. (2021). Assessing and predicting air quality in northern Jordan during the lockdown due to the COVID-19 vitus pandemic using artificial neural network. *Air Quality, Atmosphere & Health*, 643-652.

Skiriené,A.F, Stasiskiené, Z. (2021). COVID-19 and Air Pollution: Measuring Pandemic Impact to Air Quality in Five European Countries. *Atmosphere*

Tadano, Y. S., Potgieter-Vermaak, S., Kachba, Y. R., Chiroli, D. M., Casacio, L., Santos Silva, J., . . . Godoi, R. (2020). Dynamic model to predict the association between air quality, COVID-19 cases, and level of lockdown. *Environmental Pollution*.

Talbot, N., Takada, A., Bingham, A.H., Elder, D., Lay Yee, S., Golubiewski, N. E. (2021). An investigation of the impacts of a successful COVID-19 response and meteorology on air quality in New Zealand. *Atmospheric Environment.*

Tan, C. C., & Eswaran, C. (2008). Performance Comparison of Three Types of Autoencoder Neural Networks. *Second Asia International Conference on Modelling & Simulation (AMS)*, 213-218.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de*

desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá: Ediciones Universidad Cooperativa de Colombia.

Unisabana. (05 de mayo de 2020). *El futuro del big data, en mano de los ingenieros*. Universidad de la Sabana. Obtenido de https://www.unisabana.edu.co/programas/carreras/facultad-de-ingenieria/ingenieria-industrial/noticias/detalle-noticia-ingenieria-industrial/noticia/el-futuro-del-big-data-en-mano-de-los-ingenieros/

Velders, G.J.M., Willers, S.M., Wesseling, J., van den Elshout, S., van der Swaluw, E., Mooibroek, D., van Ratingen, S. (2021). Improvements in air quality in the Netherlands during the corona lockdown based on observations and model simulations. *Atmospheric Environment*.

Zhao, Y., Wang, L., Huang, T., Tao, S., Liu, J., Gao, H., . . . Ma, J. (2021). Unsupervised PM2.5 anomalies in China induced by the COVID-19 epidemic. *Science of the Total Environment* , 8.