Modelos Deep Learning en logística urbana para la predicción de la calidad del aire en la ciudad de Bucaramanga.

Paula Andrea Abril Ortiz, Edgar Leonardo Porras Ojeda

Trabajo de grado para optar el título de Ingeniero Industrial

Director

Henry Lamos Díaz

PhD. en Física-Matemática

Codirectora

Yuly Andrea Ramírez Sierra

M.Sc. en Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingeniería Físico-mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga, Santander

2019

Agradecimientos

Al grupo de Investigación OPALO, por brindarnos las herramientas y espacios necesarios para el desarrollo de este proyecto.

Al profesor Henry Lamos Díaz, por aceptar dirigir nuestro proyecto y su apoyo a pesar de las adversidades.

Agradecemos a nuestra codirectora Yuly Andrea Ramírez Sierra por creer en nosotros, por su apoyo, paciencia y comprensión durante este proceso.

A David Puentes por su tiempo, motivación y por compartir sus conocimientos con nosotros.

A nuestros familiares, por su paciencia y apoyo incondicional.

A nuestros compañeros y amigos, por sus enseñanzas y apoyo en este largo camino.

Tabla de contenido

Lista de figuras

Listado de tablas

Listado de apéndices

(Ver apéndices adjuntos en el CD)

Apéndice A. Análisis bibliométrico

Apéndice B. Artículo de carácter publicable

RESUMEN

TÍTULO: MODELOS DEEP LEARNING EN LOGÍSTICA URBANA PARA LA PREDICCIÓN DE LA CALIDAD DEL AIRE EN LA CIUDAD DE BUCARAMANGA*

AUTORES:

ABRIL ORTIZ, PAULA ANDREA**

PORRAS OJEDA, EDGAR LEONARDO**

PALABRAS CLAVE: LOGÍSTICA URBANA, DEEP LEARNING, CALIDAD DEL AIRE, REDES NEURONALES, MATERIAL PARTICULADO, CONTAMINACIÓN DEL AIRE.

DESCRIPCIÓN:

El aumento de la población de las últimas décadas y el flujo de personas de las zonas rurales a las grandes ciudades ha ocasionado un incremento del volumen de desplazamientos de pasajeros y mercancías, llevando a la sobresaturación de la circulación vehicular y al aumento de las necesidades de movilidad para satisfacer la competitividad comercial y la calidad de vida de sus habitantes, convirtiendo el tráfico en un factor relevante en la contaminación del aire y un importante elemento de riesgo medioambiental para la salud de grupos vulnerables como mujeres, niños y adultos mayores. Por ello, la gestión óptima de la movilidad urbana se enfoca un modelo capaz de cubrir tanto el aumento de la urbanización como su impacto en la calidad del aire, lo cual representa un reto para los gobiernos a nivel mundial en cuanto el diseño y ejecución de metodologías y estrategias orientadas la reducción de los niveles de contaminación del aire en pro de mejorar la salud cardiovascular y respiratoria de la población.

Teniendo en cuenta lo anterior, este proyecto se orienta a la aplicación, valoración y comparación de modelos Deep Learning (DL) entre ellos el LSTM, analizando los datos meteorológicos y de material particulado de tres estaciones de monitoreo ambiental del Área -Metropolitana de Bucaramanga – AMB, con el objetivo de obtener un modelo predictivo que apoye la planeación de estrategias orientadas a mejorar la calidad del aire.

---

* Trabajo de grado

** Facultad de Ingenierías Fisico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director PhD Henry Lamos Díaz. Codirectora MS. Yuly Andrea Ramírez Sierra.

ABSTRACT

TITLE: DEEP LEARNING MODELS IN URBAN LOGISTICS FOR AIR QUALITY PREDICTION IN THE CITY OF BUCARAMANGA [*]

AUTHORS:

ABRIL ORTIZ, PAULA ANDREA[**]

PORRAS OJEDA, EDGAR LEONARDO[**]

KEYWORDS: URBAN LOGISTICS, DEEP LEARNING, AIR QUALITY, NEURAL NETWORKS, PARTICULATE MATERIAL, AIR POLLUTION.

DESCRIPTION:

The increase in population in recent decades and the migration of people from rural areas to large cities have led to an increase in the volume of movements of passengers and goods, leading to the oversaturation of the vehicular traffic and the increase of mobility needs to satisfy business competitiveness and the quality of life of the inhabitants, thus, traffic has become an important factor of the air pollution and an important element of environmental risk for vulnerable groups health such as women, children and elderly. For this reason, the optimal urban mobility management is focused on a model that can cover the increase of urbanization and the air pollution impact, which represents a challenge to governments at a global scale in terms of design and implementation of methodologies and strategies for reducing air pollution levels to improve cardiovascular and respiratory health quality of the population.

Taking into consideration what has been previously stated, this project heads towards the application, assessment and comparison of Deep Learning models (DL) among them LSTM model, analyzing the meteorological data and particulate material in the three environment monitoring stations located in Bucaramanga with the purpose of obtaining a predictive model that supports the planning of strategies that might improve the air quality.

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Fisico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director PhD Henry Lamos Díaz. Codirectora MS. Yuly Andrea Ramírez Sierra.

Introducción

Colombia consciente de su crecimiento urbano y por ende del incremento de las necesidades de transporte y movilidad, establece el monitoreo y control de la contaminación atmosférica como uno de sus principales desafíos; el Departamento Nacional de Planeación, informó que durante el año 2015, los efectos de este fenómeno estuvieron asociados a 10.527 muertes y 67,8 millones de síntomas y enfermedades además de un costo ambiental del 1,93% del PIB ($15.4 billones de pesos), lo cual evidencia la necesidad de implementar estrategias para controlar, evaluar y monitorear las sustancias contaminantes en la atmósfera (IDEAM, n.d.-b).

En general, en las ciudades el principal responsable de la pérdida de calidad del aire lo constituyen las fuentes móviles y se estima que contribuyen entre el 75% y 80% del total de la contaminación (Tomassetti de Piacentini, n.d.). En los últimos ocho años, Bucaramanga ha tenido 65.344 vehículos más circulando por las calles y cuenta con un parque automotor de 220.993 vehículos (Dirección de Tránsito de Bucaramanga, 2019), lo cual incide negativamente en los índices de calidad del aire y por ende en la salud de sus habitantes; además la Organización Mundial de la Salud – OMS, establece que la contaminación ambiental del aire es causa de 4,2 millones de muertes prematuras por año en todo el mundo, esta mortalidad se debe a la exposición a partículas pequeñas de 2,5 micrones o menor diámetro (PM2.5), que causan enfermedades cardiovasculares, respiratorias y cáncer(Organización Mundial de la Salud - OMS, 2018).

En este sentido, el Ministerio de Ambiente y Desarrollo Sostenible, en la resolución N° 2254 del 2017, define los estándares nacionales para la calidad del aire con el fin de garantizar un ambiente sano, disminuyendo el riesgo de la contaminación ambiental sobre la salud humana, a su

vez, se resalta la importancia de incorporar equipos automáticos para realizar monitoreo constante y en tiempo real principalmente de zonas identificadas como puntos críticos, dicho monitoreo se podría complementar con la ejecución de un modelo predictivo que permita reducir los niveles de contaminación y apoye el desarrollo de acciones preventivas y no correctivas (Ministerio de Ambiente y Desarrollo Sostenible, 2017b).

Dentro de las metodologías de analítica de datos utilizadas actualmente para la predicción y control de la calidad del aire, el Deep Learning (DL) toma gran relevancia al permitir el procesamiento de gran cantidad de datos con una reducción significativa del costo del hardware de computación. La variedad de avances en aprendizaje automático y estudios de procesamiento de información, como el DL permite el monitoreo de datos temporales, realizando evaluaciones oportunas para predecir el posible resultado de acuerdo a la interacción de las variables (Xu et al., 2018).

Por otro lado, a nivel local, la Red Metropolitana de Calidad del Aire del Área Metropolitana de Bucaramanga – AMB, en su informe del índice de la calidad del aire – ICA, del 17 de marzo de 2019, mencionó que el 50% de las estaciones de monitoreo reportaron niveles perjudiciales de calidad del aire para la salud de grupos sensibles e incluso niveles peligrosos del componente PM 2.5 (AMB, 2019), teniendo en cuenta los retos mencionados inicialmente y la problemática de la región, se evidencia la oportunidad de utilizar métodos de analítica de datos como el Deep Learning para apoyar el monitoreo y predicción de la calidad del aire, disminuyendo el tiempo de respuesta y rapidez para la propuesta de medidas preventivas a corto y largo plazo.

Por consiguiente, este proyecto se enfoca en la adopción de métodos de analítica de datos como el Deep Learning y metodologías y/o herramientas tecnológicas que contribuyan a mejorar la

gestión de la información, con el objetivo de probar métodos de análisis predictivo de la calidad del aire que apoyen la toma de decisiones de las correspondientes autoridades o entidades de Bucaramanga que dirijan iniciativas al bienestar de la población.

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 desde la perspectiva de la logística urbana y el uso de Deep Learning como método de análisis de datos para la toma de decisiones.	Capítulo 3
Seleccionar y adoptar modelos de Deep Learning para relacionar variables predictivas que permitan abordar el problema objeto de estudio.	Capítulo 6
Validar los modelos de Deep Learning ajustados, utilizando métricas de evaluación de desempeño.	Capítulo 6
Elaborar un artículo de carácter publicable a partir de la investigación realizada.	Apéndice B

1. Generalidades del proyecto

1.1. Planteamiento del problema

La contaminación atmosférica deteriora gravemente el medioambiente, pero también perjudica la salud de las personas. Una de las principales causas de este tipo de contaminación son las fuentes móviles, convirtiendo el tráfico en uno de los factores más relevantes de contaminación debido a la emisión de contaminantes que se producen cerca y que se dispersan a la población generados por la dependencia excesiva de los vehículos privados, el uso de los vehículos antiguos y/o con combustibles Diesel y la falta de redes de transporte público efectivas.

Cabe resaltar que se considera contaminación del aire a la presencia de materiales y formas de energía que no forman parte de la composición natural del aire y que representan una potencial fuente de daños y molestias para la vida, al acarrear reacciones químicas impredecibles e inconvenientes (Raffino, n.d.)

Según la nota descriptiva de mayo de 2018 de la OMS, en el año 2012, 6,5 millones de muertes (11,6% de todas las muertes mundiales) estuvieron relacionadas con la contaminación del aire y además fue la causa del 25% de muertes por cardiopatías, 34% de muertes por accidentes cerebrovasculares y 36% de muertes por cáncer de pulmón, de los cuales el 91% se producen en países de ingresos bajos o medios, para el años 2013 el Centro Internacional de Investigaciones sobre el Cáncer de la OMS, determinó que la contaminación del aire es cancerígena para el ser humano y que las partículas del aire contaminado están estrechamente relacionadas con la creciente incidencia del cáncer, especialmente el cáncer de pulmón (Organización Mundial de la Salud - OMS, 2018).

Por ello, la reducción de los niveles de contaminación del aire representa un reto para los gobiernos a nivel mundial, ya que repercute en la mejora de la salud cardiovascular y respiratoria de la población, disminuyendo la morbilidad ocasionada por accidentes cerebrovasculares, cáncer de pulmón, hígado o vejiga y neumopatías crónicas y agudas.

> *"Aunque los estudios epidemiológicos de la OMS no han identificado umbrales de contaminación que no representen riesgo para la salud humana, en 2005 la OMS estableció una serie de directrices respecto a los umbrales y límites frente a contaminantes atmosféricos como: material particulado (PM), ozono (O3), dióxido de nitrógeno (NO2) y dióxido de azufre (SO2)." (Organización Mundial de la Salud - OMS, 2018).*

La problemática de la contaminación atmosférica es un tema de gran importancia a nivel mundial, pero su control y monitoreo depende en su mayoría de políticas y normativas nacionales; aunque las directrices de la OMS son de alcance mundial, cada país es libre de ajustar la normativa según sus condiciones específicas, pero, la mayoría de normativas están enfocadas hacia el monitoreo y control correctivo y las medidas de acción preventiva no se enfocan a reducir la contaminación a largo plazo, por ello, los métodos de predicción de la contaminación del aire representan una oportunidad para mejorar la salud de las personas y contribuir a la toma de decisiones efectivas por parte del gobierno.

Colombia no es ajena a la importancia de la calidad del aire y conforme a la normativa internacional, en el Informe del Estado de la Calidad del Aire en Colombia (2016) se presentan parámetros regulatorios a nivel nacional orientados a mantener un ambiente sano y reducir los índices de morbilidad y mortalidad generados por la contaminación del aire. Para ello, corporaciones autónomas regionales y las autoridades ambientales de los grandes centros urbanos,

han instalado sistemas de vigilancia de calidad del aire para obtener información sobre las concentraciones de las sustancias con potencial de afectar este recurso, esta información, es reportada y divulgada a través del Subsistema de Información de Calidad del Aire (SISAIRE) que se encarga de monitorear la calidad del aire en gran parte del país por medio de estaciones localizadas en diferentes municipios. (IDEAM - Instituto de Hidrología Meteorología y Asuntos Ambientales, 2017).

De acuerdo con el informe del estado de la calidad del aire emitido por el IDEAM en el año 2012 y que corresponde al periodo de monitoreo del 2007 al 2010, se ha encontrado que los municipios de Colombia con mayor contaminación ambiental son Bogotá DC, Medellín, Cali, Barranquilla, el Valle de Aburrá, Bucaramanga, Cartagena y Pereira; en estas ocho regiones se genera el 41% del material contaminante del aire en Colombia.

En Bucaramanga, según el AMB (2019), desde octubre del 2018 la tendencia de la calidad del aire en el área metropolitana ha presentado variaciones constantes en los niveles de calidad del aire y han aumentado las concentraciones de partículas contaminantes en algunos sectores, por ejemplo, en los resultados del índice de calidad del aire del mes de marzo de 2019, la lectura de la estación de medición del AMB localizada en el Instituto Caldas (Bucaramanga), presentó durante un periodo de 22 días un rango dañino para la salud de grupos sensibles tales como: niños, mujeres embarazadas y adultos mayores (AMB, 2019).

Para contrarrestar la emisión de material particulado en el área metropolitana de Bucaramanga, durante marzo y abril del 2019 se incrementaron los operativos de control a las fuentes móviles, sin embargo, dicho control presenta varias limitaciones, ya que aunque la mayoría de vehículos inspeccionados cumplen con el certificado de gases y control de emisiones, algunos no pasan la

prueba de emisiones realizados en el AMB, lo que dificulta la ejecución de medidas de monitoreo y control con las cuales se pueda mitigar el impacto de las concentraciones de material particulado en la calidad del aire en Bucaramanga.

La ventaja de la implementación de un modelo predictivo radica en que permite a los entes de control apoyar la toma de decisiones para la planeación de estrategias a corto y largo plazo orientadas a promover y mantener la calidad del aire. En respuesta a lo anterior, metodologías como el Deep Learning presentan una alternativa confiable y eficaz para la predicción de la calidad del aire, a pesar de ser un tema de reciente aplicación en la calidad del aire ha demostrado superar otras metodologías de análisis de datos debido su capacidad de procesamiento en cuanto a tamaño y diversidad de datos.

Por consiguiente, en este proyecto se busca adoptar métodos de Deep Learning para apoyar la toma de decisiones de las autoridades en la ciudad de Bucaramanga que realizan iniciativas dirigidas al análisis de la calidad de aire, con el fin de contribuir al bienestar de la población.

2. Objetivos

2.1. Objetivo General

Aplicar modelos Deep Learning en logística urbana para predecir la calidad del aire en la ciudad de Bucaramanga.

2.2. Objetivos Específicos

Realizar una revisión de literatura orientada al problema de la calidad del aire desde la perspectiva de la logística urbana y el uso de Deep Learning como método de análisis de datos para la toma de decisiones.

Seleccionar y adoptar modelos de Deep Learning para relacionar variables predictivas que permitan abordar el problema objeto de estudio.

Validar los modelos de Deep Learning ajustados, utilizando métricas de evaluación de desempeño.

Elaborar un artículo de carácter publicable a partir de la investigación realizada.

3. Revisión de la literatura

Para la revisión de literatura se realiza un análisis bibliométrico (Apéndice A) de los documentos que soportan el problema objeto de estudio, y a continuación se mencionan los hallazgos más

importantes.

En los últimos años, el concepto de logística urbana se ha convertido en un tema de investigación en áreas como la salud, la industria y el comercio, por lo cual las ciudades han optado por implementar procesos inteligentes para la gestión eficiente de recursos y la gestión de una vida urbana sostenible (Alvarez M. & Eslava S., 2016). A su vez, el constante crecimiento de la población y el rápido avance de la tecnología propicia un considerable aumento en la cantidad de datos producidos obligando a los gobiernos y las partes interesadas de la ciudad a tomar precauciones tempranas para procesar estos datos y predecir los efectos futuros garantizando el desarrollo sostenible (Kök, Şimşek, & Özdemir, 2018).

En este contexto de la predicción, las técnicas de análisis de datos como el Deep Learning, se han utilizado para varios problemas de pronóstico en Big Data, según (Ong, Sugiura, & Zettsu, 2016) las redes neuronales (NN, por su siglas en inglés) y sus variaciones, son los algoritmos más usados para la predicción de la calidad del aire en el mundo ya que poseen alta precisión en las tareas de predicción.

En particular, se ha demostrado que una forma de NN conocida como redes neuronales recurrentes (RNN, por sus siglas en inglés), presentan un rendimiento destacable en el modelado de estructuras temporales, el éxito de las redes profundas se atribuye a una mayor capacidad de cálculo y a nuevos métodos de capacitación que aprovechan gran cantidad de datos para entrenar con avidez capa por capa de la red de forma no supervisada (Ong et al., 2016).

En Tabla 2 se muestra que la mayoría de los modelos implementados para la predicción de la calidad del aire son aquellos que involucran redes neuronales y modelos de memoria a corto y

largo plazo, debido a que las características y variables usadas para la predicción de la calidad del aire son altamente variables y dependen en gran parte de sus relaciones espacio temporales.

Tabla 2

Resumen de la revisión de la literatura.

Citación	Modelo	Observaciones
Soh et al. (2018)	Combinación de múltiples redes neuronales, incluyendo redes neuronales artificiales, redes neuronales convolucionales y redes de memoria a largo y corto plazo.	Se realiza una combinación de redes neuronales, incluida una red neuronal artificial, una red neuronal convolucional y una red de memoria a corto y largo plazo para extraer relaciones espaciotemporales con el objetivo de pronosticar la calidad del aire para las próximas 48 horas.
Athira, Geetha, Vinayakumar, & Soman (2018)	Red neuronal recurrente (RNN), redes de memoria a corto y largo plazo (LSTM) y unidad recurrente cerrada (GRU).	Se ha encontrado un mejor desempeño en el uso de algoritmos híbridos en comparación con el desempeño de cada modelo, además se compararon los tres métodos y se concluyó que el GRU es el modelo con mejor desempeño.
Ong et al. (2016)	Redes neuronales recurrentes profundas (DRNN).	Se analiza la red de sensores mediante una red elástica, con el fin de reducir costos, al descartar sensores innecesarios para trabajar con sensores altamente correlacionados preservando la precisión de los resultados.
Qi et al. (2018)	Deep Air Learning (DAL).	Para una efectiva predicción del modelo se debe realizar una selección preliminar de las variables a utilizar para identificar las más relevantes para el estudio.

Continuación de la Tabla 1.

Resumen de la revisión de la literatura

Citación	Modelo	Observaciones
Li, Peng, Hu, Shao, & Chi (2016)	Modelo de regresión espaciotemporal (STDL).	El modelo de regresión lineal múltiple (MLR) y el modelo autorregresivo de media móvil (ARMA, por sus siglas en inglés) se utilizan comúnmente para la predicción de la calidad del aire. Sin embargo, estos métodos generalmente producen una precisión limitada debido a su incapacidad para modelar patrones no lineales.
Li et al. (2017)	Nueva red de memoria a corto y largo plazo extendida (LSTME).	La efectividad de la predicción alcanzo un 90.3%, se utilizaron tres algoritmos Deep Learning; un modelo espaciotemporal (STDL), una red neuronal de retardo de tiempo (TDNN) y un modelo de promedio móvil autorregresivo (ARMA), los cuales demostraron tener un mejor desempeño, corroborando lo dicho en estudios anteriores. Además, se identificó que el uso de datos meteorológicos y espaciotemporales pueden mejorar el desempeño en la predicción.
Wen et al., (2019)	C-LSTME.	Al comparar la precisión de la predicción de los modelos Deep Learning; de redes de memoria a largo y corto plazo extendidas (LSTME) y el modelo C-LSTME en la predicción a largo plazo, se puede encontrar que la precisión de ambos modelos disminuye a medida que se extiende el tiempo de predicción.

Continuación de la Tabla 1.

Resumen de la revisión de la literatura

Citación	Modelo	Observaciones
Bai, Zeng, Li, & Zhang (2019)	Ensamble de una red de memoria a corto y largo plazo (E-LSTM).	El modelo propuesto fue puesto a prueba usando dos sensores o estaciones de monitoreo, además se compara con dos modelos, el E-LSTM y el LSTM simple, los resultados en el pronóstico demuestran que, siguiendo la estrategia propuesta en el modelo E-LSTM, este tiene el mejor desempeño realizando el pronóstico.
Liu et al. (2019)	Nuevo modelo hibrido, (EWT-MAEGA-NARX)	Se diseña un modelo hibrido combinando el transformada de ensamble de Wavelet (EWT), algoritmo genético evolutivo multiagente (MAEGA) y la red autorregresiva con entradas exógenas (NARX). En conclusión, el modelo propuesto ofrece una metodología nueva para la predicción y prevención de concentración de contaminantes mejorando el desempeño de la predicción.
Lu, Song, Di, Kurdestany, & Wang (2018)	Nueva red de creencias profundas (DBN) basada en la combinación de máquinas de restricción multicapa de Boltzman y una red de retro propagación de capas simple	El coeficiente de correlación entre el valor de turbidez real y la predicción del modelo propuesto es de 0,8, y el error absoluto medio (MAE) es de 26 %. En comparación con los algoritmos de predicción tradicionales, el CC mejora en promedio un 18%, mientras que el MAE se reduce en 15.7 %.

Teniendo en cuenta la información recopilada, un factor clave para la precisión del modelo implementado es la correcta elección de variables y la adecuada interpretación de las mismas; generalmente se consideran tres conjuntos de variables para analizar la calidad del aire: partículas contaminantes presentes en el aire (PM10, PM2.5, O3, CO, CO2, NO2 y SO2), variables meteorológicas (temperatura, humedad, precipitación, velocidad y dirección del viento) y datos de las estaciones de monitoreo (distancia entre estaciones y ubicación).

Entre las variables de la calidad del aire se consideran principalmente peligrosas las concentraciones de material particulado (PM) que es una mezcla de partículas de sustancias orgánicas e inorgánicas, que se encuentran en suspensión en el aire, se catalogan en función de su tamaño que se expresa en unidades de micrones de metro (μm), es decir, una millonésima parte del metro. Las partículas de mayor tamaño son las PM10, 10 μm o menores, son perjudiciales para la salud ya que pueden quedar retenidas en las vías respiratorias, produciendo efectos a nivel de sistema respiratorio. A su vez, las partículas de menor tamaño las PM2.5, 2.5 μm o menores, al ser más pequeñas, tienen la capacidad de pasar al torrente sanguíneo y potencialmente pueden dañar cualquier órgano o sistema (Fundación para la Salud Geoambiental, n.d.).

Los modelos con fines predictivos deben someterse a un proceso de verificación y validación, que implica determinar los parámetros de validación para comprobar la exactitud de un método respecto a otro, en donde comúnmente se utilizan tres métricas de validación para métodos con fines de regresión: el error absoluto medio de los resultados del pronóstico (MAE, por sus siglas en inglés), el error cuadrático medio de las predicciones (MSE, por sus siglas en inglés) y la raíz cuadrada del promedio de las diferencias cuadradas entre la predicción y la observación real (RMSE, por sus siglas en inglés).

Específicamente el MAE calcula el promedio de los errores en un conjunto de predicciones, sin tener en cuenta su dirección, es decir, determina la diferencia entre la muestra de prueba y la observación real. El RMSE también determina el promedio de los errores, pero a su vez, les otorga un peso relativamente alto a los errores grandes, por ello es muy útil cuando se quiere evitar o identificar errores grandes. El MSE por su parte, calcula la diferencia cuadrada entre las predicciones y el objetivo para cada punto y luego promedia esos valores, cuanto mayor sea este valor, peor es el ajuste del modelo.

Teniendo en cuenta lo anterior y considerando que predecir la concentración de contaminantes del aire es fundamental para fortalecer la prevención de la contaminación y contribuir a mejorar la salud en la población, el presente proyecto se enfoca en la aplicación de modelos Deep Learning para la predicción del material particulado PM2.5, implementando redes LSTM que proporcionan mayor precisión y menor error en el análisis de variables, para poder brindar a las entidades gubernamentales una herramienta que apoye la toma de decisiones y el planteamiento de estrategias ambientales a corto y largo plazo.

4. Marco de antecedentes

José Manuel Buendía Martínez en su trabajo de maestría, LOGÍSTICA SOSTENIBLE: ESTUDIO DE LA CALIDAD DEL AIRE E INTERACCIÓN SOBRE LA MOVILIDAD URBANA (Buendía Martínez, n.d.), refiere que el tráfico es una de las principales fuentes de contaminación debido a que la emisión se produce muy cerca a la población y de forma muy dispersa, a su vez, establece que las emisiones de motores Diesel, principalmente provenientes de las motocicletas, y

el envejecimiento del parque automotor son los mayores contaminantes atmosféricos generados por el parque automotor.

De este trabajo se destaca la relación que establece entre la logística urbana y la calidad del aire, determinando por ejemplo las horas pico del tráfico; entrada al trabajo y centros escolares a primera hora de la mañana y con la salida del trabajo entre las 7 y 8 de la noche, de lunes a viernes. A su vez, presenta algunas alternativas y estrategias de transporte que podrían contribuir a la reducción de la contaminación del aire.

Amaya Martínez Manuel Ignacio, Gómez Ordoñez Nancy & Rey Estupiñan Isabel Cristina en su trabajo de grado titulado: RECOPILACIÓN Y ANÁLISIS DE LA INFORMACIÓN DE LA CALIDAD DEL AIRE EN EL ÁREA METROPOLITANA DE BUCARAMANGA (2009) (Amaya, M. Gómez, Nancy. Rey, 2009) manifiestan que la evaluación de los impactos de la contaminación del aire en la salud permite la asociación cuantitativa de los criterios de valoración para un contaminante específico y su efecto en la salud, dicha información favorece la promoción de mejoras costo-efectivas en la salud pública. Para el análisis cuantitativo del impacto de la contaminación del aire en la salud, se debe conocer las concentraciones de la calidad del aire, cantidad de exposición y población expuesta y su incidencia en la mortalidad y morbilidad ya que las concentraciones de PM pueden verse afectadas por la concentración de la población, dirección y desplazamiento del viento, tráfico vehicular y aumento de la construcción inmobiliaria.

A su vez menciona que la base de datos de la CDMB para el 2009, utilizaba archivos planos para exportar los datos mensuales de las estaciones, esos datos eran importados a archivos de Excel, uno por estación. Estas bases de datos incluyen los datos de concentración y otros datos de validación, que presentaban los datos separados por día y se realizaban los cálculos de los

promedios requeridos para compararlo con la norma de calidad vigente.

Se consideró esta investigación ya que establece el monitoreo y control de los contaminantes del aire que realizaba la CDMB para el año 2009, define la modelación y sistemas de control usados, así como las herramientas de apoyo para el análisis de los datos y la incidencia de las emisiones del tráfico de vehículos en la calidad del aire de Bucaramanga y su área metropolitana.

5. Marco Teórico

5.1. Logística urbana

La logística urbana o logística de la última milla, abarca todos los movimientos relacionados con la actividad comercial y el suministro y distribución de bienes en las ciudades, aspecto fundamental para su desarrollo económico y uno de los principales causantes de la congestión del tránsito y de la emisión de contaminantes (Observatorio regional de logística, n.d.).

5.2. Movilidad urbana

La Movilidad Urbana es el conjunto de desplazamientos, de personas y mercancías, que se producen en una ciudad independiente de su medio o sistema de transporte, un factor determinante tanto para la productividad económica de la ciudad como para la calidad de vida de sus ciudadanos, sin embargo, los problemas que genera la movilidad referente a perdida en tiempo y contaminación medioambiental son unos de los principales retos para las ciudades al buscar equilibrio entre la eficiencia de los traslados y uso de medios de transporte y un ambiente sano para sus habitantes.

5.3. Ciudades inteligentes

Las Ciudades Inteligentes, son ciudades que aplican las TIC (Tecnologías de Información y Comunicación) para la gestión y prestación de sus diferentes servicios, como gobernanza, economía, asuntos sociales, movilidad, seguridad, energía, cultura, medio ambiente, etc., lo que conlleva a un mejor desarrollo económico, social, mayor eficiencia administrativa y mejor calidad medioambiental. (Grupo Tecma Red S.L., n.d.)

5.4. Calidad del Aire

La calidad del aire se define como la concentración de contaminante que llega a un receptor, más o menos alejado de la fuente de emisión, así, para determinar de forma inequívoca los distintos rangos de Calidad del Aire se ha desarrollado el índice de Calidad del Aire, a través del cual se define si dicha Calidad en un lugar es buena, admisible, mala o muy mala. (Buendía Martínez, n.d.)

5.4.1. Índice de Calidad del aire. El Índice de Calidad del Aire – ICA, corresponde a un valor adimensional, que permite de manera cualitativa identificar la calidad del aire de la ciudad y su efecto en la salud humana (Observatorio Ambiental de Cartagena de Indias, n.d.)

La contaminación ambiental en un tema de interés general, por ello, los responsables de las redes de vigilancia deben establecer el mejor método para informar al público sobre los niveles de concentración registrados en las estaciones sobre los distintos contaminantes del aire, sin embargo, debido al desconocimiento general que existe sobre el tema de la contaminación atmosférica, estos datos no cumplen su objetivo en la mayoría de los casos, ya que el público al que llegan puede no estar preparado para interpretarlos. (Xunta de Galicia, 2014).

Para facilitar la compresión de los índices de calidad, se han establecido etiquetas (“bueno”, “malo”, “regular”) y un código de colores para facilitar su interpretación, los valores del ICA se ubican en una escala adimensional de 0 a 500, agrupadas en 6 rangos cada uno de estos rangos ha sido asociado a un color y etiqueta que sirve de alerta, así se podrá informar de forma comprensible una idea del estado de contaminación, ver tabla 3.

Tabla 3.

Índice de Calidad del Aire y sus efectos en la salud. Adaptado de Índice de calidad del aire (ICA) (IDEAM, 2012).

Rango	**Color**	**Estado de la calidad del aire**	**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 realizan constantemente actividad física al aire libre, deben 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 de 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	Peligroso	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.

5.5. Material particulado

El material particulado (MP) es un conjunto de partículas sólidas y líquidas emitidas directamente al aire, tales como el hollín de Diesel, polvo de vías, el polvo de la agricultura y las partículas resultantes de procesos productivos (Arciniégas Suárez, 2012); son de tamaño, forma y composición variada y se pueden clasificar según su diámetro como finas y gruesas siendo las partículas PM10, las de mayor tamaño con diámetro igual o inferior a 10 µm (micrones de metro = millonésima parte del metro) y las partículas finas PM 2.5 cuyo diámetro sería de 2.5 µm. (Fundación para la Salud Geoambiental, n.d.).

5.6. Variables meteorológicas

A continuación, se describen brevemente las variables meteorológicas objeto de estudio:

5.6.1. Temperatura ambiente. Indica la cantidad de energía calorífica que hay acumulada en el aire en un momento y lugar determinados, (Jóvenes frente al cambio climático, n.d.), es decir, indica el nivel de calor que posee el aire en un lugar y momento determinados.

5.6.2. Lluvia. La lluvia es la caída de agua desde la atmósfera hacia la superficie terrestre (“Precipitación,” n.d.)**.**

5.6.3. Humedad Relativa. La humedad relativa es un valor porcentual, entendido como la capacidad que tiene el aire de absorber más humedad, es decir, cuánta agua puede contener el aire (Leite, n.d.).

5.6.4. Dirección del viento. La dirección indica de dónde proviene el viento, su unidad de medición es en grados Dextrorsum (giro en sentido de las manecillas del reloj) donde 0° es norte

verdadero (Gobierno de Mexico, n.d.).

5.6.5. Radiación Solar. La radiación solar es la energía emitida por el Sol, que se propaga en todas las direcciones a través del espacio mediante ondas electromagnéticas (IDEAM, n.d.-c).

5.7. Aprendizaje automático

La capacidad de los ordenadores de "aprender" sin que se les indique las reglas que debe seguir, es conocido como Aprendizaje automático como cita Lugo-Reyes, Maldonado-Colín, & Murata, 2014).

> *"El aprendizaje automático es un subcampo de la inteligencia artificial (Singh et al., 2018) y comprende un conjunto de métodos que pueden detectar patrones automáticamente en datos y entonces utiliza los patrones descubiertos para predecir datos fututos. Según Tom Mitchell "una computadora aprende de la experiencia (E) con respecto a alguna clase de tareas (T) y medida de desempeño (P), si su desempeño en las tareas T, medida mediante P, mejora con la experiencia E"*

5.7.1. Categorías del aprendizaje automático. Según, Torres, J (2018), el aprendizaje automático es un amplio campo con una compleja taxonomía de algoritmos que se agrupan, en general, en tres grandes categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado.

5.7.1.1. *Aprendizaje supervisado.* Se presenta cuando los datos que se usan para el entrenamiento incluyen la solución deseada, llamada "etiqueta" que representa lo que se está intentando predecir a través de una variable de entrada, estos algoritmos son aquellos que usan la

regresión lineal, la regresión logística, las máquinas de soporte vectorial (SVM), árbol de decisión, bosques aleatorios y redes neuronales (NN).

5.7.1.2. ***Aprendizaje no supervisado.*** En este tipo de aprendizaje los datos de entrenamiento no incluyen etiquetas y el algoritmo utilizado intenta clasificar la información de acuerdo con los patrones identificados en los datos de entrada.

5.7.1.3. ***Aprendizaje reforzado.*** En el aprendizaje reforzado, el modelo debe explorar un espacio desconocido y determinar las acciones a llevar a cabo mediante prueba y error, es decir que el algoritmo aprenderá por sí mismo debido a las recompensas y penalizaciones que obtiene de sus acciones.

5.7.2. Fases del Aprendizaje Automático. Según, Torres, J. (2018) el aprendizaje automático se caracteriza por dos fases claramente diferenciadas:

5.7.2.1. ***Fase de entrenamiento.*** Se presenta cuando se crea o se "enseña" al modelo, mostrándole los ejemplos de entrada que se tienen etiquetados; de esta manera se consigue que el modelo aprenda iterativamente las relaciones entre las características y etiquetas de los ejemplos.

5.7.2.2. ***Fase de inferencia.*** Fase de inferencia o predicción, se refiere al proceso de hacer predicciones aplicando el modelo ya entrenado a ejemplos no etiquetados.

5.7.3. Conjunto de datos para entrenamiento, validación y prueba. Para la configuración y evaluación de un modelo en aprendizaje automático y por ende del Deep Learning se divide en tres conjuntos de datos: datos de entrenamiento que se usan para que el algoritmo de aprendizaje obtenga los parámetros del modelo, datos de validación y datos de prueba.

Si el modelo no se adapta a los datos de entrada, se modifican los hiper parámetros y después de entrenar nuevamente con los datos de entrenamiento se volvería a evaluar con los de validación, hasta obtener resultados que se consideren correctos (Torres, 2018).

5.8. Deep Learning

Es un subcampo específico del aprendizaje automático; realiza representaciones de aprendizaje a partir de capas sucesivas (profundidad) y de representaciones (jerarquías) cada vez más significativas, dichas capas son entrenadas, por lo general mediante modelos de redes neuronales.

El Deep Learning se basa en algoritmos de redes neuronales artificiales tradicionales estándar que han demostrado ser predictores de alta precisión para problemas de clasificación y regresión, utiliza redes más amplias y profundas, entrenadas con grandes conjuntos de datos. El proceso de capacitación en conjunto con la profundidad de las redes permite el aprendizaje de las abstracciones de datos en diferentes profundidades para desenredar características complejas (Bellinger, Mohomed Jabbar, Zaïane, & Osornio-Vargas, 2017); por tanto, puede llevar a un rendimiento destacable en las predicciones de la calidad del aire al extraer características del proceso sin conocer información previa (Athira et al., 2018).

Básicamente el aprendizaje automático se encarga de asignar entradas a objetivos y las redes neuronales se encargan de realizar dicha asignación a través de una serie de transformaciones de los datos (capas) mediante la exposición a ejemplos (aprendizaje). Este aprendizaje significa encontrar un grupo de valores para los “parámetros o pesos” de cada una de las capas de la red, pero cada red puede tener millones de parámetros y modificar uno de esos parámetros puede afectar los demás. Inicialmente las capas tienen parámetros aleatorios, por ello se deben controlar

las salidas, es decir, que tan lejos o cerca está la salida resultante de la respuesta esperada, este es el trabajo de la función de pérdida o función objetivo; usar dicha "distancia" para ajustar los parámetros y orientarlos en una dirección que disminuya la brecha entre la salida esperada y la resultante.

5.9. Redes neuronales artificiales

Las redes neuronales artificiales (NNA, por sus siglas en inglés) son algoritmos basados en el funcionamiento de las redes neuronales biológicas. Consisten en un gran número de elementos simples de procesamiento llamados nodos o neuronas que están organizados en capas. Cada neurona está conectada con otras neuronas mediante enlaces de comunicación, cada uno de los cuales tiene asociado un peso. Los pesos representan la información que será usada por la red neuronal para resolver un problema determinado, las NNA aprenden de la experiencia, es decir, aprenden a llevar a cabo ciertas tareas mediante un entrenamiento con ejemplos ilustrativos lo que les permite crear su propia representación interna del problema, es decir, las NNA son capaces de generalizar de casos anteriores a casos nuevos. Esta característica es fundamental ya que permite a la red responder correctamente ante información novedosa e información distorsionada o incompleta (Palmer Pol & Montaño Moreno, 1999).

En la figura 1, se representa gráficamente una red neuronal artificial con 3 capas: una de entrada que recibe los datos de entrada y una de salida que devuelve la predicción realizada. Las capas que se tienen en medio se llaman capas ocultas y se pueden tener muchas, cada una con distinta cantidad de neuronas (Torres, 2018).

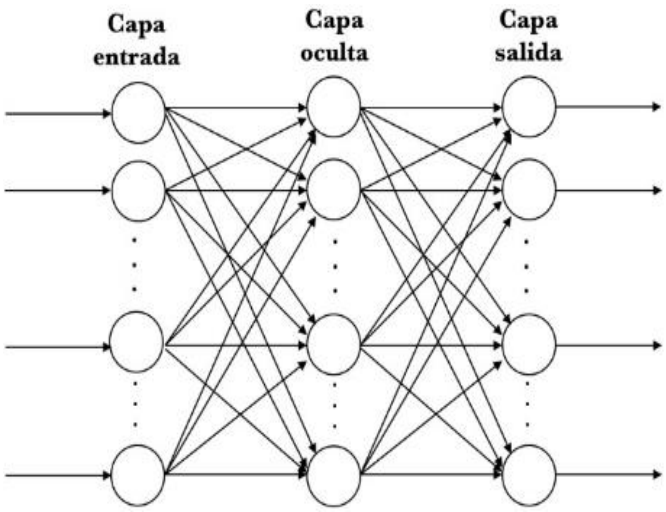

Figura 1. Aproximación gráfica a las redes neuronales. Adaptado de Torres, 2018

Dentro de los modelos de las Redes Neuronales se encuentran:

5.9.1. Redes de Avance Profundo. Las redes de avance profundo, también llamadas redes neuronales de avance de alimentación, o perceptrones de múltiples capas (MLP, por sus siglas en inglés), son los modelos de aprendizaje profundo por excelencia.

El objetivo de una red de avance es aproximarse a alguna función *f*. Por ejemplo, para un clasificador, $y = f*(x)$ asigna una entrada x a una categoría. Una red de avance de datos define una $correlación = f(x; \theta)$ y aprende el valor de los parámetros, que resultan en la mejor aproximación de la función. Estos modelos se denominan retroalimentación porque la información fluye a través de la función que se evalúa desde x, realizando los cálculos intermedios que se utilizan para definir f, finalmente proporcionan una salida. Cuando se extienden las redes neuronales de avance profundo para incluir conexiones de retroalimentación, se denominan redes neuronales recurrentes (Goodfellow, Bengio, & Courville, n.d.).

5.9.2. Redes neuronales convolucionales. Las redes neuronales convolucionales son muy

similares a las redes neuronales ordinarias; se componen de neuronas que tienen pesos y sesgos que pueden aprender. Cada neurona recibe algunas entradas, realiza un producto escalar y luego aplica una función de activación. Lo que diferencia a las redes neuronales convolucionales es que suponen explícitamente que las entradas son imágenes, lo que permite codificar ciertas propiedades en la arquitectura, permitiendo ganar en eficiencia y reducir la cantidad de parámetros en la red.

Las redes neuronales convolucionales trabajan modelando de forma consecutiva pequeñas piezas de información, y luego combinando esta información en las capas más profundas de la red (Lopez Briega, 2016). A su vez, estructuras algorítmicas permiten modelos que están compuestos de múltiples capas de procesamiento para aprender representaciones de datos, con múltiples niveles de abstracción que realizan una serie de transformaciones lineales y no lineales que a partir de los datos de entrada generen una salida próxima a la esperada (Torres, 2018).

5.9.3. Redes neuronales recurrentes. Las redes neuronales recurrentes (RNN, por sus siglas en inglés) son una clase de redes neuronales que poseen conexiones de retroalimentación entre unidades formando así un ciclo dirigido, han demostrado ser especialmente adecuadas para capturar la evolución espaciotemporal de las distribuciones de contaminantes del aire porque los RNN pueden manejar secuencias arbitrarias de entradas, garantizando así la capacidad de aprender secuencias temporales.

Existe además una arquitectura RNN especial denominada red neuronal de memoria a largo plazo (LSTM NN) que a diferencia de las RNN tradicionales, las LSTM son capaces de aprender largas series de tiempo, esta característica es especialmente importante para el modelado de contaminantes atmosféricos espaciotemporales (Li et al., 2017).

5.9.3.1. ***Red neuronal de memoria a largo plazo.*** Son una clase de redes neuronales conectadas entre sí mediante enlaces de comunicación, generando conexiones de retroalimentación entre ellas formando un ciclo dirigido. En los últimos años, las redes de memoria a largo plazo (LSTM) se ha aplicado con éxito a muchos estudios relacionados con la predicción de series de tiempo, como la predicción del flujo de tráfico, la predicción de la energía eólica, predicción de la trayectoria humana, etc.

5.9.3.2. ***Modelo secuencia a secuencia.*** Un modelo de secuencia a secuencia – seq2seq tiene como objetivo asignar una entrada de longitud fija con una salida de longitud fija donde la longitud de la entrada y la salida pueden diferir. El modelo consta de 3 partes; codificador, vector codificador y decodificador (Kostadinov, n.d.).

5.9.3.2.1. *Codificador*. Es un conjunto de varias unidades recurrentes LSTM donde cada una acepta un solo elemento de la secuencia de entrada, recopila información para ese elemento y la propaga hacia adelante.

5.9.3.2.2. *Vector codificador*. Este vector tiene como objetivo encapsular la información de todos los elementos de entrada para ayudar al decodificador a hacer predicciones precisas.

5.9.3.2.3. *Decodificador*. Una pila de varias unidades LSTM donde cada una predice una salida *y_t* en un paso de tiempo *t*. Cada unidad recurrente acepta un estado oculto de la unidad anterior y produce y genera su propio estado oculto.

5.10. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) inicialmente fueron utilizadas para resolver problemas de clasificación binaria, la idea es seleccionar un hiperplano de separación

que equidista de los ejemplos más cercanos de cada clase para conseguir lo que se denomina un margen máximo a cada lado del hiperplano (Carmona Suárez, 2013), actualmente ha demostrado ser una buena herramienta para el análisis predictivo de datos para resolver problemas de regresión y clasificación con múltiples clases. Además, han sido utilizadas con éxito en diversos campos, tales como: visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, procesamiento de lenguaje natural, y análisis de series temporales.

5.11. Imputación de datos

Los datos faltantes, son un problema común en los procesos de investigación, ya sea por muestreo insuficiente, errores en las mediciones o fallas en la adquisición de datos, estas discontinuidades representan un obstáculo significativo para los esquemas de predicción de series temporales (Kök et al., 2018).

Uno de los métodos más comunes para la sustitución de valores es el método de aproximación simple no paramétrica es el basado en la regla del vecino más cercano, que consiste en estimar el valor de un dato desconocido a partir de las características del dato más próximo, método del vecino más cercano se puede extender utilizando no uno, sino un conjunto de datos más cercanos para predecir el valor de los nuevos datos, en lo que se conoce como los k-vecinos más cercanos (k-NN o k-Nearest Neighbors). Al considerar más de un vecino, se brinda inmunidad ante ruido y se suaviza la curva de estimación (Morales, German. Mora, Juan. Vargas, 2008)

5.12. Análisis de componentes principales

El Análisis de componentes principales (PCA por sus siglas en inglés) es una técnica estadística descriptiva utilizada para describir un conjunto de datos en términos de nuevas variables

("componentes") no correlacionadas, es efectivo cuando la correlación entre variables es alta y toma como punto de partida una matriz de datos ordenados por la cantidad de varianza original que describen, es una técnica útil para reducir la dimensionalidad de un conjunto de datos.

6. Metodología

El desarrollo de la investigación apoyó el proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases"KDD", por sus siglas en inglés), adaptado según las necesidades específicas del proyecto como se observa en la figura 2.

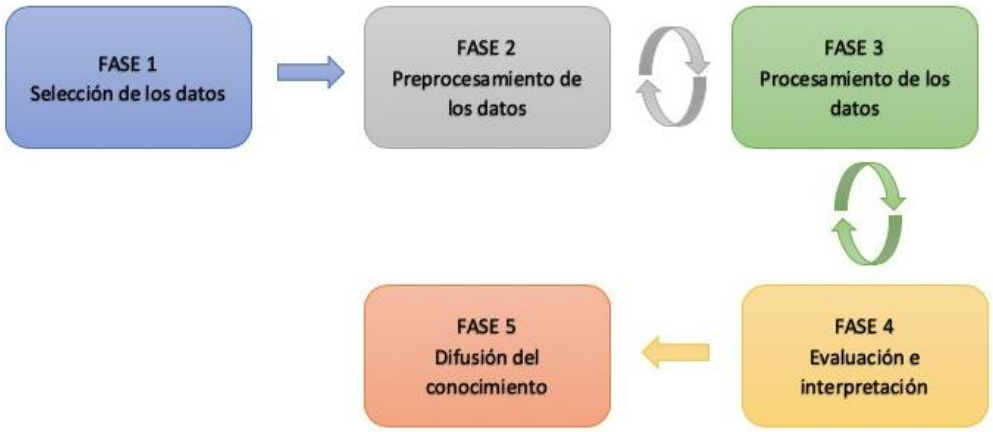

Figura 2. Relación grafica de la metodología

A continuación, se detallan las actividades realizadas en las diferentes etapas del proceso de descubrimiento de conocimiento que se apoyan en el lenguaje de programación Python.

6.1. Selección de los datos

Los datos que apoyan el objeto de esta investigación son suministrados por la Subdirección

Ambiental de la AMB, se obtienen datos de tres estaciones del Sistema de Vigilancia de la Calidad del Aire (SVCA), ubicadas en el Instituto Caldas, el La Normal y el Colegio el Pilar (ver figura 3).

En cada estación se capturan datos cada diez (10) minutos desde agosto de 2018 hasta junio de 2019, en total se encuentran 136.614 registros distribuidos así: 45.413 de la Estación Caldas, 45.462 de la Estación La Normal y 45.739 de la Estación El Pilar; estas bases de datos reportan siete variables ambientales (Temperatura interna de los sensores, Temperatura aire, Lluvia, Humedad relativa, Velocidad y Dirección del viento y Radiación solar) y dos variables de material particulado (PM10 y PM2,5).

El repositorio objeto de estudio está conformado por seis variables independientes (TA, LL, HR, WD, RS Y PM10) y una variable dependiente (PM2.5). Se toma PM2.5 como variable objetivo al ser la de mayor influencia en la salud de las personas.

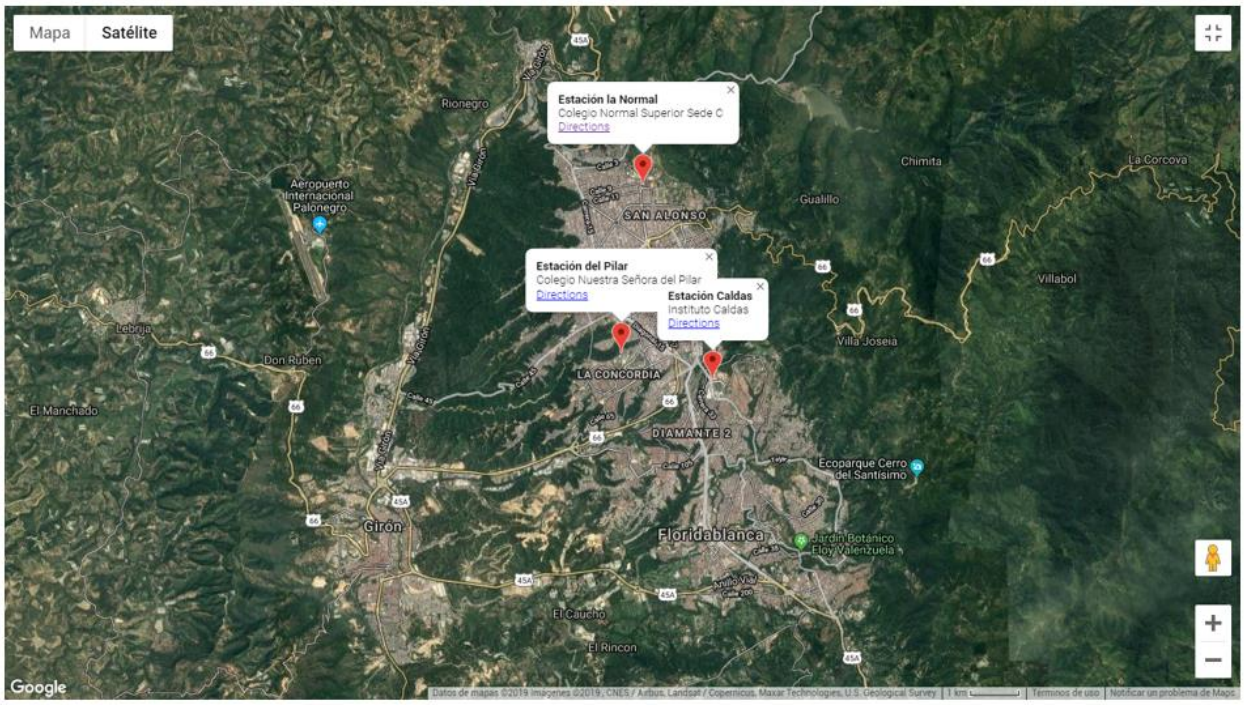

Figura 3. Mapa estaciones de la Subdirección Ambiental de la AMB.

En la tabla 4 se presenta la nomenclatura y unidades para las variables objeto de estudio.

Tabla 4

Nomenclatura de las variables utilizadas

Variable	Unidades de medida	Nomenclatura
Temperatura Ambiente	°C	*TA*
Lluvia	mm	*LL*
Humedad Relativa	%	*HR*
Dirección del Viento	Grados	*WD*
Radiación Solar	W/m^2	*RS*
Material particulado PM10	ug/m^3	*PM10*
Material particulado PM2,5	ug/m^3	*PM2.5*

6.2. Preprocesamiento de los datos

A continuación, se detalla el proceso de limpieza e imputación de datos que permiten obtener el repositorio final objeto de estudio:

Limpieza de los datos

Después de la revisión preliminar de la información se realiza la limpieza que consiste en eliminar 19.120 datos no válidos que corresponden al 13,99% de los datos suministrados (ver Tabla 5), según la AMB, estos datos no válidos pueden ser causa de actividades de mantenimiento, calibración de los sensores o fallas en la medición de los mismos con relación a alguna de las variables, razón por la cual la mayoría de datos perdidos se encuentran en los primeros meses de funcionamiento de los sensores.

Tabla 5

Cantidad de datos del repositorio objeto de estudio y datos perdidos.

	Datos totales	**Datos después de eliminar datos no válidos**	**Datos perdidos**	**% Datos perdidos**
Caldas	45.413	38.260	7.153	15,7510
Normal	45.462	36.601	8.861	19,4910
Pilar	45.739	42.633	3.106	6,7907
Total	***136.614***	***117.494***	***19.120***	***13,9956***

Imputación de datos

Después, se realiza la imputación de los datos utilizando el método del promedio de los k-vecinos más próximos, en dónde se reemplazan valores para 112 datos (ver figura 4); se resalta que para este estudio no se considera la Temperatura interna del sensor, dado que no es relevante para el problema objeto de estudio, también se descarta la Velocidad del viento, debido a la falta de datos continuos por largos periodos de tiempo, del 25 de marzo al 13 de mayo de 2019, cerca de 7.200 datos.

18/08/2018 08:00	23,8	24,8	0	62,9	340,6	1,5	468,4	30,4	15,5
18/08/2018 08:10	24	25,5	0	59,2	352	2,2	528,5	27,1	15,1
18/08/2018 08:20	24,1	25,7	0	58,4	346,3	2,1	568,3	24,7	14,2
18/08/2018 08:30	24,1	26	0	55,5	334,2	1,4	616,7	23,5	12,5
18/08/2018 08:40	24,3	27,1	0	50,7	342,7	1,9	652,5	22,2	12,2
18/08/2018 08:50	24,4	27,2	0	51,6	335,9	<Samp	685,7	20,7	11,8
18/08/2018 09:00	24,3	27	0	50,3	343,8	1,8	712,4	20,9	11,1
18/08/2018 09:10	24,4	27,5	0	48,7	340,9	1,6	740,1	17,3	10,8
18/08/2018 09:20	24,5	28	0	47,6	335,4	1	764,8	24,1	11,6
18/08/2018 09:30	24,5	28,3	0	46,4	333,3	1,2	794	24,4	11,9
18/08/2018 09:40	24,5	28,3	0	46	342	1,8	844,2	18,8	10,5

Figura 4. Dato para imputación.

La imputación del dato seleccionado se realiza promediando los 5 datos superiores y los 5 datos inferiores al dato objetivo obteniendo un valor cercano a los datos circundantes (ver ecuación 1).

$$Dato_x = \frac{1.5 + 2.2 + 2.1 + 1.4 + 1.9 + 1.8 + 1.6 + 1 + 1.2 + 1.8}{10} = 1.65 \approx 1.7 \qquad (1)$$

A continuación, se realiza el análisis descriptivo respecto al cálculo de la media, desviación, varianza, análisis de correlación, graficas entre otros, de las variables que conforman el repositorio objeto de estudio.

En las Tablas 6, 7 y 8, se observa los valores promedio obtenidos para las variables meteorológicas los cuales se encuentran dentro del rango normal para la ciudad de Bucaramanga según el Atlas Interactivo del IDEAM, (IDEAM, n.d.-a); a su vez, el material particulado, se encuentra dentro del rango promedio para un tiempo de exposición de 24 horas, teniendo en cuenta que el máximo permitido es 75 $\mu g/m^3$ para PM10 y 37 $\mu g/m^3$ para PM2.5 (Ministerio de Ambiente y Desarrollo Sostenible, 2017a).

Cabe resaltar que valores altos de PM10 y PM2,5 que se desvían de la media, corresponden a horas pico de flujo vehicular, de manera similar, los valores altos y bajos presentados en radiación solar (RS) corresponden al comportamiento natural de la variable, reflejado por alta intensidad solar cerca al medio día y su descenso en la noche.

Tabla 6

Descriptivo de datos Estación Caldas.

Estación Caldas							
	TA	**LL**	**HR**	**WD**	**RS**	**PM10**	**PM2,5**
Cantidad de Datos	38.261	38.261	38.261	38.261	38.261	38.261	38.261
Promedio	24,44	0,01	71,13	255,92	197,87	30,21	16,14
Desviación estándar	3,09	0,16	13,82	91,80	282,34	19.80	12.86
Mínimo	18,00	0,00	25,80	0,00	0,00	2.1	1,10
25%	21,90	0,00	60,70	180,30	0,00	16,50	7,30
50%	23,90	0,00	74,10	298,00	8,00	24,30	11,60
75%	26,80	0,00	82,10	330,30	350,30	37,60	20,10
Máximo	33,90	7,60	97,20	360,00	1.238,90	332.20	215.7

Tabla 7

Descriptivo de datos Estación Normal.

Estación Normal							
	TA	**LL**	**HR**	**WD**	**RS**	**PM10**	**PM2,5**
Cantidad de Datos	36.601	36.601	36.601	36.601	36.601	36.601	36.601
Promedio	24,06	0,01	70,60	252,28	202,63	30,09	15,13
Desviación estándar	2,30	0,12	12,65	93,77	293,00	14.84	8,81
Mínimo	18,70	0,00	29,30	0,10	0,00	1,80	1,00
25%	22,20	0,00	61,10	172,30	0,00	19,10	8,40
50%	23,80	0,00	72,20	295,20	8,00	27,50	13,10
75%	25,80	0,00	80,70	324,80	359,60	38,50	19,90
Máximo	31,60	7,00	98,10	360,00	1.344,00	478.5	323,80

Tabla 8

Descriptivo de datos Estación Pilar.

Estación Pilar							
	TA	**LL**	**HR**	**WD**	**RS**	**PM10**	**PM2,5**
Cantidad de Datos	42.633	42.633	42.633	42.633	42.633	42.633	42.633
Promedio	22,30	0,01	71,98	203,41	202,12	30,22	16.44
Desviación estándar	3,01	0,16	13,66	90,95	283,51	15.72	10.77
Mínimo	15,90	0,00	26,80	0,00	0,00	0.1	0.10
25%	19,80	0,00	61,70	139,20	1,00	18.5	8.7
50%	21,70	0,00	74,20	185,60	18,80	27,1	13.7
75%	24,60	0,00	82,80	295,60	364,00	38,6	21.3
Máximo	36,70	7,80	98,90	360,00	1.257,90	365.5	258.8

Como se puede ver en la tabla 9, la variable Lluvia tiene una moda de cero, lo cual se explica por la tendencia climática de Bucaramanga en la cual se presentan largos periodos de tiempo sin lluvia, en cuanto a RS debido a su comportamiento cíclico natural es de esperarse que durante la noche y madrugada sus valores sean cercamos a cero.

Tabla 9

Cálculo de la moda.

Cálculo de la Moda			
Variable	**Caldas**	**Normal**	**Pilar**
TA	22.5	22.1	19.8
LL	0	0	0
HR	78	82.1	83.2
WD	337.8	328.2	153
RS	0	0	0.1
PM10	18.3	21.3	18.1
PM2.5	5.9	7.6	8.7

En la matriz de correlación (tabla 10 y figura 5) se presentan las relaciones entre variables, respecto a nuestra variable objetivo (PM2.5), existe una relación fuerte con PM10 (0,8890), después de la relación fuerte negativa entre humedad relativa y temperatura del aire que está alrededor de 0,89.

Por otro lado, hay una relación negativa débil entre la lluvia y el material particulado, lo que lleva a deducir que debe haber efecto de otras variables para que las concentraciones de material particulado aumenten o disminuyan.

A pesar de su alta correlación, no se descarta PM10 ya que se pueden utilizar datos de PM10 para predecir con confianza la concentración de PM2.5 (Rojas & Galvis, 2005)

Tabla 10

Tabla de correlación de las variables.

Correlación							
	TA	**LL**	**HR**	**WD**	**RS**	**PM10**	**PM2,5**
TA	1	-0,0833	-0,8901	0,0543	0,7246	0.1448	0,1201
LL	-0,0833	1	0,1032	0,0132	-0,0506	-0,0388	-0,0315
HR	-0,8901	0,1032	1	0,0187	-0,7736	-0,1825	-0,1455
WD	0,0543	0,0132	0,0187	1	-0,0150	-0,0446	-0,0626
RS	0,7246	-0,0506	-0,7736	-0,0150	1	0,0520	-0,0030
PM10	0,1448	-0,0388	-0,1825	-0,0446	0,0520	1	0,8890
PM25	0,1201	-0,0315	-0,1455	-0,0626	-0,0030	0,8890	1

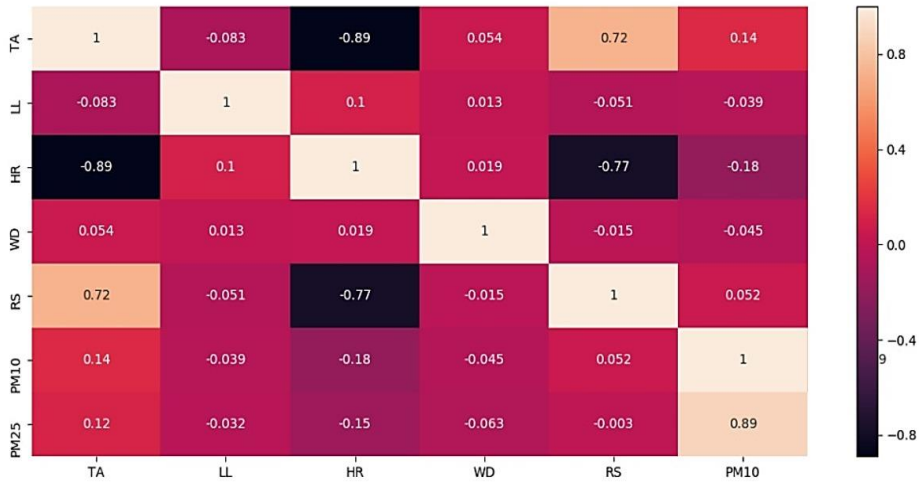

Figura 5. Relación grafica de correlación.

A continuación, se encuentra el análisis por cada variable por estación y por periodo de tiempo considerando el total de datos de las estaciones:

6.2.1. Temperatura ambiente

En las tablas 6, 7 y 8 se observa que las tres estaciones presentan valores que oscilan entre los 15.9 y 36.7 grados centígrados, en general manteniéndose cerca a la media (23,54 °C), dichos valores se mantienen dentro del rango normal para la temperatura de Bucaramanga la cual se

encuentra cerca a los 23 °C para temporada fresca y 27 °C en temporada templada.

En la figura 6, se observa una dispersión de datos similar por encima y por debajo de la mediana (23.2), presentando datos atípicos para temperaturas superiores a 32 grados, sin embargo, corresponden a periodos de tiempo comprendidos entre las (11) once de la mañana y el medio día, a su vez, son casos eventuales que se presentan en Bucaramanga, donde su temperatura en el transcurso del año rara vez baja a menos de 18 °C o sube a más de 30°C (Weather Spark, n.d.).

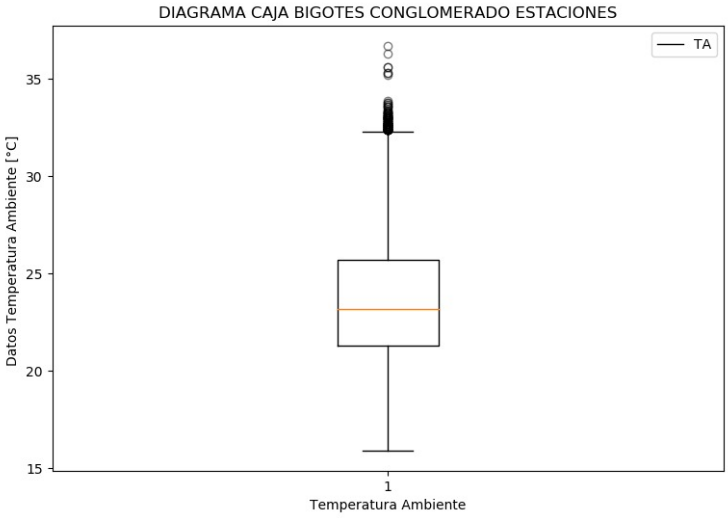

Figura 6. Diagrama de cajas y bigotes para Temperatura Ambiente, estación Pilar Adaptado de Python.

6.2.2. Lluvia. En las figuras 7, 8 y 9 se presenta la relación gráfica de la lluvia a través del tiempo, como se mencionó anteriormente, las condiciones climáticas de Bucaramanga favorecen un entorno de lluvias cero, por ejemplo, para la estación El Pilar, la proporción de datos cero corresponde al 97.19%, por ende, los intervalos o días con precipitaciones fuertes o leves serán considerados datos atípicos aumentando la variabilidad y dispersión de estos.

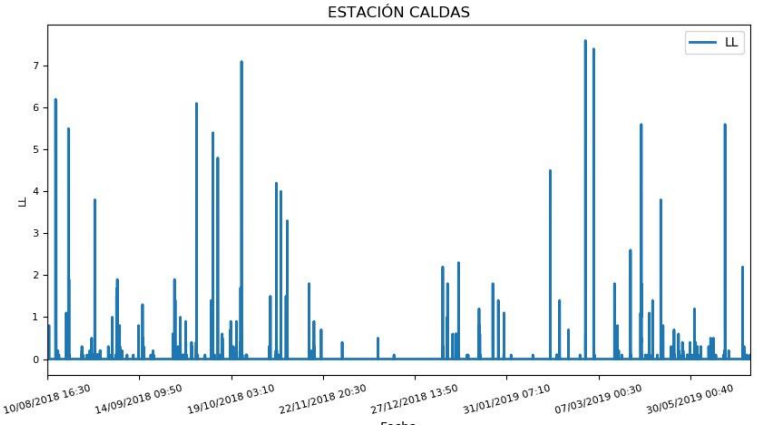

Figura 7. Gráfica Lluvia, estación Caldas. Adaptado de Python.

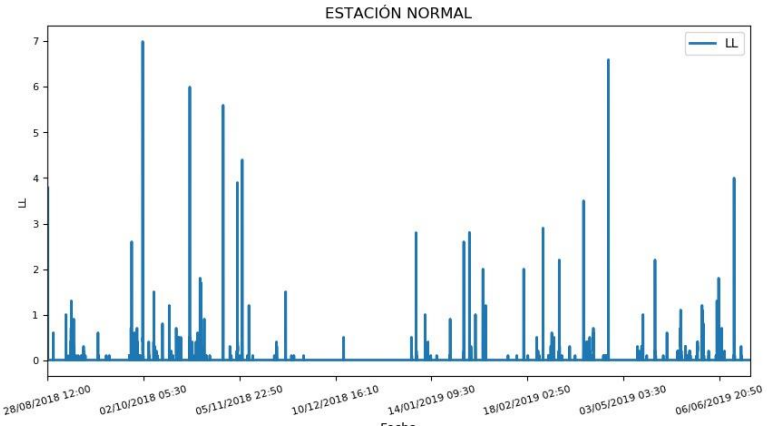

Figura 8. Gráfica Lluvia, estación Normal. Adaptado de Python.

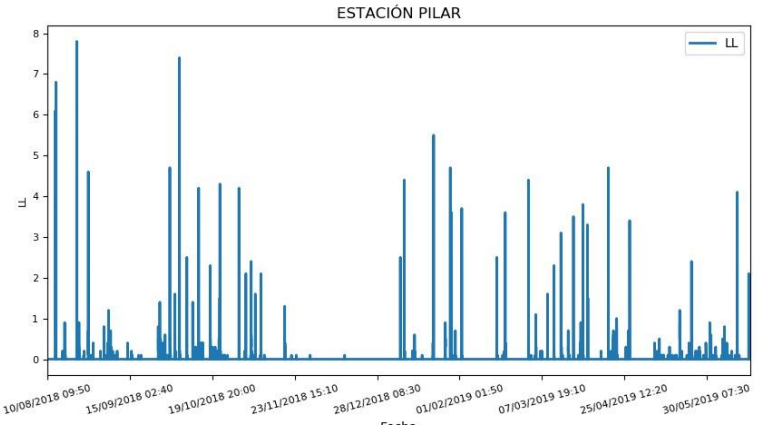

Figura 9. Gráfica Lluvia, estación Pilar. Adaptado de Python.

6.2.3. Humedad relativa. En la figura 10 se presentan los datos de humedad relativa (HR) correspondientes a dos días de observación, se tienen valores desde 25.8% hasta 98.9% donde se evidencia una alta variabilidad en el comportamiento de esta variable para todas las estaciones.

En la figura 11 se observa mayor dispersión para los datos por debajo de la mediana presentando valores atípicos inferiores a 30, pero que permanecen dentro del rango normal para la ciudad de Bucaramanga la cual mantiene un promedio de 70% a 80% para meses como agosto, enero y febrero, se debe considerar que en Bucaramanga la humedad percibida no varía significativamente; el período más húmedo del año dura 9,7 meses, por lo general del 5 de marzo al 27 de diciembre. (Weather Spark, n.d.)

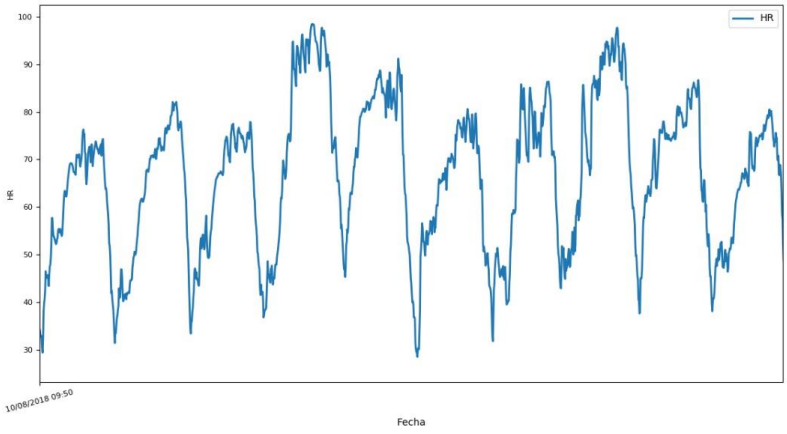

Figura 10. Gráfica Humedad relativa para dos días. Adaptado de Python.

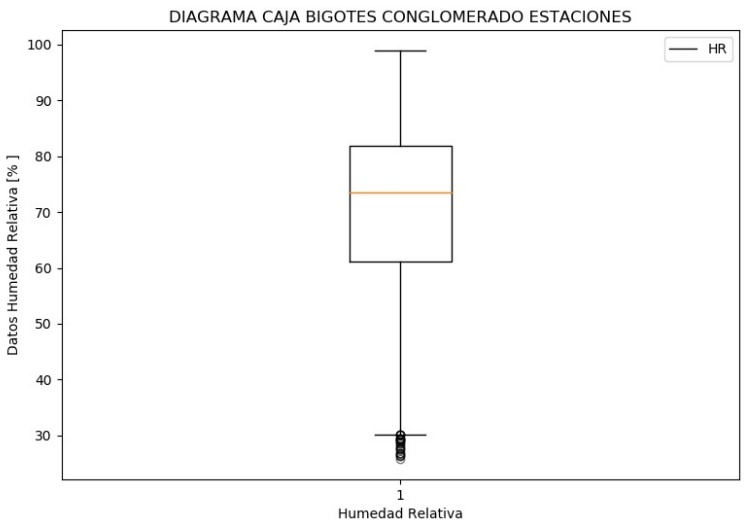

Figura 11. Diagrama cajas y bigotes para Humedad Relativa. Adaptado de Python

6.2.4. Dirección del viento. Una de las variables más complejos de analizar es la de dirección del viento WD dada su alta variabilidad, ya que no presenta un comportamiento claramente cíclico sino variable a lo largo del día como se puede ver en la figura 12, por ende, no se puede establecer franjas horarias en las que haya mayor o menor WD.

En la figura 13 se observa que a pesar de presentar una alta dispersión para los datos por debajo de la mediana (269.2) no se presentan valores atípicos por fuera del rango inter cuartil.

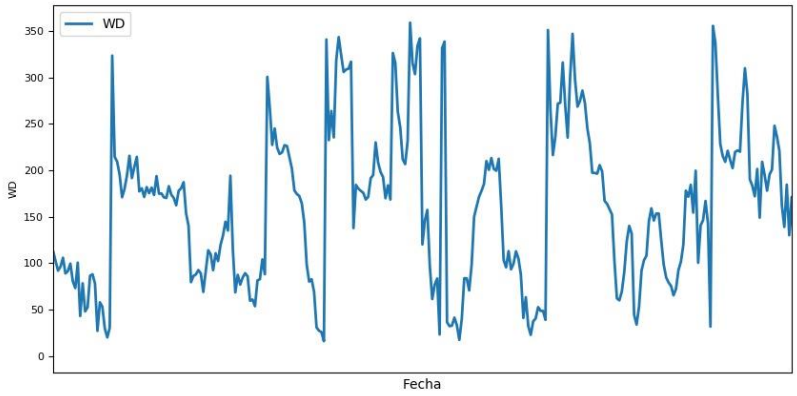

Figura 12. Gráfica Dirección del viento, dos días. Adaptado de Python.

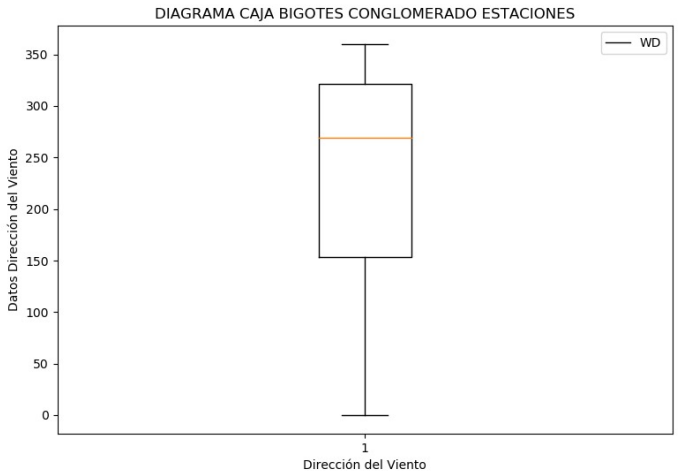

Figura 13. Diagrama cajas y bigotes para Dirección del viento. Adaptado de Python

6.2.5. Radiación Solar. Como se observa en las figuras 14, 15 y 16, la variable de radiación solar (RS) presenta un comportamiento cíclico con picos altos para valores sobre los 800 W/m^2, datos que corresponden a períodos de mayor influencia solar comprendidos entre las diez de la mañana y dos de la tarde, a su vez los valles presentados representan períodos de tiempo nocturnos de seis de la tarde a cinco de la mañana mostrando valores min de RS como comportamiento normal de la variable evidenciado para las tres estaciones.

Teniendo en cuenta el valor de la desviación estándar de las tres estaciones; 286,13 y el promedio ponderado de las mismas; 200,894 se muestra una gran dispersión de los datos superiores a la mediana y existencia de valores atípicos como se observa en la figura 17.

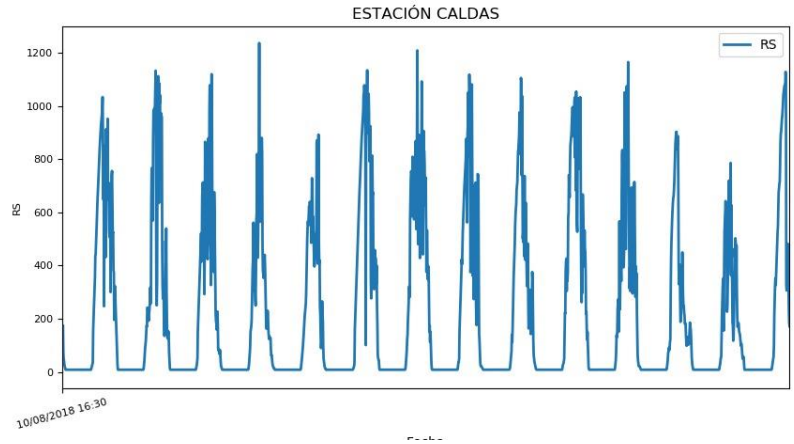

Figura 14. Gráfica Radiación solar, para dos días estación Caldas. Adaptado de Python.

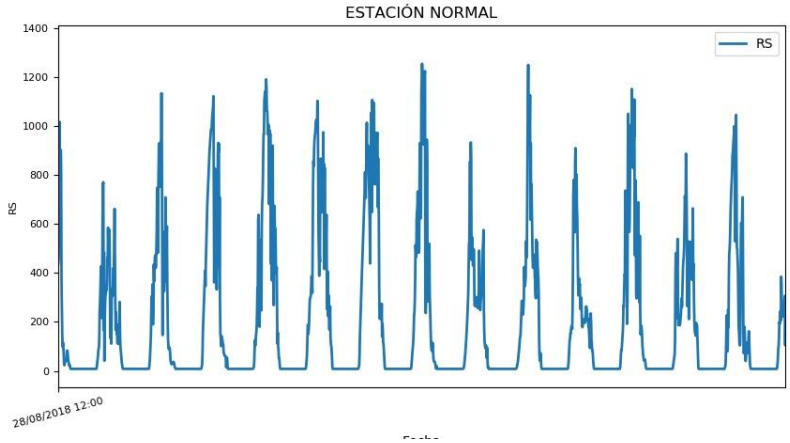

Figura 15. Gráfica Radiación solar, para dos días estación Normal. Adaptado de Python.

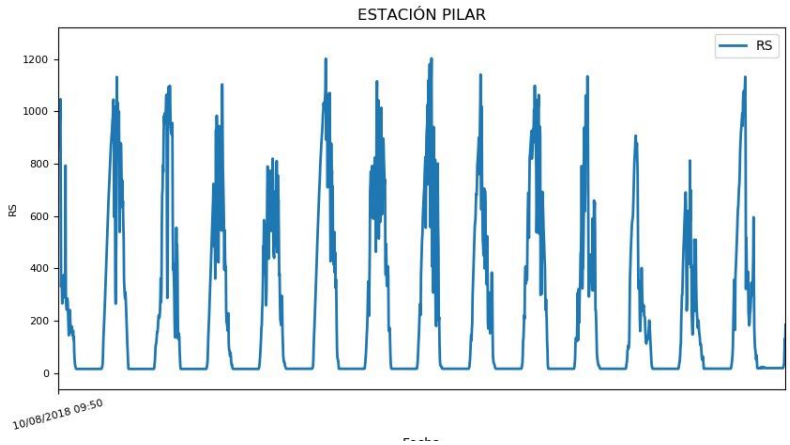

Figura 16. Gráfica Radiación solar, para dos días estación Pilar. Adaptado de Python.

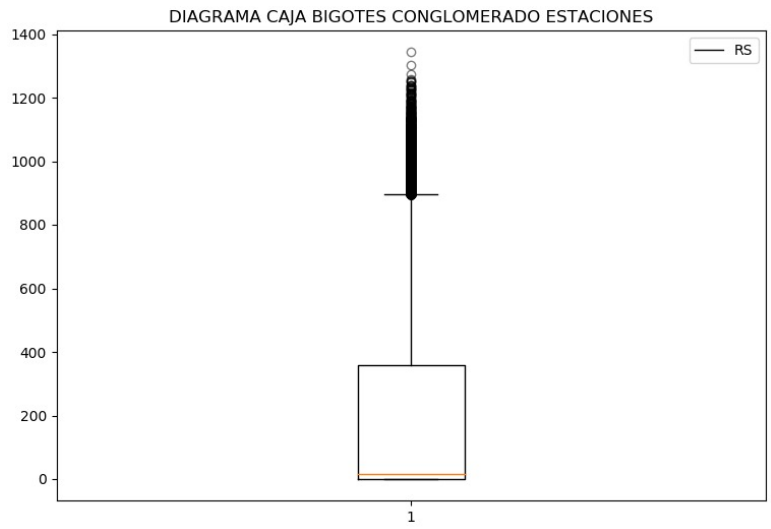

Figura 17. Diagrama cajas y bigotes para Radiación solar. Adaptado de Python.

6.2.6. Material Particulado PM10 y PM2,5. El comportamiento de material particulado PM10 para las tres estaciones, es más o menos constante en el transcurso del día, teniendo en cuenta que el tráfico, el asentamiento poblacional y estructura inmobiliaria es diferente para cada estación, se presenten datos atípicos en diferentes periodos de tiempo como se muestra en las figuras 18 a 25.

La estación Caldas en general presenta valores más altos de material particulado frente a las demás estaciones, a su vez, presenta valores atípicos los cuales son causales de alarmas como se evidencia en los reportes presentados por el AMB durante febrero y marzo del presente año (Área Metropolitana de Bucaramanga - AMB, n.d.).

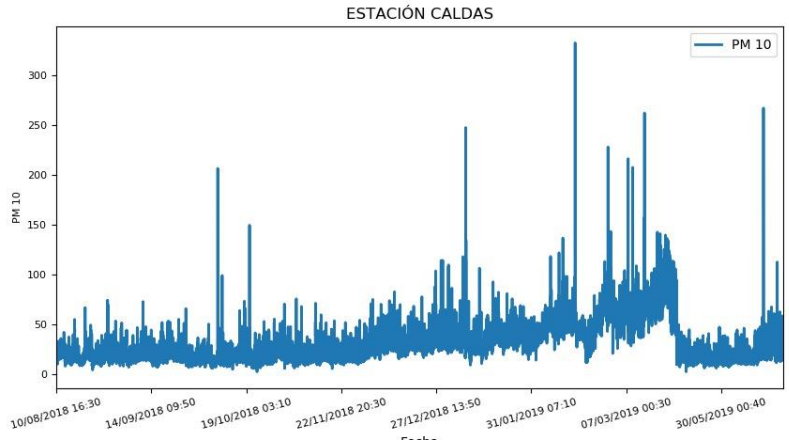

Figura 18. Grafica de PM 10, estación Caldas. Adaptado de Python.

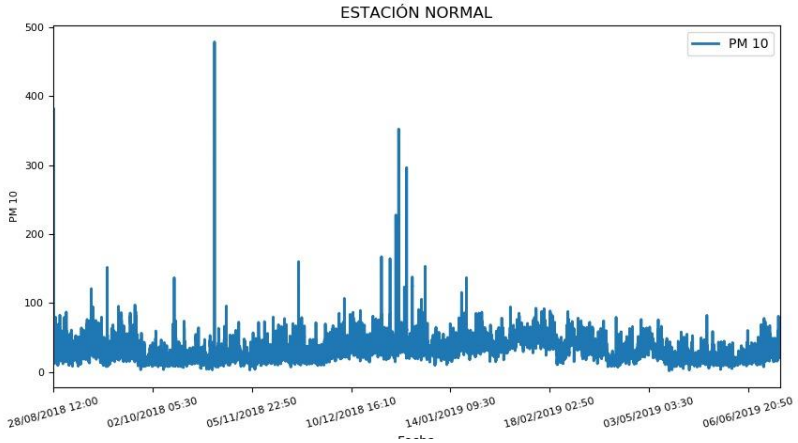

Figura 19. Grafica de PM 10, estación Normal. Adaptado de Python

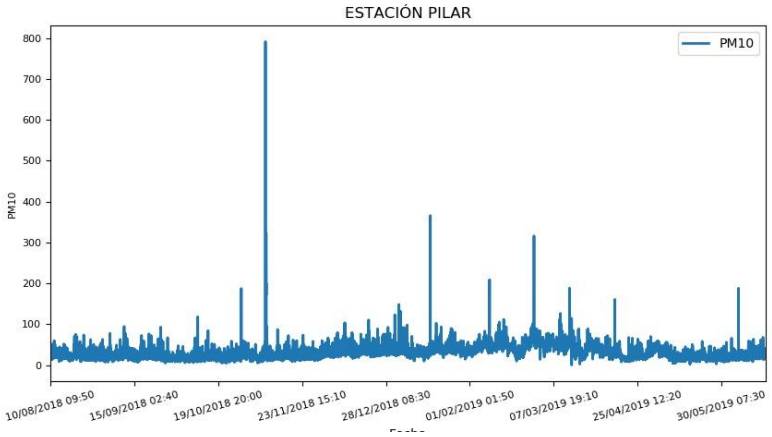

Figura 20. Grafica de PM 10, estación Pilar. Adaptado de Python.

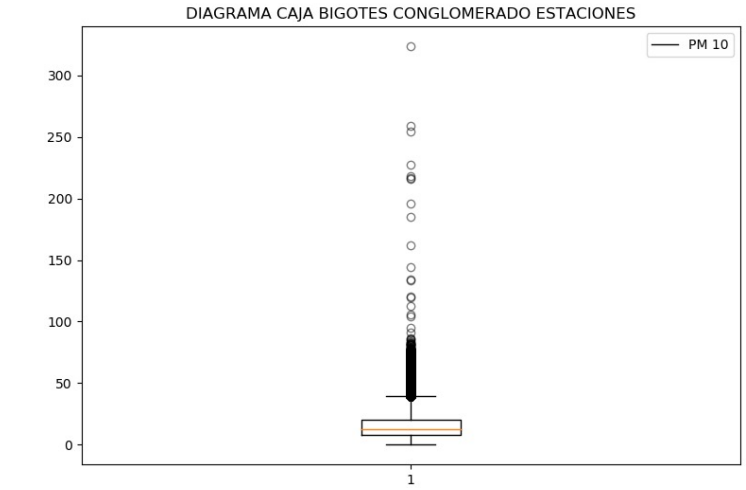

Figura 21. Diagrama caja y bigotes para PM 10. Adaptado de Python.

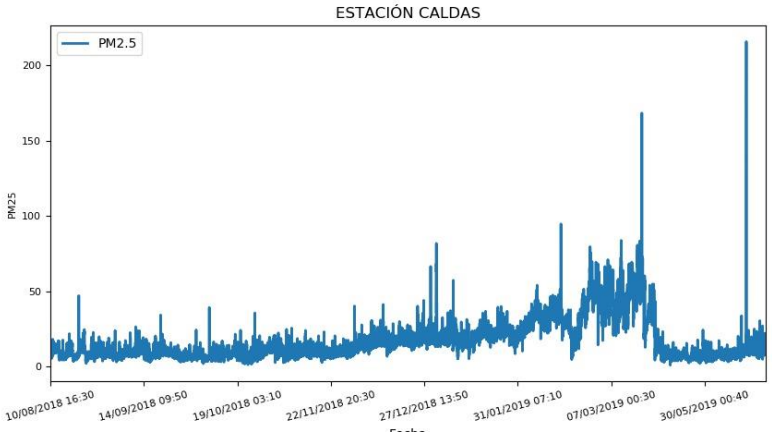

Figura 22. Grafica de PM 2.5, estación Caldas. Adaptado de Python

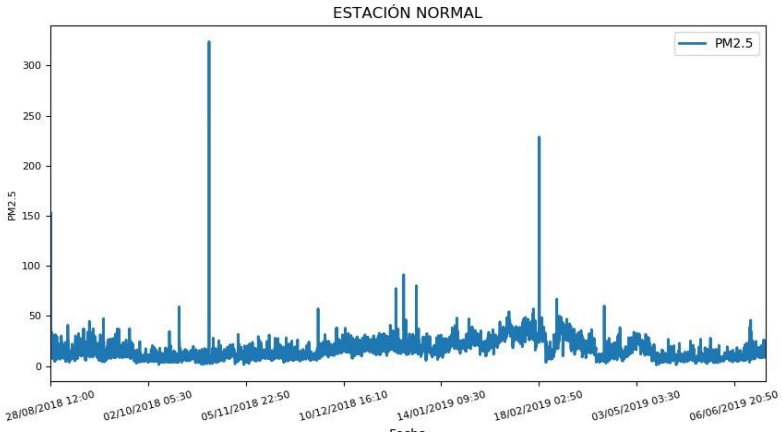

Figura 23. Grafica de PM 2.5, estación Normal. Adaptado de Python.

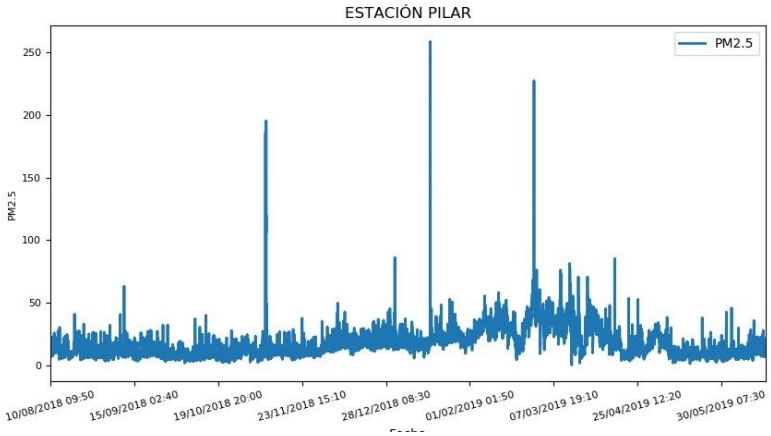

Figura 24. Grafica de PM 2.5, estación Pilar. Adaptado de Python

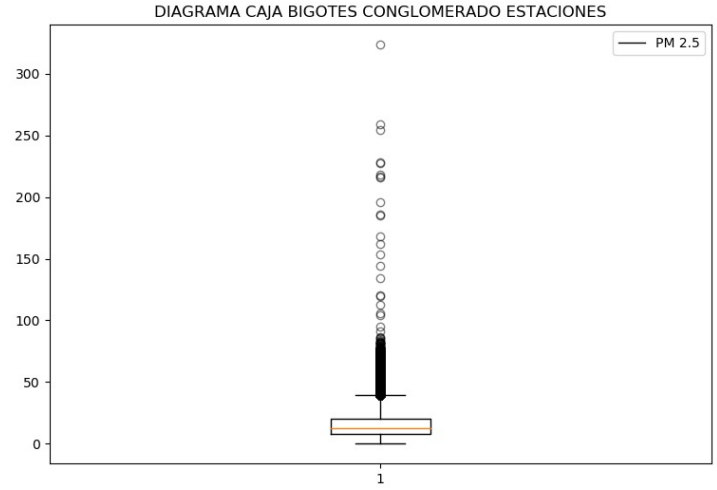

Figura 25. Diagrama caja y bigotes PM 2.5. Adaptado de Python

6.3. Procesamiento de los datos

Después de la limpieza e imputación de datos correspondiente, se aplica análisis de componentes principales - PCA, para solucionar el problema de multicoleanidad de las variables predictoras y proceder a ajustar los modelos de Deep Learning. Para el desarrollo de esta investigación se aplicó para cada estación, un modelo LSTM, un Seq2Seq y un SVM como se muestra a continuación:

6.3.1. Modelo LSTM. A partir de los datos preprocesados se realiza el ajuste de los modelos

de Deep Learning con el apoyo de Python, utilizando LSTM como modelo base, debido a su capacidad de aprender largas series de tiempo, característica especialmente importante para el modelado de contaminantes atmosféricos espaciotemporales (Li et al., 2017), además de su aplicación en estudios relacionados con la predicción de series de tiempo, como la predicción del flujo de tráfico, la predicción de la energía eólica, la predicción de la trayectoria humana, etc.

Se ajusta un modelo LSTM considerando tangente hiperbólica (tanh) como función de activación, al ser una de más conocidas y empleadas por la literatura y se varía su configuración con 15, 20, 50, 75 y 100 capas, cada una con densidad de 1 neurona y se especifican 50 iteraciones, para 12.960 datos de entrenamiento, correspondientes a los tres primeros meses de la base de datos de cada estación; cabe resaltar que se consideran 50 capas ocultas, porque un número menor de capas distorsiona los resultados de la predicción, generando mayor error, y con un número mayor de capas el modelo puede sobre ajustarse, sin embargo la diferencia entre los valores finales de error no son significativas. Se obtiene una gráfica para identificar el comportamiento del valor de la función de costo (val_loss) para los datos de prueba y el valor de la función de costo para los datos de entrenamiento (loss), en las figuras 26 a 28 se muestra un ajuste de los datos a partir de la décima iteración.

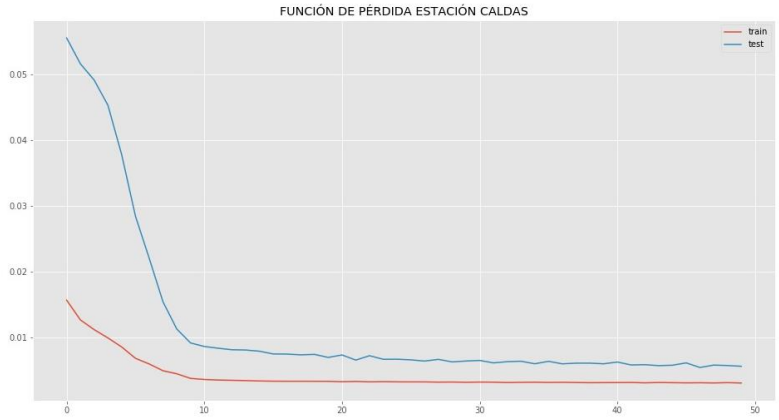

Figura 26. Función de pérdida Estación Caldas, LSTM. Adaptado de Python

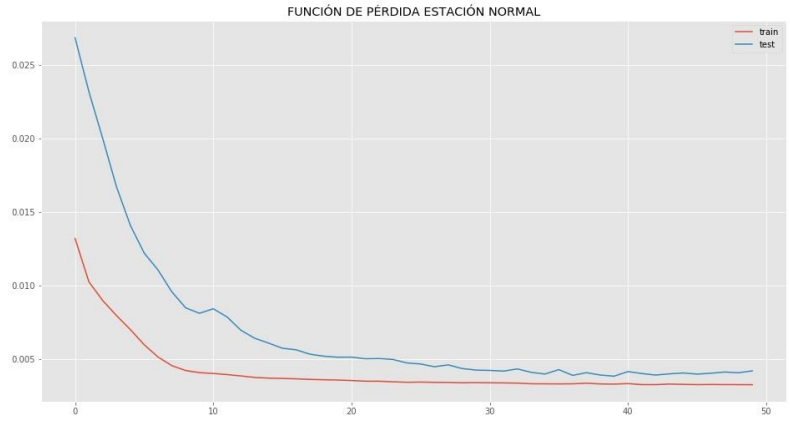

Figura 27. Función de pérdida Estación Normal, LSTM. Adaptado de Python

Figura 28. Función de pérdida Estación Pilar, LSTM. Adaptado de Python

Finalmente, en las figuras 29, 30 y 31 se muestra la comparación entre la predicción del modelo y los datos reales de PM 2,5 para los primeros 500 datos, se evidencia un buen ajuste para el modelo predictivo propuesto, cabe resaltar que para las estaciones Caldas y Normal la predicción presenta un comportamiento inferior a los datos reales, caso contrario de la estación Pilar cuyo comportamiento es levemente superior a los datos reales.

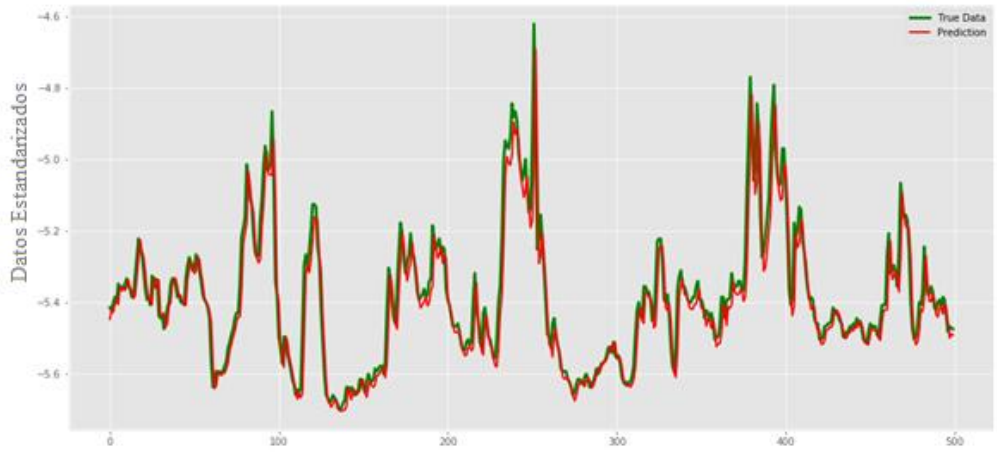

Figura 29. Gráfica de predicción LSTM, estación Caldas. Adaptado de Pyhton.

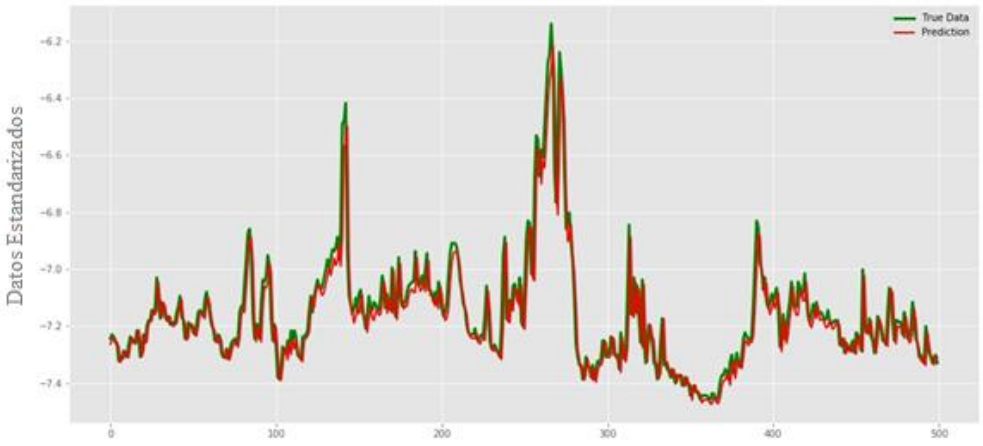

Figura 30. *Gráfica de predicción LSTM, estación Normal. Adaptado de Pyhton.*

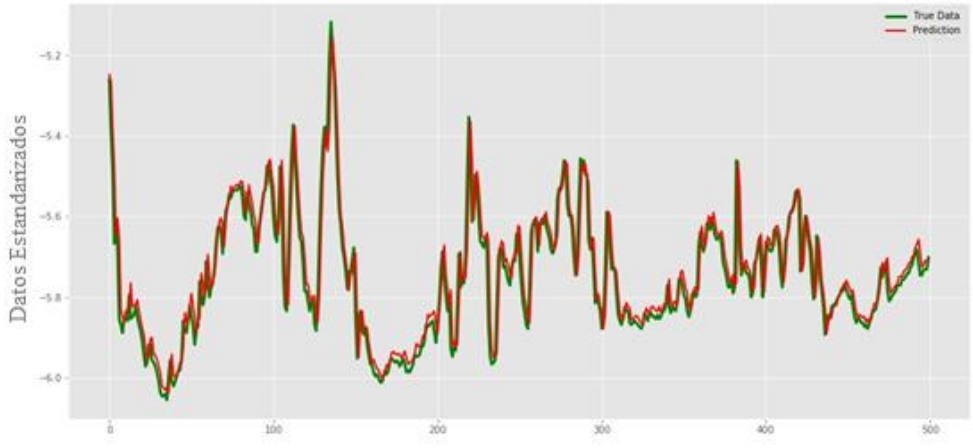

Figura 31. Gráfica de predicción LSTM, estación Pilar. Adaptado de Python.

6.3.2. Modelo seq2seq. Basados en el desarrollo y recomendaciones del trabajo titulado Deep Air: Forecasting Air Pollution in Beijing, China (Reddy, et al , n.d.) se ajusta el modelo seq2seq con 128 capas ocultas, cada una con densidad de 50 neuronas, se especifican 20 iteraciones y 8 submodelos consecutivos de LSTM con 4 funciones de activación diferentes; “relu”, “sigmoid”, “softmax” y “tanh”, para 12.960 datos de entrenamiento, correspondientes a los tres primeros meses de la base de datos de cada estación.

Luego de 8 submodelos LSTM se obtiene una gráfica para identificar el comportamiento del valor de la función de costo y de entrenamiento (loss), en las figuras 32, 33 y 34 se puede observar un ajuste de los datos a partir de la décima iteración.

Para la estación Caldas y la estación Pilar, figura 32 y 34 respectivamente se muestra que la función “softmax” invierte más tiempo para logar el ajuste del modelo, a su vez, para la estación Normal (figura 33) no se muestran comportamiento claro de ajuste y la mayoría de las funciones de activación tardan el ajustar el modelo lo cual se evidencia también en que es la estación que presenta valores de error más grandes (ver tabla 11).

En la figura 32 se observa que la mayoría de las gráficas tienen ajuste cerca a la quinta época de iteración siendo “sigmoid” quien primero se ajusta, sin embargo, la función de activación “relu” tiende a divergir de los datos reales después de la séptima iteración.

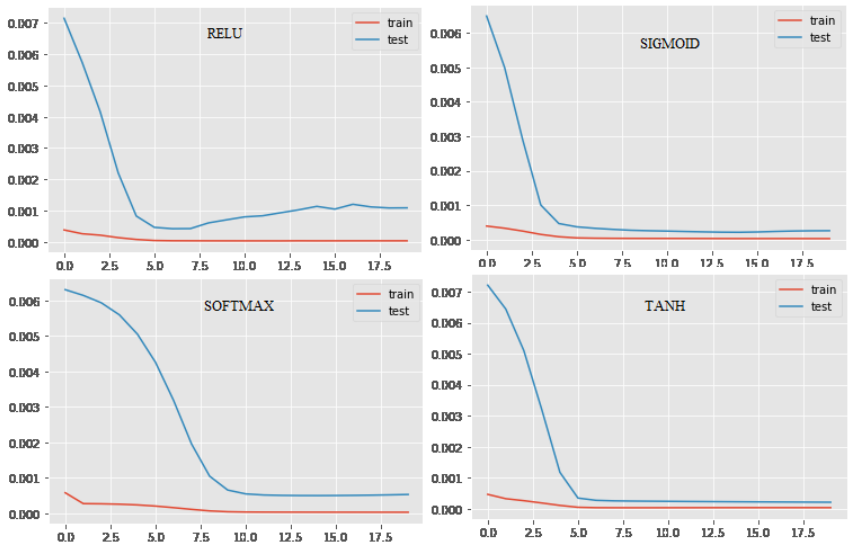

Figura 32. Función de pérdida Estación Caldas, seq2seq. Adaptado de Python.

En la figura 33, correspondiente a las gráficas de las funciones de activación implementadas para la estación Normal, se muestra que no hay un comportamiento de ajuste clara para las cuatro funciones de activación, las funciones “relu” y “softmax” luego de la décima época presentan valores de prueba inferiores a los de entrenamiento con tiempo de entrenamiento similar, 32.84 seg y 32.56 seg respectivamente.

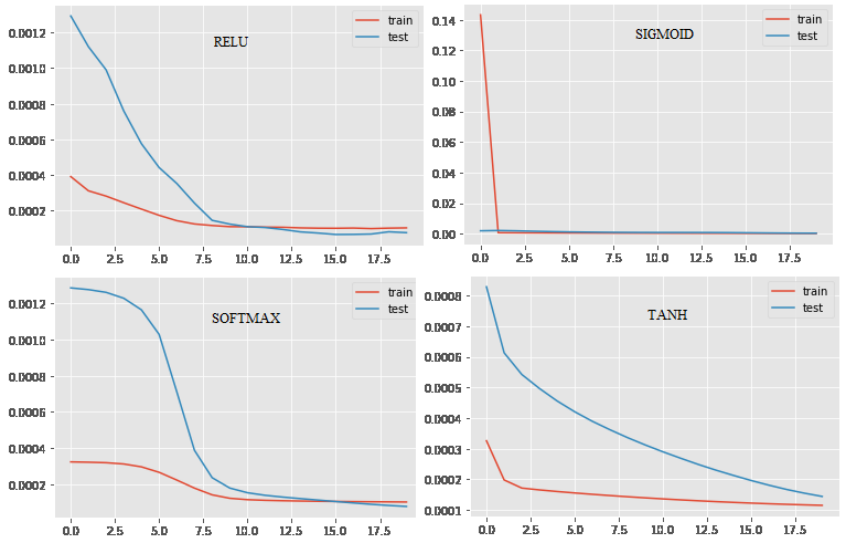

Figura 33. Función de pérdida Estación Normal, seq2seq. Adaptado de Python.

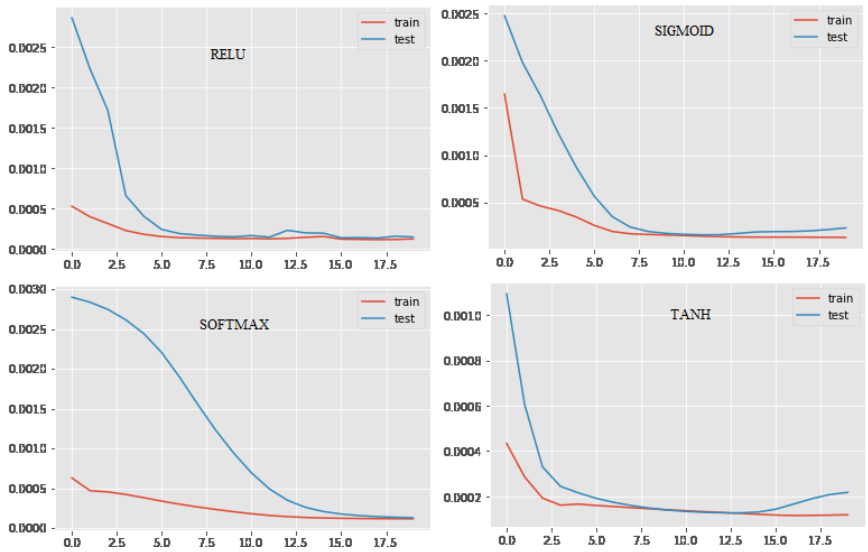

Figura 34. Función de perdida Estación Pilar, seq2seq. Adaptado de Python.

Finalmente, en las figuras 35, 36 y 37 se observa la comparación entre la predicción del modelo y los datos reales de PM 2,5 para los primeros 300 datos evidenciando un buen ajuste entre ellas, contrario al modelo LSTM se observa que los valores de los datos de predicción presentan un comportamiento levemente inferior a los datos reales (figuras 36 y 37), contrario a la estación Caldas cuya tendencia es a presentar valores superiores a los datos reales.

Figura 35. Gráfica predicción Estación Caldas, seq2seq. Adaptado de Python.

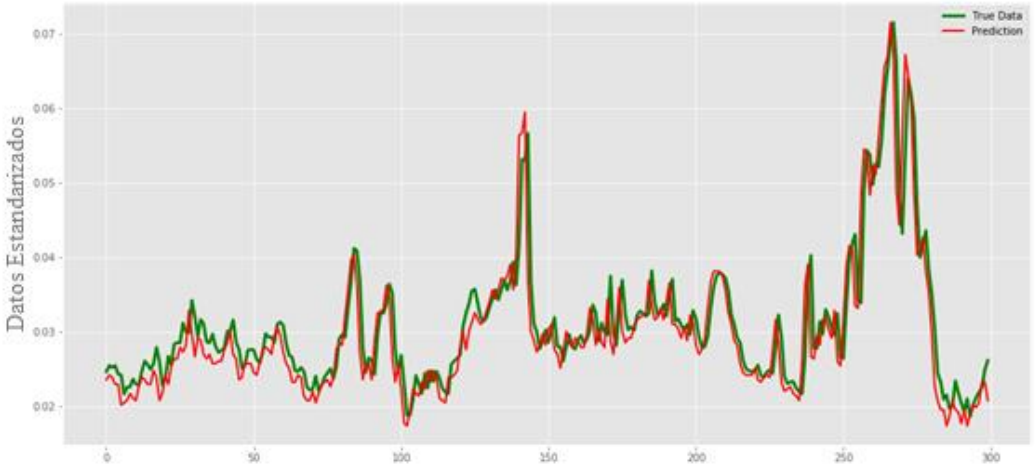

Figura 36. Gráfica predicción Estación Normal, seq2seq. Adaptado de Python.

Figura 37. Gráfica predicción Estación Pilar, seq2seq. Adaptado de Python.

6.3.3. Modelo SVM

Se ajusta un modelo SVM de Kernel lineal con las 6 variables independientes y 12.960 datos de entrenamiento, correspondientes a los tres primeros meses de la base de datos de cada estación, con ello se obtiene una gráfica entre la predicción del modelo y los datos reales de PM 2,5 para los primeros 300 días presentando buen ajuste principalmente al final de la predicción.

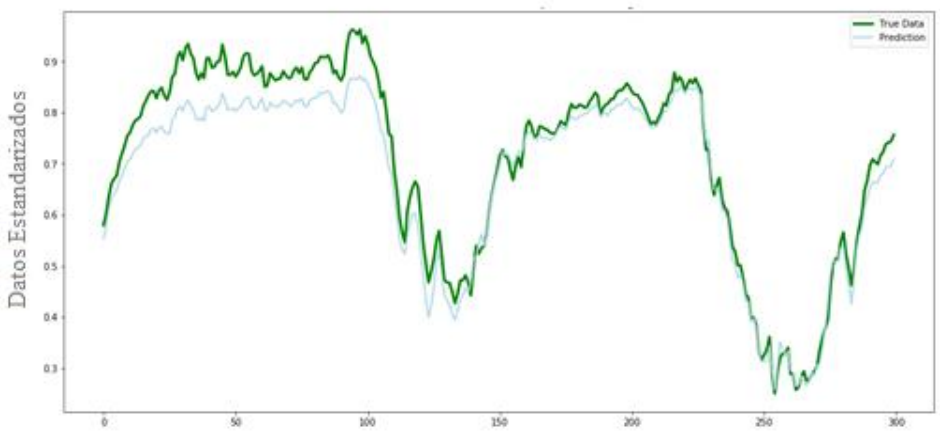

Figura 38. Gráfica predicción Estación Caldas, SVM. Adaptado Python.

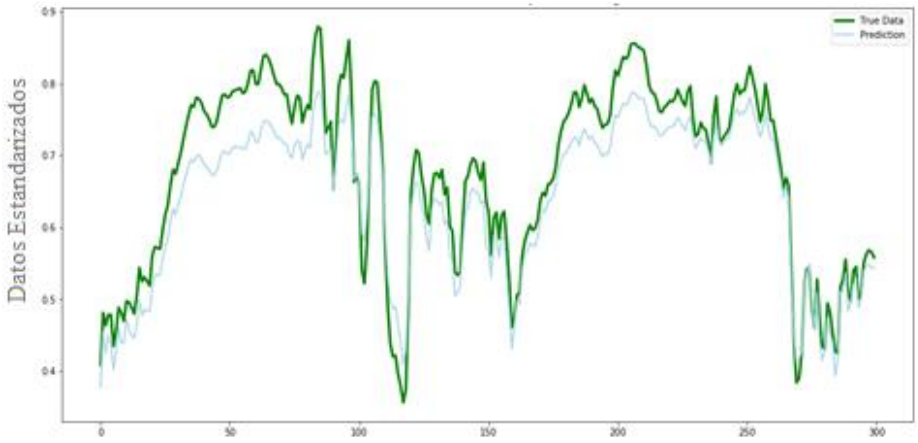

Figura 39. Gráfica predicción Estación Normal, SVM. Adaptado Python.

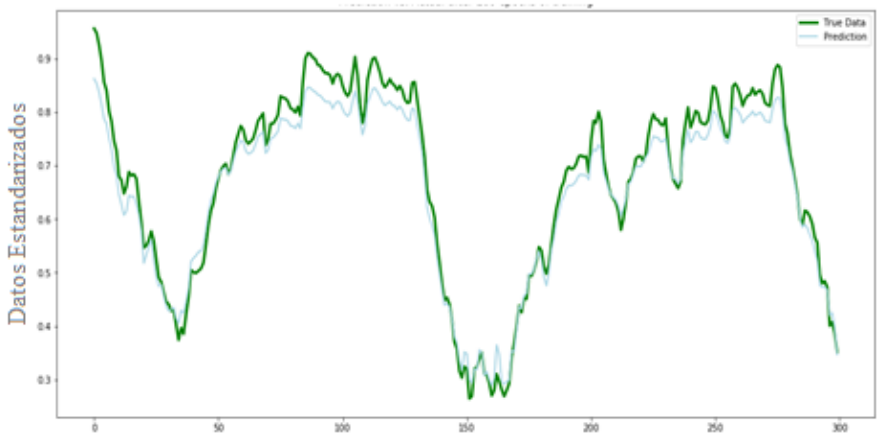

Figura 40. Gráfica predicción Estación Pilar, SVM. Adaptado Python.

6.4. Evaluación e interpretación

Como se estableció en la sección 3 del presente documento, para el proceso de verificación y validación, se utilizan 3 métricas de validación: el error absoluto medio de los resultados del pronóstico (MAE), el error cuadrático medio de las predicciones (MSE) y la raíz cuadrada del promedio de las diferencias cuadradas entre la predicción y la observación real (RMSE) (ver tabla 11 y figura 41), a su vez, se analiza el coeficiente de determinación(R^2), (ver tabla 12).

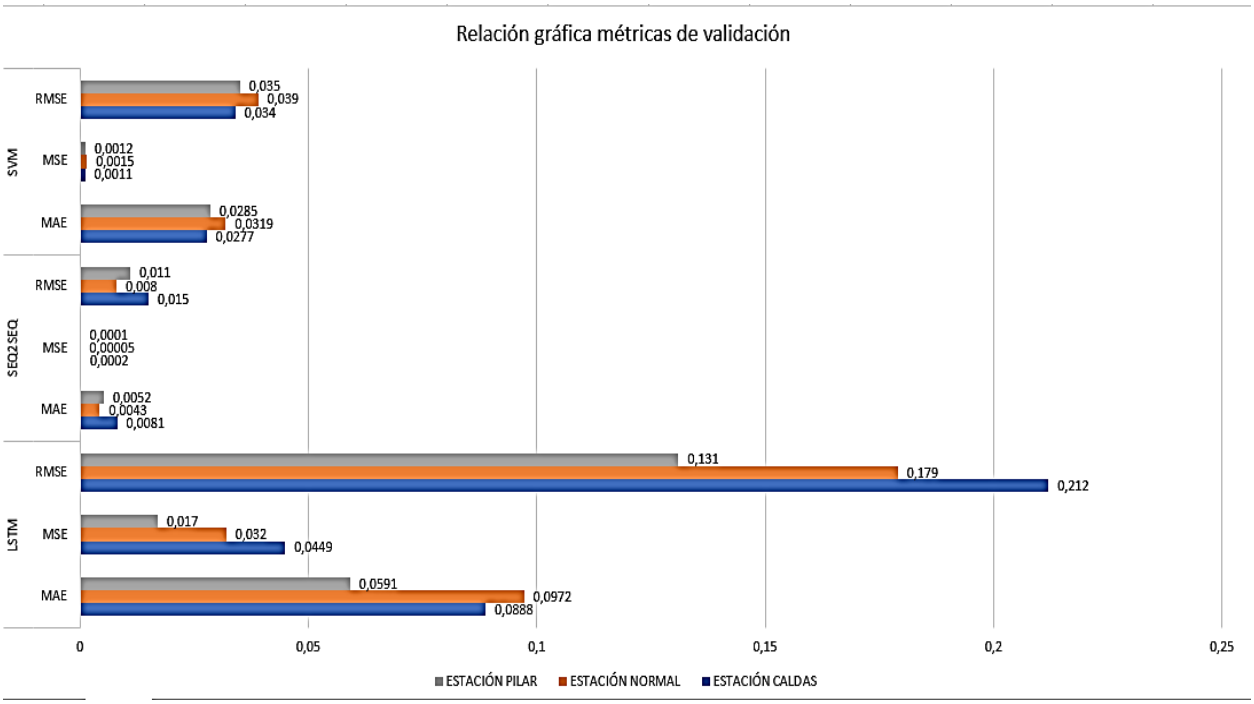

Figura 41. Relación gráfica métricas de validación.

En la tabla 11 se observa los valores de MAE, MSE y RMSE, para los modelos LSTM y el SVM la estación Normal presenta en general valores de error más altos sin embargo en la figura 28 se evidencia un buen ajuste de los datos. En general el modelo de mayor error es el LSTM con los valores más altos de RMSE para las tres estaciones. A su vez, se observa que en promedio el modelo seq2seq es el que presenta un error más bajo en las tres métricas de validación realizadas.

Tabla 11.

Relaciones métricas de validación

	LSTM			**SEQ2SEQ**			**SVM**		
Métrica de validación	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
Estación caldas	0,0888	0,0449	0,212	**0,0081**	**0,0002**	**0,015**	0,0277	0,0011	0,034
Estación normal	0,0972	0,032	0,179	**0,0043**	**0,00005**	**0,008**	0,0319	0,0015	0,039
Estación pilar	0,0591	0,017	0,131	**0,0052**	**0,0001**	**0,011**	0,0285	0,0012	0,035
Promedio	0,0817	0,0313	0,174	***0,0058***	***0,0001***	***0,0113***	0,0293	0,0012	0,036

En general la predicción de los tres modelos presenta un buen ajuste de acuerdo con los valores del coeficiente de correlación de la tabla 12, si bien con el modelo seq2seq no se obtiene el R^2 más alto, complementa los resultados obtenidos de la tabla 11 se deduce que el modelo seq2seq presenta en promedio un ajuste del 93.56% y es el modelo que presenta el menor error.

Tabla 12.

Coeficiente de determinación R^2

Coeficiente de determinación R2			
Modelo	**LSTM**	**SEQ2SEQ**	**SVM**
Estación caldas	**0,9588**	**0,946**	**0,9668**
Estación normal	0,9258	0,921	0,955
Estación pilar	0,9447	0,94	0,9633
Promedio	0,9431	0,93566667	0,9617

La combinación de una alta densidad poblacional, un alto número de espacios comerciales, con gran número de vehículos privados y vehículos de carga crean una situación compleja y difícil de administrar en todos los ámbitos, sin embargo, se considera que el transporte de carga es el principal contribuyente de las emisiones de material particulado y por ende del deterioro de la calidad del aire, por ello, una de las mayores iniciativas que han tenido éxito en algunas ciudades

europeas es la Distribución Nocturna de Mercancías (DNM) la cual tiene como objetivo mitigar los impactos causados por el movimiento de vehículos de carga en los centros urbanos, es decir consiste en crear restricciones legales para el movimiento de vehículos con carga urbana que hacen las entregas de bienes en áreas centrales de la ciudad, siendo permitido su movimiento solo durante la noche cuando la ciudad muestra gran fluidez en el tráfico (Zonalogistica, n.d.)

Teniendo en cuenta lo anterior, en Bogotá, uno de los principales centros urbanísticos nacionales, el 46 por ciento de las empresas de la Red de Logística Urbana han implementado exitosamente pilotos de buenas prácticas logísticas entre los que se destacan la carga colaborativa, cargue y descargue nocturno, seguimiento tecnológico de la operación, utilización de zonas de cargue y descargue y flota vehicular de bajas emisiones (Medina, n.d.)

Entonces, se ejecuta el algoritmo seq2seq para la estación Caldas separando los datos entre el día y la noche, teniendo en cuenta el comportamiento de variables como el RS y el WD, se busca comparar las variables y la relación entre ellas.

Se observa un cambio significativo en las correlaciones entre variables para las dos jornadas, como se muestra en la figura 42, el comportamiento de la RS se manifiesta en un cambio en el sentido de la relación frente a algunas variables, por ejemplo, con TA pasa de una relación directa de 0.78 en el día a una relación inversa de -0.23 en la noche, en el caso de HR la reacción es contraria.

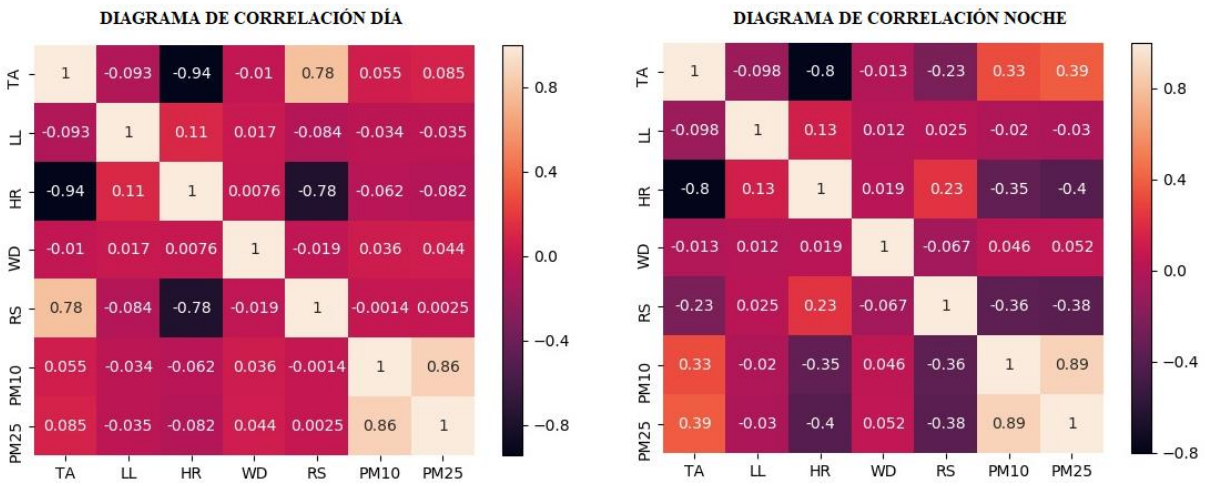

Figura 42. Diagrama de correlación Dia - Noche

Respecto a la variable respuesta (PM2,5), la TA y la HR aumentan su relación directa e indirecta respectivamente; se destaca el cambio de relación directa durante el día a indirecta durante la noche entre el RS y el PM2.5, es decir en la noche el RS tiene una relación más fuerte con el material particulado, pero no se puede afirmar que es la principal causa de presencia de material particulado.

Finalmente, en las figuras 43 y 44 se puede observar la relación entre la predicción y los datos reales de los primeros 300 datos para el día y la noche respectivamente, el modelo predictivo de la noche no solo presenta un mejor ajuste como se observa en la figura 44, a su vez, es el de menor error y mayor R2.

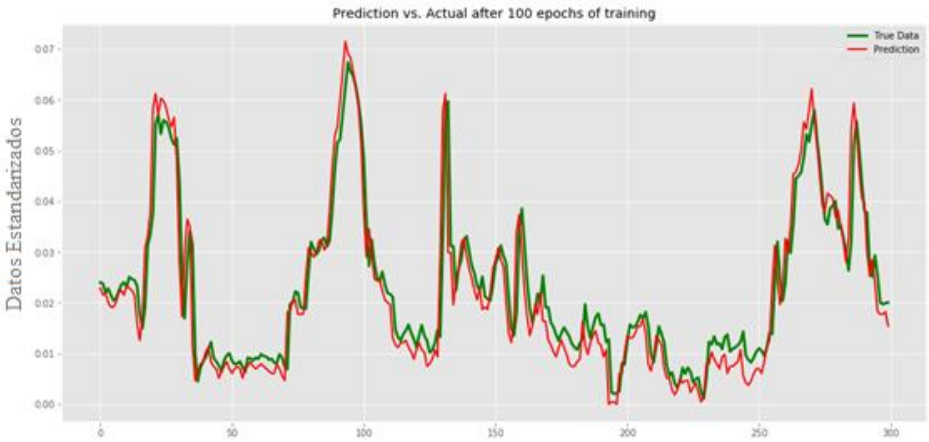

Figura 43. Predicción modelo seq2seq día.

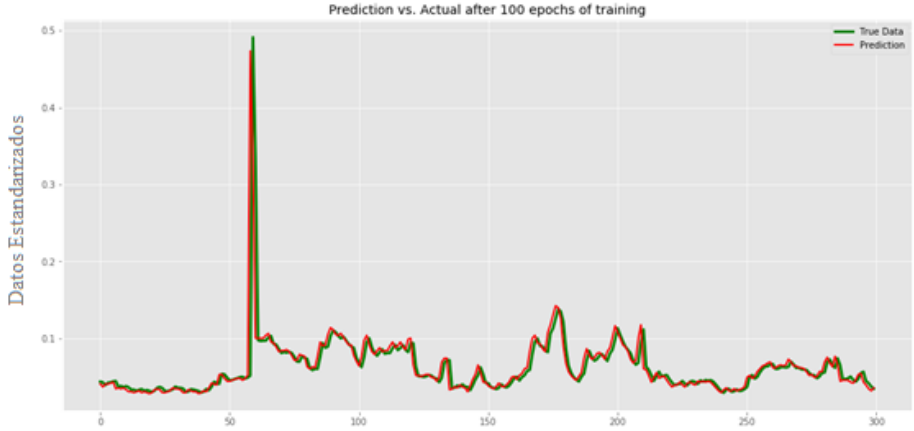

Figura 44. Predicción modelo seq2seq noche.

Tabla 13.

Métricas de validación día y noche.

	SEQ2SEQ			
Métrica de validación	**MAE**	**MSE**	**RMSE**	**R2**
Dia	0,0137	0,0005	0,023	0,88
Noche	0,0118	0,0004	0,021	0,983

6.5. Difusión de conocimiento

Para la difusión de los resultados del proyecto se elabora un artículo (Apéndice B) con los aspectos

relevantes del proyecto como la revisión de literatura, el planteamiento y justificación del problema, así como el análisis de los resultados obtenidos a partir del desarrollo de la investigación.

7. Conclusiones

El método Deep Learning se está acogiendo paulatinamente como una técnica prometedora para pronosticar series de tiempo no lineales, lo cual se evidencia en los diferentes modelos utilizados en los estudios científicos que soportan esta investigación, en dónde se integra información meteorológica y relaciona con la contaminación del aire, y que por la naturaleza misma de estos datos se presentan retos de limpieza y preprocesamiento de la información.

Los datos históricos desde agosto de 2018 a junio de 2019 de tres estaciones de monitoreo ubicadas en diferentes puntos de la ciudad y administradas por la subdirección ambiental del Área Metropolitana de Bucaramanga – AMB, para apoyar el objeto de estudio del proyecto presentó retos de un análisis de datos perdidos, datos faltantes, uso de técnicas de imputación y el entendimiento de las propias variables que podrían conformar la base de datos objeto de estudio, que consideraba el contaminante PM2,5 como variable a predecir.

La calidad del aire es una serie temporal que no depende del valor inmediatamente anterior, por ello se utilizaron dos modelos LSTM y un modelo SVM para aprender tanto a corto como a largo plazo, así que después de comparar el desempeño de estos modelos por medio de las métricas tales como: MAE, MSE, RMSE y R2, se descubre que el modelo de predicción seq2seq tiene la capacidad de predecir material particulado 2.5 con mejor desempeño que los otros modelos.

Si bien las relaciones de correlación de las variables cambian con el paso del tiempo del día a la noche presentándose valores de contaminación inferiores de PM2,5 en el periodo nocturno, no se descarta que factores como el tráfico o factores ambientales no considerados repercutan en la concentración de material particulado, sin embargo, puede ser una de las razones de la disminución de la variabilidad de los datos, ya que en el día, por el fenómeno de horas pico en cuanto al fujo vehicular se presentan valores altos en las concentraciones de material particulado.

8. Recomendaciones

Durante la selección y limpieza de datos de debe tener especial cuidado al eliminar y clasificar las variables ya que si se desconoce la naturaleza de estas se pueden descartar valores de carácter atípico para el análisis estadístico pero que son inherentes al comportamiento de la variable.

Para obtener un pronóstico más preciso, se recomienda indagar sobre otras variables diferentes a las meteorológicas y que tengan influencia en el cálculo y predicción de la calidad del aire; como la relación de la distancia entre estaciones y la distribución urbanística.

Se debe considerar un tamaño mayor para la base de datos implementada ya que esto ayuda a optimizar la predicción y devuelve valores RMSE más bajos para pronósticos futuros más largos, la idea es aprovechar tantos datos de series de tiempo como sea posible, que originen pesos más fuertes en las redes neuronales.

Referencias bibliográficas

Alvarez M., J. de J., & Eslava S., A. (2016). La logística urbana, la ciudad logística y el ordenamiento territorial logístico. *Revista RETO*, *4*(4), 21–40.

Amaya, M. Gómez, Nancy. Rey, I. (2009). *Recopilación y análisis de la calidad del aire del Área Metropolitana de Bucaramanga*.

AMB. (2019). Reporte Mensual Índice de Calidad del aire -ICA. In *22/3/2019*. Retrieved from https://www.amb.gov.co/recurso-aire/

Arciniégas Suárez, C. A. (2012). *DIAGNÓSTICO Y CONTROL DE MATERIAL PARTICULADO: PARTÍCULAS SUSPENDIDAS TOTALES Y FRACCIÓN RESPIRABLE PM 10*. 19. Retrieved from http://www.scielo.org.co/pdf/luaz/n34/n34a12.pdf

Área Metropolitana de Bucaramanga - AMB. (n.d.). Reportes Calidad del Aire. Retrieved July 2, 2019, from https://www.amb.gov.co/calidad-del-aire/

Athira, V., Geetha, P., Vinayakumar, R., & Soman, K. P. (2018). DeepAirNet: Applying Recurrent Networks for Air Quality Prediction. *Procedia Computer Science*, *132*, 1394–1403. https://doi.org/10.1016/j.procs.2018.05.068

Bai, Y., Zeng, B., Li, C., & Zhang, J. (2019). An ensemble long short-term memory neural network for hourly PM 2.5 concentration forecasting. *Chemosphere*, *222*, 286–294. https://doi.org/10.1016/j.chemosphere.2019.01.121

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 1–19. https://doi.org/10.1186/s12889-017-4914-3

Buendía Martínez, J. M. (n.d.). *Logística Sostenible: Estudio de la Calidad del Aire e interacción*

sobre la Movilidad Urbana (Universidad Politécnica de Cartagena.). Retrieved from http://repositorio.upct.es/bitstream/handle/10317/7358/tfm-bue-log.pdf?sequence=1&isAllowed=y

Carmona Suárez, E. J. (2013). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. Retrieved from http://www.ia.uned.es/~ejcarmona/publicaciones/%5B2013-Carmona%5D SVM.pdf

Dirección de Tránsito de Bucaramanga. (2019). Parque Automotor Área Metropolitana 2018. Retrieved March 18, 2019, from https://www.transitobucaramanga.gov.co/files/2019/estadisticas/parque-automotor-area-metropolitana-corte-31-diciembre-2018.png

Fundación para la Salud Geoambiental. (n.d.). Material particulado. Retrieved April 12, 2019, from https://www.saludgeoambiental.org/material-particulado

Gobierno de Mexico. (n.d.). Variables Meteorológicas. Retrieved August 4, 2019, from https://smn.conagua.gob.mx/es/variables-meteorologicas#targetText=El valor obtenido es el,de la dirección del viento.&targetText=La humedad relativa es el,unidad de medición es en %25.

Goodfellow, I., Bengio, Y., & Courville, A. (n.d.). Deep Learning. Retrieved March 2, 2019, from http://www.deeplearningbook.org/

Grupo Tecma Red S.L. (n.d.). Ciudades Inteligentes • ESMARTCITY. Retrieved May 23, 2019, from https://www.esmartcity.es/ciudades-inteligentes

IDEAM. (n.d.-a). Atlas Interactivo. Retrieved from http://atlas.ideam.gov.co/presentacion/

IDEAM. (n.d.-b). CALIDAD DEL AIRE - IDEAM. Retrieved May 13, 2019, from http://www.ideam.gov.co/web/contaminacion-y-calidad-ambiental/calidad-del-aire

IDEAM. (n.d.-c). RADIACIÓN SOLAR. Retrieved October 8, 2019, from

http://www.ideam.gov.co/web/tiempo-y-clima/radiacion-solar#targetText=La radiación solar es la,procesos atmosféricos y el clima.

IDEAM. (2012). *Hoja metodológica del indicador Índice de calidad del aire*. (p. 8). p. 8. Retrieved from http://www.epa.gov/airnow/aqikids/spanish/pdffiles/spanish_aqirefer.pdf

IDEAM - Instituto de Hidrología Meteorología y Asuntos Ambientales. (2017). *Informe del Estado de la Calidad del Aire en Colombia 2016.* Retrieved from http://www.ideam.gov.co/documents/51310/68521396/3.+Informe+del+Estado+de+la+Calidad+del+Aire+en+Colombia+2016.pdf/fb3eee92-6bcf-4979-9ea2-de0101496a2f?version=1.0

Jóvenes frente al cambio climático. (n.d.). Temperatura ambiente. Retrieved October 8, 2019, from http://www.jovenesfrentealcambioclimatico.com/glosario/

Kök, I., Şimşek, M. U., & Özdemir, S. (2018). A deep learning model for air quality prediction in smart cities. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 1983–1990. https://doi.org/10.1109/BigData.2017.8258144

Kostadinov, S. (n.d.). Understanding Encoder-Decoder Sequence to Sequence Model. Retrieved August 1, 2019, from 4/02/- website: https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

Leite, D. (n.d.). Humedad relativa. Retrieved October 8, 2019, from https://www.meteorologiaenred.com/humedad-relativa.html

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, *23*(22), 22408–22417. https://doi.org/10.1007/s11356-016-7812-9

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory

neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Liu, H., Wu, H., Lv, X., Ren, Z., Liu, M., Li, Y., & Shi, H. (2019). An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China. *Sustainable Cities and Society*, *47*(February), 101471. https://doi.org/10.1016/j.scs.2019.101471

Lopez Briega, R. E. (2016). Redes neuronales convolucionales con TensorFlow. Retrieved April 12, 2019, from https://relopezbriega.github.io/blog/2016/08/02/redes-neuronales-convolucionales-con-tensorflow/

Lu, H., Song, J., Di, T., Kurdestany, J. M., & Wang, H. (2018). A Deep Belief Network Based Model for Urban Haze Prediction. *Tehnicki Vjesnik - Technical Gazette*, *25*(2), 519–527. https://doi.org/10.17559/tv-20180204162632

Lugo-Reyes, S. O., Maldonado-Colín, G., & Murata, C. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. *Revista Alergia Mexico*, *61*(2), 110–120.

Medina, E. (n.d.). 4 iniciativas para mejorar la calidad del aire desde la Secretaría de Movilidad. Retrieved October 9, 2019, from https://bogota.gov.co/mi-ciudad/ambiente/iniciativas-ambientales-de-movilidad-en-la-alcaldia-penalosa?fbclid=IwAR1vQwC-Qa31C_LQopA0QlRVRPrkreBm98yN18VcvszKK_WC3g1t0LLdcXM

Ministerio de Ambiente y Desarrollo Sostenible. *Resolución 2254*. , Pub. L. No. 2254, 11 (2017).

Ministerio de Ambiente y Desarrollo Sostenible. (2017b). Resoluciones | Ministerio de Ambiente y Desarrollo Sostenible. Retrieved February 24, 2019, from http://www.minambiente.gov.co/index.php/normativa/resoluciones

Morales, German. Mora, Juan. Vargas, H. (2008). *Estrategia de regresión basada en el método de*

los k vecinos más cercanos para la estimación de la distancia de falla en sistemas radiales. Retrieved from http://www.scielo.org.co/pdf/rfiua/n45/n45a09.pdf

Observatorio Ambiental de Cartagena de Indias. (n.d.). ÍNDICE DE LA CALIDAD DEL AIRE. Retrieved July 28, 2019, from http://observatorio.epacartagena.gov.co/gestion-ambiental/calidad-ambiental/sistema-urbano/indice-de-la-calidad-del-aire/#targetText=El Índice de Calidad del,efecto en la salud humana.

Observatorio regional de logística. (n.d.). Logística urbana. Retrieved March 27, 2019, from http://logisticsportal.iadb.org/node/2020

Ong, B. T., Sugiura, K., & Zettsu, K. (2016). Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. *Neural Computing and Applications*, *27*(6), 1553–1566. https://doi.org/10.1007/s00521-015-1955-3

Organización Mundial de la Salud - OMS. (2018). Calidad del aire y salud. Retrieved March 18, 2019, from 2 de mayo de 2018 website: https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Palmer Pol, A., & Montaño Moreno, J. (1999). ¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones. *Adicciones*, *11*, 243–255. Retrieved from http://disi.unal.edu.co/~lctorress/RedNeu/LiRna001.pdf

Precipitación. (n.d.). Retrieved October 8, 2019, from https://www.significados.com/precipitacion/

Qi, Z., Wang, T., Song, G., Hu, W., Li, X., & Zhang, Z. (2018). Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality. *IEEE Transactions on Knowledge and Data Engineering*, *30*(12), 2285–2297. https://doi.org/10.1109/TKDE.2018.2823740

Raffino, M. E. (n.d.). Contaminación del Aire. Retrieved August 7, 2019, from https://concepto.de/contaminacion-del-aire/

Reddy, V., Yedavalli, P., Mohanty, S., & Nakhat, U. (n.d.). *Deep Air : Forecasting Air Pollution in Beijing , China*.

Rojas, N., & Galvis, B. (2005). Relación entre PM2.5 y PM10 en la ciudad de Bogotá. *Scielo - Revista de Ingeniería*, (22), 54–60. Retrieved from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-49932005000200006

Singh, G., Al'Aref, S. J., Van Assen, M., Kim, T. S., van Rosendael, A., Kolli, K. K., … Min, J. K. (2018). Machine learning in cardiac CT: Basic concepts and contemporary data. *Journal of Cardiovascular Computed Tomography*, *12*(3), 192–201. https://doi.org/10.1016/j.jcct.2018.04.010

Soh, P. W., Chang, J. W., & Huang, J. W. (2018). Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations. *IEEE Access*, *6*, 38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820

Tomassetti de Piacentini, L. Z. (n.d.). *IMPACTO AMBIENTAL DEL TRANSPORTE URBANO EN EL GRAN MENDOZA*. Cuyo, Ecuador.

Torres, J. (2018). *Deep Learning. Introducción práctica con Keras* (Primer edición). Retrieved from https://torres.ai/deep-learning-inteligencia-artificial-keras/

Weather Spark. (n.d.). Clima promedio en Bucaramanga. Retrieved September 2, 2019, from https://es.weatherspark.com/y/24381/Clima-promedio-en-Bucaramanga-Colombia-durante-todo-el-año

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of the Total*

Environment, *654*, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086

Xu, Y., Ho, H. C., Wong, M. S., Deng, C., Shi, Y., Chan, T. C., & Knudby, A. (2018). Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environmental Pollution*, *242*, 1417–1426. https://doi.org/10.1016/j.envpol.2018.08.029

Xunta de Galicia. (2014). Cálculo del índice de la calidad del aire (ica). *Calidad Do Aire de Galicia*, *2*, 17. Retrieved from http://www.meteogalicia.gal/datosred/infoweb/caire/informes/MANUALES/ES/IT_31_CALCULO_DO_ICA.pdf

Zonalogistica. (n.d.). Logística urbana de la distribución nocturna de mercancías. Retrieved from https://zonalogistica.com/logistica-urbana-de-la-distribucion-nocturna-de-mercancias/?fbclid=IwAR3TXVhkB6CPoZdGkqNsFHAkEEXnYkgkublBGJLg2wqc3M7qQ6CAwZz9F9s