MODELOS DE APRENDIZAJE PROFUNDO PARA LA IDENTIFICACIÓN DEL IMPACTO DEL COVID 19 EN LA CALIDAD DEL AIRE EN COLOMBIA

SILVIA JULIANA FLÓREZ GUERRERO[1], KAREN LISBETH NÚÑEZ SILVA[2], HENRY LAMOS DÍAZ[3]

[1] Universidad Industrial de Santander, Colombia

[2] Universidad Industrial de Santander, Colombia

[3] Universidad Industrial de Santander, Colombia

KEYWORDS

Digital Transformation

Technology Management

Scoping review

Digital technologies

Innovation

ABSTRACT

The objective of this article is to carry out an scoping review of the literature on the most important concepts and elements on the topic of digital transformation and technology management in the organizational/business field. For this, a search was carried out in the Scopus database, and 109 articles were selected. The results suggest that there is no consensus on the definition of digital transformation, however, there are similarities between existing definitions associating it with the "use of digital technologies" and relating it primarily to organizations, aspects that link it with other study topics such as dynamic capabilities, industry 4.0, innovation and knowledge management.

PALABRAS CLAVE

Transformación Digital

Gestión de la tecnología

Revisión exploratoria

Tecnologías digitales

Innovación

RESUMEN

El objetivo de este artículo es realizar una revisión exploratoria de la literatura sobre los conceptos y elementos más importantes acerca del tópico transformación digital y la gestión de la tecnología en el ámbito organizacional/empresarial. Para ello, se realizó una búsqueda en la base de datos Scopus, y se seleccionaron 109 artículos. Los resultados sugieren que no existe un consenso sobre la definición de la transformación digital, sin embargo, existen similitudes entre las definiciones existentes asociándola al "uso de tecnologías digitales" y relacionándola prioritariamente con las organizaciones, aspectos que la vinculan con otros tópicos de estudio como las capacidades dinámicas, la industria 4.0, la gestión de la innovación y el conocimiento.

1. INTRODUCCIÓN

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

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

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

Considerando lo anterior, en este proyecto de grado se analizarán antes y durante la pandemia los índices de los siguientes contaminantes atmosféricos con el propósito de determinar su comportamiento: Material Particulado (PM2.5) que se define como una mezcla de partículas sólidas y líquidas encontradas en el aire; Dióxido de Nitrógeno (NO2) formado como subproducto en los procesos de combustión a altas temperaturas como en los vehículos motorizados y las plantas eléctricas y, Ozono Troposférico u ozono superficial (O3) contaminante secundario originado a partir de reacciones que se activan por la luz solar entre contaminantes primarios como los óxidos de nitrógeno. De igual manera, se pretende aplicar 3 modelos de Aprendizaje Profundo para analizar el comportamiento de las variables escogiendo el de mayor rendimiento, con el fin de apoyar a las autoridades de salud y ambientales a diseñar políticas que contribuyan con la disminución de estos contaminantes atmosféricos, y de esta manera, mejorar la calidad de vida. El modelo de predicción a elegir se basa en los modelos de Aprendizaje Profundo, específicamente en las redes neuronales MLP, LSTM y Seq2Seq, las cuales tienen la capacidad de procesar grandes volúmenes de datos.

2. REVISIÓN DE LITERATURA

2.1. Análisis bibliométrico

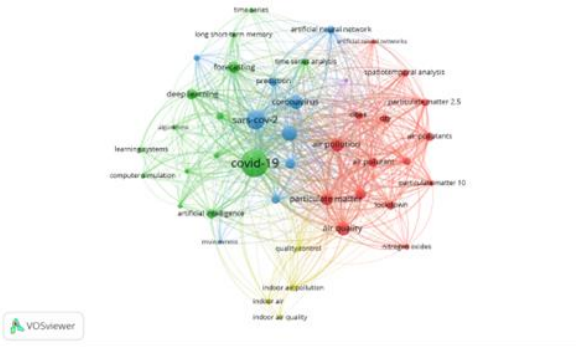
Con el fin de analizar el impacto en la calidad del aire a causa de las restricciones debidas a la pandemia provocada por el COVID 19 desde el 2020 al 2021, se ha desarrollado la siguiente ecuación de búsqueda (Figura 1), la cual fue ingresada en la base de datos SCOPUS.

Figura 1. Descriptores de búsqueda

Ecuación de búsqueda = Scopus: (“Deep Learning" OR "neural network”) AND ("air quality" OR "air pollution" OR "environmental quality") AND (covid$19 OR sars$cov$ OR coronavirus)
Tipo de documento = Artículo
Período de tiempo = 2020–2021

Se decide utilizar la base de datos Scopus, ya que ahonda a profundidad en el área profesional de estudio y la que más resultados acertados arrojó. Inicialmente se obtuvieron 322 artículos, cuyos datos fueron insertados en el software VosViewer mediante el cual se hizo un análisis de palabras clave con ocurrencia mínima de 5 menciones, seleccionando las de mayor concordancia con la ecuación de búsqueda como se muestra en la Figura 2.

Figura 2. Mapeo de palabras clave.

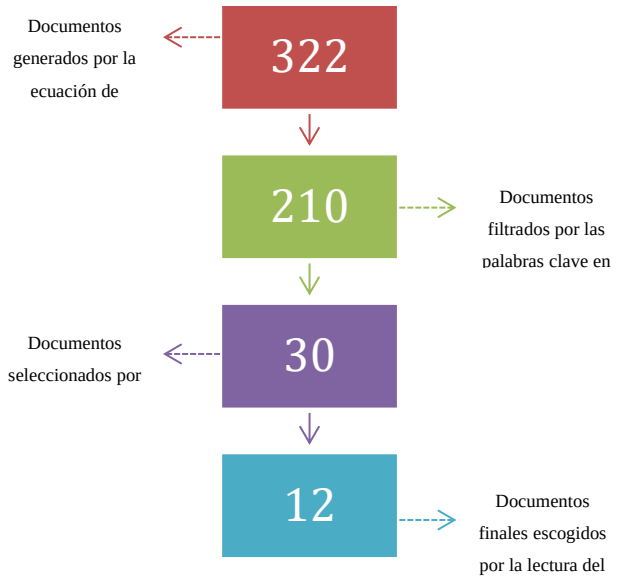


Fuente: VosViewer

Teniendo en cuenta el análisis hecho anteriormente se procede a filtrar la ecuación de búsqueda ejecutada en Scopus por las palabras claves: "COVID-19", "SARS-CoV-2", "Coronavirus Disease 2019", "Air Quality", "Coronavirus", "Air Pollution", "Particulate Matter”, "Forecasting", "Atmospheric Pollution", "Artificial Intelligence", "SARS Coronavirus", "Deep Learning", "Controlled Study", "Air Pollutant", "Prediction", "Lockdown", "Air Pollutants", "Artificial Neural Network", "Cities", "City", "Nitrogen Dioxide", "Indoor Air Pollution", "Quality Control", "Particulate Matter 2.5", "Covid-19", "Statistical Model", "Nitrogen Oxides", "Predictive Analytics", "Indoor Air Quality", "Learning Systems", "Mathematical Model", "Neural Networks", "Quarantine", "Time Series Analysis", "Long Short-term Memory", "Particulate Matter 10", "SARS", "Simulation", "Time Series", "Artificial Neural Networks", "Carbon Dioxide", "Environment", "pm2.5", "Indoor Air", "air pollution control", "computer simulation", "spatio$temporal analysis", "algorithms", "Indoor Environment" y limitado también por “artículos”, emitiendo 210 documentos académicos.

Luego, los documentos fueron depurados mediante la lectura de los títulos, dando como resultado 30 de estos y, finalmente se escogieron 12 artículos para la investigación con base en sus resúmenes, proceso presentado en la figura 3.

Figura 3. Proceso de selección de artículos



2.2. Análisis preliminar de la literatura

Durante el análisis de la literatura se pudo observar que los modelos más utilizados en los artículos son aquellos que implican redes neuronales artificiales como MLP, LSTM, ELM, ESN, RBF ya que son capaces de proporcionar mayor precisión en la predicción de los parámetros de calidad del aire. En la Tabla 1, se muestra la síntesis de la revisión de literatura donde se destacan las características de los modelos utilizados en cada artículo.

Tabla 1. Síntesis de la revisión de literatura.

Citación	Modelo	Observaciones
(Zhao, y otros, 2021)	Método CVAE (Codificador automático variacional condicional)	El método se basa en Deep Learning y tiene como objetivo identificar y evaluar de manera más realista los cambios anómalos y abruptos en las PM2.5, así proporciona un nuevo enfoque en la distribución de las fuentes contaminantes, además de detectar el impacto de las condiciones meteorológicas y actividades humanas en las anomalías de los contaminantes del aire y gases de efecto invernadero.
(Ekinci, Omurca, & Ozbay, 2021)	Modelos LSTM (Long-Short Term Memory), BILSTM (Bidirectional Long-Short Term Memory), STACKED LSTM, CNN LSTM Y CONV LSTM	Los modelos Deep Learning proporcionan una precisión adecuada en la resolución de problemas complejos, por lo tanto, usar un modelo de estos puede considerarse como una herramienta prometedora para predecir las concentraciones de O3 a nivel del suelo. El modelo Stacked LSTM tiene la capacidad de modelado más potente para clasificar datos.
(Etchie, Etchie, Jauro, Pinker, & Swaminathan, 2021)	MLPNN	Se utilizan redes neuronales para derivar el promedio mensual a nivel del suelo de los aerosoles.
(Tadano, y otros, 2020)	ELM (Extreme Learning Machine), ESN (Echo State Network), MLP (Multilayer Perceptron), RBF (The Radial Basis Function Networks)	Las redes neuronales artificiales demostraron ser herramientas de predicción robustas para estimar el mejor equilibrio entre los casos de COVID-19, el porcentaje de cierre y el nivel de contaminantes atmosféricos.
(Shatnawi & Abu-Qdais, 2021)	RNA	Una RNA adecuadamente entrenada y estructurada puede ser una herramienta útil para predecir los parámetros de calidad del aire con una precisión adecuada.

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

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

Entrenamiento: para ajustar los modelos.

Validación: con el fin de verificar el sobreentrenamiento y definir el número de neuronas en la capa intermedia.

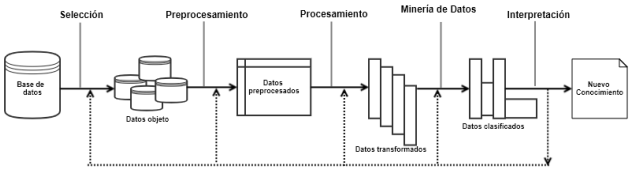
Prueba: se utiliza para evaluar el rendimiento de los modelos mediante el error cuadrático medio (MSE por sus siglas en inglés).

En este artículo de investigación, se presentarán los resultados obtenidos utilizando los modelos de redes neuronales artificiales MLP, LSTM y seq2seq, para la comparación y posterior predicción de la calidad del aire analizando los contaminantes NO2, PM2.5 y O3.

3. METODOLOGIA

Para el desarrollo de este artículo de investigación se adoptó la metodología "Descubrimiento de Conocimiento en Bases de Datos" o "Knowledge Discovery in Database" (KDD), la cual se presenta en la figura 4.

Figura 4. Metodología del artículo.



3.1. Etapa de selección

Para la etapa de selección, se hizo una búsqueda de los datos que apoyan el desarrollo de la investigación por medio de la fuente de información pública IDEAM. Se seleccionó como conjunto de datos las 3 ciudades: Bucaramanga (para efectos del presente artículo se hará referencia al Área Metropolitana de Bucaramanga), Bogotá y Cali, estableciendo una ventana de tiempo desde el 01/01/2019 al 31/07/2021 con el objetivo de comparar 3 escenarios: antes, durante y después del confinamiento obligatorio comprendido entre los días 12 de marzo al 31 de agosto del 2020.

Se encontraron inicialmente 116.467 registros, distribuidos así: 8.501 Bucaramanga, 83.236 Bogotá y 24.730 Cali; las variables PM2,5 (Material Particulado 2,5), O3 (Ozono Troposférico), NO2 (Dióxido de Nitrógeno), DV (Dirección del viento), VV (Velocidad del Viento), HA (Humedad del Aire), P (Presión atmosférica), PM10 (Material Particulado 10), TA (Temperatura del aire) fueron encontradas en dicho rango de tiempo.

Con el fin de identificar cada una de las variables en la tabla 2 se describen las nomenclaturas para ellas.

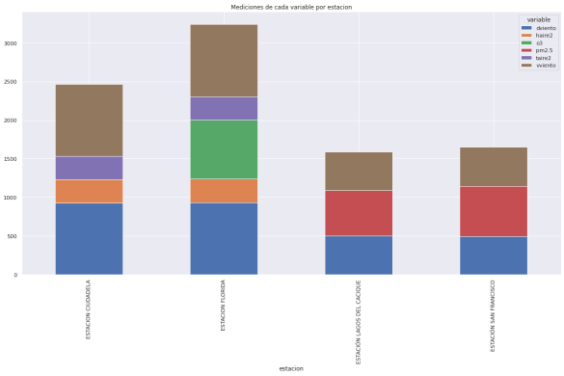
Tabla 2. Nomenclatura de las variables

Variable	Unidad de medida	Nomenclatura
Material Particulado 2.5	µg/m3	PM2.5
Ozono Troposférico	µg/m3	O3
Dióxido de Nitrógeno	µg/m3	NO2
Dirección del Viento	° (Grados)	DV
Velocidad del Viento	m/s	VV
Humedad Relativa	[%]	HR
Presión atmosférica	hPa	P
Material Particulado 10	µg/m3	PM10
Temperatura del Aire	°C	TA

En la base de datos se observó que Bucaramanga tiene 4 estaciones denominadas "Estación San Francisco", "Estación Ciudadela", "Estación Lagos del Cacique" y "Estación Lagos 1 - Floridablanca" de las cuáles sus registros respectivamente son de: 1.650,

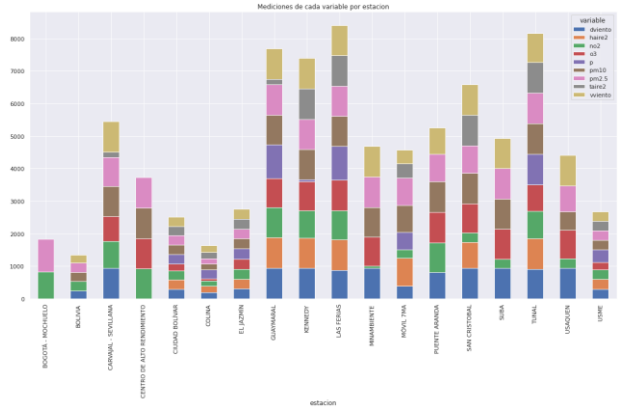
2.294, 1.592 y 2.965. En la figura 5 se muestra la distribución de las variables por estación.

Figura 5. Distribución de las variables para Bucaramanga.



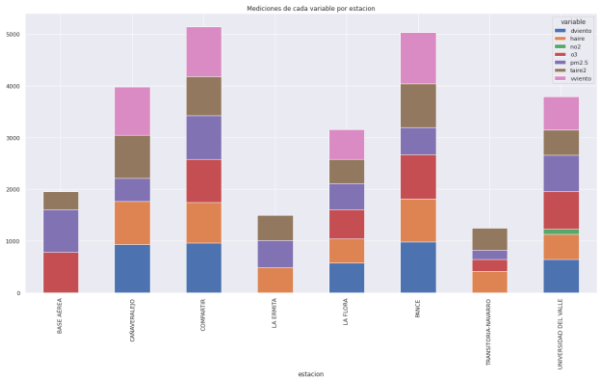
Para la ciudad de Bogotá, se encontraron 18 estaciones debido a su área y población, denominadas así con su respectiva cantidad de datos: “Estación Bolivia” (1.341), “Estación Carvajal” (5.447), “Estación Centro de alto rendimiento” (3.727), “Estación Ciudad Bolívar” (2.509), “Estación Colina” (1.538), “Estación El Jazmín” (2.722), “Estación Guaymaral” (7.593), “Estación Kennedy” (7.393), “Estación Las Ferias” (8.307), “Estación MinAmbiente” (4.694), “Estación Mochuelo” (1.646), “Estación Móvil” (4.482), “Estación Puente Aranda” (5.251), “Estación San Cristóbal” (6.583), “Estación Suba” (4.938), “Estación Tunal” (8.067), “Estación Usaquén” (4.412) y “Estación Usme” (2.586), observándose en la figura 6 la distribución de las variables por estación.

Figura 6. Distribución de las variables para Bogotá.



En la ciudad de Cali se encontraron las siguientes 8 estaciones con sus respectivas cantidades: “Estación Base aérea” (1.846), “Estación Cañaveralejo” (3.882), “Estación Compartir” (4.923), “Estación La Ermita” (1.440), “Estación La Flora” (2.978), “Estación Pance” (4.816), “Estación Transitoria Navarro” (1.248) y “Estación Univalle” (3.597). En la figura 7 se presenta la distribución de las variables.

Figura 7. Distribución de las variables para Cali.



3.2. Etapa de preprocesamiento

Con el fin de obtener los datos preprocesados se realizó la limpieza de estos, la cual consistió inicialmente en unificar los nombres de las columnas y las filas, dejar los formatos adecuados de las fechas, quitar los caracteres especiales, cambiar signos decimales, y convertir las variables en columnas

Posteriormente, se eliminaron las estaciones, mostradas en la tabla 3, que por su baja calidad y cantidad de datos afectarían directamente al proceso de procesamiento y, por ende, de predicción, las cuales representan 21.496 datos.

Tabla 3. Estaciones eliminadas por ciudad.

Ciudad	Estación eliminada
Bogotá	"Mochuelo", "Colina", "Bolivia", "Ciudad Bolívar", "El Jazmín", "Móvil 7ma", "Usaquén" y "Usme"
Cali	"Base aérea", "La Ermita" y "Transitoria"

Asimismo, se realizó la eliminación de datos atípicos por medio del método IQR, modificándose como se muestra en la tabla 4, debido a que el método original quita buenos datos que contribuyen con el estudio. De esta manera, la cantidad de datos se reduce a 94.629.

Tabla 4. Método IQR.

Método original	Método utilizado
$\boldsymbol{Q_1 - 1,5 * IQR}$	$Q_1 - 2{,}5 * IQR$
$\boldsymbol{Q_3 + 1,5 * IQR}$	$Q_3 + 2{,}5 * IQR$

De los datos obtenidos anteriormente se eliminaron 16.456 ruidosos, correspondientes a las variables PM10, P y NO2 (únicamente para la ciudad de Cali), quedando así 78.173 datos que se seguirán tratando.

Ya que existen estaciones con datos faltantes para las fechas establecidas en la ventana de tiempo, estos datos afectan la serie de tiempo, razón por la cual se decide no trabajar por estaciones sino por ciudad, promediando los valores registrados, y, así obtener datos más organizados y de mayor calidad a la hora de hacer el procesamiento. Adicionalmente, para la ciudad de Bucaramanga en las variables TA y HA, al observar su comportamiento se notó una tendencia constante, por lo cual, se decidió rellenar los datos faltantes con los datos del año anterior, generando un total de 17.192 datos.

Seguidamente, se procedió a realizar la imputación de datos utilizando el método k-means o k-vecino más próximo. En este, se iteró sobre un valor mínimo de 3 y un valor máximo de la décima parte del total de datos, eligiendo el mejor k próximo. Teniendo el caso de que, si el mejor k próximo es mayor o igual a 300, se establece un valor máximo de 300 datos vecinos. De este proceso se reemplazaron 725 datos, sin embargo, algunos de estos no presentan valores adecuados, por lo cual, se decidió imputarlos nuevamente por medio de otras estrategias mostradas en la tabla 5, obteniendo finalmente 17.917 datos preprocesados.

Tabla 5. Estrategias de imputación.

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

Posteriormente, se realiza un análisis descriptivo de la media, la desviación estándar, mínimos, máximos y gráficas de las variables objeto de estudio, con el fin de conocer y estudiar su comportamiento. En la tabla 6 se muestran dichos resultados para los contaminantes.

Tabla 6. Estadísticas obtenidas para cada variable en cada ciudad.

	NO2	O3	PM2.5
Cantidad de Datos	943	943	943
Promedio	24,540	48,904	33,214
Desviación estándar	8,685	12,265	10,684

Valor mínimo	6,078	18,034	12,496
25%	17,662	40,408	25,148
50%	23,725	48,499	32,098
75%	30,754	57,188	39,440
Valor máximo	49,960	88,527	84,639
	NO2	O3	PM2.5
Cantidad de Datos	0	943	943
Promedio	0	22,315	15,985
Desviación estándar	0	8,992	8,550
Valor mínimo	0	4,996	4,435
25%	0	15,132	9,983
50%	0	22,071	13,721
75%	0	28,743	18,953
Valor máximo	0	49,342	47,836
	NO2	O3	PM2.5
Cantidad de Datos	0	943	943
Promedio	0	28,804	16,128
Desviación estándar	0	7,910	6,099
Valor mínimo	0	9,988	3,588
25%	0	22,882	11,487
50%	0	28,045	15,739
75%	0	34,239	20,302
Valor máximo	0	58,400	40,460

3.3. Etapa de procesamiento

Una vez realizada la limpieza e imputación de datos, llega la etapa de procesamiento en la cual se procede a ajustar los modelos de Deep Learning. Para el desarrollo de este artículo se aplicó para cada ciudad un modelo MLP, LSTM y Seq2Seq como se muestra a continuación.

3.3.1. Modelo MLP

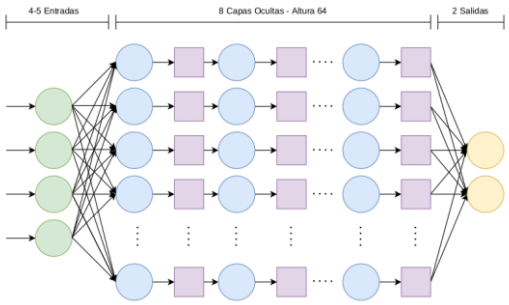
Una vez obtenidos los datos preprocesados, estos se utilizan para el ajuste de los modelos, en esta red cada neurona en una capa está conectada a todas las neuronas de la capa anterior y de la capa siguiente, formando una red de conexiones ponderadas. Durante el entrenamiento, el modelo ajusta los pesos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

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

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

Se estableció un tamaño de lote de 8, lo que significa que se utilizan 8 ejemplos de entrenamiento en cada iteración del algoritmo. La tasa de aprendizaje se estableció en 0.01 para controlar la cantidad en que los pesos de la red neuronal se ajustan durante el entrenamiento. La función de pérdida seleccionada fue la función de error cuadrático medio (MSE), adecuada para problemas de regresión en los que la salida esperada es un valor numérico y comúnmente utilizada en la literatura (Martinez H, 2020). Con esta, se mide la discrepancia entre las predicciones del modelo y los datos reales.

Figura 8. Arquitectura del modelo MLP.

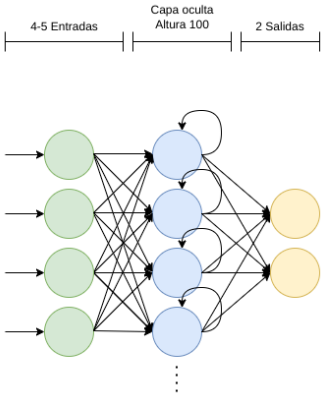


3.3.2. Modelo LSTM

Para predecir datos a partir de una secuencia temporal de datos pasados por una capa oculta de 100 unidades, una función de activación Tanh, sin Dropout, sin Batch Normalization, una tasa de aprendizaje de 0.04, una función de pérdida MSE, 120 Epochs y un tamaño de lote de 8. Estos parámetros se seleccionaron después de realizar pruebas en diferentes combinaciones y seleccionar los que proporcionaron los mejores resultados. La arquitectura de esta red se presenta en la Figura 9.

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

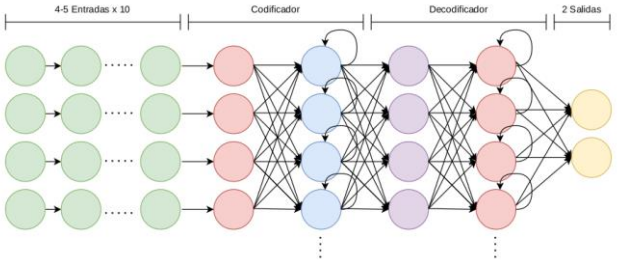
Figura 9. Arquitectura del modelo LSTM.



3.3.3. Modelo Seq2Seq

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

Figura 10. Arquitectura del modelo Seq2Seq.



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

ese vector de características y lo utiliza para generar una secuencia de salida de longitud variable que corresponde a la secuencia de entrada.

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

3.4. Etapa de minería de datos

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

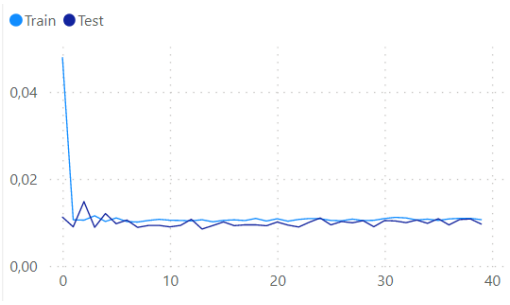
3.4.1. Modelo MLP

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

3.4.1.1. Bogotá

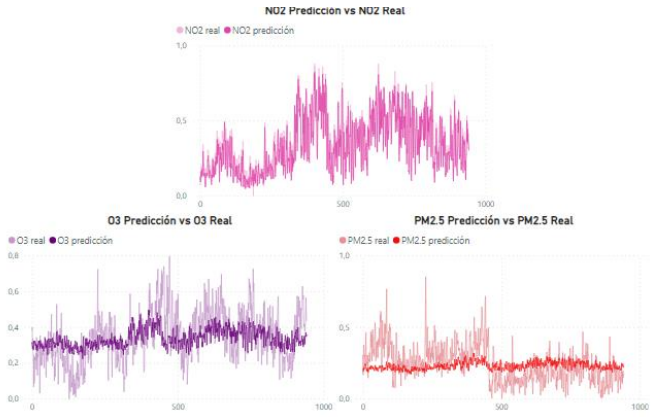
La cantidad de datos utilizada fue de 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 11 se puede observar que ocurre un ajuste de los datos a partir de la séptima iteración, presentando un error promedio de 0,01154 para entrenamiento y 0,00996 para la prueba.

Figura *11. Función de pérdida modelo MLP para Bogotá.*



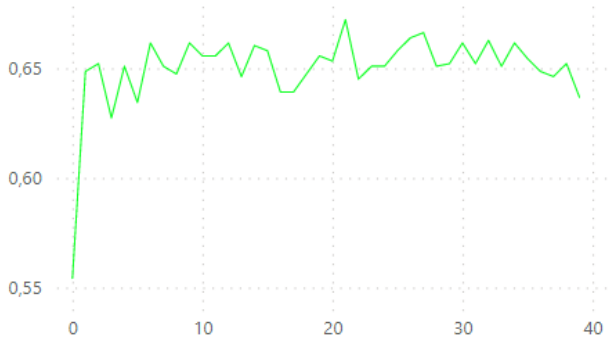
En la Figura 12 se observan las predicciones para los 943 datos con cada uno de los contaminantes (NO2, PM2.5, O3), allí se puede evidenciar que el contaminante NO2 tuvo un buen ajuste a los datos reales, sin embargo, la predicción de PM2.5 y O3 presenta un resultado centrado en el promedio de los datos reales, es decir, el modelo no logra predecir adecuadamente estos contaminantes.

Figura 12. Predicción de los contaminantes modelo MLP para Bogotá.



De acuerdo con lo anterior, el grado de concordancia entre la predicción y los datos reales se representa con la función de precisión mostrada en la Figura 13, la cual evidencia que existe en promedio una precisión de 0,65 que puede ser debida a PM2.5 y O3 por presentar una baja calidad en la predicción.

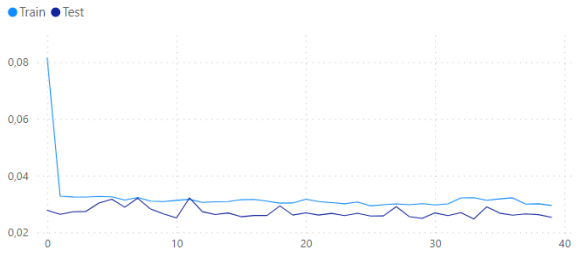
Figura 13. Función de precisión modelo MLP para Bogotá.



3.4.1.2. Bucaramanga

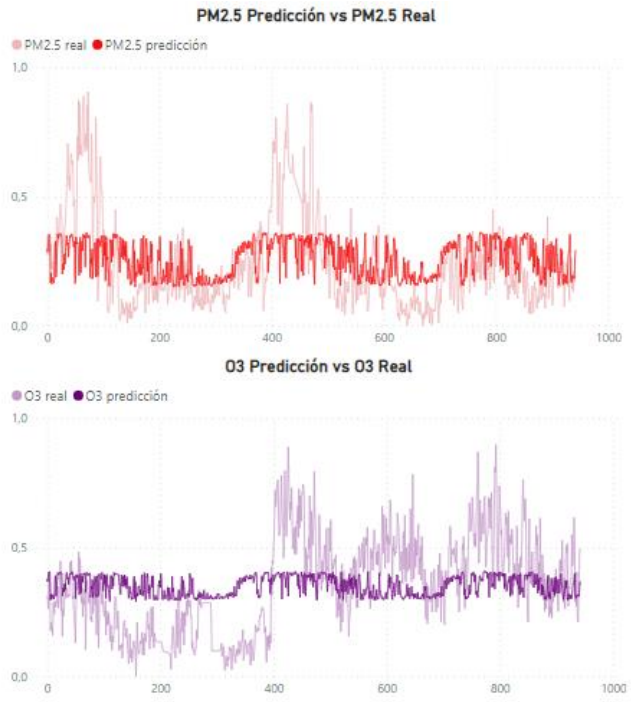
Para el entrenamiento se utilizaron 5.092 datos y para la validación 566, en esta ocasión el ajuste de los datos ocurre a partir de la iteración 12 como se representa en la Figura 14. El error cuadrático medio tuvo un promedio de 0,03226 para entrenamiento y 0,02709 para la prueba.

Figura 14. Función de pérdida modelo MLP para Bucaramanga.



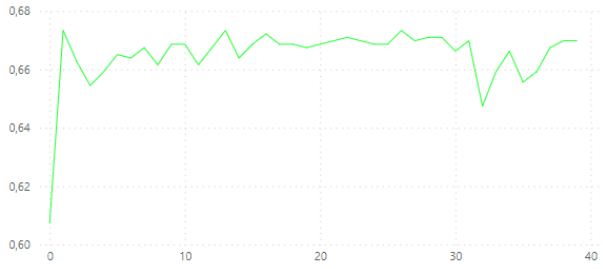
En la Figura 15 se observan las predicciones para la ciudad de Bucaramanga con los contaminantes PM2.5 y O3, en los cuales se presenta un bajo rendimiento del modelo del mismo modo que ocurre en la ciudad de Bogotá, la predicción se centra en el promedio de los datos reales, pero no es capaz de predecir su variación.

Figura 15. Predicción de los contaminantes modelo MLP para Bucaramanga.



La calidad de esta predicción se ve reflejada en la Figura 16, en donde la precisión del modelo tiene como promedio 0,66, siendo aún muy pobre para considerar que este modelo sea el óptimo.

Figura 16. Función de precisión modelo MLP para Bucaramanga.

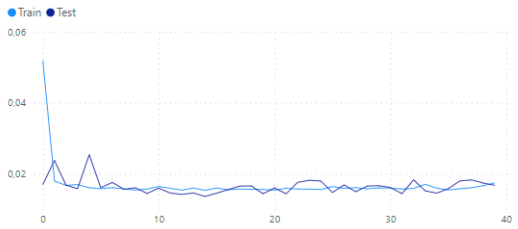


3.4.1.3. Cali

Para el entrenamiento se utilizaron 5.092 datos y para la validación 566, en la Figura 17 se evidencia un ajuste de los datos a partir de la décima iteración. El error cuadrático medio tuvo un promedio de 0,016925

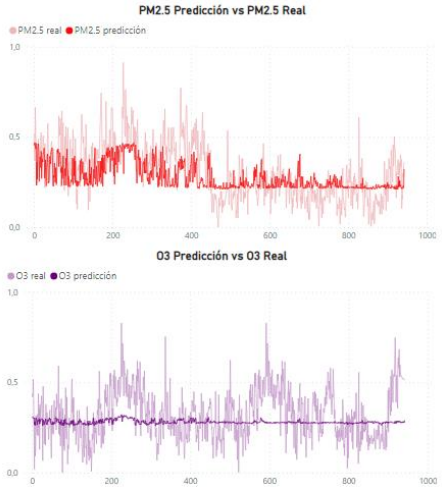
para entrenamiento y 0,016462 para la prueba. A pesar de que este es bastante bajo, se observa que los datos de prueba están presentando variaciones en comparación con los de entrenamiento.

Figura 17. Función de pérdida modelo MLP para Cali.



Del mismo modo que ocurre con las ciudades de Bogotá y Bucaramanga, en la predicción de los contaminantes PM2.5 y O3 para Cali se observa en la Figura 18 que el comportamiento de los datos predichos se ubica hacia el promedio de los datos reales, lo cual, se considera que el modelo no se ajusta al tipo de problema que se está abarcando.

Figura 18. Predicción de los contaminantes modelo MLP para Cali.



Por último, en la Figura 19 se muestra la función de precisión para esta ciudad, donde se corrobora que la precisión del modelo es pobre con un promedio de 0,62.

Figura 19. Función de precisión modelo MLP para Cali.

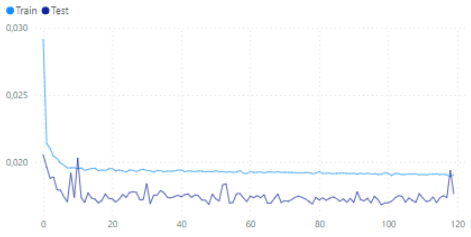
3.4.2. Modelo LSTM

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

3.4.2.1. Bogotá

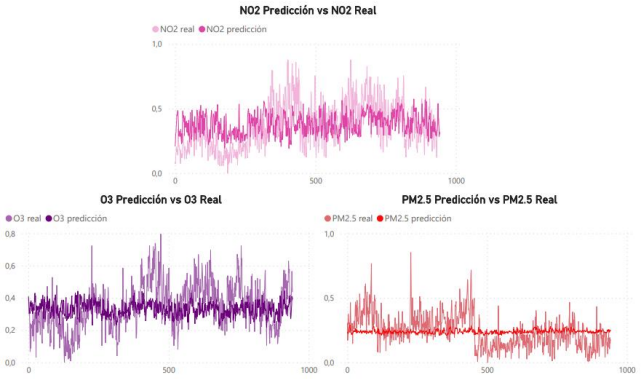
La cantidad de datos utilizada fue de 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 20 se puede observar que ocurre un ajuste de los datos a partir de la décima tercera iteración, presentando un error promedio de 0,01938 para entrenamiento y 0,01744 para la prueba. Se evidencia que en la iteración 118 la validación de los datos aumenta de nuevo, traduciéndose en un posible sobreajuste por una cantidad grande de Epochs.

Figura 20. Función de pérdida modelo LSTM para Bogotá**.**



En la predicción presentada en la Figura 21, se analiza que los contaminantes NO2 y O3 intentan ajustarse más a la variación de los datos reales, sin embargo, no logran acercarse a los límites superiores e inferiores. Por otra parte, el contaminante PM2.5 continúa presentando un comportamiento poco eficiente debido a la naturaleza de sus datos.

Figura 21.1 Predicción de los contaminantes modelo LSTM para Bogotá.



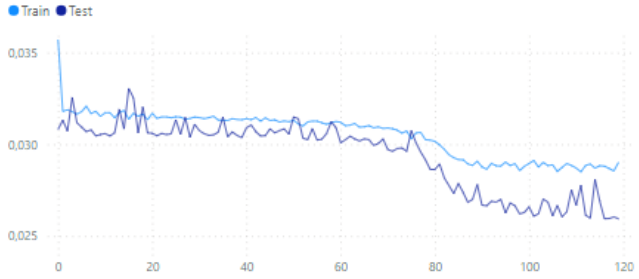
Teniendo en cuenta lo anterior, la función de precisión para esta ciudad se presenta en la Figura 22, de la cual se puede decir que no se observa gran variación y tiende a estabilizarse entre un 57 y 60 % a partir del Epoch 5. Se deduce que no se obtiene un valor más alto a causa del contaminante PM2.5.

Figura 22. Función de precisión modelo LSTM para Bogotá.

3.4.2.2. Bucaramanga

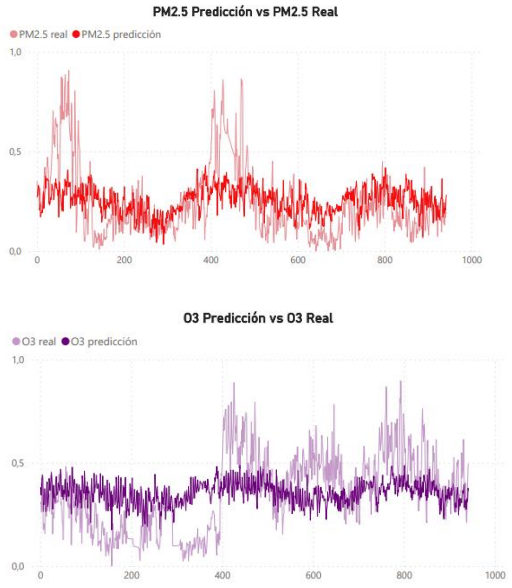
Para la ciudad de Bucaramanga, se utilizaron 5.092 datos para el entrenamiento y para la prueba 566, en esta ocasión el ajuste de los datos ocurre a partir de la iteración 20 como se representa en la Figura 23, no obstante, se evidencia que a partir de la iteración 75 disminuye la pérdida tanto para el entrenamiento como para la validación en un valor aproximado a 0,002, el cual no es significativo. El error cuadrático medio tuvo un promedio de 0,03054 para entrenamiento y 0,02938 para la validación.

Figura 23. Función de pérdida modelo LSTM para Bucaramanga.



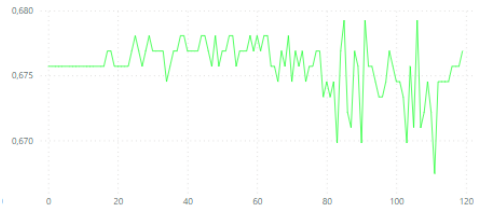
En la Figura 24, se muestra la predicción correspondiente a los contaminantes PM2.5 y O3, en la cual se evidencia que los datos predichos no se ajustan a los datos reales, es decir, no alcanzan a ajustarse a la variación real siendo más notorio este caso en O3. Adicionalmente, en PM2.5 se observa que la predicción intenta imitar el comportamiento de los datos reales contrario a lo sucedido en O3.

Figura 24. Predicción de los contaminantes modelo LSTM para Bucaramanga.



El análisis anterior se ve respaldado con la Figura 25, donde la precisión de la predicción presenta un valor de 0,68 en promedio, notándose una variación insignificante entre 0,67 y 0,68 a partir de la iteración 75, misma incidencia con la función de pérdida presentada en la Figura 23.

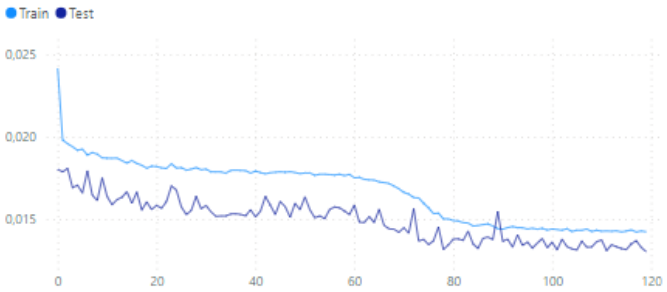
Figura 25. Función de precisión modelo LSTM para Bucaramanga.



3.4.2.3. Cali

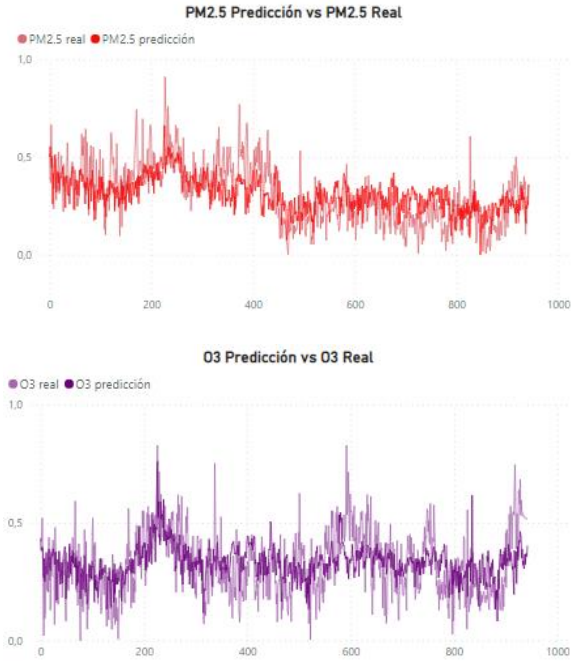
En Cali, 5.092 datos fueron utilizados para entrenar al modelo y 566 para validarlo, en la Figura 26 se evidencia un mayor ajuste de los datos y una disminución en la pérdida a partir del Epoch 80. El error cuadrático medio tuvo un promedio de 0,01670 para entrenamiento y 0,01490 para la prueba.

Figura 26. Función de pérdida modelo LSTM para Cali.



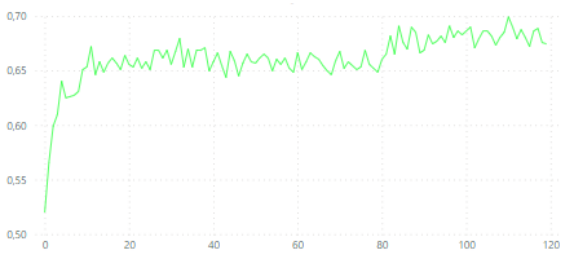
En este sentido, en la Figura 27 se evidencia que los datos predichos a comparación con las ciudades anteriores se ajustan un poco más a los datos reales para ambos contaminantes, siguiendo su comportamiento, aunque no logran con éxito una predicción óptima.

Figura 27. Predicción de los contaminantes modelo LSTM para Cali.



Lo anterior se ve reflejado también en la Figura 28, donde la precisión toma valores mayores a comparación de las otras ciudades analizadas, aunque no son altamente significativos, oscilando entre 0,65 y 0,7.

Figura 28. Función de precisión modelo LSTM para Cali.



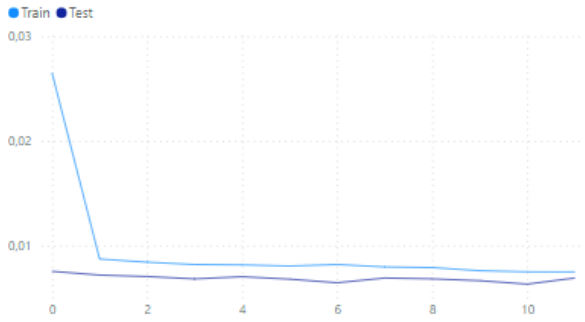
3.4.3. Modelo Seq2Seq

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

3.4.3.1. Bogotá

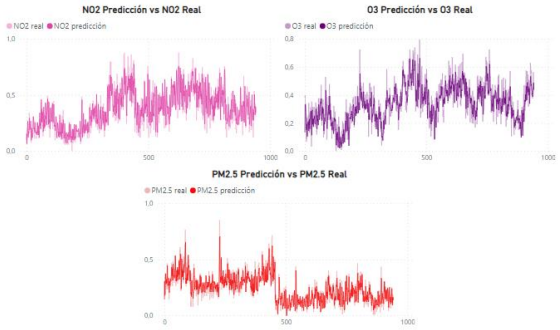
En la ciudad de Bogotá, se utilizaron 5.941 para el entrenamiento y 660 para la validación del modelo, en la Figura 29 se muestra que ocurre un ajuste de los datos a partir de la primera iteración, presentando un error promedio de 0,00887 para entrenamiento y 0,00685 para la prueba.

Figura 29. Función de pérdida modelo Seq2Seq para Bogotá.



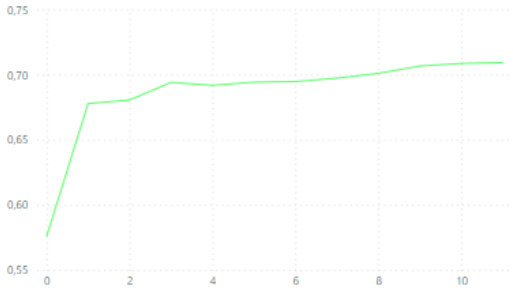
En la Figura 30, se evidencia que el modelo está realizando una predicción óptima para los 3 contaminantes, ya que se ajusta a los datos reales y sigue el comportamiento de la naturaleza de estos, a diferencia de lo ocurrido en los modelos anteriores.

Figura 30. Predicción de los contaminantes modelo Seq2Seq para Bogotá.



A continuación, en la Figura 31, se observa que la precisión de este modelo para Bogotá se encuentra entre 0,68 y 0,71 a partir de la primera iteración, respaldando lo analizado anteriormente.

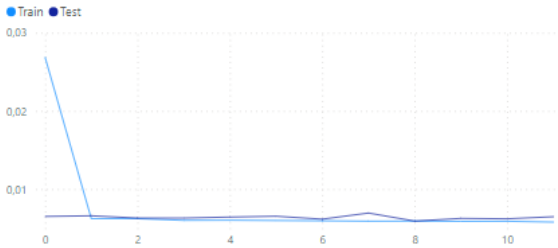
Figura 31. Función de precisión modelo Seq2Seq para Bogotá.



3.4.3.2. Bucaramanga

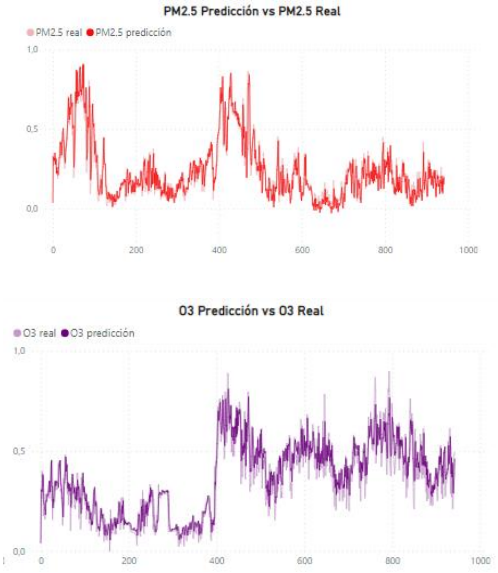
Los datos utilizados en esta ocasión fueron 5.092 datos para el entrenamiento y 566 para la validación. En la Figura 32 se refleja que el ajuste de los datos ocurre a partir de la primera iteración, siendo el promedio del error cuadrático medio de 0,007723 para entrenamiento y 0,006386 para la prueba. Estos son los más bajos para esta ciudad en comparación con los anteriores modelos analizados.

Figura 32. Función de pérdida modelo Seq2Seq para Bucaramanga.



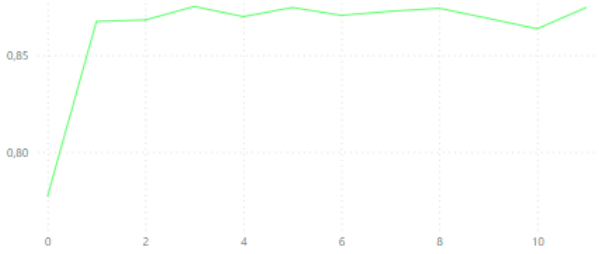
En cuanto a la visualización de la predicción presentada en la Figura 33, su comportamiento es consecuente al análisis realizado anteriormente. Se destaca el contaminante PM2.5 al ajustarse adecuadamente a los datos reales en comparación con O3, el cual, no toma por completo los límites inferiores.

Figura 33. Predicción de los contaminantes modelo Seq2Seq para Bucaramanga.



La precisión para esta ciudad toma valores aproximados a 0,86, siendo altos en contraste con los demás modelos. Esto se representa en la Figura 34.

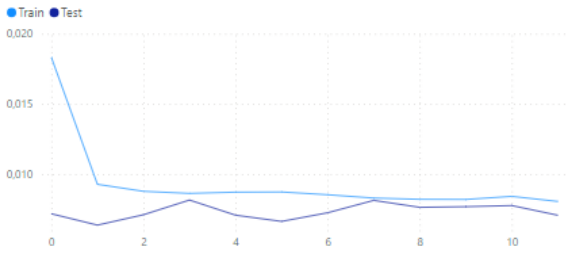
Figura 34. *Función de precisión modelo Seq2Seq para Bucaramanga.*



3.4.3.3. Cali

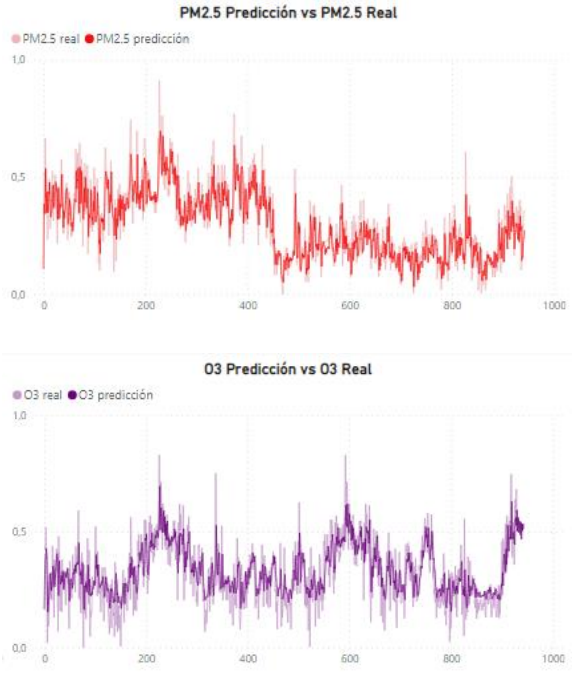
5.092 datos fueron utilizados para entrenar al modelo y 566 para validarlo en la ciudad de Cali, en la Figura 35 se evidencia un mayor ajuste de los datos a partir del Epoch 7. El error cuadrático medio tuvo un promedio de 0,009351 para entrenamiento y 0,007342 para la prueba.

Figura 35. Función de pérdida modelo Seq2Seq para Cali



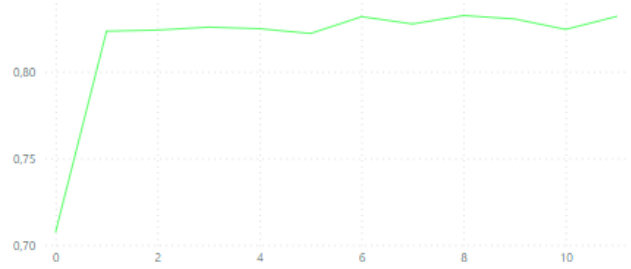
En la predicción de los contaminantes se demuestra que a pesar de que no es óptima, ya que, no predice acertadamente los límites, es mejor que los modelos LSTM y MLP al entender el comportamiento de los datos reales y su función de pérdida es mínima como se muestra en la Figura 36.

Figura 36. Predicción de los contaminantes modelo Seq2Seq para Cali.



Para corroborar lo dicho anteriormente, en la Figura 37 se observa que la función de precisión se encuentra en 0,82 aproximadamente, siento este valor el más alto para la función de precisión en comparación con el resultado de los anteriores modelos.

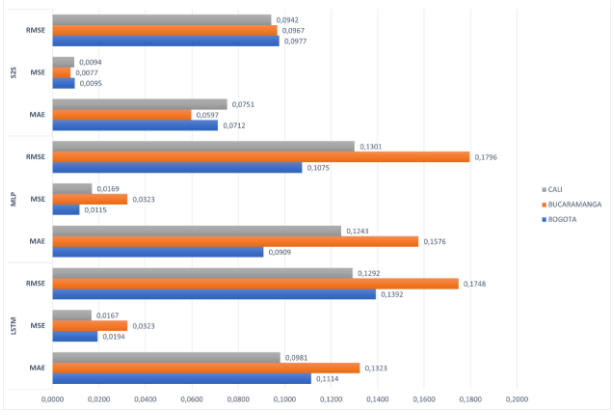
Figura 37. Función de precisión modelo Seq2Seq para Cali.



3.5. Etapa de interpretación

Como paso final, se realiza el proceso de verificación y validación de los modelos tratados anteriormente utilizando 3 métricas conocidas como: el error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz cuadrada del promedio de las diferencias cuadradas (RMSE), como se muestra en la Figura 38.

Figura 38. Relación de métricas de validación.



Para los modelos MLP y LSTM, en general la ciudad con valores de error más altos es Bucaramanga, sin embargo, en el modelo SEQ2SEQ es la que presenta mayor valor de precisión y menor MSE en promedio como se observó en la Figura 38, de igual manera, se destaca que las demás ciudades no presentan una diferencia significativa como sucede en los demás modelos. Por otra parte, la ciudad de Bogotá presentó

el menor valor para las tres métricas de validación en el modelo MLP y en la ciudad de Cali se evidenciaron valores más bajos en las métricas para el modelo LSTM respecto a las demás ciudades.

Adicionalmente, el modelo de mayor error es el LSTM para las métricas de MSE y RMSE y, en general, el modelo SEQ2SEQ presenta en promedio los valores más bajos de las métricas de validación, destacándose como el mejor. El anterior análisis se presenta en la Tabla 7.

Tabla 7. Relaciones métricas de validación.

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

4. RESULTADOS Y ANÁLISIS

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

Por otra parte, el modelo LSTM, aunque superó el problema del MLP y logró obtener resultados aceptables en la predicción de los datos, no son los mejores ya que, no consiguió interpretar adecuadamente su variación.

Finalmente, el modelo Seq2Seq tuvo la capacidad de predecir los contaminantes con mejor desempeño que los otros modelos sin presentar sobreajuste. Esto sugiere que el modelo Seq2Seq es altamente efectivo para la predicción de series de tiempo a pesar de que tiene un costo computacional más alto.

Por otra parte, teniendo cuenta los valores máximos permitidos a partir del 2018 según la resolución 2254 del 01 de noviembre de 2017 (Ministerio de Ambiente y Desarrollo Sostenible, 2017) en la Tabla 8 se resumen los valores para los contaminantes trabajados en el presente artículo.

Tabla 8. Valores máximos permitidos según la Resolución.

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

Adicionalmente, en la Tabla 9, se reportan los valores promedios y máximos presentados en los datos preprocesados con el fin de identificar si alguno de estos supera o se aproxima a los niveles máximos permitidos por las autoridades ambientales.

Tabla 9. Valores promedios y máximos de los datos preprocesados.

	2019					
	NO2	Máximo reportado	O3	Máximo reportado	PM2.5	Máximo reportado
Bogotá	18,7808	45,9985	41,5860	82,0855	38,9908	84,6395
Bucaramanga	N/A	N/A	22,3153	28,7833	15,9850	47,8375
Cali	N/A	N/A	28,9454	58,3996	20,6371	40,4605
	2020					
	NO2	Máximo	O3	Máximo	PM2.5	Máximo

		report ado		report ado		report ado
Bogotá	29,8500	49,9599	55,2354	88,5274	30,7479	73,1236
Bucaramanga	N/A	N/A	26,7381	48,8792	16,8541	45,7792
Cali	N/A	N/A	29,0908	58,3996	14,0181	34,7292

Dicho lo anterior, se evidencia que los contaminantes NO2 y O3 para las tres ciudades no superan el valor máximo permitido por las autoridades, por ende, se encuentran siempre en un rango de "Bueno". Por otra parte, el contaminante PM2.5 al menos un día para las distintas ciudades presenta un valor que supera el máximo permitido en el 2019 y para el 2020, solo Bogotá y Bucaramanga superan al menos un día los límites.

Ahora, con el fin de comprobar la hipótesis del planteamiento del problema, la cual describe que la contaminación atmosférica (PM2.5, NO2 y O3) disminuyó después de la etapa del confinamiento provocado por el COVID 19, se realiza inicialmente una gráfica de boxplot donde se comparan visualmente los tres momentos: Antes, Durante y Después del cierre. Posteriormente, para verificar si existen diferencias significativas en el valor medio de los tres escenarios, se aplica un ANOVA [1] estableciendo un nivel de significancia de 1% donde primero se corroboran los supuestos de normalidad con la prueba Kolmogorov-Smirnov[2]; de homocedasticidad[3] con la prueba Levene [4] y se asume que los datos son independientes entre sí. Si efectivamente las medias son diferentes, se realiza una prueba Tukey para determinar cuáles de los escenarios difieren.

Por otra parte, si los supuestos descritos anteriormente no se cumplen, se debe implementar una prueba no paramétrica como Kruskal-Wallis [5] para demostrar si la media de los grupos es igual y, ya que esta no proporciona una respuesta a la pregunta de cuáles de los grupos difieren, se requiere hacer una prueba Dunn-Bonferroni[6].

En la Figura 39. Boxplot escenarios de los contaminantes para la ciudad de Bogotá.Figura 39 se muestra la comparación del antes, durante y después del confinamiento para los contaminantes NO2, O3 y PM2.5 en la ciudad de Bogotá, allí se comprueba que en efecto, para el contaminante NO2 después del cierre tuvo mayor concentración. Para el contaminante O3 se presenta un aumento significativo durante el confinamiento, contrario a lo ocurrido con PM2.5, pues este disminuye durante ese mismo escenario.

[1] Análisis de la Varianza es una fórmula estadística que se utiliza para comparar las varianzas entre las medias de diferentes grupos.

[2] La prueba Kolmogorov-Smirnov se utiliza para contrastar si un conjunto de datos se ajusta o no a una distribución normal.

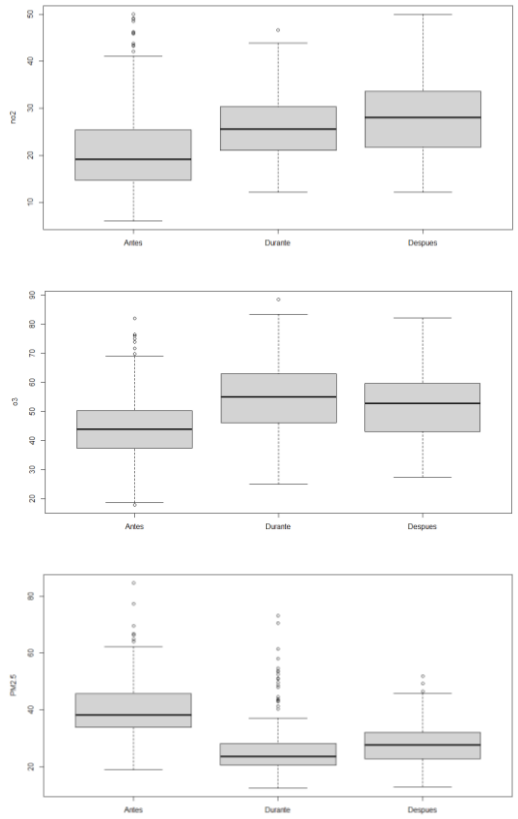
[3] El supuesto de Homocedasticidad considera que la varianza es constante (no varía) en los diferentes niveles de los grupos.

[4] La prueba de Levene comprueba si varios grupos tienen la misma varianza en la población.

[5] La prueba de Kruskal-Wallis es una prueba de hipótesis no paramétrica para muestras múltiples independientes que se utiliza cuando no se cumplen los supuestos de un ANOVA.

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

Figura 39. Boxplot escenarios de los contaminantes para la ciudad de Bogotá.



En la Tabla 10 se presentan los valores de p[7] para las distintas pruebas realizadas en los contaminantes junto con el resultado arrojado por la hipótesis nula de cada prueba. Se puede concluir que los supuestos no se cumplieron en su totalidad para el contaminante NO2, por ende, se aplica la prueba de Kruskal-Wallis, dando como resultado que los valores medios de los tres escenarios son diferentes. Por el contrario, los contaminantes PM2.5 y O3 presentaron el cumplimiento de los supuestos, por ende, se aplica ANOVA.

Tabla 10. Pruebas estadísticas para el contaminante NO2 en Bogotá.

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

Ahora, dependiendo del resultado de las pruebas realizadas anteriormente, se aplica para cada caso las pruebas post-hoc[8] correspondientes en la Tabla 11.

Tabla 11. Pruebas post-hoc para los contaminantes de la ciudad de Bogotá.

Contaminante	Prueba	Par de escenarios	Valor p
NO2	Dunn-Bonferroni	Antes – Después	$1{,}25x10^{-33}$
		Antes – Durante	$5{,}14x10^{-11}$
		Después – Durante	$2{,}42x10^{-3}$
O3	Tukey	Después – Antes	0
		Durante – Antes	0
		Durante – Después	0,0081
PM2.5	Tukey	Antes – Después	0
		Antes – Durante	0
		Después – Durante	0,1616

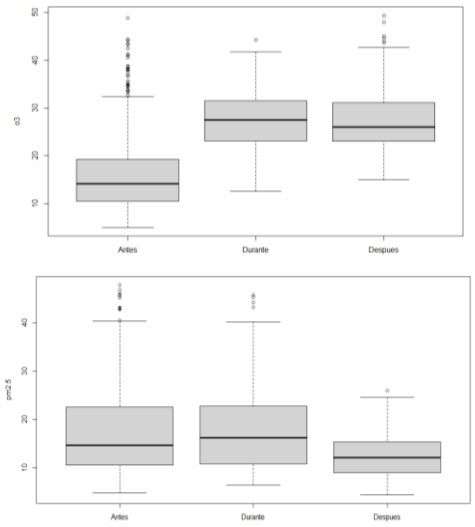
[7] El valor p es la probabilidad de que un valor estadístico calculado sea posible dada una Hipótesis nula cierta.

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

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

En la Figura 40 se muestra la comparación del antes, durante y después del confinamiento para los contaminantes PM2.5 y O3 en la ciudad de Bucaramanga, se puede observar que las medias son similares en los momentos Durante y Después del confinamiento del contaminante O3 y Antes y Durante del contaminante PM2.5, es decir, ambos presentaron mayor concentración en estos escenarios.

Figura 40. Boxplot escenarios de los contaminantes para la ciudad de Bucaramanga.



En la Tabla 12 se presentan los valores de p para las distintas pruebas estadísticas realizadas, se concluye que los supuestos no se cumplieron en su totalidad para ambos contaminantes, por lo tanto, la prueba de Kruskal-Wallis fue aplicada para estos, dando como resultado que hay evidencia significativa de que las medias en los tres escenarios son diferentes.

Tabla 12. Pruebas estadísticas para los contaminantes en la ciudad de Bucaramanga.

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

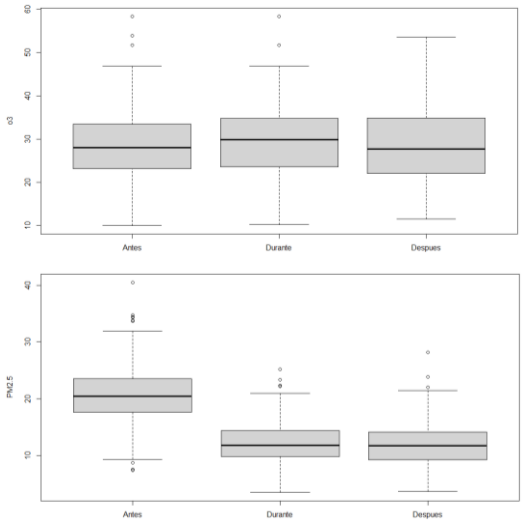
Posteriormente, se aplica para ambos contaminantes la prueba post-hoc Dunn-Bonferroni como se muestra en la Tabla 13. De esta se puede concluir que los pares de escenarios del contaminante PM2.5 "Antes - Después" y "Después - Durante" son aquellos que tienen diferencias entre sí. En cuanto al contaminante O3 se deduce que los grupos que presentan diferencias en sus valores medios son "Después – Antes" y "Durante – Antes".

Tabla 13. Prueba Dunn-Bonferroni para las contaminantes de la ciudad de Bucaramanga.

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

Por último, para los contaminantes de la ciudad de Cali se presenta la Figura 41 donde se puede evidenciar que para el contaminante O3 no hay diferencias significativas en los valores medios de cada escenario, mientras que para el PM2.5 las medias de los momentos Durante y Después son similares.

Figura 41. Boxplot escenarios de los contaminantes para la ciudad de Cali.



Los valores de p para las pruebas estadísticas seleccionadas son tabulados en la Tabla 14, se concluye de igual manera que ocurre en la ciudad de Bucaramanga, que los supuestos no se cumplen para ambos contaminantes, por lo tanto, la prueba de Kruskal-Wallis se aplica y como resultado, se comprueba lo visto en la figura anterior para el contaminante O3, pues no hay evidencia significativa para afirmar que los valores medios de los escenarios son diferentes, por ello, no es necesario aplicar la prueba post-hoc para este. Por otra parte, el PM2.5 presenta diferencias en las medias de sus grupos.

Tabla 14. Pruebas estadísticas para los contaminantes en la ciudad de Cali.

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

Considerando las pruebas anteriores, se aplica la prueba Dunn-Bonferroni únicamente para el contaminante PM2.5 como se observa en la Tabla 15. De esta se puede concluir que los escenarios que difieren entre sí son "Antes – Después" y "Antes – Durante" como también se evidencia en Figura 41.

Tabla 15. Prueba Dunn-Bonferroni para el contaminante PM2.5 en la ciudad de Cali.

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

5. CONCLUSIONES

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

En la implementación de los modelos de Aprendizaje Profundo se concluye que, para el conjunto de datos utilizados en el artículo, el modelo óptimo es el Seq2Seq, dado que arrojó predicciones ajustadas a los datos reales con porcentajes bajos de las métricas de validación RMSE, MAE y MSE con valores de 9.62, 0.88 y 6.86 para todas las ciudades en cuestión.

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

Por último, con el fin de apoyar a las autoridades de salud y ambientales a diseñar políticas que contribuyan con la disminución de estos contaminantes atmosféricos, y de esta manera, mejorar la calidad de vida, se plantean las siguientes propuestas:

- Incentivar el uso de bicicletas y transporte público cuando el nivel de los contaminantes de aproximen al límite establecido por el ICA, reduciendo sus tarifas, estableciendo convenios con empresas comprometidas con el medio ambiente para la generación de bonos de descuento o implementando alternativas de cambio de reciclaje por pasajes.

- Realizar talleres de formación y divulgación de educación ambiental a la comunidad.

- Regular estrictamente el parque automotor aplicando sanciones a los que no cumplen con las características ambientales para su movilización, impulsando el uso de transporte con energías renovables como el biocarburante.

- Mejorar la red de vigilancia de la calidad del aire y la detección de incendios forestales para mayor control y prevención.

- Establecer jornadas de plantación de especies arbóreas y arbustivas en los espacios urbanos.

- Fomentar del compostaje colectivo de bioresiduos y el reciclaje de los productos que no han terminado su ciclo de vida.

6. RECOMENDACIONES

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

Por otra parte, debido a que se actualizó la plataforma del IDEAM a partir de abril de 2023, donde se agrupan todos los datos de las estaciones de las diferentes ciudades, es conveniente utilizar mayor

cantidad de datos para tener mejores predicciones y escoger estratégicamente las variables.

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

7. REFERENCIAS

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

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

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

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

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

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

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

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

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

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

EPA. (26 de Mayo de 2021). *Particulate Matter (PM) Basics*. EPA. Obtenido de

https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM

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

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

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

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

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

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

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

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

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

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

IDEAM. (22 de 01 de 2021). *Proporción de datos del Índice de la Calidad del Aire por autoridad ambiental.* MinAmbiente y Desarrollo. Obtenido de http://www.ideam.gov.co/documents/11769/6

41368/2.01+HM+Indice+calidad+aire.pdf/5130ffb3-a1bf-4d23-a663-b4c51327cc05#:~:text=El%20%C3%8Dndice%20de%20calidad%20del,por%20parte%20de%20la%20poblaci%C3%B3n.

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

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

Mañas, A. M. (16 de junio de 2019). *Notas sobre pronóstico del flujo de tráfico en la ciudad de Madrid*. Bookdown.org. Obtenido de https://bookdown.org/amanas/traficomadrid/m%C3%A9todos-basados-en-deep-learning.html#lstm-univariado

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

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

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

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

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

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

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

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

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

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

Shatnawi, N., & Abu-Qdais, H. (2021). Assessing and predicting air quality in northern Jordan. *Air Quality, Atmosphere & Health*, 643-652.

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

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

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

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*. Bogotá: Ediciones Universidad Cooperativa de Colombia.

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

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