Prototipo de red neuronal artificial para el pronóstico de eventos críticos por partículas PM2.5 en el centro de la ciudad de Manizales.

Artificial neural network prototype for the forecast of critical events due to PM 2.5 particles in the center of the city of Manizales.

Estudiante:
David Eduardo García Correa
degarciac@libertadores.edu.co
Facultad de Ingeniería y Ciencias Básicas
Fundación Universitaria Los Libertadores
Bogotá D.C

Asesor:
Msc John González Veloza
jjgonzalezv02@libertadores.edu.co
Docente

Received: September, 2021

Resumen. El siguiente trabajo pretende desarrollar un modelo de pronóstico de calidad de aire en el centro de la ciudad de Manizales – Caldas, partiendo de un conjunto de datos más o menos procesado , se desea crear un modelo que permita predecir con éxito el comportamiento o valor que toma la variable PM2.5 a partir de 8 variables (Fecha, PM2.5, Temperatura, Velocidad del Viento, Precipitación, SO_2, O_3 y CO) y 603 datos por cada variable para un total de 264.299 datos, comprendidos entre los años 2019 y 2020, con el fin de generar medidas de tipo preventivo de enfermedades respiratorias que se diagnostiquen a partir de los altos índices de contaminación atmosférica. El procedimiento de pronóstico se realizó mediante la creación de una red neuronal artificial, para lo cual se subdividió el conjunto de datos en train y en test, en una subdivisión de 567 días consecutivos para entrenamiento de la red y los siguientes 30 días para su validación. Esta es una proporción que se eligió debido a que se está pronosticando un mes adelante por lo cual pareció conveniente. Por último, este trabajo presenta el uso de series de tiempos mediante técnicas de Machine Learning para la predicción de PM2.5, dicha herramienta y metodología no había sido utilizada hasta ahora con ese fin por parte de la Autoridad Ambiental – CORPOCALDAS. Los resultados obtenidos permiten tener confianza en las predicciones de PM2.5 con un error de 3%, con lo cual se puede tener la confianza suficiente para informar la comunidad sobre los cambios en los niveles de la calidad del aire y concentración de PM2.5 en un futuro cercano.

Palabras clave: Machine Learning, Redes Neuronales Artificiales, Series de Tiempo, ForestCasting, Feedforward, backpropagation, Embeddings.

ABSTRACT. The following work aims to develop an air quality forecast model in the center of the city of Manizales - Caldas, starting from a more or less processed data set, it is desired to create a model that allows to successfully predict the behavior or value that takes the variable PM2.5 from 8 variables (Date, PM2.5, Temperature, Wind Speed, Precipitation, SO2, O3 and CO) and 603 data for each variable for a total of 264,299 data, between the years 2019 and 2020, in order to generate preventive measures for respiratory diseases that are diagnosed based on high levels of air pollution. The forecasting procedure was carried out by creating an artificial neural network, for which the data set was subdivided into train and test, in a subdivision of 567 consecutive days for network training and the next 30 days for its validation. . This is a proportion that was chosen because it is forecasting a month ahead, which is why it seemed convenient. Finally, this work presents the use of time series using Machine Learning techniques for the prediction of PM2.5, this tool and methodology had not been used until now for that purpose by the Environmental Authority - CORPOCALDAS. The results obtained allow to have confidence in the predictions of PM2.5 with an error of 3%, with which it is possible to have sufficient confidence to inform the community about the changes in the levels of air quality and concentration of PM2.5 in the near future.

Keywords: Machine Learning, Artificial Neural Networks, Time Series, ForetCasting, Redes neuronales, Keras, Tensorflow, Embeddings.

1. INTRODUCCIÓN

Según información reportada por la OMS, cada año se reportan un total de 3,8 millones de muertes prematuras debido a enfermedades no transmisibles, particularmente accidentes cerebro vasculares, cardiopatía isquémica, neumopatía al aire de interiores contaminados. A nivel mundial más del 50% de las muertes por neumonía en menores de cinco años son generadas por material particulado inhalado del aire. La contaminación atmosférica se visualizó como un problema para los científicos al presentarse eventos como los de Meuse Valley en 1930, donde murieron más de 60 personas por emisiones de SO_2 y fluorocarbonados; el de Donora Pensylvania en 1948, dando muerte a más de 20 personas por emisiones de material particulado, y el más importante, en Londres en 1952 con la muerte de más de 4.000 personas también por presencia de partículas en exceso en el ambiente. Esto dio la alerta para tomar medidas radicales a nivel mundial en términos políticos y científicos (De Nevers, 1998).

En Colombia, el 74% de la población a identificado la contaminación del aire como uno de los problemas medioambientales y de salud pública más trascendentales en el país, ya que afecta directamente a la población, generando 7000 casos de muertes prematuras anuales, 7400 casos de bronquitis crónica, 13000 hospitalizaciones por causas de enfermedades respiratorias agudas y 255000 visitas a salas de urgencia (Larner, 2004).

El siguiente trabajo pretende desarrollar un modelo de pronóstico de calidad de aire en el centro de la ciudad de Manizales – Caldas, partiendo de un conjunto de datos más o menos procesado , se desea crear un modelo que permita predecir con éxito el comportamiento o valor que toman las partículas inferiores a 2,5 micras (variable PM2.5) a partir de un cierto grupo de observaciones, con el fin de generar medidas de tipo preventivo de enfermedades respiratorias que se diagnostiquen a partir de los altos índices de contaminación atmosférica. Este tema innovador para la ciudad de

Manizales, en la cual la industria y el alto flujo vehicular en el centro de la ciudad, causan un incremento de las partículas inferiores a 2,5 micras, lo que repercute en el padecimiento de enfermedades como rinitis alérgica, asma y otras enfermedades clasificadas como enfermedad respiratoria aguda (IRA). Generalmente los habitantes de los centros poblados desconocen la calidad del aire que respiran, así como las medidas de tipo preventivo que pueden ser tomadas en cuenta para disminuir el deterioro de la salud por baja calidad del aire. El set de datos utilizado para el desarrollo de este trabajo de investigación, contiene información sobre la concentración de material particulado PM 2.5 y variables climatológicas como son: Temperatura, Velocidad del Viento, Precipitación, y otras variables de calidad del aire como SO_2, O_3 y CO. La base de datos presentaba valores ausentes, por lo cual se utilizados datos capturados de equipos cercanos a la estación de la Gobernación de Caldas, como son la estación del aeropuerto La Nubia, La central hidroeléctrica de Caldas – CHEC, Aguas de Manizales y el Sistema de Vigilancia de Calidad del Aire (SVCA) de Corpocaldas, información que es recopilada por parte del Instituto de Estudios Ambientales IDEA de la Universidad Nacional sede Manizales (Cdiac).

2. REFERENTES TEORICOS

Se han propuesto diversos métodos de Inteligencia Artificial para la predicción de partículas contaminantes PM10 y PM2.5, alguno de los cuales se han utilizado para la predicción de la contaminación del aire por parte de autoridades ambientales.

A continuación, se presentan algunos antecedentes teóricos de las investigaciones realizadas en el tema:

La contaminación del aire es uno de los problemas más graves de la sociedad moderna, enmarcados de manera significante por el desarrollo urbanístico e industrial de la sociedad. Los efectos que puede tener la contaminación en la salud humana han sido tema de un intenso estudio; dado, su asociación con el aumento en la mortalidad y en enfermedades respiratorias y cardiovasculares (Bert, B. and Stephen, H.T, 1999). Estos efectos pueden ser vistos a corto y largo plazo, con distintas repercusiones y afectaciones en cualquier etapa de la vida, desde la concepción hasta la vejez (Schneider, A., et al., 2011).

Los modelos de RNA funcionan mejor cuando la información disponible corresponde a las distintas variables relacionadas con el fenómeno en cuestión, son más larga y completa. En el caso de la predicción de eventos críticos de contaminación urbana, son importantes tanto las variables meteorológicas como los registros de los contaminantes. Por su sencillez, los modelos estadísticos pueden proveer una primera aproximación al pronóstico de eventos críticos de contaminación atmosférica debida a diferentes contaminantes [material particulado, dióxido de azufre (SO_2), dióxido de nitrógeno (NO_2), monóxido de carbono (CO) y ozono (O_3)], y por esta vía mitigar su efecto (Guevara et al., 2019).

Al mismo tiempo, (Guevara et al., 2019) determinan que el modelo, identifica dentro del periodo de validación los posibles eventos de contaminación crítica, pero tiene una margen de error entre los valores pronosticados y los valores observados. Aunque presentó una eficiencia del 70 % una falencia de este modelo es que basado en los días anteriores, no ha tenido mucha efectividad en pronosticar el primer día de eventos críticos.

En Colombia, por su parte se han llevado a cabo investigaciones importantes en la materia, cuyo objeto ha sido determinar la relación entre la concentración del material particulado PM10 atmosférico respecto a la carga de material en resuspensión PM10 en algunas vías de Bogotá. Para el desarrollo de este trabajo se propuso un modelo ARIMA de correlaciones, identificando los parámetros óptimos de ajuste, con el fin de relacionar los datos de concentraciones de material particulado PM10 atmosférico que se obtuvieron durante los muestreos realizados, con respecto a la cantidad de material en resuspensión, este trabajo permitió concluir que los valores de concentración más altos para PM10 se presentaron durante el periodo 2002-2007 con valores que exceden los 100 $\mu g/m^3$, siendo esta la concentración máxima establecida por el Ministerio de Ambiente y Desarrollo Sostenible (MADS) para un tiempo de exposición 24 horas, principalmente en las zonas Occidental y Suroccidental de la ciudad de Bogotá; mientras que, del año 2008 al año 2012, los valores de concentración reportaron una disminución gradual con el paso del tiempo siendo el año 2012 el periodo en el cual se registraron las menores concentraciones de material particulado existentes hasta la fecha con promedios máximos de 80 $\mu g/m^3$; niveles de inmisión que pertenecían a las estaciones ubicadas en la zona Suroccidental de la ciudad (Castañeda y Méndez, 2018).

(Salazar et al., 2018) propusieron un caso interesante con el fin de desarrollar y probar un modelo de redes neuronales artificiales (RNA) para pronosticar la concentración diaria del material particulado menor a 2.5 micras (PM2.5) en el Valle de Aburrá (Colombia), con un día de anticipación, a partir de información de tres estaciones de la Red de Monitoreo de Calidad del Aire del Área Metropolitana. Mediante este trabajo se logró la previsión de las concentraciones diarias de PM2.5, teniendo en cuenta información de variables meteorológicas y de calidad del aire para el Valle de Aburrá en escala diaria. Las variables de entrada se definen con base en análisis de correlación. Sus señales son transformadas con las funciones de activación que no suavizaran excesivamente las salidas del modelo. La precisión de la predicción del número de días con eventos de contaminación critica, es relativamente buena, pero sin embargo no se descarta la posibilidad de que el modelo pueda mejorar si se toman medidas estadísticas más estrictas de descarte. Ejemplo es. PM2.5 T-2 variable que el modelo omitió, probablemente por su alta relación con el PM2.5. Los modelos de RNA funcionan mejor cuando la información disponible corresponde a las distintas variables relacionadas con el fenómeno en cuestión, son más largas y completas. En el caso de la predicción de eventos críticos de contaminación urbana, son importantes tanto las variables meteorológicas como los registros de los contaminantes. Por su sencillez, los modelos estadísticos pueden proveer una primera aproximación al pronóstico de eventos críticos de contaminación atmosférica debida a diferentes contaminantes [material particulado, dióxido de azufre (SO_2), dióxido de nitrógeno (NO_2), monóxido de carbono (CO) y ozono (O_3)], y por esta vía mitigar su efecto. El modelo, identifica dentro del periodo de validación los posibles eventos de contaminación crítica, pero tiene una margen de error entre los valores pronosticados y los valores observados. Aunque presentó una eficiencia del 70% una falencia de este modelo es que basado en los días anteriores, no ha tenido mucha efectividad en pronosticar el primer día de eventos críticos.

(González, C., et al., 2018), utilizaron coeficientes de correlación de Pearson para determinar las relaciones entre PM10 y variables meteorológicas mediante regresión simple modelo en la ciudad de Manizales. El análisis de varianza (ANOVA) fue aplicado para determinar los niveles de confianza entre estas variables. Baja significativa diferencia (LSD). Se utilizó la prueba de Fisher para estimar diferencias entre las concentraciones medias de PM10 para períodos húmedos versus secos.

Distribución estacional de las concentraciones de PM10. Como resultado se obtuvo que el promedio más alto de PM10 se asoció con alto tráfico urbano y alta densidad de transporte público en la estación Liceo del centro (44 µg m-3). En términos de las estaciones de HVS, el promedio de PM10 en el centro Liceo fue un 75% más alto que las otras tres estaciones combinadas (n = 468). En estudios anteriores, la combustión de diesel y gasolina se informó como principales fuentes de emisiones en el centro de la ciudad. Entre las otras estaciones, se observó poca diferencia, con promedios de PM10 que van desde 24 µg m3 (Gobernación) a 26 µg m3 (Palogrande y Nubia).

(Chiarvetto, L., et al., 2008), experimentaron tres aspectos en particular del diseño de una red neuronal: la normalización de los datos, la selección de la arquitectura y la selección de la función de activación. En base a nueve variables de entrada: dos estacionales, y siete meteorológicas; se determinó que la mejor candidata es una red natural compuesta por: una capa de entrada lineal de nueve neuronas artificiales (NA), una capa oculta de catorce NA y una capa de salida de una NA; ambas con una función de activación tangente hiperbólica. Mediante este trabajo se determinan los siguientes resultados: Si se hubiese construido una RN para cada combinación de todas las posibles decisiones de diseño, se hubiesen construido 132 RNs (seis opciones de normalización, once opciones en la cardinalidad de la capa oculta, dos funciones de normalización y tres algoritmos de aprendizaje). En lugar de esto, solo 22 prototipos fueron construidos y una RN. De esta forma se reduce los costos de desarrollo de la red neuronal.

(Ramirez y Londoño, 2018), realizaron la investigación que permitiera la caracterización espacial de la concentración de PM2.5 en la ciudad de Medellín entre los años 2016 y 2018, mediante la implementación de una metodología de modelos de interpolación espacial. Para lo cual utilizaron el algoritmo de regresión espacial permite realizar estimaciones locales de una variable de estudio Z en las coordenadas (xi, yi) mediante la implementación del método de mínimos cuadrados ordinarios. (Londoño Ciro, 2018). Luego de revisada la dependencia de las variables mediante el coeficiente de correlación (Londoño y Cañon, 2018), se procedió a usar el algoritmo de regresión geográficamente ponderado (GWR). Los resultados obtenidos mediante los datos medidos de PM2.5 para los años 2016, 2017 y 2018, permitieron identificar que, para el primer cuatrimestre del año, es decir los meses enero, febrero, marzo y abril, las mediciones de PM2.5 superan el promedio (27,69 µg/m^3) de los datos mensuales y llegan a superar los rangos máximos aceptables (37 µg/m^3) según la resolución 2254 del 1 de noviembre 2017 (Ministerio de Ambiente y Desarrollo Sostenible).

En el año 2020, (Caso, M. 2020), realizó un estudio sobre la predicción de la calidad del aire de la ciudad de Madrid mediante técnicas de Machine Learning, para lo cual selecciono modelos de aprendizaje profundo, redes neuronales artificiales, redes neuronales recurrentes, redes neuronales LSTM, LSTM vanilla, LSTM bidireccionales y apiladas, Redes GRU simple y apilada. Tras realizar un exhaustivo examen a los datos conjugando diferentes modelos de minería de datos para diferentes configuraciones y conjunto de variables predictoras, se pudo determinar que, a pesar de haber utilizado modelos más complejos de redes neuronales recurrentes, como los de capas apiladas LSTM o CRU, se ha obtenido un mejor resultado con modelos más simples basados en una única capa con neuronal LSTM.

(Pedraza, J. 2018), realizó un trabajo de investigación mediante la construcción de un modelo de machine learning para la predicción de partículas de contaminación atmosférica finas, en la localidad de Kennedy en Bogotá, mediante el cual se concluyó que como resultado de esta investigación se puede observar el uso de redes neuronales artificiales modeladas en un prototipo

el cual permite la predicción del contaminante PM 2.5, con los datos de la localidad de Kennedy en la ciudad de Bogotá. Tras realizar el prototipo de la red neuronal aplicada para predecir el contaminante PM 2.5, se pueden comprobar las posibilidades sobre las que se puede trabajar al aplicar métodos de machine learning en los distintos aspectos que nos afectan como seres humanos, también se pudo observar la implementación dentro de un prototipo y los resultados obtenidos con este.

(Roncancio y Trujillo., 2019), crearon un modelo de predicción de material particulado de 10µm en la ciudad de Bogotá, estudio del caso y posibles mejoras, como resultado de este estudio se concluyó que finalmente, de los resultados hallados es posible concluir que a pesar de que las características más relevantes para el modelo provienen de la red de monitoreo, utilizar únicamente la información recopilada por la RMCAB no es suficiente para que un modelo de machine learning aprenda las dinámicas del PM10 en toda la ciudad. Sin embargo, usar esta información en conjunto con bases de datos de otras entidades como el IDEAM o el sistema de transporte público, permite a estos algoritmos hallar relaciones más complejas y útiles. No obstante, el mejor modelo hallado utilizando redes neuronales completamente conectadas, y poco profundas, no logró capturar más del 40% de la varianza del modelo real. Para futuros modelos es deseable aplicar nuevas arquitecturas de redes neuronales, como las redes neuronales recurrentes (RNN) o las redes neuronales convolucionales que han demostrado lograr mejores resultados en este tipo de problemáticas. De igual forma, es de gran interés evaluar el desempeño de otros modelos, como los árboles de decisión o técnicas de estimación por máxima verosimilitud.

(Fajardo, S. M., 2020), establecido un modelo de pronóstico de la calidad de aire respecto al material particulado en Bogotá, por medio de minería de datos, para ello realizó un análisis de Componentes Principales (PCA), técnica que fue implementada en cada una de las estaciones de monitoreo. Posteriormente se realizó un modelo estadístico mediante Redes Neuronales Artificiales (RNA) mediante técnicas de Backpropagation.

(Vásquez, D., et al., 2017), desarrollaron un modelo de Red neuronal Backpropagation para la predicción de datos de contaminación y prevención de ataques a personas con padecimientos de rinitis alérgica y asma. El diseño especifico de este trabajo de investigación se fundamenta en cinco contaminantes ($PM_{10,}$ O_3, $NO_{2,}$ SO_2 Y CO). Los autores seleccionaron 120 datos por cada contaminante correspondiente, a partir de los cuales entrenaron 85 datos y 35 datos para prueba, con lo cual los autores concluyeron que el comportamiento del error generado en el modelo es bueno para la predicción de $PM_{10,}$ $CO2,$ SO_2 y O_3, determinando así que las redes neuronales tienen un buen desempeño a la hora de realizar modelos predictivos.

3. METODOLOGÍA

La metodología sistemática utilizada para el desarrollo del modelo de pronóstico de concentración de PM2.5, se basa en Series Temporales con Redes Neuronales - Forecasting, ya que este método permite trabajar con un conjunto de muestras tomadas a intervalos de tiempo regulares, donde el interés en el predictor seleccionado surge de la necesidad de determinar con anticipación la calidad del aire en el centro de la ciudad de Manizales, la predicción de PM2.5 constituye la fase inicial de proyectos futuros de gran magnitud.

Las Redes Neuronales han mostrado ser un método eficiente y universal en la aproximación de funciones para cualquier tipo de dato (Lek y Guégan, 1999). Son de especial utilidad cuando dicha función es desconocida, son capaces de resolver complejos patrones entre la fuente de emisión y la concentración (Gardner y Dorling, 1999). Por otro lado, han mostrado ser superiores en predicción de calidad de aire en comparación a métodos tradicionalmente estadísticos (Grivas y Chaloukou, 2006). La metodología Forecasting, consiste en estimar y prever la demanda futura de un producto o servicio, la cual es aplicable a los modelos de predicción de contaminantes atmosféricos, para lo cual se utilizarán los históricos de concentración de PM2.5 en la ciudad de Manizales.

Singh y Kumar (2010) dieron a conocer un método que permite diseñar pronósticos mediante el uso de redes neuronales para series de tiempo, el cual puede ser usado en el pronóstico de contaminantes para determinado periodo de tiempo. Los pronósticos generados con la red neuronal pueden ser comparados con la técnica de regresión múltiple. La capacidad predictiva de ambos modelos se evalúa mediante técnicas de medición del error (MAPE, MSE y RMSE). Los resultados permiten determinar que las redes neuronales entrenadas con suficientes datos de entradas y seleccionando una topología adecuada se obtendrán predicciones de la concentración de PM2.5 con mayor precisión que la técnica estadística. Un modelo de red neuronal multicapa (RNM) con p nodos de entradas, h nodos en la capa oculta y un nodo de salida poder ser usado para analizar una serie de tiempo. Una serie de tiempo y_t ; t = 1,2,3,….,n se puede moldear con una RNM con p variables de entradas y_{t-i} = (i = 1,2,….,p) y relacionarla con una variable de salida y_t. El modelo RNM para el análisis de series de tiempo se indica mediante la siguiente expresión:

$$y_t = \alpha_0 + \sum_{j=1}^{P} \alpha_1 G\left(\beta_{oj} + \sum_{i=1}^{h} \beta_{ij}\, y_{t-i}\right) + e_t \qquad (1)$$

Donde:

y_t: Valor observado de la serie de tiempo en el instante t. Representa el valor para la variable de salida de la RNM,
y_{t-i}: Son los valores rezagados de la serie de tiempo en el instante t. Representan los valores de las p variables de entrada en la RNM,
α_j: Representan los pesos de la capa oculta a la capa de salida. Siendo el respectivo sesgo,
β_{ij}: Son coeficientes que representan los pesos de la capa de entrada a la capa oculta. Siendo el respectivo sesgo,
G: Representa la función de activación o transferencia de la capa de entrada, la cual determina la salida de la capa oculta. Las funciones de activación que generalmente son usadas son la función logística para la capa de entrada y la función identidad para la capa de salida,
h: Número de neuronas en la capa oculta,
p: Número de neuronas en la capa de entrada, Su valor determina el número de rezagos con que se analizará la serie de tiempo,
e_t: Son los errores aleatorios del modelo, los cuales se asumen que son independientes e idénticamente distribuidos con media cero y varianza constante.

En este caso la función está determinada por un conjunto de variables meteorológicas, estacionales y de concentraciones de otros contaminantes, para lo cual se utilizó una base de datos de 8 variables y 603 datos para cada variable, con un total de 264.229 datos. Para esto es importante contar con

un entorno de ejecución de Python, en este caso se utilizó Google colab, el cual permite ejecutar y programar en Python en tu navegador con las siguientes ventajas:

- No requiere configuración,
- Da acceso gratuito a GPUs, y
- Permite compartir contenido fácilmente.

Como primera fase se procedió a cargar la librería Pandas, especializada en el manejo y análisis de estructuras de datos. Posteriormente se carga la base de datos en este caso en Excel:

Figura 8
Base de datos utilizada para el trabajo de investigación

```
#https://drive.google.com/file/d/1dmRx_m-MnB9FSVb0Q9LYul9KPJEJlV3R/view?usp=sharing
url = 'https://drive.google.com/uc?id=1dmRx_m-MnB9FSVb0Q9LYul9KPJEJlV3R'
Data = pd.read_excel(url)
Data
```

	Fecha	PM2_5	Temperatura	Velocidad_Viento	Precipitación	SO2	O3	CO
0	2019-01-01	14.0	19.5	2.1	14.2	5.21	13.36	865.83
1	2019-01-02	12.0	18.5	1.9	0.0	4.59	10.35	591.11
2	2019-01-03	8.0	17.1	2.2	8.9	5.12	14.10	305.53
3	2019-01-04	16.0	17.6	0.7	0.0	5.23	14.92	672.64
4	2019-01-05	10.0	18.9	3.3	0.0	4.86	17.78	357.10
...	...	...	...	...	...	...	...	...
598	2020-08-21	11.5	6.9	1.0	0.4	1.64	15.09	395.27
599	2020-08-22	10.6	3.6	1.1	0.2	1.45	14.81	567.00
600	2020-08-23	8.0	9.4	1.6	1.6	18.30	15.07	851.20
601	2020-08-24	8.2	8.9	1.4	2.0	25.85	10.52	544.68
602	2020-08-25	8.9	7.5	1.7	0.0	18.52	15.07	629.35

603 rows × 8 columns

Una vez, se cargó la base de datos se procedió a realizar la visualización y el análisis descriptivo de las variables de interés. Primero se visualizó que las variables no presentaran observaciones ausentes y se verifico que el dataframe que se utilizó con pandas tiene como índice en la primera columna el año y el día, lo que permite hacer filtrados directamente y algunas operaciones especiales. Para el análisis de patrones de características individuales mediante visualización, se instaló la librería Seaborn y los paquetes "Matplotlib" y "Seaborn". Como se trabajó con variables numéricas continuas que pueden contener cualquier valor dentro de cierto rango y al mismo tiempo pueden tener el tipo "int64" o "float64", se utilizaron diagramas de dispersión con líneas ajustadas con el fin de visualizar este tipo de variables. Para comenzar a comprender la relación (lineal) entre una variable individual y el PM2.5. Este paso se llevó a cabo usando "regplot", que traza el diagrama de dispersión más la línea de regresión ajustada para los datos. A partir de allí, se evidencio que las variables climatológicas no parecen ser buenos predictores de la concentración de PM2.5, las líneas de regresión están cerca del eje horizontal. Además, los puntos están muy dispersos y lejos de la línea ajustada y dan como resultado una regresión lineal débil. Variables categóricas Estas son variables que describen una 'característica' de una unidad de datos y se seleccionan de un pequeño grupo de categorías. Las variables categóricas pueden tener el tipo "objeto" o "int64". Una buena forma de visualizar variables categóricas es mediante el uso de diagramas de caja.

Figura 9
Diagrama de dispersión para PM2.5 y SO2

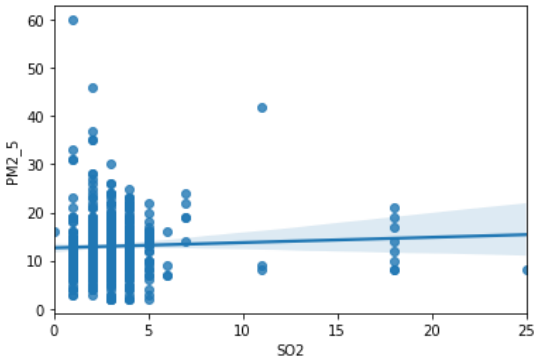


Figura 10
Diagrama de dispersión PM2.5 y CO

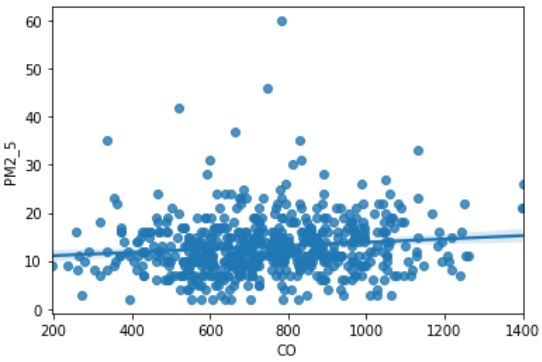
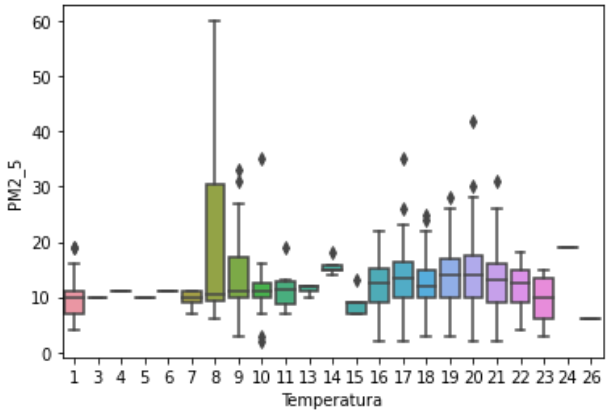


Figura 11
Diagrama de caja PM2.5 y Temperatura



De acuerdo a la figura 11, podemos observar que las distribuciones de PM2.5 entre los diferentes grados de temperatura tiene una superposición significativa, por lo que aparentemente no se evidencia cambio en la mediana del PM2.5 para diferentes rangos de temperatura.

Una vez, realizado la visualización descriptiva de los datos, se procedió a realizar la segunda fase correspondiente a un primer pronóstico de series temporales univariable con redes neuronales en Python, como se observó una baja correlación entre las variables se decidió realizar el primer modelo de pronóstico utilizando solo la variable de interés PM2.5, tal y como se observa en la figura 12:

Figura 12
Variable de interés PM2.5 en µg/m^3

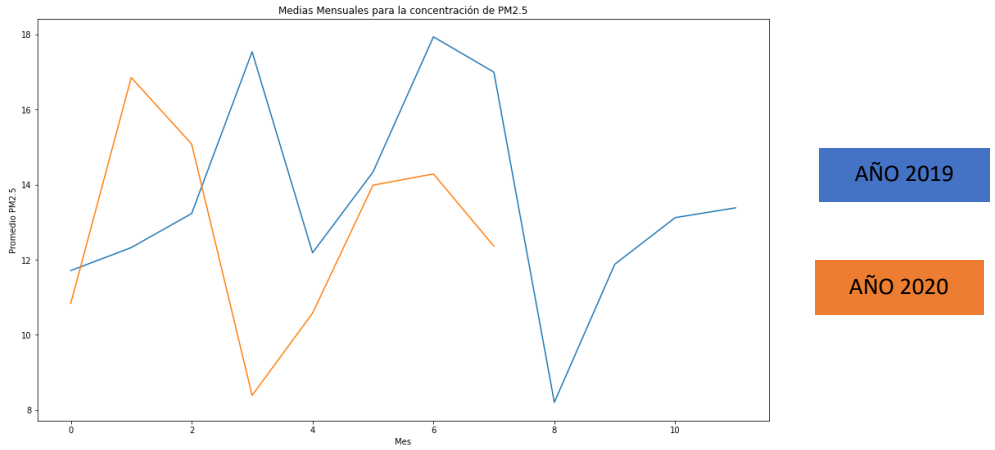
```
[ ] Variable_PM2_5 = df[['PM2_5']]
    Variable_PM2_5
```

Fecha	PM2_5
2019-01-01	14.0
2019-01-02	12.0
2019-01-03	8.0
2019-01-04	16.0
2019-01-05	10.0
...	...

En total se tenían 603 registros, la media de la concentración de PM2.5 es de 13.26 y una desviación estándar de 5.97, es decir por lo general estaremos entre 19.23 y 7.29.

A continuación, se visualizaron esas medias mensuales para la concentración de PM2.5:

Figura 13
Medias mensuales para la concentración de PM2.5 año 2019 y 2020



Vemos que en 2019 (en azul) tenemos un inicio de año con un ascenso en la concentración de PM2.5, luego comienza a subir de manera exponencial hasta la llegada del periodo escolar y de universidades en donde en los meses marzo, abril se tiene el mayor flujo vehicular. Para los meses de junio y julio se presenta un descenso vertiginoso el cual puede deberse al periodo de vacaciones donde se disminuye el tránsito vehicular en la ciudad de Manizales. Finalmente vuelve a disminuir y tiene un pequeño pico entre agosto y octubre. También vemos que 2020 (naranja) se comporta diferente del año inmediatamente anterior. Para ese año se observa un incremento vertiginoso de la concentración de PM2.5 en los primeros meses del año, pero desciende rápidamente debido posiblemente a la pandemia generada por el Covid-19 y al confinamiento decretado por el Gobierno

Nacional y Local. No se cuenta con datos para la concentración de PM2.5 para los últimos meses de 2020.

Una vez confirmado que la serie es estacionaria, se procede a realizar el pronóstico para lo cual existen diversos métodos. En nuestro caso, las concentraciones de PM2.5 parecen comportarse similares en algunos periodos del año, pero presenta diferencias muy marcadas para otros periodos, para lo cual un método sencillo si por ejemplo quisiéramos proveer la concentración de PM2.5 que se tiene para un periodo determinado, sería decir "Si en noviembre de 2019 la concentración promedio de PM2.5 es de 13.38µg/m3, pronóstico que en diciembre será similar". En esta oportunidad utilizaremos Machine Learning: una red neuronal para hacer el pronóstico. Esta red es relativamente sencilla de crear, y seguramente en algún momento se estará utilizando un modelo más moderno para hacer el pronóstico. Para este momento se utilizó una arquitectura sencilla de red neuronal FeedForward (también llamada MLP por sus siglas Multi-Layered Perceptron), con pocas neuronas y como método de activación tangente hiperbólica pues entregaremos valores transformados entre -1 y 16,7,8.

La tercera fase corresponde a la preparación de los datos: Este puede que sea uno de las etapas más importantes de este ejercicio. Lo que se realiza es alterar el flujo de entrada del archivo en nuestro caso en Excel que contiene una columna con las unidades despachadas, y lo convertiremos en varias columnas. ¿Y por qué hacer esto?, en realidad, lo que se llevó a cabo es tomar la serie temporal y convertirla en un "problema de tipo supervisado" para poder alimentar nuestra red neuronal y poder así entrenarla con backpropagation ("como es habitual"). Para hacerlo, se debe tener unas entradas y unas salidas para entrenar al modelo. Lo que haremos en este modelamiento, es tomar los 7 días previos para "obtener" el octavo.

Entradas: serán "7 columnas" que representan las concentraciones de PM2.5µg/m^3 de los 7 días anteriores.

Salida: El valor del "8vo día". Es decir, la concentración de PM2.5 µg/m^3 de ese día.

Para hacer esta transformación se utilizó una función llamada series_to_supervised(). (La verás en el código, a continuación). Antes de usar la función, se utilizó el MinMaxScaler para transformar el rango de nuestros valores entre -1 y 1.

Figura 14
Resultados generados con la función series_to_supervied()

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
7	-0.590444	-0.658703	-0.795222	-0.522184	-0.726962	-0.556314	-0.829351	-0.317406
8	-0.658703	-0.795222	-0.522184	-0.726962	-0.556314	-0.829351	-0.317406	-0.522184
9	-0.795222	-0.522184	-0.726962	-0.556314	-0.829351	-0.317406	-0.522184	-0.829351
10	-0.522184	-0.726962	-0.556314	-0.829351	-0.317406	-0.522184	-0.829351	-0.556314
11	-0.726962	-0.556314	-0.829351	-0.317406	-0.522184	-0.829351	-0.556314	-0.726962

Se utilizó como entradas las columnas encabezadas como var1(t-7) a (t-1) y nuestra salida (lo que sería el valor "Y" de la función) será el var1(t) -la última columna-.

En la cuarta fase se creó la red neuronal artificial, para lo cual se subdividió el conjunto de datos en train y en test. Algo importante de este procedimiento, a diferencia de otros problemas en los que podemos "mezclar" los datos de entrada, es que en este caso es importante mantener el orden en

el que alimentaremos la red. Para el desarrollo del ejercicio se trabajó con una subdivisión de los primeros 567 días consecutivos para entrenamiento de la red y los siguientes 30 para su validación. Esta es una proporción que se eligió debido a que se está pronosticando un mes adelante por lo cual pareció conveniente, pero definitivamente, puede no ser la óptima (queda propuesto, variar esta proporción por ejemplo a 80-20 y comparar resultados). Una vez se transformó la entrada en un arreglo con forma (567,1,7) esto al castellano significa algo así como "567 entradas con vectores de 1×7", se desarrolló la arquitectura de la red neuronal de la siguiente forma:

- Entrada 7 inputs, como se determinó en el parágrafo anterior,
- 1 capa oculta con 7 neuronas,
- La salida será 1 sola neurona,
- Como función de activación utilizamos tangente hiperbólica puesto que utilizaremos valores entre -1 y 1,
- Utilizaremos como optimizador Adam y métrica de pérdida (loss) Mean Absolute Error,
- Como la predicción será un valor continuo y no discreto, para calcular el Accuracy utilizaremos Mean Squared Error y para saber si mejora con el entrenamiento se debería ir reduciendo con las épocas (EPOCHS).

Hasta ahora se ha explicado el referente teórico que se utilizó para crear la red neuronal MLP (Perceptrón multicapa) feedforward de pocas capas. A partir de esta etapa se utilizaron metodologías que permiten mejorar el modelo de series temporales, a partir de dos (2) modelos con redes neuronales Feedforward para intentar mejorar los pronósticos de emisión de material particulado PM2.5:

- Un primer modelo tomando la fecha como nueva variable de entrada valiosa y que aporta datos,
- Un segundo modelo también usando la fecha como variable adicional, pero utilizándola con variables dummy.

Primer Mejora: Serie Temporal de múltilples Variables. Un clásico ejemplo para entender lo que son las Series Temporales de Múltiples Variables sea el pronóstico del tiempo, en donde se tienen varias "columnas de entrada" con la temperatura, la presión atmosférica, humedad. Con esas tres variables se tendrá una mejor predicción de "la temperatura de mañana" que si tan sólo usásemos una sola feature.

Segunda mejora: Se trabajo mediante "One hot encoding" o "Transformación de variables dummy" en variables categóricas. Para el segundo modelo, se variables dummy en las variables categóricas, es decir, en la columna de día y de mes. Los valores de día van del 0 al 6 representando los días de la semana. Pero no quiere decir que el día 6 "vale" más que el día 0. Son identificadores. No tendría sentido decir que jueves es mayor que domingo. Sin embargo, la red neuronal esto no lo sabe y podría interpretar erróneamente esos valores (categóricos). Con los meses lo mismo; van del 1 al 12 pero no quiere decir que "diciembre valga más que agosto". Para intentar resolver esta problemática, es que aparecen las transformaciones de variables dummy.

4. RESULTADOS

A lo largo del desarrollo del trabajo de investigación, se creó una red neuronal perceptrón multicapa (MLP) feedforward de pocas capas, trabajando principalmente en los datos de entrada. El archivo Excel que se cargó contaba con dos (2) columnas: Fecha y PM2_5, las cuales se transformaron en un "problema de aprendizaje supervisado". Para lo cual se generó un archivo nuevo de 7 columnas de entrada en la cual se puso la concentración de material particulado PM2.5 en los 7 días anteriores y de salida la concentración de PM2.5 en la fecha actual. A partir de allí, se alimentó la red y ésta pudo realizar los pronósticos aceptables. Sólo se utilizó la columna de PM2_5, no se utilizó la columna Fecha quedando la pregunta si esta columna podría ser un dato importante para la predicción de la concentración de PM2.5. Para mejorar el modelo de series temporales, se propusieron dos (2) nuevos modelos con redes neuronales feedforward con el cual se buscó mejorar el pronóstico de concentración de PM2.5.

- Se realizó un primer modelo incluyendo la variable Fecha como nueva variable de entrada,
- Un segundo modelo usando nuevamente la variable Fecha como variable adicional, pero utilizándola con variables dummy, con el fin de ver si se genera mejoría en el modelo.

A continuación, se comparan los resultados de los tres (3) modelos:

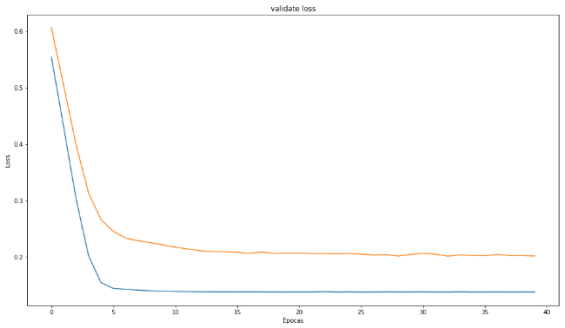
Tabla 1
Valores finales de las métricas tras las 40 épocas (EPOCHS):
ST1: Series de Tiempo Univariada,
STMV: Series de Tiempo con Múltiples Variables, y
STE: Serie de Tiempo con One hot encoding.

Modelo	loss	val_loss	MSE	val_MSE
ST1	0.1197	0.2066	0.0268	0.177
STMV	0.1338	0.2045	0.0324	0.1538
STE	0.1384	0.2062	0.0396	0.1651

Como se puede observar en la tabla 1, los modelos STMV y STE quedan prácticamente iguales, es decir que tras realizar el ejercicio con múltiples variables y agregando variables dummy no pareciera haber mejoría significativa en las métricas. Por su parte el modelo ST1, presenta un resultado de MSE de 0.0268 siendo inferior con respecto a los resultados generados en los modelos STMV y STE.

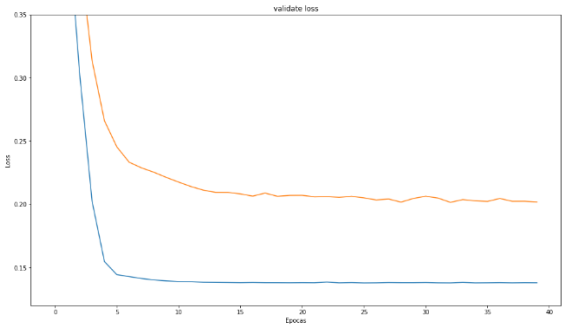
Si comparamos las siguientes gráficas de pérdida (loss), se puede observar que las métricas de los en los sets de entrenamiento descienden y se mantienen estables, aun así, en el modelo STMV la curva de validación es algo errática. Las curvas de los modelos 1 y 3 se mantienen sobre el 0.2, mientras que la curva del modelo 2 tiende a oscilar. Mediante las figuras 15, 16 y 17, se puede concluir que el modelo está aprendiendo y no memorizando, con lo cual se evita el sobreajuste de parámetros.

Figura 15
Modelo 1. ST1: En azul el entrenamiento y naranja el set de validación



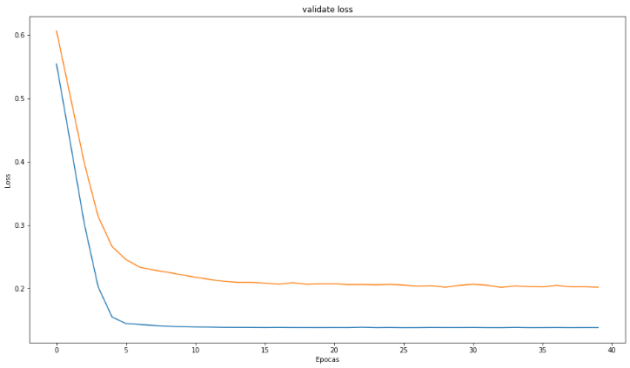
Entrenamiento Validación

Figura 16
Modelo 2. STMV: En azul el entrenamiento y naranja el set de validación



Entrenamiento Validación

Figura 17
Modelo 3. STE: En azul el entrenamiento y naranja el set de validación



Entrenamiento Validación

Ahora, se comparan los pronósticos y sus aciertos para cada uno de los modelos propuestos:

Figura 18
Comparación Modelo 1. ST1:
Con aciertos, pero pronóstico conservador

Figura 19
Comparación Modelo 2. STMV:
Presenta mayor mejoría en el pronóstico ST1

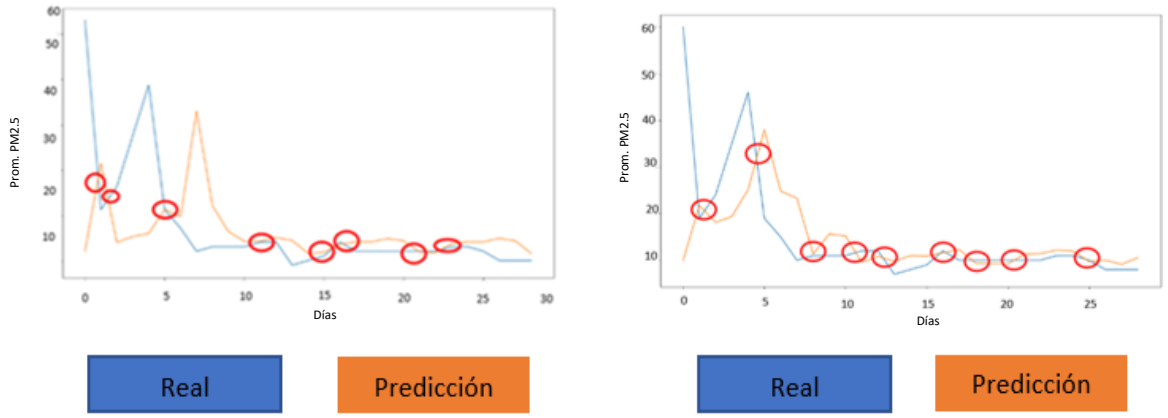
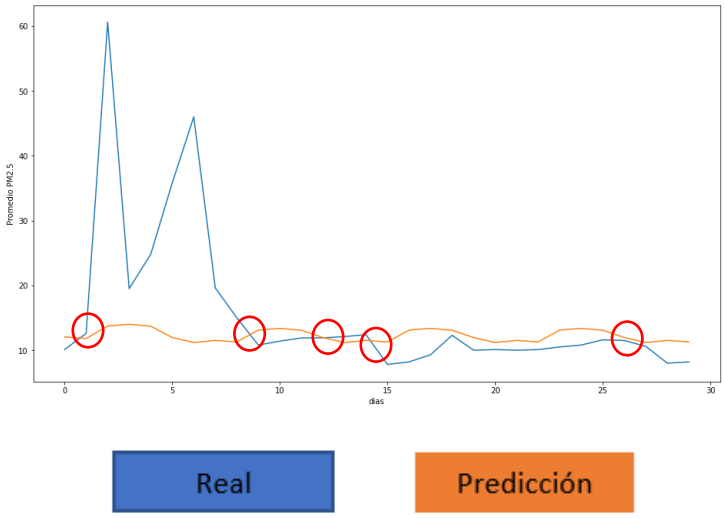


Figura 20
Comparación Modelo 3. STE:
Los “One hot encoding” disminuyen los aciertos a la curva de pronóstico



5. CONCLUSIONES

Se puede observar que la red neuronal para el primer modelo ST1 es más conservadora, manteniéndose la media de 11.45, con un pico brusco al inicio. El segundo modelo STMV tiene una media de 15.06 y algo más de amplitud en algunas de sus predicciones y la red neuronal que presentó menos amplitud, pero un comportamiento que presenta menos picos o valores más alejados de la media de 15.45 correspondiente al modelo STE.

Como segunda conclusión se puede decir que mejoran las predicciones al agregar más variables de entrada a la red, cómo es el resultado del modelo 2. STMV.

Para el desarrollo del modelo de múltiples variables, se eligió la variable “Fecha” cómo la variable adicional, pero es posible utilizar otras variables de calidad del aire con datos que pueden ser utilizados para el ejercicio de pronóstico. Al mismo tiempo, se pueden mejorar los tres modelos

propuestos en este trabajo, para ello es viable aumentar variable de DIAS = 7 se puede probar con 5 ó 10 DIAS.

Por último, este trabajo presenta el uso de series de tiempos mediante técnicas de Machine Learning para la predicción de PM2.5, dicha herramienta y metodología no había sido utilizada hasta ahora con ese fin por parte de la Autoridad Ambiental - CORPOCALDAS. De esta manera el trabajo de investigación tiene un impacto positivo, ya que en el estado del arte no se reporta ningún trabajo que aborde el tema desarrollado para la ciudad de Manizales. Las redes neuronales tienen un buen desempeño para la predicción de series de tiempo. Los resultados obtenidos permiten tener confianza en las predicciones de PM2.5 con un error de 3% de error y así informar la comunidad sobre los cambios en los niveles de la calidad del aire y concentración de PM2.5 en un futuro cercano, este error del 3% puede ser considerado adecuado desde el punto de vista de manejo preventivo de enfermedades respiratorias originadas por contaminación del aire. Es viable a partir de este error informar a la comunidad sobre posibles afectaciones a la salud humana por contaminación del aire en el centro de la ciudad de Manizales.

6. REFERENCIAS BIBLIOGRÁFICAS

- Lucila L. Chiarvetto Peralta, Fernando A. Rey Saravia, y Nélida B. Brignole. (2008). Aplicación de redes neuronales artificiales para la predicción de calidad de aire. Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC).
- Daniela Baena-Salazar, José F. Jiménez, Carmen E. Zapata & Álvaro Ramírez-Cardona. (2018). Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá.
- Carlos Lino-Ramírez, Rogelio Bautista-Sánchez, Sandra P. Bombela-Jiménez. (2019). Utilización de un sistema en tiempo real para la predicción de contaminación del aire. Tecnológico Nacional de México en León, Guanajuato.
- Riveros, C.A.; Melgarejo, M.; Riveros, A. y Alvarado L. (2012). Sistema Difuso Evolutivo para la Predicción del Nivel de Contaminación del Aire por Material Particulado: Caso Puente Aranda (Bogotá). *Ingeniería, Vol. 17, No. 2, pág.* 55 - 62.
- Instituto para la salud Geoambiental. Contaminación del aire de interiores y salud. https://www.saludgeoambiental.org/contaminacion-aire-interiores-salud?gclid=CjwKCAjw9aiIBhA1EiwAJ_GTSuJ3jc0YG631oFBm3hZthcGu09SGByrrUb6zq8T_c8TnZSkEWohHmxoCbwQQAvD_BwE.
- University College London – Universidad de los Andres. Caracterización de la contaminación atmosférica en Colombia. https://prosperityfund.uniandes.edu.co/site/wp-content/uploads/Caracterizaci%C3%B3n-de-la-contaminaci%C3%B3n-atmosf%C3%A9rica-en-Colombia.pdf
- Universidad Nacional de Colombia. Centro de Datos e Indicadores Ambientales de Caldas CDIAC. http://cdiac.manizales.unal.edu.co/.