***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**](mailto:degarciac@libertadores.edu.co)

**Facultad de Ingeniería y Ciencias Básicas**

**Fundación Universitaria Los Libertadores**

**Bogotá D.C**

**Director:**

**Msc John González Veloza**

[**jjgonzalezv02@libertadores.edu.co**](mailto: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 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 SO2, O3 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).

**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

a certain group of observations, in order to generate preventive measures of respiratory diseases that are diagnosed from the high levels of atmospheric pollution. This innovative theme for the city of Manizales, in which the industry and the high traffic flow in the city center, cause an increase in particles of less than 2.5 microns, which affects the suffering of diseases such as allergic rhinitis, asthma and other diseases classified as acute respiratory disease (ARI). Generally, the inhabitants of the populated centers do not know the quality of the air they breathe, as well as the preventive measures that can be taken into account to reduce the deterioration of health due to low air quality. The data set used for the development of this research work contains information on the concentration of particulate matter PM 2.5 and climatological variables such as: Temperature, Wind Speed, Precipitation, and other air quality variables such as SO2, O3 and CO. The database presented absent values, for which data captured from equipment near the Caldas Government station, such as the La Nubia airport station, the Caldas hydroelectric plant - CHEC, Aguas de Manizales, and Air Quality Surveillance System (SVCA) of Corpocaldas, were used, information that It is compiled by the Institute of Environmental Studies IDEA of the National University, Manizales headquarters.

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

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 aire1. 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 SO2 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)2.

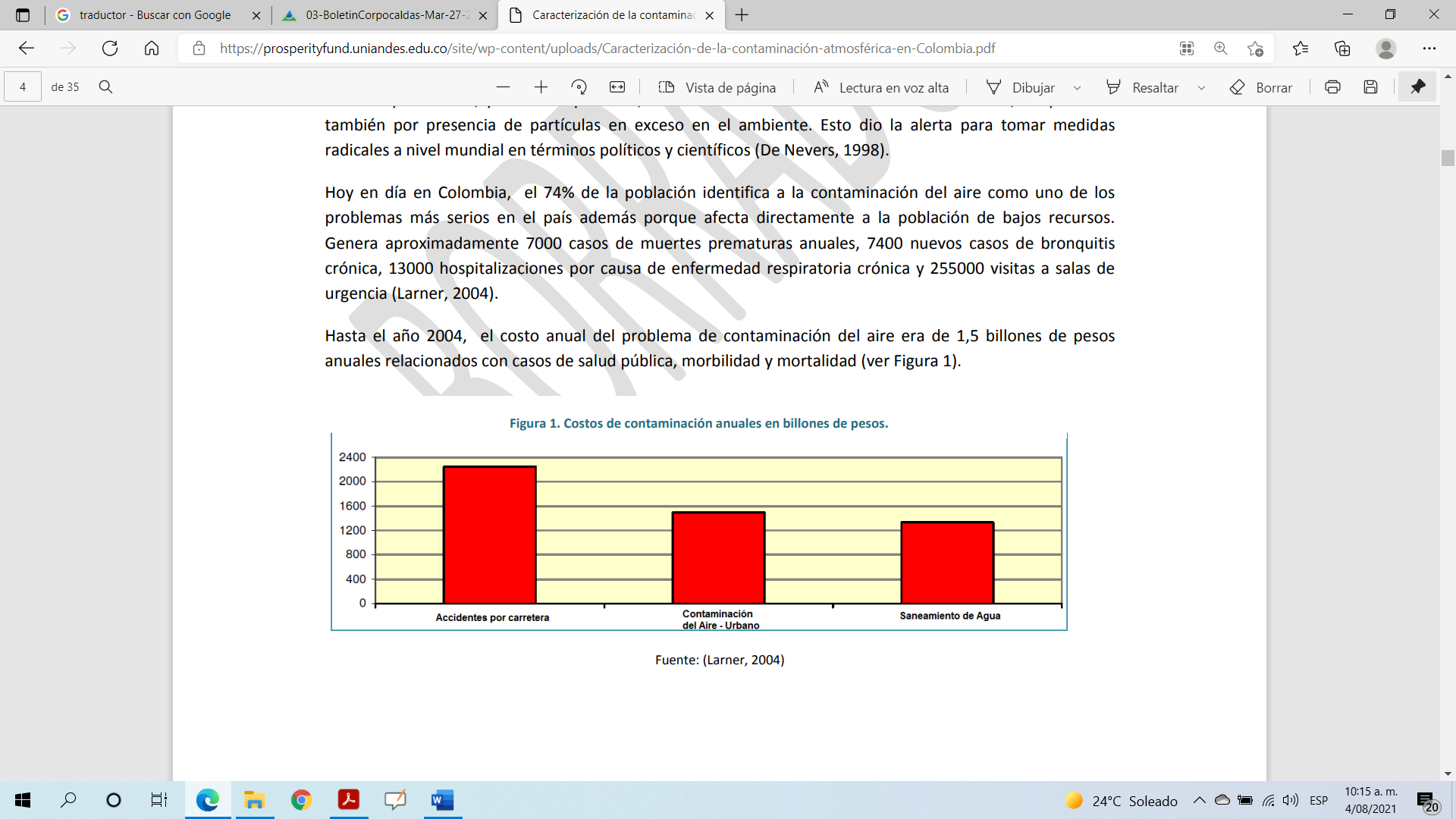
1. <https://www.saludgeoambiental.org/contaminacion-aire-interiores-salud?gclid=CjwKCAjw9aiIBhA1EiwAJ_GTSuJ3jc0YG631oFBm3hZthcGu09SGByrrUb6zq8T_c8TnZSkEWohHmxoCbwQQAvD_BwE>

2. <https://prosperityfund.uniandes.edu.co/site/wp-content/uploads/Caracterizaci%C3%B3n-de-la-contaminaci%C3%B3n-atmosf%C3%A9rica-en-Colombia.pdf>

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).

Hasta el año 2004, el costo anual del problema de contaminación del aire era de 1,5 billones de pesos anuales relacionados con casos de salud pública, morbilidad y mortalidad (ver Figura 1).

Figura 1. Costos de contaminación anuales en billones de pesos.



Fuente: (Larner, 2004)

En términos de Sostenibilidad, la contaminación del aire genera repercusiones importantes en materia social, ambiental y económica, por lo cual se busca crear un modelo preciso y confiable que permita pronosticar las concentraciones de material particulado PM2.5 en el centro de la ciudad de Manizales, el cual puede ser extrapolable a otras zonas de la ciudad, con el fin de informar de manera temprana las condiciones de la calidad del aire para que la autoridad ambiental CORPOCALDAS y las Instituciones de Salud Pública puedan tomar decisiones en materia de protección de la calidad del aire y de la salud pública.

La situación de aislamiento preventivo y obligatorio establecido por el Gobierno Nacional y las Administraciones Locales a causas del SARS-COVID 19, en el año 2020, dejó en evidencia que la contaminación atmosférica producida por material particulado PM10 y PM2.5 en Manizales se redujo en más del 40% durante los días de confinamiento. El mejoramiento de la calidad del aire se debe a la reducción del tráfico vehicular y otras actividades de tipo industrial.

3. <http://cdiac.manizales.unal.edu.co/>

4. https://es.wikipedia.org/wiki/Manizales

1. **METODOLOGÍA**

La zona de interés para el presente trabajo de investigación se ubica en el Centro de la ciudad de Manizales departamento de Caldas. Se utilizó la información contenida en el Centro de datos e Indicadores Ambientales de Caldas (CDIAC)3, plataforma online que permite visualizar la información ambiental del departamento de Caldas, utilizando los datos capturados por la estación instalada en la Gobernación del departamento de Caldas, identificada con el ID 170010207054, la cual entro en funcionamiento a partir del año 2009.

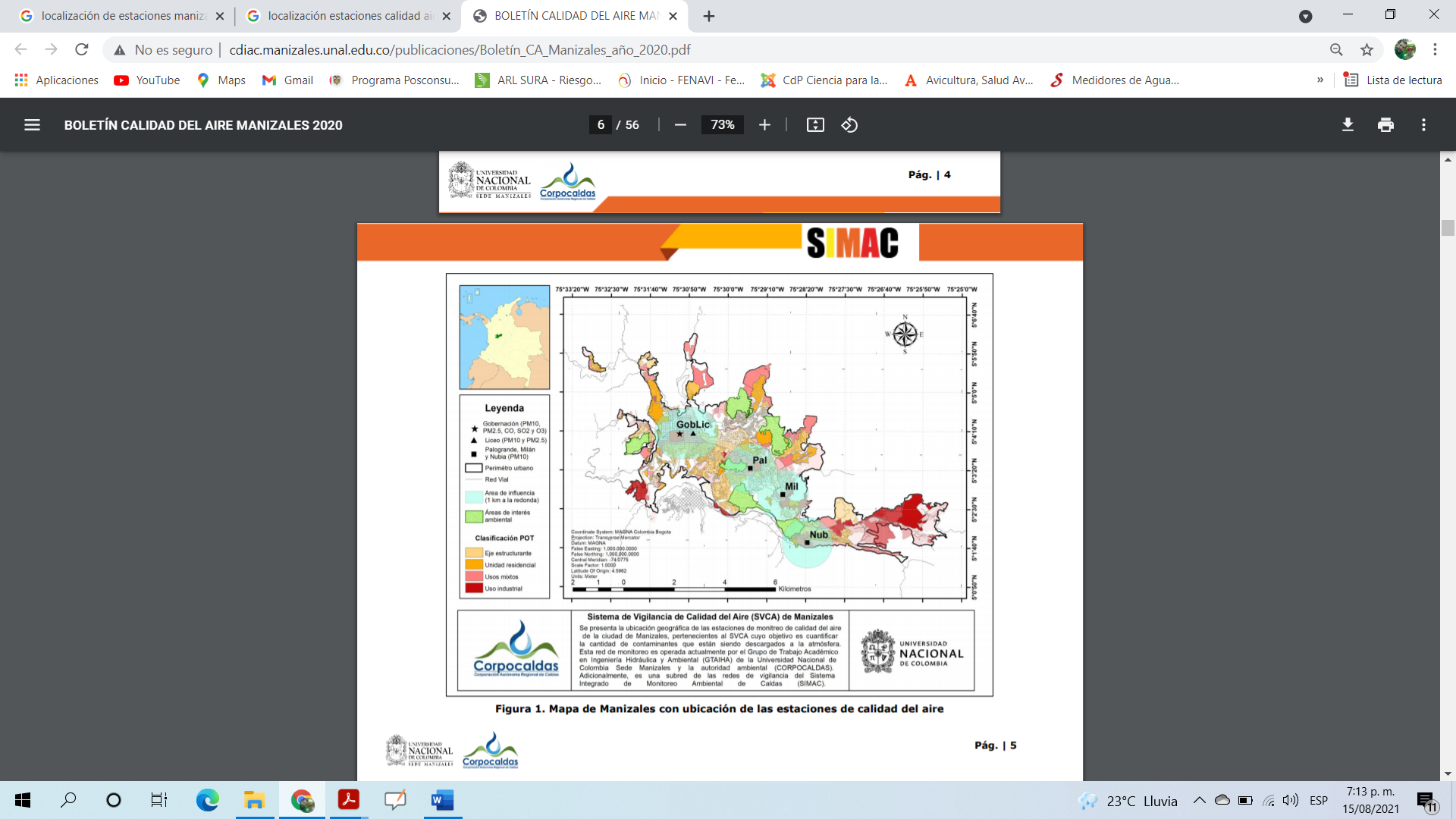
La metodología sistemática utilizada para el desarrollo del modelo de pronóstico de concentración de PM2.5, se basa en Series Temporales con Redes Neuronales - Forecasting, el cual se explicará en detalle posteriormente; incluyendo los pasos que se sugiere aplicar para el desarrollo del modelo.

* 1. **Zona de Estudio**

Manizales es [municipio](https://es.wikipedia.org/wiki/Municipios_de_Colombia) [colombiano](https://es.wikipedia.org/wiki/Colombiano), capital del [departamento](https://es.wikipedia.org/wiki/Departamentos_de_Colombia) de [Caldas](https://es.wikipedia.org/wiki/Caldas_(Colombia)). Está ubicado en el centro occidente de [Colombia](https://es.wikipedia.org/wiki/Colombia) en el [eje cafetero](https://es.wikipedia.org/wiki/Eje_cafetero), sobre la Cordillera Central de los Andes, cerca del [Nevado del Ruiz](https://es.wikipedia.org/wiki/Nevado_del_Ruiz). Tiene una población de 434,000 habitantes (2018). Es la región más poblada y competitiva del departamento con un aporte del 68% de su PIB total. Fundada en 1849 por colonos [antioqueños](https://es.wikipedia.org/wiki/Antioque%C3%B1o), hoy es una ciudad con actividades económicas, industriales, culturales y turísticas.[8](https://es.wikipedia.org/wiki/Manizales#cite_note-8)​ De su actividad cultural son de resaltar la [Feria de Manizales](https://es.wikipedia.org/wiki/Feria_de_Manizales), el Festival Internacional de Teatro de Manizales, siendo la ciudad teatral más importante en la historia de Colombia, y el Festival Manizales Grita Rock. Como capital de departamento, Manizales alberga las sedes de la [Gobernación de Caldas](https://es.wikipedia.org/wiki/Caldas), la [Asamblea Departamental](https://es.wikipedia.org/wiki/Asamblea_Departamental_(Colombia)), el Tribunal Administrativo de Caldas, la [Fiscalía General de la Nación](https://es.wikipedia.org/wiki/Fiscal%C3%ADa_General_de_la_Naci%C3%B3n_(Colombia)), así como diferentes empresas públicas e instituciones y organismos del estado4.

Figura 2

Mapa de Manizales con ubicación de las estaciones de calidad del aire



Nota: La figura 2 muestra la cantidad y ubicación de las estaciones de calidad del aire en la ciudad de Manizales.

Para el desarrollo de la presente investigación, se hace uso de los datos capturados por medio de la estación ubicada en el edificio de la Gobernación de Caldas, este equipo permite recopilar datos para PM10-Manual, PM2.5-Manual, SO2-Automático, O3-Automático, CO-Automático, PM10 y PM2.5-Automático. Dicha estación de monitoreo de la calidad del aire es propiedad de la Corporación Autónoma Regional de Caldas – CORPOCALDAS, es importante tener claridad que en el año 2019 se instaló la unidad que permite recopilar datos sobre la calidad del aire medido en términos de PM2.5-Automático.

**2.2 Pronóstico de Series Temporales con Redes Neuronales en Python - ForesCasting**

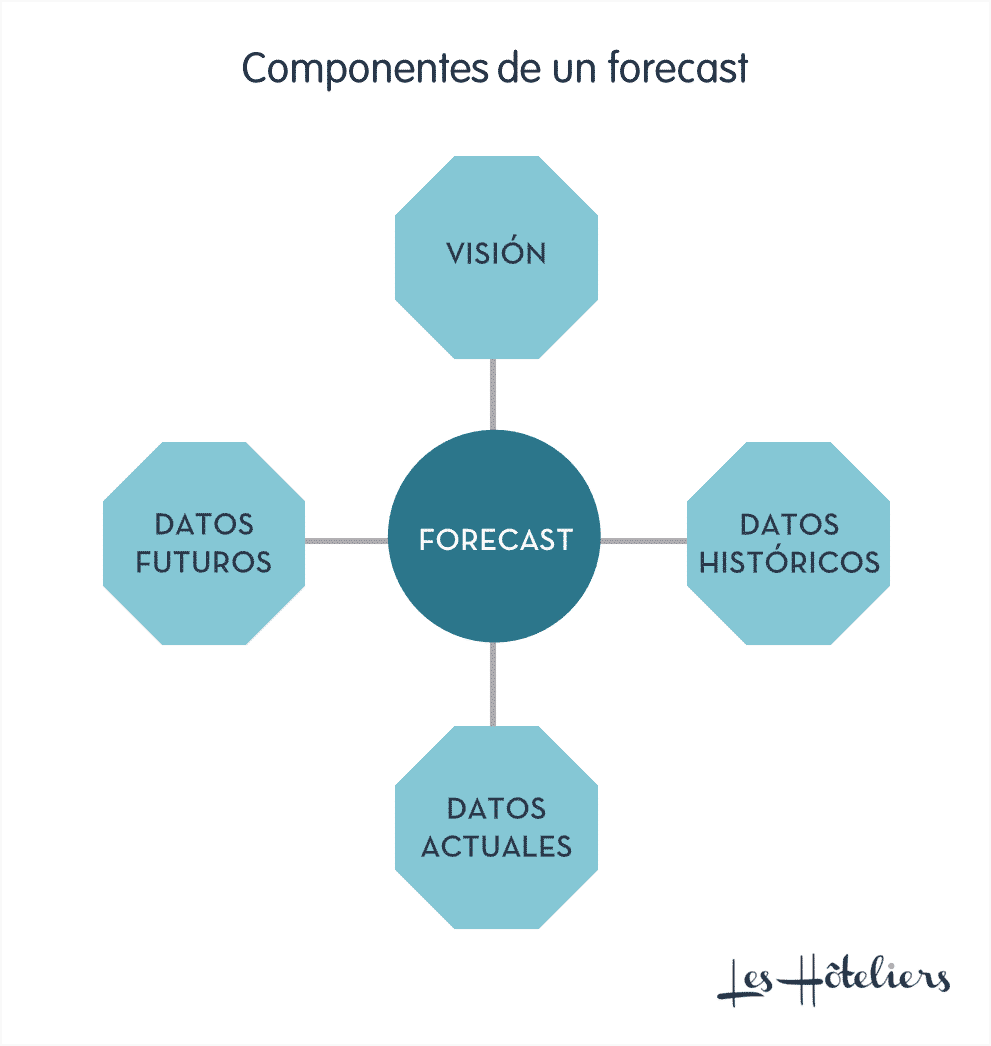
Para la presente investigación se desarrolló un modelo predictivo para un determinado contaminante de manera individual (PM2.5). El interés en el predictor seleccionado surge de la necesidad de determinar con anticipación la calidad del aire en el centro de la ciudad de Manizales, la predicción de PM2.5 constituye la fase inicial de proyectos futuros de gran magnitud.

5. <https://www.minambiente.gov.co/images/normativa/app/resoluciones/96-res%202254%20de%202017.pdf>

Con el desarrollo y validación del modelo de pronóstico, se llevará a cabo un análisis de la construcción, con el fin de extraer los conocimientos y resultados que más beneficios generan, y poder así, aplicarlos a la predicción de otros tipos de contaminantes que disminuyen la calidad del aire. En Colombia, existe normatividad de calidad del aire ambiente que debe ser cumplida para lograr así un predictor operacional (Ministerio de Ambiente y Desarrollo Sostenible (2017). Resolución 2254)5. Para llevar a cabo un pronóstico de calidad del aire basado en la concentración de un parámetro como PM2.5 requiere de experimentación en el desarrollo de la creación de modelos. En un principio se eligió realizar el modelo de pronóstico de contaminación del aire mediante Series de Tiempo con Redes Neuronales, ya que este método permite trabajar con un conjunto de muestras tomadas a intervalos de tiempo regulares, siendo interesante analizar el comportamiento de los contaminantes en el mediano y largo plazo, intentando detectar patrones y llagar así a realizar pronósticos de cómo será su comportamiento futuro. Las Redes Naturales han mostrado ser un método eficiente y universal en la aproximación de funciones para cualquier tipo de dato (Lek y Guégan, 1999). En este caso la función está determinada por un conjunto de variables meteorológicas, estacionales y de concentraciones de otros contaminantes al momento de la predicción que determinan el aumento o disminución de un contaminante en particular en este caso PM2.5. Las Redes Neuronales 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). Lo que hace aún más interesante el uso de series temporales con Redes Neuronales es que estas son independientes del tiempo, lo que rompe con el requerimiento que tiene la regresión lineal de que sus observaciones sean independientes, y por último suelen tener algún tipo de estacionalidad, ó de tendencias de crecer ó decrecer. Para realizar un pronóstico de series temporales, es importante confirmar que la serie es estacionaria, a partir de dicha confirmación es viable realizar el pronóstico. 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. La finalidad del forecasting es estimar el nivel de contaminación que tendrá el centro de la ciudad en determinadas fechas de forma que permita plantear estrategias que aseguren un bajo impacto en la salud de las habitantes y transeúntes del sector. El forecasting, plantea cuatro componentes para una adecuada previsión de la contaminación del aire:

Figura 3

Componentes de un Forecasting



Nota: La figura representa los componentes de un forecasting. Tomado de https://www.leshoteliers.com/que-es-un-forecast/.

En la literatura se proponen diversas estructuras de redes neuronales para analizar y generar pronósticos de las series de tiempo. Una de estas estructuras es el modelo de red neuronal multicapa (RNM), siendo el más utilizado, este modelo comprende un conjunto de unidades de procesamientos o nodos, los cuales se organizan en tres capas: una capa de entrada, al menos una capa oculta y una capa de salida (Zhang et al., 1998). Se han realizado trabajos investigativos en los cuales se comparan las redes neuronales artificiales con los métodos de suavización exponencial Winter, modelos de Box-Jenkins ARIMA y regresión multivariada para analizar series de tiempo. Los resultados indican que las redes neuronales obtienes pronósticos más precisos que los métodos tradicionales, siguiendo los modelos ARIMA. Las redes neuronales permiten identificar patrones de tendencia y estacionalidad no lineales en las series de tiempo analizadas.

Singh y Kumar (2010) dieron a conocer run 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 yt ; t = 1,2,3,….,n se puede moldear con una RNM con p variables de entradas yt-i = (i = 1,2,….,p) y relacionarla con una variable de salida yt. El modelo RNM para el análisis de series de tiempo se indica mediante la siguiente expresión:

h p

yt = α0 + ∑ α1 G (βoj + ∑ βij yt-i) + et (1)

j=1 i=1

Donde:

**yt:** Valor observado de la serie de tiempo en el instante t. Representa el valor para la variable de salida de la RNM,

**yt−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,

**αj:** Representan los pesos de la capa oculta a la capa de salida. Siendo el respectivo sesgo,

**βij:** 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,

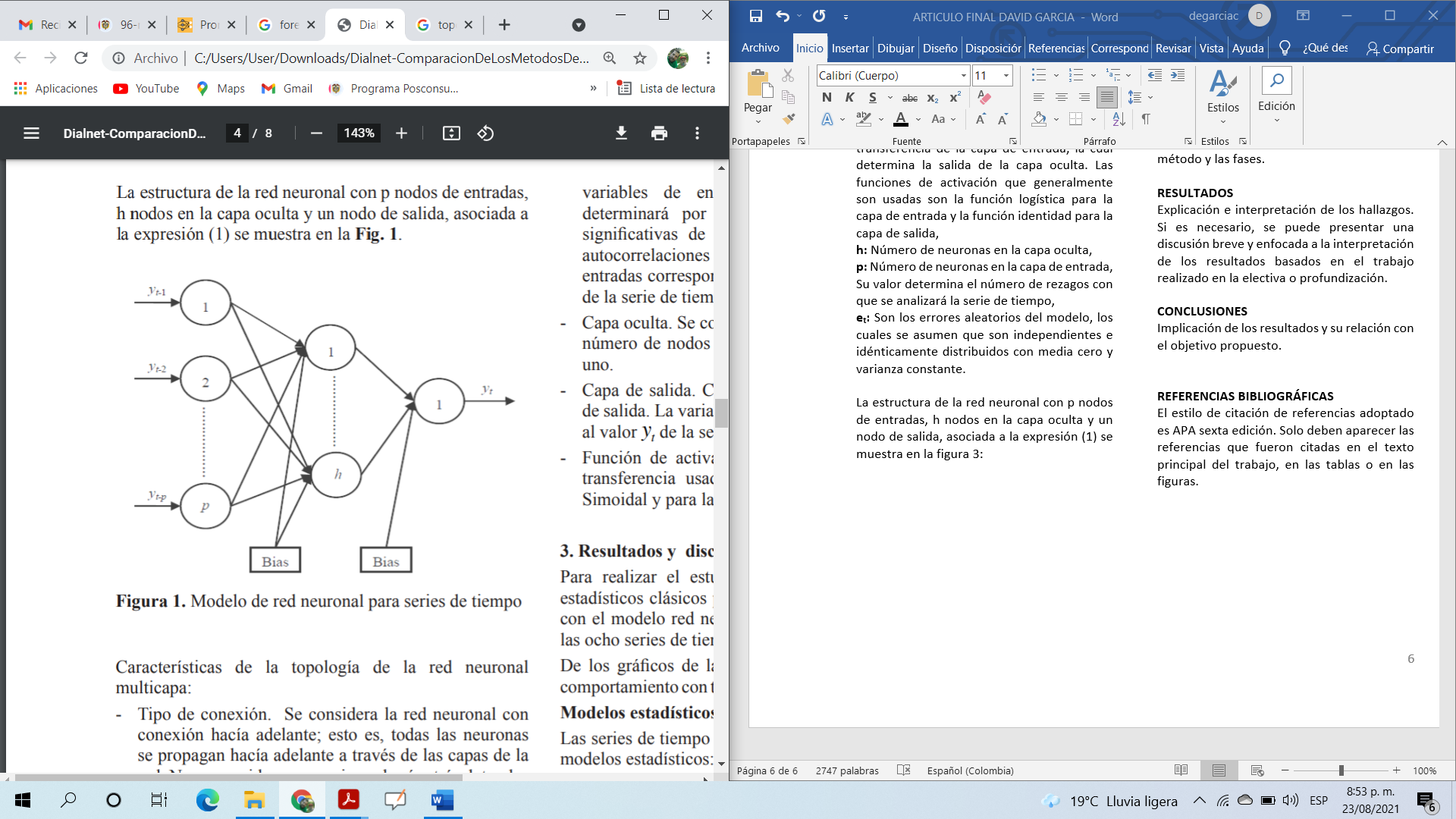
**p:** Número de neuronas en la capa de entrada, Su valor determina el número de rezagos con que se analizará la serie de tiempo,

**et:** Son los errores aleatorios del modelo, los cuales se asumen que son independientes e idénticamente distribuidos con media cero y varianza constante.

La estructura de la red neuronal con p nodos de entradas, h nodos en la capa oculta y un nodo de salida, asociada a la expresión (1) se muestra en la figura 4:

Figura 4

Modelo de red neuronal para series de tiempo



La red neuronal multicapa tiene ciertas características según su topología:

* Tipo de conexión: Se considera una red neuronal con conexión hacia adelante; es decir, todas las neuronas se propagan hacía adelante a través de las capas de la red. No se consideran conexiones hacía atrás, laterales o autorecurrentes (Una neurona de salida se aplica como su propia entrada),
* Capa de entrada: Es la capa en la cual se reciben los datos históricos de la serie de tiempo. El número de nodos o variables de entrada (número de rezagos), se determinará por el número de autocorrelaciones significativas de la serie de tiempo. Si existen p autocorrelaciones significativas, las variables de entradas corresponden a los valores de las posiciones de la serie de tiempo: yt-1, yt-2, yt-3,…..,yt-p.
* Capa oculta: Se considera una sola capa oculta, con el número de nodos igual al número de rezagos menos uno,
* Capa de salida: Conformada por un nodo en la capa de salida. La variable en la capa oculta de salida corresponde al valor yt de la serie de tiempo,
* Función de activación: La función de activación o transferencia usada para la capa de entrada es la Simoidal y para la capa de salida la identidad.

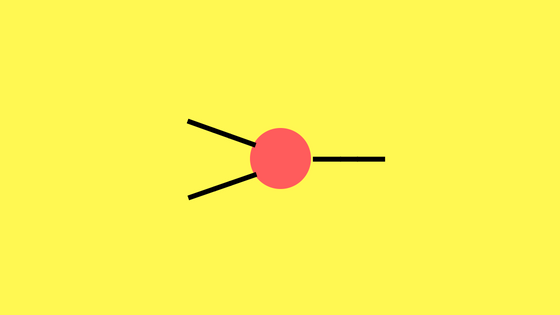
1. **REFERENTES TEORICOS**

A continuación, se describe brevemente la evolución de las redes neuronales, recopilando las que posiblemente sean las consideraciones de mayor importancia que han permitido llegar al nivel actual de aprendizaje profundo.

El inicio de todo se considera la neurona artificial – Perceptrón:

Figura 5

Perceptrón 1958



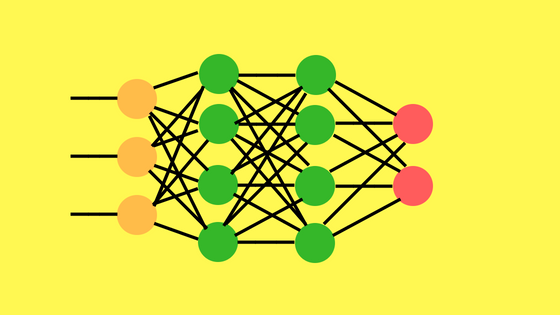
Entre las décadas de 1950 y 1960, el científico Frank Rosenblatt, inspirado en el trabajo de Warren McCulloch y Walter Pitts creó el Perceptron, la unidad desde donde nacería y se potenciarían las redes neuronales artificiales.

Un perceptrón toma varias entradas binarias x1, x2, etc y produce una sola salida binaria. Para calcular la salida, Rosenblatt introduce el concepto de “pesos” w1, w2, etc, un número real que expresa la importancia de la respectiva entrada con la salida. La salida de la neurona será 1 o 0 si la suma de la multiplicación de pesos por entradas es mayor o menor a un determinado umbral.

En el año de 1965, se introdujo el concepto de perceptrón multicapa:

Figura 6

Perceptrón multicapa

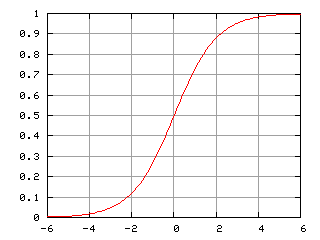


El perceptrón multicapa es una “ampliación” del perceptrón de una única neurona a más de una. Además, aparece el concepto de capas de entrada, oculta y salida. Pero con valores de entrada y salida binarios. Por lo cual el valor de los pesos como el de umbral de cada neurona lo asignaba manualmente el científico. Cuantos más perceptrones en las capas, mucho más difícil conseguir los pesos para obtener salidas deseadas.

A partir de los años 1980s, se empezó a hablar con mayor fuerza del concepto de aprendizaje automático y las neuronas Sigmoides. Con el objetivo de lograr que las redes de neuronas aprendieran solas fue necesario introducir un nuevo tipo de neuronas, llamadas Neuronas Sigmoides son similares al perceptrón, pero permiten que las entradas, en vez de ser ceros o unos, puedan tener valores reales como 0,5 ó 0,377 ó lo que sea. También aparecen las neuronas “bias” que siempre suman 1 en las diversas capas para resolver ciertas situaciones. Ahora las salidas en vez de ser 0 ó 1, será d(w . x + b) donde d será la función sigmoide definida como d(z) = 1/( 1 +e-z). ¡Esta es la primera función de activación!

Figura 7

Curva Logística Normalizada de Wikipedia



Con esta nueva fórmula, se puede lograr que pequeñas alteraciones en valores de los pesos (deltas) produzcan pequeñas alteraciones en la salida. Por lo tanto, se puede ir ajustando muy de a poco los pesos de las conexiones e ir obteniendo las salidas deseadas.

Por su parte el concepto de Redes Feedforward, se utiliza para llamar así a las redes en que las salidas de una capa son utilizadas como entradas en la próxima capa. Esto quiere decir que no hay loops “hacia atrás”. Siempre se “alimenta” de valores hacia adelante. Además, existe el concepto de “fully connected Feedforward Networks” y se refiere a que todas las neuronas de entrada, están conectadas con todas las neuronas de la siguiente capa.

Para el periodo de 1986, se empieza hablar sobre el concepto de Backpropagation, siendo un algoritmo mediante el cual se hizo posible entrenar redes neuronales de múltiples capas de manera supervisada. Al calcular el error obtenido en la salida e ir propagando hacia las capas anteriores se van haciendo ajustes pequeños (minimizando costo) en cada iteración para lograr que la red aprenda consiguiendo que la red pueda -por ejemplo- clasificar las entradas correctamente.

Hasta ahora se ha explicado el referente teórico que se utilizo 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 Embeddings.

**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:** Embeddings en variables categóricas. Para el segundo modelo, se utilizaron embeddings 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 los Embeddings.

La palabra “Embed” es incrustrar, pero en este caso se puede definir como “incorporaciones”. A partir de allí, se puede decir que los embeddings son una manera de dar valoración útil a datos categóricos. Para ello se asigna una profundidad a cada “identificador”, es decir un vector con valores continuos inicialmente aleatorios. Esos valores se ajustarán con backpropagation al igual que nuestra red neuronal. Y finalmente nuestros datos categóricos quedan enriquecidos y dejan de ser “lunes” para ser unos vectores con valores que “significan algo”. ¿Qué significan? para simplificar e intentar entenderlo podemos decir que esos vectores “acercan identificadores similares entre sí y distancia a los opuestos”. Un ejemplo: cuando se utiliza en Natural Language Processing (NLP) con un gran número de palabras, los Embeddings logran hacer que palabras sobre sentimientos positivos “alegría”,”felicidad” queden cercanas pero distanciadas de las que significan sentimientos negativos “odio”,”tristeza”.

Como conclusión de los embeddings, al asignarle vectores con valor numérico continuo a entradas categóricas, estos terminan funcionando como “una mini red neuronal” dentro de la red principal. Aprenden con backpropagation. Y resuelven como valores continuos esos identificadores discretos, acentuando su valor intrínseco. Una ventaja de usar embeddings es que se pueden reutilizar. Una vez entrenados podemos guardarlos para utilizarlos en otra red.

1. **METODOLOGÍA**

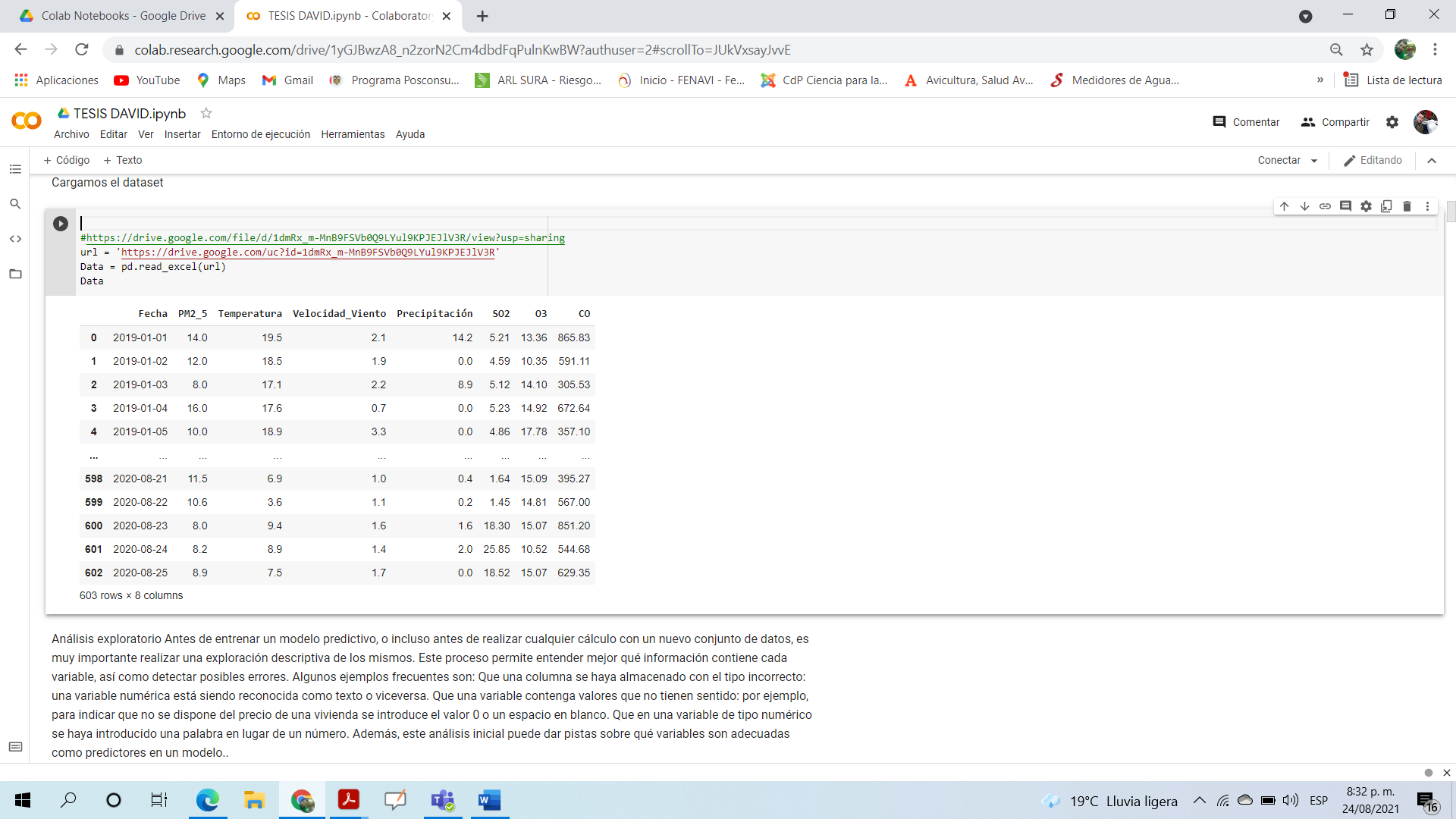
Para llevar a cabo el trabajo de investigación y llevarlo a la práctica, es importante contar con un entorno de ejecución de Python, en este caso se utilizó Google colab, el cual permite ejecutar y programar en Python en tu navegador con las siguientes ventajas:

* 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



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

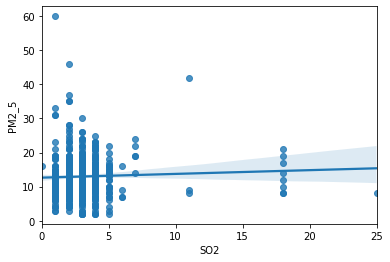


Figura 10

Diagrama de dispersión PM2.5 y CO

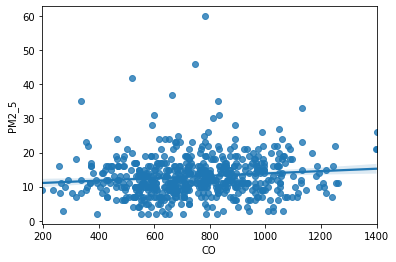
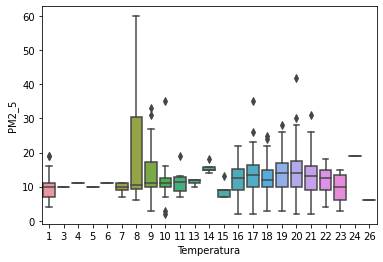


Figura 11

Diagrama de caja PM2.5 y Temperatura

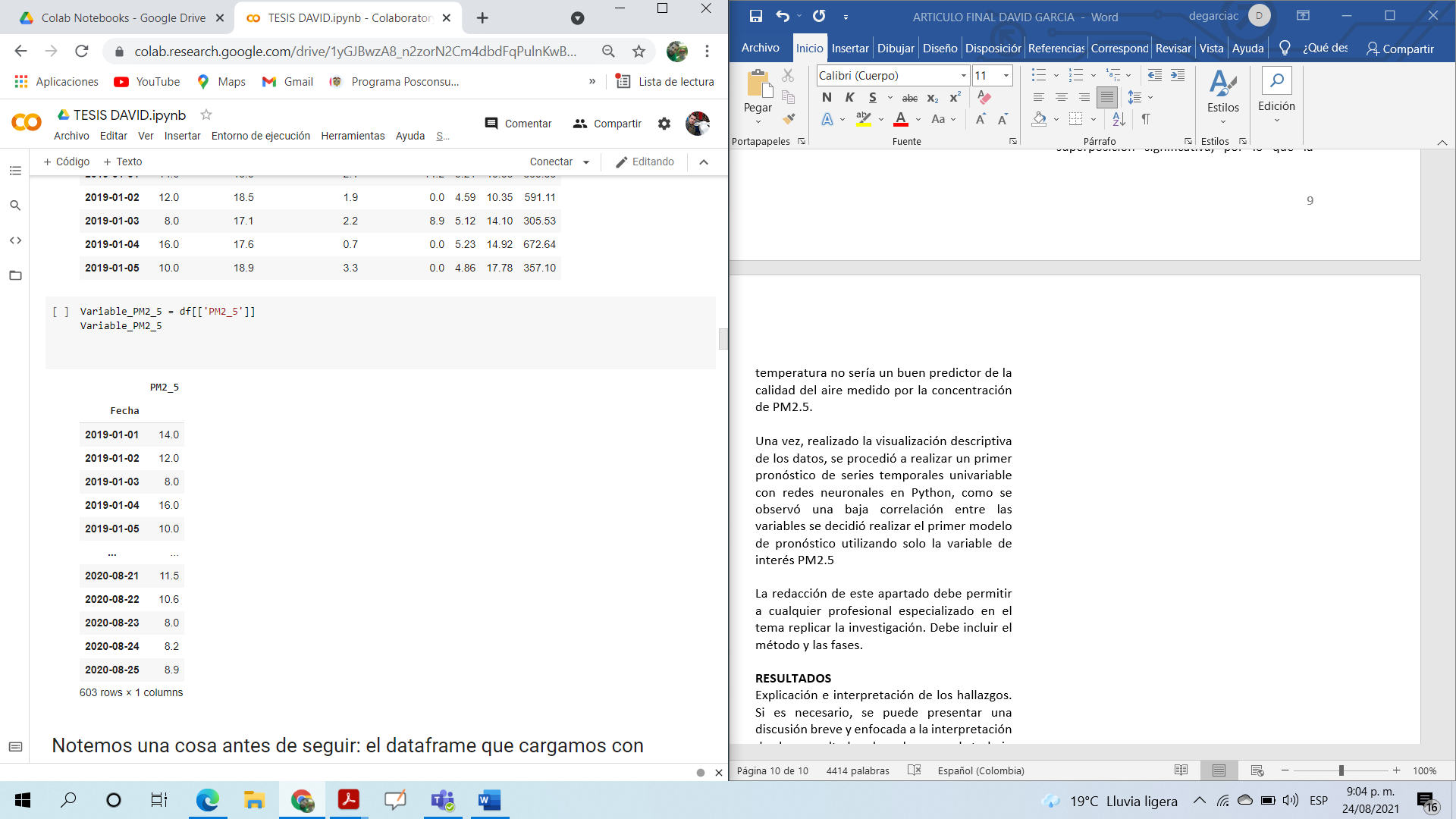


De acuerdo a la figura 10, podemos observar que las distribuciones de PM2.5 entre los diferentes grados de temperatura tiene una superposición significativa, por lo que la temperatura no sería un buen predictor de la calidad del aire medido por la concentración de PM2.5.

Una vez, realizado la visualización descriptiva de los datos, se procedió a realizar 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

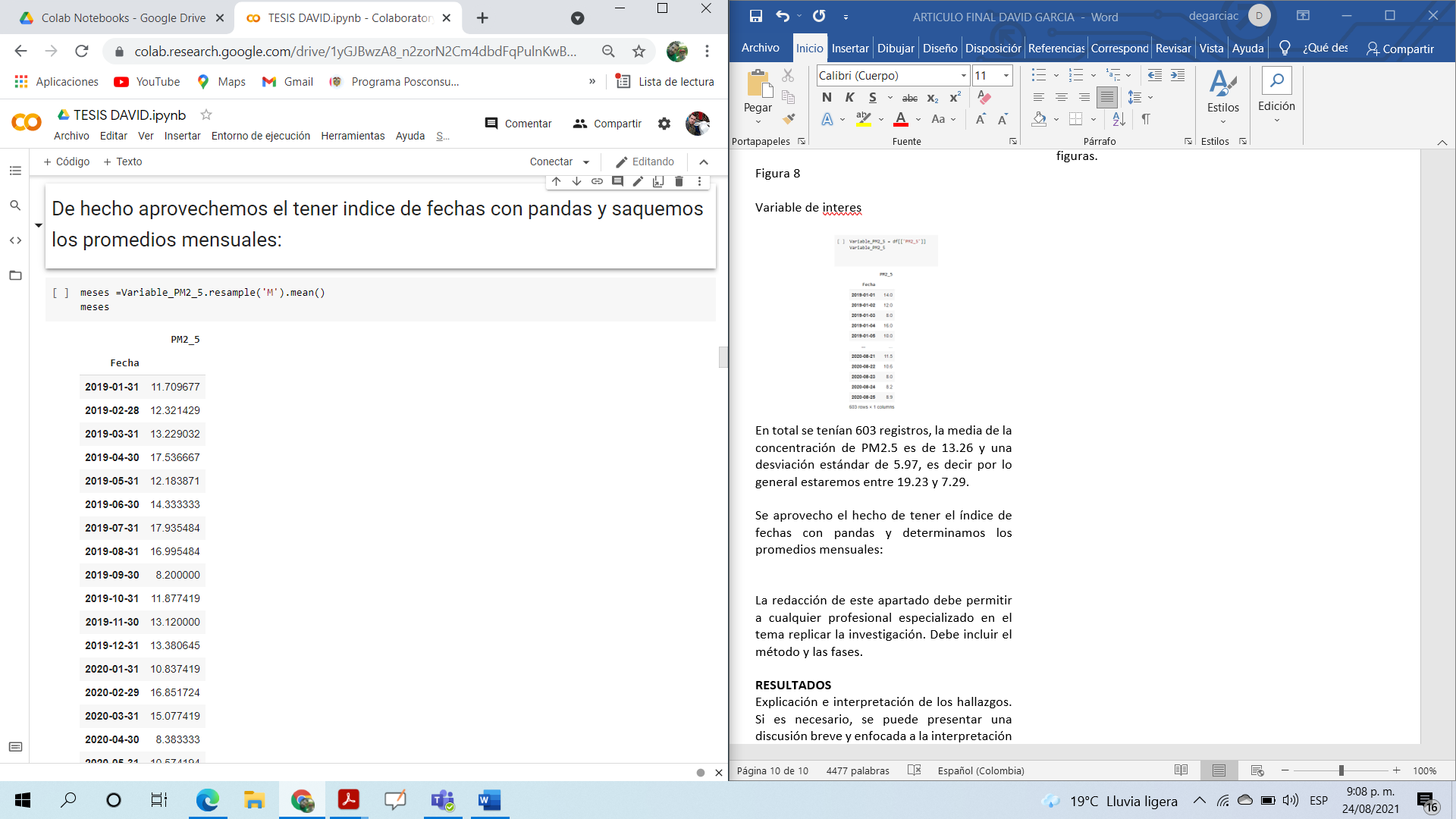


En total se tenían 603 registros, la media de la concentración de PM2.5 es de 13.26 y una desviación estándar de 5.97, es decir por lo general estaremos entre 19.23 y 7.29.

Se aprovecho el hecho de tener el índice de fechas con pandas y determinamos los promedios mensuales:

Figura 13

Promedios mensuales PM2.5

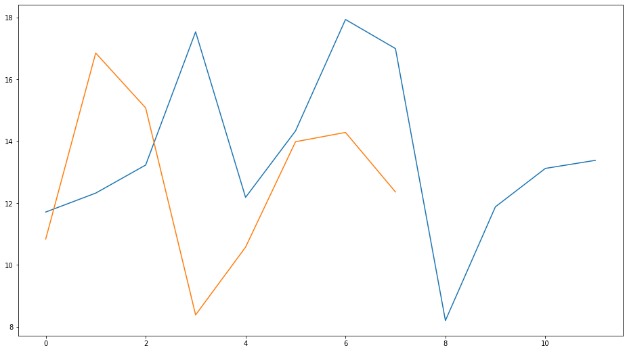


A continuación, se visualizaron esas medias mensuales para la concentración de PM2.5:

Figura 14

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

18



10

8

6

4

2

0

8

10

12

14

16

2020

2019

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.

Preparando los datos: Este puede que sea uno de los pasos 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/m3 de los 7 días anteriores.

**Salida:** El valor del “8vo día”. Es decir, la concentración de PM2.5 µg/m3 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 (pues sabemos que, a nuestra red neuronal, le favorece para realizar los cálculos).

Figura 15

Código función series\_to\_supervised()

6. https://www.aprendemachinelearning.com/aprendizaje-profundo-una-guia-rapida/

7. https://www.aprendemachinelearning.com/una-sencilla-red-neuronal-en-python-con-keras-y-tensorflow/

8- https://www.aprendemachinelearning.com/crear-una-red-neuronal-en-python-desde-cero/

https://www.aprendemachinelearning.com/breve-historia-de-las-redes-neuronales-artificiales/

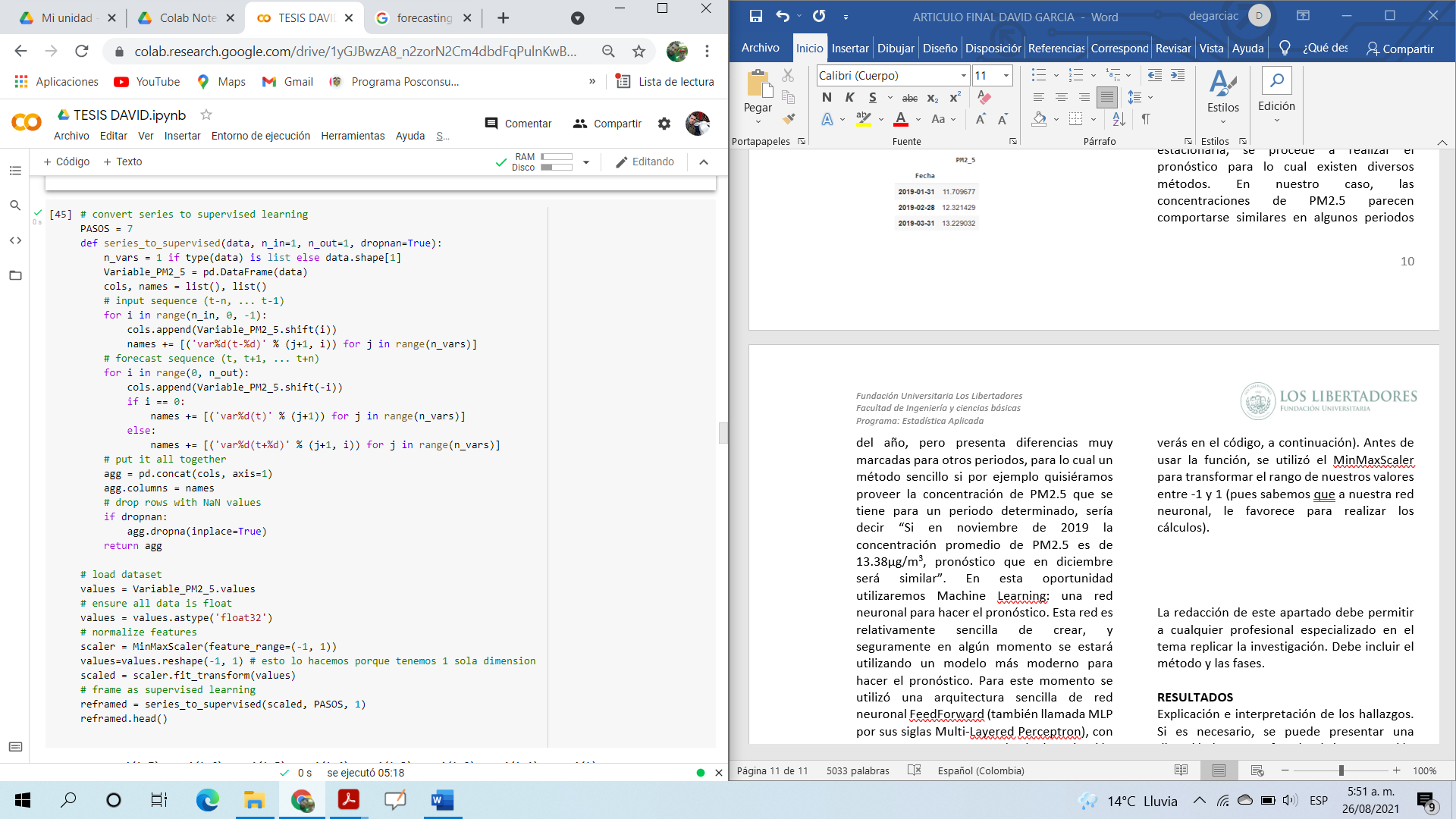
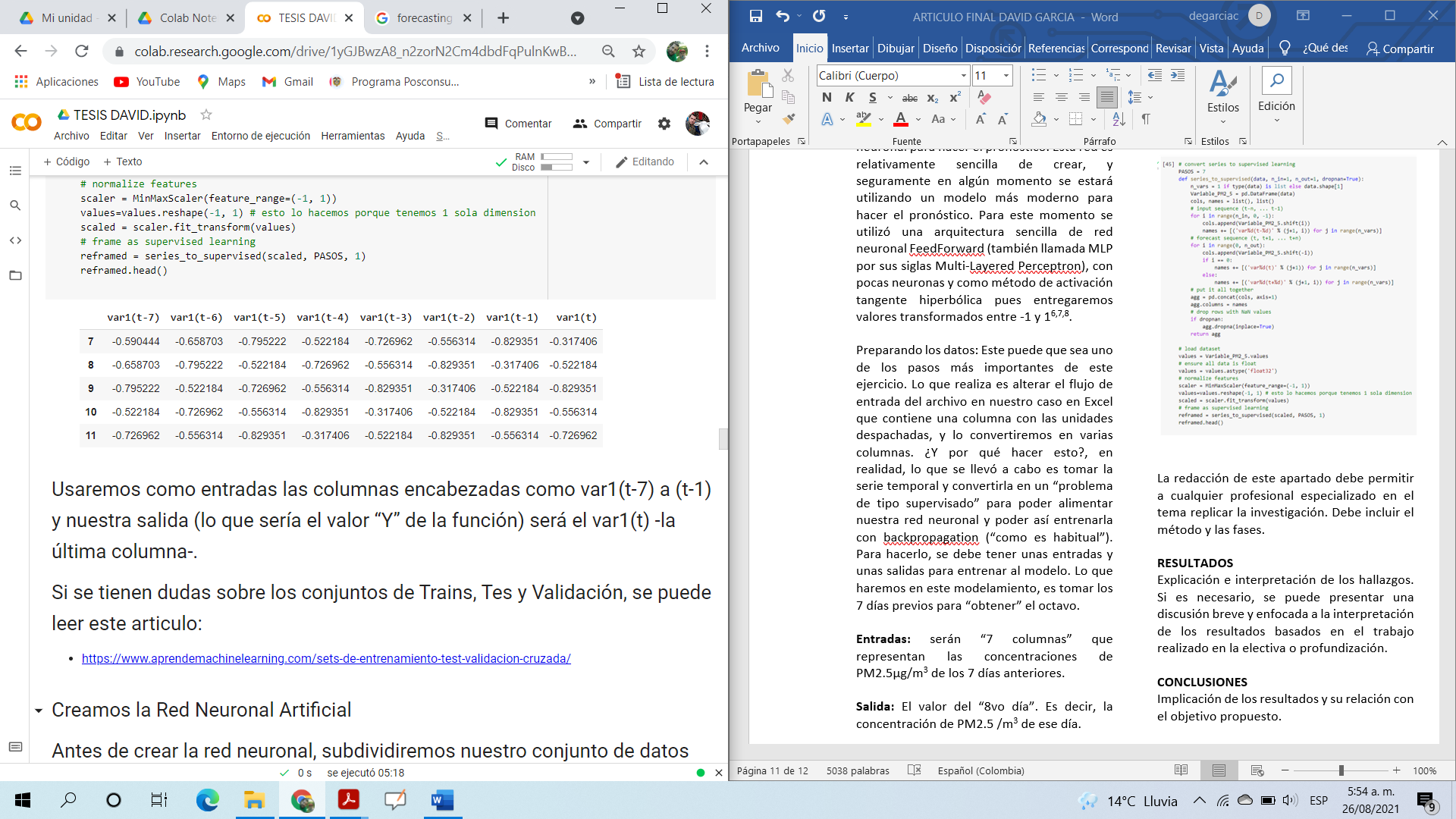


Figura 16

Resultados generados con la función series\_to\_supervied()



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-.

Seguido 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 EPOCHS.

**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 Embeddings, con el fin de ver si se genera mejoría en el modelo.

0.2

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

Tabla 1

Valores finales de las métricas tras las 40 EPOCHS

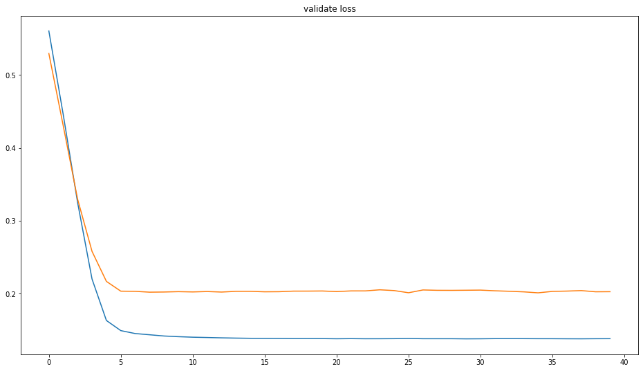


Como se puede observar en la tabla 1, los modelos STMV y STE quedan prácticamente iguales, es decir que tras realizar el ejercicio con múltiples variables y agregando Embeddings no pareciera haber mejoría significativa en las métricas. Por su parte el modelo ST1, presenta un resultado de MSE de 0.0268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

Figura 17

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



40

35

30

25

20

15

10

5

0

0.5

0.3

0.4

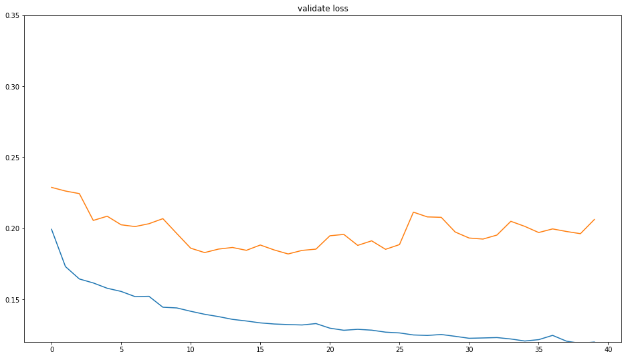
Validación

Entrenamiento

Figura 18

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

0.35



0.06

0.04

0.08

0.18

0.16

0.14

0.12

0.30

0.25

0.20

0.15

40

35

30

25

20

15

10

5

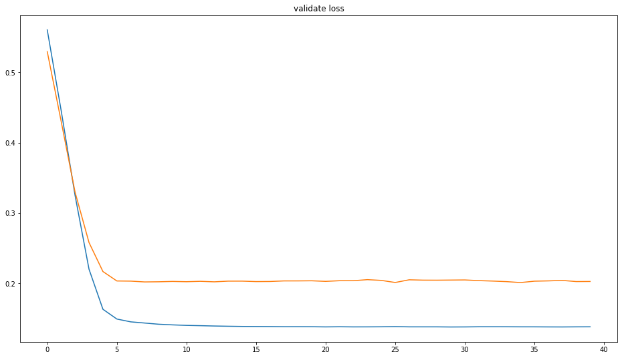
0.0

Validación

Entrenamiento

Figura 19

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



40

35

30

25

20

15

10

5

0.0

0.2

0.3

0.4

0.5

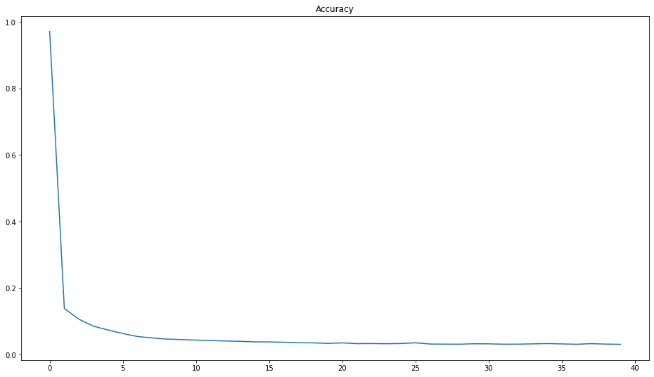
Entrenamiento

Validación

Al comparar las gráficas de accuracy, utilizando la métrica de MSE, se puede observar que el modelo 3 STE desciende hasta 0.05 indicando una mejora en el modelo.

Figura 20

Accuracy Modelo 1. ST1



0.2

0.4

0.6

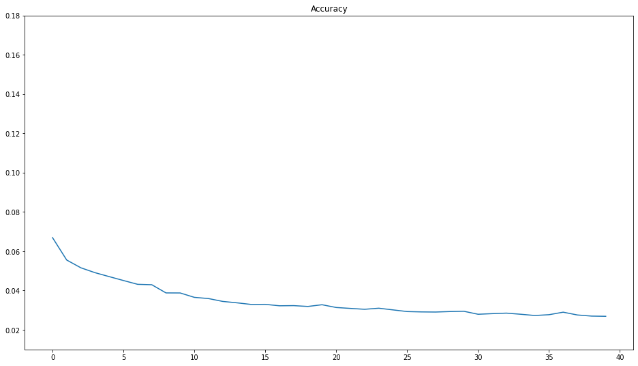
0.8

10

0 5 10 15 20 25 30 35 40

Figura 21

Accuracy Modelo 2. STMV



40

35

30

25

20

15

10

5

0.0

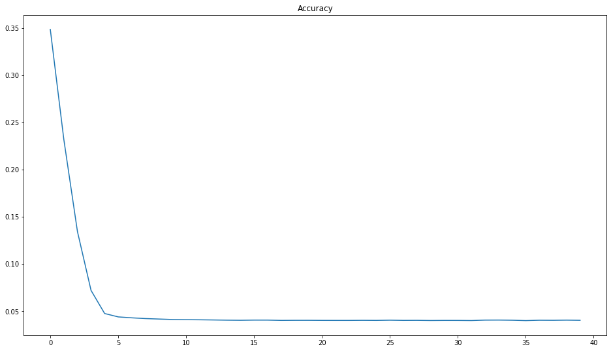
0.02

0.10

Figura 22

Accuracy Modelo 3. STE

0.35



40

35

30

25

20

15

10

5

0.0

0.05

0.10

0.15

0.20

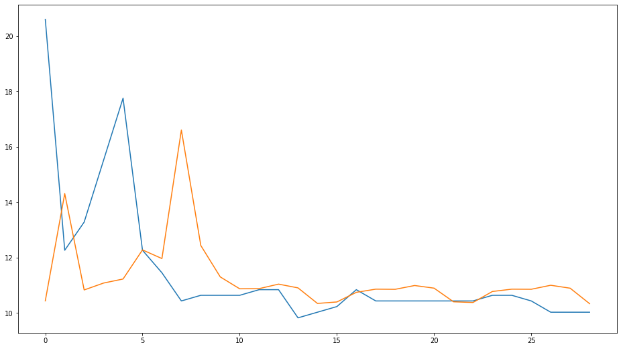
0.25

0.30

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

Figura 23

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



10

12

14

16

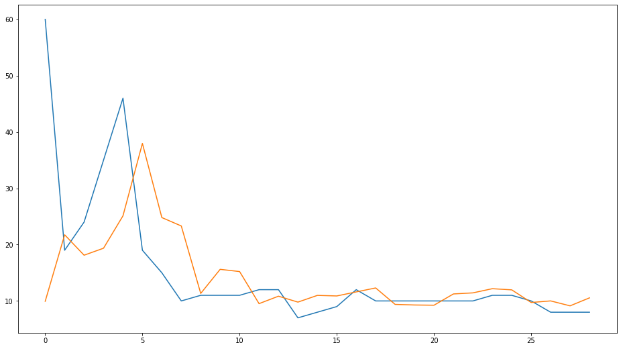
18

20

Figura 24

Comparación Modelo 2. STMV: Presenta mayor mejoría en el pronóstico que el modelo 1.

60



25

20

15

10

5

0

10

20

30

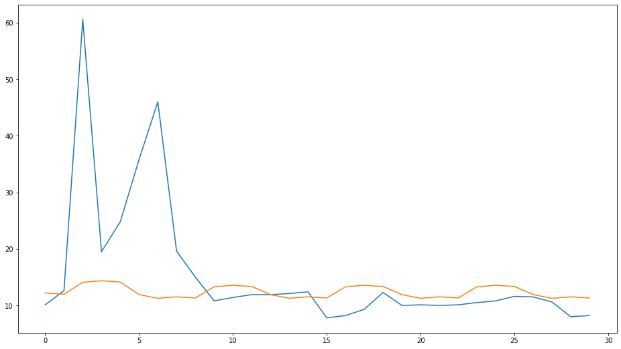
40

50

Figura 25

Comparación Modelo 3. STE: Los Embeddings disminuyen los aciertos a la curva de pronóstico

60



30

25

20

15

10

5

0

50

40

30

20

10

**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 PASOS = 7 se puede probar con 5 ó 10 PASOS.

Las herramientas de Machine Learning, tienen la capacidad a partir de sus librerías de adaptar y mejorar el modelo. Por ejemplo, para usar variables de tipo categórico se puede realizar el ejercicio a través de Embeddings con el fin de optimizar los resultados.

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. Se utilizó un importante volumen de datos y de una gran cantidad de iteraciones para determinar que los resultados no son por la naturaleza estocástica de las redes neuronales; al hacer esto se asegura que los modelos propuestos son certeros y son robustos. 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.

**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/>.