ANÁLISIS DE CORRELACIÓN ENTRE EL ÍNDICE DE CALIDAD DEL AIRE Y EL IMPACTO EN TWITTER PARA LA CIUDAD DE BUCARAMANGA APLICANDO ANÁLISIS DE SERIES TEMPORALES, EXTRACCIÓN Y PROCESAMIENTO DE LENGUAJE NATURAL

KEVIN ALEJANDRO NARANJO REYES

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA

FACULTAD DE INGENIERÍAS

PROGRAMA DE INGENIERÍA MECATRÓNICA

BUCARAMANGA

2021

ANÁLISIS DE CORRELACIÓN ENTRE EL ÍNDICE DE CALIDAD DEL AIRE Y EL IMPACTO EN TWITTER PARA LA CIUDAD DE BUCARAMANGA APLICANDO ANÁLISIS DE SERIES TEMPORALES, EXTRACCIÓN Y PROCESAMIENTO DE LENGUAJE NATURAL

KEVIN ALEJANDRO NARANJO REYES

TRABAJO DE GRADO PRESENTADO COMO REQUISITO PARCIAL PARA OPTAR EL TÍTULO DE INGENIERO MECATRÓNICO

DIRECTOR:

PHD. CARLOS JULIO ARIZMENDI PEREIRA

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA

FACULTAD DE INGENIERÍAS

PROGRAMA DE INGENIERÍA MECATRÓNICA

BUCARMANGA

2021

AGRADECIMIENTOS

Después de 32 semanas, hoy es el día en que escribo este apartado para concluir mi proyecto de fin de grado. Este es un período que requiere mucho aprendizaje no solo en el ámbito académico sino también a nivel individual. Escribir este trabajo ha tenido un gran impacto en mí, por lo que quiero agradecer a todos los que me ayudaron y apoyaron en este proceso.

En primer lugar, me gustaría agradecer a mis tutores Carlos Arizmendi y la empresa K2 ingeniería por su valiosa ayuda. Definitivamente me han ofrecido todas las herramientas necesarias para completar mi proyecto de fin de grado de forma satisfactoria.

De igual manera quiero dar las gracias a la Universidad Autónoma de Bucaramanga, a toda la Facultad de ingenierías, a mis profesores quienes con el aporte de sus valiosos conocimientos me hicieron crecer día a día como profesional. Gracias a cada uno de ustedes por su paciencia, dedicación, apoyo incondicional y amistad.

En especial, me gustaría agradecer a mi padre, a mi madre y mi abuela por sus sabios consejos y su comprensión. Siempre han estado ahí cuando los he necesitado. No solo fueron los primeros en traer internet a Palmas, también son los primeros en regalarle el primer ingeniero mecatrónico al pueblo.

Mi agradecimiento infinito a mi novia, quien supo apoyarme en todo momento dándome mucha alegría día a día. Gracias a mis amigos. No solo han estado a mi lado para apoyarnos entre nosotros en los momentos más complicados, sino que también hemos tenido conversaciones sobre otras cosas no relacionadas con universidades y artículos académicos que han dado mucho valor a mi vida.

¡Muchas gracias a todos!

Kevin Alejandro Naranjo Reyes

RESUMEN

Se aplicó una serie de modelos que predicen el Índice de Calidad del Aire (ICA) a partir de las publicaciones realizadas por los habitantes de Bucaramanga en Twitter, para determinar la calidad del aire en la ciudad. Se seleccionó el mejor modelo a partir de diferentes métricas con base en el accuracy de la clasificación. Para ello, se recopilaron mediciones reales del ICA en varios puntos de la ciudad y se hizo la extracción de Tweets para la misma serie temporal. Al emplear distintos algoritmos de reducción de dimensionalidad junto con técnicas de clasificación, tales técnicas son *FSCNCA*, *Forward Selection*, *LDA* y *Redes Neuronales*, se desarrollaron los modelos de predicción y se seleccionó el de mayor porcentaje de clasificación. El modelo que emplea *FSCNCA* como técnica de reducción de dimensionalidad y *LDA* como clasificador fue el que obtuvo el mejor porcentaje de clasificación, con un *Accuracy* de 69.07% en el conjunto *Validation*.

ABSTRACT

A series of models that predict the Air Quality Index (AQI) from Twitter posts made by Bucaramanga residents were applied to determine air quality in the city. The best model was selected different metrics based (accuracy of the classification). For this purpose, real AQI measurements were collected at several points in the city and Tweets were extracted for the same time series. By employing different dimensionality reduction algorithms along with classification techniques, as FSCNCA, Forward Selection, LDA and Neural Networks, the prediction models were developed and the one with the highest classification percentage was selected. The model using FSCNCA as dimensionality reduction technique and LDA as classifier was the one that obtained the best classification percentage, with an Accuracy of 69.07% in the Validation set.

LISTADO DE TABLAS

LISTADO DE FIGURAS

LISTADO DE SÍMBOLOS

EIC: Enfermedad Isquémica del Corazón

EPOC: Enfermedad Pulmonar Obstructiva Crónica

IRAB: Infección Respiratoria Aguda Baja

ECV: Enfermedad Cerebrovascular

ICA: Índice de Calidad del Aire

NN: Neural Network

FSCNCA: Feature Selection Using Neighborhood Component Analysis

ADL: Análisis Discriminante Lineal

NAR: Nonlinear autoregressive neural network

NARX: Nonlinear autoregressive with exogenous inputs

NLIO: Nonlinear Input/Output

NLP: Natural language processing

CONTENIDO

INTRODUCCIÓN

La empresa K2 Ingeniería tiene como objetivo analizar la relación que tiene el índice de calidad del aire con las publicaciones en Twitter para generar un modelo predictivo a partir de las publicaciones hechas. Las redes sociales permiten canalizar parte de la percepción de las personas sobre diversos temas. El presente trabajo de investigación define y desarrolla una metodología para el desarrollo de un modelo de Machine Learning que permite predecir el índice de calidad del aire a partir de comentarios realizados en Twitter en la ciudad de Bucaramanga aplicando análisis de series temporales para completar los datos faltantes de material particulado, extracción y procesamiento de Tweets, y clasificadores de análisis discriminante y redes neuronales con reducción de características por FSCNCA (*feature selection using neighborhood component analysis for classification*) y Forward Selection. La solución a este problema es desglosada en tres procesos principales:

El primer proceso se basa en la recopilación de datos que están relacionados con la calidad del aire como lo son las partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 µm. Para poder realizar la clasificación de la calidad del aire en categorías relacionadas con el daño que los contaminantes ocasionan en la salud pública se toman como base los estándares de la Resolución 2254 de 2017 (“Por la cual se adopta la norma de calidad del aire ambiente y se dictan otras disposiciones”). Además, se emplean las mediciones del índice de calidad del aire tomadas por la empresa K2 Ingeniería cuya estación de medición está ubicada en la Calle 36 con Carrera 36 (Bucaramanga), y por el Acueducto Metropolitano de Bucaramanga (AMB), el cual cuenta con sistemas de vigilancia de calidad del aire ubicados en la ciudad de Bucaramanga, Girón y Floridablanca.

El segundo proceso consiste en la recopilación de publicaciones y noticias expuestas en Twitter que se tienen sobre la calidad del aire. Esto se logra extrayendo y estructurando los textos (Tweets) para posteriormente aplicar técnicas de procesamiento natural. Seguido a esto, se separan las palabras y emojis en los diversos Tweets utilizando técnicas de Inteligencia Artificial. Posteriormente, se desarrollan modelos para reducir la dimensionalidad en las palabras relevantes empleando FSCNCA y Forward Selection. Los clasificadores implementados son el Análisis Discriminante (*diaglinear, diagquadratic, pseudolinear, pseudoquadratic*) y redes neuronales *patternet* con *Early Stopping.*

Por último, en el tercer proceso se analiza la correlación existente entre la perspectiva general de las personas según los datos extraídos de Twitter y los datos reales medidos por los sensores de los sistemas de vigilancia de calidad del aire.

Durante el desarrollo del proyecto de plantean las siguientes preguntas de investigación:

Pregunta de investigación general

- ¿Cuál es la relación del índice de calidad del aire y la percepción de la contaminación del aire en las publicaciones realizadas en Twitter por las personas en Bucaramanga?

Preguntas de investigación específicas

- ¿Cuál es la información científica relacionada con contaminación del aire y sus principales influyentes?
- ¿Cómo determinar el índice de calidad del aire?
- ¿Cómo determinar cuál es la percepción de las personas en la ciudad de Bucaramanga sobre la calidad del aire mientras se han usado los sensores?
- ¿Cómo se correlacionan las percepciones en las publicaciones hechas en Twitter con la calidad del aire que hay en la ciudad?

OBJETIVOS

Objetivo General

Generar un modelo que determine el índice de calidad del aire usando las publicaciones realizadas por las personas en Bucaramanga a partir de datos extraídos en Twitter.

Objetivos Específicos

- Registrar la información relacionada con contaminación del aire por medio de una revisión sistemática de la literatura.
- Obtener los datos históricos de material particulado que han obtenido con las estaciones meteorológica distribuidas por la ciudad.
- Implementar un algoritmo automatizado para la extracción y transformación de Tweets haciendo una comparación entre las diferentes posibilidades que se encuentran.
- Analizar la percepción de las personas en la ciudad de Bucaramanga sobre la calidad del aire usando técnicas de procesamiento de lenguaje natural separando las publicaciones en palabras.
- Correlacionar las palabras de las publicaciones con la calidad del aire que hay en la ciudad de Bucaramanga para identificar las palabras más relevantes.
- Seleccionar las palabras y el clasificador de Deep Learning para la implementación del modelo.

1. MARCO TEÓRICO

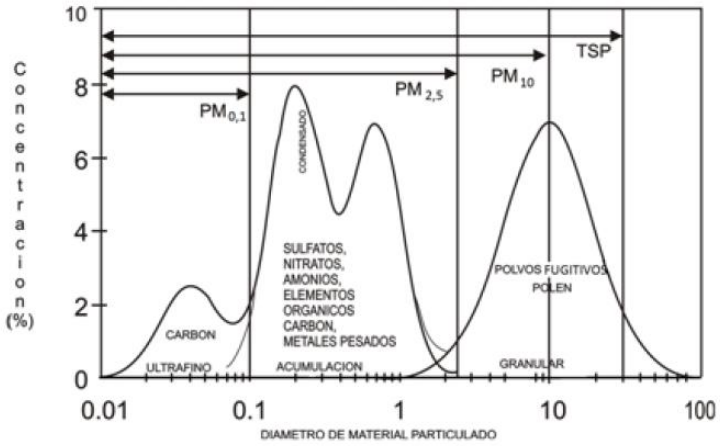
1.1. Calidad del aire

1.1.1. Material particulado

El material particulado (MP) es un conjunto de partículas sólidas y líquidas emitidas directamente al aire, tales como el hollín de diésel, polvo de vías, el polvo de la agricultura y las partículas resultantes de procesos productivos. El MP no sedimenta en períodos cortos, sino que permanece suspendido en el aire debido a su tamaño y densidad[1]. Estas partículas en suspensión son una compleja mezcla de productos químicos y/o elementos biológicos como metales, sales, materiales carbonosos, orgánicos, volátiles, compuestos volátiles, hidrocarburos aromáticos policíclicos y endotoxinas que pueden interactuar entre sí formando otros compuestos[2].

Debido a que son de tamaño, forma y composición variada, para su identificación se han clasificado en términos de su diámetro aerodinámico que corresponde al diámetro de una esfera uniforme en unidad de densidad que alcanza la misma velocidad terminal de asentamiento que la partícula de interés y que está determinado por la forma y densidad de la partícula. De acuerdo con esto, puede ser clasificadas como finas y gruesas, como se presenta en la Fig. 1[3].

Fig. 1: Distribución típica de las partículas en la atmósfera.



[1] FANG, Guor-Cheng *et al.*, "Characterization of particulate, metallic elements of TSP, PM2.5 and PM2.5-10 aerosols at a farm sampling site in Taiwan, Taichung," *Sci. Total Environ.*, Jun. 2003, vol. 308, no. 1–3, pp. 157–166, doi: 10.1016/S0048-9697(02)00648-4.
[2] BILLET. Sylvain *et al.*, "Ambient particulate matter (PM2.5): physicochemical characterization and metabolic activation of the organic fraction in human lung epithelial cells (A549)," *Environ. Res.*, Oct. 2007, vol. 105, no. 2, pp. 212–223, doi: 10.1016/J.ENVRES.2007.03.001.
[3] ARCINIÉGAS, Cesar S., "Diagnóstico y control de material particulado: partículas suspendidas totales y fracción respirable PM10," 2012, doi: 10.17151/luaz.2012.34.12.

Las partículas finas son las que presentan diámetro aerodinámico inferior a 2,5 µm (PM2.5). Estas partículas son tan pequeñas que pueden ser detectadas solo con un microscopio electrónico. Las fuentes de las partículas finas incluyen todo tipo de combustiones, incluidos los vehículos automóviles, plantas de energía, la quema residencial de madera, incendios forestales, quemas agrícolas y algunos procesos industriales. Las partículas gruesas son aquellas cuyo diámetro aerodinámico está entre 2,5 y 10 µm (PM2.5 y PM10, respectivamente). Entre las fuentes de las partículas gruesas se encuentran las operaciones de trituración o molienda y el polvo levantado por los vehículos que circulan en las carreteras[4].

1.1.2. Sistema de Vigilancia de Calidad del Aire (SVCA Tipo III)

Un Sistema de Vigilancia de Calidad del Aire (SVCA) se puede definir como un conjunto de equipos de monitoreo de los contaminantes atmosféricos, que se encuentran instalados en un lugar de interés con un propósito determinado; esta infraestructura debe estar acompañada de todas las actividades necesarias para su correcto funcionamiento, dentro de las cuales se puede mencionar la operación por personal calificado, programas de mantenimiento preventivo y correctivo, un sistema de administración de información que permita una correcta validación de los datos, entre otras. Su implementación se realiza bien sea cuando se presenta una problemática específica relacionada con la calidad del aire, o en zonas con población por encima de los cincuenta mil habitantes. Los criterios para su diseño y operación se encuentran en el Protocolo para el Monitoreo y Seguimiento de la Calidad del Aire[5].

El SVCA en el área metropolitana de Bucaramanga se puede clasificar como categoría III, ACLS intermedio, que se caracteriza por una población entre 500 000 y 1 500 000 habitantes. En cuanto a los parámetros a medir, se determina el monitoreo de las fracciones de materia particulada PM10 y PM2.5.

La protección del medio ambiente atmosférico incluye el control de emisión de contaminantes y el monitoreo de los niveles de emisión, es decir, de su concentración en aire ambiente. Para monitorear estos niveles se utilizan las Redes de Monitoreo de la Calidad del Aire. Estas redes se encargan de vigilar las concentraciones de contaminantes criterio (material particulado, O_3, SO_2, NO_2 y CO), de carbono negro y variables meteorológicas (precipitación, temperatura, presión atmosférica, radiación solar, velocidad y dirección del viento), por medio de estaciones de monitoreo instaladas en ubicaciones específicas de un territorio. Otros objetivos de las Redes de Monitoreo son:

- Localizar áreas con problemas de contaminación y comprender la evolución temporal y espacial.

[4] “Contaminantes del aire: Materias particuladas.” http://www.murciasalud.es/pagina.php?id=244308&# (accessed Oct. 04, 2021).
[5] “Monitoreo aire - IDEAM.” http://www.siac.gov.co/monitoreoaire (accessed Oct. 04, 2021).

- Dar cumplimiento a la legislación de protección del aire atmosférico.
- Obtener la información necesaria para definir un plan de acción marcado por directivas europeas u otros estándares internacionales en caso de sobrepasar el umbral de alarma.

Las Redes de Monitoreo de la Calidad del Aire permiten la medición, operación y análisis predictivo de la evolución de la contaminación del aire en diferentes áreas.

1.1.3. Resolución 2254 de 2017

El objeto y ámbito de aplicación de esta Resolución es el siguiente: "establece la norma de calidad del aire o nivel de inmisión y adopta disposiciones para la gestión del recurso aire en el territorio nacional, con el objeto de garantizar un ambiente sano y minimizar el riesgo sobre la salud humana que pueda ser causado por la exposición a los contaminantes en la atmósfera."

La Resolución establece los niveles máximos permisibles a condiciones de referencia de contaminantes criterio, tal como lo muestra la Tabla 1. Sus principales objetivos son el de crear gestión preventiva, estándares progresivos y fortalecer la medición de calidad del aire, lo cual es algo inminente, debido a que 15681 muertes fueron atribuidas a la mala calidad del aire en 2017, principalmente por EIC (Enfermedad Isquémica del Corazón) y EPOC (Enfermedad Pulmonar Obstructiva Crónica)[6], como lo muestra la Tabla 2.

Tabla 1: Niveles máximos permisibles de contaminantes criterio.

Contaminante	Nivel máximo Permisible ($\mu m/m^3$)	Tiempo de Exposición
PM_{10}	50	Anual
	100	24 horas
$PM_{2.5}$	25	Anual
	50	24 horas
SO_2	50	24 horas
	100	1 hora
NO_2	60	Anual
	200	1 hora
O_3	100	8 horas
CO	5000	8 horas
	35000	1 hora

[6] "Informe Carga de Enfermedad Ambiental en Colombia." https://www.ins.gov.co/Noticias/Paginas/Informe-Carga-de-Enfermedad-Ambiental-en-Colombia.aspx (accessed Oct. 04, 2021).

Tabla 2: Muerte atribuibles por factor de riesgo ambiental aire.

Enfermedad	Muertes Atribuibles
EIC	7230
EPOC	3873
IRAB	1809
ECV	1929
Cáncer de Pulmón	840
Cataratas	0
Totales	15681

1.1.4. Índice de Calidad del Aire (ICA)

El Índice de Calidad del Aire (ICA) es un valor adimensional asociado a un código de colores para reportar el estado de la calidad del aire al que están asociados unos efectos generales que deben ser tenidos en cuenta para reducir la exposición a altas concentraciones por parte de la población. Este índice también es utilizado en el pronóstico de la calidad del aire (Resolución 2254 de 2017). El ICA se calcula para los seis contaminantes criterios contemplados en la normativa nacional (Ozono-O_3, Material Particulado menor a PM10, Material Particulado menor a PM2.5, Monóxido de Carbono-CO, Dióxido de Azufre-SO_2 y Dióxido de Nitrógeno-NO_2) en tiempos de exposición que oscilan entre 1 hora y 24 horas de acuerdo con los puntos de corte establecidos en la Resolución 2254 de 2017[7].

El ICA para Colombia busca determinar el estado de la calidad del aire relacionado con los riesgos a la salud a los que puede estar expuesta la población, la presión sobre los ecosistemas y las acciones preventivas a tomar. El ICA, adoptado por el Ministerio de Ambiente y Desarrollo Sostenible (MADS), mediante la Resolución 2254 de 2017, es una metodología desarrollada por la Agencia de Protección Ambiental de los Estados Unidos (Technical Assistance Document for the Reporting of Daily Air Quality the Air Quality Index (AQI) – Documento EPA-454/B-09-001 de febrero de 2009) que permite evaluar rápidamente los efectos a la salud y las medidas preventivas a implementar con base en la concentración de un contaminante específico.

El ICA corresponde a una numérica a la cual se le asigna un color, el cual a su vez tiene una relación con los efectos sobre la salud. El índice es adimensional y su valor oscila entre 0 y 500, y está asociado a un código de colores para reportar el estado de la calidad del aire al que están asociados unos efectos generales que

[7] M. y E. A. IDEAM, Instituto de hidrología, "Índice de calidad del aire (ICA)," no. 571, p. 2013, 2012, [Online]. Available: http://www.ideam.gov.co/documents/11769/641368/2.01+HM+Indice+calidad+aire.pdf/5130ffb3-a1bf-4d23-a663-b4c51327cc05.

deben ser tenidos en cuenta para reducir la exposición a altas concentraciones de contaminantes por parte de la población.

En la Tabla 3 se presenta la descripción general del Índice de Calidad del Aire.

Tabla 3: Descripción general del Índice de Calidad del Aire.

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

En la Tabla 4 se muestran las acciones preventivas de acuerdo con la categoría del ICA.

Tabla 4: Acciones preventivas recomendadas según la categoría del ICA.

VERDE (Buena)	Ninguna
AMARILLO (Moderada)	
ANARANJADO (Dañina a la salud para grupos sensibles)	Niños activos, adultos y personas con enfermedades respiratorias como asma, deben reducir la actividad física al aire libre
ROJO (Dañina a la salud)	Niños y adultos activos y personas con enfermedades respiratorias, deben reducir la actividad física prolongada al aire libre, especialmente los niños.
MORADO (Muy dañina a la salud)	Niños activos, adultos y personas con enfermedades respiratorias como asma, deben reducir la actividad física al aire libre; especialmente los niños
MARRÓN (Peligrosa)	Todas las personas deben evitar el esfuerzo al aire libre

1.1.4.1. Cálculo del ICA

Según los datos recopilados de los Sistemas de Vigilancia de Calidad del Aire es posible determinar el valor del ICA en base al artículo 21 de la Resolución 2254 de 2017. Para determinar el Índice de Calidad del Aire se emplea la siguiente ecuación:

$$ICA_P = \frac{I_{Alto} - I_{Bajo}}{PC_{Alto} - PC_{Bajo}} * \left(C_P - PC_{Bajo}\right) + I_{Alto} \quad \text{(Ec. 1)}$$

donde:

- ICA_P: Índice de Calidad del Aire para el contaminante *P*.
- C_P: Concentración medida para el contaminante *P*.
- PC_{Alto}: Punto de corte mayor o igual a C_P.
- PC_{Bajo}: Punto de corte menor o igual a C_P.
- I_{Alto}: Valor del ICA correspondiente al PC_{Alto}.
- I_{Bajo}: Valor del ICA correspondiente al PC_{Bajo}.

1.2. Completar datos faltantes de Material Particulado

1.2.1. Series Temporales

Una serie temporal se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados. Si los datos se recogen en instantes temporales de forma continua, se debe o bien digitalizar la serie, es decir, recoger solo los valores en instantes de tiempo equiespaciados, o bien acumular los valores sobre intervalos de tiempo. La característica fundamental de las series temporales es que las observaciones sucesivas no son independientes entre sí, y el análisis debe llevarse a cabo teniendo en cuenta el orden temporal de las observaciones. Los métodos estadísticos basados en la independencia de las observaciones no son válidos para el análisis de series temporales porque las observaciones en un instante de tiempo dependen de los valores de la serie en el pasado[8].

El estudio de una serie temporal permite analizar, describir y explicar un fenómeno en el tiempo y extraer consecuencias para la toma de decisiones. Uno de los principales objetivos del estudio de series temporales es la predicción de valores futuros de la serie a partir de sus valores observados. En este trabajo los valores

[8] "Series Temporales: Introducción," Accessed: Oct. 04, 2021. [Online]. Available: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema7.pdf.

observados son las concentraciones de Material Particulado PM10 y PM2.5 recopiladas en los Sistemas de Vigilancia de Calidad del Aire.

Para completar los datos faltantes de concentración de Material Particulado se emplean modelos de Redes Neuronales. Para ello, se hace uso de la Toolbox Machine Learning de MATLAB, la cual cuenta con técnicas diseñadas e implementadas por los ingenieros de MATLAB.

Las técnicas implementadas y comparadas para el proceso de completado de datos son:

- NAR (*Nonlinear autoregressive neural network*)
- NARX (*Nonlinear autoregressive network with exogenous inputs*)
- Nonlinear Input-Output

Las Redes Neuronales dinámicas son generalmente más poderosas que las estáticas (aunque un poco más difíciles de entrenar). Dado que las redes dinámicas tienen memoria, pueden ser entrenadas para aprender patrones secuenciales o variables en el tiempo. Esto tiene aplicaciones en áreas tan dispares como la predicción en mercados financieros, ecualización de canales en sistemas de comunicación, detección de fase en sistemas de potencia, clasificación, detección de fallas, reconocimiento de voz, entre otras[9].

Las Redes Neuronales funcionan mejor en análisis predictivo debido a las capas ocultas. Los modelos de regresión lineal usan solo nodos de entrada y salida para hacer predicciones. La red neuronal también usa la capa oculta para hacer predicciones más precisas. Eso es porque “aprende” de forma similar a como lo hace un humano.

Es muy difícil saber cuál algoritmo de entrenamiento será el más rápido para un problema dado. Esto depende de muchos factores, incluyendo la complejidad del problema, el número de datos en el conjunto de Training, el número de pesos y sesgos en la red, el error objetivo, y si la red es empleada para reconocimiento de patrones (análisis discriminante) o para regresión[10]. Como la naturaleza de las series de tiempo analizadas en este trabajo no está definida, se comparan múltiples técnicas de series de tiempo utilizando Redes Neuronales dinámicas, incluidas las redes con retroalimentación.

El software Deep Learning Toolbox está diseñado para entrenar una clase de red llamada Red Dinámica Digital en Capas (*Layered Digital Dynamic Network, LDDN*). Cualquier red que se pueda configurar en la forma de una LDDN puede ser entrenada con la Toolbox. A continuación, se presenta una descripción básica de las LDDN[9]:

---

[9] “How Dynamic Neural Networks Work - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ug/how-dynamic-neural-networks-work.html (accessed Oct. 05, 2021).

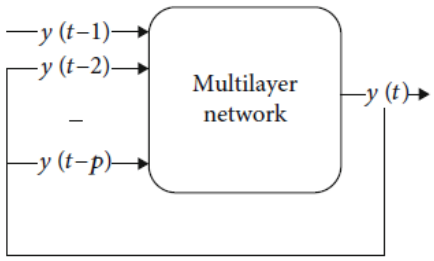
[10] “Choose a Multilayer Neural Network Training Function - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ug/choose-a-multilayer-neural-network-training-function.html (accessed Oct. 05, 2021).

- Conjunto de matrices de peso que entran a la capa (que pueden estar conectadas a otras capas o a entradas externas), regla de función de peso asociada utilizada para combinar la matriz de peso con su entrada (normalmente es una multiplicación de matrices estándar), y línea de retraso asociada.
- Vector de sesgo.
- Regla de función de entrada neta empleada para combinar las salidas de las distintas funciones de peso con el sesgo para producir la entrada neta (normalmente una unión de suma).
- Función de transferencia.

1.2.2. NAR (*Nonlinear autoregressive neural network*)

La Red Neuronal Autorregresiva No Lineal (NAR), como lo muestra la Fig. 2, puede ser entrenada para predecir series temporales a partir de valores pasados de esa misma serie $Y(t-1)$, $Y(t-2)$, …, $Y(t-d)$ llamados retrasos de retroalimentación (*feedback delays*), donde d es el parámetro de retardo (*delay parameter*).

Fig. 2: Arquitectura simplificada de una Red Neuronal Autorregresiva No Lineal (NAR).



La red es creada y entrenada en un lazo abierto, usando datos reales como respuesta. Posterior al entrenamiento, la red es convertida en un lazo cerrado y los valores predichos son utilizados como nuevas entradas a la red. Una NAR aplicada a predicción de series describe un modelo discreto, no lineal y autorregresivo, que puede ser escrito de la siguiente forma:

$$Y_t = h(Y_{t-1}, Y_{t-2}, \dots, Y_{t-d}) + \varepsilon_t \quad \text{(Ec. 2)}$$

La función $h()$ es desconocida de antemano, y el entrenamiento de la red está enfocado en aproximar la función por medio de la optimización de los pesos de la red y los sesgos (*bias*) de cada neurona.

Por tanto, un modelo NAR es definido precisamente por una ecuación del tipo:

$$Y_t = \alpha_0 + \sum_{j=1}^{k} \alpha_j \phi \left(\sum_{i=1}^{a} \beta_{ij} Y_{t-i} + \beta_{0j} \right) + \varepsilon_t \quad \text{(Ec. 3)}$$

donde a es el número de entradas, k es el número de capas ocultas con función de activación ϕ, y β_{ij} es el parámetro correspondiente al peso de la conexión entre la unidad de entrada i y la unidad oculta j, α_j es el peso de la conexión entre la unidad oculta j y la unidad de salida, y β_{0j} y α_0 son constantes que corresponden, respectivamente, a la unidad oculta j y la unidad de salida[11].

1.2.3. NARX (*Nonlinear autoregressive network with exogenous inputs*)

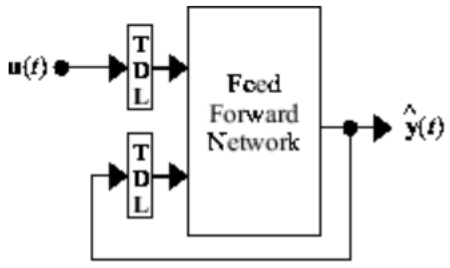
La Red Neuronal No Lineal Autorregresiva con Entradas Exógenas (NARX) es una red dinámica recurrente, con conexiones de retroalimentación que encierran varias capas de la red. El modelo NARX es basado en el modelo lineal ARX, que comúnmente es utilizado en el modelado de series temporales. La ecuación definitoria del modelo NARX es la siguiente:

$$y(t) = f\left(y(t-1), y(t-2), \ldots, y\left(t-n_y\right), u(t-1), u(t-2), \ldots, u(t-n_u)\right) \quad \text{(Ec. 4)}$$

donde el siguiente valor de la señal de salida dependiente $y(t)$ se hace depender de los valores anteriores de la señal de salida y de los valores anteriores de una señal de entrada independiente (exógena). Es posible implementar un modelo NARX al utilizar una red neuronal con retroalimentación para aproximar la función f[12].

La Fig. 3 muestra la arquitectura simplificada de una Red Neuronal No Lineal Autorregresiva con Entradas Exógenas.

Fig. 3: Arquitectura simplificada de una Red Neuronal No Lineal Autorregresiva con Entradas Exógenas (NARX).



[11] BENRHMACH, G., NAMIR, K., NAMIR, A. and BOUYAGHROUMNI, J., "Nonlinear Autoregressive Neural Network and Extended Kalman Filters for Prediction of Financial Time Series," *J. Appl. Math.*, 2020, vol. 2020, doi: 10.1155/2020/5057801.

[12] "Design Time Series NARX Feedback Neural Networks - MATLAB & Simulink - MathWorks América Latina." https://la.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html (accessed Oct. 05, 2021).

1.2.4. Nonlinear Input-Output

En este modelo se predice la serie $y(t)$ dados valores pasados de una serie $x(t)$. La ecuación definitoria de este modelo es:

$$y(t) = f\big(x(t-1), \dots, x(t-d)\big) \quad \text{(Ec. 5)}$$

donde d es el parámetro de retardo.

En la Fig. 4 se muestra la arquitectura simplificada de una Red Neuronal Nonlinear Input-Output.

Fig. 4: Arquitectura simplificada de una Red Neuronal Nonlinear Input-Output.

1.2.5. Trainlm

Trainlm es una función de entrenamiento de red que actualiza los valores de los pesos y los sesgos acorde con la optimización de Levenberg-Marquardt. *Trainml* es comúnmente el algoritmo de *backpropagation* más rápido en la Deep Learning Toolbox de Matlab, y es altamente recomendado como una primera opción de algoritmo supervisado, aunque requiere más memoria que otros algoritmos[13].

La función de rendimiento de la red que se utiliza es el error cuadrático medio. Mide el rendimiento de la red de acuerdo con la media de errores cuadráticos. Es un promedio de los cuadrados de las diferencias entre las observaciones reales y las predichas. *Trainlm* puede entrenar cualquier red siempre que sus funciones de peso, entrada neta y transferencia tengan funciones derivadas.

1.2.6. Trainbr

Trainbr es una función de entrenamiento de red que actualiza los valores de pesos y los sesgos acorde con la optimización de Levenberg-Marquardt. Minimiza una

[13] "Levenberg-Marquardt backpropagation - MATLAB trainlm - MathWorks América Latina." https://la.mathworks.com/help/deeplearning/ref/trainlm.html (accessed Oct. 05, 2021).

combinación de errores cuadrados y pesos, y luego determina la combinación correcta que hace que la red tenga buena generalización. El proceso es denominado regularización Bayesiana[14].

1.2.7. Trainscg

Trainscg es una función de entrenamiento de red que actualiza los valores de pesos y sesgos según el método del gradiente conjugado escalado[15].

1.3. Procesamiento de Lenguaje Natural (NLP)

El Procesamiento de Lenguaje Natural (NLP) se refiere a la rama de la ciencia computacional – y más específicamente, a la rama de la Inteligencia Artificial – que se enfoca en dotar a los computadores con la habilidad de entender texto y las palabras habladas de una manera similar en que los humanos lo hacen. NLP combina la lingüística computacional – modelado del lenguaje humano basado en reglas – con modelos estadísticos, de Machine Learning y Deep Learning. Juntas, estas tecnologías permiten a los computadores procesar el lenguaje humano en la forma de texto o datos de voz para "entender" su significado, con la intención y sentimiento del hablante o escritor. NLP impulsa los programas computacionales que traducen texto de un lenguaje a otro, responden comandos hablados, y resumen grandes volúmenes de texto[16].

Dada la complejidad intrínseca que acompaña a cualquier proceso de NLP, muchas de las técnicas automatizadas están relacionadas con el uso de la Inteligencia Artificial y las ciencias cognitivas. Para aplicar las técnicas más comunes en NLP, la librería NLTK (*Natural Language Tool Kit*) permite que cualquier programa escrito en lenguaje Phyton pueda invocar a un amplio conjunto de algoritmos que sustentan las principales técnicas de NLP para la generación de métricas, frecuencia de términos, polaridad de frases y textos, entre otras[17].

---

[14] "Bayesian regularization backpropagation - MATLAB trainbr - MathWorks América Latina." https://la.mathworks.com/help/deeplearning/ref/trainbr.html?s_tid=doc_ta (accessed Oct. 05, 2021).

[15] "Scaled conjugate gradient backpropagation - MATLAB trainscg - MathWorks América Latina." https://la.mathworks.com/help/deeplearning/ref/trainscg.html?searchHighlight=trainscg&s_tid=srchtitle (accessed Oct. 05, 2021).

[16] "What is Natural Language Processing? | IBM." https://www.ibm.com/cloud/learn/natural-language-processing#toc-what-is-na-jLju4DjE (accessed Oct. 05, 2021).

[17] "Procesamiento del lenguaje natural con NLTK para Ingeniería social automatizada – Seguridad en Sistemas y Técnicas de Hacking. TheHackerWay (THW)." https://thehackerway.com/2015/02/17/procesamiento-del-lenguaje-natural-con-nltk-para-ingenieria-social-automatizada/ (accessed Oct. 05, 2021).

Existen algunos términos comunes en NLP que se deben comprender antes de poder aplicar cualquier técnica y entender los resultados que arrojan:

- **Token:** Se trata de la unidad más simple de procesamiento y representa una palabra en el texto.
- **Sentencia:** Secuencia ordenada de tokens.
- **Tokenización:** Se trata del proceso de segmentar una sentencia en cada uno de los tokens que la componen.
- **Corpus:** Cuerpo del mensaje que se encuentra compuesto por un conjunto de sentencias.
- **Part-of-speech (POS):** Dependiendo de la semántica del lenguaje, cada token que compone una sentencia puede ser un verbo, un adjetivo, un pronombre, un artículo, etc. Un POS es una clasificación para cada token dentro de una sentencia, de manera que sea posible identificar el significado de cada token y las partes clave de cada sentencia.
- **Árbol:** Todos los textos están compuestos por varias sentencias, y cada sentencia tiene varios tokens con sus respectivos POS. Un árbol parseado incluye cada una de las dependencias de las sentencias y cada parte del texto. Es una forma de ordenar cada uno de los elementos del "corpus" de una forma que sea fácil de consultar.

1.3.1. Minería de texto: Extracción de Tweets

La minería de texto es una combinación de enfoques cuantitativos y cualitativos cuyo análisis permite establecer claras respuestas en virtud del análisis de textos objetivos y su posible relación a enfoques de planeación, políticos, regulatorios, entre otros. La minería de texto utiliza métodos de NLP que permite el tratamiento de textos no estructurados; generalmente comprende las siguientes etapas:

- Selección
- Preprocesamiento
- Transformación
- Minería de datos
- Interpretación
- Evaluación

Posteriormente, se emplean las siguientes técnicas:

- Remoción de ruido
- Determinación de características para representación de datos
- Reducción de la dimensionalidad
- Análisis de datos

La definición de términos más frecuentes puede ser realizada mediante la aplicación del método de análisis semántico latente que se basa en el modelo de bolsa de palabras, en el cual se analiza la estructura del conjunto de datos en términos de sus componentes principales.

Por otra parte, en el análisis de textos se emplea además el coeficiente de correlación al aplicar métodos jerárquicos de análisis de clúster de textos. Este permite establecer el grado de asociación de términos en virtud del proceso de agrupación previa de términos analizados. Al poseer un mayor coeficiente, el modelo presentará menor distorsión de las relaciones originales existentes.

Las herramientas que se compararon para la extracción de Tweets son:

- **API Twitter**

Twitter admite el acceso a sus servidores a través de API, lo que permite a las personas desarrollar software que se integra con Twitter como una solución para ayudar a las empresas a responder a los comentarios de los clientes en Twitter.

- **Octoparse**

Octoparse es un moderno software de extracción de datos de redes visuales. Puede ser usado para extraer información en lotes de sitios web. La mayoría de las tareas de raspado no requieren codificación.

- **Twint**

Twint es una herramienta avanzada de raspado de Twitter escrita en Phyton que permite raspar Tweets de perfiles de Twitter sin usar su API. Utiliza los operadores de búsqueda de Twitter para permitirle raspar Tweets de usuarios específicos, raspar Tweets relacionados con ciertos temas, hashtags y tendencias, o clasificar información confidencial como correos electrónicos y números de teléfono.

1.3.2. Data Wrangling para Tweets

Limpieza de Tweets. Para ello se crean dos funciones: una para eliminar URLS, menciones y hashtags, y otra para limpiar el texto restante – eliminar palabras

vacías y puntuación-. En este paso, se encuentran las palabras más frecuentes en los datos, extrayendo información sobre su contenido y temas. Por último, se eliminan las palabras vacías ya que aparecen con frecuencia, pero no son muy informativas.

1.4. Técnicas de Inteligencia Artificial (Machine Learning)

Las técnicas de análisis de datos permiten estudiar el comportamiento de diferentes procesos, para aplicar diferentes acciones con el fin de tener mejores resultados, a partir de gráficas que permiten facilidad de comprensión en cualquier situación que se presente. Hay una gran variedad de técnicas, entre las que se pueden encontrar la visualización de datos, la minería de datos y Redes Neuronales. La visualización de datos es una representación con elementos visuales como gráficos o mapas para detectar tendencias o patrones significativos respecto a algún proceso. A medida que el Big Data va avanzando, esta estrategia permite resaltar los datos que tienen información útil para una correcta toma de decisiones, sin embargo, más allá del gráfico, debe haber un equilibrio entre forma y función, es decir, que los datos que van llegando tengan relación con la gráfica que se está generando.

De igual manera, la minería de datos busca detectar información de los grandes volúmenes de datos que llegan constantemente para deducir tendencias o patrones repetitivos que determinen el comportamiento de un proceso. Esta técnica se considera un proceso dinámico e iterativo que hace uso de métodos estadísticos, y, en ocasiones, requiere de la Inteligencia Artificial para realizar su objetivo. Este proceso se puede desglosar en 4 etapas. En primer lugar, está la determinación de objetivos. Este delimita la información respecto a lo que se busca analizar. En segundo lugar, el preprocesamiento de datos actúa como filtro de la información que se considera importante. En tercer lugar, se encuentra la determinación del modelo. En esta parte se realizan los análisis estadísticos o desarrollo de algoritmos para dar paso a la última fase, la cual es el análisis de resultados, en la cual se valida la coherencia de estos con el fin de obtener respuestas que permitan considerar ciertas decisiones.

Las Redes Neuronales son una técnica de la Inteligencia Artificial empleadas para resolver problemas referentes al monitoreo del sistema, los cuales mantienen errores que se determinan por medio de la desviación absoluta media, y el error porcentual absoluto medio, que se basan en la diferencia entre el modelo y los valores observados, cuyo propósito es centrarse en la capacidad de predecir resultados para tener un sistema de gran eficiencia como lo es el uso del Deep Learning.

A medida que la contaminación del aire se vuelve cada vez más grave, la predicción de la calidad del aire se ha convertido en un enfoque importante para la gestión y prevención de la contaminación del aire. En los últimos años, se han propuesto

varios métodos para predecir la calidad del aire, como los métodos deterministas, métodos estadísticos y métodos de aprendizaje automático. Sin embargo, estos métodos tienen algunas limitaciones. Los primeros requieren cálculos costosos y conocimientos específicos para la identificación de parámetros, mientras que el rendimiento del pronóstico de los métodos estadísticos es limitado debido al supuesto lineal y al problema de la multicolinealidad. La mayoría de los métodos de aprendizaje automático, por otro lado, no pueden capturar los patrones de series de tiempo o aprender de las dependencias a largo plazo de las concentraciones de contaminantes del aire. Además, hay una falta de métodos que sean capaces de generar una alta precisión de predicción para el pronóstico de la calidad del aire. Según el marco de la metodología se hace una comparación con dos modelos diferentes de Machine Learning: uno clásico de Análisis Discriminante y el otro de Redes Neuronales. Para implementar estos modelos, los datos tienen que pasar por las siguientes etapas:

- **Exploración/Transformación**

Los datos nunca llegan limpios. Por eso es importante detectar si existen anomalías, limpiarlas y preparar el conjunto de datos para que el análisis no se sesgue o arroje resultados erróneos. En esta fase se analiza cómo se distribuyen los datos, se calculan variables estadísticas descriptivas como la media, varianza y desviación estándar, y se transforma el conjunto para que los modelos predictivos de clasificación puedan operar sobre él.

- **Agregado/Etiquetado**

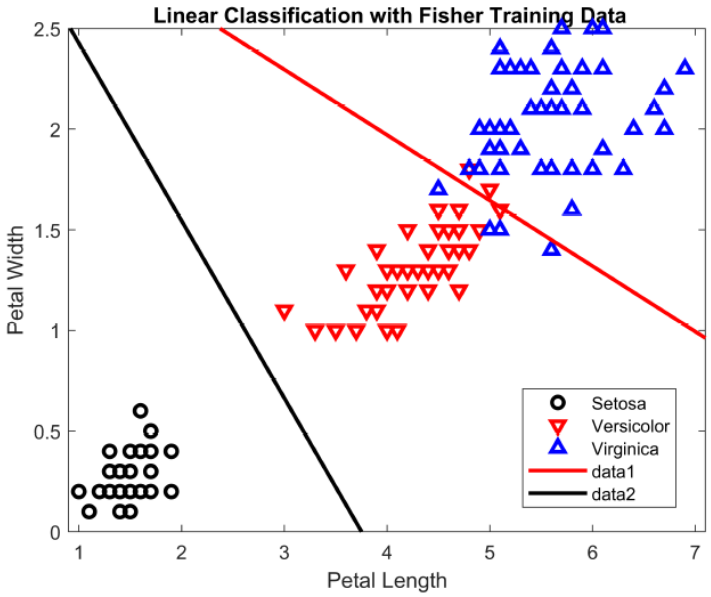
Con un conjunto de datos limpio y con una idea general de su comportamiento, es posible realizar un análisis más a fondo. Se pueden segmentar los datos, definir y obtener métricas, generar agregaciones, extraer características y prepararlos para el aprendizaje automático. También es momento de generar visualizaciones para facilitar la detección de patrones, correlaciones y relaciones de causalidad que estaban escondidas en los datos crudos.

1.4.1. Análisis Discriminante

El Análisis Discriminante es un método de clasificación. Asume que diferentes clases generan datos basados en diferentes distribuciones gaussianas. El Análisis Discriminante Lineal (*Linear Discriminant Analysis, LDA*) es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son

conocidos a priori, y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y = k\,|X = x)$. Finalmente, se asigna la observación a la clase k para la que la probabilidad predicha es mayor[18]. En la Fig. 5 se muestra una aplicación del LDA en la cual se clasifican 3 tipos de Iris – Setosa, Versicolor y Virginica – en base al ancho y largo de sus pétalos.

Fig. 5: Clasificación lineal de 3 tipos de Iris en base a 2 variables: ancho y largo del pétalo.



1.5. Algoritmos de selección de características

La selección de características reduce la dimensionalidad de los datos seleccionando únicamente un subconjunto de características medidas (variables predictoras) para crear un modelo. Los algoritmos de selección de características buscan un subconjunto de predictores que modelen óptimamente las respuestas medidas, sujeto a restricciones tales como variables requeridas o excluidas y al tamaño del subconjunto. Los principales beneficios de la selección de características son mejorar el rendimiento de predicción, proporcionar predictores

[18] "Análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA)." https://www.cienciadedatos.net/documentos/28_linear_discriminant_analysis_lda_y_quadratic_discriminant_analysis_qda (accessed Oct. 05, 2021).

más rápidos y rentables, y brinda un mejor entendimiento del proceso de generación de datos[19].

Es posible categorizar los algoritmos de selección de selección de características en tres tipos:

- **Selección de características tipo filtro (*Filter Type*):** Mide la importancia de las variables basado en sus características, dales como la varianza de la variable y su relevancia a la respuesta. La selección de características tipo filtro no está relacionada con el algoritmo de entrenamiento.
- **Selección de características tipo "envoltura" (*Wrapper Type*):** Los algoritmos de selección de características de este tipo empiezan entrenando un subconjunto de variables y luego agregan o remueven empleando un criterio de selección. Este criterio directamente mide el cambio en el rendimiento del modelo que resulta de añadir o remover una variable. El algoritmo repite el entrenamiento y mejora el modelo hasta que el criterio de parada sea satisfecho.
- **Selección de características tipo embebido (*Embedded Type*):** Este tipo de algoritmos aprenden la importancia de las variables como parte del proceso de entrenamiento del modelo. Una vez se entrena un modelo, se obtiene la relevancia de las características en el modelo entrenado. Estos algoritmos seleccionan variables que trabajan bien con un proceso de aprendizaje particular.

La selección de características aplicada de forma adecuada no solo reduce la dimensión de las características, y, por tanto, la cantidad de datos utilizados en el aprendizaje, sino que también alivia los efectos de la maldición de la dimensionalidad para mejorar el rendimiento de la generalización del algoritmo. Además, también incrementa la velocidad de ejecución y la interpretabilidad del modelo.

La selección de características es preferible a la transformación de características cuando las variables originales y sus unidades son importantes, y el objetivo del modelado es identificar un subconjunto influyente. Cuando características categóricas están presentes, y la transformación numérica es inapropiada, la selección de características se convierte en la vía primaria de reducción de dimensionalidad[20].

1.5.1. FSCNCA (*Feature Selection Using Neighborhood Component Analysis*)

[19] "Introduction to Feature Selection - MATLAB & Simulink - MathWorks América Latina." https://la.mathworks.com/help/stats/feature-selection.html (accessed Oct. 06, 2021).

[20] YANG, W., WANG, K., and ZUO, W. "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, 2012, vol. 7, no. 1, pp. 162–168, doi: 10.4304/JCP.7.1.161-168.

El análisis de componentes de vecindad (NCA) es un método no paramétrico para la selección de características con el objetivo de maximizar la precisión de predicción de los algoritmos de regresión y clasificación. La función *fscnca* del software Statistics and Machine Learning Toolbox de MATLAB ejecuta la selección de características mediante NCA con regularización para aprender los pesos de las variables para la minimización de una función objetivo que mide el error de "dejar uno afuera" de la clasificación o regresión de los datos de entrenamiento[21].

El algoritmo propuesto por Yang et al. [20] para la selección de características mediante el análisis de componentes de vecindad se describe a continuación:

Sea $T = \{(x_1, y_1, \dots, (x_i, y_i), \dots, (x_N, y_N))\}$ el conjunto de patrones de entrenamiento, donde $x_i$es una característica *d*-dimensional, el vector $y_i \in \{1, \dots, C\}$ su correspondiente etiqueta de clase y N el número de patrones. El objetivo es encontrar un vector de pesos w que seleccione el subconjunto de características que optimicen la clasificación del vecino más cercano (*nearest neighbor classification*). En términos del vector de pesos w, se denota la distancia ponderada entre dos patrones x_i y x_j de la siguiente manera:

$$D_w(x_i, x_j) = \sum_{l=1}^{d} \omega_l^2 |x_{il} - x_{jl}| \quad \text{(Ec. 6)}$$

donde ω_l es el peso asociado a la característica l.

Para que la clasificación del vecino más cercano tenga éxito, una estrategia es maximizar la precisión de la clasificación de "dejar uno afuera" (*leave-one-out classification*) del conjunto de entrenamiento T. Sin embargo, debido a que la precisión verdadera de "dejar uno afuera" que selecciona el vecino más cercano como punto de referencia de clasificación no es una función diferenciable, una aproximación efectiva es que se determine el punto de referencia utilizando una distribución de probabilidad. Aquí, la probabilidad que x_i seleccione a x_j como su punto de referencia se define así:

$$p_{ij} = \begin{cases} \frac{\kappa(D_w(x_i, x_j))}{\sum_{k \neq i} k(D_w(x_i, x_j))}, & si\ i \neq j \\ 0, & si\ i = j \end{cases} \quad \text{(Ec. 7)}$$

donde $\kappa(z) = \exp\left(\frac{-z}{\sigma}\right)$ es una función de núcleo y el ancho del núcleo σ es un parámetro de entrada que influencia la probabilidad de que cada punto sea seleccionado como punto de referencia. En particular, si $\sigma \to 0$, solo el vecino más cercano del patrón de muestra puede ser seleccionado como su punto de referencia. Por otra parte, si $\sigma \to \infty$, todos los puntos tienen la misma probabilidad de ser

[21] "Neighborhood Component Analysis (NCA) Feature Selection - MATLAB & Simulink - MathWorks América Latina." https://la.mathworks.com/help/stats/neighborhood-component-analysis.html (accessed Oct. 06, 2021).

seleccionados, excepto el punto de muestra. Basado en la definición anterior, la probabilidad de que el punto de muestra x_i sea clasificado correctamente está dada por:

$$p_i = \sum_j y_{ij} p_{ij} \quad \text{(Ec. 8)}$$

donde $y_{ij} = 1$ si y solo si $y_i = y_j$, y en caso contrario, $y_{ij} = 0$. Por tanto, la precisión aproximada de la clasificación de "dejar uno afuera" puede ser escrito así:

$$\xi(w) = \frac{1}{N}\sum_i p_i = \frac{1}{N}\sum_i \sum_j y_{ij} p_{ij} \quad \text{(Ec. 9)}$$

Con el fin de realizar la selección de características y aliviar el sobreajuste (*overfitting*), se introduce un término de regularización y por tanto se obtiene la siguiente función objetivo:

$$\xi(w) = \sum_i \sum_j y_{ij} p_{ij} - \lambda \sum_{l=1}^{d} \omega_l^2 \quad \text{(Ec. 10)}$$

donde $\lambda > 0$ es un parámetro de regularización que puede ser ajustado mediante *cross validation*. Como la función objetivo (Ec. 10) es diferenciable, su derivada con respecto a ω_l puede ser computada:

$$\frac{\partial \xi(w)}{\partial \omega_l} = 2\left(\frac{1}{\sigma}\sum_i \left(p_i \sum_{j \neq i} p_{ij} |x_{il} - x_{jl}| - \sum_j y_{ij} p_{ij} |x_{il} - x_{jl}|\right) - \lambda\right)\omega_l \quad \text{(Ec. 11)}$$

El uso de la derivada anterior conduce a la correspondiente ecuación de actualización de ascenso del gradiente.

1.5.2. Forward Selection

Forward Selection es un algoritmo de búsqueda ascendente que comienza con un conjunto vacío de variables y gradualmente añade características seleccionadas por alguna función de evaluación. En cada iteración, la variable a ser incluida en el conjunto es seleccionada del conjunto restante de características disponibles que anteriormente no ha sido empleada. Por lo tanto, el nuevo conjunto de características ampliado debería producir un error de clasificación mínimo en comparación con la adición de cualquier otra. Forward Selection es ampliamente usada por su simplicidad y velocidad de cómputo[22].

Los pasos seguidos por el algoritmo son los siguientes:

I. La característica más importante $S_1 = f_1$ es seleccionada primero utilizando algún criterio.

[22] MARCANO, A. C., QUINTANILLA, J. D., CORTINA M. G. J., and ANDINA, D. "Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network," *IECON Proc.* 2010 *(Industrial Electron. Conf.*, pp. 2845–2850, doi: 10.1109/IECON.2010.5675075.

II. Luego pares de características son formadas con f_i y el mejor par es seleccionado como $S_2 = \{f_i, f_j\}$.
III. Un conjunto de tres características es formado utilizando S_2 y el mejor conjunto de tres características es seleccionado como $S_3 = \{f_i, f_j, f_k\}$.
IV. Este proceso es repetido hasta que un número predefinido de características es seleccionado.

La desventaja de Forward Selection es que las nuevas características son añadidas continuamente en el conjunto de características seleccionadas. Este no brinda flexibilidad para remover aquellas características que ya han sido agregadas, en el caso de que se vuelvan obsoletas tras la adición de nuevas características[23].

[23] CHANDRA, B. "Gene Selection Methods for Microarray Data," *Appl. Comput. Med. Heal.*, Jan. 2016 pp. 45–78, doi: 10.1016/B978-0-12-803468-2.00003-5.

2. METODOLOGÍA

La metodología en la que se basa el proyecto es una unión y adaptación entre la metodología de Standford d. School Design Thinking Process y la metodología ETL (*Extract-Transform-Load*) que es muy utilizada en el análisis de datos porque cumple con las 3 funciones principales de un proyecto de ciencia de datos la cual consta de 3 etapas para cumplir con los objetivos.

En la Fig. 6 se muestra la metodología empleada para cumplir con el objetivo general de este proyecto.

2.1. Etapa 1: Problema

La empresa K2 Ingeniería desea determinar la posible relación entre la información publicada en Twitter y el comportamiento contaminante del aire registrado por las estaciones de monitoreo ubicadas en Bucaramanga y el Área Metropolitana.

2.2. Etapa 2: Definir

2.2.1. Diseño de la investigación

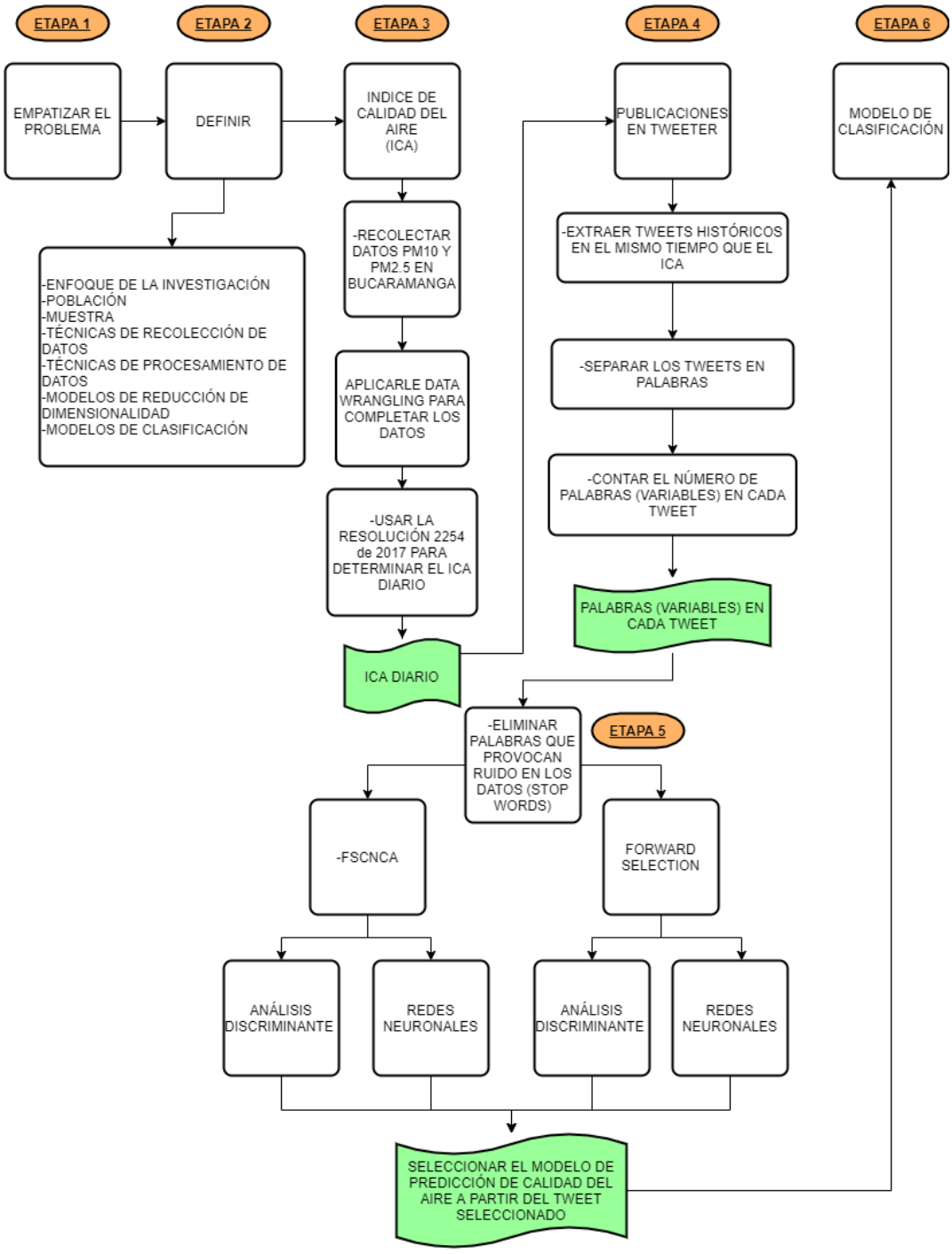
Dado que el objetivo de estudio será analizar la relación que tiene el ICA de la ciudad de Bucaramanga y el Área Metropolitana con la percepción de calidad del aire que tienen los habitantes de estos territorios, según sus publicaciones en Twitter, se recurre a un diseño de ciencia de datos del tipo ETL (*Extract-Transform-Load*). Se realiza una investigación mixta, la cual presenta un enfoque cualitativo que analiza las intenciones de dar la percepción sobre la calidad del aire en los textos de las publicaciones, y un enfoque cuantitativo que busca encontrar la correlación que hay entre la percepción dada en los textos publicados en internet y el valor real del ICA. Para ello, se realiza una investigación longitudinal.

La investigación longitudinal es un estudio de tipo observacional en el que se recopilan datos de la misma muestra repetidamente durante un periodo de tiempo prolongado. El estudio longitudinal puede durar algunos años, dependiendo de qué tipo de información se desea obtener. El beneficio de realizar un estudio longitudinal es que se puede detectar cualquier tipo de cambio que ocurra en las características de las variables muestreadas. Un aspecto importante es que el estudio longitudinal se extiende más allá de un solo marco en el tiempo. Como resultado, se establece una secuencia adecuada de los eventos ocurridos.

2.2.2. Enfoque de la investigación

El enfoque que se trabaja en este proyecto es del tipo mixto, debido a que es el que mejor se adapta a las características y necesidades de la investigación. El enfoque cualitativo recolecta y procesa las publicaciones haciendo uso de técnicas de Procesamiento de Lenguaje Natural (NLP) para la extracción de la percepción de la calidad del aire de las personas en Bucaramanga. Por otra parte, el enfoque cuantitativo utiliza la recolección y análisis de datos proporcionados por la empresa K2 Ingeniería y AMB para la clasificación del Índice de Calidad del Aire según lo dictamina la Resolución 2254 de 2017.

Fig. 6: Metodología.



2.2.3. Población

La población se define como el conjunto de todos los elementos estudiados, sobre los cuales se intenta sacar conclusiones. En este caso la población estará conformada por dos grupos:

I. Las personas de Bucaramanga con acceso a internet que publican temas relacionados con la calidad del aire.
II. Mediciones de PM10 y PM2.5 en la ciudad de Bucaramanga y el Área Metropolitana.

2.2.4. Muestra y muestreo

En este trabajo se utiliza el método de muestreo no probabilístico, en el cual se toman los datos que estén disponibles en un momento dado, puesto que se extraen los datos que estén al alcance en Twitter a partir de enero 1 de 2018. De la misma forma, se solicita a la empresa K2 Ingeniería y a AMB los datos históricos medidos de Material Particulado PM10 y PM2.5.

2.2.5. Técnicas de recolección de datos

La técnica de recolección de datos que se utiliza en el presente trabajo es la observación sistematizada aplicando técnicas de *Web Scraping*. Para extraer datos de Twitter se puede emplear la API, pero presenta el inconveniente de que no se pueden extraer datos más allá de 7 días anteriores a las solicitudes. Por esta razón, la única opción viable para la recolección de datos es la de hacer uso de técnicas de *Web Scraping*.

2.2.6. Instrumento de recolección de datos

Es cualquier herramienta con el que se puede extraer información. De este modo, el instrumento sintetiza en si toda la labor de la investigación, resume los aportes del marco teórico al seleccionar datos que corresponden a los indicadores y, por lo tanto, a las variables o conceptos utilizados. El instrumento será un script diseñado para recibir las direcciones de las páginas web que se deben extraer usando el lenguaje Phyton y las librerías especializadas en la extracción de información de páginas web.

2.2.7. Técnicas de procesamiento de datos

Las técnicas empleadas en el procesamiento de los datos serán la estadística descriptiva con técnicas de Machine Learning para el Procesamiento de Lenguaje Natural, y herramientas de análisis estadístico, que consisten en conjuntos de procedimientos que tienen por objeto presentar masas de datos por medio de tablas, gráficos y/o medidas de resumen.

2.2.8. Herramientas para el procesamiento de datos

Para llevar a cabo el procesamiento de los datos obtenidos se utilizan diversos softwares como lo son Octoparse, MATLAB y Phyton, con sus respectivas librerías o paquetes como lo son The Statistics and Machine Learning Toolbox de MATLAB y NLTK de Phyton, entre otras. Esta decisión fue tomada luego de realizar un análisis de funcionalidad.

2.3. Etapa 3: Índice de Calidad del Aire (ICA)

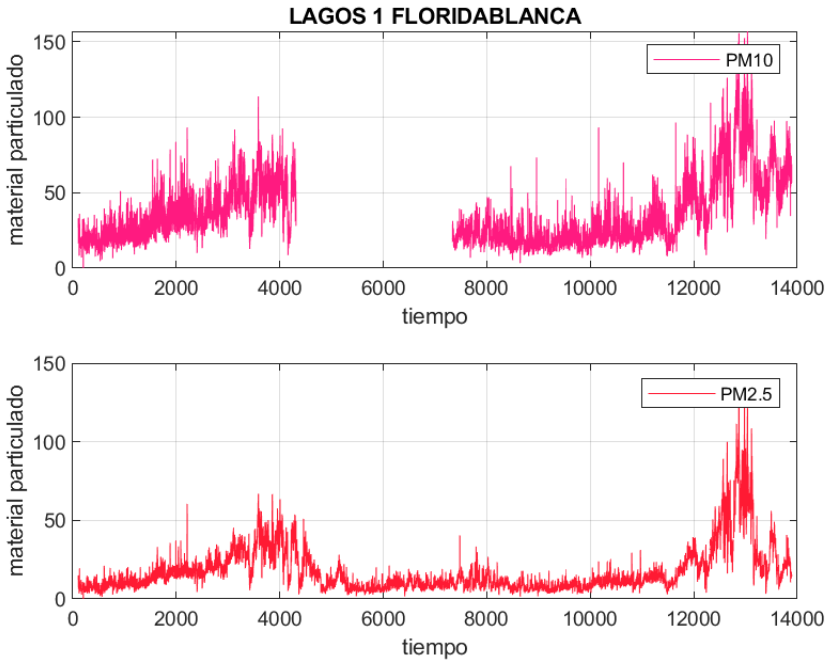
2.3.1. Datos recogidos por los Sistemas de Vigilancia de Calidad del Aire

Las Fig. 7 – 10 detallan los valores de Material Particulado PM10 y PM2.5 que fueron muestreados, hora a hora, desde octubre 1 de 2018 hasta octubre 15 de 2020.

2.3.1.1. Lagos 1 Floridablanca

Estación Lagos 1, Calle 29 No. 10-13 (Floridablanca).

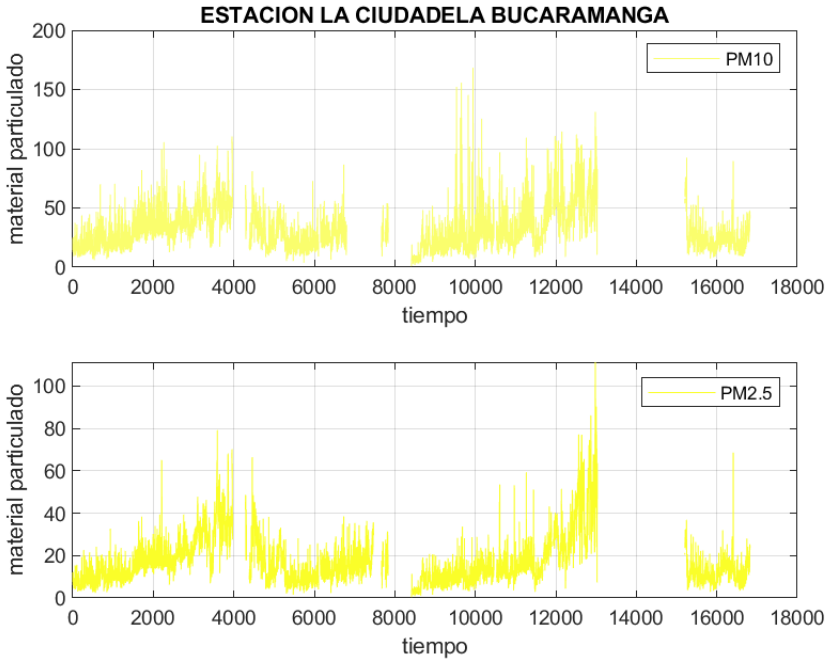
Fig. 7: PM10 y PM2.5 medido por la Estación Lagos 1, Calle 29 No. 10-13 (Floridablanca).



2.3.1.2. Estación Ciudadela

Estación Ciudadela, Calle de los Estudiantes 9-333 (Bucaramanga).

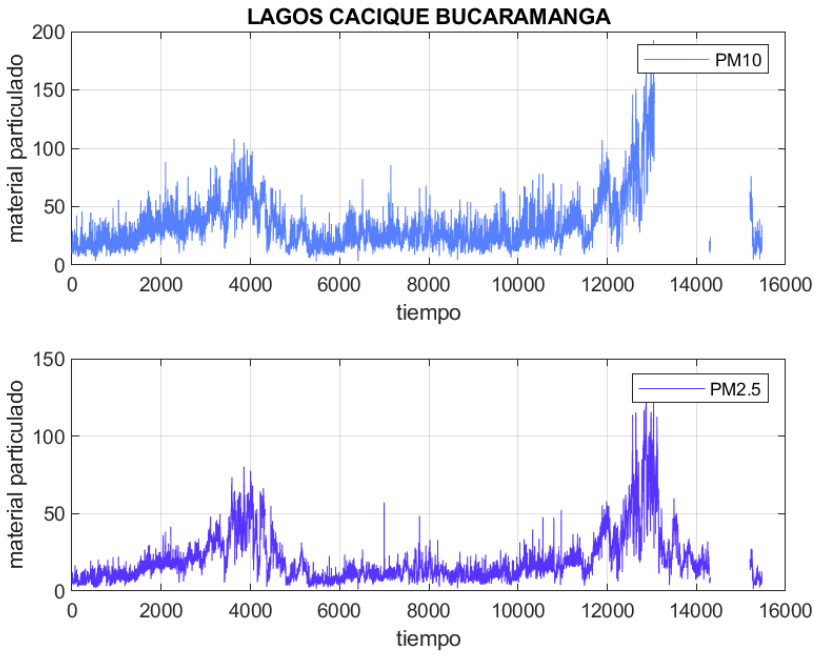
Fig. 8: PM10 y PM2.5 medido por Estación Ciudadela, Calle de los Estudiantes 9-333 (Bucaramanga).



2.3.1.3. Estación Lagos del Cacique Bucaramanga

Estación Lagos del Cacique, Transversal Oriental Lagos del Cacique (Bucaramanga)

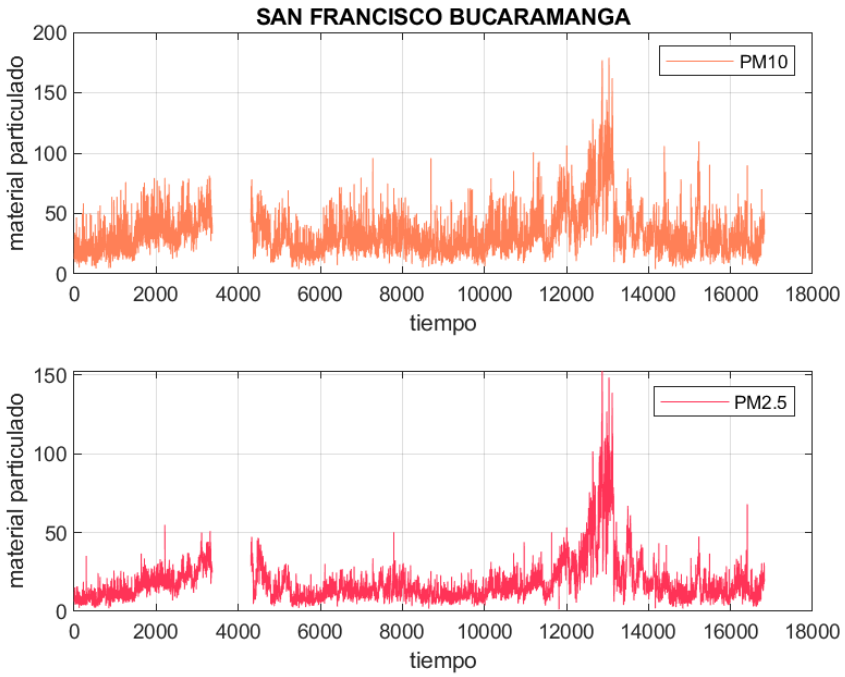
Fig. 9: PM10 y PM2.5 medido por la Estación Lagos del Cacique, Transversal Oriental Lagos del Cacique (Bucaramanga).



2.3.1.4. Estación San Francisco Bucaramanga

Estación San Francisco, Carrera 27 No. 29-69, Bucaramanga.

Fig. 10: PM10y PM2.5 medidos por la Estación San Francisco, Carrera 27 No. 29-69 (Bucaramanga).

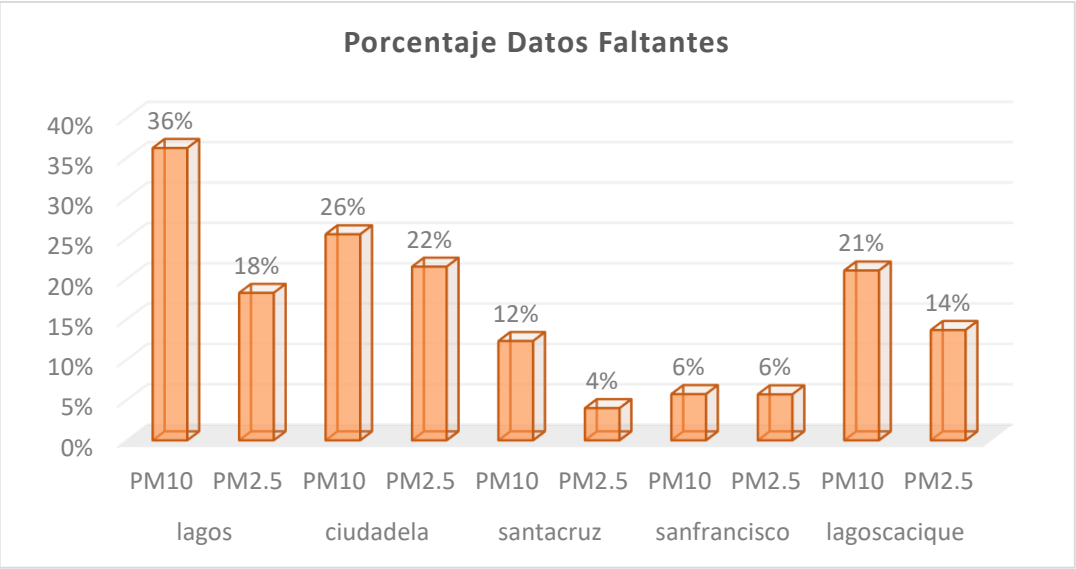


La Tabla 5 resume los datos faltantes con sus respectivos porcentajes según el lugar donde están ubicados los sensores. La comparación gráfica del porcentaje de datos faltantes en cada estación se muestra en la Fig. 11.

Tabla 5: Número de datos faltantes en el historial de calidad del aire.

	ESTACIÓN									
	Lagos		**Ciudadela**		**Santacruz**		**San Francisco**		**Lagos del Cacique**	
	PM10	**PM2.5**	**PM10**	**PM2.5**	**PM10**	**PM2.5**	**PM10**	**PM2.5**	**PM10**	**PM2.5**
Faltantes	6094	3074	4295	3620	2071	671	964	959	3540	2300
Totales	16824	16824	16824	16824	16824	16824	16824	16824	16824	16824
Porcentaje	36%	18%	26%	22%	12%	4%	6%	6%	21%	14%

Fig. 11: Contraste de datos faltantes en el historial de calidad del aire.



Debido a que los datos históricos de Material Particulado están incompletos, se usan técnicas de Deep Learning empleando neuronas LSTM para completarlos, para lo cual se deben estructurar los datos para cada uno de los diferentes modelos.

2.3.2. Data Wrangling para completar los datos de PM10 y PM2.5.

Debido a que los datos de PM10 y PM2.5 están incompletos, se implementan 3 modelos diferentes de regresión usando Redes Neuronales -NAR, NARX, Nonlinear Input/Output-. Debido a que cada uno de estos modelos puede usar distintas funciones de entrenamiento y distinto número de neuronas, se realiza una comparación entre las posibles combinaciones, iterando según el número de neuronas y la función de entrenamiento. Finalmente, en base al error cuadrático medio se selecciona el mejor modelo.

En la Tabla 6 se muestran los parámetros de cada experimento realizado para la obtención del modelo de completado de datos.

Tabla 6: Diseño de experimentos para obtención de modelo de completado de datos.

Estructura de predicción	Funciones de entrenamiento	No. Neuronas
NAR	Trainlm	5
NARX	Trainbr	10
NLIO	Trainscg	15
-	-	20
-	-	25
-	-	30
-	-	35
-	-	40

Para cada estructura de predicción se realizan 24 modelos al variar las funciones de entrenamiento y el número de neuronas utilizadas. Esto resulta en 72 modelos en total, de los cuales se escoge aquél con menor error cuadrático medio del conjunto Validation.

El error cuadrático medio obtenido para la estructura NAR, con cada función de entrenamiento y número de neuronas considerado, se muestra en la Fig. 12 y Tabla 7. Estos mismos resultados para las estructuras NARX y NLIO, se muestran la Fig. 13 y Tabla 8, y en la Fig. 14 y Tabla 9, respectivamente.

Fig. 12: Error cuadrático medio según el número de neuronas y función de entrenamiento para NAR.

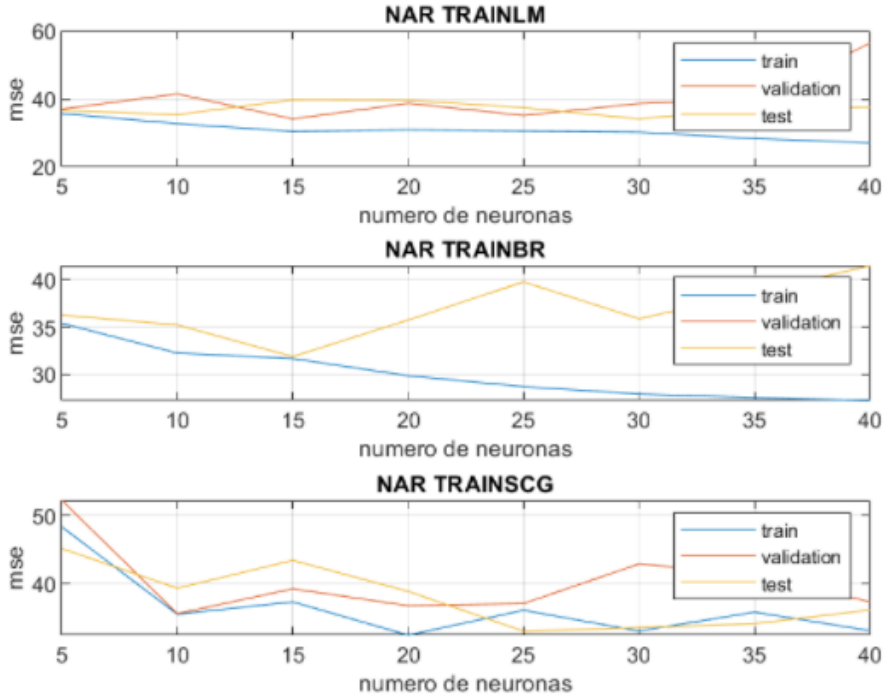


Tabla 7: Error cuadrático medio según el número de neuronas y función de entrenamiento para NAR.

NAR ERROR CUADRATICO MEDIO MSE											
TRAINLM				TRAINBR				TRAINSCG			
NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]
5	35.69	36.99	36.63	5	35.40	36.29	36.29	5	48.31	52.19	45.14
10	32.70	41.47	35.35	10	32.26	35.24	35.24	10	35.50	35.59	39.32
15	30.41	34.08	39.72	15	31.68	31.89	31.89	15	37.31	39.23	43.41
20	30.88	38.69	39.61	20	29.87	35.76	35.76	20	32.43	36.75	38.86
25	30.54	35.12	37.42	25	28.71	39.79	39.79	25	36.09	37.07	33.01
30	30.22	38.63	34.08	30	27.94	35.89	35.89	30	33.02	42.89	33.53
35	28.31	40.00	37.22	35	27.54	38.64	38.64	35	35.81	41.42	34.11
40	27.06	56.37	37.56	40	27.26	41.50	41.50	40	33.09	37.32	36.12

Fig. 13: Error cuadrático medio según el número de neuronas y función de entrenamiento para NARX.

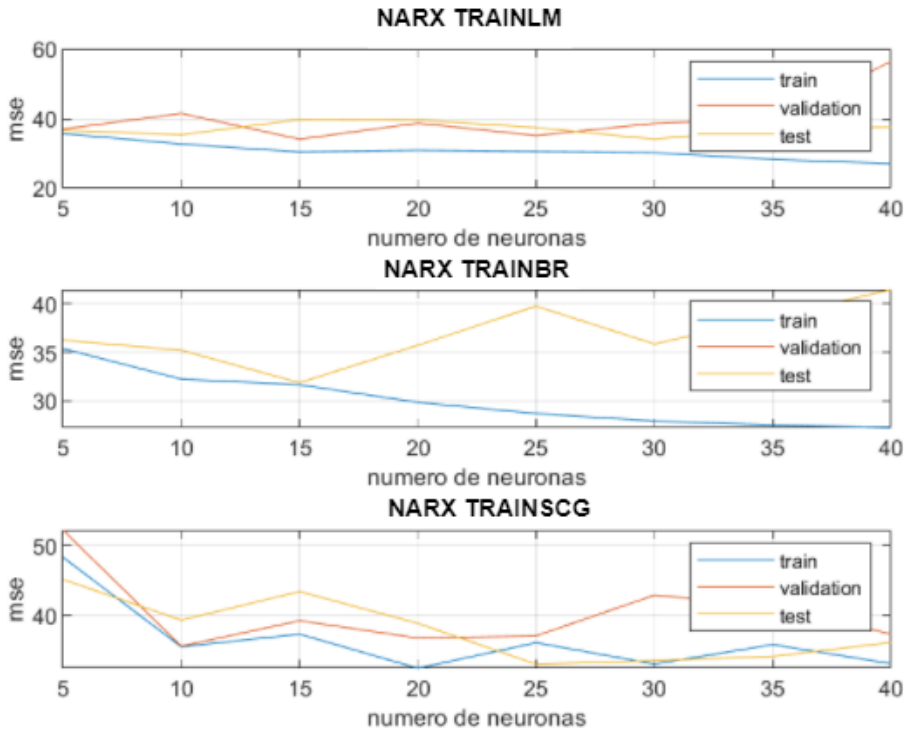


Tabla 8: Error cuadrático medio según el número de neuronas y función de entrenamiento para NARX.

NARX ERROR CUADRATICO MEDIO MSE											
TRAINLM				TRAINBR				TRAINSCG			
NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]
5	17.97	21.45	23.56	5	18.33	23.25	23.25	5	27.40	24.75	25.98
10	18.87	21.18	20.72	10	16.88	22.46	22.46	10	21.45	23.51	21.21
15	16.21	23.88	22.53	15	15.07	25.31	25.31	15	22.89	24.28	22.38
20	17.74	20.75	22.19	20	14.67	26.05	26.05	20	23.95	31.99	27.45
25	17.22	22.74	22.22	25	13.82	31.03	31.03	25	33.44	33.76	34.89
30	20.05	27.19	31.64	30	13.09	41.22	41.22	30	25.62	27.87	23.65
35	14.84	25.07	25.31	35	12.18	80.23	80.23	35	24.73	29.19	29.12
40	15.39	24.42	28.92	40	17.03	21.25	21.25	40	30.41	37.62	27.42

Fig. 14: Error cuadrático medio según el número de neuronas y función de entrenamiento para NLIO.

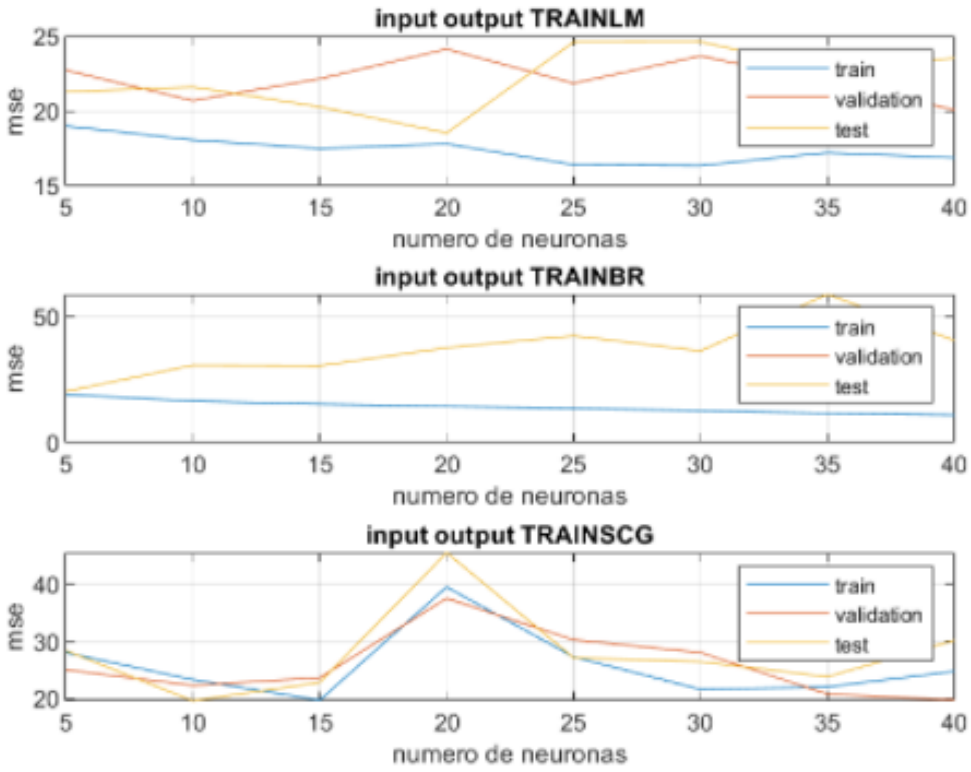


Tabla 9: Error cuadrático medio según el número de neuronas y función de entrenamiento para NLIO.

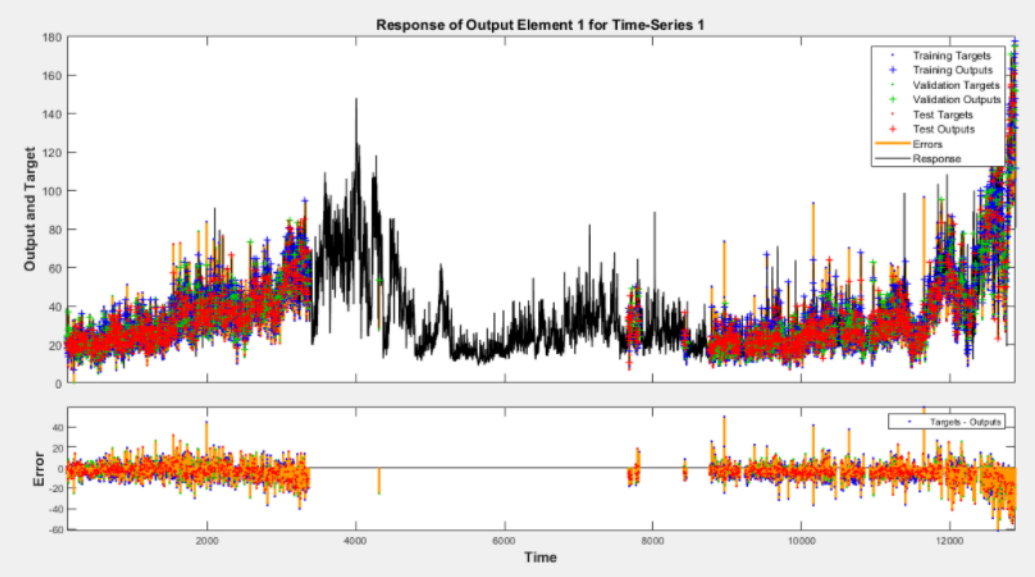
NLIO ERROR CUADRATICO MEDIO MSE											
TRAINLM				TRAINBR				TRAINSCG			
NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]	NO.	TRAINING [%]	VALIDATION [%]	TEST [%]
5	18.99	22.72	21.28	5	19.06	20.29	20.29	5	28.01	25.01	28.37
10	18.06	20.68	21.60	10	16.50	30.67	30.67	10	23.28	22.26	19.58
15	17.48	22.15	20.27	15	15.23	30.48	30.48	15	19.68	23.61	22.78
20	17.79	24.13	18.52	20	14.33	37.63	37.63	20	39.52	37.54	45.54
25	16.40	21.85	24.60	25	13.55	42.39	42.39	25	27.26	30.26	27.11
30	16.34	23.67	24.62	30	12.64	36.39	36.39	30	21.59	28.03	26.45
35	17.21	21.93	22.77	35	11.61	58.82	58.82	35	22.01	20.75	23.78
40	16.86	20.08	23.53	40	10.90	40.57	40.57	40	24.74	19.73	30.16

2.3.2.1. Modelo seleccionado para completar los datos faltantes

Según el error cuadrático medio de los modelos evaluados se selecciona el modelo cuya estructura de predicción es NARX, con *Trainlm* como función de entrenamiento, y con 25 neuronas en la capa oculta. El modelo seleccionado es usado para completar los datos de PM10 y PM2.5. En la Fig. 15 se hace un análisis de *forecasting* que detalla los resultados para la serie de tiempo, donde:

- El punto azul indica los valores reales con los que se entrenó el modelo.
- La cruz azul indica los valores reales con los que se clasificó el modelo.
- El punto verde y el rojo indican los valores reales con los que se contrastó el entrenamiento del modelo.
- La cruz verde y roja indican los valores reales con los que se clasificó el conjunto de Validation y Test del modelo.
- La línea naranja indica el error que hay entre los valores reales y los valores predichos por el modelo.

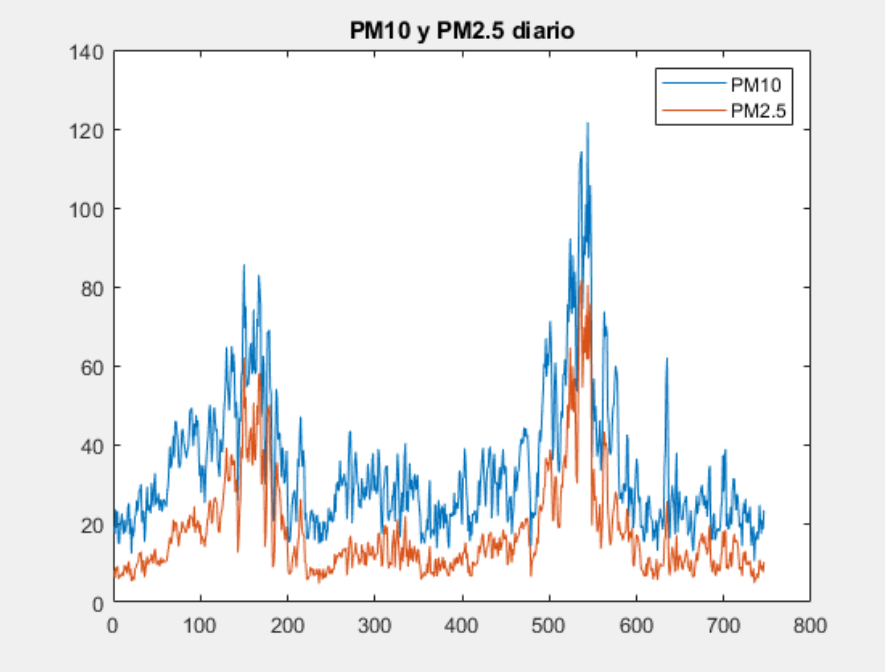
Fig. 15: Análisis de *forecasting* y error cuadrático.



2.3.3. PM10 y PM2.5 diario en Bucaramanga

El histórico de datos medidos de PM10 y PM2.5 está registrado de forma horaria. Como la metodología de cálculo del ICA según la Resolución 2254 de 2017 especifica que para el cálculo del ICA los datos de Material Particulado deben estar en base diaria, se calculó un valor promedio diario a partir de los datos horarios originales, empezando desde octubre 1 de 2018. En la Fig. 16 se muestran los valores horarios promedio reales de PM10 y PM2.5 en la ciudad de Bucaramanga.

Fig. 16: PM10 y PM2.5 promedio diario en Bucaramanga.



2.3.4. Determinación del ICA según Resolución 2254 de 2017

Haciendo uso de la Resolución 2254 de 2017, la cual dicta las pautas para el cálculo del Índice de Calidad del Aire en función de los niveles de Material Particulado, se desarrolla un algoritmo en MATLAB que calcula el ICA y lo clasifica según la escala mostrada en la Tabla 3, pero empleando otros valores numéricos para indicar el estado de calidad del aire, tal como lo muestra la Tabla 10.

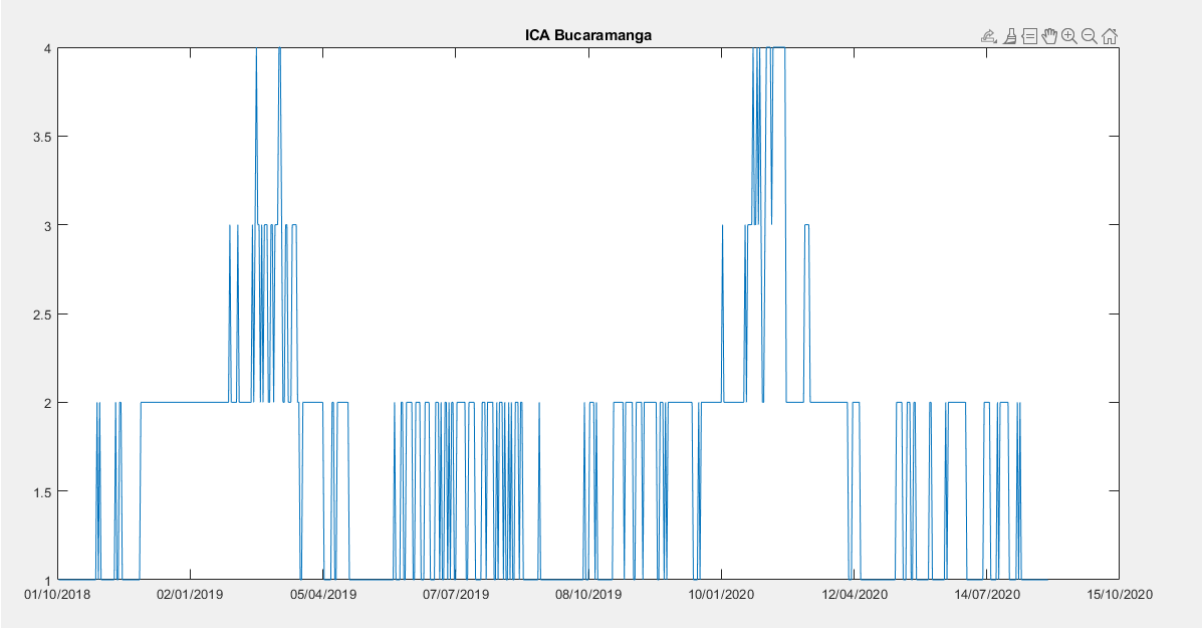
Tabla 10: Indicador Modificado para expresar el estado de calidad del aire.

Rango Original	Color	Estado	Indicador Modificado
0 - 50	Verde	Buena	1
51 - 100	Amarillo	Aceptable	2
101 - 150	Naranja	Dañina a la salud de grupos sensibles	3

151 - 200	Rojo	Dañina para la salud	4
201 - 300	Púrpura	Muy dañina para la salud	5
301 - 500	Marrón	Peligrosa	6

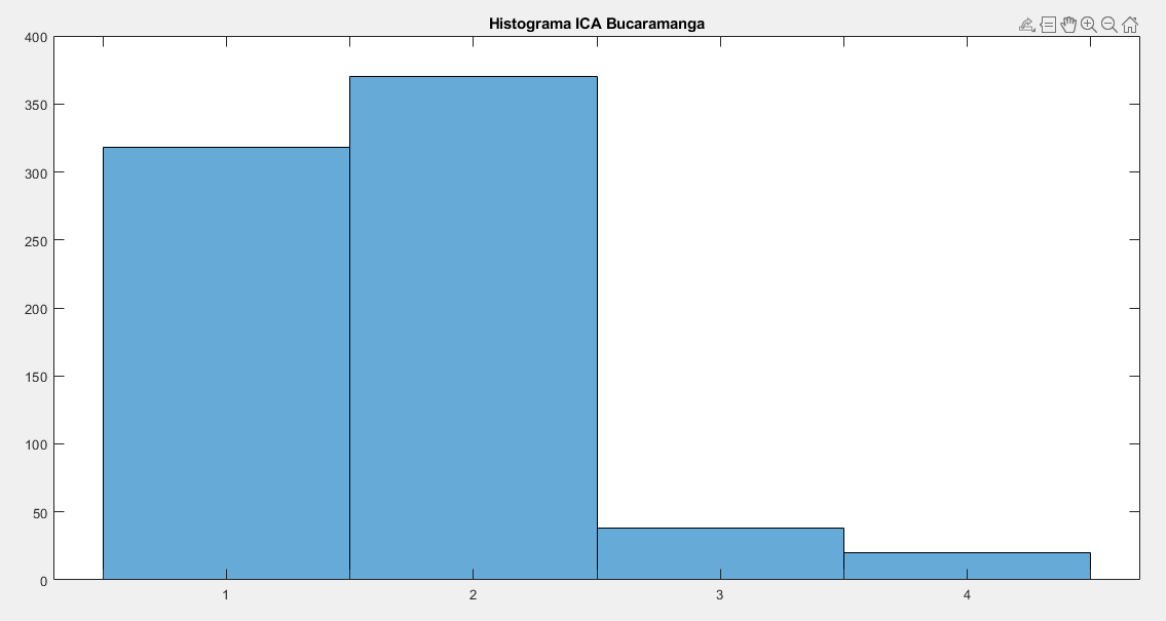
De esta manera, el ICA -en base al Indicador Modificado- para la ciudad de Bucaramanga calculado en función de las mediciones de Material Particulado PM10 y PM2.5, se muestra en la Fig. 17, donde el eje X representa el día de medida y el eje Y el Indicador Modificado.

Fig. 17: ICA (expresado como el Indicador Modificado) en la ciudad de Bucaramanga.



En la Fig. 18 se muestra un histograma para los diferentes valores de ICA según su frecuencia en los días evaluados. Con esto es posible identificar que la calidad del aire en Bucaramanga es generalmente Buena y Aceptable la mayoría de días. Sin embargo, en algunas épocas, la calidad del aire se vuelve dañina para la salud, indicado por los valores de 3 y 4 de ICA.

Fig. 18: Histograma niveles de ICA (expresado como Indicador Modificado) en la ciudad de Bucaramanga.



2.4. Etapa 4: Publicaciones en Twitter

2.4.1. API Twitter

A diario, miles de desarrolladores realizan solicitudes a la plataforma de desarrolladores de Twitter. Para ayudar a administrar el gran volumen de estas solicitudes se establecen límites en la cantidad permisible. Estos límites sirven para aliviar la carga de la infraestructura de Twitter y reducir el tiempo de interrupción del servicio y las páginas de error.

La API de Twitter se evaluó como primera opción para la extracción de Tweets, la cual presentó un inconveniente en particular: no es posible extraer la información de años anteriores, lo que hace que la muestra sea muy pequeña -apenas de 1 semana-.

2.4.2. Octoparse

La solución de Octoparse permite crear un *Web Scraping* automático sin necesidad de usar un lenguaje de programación. Octoparse permite extraer los datos de una

manera sencilla y sin importar la fecha ni el número de publicaciones. Su desventaja más notoria es el tiempo de extracción.

Comparando la API de Twitter con Octoparse, este último cumpliría los requerimientos de poder descargar los datos estructurados, pero teniendo en cuenta el coste computacional y el tiempo en que se ejecuta la extracción de datos, no es una opción totalmente viable.

2.4.3. Twint

Esta librería desarrollada en Phyton e impulsada por la comunidad no tiene límites en cuanto a la extracción de datos ni a su antigüedad. Usando esta librería se implementó un extractor de mensajes en donde las personas en Bucaramanga publicaban sobre su percepción de la calidad del aire. Este método permite descargar los datos en un formato apto para Excel, SQL y NOSQL.

Las características de cada Tweet extraído son:

- Fecha
- Hora exacta
- Tweet
- Hora
- Día
- Mes
- Año
- Longitud de Tweet
- Hashtags
- Menciones
- Número de Hashtags
- Número de Menciones

Estos datos son estructurados en una tabla de Excel que permite su fácil visualización, y son comparados con el ICA Real de ese día en particular. La Tabla 11 muestra un ejemplo de algunos Tweets extraídos con sus respectivas características. En total fueron extraídos 852 Tweets entre octubre 1 de 2018 y octubre 12 de 2020.

Tabla 11: Tweets extraídos y estructurados.

Fecha	Hora Exacta	Tweet	Tweet Limpio	Hora	Día	Mes	Año	Longitud	Hashtags	Menciones	No. hashtags	No. Menciones	ICA REAL
12/10/2020	12:26:08	.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-12T12:25:51.702104-05:00	condiciones hora local	12	12	10	2020	27	['#bucaramanga', '#colombia', '#calidad', '#aire']	[]	4	0	1
3/10/2020	13:28:28	.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-03T13:28:12.447280-05:00	condiciones hora local	13	3	10	2020	27	['#bucaramanga', '#colombia', '#calidad', '#aire']	[]	4	0	1
1/10/2020	10:07:20	¿Cuál te queda más cerca? Aprovecha y movilízate en 🚲. Juntos podemos hacer de una ciudad sostenible, pedalea para respirar la calidad del aire que te mereces 🍃.	cu queda aprovecha movil zate juntos ciudad sostenible pedalea respirar mereces	10	1	10	2020	32	['#bucaramanga', '#yovoyenclobi', '#gobernareshacer']	['@clobibga']	3	1	1
1/10/2020	08:26:18	.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 76 -> ☺ Hora Local: 2020-10-01T08:26:02.738844-05:00	condiciones hora local	8	1	10	2020	27	['#bucaramanga', '#colombia', '#calidad', '#aire']	[]	4	0	1

2.4.4. Limpieza de Tweets

No existe una lista universal de palabras vacías. Estas se pueden preseleccionar o construir desde cero. Un enfoque potencial es comenzar adoptando palabras vacías predefinidas y agregar palabras a la lista más adelante. No obstante, la eliminación de palabras vacías puede borrar información relevante y modificar el contexto en una oración determinada. Por ejemplo, el algoritmo se puede desviar si se elimina una palabra de parada como “bueno”. Bajo estas condiciones, se selecciona una lista mínima de palabras vacías ý se agregan términos adicionales según su objetivo específico.

Las palabras seleccionadas para la limpieza de *Stop Word* son:

['abril', 'actualmente', 'acuerdo', 'adelante', 'además', 'adrede', 'afirmó', 'agosto', 'agregó', 'ahí', 'ahora', 'ahí', 'aire', 'al', 'algo', 'alguna', 'algunas', 'alguno', 'algunos', 'algún', 'allí', 'allí', 'alrededor', 'ambos', 'empleamos', 'amable', 'antaño', 'antaño', 'ante', 'anterior', 'antes', 'apenas', 'aproximadamente', 'aquel', 'aquella', 'aquellas', 'aquello', 'aquellos', 'aquí', 'aquél', 'aquélla', 'aquéllas', 'aquéllos', 'aquí', 'arriba', 'arriba abajo', 'as', 'aseguró', 'así', 'así', 'atrás', 'aun', 'aunque', 'ayer', 'añadió', 'aún', 'bajo', 'bastante', 'bien', 'Bogotá', 'breve', 'Bucaramanga', 'buen', 'buena', 'buenas', 'bueno', 'buenos', 'cada', 'calidad', 'casi', 'cerca', 'cierta', 'ciertas', 'cierto', 'ciertos', 'cinco', 'claro', 'comentó', 'como', 'con', 'conmigo', 'conocer', 'conseguimos', 'conseguir', 'considera', 'consideró', 'consigo', 'consigue', 'consiguen', 'consigues', 'contigo', 'contra', 'cosas', 'creo', 'cual', 'cuales', 'cualquier', 'cuando', 'cuanta', 'cuantas', 'cuanto', 'cuantos', 'cuatro', 'cuenta', 'cuál', 'cuáles', 'cuándo', 'cuánta',

'cuántas', 'cuánto', 'cuántos', 'cómo', 'da', 'dado', 'dan', 'dar', 'de', 'debajo', 'debe', 'deben', 'debido', 'decir', 'dejó', 'del', 'delante', 'demasiado', 'demás', 'dentro', 'deprisa', 'desde', 'despacio', 'después', 'después', 'detrás', 'detrás', 'día', 'días', 'dice', 'dicen', 'dicho', 'dieron', 'diferente', 'diferentes', 'dijeron', 'dijo', 'dio', 'donde', 'dos', 'durante', 'día', 'días', 'dónde', 'ejemplo', 'el', 'ella', 'ellas', 'ello', 'ellos', 'embargo', 'empleáis', 'emplean', 'emplear', 'empleas', 'empleo', 'en', 'encima', 'encuentra', 'enero', 'enfrente', 'enseguida', 'entonces', 'entre', 'era', 'éramos', 'eran', 'eras', 'eres', 'es', 'esa', 'esas', 'ese', 'eso', 'esos', 'esta', 'estaba', 'estaban', 'estado', 'estados', 'estáis', 'estamos', 'están', 'estar', 'estará', 'estas', 'este', 'esto', 'estos', 'estoy', 'estuvo', 'está', 'están', 'ex', 'excepto', 'existe', 'existen', 'explicó', 'expresó', 'febrero', 'fin', 'final', 'fue', 'fuera', 'fueron', 'fui', 'fuimos', 'general', 'gran', 'grandes', 'gta.', 'gueno', 'ha', 'haber', 'había', 'habla', 'hablan', 'habrá', 'había', 'habían', 'hace', 'hacéis', 'hacemos', 'hacen', 'hacer', 'hacerlo', 'haces', 'hacia', 'haciendo', 'hago', 'han', 'hasta', 'hay', 'haya', 'he', 'hecho', 'hemos', 'hicieron', 'hizo', 'horas', 'hoy', 'hubo', 'igual', 'incluso', 'índice', 'indicó', 'informo', 'informó', 'intenta', 'intentáis', 'intentamos', 'intentan', 'intentar', 'intentas', 'intento', 'ir', 'julio', 'junio', 'junto', 'la', 'lado', 'largo', 'las', 'le', 'lejos', 'les', 'llegó', 'lleva', 'llevar', 'lo', 'los', 'luego', 'lugar', 'mal', 'manera', 'manifestó', 'marzo', 'mas', 'mayo', 'mayor', 'me', 'Medellín', 'mediante', 'medio', 'mejor', 'mencionó', 'menos', 'menudo', 'metropolitana', 'mi', 'mía', 'mías', 'mientras', 'mío', 'míos', 'mis', 'misma', 'mismas', 'mismo', 'mismos', 'modo', 'momento', 'mucha', 'muchas', 'mucho', 'muchos', 'muy', 'más', 'mí', 'mía', 'mías', 'mío', 'míos', 'nada', 'nadie', 'ni', 'ninguna', 'ningunas', 'ninguno', 'ningunos', 'ningún', 'no', 'nos', 'nosotras', 'nosotros', 'nuestra', 'nuestras', 'nuestro', 'nuestros', 'nueva', 'nuevas', 'nuevo', 'nuevos', 'nunca', 'ocho', 'os', 'otra', 'otras', 'otro', 'otros', 'país', 'para', 'parece', 'parte', 'partir', 'pasada', 'pasado', 'país', 'peor', 'pero', 'pesar', 'poca', 'pocas', 'poco', 'pocos', 'podéis', 'podemos', 'poder', 'podría', 'podríais', 'podríamos', 'podrían', 'podrías', 'podrá', 'podrán', 'podría', 'podrían', 'poner', 'por', 'porque', 'posible', 'primer', 'primera', 'primero', 'primeros', 'principalmente', 'pronto', 'propia', 'propias', 'propio', 'propios', 'próximo', 'próximo', 'próximos', 'pudo', 'pueda', 'puede', 'pueden', 'puedo', 'pues', 'que', 'que', 'quedó', 'queremos', 'quien', 'quienes', 'quiere', 'quizá', 'quizás', 'quizá', 'quizás', 'quién', 'quiénes', 'qué', 'raras', 'rea', 'realizado', 'realizar', 'realizó', 'repente', 'respecto', 'sabe', 'sabéis', 'sabemos', 'saben', 'saber', 'sabes', 'salvo', 'se', 'sea', 'sean', 'según', 'segunda', 'segundo', 'según', 'seis', 'septiembre', 'ser', 'será', 'será', 'serán', 'sería', 'señaló', 'si', 'sido', 'siempre', 'siendo', 'siete', 'sigue', 'siguiente', 'sin', 'sino', 'sobre', 'sois', 'sola', 'solamente', 'solas', 'solo', 'solos', 'somos', 'son', 'soy', 'soyos', 'su', 'supuesto', 'sus', 'suya', 'suyas', 'suyo', 'sé', 'sí', 'sólo', 'tal', 'también', 'también', 'tampoco', 'tan', 'tanto', 'tarde', 'te', 'temprano', 'tendrá', 'tendrán', 'tenéis', 'tenemos', 'tener', 'tenga', 'tengo', 'tenido', 'tenía', 'tercera', 'ti', 'tiempo', 'tiene', 'tienen', 'toda', 'todas', 'todavía', 'todavía', 'todo', 'todos', 'total', 'trabaja', 'trabajáis', 'trabajamos', 'trabajan', 'trabajar', 'trabajas', 'trabajo', 'tras', 'trata', 'través', 'tres', 'tu', 'tus', 'tuvo', 'tuya', 'tuyas', 'tuyo', 'tuyos', 'tú', 'ultimo', 'un', 'una', 'unas', 'uno', 'unos', 'usa', 'usáis', 'usamos', 'usan', 'usar', 'usas', 'uso', 'usted', 'ustedes', 'va', 'vais', 'valor', 'vamos', 'van', 'varias', 'varios', 'vaya', 'veces', 'ver', 'verdad', 'verdadera', 'verdadero', 'vez', 'vosotras', 'vosotros', 'voy', 'vuestra', 'vuestras', 'vuestro', 'vuestros', 'ya', 'yo',

'él', 'ésa', 'ésas', 'ése', 'ésos', 'ésta', 'éstas', 'éste', 'éstos', 'última', 'últimas', 'último', 'últimos'].

Estas palabras vienen dadas por modelos trabajados por los equipos de NLP de la librería NLTK.

2.4.5. Conteo del número de palabras (variables) en cada Tweet

En esta etapa se desarrolla un algoritmo encargado de contar y separar las palabras únicas. Seguido a esto, recorre cada Tweet y calcula el número de veces que la palabra hace parte del texto. En la Tabla 12 se muestra un ejemplo de cómo se estructuran las palabras. En total, son 2739 palabras diferentes.

Tabla 12: Conteo de palabras presentes en los Tweets.

Tweet	área	mala	ambiental	placa	pico	mejorar	estaciones	ciudad	más	salud	alerta	rango	está	monitoreo	contaminación
.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-12T12:25:51.702104-05:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-03T13:28:12.447280-05:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
¿Cuál te queda más cerca? Aprovecha y movilízate en 🚲. Juntos podemos hacer de una ciudad sostenible, pedalea para respirar la calidad del aire que te mereces 🍃.	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 76 -> ☺ Hora Local: 2020-10-01T08:26:02.738844-05:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Debido a que muchas de las palabras están presentes en muy pocos Tweets, se escoge un umbral de frecuencia mínima en la palabra, para que esta sea tenida en cuenta. Este umbral consiste en que aparezca en al menos 1% de los Tweets, lo que deja 194 palabras. Las palabras por escoger son:

['área', 'mala', 'ambiental', 'placa', 'mejorar', 'estaciones', 'ciudad', 'más', 'salud', 'alerta', 'rango', 'está', 'monitoreo', 'contaminación', 'según', 'condiciones', 'su', 'sobre', 'día', 'hora', 'días', 'buena', 'sin', 'Girón', 'local', 'Bogotá', 'dañino', 'Floridablanca', 'todos', 'índice', 'moderado', 'son', 'marzo', 'sigue', 'mejora', 'resultados', 'medidas', 'Medellín', 'medición', 'esta', 'deterioro', 'ambiente', 'cinco', 'están', 'mejoró', 'estado', 'hoy', 'ha', 'medida', 'durante', 'ciudades', 'mediciones',

'niveles', 'hay', 'AMB', 'autoridades', 'alcalde', 'vehículos', 'pero', 'medio', 'nos', 'también', 'nuestra', 'mes', 'bueno', 'reporte', 'si', 'riesgo', 'lluvias', 'pese', 'entre', 'mejor', 'dos', 'material', 'prevención', 'público', 'transporte', 'mantiene', 'particulado', 'ambientales', 'debe', 'han', 'desde', 'así', 'qué', 'vida', 'cuatro', 'ciudadanos', 'carro', 'eso', 'cuando', 'ahora', 'contaminantes', 'buses', 'semana', 'sus', 'ya', 'último', 'solo', 'partir', 'fue', 'respiramos', 'control', 'esto', 'porque', 'árboles', 'últimos', 'hace', 'tiene', 'movilidad', 'incendios', 'naranja', 'alcaldía', 'sistema', 'pm', 'preocupación', 'aunque', 'vía', 'mejoría', 'índice', 'muy', 'le', 'cuarentena', 'ante', 'decisiones', 'ser', 'febrero', 'decretan', 'rojo', 'mayo', 'hacer', 'proyecto', 'concentraciones', 'sensibles', 'estos', 'preocupa', 'problemas', 'debido', 'tomar', 'siendo', 'zonas', 'decir', 'red', 'excelente', 'tiempo', 'mayor', 'pésima', 'vamos', 'estamos', 'grupos', 'noticias', 'foro', 'va', 'acciones', 'tenemos', 'habrá', 'personas', 'reportan', 'acuerdo', 'mejorado', 'tema', 'país', 'tendencia', 'enfermedades', 'puede', 'bicicleta', 'estación', 'semanas', 'ver', 'lagos', 'lunes', 'mantuvo', "registra', 'gracias', 'reciente', 'arrojaron', 'cada', 'empeoró', 'atmosférica', 'sino', 'problema', 'tienen', 'alarmas', 'coronavirus', 'gente', 'situación', 'moderada', 'vigilancia', 'arrojó, -_-, --, ``,].

Cada una de las palabras anteriores corresponde a una variable empleada en el clasificador. Las otras variables agregadas son metadatos específicos de cada Tweet -hora, día, mes, año, longitud, no. Hashtags, no. Menciones-. Es necesario hacer un cruce de tablas entre el Índice de Calidad del Aire y el Tweet para comparar su valor con la percepción que se tenía en ese momento por las personas. En la Tabla 13 se muestra un ejemplo de lo mencionado.

Tabla 13: ICA Real versus Tweets.

Tweet	ICA Real	Hora	Día	Mes
.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-12T12:25:51.702104-05:00	1	12	12	10
.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 73 -> ☺ Hora Local: 2020-10-03T13:28:12.447280-05:00	1	13	3	10
¿Cuál te queda más cerca? Aprovecha y movilízate en 🚴♂□. Juntos podemos hacer de una ciudad sostenible, pedalea para respirar la calidad del aire que te mereces 🍃.	1	10	1	10

.:-_-:. Condiciones de Calidad del Aire: Bucaramanga: Buena calidad de aire Indice de Calidad del Aire: 76 -> ☺ Hora Local: 2020-10-01T08:26:02.738844-05:00	1	8	1	10

2.5. Etapa 5: Eliminación de palabras que provocan ruido en el clasificador

2.5.1. FSCNCA

El algoritmo de FSCNCA indica los pesos que tienen las variables según las etiquetas. Entre mayor es el peso de la variable, mayor relevancia tiene para el clasificador. Para seleccionar las variables más relevantes se fija un límite inferior de 0.1, de tal manera que aquellas variables cuyos pesos sean mayores a 0.1 serán seleccionadas. En la Fig. 19 se grafican los pesos de cada una de las variables según FSCNCA. En la Tabla 14 se muestra cada palabra con su respectivo índice.

Fig. 19: Pesos de cada palabra según FSCNCA.

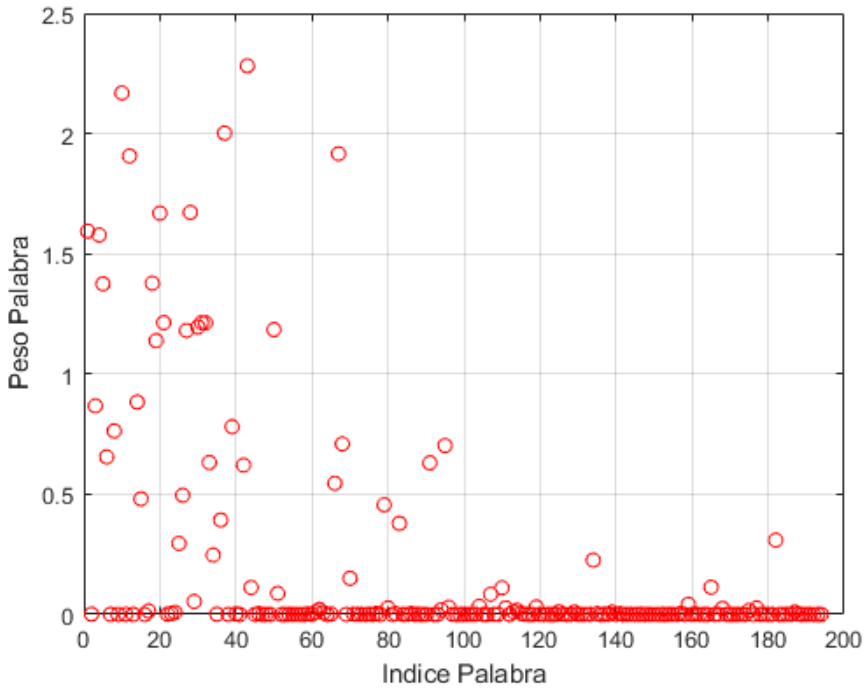


Tabla 14: Palabras seleccionadas por FSCNCA.

Índice de Palabra	Palabra Seleccionada
5	área
7	ambiental
8	placa
9	pico
10	mejorar
12	ciudad
14	salud
16	rango
18	monitoreo
19	contaminación
22	su
23	sobre
24	día
25	hora
29	girón
30	local
31	Bogotá
32	dañino
34	todos
35	-_-
36	índice
37	moderado
38	son
40	sigue
41	mejora
43	medidas
46	esta
47	deterioro
54	ha
70	mes
71	bueno
72	``
83	público
87	ambientales
95	ciudadanos
99	cuando
186	cada

2.5.1.1. FSCNCA + Análisis Discriminante

Con las palabras seleccionadas por FSCNCA indicadas en la Tabla 14, se aplica un clasificador de Análisis Discriminante, a 30 *runs*, variando los conjuntos de Training, Validation y Test. Los resultados de cada *run* se describen en la Tabla 15.

Tabla 15: Accuracy FSCNCA + Clasificador ADL para 30 runs.

Accuracy (FSCNCA+Análisis Discriminante Lineal)			
Run	**Training**	**Validation**	**Test**
1	75.79%	67.47%	68.07%
2	70.87%	71.69%	70.48%
3	71.85%	67.47%	74.10%
4	72.64%	66.27%	64.46%
5	72.05%	68.07%	65.66%
6	70.47%	68.67%	70.48%
7	71.06%	66.27%	67.47%
8	74.21%	71.08%	61.45%
9	74.61%	65.66%	69.28%
10	72.44%	74.10%	63.25%
11	71.85%	71.69%	68.67%
12	72.05%	67.47%	68.67%
13	72.44%	63.86%	62.05%
14	74.80%	70.48%	68.07%
15	75.00%	70.48%	66.87%
16	74.61%	68.07%	65.06%
17	72.64%	69.88%	68.67%
18	71.85%	69.88%	68.07%
19	72.24%	69.88%	68.07%
20	71.26%	65.66%	67.47%
21	71.46%	69.28%	63.86%
22	71.85%	69.28%	71.08%
23	72.83%	65.66%	71.08%
24	75.79%	68.07%	66.87%
25	72.83%	69.28%	72.89%
26	75.39%	70.48%	65.66%
27	73.43%	72.29%	66.87%
28	73.43%	74.70%	68.67%
29	70.28%	67.47%	69.88%
30	71.46%	71.69%	65.66%

El valor de Accuracy promedio para cada conjunto Training, Validation y Test, así como su desviación estándar, se muestran en la Tabla 16. El modelo de clasificación descrito cuenta con un Accuracy promedio de 69.07% $\pm$ 0.02%, el cual equivale al del conjunto Validation.

Tabla 16: Media y desviación estándar para FSCNCA + Clasificador ADL para 30 runs.

FSCNCA+Análisis Discriminante Lineal		
Media		
Training	**Validation**	**Test**
72.78%	69.08%	67.63%
Desviación Estándar		
1.58%	2.57%	2.97%

2.5.1.2. FSCNCA + Redes Neuronales

Con las palabras seleccionadas por FSCNCA indicadas en la Tabla 14, se aplica un clasificador de Redes Neuronales, a 30 *runs*, variando los conjuntos de Training, Validation y Test. Los resultados por cada *run* se muestran en la Tabla 17.

Tabla 17: Accuracy FSCNCA + Redes Neuronales para 30 runs.

Accuracy (FSCNCA + NN)			
Run	**Training**	**Validation**	**Test**
1	71.46%	62.05%	64.46%
2	70.08%	57.83%	53.61%
3	59.06%	54.22%	55.42%
4	75.20%	62.05%	63.25%
5	70.87%	61.45%	62.05%
6	69.69%	59.64%	61.45%
7	64.96%	62.05%	57.83%
8	78.54%	58.43%	61.45%
9	73.03%	59.64%	61.45%
10	77.36%	60.84%	59.04%
11	69.49%	58.43%	56.02%
12	63.58%	52.41%	53.61%
13	64.96%	45.78%	56.02%
14	70.87%	61.45%	59.64%

15	74.41%	62.05%	67.47%
16	66.34%	58.43%	58.43%
17	66.34%	54.22%	54.22%
18	64.96%	59.04%	62.65%
19	71.46%	62.05%	62.05%
20	66.93%	63.25%	59.04%
21	67.72%	60.24%	61.45%
22	70.47%	59.04%	62.05%
23	69.29%	56.63%	57.23%
24	67.52%	62.65%	60.24%
25	67.72%	61.45%	60.84%
26	66.34%	60.24%	53.61%
27	71.65%	62.05%	62.05%
28	75.79%	65.66%	65.66%
29	65.16%	56.63%	63.86%
30	70.08%	65.66%	56.63%

En la Tabla 18 se muestra la media y desviación estándar, para cada conjunto, de los resultados de Accuracy mostrados en la Tabla 17. El modelo de clasificación consistente de FSCNCA con Redes Neuronales presenta, según el conjunto Validation, un Accuracy de 59.52% $\pm$ 4.03%.

Tabla 18: Media y desviación estándar FSCNCA + Redes Neuronales para 30 runs.

FSCNCA + NN		
Media		
Training	**Validation**	**Test**
69.38%	59.52%	59.76%
Desviación Estándar		
4.32%	4.03%	3.72%

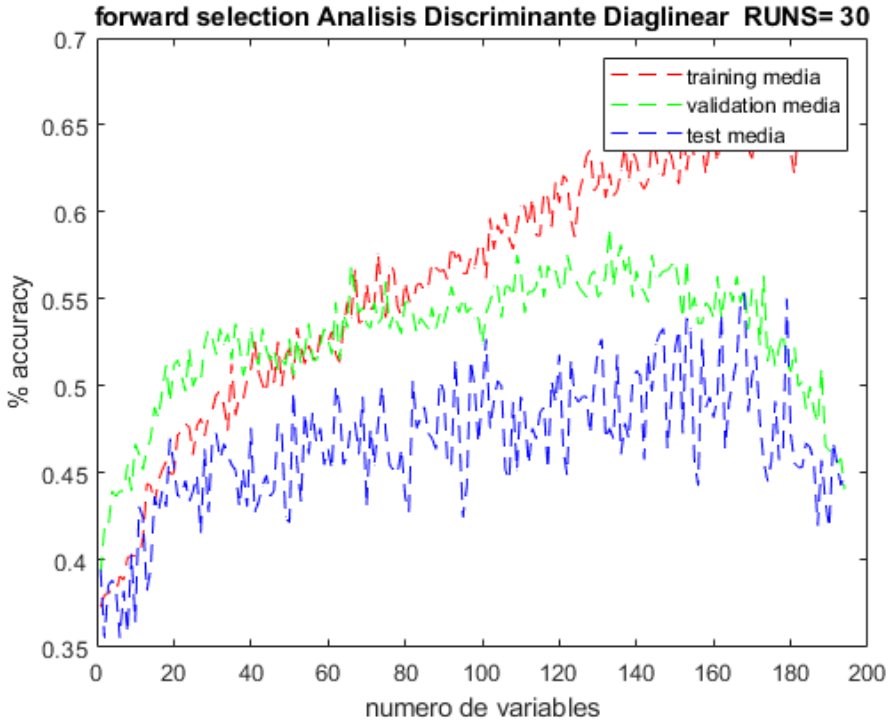
2.5.2. Forward Selection

El algoritmo de Forward Selection organiza las variables según la importancia que le da el clasificador. Debido a que se utilizan dos clasificadores diferentes, el algoritmo debe ser adaptado para cada uno, aumentando las variables según la relevancia que el clasificador le da, con el objetivo de aumentar el porcentaje de clasificación para los conjuntos de Validation y Test.

2.5.2.1. Forward Selection + Análisis Discriminante

Al emplear Forward Selection con un Análisis Discriminante Lineal -*diaglinear* para 30 *runs*-se obtienen los porcentajes de Accuracy para cada número de variables añadidas (ver Fig. 20).

Fig. 20: Accuracy promedio Forward Selection + ADL para 30 runs.



Las variables -palabras- se organizan según la importancia que le da el clasificador para mantener el mayor porcentaje de Validation y Test. La Tabla 19 resume en la primera columna el número que se le otorga en la Fig. 21 a cada palabra. La columna 3 indica la variable según la posición anterior que tenía, y las últimas 3 últimas columnas indican el porcentaje de clasificación en los conjuntos de Training, Validation y Test.

Tabla 19: Palabras seleccionadas por Forward Selection + Análisis Discriminante Lineal.

N°	Palabra seleccionada	Training	Validation	Test
1	año	37.30%	39.16%	39.46%
2	mes	37.99%	41.87%	35.54%

3	últimos	38.09%	41.87%	38.55%
4	cuando	38.29%	43.98%	38.86%
5	medellín	38.19%	43.67%	38.25%
6	local	39.07%	43.98%	35.54%
7	reciente	38.88%	43.98%	38.25%
8	girón	40.16%	45.48%	35.84%
9	aunque	40.26%	43.98%	40.36%
10	acuerdo	40.26%	46.69%	36.45%
11	eso	40.26%	46.08%	43.07%
12	red	41.24%	45.18%	42.47%
13	niveles	44.39%	46.39%	38.25%
14	medidas	44.29%	46.99%	39.46%
15	febrero	43.21%	48.80%	43.67%
16	marzo	44.59%	48.19%	42.77%
17	días	45.18%	49.70%	43.98%
18	ambiental	45.67%	51.20%	43.07%
19	reporte	44.98%	49.10%	46.99%
20	floridablanca	44.98%	51.20%	45.48%
21	ciudadanos	47.24%	51.51%	43.67%
22	sobre	47.34%	50.60%	43.67%
23	tomar	47.83%	49.70%	44.58%
24	cuarentena	47.64%	52.11%	43.37%
25	rojo	46.06%	50.00%	43.98%
26	indice	47.74%	50.60%	44.88%
27	árboles	48.13%	50.30%	41.57%
28	ver	47.15%	52.41%	46.39%
29	hacer	48.33%	53.31%	42.77%
30	medición	48.62%	52.71%	46.99%
31	material	49.51%	51.81%	47.29%
32	ser	49.70%	53.31%	45.78%
33	gente	48.92%	52.41%	46.69%
34	durante	47.44%	53.01%	46.08%
35	tienen	51.28%	51.51%	45.48%
36	más	48.33%	53.61%	45.18%
37	cuatro	49.41%	52.41%	43.07%
38	movilidad	49.21%	50.60%	45.18%
39	así	50.39%	51.51%	43.07%
40	estos	50.98%	53.31%	47.59%
41	tiene	52.56%	51.81%	42.77%
42	empeoró	50.79%	51.20%	43.37%
43	excelente	50.10%	53.31%	44.28%

44	qué	49.61%	52.11%	44.88%
45	cada	51.77%	51.51%	43.67%
46	personas	50.20%	52.11%	43.98%
47	ante	52.56%	50.90%	47.89%
48	partir	50.00%	51.51%	47.29%
49	área	51.77%	51.20%	42.47%
50	debido	52.17%	52.71%	42.17%
51	hora	50.39%	50.60%	49.70%
52	``	53.35%	52.71%	45.48%
53	hora	52.17%	51.20%	43.37%
54	decisiones	51.28%	52.11%	48.49%
55	--	52.36%	53.61%	46.08%
56	buena	51.67%	51.51%	48.49%
57	pico	51.48%	51.51%	46.69%
58	día	52.95%	52.41%	47.59%
59	resultados	52.85%	53.31%	45.48%
60	semana	52.56%	52.71%	46.39%
61	dañino	53.15%	51.81%	45.18%
62	particulado	51.87%	54.82%	50.00%
63	hoy	51.38%	51.51%	48.80%
64	público	53.35%	54.22%	45.48%
65	atmosférica	54.82%	53.31%	46.08%
66	moderado	53.94%	56.93%	47.89%
67	alarmas	56.69%	55.42%	47.89%
68	pm	53.64%	54.22%	46.08%
69	le	53.74%	54.52%	49.40%
70	puede	55.91%	53.61%	43.07%
71	pero	54.33%	53.31%	47.59%
72	estaciones	54.72%	54.82%	45.48%
73	mejor	57.58%	53.61%	47.89%
74	registra	55.02%	53.92%	50.00%
75	mayor	53.94%	56.02%	47.29%
76	zonas	56.89%	53.31%	45.48%
77	riesgo	56.69%	54.22%	45.18%
78	sigue	54.63%	53.61%	44.88%
79	bueno	54.04%	53.01%	46.39%
80	monitoreo	56.00%	54.82%	43.67%
81	está	54.92%	54.82%	42.77%
82	estación	54.82%	54.22%	50.30%
83	transporte	55.81%	53.01%	47.59%
84	entre	55.91%	53.92%	48.19%

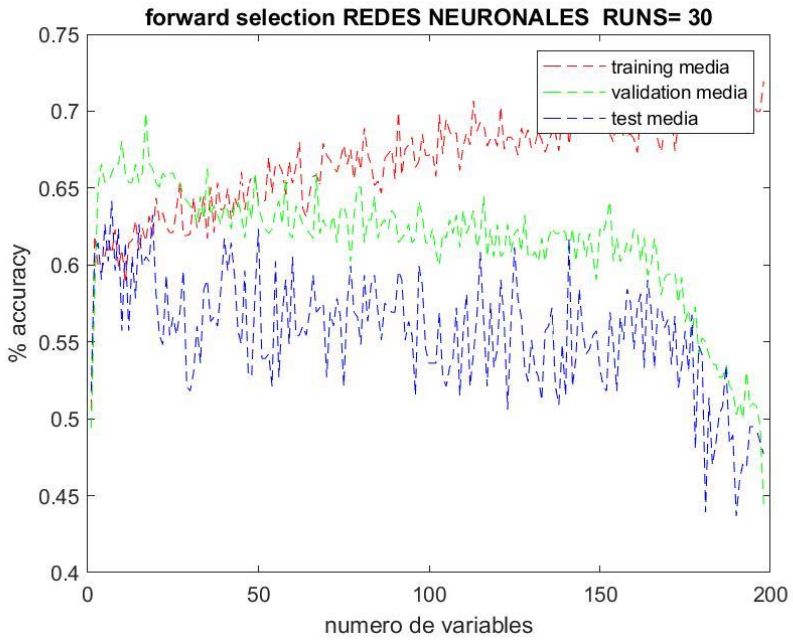
85	mantiene	55.41%	53.61%	48.80%
86	preocupa	55.71%	53.31%	47.29%
87	ya	57.19%	54.52%	46.99%
88	dos	56.79%	53.31%	46.39%
89	sistema	56.59%	54.22%	50.30%
90	alcalde	56.69%	53.61%	49.70%
91	ahora	56.10%	54.52%	45.48%
92	proyecto	58.07%	55.72%	45.48%
93	nuestra	57.38%	54.82%	51.51%
94	carro	57.58%	54.22%	47.59%
95	último	56.50%	55.12%	42.47%
96	prevención	56.50%	54.22%	44.28%
97	dia	57.38%	54.52%	51.51%
98	porque	56.79%	53.92%	50.00%
99	contaminantes	57.97%	53.92%	46.69%
100	mes	58.27%	52.41%	49.40%
101	han	56.20%	54.22%	52.71%
102	lluvias	59.65%	54.82%	47.59%
103	problema	57.87%	55.72%	49.10%
104	tendencia	59.25%	55.72%	50.30%
105	-_-	58.76%	53.92%	50.00%
106	concentraciones	59.94%	55.72%	45.18%
107	mejorado	58.46%	55.72%	44.58%
108	decretan	57.87%	55.12%	49.10%
109	estado	59.25%	57.53%	45.18%
110	ciudad	60.33%	56.63%	49.40%
111	va	60.33%	54.22%	48.49%
112	semanas	59.15%	55.42%	45.48%
113	habrá	60.83%	55.42%	47.59%
114	fue	58.66%	55.72%	46.99%
115	arrojaron	58.66%	56.93%	48.49%
116	debe	61.22%	55.12%	48.80%
117	problemas	60.83%	56.93%	45.18%
118	país	59.15%	56.33%	50.30%
119	sensibles	61.71%	56.63%	49.10%
120	mantuvo	60.53%	55.72%	51.81%
121	sin	62.11%	55.42%	46.39%
122	preocupación	61.81%	57.53%	44.88%
123	contaminación	59.55%	56.63%	51.51%
124	bicicleta	58.56%	56.33%	49.40%
125	mala	60.93%	55.72%	49.10%

126	pese	61.91%	55.42%	49.40%
127	estamos	63.19%	55.12%	49.40%
128	respiramos	63.58%	56.33%	47.29%
129	foro	61.22%	56.63%	50.00%
130	desde	61.52%	55.12%	52.11%
131	alcaldía	62.89%	55.42%	52.71%
132	mejora	60.83%	56.33%	46.99%
133	enfermedades	62.20%	59.04%	47.59%

2.5.2.2. Forward Selection + Redes Neuronales

La Fig. 21 describe la evolución que tiene el Accuracy del clasificador de Redes Neuronales a 30 *runs* según el número de variables que utiliza.

Fig. 21: Accuracy promedio Forward Selection + Redes Neuronales para 30 runs.



Las palabras que aparecen en la Fig. 22 se organizan según la importancia que le da el clasificador para mantener el mayor porcentaje de clasificación en los conjuntos Validation y Test. La Tabla 20 resume en la primera columna el número que se le otorga en la Fig. 22. La columna 3 indica la variable según la posición

anterior que tenía y las 3 últimas columnas indican el porcentaje de clasificación en los conjuntos de Training, Validation y Test.

Tabla 20: Palabras seleccionadas por Forward Selection + Redes Neuronales.

Palabra seleccionada	N° Variables	Training	Validation	Test
mes	3	50.59%	49.40%	51.81%
año	4	61.91%	58.13%	60.84%
sin	28	59.94%	65.06%	61.45%
esta	46	59.55%	66.57%	59.04%
prevención	82	60.53%	65.36%	62.65%
día	24	61.61%	65.36%	60.24%
últimos	115	60.73%	65.96%	64.16%
deterioro	47	62.30%	66.57%	59.64%
particulado	86	59.84%	66.57%	62.35%
último	106	61.02%	68.07%	55.72%
proyecto	140	58.86%	65.96%	61.45%
buses	102	61.42%	65.36%	55.72%
ciudadanos	95	61.61%	65.36%	60.54%
hora	25	62.89%	66.57%	57.83%
qué	92	60.93%	65.36%	62.65%
mediciones	58	62.40%	65.96%	59.94%
contaminantes	101	61.52%	69.88%	60.54%

En la Tabla 21 se muestra el porcentaje de Accuracy para cada subconjunto y cada Run, al emplear Forward Selection más Redes Neuronales. La Tabla 22 detalla la muestra y la desviación estándar de estos resultados.

Tabla 21: Accuracy para Forward Selection + Redes Neuronales para cada Run.

Run	Training	Validation	Test
1	74.33%	71.20%	73.61%
2	73.15%	74.82%	66.99%
3	70.59%	74.22%	69.40%
4	73.54%	65.78%	71.81%
5	75.12%	69.40%	68.19%
6	73.74%	69.40%	65.78%
7	72.36%	67.59%	72.41%
8	72.76%	68.80%	68.19%
9	73.15%	65.78%	71.20%
10	69.80%	68.19%	70.00%

11	74.72%	67.59%	64.58%
12	73.54%	68.19%	72.41%
13	73.15%	66.99%	68.80%
14	75.91%	65.78%	63.37%
15	74.53%	73.61%	69.40%
16	72.76%	68.80%	68.19%
17	72.56%	70.60%	66.39%
18	70.79%	64.58%	76.02%
19	68.23%	63.98%	65.18%
20	72.56%	68.80%	77.83%
21	73.35%	60.36%	65.78%
22	72.95%	63.98%	73.61%
23	75.31%	62.77%	68.19%
24	73.94%	69.40%	72.41%
25	70.79%	71.20%	77.23%
26	74.13%	71.81%	70.00%
27	75.31%	68.19%	68.80%
28	74.92%	65.18%	65.18%
29	71.18%	73.01%	72.41%
30	73.54%	75.42%	68.80%

Tabla 22: Media y desviación estándar para Forward Selection + Redes Neuronales.

Media		
Training	**Validation**	**Test**
73.09%	68.51%	69.74%
Desviación Estándar		
Training	**Validation**	**Test**
1.77%	3.68%	3.70%

2.6. Etapa 6: Modelo de Clasificación

Para seleccionar el modelo de clasificación final, se comparan los resultados de las técnicas de reducción de dimensionalidad con sus respectivos clasificadores. Según el porcentaje de clasificación en el conjunto Validation se escoge el modelo final a ser implementado para la clasificación de la calidad del aire según los Tweets. La Tabla 23 resume los resultados de los modelos implementados. El modelo seleccionado es el clasificador de Análisis Discriminante Lineal con las variables seleccionadas por FSCNCA: ['área', 'ambiental', 'placa', pico', 'mejorar', 'ciudad', 'salud', 'rango', 'monitoreo', 'contaminación', 'su', 'sobre', 'día', 'hora', 'girón', 'local', 'Bogotá', 'dañino', 'todos', -_-, 'índice', 'moderado', 'son', 'sigue', 'mejora', 'medidas', 'esta', 'deterioro', 'ha', 'mes', 'bueno', ``, 'público', 'ambientales', 'ciudadanos', 'cuando', 'cada'].

Tabla 23: Resumen de resultados de los modelos implementados.

Algoritmo de reducción de dimensionalidad	Clasificador	Conjunto:	Training	Validation	Test
FSCNCA	ADL	Media	72.78%	69.08%	67.63%
		STD	1.58%	2.57%	2.97%
	NN	Media	69.38%	59.52%	59.76%
		STD	4.32%	4.03%	3.72%
Forward Selection	ADL	Media	62.89%	55.42%	52.71%
		STD	1.91%	2.82%	2.31%
	NN	Media	73.09%	68.51%	69.74%
		STD	1.77%	3.68%	3.70%

2.6.1. Algoritmos obtenidos e implementados

Los algoritmos obtenidos e implementados durante el desarrollo del proyecto son los siguientes:

- Extractor de Tweets sin límites.
- 3 modelos de Redes Neuronales para completar datos en series de tiempo (NAR, NARX, NLIO).
- Reductor de palabras relevantes aplicando Forward Selection y FSCNCA.
- Modelo clasificador del Índice de Calidad del Aire (ICA) según la frecuencia de palabras en Tweets aplicando Análisis Discriminante Lineal y Redes Neuronales.

2.6.2. Resultados obtenidos e implementados

Para visualizar los resultados obtenidos se hace una prueba en un conjunto de datos de Test en el que se conoce el ICA real correspondiente a cada día. La Fig. 22 describe para cada Tweed el ICA real correspondiente. La Fig. 23 es el modelo implementado en ese conjunto de

Fig. 22: Índice de calidad del aire correspondiente a cada Tweet

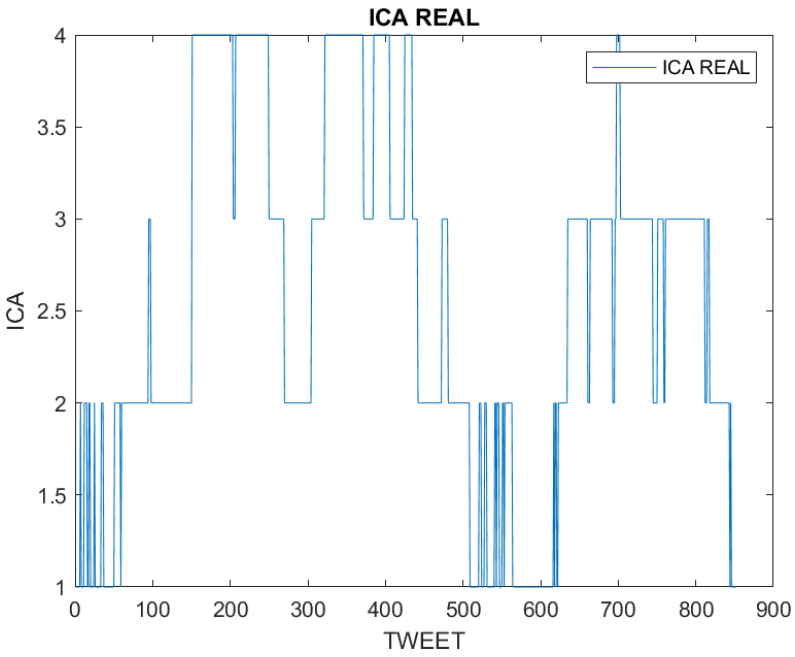


Fig. 23: Índice de calidad del aire (PREDICT) correspondiente a cada Tweet

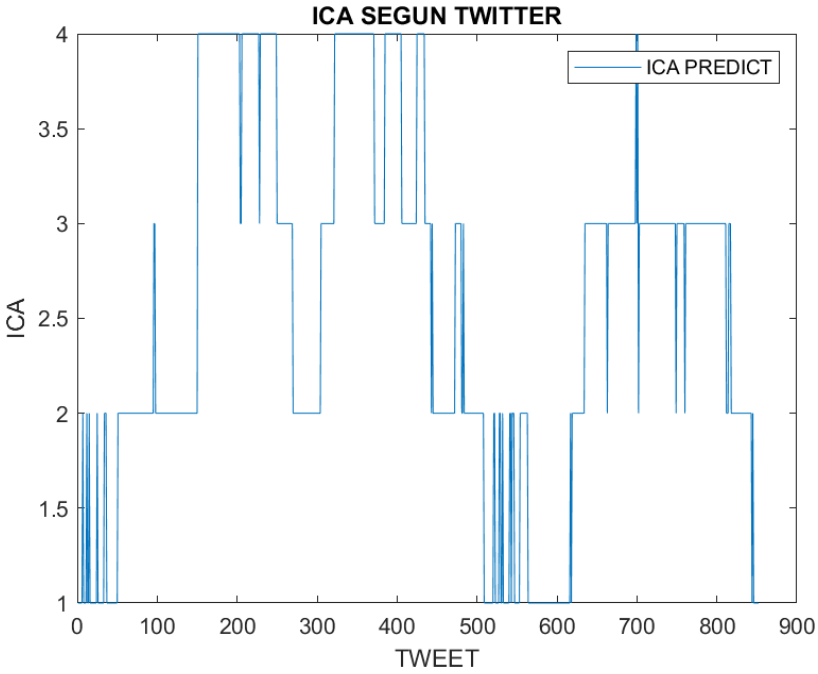
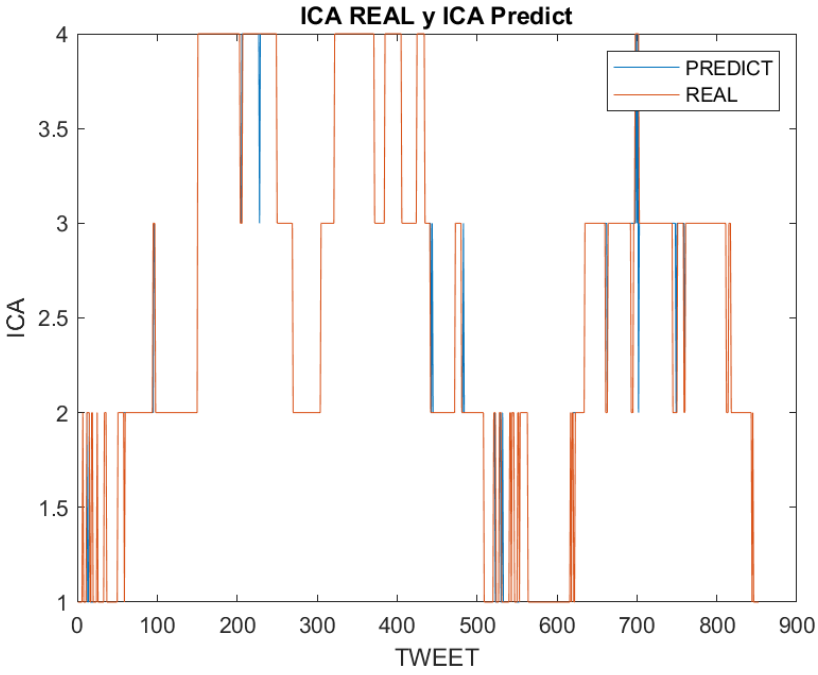


Fig. 24: Comparación de resultados Reales/Predichos



3. CONCLUSIONES

El objetivo principal de este proyecto es el de abordar el problema de obtención de un modelo que permita clasificar la calidad del aire en base a las publicaciones realizadas en Twitter y aportar una solución para el proceso de identificación de palabras relevantes en los modelos de clasificación de Análisis Discriminante Lineal y Redes Neuronales. De esta manera, el principal aporte de este proyecto es el de diseñar e implementar una metodología para el trabajo en el área del Procesamiento de Lenguaje Natural para la empresa K2 Ingeniería. Los algoritmos diseñados en este proyecto, además de ser novedosos, aumentan las posibilidades de trabajar con cualquier tipo de tema en Twitter donde la variable objetivo este siendo muestreada.

Debido a que los datos de Material Particulado en Bucaramanga estaban incompletos, el modelo de Redes Neuronales NARX pudo completarlos dando resultados muy buenos en comparación con los de los modelos NAR Y NLIO, con un error cuadrático medio de 20. Gracias a las fechas de muestreo de datos se pudo relacionar el ICA Real en Bucaramanga con cada Tweet.

El modelo de reducción de dimensionalidad escogido fue FSCNCA que seleccionó las siguientes palabras relevantes para el clasificador: ['área', 'ambiental', 'placa', pico', 'mejorar', 'ciudad', 'salud', 'rango', 'monitoreo', 'contaminación', 'su', 'sobre', 'día', 'hora', 'girón', 'local', 'Bogotá', 'dañino', 'todos', -_-, 'índice', 'moderado', 'son', 'sigue', 'mejora', 'medidas', 'esta', 'deterioro', 'ha', 'mes', 'bueno', ``, 'público', 'ambientales', 'ciudadanos', 'cuando', 'cada'].

Se generó un modelo de contaminación del aire usando las publicaciones realizadas por las personas en Bucaramanga a partir de datos extraídos en Twitter con un porcentaje de clasificación de 69% $\pm$ 0.02%.

REFERENCIAS

FANG, Guor-Cheng *et al.*, “Characterization of particulate, metallic elements of TSP, PM2.5 and PM2.5-10 aerosols at a farm sampling site in Taiwan, Taichung,” *Sci. Total Environ.*, Jun. 2003, vol. 308, no. 1–3, pp. 157–166, doi: 10.1016/S0048-9697(02)00648-4.

BILLET. Sylvain *et al.*, “Ambient particulate matter (PM2.5): physicochemical characterization and metabolic activation of the organic fraction in human lung epithelial cells (A549),” *Environ. Res.*, Oct. 2007, vol. 105, no. 2, pp. 212–223, doi: 10.1016/J.ENVRES.2007.03.001.

ARCINIÉGAS, Cesar S., “Diagnóstico y control de material particulado: partículas suspendidas totales y fracción respirable PM10,” 2012, doi: 10.17151/luaz.2012.34.12.

“Contaminantes del aire: Materias particuladas.” http://www.murciasalud.es/pagina.php?id=244308&# (accessed Oct. 04, 2021).

“Monitoreo aire - IDEAM.” http://www.siac.gov.co/monitoreoaire (accessed Oct. 04, 2021).

“Informe Carga de Enfermedad Ambiental en Colombia.” https://www.ins.gov.co/Noticias/Paginas/Informe-Carga-de-Enfermedad-Ambiental-en-Colombia.aspx (accessed Oct. 04, 2021).

M. y E. A. IDEAM, Instituto de hidrología, “Índice de calidad del aire (ICA),” no. 571, p. 2013, 2012, [Online]. Available: http://www.ideam.gov.co/documents/11769/641368/2.01+HM+Indice+calidad+aire.pdf/5130ffb3-a1bf-4d23-a663-b4c51327cc05.

“Series Temporales: Introducción,” Accessed: Oct. 04, 2021. [Online]. Available: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema7.pdf.

“How Dynamic Neural Networks Work - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ug/how-dynamic-neural-networks-work.html (accessed Oct. 05, 2021).

“Choose a Multilayer Neural Network Training Function - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ug/choose-a-multilayer-neural-network-training-function.html (accessed Oct. 05, 2021).

BENRHMACH, G., NAMIR, K., NAMIR, A. and BOUYAGHROUMNI, J., “Nonlinear Autoregressive Neural Network and Extended Kalman Filters for Prediction of Financial Time Series,” *J. Appl. Math.*, 2020, vol. 2020, doi:

10.1155/2020/5057801.

“Design Time Series NARX Feedback Neural Networks - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html (accessed Oct. 05, 2021).

“Levenberg-Marquardt backpropagation - MATLAB trainlm - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ref/trainlm.html (accessed Oct. 05, 2021).

“Bayesian regularization backpropagation - MATLAB trainbr - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ref/trainbr.html?s_tid=doc_ta (accessed Oct. 05, 2021).

“Scaled conjugate gradient backpropagation - MATLAB trainscg - MathWorks América Latina.” https://la.mathworks.com/help/deeplearning/ref/trainscg.html?searchHighlight=trainscg&s_tid=srchtitle (accessed Oct. 05, 2021).

“What is Natural Language Processing? | IBM.” https://www.ibm.com/cloud/learn/natural-language-processing#toc-what-is-na-jLju4DjE (accessed Oct. 05, 2021).

“Procesamiento del lenguaje natural con NLTK para Ingeniería social automatizada – Seguridad en Sistemas y Técnicas de Hacking. TheHackerWay (THW).” https://thehackerway.com/2015/02/17/procesamiento-del-lenguaje-natural-con-nltk-para-ingenieria-social-automatizada/ (accessed Oct. 05, 2021).

“Análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA).” https://www.cienciadedatos.net/documentos/28_linear_discriminant_analysis_lda_y_quadratic_discriminant_analysis_qda (accessed Oct. 05, 2021).

“Introduction to Feature Selection - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/stats/feature-selection.html (accessed Oct. 06, 2021).

YANG, W., WANG, K., and ZUO, W. “Neighborhood component feature selection for high-dimensional data,” *J. Comput.*, 2012, vol. 7, no. 1, pp. 162–168, doi: 10.4304/JCP.7.1.161-168.

“Neighborhood Component Analysis (NCA) Feature Selection - MATLAB & Simulink - MathWorks América Latina.” https://la.mathworks.com/help/stats/neighborhood-component-analysis.html (accessed Oct. 06, 2021).

MARCANO, A. C., QUINTANILLA, J. D., CORTINA M. G. J., and ANDINA, D. “Feature selection using Sequential Forward Selection and classification

applying Artificial Metaplasticity Neural Network," *IECON Proc.* 2010 *(Industrial Electron. Conf.*, pp. 2845–2850, doi: 10.1109/IECON.2010.5675075.

CHANDRA, B. "Gene Selection Methods for Microarray Data," *Appl. Comput. Med. Heal.*, Jan. 2016 pp. 45–78, doi: 10.1016/B978-0-12-803468-2.00003-5.