Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá

Juan Pablo Vásquez Arenas

Monografía presentada para optar al título de Especialista en Gestión Ambiental

Asesor
Julio Eduardo Cañón Barriga, MSc., Ph.D.

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Gestión Ambiental
Medellín, Antioquia, Colombia
2022

Cita	(Vásquez Arenas, 2022)
Referencia **Estilo APA 7 (2020)**	Vásquez Arenas, J. P (2022). *Beneficios de los modelos basados en algoritmos de machine learning para la predicción de concentraciones de PM2.5 en el Valle de Aburrá* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Gestión Ambiental.

Grupo de Investigación Ingeniería y Gestión Ambiental (GIGA).

Centro de Investigación Ambientales y de Ingeniería (CIA).

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: John Dairo Zapata Ochoa

Agradecimientos

En primer lugar, quiero agradecerle a mi asesor, Dr Julio Eduardo Cañón Barriga por su apoyo, dedicación y paciencia brindada para el desarrollo de esta monografía, gracias por la comprensión de mis dudas y sus valiosos aportes.

También quiero agradecer a mis profesores y compañeros, por todo el conocimiento que pusieron a mi disposición de forma desinteresada durante todos estos meses.

Por último y no menos importante, quiero agradecer a mi familia y compañeros de trabajo por acompañarme durante esta etapa tan importante en mi vida, por su amor y comprensión, mil gracias.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

ANN	Artifical Neural Networks
CNN	Convolutional Neural Network
OMS	Organización Mundial de la Salud
PIGECA	Plan Integral de Gestión de la Calidad del Aire
PM10	Material particulado inferior a diez 10 micras
PM2,5	Material particulado inferior a diez 2.5 micras
POECA	Protocolo en épocas de contingencia ambiental
RNN	Recurrent Neural Networks
SIATA	Sistema de Alerta Temprana del Valle de Aburrá

Resumen

Esta monografía hace una revisión crítica de la utilización de algoritmos de machine learning para la predicción de concentraciones de contaminantes atmosféricos, identificando cuáles serían sus beneficios y limitaciones para la predicción de concentraciones de PM2.5 en el Valle de Aburrá. Se hace una revisión general de casos de estudio publicados en revistas indexadas desde el 2019 y el análisis de las condiciones particulares del Valle de Aburrá que inciden en las concentraciones de contaminantes atmosféricos. Se seleccionaron y analizaron algunos casos de estudio en los que se identificaron las principales características de desarrollo de los modelos como datos de entrada, tipo de algoritmo utilizado, métricas de desempeño y algoritmo de mejor desempeño. El análisis dio como resultado que los modelos desarrollados hasta la fecha permiten predecir concentraciones de contaminantes a un nivel que pueda ser útil para un sistema de alerta temprana, permitiendo también representar las condiciones de inmisión del Valle de Aburrá y sus particularidades topográficas, meteorológicas y de emisión que inciden en las concentraciones de PM2.5, e indican la presencia de fenómenos como picos abruptos de concentración. Sin embargo aún no se pudo identificar un algoritmo particular que tuviese un mejor desempeño para este tipo de aplicaciones, debido a que se suelen usar algoritmos específicos en cada caso de estudio, dificultando su reusabilidad y repetibilidad. También se estableció que el Valle de Aburrá cuenta con la información necesaria para construir los conjuntos de datos de entrenamiento para la implementación de modelos de machine learning para la predicción de las concentraciones de material particulado.

Palabras clave: artículo de revisión, calidad del aire, machine learning, Valle de Aburrá.

Abstract

This work carried out a critical review of the use of machine learning algorithms for the prediction of concentrations of air pollutants, identifying what their benefits are for the prediction of PM2.5 concentrations in the Aburrá Valley. We reviewed case studies published in journals indexed since 2019 and studies on the conditions of the Aburrá Valley that affect the concentrations of atmospheric pollutants. We selected case studies that use machine learning, identifying the main characteristics of the models, such as input data, type of algorithm used, performance metrics and the best performing algorithm. Results show that the machine learning models developed to date allow predicting pollutant concentrations at a level that can be useful for an early warning system of PM2.5 concentrations, even in abrupt concentration peaks, allowing the representation of the immission conditions of the Aburrá Valley with its topographical, meteorological and environmental characteristics. However, we did not identify a particular algorithm that would have a better performance for this type of applications, because the algorithms are very specific and case-oriented, making their reusability and repeatability difficult. Currently, the Valle de Aburrá has the necessary data and information to build the training sets for the implementation of these kinds of algorithms to predict concentrations of particulate matter.

*Keywords***:** Review article, air quality, machine learning, Aburrá Valley.

1. Introducción

De acuerdo con los resultados de los monitoreos publicados por el Sistema de Alerta Temprana del Valle de Aburrá (SIATA), en las últimas dos décadas la mayoría de estaciones de monitoreo presentaron concentraciones promedio anuales de PM2.5 superiores tanto al límite propuesto por la organización mundial de la salud (OMS) como por la normatividad colombiana vigente, convirtiéndolo en el contaminante crítico de evaluación, tal y como se resolvió en el Protocolo en épocas de contingencia ambiental (POECA) (AMVA, 2021), acuerdo metropolitano 16 del 2017 (AMVA, 2017), y Plan Integral de Gestión de la Calidad del Aire (PIGECA) (AMVA, 2017).

Dichas concentraciones de PM2.5 vienen influenciadas en buena parte por fenómenos meteorológicos típicos de un valle como inversiones térmicas o bajas velocidades del viento, que en el caso específico del Valle de Aburrá se suelen presentar al inicio de las dos temporadas de lluvias de cada año, donde formaciones de nubes sobre el valle afectan la entrada de radiación solar y alteran la forma en la que se calienta la masa de aire más cercana a la superficie, enfriándose, e impidiendo los fenómenos de dispersión convectiva (Bernal Manrique & Rendón Pérez, 2019).

Estos problemas han obligado a las autoridades competentes a desarrollar una serie de herramientas, sistemas y protocolos, que permitan identificar, monitorear y gestionar la calidad del aire de la ciudad (AMVA, 2021). Entre estos desarrollos están: inventarios de emisiones, sistemas de monitoreo de calidad del aire y modelos de dispersión.

En los últimos años, el uso de modelos de machine learning para predicción de contaminantes atmosféricos ha estado en aumento (Iskandaryan, Ramos, & Trilles, 2020). Según la revisión bibliográfica realizada, se han encontrado casos de estudio de implementación de este tipo de modelos tanto para el Valle de Aburrá (Mogollón-Sotelo et al., 2020; Murillo-Escobar et al. 2019; Becerra et al., 2021), como para otras ciudades en Colombia (Casallas et al., 2021).

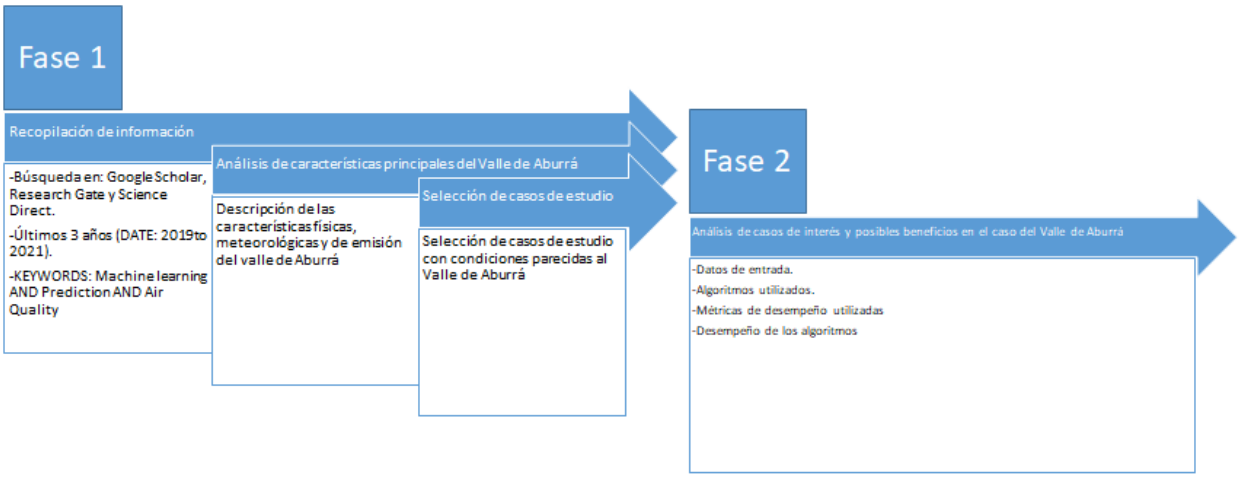
Esta situación lleva a plantearse la siguiente pregunta: ¿Qué ventajas tienen los modelos basados en algoritmos de machine learning para predecir o complementar la predicción de concentraciones de PM2.5 en el Valle de Aburrá?, para resolverla, primero se analizaron las metodologías utilizadas en casos de estudio de publicaciones recientes respecto al tema, on el fin de identificar casos de interés donde las condiciones topográficas, meteorológicas, de

inmisión/emisión, o de disponibilidad de datos se parezcan a las condiciones del Valle de Aburrá.

A dichos casos de interés se les realizó un análisis a profundidad donde se examinaron los materiales y métodos utilizados, en función de los resultados presentados, para así determinar los beneficios y eventuales desafíos del uso de este tipo de aplicaciones para la predicción de concentraciones de PM2.5 en el Valle de Aburrá.

2. Metodología

La monografía se desarrolló con base en una revisión bibliográfica en dos fases (ver Figura 1):



Nota. Fuente: Elaboración propia, 2022

Figura 1. Diagrama de flujo metodología.

2.1 Primera fase – Revisión general de artículos

Dentro de esta fase se realizaron tres actividades

1. Recopilación de información: Fase inicial en la cual se recopilaron artículos de investigación respecto a implementaciones de modelos de machine learning para predicción de PM2.5 en los últimos 3 años (DATE: 2019 to 2021). Estos artículos se buscaron en las bases de datos Google Scholar y Science Direct, las cuales presentaron mayor número de artículos disponibles respecto al tema de Machine learning y

predicción de concentración de contaminantes (KEYWORDS: PM2.5 AND concentration AND Prediction AND Machine Learning AND urban).

2. Análisis de características principales del Valle de Aburrá: Se llevó a cabo una recopilación de artículos o bibliografía disponible, que describiera las características físicas y de emisión de mayor incidencia para diagnosticar o predecir las condiciones de calidad del aire por PM2.5 en el Valle de Aburrá. De igual forma, se analiza la disponibilidad de datos de monitoreo de calidad del aire y otras variables relacionadas para el entrenamiento de algoritmos.

3. Selección de casos de estudio: Ya identificadas las características de los casos de estudio recopilados y definidas las características físicas de mayor incidencia en la calidad del aire del Valle de Aburrá, se seleccionaron los casos de estudios cuyas condiciones se parecieron más a las identificadas para el Valle de Aburrá, para analizar los beneficios en la segunda fase.

2.1 Segunda fase: Análisis de casos con similaridades

Se analizaron los artículos seleccionados con el fin de determinar las condiciones y metodologías de modelación de los casos que tuvieron mejor desempeño en condiciones similares a las del Valle de Aburrá. Para dicho análisis se identificaron las principales características que condicionaron el entrenamiento y evaluación del modelo de predicción

- Datos de entrada: Tipos de monitoreo, resolución temporal, disponibilidad y fuente de datos, periodo de entrenamiento y testeo.
- Datos de salida: Tipos de datos predichos como concentraciones, índice de calidad del aire (ICA), alerta de episodios de mala calidad del aire.
- Algoritmos utilizados: Conjunto de ecuaciones asociadas al modelo que, sirviéndose de unos datos de entrenamiento previamente suministrados, predicen la concentración de un contaminante, valiéndose de un conjunto de datos de entrada correspondientes a periodos anteriores al dato a predecir.
- Métricas de desempeño utilizadas: Cuáles fueron los indicadores de desempeño usados para evaluar la precisión de los datos predichos por los algoritmos estudiados.
- Desempeño de los algoritmos: Valiéndose de los resultados de las métricas de desempeño de cada algoritmo se identificaron los de mayor y menor precisión.

3. Revisión General

3.1 Información recopilada

Actualmente en contextos urbanos, los modelos determinísticos presentan limitaciones en representar el comportamiento no lineal de la concentración de contaminantes respecto a sus fuentes de emisión, dichas limitaciones afectan los ejercicios de predicción de concentración de contaminantes, sin embargo, se están realizando aproximaciones utilizando modelos de machine learning, los cuales han tenido mejores resultados (Masih, 2019).

Los modelos de dispersión determinísticos han sido desarrollados e implementados por agencias gubernamentales o empresas asociadas a ellas, estos modelos son sistemas numéricos los cuales hacen uso de un conjunto de asunciones y aproximaciones matemáticas para representar el fenómeno de dispersión de contaminantes en ciertas condiciones según ciertas condiciones.

Los modelos de dispersión de contaminantes más usados (Rybarczyk & Zalakeviciute, 2018) se pueden agrupar en cuatro siguientes grupos: Gaussianos, Lagrangianos, Eulerianos y Fotoquímicos (CTMs).

Todos estos modelos tienen en común dos cosas: dependen de la calidad de una gran cantidad de datos detallados (como meteorología e inventarios de emisiones) y el alto costo computacional que implica correrlos, lo que genera una serie de limitaciones tanto en representatividad como en dinamismo del proceso mismo. Sin embargo, se ha demostrado en varios casos de estudio alrededor del mundo que estos modelos pueden caracterizar adecuadamente un escenario de dispersión siempre y cuando se sepan las características y restricciones de estos modelos.

Las condiciones mencionadas anteriormente no suelen cumplirse en el contexto de entornos urbanos, donde no sólo las variables mencionadas no poseen información detallada, sino también donde las condiciones de emisión y meteorológicas pueden presentar variaciones respecto a su representación original (Minghao, Zigler, & Selin, 2022).

3.2 Uso del machine learning en calidad del aire

El machine learning es una de las ramas de la inteligencia artificial en la cual se estudian métodos para permitir a una máquina desarrollar acciones basadas en datos y algoritmos en vez

de instrucciones explícitamente indicadas (Ray, 2019). Para el caso de calidad del aire en los últimos 5 años el uso del machine learning para la predicción de concentraciones de contaminantes atmosféricos ha aumentado, presentándose diversos casos de estudio (Iskandaryan, Ramos, & Trilles, 2020) en los cuales se suele seguir el mismo patrón de aprendizaje y predicción.

Si bien los casos de estudio suelen ser diversos, por lo general siguen el mismo esquema de investigación, en el cual se seleccionan conjuntos de datos observados de calidad del aire, meteorología y en algunos casos fuentes de emisión, los cuales son utilizados para entrenar y evaluar un conjunto de modelos con los cuales se realizan predicciones de un conjunto de valores observados para posteriormente verificar su desempeño mediante métricas estadísticas (Bozdağ, Dokuzb, & Gökçekc, 2020) y así elegir cuál fue el modelo que tuvo mayor éxito realizando las predicciones y qué tan precisas fueron las mismas.

En este esquema, el proceso consta de 5 pasos:

1. Adquisición de datos: En este paso se reúnen los datos con los cuales se planea entrenar el modelo. Para el caso de calidad del aire, estos datos suelen ser registros de monitoreo de calidad, meteorología, comportamiento y composición de fuentes.

2. Preprocesamiento de datos: Proceso mediante el cual se consolidan todos los datos en una sola matriz, identificando datos faltantes, formatos de fechas e incoherencias en la resolución temporal, para posteriormente estimar los valores de los datos faltantes, interpolar periodos sin información o realizar técnicas de remuestreo.

3. Ingeniería de características: Después de procesados los datos, se realiza un análisis de cada una de las variables en el conjunto, identificando el tipo (categóricas o numéricas), estandarizando/normalizando datos numéricos, y realizando un análisis de correlación entre variables, con el fin de identificar las variables que tienen más correlación entre sí, excluirlas del aprendizaje y así reducir tiempos de computación y la complejidad de la matriz. En este paso se divide la matriz de datos en dos, una para entrenar el modelo (x-train/y train) y otra para predecir y evaluar el desempeño de los modelos (x-test, y-test).

4. Selección de modelo: Una de las secciones finales del proceso en la cual se evalúa el desempeño de los algoritmos, comparando los valores predichos con los valores

observados (matriz de prueba) mediante el uso de una serie de métricas (Bozdağ, Dokuzb, & Gökçekc, 2020).

5. Interpretación de modelo y predicciones: En este paso se entrenan los algoritmos con los conjuntos de datos mencionados en el anterior numeral y se realizan las predicciones con cada uno de ellos, ajustando los hiperparámetros de los algoritmos para mejorar el desempeño de los mismos. Estos hiperparámetros son parte del algoritmo y no del conjunto de datos de entrenamiento, y se encargan de regular la manera en la cual aprende el algoritmo.

Los modelos de predicción basados en machine learning revisados suelen usar un amplio conjunto de algoritmos. Los más usados suelen ser las redes neuronales, seguidos de regresiones, modelos de ensamble y modelos híbridos (Iskandaryan, Ramos, & Trilles, 2020).

3.2.1 Redes neuronales

Es uno de los conjuntos de algoritmos más usado en la actualidad, su funcionamiento se basa en el comportamiento de las neuronas humanas, donde existe una capa de entrada la cual es precedida por una serie de n-capas ocultas cuyos procesos convergen finalmente en una capa de salida el funcionamiento de este algoritmo es secuencial y redundante, donde a la capa de entrada ingresan las variables a ser procesadas por el modelo para realizar la predicción, en las capas ocultas se asignan una serie de ponderaciones a cada variable las cuales son aplicadas a las funciones de activación de cada neurona con las que posteriormente se realiza la predicción del valor mediante el promedio de los valores predichos por todas las capas de salida, los cuales se comparan con el valor observado para volver a realizar un ajuste de las ponderaciones de las capa ocultas hasta que la predicción sea lo suficientemente cercana al valor observado (Kuiying, Zhou, Sun, Zhao, & Liu , 2019).

Respecto a la calidad del aire, los algoritmos más estudiados según la literatura consultada, suelen ser Redes Neuronales artificiales (ANN por sus siglas en inglés), redes neuronales profundas (Deep learning), Redes Neuronales Recurrentes (RNN por sus siglas en inglés), y Redes Neuronales Convolutivas (CNN por sus siglas en inglés).

3.2.2 Regresiones

Los algoritmos basados en regresiones establecen una relación entre variables independientes y dependientes, tomando como referencia el conjunto de datos de entrenamiento. Este tipo de algoritmos se suele usar para la predicción de datos. La forma más primitiva de este tipo de algoritmos es la regresión lineal, donde la relación entre la variable independiente y dependiente viene dada por la ecuación de la recta y = mx+ b, función que se ajusta a la recta entre dos puntos (Castelli, Martins Clemente, Popovic, Silva, & Vanneschi, 2020).

Entre los algoritmos considerados en los casos de estudio investigados están Potenciación del gradiente (Gradient Boosting), Árboles de decisión, bosques aleatorios, y Máquinas de vectores de soporte.

3.2.3 Modelos de ensamble

Los modelos de ensamble consisten en la utilización en paralelo de varios algoritmos de machine learning con el fin de mejorar la precisión del modelo mediante distintas salidas proporcionadas por cada algoritmo. La principal razón de ser de esta práctica es la necesidad de reducir fenómenos como la alta varianza y posible sesgo que pueda tener una predicción obtenida a partir de un solo algoritmo (Chi-Yeh, Chang, & Abimannan, 2021). Entre los modelos de ensamble más usados para la predicción de calidad del aire, según la bibliografía consultada, se encuentran Adaboost y XGBoost.

3.2.4 Modelos híbridos

Los modelos híbridos, al igual que los modelos de ensamble, suelen utilizar una combinación de varios tipos de algoritmos. Sin embargo, para el caso de los modelos híbridos, estos son usados en serie, siendo implementados principalmente para disminuir la complejidad y la incertidumbre de los datos de entrada para una mejor predicción de resultados (Gu, Li, & Meng, 2022). Las distintas configuraciones de los modelos usados para la predicción de concentraciones de contaminantes atmosféricos suelen ser particulares para cada caso de estudio.

3.2.5 Métricas de desempeño

En la Tabla 1 se presentan las principales métricas de desempeño utilizadas en los casos de estudio presentados más adelante.

Tabla 1. *Métricas de desempeño usadas en los casos de estudio analizados*

Indicador	Siglas	Descripción
Fracción de predicciones con un factor de 2	FAC2	Fracción de los datos donde la razón entre dato predicho y observado se encuentra entre 0.5 y 2
Sesgo Promedio Normalizado (Normalised Mean Bias)	NMB	Diferencia promedio normalizada entre los valores predichos y observados
Error Cuadrático Medio (Root Mean Square Error)	RMSE	Indica la raíz de la desviación entre promedio entre valores pronosticados y observados
R2	R2	Indica el ajuste entre los valores observados y los predichos
Coeficiente de eficiencia (Coefficient of efficiency)	COE	Indica la capacidad de un modelo de representar un valor observado
Índice de acuerdo (Index of Agreement)	IOA	También llamado índice de Willmott, es una medida entre 0 y 1 que representa la razón entre el error cuadrático medio y el error potencial
Erros Absoluto medio (Mean Absolute Error)	MAE	Indica la diferencia entre los valores predichos y los observados
GeoMean	GeoMean	Media geométrica

Nota. Fuente: Elaboración propia (2022)

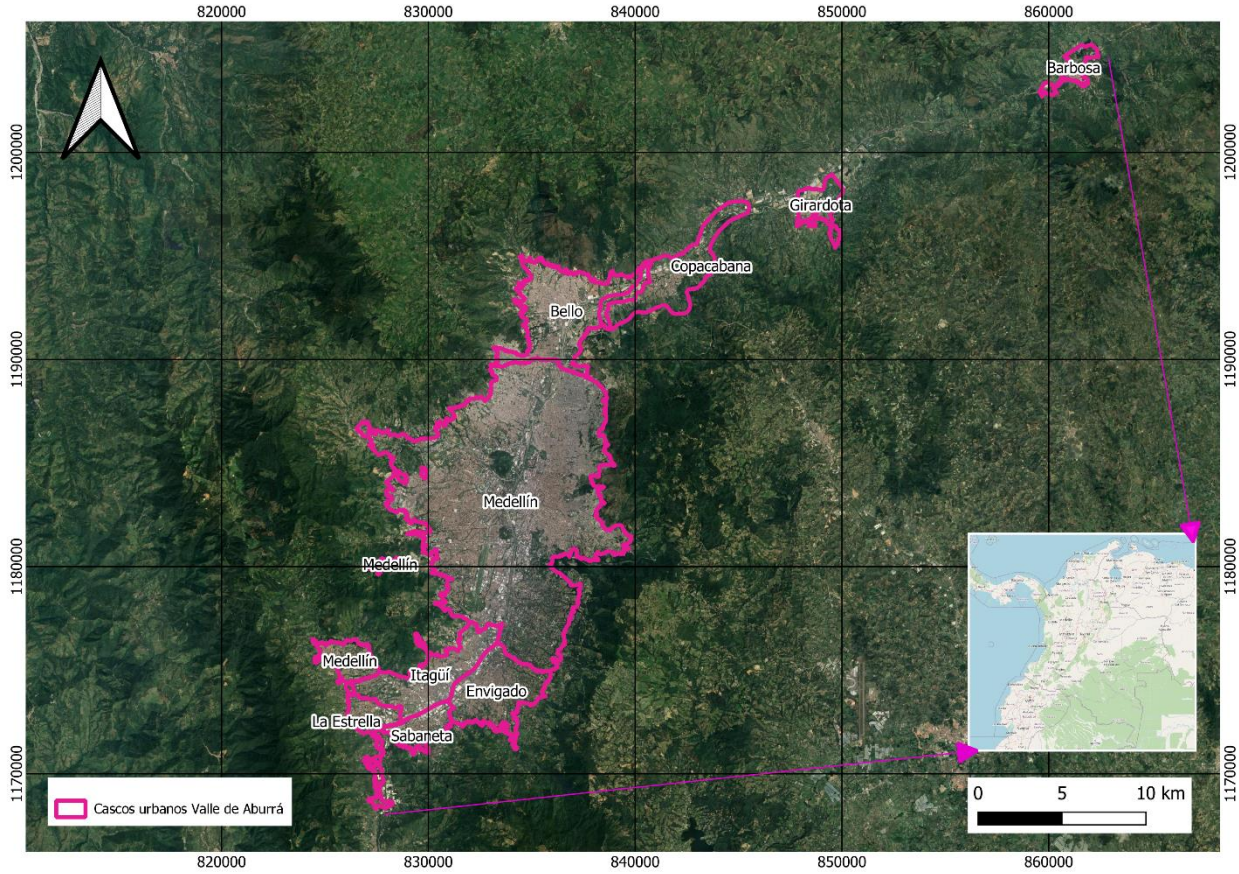
4. Condiciones del Valle de Aburrá que inciden en la calidad del aire

Para identificar con qué datos cuenta el Valle de Aburrá, con el fin de representar en los modelos las condiciones de emisión y dispersión del valle, se realizó un análisis de los fenómenos que lo caracterizan y las variables sobre las cuales se manifiestan.

4.1 Condiciones de dispersión

La calidad del aire en el Valle de Aburrá en gran parte viene definida por sus particularidades meteorológicas, que a su vez son condicionadas por las características morfológicas del área (ver Figura 2), la presencia de fenómenos como la inversión térmica (Bernal Manrique & Rendón Pérez, 2019), la isla de calor urbano (Henao, Rendón, & Salazar, 2020) y la propia

dinámica de los vientos dentro del valle, hace que sus condiciones de estabilidad varíen de forma cíclica durante el año



Nota. Fuente: Elaboración Propia (2022).

Figura 2. *Ubicación geográfica del Valle de Aburrá y sus aglomeraciones urbanas*

Dichos fenómenos afectan las condiciones de estabilidad de la atmósfera, alterando el comportamiento de los componentes convectivo y mecánico de la altura de capa de mezcla. Dichas alteraciones pueden identificarse por métodos de medición directos como el uso de radiosondas para el estudio de los perfiles de la turbulencia de la zona, o mediante el análisis de los patrones en las variables meteorológicas correspondientes a los componentes de la altura de capa de mezcla, así:

- Movimiento convectivo: Todas las variables relacionadas al movimiento vertical de la masa de aire, dado principalmente por cambios de densidad (temperatura, radiación, humedad, nubosidad y precipitación) (Hujia, y otros, 2019).

- Movimiento mecánico: Aquellas variables relacionadas al movimiento horizontal de la masa de aire y el flujo del mismo sobre la superficie (velocidad y dirección del viento) (Cuesta-Mosquera, Wahla, Acosta-López, García-Reynoso, & Aristizábal-Zuluaga, 2020).

Estas variables meteorológicas por lo general son medidas por los sistemas de monitoreo de calidad del aire del Valle de Aburrá, por lo que pueden ser descriptores de las condiciones particulares de dispersión del mismo.

4.2 Condiciones de emisión

El Valle de Aburrá presenta una gran cantidad y variedad de fuentes emisoras de contaminantes atmosféricos de origen antrópico, que pueden ser clasificadas como fuentes móviles o industriales, las cuales, según el inventario de emisiones atmosféricas más reciente, han generado emisiones considerables de PM10, PM2.5, VOC, NOx, SOx y NOx (AMVA, 2018).

Según este mismo inventario, las principales fuentes de PM2.5 dentro del Valle de Aburrá son las fuentes móviles, las cuales realizan alrededor del 91% de los aportes. Medios de transporte como los camiones, volquetas y buses especiales son los principales aportantes con un 91% de las emisiones por fuentes móviles (AMVA, 2018).

Teniendo en cuenta esto, es de resaltar que las distintas autoridades municipales del Valle de Aburrá poseen información horaria de tráfico vehicular de sus principales vías, clasificándose por tipo de vehículo y sentido de circulación, por lo que se puede afirmar que también existe información de la ubicación y comportamiento de lo que, según la autoridad ambiental competente, son las principales fuentes de emisiones atmosféricas del área de interés.

Aparte de las fuentes ya mencionadas, también se suelen presentar sucesos de emisión a escala regional y global que afectan directamente las concentraciones de contaminantes atmosféricos del Valle, incidiendo principalmente en la concentración de fondo, la cual es el nivel base sobre el que las fuentes de origen antropogénico realizan el aporte. Dichos sucesos pueden ser desde quema de biomasa (Mendez-Espinosa, Belalcazar, & Morales , 2019) hasta transporte de arenas desde el Sahara (Mendez, Pinto Herrera, & Belalcázar Cerón, 2018). El impacto de estos sucesos se suele manifestar como un aumento en las concentraciones registradas por todas las estaciones de monitoreo de calidad del aire del Valle de Aburrá, independientemente de la

presencia o no de fenómenos meteorológicos cíclicos como los relacionados con los procesos de inversión térmica.

4.3 Disponibilidad de datos

Entre las herramientas disponibles para acceder a datos de calidad del aire y variables relacionadas a las condiciones atmosféricas del Valle de Aburrá, se encuentra el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA), el cual es un sistema que opera la red de monitoreo de calidad del aire y estaciones meteorológicas. Dichas redes reportan en tiempo real las variables monitoreadas ante la plataforma de visualización y descarga de datos del SIATA, desde donde se puede descargar el histórico de los valores monitoreados en un periodo, en alguna de las estaciones de monitoreo activas.

La red de monitoreo de calidad del aire y meteorología está compuesta por aproximadamente 36 estaciones de monitoreo (mayoritariamente automáticas) distribuidas en sitios representativos de todos los municipios de área metropolitana del Valle de Aburrá, esta red monitorea los siguientes contaminantes criterio: PM10, PM2.5, Ozono, CO, SO2, NO, NO2 y NOx. La red también publica las concentraciones monitoreadas con una resolución horaria.

La mayoría de estas estaciones de monitoreo vienen acompañadas de estaciones meteorológicas que constantemente monitorean humedad, precipitación, presión, velocidad y dirección del viento, y radiación solar. Dicha red reporta los valores monitoreados con una resolución de mínimo 1 segundo.

5. Análisis de casos de interés

Para la selección de casos de estudio se tuvieron en cuenta las siguientes características de dispersión, emisión, inmisión y disponibilidad de datos para el Valle de Aburrá:

1. Representatividad meteorológica: Casos en los cuales el conjunto de datos incluyese algún conjunto de variables relacionadas a los fenómenos convectivos (temperatura, radiación, humedad, precipitación) y mecánicos (dirección y velocidad del viento) del transporte de contaminantes.

2. Características de emisión: Casos que el conjunto de datos de entrada incluyese alguna variable que indicara la presencia, distancia, emisión o nivel de actividad de las fuentes

generadoras de las concentraciones estudiadas durante el periodo con disponibilidad de datos de concentración monitoreados

3. Características de inmisión: Casos en los cuales los datos de entrada incluyesen información histórica de concentraciones de contaminantes atmosféricos (PM10, PM2.5, NOx, SOx y CO), donde se indicase la concentración en ug/m^3, el periodo de exposición (minutal, horario, diario, mensual), la identificación de pico de concentración y la fecha y hora del registro.

4. Disponibilidad de datos oficiales: Casos en los cuales los datos utilizados para el entrenamiento y prueba de los modelos hubiesen sido medidos y publicados por entidades oficiales encargadas del monitoreo de calidad del aire y meteorología, como lo pueden ser sistemas de monitoreo municipales, redes de vigilancia de corredores viales o zonas industriales.

En la Tabla 2 se presentan los casos de estudio de mayor interés por su similaridad con alguna de las condiciones identificadas para el caso del Valle de Aburrá, junto con sus principales características.

Tabla 2. Análisis de casos de interés

Nombre del estudio	Datos de entrada	Algoritmos utilizados	Métricas usadas	Algoritmo de mayor desempeño
Using Machine Learning to estimate the impact of ports and cruise ship traffic on urban air quality: The case of Barcelona.	*Concentraciones diarias de PM10, NO, NO2, Nox, SO2, CO y O3 *Tráfico rodado, marítimo y aéreo *Condiciones meteorológicas (Costa)	GBM, SVM, , Random Forest, ANN (Retroalimentada-multicapa), Multivariate Adaptive Regression Splines (earth), ANN (Retroalimentada-multinomial para modelos log-lineales).	GeoMean, R, MB, RMSE	GBM
Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey	*Concentraciones de PM10	LASSO, SVM, Random Forest, k-Nearest Neighbor, eXtreme Gradient Boosting algorithms, ANN	R2, RMSE, MAE	ANN

Nombre del estudio	Datos de entrada	Algoritmos utilizados	Métricas usadas	Algoritmo de mayor desempeño
Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5)	Concentraciones diarias de PM10 y PM2.5 *Tráfico rodado, marítimo y aéreo *Variables meteorológicas	ANN, BRT y SVM	FAC2, NMB, NMGE, RMSE, R, COE, IOA	BRT, ANN
Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2.5 using boosted gradient machine models	*Concentraciones diarias de PM10, PM2.5, NO, NO2, SO2 y CO. *Picos de concentración identificados *Condiciones meteorológicas	GBM	MSE, RMSE	
A Machine Learning Approach to Predict Air Quality in California	*Concentraciones horarias (CO, SO2, NO2, O3 y PM2.5). *Meteorología	SVR, PCA SVR	MAE, R2, RMSE, nRMSE	Ambos similares
Smart City Air Quality Prediction using Machine Learning	*Concentraciones de PM2.5. *Ubicación de las estaciones de monitoreo	Multilayer perceptrón, Random forest	Matriz de confusión (Accuracy, Precision, Recall)	Random Forest
Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia	*Concentraciones horarias de PM10, PM2.5, NO, NO2 y O3. *Variables meteorológicas	ANN (Retroalimentada), SVR-PSO	RMSE	SVR-PSO

Como base de análisis se tiene un caso de estudio de aplicación de herramientas de machine learning para la estimación de concentraciones de material particulado PM2.5 en el Valle de Aburrá (Murillo-Escobar, Sepulveda-Suescun, Correa, & Orrego Matute, 2019), en el que utilizó un conjunto de datos conformado por variables meteorológicas (Dirección del viento, velocidad del viento, temperatura, humedad relativa y radiación solar) y características de inmisión (concentraciones horarias, diarias y semanales de PM10, PM2.5, NO, NO2 y O3). Dicha aplicación, si bien indicó una alta capacidad para predecir O_3, registró una tendencia a

subestimar concentraciones de material particulado (PM10 y PM2.5), por lo que se indica que el modelo debe mejorar, ya que tiende a subestimar este contaminante.

Si se compara el estudio base con otras aplicaciones como la desarrollada por Bozdağ, Dokuzb, & Gökçekc (2020), en Ankara Turquía, o por Murugan & Palanichamy (2021) en distintas ciudades de Malasia, se puede observar el impacto que puede tener la no inclusión de información de meteorología, ya que el RMSE de algoritmos similares (ANN retroalimentado y Support Vector Machine/Regression) suele dar más bajo en el caso base, el cual sí tiene en cuenta variables meteorológicas.

Por otro lado, el estudio base no cuenta con datos de emisión. Sin embargo, el caso de estudio analizado por Suleiman, Tight, & Quinn (2018), muestra cómo la variable relacionada con la emisión de los vehículos (g/km) tiene un alto impacto en la variabilidad de los datos de PM2.5. De forma similar a la del caso de estudio analizado por Fabregat, Vázquez, & Vernet (2020), la variable relacionada con la intensidad del tráfico es de las de mayor importancia en las estaciones de monitoreo de tráfico urbano.

Una de las principales capacidades que debe de tener un algoritmo de predicción de concentración de PM2.5 para el Valle de Aburrá es la de predecir picos de concentración. El estudio base no tiene esta capacidad. Sin embargo, como lo indica Miskell, Pattinson, & Weissert (2019) en su estudio en la ciudad de Christchurch, es posible realizar predicciones de picos de concentración de distintas duraciones con una precisión de entre el 80% y el 90%. Para esto se debe incluir en el conjunto de datos una variable que identifique si el registro de concentración se clasifica como pico o no. Las variables que tuvieron una mayor influencia en la predicción de esta variable fueron concentraciones de NO, NO_2, temperatura y vientos (para picos de corta duración), presión atmosférica y temperatura (para picos de larga duración).

Un análisis integral de todos los casos de estudio muestra el uso de una gran variedad de algoritmos de machine learning para la predicción de contaminantes atmosféricos, presentando también una variedad en los análisis de resultados respecto a cuáles algoritmos presentaron el mejor desempeño según las métricas. De igual forma, se logra ver cómo cada vez más se implementan modelos de ensamble y modelos híbridos, y con ello la aparición de algoritmos cada vez más complejos y de uso no tan amplio en todos los casos de estudio.

Todos los casos de estudio presentados en esta sección utilizaron datos de meteorología, inmisión y emisión tomados de fuentes oficiales como sistemas urbanos de monitoreo, redes de vigilancia de corredores viales o zonas industriales, ya sea como información primaria o como insumo para el cálculo o generación de más variables complementarias como la identificación de picos.

6. Discusión

6.1 Posibles beneficios de estas tecnologías en el contexto del Valle de Aburrá

Alrededor del mundo se están realizando pilotos de implementación de modelos basados en algoritmos de machine learning sobre sistemas de información relacionados con el monitoreo de calidad del aire. Estos pilotos abarcan desde casos de estudio que hacen predicciones mediante modelos entrenados únicamente con datos históricos de concentración con altos índices de precisión (Murugan & Palanichamy, 2021), hasta modelos con conjuntos de datos de entrada más complejos que utilizan concentraciones de material particulado, registros de meteorología e información de tráfico de las fuentes móviles más cercanas (datos proporcionados por las redes de monitoreo locales), para predecir no solo concentraciones diarias de material particulado sino también identificar el conjunto de fuentes que genera mayores aportes (Suleiman, Tight, & Quinn, 2018). Estos son ejemplos de casos de estudio extranjeros que se asemejan al tipo de información que brinda el SIATA, y que cuentan con herramientas de datos abiertos similares a las del área metropolitana. Por lo tanto, se tiene la infraestructura de datos necesaria para construir conjuntos de datos de entrenamiento para implementar modelos de predicción de calidad del aire mediante herramientas de machine learning.

Lo anterior se corrobora con el estudio realizado por Murillo-Escobar et al. (2019), en el que se entrenaron modelos de predicción de concentraciones diarias y semanales de contaminantes (PM10, PM2.5, NO_2, NO, O_3) a partir de información meteorológica y de calidad del aire tomada de las estaciones del SIATA. Según estos autores, estos modelos fueron capaces de predecir las tendencias pero subestimaron las concentraciones de los posibles picos abruptos que se presentan a lo largo del año, de forma similar al caso de estudio planteado por Mogollón-Sotelo et al. (2020) para Bogotá, donde el error aumentaba con los cambios abruptos de concentración.

La correcta predicción de picos de contaminantes con estos modelos puede mejorar en la medida en que se incorporen más datos históricos de concentraciones pico identificadas y etiquetadas (Miskell, Pattinson, & Weissert, 2019), algo que aún no ha sucedido con los casos de estudio encontrados hasta el momento para el Valle de Aburrá u otras ciudades de Colombia, pero que eventualmente se puede alcanzar.

6.2 Desafío en la selección y desarrollo de algoritmos

Según el análisis de casos de interés, no hay una tendencia clara respecto al tipo de algoritmos que presenta un mejor desempeño. Esto se debe a la aparición exponencial de algoritmos disponibles en los últimos años, que incluyen los modelos de ensamble y modelos híbridos, por lo que cada algoritmo responde más a un diseño específico del marco experimental del caso de estudio que a la elección razonada de un algoritmo ya definido en su totalidad, generando una dispersión de datos y metodologías utilizadas para la elaboración de dichos algoritmos, dificultando la repetibilidad y reusabilidad de los mismos (Hartley & Olsson, 2020).

Sin embargo, se puede afirmar que los tipos de algoritmo de mejor desempeño están relacionados con aplicaciones de modelos de ensamble basadas en redes neuronales (NN) y modelos de regresión.

Varios casos de estudio referenciados en esta monografía mencionan la posibilidad de realizar una implementación de modelos basados en machine learning para predicción, ya sea en sistemas de alerta temprana o para predicción en general. Sin embargo, no se encuentra hasta la fecha un caso de estudio que compare el tiempo y los recursos computacionales necesarios para realizar una predicción con modelos de machine learning con los necesarios para correr los modelos determinísticos usados actualmente.

6.3 Disponibilidad y validez de datos de entrada

En todos los casos de interés estudiados, los conjuntos de datos utilizados estaban compuestos por variables que son actualmente monitoreadas por el SIATA (meteorología e inmisión), o que están disponibles en portales de datos abiertos (emisión de fuentes móviles, volumen de tráfico vehicular y composición del tráfico), a excepción de la identificación de picos de concentración que, si bien se pueden deducir de la información disponible, implican un esfuerzo de preprocesamiento extra a la hora de hacer este tipo de aplicaciones para el Valle de Aburrá.

Si bien hay estudios como el de Mogollón-Sotelo et al. (2020) que resaltan la capacidad de estos algoritmos para describir relaciones complejas entre factores como la topografía, la meteorología y la concentración de contaminantes, es necesario ampliar la información disponible sobre el tema, con casos de estudio donde se compare el desempeño de los modelos con y sin el uso de datos de meteorología en zonas de topografía compleja y propensa a fenómenos de inversión térmica como en el Valle de Aburrá.

7. Conclusiones

En los últimos años, los avances y usos en los algoritmos de machine learning han aumentado en todos los campos. Para el campo ambiental, específicamente para la predicción de las condiciones de calidad del aire, también hay un aumento en la publicación de casos de estudio. Está monografía discute los posibles beneficios que puede brindar la aplicación de este tipo de modelos para la predicción de concentraciones de PM2.5 en el Valle de Aburrá, el cual presenta problemas de calidad del aire en la actualidad y precisa de más herramientas que permitan una gestión ambiental del mismo.

Según los casos de estudio analizados, es posible implementar un modelo de predicción de concentraciones de contaminantes atmosféricos para el Valle de Aburrá que sea capaz de estimar con bajos porcentajes de error niveles de concentración e identificar picos abruptos de inmisión, mediante modelos que representen las complejas relaciones que se dan entre las particularidades topográficas y meteorológicas de la zona, con las características de emisión e inmisión de las fuentes del área estudiada.

Posterior al análisis de los casos de estudio de interés, no se pudo identificar un algoritmo específico como el de mejor desempeño para este tipo de aplicaciones. Por lo general, los algoritmos correspondían a una configuración específica realizada para el caso de estudio, por lo que en la mayoría de casos los algoritmos de mejor desempeño no fueron repetidos o reutilizados en otros casos. Sin embargo, sí fue posible identificar los tipos de algoritmo de mayor desempeño, los cuales corresponden a redes neuronales (RN) y modelos de regresión.

Las particularidades topográficas del Valle de Aburrá y fenómenos meteorológicos como la inversión térmica, perturbación en la dinámica de vientos, o cambios en la temperatura

superficial por fenómenos como la isla urbana de calor se reflejan en las variables meteorológicas monitoreadas para las zonas de estudio, por lo que pueden ser incluidos en los conjuntos de datos de entrenamiento, pudiendo tener incidencia directa en un mejor desempeño de los modelos de predicción.

Para el desarrollo de aplicaciones de alerta temprana cuyos algoritmos sean capaces de predecir picos de concentración de contaminantes es necesario preprocesar información de concentraciones históricas el conjunto de datos de entrenamiento, mediante la debida identificación y clasificación de la concentración en caso de que este sea un pico o no.

Actualmente el Valle de Aburrá cuenta con todos los datos necesarios para el entrenamiento e implementación de este tipo de modelos de predicción en aplicaciones de alerta temprana, ya que tanto condiciones de inmisión (concentración histórica y actual de contaminantes atmosféricos) como meteorológicas son actualmente monitoreadas y reportadas por el SIATA, aparte de también tener información adicional sobre fuentes de emisión como factores de emisión de fuentes móviles, información de volumen y composición de tráfico vehicular y comportamiento.

Referencias

AMVA. (2017). Acuerdo metropolitano no 16. Medellín.

AMVA. (2017). Documento del plan integral de gestión de la calidad del aire - PIGECA. Medellín.

AMVA. (2018). Actualización inventario de emisiones atmosféricas del Valle de Aburrá – AÑO 2018. Medellín.

AMVA. (2021). Plan de acción para la implementación del plan operacional para enfrentar episodios de contaminación atmosférica (poeca) en el área metropolitana del Valle de Aburrá. Medellín.

Becerra, M. A., Uribe, Y., Peluffo Ordóñez, D. H., Álvarez Uribe, K. G., & Tobón, C. (2021). Information fusion and information quality assessment for environmental forecasting. *Urban Climate*.

Bernal Manrique, N., & Rendón Pérez, A. M. (2019). Analysis of surface meteorological conditions in the Aburrá Valley and its possible effects on air quality. *Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP)*.

Bozdağ, A., Dokuzb, Y., & Gökçekc, Ö. B. (2020). Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environmental Pollution*.

Casallas, A., Ferro, C., Celis, N., Guevara-Luna, M. A., Mogollón-Sotelo, C., Guevara-Luna, F. A., & Merchán , M. (2021). Long short-term memory artificial neural network

approach to forecast meteorology and PM2.5 local variables in Bogotá, Colombia. *Modeling Earth Systems and Environment*.

Castelli, M., Martins Clemente, F., Popovic, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*.

Chi-Yeh, L., Chang, Y.-S., & Abimannan, S. (2021). Ensemble multifeatured deep learning models for air quality forecasting. *Atmospheric Pollution Research*.

Cuesta-Mosquera, A., Wahla, M., Acosta-López, J. G., García-Reynoso, J. A., & Aristizábal-Zuluaga, B. H. (2020). Mixing layer height and slope wind oscillation: Factors that control ambient. *Sustainable Cities and Society*.

Fabregat, A., Vázquez, L., & Vernet , A. (2020). Using Machine Learning to estimate the impact of ports and cruise ship traffic on urban air quality: The case of Barcelona. *Environmental Modelling and Software*.

Gu, Y., Li, B., & Meng, Q. (2022). Hybrid interpretable predictive machine learning model for air pollution. *Neurocomputing*.

Hartley, M., & Olsson, T. (2020). dtoolAI: Reproducibility for Deep Learning. *Patterns*.

Henao, J. J., Rendón, A. M., & Salazar, J. F. (2020). Trade-off between urban heat island mitigation and air quality in urban valleys. *Urban Climate*.

Hujia, Z., Huizheng, C., Xiangao, X., Yaqiang, W., Hong, W., Peng, W., . . . Xiaoye, Z. (2019). Climatology of mixing layer height in China based on multi-year. *Atmospheric Environment*.

Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. *Applied Sciences*.

Kuiying, G., Zhou, Y., Sun, H., Zhao, L., & Liu , S. (2019). Prediction of air quality in Shenzhen based on neural network. *Neural Computing and Applications*.

Masih, A. (2019). Machine learning algorithms in air quality modeling. *Global Journal of Environmental Science and Management*.

Mendez, J. F., Pinto Herrera, L. C., & Belalcázar Cerón, L. C. (2018). Study of a Saharan Dust Intrusion into the Atmosphere of Colombia. *Revista Ingenierías Universidad de Medellín*.

Mendez-Espinosa, J. F., Belalcazar, L. C., & Morales , R. (2019). Regional air quality impact of northern South America biomass burning emissions. *Atmospheric Environment*.

Minghao, Q., Zigler, C., & Selin, N. (2022). Statistical and Machine Learning Methods for Evaluating Trends in Air Quality under Changing Meteorological Conditions. *Atmospheric Chemistry and Physics*.

Miskell, G., Pattinson, W., & Weissert, L. (2019). Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2.5 using boosted gradient machine models. *Journal of Environmental Management*.

Mogollón-Sotelo, C., Casallas García, A., Vidal, S., Celis Mayorga, N., Ferro, C., & Belalcazar, L. (2020). A support vector machine model to forecast ground-level PM 2.5 in a highly populated city with a complex terrain. *Air Quality, Atmosphere & Health*.

Murillo-Escobar, J., Sepulveda-Suescun, J. P., Correa, M., & Orrego Matute, D. (2019). Forecasting concentrations of air pollutants using support vector regression improved

with particle swarm optimization: Case study in Aburrá Valley, Colombia. *Urban Climate*.

Murugan, R., & Palanichamy, N. (2021). Smart City Air Quality Prediction using Machine Learning. *Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)*.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*.

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. *Applied sciences*.

Suleiman, A., Tight, M. R., & Quinn, A. D. (2018). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*.