Identificación de variables influyentes en la superación de niveles máximos permisibles de PM 2.5 en Bogotá y su uso en modelos de Machine Learning.

Cristian Duvan Laverde Quiroga

Cristian Rubén Herrera Cuartas

Asesor

Fernando Luis Carrascal Porras

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Dedicatoria

Dedicamos esta monografía a nuestras familias, y a todas las personas que, al igual que nosotros, sienten la curiosidad y el compromiso de entender más sobre algo tan vital como es la calidad de nuestro aire. A quienes, con su pasión por el conocimient o y la investigación, continuarán explorando y ampliando el camino que hoy comenzamos a trazar. Esperamos que este trabajo sirva como un humilde peldaño para alcanzar nuevas preguntas, hallazgos y soluciones que impacten positivamente a nuestra sociedad y al mundo.

Agradecimientos

Agradecemos profundamente a nuestras familias, por su apoyo incondicional y paciencia en cada etapa de este proceso. A nuestro tutor, el profesor Fernando Luis Carrascal Porras, por su guía y sabiduría, que enriquecieron enormemente este trabajo. A nuestros compañeros de clase y amigos, por los momentos compartidos y su estímulo constante, y a la Universidad Nacional Abierta y a Distancia, por los recursos y el ambiente académico brindado. Finalmente, dedicamos un especial reconocimiento a todos aquellos que, con su labor diaria, contribuyen al avance del conocimiento científico y la investigación.

Resumen

Esta monografía propone el estudio de modelos de Machine Learning (ML) con el objetivo de identificar las variables que intervienen cuando la concentración de PM2.5 supera los límites permisibles en Bogotá, considerando que la mayoría de las ciudades superan los estándares recomendados de calidad del aire.

El planteamiento del problema destaca los efectos adversos de la contaminación del aire en la salud y la necesidad de encontrar formas de anticiparse y encontrar estrategias predictivas y preventivas a la ocurrencia de eventos críticos, resaltando el potencial del ML en este ámbito.

A través de una revisión bibliográfica se examinarán distintos modelos de ML aplicados o estudiados en diferentes ciudades del mundo para evaluar su aplicabilidad en Bogotá. Con ello se busca determinar el modelo que mejor se adecue en la predicción de la ocurrencia o no ocurrencia de la superación de los limites permisibles y las variables más influyentes en el evento.

***Palabras clave*:** Contaminación, PM 2.5, PM 10, Machine learning, Bogotá

Abstract

This monograph proposes a study of Machine Learning (ML) models with the aim of identifying the variables involved when PM2.5 concentrations exceed permissible limits in Bogotá, acknowledging that most cities surpass recommended air quality standards.

The problem statement highlights the adverse effects of air pollution on health and the need to proactively develop predictive and preventive strategies for critical events, emphasizing ML's potential in this domain.

Through a literature review, various ML models applied or studied in different cities worldwide will be examined to assess their applicability to Bogotá. This will help determine the model best suited to predict whether permissible limits will be exceeded and identify the most influential variables in these events.

***Keywords*:** Pollution, PM 2.5, PM 10, Machine Learning, Bogotá

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

La exposición prolongada a contaminantes del aire es uno de los principales factores en el desarrollo de Enfermedades No Transmisibles (ENT), como las enfermedades vasculares y respiratorias, que, según la Organización Mundial de la Salud – OMS - causan cerca de 5,7 millones de muertes al año (Organización Mundial de la Salud, 2023). Estas enfermedades tienden a ser más graves según la intensidad, frecuencia y duración de la exposición (Medina, 2019). Increíblemente el 80% de las ciudades del mundo superan los estándares recomendados de calidad del aire (Mathew et al., 2023).

El PM2.5 y otros contaminantes presentes en el aire afectan el desarrollo mental y motor de los niños, así como la función pulmonar, incluso en bajas concentraciones (Vargas et al., 2023) y en cuanto a la morbilidad, las partículas de PM2.5 se encuentran entre los desencadenantes de síntomas de diferentes enfermedades que afectan las vías respiratorias (Álvarez et al., 2023).

En diferentes ciudades del territorio colombiano se ha estudiado el contaminante PM 2.5 logrando caracterizarlo y analizarlo con una visión descriptiva de las concentraciones de este según la data recolectada, además de estimar sus efectos en la salud de las personas que habitan allí y la presión al sistema de salud.

Sin embargo, se pretende entender esta problemática desde una perspectiva del Machine Learning identificando modelos aplicados en diferentes lugares del mundo y su posible aplicación en Bogotá, de tal manera que se pueda identificar las variables más influyentes en el evento que las concentraciones de PM2.5 superan los límites permisibles en Bogotá, y su uso en modelos de ML teniendo en cuenta la alta volatilidad de este contaminante (Zhang et al., 2021). Sabiendo que muchos datos son capturados de manera constante por la Red de Monitoreo de

Calidad del Aire de Bogotá (RMCAB) (Secretaría Distrital de Ambiente, 2024), la Secretaría Distrital de Ambiente y la Corporación Autónoma Regional de Cundinamarca.

En ese sentido, ¿Cuáles son las variables más relevantes cuando las concentraciones de PM2.5 superan los límites permisibles en Bogotá, y como podemos usarlas en modelos de ML para predecir cuándo se podría dar este evento?

Justificación

Se denomina como PM 2.5 a las pequeñas partículas suspendidas en el aire con un diámetro aerodinámico inferior a 2.5 micras, (Gregorio et al., 2022; Mathew et al., 2023). Este ha sido bastante estudiado debido a su capacidad de penetración en el tracto respiratorio (Grisales et al., 2021), siendo un factor de riesgo importante para poblaciones vulnerables como los infantes, quienes son especialmente susceptibles a los efectos de los contaminantes en el aire (Pan et al., 2020) o las poblaciones de escasos recursos las cuales experimentan una mayor exposición a la contaminación en todas sus formas (Song et al., 2024), lo cual aumenta significativamente el riesgo para ellos debido a que este aumenta a medida que lo hace la exposición (Rodriguez et al., 2022).

Cuando podemos predecir ya no estamos en el plano de las acciones reactivas, sino que por el contrario podemos tomar medidas a nivel comunitario antes de que ocurra un evento para evitarlo o atenuarlo (Dhake et al., 2024). Así las cosas, es importante plantear mecanismos de pronóstico que puedan utilizarse, de modo que se pueda anticipar qué medidas tomar cuando se prevean altas concentraciones de ciertos contaminantes en un futuro próximo entendiendo, por supuesto, las variables que podrían desencadenar niveles por fuera de lo recomendado y de lo permisible.

Los modelos de Machine Learning han demostrado un potencial sólido al permitir la gestión de grandes volúmenes de datos y brindar un mayor nivel de exactitud en los datos estimados (Aram et al., 2023). Y con la recopilación y estudio de datos continuos a largo plazo, es posible la investigación y el desarrollo de políticas y estrategias de gestión (Wright et al., 2023). En este contexto se muestra como una opción robusta frente a la previsión del evento de que ocurra la superación de niveles máximos permisibles de PM 2.5.

En ese sentido, los datos recolectados por la Red de Monitoreo de calidad del Aire de Bogotá, y demás entidades, no solo servirán para recomendar a la población tomar medidas preventivas cuando se determine concentraciones de contaminantes nocivas, sino que también posibilita la aparición de nuevas perspectivas para abordar los problemas de concentraciones nocivas de contaminantes, y proteger la salud y bienestar de la comunidad; esto debe hacerse de forma oportuna, de manera que se eviten incidentes graves de contaminación (Vieru y Cărbureanu, 2024; Liu et al., 2019).

Las mejoras en los métodos de gestión de la calidad del aire pueden tener impactos positivos en diferentes áreas, pues de mantenerse los niveles del contaminante por debajo de los permitidos por la normatividad Colombiana, Resolución MADS 2254 de 01 de noviembre de 2017 (Ministerio de Ambiente y Desarrollo Sostenible, 2017), se reduciría el desarrollo de enfermedades respiratorias asociadas a la exposición de niveles nocivos PM2.5 y cientos de vidas podrían salvarse cada año con efectos positivos incalculables desde el punto de vista social y económico (Álvarez et al., 2023). Un efecto aún mayor si se mantiene por debajo de los niveles recomendados por la Organización Mundial de la Salud.

A través de una revisión sistemática, se podrá determinar cuáles variables influyen significativamente en el aumento de la concentración de PM 2.5. Luego, se podrá determinar el modelo (o modelos) de ML con mayor potencial para estimar las variables que influyen significativamente en el evento que se superan los límites permisibles por la normatividad.

Esto es importante, porque sienta las bases para el futuro desarrollo de un modelo que pueda predecir o clasificar de manera efectiva el comportamiento de la variable PM2.5, y sobre todo predecir eventos en los que se pueda superar los límites, para crear planes o políticas públicas encaminadas a evitar la ocurrencia de estos eventos.

Objetivos

Objetivo General

Analizar los Modelos de Machine Learning aplicados en diferentes ciudades haciendo énfasis en las variables independientes utilizadas y las condiciones del área de estudio y su aplicación en el contexto de Bogotá.

Objetivos Específicos

Identificar a través de una revisión bibiliografica las variables que más influyen en el evento de la superación de los niveles máximos permisibles de PM 2.5, y su aplicación en modelos de Machine Learning.

Realizar una revisión bibliográfica sobre los factores que influyen en los niveles de PM 2.5 y las técnicas de Machine Learning aplicadas a la predicción de contaminantes atmosféricos.

Reconocer las técnicas estadísticas y de Machine Learning usadas para la identificación de las variables clave que influyen en la concentración PM 2.5.

Marco Conceptual y Teórico.

La calidad del aire afecta a todos los miembros de la sociedad y la creciente preocupación por ella ha dado lugar a la necesidad de estrategias en la gestión de la calidad del aire y el control de la contaminación (Aram et al., 2023). Contaminantes atmosféricos como el monóxido de carbono (CO), compuestos organicos volátiles (COV), el dióxido de azufre (SO2), el dióxido de nitrógeno (NO2), el ozono (O3), y las partículas en suspensión (PM10 y PM2.5) afectan negativamente a la calidad de la atmósfera y tienen un impacto negativo en la salud humana (Aram et al., 2023) y están directamente ligados a la ocurrencia de fenómenos meteorológicos igualmente nocivos como las lluvias ácidas (Zhu et al., 2019).

El Índice de Calidad del Aire (AQI, por sus siglas en inglés) es una función lineal fragmentada de las concentraciones de contaminantes como el ozono (O3), material particulado (PM2,5, PM10), monóxido de carbono (CO), dióxido de azufre (SO2) y dióxido de nitrógeno (NO2), que puede verse afectada por condiciones climatológicas o geográficas que cambian su comportamiento (Xu et al., 2023).

No existe un estándar global y cada país tiene su propio índice; las normas y regulaciones de calidad del aire varían según la región y el país (Méndez et al., 2023). Para Colombia, los límites máximos permisibles se regulan por la Resolución 2254 de 2017 del Ministerio de Ambiente y Desarrollo Sostenible (Ministerio de Ambiente y Desarrollo Sostenible, 2017), pero discrepan de los límites recomendados por la Organización Mundial de Salud, siendo más laxos (Grisales et al., 2021).

Es muy importante tener en cuenta que las concentraciones de los contaminantes varían temporalmente dependiendo de meteorología local y de las emisiones de esa zona geográfica (Mathew et al., 2023). Para el caso particular de Bogotá las emisiones se deben en gran medida a

fuentes fijas industriales debido a una falta de actualización de los equipos utilizados, a fuentes móviles como lo son los vehículos motorizados (Sanchez et al., 2022) .

Comúnmente las ciudades de países emergentes presentan altas emisiones debido a vehículos en mal estado y vías inadecuadas (Janarthanan et al., 2021), agravado con el hecho de no manejar métodos alternativos de transporte menos contaminantes (Cao et al., 2023), algo que en efecto se puede relacionar para el caso particular de Bogotá por la elevada concentración de residuos metálicos en agua, terreno y aire en la ciudad (Zafra et al., 2019), lo que lleva a medidas tales como el "pico y placa ambiental" donde se intenta frenar el flujo vehicular (Alcaldía de Bogotá, 2021).

Los modelos ML se han utilizado ampliamente para predecir tanto el AQI como el nivel de concentración de contaminantes específicos relacionados con la calidad del aire (Méndez et al., 2023). A raíz de ello surgen corrientes altamente efectivas, como los modelos híbridos donde se aplican los beneficios de las técnicas de descomposición de los métodos de ML y los algoritmos de optimización (Wu y Lin, 2019). Además, los métodos de ensamble tienden a suplir las debilidades de otros modelos (Zhai y Chen, 2018), generando una optimización en los resultados y mejorando los resultados conseguidos (Medhi et al., 2024)

Dentro de los algoritmos de ML más utilizados para la predicción de concentraciones de contaminantes tenemos las Regresiones Lineales (ya sean simples o múltiples), algoritmos basados en regresiones como la Regresión Vectorial de Soporte – SVR-, Arboles de Decisión y Bosques Aleatorios y K vecinos más cercanos (KNN).

Cabe destacar que también se ha utilizado ampliamente algoritmos de Deep Learning -DL- que puede verse como una evolución del ML y que utiliza una estructura de múltiples capas llamada red neuronal artificial (Méndez et al., 2023).

Regresión Lineal: Se utiliza para inferir la relación entre una variable dependiente y un conjunto de variables independientes ayudando a predecir un valor continuo (Méndez et al., 2023). Lo que se busca es construir una ecuación lineal que minimice el error cuadrático medio, es decir, una línea recta que se adapte a la nube de puntos.

Regresión Lineal Simple: Cuando hablamos de Regresión Lineal Simple nos referimos a que se usa una única variable independiente y una variable objetivo o dependiente (Liang et al., 2020).

Regresión Lineal Múltiple: Para este caso, lo único que cambia es que el número de variables predictoras es superior a una, es decir, tenemos más de una variable predictora para una variable objetivo.

En estos modelos la correlación de las variables indica el grado de dependencia o relación entre dos variables siempre que estas sean cuantitativas (Cuellar et al., 2022)

Media Móvil Autorregresiva (ARIMA): Como indican Méndez et. al (2023), este algoritmo utiliza series temporales para la predicción, lo que quiere decir que realiza predicciones de acuerdo con valores pasados y presentes. Este modelo incluye el componente autorregresivo que es el número de retrasos utilizados, el componente integrado que representa el grado de diferenciación requerido para convertir la serie temporal en estacionaria, y, por último, el componente la de media móvil, se refiere al número de errores pasados necesarios para explicar el error actual (Zhu et al., 2017).

Máquinas de Vectores de Soporte (SVM): Son modelos de aprendizaje supervisado con algoritmos de aprendizaje asociados que analizan los datos proporcionados con el objetivo de realizar una clasificación o una regresión (Liu et al., 2019).

Regresión Vectorial de Soporte (SVR):Crea modelos de regresión precisos al encontrar un hiperplano optimo mientras minimiza la diferencia entre los valores predichos y los reales (Vieru y Cărbureanu, 2024). El objetivo es encontrar una función de regresión lineal, luego se considera un margen de tolerancia esperando que todos los datos estén a una distancia máxima del hiperplano y que no se desvíe de las etiquetas (Liu et al., 2019).

Es más interesante cuando se introducen funciones no lineales, pues se introducen funciones Kernel para transformar los datos en un espacio dimensional superior a este y así desarrollar una transformación de regresión lineal (Méndez et al., 2023). Los métodos Kernel son una clase de técnicas de aprendizaje automático cada vez más popular para tareas como el reconocimiento de patrones, la clasificación o la detección de novedades ya que estos proporcionan un puente entre la linealidad y la no linealidad (Mahalingam et al., 2019).

Arboles de Decisión: Como sintetizan Méndez et al. (2023), este algoritmo se basa en un particionamiento recursivo con el objetivo de diseñar un modelo de predicción de una variable cuantitativa a partir de un conjunto de variables independientes. La estructura resultante se asemeja a un árbol, donde se parte de un nodo raíz que se divide en función de la variable independiente más relevante y luego los demás nodos se dividen de nuevo considerando la variable con la menor suma de la estimación cuadrática de errores como nodo de decisión. Los últimos nodos se conocen como nodos de salidas u hojas.

Bosques Aleatorios: Este es un método de aprendizaje de conjuntos para la clasificación, regresión y otras tareas que se basa en la construcción de múltiples arboles de decisión en diferentes momentos de entrenamiento y generando la clase que representa el modo de clases (clasificación) o la predicción media (regresión) de árboles individuales (Liu et al., 2019) y la predicción será el promedio de las predicciones proporcionadas por los diferentes árboles. Este

método sobresale con conjuntos de datos complejos pues combina múltiples árboles para proporcionar predicciones fiables, mitiga el sobre ajuste y es un experto en capturar relaciones intrincadas en los datos de calidad del aire (Vieru y Cărbureanu, 2024).

K- vecinos más cercanos (KNN): Como mencionan Méndez et. al (2023), el algoritmo KNN se aplica a comúnmente a problemas de clasificación, aunque también se puede aplicar a problemas de regresión. Este algoritmo se basa en el principio de similitud de los datos y las predicciones se realizan considerando la clase mayoritaria de los k-vecinos más cercanos (Vieru y Cărbureanu, 2024). En términos más simples el algoritmo calcula la distancia entre un punto y los puntos del conjunto de datos de entrenamiento para seleccionar los k vecinos más cercanos y establece el promedio de ellos (una votación) como predicción (Zhang S. , 2022).

Aumento de gradiente basado en histogramas (HGBoost): La combinación de algoritmos basados en histogramas en coordinación con aumentos del gradiente construyen modelos de aprendizaje automático de alto rendimiento (Nhat-Duc y Van-Duc, 2023), haciendo que el algoritmo HGBoost se aferre a estructuras de datos basadas en valores enteros en lugar de depender de valores continuos ordenados convirtiéndolo en el modelo adecuado para capturar relaciones no lineales complejas en los datos (Mathew et al., 2023).

AdaBoost (Adaptative Bosting): Este modelo mejora la precisión de la predicción a través de la combinación de múltiples modelos débiles y la asignación de más peso a los puntos de datos mal clasificados (Vieru y Cărbureanu, 2024), cambiando la distribución de los datos de entrenamiento según el error de predicción después de cada iteración para obtener múltiples predictores (Liu y Chen, 2020)

Es relevante hacer la anotación de que si bien los modelos de ML resultan de vital importancia en la predicción y gestión de los índices de contaminantes en el aire también se

necesita una correcta gestión de la data que alimentará dichos modelos, esta normalmente se gestiona por medio de técnicas de preprocesamiento propias de Big Data que optimizan los modelos y permiten la adquisición de resultados y estimaciones adecuadas (Gangwar et al., 2023).

Al verlo desde la perspectiva de predicción del contaminante PM2.5, entendemos a esta como la variable objetivo y variables como la temperatura, humedad, tráfico, incluso otros compuestos que se miden en el aire como el SO2, NO2, CO entre otros, como las variables independientes o predictoras (Mathew et al., 2023).

Metodología

1. Enfoque Metodológico: Se emplea un enfoque cualitativo y descriptivo con un diseño de revisión documental. Con este enfoque se logra analizar y sintetizar estudios existentes sobre los factores que afectan las concentraciones de PM2.5 y también evaluar el potencial de los modelos de Machine Learning (ML) para predecir la ocurrencia de la superación de los limites permisibles de este contaminante.

Dado que el objetivo es generar una comprensión profunda de las variables implicadas y los modelos aplicados en contextos similares, se opta por un análisis no experimental, ya que no se recogen datos originales ni se realizan intervenciones experimentales.

2. Revisión de la Literatura: La base de esta monografía es la revisión detallada de la literatura científica relacionada con los factores que influyen en el comportamiento del PM2.5, así como con el uso de algoritmos de ML en el contexto de la predicción de eventos en los que se superen los límites máximos permisibles por la normativa colombiana. Esta revisión incluirá artículos de bases de datos académicas como Scopus, Web of Science, Google Scholar, Scielo, publicaciones de organizaciones gubernamentales entre otras fuentes, focalizándose en estudios sobre:

- Impacto del PM2.5 en la salud pública, especialmente en entornos urbanos.
- Variables ambientales y de origen humano que influyen en las concentraciones de PM2.5.
- Modelos de ML aplicados para predecir la concentración de PM2.5 y su precisión.
- Normativas nacionales y extranjeras sobre la gestión de contaminantes PM2.5.

3. Definición de Variables Relevantes que influyen en las concentraciones de PM 2.5: Con base en la literatura revisada, se identificará y analizará un conjunto de variables clave

que influyen en la concentración de PM2.5. Estas variables incluyen factores climáticos, fuentes de emisión, entre otros.

4. Análisis Modelos de Machine Learning: Se analizará críticamente los enfoques y modelos de ML empleados en estudios de predicción de contaminantes atmosféricos y calidad del aire en general, tales como regresiones lineales, árboles de decisión, bosques aleatorios y métodos de ensamble, la revisión se enfocará en:

- Ventajas, limitaciones y métricas de rendimiento de cada modelo en la predicción de PM2.5.
- Adaptabilidad de los modelos al contexto de la ciudad de Bogotá.

5. Documentación y Síntesis de Hallazgos: Se procederá con una síntesis de los hallazgos obtenidos mediante la revisión de la literatura y el análisis de modelos ML; la información será organizada de manera que se destaquen los hallazgos principales de cada estudio analizado, las variables determinantes en el aumento de PM2.5 y la aplicabilidad de los modelos revisados.

Para ello, se emplearán también tablas y gráficos que permitan sintetizar los estudios analizados y las características más relevantes de cada modelo.

6. Propuestas para Investigaciones Futuras: Con base en los hallazgos y limitaciones identificadas, se ofrecerán recomendaciones para investigaciones futuras que deseen aplicar ML para la predicción de PM2.5 en Bogotá.

Desarrollo

Análisis de las Fuentes de Contaminación

Bogotá es uno de los mayores centros urbanos de Colombia. A enero de 2024 la población había llegado a los 7.929.539 habitantes (Departamento Administrativo Nacional de Estadistica - DANE, 2024). Esta alta densidad va de la mano con un alto flujo vehicular e industrial, que funcionan como un detonante para el aumento en los índices de contaminación como consecuencia de las actividades industriales y las actividades cotidianas de la población (Ministerio de Salud y Protección Social, 2024).

Las fuentes de contaminación se clasifican normalmente en 2 tipos diferentes, móviles y fijas (Granada-Aguirre et al., 2014). Una fuente móvil es toda fuente de emisión que, por razón de su uso o propósito, es susceptible de desplazarse (Vazquez, 2021). Las fuentes móviles normalmente son las que emiten directa e indirectamente la mayor cantidad de contaminantes a la atmosfera . Estas se originan por la combustión interna producto de la quema de combustibles fósiles en vehículos particulares, motocicletas, aviones y sobre todo vehículos que utilizan Diesel como combustible (Sanchez et al., 2022).

Tanto así que se estima que los vehículos particulares livianos contribuyen con más del 70 % de las emisiones de CO y SO2, y más del 50% de la emisión de COV y CO2. Los vehículos de transporte masivo e intermunicipal, y los camiones contribuyen con un 26% y 28% respectivamente, y además producen las mayores emisiones de PM2.5 (Vazquez, 2021). Recordemos que dentro de los componentes de las partículas de PM2.5 pueden encontrarse varios elementos como el plomo, hierro, níquel cobre, platino, cromo, sodio, carbón y otros compuestos orgánicos (Gutierrez et al., 2020).

Para estimar la cantidad de contaminante liberado se usan factores de emisión e información de actividad vehicular; el primero representa un valor que relaciona la cantidad de compuesto que es emitido a la atmosfera y una unidad de actividad (Velasco & Bernabe, 2004; Vazquez, 2021). Este factor se expresa como la relación entre la cantidad de contaminante liberado a la atmósfera y la distancia recorrida, o la cantidad de combustible consumido (Vazquez, 2021).

Algo muy importante, es que en Colombia a falta de mediciones directas de los factores de emisión (FE), los inventarios de emisiones obtenidos a nivel nacional a lo largo de los años se han basado en factores de emisión medidos en otros países o factores simulados con modelos de emisión (Ramirez et al., 2019; Vazquez, 2021). Esto reduce la confiabilidad de los resultados, ya que son necesarios valores y/o funciones de tasas de emisión precisas que permitan establecer puntos de referencia reales para el planteamiento de programas de gestión de la calidad del aire (Vazquez, 2021).

Ahora, de acuerdo con el Inventario de Emisiones de Bogotá, se estima que para el año 2022 se emitieron alrededor de 4628 toneledas de PM2.5. Se destaca como fuentes la resuspensión de Material Particulado por el transito en vias no pavimentadas (40%), emisiones por fuentes móviles por combustión (31%), resuspensión por canteras y construcciónes (15%), maquinaria amarilla (9%) y por ultimo fuentes fijas industriales, residenciales y comerciales (2%), visible en la tabla 1.

Tabla 1

Emisiones de PM2.5 en toneladas por año y detalle de fuente de emisión.

Fuentes de emisión		Toneladas	% del total
Móviles	En Carretera	1435.8	31%
	Fuera de carretera	430.4	9%

Fijas	Industriales	74	2%
	Comerciales	67.6	1%
	Residenciales	2.9	0%
Fuentes naturales y forestales	Incendios forestales	55	1%
Re suspendido	Desgaste de frenos y llantas	99	2%
	Vías pavimentadas	319.7	7%
	Vías no pavimentadas	1446.2	31%
	Rehabilitación de vías	122.9	3%
	Canteras	280.7	6%
	Construcción de edificaciones	293.2	6%
Total		4627.4	100%

Nota. Estimación de emisiones de contaminante PM2.5 a la atmosfera en el año 2022 y detalle de fuentes de emisión. *Fuente.* Inventario de Emisiones de Bogotá, 2022.

La re suspensión de partículas en vías no pavimentadas y las emisiones producto de la combustión interna en vehículos tienen un impacto considerablemente alto en la generación de contaminación por PM2.5, en comparación con otras fuentes. En consecuencia, podemos establecer una correlación significativa entre el flujo vehicular y este contaminante, ya que la suma del impacto de ambas fuentes representa más del 70% del total de emisiones de *PM2.5*.

Ahora, el número de vehículos que transitan en la ciudad es algo que se debe tener presente. De acuerdo con las cifras presentadas por el Observatorio de movilidad, para noviembre de 2024, se registraron cerca de 2.590.000 vehículos en Bogotá, una cifra que ha tendido al crecimiento en los últimos años, tendencia que se evidencia en la figura 1 y la tabla 2 en conjunto.

Tabla 2

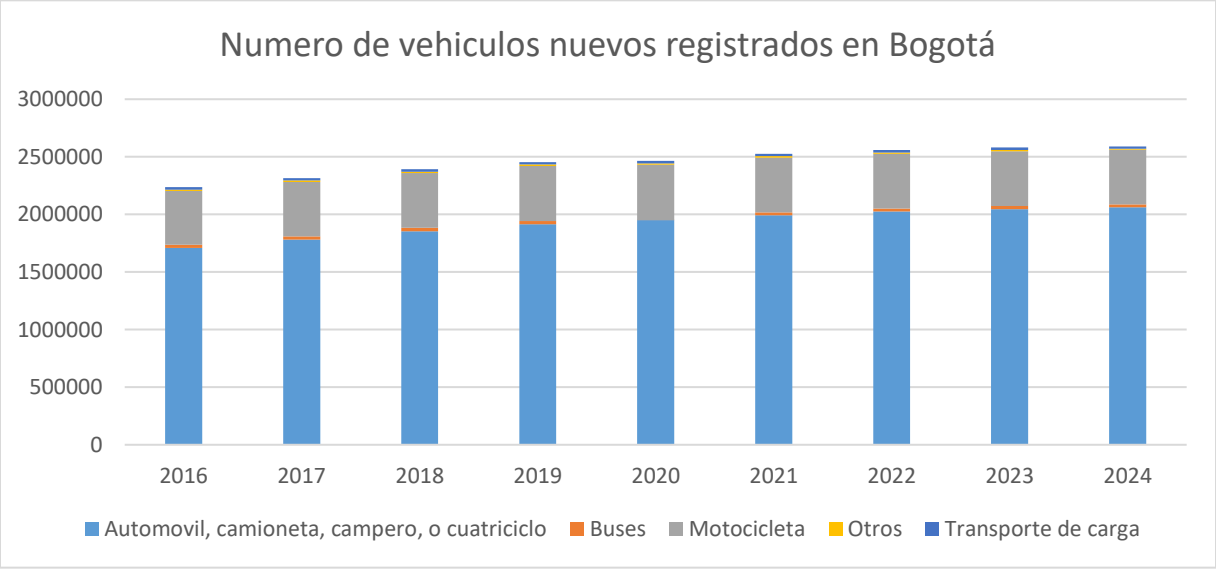
Numero de vehículos registrados en Bogotá por año.

Año	Automóvil, camioneta, campero, o cuatriciclo	Buses	Motocicleta	Otros	Transporte de carga	Total
2016	1 708.307	29.502	466.046	11.340	21.168	2 236.363
2017	1 781.026	29.286	472.239	11.715	20.984	2 315.250
2018	1 854.175	28.573	476.159	12.035	21.426	2 392.368
2019	1 915.756	27.766	477.715	12.358	21.105	2 454.700
2020	1 951.796	28.200	477.927	12.504	20.896	2 463.405
2021	1 992.133	25.476	475.735	12.636	20.541	2 526.521
2022	2 024.665	25.719	474.977	12.782	20.318	2 558.461
2023	2 046.515	25.224	475.776	11.792	21.303	2 580.610
2024	2 061.301	24.906	476.923	7.397	19.277	2 589.804

Nota. Datos obtenidos del Observatorio de Movilidad de Bogotá, 2024.

Figura 1

Numero de vehículos automotores registrados desde 2016



Nota. Adaptado de datos del Observatorio de Movilidad de Bogotá

Por otro lado, existe también evidencia de la existencia de correlación entre contaminantes atmosféricos y variables meteorológicas (humedad, radiación solar, temperatura) de tal manera que se sabe que los factores meteorológicos influyen, modifican, aumentan o disminuyen la concentración de los contaminantes (Gutierrez et al., 2020). Además, algunos estudios han demostrado correlación entre partículas pesadas y ligeras, por lo tanto, se puede establecer que la concentración y existencia de alguna de ellas depende de la concentración y existencia de otra (Gutierrez et al., 2020).

Pero también hay una característica que se debe tener en cuenta y es que, aunque existe una correlación entre el comportamiento de los pares de partículas PM10 y PM2.5 esto puede deberse a que las segundas son parte del rango definido para las partículas PM10 (Gutierrez et al., 2020). Tanto así, que en el estudio realizado por López et al. (2024) al analizar el comportamiento del PM2.5 se vio que era muy similar al del PM10, y se estableció que esto pudo deberse a que el primero constituye cerca del 50% del segundo. Lo que evidencia la influencia que tienen sobre el PM2.5 otras fuentes de material particulado y las emisiones derivadas por la combustión en las fuentes móviles (López et al., 2024).

Este mismo estudió demostró que en presencia de escasas precipitaciones y vientos débiles se sobrepasan las concentraciones máximas admisibles para varios contaminantes, incluido el PM2.5.

Análisis de Modelos de Machine Learning

A continuación, se relacionan 10 diferentes estudios relacionados con predicción de calidad del aire en diferentes ciudades del mundo con diferentes enfoques:

En el artículo desarrollado por Méndez et al. (2023), Machine Learning Algorithms to Forecast Air Quality: A Survey, se realiza una revisión integral de los principales algoritmos de Machine Learning (ML) aplicados a la predicción de la calidad del aire durante el periodo 2011-2021.

El estudio se centra en el Índice de Calidad del Aire (AQI) y su relación con contaminantes clave como PM2.5, PM10, SO2, NO2, y O3; aunque no existe un estándar global para el AQI, los autores destacan su uso generalizado como medida para determinar los niveles de contaminación y orientar políticas públicas.

Los autores analizaron 155 estudios, clasificando las contribuciones en función de los algoritmos utilizados, las áreas geográficas estudiadas, las variables predictoras y las métricas de evaluación, en este proceso se identificaron dos grandes categorías de modelos; algoritmos de aprendizaje profundo (DL) y algoritmos de regresión. Entre los algoritmos más utilizados destacan:

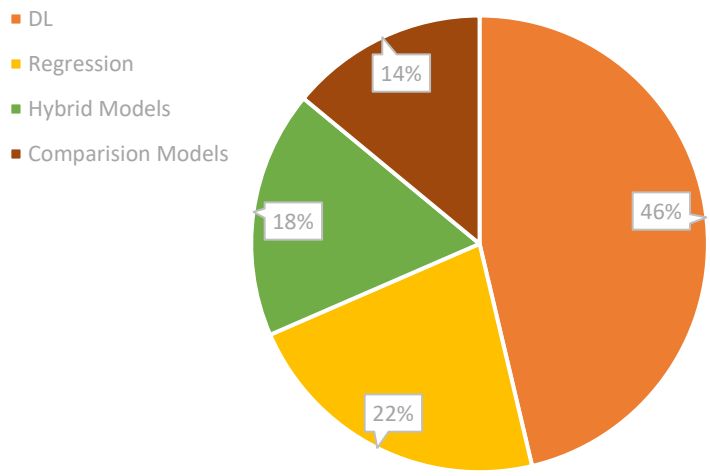
- LSTM y MLP (Redes Neuronales): Predominantes en la predicción de series temporales debido a su excelente capacidad para manejar dependencias temporales y no linealidades.
- Random Forest (RF): De amplio uso debido a su adaptabilidad y robustez frente a datos ruidosos.
- Support Vector Regression (SVR): Este es reconocido por su precisión en escenarios de alta dimensionalidad.

El estudio concluye que los algoritmos de DL suelen superar a los de regresión en más del 60% de los casos, particularmente en escenarios con múltiples variables predictoras y

correlaciones complejas. Por ello se convierte en uno de los más usados, como podemos ver en la figura 2.

Figura 2

Conteo algoritmos utilizados



Nota: Elabodo a partir de la obra de Méndez et al., 2023.

En el estudio se refuerza que las variables relacionadas con los contaminantes (PM2.5, NO2, y O3) y las condiciones meteorológicas (velocidad del viento, temperatura y humedad relativa) son las más utilizadas en los modelos.

Las métricas como el RMSE, MAE y R^2 son comúnmente empleadas para evaluar el rendimiento de los modelos, siendo R^2 y MAPE las más indicativas en el contexto de la predicción de calidad del aire y por tanto las más usadas.

La clasificación estructurada que los autores presentan, agrupando algoritmos, métricas de evaluación y variables predictoras, es uno de los puntos fuertes del estudio, ya que facilita no solo la interpretación de los hallazgos, sino también su aplicabilidad en diferentes contextos.

El estudio realizado por Mihai-Claudiu Vieru y Mădălina Cărbureanu (2024), Machine Learning Methods Applied In Air Quality Prediction, aborda la aplicación de técnicas de Machine Learning (ML) para predecir la calidad del aire (AQI), con un enfoque al ML que combina datos históricos y algoritmos avanzados para realizar predicciones en tiempo real.

En el desarrollo del estudio se evalúan cinco algoritmos de ML: K-Nearest Neighbors (K-NN), Random Forest, Gradient Boosting, Support Vector Regression (SVR) y AdaBoost. Los hallazgos destacaron a AdaBoost como el algoritmo más efectivo, seguido de cerca por Gradient Boosting. Los modelos fueron evaluados mediante técnicas de validación cruzada, lo que garantizó su robustez y capacidad de generalización (Vieru y Cărbureanu, 2024).

En el estudio las variables como PM2.5, SO2, y NO2 fueron identificadas como muy influyentes en el Índice de Calidad del Aire (AQI). Las métricas de desempeño, como el MAE y R2, mostraron que AdaBoost tiene el menor error absoluto medio y el mayor coeficiente de determinación, lo que refuerza su capacidad predictiva.

Los autores concluyen que AdaBoost obtuvo mejor rendimiento debido a su enfoque iterativo, que asigna mayores pesos a los datos mal clasificados. Esto permite manejar conjuntos de datos desbalanceados y ruidosos con eficacia, además de su capacidad para capturar patrones complejos y generalizar correctamente a datos no vistos. Al ejecutar la validación cruzada, resultó en menores errores y mayor precisión en métricas clave como MAE y R^2.

Adicional, se concluye que la incorporación de validación cruzada mejoraría la confiabilidad de los modelos en contextos donde los datos pueden ser inconsistentes o incompletos, permitiendo una mayor maniobrabilidad para la construcción de los modelos.

El estudio Machine Learning-Based Prediction of Air Quality Index and Air Quality Grade: A Comparative Analysis realizado por Aram et al. (2024), analiza y compara diferentes

modelos de Machine Learning (ML) para la predicción del Índice de Calidad del Aire (AQI) y la clasificación de los grados de calidad del aire (AQG) en Beijing, China.

El objetivo principal es identificar el modelo más eficiente y preciso para predecir tanto las concentraciones de contaminantes atmosféricos como sus niveles de impacto categorizados. El trabajo se centra en seis contaminantes clave: PM2.5, PM10, NO2, SO2, CO y O3, utilizando datos diarios recopilados entre 2014 y 2019.

El estudio adopta un enfoque bastante interesante al combinar modelos individuales, como Random Forest, Gradient Boosting y LASSO, en una arquitectura de aprendizaje por apilamiento (stacked models), lo que permite mitigar las debilidades de los modelos individuales y mejorar la precisión general.

Principales hallazgos:

1. Modelos Apilados: Superaron consistentemente a los modelos individuales en todas las métricas clave, como R^2, RMSE y MAE para el AQI, y ACC, MCC y F1 para el AQG, esto se puede visualizar en la tabla 3.

Tabla 3

Métricas y resultados de estudio

Modelos	Data de entrenamiento			Data de prueba			Porcentaje de cambio		
	R2	RMSE	MAE	R2	RMSE	MAE	R2%	RMSE%	MAE%
RF	0.0991	6.590	1.048	0.975	7.305	2.099	1.6	11	100
LASSO	0.913	20.781	13.112	0.827	19.073	12.903	10	9	1.6
GB	0.977	10.627	2.663	0.979	6.707	3.241	0.2	37	22
Stack	0.968	12.667	2.471	0.981	6.253	3.040	1.3	102	23

Fuente: Adaptado de Aram et al., (2023, p. 1355)

2. Variables más Influyentes: Se identificó el PM2.5 como la variable con mayor peso predictivo, seguida por PM10, O3, CO, NO2 y SO2, como se evidencia en la tabla 4.

Tabla 4

Correlación variables identificada en estudio

	PM2.5	**PM10**	**SO2**	**CO**	**NO2**	**O3**
PM2.5	1.000	0.847	0.550	0.844	0.782	-0.094
PM10	0.847	1.000	0.533	0.714	0.730	-0.048
SO2	0.550	0.533	1.000	0.626	0.615	-0.272
CO	0.844	0.714	0.626	1.000	0.804	-0.297
NO2	0.782	0.730	0.615	0.804	1.000	-0.339
O3	-0.094	-0.048	-0.272	-0.297	-0.034	1.000

Nota: Adaptado de Aram et al., (2023, p. 1347)

3. Predicción de AQI: El modelo apilado logró un R^2 de 0.973 y un RMSE de 7.568 en su versión final, lo que indica alta precisión y capacidad de generalización.

4. Clasificación de AQG: Para categorizar el AQG, el modelo apilado obtuvo una precisión (ACC) del 97%, destacándose como el mejor modelo para clasificar niveles de contaminación.

Al incorporar modelos apilados se resuelve limitaciones comunes como el sobreajuste y la falta de precisión en datos no estacionarios, algo importante al tratar con la calidad del aire y contaminantes, donde las condiciones pueden variar ampliamente en el tiempo y el espacio.

El estudio Air Quality Analysis and PM2.5 Modelling Using Machine Learning Techniques: A Study of Hyderabad City in India, Mathew et al. (2023), aborda los desafíos

relacionados con la calidad del aire en Hyderabad, una ciudad caracterizada por un crecimiento urbano e industrial acelerado. Su propósito principal es analizar las concentraciones de PM2.5 y desarrollar modelos predictivos utilizando técnicas de ML para identificar patrones y relaciones entre variables clave.

El estudio se basa en un marco que combina datos de contaminantes y meteorología, destacando el impacto significativo de factores climáticos y emisiones antropogénicas en la contaminación del aire.

En este estudio destaca el uso de modelos como la Regresión Lineal Múltiple (MLR), Regresión Basada en K-Nearest Neighbors (KNN) e Histograma Combinado con Gradient Boosting (HGBoost)

El estudio revela patrones estacionales marcados en las concentraciones de PM2.5, observándose un aumento promedio del 68% durante el invierno en comparación con otras estaciones. Este incremento se atribuye a la inversión térmica un fenómeno natural que se produce cuando la temperatura del aire aumenta a medida que lo hace la altura. Este fenómeno atmosférico reduce la dispersión de contaminantes, atrapándolos cerca del suelo y aumentando su concentración.

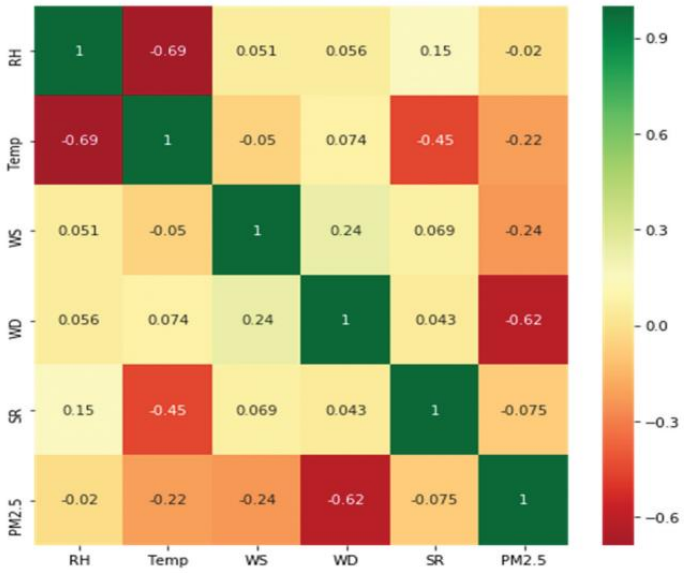
El análisis geoespacial también identificó que las zonas urbanas con tráfico denso y actividades industriales son los principales focos de contaminación, lo que refleja la influencia significativa de las fuentes antropogénicas en los niveles de PM2.5.

Entre los tres modelos evaluados, HGBoost demostró ser el más preciso, con un R2 de 0.859, un error absoluto medio (MAE) de 5.717 $\mu g/m^3$ y una raíz del error cuadrático medio (RMSE) de 7.647 $\mu g/m^3$.

Adicionalmente el análisis de correlaciones identificó relaciones significativas entre las variables meteorológicas y las concentraciones de PM2.5 como se puede observar en la figura 3. La dirección y velocidad del viento mostraron correlaciones negativas (-0.62 y -0.24, respectivamente), lo que indica que una mayor dispersión del viento puede reducir las concentraciones de contaminantes.

Figura 3

Correlación variables metereologicas con PM2.5

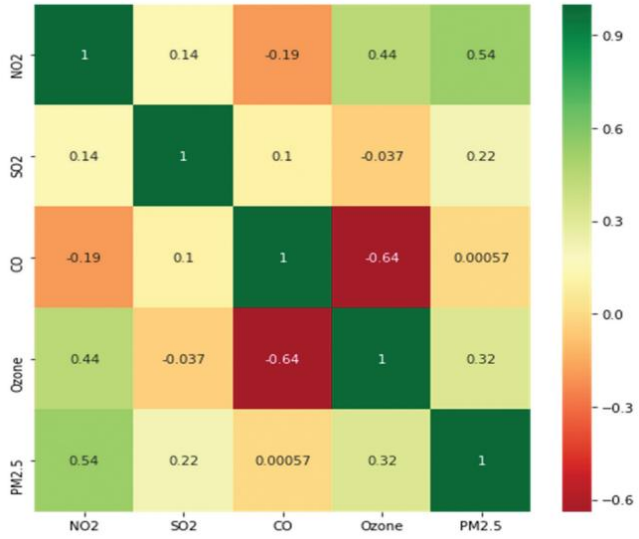


Fuente: Tomada de Mathew et al., 2023.

Entre los contaminantes, el NO2 tuvo la correlación positiva más alta (0.54), lo que sugiere su papel importante en la formación y acumulación de PM2.5, probablemente debido a su relación con emisiones vehiculares e industriales, evidenciable en la figura 4.

Figura 4

Correlación entre contaminantes



Fuente: Tomado de Mathew et al., 2023.

Estos hallazgos subrayan la importancia de integrar tanto factores meteorológicos como contaminantes específicos en los modelos predictivos para capturar las dinámicas subyacentes de la contaminación y las complejidades de los sistemas ambientales.

A Machine Learning Model for Air Quality Prediction for Smart Cities, estudio realizado por Mahalingam et al. (2019), busca desarrollar un modelo de Machine Learning (ML) para predecir el Índice de Calidad del Aire (AQI) en ciudades inteligentes. Esta metodología se centra en mejorar la precisión y robustez de las predicciones mediante el uso combinado de Redes Neuronales Artificiales (ANN) y Máquinas de Soporte Vectorial (SVM).

El estudio utilizó datos de diciembre de 2019, mes caracterizado por niveles elevados de contaminación debido a niebla y smog. Se recopilaron registros de 37 estaciones de monitoreo, incluyendo parámetros críticos como PM2.5, PM10, NO2, SO2 y CO.

Los algoritmos utilizados fueron Redes Neuronales Artificiales (ANN), Máquinas de Soporte Vectorial (SVM), y un modelo híbrido que combina ANN y SVM en dos etapas, ANN realiza predicciones iniciales del AQI, mientras que SVM refina los resultados, maximizando la separación entre categorías.

Como resultado se obtuvo que ANN superó a los métodos tradicionales en términos de error medio cuadrático (RMSE), demostrando buena capacidad predictiva. Sin embargo, SVM, utilizando el kernel gaussiano medio, logró una mayor precisión, del 97.3%. Este rendimiento destaca la efectividad de SVM para modelar relaciones complejas en los datos. La combinación híbrida de ambos algoritmos mejoró aún más la precisión global del sistema.

Se identificó que las variables PM2.5 y PM10 fueron las más influyentes en la predicción del AQI, mostrando correlaciones altas con los valores de calidad del aire.

La combinación de estos algoritmos permite capturar relaciones no lineales y dinámicas complejas presentes en los datos, lo que representa un avance significativo frente a los métodos tradicionales.

La validación del modelo mediante funciones kernel optimizadas y métricas como la validación cruzada, refuerza la confiabilidad de los resultados, asegurando que las predicciones sean precisas y generalizables, además, la propuesta de un modelo híbrido demuestra la capacidad de ANN para procesar grandes volúmenes de datos y la fortaleza de SVM en la clasificación precisa, logrando con ellos un sistema complementario y eficiente.

Sin embargo, el estudio también presenta limitaciones, ya que el uso de un data set limitado a un solo mes de datos restringe la capacidad del modelo para capturar patrones estacionales, como las variaciones en la calidad del aire durante los períodos de monzón o verano.

En el estudio Predicting PM2.5 Concentrations Across USA Using Machine Learning, Preetham Vignesh et al. (2023), se aborda la predicción de concentraciones de PM2.5 en Estados Unidos mediante modelos de machine learning. En un contexto donde la contaminación del aire sigue siendo un desafío crítico para la salud pública y el medio ambiente, este trabajo se centra en evaluar la eficacia de múltiples algoritmos de aprendizaje automático para proporcionar predicciones precisas y confiables. Los datos utilizados abarcan una variedad de características meteorológicas y contaminantes, lo que permite analizar las complejas relaciones entre estas variables y los niveles de PM2.5.

Está fundamentado en la necesidad de desarrollar modelos predictivos robustos que permitan una gestión más eficaz de la calidad del aire, contribuyendo a la identificación de variables críticas y estrategias preventivas (Preetham Vignesh et al., 2023).

Para su desarrollo se implementan una amplia gama de modelos de machine learning para predecir concentraciones de PM2.5. Entre ellos:

1. Regresión Lineal (Linear Regression): Se utilizó como modelo básico de referencia para analizar relaciones lineales entre variables predictoras y la concentración de PM2.5.
2. Árboles de Decisión (Decision Tree): Aplicados para dividir los datos en subconjuntos homogéneos, facilitando la identificación de patrones clave en las variables predictoras.
3. Random Forest (RF): Utilizado como modelo ensamblado para combinar múltiples árboles de decisión y capturar relaciones más complejas en los datos.
4. XGBoost y AdaBoost: Implementados para mejorar las predicciones mediante boosting, ajustando iterativamente modelos débiles para corregir errores.

5. K-Nearest Neighbors (KNN): Se empleó para detectar patrones locales en los datos basándose en la proximidad entre puntos.

6. Gradient Boosting Regression: Usado para realizar predicciones continuas optimizando errores sucesivos mediante árboles de decisión.

7. Support Vector Regression (SVR): Se utilizó para modelar relaciones no lineales entre las variables predictoras y la concentración de PM2.5.

8. LSTM (Long Short-Term Memory): Aplicado para capturar dependencias temporales y analizar tendencias en series históricas de datos.

El análisis reveló, como se puede observar en la tabla 5, que los modelos basados en aprendizaje ensamblado, como Random Forest y Gradient Boosting, superaron consistentemente a otros en términos de precisión y generalización. Random Forest alcanzó la mejor relación entre simplicidad computacional y exactitud, con valores bajos de error cuadrático medio (RMSE) y alta correlación con los datos reales. Por otro lado, aunque el modelo LSTM mostró potencial para capturar tendencias temporales, no logró superar el rendimiento de los modelos basados en Machine Learning clásico.

Tabla 5

Diferentes métricas de modelos para el estado de Nueva York

Modelo	RMSE	MAE	MAPE	R^2	NSE	NORM	PBIAS	RSR
Linear regression	3.883	2.309	0.285	0.688	0.613	60.156	11.24	0.561
Decision tree	5.136	3.109	0.254	0.454	0.533	79.58	13.44	0.691
Gradient boost regressor	3.822	2.394	0.545	0.698	0.683	59.207	8.210	0.546

AdaBoost regressor	3.961	2.316	0.188	0.676	0.683	61.369	9.653	0.576
XGBoost	3.898	2.501	0.202	0.686	0.681	60.393	8.342	0.559
KNeighbors regressor	3.919	2.379	0.195	0.683	0.677	60.711	7.515	0.562
LSTM	7.487	3.359	0.218	0.158	0.455	115.991	6.002	0.812
Random forest	3.121	2.122	0.182	0.899	0.811	38.671	2.989	0.338
SVM	3.125	2.145	0.183	0.857	0.820	39.161	3.011	0.338

Nota. Tomada de Preetham Vignesh et al., (2023, p. 12)

Adicionalmente, se identificaron varias variables predictoras importantes:

- PM2.5: Como variable objetivo, se midió en función de otros contaminantes y factores meteorológicos.
- PM10, NO2 y O3: Fueron las variables más influyentes en la predicción de PM2.5.
- Velocidad del viento y temperatura: También mostraron correlaciones significativas, indicando que las condiciones meteorológicas desempeñan un papel clave.

En términos de aplicabilidad, la investigación demuestra que los modelos basados en Machine Learning son herramientas efectivas para predecir concentraciones de PM2.5 y podrían adaptarse a contextos urbanos complejos como Bogotá. La integración futura de enfoques híbridos, combinando la robustez de algoritmos clásicos con la capacidad de deep learning para capturar patrones temporales, podría potenciar significativamente la precisión y la utilidad práctica de estas soluciones predictivas, contribuyendo de manera más eficaz a la gestión preventiva de la calidad del aire.

En el estudio Spatial Air Quality Index Prediction Model Based on Decomposition Adaptive Boosting and Three-Stage Feature Selection: A Case Study in China, Liu y Chen (2020), se centran en la predicción del Índice de Calidad del Aire (AQI) mediante un modelo híbrido que combina técnicas avanzadas de descomposición temporal, selección de características y algoritmos de aprendizaje automático basados en boosting adaptativo. El objetivo principal es mejorar la precisión de las predicciones del AQI al optimizar la selección de variables relevantes y manejar datos no estacionarios.

Para abordar estas complejidades, el estudio propone un modelo que integra una selección de características en tres etapas, la descomposición de series temporales mediante la transformada de paquetes de onda discreta con solapamiento máximo (MODWPT) y el algoritmo de boosting adaptativo combinado con máquinas de aprendizaje extremas robustas a valores atípicos (ORELM) .

El modelo propuesto superó a cinco modelos base y tres híbridos previos en términos de precisión, demostrando su capacidad para manejar la complejidad de la predicción del Índice de Calidad del Aire (AQI). En seis estaciones de monitoreo, los valores de MAPE, MAE y RMSE fueron consistentemente más bajos, lo que refleja una reducción significativa de los errores en comparación con otros enfoques. La combinación de técnicas avanzadas, como la descomposición temporal mediante MODWPT y el boosting adaptativo con ORELM, permitió al modelo capturar patrones complejos y manejar valores atípicos de manera efectiva.

El análisis de variables mostró que PM2.5 y PM10 son los factores más influyentes en la predicción del AQI, debido a su alta correlación con los valores de contaminación y su impacto directo en la calidad del aire. En contraste, el monóxido de carbono (CO) tuvo una influencia menor, lo que sugiere que su papel en el AQI es más limitado en el contexto estudiado. Este

hallazgo valida la prioridad de PM2.5 y PM10 en los modelos predictivos y en la formulación de políticas de gestión de la calidad del aire, dado su efecto predominante en los niveles de contaminación.

La integración de datos espaciales mejoró significativamente la precisión del modelo. Al seleccionar estaciones con alta correlación espacial, se capturaron de manera más efectiva las dinámicas locales de contaminación. Esto permitió al modelo representar con mayor precisión las variaciones geográficas en los niveles de AQI, lo que refuerza la importancia de incluir datos espaciales para abordar las diferencias locales en la calidad del aire, especialmente en áreas urbanas densamente pobladas o con emisiones industriales relevantes.

Medhi et al. (2024) en su estudio Homogenous Ensemble Learning for Air Quality Index Prediction destacan la aplicación de técnicas de aprendizaje en conjunto homogéneo (ensemble learning). El propósito principal es realizar un análisis comparativo entre modelos ensamblados y no ensamblados, optimizando hiperparámetros para cada caso y evaluando el impacto de las variables en la predicción del AQI.

El estudio implementó bagging (Bootstrap Aggregating) como la principal técnica de ensamblado, utilizando modelos homogéneos como regresión lineal múltiple (MLR), máquinas de soporte vectorial (SVM), árboles de decisión (DT) y perceptrón multicapa (MLP). Los autores indican que Bagging funciona al crear múltiples subconjuntos de datos de entrenamiento mediante bootstrapping, es decir, muestreos aleatorios con reemplazo, permitiendo que cada modelo base se entrene en una versión un poco diferente del conjunto de datos original.

Esta estrategia reduce la varianza al promediar las predicciones de los modelos base, y también mejora la capacidad de generalización del modelo final, ya que cada modelo base captura diferentes patrones y relaciones en los datos.

Este enfoque es útil para manejar problemas de sobreajuste. Por ejemplo, en los árboles de decisión es capaz de suavizar las predicciones mediante el ensamblado de múltiples árboles. Para cada modelo base, el bagging permitió explotar sus fortalezas específicas, por ejemplo, en el caso de los árboles de decisión, el ensamblado mejoró la precisión al minimizar errores residuales que podrían ser amplificados en un único árbol. En SVM, el bagging ayudó a suavizar las predicciones al generar hiperplanos que representaran mejor la distribución de los datos.

Aplicaron técnicas avanzadas de validación cruzada (k-fold) para evaluar la capacidad de generalización de los modelos y ajustar sus hiperparámetros. Este método divide los datos en k subconjuntos (folds), utilizando uno como conjunto de prueba y los restantes como conjunto de entrenamiento en cada iteración, asegurando que cada dato sea utilizado tanto para entrenamiento como para validación. Esto permite obtener una evaluación del desempeño del modelo de mayor calidad y menos sesgada.

La evaluación de los modelos ensamblados y no ensamblados se realizó utilizando métricas estándar de desempeño, lo que permitió una comparación entre ambos enfoques. Los modelos ensamblados demostraron un desempeño superior en todas las métricas de error comparados con sus contrapartes no ensambladas, en particular, el modelo ensamblado basado en árboles de decisión (enDT) sobresalió significativamente, logrando valores de RMSE casi nulos y un coeficiente de determinación (R^2) de 1.0 cuando se utilizaron 100 estimadores.

Este resultado refleja la capacidad del enfoque de bagging para capturar patrones complejos en los datos al combinar múltiples predicciones individuales de árboles de decisión. Cada árbol contribuye a reducir el error residual, mejorando la precisión general y aumentando la robustez del modelo frente a variaciones en los datos. Este enfoque es especialmente efectivo en escenarios donde las relaciones no lineales entre variables dominan la dinámica de los datos.

El análisis de las variables predictoras destacó a PM2.5 y PM10 como los factores más influyentes en la predicción del Índice de Calidad del Aire (AQI). Estas partículas finas, debido a su tamaño microscópico y su composición química, tienen un impacto directo en la calidad del aire y la salud humana. Su alta relevancia en el modelo subraya su papel fundamental en la contaminación del aire, especialmente en áreas urbanas e industriales.

En contraste, variables como el monóxido de carbono (CO) mostraron una contribución significativamente menor en las predicciones, lo que sugiere que, aunque relevantes en ciertos contextos, tienen un impacto más limitado en la determinación del AQI en los escenarios estudiados. Este hallazgo refuerza la necesidad de priorizar la gestión de partículas finas en las políticas de calidad del aire.

Entre los modelos ensamblados, el desempeño del modelo basado en máquinas de soporte vectorial (enSVM) fue el menos destacado, mostrando una capacidad limitada para ajustarse a los datos y capturar relaciones complejas, esto puede deberse a la dificultad de SVM para manejar datos de alta dimensionalidad y ruido cuando se utiliza en un esquema de bagging. En contraste, los modelos ensamblados basados en perceptrones multicapa (enMLP) y regresión lineal múltiple (enMLR) presentaron mejoras moderadas en precisión y consistencia, destacándose como alternativas intermedias.

El modelo MLP aprovechó su capacidad para capturar relaciones no lineales, enMLR se benefició del ensamblado para reducir la varianza inherente a los datos. Estas comparaciones resaltan que el desempeño del enfoque de bagging depende no solo del ensamblado, sino también de las características del modelo base utilizado.

El estudio Unmasking the Sky: High-Resolution PM2.5 Prediction in Texas Using Machine Learning Techniques de Zhang et al. (2024), se centra en la aplicación de técnicas

avanzadas de machine learning para mejorar las predicciones de concentraciones de PM2.5 con alta resolución espacial, aborda la incapacidad de los modelos tradicionales para capturar adecuadamente la variabilidad espacial y las complejas interacciones entre los factores meteorológicos, las emisiones locales y las características geográficas.

Se utilizaron datos provenientes de sensores de calidad del aire, combinados con información meteorológica y geoespacial, para entrenar y evaluar una variedad de modelos de machine learning.

Se utilizaron los modelos Random Forest (RF), Gradient Boosting Machines (GBM), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), y Redes Neuronales Artificiales (ANN).

Los modelo RF y GBM destacaron como los modelos más precisos, lo cual se puede ver en la tabla 6, mostrando menor error cuadrático medio (RMSE) y alta correlación con los datos reales. ANN también presentó resultados competitivos, pero requirió más tiempo de ajuste.

Tabla 6

Desempeño de predicción de los modelos Gradient Boosting y Random Forest

Año	Gradient Boosting - In-sample	Gradient Boosting - Out-of-sample	Random Forest - In-sample	Random Forest - Out-of-sample
	R^2	RMSE	R^2	RMSE
2013	0.969	0.975	0.681	3.118
2014	0.968	0.850	0.642	2.669
2015	0.966	0.840	0.612	2.996
2016	0.952	0.948	0.754	1.997
2017	0.936	1.103	0.518	2.544

Nota. Tomada de Zhang et al., (2024, p. 817)

Se identificaron las variables PM10, NO2 y la densidad de tráfico como las predictores más influyentes en la concentración de PM2.5, seguidos por variables meteorológicas como la velocidad del viento y la humedad relativa. Adicionalmente, la integración de datos geoespaciales mejoró significativamente las predicciones, destacando la importancia de considerar factores locales y regionales en los modelos.

El estudio A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5) de Li et al (2020), se centra en la predicción de concentraciones de PM2.5 utilizando un modelo híbrido CNN-LSTM, combinando redes neuronales convolucionales (CNN) y redes neuronales de memoria a corto y largo plazo (LSTM).

La investigación tiene como objetivo superar las limitaciones de precisión y eficiencia de métodos anteriores, abordando las complejidades inherentes de las series temporales relacionadas con la calidad del aire, caracterizadas por alta volatilidad y no linealidad, este modelo se aplicó a datos obtenidos en Beijing, donde el aumento de PM2.5 está relacionado con impactos significativos en la salud y el medio ambiente.

El modelo aprovecha las capacidades de las CNN para extraer características relevantes de los datos de entrada y de las LSTM para capturar patrones temporales. El uso de un enfoque multivariado, incluyendo variables como temperatura, presión atmosférica y velocidad del viento, permite una evaluación integral de los factores que influyen en la calidad del aire. Este marco contribuye significativamente al desarrollo de modelos más precisos para abordar problemas de contaminación.

Se emplea el modelo híbrido CNN-LSTM que se centra en la predicción de las concentraciones de PM2.5 para las próximas 24 horas, los datos de los últimos 7 días sirven

como entrada, mientras que la concentración de PM2.5 del día siguiente es el objetivo de predicción, los pasos clave incluyen:

1. Descomposición de Características con CNN: Utiliza dos capas convolucionales 1D para extraer características significativas, seguidas de una capa de MaxPooling y una capa de Flatten para preparar los datos para la LSTM.

2. Predicción Temporal con LSTM: Procesa los datos normalizados en series temporales, optimizando la captura de patrones diarios y semanales.

3. Validación y Optimización: Las métricas utilizadas incluyen MAE y RMSE, destacando la efectividad del modelo en comparación con alternativas univariadas y otros híbridos.

El modelo multivariado CNN-LSTM superó significativamente a las variantes univariadas y modelos independientes, mostrando el menor error absoluto medio (MAE: 13.97) y error cuadrático medio (RMSE: 17.93), lo cual se evidencia en la tabla 7. Además, su tiempo de entrenamiento fue más corto, con 50-60 segundos por época, frente a los 90-100 segundos de modelos alternativos. Este rendimiento se atribuye por los autores a la integración eficiente de datos multivariados y a la capacidad del modelo para procesar patrones complejos y no lineales.

Tabla 7

RMSE de los resultados experimentales

Muestras	**Multivariante CNN-LSTM**	**Multivariante LSTM**	**Univariante CNN-LSTM**	**Univariante LSTM**
1	8.872	14.407	8.119	20.853
2	16.926	31.458	23.716	45.033
3	11.558	9.384	24.462	14.841

4	14.349	23.283	17.790	36.841
5	11.544	14.066	23.733	23.013
6	33.689	22.162	16.477	21.562
7	17.813	19.403	23.638	20.009
8	26.917	19.477	22.064	20.645
9	19.708	19.075	22.064	20.645
10	16.251	10.76	21.289	15.4
Promedio	17.9306	18.0852	19.7692	23.2646

Nota: Tomada de Li et al., (2020, p. 26938)

Los resultados destacan a PM10, la velocidad del viento y la presión atmosférica como predictores esenciales. La inclusión de múltiples características permite un análisis más robusto y una mejora sustancial en la precisión de las predicciones, validando la importancia de enfoques multivariados para modelos de calidad del aire.

Discusión

Si bien existe una considerable cantidad de literatura sobre el Índice de Calidad del Aire (AQI), la investigación en torno al PM 2.5 toma protagonismo puesto que el AQI ya es una medida ampliamente estudiada y utilizada, y el PM 2.5, al ser un componente clave del AQI, ofrece un enfoque más específico y detallado, permitiendo identificar directamente las fuentes y los efectos de las partículas más dañinas en lugar de limitarse al índice general.

El PM 2.5, debido a su diminuto tamaño, es un contaminante atmosférico sumamente peligroso, ya que puede penetrar profundamente en los pulmones y entrar en el torrente sanguíneo, afectando múltiples sistemas del cuerpo humano, esto lo diferencia de partículas más grandes, que suelen ser filtradas por las defensas naturales del sistema respiratorio.

Su capacidad de penetración y los riesgos que representa para la salud pública han sido bien documentados, siendo vinculados a enfermedades respiratorias, cardiovasculares y un aumento en la mortalidad prematura, y debido a su relevancia, sigue siendo un tema de estudio prioritario debido a la necesidad de entender a fondo sus patrones de comportamiento y sus fuentes de emisión.

En el contexto del PM 2.5, la integración de modelos avanzados de machine learning permite aprovechar metodologías desarrolladas inicialmente para la predicción del AQI, adaptándolas a las necesidades específicas de este contaminante PM2.5. Esto incluye el uso de variables meteorológicas, geográficas y de fuentes de emisión, que son cruciales para mejorar la precisión y robustez de las predicciones. Adicionalmente el uso de modelos híbridos, usando Deep learning en conjunto con modelos tradicionales, se convierte en una herramienta aun más poderosa para poder capturar la complejidad inherente a los datos y su posible comportamiento no lineal.

Conclusiones

Los factores que más influyen en el comportamiento del contaminante PM 2.5 incluyen la densidad vehicular (relacionada con emisiones por fuentes móviles), la velocidad del viento, las precipitaciones y la resuspensión de partículas en vías no pavimentadas, y en menor medida, otros contaminantes como el PM10 y el NO2 también tienen un impacto relevante, dado que el PM 2.5 forma parte significativa del rango definido para el PM10, estableciendo una fuerte correlación entre ambos.

Para estimar la concentración del PM 2.5 y determinar eventos de superación de los límites máximos permisibles, es esencial que los modelos predictivos incorporen estas variables clave, además de contar con suficiente cantidad de datos para capturar patrones estacionales, como las variaciones en la calidad del aire durante el verano o invierno; es crucial para garantizar la precisión del modelo.

Luego del análisis de varios modelos, se determina que el modelo híbrido basado en CNN-LSTM combinado con Gradient Boosting es la opción más prometedora para modelar relaciones complejas y no lineales en la predicción de PM 2.5.

Este enfoque combina la capacidad de las redes convolucionales (CNN) para extraer características espaciales relevantes relacionadas con la dispersión de contaminantes en diferentes áreas, con la habilidad de las redes LSTM para capturar patrones temporales, como tendencias y estacionalidad. Gradient Boosting complementa estas capacidades refinando las predicciones mediante el ajuste iterativo de los errores residuales, lo que mejora aún más la precisión y robustez del modelo.

Este modelo es especialmente adecuado para contextos urbanos como Bogotá, siempre que se ajuste con datos locales de la Red de Monitoreo de Calidad del Aire (RMCAB). La

integración de CNN y LSTM permite procesar datos multivariados, mientras que Gradient Boosting optimiza las predicciones al manejar relaciones no lineales entre variables. Para garantizar su precisión y aplicabilidad, el modelo debe incorporar las variables más significativas para el PM2.5, además de parámetros meteorológicos como velocidad del viento, temperatura y humedad.

Esta combinación no solo aborda las complejidades inherentes de las dinámicas del PM 2.5, sino que también permite adaptar el análisis a las características específicas de la región, proporcionando una herramienta más robusta y efectiva para la gestión de la calidad del aire, además que puede ajustarse a la necesidad de generar un output donde se indique si sobrepasa o no el umbral permitido para el contaminante PM2.5 teniendo en cuenta umbrales de clasificación preestablecidos y una transformación de la salida del modelo.

Referencias

Alcaldía de Bogotá. (30 de 12 de 2021). *ALCALDÍA DE BOGOTÁ D.C*. Bogotá: https://bogota.gov.co/mi-ciudad/movilidad/abece-de-las-nuevas-medidas-de-pico-y-placa-en-bogota-para-el-2022

Álvarez, J., Quiñones, E., Fernandez, A., Saba, M., y Caraballo, L. (2023). Environmental and Health Benefits Assessment of Reducing PM2.5 Concentrations in Urban Areas in Developing Countries: Case Study Cartagena de Indias. *Environments, 10*(3), Article 42. https://doi.org/https://doi.org/10.3390/environments10030042

Aram, S., Nketiah, E., y Saalidong, B. (2023). Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. *International Journal of Environmental Science and Technology, 21*, 1345-1360. https://doi.org/https://doi.org/10.1007/s13762-023-05016-2

Cao, G., Zhou, L.-A., Liu, C., y Zhou, J. (2023). The effects of the entries by bike-sharing platforms on urban air quality. *China Economic Quarterly International, 3*, 213-224. https://doi.org/https://doi.org/10.1016/j.ceqi.2023.09.003

Cuellar, J., Isaza, L., y Hernandez, E. (2022). Correlation model between PM2.5 and atmospheric variables in the city of Villavicencio. *Ingeniería Solidaria, 18*(2), 1-16. https://doi.org/https://doi.org/10.16925/2357-6014.2022.02.09

Departamento Administrativo Nacional de Estadistica - DANE. (2024). *Proyecciones de Población*. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

Dhake, R., Rane, M., Bari, R., Patil, A., Patil, P., y Naikwade, M. (2024). Air Quality Monitoring and Predictions. *Grenze International Journal of Engineering and Technology, 10*(1), 1986-1991. https://doi.org/01.GIJET.10.1.563_1

Gangwar, A., Singh, S., Mishra, R., y Prakash, S. (2023). The State-of-the-Art in Air Pollution Monitoring and Forecasting Systems Using IoT, Big Data, and Machine Learning. *Wireless Personal Communications, 130*, 1699–1729. https://doi.org/https://doi.org/10.1007/s11277-023-10351-1

Granada-Aguirre, L. F., Pérez-Vergara, I., y Valencia-Rodríguez, M. (2014). Sistema para el manejo de la calidad del aire en la ciudad de Cali -- Colombia. *Ingeniería Industrial, 35*(1), 13–24.

Gregorio, J., Gouveia, C., y Caridabe, P. (2022). Modeling PM2.5 and PM10 Using a Robust Simplified Linear Regression Machine Learning Algorithm. *Atmosphere, 13*, 1334. https://doi.org/https://doi.org/10.3390/atmos13081334

Grisales, H., Montealegre, N., Piñeros, J., Ospina, D., y Nieto, E. (2021). Relación de PM2,5 y Enfermedad Respiratoria Aguda en un territorio de Colombia: Modelo Aditivos Generalizados. *Universidad y Salud, 24*(1), 45-54. https://doi.org/https://doi.org/10.22267/rus.222401.256

Gutierrez, I., Calderón, D., Gutierrez, D., y Vergara, E. A. (2020). Correlación entre diferentes contaminantes atmosféricos de la Ciudad de México y el Área Metropolitana. *CIENCIA Ergo-Sum*, *27*(3), 364-375. https://doi.org/https://doi.org/10.30878/ces.v27n3a5

Janarthanan, P., Partheeban, P., Somasundaram, K., y Elamparithi, N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society, 67*, 102720. https://doi.org/https://doi.org/10.1016/j.scs.2021.102720

Li, T., Hua, M. H., & Wu, X. (2020). A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). *IEEE, 8*, 26820–26829. https://doi.org/10.1109/ACCESS.2020.2971348

Liang, Y.-C., Maimury, Y., Hsiang-Ling Chen, A., y Cuevas Juarez, J. R. (2020). Machine Learning-Based Prediction of Air Quality. *Applied Sciences, 10*(24), 9151. https://doi.org/https://doi.org/10.3390/app10249151

Liu, H., & Chen, C. (2020). Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China. *Journal of Cleaner Production, 265*, 121777. https://doi.org/https://doi.org/10.1016/j.jclepro.2020.121777

Liu, H., Li, Q., Yu, D., y Gu, Y. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Applied Sciences, 9*(19), 4096. https://doi.org/https://doi.org/10.3390/app9194069

López, R., Castillo, I., Collazo, A., y Martínez, R. (2024). Relación entre factores meteorológicos y contaminantes criterio en La Habana. *Revista Cubana de Metereología, 29*(4), 196-205. https://doaj.org/article/c15e71d6f77c4b35ae2492671fbcd725

Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., y Kedam, G. (2019). A Machine Learning Model for Air Quality Prediction for Smart Cities. *2019 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 452–457). IEEE. https://doi.org/https://doi.org/10.1109/WiSPNET45539.2019.9032734

Mathew, A., Gokul, P., Raja, P., Aranub, K., Ghassan, H., Almohamad, H., y Abdullah, A. (2023). Air quality analysis and PM2.5 modelling using machine learning techniques: A

study of Hyderabad city in India. *Cogent Engineering*, *10*(1). https://doi.org/https://doi.org/10.1080/23311916.2023.2243743

Medhi, S., Boruah, R., Pratim, S., y Kumar, H. (2024). Homogenous Ensemble Learning for Air Quality Index Prediction. *Grenze International Journal of Engineering and Technology,*, *10*(1), 424–429. https://research-ebsco-com.bibliotecavirtual.unad.edu.co/linkprocessor/plink?id=b7b23b96-4b89-3175-86ff-0da2dc2337df

Medina, E. C. (2019). La contaminación del aire, un problema de todos. *Revista de la Facultad de Medicina*, *67*(2), 189–191. https://doi.org/https://doi.org/10.15446/revfacmed.v67n2.82160

Méndez, M., Merayo, M., y Nuñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, *56*(9), 10031–10066. https://doi.org/https://doi.org/10.1007/s10462-023-10424-4

Ministerio de Ambiente y Desarrollo Sostenible. (1 de Noviembre de 2017). *Resolución 2254 de 2017.* Minambiente: https://www.minambiente.gov.co/documento-normativa/resolucion-2254-de-2017/

Ministerio de Salud y Protección Social. (2024). *Minsalud.* Minsalud: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/INTOR/contexto-migratorio-bogota-2024.pdf

Nhat-Duc, H., & Van-Duc, T. (2023). Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification. *Automation in Construction*, *148*, 104767. https://doi.org/https://doi.org/10.1016/j.autcon.2023.104767

Observatorio de movilidad de Bogotá. (s.f.). *Registro Distrital Automotor (RDA)*. Retrieved 2024 de noviembre de 2024, from Observatorio de Movilidad: https://observatorio.movilidadbogota.gov.co/indicadores/registro-distrital-automotor

Organización Mundial de la Salud. (16 de Septiembre de 2023). *OMS*. OMS: https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases

Pan, R., Wang, X., Yi, W., Wei, Q., Gao, J., Xu, Z., . . . Yanhu, J. (2020). Interactions between climate factors and air quality index for improved childhood asthma self-management. *Science of the Total Environment, 723*, 137804. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.137804

Preetham Vignesh, P., H. Jiang, J., y Kishore, P. (2023). Predicting PM2.5 Concentrations Across USA Using Machine. *Earth and Space Science, 10*(10), e2023EA002911. https://doi.org/https://doi.org/10.1029/2023EA002911

Ramirez, J., Pachon, J., Casas, O., y Gonzalez, S. (2019). A NEW DATABASE OF ON-ROAD VEHICLE EMISSION FACTORS FOR COLOMBIA: A CASE STUDY OF BOGOTA. *CT&F - Ciencia, Tecnología y Futuro, 9*(1), 73-82. https://doi.org/https://doi.org/10.29047/01225383.154

Rodriguez, L. A., Belalcazar, L. C., Castillo, M. P., Sanchez, E. R., Herrera, V., y Agudelo, D. M. (2022). Avoidable mortality due to long-term. *Environmental Health, 21*(1), 137. https://doi.org/https://doi.org/10.1186/s12940022009478

Sanchez, J., Gaona, A., y Dallos, D. (2022). Modelo de simulación para evaluación de políticas ambientales mediante la caracterización de la contaminación del aire en la ciudad de Bogotá usando dinámica de sistemas. *Ingeniería y competitividad: revista científica y tecnológica, 24*(2), 1-19. https://doi.org/https://doi.org/10.25100/iyc.v24i2.11573

Secretaría Distrital de Ambiente. (2024). *Red de Monitoreo de Calidad del Aire de Bogotá - RMCAB.* https://www.ambientebogota.gov.co/red-de-monitoreo-de-calidad-del-aire-de-bogota-rmcab

Secretaria Distrital de Ambiente -SDA-. (2022). *Inventario de Emisiones de Bogotá Contaminantes Criterio y Carbono negro.*

Song, W., Kwan, M.-P., y Huagn, J. (2024). Assessment of air pollution and air quality perception mismatch using mobility-based real-time exposure. *PLOS ONE, 19*(2), e0294605. https://doi.org/https://doi.org/10.1371/journal.pone.0294605

Vargas, Y., Toro, A., Rojas, N., y Fajardo, O. (2023). School Children's exposure to PM 2.5 in a high pollution area of Bogotá, Colombia. *Ingeniería e Investigación, 43*(2), e96125. https://doi.org/https://doi.org/10.15446/ing.investig.96125

Vazquez, D. F. (2021). *Inventario de emisiones de fuentes móviles bajo metodología Bottom - Up para Bogotá, Colombia y análisis de escenarios de oferta para su reducción.* Bogotá, Colombia: Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/81579

Velasco, E., & Bernabe, R. (2004). *Biogenic Emissions: Emissions of volatile organic compounds no methane from vegetation and nitric oxide from soil (in Spanish).* México: Instituto Nacional de Ecología.

Vieru, M.-C., y Cărbureanu, M. (2024). Machine Learning Methods Applied in Air Quality Prediction. *Romanian Journal of Petroleum & Gas Technology, 5*(1), 5-18. https://doi.org/https://doi.org/10.51865/jpgt.2024.01.01

Wright, C., Benyon, M., Mahlangen, N., Kapwata, T., Laban, T., y Garland, R. (2023). Data gaps will leave scientists 'in the dark': How load shedding is obscuring our

understanding. *South African Journal of Science, 119*(9/10), 1-5. https://doi.org/https://doi.org/10.17159/sajs.2023/16009

Wu, Q., & Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Science of the Total Environment, 683*, 808-821. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.05.288

Xu, Y., You, T., Wen, Y., Ning, J., Xiao, Y., y Shen, H. (2023). Air Quality Research Based on B-Spline Functional Linear Model: A Case Study of Fujian Province, China. *Applied Sciences, 13*(20), Artículo 11206. https://doi.org/https://doi.org/10.3390/app132011206

Zafra, C., Alvaro, G., y Hernandez, Y. (2019). Correlation between vehicular traffic and heavy metal concentrations in road sediments of Bogotá, Colombia. *Revista de la facultad de medicina, 67*, 193-199. https://doi.org/http://dx.doi.org/10.15446/revfacmed.v67n2.68269

Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China. *Science of the Total Environment, 635*, 644-658. https://doi.org/Science of the Total Environment

Zhang, K., Lin, J., Li, Y., Sun, Y., Tong, W., Li, F., . . . Craft, E. (2024). Unmasking the sky: high-resolution PM2.5 prediction in Texas. *f Exposure Science & Environmental Epidemiology, 34*, 814–820. https://doi.org/https://doi.org/10.1038/s41370-024-00659-w

Zhang, S. (2022). Challenges in KNN Classification. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 34*, 4663-4675. https://doi.org/10.1109/TKDE.2021.3049250

Zhang, Z., Zeng, Y., y Yan, K. (2021). A hybrid deep learning technology for PM2.5 air quality forecasting. *Environmental Science and Pollution Research, 28*, 39409-39422. https://doi.org/https://doi.org/10.1007/s11356-021-12657-8

Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., y Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. *Environmental Pollution, 231*, 1232-1244. https://doi.org/http://dx.doi.org/10.1016/j.envpol.2017.08.069

Zhu, S., Qiu, X., Yin, Y., Fang, M., Liu, X., Zhao, X., y Shi, Y. (2019). Two-step-hybrid model based on data preprocessing and intelligent optimization algorithms (CS and GWO) for NO2 and SO2 forecasting. *Atmospheric Pollution Research, 10*, 1326-1335. https://doi.org/https://doi.org/10.1016/j.apr.2019.03.004.