UNIVERSIDAD EAFIT

Análisis Comparativo de Modelos Predictivos para la Estimación de $PM_{2.5}$: Un Enfoque Basado en Aprendizaje Automático y Predicción Conformal

Comparative Analysis of Predictive Models for $PM_{2.5}$ Estimation: A Machine Learning and Conformal Prediction Approach

Matías Camelo Valera

Tesis de grado

Juan David Martínez Vargas

Lina Maria Sepúlveda Cano

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2024

CONTENIDO

LISTA DE FIGURAS

RESUMEN

La contaminación por material particulado fino ($PM_{2.5}$) representa un desafío ambiental y de salud pública, requiriendo modelos predictivos precisos para su monitoreo y control. En este trabajo, se comparan diferentes enfoques de aprendizaje automático, incluyendo Regresión Lineal, Random Forest y XGBoost, con y sin la inclusión de variables de movilidad, para estimar los niveles de $PM_{2.5}$. Además, se implementa la predicción conformal inductiva para cuantificar la incertidumbre en las estimaciones y proporcionar intervalos de confianza con $\alpha = 0.05$.

Los resultados evidencian que XGBoost, pese a experimentar un deterioro en la fase de entrenamiento al incluir variables de movilidad, logra el mejor desempeño en validación con un menor error absoluto medio y mayor coeficiente de determinación. La predicción conformal permitió establecer intervalos de confianza con una cobertura del 89.26%, cercana al 95% esperado, lo que garantiza la fiabilidad del modelo en distintos escenarios espaciales y temporales.

En conclusión, el uso de modelos de aprendizaje automático en combinación con técnicas avanzadas de validación y calibración, como la predicción conformal, permite mejorar la precisión y confiabilidad en la estimación de $PM_{2.5}$. Sin embargo, la calidad de las variables de entrada, especialmente las de movilidad, sigue representando un desafío, lo que sugiere la necesidad de incorporar información meteorológica y mejorar la resolución de los datos. Estos hallazgos contribuyen al desarrollo de herramientas predictivas más fiables para la gestión ambiental y la toma de decisiones en políticas de calidad del aire.

Palabras clave: calidad del aire; $\mathrm{PM}_{2.5}$; predictor; aprendizaje automático; serie de tiempo; predicción conformal

ABSTRACT

Fine particulate matter ($PM_{2.5}$pollution poses a significant environmental and public health challenge, requiring accurate predictive models for its monitoring and control. This study compares different machine learning approaches, including Linear Regression, Random Forest, and XGBoost, with and without the inclusion of mobility variables, to estimate $PM_{2.5}$ levels. Additionally, inductive conformal prediction is implemented to quantify uncertainty in the estimates and provide confidence intervals with $\alpha = 0.05$.

The results show that while XGBoost experiences performance deterioration during training when mobility variables are included, it achieves the best validation performance with the lowest mean absolute error and the highest coefficient of determination. Conformal prediction enabled the establishment of confidence intervals with 89.26% coverage, close to the expected 95%, ensuring model reliability across different spatial and temporal scenarios.

In conclusion, the use of machine learning models combined with advanced validation and calibration techniques, such as conformal prediction, enhances the accuracy and reliability of $PM_{2.5}$ estimation. However, the quality of input variables, particularly mobility-related data, remains a challenge, highlighting the need to incorporate meteorological information and improve data resolution. These findings contribute to the development of more reliable predictive tools for environmental management and air quality policy decision-making.

Keywords: Air quality; $\mathrm{PM}_{2.5}$; Predictive model; Machine Learning; Time series; Conformal prediction

INTRODUCCIÓN

Según la Organización Panamericana de la Salud (2024), para el año 2019, se registraron 55.4 millones de muertes a nivel global, lo que representó un aumento del 8% en la mortalidad en el período 2000-2019. A nivel regional, en las Américas, este incremento fue más pronunciado, alcanzando el 31 % en el mismo período. Adicionalmente, se menciona en el informe, que uno de los principales desafíos en materia de salud pública son las enfermedades no transmisibles y de causa externa, con un enfoque particular en las enfermedades cardiovasculares, las infecciones respiratorias y los cánceres de las vías respiratorias. La Oficina de Evaluación de Peligros para la Salud Ambiental de California (2025), estima que las principales causas de estas enfermedades incluyen la contaminación del aire, principalmente el material particulado y la exposición al humo del tabaco, lo que afecta la calidad de vida de las personas y por consiguiente genera un impacto significativo en el desarrollo social y económico de los países.

La predicción de la calidad del aire, en particular de la concentración de $PM_{2.5}$, ha cobrado gran relevancia a nivel global debido a su impacto en la salud pública y el medio ambiente (Pasquier. et al., 2017). El estudio del material particulado y otros contaminantes ha sido un área de investigación creciente en ciudades principales como Buenos Aires, Bogotá, Santiago de Chile y Asunción (Salas, 2022; Westerlund, et al., 2014; Carlés, et al., 2023), así como en ciudades de mediano tamaño como Manizales (Gónzalez, et al., 2018).

En el Valle de Aburrá, el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá conocido comoSIATA, a pesar de contar con existencia de redes de monitoreo en algunas zonas, tiene una limitada capacidad de predicción y alerta temprana, existe la necesidad de fortalecer la infraestructura con modelos de pronóstico ajustados a las condiciones específicas del territorio que permita alertar a las autoridades cualquier eventualidad que pueda ser nociva para la salud pública (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá, 2024).

En este contexto, en el presente estudio se desarrolló un modelo de predicción de material particulado ($PM_{2.5}$) basado en datos históricos locales para las comunas 4, 10, 14 y 16 de la ciudad de Medellín, además se evaluó el uso de variables de movilidad para mejorar la precisión de las predicciones. Este modelo incorpora intervalos de confianza mediante predicción conformal lo que permite cuantificar la incertidumbre asociada a cada predicción. De esta manera, se espera mejorar la capacidad de respuesta ante contingencias ambientales en el Valle de Aburrá y contribuir a una gestión más efectiva de la calidad del aire, ayudando a la generación de nuevas propuestas de mitigación de la contaminación en la ciudad (Casali. et al, 2022).

PLANTEAMIENTO DEL PROBLEMA

La predicción del material particulado se ha abordado desde diversas perspectivas en los últimos años. Los enfoques más comunes han sido modelos estadísticos, tanto univariantes como multivariantes, los cuales analizan series de tiempo bajo supuestos específicos sobre la distribución de los datos y las relaciones entre las variables (Kokkinos et al., 2021). Sin embargo, estos modelos presentan ciertas limitaciones: restringen la cantidad de variables que pueden incorporarse y asumen que el comportamiento del fenómeno a modelar es predominantemente lineal. Como consecuencia, pueden no ser capturadas de manera efectiva relaciones no lineales entre las variables (Salas, 2022).

Por otro lado, se ha hecho uso de modelos de aprendizaje automático principalmente métodos basados en árboles de decisión, regresión lineal y diversas configuraciones de redes neuronales. Sin embargo, estos enfoques presentan sus propias limitaciones, entre ellas, la reducción del nivel de interpretabilidad a medida que aumenta la complejidad del fenómeno estudiado, la dificultad para cuantificar la incertidumbre asociada a sus predicciones y una marcada tendencia al sobreajuste (Das, et al. 2022; Molnar, 2023). Además, que las redes neuronales suelen requerir una gran cantidad de datos para generar predicciones precisas, aunque, el impacto de esta necesidad en la precisión de las predicciones varía según el modelo específico y la calidad de los datos disponibles (Cao, 2024).

En este contexto, surge la necesidad de desarrollar un modelo predictivo que permita su interpretación, que, además pueda cuantificar la incertidumbre de sus predicciones y como resultado permita anticipar los niveles futuros de $\mathrm{PM}_{2.5}$ con un enfoque especial en aquellas concentraciones que representen un potencial riesgo a la salud de las personas en la ciudad de Medellín y haciendo uso exclusivo de la información disponible limitada al periodo 2021-2023.

Definiendo así, la pregunta central de esta investigación como: ¿Es posible entrenar un modelo que prediga con intervalos de confianza los niveles de $PM_{2.5}$ utilizando datos históricos de calidad del aire específicamente para las comunas 4, 10, 14 y 16 de la ciudad de Medellín? Además, se evalúa el impacto que tiene el uso de variables relacionadas con el tráfico vehicular en la precisión del modelo.

JUSTIFICACIÓN

En el Valle de Aburrá, la baja calidad del aire se presenta como una problemática recurrente que ha motivado la implementación de diversas medidas para mitigar su impacto. Entre estas acciones se destacan iniciativas como el "pico y placa ambiental" y la creación de espacios verdes en la ciudad (de Andrade, 2023) (Rodas, 2024). Sin embargo, es importante señalar que en los últimos años se ha observado un crecimiento significativo del parque automotor en Antioquia; específicamente, entre 2016 y 2018, la cantidad de vehículos en circulación en Medellín aumentó en un 20% (González, M et al. 2023). Este incremento ha contribuido a una disminución en la calidad del aire en comparación con años anteriores, exacerbado por condiciones meteorológicas específicas que caracterizan la región.

La dispersión de contaminantes, especialmente del $\mathrm{PM}_{2.5}$, se ve obstaculizada por factores climáticos, siendo particularmente notoria esta dificultad durante los meses de transición de temporada húmeda a seca, específicamente entre febrero y marzo y de septiembre a octubre (Ramírez, et al, 2023). Durante estos períodos, se presenta un fenómeno de estancamiento del aire en el Valle de Aburrá, lo que agrava aún más la situación. Bajo esta perspectiva, la aplicación de modelos de predicción se vuelve esencial para abordar la problemática de la contaminación del aire. Los modelos predictivos tradicionales suelen ofrecer estimaciones basadas en patrones históricos sin considerar adecuadamente la incertidumbre asociada a sus predicciones (Molnar, 2023). En contraste, la técnica de predicción conformal se presenta como una solución innovadora al proporcionar intervalos de confianza que reflejan la incertidumbre inherente a los datos, permite su aplicación a una amplia variedad de modelos predictivos, sin requerir suposiciones específicas sobre la distribución de los datos; Además de permitir especificar un nivel de confianza para que una región de predicción con mayor probabilidad contenga el resultado verdadero, lo que facilita la toma de decisiones informadas y mejora la planificación ante escenarios futuros (McMaho, et al., 2023).

OBJETIVOS

GENERAL

Desarrollar un modelo de predicción conformal para representar el comportamiento espacio-temporal de la concentración de material particulado menor a 2.5 micras ($PM_{2.5}$), utilizando registros históricos de calidad del aire en las comunas 4, 10, 14 y 16 de Medellín.

ESPECÍFICOS

- Construir una base de datos que integre la información disponible del Sistema de Alerta Temprana del Valle de Aburrá y de movilidad de Medellín, garantizando su calidad y coherencia para el modelado predictivo.
- Caracterizar el comportamiento de los datos, mediante el análisis exploratorio, identificando patrones y características relevantes para seleccionar los algoritmos predictivos más adecuados.
- Diseñar modelos predictivos, evaluando su desempeño mediante la comparación de métricas como MAE, $RSME$ y R^2, asegurando su capacidad de generalización.
- Optimizar el modelo con mejor desempeño, aplicando técnicas de ajuste de hiperparámetros, predicción conformal y validación cruzada para cuantificar la incertidumbre de las predicciones.

ESTADO DEL ARTE

Existen dos enfoques principales para la predicción de contaminantes: los modelos estadísticos y los modelos de aprendizaje automático, Ambos pueden ser univariantes o multivariantes, siendo estos último utilizados con el objetivo de mejorar el desempeño del modelo al incluir una mayor cantidad de información (Mghouchi et al 2024). Cabe destacar que, en la bibliografía revisada para este apartado no se menciona el uso de predicción conformal ni la generación de intervalos de confianza para la predicción, No obstante, Westerlund et al. (2014) proponen una forma alternativa para cuantificar la incertidumbre de los modelos, lo que resalta la importancia de incluir este tipo de análisis.

Además, en la actualidad, existe un extenso cuerpo de investigación sobre la problemática de la calidad del aire a nivel global. A continuación, se presentan los modelos desarrollados en diversas investigaciones para la predicción de contaminantes atmosféricos, con especial énfasis en aquellos que guardan mayor similitud con el enfoque de este estudio y cuyos resultados son relevantes para el diseño de la metodología aquí planteada.

MÉTODOS ESTADÍSTICOS TRADICIONALES

Univariados

Como es el estudio publicado por Gocheva-Ilieva, et al (2014) donde se utilizaron modelos de la familia Box-Jenkins, mejor conocidos como ARIMA y SARIMA, para la predicción a corto plazo de NO, NO_2, NO_x, SO_2, O_3 y PM_{10} en Blagoevgrad, Bulgaria. Esta ciudad se caracteriza por su alto nivel de vegetación y la ausencia de tráfico vehicular significativo, siendo los hogares la principal fuente de contaminación debido a la falta de un sistema de calefacción centralizada. Los autores utilizaron datos históricos recopilados entre el 1 de septiembre de 2011 y el 31 de agosto de 2012, con mediciones horarias. Para el análisis, emplearon Análisis de Componentes Principales (PCA) y análisis de correlación, lo que permitió identificar la posibilidad de agrupar los contaminantes en tres categorías principales: Primer grupo: NO_x, NO_2, NO y PM_{10}; Segundo grupo: O_3; Tercer grupo: SO_2. Esta clasificación se basó en la alta correlación entre los contaminantes y en la capacidad de explicar más del 90% de la varianza de los datos originales, determinada por provenir de las mismas fuentes de contaminación. No obstante, a pesar de realizar este análisis factorial, se construyó un modelo Box-Jenkins para cada serie temporal individual.

Adicionalmente, dentro de la metodología utilizada se asumió que la concentración final de contaminantes en la atmósfera es el resultado de la interacción compleja de factores como meteorología, química, transporte y difusión. En consecuencia, la información combinada de estos efectos sobre la concentración de contaminantes está contenida de manera estocástica en las series temporales correspondientes.

Este enfoque han de mencionarse que en las series temporales se asumen la no presencia de datos nulos, se evidencia que estas metodologías cuentan con una gran capacidad para capturar la estacionariedad en las series e incorporar tendencias y patrones recurrentes, permitiendo realizar pronósticos con intervalos de hasta 72 horas después de la última observación registrada. Lo que, en comparación con modelos de aprendizaje automático, como modelos de ensamble y redes neuronales, se considera un horizonte de predicción corto.

Multivariados

En este apartado la literatura muestra una fuerte inclinación hacia metodologías que involucran Técnicas de Suavización Exponencial (ETS) (Roy et al. 2018; Bose et al. 2020; Ventura et al. 2019). Su popularidad radica en su relativa simplicidad, flexibilidad y buen desempeño en tareas de predicción (Smyl et al. 2025). Estas técnicas fueron desarrolladas como una heurística en la que la predicción t es dada por la predicción inmediatamente anterior $t-1$ ajustada mediante un factor de corrección proporcional al error de la predicción previa, como se observa en la configuración utilizada por Roy et al. (2018).

En dicho estudio, además de abordar la predicción de contaminantes, se propone el diseño de una infraestructura para el procesamiento en tiempo real de los niveles de contaminación ambiental. Esto pone de manifiesto la necesidad de contar con sistemas de alerta temprana en caso de altos niveles de contaminación para la protección de la población. Asimismo, se menciona que la Técnica de Suavización Exponencial Simple es adecuada para el pronóstico a corto plazo (máximo un mes) y se define matemáticamente como:

$$F_t = \alpha C_t + (1+\alpha)F_{t-1}$$

Donde F_t representa el valor de suavizado, C_t es la observación actual y α es la constante de suavizado, que toma valores entre 0 y 1 para $t \geq 3$.

Además, esta técnica puede incorporar componentes de estacionalidad y tendencias locales, dando lugar a las variantes conocidas como ETS doble y ETS triple.

Para el caso de estudio de ETS doble Bose et al. (2020) plantea el uso de está técnica que integra tendencias de las series de tiempo, este estudio es la continuación y mejora de la metodología previa con su enfoque en el pronóstico en tiempo real, para esto se define la siguiente formulación matemática:

$$F_t = \alpha C_t + (1+\alpha)(F_{t-1} + K_{t-1})$$

$$K_t = \beta(F_t - F_{t-1}) + (1-\beta)K_{t-1}$$

En esta formulación se introducen los términos K_t que representa la tendencia estimada con una constante γ asociada a α donde α y β toma valores entre 0 y 1.

De esta forma se mejora el desempeño de los modelos obteniendo mejores predicciones sobre la mayoría de contaminantes ambientales.

MÉTODOS DE APRENDIZAJE AUTOMÁTICO Y PROFUNDO

Univariados

El principal beneficio de estos modelos es su alta capacidad predictiva, ya que permiten emplear una gran cantidad de variables explicativas. En este enfoque, los modelos univariantes son poco utilizados, especialmente en el caso de series de tiempo, donde los rezagos temporales de la variable objetivo se incorporan como características explicativas. Como resultado, un modelo basado en una única variable no sería sustentable, pues carecería de la capacidad predictiva necesaria (Hyndman, et al. 2018).

Multivariados

Estas se caracterizan por permitir integrar diversos tipos de datos en el proceso de predicción. Por ejemplo, Kokkinos. et al., (2021) estudió la predicción de $PM_{10}, PM_{2.5}$ y NO_2 en el centro urbano de Cambridge, Reino Unido, utilizando métodos como ANFIS (Sistema de Inferencia Difusa Neuroadaptativo), LSTM (Redes Neuronales de Memoria a Largo y Corto Plazo) y ELM (Máquinas de Aprendizaje Extremo). Estos métodos fueron aplicados en conjunto con un preprocesamiento en el cual se incluye la imputación de datos nulos por regresión secuencial múltiple, técnica que tiene la capacidad de generar múltiples predicciones por cada valor faltante basado en las características y la relación entre diferentes observaciones, de forma que permite tener en cuenta la incertidumbre en las imputaciones logrando errores estándar precisos (Azur. et al., 2011).

Además, se aplicaron metodologías de eliminación de tendencias y estacionaridad de las series de tiempo $PM_{10}, PM_{2.5}$ y NO_2, según los autores evitando que ciertas series específicas influyan desproporcionadamente en el proceso de predicción y segundo dado que algunas estaciones de monitoreo muestran una tendencia en la distribución, ya que constantemente registran valores más altos de PM en ciertos intervalos de tiempo del día. Para la selección de variables relevantes, se utilizó PCA, identificando como principales características para el entrenamiento: temperatura, precipitación, promedio de velocidad del viento, humedad y Volumen de Tráfico. Se concluye en el estudio que los tres métodos utilizados logran métricas de desempeño aceptables, para el caso de $PM_{2.5}$ se obtienen valores de MAE entre 0.07 y 0.08, $MAPE$ entre 18% y 21%, $RMSE$ entre 0.07 y 0.09 y finalmente R^2 no menores a 0.89. Los autores identifican que el modelo con mejor rendimiento corresponde a una configuración basada en ANFIS. Sin embargo, a lo largo del estudio se evidencia la alta complejidad computacional inherente a los métodos basados en redes neuronales. Además, estos modelos presentan una interpretabilidad limitada, ya que sus predicciones resultan de múltiples capas de multiplicación de pesos y transformaciones no lineales, lo que dificulta la comprensión del proceso subyacente (Molnar, 2021).

Por su parte, Westerlund, J. et al (2014) proponen un enfoque diferente en su estudio sobre la predicción de contaminantes atmosféricos en la ciudad de Bogotá, específicamente para PM_{10}, CO, NO_2, NO_x, SO_2 y O_3 Además de desarrollar modelos multivariantes, emplean la técnica de combinación predictiva. Este parte de la premisa de que, al contar con múltiples modelos en competencia, cada uno con sus propias fortalezas y debilidades, en lugar de seleccionar un único modelo con el mejor desempeño, se combinan las predicciones individuales en una única estimación. De este modo, la predicción final integra la incertidumbre asociada a cada modelo, logrando un desempeño comparable al de las mejores predicciones individuales, pero con mayor estabilidad y confiabilidad.

En este estudio se hace uso de redes neuronales usando como función de transferencia la función logística definida como:

$$G(x) = \frac{1}{(1 + e^{-x})}$$

Asimismo, se utiliza como función objetivo la ecuación de mínimos cuadrados no lineales (NLS), comparando el desempeño de este método con la regresión lineal.

Al trabajar con redes neuronales bajo esta configuración, se destacan dos aspectos clave:

1. Sobreajuste del modelo: A medida que se incrementa el número de capas ocultas, existe una mayor tendencia al sobreajuste, lo que puede afectar la capacidad de generalización del modelo.
2. Múltiples mínimos locales: La función objetivo NLS presenta varios mínimos locales por naturaleza. Como resultado, la convergencia del modelo no garantiza alcanzar un mínimo global, lo que dificulta la selección de un modelo óptimo en relación con los demás entrenados.

Adicionalmente, se lleva a cabo la imputación de datos nulos en el conjunto de datos utilizando dos enfoques distintos:

1. Método del Efecto Dependiente del Lugar (SDEM): Esta técnica considera tanto la correlación espacial como temporal de los datos y ha demostrado superar los métodos de imputación simple y múltiple (Westerlund, J. et al, 2014). Para su aplicación, se emplean datos de estaciones de monitoreo adicionales, lo que permite mejorar la precisión de la imputación.
2. Método de Covariabilidad Temporal Cicloestacionaria: A diferencia del SDEM, este enfoque solo tiene en cuenta la correlación temporal de los datos, lo que lo hace menos exigente en términos de información requerida. Sin embargo, al no considerar la dimensión espacial, su desempeño puede ser inferior en ciertos escenarios.

Se concluye que, en todos los casos, los modelos de predicción combinada superan en desempeño a los modelos individuales basados en redes neuronales y regresión lineal. Además, se observa que el efecto de las variables meteorológicas es

predominantemente lineal, ya que, a pesar de experimentar con términos cuadráticos, los resultados indican que su inclusión no mejora significativamente el desempeño del modelo**.**

Finalmente un modelo similar al planteado en este trabajo es realizado por Ramírez, et al. (2023) para la predicción de $PM_{2.5}$ en el Valle de Aburrá en este se prueban modelos como la Regresión Lineal Múltiple (MLR), Random Forest (RF), Máquinas de Soporte Vectorial (SVM), gradient-boosting (GB) y K-Vecinos Más Cercanos (KNN), con datos disponibles para entrenamiento pertenecientes al periodo 2017-2020 (previo al inicio de pandemia) y un conjunto de prueba comprendido entre los meses de febrero hasta abril del 2022, su enfoque difiere al aquí empleado integrando características explicativas de una variedad de condiciones metereológicas como precipitación, temperatura, humedad, enter otros; Acompañados de rezagos de la serie temporal hasta 24 horas previas, además de un índice de afectación por incendios. Para la predicción se define metodológicamente que la última predicción $t+1$ se agrega como insumo para la predicción $t+2$ y así repetidamente hasta $t+96$, los resultados muestran valores obtenidos de $RMSE\ (\mu g/m^3)$ entre 12.6 y 9.6, se destaca que los modelos con mejor rendimiento son RF y SVM demostrando que este tipo de algoritmos son capaces de modelar el fenómeno de estudio con un desempeño notable.

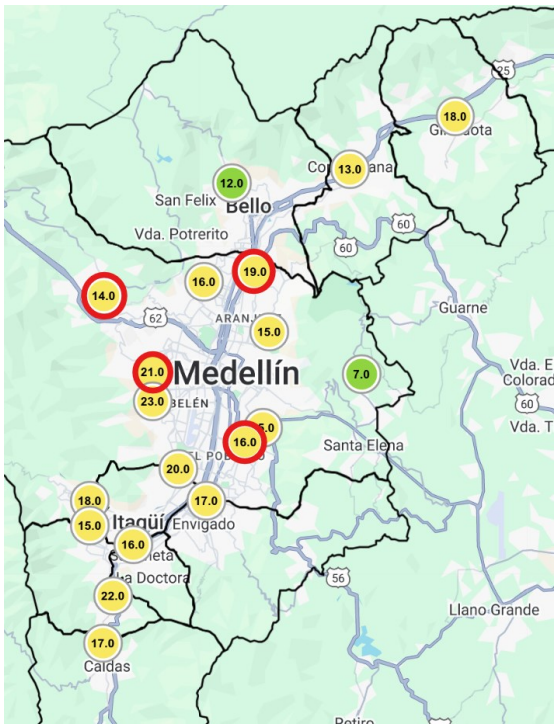
MARCO TEÓRICO

CONTEXTO CLIMATOLÓGICO Y DE MONITOREO DE CALIDAD DEL AIRE EN MEDELLÍN, COLOMBIA

Medellín tiene un clima templado-seco con las siguientes características (IDEAM, 2021): Temperatura promedio: $21.5\ ºC$; Promedio de lluvia total anual: $1685\ mm$.

Con una patrón estacional definido para la precipitación donde se evidencian: (1) Dos temporadas secas: enero y febrero, y de finales de junio a principios de septiembre. Y (2) Dos temporadas lluviosas: finales de marzo a principios de junio, y finales de septiembre a principios de diciembre.

Figura 1. Estaciones de monitoreo de calidad de aire del Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá



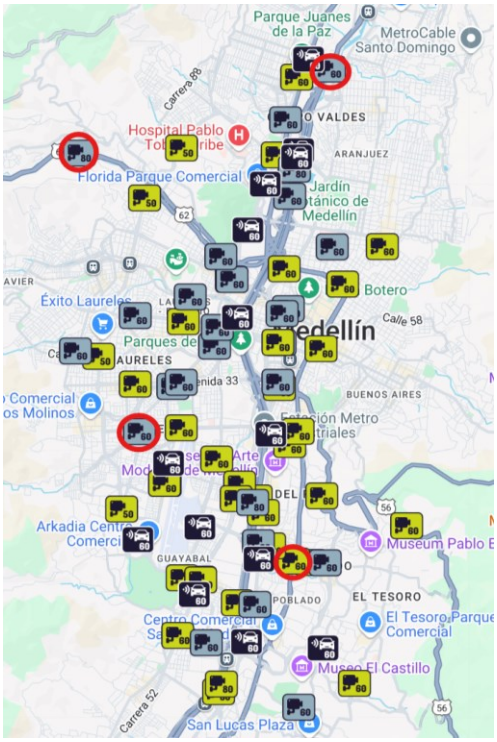
Fuente:
Figura tomada de Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá, 2024.

Por su parte la estaciones de monitoreo presentes en la ciudad cuentan con una amplia cobertura espacio-temporal, en la Figura # pueden verse los puntos amarillos y verdes donde se encuentran todas las estaciones de monitoreo disponibles,

resaltando con un círculo rojo aquellas seleccionadas para este estudio, estas son seleccionadas por presentar una mayor calidad y completitud en la información.

Por otro lado, las cámaras de movilidad de las cuales se obtienen los datos referentes al tráfico vehicular se encuentran ubicadas en los puntos mostrados en la Figura 2 se resaltan nuevamente con un círculo rojo aquellas con mayor proximidad a las estaciones de calidad de aire dado que se estima tienen mayor influencia sobre la contaminación en cada respectiva zona.

Figura 2 Cámaras de tráfico Vehicular SIMM



Fuente:
Figura tomada de Alcaldía de Medellín, 2025.

MATERIAL PARTICULADO $PM_{2.5}$

El Material Particulado 2.5 ($PM_{2.5}$), con un valor de Error Absoluto Medio (MAE) menor a 10 micras, hace referencia a partículas que tienen un diámetro igual o inferior a 2.5 micras. Estas partículas son tan diminutas que no son perceptibles a simple vista, entre todas las mediciones de contaminación del aire, esta representa la mayor preocupación para la salud, su tamaño permite que permanezcan suspendidas en el aire durante largos periodos, pudiendo ser inhaladas o penetrar profundamente en el sistema circulatorio humano o de estos especies animales (IQAir, 2022).

Según (IDEAM, 2021) una de las regiones a nivel nacional que muestra niveles significativos de contaminación atmosférica es el Área Metropolitana del Valle de Aburrá. Esto, adicional a los costos ambientales relacionados con la contaminación atmosférica en Colombia, resalta la importancia de continuar implementando estrategias para controlar, evaluar y monitorear este índice de contaminación.

La Tabla 1 presenta los niveles de prevención de la concentración de $\mathrm{PM}_{2.5}$, los cuales proporcionan una guía para evaluar el impacto de la contaminación en la salud pública y tomar medidas preventivas adecuadas.

Tabla 1 Niveles de prevención de la concentración de $PM_{2.5}$

Concentración de $PM_{2.5}$ (µg/m3)	**Nivel de Prevención**
0-12	Bueno
13-37	Aceptable
38-55	Moderado
56-150	No Saludable para Grupos Sensibles
>150	No Saludable

Fuente: Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá

SERIES DE TIEMPO

Las series de tiempo son estructuras de datos no rectangulares, las cuales registran medidas sucesivas de una misma variable en un período de tiempo específico (Bruce et al. 2020; Hodeghatta et al. 2017). Este tipo de datos tienen como característica principal que el ordenamiento en función del tiempo es determinante para el análisis y la predicción de estos (Montegro, 2011). Otras características relevantes de las series de tiempo según (Rojas-Jímenez, 2022) son: (1) Tendencia**:** Describe el comportamiento o patrón subyacente a largo plazo de una serie temporal, no es necesariamente lineal. (2) Estacionalidad: Movimiento similar del fenómeno que se presenta en la misma época del año como un día, mes, trimestre específico. Formalmente la recurrencia de un patrón en dos o más períodos (Brockwell, 2006). (3) Estacionaridad**:** Refleja comportamientos recurrentes que no son necesariamente periódicos, es decir que no se presentan en un período específico en la serie de tiempo. Y (4) Aleatoriedad**:** Parte no explicable de los datos.

En este estudio se propone hace uso de métodos de ensamble en conjunto con predicción conformal en aras de obtener una predicción dentro de un rango de confianza disminuyendo así la incertidumbre asociada a los valores predichos y asegurando su calidad.

MÉTODOS DE ENSAMBLES BASADO EN ÁRBOLES Y APLICADOS A SERIES DE TIEMPO

Los métodos de ensamble se basan en algoritmos supervisados como lo son los árboles de decisión. Estos consisten en una o más declaraciones condicionales

anidadas para los predictores que dividen los datos. Dentro de estas particiones, se utiliza un modelo para predecir el resultado (Kuhn, M et al. 2013).

Hay dos modelos de ensamble que han demostrado ser efectivos en una amplia gama de conjuntos de datos para clasificación y regresión (Hajirahimi, Z et al. 2019), ambos de los cuales utilizan árboles de decisión como sus bloques de construcción: los bosques aleatorios y los árboles de decisión potenciados por gradiente (Müller, A et al. 2016). La necesidad de hacer uso de una colección de árboles nace del principio de que el promedio (o la mayoría de los votos) de múltiples modelos, es decir, un conjunto de modelos resulta ser más preciso que simplemente seleccionar un modelo (Bruce, P et al, 2020).

A su vez, permiten superar los problemas de previsión de datos de series temporales no lineales y no estacionarias (Houssainy, R et al. 2021).

Algunas de las características principales de los árboles de decisión mencionadas por Bruce, P et al. (2020), se listan a continuación:

División recursiva: división repetida y subdivisión de los datos con el objetivo de hacer que los resultados en cada subdivisión final sean lo más homogéneos posible.

Valor de división: un valor predictor que divide los registros en aquellos donde ese predictor es menor que el valor de división y aquellos donde es mayor.

Nodo: en el árbol de decisión, o en el conjunto de reglas de ramificación correspondientes, un nodo es la representación gráfica o de reglas de un valor de división.

Hoja: el final de un conjunto de reglas *if-then*, o ramas de un árbol. Las reglas que te llevan a esa hoja proporcionan una de las reglas de clasificación para cualquier registro en un árbol.

Pérdida: el número de clasificaciones incorrectas en una etapa del proceso de división; cuantas más pérdidas, más impureza.

Impureza: el grado en que se encuentra una mezcla de clases en una subpartición de los datos (cuanto más mezclada, más impura).

Ensamble: formación de una predicción mediante el uso de una colección de modelos.

Importancia de la variable: una medida de la importancia de una variable predictora en el rendimiento del modelo.

Metodología base de los modelos de ensamble

1. Desarrollar un modelo predictivo y registrar las predicciones para un conjunto de datos dado.

2. Repetir para varios modelos en los mismos datos.

3. Para cada registro a predecir, tomar un promedio (o un promedio ponderado, o una mayoría de votos) de las predicciones.

VARIANTES DE LOS MODELOS DE ENSAMBLE

Bagging (Bruce et al. 2020)

También llamado *agregación Bootstrap*, fue introducido por Leo Breiman en 1994. Supongamos que tenemos una variable de respuesta Y y P variables predictoras $X = X_1, X_2, \ldots, X_P$ con N registros. En lugar de ajustar los diversos modelos a los mismos datos, cada nuevo modelo se ajusta a una muestra *Bootstrap*. A continuación, se explica formalmente cómo funciona el algoritmo:

1. Inicializar M, el número de modelos a ajustar, y n, el número de registros a elegir ($n < N$). Establecer la iteración $m = 1$.

2. Tomar una muestra *Bootstrap* (es decir, muestreo con reemplazo) de n registros de los datos de entrenamiento para formar un subconjunto Y_m y X_m (la bolsa).

3. Entrenar un modelo usando Y_m y X_m para crear un conjunto de reglas de decisión $f_m(X)$.

4. Incrementar el contador de modelos $m = m + 1$. Si $m \leq M$, volver al paso 2. En el caso en que f_m predice la probabilidad $Y = 1$, la estimación bagged se calcula como:

$$\hat{f} = \frac{1}{M}\left(f_1(X) + f_2(X) + \ldots + f_M(X)\right)$$

Random Forest

El método se basa en aplicar *bagging* a árboles de decisión, con una extensión importante: además de muestrear los registros, el algoritmo también muestrea las variables. En los árboles de decisión tradicionales, para determinar cómo crear una subpartición de una partición A, el algoritmo elige la variable y el punto de división minimizando un criterio como la impureza de Gini. Con los bosques aleatorios, en cada etapa del algoritmo, la elección de la variable se limita a un subconjunto aleatorio de variables. El algoritmo formalmente es como se explica a continuación.

1. Tomar una submuestra *Bootstrap* (con reemplazo) de los registros.

2. Para la primera división, muestrear $p < P$ variables al azar sin reemplazo.

3. Para cada una de las variables muestreadas $X_{j1}, X_{j2}, \ldots, X_{jp}$, aplicar el algoritmo de división:

    a) Para cada valor s_{jk} de X_{jk}:

        1) Dividir los registros en la partición A, con $X_{jk} < s_{jk}$ como una partición y los registros restantes donde $X_{jk} \geq s_{jk}$ como otra partición.

2) Medir la homogeneidad de las clases dentro de cada subpartición de A.

b) Seleccionar el valor de s_{jk} que produce la máxima homogeneidad dentro de la partición.

4. Seleccionar la variable X_{jk} y el valor de división s_{jk} que produce la máxima homogeneidad dentro de la partición.

5. Proceder a la siguiente división y repetir los pasos anteriores, comenzando con el paso 2.

6. Continuar con divisiones adicionales, siguiendo el mismo procedimiento hasta que el árbol esté completamente desarrollado.

7. Regresar al paso 1, tomar otra submuestra Bootstrap, y comenzar el proceso nuevamente.

Boosting

El *boosting* es una técnica general para crear un conjunto de modelos. Fue desarrollado aproximadamente al mismo tiempo que el *bagging*. Al igual que el *bagging*, el *boosting* se usa más comúnmente con árboles de decisión. A pesar de sus similitudes, el *bagging* se puede realizar con relativamente poca afinación, mientras que el *boosting* requiere mucho más cuidado en su aplicación.

Una técnica general para ajustar una secuencia de modelos dando más peso a los registros con residuos grandes para cada ronda sucesiva, en las cuales cada modelo sucesivo busca minimizar el error del modelo anterior. Varias variantes del algoritmo son comúnmente utilizadas: *Adaboost, gradient boosting y stochastic gradient boosting*. Se hace uso del algoritmo de *Adaboost* para formalmente explicar cómo funciona el *Boosting* a continuación:

1. Inicializar M, el número máximo de modelos a ajustar, y establecer el contador de iteraciones $m = 1$. Inicializar los pesos de las observaciones $w_i = 1/N$ para $i = 1, 2, \ldots, N$. Inicializar el modelo de conjunto $F_0 = 0$.

2. Utilizando los pesos de las observaciones $w_1, w_2, \ldots, w_N$, entrenar un modelo f_m que minimice el error ponderado e_m definido como la suma de los pesos para las observaciones mal clasificadas.

3. Agregar el modelo al conjunto: $F_m = F_m - 1 + \alpha_m f_m$, donde $\alpha_m = log\,1 - e_m\,/e_m$.

4. Actualizar los pesos $w_1, w_2, \ldots, w_N$ para que los pesos se incrementen para las observaciones que fueron mal clasificadas. El tamaño del incremento depende de α_m, con valores más grandes de α_m que conducen a pesos más grandes.

5. Incrementar el contador de modelos $m = m + 1$. Si $m \leq M$, ir al paso 2.

La estimación boosteada se da por:

$$F = \alpha_1 f_1 + \alpha_2 f_2 + \ldots + \alpha_M f_M$$

INGENIERÍA DE CARACTERÍSTICAS

Para que un modelo de aprendizaje automático pueda predecir datos a partir de una serie temporal y utilizar variables exógenas, es necesario transformar la serie temporal creando características adicionales a partir de los rezagos de la variable. Esto implica utilizar los valores anteriores de la variable como características para la predicción, evitando la fuga de información futura durante el entrenamiento y la validación, ya que incluirlos generaría un sesgo en las predicciones al enfrentarse a datos reales. (Hyndman, et al. 2018).

En el caso específico de la concentración de material particulado $PM_{2.5}$, se analizan los gráficos de autocorrelación y autocorrelación parcial para identificar los rezagos relevantes en la serie temporal. Además, se aplica la prueba de Dickey-Fuller aumentada para determinar si la serie es estacionaria. La estacionariedad indica que las propiedades estadísticas de la serie, como la media y la varianza, son constantes en el tiempo, lo cual es crucial para muchos modelos predictivos. Si la serie no es estacionaria, puede ser necesario diferenciarla o aplicar transformaciones adicionales para lograr estacionariedad (Barbieri, 2005).

Al identificar los rezagos significativos y asegurar la estacionariedad de la serie, se puede construir un conjunto de características adecuado para que los modelos de aprendizaje automático procesen la información temporal de manera efectiva. Este enfoque permite que los modelos, que no manejan series temporales de forma directa, capturen las dependencias temporales y realicen predicciones más precisas (Amat et al, 2023).

PREDICCIÓN CONFORMAL

En muchas ocasiones, las predicciones generadas por modelos estadísticos, de aprendizaje automático o profundo pueden llegar a ser incorrectas o poco fiables para la toma de decisiones. Esta falta de fiabilidad se debe a la incapacidad de ciertos modelos para generar medidas de confianza en predicciones individuales. Aunque pueden mostrar un excelente desempeño predictivo, no cuantifican la incertidumbre asociada a una entrada específica. Esto puede deberse a diversos factores como: (1) la insuficiencia o incompletitud de los datos, (2) problemas surgidos durante el proceso de modelado o (3) la aleatoriedad y la complejidad inherente del fenómeno en estudio (Molnar, 2023).

En este contexto, la predicción conformal emerge como una herramienta clave para estimar la incertidumbre proporcionando regiones de predicción y medidas de confianza con validez estadística para las predicciones individuales obtenidas mediante modelos predictivos de cualquier naturaleza. Su aplicación resulta especialmente relevante en escenarios donde las decisiones dependen de la precisión del modelo, se enfrentan situaciones imprevistas o se busca avanzar en la automatización de procesos (Manokhin, 2023).

Las técnicas de predicción conformal se diferencian de otros enfoques como los métodos Monte Carlo o el *Bootstraping*, principalmente porque no asume un distribución especifica de los datos, siendo un modelo probabilístico no paramétrico, aplicable a una mayor variedad de modelos predictivos en escenarios reales según Molnar (2023). Por su parte, Manokhin (2023) define matemáticamente la validez de la predicción conformal a través de la probabilidad de cobertura de los intervalos o regiones de predicción generados. Un predictor conformal se considera válido si, para cualquier nivel de confianza deseado $(1-\alpha)$, el porcentaje de valores reales de la variable objetivo que se encuentran dentro de sus respectivos intervalos de predicción es, en promedio, al menos $(1-\alpha)$ a lo largo de múltiples instancias.

Recientemente, los modelos de predicción conformal se han ampliado a contextos en los que la suposición de intercambiabilidad de los datos no siempre se cumple, logrando aplicaciones exitosas en el pronóstico de series temporales, relevante para la metodología planteada en este trabajo. Sus principales ventajas incluyen su adaptabilidad a cualquier modelo predictivo, su simplicidad y eficiencia computacional, así como su capacidad para generar estimaciones de incertidumbre sin depender del tamaño del conjunto de datos.

Incertidumbre (Manokhin, 2023)

Incertidumbre aleatoria: Se refiere al tipo de incertidumbre causada por la aleatoriedad inherente e impredecible en un sistema.

Incertidumbre epistémica: Se refiere al tipo de incertidumbre que surge de la falta de conocimiento o comprensión sobre un sistema.

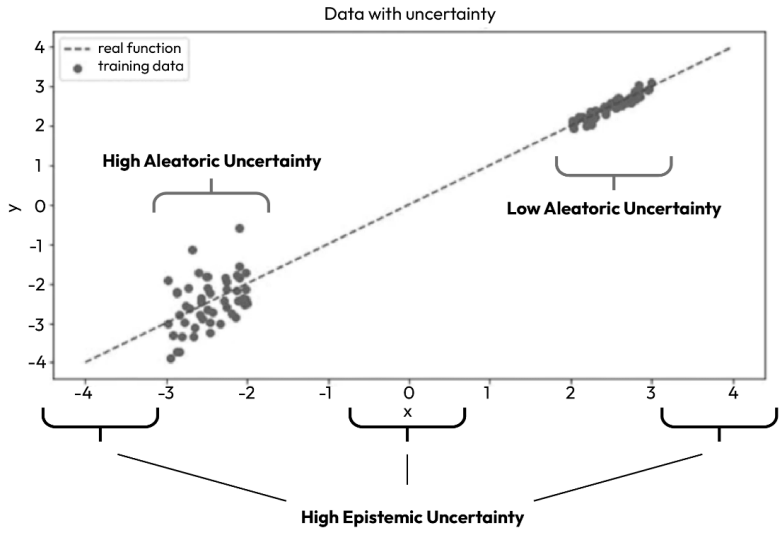
En la Figura 3 se evidencia la diferencia entre ambos tipos de incertidumbre, anteriormente mencionados, el grupo de puntos de la izquierda demuestra una alta incertidumbre aleatoria, en contraste con la baja incertidumbre aleatoria en el grupo de la derecha, atribuida a las regularidades de los datos. En la zona inferior se encuentran tres áreas, estas muestran una alta incertidumbre epistémica puesto que, corresponden a vacíos en los datos, lo que indica brechas en nuestra comprensión o conocimiento del sistema.

Componentes de la predicción conformal

Medida de no conformidad: es una función que evalúa cuánto difiere un nuevo punto de datos de los puntos de datos existentes. Se compara la nueva observación con todo el conjunto de datos (en la versión transductiva completa de la predicción conformal) o con el conjunto de calibración (en la variante más popular, la ICP). Existen dos tipos de medidas de no conformidad, las dependientes del modelo y las independientes del modelo. Las medidas de no conformidad dependientes del modelo son específicas de un tipo particular de modelo subyacente utilizado en la predicción conformal, aprovechando sus características internas para calcular las puntuaciones de no conformidad. A diferencia de las medidas independientes del

modelo, estas se adaptan al modelo utilizado y pueden personalizarse según las características y salidas del modelo, como las estimaciones de probabilidad o los límites de decisión.

Figura 3 Incertidumbre aleatoria y epistémica



Fuente:
Figura tomada de Manokhin, V. (2023) Practical Guide to Applied Conformal Prediction in Python (1^{st} ed.) Packt.

Conjunto de calibración: es una parte del conjunto de datos utilizado para calcular las puntuaciones de no conformidad para los puntos de datos conocidos.

Estas puntuaciones son una referencia para establecer intervalos o regiones de predicción para nuevos puntos de datos de prueba. El conjunto de calibración debe ser una muestra representativa de toda la distribución de datos y típicamente se selecciona al azar. Debe contener un número suficiente de puntos de datos. Si el conjunto de datos es pequeño se considera una variante llamada Predicción Conformal Transductiva (TCP) (Manokhin, 2023), explicada en las subsecciones posteriores.

El conjunto de entrenamiento se utiliza exclusivamente para entrenar el modelo de predicción base y no debe emplearse para construir el predictor conformal. De manera similar, el conjunto de calibración se reserva únicamente para el predictor conformal y no para entrenar el modelo base.

Conjunto de pruebas: contiene nuevos puntos de datos para generar predicciones. Para cada punto de datos en el conjunto de pruebas, el modelo de predicción conformal calcula una puntuación de no conformidad utilizando la medida de no conformidad y la compara con las puntuaciones del conjunto de calibración. Utilizando esta comparación, el predictor conformal genera una región de predicción que incluye el valor objetivo con un nivel de confianza definido por el usuario.

Metodología base de la predicción conformal

Existen dos variantes de la predicción conformal: la Predicción Conformal Inductiva (ICP) y Predicción Conformal Transductiva (TCP). A continuación, se explica brevemente el paso a paso de cómo funciona la Predicción Conformal Inductiva la cual será utilizada en este trabajo.

1. Un predictor conformal utiliza ejemplos de entrenamiento previos para medir la incertidumbre en torno a las predicciones para nuevas observaciones.

2. Al medir esta incertidumbre, el predictor calcula puntajes de no conformidad, que indican cuán diferente es la nueva observación en comparación con el conjunto de entrenamiento (en la versión clásica de la predicción conformal) o con la calibración (en la versión inductiva).

3. Estos puntajes de no conformidad ayudan a determinar si la nueva observación se encuentra dentro del rango esperado según los datos de entrenamiento.

4. El modelo genera medidas de confianza personalizadas y conjuntos de predicción para problemas de clasificación, o intervalos de predicción para problemas de regresión y pronóstico de series temporales.

METODOLOGÍA

Este proyecto se realiza bajo una metodología adaptada basada en las etapas del estándar CRISP-DM (de sus siglas en inglés *Cross-Industry Standard Process for Data Mining*) (IBM, 2021)

COMPRENSIÓN DEL TEMA: en primer lugar, se definen los objetivos y el alcance del proyecto, cuyo propósito simplificado es desarrollar un modelo entrenado con variables históricas de $\mathrm{PM}_{2.5}$. Incluyendo, experimentos donde se estima si la incorporación de algunas variables de movilidad impacta al desempeño del modelo predictivo.

COMPRENSIÓN DE LOS DATOS: se recolectaron y preprocesaron los datos relevantes, incluyendo los registros históricos de concentraciones de $\mathrm{PM}_{2.5}$ y los datos de tráfico vehicular en la ciudad de Medellín. Posteriormente, se realizó un análisis exploratorio para comprender la distribución, las tendencias y las relaciones entre las diferentes variables. Además, se evaluó la calidad de los datos y se abordaron problemas asociados, como valores faltantes o inconsistentes. Este último punto fue abordado con ayuda de técnicas imputación de datos, dado que los conjuntos de datos disponibles contaban con datos nulos desde su ingesta en los sitios oficiales de datos públicos de la Alcaldía de Medellín.

Descripción de los datos

Se cuenta con bases de datos de uso abierto de la calidad del aire, estas bases son generadas por diferentes estaciones de monitoreo y cámaras vehiculares en varios puntos de la ciudad, detallados anteriormente en la Figuras 1 y 2, constan de archivos tipo csv para los datos de calidad del aire (Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá, 2024) y de archivos tipo xslx para las variables de movilidad, se identifica que es necesario asignar un único carril por cada estación de monitoreo de calidad de aire en aras de obtener un conjunto de datos unificado por comuna, los datos van desde 01-01-2021 hasta el 31-10-2023.

Del conjunto de datos de calidad de aire se incluyen las siguientes variables:

Material Particulado menor a 2.5 micras $\mathrm{PM}_{2.5}$ $\mu\mathrm{g}/\mathrm{m}^3$

Código serial: representa el código de la estación con esta podemos saber dónde se encuentra ubicada, para poder comparar los datos de movilidad.

Con ayuda de la Tabla 2 se crea una columna denominada comuna la cual nos permite hacer la unión más adelante con el conjunto de movilidad.

Valor de bandera: indica la calidad de la medición (qué tan confiable es el dato reportado) que está entre 1 y 4,9. Tal como se muestra en la Tabla 3.

Tabla 2 Generalidades de la red de calidad de aire del Valle de Aburrá.

Serial	Comuna	Ubicación
83	16	Belén - I.E Pedro Justo Berrio
84	14	El Poblado - I.E INEM Sede Santa Catalina
85	10	San Cristóbal - Parque Biblioteca Fernando Botero
86	4	Aranjuez - I.E Ciro Mendía

Fuente:
Tomado de Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá 2024.

Tabla 3 Generalidades calidad de los datos Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá.

Valor Bandera	Calidad del dato
1	Dato válido
-1	Dato válido por el operador anterior
1.8 - 2.5	Dato dudoso
2.6 - 3.9	Dato malo
≥ 4.0	Dato faltante
Dato -9999 y Calidad (Flag) 1	Equipo fuera de operación

Fuente:
Tomado de Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá 2024

Con la información proporcionada por la Tabla 3, Se eliminaron los datos con valores bandera mayores a 1. Para los valores faltantes generados por esta modificación (aproximadamente entre el 5 % y el 7 % del conjunto de datos), se aplicó la imputación mediante la media por hora de cada día. Esta técnica es ampliamente utilizada en la imputación de series de tiempo, ya que permite preservar los patrones presentes en la serie, los cuales están influenciados por actividades diarias como el tráfico vehicular y la actividad industrial propias de la ciudad.

Además, este método ayuda a mantener la correlación temporal al utilizar información de días con condiciones similares, evitando interpolaciones que podrían distorsionar la dinámica real del fenómeno (Hua, 2024). Es importante destacar que la imputación se llevó a cabo considerando la baja presencia de datos atípicos y con plena conciencia del posible sesgo que podría introducirse al entrenar el modelo

Seguidamente, se observó que aún persistían algunos valores nulos debido a la ausencia de registros en ciertas fechas de la serie temporal. Para completar estos datos, se aplicó una imputación hacia adelante, que consiste en utilizar la última observación disponible antes del dato nulo para estimarlo. Este método se considera adecuado, ya que las tendencias de $PM_{2.5}$ se manifiestan en períodos temporales más amplios, desde semanas hasta meses, según el análisis

exploratorio realizado. Por ello, esta técnica permite preservar la coherencia en las tendencias a mediano y largo plazo (Echeverri, 2008).

Además, según Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá (2024), estos valores nulos suelen generarse debido a fallas o mantenimientos en los sensores, los cuales quedan temporalmente fuera de funcionamiento. Dado que la variabilidad en escalas cortas es limitada, el uso del último valor registrado para completar la serie resulta justificado y apropiado.

Del conjunto de datos de movilidad se incluyen las siguientes variables:

- Carril: carril configurado para cada dispositivo del proyecto SIMM.
- Fecha Tráfico: fecha y hora de registro.
- Fecha: fecha de registro.
- Hora: hora de registro.
- Día-núm: día de la semana de registro en número entre lunes (1) y domingo (7). Mes-núm: mes de registro en número entre enero (1) y diciembre (12).
- Año: año de registro. Velocidad: velocidad promedio de todos los vehículos en la hora de registro en km/h.
- Corredor: corredor en el que se encuentra ubicado el dispositivo.
- Sentido: sentido de circulación del corredor en el que se encuentran los vehículos registrados.
- Operación: tipo de flujo vehicular del corredor (ininterrumpido - continuo, semi interrumpido, interrumpido).
- Intensidad: volúmenes vehiculares en vehículos/hora.
- Categoría 1: volúmenes vehiculares para la categoría 1 (vehículos con longitudes entre 0 y 6 metros) en vehículos/hora.
- Categoría 2: volúmenes vehiculares para la categoría 2 (vehículos con longitudes entre 6 y 12 metros) en vehículos/hora.
- Categoría 3: volúmenes vehiculares para la categoría 3 (vehículos con longitudes mayores a 12 metros) en vehículos/hora.
- Ocupación: porcentaje promedio de la hora que permanecieron los vehículos ocupando la zona de conteo.
- Corredor-Sentido: corredor en el que se encuentra ubicado el dispositivo y el sentido de circulación en el que se encuentran los vehículos registrados.
- Longitud: coordenada Y en sistema de coordenadas Magna Sirgas.
- Latitud: coordenada X en sistema de coordenadas Magna Sirgas.
- Código comuna: código de la comuna o corregimiento de Medellín por el cual cruza el corredor.

Como se mencionó anteriormente, se debe seleccionar un único carril cercano a las estaciones de monitoreo de calidad del aire. Esto se debe a la existencia de varias series de tiempo diferentes para cada carril dentro de una misma vía. Para unificar los conjuntos de datos, es necesario delimitar esta información. En la Tabla 4, se presentan los carriles seleccionados.

Tabla 4 Carriles seleccionados para la ingesta de datos

Carril	Serial	Comuna
XC-1-AV80-30A-ENT:ORI-OCC	83	16
XC-1-ORI-52-ENT:NOR-SUR 1	85	10
DAI-INFTESORO-P1-L0	84	14
HK-3-C53-C94-ENT:NOR-SUR	86	4

Fuente:
Elaboración propia

En este caso, el tratamiento de datos nulos se realiza mediante un imputador de regresión lineal multivariante (IterativeImputer), similar al utilizado en el estudio de Kokkinos et al (2021). Este imputador estima los valores faltantes en una o varias columnas utilizando la información disponible en las filas sin datos nulos. Su aplicación es adecuada por varias razones, entre ellas la alta correlación entre variables de tráfico, como velocidad, flujo vehicular e intensidad, relaciones que el imputador aprovecha para realizar estimaciones más precisas (Li et al., 2013).

Además, este método permite capturar patrones horarios, variaciones según el día de la semana y diferencias espaciales presentes en los datos de tráfico vehicular, lo que contribuye a mejorar la coherencia de la imputación (Jiang et al, 2023). Estas características justifican su uso en el presente estudio.

Para las variables categóricas, como operación, corredor y sentido, se procede primero a codificarlas y posteriormente a imputar los valores nulos utilizando el dato inmediatamente anterior. Esta estrategia se justifica de manera similar a la imputación hacia adelante aplicada previamente en el conjunto de datos de calidad del aire.

Estadísticos Básicos

Tabla 5 Estadísticos básicos conjunto de datos de calidad del aire

Código Serial	Conteo	Media	Des. Estándar	Mínimo	25%	50%	75%	Máximo
83	27720	19.9613	11.3848	0.0	11.1151	18.1950	27.1077	78.2079
84	27720	15.8967	8.2477	0.0	9.9172	14.8181	20.4503	75.6756
85	27720	15.7275	7.5588	0.0	10.3345	14.7336	19.7552	65.9214
86	27720	18.2769	8.5221	0.0	12.1652	17.4652	22.8561	110.5340

Fuente:
Elaboración propia

En el análisis descriptivo se calcularon estadísticas básicas para cada conjunto de datos, incluyendo la media, mediana, desviación estándar, percentiles (25%, 50%

y 75%), así como los valores mínimo y máximo. En las Tablas 5 y 6, se presenta un resumen de dichas estadísticas por cada uno de los conjuntos de datos utilizados

Tabla 6 Estadísticos básicos - Variables continuas del conjunto de datos de movilidad

Variable	Conteo	Media	Des. Estándar	Mínimo	25%	50%	75%	Máximo
Velocidad	92031	16.5951	12.9634	1.0	10.0	13.0	19.0	140.0
Intensidad	92031	166.6582	130.5702	0.0	64.0	138.0	225.0	1158.0
Vehículos long 1	92031	116.0651	96.8396	-49.7	32.0	96.95	176.85	717.0
Vehículos long 2	92031	24.2856	31.0422	0	2.0	13.0	31.0	248.0
Vehículos long 3	92031	9.8918	16.9132	-2.9	0.0	3.5	10.0	307.0
Ocupación	92031	36.0138	27.4545	-73.3	11.0	32.0	58.0	100.0

Fuente:
Elaboración propia

PREPARACIÓN DE LOS DATOS: se seleccionaron las características relevantes de los conjuntos de datos disponibles para la predicción de $\mathrm{PM}_{2.5}$, como la ocupación de las vías principales y la densidad de vehículos en las mismas. Adicionalmente con ayuda de una matriz de correlación se identificaron posibles colinealidades entre las variables de movilidad y la variable objetivo. Así mismo, se llevó a cabo la prueba de Dickey-Fuller Aumentada (ADF), para evaluar la necesidad de aplicar transformaciones adicionales a la serie temporal objetivo para que sea estacionaria.

Posteriormente, los datos fueron normalizados, las variables categóricas codificadas, y los valores nulos tratados mediante imputación con regresión lineal multivariante (IterativeImputer), a fin de evitar la pérdida de registros en la serie temporal como se mencionó anteriormente. Finalmente, se unificaron las bases de datos y se dividió el conjunto de datos en los subconjuntos de entrenamiento, validación y prueba.

Análisis de correlación, estacionaridad, ACF y PACF

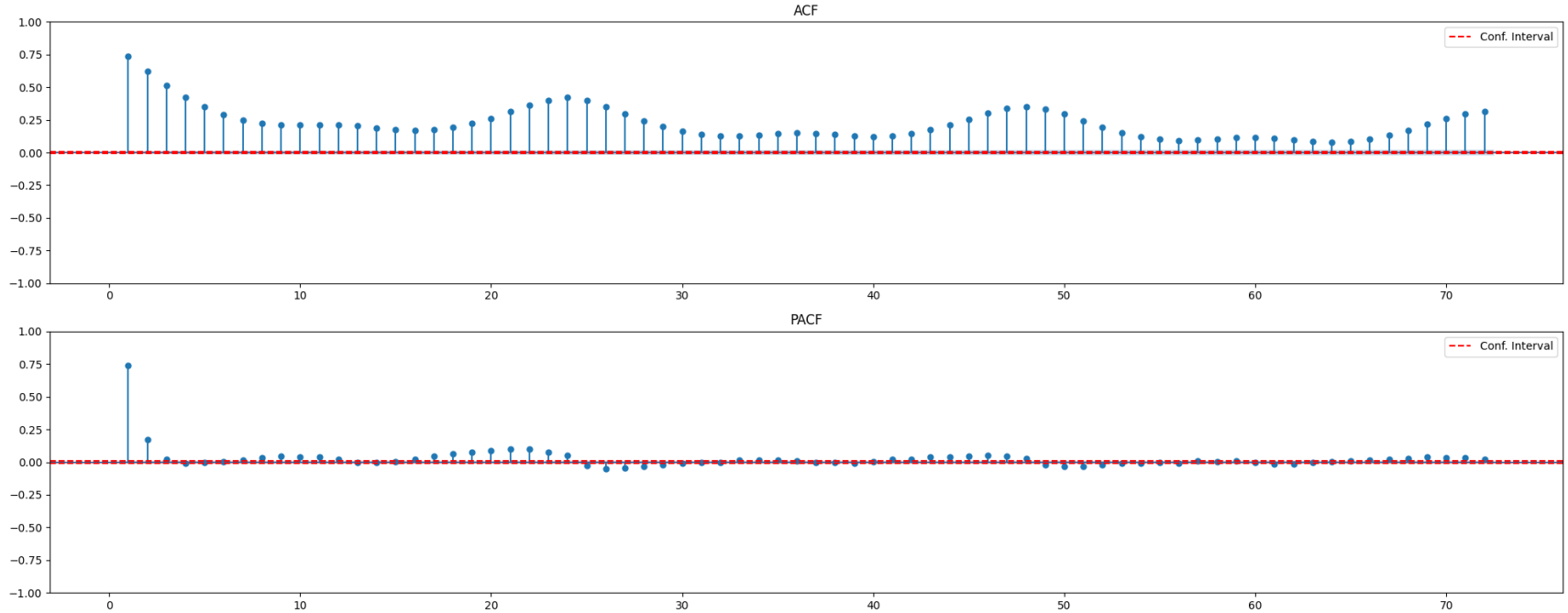
En la Figura 4 se muestra las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie temporal de $PM_{2.5}$ para 72 rezagos en la gráfica superior, 168 rezagos para la gráfica en el medio y 672 rezagos para la gráfica inferior, estas permiten analizar la dependencia entre los valores pasados y presentes de la variable.

La gráfica de autocorrelación (ACF) muestra una correlación alta en los primeros rezagos, lo que sugiere que la serie tiene una fuerte relación con sus valores pasados inmediatos, Además, se observa una disminución gradual y persistente, lo que sugiere la posible presencia de componentes estacionales o tendencias a largo plazo en los datos. Sumado a lo anterior, en la totalidad de las gráficas se pueden identificar patrones cíclicos bien definidos, con picos de autocorrelación aproximadamente cada 24 horas, lo que evidencia una fuerte componente estacional diaria en la serie.

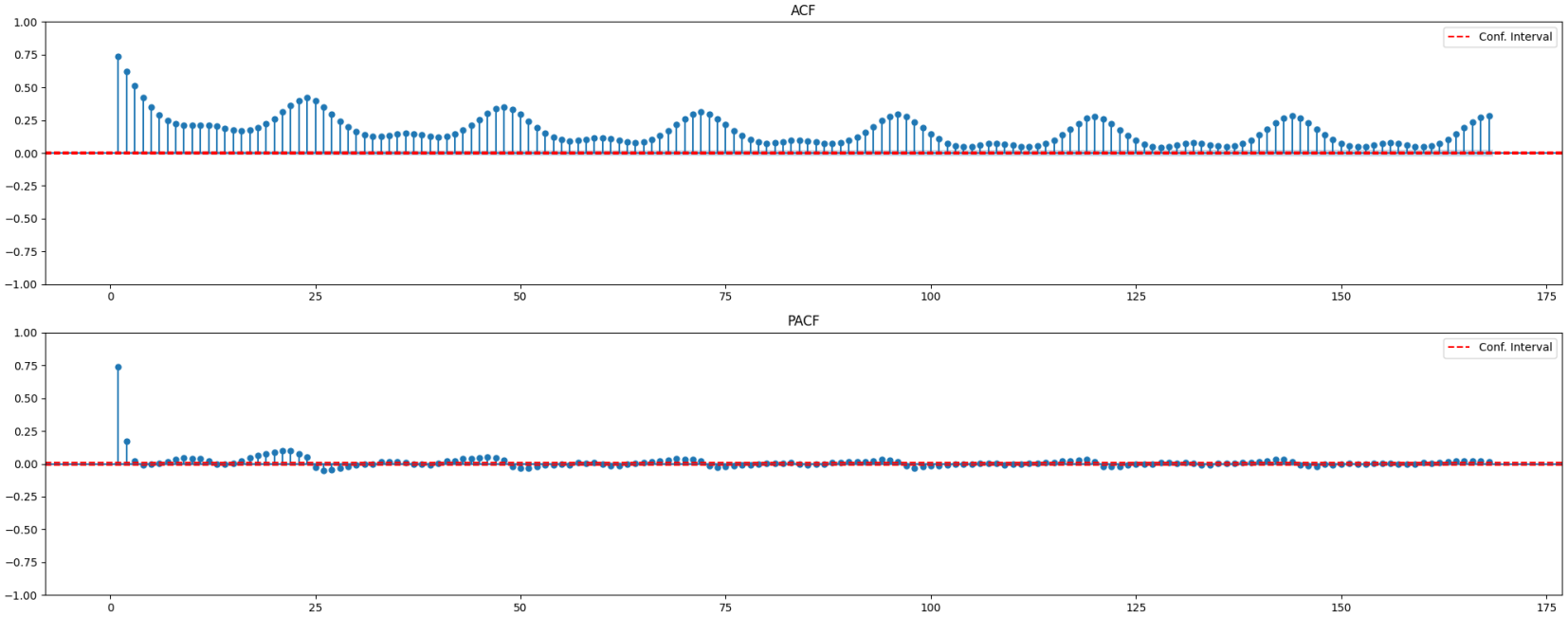
Por otro lado, la gráfica de autocorrelación parcial (PACF) muestra que, tras el primer rezago, los valores disminuyen rápidamente y se estabilizan cerca de cero. Los primeros rezagos son los más significativos, ya que se encuentran claramente fuera del intervalo de confianza (representado por la línea punteada roja) en las gráficas. Esto refuerza la idea de que el valor actual de la serie depende principalmente de sus valores inmediatos anteriores, aunque algunos rezagos más lejanos también muestran una influencia no despreciable, lo que indica la posible presencia de estructuras estacionales más complejas. En el caso de los rezagos más largos (168 y 672 horas), la periodicidad en la ACF se vuelve más evidente, con picos que confirman la existencia de una tendencia cíclica diaria recurrente.

Figura 4 Gráficos ACF y PACF

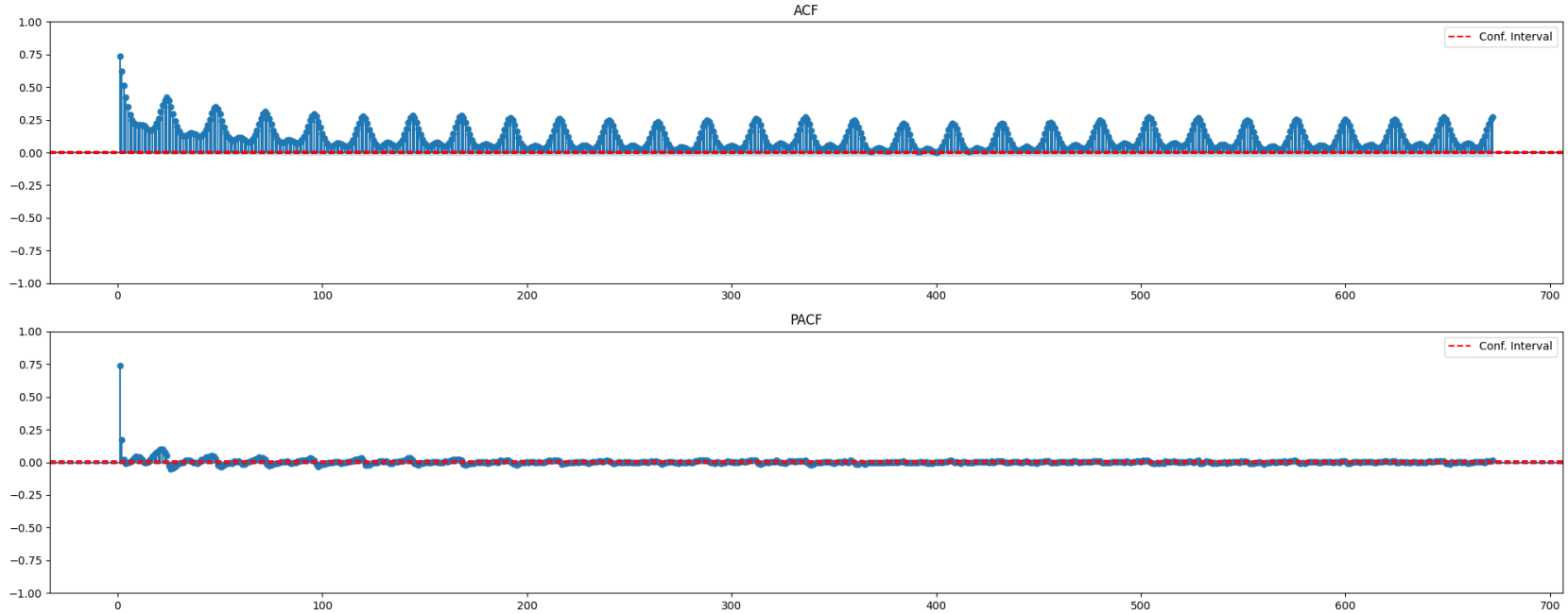
72 rezagos*



168 rezagos*


ACF
Conf. Interval
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
25
50
75
100
125
150
175
PACF
Conf. Interval
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
25
50
75
100
125
150
175

672 rezagos*



**Cada rezago corresponde a una unidad de tiempo en horas*

Fuente:

Elaboración propia.

Transformación de variables con *Temporian*

Con ayuda de la librería *Temporian*, disponible en Google Colab, se generó la variable objetivo como una serie temporal. Además, se crearon nuevas variables que representan rezagos, horizontes y estadísticos móviles, como la media, desviación estándar y suma de $PM_{2.5}$, calculados sobre diferentes ventanas de tiempo. Se seleccionaron únicamente rezagos desde tres horas previas a la predicción, ya que, según el análisis de las gráficas de ACF y PACF, estos son los más relevantes para la predicción.

Por otro lado, dado que el modelo no puede utilizar las fechas directamente, estas se transformaron en características numéricas que el modelo puede procesar. En particular, se incluyeron las variables día-num, mes-num y año, que permiten al modelo incorporar nociones de temporalidad.

Finalmente, los datos se dividieron en conjuntos de entrenamiento, validación y prueba, organizados como se muestra en la Tabla 7.

Tabla 7 Particiones del conjunto de datos

Partición	**Fechas de inicio y fin**
Entrenamiento	01-01-2021 – 30-05-2023
Validación	31-05-2023 – 30-08-2023
Prueba	31-08-2023 – 31-10-2023

Fuente:
Elaboración propia.

MODELADO: se entrenaron diferentes algoritmos de ensamble basados en árboles, específicamente Random Forest y XGBoost. Ambos modelos fueron comparados con un modelo de regresión lineal como referencia. Los modelos se entrenaron utilizando todas las variables disponibles (calidad de aire y movilidad), así como solo la serie temporal de la variable de interés, junto con algunas variables de estadísticas móviles. Esto permitió determinar que la adición de las variables de los datos de movilidad no representa una mejora para el desempeño del modelo.

Una vez entrenados los modelos, se ajustaron los hiperparámetros del mejor utilizando técnicas como validación cruzada con un enfoque temporal y GridSearch. El rendimiento de los modelos se evaluó utilizando métricas relevantes, como la Raíz cuadrada del Error Cuadrático Medio (RMSE), la Media del Error Absoluto (MAE) y el índice de correlación (R^2). Finalmente, se empleó la predicción conformal para calcular los intervalos de confianza, basados en un nivel de incertidumbre previamente seleccionado. Las predicciones del mejor modelo optimizado fueron graficadas junto con sus intervalos de confianza.

Predicciones de cada modelo

En total, se entrenaron tres tipos de algoritmos para su comparación: regresión lineal, Random Forest y XGBoost. En primera instancia, se utilizaron todas las variables disponibles, y luego, para la comparación, se entrenaron los modelos utilizando únicamente las variables propias de la serie temporal de $\mathrm{PM}_{2.5}$ y las variables generadas con *Temporian* sobre estadísticas móviles de la misma.

A continuación, se presentan los resultados obtenidos de la validación cruzada, en la que, debido a la temporalidad de los datos, se utilizó TimeSeriesSplit para evitar la filtración de información futura. Adicionalmente, se evaluó el rendimiento en el conjunto de validación para asegurar que las métricas fueran consistentes, los resultados obtenidos se muestran en las Tablas 8 y 9.

Tabla 8 Resultados de entrenamiento sin variables de movilidad

Modelo	Media RMSE	Media MAE	Media MAPE	Media R^2	Val RMSE	Val MAE	Val MAPE	Val R^2
Regresión lineal	8.0212	6.1946	35.4674	0.2232	6.7117	5.1495	29.4837	0.2174
Random Forest	8.1485	6.3506	36.3606	0.2011	6.7511	5.1690	29.5953	0.2082
XGboost	**8.2214**	**6.3430**	**36.3171**	**0.1872**	**6.7003**	**5.1073**	**29.2420**	**0.2201**

Fuente:
Elaboración propia

Tabla 9 Resultados de entrenamiento con variables de movilidad

Modelo	Media RMSE	Media MAE	Media MAPE	Media R^2	Val RMSE	Val MAE	Val MAPE	Val R^2
Regresión lineal	8.2210	6.3135	36.1482	0.1834	6.7089	5.1548	29.5140	0.2181
Random Forest	8.5244	6.6165	37.8830	0.1118	6.6026	5.0945	29.1688	0.2427
XGboost	**9.1583**	**7.0633**	**40.4412**	**-0.0441**	**6.4890**	**4.9722**	28.4685	**0.2685**

Fuente:
Elaboración propia

RESULTADOS

Comparación de modelos

El análisis comparativo de los modelos presentados en las Tablas 8 y 9 permite evaluar el impacto de la inclusión de variables de movilidad en el desempeño predictivo de los algoritmos de regresión lineal, Random Forest y XGBoost. Se analizaron cuatro métricas de error $(RMSE, MAE, MAPE\ y\ R^2)$ tanto en la fase de entrenamiento como en la de validación, lo que permite medir la capacidad de generalización de cada modelo.

En primer lugar, al observar los resultados sin variables de movilidad (Tabla 8), se evidencia que la regresión lineal obtiene el mejor desempeño en términos

generales, con un $RMSE$ medio de 8.0212 en entrenamiento y 6.7117 en validación, además de un R^2 de 0.2232 y 0.2174, respectivamente. XGBoost presenta un rendimiento similar, con un $RMSE$ en validación de 6.7003 y un R^2 de 0.2201, lo que indica que es un modelo competitivo frente a la regresión lineal. Random Forest, en cambio, muestra los valores de error más altos, con un $RMSE$ de 8.1485 en entrenamiento y 6.7511 en validación, y el menor coeficiente de determinación (R^2) en ambos casos, lo que sugiere que es el modelo menos efectivo en este escenario.

Por otro lado, al incluir las variables de movilidad (Tabla 9), se observa un deterioro en el desempeño de los modelos en la fase de entrenamiento, especialmente en XGBoost, cuyo R^2 disminuye a valores negativos (-0.0441), acompañado de un $RMSE$ de 9.1583 y un $MAPE$ de 40.4412, lo que indica que el modelo no logra capturar adecuadamente la relación entre las variables. Sin embargo, en la fase de validación, XGBoost logra mejorar su desempeño con respecto a la configuración sin movilidad, reduciendo su $RMSE$ a 6.4890 y alcanzando el mayor R^2 (0.2685) entre los tres modelos, lo que sugiere que, pese a la sobrecarga de información en el entrenamiento, logra generar mejores predicciones en validación.

En el caso de la regresión lineal, la inclusión de variables de movilidad no representa una mejora significativa, ya que sus métricas se mantienen prácticamente constantes en ambas configuraciones. Su $RMSE$ en validación apenas varía de 6.7117 a 6.7089, y su R^2 pasa de 0.2174 a 0.2181, lo que indica que la adición de estas variables no aporta una ganancia relevante en términos predictivos para este modelo.

Random Forest, en cambio, experimenta una mejora notable con la incorporación de las variables de movilidad. Aunque su desempeño en entrenamiento empeora ($RMSE$ de 8.5244 y R^2 de 0.1118), en validación logra reducir su $RMSE$ a 6.6026 y mejorar su R^2 a 0.2427, lo que sugiere que las variables adicionales le permiten capturar mejor las tendencias de los datos en la fase de prueba.

Optimización del modelo

Una vez seleccionado XGBoost como el mejor modelo, se procede a optimizarlo utilizando GridSearch, aplicando un enfoque temporal para las diferentes divisiones del conjunto de entrenamiento. Se experimenta con varios hiperparámetros, entre los cuales se incluyen: número de estimadores, máxima profundidad, tasa de aprendizaje, submuestra y gamma, que permite emplear la regularización L_2 para prevenir el sobreajuste. Los hiperparámetros obtenidos se muestran en la Tabla 10.

XGBoost se considera el mejor modelo debido a su capacidad avanzada de manejo de parámetros y técnicas de regularización, lo que le permite evitar el sobreajuste y lograr un mejor desempeño en la fase de validación. En las pruebas iniciales, se evidenció que la regresión lineal presentaba un sobreajuste considerable, mostrando un buen desempeño en entrenamiento, pero un deterioro en validación, lo que indicaba que no generalizaba bien a datos nuevos. Para mitigar este problema, se optó por agregar ruido a las variables de entrada, con el objetivo de

reducir su sensibilidad a patrones espurios en los datos de entrenamiento. Sin embargo, esta estrategia no logró mejoras significativas en la capacidad predictiva del modelo, lo que sugiere que la regresión lineal no era lo suficientemente flexible para capturar las relaciones subyacentes en los datos, incluso con la adición de ruido.

Tabla 10 Resultados optimización del modelo XGBoost

Hiperparámetro	Valor
gamma	1
Learning_rate	0.05
Max_depth	5
N_estimators	100
Subsample	0.8

Fuente:
Elaboración propia

En contraste, XGBoost demostró una mayor capacidad de generalización gracias a su combinación de técnicas como la reducción de la varianza mediante boosting, la regularización L1 y L2, y su capacidad para manejar relaciones no lineales en los datos. A pesar de que su desempeño en la fase de entrenamiento fue menos favorable que el de otros modelos debido a su proceso de regularización, en validación logró los mejores resultados, con un menor $RMSE$ y un mayor coeficiente de determinación (R^2). Esto indica que, aunque XGBoost puede parecer menos ajustado a los datos de entrenamiento en una evaluación superficial, en realidad está evitando el sobreajuste y capturando mejor la estructura de los datos de validación.

Adicionalmente, la flexibilidad de XGBoost en el ajuste de hiperparámetros permitió optimizar su rendimiento mediante GridSearch, afinando parámetros clave como la tasa de aprendizaje, la profundidad máxima y el número de estimadores. Gracias a este proceso de optimización, el modelo logró mejorar su capacidad predictiva, demostrando ser la mejor opción para abordar la tarea propuesta.

Reentrenamiento del modelo y predicción conformal

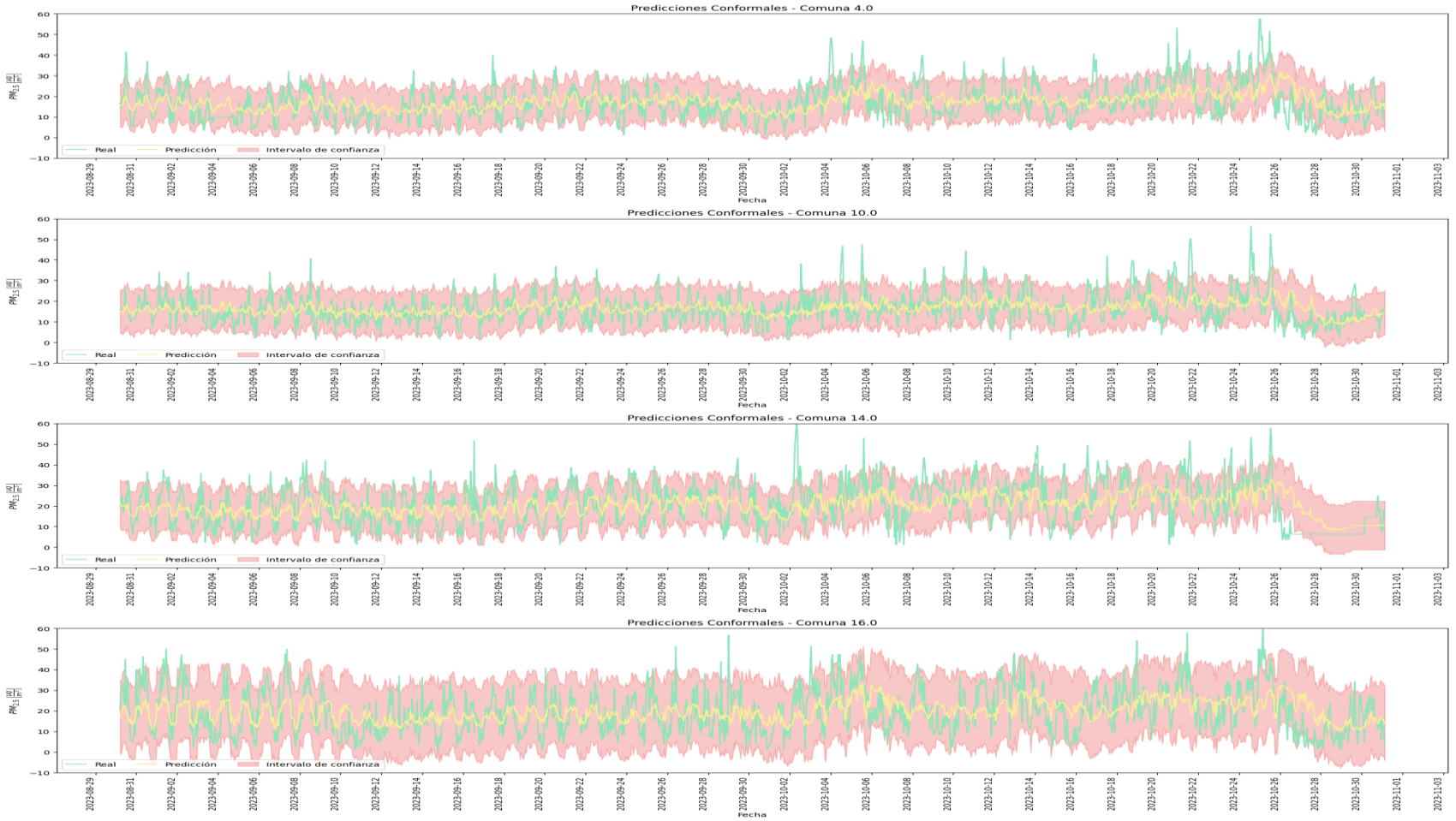
Finalmente, el modelo se reentrenó utilizando los hiperparámetros de la Tabla 10 y se incorporó la predicción conformal, siguiendo la metodología inductiva. Esta metodología emplea un conjunto de calibración para calcular los intervalos de confianza, como se muestra en la Figura 5. En dicha figura, se grafican tres componentes: la línea verde representa los valores reales de $\mathrm{PM}_{2.5}$ entre las fechas 31-08-2023 y 31-10-2023. La línea amarilla muestra las predicciones puntuales generadas por el modelo, y la franja roja indica los intervalos de confianza superior e inferior para cada predicción. Estos intervalos de confianza fueron construidos con un nivel de significancia de α=0.05, lo que implica que se espera que aproximadamente el 95% de los valores reales se encuentren dentro del rango predicho. En este caso, se obtiene una cobertura del 89.26%, lo que sugiere que,

aunque la metodología proporciona una alta certeza en las predicciones, aún existen algunas observaciones que quedan fuera de los intervalos.

En la imagen se presentan los resultados de la predicción conformal para cuatro comunas diferentes: 4, 10, 14 y 16 en Medellín, Colombia. Se observa que las predicciones del modelo siguen de manera consistente la tendencia de los valores reales, lo que sugiere un buen ajuste. Sin embargo, se pueden notar algunas fluctuaciones y picos en los datos reales que no siempre son capturados con precisión por el modelo, lo que indica la presencia de posibles eventos atípicos o variaciones que no han sido completamente explicadas por las variables utilizadas en la modelación.

Además, los intervalos de confianza, representados por la franja roja, se mantienen relativamente estables en la mayoría de los casos, aunque en ciertas fechas específicas se amplían, lo que indica una mayor incertidumbre en la predicción. Esto podría estar relacionado con la variabilidad de los datos en esas fechas o con la escasez de información en el conjunto de entrenamiento para esos periodos. En particular, en la comuna 16 se observa una mayor dispersión en los valores reales y predichos en comparación con las otras comunas, lo que sugiere que las características de esta zona pueden estar generando una mayor dificultad en la predicción de $PM_{2.5}$.

Figura 5 Comparación predicción conformal y valores reales por comuna



Fuente:
Elaboración propia

CONCLUSIONES

Los resultados obtenidos en este estudio evidencian el impacto diferenciado de la inclusión de variables de movilidad en los modelos de predicción de PM_2.5. Mientras que el modelo XGBoost experimenta un deterioro significativo en su desempeño durante el entrenamiento, logra obtener la mejor precisión en la fase de validación. En contraste, Random Forest muestra una mejora moderada en su capacidad predictiva, mientras que la regresión lineal permanece prácticamente inalterada. Estos hallazgos sugieren que la relación entre la variable objetivo y las variables de movilidad no es trivial, y que modelos más complejos pueden beneficiarse de esta información en validación, pese a dificultades en su ajuste durante el entrenamiento. Mediante el uso de técnicas avanzadas de ingeniería de características, es posible mejorar significativamente la utilidad de estas variables. Sin embargo, se identificaron limitaciones clave, como la presencia de datos nulos y la imputación basada en regresión lineal, que pueden afectar la calidad de las variables de movilidad, disminuyendo su impacto en el entrenamiento de los algoritmos. Además, la falta de cámaras vehiculares exclusivas asociadas a estaciones específicas de calidad del aire y la necesidad de seleccionar un carril adecuado representan desafíos críticos para maximizar el aporte de estas variables al modelo.

Desde una perspectiva metodológica, el modelo de regresión lineal con variables exógenas se comporta de manera similar a un modelo SARIMAX, al incorporar rezagos y tendencias móviles en la predicción. A su vez, los algoritmos de aprendizaje automático, como Random Forest y XGBoost, han demostrado ser capaces de obtener precisiones comparables a las de los modelos estadísticos tradicionales en la predicción de series temporales. Este resultado es particularmente relevante cuando los datos de entrada se preparan adecuadamente para evitar la filtración de información, subrayando la importancia de un correcto preprocesamiento en el diseño de los modelos.

Por otro lado, la aplicación de la predicción conformal permitió establecer límites de confianza para evaluar la fiabilidad del modelo en distintos escenarios espaciales y temporales. La cobertura del 89.26%, aunque cercana al 95% esperado con α=0.05, indica que aún hay margen de mejora en la estimación de los intervalos, especialmente en el tratamiento de eventos extremos o en la inclusión de nuevas variables explicativas que puedan reducir la incertidumbre en la predicción. Esta metodología, que emplea un conjunto de calibración para calcular intervalos de confianza, ha permitido construir un modelo altamente fiable, con baja incertidumbre y granularidad horaria. Gracias a esto, se obtiene un conocimiento más detallado del comportamiento de $PM_{2.5}$ y se facilita la predicción en ventanas temporales amplias, como los dos meses evaluados en este estudio, logrando mantener un error reducido.

Finalmente, se estima que la inclusión de variables meteorológicas, como velocidad y dirección del viento, temperatura y precipitaciones, podría mejorar significativamente la capacidad predictiva del modelo. Estas variables no solo contribuirían a aumentar la precisión en las estimaciones, sino que también aportarían mayor interpretabilidad, al permitir identificar los factores que inciden en los niveles elevados de $PM_{2.5}$.

En conclusión, la combinación de modelos de aprendizaje automático con técnicas avanzadas de calibración y validación ofrece un enfoque prometedor para la predicción de contaminantes atmosféricos. No obstante, la calidad de los datos de entrada y la adecuada selección de características juegan un papel crucial en la precisión de los modelos. Por lo tanto, futuras investigaciones deberían centrarse en mejorar la disponibilidad y precisión de las variables explicativas, así como en el desarrollo de estrategias de modelación que minimicen la incertidumbre y optimicen la capacidad predictiva de los algoritmos utilizados.

ANEXOS

Con el fin de permitir la replicación y difusión de los resultados obtenidos en este trabajo de grado, se ha creado un repositorio en GitHub donde se encuentran los diferentes scripts utilizados para el desarrollo de este proyecto, incluyendo preprocesamiento de cada base de datos individualmente, unificación de los datos, uso de *Temporian* para ingeniería de características y modelación.

El repositorio está disponible en el siguiente enlace:

https://github.com/Matu10100/AirQuality_Prediction

REFERENCIAS

Alcaldía de Medellín (2025). *Mapa Cámara de Fotodetección*. Obtenido de https://www.medellin.gov.co/SIMM/mapas/index.html?map=camarasFotodeteccion

Amat, R. et al. (2023). *Predicción de series temporales con gradient boosting: Skforecast, XGBoost, LightGBM y Catboost*. Obtenido de skforecast (Version 0.14.0): https://doi.org/10.5281/zenodo.8382787

Azur, M et al. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 40-49.

Barbieri, L. (2005). Panel Unit Root Tests: A Review. *Università Cattolica del Sacro Cuore*.

Bose, R et al. (2020). Time series forecasting using double exponential smoothing for predicting the major ambient air pollutnats. *Information and communication technology for sustainable developments* (págs. 603-613). Singapur: Springer.

Brockwell, P. (2006). Time Series: Theory and Methods. En P. Brockwell, *Time Series: Theory and Methods* (págs. 14-38). New York: Springer.

Bruce, P et al. (2020). Nonrectangular Data Structures. En P. Bruce, A. Bruce, & P. Gedeck, *Practical statistics for data scientist* (págs. 6-7). O'reilly.

Cao, C. (2024). How to better predict the effect of urban traffic and weather on air pollution? Norwegian evidence from machine learning approaches. *Journal of Economic Behavior & Organization*, 544-569. doi:https://doi.org/10.1016/j.jebo.2024.03.018

Carlés, F et al. (2023). Air Quality Time Series Forecasting Using Machine Learning Algorithms. *XLIX Latin American Computer Conference.* Bolivia: IEEE.

Casali, Y et al. (2022). Machine learning for spatial analyses in urban areas: a scoping review. *Sustainable Cities and Society*.

Chaudhary, V et al. (2018). Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India. *Samsung Research Institute India*.

Chengyong, J et al. (2021). Road traffic and air pollution: Evidence from a nationwide traffic control during coronavirus disease 2019 outbreak. *Science of The Total Environment*.

Das, R et al. (2022). High granular and short term time series forecasting of PM2.5 air pollutant - a comparative review. *Springer Nature*, 1253-1287.

de Andrade, M. G. (2023). Medellín, la ciudad colombiana que logró reducir el calor con un entramado de corredores verdes. *BBC News Mundo*. Obtenido de https://www.bbc.com/mundo/articles/cp3d1v0rryro

Echeverri, C. et al. (2008). Relación entre las partículas finas (PM 2.5) y respirables PM 10) en la ciudad de Medellín. *Revista Ingenierías Universidad de Medellín*.

Elangasinghe, M et al. (2014). Complex time series analysis of pm10 and pm2.5 for coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment*, 106-116.

Ghasemi, A et al. (2019). Integration of ANFIS model and forward selection method for air quality forecasting. *Air Quality, Atmosphere Health* (págs. 59-72). Springer.

Gocheva-Ilieva, S et al. (2014). Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment*, págs. 1045-1060.

Gónzalez, C et al. (2018). High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM10 dynamics. Atmospheric Pollution Research, 934-948.

González, M et al. (2023). Estimación del aporte por fuentes a la concentración de pm2.5 en el valle de aburrá. *Calidad de aire, cambio climático y salud pública* (págs. 367-372). Santa Marta: Hill Consulting.

Hajirahimi, Z et al. (2019). Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence*.

He, H et al. (2020). Study of LSTM air quality index prediction based on forecasting timelines. *IOP Conference Series: Earth and Environmental Science* (pág. 032113). IOP Publishing.

Hodeghatta, D et al. (2017). Business analytics process and data exploration. En D. U. Hodeghatta, & U. Nayak, *Business Analytics Using R - A practical approach* (págs. 96-97). Bangalore: Apress.

Houssainy, R et al. (2021). Time Series Forecasting Using Tree Based Methods. *Journal of Statistics Applications & Probability*.

Hua, V. (2024). The impact of data imputation on air quality prediction problem. *PLOS ONE*.

Hyndman, R et al. (2018). *Forecasting principles and practice.* Melbourne: OTexts.

IBM (2021). Conceptos básicos de ayuda de CRISP-DM. Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Instituto de Hidrología, Meteorología y Estudios Ambientales. (2021). *Características climatológicas de ciudades principales y municipios turísticos.* Obtenido de Instituto de Hidrología, Meteorología y Estudios Ambientales: http://archivo.ideam.gov.co/documents/21021/21789/1Sitios+turisticos2.pdf/cd4106e9-d608-4c29-91cc-16bee9151ddd

IQAir (2022). *PM2.5*. Obtenido de IQAir: https://www.iqair.com/es/newsroom/pm2-5

Jiang, L. et al. (2023). *A Deep Learning Framework for Traffic Data Imputation Considering Spatiotemporal Dependencies.* Retrieved from Cornell University: https://arxiv.org/abs/2304.09182

Jiao, Y et al. (2019). Prediction of air quality index based on LSTM. *8th joint international information technology and artificial intelligence conference* (págs. 17-20). Karlsruhe: Morgan Kaufmann Publishers Inc.

Kokkinos, K et al. (01 de 11 de 2021). A comparative analysis of statistical and computational intelligence methodologies for the prediction of traffic-induced fine particulate matter and NO2. *Journal of Cleaner Production*, 1-17. doi:https://doi.org/10.1016/j.jclepro.2021.129500

Kuhn, M et al. (2013). *Applied Predictive Modeling.* Springer. doi:DOI:10.1007/978-1-4614-6849-3

Li, J et al. (2019). Prediction of pm2.5 concentration based on CEEMD-LSTM model. *Chinese control conference* (págs. 8439-8444). Guangzhou: IEEE.

Li, L., et al. (2013). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, 108-120.

Manokhin, V. (2023). *Practical Guide to Applied Conformal Prediction in Python.* Birmingham: Packt Publishing Ltd.

McMahon, L et al. (15 de 11 de 2023). How factories are deploying AI on production lines. *BBC*.

Mghouchi, Y et al. (2024). Multivariable Air-Quality Prediction and Modelling via Hybrid Machine Learning: A Case Study for Craiova, Romania. *Sensors*, 1532.

Mihalache, S et al. (2016). Development of ANFIS models for pm short-term prediction. Case study. *8th international conference on electronics, computers and artificial intelligence*, (págs. 1-6). Ploiesti.

Molnar, C. (2021). Aprendizaje automático interpretable, Capítulo 7: Interpretación de redes neuronales. Bookdown.

Molnar, C. (2023). *Introduction To Conformal Prediction With Python.* München: c/o MUCBOOK.

Montegro, Á. (2011). Análisis de series de tiempo. Bogotá. Pontificia Universidad Javeriana.

Müller, A et al. (2016). *Engineering Applications of Artificial Intelligence.* O'Reilly.

Oficina de Evaluación de Peligros para la Salud Ambiental OEHHA, Estado de California (2025). Obtenido de PM2.5: https://oehha.ca.gov/calenviroscreen/indicator/pm25

Oprea, M et al. (2017). Data mining and ANFIS application to particulate matter air pollutant prediction. A comparative study. *Proceedings of the 9th international conference on agents and artificial intelligence* (págs. 551-558). Porto: SciTePress.

Pasquier, A et al. (2017). Considering criteria related to spatial variabilities for the assessment of air pollution from traffic. *Transportation Research Procedia*, 3354-3369.

Ramírez, M et al. (2023). Implementación y evaluación de un modelo de pronóstico estadístico de pm.2.5 para el Valle de Aburrá. *Calidad de aire, cambio climático y salud pública.* (págs. 373-378). Santa Marta: Hill Consulting.

Reddy, V et al. (2017). Deep air: Forecasting air pollution in Beijing, China.

Rodas, M. (2024). *Este lunes 4 de marzo inicia la medida de Pico y Placa ambiental para vehículos de carga y volquetas en Medellín.* Medellín: Alcaldía de Medellín.

Rojas-Jímenez, K. (2022). Ciencia de Datos para Ciencias Naturales. Obtenido de https://bookdown.org/keilor_rojas/CienciaDatos/

Roy, S et al. (2018). Time series forecasting using exponential smoothing to predict the major atmospheric pollutants. *International conference on advances in computing, communication, control and networking*, (págs. 679-684). Gretaer Noida.

Organización Panamericana de la Salud (2024). *Leading causes of death and disease burden in the Americas.* Washington, D.C: World Health Organization.

Salas, L. F. (2022). Modelo de Predicción Material Particulado (PM2.5) en Bogotá. *Avances Investigación En Ingeniería*, 57-65.

Sistema De Alerta Temprana De Medellín Y El Valle De Aburrá. (2024). Obtenido de https://SIATA.gov.co/SIATA_nuevo/

Smyl, S. et al. (2025). Local and global trend Bayesian exponential smoothing models. *International Journal of Forecasting*, 111-127.

Unnikrishnan, R et al. (2019). Comparative study on the effects of meteorological and pollutant parameters on ANN modelling for prediction of SO2. *Applied Sciences*, 1394. doi:https://doi.org/10.1007/s42452-019-1440-1

Ventura, L et al. (2019). Forecast of daily pm2.5 concentrations applying artificial networks and holt-winters models. *Air Quality, Atmosphere Health*, 317-325.

Westerlund et al. (2014). Application of air quality combination forecasting to Bogota. *Atmospheric Environment* 89, 22-28.

Yeganeh, B et al. (2017). A satellite-based model for estimating pm2.5 concentration in a sparsely populated environment using soft computing techniques. *Environmental Modelling & Software*, 84-92.