

Article

Análisis de la correlación espacio-temporal de las partículas 2.5 (PM2.5) y su impacto en la salud pública utilizando IA

Sonia Meneses^{1*}, David Delgado², Miguel Reyes³

Citation: Meneses, S., Delgado, D. & Reyes, M. (2025). Análisis de la correlación espacio-temporal de las partículas 2.5 (PM2.5) y su impacto en la salud pública utilizando IA. Proceedings of the 2025 Academy of Latin American Business and Sustainability Studies (ALBUS), San Miguel, El Salvador. <https://doi.org/10.70469/ALBUS.15>



Copyright: © with the authors. This Open Access article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0).

Universidad Libre de Colombia, Colombia, sonial.menesesv@unilibre.edu.co
Universidad Libre de Colombia, Colombia, davidr-delgadom@unilibre.edu.co,
Universidad Libre de Colombia, Colombia, miguell-reyesb@unilibre.edu.co

Resumen: El material particulado de 2.5 micras (PM2.5) representa un riesgo significativo para la salud en Bogotá D.C., Colombia, especialmente en la localidad de Kennedy, donde se registran altas concentraciones. Este estudio entrena y compara modelos predictivos univariados utilizando herramientas de análisis de datos e inteligencia artificial (IA), y presenta formas de contribuir a la gestión localizada de la calidad del aire. La metodología incluyó el uso de cinco años de datos históricos, validación cruzada, técnicas de optimización de hiperparámetros y pronósticos de un mes, empleando librerías especializadas. Los hallazgos principales indican que, aunque los modelos univariados enfrentan limitaciones dada la complejidad del fenómeno, pueden ofrecer predicciones razonables, son de bajo costo y fáciles de implementar. La conclusión clave es que estos modelos constituyen herramientas viables para apoyar políticas ambientales específicas a corto plazo. Esto implica que los gestores y responsables de políticas públicas pueden utilizar estos modelos accesibles para una acción inmediata, mientras que se sugiere evaluar el costo-beneficio de implementar modelos multivariados más complejos en aplicaciones futuras.

Palabras clave: darts library; políticas ambientales; machine learning; time series; univariate models

1. Introducción

La contaminación del aire es un problema crítico en muchas ciudades del mundo, y Bogotá no es la excepción, según GreenPeace en el 2021 la ciudad excedía los niveles mínimos de PM2.5 en 400% en comparación con los parámetros de la Organización Mundial de la Salud (OMS) (Greenpeace, 2021). En particular, la localidad de Kennedy ha sido identificada como una de las zonas más críticas, superando los estándares del Índice Bogotano de Calidad del Aire (IBOCA). Mientras que la media anual es de aproximadamente 14.3 $\mu\text{g}/\text{m}^3$ en toda la ciudad, en la localidad de Kennedy se han registrado promedios cercanos a 20 $\mu\text{g}/\text{m}^3$ y se define el nivel de riesgo asociado a esa concentración según el Tabla 1 (IBOCA, 2025). Estas emisiones se atribuyen principalmente a factores como la carga logística, el tráfico vehicular y las actividades cotidianas de la población local (SEMARNAT, 2021).

El PM2.5, compuesto por partículas finas suspendidas en el aire, es especialmente peligroso debido a su capacidad para penetrar profundamente en los pulmones y entrar en el torrente sanguíneo (SEMARNAT, 2021). La exposición prolongada a estos contaminantes ha sido asociada con un aumento significativo en enfermedades respiratorias y cardiovasculares, especialmente en poblaciones vulnerables como niños y adultos mayores. El Instituto Nacional de Salud (INS) indica que el 89.36 % de las muertes por riesgos ambientales en Colombia están

relacionadas con la mala calidad del aire, mientras que el Departamento Nacional de Planeación (DNP) estima que los costos asociados ascienden al 1.5 % del Producto Interno Bruto (PIB), con alrededor de 8000 muertes anuales y más de 67.8 millones de casos de enfermedades respiratorias (Hernández, 2021).

Tabla 1. Atributos del IBOCA

Color	Nivel de riesgo por exposición a la calidad del aire	Nivel de actuación o respuesta	PM2.5, 12h($\mu\text{g}/\text{m}^3$)
Verde	Bajo	Prevención	0 – 12
Amarillo	Moderado	Prevención	12.1 – 35.4
Naranja	Regular	Alerta Naranja	35.5 – 55.4
Rojo	Alto	Alerta Roja	55.5 – 151.2
Morado	Peligroso	Emergencia	151.2 – 250.4
Morado	Peligroso	Emergencia	250.5 – 500.4

Fuente: Elaboración propia. Datos recuperados del IBOCA 2025.

Kennedy, al ser una de las localidades más densamente pobladas de Bogotá (SDP, 2018), enfrenta un reto particular en la gestión de su calidad del aire. La Universidad Nacional de Colombia ha destacado que la alta densidad poblacional, combinada con la proximidad a corredores industriales y vías principales, incrementa considerablemente las concentraciones de material particulado en esta zona, como indica el informe trimestral de enero a marzo (RMCAB, 2024). Asimismo, la Secretaría Distrital de Ambiente (SDA) ha identificado niveles críticos de PM2.5 durante las horas pico de tráfico (UNAL, 2024).

La ciudad dispone de la Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), que en alianza con el IBOCA forma una infraestructura administrada por la SDA (RMCAB, 2024) (SDA, 2022). Esta red, con más de 20 años de antigüedad, está compuesta por 19 estaciones automáticas que envían datos cada hora, distribuidas estratégicamente en las localidades para medir y reportar en tiempo real las concentraciones de contaminantes atmosféricos. Los datos recopilados son de acceso público y están disponibles a través de plataformas digitales gestionadas por la SDA. Sin embargo, los pronósticos generados a partir de esta información presentan un potencial limitado al no considerar características particulares de cada zona y al utilizar herramientas poco optimizadas, lo que reduce su capacidad para aportar valor a la propuesta de soluciones focalizadas para cada localidad de la ciudad. Dada la amplitud de la información, es crucial maximizar el aprovechamiento de los datos y las herramientas actuales para apoyar a las comunidades, generar alertas tempranas, definir soluciones estratégicas para los empresarios afectados y hacerlo de manera versátil, eficiente y en el menor tiempo posible.

En este contexto, el presente estudio tiene como objetivo entrenar diversos modelos predictivos univariados para estimar las concentraciones de PM2.5 en Kennedy. La investigación combina técnicas de análisis tradicionales y herramientas avanzadas de IA, aprovechando la disponibilidad de datos para respaldar la toma de decisiones de las partes interesadas. Además, se eligió un enfoque univariado para reducir la cantidad de información a procesar y considerar las dificultades de ampliar a otros tipos de contaminantes ambientales.

En los últimos años, el uso de técnicas avanzadas de IA y aprendizaje automático ha cobrado relevancia en el análisis ambiental, proporcionando herramientas más precisas y adaptativas para el monitoreo y el pronóstico de contaminantes atmosféricos, como el PM2.5. Un ejemplo destacado fue la investigación de “AlreySalud” en México, donde se utilizó IA para modelar la calidad del aire y pronosticar el nivel de riesgo (Ortuño Mojica et al., 2020). De forma similar, en Callao, Perú, se implementó un sistema de predicción basado en redes neuronales para monitorear PM10 y PM2.5 (Astocondor, 2024), mientras que, en Manizales, Colombia, se diseñó un prototipo de redes neuronales adaptado a variables locales, logrando un margen de error del 3 % (García Correa, 2021).

A nivel local, se utilizó regresión lineal múltiple para predecir el comportamiento de PM2.5 en Kennedy, Bogotá, con un error del 5.6 %, además se determinó que el monóxido de carbono por tráfico vehicular es el mayor influyente en la densidad de las emisiones (Salas Nuñez, 2022). Aunque estos estudios demuestran el potencial de la IA para predecir la calidad del aire, el rápido desarrollo de estas tecnologías mantiene abiertas las posibilidades de mejora. Sin embargo, persiste un vacío en la integración de estas herramientas con las entidades de control ambiental, lo que limita su capacidad para impactar en las comunidades afectadas.

Para el análisis se utilizó RStudio, mientras que el entrenamiento de modelos predictivos se realizó en Python. Se emplearon modelos tradicionales de predicción y algoritmos de machine learning, con el uso de librerías especializadas y herramientas específicas de cada uno. Además, se empleó la validación cruzada para garantizar la capacidad predictiva. Para optimizar los resultados de los mejores modelos se integraron métodos de optimización de hiperparámetros (Mango y Optuna) con el fin de anticipar picos de contaminación, alertar con suficiente antelación a la población y generar información valiosa para la gestión ambiental focalizadas en la

población de Kennedy destacando la importancia de sectorizar políticas ambientales particulares basadas en las características de cada área. Además de aprovechar los avances en IA para el análisis ambiental, se enfatiza la importancia de desarrollar soluciones innovadoras, ágiles y sostenibles para mitigar problemas de salud pública y mejorar la calidad de vida en comunidades vulnerables, como la de Kennedy, de forma focalizada.

2. Métodos

2.1 Área De Estudio

La localidad de Kennedy fue seleccionada debido a sus elevados niveles de PM_{2.5}, como lo indica el Observatorio Ambiental de Bogotá (OAB, 2017). La estación de monitoreo en esta área cuenta con sensores que han registrado un volumen significativo de datos históricos desde el 2015, lo suficiente para identificar patrones estacionarios en las concentraciones de PM_{2.5}. Ubicada en las inmediaciones del Parque Metropolitano Cayetano Cañizares, es una zona densamente poblada y presenta alta actividad vehicular, industrial y comercial (UNAL, 2024).

2.2 Recolección De Datos

Se recopilaron datos históricos obtenidos de la página oficial del IBOCA (IBOCA, 2025) desde el 1 de enero de 2018 (00:00) hasta el 31 de junio de 2024 (23:00), mediante descargas por períodos semestrales obteniendo 12 archivos que mediante scripts semiautomáticos se consolida una base de datos con 58,440 registros en los cuales se hayan la hora, la fecha, la concentración de PM_{2.5} en microgramos por metro cúbico ($\mu\text{g}/\text{m}^3$) y el índice IBOCA. Es importante señalar que los datos de julio y agosto de 2024 fueron excluidos del estudio debido al mantenimiento de la página IBOCA. Dado que estos registros nunca estuvieron disponibles, no se incluyeron en el entrenamiento, la validación ni el pronóstico, por lo que su ausencia no introduce sesgos en los resultados, ya que el corte de validación se estableció hasta el 31 de junio de 2024.

2.3 Procesamiento Y Transformación De Datos

El análisis exploratorio, la transformación y limpieza de los datos se llevaron a cabo utilizando RStudio e integrando librerías especializadas en manejo de datos y series de tiempo. Se emplearon DataExplorer, PMCMRplus y tseries para el análisis exploratorio; imputeTS para la imputación de datos faltantes, y forecast para el manejo de los datos atípicos. Además, se integraron las librerías broom, car, ggplot2, lubridate, Metrics, openxlsx, readxl, reshape2, tidyverse y xts para tareas generales como transformación de datos, ajustes estadísticos, evaluación de métricas, visualización avanzada y manejo de archivos. Véase el detalle específico del uso de cada librería en la Tabla 2.

Tabla 2. Detalles de las librerías utilizadas en RStudio.

Librería	Descripción	Uso específico
Generales		
Metrics	Evaluación de métodos de imputación de datos	Cálculo de métricas como MAE, RMSE, MAPE
reshape2	Transformación de estructuras de datos	Conversión entre formatos wide y long
xts	Manejo eficiente de series temporales indexadas	Estructuración de los datos por tiempo
openxlsx	Escritura y creación de archivos Excel	Exportación de resultados, generación de tablas para análisis o anexos
lubridate	Manipulación de fechas y horas	Extraer componentes de fechas (hora, día, mes, día de la semana, trimestre del año, etc.)
readxl	Lectura de archivos Excel (.xls y .xlsx)	Importación de datos desde fuentes externas en formato Excel
tidyverse	Conjunto de herramientas para manipulación y visualización de datos	Limpieza, transformación, filtrado, análisis general

car	Pruebas estadísticas y regresión	Mejoramiento de las agrupaciones en los gráficos de bigotes, histogramas y densidad
ggplot2	Visualización de datos	Mejoramiento de los gráficos en los casos pertinentes
broom	Conversión de objetos estadísticos a data frames	Organización de resultados y resúmenes para visualización
Análisis exploratorio		
PMCMRplus	Pruebas estadísticas no paramétricas (post-hoc)	Prueba Games-Howell
tseries	Herramientas de análisis de series temporales	Pruebas ADF y estacionariedad
DataExplorer	Exploración automatizada de datos	Resumen estadístico y visualizaciones de datos faltantes
Imputación		
imputeTS	Imputación de datos faltantes en series temporales	Imputar datos faltantes del PM2.5
Datos atípicos		
forecast	Modelado y predicción de series temporales	Manejo automatizado de valores atípicos con la función tsoutliers()

Fuente: Elaboración propia, 2025.

Para evaluar la estacionariedad de la serie temporal, se emplearon gráficos de descomposición estacional y la prueba Aumentada de Dickey-Fuller (ADF) sin descomponer la serie, la cual constituye una extensión de la prueba de Dickey-Fuller convencional que corrige posibles problemas de autocorrelación. La hipótesis nula plantea no estacionariedad, mientras que la hipótesis alternativa postula estacionariedad. Si el valor p es inferior a un umbral de significancia de 0.05, se rechaza la hipótesis nula, concluyéndose que la serie presenta características de estacionariedad (Said & Dickey, 1984).

Se identificaron 2882 datos faltantes (~4.93 %) distribuidos aleatoriamente, cumpliendo con el supuesto de Missing Completely At Random (MCAR), según el enfoque propuesto en Review for Handling Missing Data with special missing mechanism (Zhou, Aryal & Bouadjenek, 2024). Para imputarlos se utilizó la librería especializada imputeTS, se consideraron dos enfoques:

- Primero, Agrupación Por Hora Del Día. Se propone un supuesto en el que hay diferencias significativas entre los datos separados por hora y se verifica con una prueba de hipótesis Games-Howell (Games & Howell, 1976). La imputación se hizo para cada grupo con algoritmos de interpolación, Last Observation Carried Forward (LOCF), moving average (MA), mean y aleatorio, seleccionando el método con el menor Median Absolute Error (MDAE), el mayor p -valor de las pruebas de Kolmogorov-Smirnov (Massey, 1951) y coherencia continua de los datos.
- Segundo, Sin Agrupación. Se utiliza la librería imputeTS, que ofrece algoritmos especializados para la imputación de series temporales. Se aplicaron los métodos interpolación (en sus variantes lineal, spline y stine) para una imputación suavizada de datos; Kalman StructTS, el cual ajusta modelos estructurales que descomponen la serie en componentes como tendencia y estacionalidad, permitiendo una imputación basada en la estimación por máxima verosimilitud de dichos componentes y Kalman auto.arima, que identifica automáticamente la mejor estructura ARIMA para los datos y ajusta sus parámetros; y métodos avanzados como seadec y seasplit, cada uno combinado con variantes internas de imputación: interpolación, Kalman y media móvil, que permiten capturar estructuras estacionales de forma explícita (Moritz & Bartz-Beielstein, 2017). El mejor método se selecciona con los criterios del enfoque anterior.

Finalmente, se manejaron los datos atípicos de ambos conjuntos de datos. Para ello se usó la librería forecast, que clasifica e imputa automáticamente los outliers mediante descomposiciones estadísticas. Específicamente se usa la función tsoutliers(), este método divide la serie en componentes de tendencia, estacionalidad y ruido, identificando aquellos con desviaciones significativas en el componente de ruido. Posteriormente, los imputa utilizando interpolación o extrapolación, manteniendo la coherencia con la tendencia y estacionalidad de la serie para evitar sesgos en los análisis posteriores (Hyndman & Khandakar, 2008).

2.4 Entrenamiento, Validación, Evaluación Y Comparación De Modelos Predictivos

Antes de iniciar se dio reproducibilidad al experimento utilizando la librería Random, y específicamente la función Random.seed(0) que fija la semilla del generador aleatorio nativo de Python. Todo el proceso de entrenamiento, validación, evaluación y comparación de los modelos se realizó en Python 3.11.3 utilizando las librerías especializadas; Darts, dado que ofrece una estructura unificada y versátil para el entrenamiento y evaluación de múltiples modelos de predicción en series temporales, integrando tanto enfoques clásicos como redes neuronales modernas (Herzen et al., 2021). Prophet, utilizado directamente desde su librería oficial por su capacidad para modelar tendencias y estacionalidades de manera transparente e intuitiva, especialmente útil para datos con ciclos múltiples o discontinuidades (Taylor & Letham, 2017). Finalmente, se integró XGBoostRegressor desde su paquete específico, dada su eficiencia computacional, alto rendimiento en tareas de predicción, su capacidad para capturar relaciones no lineales y manejar automáticamente interacciones entre características (Chen & Guestrin, 2016). Aunque algunos modelos permiten incluir variables relacionadas con días festivos, en este estudio se optó por no considerarlas, dado que su incorporación añade complejidad y tiempo de procesamiento. El objetivo principal es contar con un modelo ágil que apoye la toma de decisiones ambientales, por lo que se priorizó la simplicidad y la eficiencia.

Se entrenan un total de ocho modelos de pronósticos, agrupados en dos categorías: métodos tradicionales y de machine learning. La selección de estos modelos se basó en su capacidad para capturar diferentes estructuras presentes en series temporales, como la tendencia, la estacionalidad y los patrones no lineales. Entre los métodos tradicionales se incluyen Prophet, Exponential Smoothing, ARIMA, Linear Regression y así como enfoques más recientes como Theta y FourTheta. Por otra parte, los modelos de machine learning fueron N-BEATS y XGBoostRegressor por su potencial para aprender relaciones complejas sin supuestos estrictos sobre la forma de la serie. Revise en detalle estos modelos en la Tabla 3.

Tabla 3. Detalles de los modelos utilizados en Python.

Categoría	Modelo	¿Por qué?	Biblioteca
Tradicional	Prophet	Modelo aditivo que descompone en tendencia, estacionalidad y días festivos, muy útil para series con ciclos múltiples y cambios de nivel. (Taylor & Letham, 2017)	Prophet
	Exponential Smoothing	Captura de forma simple los componentes de nivel, tendencia y estacionalidad, excelente como baseline y para datos con patrones estacionales. (Hyndman & Athanasopoulos, 2021)	Darts
	Arima	Modelo estadístico que combina auto regresión, diferenciación de medias móviles, muy efectivo para series estacionarias tras diferenciación. (Hyndman & Athanasopoulos, 2021)	
	Linear Regression Model	Permite modelar tendencias lineales y sirve como referencia sencilla antes de usar métodos más complejos. (Hyndman & Athanasopoulos, 2021)	
	Theta	combina dos variantes de la serie una centrada en tendencia y otra en curvatura para ofrecer pronósticos estables que capturan tendencia y estacionalidad de forma sencilla y robusta. (Assimakopoulos & Nikolopoulos, 2000)	
	FourTheta	Extiende al método Theta con cuatro linealizaciones distintas para mejorar la captura de patrones de tendencia y estacionalidad. (Spiliotis & Assimakopoulos, 2020)	
Machine Learning	N-beats	Red neuronal específica para series temporales, capaz de aprender patrones complejos de forma jerárquica sin necesidad de ingeniería manual. (Oreshkin et al., 2020)	

XGBoost Regressor	Modelo de boosting de gradiente que captura relaciones no lineales y alta interacción entre rezagos muy potente tras ajuste de hiperparámetros. (Chen & Guestrin, 2016)	XGBoost Regressor
----------------------	---	----------------------

Fuente: Elaboración propia, 2025.

El ajuste inicial de cada modelo se realizó sin optimización de hiperparámetros y un intervalo de confianza del 95%. Los ocho modelos fueron entrenados y validados con los datos históricos hasta el 31 de mayo del 2024, y la evaluación se dio con un intervalo de predicción en el horizonte de un mes en comparación con los datos originales, es decir, desde 01 al 30 de junio, considerando las dos perspectivas de imputación que se tomaron en cuenta, utilizando las métricas Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) y Symmetric Mean Absolute Percentage Error (SMAPE), junto con una valoración de consistencia visual basada en la alineación gráfica entre los pronósticos y los valores reales el cual tuvo mayor peso en la selección del modelo en este estudio.

Se compararon los ocho modelos desde sus métricas cuantitativas y de una clasificación base a las visualizaciones de los pronósticos, seleccionando los de mejor desempeño inicial para comenzar un proceso de optimización de hiperparámetros en Google Colab dado el tiempo de procesamiento requerido (aproximadamente 20 horas por modelo). Para ello se usó la librería Mango (Sandha, Aggarwal, Fedorov & Srivastava, 2020) para Prophet y XGBoostRegressor, mientras que para N-BEATS se recurrió a Optuna (Akiba, Sano, Yanase, Ohta & Koyama, 2019). Tras la optimización, los modelos ajustados se sometieron nuevamente a entrenamiento y validación cruzada, asegurando una comparación adecuada. Finalmente, se seleccionó el modelo XGBoostRegressor por presentar el mejor desempeño en el conjunto de evaluación, destacándose por su precisión predictiva, la coherencia con los patrones estacionales observados, sus métricas cuantitativas y la consistencia visual.

Adicionalmente a las métricas cuantitativas, se evaluó la consistencia visual de los pronósticos, un factor cualitativo al que se le otorgó un peso significativo en la selección final del modelo. Para convertir este criterio en una evaluación sistemática y reproducible, se diseñó una rúbrica con tres criterios específicos: la captura de la estacionalidad diaria, la tendencia semanal y la coherencia de la magnitud. A continuación, en la Tabla 4, se detalla la rúbrica empleada para asignar una calificación de 'Alta', 'Media' o 'Baja' a cada modelo.

Tabla 4. Rúbrica para la Evaluación de la Consistencia Visual de los Pronósticos.

Criterio de Evaluación	Nivel de Desempeño	Descripción Objetiva del Pronóstico
1. Captura de Estacionalidad Diaria	Alto	Replica consistentemente los dos picos de concentración diarios identificados: el matutino (aprox. 06:00-09:00) y el nocturno (aprox. 20:00-01:00). La sincronización y la altura relativa de los picos son similares a las de los datos reales. Muestra un patrón cíclico diario, pero puede fallar en capturar ambos picos, presentar un desfase horario significativo o subestimar/sobrestimar consistentemente su magnitud.
	Medio	El pronóstico es una línea plana o no muestra un patrón cíclico diario reconocible que corresponda a los datos de validación.
	Bajo	Refleja visiblemente la disminución de las concentraciones durante el fin de semana, especialmente el domingo, en comparación con los días de lunes a sábado.
2. Captura de Tendencia Semanal	Alto	Muestra alguna variabilidad a lo largo de la semana, pero no logra capturar de forma clara y consistente la caída del domingo o la tendencia general de los días laborables.
	Medio	

3. Coherencia de Magnitud y Rango	Bajo	El pronóstico no presenta un patrón semanal discernible; todos los días son tratados de manera similar.
	Alto	Los valores del pronóstico se mantienen dentro del rango general de los datos reales. Captura tanto los niveles base de concentración como la amplitud de los picos con una precisión razonable.
	Medio	El pronóstico es visualmente coherente en patrón pero presenta un sesgo constante (ej. siempre por debajo de los valores reales) o es demasiado "suave", perdiendo la variabilidad (picos y valles) de la serie.
	Bajo	Los valores pronosticados se encuentran en un rango completamente diferente al de los datos reales o se mantienen casi constantes, ignorando la dinámica de la serie.

Fuente: Elaboración propia (2025).

Para la evaluación del rendimiento de los modelos, se utilizó un conjunto de métricas complementarias. Si bien se reporta el MAPE por su interpretación intuitiva en términos porcentuales, se reconoce su conocida inestabilidad ante valores reales cercanos a cero, una condición frecuente a las características en esta serie temporal. Para contrarrestar esta limitación, se dio mayor énfasis al sMAPE (Symmetric Mean Absolute Percentage Error), que penaliza los errores de forma simétrica y es más robusto en estos escenarios, y al MAE y RMSE, que miden el error absoluto en las unidades originales de la serie ($\mu\text{g}/\text{m}^3$).

3. Resultados y discusión

3.1 Implicaciones En La Analítica De Datos

La prueba ADF se empleó para evaluar la presencia de una raíz unitaria en la serie temporal, esta mostró un p-valor de 0.01 menor que 0.05 que permite rechazar la hipótesis nula, indicando la ausencia de una tendencia estocástica, confirmando estacionariedad en la serie temporal y validando la viabilidad de emplear modelos predictivos. Este hallazgo se complementa con el gráfico de descomposición estacionaria que revela patrones diarios, semanales y mensuales, junto a una tendencia decreciente del último año, como se muestra en la Figura 1. Estos patrones refuerzan la naturaleza predecible y estacional de las concentraciones de PM2.5, destacando la relevancia de aplicar enfoques predictivos robustos, entendidos como aquellos capaces de capturar la estacionalidad (diaria, semanal y mensual), adaptarse a variaciones no lineales, manejar el ruido en los datos, mantener la coherencia contextual y ofrecer estimaciones útiles.

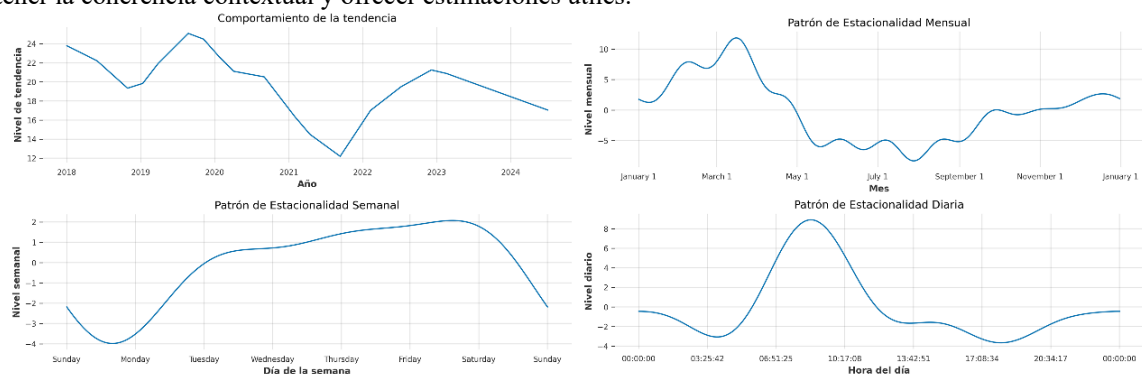


Figura 1. Componentes de descomposición estacional y tendencia.

Fuente: Elaboración propia (2024), con todos los datos de concentración de PM2.5 del conjunto de análisis.

Las concentraciones de PM2.5 durante la semana muestran patrones crecientes desde el lunes al sábado, siendo el día menos contaminado el domingo. Se observa un incremento crítico en las concentraciones diarias entre las 06:00 y las 09:00 horas, seguido de un segundo pico entre las 20:00 y las 01:00 horas, lo que sugiere que

el tráfico vehicular, las actividades comerciales y la rutina poblacional tienen un papel determinante en la contaminación del aire. En los patrones anuales, las concentraciones más altas se registraron durante los meses secos (diciembre a marzo), mientras que los meses lluviosos presentaron valores más bajos, lo cual es consistente con la investigación de Castro Camberos y Mancipe Díaz (Castro Camberos & Mancipe Díaz, 2023).

Para el primer enfoque de imputación, en el análisis se identifican diferencias significativas en las concentraciones de PM_{2.5} entre las 06:00 y las 13:00 horas, soportado por la prueba Games-Howell como se observa en la Figura 2.

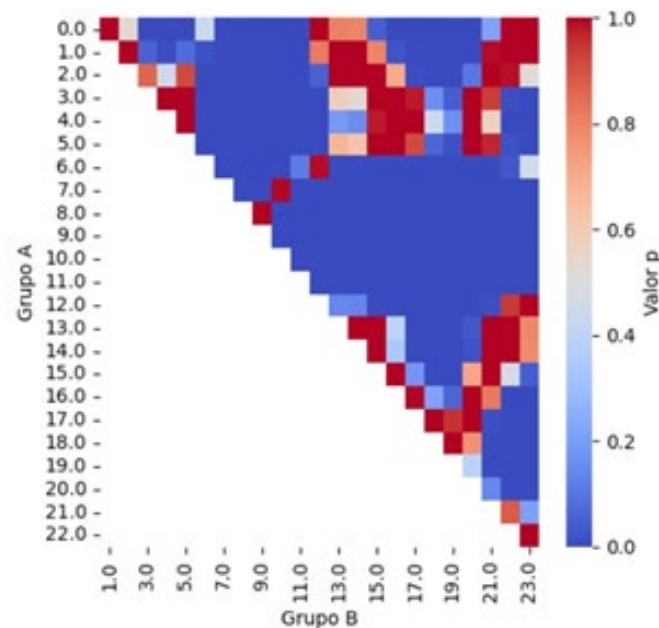


Figura 2. Mapa de calor de los valores p de la prueba Games-Howell por hora.
Fuente: Elaboración propia (2024).

Se agrupan los datos según la hora del día y al imputar se observó que los métodos aleatorios y de media presentaron inconsistencias significativas. En contraste, la media móvil se destacó como el método más robusto en la mayoría de las horas, sin embargo, para las horas 9, 10 y 11, se prefirió la interpolación, y para la hora 22, el método LOCF (imputación según la última observación disponible previa en la serie de tiempo) mostró mejor desempeño como se evidencia en la Tabla 5.

Tabla 5. Método de imputación seleccionado para cada hora del día.

Hora	Métricas de evaluación							Método seleccionado
	PH Kolmogorov Smirnov	MDAE						
	Mejor método	P-value	Interpolación	LOCF	MA	Media	Random	
0	MA	0.881	10.00	10.00	10.00	10.00	10.00	MA
1	MA	0.997	10.00	10.00	10.00	10.00	10.00	MA
2	MA	0.858	9.00	9.00	9.00	9.00	10.00	MA
3	MA	0.971	9.00	9.00	9.00	9.00	9.00	MA
4	MA	0.884	10.00	10.00	10.00	10.00	11.00	MA
5	MA	0.884	9.00	9.00	9.00	9.00	9.00	MA
6	MA	0.867	10.00	10.00	9.96	10.00	10.00	MA
7	MA	0.893	11.00	11.00	11.00	11.00	11.15	MA
8	MA	0.946	13.00	13.00	13.00	13.00	14.00	MA
9	MA	0.727	13.00	14.00	13.00	13.93	14.00	Interpolación
10	Interpolación	0.908	12.00	12.00	12.00	11.92	12.00	Interpolación
11	Interpolación	0.837	11.00	11.00	11.00	11.00	12.00	Interpolación

12	MA	0.570	9.18	9.00	9.00	9.00	10.00	MA
13	MA	0.886	9.00	9.00	8.97	9.00	9.00	MA
14	MA	0.539	10.00	10.00	10.00	10.00	11.00	MA
15	MA	0.704	9.00	9.00	9.00	9.00	9.00	MA
16	MA	0.734	9.00	9.00	9.00	9.00	9.00	MA
17	MA	0.849	9.00	9.00	9.00	9.00	9.00	MA
18	MA	0.862	8.00	8.00	8.00	8.00	8.00	MA
19	MA	0.894	9.00	9.00	9.00	9.00	9.00	MA
20	MA	0.982	9.00	9.00	9.00	9.00	9.00	MA
21	MA	0.944	9.00	9.00	9.00	9.00	9.00	MA
22	LOCF	0.960	8.20	9.00	8.83	9.00	9.00	LOCF
23	MA	0.959	10.00	10.00	10.00	10.00	10.00	MA

Fuente: Resultados de test kolmogorov y MDAE.

Nota. MDAE: Median Absolute Error. MA: Moving Average. LOCF: Last Observation Carried Forward.

En el conjunto imputado sin agrupación por horas, el método Seasplit- Interpolación destacó por su coherencia visual. Sin embargo, las pruebas de Kolmogorov-Smirnov aplicadas a todos los métodos revelaron diferencias significativas respecto a la distribución de densidades de probabilidad de los datos originales. Aunque la literatura no presenta un consenso claro sobre los métodos óptimos para imputar series temporales con estacionalidades diarias, semanales y mensuales, el primer enfoque de agrupamiento por hora mostró una mayor similitud con la distribución original, mientras que el enfoque global sobresalió en coherencia temporal, como se ilustra en la Figura 3.

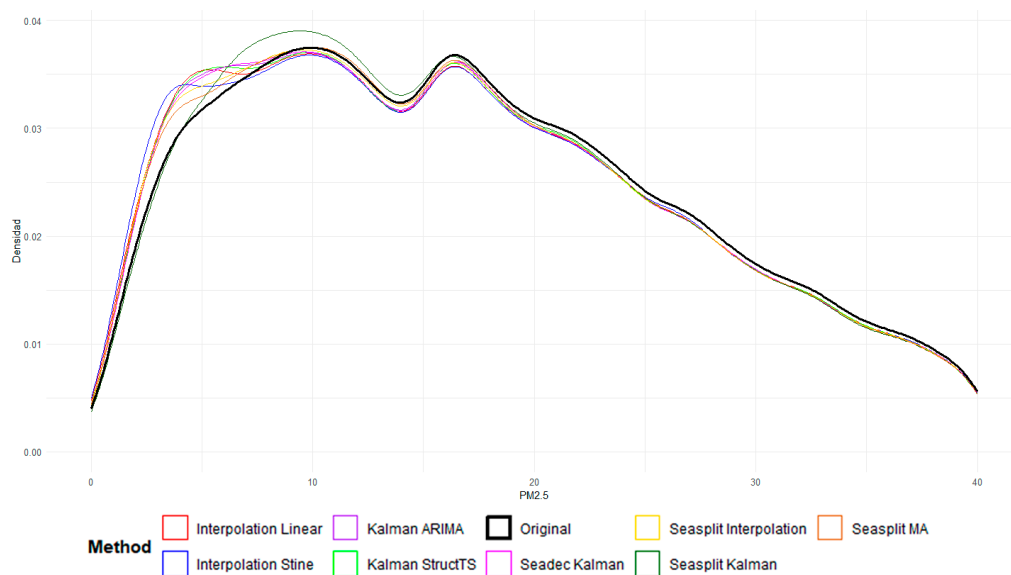


Figura 3. Comparación de densidades de las concentraciones de PM2.5 con imputación global respecto al original.

Fuente: Elaboración propia (2024). Eje X: concentraciones de PM2.5 $\mu\text{g}/\text{m}^3$.

En cuanto a los datos atípicos, se realizó una verificación minuciosa mediante la comprobación local de las concentraciones reportadas y su coherencia con las tendencias subyacentes en la curva diaria. Se concluyó que los datos atípicos detectados probablemente eran errores en el aparato de medición. Al separar estos mismos por las horas del día se logra mejorar la coherencia de la serie de tiempo como se muestra en la Figura 4. Análogamente se hizo lo mismo con ambos conjuntos de datos imputados.

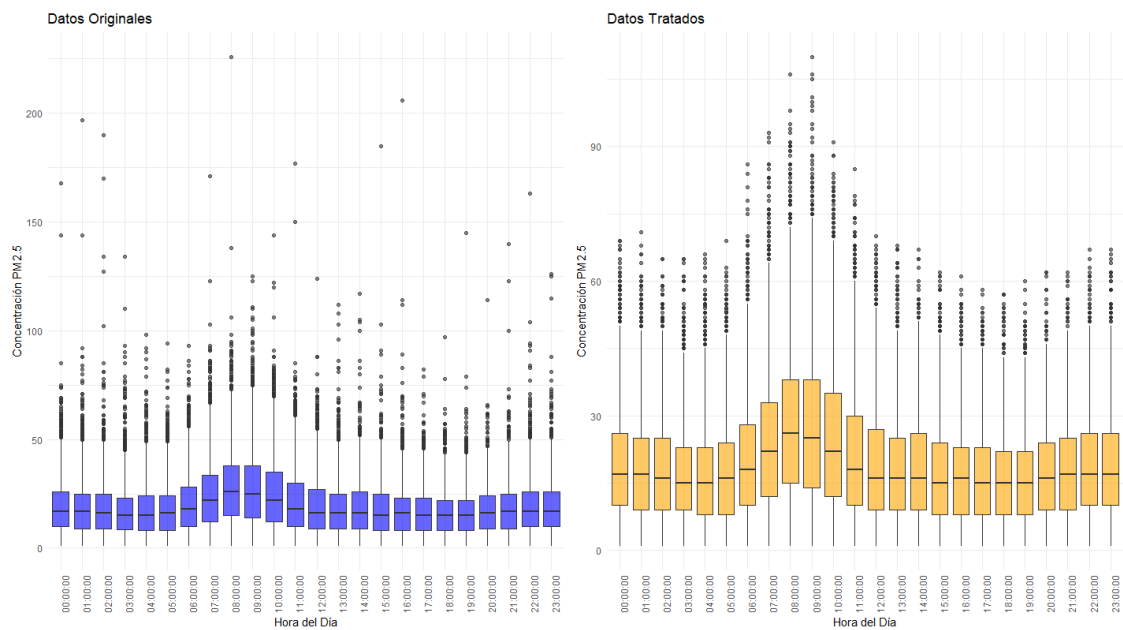


Figura 4. Comparación de ruido: datos originales vs datos tratados (reducción de atípicos).
Fuente: Elaboración propia (2024). Eje Y: concentraciones de PM2.5 µg/m³. Eje X: horas del día.

La elección de la librería Darts resultó estratégica para este estudio, ya que permitió comparar de forma eficiente diversos enfoques de pronóstico bajo condiciones experimentales homogéneas. Su compatibilidad tanto con modelos tradicionales como con métodos de aprendizaje automático facilitó su entrenamiento, validación y comparación coherente, reduciendo significativamente el tiempo de iteración entre pruebas. Esta flexibilidad operativa fue clave para evaluar el rendimiento relativo de seis modelos (Ver tabla 3.) diferentes con una base metodológica uniforme, lo que fortaleció la confiabilidad de los resultados presentados.

En cuanto a los modelos predictivos, los resultados iniciales no evidenciaron diferencias significativas entre los conjuntos imputados, lo que sugiere que los modelos son poco sensibles a los enfoques de imputación propuestos en esta investigación. Por lo que, a partir de aquí, solamente se tuvo en cuenta el primer grupo para el entrenamiento, validación y evaluación de los modelos.

Los modelos ARIMA y LinearRegression no lograron capturar adecuadamente los patrones estacionales, mientras que los modelos Exponential Smoothing, Theta y FourTheta lograron capturar únicamente la estacionalidad diaria, evidenciando una forma cíclica en sus predicciones. Por su parte, los modelos Prophet, N-BEATS y XGBoostRegressor fueron seleccionados con base en su rendimiento predictivo y coherencia, tanto en sus métricas cuantitativas (Tabla 6) y la rúbrica de consistencia visual (Tabla 4) como en las visualizaciones presentadas en la Figura 5.

Tabla 6. Métricas de evaluación general de los modelos entrenados.

Modelo	Métricas de evaluación				
	MAE	MAPE	RMSE	SMAPE	Consistencia visual
Sin optimización					
Prophet	6.74	163.93	8.32	70.14	Alta
ExponentialSmoothing	5.87	98.78	8.11	66.22	Media
ARIMA	5.84	113.86	7.95	66.09	Baja
LinearRegressionModel	10.78	293.59	11.79	90.33	Baja
Theta	6.50	80.07	9.32	82.36	Media
FourTheta	6.50	80.07	9.32	82.36	Media
N-BEATS	22.76	71.72	20.03	119.58	Alta
XGBoostRegressor	7.74	154.55	9.87	86.70	Alta

Fuente: Resultados de métricas de los pronósticos.
Nota. MAE: Mean Average Error. MAPE: Mean Absolute Percentage Error. RMSE: Root Mean Squared Error. SMAPE: Symmetric Mean Absolute Percentage Error.

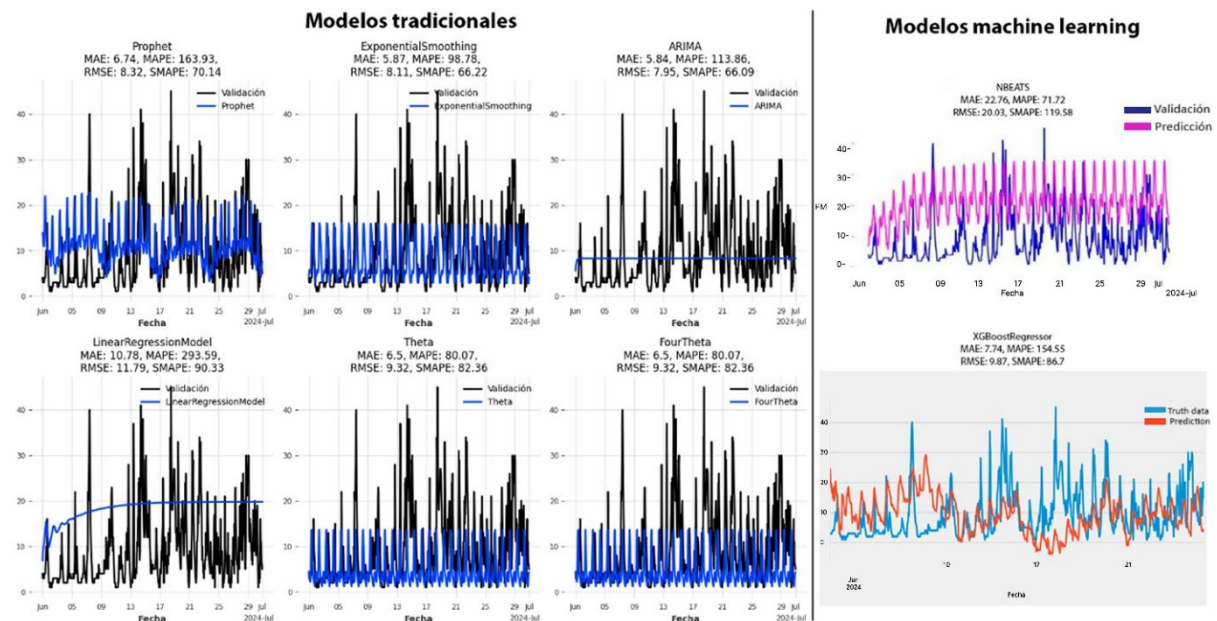


Figura 5. Pronóstico de los 8 modelos entrenados desde el 01 al 30 de junio del 2024.

Fuente: Elaboración propia (2024). Eje Y: concentraciones de PM_{2.5} $\mu\text{g}/\text{m}^3$ predicción vs real. Eje X: días del mes.

Para ajustar los hiperparámetros y mejorar los resultados de los modelos seleccionados se recurrió a librerías especializadas en optimización. Para Prophet y XGBoostRegressor se utilizó Mango, una librería ligera de optimización bayesiana diseñada para ejecutar múltiples evaluaciones en paralelo, priorizando la velocidad y la escalabilidad en la búsqueda de combinaciones óptimas (Sandha, Aggarwal, Fedorov & Srivastava, 2020). Mientras que para N-BEATS se empleó Optuna, una herramienta flexible que implementa búsqueda bayesiana avanzada mediante Tree-structured Parzen Estimator para encontrar los mejores parámetros (Akiba, Sano, Yanase, Ohta & Koyama, 2019). Los resultados obtenidos se muestran en la Figura 6 y las métricas en la Tabla 7, destacando al modelo XGBoostRegressor como el mejor debido a su capacidad para detectar y proyectar patrones estacionales diarios, semanales y mensuales, lo que le permite realizar pronósticos razonables para un horizonte temporal de un mes.

Tabla 7. Métricas de evaluación general de los modelos entrenados optimizados.

Modelo	Métricas de evaluación				
	MAE	MAPE	RMSE	SMAPE	Consistencia visual
Optimizados					
Prophet	6.76	164.56	8.33	70.19	Media
N-BEATS	2.20	99.00	5.20	82.31	Baja
XGBoostRegressor	6.52	154.52	7.90	69.29	Alta

Fuente: Resultados de métricas de los pronósticos.

Nota. MAE: Mean Absolute Error. MAPE: Mean Absolute Percentage Error. RMSE: Root Mean Squared Error. SMAPE: Symmetric Mean Absolute Percentage Error.

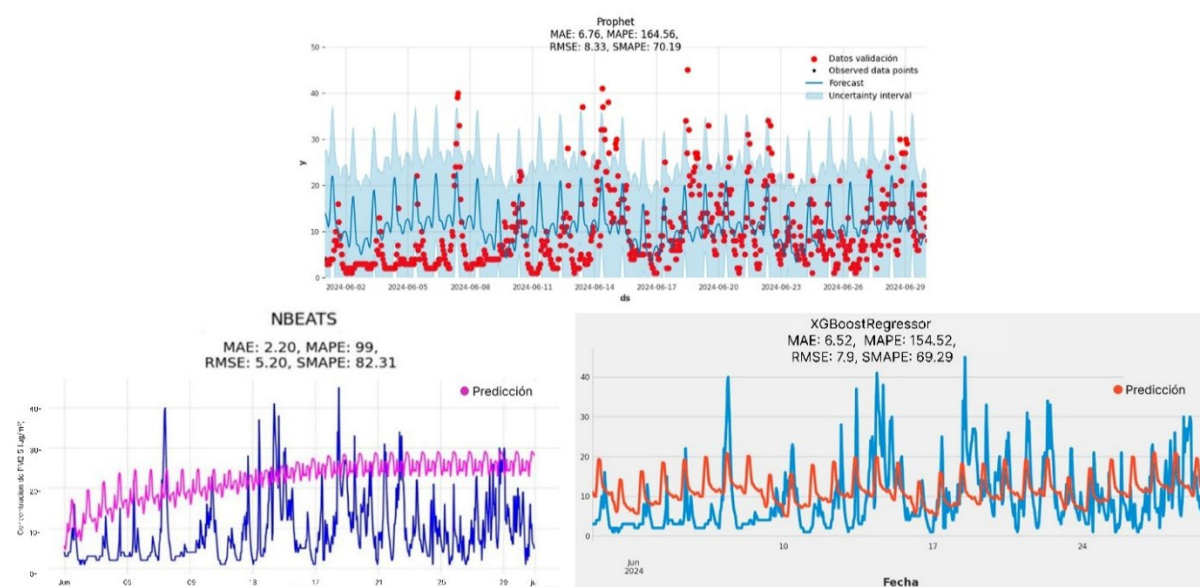


Figura 6. Modelos seleccionados optimizados en Google Colab.

Fuente: Elaboración propia (2024). Eje Y: concentraciones de PM_{2.5} µg/m³ predicción vs real. Eje X: días del mes.

El modelo XGBoostRegressor entrenado se presenta como una herramienta práctica y de bajo costo que aprovecha datos accesibles y técnicas de IA, logrando predicciones razonables para la toma de decisiones, demostrando la viabilidad de utilizar recursos limitados para abordar problemas críticos de calidad del aire. Su implementación impactaría directamente en la comunidad, incluso mientras se desarrollan modelos más complejos en paralelo y puede mejorar significativamente la capacidad de respuesta ante episodios de contaminación en corto y mediano plazo en la localidad de Kennedy, protegiendo a la población vulnerable y optimizando los recursos disponibles.

En la fase sin optimización, XGBoostRegressor mostró un desempeño competitivo y consistente visualmente, superando a modelos como ARIMA y Linear Regression en la mayoría de métricas evaluadas (Ver Tabla 6.). Tras la optimización de hiperparámetros, mejoró su rendimiento manteniendo alta consistencia visual y un desempeño equilibrado. Aunque el modelo N-BEATS optimizado obtuvo los valores más bajos de MAE y RMSE, presentó alta variabilidad en otras métricas y baja coherencia visual, lo que afectó su fiabilidad. Por estas razones, la selección final del modelo se fundamentó en un balance entre métricas cuantitativas y consistencia visual.

3.2 Implicaciones En La Salud Pública

Durante el análisis, se identificaron franjas horarias (entre las 06:00 y las 13:00 horas), así como los días de la semana (jueves a sábado) y los meses del año (septiembre a mayo) en las que es más probable que se registren picos de concentración de PM_{2.5} peligrosos para la salud. Esta evidencia refuerza la necesidad de considerar estos patrones en la formulación de políticas públicas focalizadas específicas y dinámicas, orientadas a mitigar el daño a la salud de la población.

Entre las estrategias derivadas de estos hallazgos se sugiere: restricciones temporales al tránsito en horarios o días críticos, mayor vigilancia ambiental durante los periodos mencionados, y el fortalecimiento de campañas de promoción del uso de transporte limpio. Además, podría permitir la emisión de alertas en tiempo real a través de una página web o aplicación móvil que se apoyen en modelos predictivos como los desarrollados en este estudio. Estas herramientas permitirían respaldar los mecanismos de control ambiental existentes en Bogotá, como la SDA, RMCAB, IBOCA, MinSalud y el OAB.

El modelo también puede contribuir a la planificación de rutas vehiculares en función de los niveles de contaminación, la justificación de proyectos de transición energética en camiones de carga, la adecuación de horarios laborales o escolares, y la emisión de alertas tempranas para la población. Este estudio resalta el valor de integrar herramientas emergentes en el diseño de políticas locales específicas, considerando la variabilidad espacial de los niveles de PM_{2.5}, influenciada por factores como el horario, el tráfico vehicular, la cultura local y otras condiciones propias de cada zona.

En consecuencia, los hallazgos permiten fundamentar la formulación de intervenciones localizadas, combinando acciones de corto plazo como la reorganización de actividades en horarios críticos junto a estrategias de mediano y largo plazo, como la inversión en infraestructura verde y zonas de amortiguamiento.

Este trabajo abordó el desafío de modelar las concentraciones de PM_{2.5} en la localidad de Kennedy con un enfoque univariado, logrando avances en: la comprensión del fenómeno y sus implicaciones en la salud pública, el análisis de la calidad de los datos ambientales con estacionalidades, la propuesta de políticas específicas para la localidad, y la promoción de modelos predictivos accesibles como parte de una gobernanza de datos más efectiva.

Finalmente, se identificaron retos significativos relacionados con la cobertura de los sensores actuales, cuya precisión no supera los 100 metros (TSI, 2023). Dada la extensión territorial de Kennedy, resulta difícil generalizar un estado representativo para toda la localidad. Por ello, se recomienda a la SDA evaluar los requerimientos establecidos en la normatividad vigente (SDA, 2022), considerando las características de cada sector y la validez de la información según el alcance de los medidores.

5. Conclusiones

La calidad de los datos representó un desafío significativo; los algoritmos de imputación seleccionados lograron preservar las tendencias estacionales manteniendo la coherencia temporal, mientras que la agrupación por hora del día resultó eficaz para minimizar el impacto de anomalías, permitiendo un tratamiento contextualizado de los valores atípicos sin comprometer las características estacionales.

Los resultados sugieren que el modelo XGBoostRegressor puede constituir una herramienta sólida para la toma de decisiones informada al proporcionar estimaciones suficientemente consistentes que fortalezcan la justificación técnica de proyectos orientados a la mitigación de la exposición al PM_{2.5}. Aunque los modelos multivariados podrían ofrecer una mayor precisión, este enfoque es ágil, de bajo costo y escalable.

La flexibilidad de la librería Darts fue clave en la comparación eficiente y coherente entre múltiples modelos de pronóstico (modelos tradicionales y machine learning), mejorando la calidad del análisis y optimizando recursos técnicos. Por su parte, la incorporación de las librerías Mango y Optuna permitió ejecutar búsquedas de hiperparámetros en paralelo y reproducible, agilizando el ajuste de los modelos.

La metodología empleada podría adaptarse a otras localidades que cuenten con información histórica suficiente. Este estudio sugiere que no es viable aplicar soluciones generalizadas para la gestión de la calidad del aire en Bogotá D.C. La agilidad que ofrecen las tecnologías de analítica y pronóstico permitiría responder con mayor rapidez a las diferencias territoriales.

Esta herramienta puede facilitar diversas acciones estratégicas para las entidades de control ambiental en Bogotá como priorización de intervenciones en zonas críticas, planificación de políticas públicas basadas en evidencia, y optimización recursos. Además de fortalecer la gobernanza de datos ambientales al integrar modelos predictivos accesibles en la toma de decisiones locales. Este enfoque se alinea con varios Objetivos de Desarrollo Sostenible (ODS), en particular el 9 (Industria, innovación e infraestructura), 11 (Ciudades y comunidades sostenibles) y 13 (Acción por el clima) (United Nations, 2015).

Finalmente, se resalta el desafío de mejorar la precisión mediante el entrenamiento de un modelo multivariado analizando la agilidad, los costo-beneficio y la velocidad de implementación, así como el perfeccionamiento de algoritmos de imputación que se adapten a patrones estacionarios. Este esfuerzo colectivo consolidará el uso de IA en la gestión de la calidad del aire y transformará la manera en que se aborda este problema en ciudades con desafío similares a Bogotá.

Referencias

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. arXiv. <https://doi.org/10.48550/arXiv.1907.10902>
- Assimakopoulos, V., & Nikolopoulos, K. (2000). The Theta model: A decomposition approach to forecasting. *International Journal of Forecasting*, 16(4), 521–530. [https://doi.org/10.1016/S0169-2070\(00\)00066-2](https://doi.org/10.1016/S0169-2070(00)00066-2)
- Astocondor, J. (2024). Sistema de medición y predicción con redes neuronales artificiales de la calidad del aire en la provincia del Callao [Tesis de licenciatura, Facultad de Ingeniería Eléctrica y Electrónica, Universidad Nacional del Callao]. <https://hdl.handle.net/20.500.12952/8626>
- Castro Camberos, Y. K., & Mancipe Díaz, J. X. (2023). Evaluación de la concentración del material particulado (PM_{2.5} y PM₁₀) para la gestión ambiental enfocada en la calidad del aire. Estudio de caso Parque Juan Amarillo - Siete canchas localidad de Engativá [Tesis de licenciatura, Universidad El Bosque]. Facultad de Ingeniería Ambiental, Universidad El Bosque.

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. arXiv. <https://doi.org/10.48550/arXiv.1603.02754>
- Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1(2), 113–125. <https://doi.org/10.3102/10769986001002113>
- García Correa, D. E. (2021). Prototipo de red neuronal artificial para el pronóstico de eventos críticos por partículas PM2.5 en el centro de la ciudad de Manizales [Tesis de licenciatura, Facultad de Ingeniería y Ciencias Básicas, Estadística Aplicada].
- Greenpeace. (2021). Según nuevos parámetros de la OMS, Bogotá necesita aire limpio ya. [Reporte de ONG – literatura gris]. <https://www.greenpeace.org/colombia/noticia/uncategorized/segun-nuevos-parametros-de-la-oms-bogota-necesita-aire-limpio-ya/>
- Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Van Pottelbergh, T., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., Kosczysz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., & Grosch, G. (2021). Darts: User-friendly modern machine learning for time series. arXiv. <https://doi.org/10.48550/arXiv.2110.03224>
- Hernández, J. (2021, 23 de agosto). Minsalud comprometido con la calidad del aire, Boletín de Prensa No.872 de 2021. Ministerio de Salud y Protección Social [Discurso principal]. <https://www.minsalud.gov.co/Paginas/Minsalud-comprometido-con-la-calidad-del-aire-.aspx>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/prophet.html>
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 1–22. <https://doi.org/10.18637/jss.v027.i03>
- IBOCA. (2025). Mapa de monitoreo. Índice Bogotano de Calidad del Aire y Riesgo en Salud – IBOCA. [Reporte técnico – literatura gris]. <http://iboca.ambientebogota.gov.co/mapa/>
- Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68–78. <https://doi.org/10.1080/01621459.1951.10500769>
- Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal*, 9(1), 207–211. <https://doi.org/10.32614/RJ-2017-009>
- OAB. (2017). Kennedy, la localidad de Bogotá con la más baja calidad de aire. Observatorio Ambiental de Bogotá. [Reporte técnico gubernamental – literatura gris]. <https://oab.ambientebogota.gov.co/kennedy-la-localidad-de-bogota-con-la-mas-baja-calidad-de-aire/>
- Ortuño Mojica, C., Díaz, L. R., Trujillo, C., & Coello, G. (2020). AlreySalud: Modelación de la calidad del aire con inteligencia artificial. *Saturdays.AI*. <https://saturdays.ai/2022/03/16/inteligencia-artificial-para-modelar-calidad-de-aire/>
- Oreshkin, B. N., Carpow, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1905.10437>
- RM CAB. (2024). Informe trimestral de calidad del aire de Bogotá. Enero-Febrero-Marzo 2024. Red de Monitoreo de Calidad del Aire de Bogotá. [Reporte técnico – literatura gris]. <http://rmcab.ambientebogota.gov.co/Pagesfiles/Informe%20trimestral%20EFM%202024.pdf>
- Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3), 599–607. <https://doi.org/10.1093/biomet/71.3.599>
- Salas Nuñez, L. F. (2022). Modelo de predicción material particulado (PM2.5) en la ciudad de Bogotá. *Avances Investigación en Ingeniería*, 19(2), 57–65. <https://doi.org/10.18041/1794-4953/avances.2.8574>
- Sandha, S. S., Aggarwal, M., Fedorov, I., & Srivastava, M. (2020). MANGO: A Python library for parallel hyperparameter tuning. arXiv. <https://doi.org/10.48550/arXiv.2005.11394>
- SDA. (2022). Herramientas para medir la calidad del aire en Bogotá. Secretaría Distrital de Ambiente. [Reporte técnico gubernamental – literatura gris]. <https://www.ambientebogota.gov.co/calidad-del-aire>
- SDP. (2018). Pasado, presente y futuro de Bogotá. Secretaría Distrital de Planeación. [Reporte técnico gubernamental – literatura gris]. https://www.sdp.gov.co/sites/default/files/bogota_pasado_presente_y_futuro.pdf
- Spiliotis, E., & Assimakopoulos, V. (2020). Forecasting with Theta models: A review and comparative study. *International Journal of Forecasting*, 36(1), 95–117.
- Taylor, S. J., & Letham, B. (2017). Forecasting at scale. *PeerJ*. <https://doi.org/10.7287/peerj.preprints.3190v2>
- TSI. (2023). BlueSky Air Quality Monitor Model 8143: Operation and maintenance manual. Thermo-Systems Engineering. [Manual técnico – literatura gris]. https://tsi.com/getmedia/a9299d7b-de37-4177-ab3b-488dfbfd2d07/BlueSky_Op_Maint_Manual_6013929?ext=.pdf

- UNAL. (2024). Kennedy y Bosa con más contaminación por falta de zonas verdes. Universidad Nacional de Colombia. [Reporte técnico – literatura gris]. <https://agenciadenoticias.unal.edu.co/detalle/kennedy-y-bosa-con-mas-contaminacion-por-falta-de-zonas-verdes>
- United Nations. (2015). Objetivos de desarrollo sostenible. <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>
- Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for handling missing data with special missing mechanism. arXiv. <https://doi.org/10.48550/arXiv.2404.04905>