Integration of Detection Techniques and Machine Learning to Improve Data Quality in Atmospheric Monitoring

Eladio Quintero[1], Jonathan González[1], Felisindo García[1], Edwin Collado, Ph.D. [1,2], Antony García[1], Yessica Sáez, Ph.D.[1,2,*]

[1] Universidad Tecnológica de Panamá, Panamá, {eladio.quintero1, jonathan.gonzalez14, felisindo.garcia, edwin.collado, antony.garcia, yessica.saez}@utp.ac.pa

[2] Centro de Estudios Multidisciplinarios en Ciencias, Ingeniería y Tecnología AIP (CEMCIT AIP, Panamá

*Corresponding author: yessica.saez@utp.ac.pa

Abstract*– Concentrations of particulate matter (PM_{10} and $PM_{2.5}$) in the air pose a significant risk to human health and the environment. Accuracy in the measurement of these pollutants is critical for effective air quality management, however, monitoring stations, especially low-cost ones, present errors and inconsistent data that affect the reliability of the analysis. In this study, different methods based on data science and machine learning (ML)are presented and compared to correct and improve the quality of PM measurements. The proposed approach includes an exploration data analysis to identify temporal patterns in air pollution specifically of PM, detection and removal of outliers using the interquartile range method, normalization and transformation of temporal variables, and implementation of a convolutional autoencoder model for missing data correction. The methodology was applied to a dataset collected by a monitoring station in Panama, and the results showed that the removal of outliers significantly reduced the distortion in the data, while the autoencoder achieved a moderate reconstruction of missing values, with a MAE of 0.1322 and a coefficient of determination R^2 of 0.5770. The findings suggest that the combination of statistical techniques and ML models allows to improve the reliability of PM monitoring data, providing more accurate information for environmental decision-making. In addition, this study opens new lines of research, such as the development of low-cost correction models for community stations, the analysis of the impact of meteorological events on particulate matter concentrations, and the comparison of pollution patterns in different urban environments. These advances will allow a better assessment of air pollution and contribute to the design of more effective strategies for its mitigation.

***Keywords*-- Data analysis, convolutional autoencoder, Machine Learning, particulate matter, environmental monitoring.**

Integración de Técnicas de Detección y Aprendizaje Automático para Mejorar la Calidad de Datos en Monitoreo Atmosférico

Eladio Quintero[1], Jonathan González[1], Felisindo García[1], Edwin Collado, Ph.D. [1,2], Antony García[1], Yessica Sáez, Ph.D.[1,2,*]
[1] Universidad Tecnológica de Panamá, Panamá, {eladio.quintero1, jonathan.gonzalez14, felisindo.garcia, edwin.collado, antony.garcia, yessica.saez}@utp.ac.pa
[2] Centro de Estudios Multidisciplinarios en Ciencias, Ingeniería y Tecnología AIP (CEMCIT AIP, Panamá
*Corresponding author: yessica.saez@utp.ac.pa

Resumen– Las concentraciones de material particulado (PM_{10} y $PM_{2.5}$) en el aire representan un riesgo significativo para la salud humana y el medio ambiente. La precisión en la medición de estos contaminantes es fundamental para una gestión efectiva de la calidad del aire, sin embargo, las estaciones de monitoreo, especialmente las de bajo costo, presentan errores y datos inconsistentes que afectan la confiabilidad del análisis. En este estudio, se presentan y comparan diferentes métodos basados en ciencia de datos y aprendizaje automático (machine learning, ML) para corregir y mejorar la calidad de las mediciones de PM. El enfoque propuesto incluye un análisis exploratorio de datos para identificar patrones temporales en la contaminación del aire específicamente de PM, la detección y eliminación de valores atípicos mediante el método del rango intercuartílico, la normalización y transformación de variables temporales, y la implementación de un modelo de autoencoder convolucional para la corrección de datos faltantes. La metodología fue aplicada a un conjunto de datos recopilados por una estación de monitoreo en Panamá, y los resultados mostraron que la eliminación de valores atípicos redujo significativamente la distorsión en los datos, mientras que el autoencoder logró una reconstrucción moderada de valores perdidos, con un MAE de 0.1322 y un coeficiente de determinación R^2 de 0.5770. Los hallazgos sugieren que la combinación de técnicas estadísticas y modelos de ML permite mejorar la confiabilidad de los datos de monitoreo de PM, proporcionando información más precisa para la toma de decisiones ambientales. Además, este estudio abre nuevas líneas de investigación, como el desarrollo de modelos de corrección de bajo costo para estaciones comunitarias, el análisis del impacto de eventos meteorológicos en las concentraciones de material particulado y la comparación de patrones de contaminación en diferentes entornos urbanos. Estos avances permitirán una mejor evaluación de la contaminación del aire y contribuirán al diseño de estrategias más efectivas para su mitigación.

Palabras clave— Análisis de datos, autoencoder convolucional, Machine Learning, material particulado, monitoreo ambiental.

I. Introducción

En las últimas décadas, la contaminación del aire se ha consolidado como un problema global de graves consecuencias para la salud humana y el medio ambiente. El deterioro progresivo de la calidad del aire, impulsado por el aumento de las emisiones contaminantes, representa un riesgo significativo para la población mundial. Según la Organización Mundial de la Salud (OMS), el 99% de la población respira aire insalubre, y la contaminación atmosférica es responsable de aproximadamente 7 millones de muertes prematuras anuales [1]. Entre los contaminantes más preocupantes se encuentra el material particulado fino, que por su tamaño diminuto puede penetrar el sistema respiratorio, alcanzar el torrente sanguíneo y provocar enfermedades cardiovasculares, accidentes cerebrovasculares, enfermedades pulmonares y cáncer [1], [2].

Los efectos de la contaminación del aire no se limitan a la salud humana; también están estrechamente vinculados al cambio climático. Muchos contaminantes atmosféricos y gases de efecto invernadero comparten fuentes comunes, como la quema de combustibles fósiles en vehículos y centrales eléctricas. Además, algunos contaminantes, como el metano y el carbono negro, poseen un alto potencial de calentamiento global [2]. Ante estos riesgos, la OMS ha actualizado sus directrices de calidad del aire con el objetivo de reducir la exposición a contaminantes y mitigar sus efectos en la salud pública [3], [4].

Para abordar este problema, es esencial contar con métodos precisos y confiables para evaluar la calidad del aire. Actualmente, se emplean diversas técnicas de monitoreo y control del material particulado, que incluyen estaciones de medición fijas, dispositivos personales de muestreo y modelos matemáticos de dispersión. Estas herramientas permiten predecir las concentraciones de contaminantes y evaluar su impacto [5], [6]. Sin embargo, la falta de infraestructura adecuada y el acceso limitado a datos confiables dificultan la gestión efectiva de la contaminación en muchas regiones del mundo [1], [6].

El material particulado en suspensión no solo representa una amenaza por su tamaño, sino también por su composición química. Estas partículas contienen metales tóxicos como plomo (Pb), cadmio (Cd), arsénico (As) y níquel (Ni), cuya detección y cuantificación han sido ampliamente estudiadas. Los métodos tradicionales, como la espectrometría de masas con plasma acoplado inductivamente (ICP-MS) después de digestión por microondas, han sido valorados por su alta precisión. No obstante, estas técnicas enfrentan desafíos en términos de costos y generación de residuos químicos. Como alternativa, técnicas como la fluorescencia de rayos X y la ablación láser acoplada a espectrometría de masas han surgido

como opciones prometedoras, ofreciendo un análisis más rápido y con menor impacto ambiental [7].

Para garantizar la calidad de los datos obtenidos en el monitoreo del aire, la Unión Europea ha implementado programas de control en laboratorios de referencia y estaciones de monitoreo. Estos esfuerzos han revelado que las mediciones realizadas por las redes nacionales pueden subestimar las concentraciones reales de material particulado, con diferencias de hasta un 11 %. Además, se ha observado que los analizadores automáticos presentan mayores incertidumbres en comparación con los métodos gravimétricos, lo que subraya la necesidad de aplicar factores de corrección adecuados [8].

El monitoreo de la calidad del aire a través de sensores de bajo costo ha generado un gran interés en la comunidad científica debido a su potencial para proporcionar datos con alta resolución espacial y temporal. Sin embargo, estos sensores enfrentan desafíos en términos de precisión y confiabilidad, lo que ha impulsado el desarrollo de métodos de corrección basados en aprendizaje automático. En este contexto, el uso de técnicas estadísticas y de ML ha demostrado ser una herramienta eficaz para mejorar la precisión de los datos ambientales. Modelos como los bosques aleatorios han permitido corregir errores en sensores de bajo costo, reduciendo significativamente los errores de medición. Asimismo, la detección de valores atípicos mediante clasificación espaciotemporal y análisis funcional de datos ha sido clave para identificar errores y eventos de contaminación inusuales, optimizando así la gestión de la calidad del aire [9]-[11]. Estos avances resaltan la importancia de combinar técnicas analíticas, modelos estadísticos y herramientas de inteligencia artificial para mejorar la precisión del monitoreo y garantizar la fiabilidad de los datos en la toma de decisiones ambientales [10]-[12]. En [13] y [14], se propusieron estrategias de control de calidad (QC) utilizando datos temporales homogéneos, datos meteorológicos diversos y características espaciotemporales. Se evaluaron modelos como regresión de vectores de soporte (SVR), regresión lineal múltiple (MLR) y bosques aleatorios (RFR), demostrando que RFR ofrecía el mejor desempeño al reducir el error en un 85 %. Además, se encontró que la inclusión de datos de estaciones meteorológicas automáticas mejoró significativamente la precisión del modelo.

En la misma línea, en [15] y [16] se evaluaron diferentes modelos de corrección para sensores de bajo costo de $PM_{2.5}$ en América del Norte. Se encontró que el uso de modelos que incorporan humedad relativa (HR) mejoró la precisión de las mediciones en concentraciones moderadas y altas, logrando una reducción significativa del error cuadrático medio (RMSE) de 8 a 3 µg/m³. Estos avances han sido implementados en plataformas de monitoreo ambiental como el Mapa de Incendios y Humo de AirNow, lo que demuestra su aplicabilidad en tiempo real.

Por otro lado, en [17] se abordó la mejora del rendimiento de sensores de bajo costo mediante modelos de corrección de bosque aleatorio aplicados a la detección de NO_2, PM_{10} y $PM_{2.5}$. Logrando reducir la incertidumbre en las mediciones, cumpliendo con los estándares de calidad del aire establecidos en la Unión Europea. En [18],[19],[20] se analizó la confiabilidad de sensores de bajo costo en entornos interiores y se aplicaron técnicas de ML para calibrar sus mediciones de PM_1, $PM_{2.5}$ y PM_{10}. Se observó que la precisión variaba en función del microambiente, lo que sugiere la necesidad de calibraciones específicas.

Por otra parte, en [21] y [22] se exploraron estrategias de corrección de sesgos en sensores de PM_{10}, comparando modelos tradicionales de transporte químico con enfoques de ML. Se encontró que el aprendizaje automático permitió estimaciones más precisas y confiables, facilitando su uso en sistemas de monitoreo atmosférico de gran escala. Estos estudios resaltan el impacto positivo del ML en la mejora de la calidad de datos obtenidos de sensores de bajo costo, permitiendo su integración efectiva en redes de monitoreo ambiental y en la toma de decisiones para la gestión de la calidad del aire.

Los métodos antes mencionados han demostrado ser altamente efectivos para calibrar y corregir datos en redes de monitoreo ambiental. Con base en esto, este trabajo busca desarrollar un modelo basado en ML para mejorar la precisión de los datos de material particulado PM_{10} y $PM_{2.5}$ obtenidos de estaciones de monitoreo. Dado que los sensores de bajo costo presentan limitaciones en la exactitud de sus mediciones debido a condiciones ambientales y fallos técnicos, es necesario aplicar métodos de corrección que permitan garantizar la fiabilidad de los datos. Para ello, se implementarán técnicas de detección y eliminación de valores atípicos, normalización y estandarización de los datos, así como la aplicación de un modelo de autoencoder convolucional para corregir mediciones erróneas o faltantes. En comparación con los métodos consultados en la literatura, el modelo de autoencoder convolucional propuesto en este estudio logró una reconstrucción moderada con un MAE de 0.1322 y un R^2 de 0.5770, representando una mejora cuantitativa significativa respecto a los datos sin procesar, y destacando su aplicabilidad en condiciones de infraestructura limitada.

El objetivo de este proyecto es optimizar la calidad de los datos de monitoreo ambiental y facilitar su uso en la toma de decisiones relacionadas con la gestión de la contaminación del aire. A través del uso de modelos avanzados, se espera reducir la incertidumbre en las mediciones y mejorar la precisión de los reportes sobre calidad del aire. Además, los resultados obtenidos podrían aplicarse en el desarrollo de herramientas accesibles para estaciones de monitoreo comunitarias, permitiendo la expansión de redes de medición de bajo costo con datos más confiables. De esta manera, este trabajo contribuye a fortalecer las estrategias de vigilancia ambiental y mitigación de la contaminación atmosférica.

El artículo está organizado de la siguiente manera: La sección II describe el diseño y metodología utilizada en el proyecto. La sección III los resultaos y discusión. La sección IV presenta las conclusiones.

II. Diseño y Metodología

A. *Formulación del problema*

La contaminación del aire, particularmente la causada por el material particulado, representa un desafío significativo para la salud pública y el medio ambiente. Aunque los sistemas de monitoreo atmosférico, como el desarrollado por [23], permiten medir las concentraciones de estos contaminantes en tiempo real, los datos recolectados a menudo presentan errores debido a mediciones corruptas, fallas técnicas o valores no deseados. Estos errores pueden distorsionar los resultados del análisis y dificultar la toma de decisiones informadas para la gestión de la calidad del aire.

Además, la falta de infraestructura adecuada y el acceso limitado a datos confiables en muchas regiones agravan este problema. Por lo tanto, se requiere una herramienta basada en ciencia de datos y ML que permita procesar, corregir y analizar los datos de manera efectiva, reduciendo así la incertidumbre en las mediciones y mejorando la precisión de los modelos predictivos. Esto nos hizo cuestionarnos la siguiente pregunta "*¿Cómo podemos desarrollar un modelo de ML que permita corregir errores en los datos de concentración de material particulado, mejorando así la calidad y confiabilidad de los sistemas de monitoreo atmosférico?"*

Para abordar el objetivo de este estudio, se diseñó una metodología estructurada que combina técnicas de ciencia de datos y aprendizaje automático para procesar, corregir y analizar los datos de concentración de material particulado (PM_{10} y $PM_{2.5}$). Esta metodología se divide en varias etapas clave, que incluyen el análisis exploratorio de datos (EDA), la detección y corrección de outliers, la normalización y estandarización de los datos, y la creación de un modelo de autoencoder convolucional para corregir datos faltantes o erróneos. Cada una de estas etapas fue cuidadosamente planificada y ejecutada con el fin de garantizar la calidad y confiabilidad de los datos, lo cual es fundamental para obtener resultados precisos y útiles en el análisis de la calidad del aire. En la Figura 1 se muestra de forma simplificada del flujo de trabajo implementado. En la fase de preprocesamiento, se aplicaron técnicas de normalización (MinMaxScaler) y codificación cíclica de variables horarias (utilizando funciones seno y coseno) para conservar la periodicidad de los datos temporales. También se utilizó codificación one-hot para los días de la semana. Posteriormente, se entrenaron tres modelos de autoencoder convolucional utilizando PyTorch. Estos modelos incluían capas Conv1D, MaxPooling1D y ReLU para la extracción de patrones temporales. El entrenamiento se realizó con el optimizador Adam y la función de pérdida MSELoss. Se empleó validación cruzada temporal y evaluación con métricas MAE y R^2 para seleccionar el mejor modelo. El Modelo 1 presentó los mejores resultados: MAE = 0.1322 y R^2 = 0.5770.

Fig. 1 Metodología estructurada para procesar, corregir y analizar los datos de concentración de material particulado.

B. *Fuente de Datos y Contexto del Estudio*

El estudio se basó en el sistema de monitoreo de contaminación atmosférica desarrollado por [23], el cual está compuesto por los sensores Nova PM SDS011 (partículas suspendidas), DHT22 (temperatura y humedad) y un microcontrolador ESP32 encargado de la recolección y transmisión de datos vía WiFi a una plataforma de visualización en línea para su análisis en tiempo real. Durante el periodo de validación de este sistema, se observó que parte de los datos recolectados presentaban errores. Estos errores podían deberse a mediciones corruptas, fallas en el funcionamiento del sistema o la captura de valores no deseados. Esta problemática motivó el desarrollo de una herramienta basada en ciencia de datos y ML para procesar y corregir adecuadamente los datos, reduciendo así la cantidad de errores en las mediciones.

Para garantizar la calidad de los datos y entrenar un modelo de manera confiable, se utilizó como referencia la estación de monitoreo AirSENCE. Esta estación ha demostrado una alta correlación con los equipos de referencia, proporcionando mediciones robustas y confiables de la calidad del aire [24]. La estación AirSENCE fue ubicada en la ciudad de Chitré, Panamá, y durante un periodo de monitoreo de 8 meses recopiló un total de 471,541 datos, con mediciones tomadas cada minuto. Estos datos servirán como base para crear un modelo de ML capaz de corregir errores en las mediciones de material particulado y ser funcional en estaciones de monitoreo de PM, mejorando así la precisión y confiabilidad de los datos.

C. *Herramienta y Entorno de Trabajo*

Para el análisis del conjunto de datos (Dataset), se utilizó la herramienta Jupyter Notebook junto con el lenguaje de programación Python. Esta combinación permitió implementar los códigos necesarios y mantener un registro detallado de cada paso realizado durante el proceso.

D. *Análisis Exploratorio de Datos*

El primer paso en el procesamiento de los datos fue realizar un análisis exploratorio de datos. Este proceso es esencial para

comprender la distribución, tendencias y correlaciones entre las variables del conjunto de datos. A través del EDA, se identificaron patrones en las concentraciones de material particulado y se detectaron posibles anomalías o errores en las mediciones.

Para ello, se llevaron a cabo los siguientes análisis:

- Distribución de los datos: Se generaron histogramas de las concentraciones de PM_{10} y $PM_{2.5}$ para evaluar si los datos seguían una distribución normal o presentaban sesgos.
- Tendencias temporales: Se visualizaron las concentraciones de material particulado a lo largo del tiempo para identificar patrones diarios, semanales y estacionales.
- Patrones de emisión: Se analizaron las variaciones horarias y diarias en la concentración de material particulado, considerando la influencia de actividades humanas, como el tráfico vehicular en horas pico y días laborales.
- Correlaciones entre variables: Se construyó una matriz de correlación para evaluar la relación entre la humedad relativa, la temperatura y las concentraciones de PM_{10} y $PM_{2.5}$, permitiendo identificar posibles relaciones lineales entre estas variables.

E. Detección y Corrección de Valores Atípicos

Tras el análisis exploratorio de datos, se procedió a la detección y corrección de valores atípicos (outliers), ya que estos pueden distorsionar el análisis y afectar la precisión de los modelos de ML. Para ello, se utilizó el rango intercuartílico (IQR), definido como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) [25]. Los valores atípicos se identificaron según el siguiente criterio:

$$IQR \in (Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR). \quad (1)$$

Los pasos seguidos fueron:

- Cálculo del IQR para las concentraciones de PM_{10} y $PM_{2.5}$.
- Identificación de valores atípicos mediante la aplicación del criterio del IQR.
- Eliminación de valores atípicos en los casos donde se consideraron errores de medición o datos corruptos.
- Visualización mediante box plots, permitiendo observar los datos antes y después del proceso de limpieza para verificar la eliminación de outliers.

F. Normalización y Estandarización de los Datos

Antes del modelado, fue necesario aplicar normalización y estandarización a los datos para garantizar una escala homogénea entre las variables y mejorar el rendimiento del modelo. Se implementaron los siguientes pasos:

- Normalización con MinMaxScaler a los valores de PM_{10} y $PM_{2.5}$ al rango [0, 1].
- Transformación cíclica de la variable "hora del día" utilizando una codificación basada en funciones seno y coseno para reflejar la periodicidad de la variable.
- Codificación one-hot- para los días de la semana. Esta técnica convirtió la variable categórica en variables binarias, facilitando su interpretación en los modelos de ML.

G. Creación del modelo de Corrección de Datos Faltantes

Para abordar la problemática de datos faltantes o erróneos en las mediciones de PM_{10} y $PM_{2.5}$, se desarrolló un modelo de autoencoder convolucional utilizando la biblioteca PyTorch. Este modelo, diseñado para trabajar con series temporales, fue entrenado con datos con valores faltantes simulados, con el objetivo de reconstruir las mediciones originales.

Finalmente, el desempeño del modelo se evaluó en un conjunto de prueba utilizando las siguientes métricas:

- Error Absoluto Medio (MAE): Mide la diferencia promedio entre los valores reales y los valores reconstruidos por el modelo.
- Coeficiente de Determinación (R^2): Evalúa la capacidad del modelo para explicar la variabilidad en los datos.

Estas métricas permitieron cuantificar la precisión del modelo en la reconstrucción de los datos, validando su efectividad en la corrección de datos faltantes en mediciones de material particulado.

III. Resultados y Discusión

En este estudio, se realizó un análisis exhaustivo de los datos de material particulado PM_{10} y $PM_{2.5}$ obtenidos de una estación de monitoreo ambiental, con el objetivo de comprender su distribución, identificar patrones temporales y aplicar técnicas de ML para la corrección y reconstrucción de datos. A continuación, se presenta un análisis detallado de los hallazgos para cada etapa.

A. Resultados del Análisis Exploratorio de Datos

El análisis exploratorio permitió identificar patrones en los datos de material particulado y evaluar su comportamiento temporal y correlación con variables meteorológicas.

Como se observa en la Figura 2, los histogramas revelan una distribución sesgada hacia la derecha, con la mayoría de las concentraciones ubicadas en el rango de 0-50 µg/m³. Sin embargo, la presencia de valores extremos de hasta 3500 µg/m³ sugiere la existencia de outliers que podrían distorsionar el análisis y afectar el rendimiento del modelo.

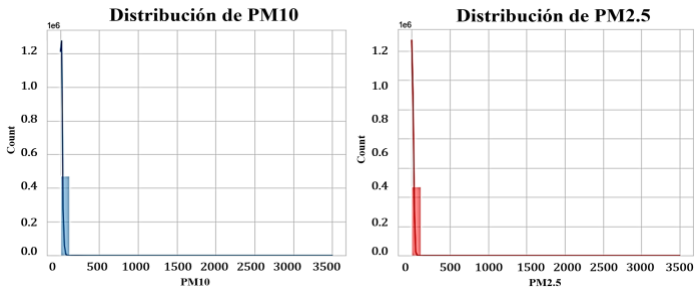

Fig. 2 Distribución de PM con la data completa.

En la Figura 3 se presentan los valores promedio de PM_{10} y $PM_{2.5}$ por hora del día (ambos valores resultaron con el mismo patrón). Se identificó un aumento de las concentraciones en las primeras horas de la mañana (6:00 - 8:00) y en la noche (20:00 - 22:00), lo que podría estar asociado a la actividad vehicular y las condiciones meteorológicas que favorecen la acumulación de contaminantes. En contraste, los niveles más bajos se registraron entre las 10:00 y 14:00, posiblemente debido a la mayor dispersión de los contaminantes por la radiación solar y la convección térmica.

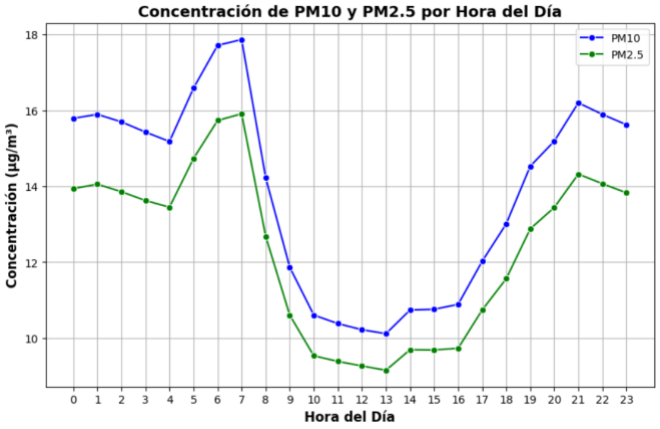

Fig. 3 Concentración de PM_{10} por hora del día.

La Figura 4 muestra la concentración promedio de PM_{10} y $PM_{2.5}$ por día de la semana (ambos valores resultaron con el mismo patrón). Se observa que los niveles más altos ocurren los lunes y viernes, posiblemente relacionados con el aumento del tráfico y la actividad industrial al inicio y final de la semana laboral. En contraste, los valores más bajos se registraron los miércoles y jueves, lo que podría indicar una reducción en la actividad emisora de contaminantes.

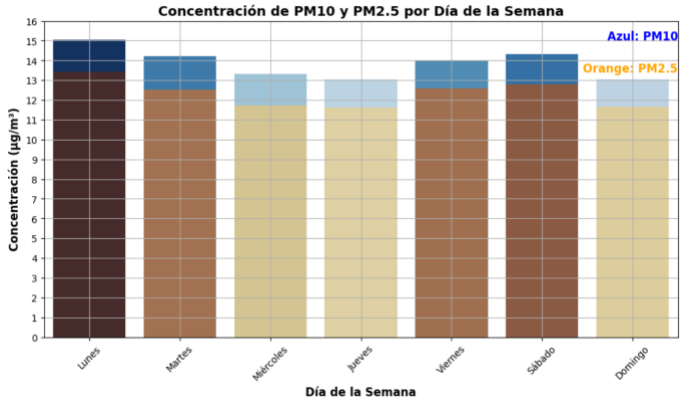

Fig. 4 Concentración de PM por día de la semana.

La matriz de correlación observada en la Figura 5 indica una relación muy fuerte (0.994) entre PM_{10} y $PM_{2.5}$, lo que sugiere que ambos contaminantes comparten fuentes similares. Sin embargo, la correlación entre estas variables y los factores meteorológicos fue débil, con valores cercanos a -0.07, lo que indica que temperatura y humedad no tienen un impacto significativo en las concentraciones de material particulado en este conjunto de datos. Por lo que no se utilizaron estas variables para entrenar el modelo ya que no aportan un peso significativo.

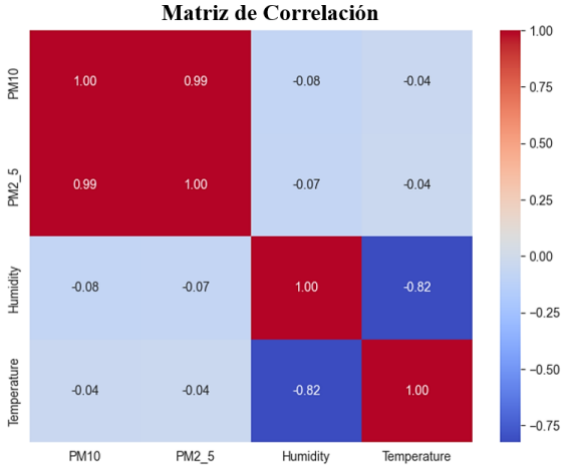

Fig. 5 Matriz de correlación entre variables.

B. Detección y Corrección de Outliers

Los valores atípicos identificados en los datos de PM pueden deberse a múltiples causas. Desde una perspectiva ambiental, eventos como intrusiones de polvo del Sahara, incendios forestales o quemas agrícolas pueden provocar elevaciones abruptas en las concentraciones de partículas. Técnicamente, fallos en los sensores, interferencias electromagnéticas o problemas de calibración también generan lecturas erróneas. Estas anomalías fueron tratadas mediante la eliminación de outliers usando el método del rango intercuartílico, lo cual permitió obtener distribuciones más centradas y facilitar el modelado confiable de la calidad del aire.

El análisis de outliers mostró que los valores extremos pueden influir negativamente en la interpretación de los datos y en el desempeño del modelo. Se identificaron 23,528 outliers en PM_{10} y 21,593 en $PM_{2.5}$ mediante el método del rango intercuartílico (IQR), los limites se prestan en la TABLA 1.

TABLA 1
LIMITES INFERIORES Y SUPERIORES PARA PM_{10} Y $PM_{2.5}$

IQR		
Variables	**Límite Inferior**	**Límite superior**
PM_{10}	-12.93	36.95
$PM_{2.5}$	-11.84	33.43

Tras la eliminación de estos valores, el número total de registros se redujo de 471,540 a 437,802. La comparación entre los boxplots antes y después de la limpieza mostrada en las Figuras 6 y 7 confirma que las distribuciones de PM_{10} y $PM_{2.5}$ se volvieron más centradas y simétricas, lo que mejora la calidad de los datos y reduce el impacto de valores anómalos en el modelado.

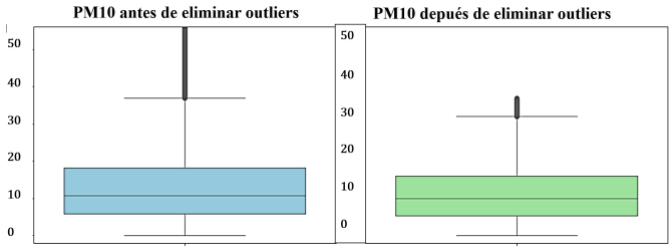

Fig. 6 Box plot de PM_{10} antes y después de outliers.

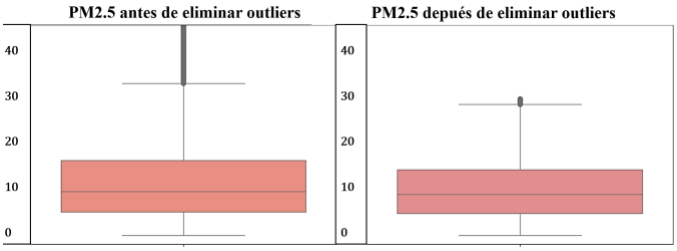

Fig. 7 Box plot de $PM_{2.5}$ antes y después de outliers.

A su vez en la Figura 8 se puede apreciar los histogramas mostrado en el cual después de eliminar los outliers, las distribuciones de PM_{10} y $PM_{2.5}$ se vuelven más centradas y simétricas, lo que facilita el análisis y el modelado.

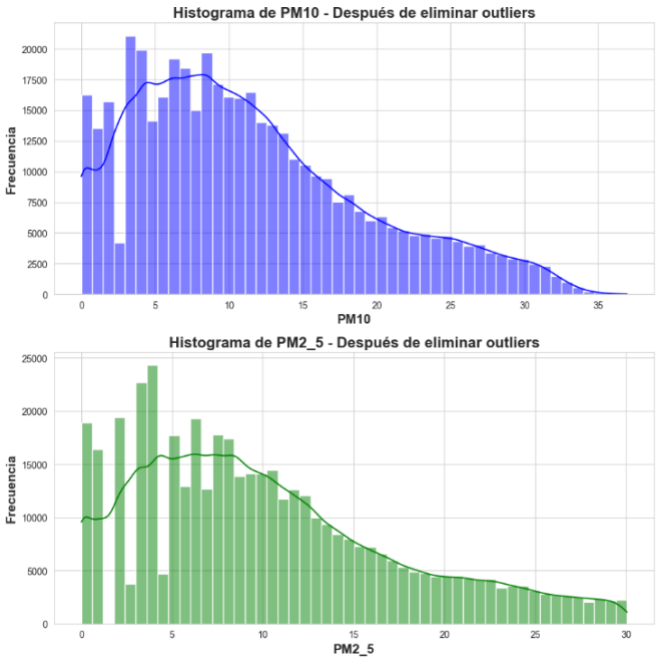

Fig. 8 Histograma de PM después de eliminar los outliers.

C. *Entrenamiento del Modelo de Autoencoder Convolucional*

Se evaluaron tres modelos de autoencoder convolucional entrenados en PyTorch para la reconstrucción de datos faltantes. Los resultados obtenidos se presentan en la TABLA 2.

TABLA 2
DESEMPEÑO DE LOS MODELOS DE AUTOENCODER CONVOLUCIONAL

MODELO	PERDIDA	MAE	R^2
MODELO 1	0.0824	0.1322	0.5770
MODELO 2	0.0854	0.3553	-0.3895
MODELO 3	0.0580	0.3643	-0.5838

El Modelo 1 mostró el mejor desempeño, con un MAE de 0.1322 y un coeficiente de determinación (R^2) de 0.5770, indicando una reconstrucción moderada de los datos. Esto se puede apreciar mejor en la Figura 9, donde se muestra una comparación de los valores reconstruidos y originales del modelo. En contraste, los Modelos 2 y 3 tuvieron valores negativos de R^2, lo que sugiere que no lograron generalizar bien los patrones en los datos.

Estos resultados sugieren que:

1. La arquitectura del modelo puede requerir ajustes para mejorar su capacidad de aprendizaje. Explorar redes más profundas o modificar los hiperparámetros podría optimizar el rendimiento.
2. La transformación de los datos es crucial, pero podrían requerirse estrategias adicionales, como reducción de dimensionalidad o extracción de características temporales más elaboradas.
3. El manejo de outliers sigue siendo un desafío, ya que su eliminación puede mejorar la distribución de los

datos, pero también afectar la capacidad del modelo para aprender patrones complejos.

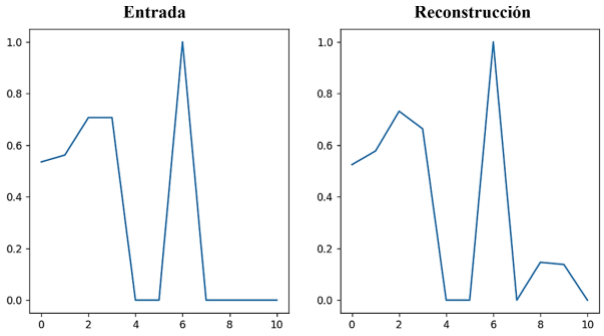

Fig. 9 Gráfica de reconstrucción de datos faltantes de PM del Modelo 1

IV. Conclusiones

La contaminación del aire, especialmente por material particulado de PM_{10} y $PM_{2.5}$, es un problema ambiental y de salud pública. Estas partículas pueden penetrar en el sistema respiratorio y causar enfermedades. A pesar de los avances en la monitorización de la calidad del aire, las estaciones de bajo costo presentan limitaciones en la precisión de sus mediciones, afectando la confiabilidad de los datos. Por ello, este estudio se enfocó en desarrollar y evaluar métodos de análisis y corrección de datos para mejorar la calidad de las mediciones obtenidas de estaciones de monitoreo ambiental.

Se implementó un enfoque basado en ciencia de datos y ML para procesar y corregir datos de PM_{10} y $PM_{2.5}$. El análisis exploratorio reveló patrones temporales en las concentraciones de material particulado, con niveles más altos en la mañana y la noche, posiblemente asociados con la actividad vehicular y condiciones meteorológicas. Además, se identificó una fuerte correlación entre PM_{10} y $PM_{2.5}$, pero una relación débil con variables como temperatura y humedad relativa.

Para mejorar la calidad de los datos, se eliminaron más de 23,000 outliers en PM_{10} y 21,000 en $PM_{2.5}$ mediante el rango intercuartílico. Esto optimizó la distribución de los datos y redujo la influencia de valores extremos en los modelos de ML. También se aplicaron técnicas de normalización y transformación cíclica a las variables temporales.

Para corregir datos faltantes, se entrenó un autoencoder convolucional en PyTorch, obteniendo un MAE de 0.1322 y un R^2 de 0.5770. Sin embargo, otros modelos evaluados mostraron valores negativos de R^2, indicando dificultades en la generalización de patrones. Estos resultados sugieren que la arquitectura del modelo puede optimizarse mediante ajustes en los hiperparámetros y la incorporación de nuevas características.

El desarrollo de estas estrategias permitirá mejorar la precisión del monitoreo de la calidad del aire y optimizar herramientas de análisis para la detección de eventos de contaminación. La combinación de técnicas estadísticas y modelos de inteligencia artificial representa una solución prometedora para garantizar datos más confiables, facilitando la toma de decisiones basadas en evidencia y contribuyendo a la mitigación de la contaminación atmosférica y sus efectos en la salud pública.

IV. Trabajos Futuros

En futuras investigaciones, se espera desarrollar un modelo de corrección de bajo costo para estaciones comunitarias y analizar con mayor profundidad el impacto de eventos meteorológicos en las concentraciones de PM. También sería relevante comparar patrones de contaminación en distintas zonas urbanas, integrando datos de múltiples estaciones y complementándolos con información satelital.

Agradecimiento

Los autores agradecen el apoyo de la Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT) bajo la subvención No. 157-2023 / FID23-078 y del Sistema Nacional de Investigadores (SNI) de Panamá. También, agradecen al grupo de investigación ITSIAS – UTP.

Referencias

[1] "¿Cómo se mide la calidad del aire?" UNEP. [Online]. Available: https://www.unep.org/es/noticias-y-reportajes/reportajes/como-se-mide-la-calidad-del-aire.

[2] Banco Mundial, "El cambio climático y la contaminación atmosférica," World Bank. [Online]. Available: https://www.bancomundial.org/es/news/feature/2022/09/01/what-you-need-to-know-about-climate-change-and-air-pollution.

[3] IQAir Staff Writers, "Las nuevas pautas de calidad del aire de la OMS salvarán vidas," Iqair.com, Dec. 2021. [Online]. Available: https://www.iqair.com/es/newsroom/2021-who-air-quality-guidelines.

[4] "Guías actualizadas de la OMS sobre la calidad del aire y sus implicancias para los países latinoamericanos," Salud sin Daño. [Online]. Available: https://lac.saludsindanio.org/recursos/guias-actualizadas-de-la-oms-sobre-la-calidad-del-aire-y-sus-implicancias-para-los-paises.

[5] O. A. US EPA, "Conceptos básicos sobre el material particulado (PM, por sus siglas en inglés)," 2018.

[6] C. A. H. Suárez, "Diagnóstico y control de material particulado: partículas suspendidas totales y fracción respirable PM10," *Luna Azul*, no. 34, pp. 195–213, 2012.

[7] M. Ogrizek, A. Kroflič, and M. Šala, "Critical review on the development of analytical techniques for the elemental analysis of airborne particulate matter," *Trends Environ. Anal. Chem.*, vol. 33, no. e00155, p. e00155, 2022.

[8] IES: Institute for Environment and Sustainability, *A quality assurance and control program for PM2.5 and PM10 measurements in European air quality monitoring networks*. Publications Office of the European Union, 2011.

[9] J. Reina and O. Olaya, "Ajuste de curvas mediante métodos no paramétricos para estudiar el comportamiento de contaminación del aire por material particulado PM10," *Rev. EIA*, vol. 9, no. 18, pp. 19–31, 2012.

[10] E. M. Considine, C. E. Reid, M. R. Ogletree, and T. Dye, "Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network," *Environ. Pollut.*, vol. 268, no. 115833, p. 115833, 2021.

[11] V. M. van Zoest, A. Stein, and G. Hoek, "Outlier detection in urban air quality sensor networks," *Water Air Soil Pollut.*, vol. 229, no. 4, 2018.

[12] J. Martínez Torres et al., "A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland," *Mathematics*, vol. 8, no. 2, p. 225, 2020.

[13]H.-J. Kim et al., “Spatiotemporal approaches for quality control and error correction of atmospheric data through machine learning,” *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–12, 2020.
[14]W.-C. V. Wang, S.-C. C. Lung, and C.-H. Liu, “Application of machine learning for the in-field correction of a PM2.5 low-cost sensor network,” *Sensors (Basel)*, vol. 20, no. 17, p. 5002, 2020.
[15]K. K. Barkjohn, B. Gantt, and A. L. Clements, “Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor,” *Atmos. Meas. Tech.*, vol. 14, no. 6, pp. 4617–4637, 2021.
[16]B. Nilson, P. L. Jackson, C. L. Schiller, and M. T. Parsons, “Development and evaluation of correction models for a low-cost fine particulate matter monitor,” *Atmos. Meas. Tech.*, vol. 15, no. 11, pp. 3315–3328, 2022.
[17]T. Bush et al., “Machine learning techniques to improve the field performance of low-cost air quality sensors,” *Atmos. Meas. Tech.*, vol. 15, no. 10, pp. 3261–3278, 2022.
[18]H. Chojer et al., “Can data reliability of low-cost sensor devices for indoor air particulate matter monitoring be improved? – An approach using machine learning,” *Atmos. Environ. (1994)*, vol. 286, no. 119251, p. 119251, 2022.
[19]I. N. K. Wardana, J. W. Gardner, and S. A. Fahmy, “Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder,” *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16129–16154, 2022.
[20]T. T. Lai, T. P. Tran, J. Cho, and M. Yoo, “Noise-tolerant data reconstruction based on convolutional autoencoder for wireless sensor network,” *Appl. Sci. (Basel)*, vol. 13, no. 18, p. 10090, 2023.
[21]J. Jin et al., “Machine learning for observation bias correction with application to dust storm data assimilation,” *Atmos. Chem. Phys.*, vol. 19, no. 15, pp. 10009–10026, 2019.
[22]Q. Guo et al., “Correction of light scattering-based total suspended particulate measurements through machine learning,” *Atmosphere (Basel)*, vol. 11, no. 2, p. 139, 2020.
[23]E. Quintero, J. González, F. García, Y. Sáez, and E. Collado, “IoT-based system prototype for particulate matter monitoring in the city of Chitre, Panama,” in *Proc. 21st LACCEI Int. Multi-Conf. Eng., Educ. Technol.*, Buenos Aires, Argentina, Jul. 2023.
[24][24] “AirSENCE - real time ambient micro air quality monitor,” AirSENCE | Breathe Safe, Breathe Easy. [Online]. Available: https://airsence.com/
[25]J. González, E. Quintero, F. García, A. García, Y. Sáez, y E. Collado, “Data science-based tool to reduce measurement errors in atmospheric monitoring systems,” in *Proc. 22nd LACCEI Int. Multi-Conf. Eng., Educ. Technol.*, 2024.