Análisis de la Calidad del Aire en Municipios Colombianos: Identificación de Patrones y Factores Asociados a la Contaminación

1st Luisa Fernanda Carpintero Gabanzo
dept. Ingeniería
Universidad de La Sabana
Bogotá, Colombia
luisacarga@unisabana.edu.co

2nd David Eduardo Lopez Jimenez
dept. Ingeniería)
Universidad de La Sabana
Bogotá, Colombia
davidloji@unisabana.edu.co

3rd María José Melo Gonzalez
dept. Ingeniería
Universidad de La Sabana
Bogotá, Colombia
mariamelgo@unisabana.edu.co

***Abstract*—La contaminación del aire es un problema ambiental crítico que afecta la salud pública y la calidad de vida en diversas regiones del país. Este estudio analiza los niveles de contaminación atmosférica en múltiples municipios, identificando patrones, fuentes principales y variaciones geográficas. A través del análisis de datos obtenidos de estaciones de monitoreo ambiental, se examinan los niveles de material particulado (PM10 y PM2.5), dióxidos de azufre (SO) y otros contaminantes clave. Los resultados evidencian diferencias significativas entre zonas urbanas e industriales frente a áreas rurales, así como el impacto de factores climáticos y socioeconómicos.**

***Index Terms*—Contaminación del aire, material particulado, calidad del aire, salud pública, factores climáticos, impacto socioeconómico.**

I. Introduccion

La contaminación del aire representa uno de los principales desafíos ambientales y de salud pública a nivel mundial. En Colombia, diversas regiones enfrentan niveles preocupantes de contaminación atmosférica, derivados de fuentes industriales, el transporte vehicular y la quema de combustibles fósiles. La exposición prolongada a contaminantes como el material particulado (PM10 y PM2.5), los óxidos de nitrógeno (NO) y el dióxido de azufre (SO) se ha asociado con un aumento en enfermedades respiratorias, cardiovasculares y otros efectos adversos en la salud humana.

El impacto de la contaminación del aire varía significativamente entre los municipios del país, influenciado por factores geográficos, climáticos y socioeconómicos. Mientras que en zonas urbanas e industriales se registran niveles elevados de contaminantes debido a la alta densidad vehicular y las actividades productivas, en áreas rurales la problemática suele estar relacionada con la quema de biomasa y la deforestación.

Este estudio tiene como objetivo analizar la calidad del aire en distintos municipios, identificando patrones de contaminación y sus posibles causas. A partir de datos obtenidos de estaciones de monitoreo ambiental, se evaluará la distribución espacial y temporal de los principales contaminantes atmosféricos, permitiendo comprender mejor los riesgos asociados y la necesidad de estrategias de mitigación efectivas.

II. Metodología

A. comoprensión del negocio

La contaminación del aire es un problema ambiental y de salud pública que afecta a numerosas ciudades en el mundo. Según la Organización Mundial de la Salud (OMS), la contaminación del aire causa aproximadamente 7 millones de muertes prematuras cada año, debido a su relación con enfermedades respiratorias, cardiovasculares y cáncer de pulmón [1]. En el caso de Colombia, las emisiones de material particulado (PM10 y PM2.5) en diversas ciudades han superado los niveles recomendados por la OMS, lo que ha generado preocupaciones sobre la calidad del aire y sus efectos en la población [2].

Dado este contexto, el presente estudio busca analizar la calidad del aire en distintos municipios del país a través de un enfoque basado en minería de datos. Aplicando la metodología CRISP-DM, se pretende transformar los datos disponibles en conocimiento útil para la toma de decisiones ambientales. Como lo indica Han et al., la minería de datos permite descubrir patrones y relaciones en grandes volúmenes de información, lo que resulta fundamental para entender fenómenos complejos como la contaminación atmosférica [3].

El objetivo principal de esta fase es definir claramente las preguntas que guiarán el análisis. En este caso, se plantea responder: ¿Cuáles son los principales contaminantes del aire en los municipios analizados? ¿Cómo han variado sus niveles a lo largo del tiempo? ¿Existen patrones que permitan predecir episodios críticos de contaminación? La identificación de estas problemáticas permitirá desarrollar modelos adecuados en las siguientes etapas del proceso.

Asimismo, este estudio es relevante para la formulación de políticas públicas, dado que la evidencia científica sugiere que la reducción de emisiones contaminantes puede mejorar la salud pública y reducir los costos asociados a enfermedades respiratorias [4]. Por tanto, este análisis no solo tiene un propósito exploratorio, sino también un impacto potencial en la gestión ambiental y en la salud de la población.

B. Comprension de los datos

La fase de comprensión de los datos dentro del marco CRISP-DM es esencial para identificar la estructura y características de los datos antes de su procesamiento y modelado.

En este estudio, se implementaron diversas técnicas de exploración y análisis utilizando bibliotecas de Python como Pandas y NumPy para manejar y estructurar los datos, así como Matplotlib y Seaborn para la visualización gráfica.

1) Identificación y Estructura de los Datos: Inicialmente, se realizó una revisión de los tipos de datos disponibles en la base de datos utilizando las funcionalidades de Pandas y NumPy, lo cual permitió clasificar las variables en numéricas y categóricas. Posteriormente, se utilizó el método .head() para inspeccionar los cinco primeros registros y verificar la coherencia en el formato de los datos.

Para comprender el significado de cada variable, se consultó la documentación oficial de la base de datos, donde se encuentra una tabla con la descripción de cada campo. Esto facilitó la correcta interpretación de los valores registrados, asegurando que el análisis posterior fuera adecuado y alineado con la información contenida en la fuente de datos.

2) Visualización y Exploración Gráfica: Con el fin de analizar la distribución y características de los datos, se implementaron diversas representaciones gráficas según el tipo de variable:

- **Gráficos para datos cualitativos y categóricos:** Se utilizaron diagramas de barras y gráficos de pastel para visualizar la distribución de categorías dentro de las variables cualitativas.
- **Gráficos para datos cuantitativos:** Se generaron histogramas y boxplots para analizar la distribución de las variables numéricas y detectar posibles valores atípicos.
- **Análisis de relaciones entre variables:** Se utilizó un gráfico de dispersión (scatter plot) para evaluar la relación entre diferentes variables numéricas, lo que permitió identificar patrones de correlación entre los datos recolectados.

C. Preparación de los datos

Se llevó a cabo un tratamiento de valores faltantes y atípicos para mejorar la calidad del conjunto de datos. En primer lugar, se identificaron los valores ausentes en cada columna, calculando su cantidad y porcentaje de afectación. Dado que estos valores no eran representativos dentro del dataset, se optó por eliminar los registros incompletos. Posteriormente, se comparó el tamaño del dataset antes y después de la limpieza para evaluar el impacto de esta eliminación. Además, se destacaron las columnas con mayor cantidad de valores faltantes, lo que permitió comprender mejor la magnitud del problema y justificar la decisión tomada.

Luego, se realizó la detección de valores atípicos, enfocándose únicamente en las columnas numéricas. Para ello, se aplicaron dos métodos: el primero utilizó rangos intercuartílicos para identificar valores que se desviaban significativamente del resto, mientras que el segundo analizó la dispersión de los datos con base en la desviación estándar. Los resultados se visualizaron mediante diagramas de caja, lo que facilitó la identificación gráfica de los valores extremos. Durante este proceso, se observó que las columnas ”Código del Departamento” y ”Código del Municipio” fueron interpretadas como valores numéricos, por lo que se consideró necesario convertirlas en cadenas de texto en la etapa de transformación de datos.

Finalmente, se repitió el análisis de valores atípicos excluyendo las columnas mencionadas para obtener resultados más precisos. No se aplicaron modificaciones sobre los valores extremos identificados, ya que se determinó que no afectaban significativamente el análisis y podrían contener información relevante para la interpretación de los datos. Este enfoque permitió garantizar que el dataset mantuviera su integridad sin perder información valiosa.

D. Modelado

En esta fase, se realizó un análisis exploratorio de los datos para identificar patrones en la calidad del aire en distintos municipios. Se encontró que los municipios pequeños no presentan días de excedencia en los niveles de contaminación, mientras que en los municipios más grandes sí se registraron días en los que los contaminantes superaron los límites establecidos. Además, se observaron diferencias en la variabilidad de los niveles de contaminación entre ambos tipos de municipios, lo que sugiere la influencia de factores como la densidad poblacional, el tráfico vehicular y las actividades industriales. Estos hallazgos permiten formular hipótesis sobre las causas de la contaminación y su impacto en la salud pública.

E. Evaluación

Cumplimiento de los Objetivos del Negocio

Mejorar la Calidad del Aire:

- Reducción de Contaminación: El modelo ha identificado con precisión los municipios con mayores niveles de contaminación y días de excedencia, lo que permite implementar medidas específicas para reducir la contaminación en estas áreas.
- Políticas de Control de Emisiones: Basado en los hallazgos, se pueden desarrollar políticas de control de emisiones más efectivas en municipios grandes, donde se ha observado un mayor número de días de excedencia.

  Proteger la Salud Pública: Al identificar los municipios con mayores niveles de contaminación, se pueden dirigir recursos y esfuerzos para mejorar la calidad del aire en estas áreas, reduciendo así la incidencia de enfermedades respiratorias y cardiovasculares.

Desarrollar Políticas Ambientales:

Los resultados del modelo proporcionan una base sólida para el desarrollo de políticas ambientales basadas en datos. Esto asegura que las decisiones políticas estén respaldadas por evidencia científica y sean más efectivas. Las diferencias significativas encontradas entre municipios pequeños y grandes permiten adaptar las estrategias de mitigación de la contaminación según las características específicas de cada tipo de municipio[8].

La validación de los modelos es un paso crítico para asegurar la fiabilidad y precisión de los resultados obtenidos. En este análisis, se utilizó la validación cruzada para evaluar el desempeño del modelo.

Validación Cruzada: Se utilizó la validación cruzada k-fold, dividiendo el conjunto de datos en cinco particiones (folds). Esta técnica permite evaluar la estabilidad y generalización del modelo al entrenarlo y probarlo en diferentes subconjuntos de datos.

Las puntuaciones de validación cruzada obtenidas fueron [1. 1. 1. 1. 1.], lo que indica que el modelo obtuvo una precisión perfecta del 100% en cada una de las cinco particiones del conjunto de datos. Esto significa que el modelo clasificó correctamente todos los ejemplos en cada partición, demostrando una capacidad excepcional para generalizar y predecir correctamente los días de excedencia en los municipios analizados.

La fase de evaluación ha demostrado que los modelos desarrollados son efectivos para identificar patrones de contaminación y días de excedencia en los municipios analizados.

F. Despliegue

A continuación, se detallan los principales hallazgos:

Para verificar la normalidad de los datos, se realizaron las siguientes pruebas:

- Prueba de Shapiro-Wilk
- Prueba de Kolmogorov-Smirnov (K-S)
- Gráfico Q-Q (Quantile-Quantile)

Los resultados de estas pruebas indicaron que los datos no siguen una distribución normal. Los valores p obtenidos fueron significativamente menores a 0.05, lo que llevó a rechazar la hipótesis nula de normalidad. Esto es crucial ya que muchas pruebas estadísticas paramétricas asumen normalidad en los datos, y la falta de esta puede invalidar los resultados de dichas pruebas.

La homocedasticidad se evaluó utilizando las siguientes pruebas:

- Prueba de Levene
- Prueba de Breusch-Pagan

Ambas pruebas mostraron que las varianzas no son iguales (p ¡ 0.05), indicando heterocedasticidad en los datos. Esto significa que la variabilidad de los días de excedencia no es constante entre los diferentes municipios

La pregunta de investigación planteada fue: ¿En los municipios pequeños hay menos días de excedencia debido a que tienen una menor población? Para responder a esta pregunta, se formularon las siguientes hipótesis:

- H0: En los municipios pequeños, el número de días de excedencia no es significativamente diferente al de los municipios más grandes.
- H1: En los municipios pequeños, el número de días de excedencia es significativamente menor que en los municipios más grandes.

Dado que los datos no se distribuyen normalmente, se utilizó la prueba no paramétrica de Mann-Whitney U. Los resultados mostraron una diferencia significativa en el número de días de excedencia entre municipios pequeños y grandes (p-valor = 0.0), lo que llevó a rechazar la hipótesis nula y aceptar la hipótesis alternativa.

III. Recomendaciones Basadas en el Análisis

- Implementación de Políticas de Control de Emisiones en Municipios Grandes: - Reducción de Emisiones Industriales: Se deben implementar políticas más estrictas para controlar las emisiones de las industrias ubicadas en municipios grandes. Esto puede incluir la adopción de tecnologías más limpias y la regulación de las emisiones de contaminantes. - Control del Tráfico Vehicular: Dado que el tráfico vehicular es una fuente importante de contaminación en áreas urbanas, se recomienda la implementación de medidas como la promoción del transporte público, la creación de zonas de bajas emisiones y la incentivación del uso de vehículos eléctricos. - Fomento de Espacios Verdes y Áreas Naturales: - Creación de Parques y Áreas Verdes: La creación y mantenimiento de parques y áreas verdes en municipios grandes puede ayudar a mejorar la calidad del aire al actuar como sumideros de contaminantes. - Protección de Áreas Naturales: La protección y expansión de áreas naturales alrededor de municipios grandes puede contribuir a la mejora de la calidad del aire y proporcionar beneficios adicionales para la biodiversidad y el bienestar de la comunidad. - Educación y Concienciación Pública: - Campañas de Concienciación: Se deben llevar a cabo campañas de concienciación pública sobre los efectos de la contaminación del aire en la salud y las medidas que los ciudadanos pueden tomar para reducir su exposición y contribuir a la mejora de la calidad del aire. -Programas Educativos: La inclusión de programas educativos sobre la calidad del aire y la sostenibilidad ambiental en las escuelas puede fomentar una mayor conciencia y responsabilidad ambiental entre las generaciones más jóvenes.

Las recomendaciones presentadas se basan en los hallazgos del análisis y están diseñadas para abordar las diferencias significativas en la calidad del aire entre municipios pequeños y grandes. La implementación de estas recomendaciones puede contribuir a la mejora de la calidad del aire, la protección de la salud pública y la promoción de un entorno más sostenible y saludable para todos los ciudadanos [9].

References

[1] Organización Mundial de la Salud, "Contaminación del aire y salud," OMS, 2021. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] Ministerio de Ambiente y Desarrollo Sostenible de Colombia, "Informe sobre la calidad del aire en Colombia," MinAmbiente, 2023.

[3] J. Han, M. Kamber y J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[4] D. W. Dockery y C. A. Pope, "Acute Respiratory Effects of Particulate Air Pollution," Annual Review of Public Health, vol. 15, pp. 107-132, 2006.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[8] Ballester, F. (2005). Contaminación atmosférica, cambio climático y salud. *Revista Española de Salud Pública*, 79(2), 159-175.
[9] Academia Nacional de Medicina de México. (2015). La contaminación del aire y los problemas respiratorios. *Revista de la Facultad de Medicina (México)*, 58(5), 72-75.