Modelos para la validación del empleo de sensores de bajo costo en la medición de la calidad del aire

Alexis Fernando Castrillón Gutiérrez

Artículo presentado para optar al título de Magíster en Ciencias Naturales y Matemática

Director
Ferney Orlando Amaya Fernandez, Doctor (PhD) en Ingeniería Área Telecomunicaciones

Universidad Pontificia Bolivariana
Escuela de Ingenierías
Maestría en Ciencias Naturales y Matemática
Medellín, Antioquia, Colombia
2024

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Modelos para la validación del empleo de sensores de bajo costo en la medición de la calidad del aire

Alexis Fernando Castrillón Gutiérrez [a] & Ferney Amaya-Fernández[b]

[a] *Estudiante Maestría en Ciencias Naturales y Matemáticas, Universidad Pontificia Bolivariana, Medellín, Colombia.* alexis.castrillon@upb.edu.co.
[b] *Escuela de Ingenierías, Universidad Pontificia Bolivariana, Medellín, Colombia.* ferney.amaya@upb.edu.co

Abstract
Nations have invested in air monitoring stations to design environmental policies based on data. However, due to their high cost, most cases see only a few stations, even in large cities. Consequently, studies are being conducted on the effectiveness of measurements taken by low-cost sensors (LCS). This research, part of the PROMESA Project (PROcedencia del Material particulado y su Efecto en la SAlud de los niños), conducted in the cities of Bogotá and Medellín, evaluated different machine learning models to improve the quality of the measurements provided by LCS in these cities. The calibration of the LCS was performed using meteorological variable measurements from official stations belonging to the air quality network in these cities.

Keywords: Data Analytics, Data Science, Automatic Learning, Low-Cost Sensors, Environmental Pollution/.

Resumen
Las naciones han invertido en estaciones de mediciones del aire para diseñar políticas medioambientales basados en los datos. Sin embargo, al ser tan onerosas, en la mayoría de los casos se cuentan con pocas estaciones, inclusive en las grandes ciudades. Por esta razón, se están realizando estudios sobre la efectividad de las mediciones realizadas por sensores de bajo costo (LCS, *Low Cost Sensors*). En esta investigación, que hace parte del Proyecto PROMESA (PROcedencia del Material particulado y su Efecto en la SAlud de los niños) que se desarrolló en las ciudades de Bogotá y Medellín, se evaluaron diferentes modelos de aprendizaje automático para mejorar la calidad de las mediciones entregadas por los LCS que se tienen estas ciudades. El ajuste de los LCS se realizó tomando como referencia las mediciones de las variables meteorológicas entregadas por estaciones oficiales pertenecientes de la red de calidad del aire presente en estas ciudades.

Palabras clave: Analítica de los datos, ciencia de los datos, aprendizaje automático, sensores de bajo costo, contaminación ambiental.

1. Introducción

Es indudable que la humanidad está afrontando de forma creciente el impacto de la degradación ambiental causada por las decisiones relacionadas con el sostenimiento de la dinámica económica actual que requiere un uso extensivo de los recursos naturales.

Un ejemplo de lo anterior tiene que ver con la incesante demanda de bienes y servicios por la sociedad, lo que ha supuesto un exceso de residuos, entre ellos algunos tóxicos, que por su manejo inadecuado ha generado una extensa contaminación de las fuentes hídricas, de los suelos y del ambiente en general. La pérdida de biodiversidad, la escasez de agua potable y el cambio climático son otros de los efectos más evidentes de las acciones humanas sobre del medio ambiente.

Lamentablemente uno de los componentes ambientales más afectados por la contaminación es el aire, el cual es esencial para toda forma de vida en la tierra. La calidad del aire se ha visto afectada por los altos niveles de gases de efecto invernadero que existen en la actualidad, los cuales son diseminados en el ambiente por los vehículos motorizados, las industrias, por el procesamiento y uso de las energías de origen fósil, entre otras acciones humanas.

Ahora bien, esta preocupante realidad, ha sido considerada en diversas conferencias y tratados internacionales enfocados en la protección del medioambiente, como en el Protocolo de Kyoto de 1997 o en la Cumbre Mundial sobre Desarrollo Sostenible celebrada en Johannesburgo en el año 2002. En todas estas convenciones, no solo se ha realizado un diagnóstico del problema ambiental en todos los niveles, sino que además se han planteado unos compromisos específicos para las naciones firmantes con el objetivo de mitigar el impacto de la

contaminación ambiental.

Sin embargo, pese a las anteriores iniciativas, los niveles de compuestos químicos y óxidos perjudiciales para el ser humano generados por la contaminación atmosférica (tales como el dióxido de azufre (SO2), dióxido de nitrógeno (NO2), monóxido de carbono (CO), o el material particulado (PM, Particulate Matter), contaminantes que siguen impactando a la sociedad en todas sus dimensiones, en especial en la salud pública (Scagliotti & Jorge, 2021).

Estudios realizados en diferentes países han constatado la correlación positiva entre el aumento en los niveles de los contaminantes y el incremento en las enfermedades cardiacas y respiratorias. Morakinyo et al. (2016) señalan en su investigación que la exposición a contaminantes del aire está relacionada con el aumento de la presión sistólica, lesiones pulmonares, inflamación de las vías respiratorias y la aparición de síntomas de asma entre pacientes atendidos en la Columbia Británica.

Ahora bien, la contaminación ambiental no solo está impactando de forma negativa la salud pública, sino también su esfera de influencia se expande hasta la economía. El Departamento Nacional de Planeación (2015) en Colombia, indica que la degradación ambiental genera pérdidas en la productividad sectorial y en el capital humano impactando ampliamente la economía. Un ejemplo del impacto de la contaminación ambiental en la economía podemos verlo en China, cuyo fuerte crecimiento económico en varias décadas ha supuesto un fuerte daño ambiental reflejado en altos niveles de $PM_{2.5}$, PM_{10} y CO en el ambiente (Cifuentes Martínez et al., 2016), que de acuerdo con World Bank (2013), los costos de la contaminación ambiental pueden estimarse en un 9% del PIB (Producto Interno Bruto) del país, lo que se explica en el agotamiento de recursos energéticos y minerales, más los gastos generados por el sistema de salud (2.8% del PIB) para la atención de quienes sufren enfermedades cardiacas y respiratorias.

En el caso de Colombia el Departamento Nacional de Planeación (2018) estimó que para el año 2015 la degradación ambiental le costó a este país el equivalente de 2.1% de su PIB, lo que equivale a 12.2 billones de pesos.

Además de las referencias presentadas, existe evidencia científica concluyente que relaciona el aumento de los niveles de las partículas o sustancias contaminantes en el ambiente con el incremento de la presencia de enfermedades respiratorias que son atendidas por los hospitales o que incluso pueden conducir a la muerte (Sepadi y Nkosi, 2021). De hecho, esta es una realidad que está golpeando especialmente a las grandes ciudades, debido a lo mayor presencia de los factores contaminantes en estos lugares, y por esta razón ha sido motivo de numerosas investigaciones en las dos últimas décadas.

Indudablemente, cuando se presenta hechos probados sobre el impacto de la contaminación ambiental en la salud pública y en la economía, se genera posibles escenarios para el diseño de una política medioambiental basada en la evidencia, en la cual se implementen acciones que corrijan y mitiguen la degradación medioambiental.

Además, estos estudios constituyen el fundamento para las proyecciones que se pueden hacer a corto, mediano y largo plazo, tanto de la atención hospitalaria por enfermedades cardiacas y respiratorias causadas por la contaminación ambiental, así como su impacto en el PIB. Por tanto, esta información es trascendental porque constituye una herramienta necesaria para quienes tienen la responsabilidad de diseñar y aplicar planes de desarrollo locales y nacionales. Por todo lo expuesto, continuar estudiando esta problemática con los modelos predictivos permitirá proyectar mejor las decisiones pertinentes en relación con el cuidado del medioambiente.

Hasta ahora se ha valorado la importancia de estudiar la relación que existe entre los contaminantes ambientales con la presencia de enfermedades cardiacas y respiratorias, como fundamento para el diseño de políticas públicas que reduzcan el daño medioambiental; sin embargo, es necesario precisar que la pertinencia de los resultados obtenidos en las investigaciones depende de la fiabilidad de los datos que se utilicen. Por lo anterior, los países han invertido recursos para la compra de redes de monitoreo para realizar mediciones de diferentes variables medioambientales. Sin embargo, estos equipos tienen un alto costo de compra además que su mantenimiento también es oneroso, por lo que, en países como Colombia, en la mayoría de las ciudades se tienen pocas estaciones, mientras que en la mayoría de los municipios no cuenta con estaciones para medir variables medioambientales.

Por tanto, se viene estudiando alternativas más económicas y prácticas, con una efectividad equivalente. De hecho, varias naciones están realizando pruebas con estaciones que contienen sensores de bajo costo (LCS), y debido a sus bajos costos y tamaño más reducido, pueden ubicarse en diferentes ubicaciones dentro de la ciudad o municipio. Sin embargo, Scagliotti y Jorge (2021, pág. 2) mencionan varias desventajas en su utilización, incluyendo que no se cuentan con protocolos de calibración, validación y evaluación de desempeño que sean universalmente aceptados, sin embargo, afirman que existen numerosos esfuerzos al respecto.

Para resolver este problema, los investigadores se han enfocados en dos procesos: Gestión automática de Calidad de datos (QA) y el de control de calidad (QC). El primero hace referencia a los ajustes automáticos a los datos cuando se corrigen los procesos, mientras el segundo hace referencia a la verificación de los datos de acuerdo a la exigencia del usuario. (Scagliotti y Jorge, 2021, pág. 2). Para la QA se están validando los procesos a través del diseño de modelos que permitan el aprendizaje automático de los LCS.

Teniendo presente lo anterior, esta investigación está enmarcada en el proyecto PROMESA realizado en las ciudades de Bogotá y Medellín, lugares con altos niveles de contaminación ambiental y alta preocupación por su impacto en la salud. Se evaluaron diferentes modelos de aprendizaje automático para mejorar la calidad de las mediciones entregadas por los LCS que se tienen estas ciudades. El ajuste de los LCS se realizó teniendo como referencia las mediciones de las variables meteorológicas y de calidad de aire entregadas por estaciones oficiales. Se analizaron las

variables humedad, temperatura, $PM_{2.5}$ y PM_{10} provenientes de una estación LCS ubicada en la ciudad de Bogotá que cuenta con datos disponibles para el estudio. Se realizó limpieza de los datos y posteriormente se aplicaron los modelos de aprendizaje automático Regresión Lineal y Bosque Aleatorio en sus versiones univariable y multivariable. Como medida de comparación entre los diferentes modelos se seleccionó el Coeficiente de Determinación. Los mejores resultados se obtuvieron para el modelo de Bosque Aleatorio Multivariable.

En este artículo, se presentará inicialmente una revisión bibliográfica de las investigaciones que abordan el impacto ambiental y modelos de ajusta de estaciones LCS. Posteriormente se presenta la metodología, que se fundamente en la metodología CRISP- DM (*CRoss Industry Standard Process for Data Mining*) ampliamente empleada en modelos de inteligencia de negocios y análisis de datos (Parra Sánchez et al. 2020). Posteriormente se presentan los resultados y luego las conclusiones.

2. Estado del Arte

Como se evidencio en la introducción, una mayor presencia de contaminantes en el aire tiene un fuerte impacto en la salud pública. Por tanto, es necesario identificar mecanismos que permitan calibrar los LCS. En esta sección se presenta una revisión bibliográfica en los temas de monitoreo de la contaminación ambiental, impacto en la salud pública y técnicas de calibración y ajuste de LCS.

2.1. Monitoreo de la contaminación ambiental

En vista de la correlación entre los contaminantes del aire y la salud pública, ha cobrado mucha relevancia la medición de los niveles de contaminantes del aire y para ello se utilizan sistemas de monitoreo del aire. Estos sistemas pueden incluir muestreadores pasivos, activos, analizadores automáticos y sensores remotos. En nuestro país las principales redes de monitoreo están ubicadas en las dos principales ciudades:

- Red de Monitoreo de Calidad del Aire de Bogotá (RMCAB), que cuenta 20 estaciones,
- Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA), que cuenta con 36 estaciones.

Las estaciones de monitoreo miden principalmente partículas en estado líquido o sólido, que son clasificadas de acuerdo de su tamaño medido en micrómetros. De esta forma encontramos material particulado $PM_{2.5}$, que incluye partículas sólidas o líquidas que se encuentran en suspensión aerodinámica cuyo diámetro es de menor de 2.5 micrómetros. También se puede medir PM_{10} y $PM_{1.0.}$ Debido a su minúsculo tamaño, una vez absorbidas pueden alojarse en el sistema respiratorio, en especial en los pulmones. De acuerdo con OMS (2022) los "principales componentes de la materia particulada son los sulfatos, los nitratos, el amoníaco, el cloruro de sodio, el carbono negro, los polvos minerales y el agua".

Entre los principales contaminantes se encuentran el CO2 NO2 y O3. El CO, monóxido de carbono, es un gas indoloro, no visible, surgido por la combustión deficiente de combustibles tales como el petróleo, carbón, gas, entre otros. (OMS, 2022). El NO2, dióxido de nitrógeno, es un gas contaminante generado por los procesos de combustión causados en la industria y los diversos medios de transporte. (OMS, 2022). El O3 también conocido como ozono al nivel del suelo, es un gas que influye en la generación de la lluvia acida, el cual es formado en la atmosfera al reaccionar con otros contaminantes (Farrow, 2021).

2.2. Impacto en la salud publica

Diversas investigaciones han relacionado la contaminación ambiental con los efectos en la salud. Arku et al. (2018) relacionó la presencia de $PM_{2.5}$ en niveles superiores a 10 μg/m3 con el incremento en la mortalidad cardiaca y respiratoria en 0.46% y 0.47% respectivamente. En la misma línea, un informe de la publicación semanal The Economist (2017), retomando una investigación divulgada por la revista Nature, informan que se puede atribuir la muerte de cerca de 3 millones de personas en el año 2007 en el mundo a la emisión de partículas finas $PM_{2.5}$. Este estudio concluye que la mayor presencia de estas partículas se da en naciones pobres, debido a la producción bienes y servicios que luego son exportados a naciones más ricas.

En vista de lo anterior, la OMS (2022) ha establecido unos valores máximos para cada uno de los contaminantes, para que los sistemas de monitoreo generen las respetivas alertas. Los resultados se presentan en la Tabla 1.

Tabla 1.
Niveles de contaminante considerados nocivos para la salud.

Contaminante	Valor máximo anual en $\mu g/m^3$	Valor máximo en 24 horas en $\mu g/m^3$
$PM_{2.5}$	5	15
PM_{10}	15	45
CO	-	4
NO_2	10	25
O_3	60	100

Fuente: Elaboración propia basada en OMS (2022).

En esta misma línea, Parra Sánchez et al. (2020) investigaron la relación existente entre el aumento de las partículas $PM_{2.5}$ con las personas que acuden a los hospitales en la ciudad de Medellín, Colombia. Utilizando los datos arrojados por los monitores de calidad de aire del Área Metropolitana, diseñaron un modelo predictivo del impacto de la contaminación del aire y otro modelo que permite observar la asistencia a centros de salud por enfermedades respiratorias. El modelo predictivo desarrollado por Parra Sánchez et al. (2020) utilizó la metodología CRISP- DM, además de incluir un modelo de visualización de los hallazgos. Un resultado notable de esta investigación es que demuestran la correlación positiva entre aumentos de las partículas $PM_{2.5}$ y el incremento en las personas que consultan en los hospitales por enfermedades respiratorias.

También en Latinoamérica, en este caso en Brasil, (de Moraes et al., 2019) se realizó un estudio en Sao Paulo, en el que se asociaron las condiciones meteorológicas, la

contaminación ambiental y las hospitalizaciones de niños, en el periodo de 2003 hasta 2013. Para medir esta asociación se utilizó un modelo lineal generalizado con una distribución binomial negativa además de un modelo DLNM (*Distributed Lag Non-linear Model*). Con esta investigación se logró demostrar la vinculación entre la temperatura promedio y otras condiciones relacionadas con la exposición a contaminantes, con las hospitalizaciones de niños de hasta 9 años de edad en 14 localidades de Sao Paulo.

Continuando en Latinoamérica, Cifuentes Martínez et al. (2020) hicieron un análisis estadístico con el número promedio diario de personas que acuden a urgencias por enfermedades respiratorias y la concentración de $PM_{2.5}$ diaria de acuerdo con los datos extraídos obtenidos del Sistema de Información Nacional de Calidad del aire (SINCa) en dos comunas de la región de Ñuble, Chile. Los autores desarrollaron un modelo estadístico en el que aplicaron el test estadístico de Dickey-Fuller y el análisis inferencial basado en correlación de Spearman y Cross-Correlation, encontrando que, en los meses de abril a septiembre, correspondientes a los meses de temperaturas más bajas, aumenta la concentración de $PM_{2.5}$ en el aire. Específicamente cuando los días en los que los niveles de $PM_{2.5}$ están por encima de 170 μg/m3, se determinó un incremento en las consultas médicas, especialmente los días 1 y 9 después de este evento. Por tanto, se evidencia una correlación positiva entre contaminación ambiental y el número de personas que acuden al servicio médico debido a enfermedades respiratorias (Cifuentes Martínez et al., 2020).

Por otro lado, Oyana et al. (2019) realizaron una investigación con niños de 2 a 5 años que vivían en la ciudad de Memphis, Estados Unidos. A partir de un modelo de regresión logística y otro predictivo diseñado desde la analítica de datos, concluyeron que hay una fuerte relación entre niños con cuadros de asma y su exposición a ambientes con altas concentraciones de $PM_{2.5}$.

Trasladándonos al continente europeo, Wrotek et al. (2021), utilizando un modelo de regresión general por aglomeración, estudiaron en Polonia durante el periodo de 2010 a 2019 la relación de los contaminantes presentes en el aire ($PM_{2.5}$, PM_{10} y NO2) con el número de hospitalizaciones de niños debido al virus respiratorio sincitial (VRS). En este caso se demuestra que un aumento del 31.4% de hospitalizaciones por VRS se puede explicar con un incremento de la presencia en el ambiente de $PM_{2.5}$, PM_{10} y NO2. También se concluye que las condiciones climáticas influyen en la mayor presencia de los contaminantes ambientales.

Continuando en Polonia, Niewiadomska et al. (2020) aplicaron un modelo no lineal con retardo para estudiar el riesgo de hospitalización de personas adultas en la ciudad de Silesia por el incremento de los contaminantes del aire. En este caso se centraron en las siguientes enfermedades respiratorias: Crisis de asma y Bronquitis aguda. Las estimaciones de la contaminación ambiental, utilizadas por Niewiadomska et al. (2020) se realizaron de acuerdo con los datos arrojados por monitores móviles no oficiales o LCS, los cuales presentan informes distribuidos de forma masiva a los silesianos. Los investigadores hallaron una correlación positiva entre la exposición a los contaminantes ambientales y las citas médicas por asma y bronquitis aguda, con un retardo promedio de 3 días.

En cuanto a investigaciones publicadas en el año 2022, se reseñará la realizada por (Chakraborty et al., 2022), quienes utilizaron un modelo de regresión binomial negativo bayesiano, para analizar en cuatro regiones estadounidenses, el registro de muertes por COVID entre marzo y agosto de 2020. Desarrollaron un modelo espaciotemporal con los datos de 150 condados, relacionando la cantidad de muertes de COVID y la concentración de $PM_{2.5}$ en estos lugares. Uno de los hallazgos más importantes, es la correlación positiva entre una mayor presencia $PM_{2.5}$ a largo plazo con la mortalidad por contagio de COVID expuesta en los recuentos semanales.

Finalmente se retomarán dos investigaciones realizadas en China. Peng et al. (2022), utilizando un modelo aditivo con distribución de cuasi-Poisson, realizaron un estudio en la ciudad de Shanghái, en el que se relaciona la presencia de las partículas $PM_{2.5}$y PM_{10} en el aire con la cantidad de personas ingresadas al hospital por enfermedades respiratorias. La investigación está comprendida desde el 1 de enero 2008 hasta el 31 de julio de 2020, y por tanto es un trabajo investigativo con resultados de un periodo extenso de una ciudad caracterizada por su alta contaminación ambiental. Se evidenciaron en los hallazgos de los 12.5 años estudiados, un aumento en las hospitalizaciones cuando los niveles de $PM_{2.5}$ y PM_{10} superaban 10 μg/m3 cada una. Además, se determinó que los incrementos de los contaminantes impactan en mayor medida a las personas mayores de 45 años.

Por tanto, podemos concluir en relación con las investigaciones referenciadas hasta ahora, que existe evidencia contundente, sobre la relación entre el aumento de los contaminantes y el incremento en el número de enfermedades cardiacas y respiratorias registradas en diferentes centros de salud.

2.3. Calibración de LCS

En esta sección se presentan las investigaciones más relevantes en los últimos cinco años, que es el periodo en el que se han intensificado las investigaciones sobre la calibración de LCS.

En primer lugar, por lo establecido en la sección anterior, las acciones estatales y privadas para medir la calidad del aire son cada vez más apremiantes. Sin embargo, como se ha declarado anteriormente, tanto la compra como mantenimiento de las redes de monitoreo resultan una inversión muy costosa, por lo que la alternativa de adquirir LCS está siendo evaluada a nivel mundial. Ahora bien, al no existir un protocolo para su calibración, los investigadores en este campo están diseñando modelos para ajustar los datos resultantes de las mediciones de los LCS con las de las estaciones oficiales. En esta sección se presentan algunas investigaciones relevantes.

Chojer et al. (2022) realizaron la evaluación de

sensores de bajo costo en una escuela de Oporto en Portugal. Con los algoritmos de regresión se alcanzó un R2 de 0.9, corroborando la eficiencia en el proceso de calibración. Ahora bien, este proceso fue desarrollado en las siguientes etapas:

- Preprocesamiento de datos
- Clasificación de los datos en subconjuntos de entrenamiento y prueba
- Algoritmos de regresión: MLR (regresión lineal simple) y SVR (la máquina de vectores de soporte). Regresiones de refuerzo (GBR - regresión de refuerzo de gradiente y XGB - refuerzo de gradiente extremo) para la optimización de los hiperparámetros
- Modelo de entrenamiento
- Prueba del modelo
- Evaluación del modelo.

Continuando con la revisión de investigaciones Malyan et al. (2023) realizaron un muestreo de campo intensivo en cinco ubicaciones diferentes dentro del campus de IIT Bombay, India, para identificar las discrepancias en los datos de contaminación ambiental en comparación con las mediciones de sensores de bajo costo. Para la calibración de LCS, en este estudio se emplearon los siguientes algoritmos de aprendizaje automático k-vecinos más cercanos (*kNN, k-Nearest Neighbors*), bosque aleatorio (RF, *Random Forest*) y aumento de gradiente (GB, *Gradient Boosting*). Los resultados de los algoritmos reflejan una mejora en los valores de R2 pasando de 0.75 a 0.85. Adicionalmente sugieren una calibración robusta específica del sitio de LCS basada en la distribución del tamaño de las partículas y la dependencia del rendimiento de LCS.

Pasando a la China, Ghamari et al. (2022) compararon ocho enfoques de aprendizaje de máquina (ML, *Machine Learning*) para ajustar los datos obtenidos del LCS. Con los resultados concluyen que las técnicas de calibración propuestas basadas en enfoques de ML son el primer paso para avanzar en esta investigación hacia el campo del ML integrado, donde los sistemas integrados que incorporan procesadores de bajo coste pueden utilizar los datos entrenados existentes en la nube para tomar decisiones y hacer predicciones de forma independiente.

Por otro lado, Liang y Daniels (2022) realizaron un estudio en el que evaluaron exhaustivamente diez técnicas de datos ampliamente utilizadas para calibrar los LCS, a saber, regresor de aumento adaptativo (AdaBoost), Bayesiano, GB, kNN, RF, Lasso, regresión lineal multivariable (MLR, *Multiple Linear Regression*), red neuronal, regresión de cresta y máquina de soporte vectorial (SVM, *Support-Vector Machines*). Un resultado interesante de la investigación fue que la regresión lineal sigue siendo una opción viable para estudios con esfuerzo limitado en el ajuste de parámetros y selección de métodos, especialmente considerando su eficiencia computacional y sencillez (Hashmy et al., 2021).

Por otro lado, Ferrer-Cid et al. (2022) investigaron la calibración de sensores de bajo costo en Barcelona, España utilizando técnicas de aprendizaje automático. El proceso implico entrenar un modelo de calibración para traducir de mediciones de sensores de bajo costo a concentraciones para una especie de gas específica y para corregir las mediciones de sensores para mejorar su precisión. Con su trabajo, concluyeron que con un mayor periodo de muestreo se puede obtener una buena calidad de datos a cambio de un importante ahorro energético. Además, afirman que los esfuerzos se han centrado en obtener la máxima calidad de datos (calidad de calibración) sin tener en cuenta las posibles restricciones que pueda tener un nodo de internet de las cosas (IoT, *Internet of things*), especialmente en el caso de siendo alimentado por baterías.

Mientras que Patton et al. (2022) desarrollaron modelos directos de calibración de campo utilizando árboles de decisión potenciados por gradiente probabilístico (GBDT). Partieron de la hipótesis de que los sensores de bajo costo exhiben sesgos no lineales, por lo que requieren de calibración. Con los resultados de su investigación concluyeron que el uso de modelos de aprendizaje automático probabilístico de código abierto para la calibración de sensores en el lugar supera a los modelos lineales tradicionales y no requiere un paso inicial de calibración de laboratorio.

Volviendo a China, Day et al. (2022) diseñaron una Calibración con un modelo de regresión lineal multivariante (MLR). Se realizó utilizando redes neuronales artificiales (ANN, *Artificial Neural Network*). Los modelos ANN se entrenaron usando de una a tres capas ocultas. En el segundo paso de la calibración, se utilizaron modelos no lineales que utilizan aprendizaje automático (ML) para el análisis de regresión. Se utilizaron tres algoritmos de conjunto de refuerzo diferentes: AdaBoost, regresor de aumento de gradiente estocástico (GBR) y regresor de aumento de gradiente extremo (XGBoost). Se evaluó la efectividad de la recolección de los contaminantes ambientales por parte de sensores de bajo costo y se recomienda una estrategia de calibración en dos pasos.

En Pakistan, Hashmy et al. (2021) ejecutaron la metodología de pronóstico (MAQ-CaF), que elude los desafíos de la falta de confiabilidad a través de su máquina de diseño modular basado en el aprendizaje que aprovecha el potencial del marco IoT. Este estudio confirma que los sensores de bajo costo requieren de la calibración. Los investigadores hallaron resultados alentadores del pronóstico de parámetros de calidad del aire determinados por información mutua basada en puntajes método de selección de características y multivariado-LSTM.

Por otro lado, Ali et al. (2023) realizaron la calibración basada en la red neuronal convolucional unidimensional (1DCNN) para sensores de monóxido de carbono de bajo costo y comparar su rendimiento con varias técnicas de calibración basadas en aprendizaje automático. Se encontró que 1DCNN funciona de manera consistente en todos los conjuntos de datos.

En general, se encontró que el empleo de técnicas de ML ha permitido ajustar datos obtenidos con LCS para la medición de variables ambientales y de contaminación ambiental.

3. Metodología

Para abordar la solución se empleó la Metodología CRISP-DM que es un modelo de procesos ampliamente utilizado en la minería de datos y el análisis de datos y se compone de seis fases principales:

- Comprensión del Negocio: determinar los objetivos específicos del proyecto.
- Comprensión de los Datos: recolectar datos, obtener una visión general de los datos recopilados y evaluar la calidad de los datos.
- Preparación de los Datos: selección, limpieza, transformación e integración de datos.
- Modelado: selección de técnicas de modelado más apropiadas según los objetivos del proyecto, construcción y evaluación del modelo.
- Evaluación: comparar los modelos y seleccionar el mejor basado en su rendimiento y su capacidad para cumplir con los objetivos.
- Despliegue: implementación del modelo seleccionado en un entorno real.

Siguiendo la metodología CRISP-DM, en la Figura 1 se presenta un diagrama con los pasos del proceso realizado. La descripción es la siguiente:

- Definición del estudio: identificación del propósito del estudio, selección de técnicas de aprendizaje automático y selección de la métrica de comparación.
- Captura de datos.
- Análisis preliminar de los datos: se identifican las principales características de los datos capturados identificando acciones necesarias para realizar limpieza.
- Limpieza de datos: se aplican acciones para limpiar, depurar y homogeneizar los datos.
- Conjuntos de entrenamiento y prueba: separación de los datos en los conjuntos de entrenamiento y prueba.
- Técnicas de aprendizaje automático: empleo de las técnicas de aprendizaje automático seleccionadas.
- Cálculo de métricas.
- Análisis de resultados.

3.1. Definición del estudio

El propósito del estudio es probar diferentes técnicas de aprendizaje de máquina para ajustar los datos provenientes de una estación LCS teniendo como referencia los datos de una estación oficial. En adelante se llamará a la estación LCS estación PROMESA. Debe seleccionarse una estación oficial cerca a la estación LCS. De acuerdo a la revisión de la literatura se seleccionan las técnicas de regresión lineal univariable, regresión lineal multivariable, bosque aleatorio univariable y bosque aleatorio multivariable. Como parámetro de comparación se emplea el coeficiente de determinación (Pedregosa et al. (2011).

En la Figura 2 se presenta la imagen de la estación PROMESA.

Figura 2. Estación PROMESA. Fuente: Proyecto PROMESA.

En la Tabla 2 se presentan las características para las variables $PM_{2.5}$, PM_{10}, humedad y temperatura, que son las seleccionadas en este estudio.

Tabla 2.
Características de los sensores de la Estación PROMESA.

Características	Valor
Alimentación	5 V
Código del sensor	SEN0233
Rango de medida $PM_{2.5}$, PM_{10}	0 – 2000 µg / m³
Marca	DFROBOT
Humedad	0 – 99%
Temperatura	-10° - 50°C

Fuente: Proyecto PROMESA

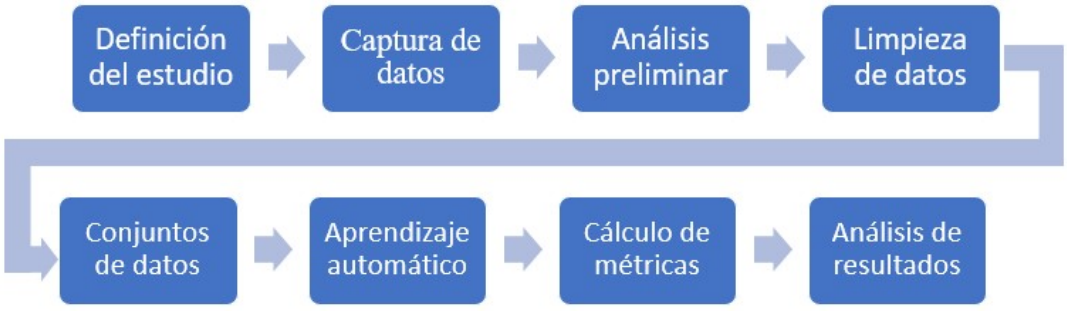

Figura 1. Diagrama del proceso realizado.

3.2. Captura de datos

El estudio se realizó con datos de la estación oficial de Las Ferias en Bogotá, debido a que en ese sitio se cuenta con una estación PROMESA y se cuenta con datos oficiales de la red de monitoreo de la ciudad. La estación PROMESA tiene datos para las fechas entre el 28 de febrero y el 31 de marzo del 2023 para las variables $PM_{2.5}$, PM_{10}, humedad y temperatura.

Para la estación PROMESA se cuenta con mediciones cada minuto y para la estación oficial se cuenta con mediciones cada hora. En este caso se obtiene el promedio aritmético por hora para las estaciones PROMESA. En la Tabla 3 se presenta un resumen de la información disponible para el inicio del estudio.

Tabla 3.
Resumen de la información disponible para el inicio del estudio.

Variable	Unidad de medida	Número de registros
$PM_{2.5}$	Concentración en microgramos por metro cúbico	684
PM_{10}	Concentración en microgramos por metro cúbico	684
Humedad	Porcentaje	684
Temperatura	Grados °C	684

3.3. Análisis preliminar de los datos

En la Figura 3 se presenta el diagrama de cajas y bigotes para las 4 variables obtenidas de la estación PROMESA.

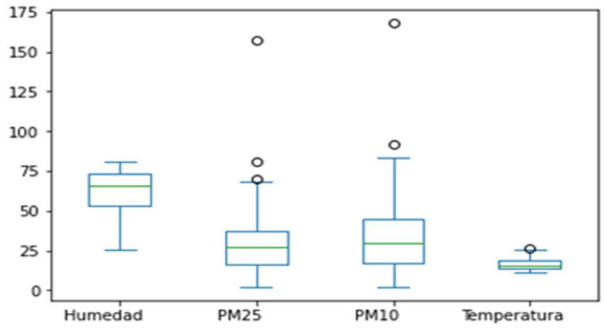

Figura 3. Diagrama de cajas y bigotes para las variables de las estaciones PROMESA.

Puede observarse en el diagrama que hay registros atípicos, los cuales deben ser ajustados. Adicionalmente, se encontró un desplazamiento temporal entre los datos adquiridos desde las estaciones PROMESA comparado con los datos adquiridas por las estaciones oficiales, esto se debe posiblemente a una diferencia en la sincronización entre los relojes de las estaciones. Este desfase también debe ser ajustado.

3.4. Limpieza de datos

Inicialmente se eliminaron los registros atípicos. Posteriormente se ajustó el desplazamiento temporal entre los datos adquiridos desde las estaciones PROMESA comparado con los datos adquiridas por las estaciones oficiales. Para encontrar el tiempo de desplazamiento se realizó una correlación cruzada encontrando los siguientes desplazamientos para las diferentes variables:

- Humedad: 1 hora
- $PM_{2.5}$: 4 horas
- PM_{10}: 6 horas
- Temperatura: 1 hora

Se realizó el ajuste temporal correspondiente para las mediciones de las estaciones PROMESA.

Se calculó la matriz de correlación entre las variables de las dos estaciones para encontrar posibles dependencias entre las variables. Los resultados se presentan en la Tabla 4.

A partir de la matriz de correlación puede observarse lo siguiente:

- Se observa una alta correlación negativa entre la humedad y la temperatura tanto en la estación oficial como en la de bajo costo.
- Se observa una alta correlación positiva entre $PM_{2.5}$ y PM_{10} tanto en la estación oficial como en la de bajo costo.
- No se observa correlación entre la temperatura o la humedad con las mediciones $PM_{2.5}$ o PM_{10} en ninguna de las dos estaciones.

Tabla 4.
Matriz de correlación entre las variables de
la estación PROMESA y le estación oficial (H:Humedad; T:Temperatura)

	H Oficial	H PROMESA	$PM_{2.5}$ Oficial	$PM_{2.5}$ PROMESA
PM_{10} Oficial	-0.08	-0.10	0.77	0.70
PM_{10} PROMESA	-0.14	-0.17	0.51	0.71
T Oficial	-0.94	-0.92	-0.06	-0.06
T PROMESA	-0.90	-0.95	-0.01	-0.01

3.5. Conjuntos de entrenamiento y prueba

Se dividieron los datos en conjunto de entrenamiento y conjunto de prueba, con el 70% de los datos para entrenar el modelo y el 30% de los datos para probar la eficiencia del modelo. La selección de los datos para los dos conjuntos se

realizó de forma aleatoria.

3.6. Técnicas de aprendizaje automático

Se empleó la biblioteca scikit-learn (Pedregosa et al. 2011) para el empleo de las siguientes técnicas:

- Regresión lineal univariable
- Regresión lineal multivariable
- Bosque aleatorio univariable
- Bosque aleatorio multivariable

4. Resultados

En esta sección se presentan los resultados de aplicar las diferentes técnicas de aprendizaje automático para ajustar los datos de la estación PROMESA empleando como referencia los datos de la estación oficial. En la Tabla 5 se presentan los valores del coeficiente de determinación antes y después de aplicar el modelo de regresión lineal univariable a las diferentes variables analizadas provenientes de la estación PROMESA.

Tabla 5.
Coeficiente de determinación para la regresión lineal univariable.

Variable	Antes de la regresión lineal univariable	Después de la regresión lineal univariable
Humedad	0.86	0.92
$PM_{2.5}$	-1.57	0.47
PM_{10}	-0.02	0.35
Temperatura	0.34	0.92

A partir de los resultados se observa lo siguiente:

- La humedad medida por la estación PROMESA tiene una similitud aceptable al comparar con la medición de la estación oficial (0.86) y se mejora la medida luego de aplicar la regresión lineal (0.92).
- Se obtiene una mejora significativa para la medición de la temperatura realizada por la estación PROMESA luego de aplicar regresión lineal univariable (0.92).
- Se mejora la medida proveniente de las estaciones $PM_{2.5}$ y PM_{10}, sin embargo, aún el coeficiente de determinación no presenta un alto valor.

En la Figura 4 se presenta la gráfica de dispersión para las diferentes variables. Se indica la distribución de puntos para las mediciones de la estación PROMESA antes y después de aplicar el modelo de regresión lineal univariable.

Para el modelo Bosque Aleatorio univariable y multivariable, se realizó un ajuste de hiper-parámetros (añadir referencia). Para esto se seleccionaron los siguientes parámetros críticos en el desempeño de la técnica:

- Máxima profundidad (*max depth*): indica la profundidad máxima del árbol. Para el ajuste se probó con los valores 3, 6 y 9.
- Máximo número de características (*max features*): cantidad de características a considerar al buscar la mejor división. Realiza una operación sobre el número de características (*n_features*), se probó con las siguientes opciones:
  - NONE: *max features* = *n_features*
  - SQRT: *max features* = raiz(*n_features*)
  - LOG2: *max features* = $\log_2$(*n_features*)
- Máximo de nodos de hoja (*max leaf nodes*): cuando este parámetro se configura, se construye el árbol de la mejor manera primero en lugar de primero en profundidad. Para el ajuste se probó con los valores 3, 6 y 9.

El ajuste de hiper-parámetros consistió en ejecutar las técnicas con las opciones indicadas seleccionando la que entregó mejor desempeño. Para la técnica multivariable, los mejores resultados se obtuvieron al incluir todas las variables, es decir: $PM_{2.5}$, PM_{10}, temperatura y humedad. Los parámetros con los que se obtuvo el mejor desempeño se presentan en la Tabla 6.

En la Tabla 7 se presenta el valor del coeficiente de determinación para los datos provenientes de la estación PROMESA antes de aplicar los modelos y luego de aplicar los modelos de regresión lineal univariable (RLU), regresión lineal multivariable (RLM), Bosque Aleatorio Univariable (BAU) y Bosque Aleatorio Multivariable (BAM). Se sombrean los máximos valores del coeficiente de determinación.

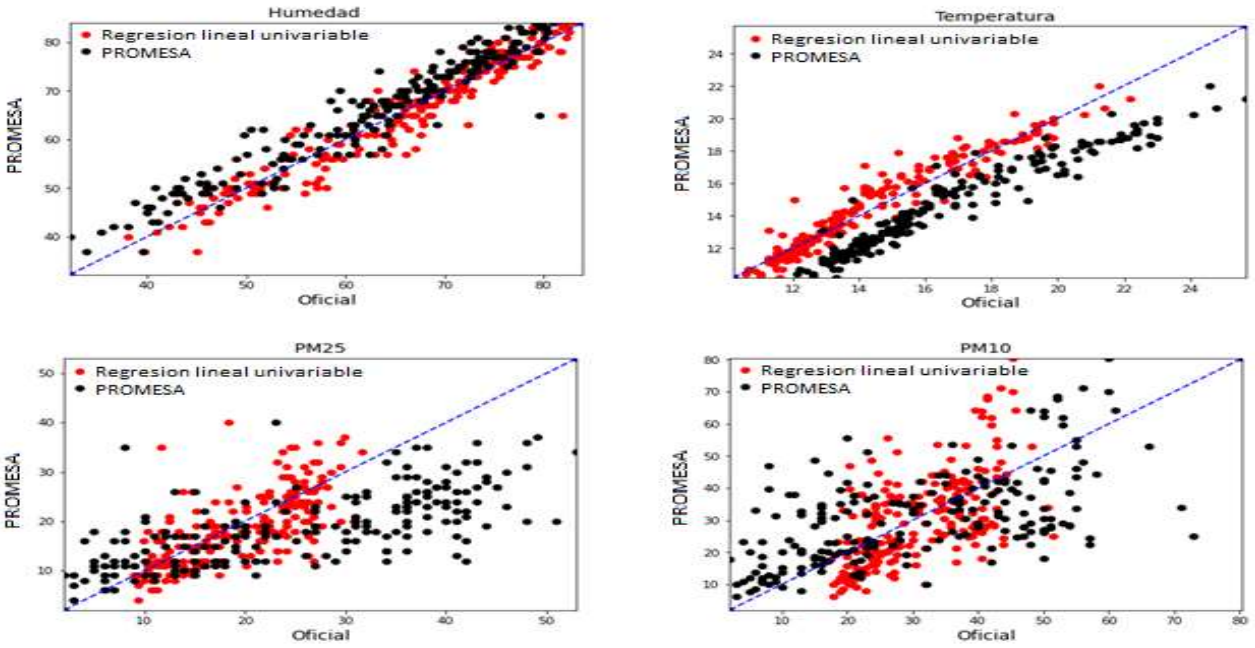

Figura 4. Diagrama dispersión para las diferentes variables antes y después de aplicar el modelo de Regresión Lineal Univariable.

Tabla 6.
Parámetros óptimos para las técnicas de Bosque Aleatorio univariable y multivariable.

	Univariable			Multivariable		
Variable	max depth	max features	max leaf nodes	max depth	max features	max leaf nodes
Humedad	3	SQRT	9	3	-	9
$PM_{2.5}$	3	NONE	6	6	LOG2	9
PM_{10}	3	SQRT	6	6	SQRT	9
Temperatura	3	SQRT	9	6	NONE	9

Tabla 7.
Coeficiente de determinación para las diferentes técnicas para todas las variables.

	PROMESA	RLU	RLM	BAU	BAM
Humedad	0.86	0.92	0.95	0.93	0.95
PM_{25}	-1.57	0.47	0.55	0.60	0.63
PM_{10}	-0.02	0.35	0.57	0.58	0.65
Temperatura	0.34	0.92	0.94	0.93	0.95

A partir de los resultados se observa lo siguiente:

- Para la humedad, se obtiene una mejora tanto con el modelo de regresión lineal multivariable como con el de Bosque Aleatorio Multivariable.
- Para la temperatura se tiene una mejora sustancial al aplicar cualquier modelo de aprendizaje automático, sin embargo, el mejor coeficiente de determinación se obtiene con el modelo de Bosque Aleatorio Multivariable.
- Para $PM_{2.5}$, PM_{10} se obtienen los mejores valores con el modelo de Bosque Aleatorio Multivariable, sin embargo, en ambos casos el valor obtenido es de 0.6.

En general puede concluirse que con el modelo de Bosque Aleatorio Multivariable se obtienen los mejores resultados.

5. Conclusiones

Es evidente que, por el aumento en la contaminación ambiental, su impacto en la salud pública ha sido notorio por la creciente atención de pacientes por enfermedades cardiacas y respiratorias. Las naciones han respondido a esta problemática con la puesta en funcionamiento de redes de monitoreo ambiental, las cuales, al ser costosas, su uso no se ha podido extender hasta abarcar todo el territorio.

El uso de sensores de bajo costo ha aumentado en todo el mundo, no solo por sus costos sino por su practicidad en la puesta de su funcionamiento. Ahora bien, en diversos estudios e investigaciones, se ha resaltado la necesidad de calibrar sus mediciones a través de modelos de entrenamiento, para equiparar sus resultados a los de las estaciones oficiales.

En los sensores de bajo costo de las estaciones de PROMESA, son necesarios los ajustes en el horario de las mediciones, porque es evidente desfases en los gráficos de señales de tiempo de las variables estudiadas.

Como conclusión de este proceso de validación, puede determinarse que el mejor desempeño se obtiene con el modelo Random forest multivariable luego de ajustar los hiperparámetros, obteniendo una mejora sustancial en la calidad de las mediciones para PM2.5 y PM10 de las estaciones LCS al comparar las mediciones con estaciones oficiales. Se recomienda realizar pruebas con todas las otras estaciones PROMESA.

6. Referencias bibliográficas

Ali, S., Alam, F., Arif, K. M., & Potgieter, J. (2023). Low-Cost CO Sensor Calibration Using One Dimensional Convolutional Neural Network. Sensors, 23(2). https://doi.org/10.3390/s23020854

Arku, R. E., Birch, A., Shupler, M., Yusuf, S., Hystad, P., & Brauer, M. (2018). Characterizing exposure to household air pollution within the Prospective Urban Rural Epidemiology (PURE) study. Environment International, 114, 307–317. https://doi.org/10.1016/j.envint.2018.02.033

Chakraborty, S., Dey, T., Jun, Y., Lim, C. Y., Mukherjee, A., & Dominici, F. (2022). A Spatiotemporal Analytical Outlook of the Exposure to Air Pollution and COVID-19 Mortality in the USA. Journal of Agricultural, Biological, and Environmental Statistics, 27(3), 419–439. https://doi.org/10.1007/s13253-022-00487-1

Chojer, H., Branco, P. T. B. S., Martins, F. G., Alvim-Ferraz, M. C. M., & Sousa, S. I. V. (2022). Can data reliability of low-cost sensor devices for indoor air particulate matter monitoring be improved? – An approach using machine learning. Atmospheric Environment, 286. https://doi.org/10.1016/j.atmosenv.2022.119251

Cifuentes Martínez, P., Rodríguez-Fernández, A., Luengo, C., & Tapia, L. (n.d.). Relación entre contaminación atmosférica y consultas por enfermedades respiratorias en atención primaria de urgencia.

Day, R. F., Yin, P. Y., Huang, Y. C. T., Wang, C. Y., Tsai, C. C., & Yu, C. H. (2022). Concentration-Temporal Multilevel Calibration of Low-Cost PM2.5 Sensors. Sustainability (Switzerland), 14(16). https://doi.org/10.3390/su141610015

de Moraes, S. L., Almendra, R., Santana, P., & Galvani, E. (2019). Meteorological variables and air pollution and their association with hospitalizations due to respiratory diseases in children: A case study in São Paulo, Brazil. Cadernos de Saude Publica, 35(7). https://doi.org/10.1590/0102-311x00101418

Ghamari, M., Kamangir, H., Arezoo, K., & Alipour, K. (2022). Evaluation and calibration of low-cost off-the-shelf particulate matter sensors using machine learning techniques. IET Wireless Sensor Systems, 12(5–6), 134–148. https://doi.org/10.1049/wss2.12043

Hashmy, Y., Khan, Z., Hafiz, R., Younis, U., & Tauqeer, T. (2021). MAQ-CaF: A Modular Air Quality Calibration and Forecasting method for cross-sensitive pollutants. http://arxiv.org/abs/2104.12594

Liang, L., & Daniels, J. (2022). What Influences Low-cost Sensor Data Calibration? - A Systematic Assessment of Algorithms, Duration, and Predictor Selection. Aerosol and Air Quality Research, 22. https://doi.org/10.4209/aaqr.220076

Malyan, V., Kumar, V., & Sahu, M. (2023). Significance of sources and size distribution on calibration of low-cost particle sensors: Evidence from a field sampling campaign. Journal of Aerosol Science, 168. https://doi.org/10.1016/j.jaerosci.2022.106114

Morakinyo et al. (2016) – Morakinyo, O., Mokgobu, M., Mukhola, M. & Hunter, R. Resultados de salud de la exposición a componentes biológicos y químicos de partículas inhalables y respirables. En t. J. Medio Ambiente. Res. Salud Pública 13, 592 (2016). https://doi.org/10.3390/ijerph13060592

Niewiadomska, E., Kowalska, M., Niewiadomski, A., Skrzypek, M., & Kowalski, M. A. (2020). Assessment of risk hospitalization due to acute respiratory incidents related to ozone exposure in silesian voivodeship (Poland). International Journal of Environmental Research and Public Health, 17(10). https://doi.org/10.3390/ijerph17103591.

Oyana, T. J., Podila, P., & Relyea, G. E. (2019). Effects of childhood exposure to PM<inf>2.5</inf> in a Memphis pediatric asthma cohort. Environmental Monitoring and Assessment, 191. https://doi.org/10.1007/s10661-019-7419-y

Parra Sánchez, J. S., Oviedo Carrascal, A. I., & Amaya Fernández, F. O. (2020). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia). Revista de Salud Pública, 22(6), 1–9. https://doi.org/10.15446/rsap.v22n6.78985

Patton, A., Datta, A., Zamora, M. L., Buehler, C., Xiong, F., Gentner, D. R., & Koehler, K. (2022). Non-linear probabilistic calibration of low-cost environmental air pollution sensor networks for neighborhood level spatiotemporal exposure assessment. Journal of Exposure Science and Environmental Epidemiology, 32(6), 908–916. https://doi.org/10.1038/s41370-022-00493-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research pp. 2825-2830. https://scikit-learn.org/stable/modules/model_evaluation.html

Peng, W., Li, H., Peng, L., Wang, Y., & Wang, W. (2022). Effects of particulate matter on hospital admissions for respiratory diseases: an ecological study based on 12.5 years of time series data in Shanghai. Environmental Health: A Global Access Science Source, 21(1). https://doi.org/10.1186/s12940-021-00828-6

Scagliotti, A., & Jorge, G. (2021). Caracterización de sensor de material particulado de bajo costo. Revista Tecnología y Ciencia, 42, 96–111. https://doi.org/10.33414/rtyc.42.96-111.2021

Sepadi, M. M., & Nkosi, V. (2021). A study protocol to assess the respiratory health risks and impacts amongst informal street food vendors in the inner city of johannesburg, South Africa. International Journal of Environmental Research and Public Health, 18(21). https://doi.org/10.3390/ijerph182111320

Wrotek, A., Badyda, A., Czechowski, P. O., Owczarek, T., Dąbrowiecki, P., & Jackowska, T. (2021). Air pollutants' concentrations are associated with increased number of rsv hospitalizations in polish children. Journal of Clinical Medicine, 10(15). https://doi.org/10.3390/jcm10153224