Prediction of PM2.5 and PM10 Concentrations Using XGBoost and LightGBM Algorithms: A Case Study in Lima, Peru

Johan Andrés Oblitas Mantilla
20191412@aloe.ulima.edu.pe
https://orcid.org/0009-0007-2183-5583
Universidad de Lima, Peru

Edwin Jhonatan Escobedo Cárdenas
eescobed@ulima.edu.pe
https://orcid.org/0000-0003-2034-513X
Universidad de Lima, Peru

Received: September 3th, 2024 / Accepted: October 12th, 2024
doi: https://doi.org/10.26439/interfases2024.n020.7417

ABSTRACT. Air pollution is a major problem that affects both human health and the environment, causing millions of premature deaths annually worldwide and severely degrading the state of the planet. Exposure to fine particulate matter, which is highly hazardous, enables these particles to penetrate deeply into the lungs and lead to serious health issues, including a reduction in life expectancy by more than two years. In response to this problem, it is crucial to identify effective ways to monitor the levels of these pollutants in our daily surroundings. This article presents a case study conducted in the district of San Borja, Lima, Peru, where prediction models for PM2.5 and PM10 were implemented using the XGBoost and LightGBM algorithms. Employing data from the SENAMHI portal and a correlation analysis of variables, two different scenarios were developed for training the models. In scenario 1, prediction models for PM2.5 and PM10 were trained using all available meteorological and pollution variables. In scenario 2, the models were trained for PM2.5 excluding the PM10 variable, and vice versa. The results showed that both models achieved high accuracy, measured by the coefficient of determination, with no statistically significant difference indicating the superiority of either model. Furthermore, the analysis of the proposed scenarios revealed that excluding key variables can result in significantly less accurate predictions, potentially undermining the effectiveness of environmental management strategies.

KEYWORDS: air pollution / air quality / meteorological data / machine learning / XGBoost / LightGBM

PREDICCIÓN DE CONCENTRACIONES DE PM2.5 Y PM10 UTILIZANDO LOS ALGORITMOS XGBOOST Y LIGHTGBM: UN ESTUDIO DE CASO EN LIMA, PERÚ

RESUMEN. La contaminación del aire es un problema importante que afecta tanto a la salud humana como al medio ambiente, causando millones de muertes prematuras anualmente en todo el mundo y degradando severamente el estado del planeta. La exposición a material particulado fino, altamente peligroso, permite que estas partículas penetren profundamente en los pulmones y provoquen problemas de salud graves, incluyendo una reducción en la esperanza de vida de más de dos años. En respuesta a este problema, es crucial identificar formas efectivas de monitorear los niveles de estos contaminantes en nuestro entorno diario. Este artículo presenta un estudio de caso realizado en el distrito de San Borja, Lima, Perú, donde se implementaron modelos de predicción para PM2,5 y PM10 utilizando los algoritmos XGBoost y LightGBM. Empleando datos del portal del SENAMHI y un análisis de correlación de variables, se desarrollaron dos escenarios diferentes para el entrenamiento de los modelos. En el escenario 1, se entrenaron modelos de predicción para PM2,5 y PM10 utilizando todas las variables meteorológicas y de contaminación disponibles. En el escenario 2, los modelos se entrenaron para PM2,5 excluyendo la variable PM10, y viceversa. Los resultados mostraron que ambos modelos lograron una alta precisión, medida por el coeficiente de determinación, sin diferencias estadísticamente significativas que indicaran la superioridad de alguno de los modelos. Además, el análisis de los escenarios propuestos reveló que excluir variables clave puede resultar en predicciones significativamente menos precisas, lo que podría comprometer la efectividad de las estrategias de gestión ambiental.

PALABRAS CLAVE: contaminación del aire / calidad del aire / datos meteorológicos / aprendizaje automático / XGBoost / LightGBM

1. INTRODUCTION

Air pollution is a global problem that affects both human health and the environment. According to the World Health Organization (WHO, 2022), the combined effects of ambient and household air pollution are responsible for approximately 6,7 million premature deaths annually worldwide. Furthermore, WHO data reveals that 99 % of the global population breathes air with pollutant concentrations exceeding the levels established by WHO guidelines, with low- and middle-income countries being the most affected.

Exposure to fine particulate matter —one of the most harmful air pollutants for human health— enables these particles to penetrate deeply into the lungs, triggering reactions on lung surfaces and in defense cells, according to the Pan American Health Organization (PAHO, 2016). As mentioned by Sloss et al. (2000), PM10 and PM2.5 refer to particulate matter with diameters of 10 microns or less and 2.5 microns or less, respectively. These particles originate from various chemical species emitted by both natural and human sources, including coal-fired power plants, industrial activities, and road transport. They can be emitted directly or formed through atmospheric chemical reactions. Increased concentrations of this particulate matter pose a serious threat to human health.

The Ministerio del Ambiente (MINAM – Ministry of the Environment of Peru) has acknowledged the severity of air pollution in the country, noting that mobile sources, mainly vehicles, account for 58 % of particulate matter emissions, followed by stationary sources (26 %), and area sources (16 %) (MINAM, 2019).

Studies by MINAM (2021) and WHO (2021) highlight that prolonged exposure to fine particles can reduce life expectancy by more than two years. Addressing this issue requires the development of accurate air quality prediction methods that can help mitigate its effects. In this context, there is a growing need to explore advanced approaches that leverage detailed meteorological data and machine learning (ML) techniques. As stated by Samad et al. (2023) and Xing et al. (2020), machine learning models have recently shown strong accuracy in predictions and have gained widespread use. These models offer significant benefits over traditional approaches, as they are both cost-effective and computationally efficient. Research on estimating pollutant concentrations remains highly active, with efforts focused on reducing reliance on sensors and networks by using approximate predictions. Thus, various machine learning models have been applied for this purpose.

Researchers like Wang et al. (2023), Amuthadevi et al. (2021), and Gryech et al. (2020) have reported improvements in pollution estimation accuracy by through the incorporation of ng more detailed meteorological data, such as wind speed and wind direction. They also further suggest that, in the presence of large datasets, deep

learning models should be prioritized over traditional ML models when working with large datasets. According to Cordova et al. (2021), while several studies have been focusing on applying machine learning methods to forecast air quality in large cities, there is a limited number of studies of such research in the context of Lima, Peru, a city that ranks among the most populated in South America.

In this research, the use of XGBoost and LightGBM, both advanced implementations of the Gradient Boosting Decision Tree (GBDT) algorithm, is crucial for accurately predicting PM2.5 and PM10 concentrations (Ma et al., 2020). Unlike traditional machine learning algorithms, such as logistic regression, which are limited to linear regression problems, GBDT algorithms can effectively tackle a wide range of regression and binary classification tasks, making them versatile tools in this context (Friedman et al., 2001). XGBoost stands out for its efficient optimization techniques that enhance performance while requiring fewer computational resources, thus facilitating superior results in complex datasets (Chen et al., 2016). Meanwhile, LightGBM employs a unique leaf-wise growth strategy that allows it to reduce loss more effectively than level-wise algorithms, making it particularly adept at handling large datasets with lower memory consumption (Narayani et al., 2020). The combination of these features makes XGBoost and LightGBM suitable for air quality prediction tasks, enabling the development of robust models that leverage detailed meteorological data while maintaining computational efficiency.

Based on the previous context, a case study is proposed to predict fine particles PM2.5 and PM10 using the XGBoost and LightGBM algorithms. Data will be collected through the Servicio Nacional de Meteorología e Hidrología (SENAMHI – National Meteorology and Hydrology Service of Peru) web portal, specifically from the San Borja station. Using this data and two scenarios developed from a correlation analysis, prediction models will be implemented and evaluated using different performance metrics to conduct a comparative analysis.

This research article adopts a classical structure, beginning with the state of the art, where key variables in pollution estimation are reviewed, various prediction algorithms are explored, and relevant data processing techniques are discussed. The methodology focuses on the study area in San Borja, using SENAMHI datasets that include meteorological and pollutant data, refined and integrated to identify significant correlations. During the experimentation phase, the integrated dataset is normalized and divided to apply methods such as XGBoost and LightGBM, evaluating the models according to specific criteria. The results are discussed in detail in relation to the hypotheses posed, and conclusions summarize the study's key findings. Finally, future research is proposed, addressing potential methodological improvements or new areas of application.

2. STATE OF THE ART

The state of the art provides an updated and comprehensive overview of the study topic. In this regard, this section presents a structured analysis of the existing literature. First, the databases commonly used in this field are described, detailing their names, variables, number of records, frequency, location, and data collection periods. Subsequently, ML algorithms applied in various research articles are discussed, including the metrics used and their respective results.

2.1 Identified Variables in Pollution Estimation

The relevance of meteorological data in developing prediction models has been well established. For example, temperature and solar radiation influence the formation of ground-level ozone, while wind direction and wind speed affect the dispersion of pollutants and their impact across different geographical areas. Moreover, precipitation significantly contributes to the dispersion and removal of atmospheric pollutants, as it washes away particles and clears the air. Additionally, relative humidity plays a role in the reactivity and chemical transformation of pollutants (Zhang et al., 2020; Gryech et al., 2020; Ameer et al., 2019).

Furthermore, besides finding a strong correlation between PM10, PM2.5, nitrogen dioxide (NO2), and carbon monoxide (CO) concentrations and meteorological conditions, a significant seasonal correlation was observed between atmospheric pollutants and temperature. Specifically, during winter, the concentrations of these pollutants were found to be double those recorded in summer (Gryech et al., 2020). However, some studies have analyzed the importance of incorporating data related to vehicular flow when predicting air quality. According to Gryech et al. (2020), greater accuracy can be achieved by combining meteorological data with traffic-related data to estimate unmeasured pollutant concentrations.

Additionally, Sulaimon et al. (2022) conducted a research experiment where several air pollution prediction models were trained based on different ML algorithms. In scenario 1, only air quality and meteorological data was used to process the dataset, whereas in scenario 2, an experimental dataset that included traffic data was employed. The results consistently showed that models trained with the experimental dataset outperformed those trained with the control dataset. A performance improvement of at least 20 % and an error reduction of at least 18.97 % were observed in 98 % of the ML algorithms when trained with a dataset containing traffic-related information. These findings underscore the significant impact of traffic data on improving the performance of ML-based air pollution prediction models (Sulaimon et al., 2022).

For this study, datasets provided by SENAMHI, which include pollutant and meteorological data, were selected. These datasets are particularly relevant because they provide specific information about the study area of interest—in this case, Lima, Peru.

2.2 Prediction Algorithms

Amuthadevi et al. (2021) mentioned that the meteorological dataset used in their research was collected over five years and used to train four different prediction models: non-linear artificial neural network (ANN), statistical multilevel regression (SMR), neuro-fuzzy systems (neuro-fuzzy), and deep learning long short-term memory (DL-LSTM). When comparing these models, the last one (DL-LSTM) was found to be particularly suitable for analyzing and predicting air pollutant concentrations, as it achieved lower root mean square error (RMSE) and mean absolute percentage error (MAPE) (0.1268 and 9.475, respectively), while demonstrating better correlations with the test data.

On the other hand, Zhang et al. (2020) achieved positive results with their proposed distributed hybrid system of fixed and IoT sensors for predicting air quality. When comparing the applied algorithms—support vector regression (SVR), random forest regression (RFR), and gradient boosting regression (GBR)—the latter yielded the best results, with an RMSE of 13.8375 for PM10 predictions and 11.225 for PM2.5 predictions.

In contrast, Gryech et al. (2020) used accuracy as the evaluation metric to compare the techniques employed, with the random forest (RF) algorithm achieving high accuracy: 94 %, 97 %, and 98 % for pollutants NO2, PM10, and PM2.5, respectively. Additionally, they demonstrated that certain pollutants can be accurately predicted based on the concentrations of other pollutants. A key finding was that NO2 concentration could be predicted based on PM10 concentration and vice versa. They concluded that the RF algorithm is highly flexible and shows less variation compared to individual decision trees.

For Gokul et al. (2023), the gradient-boosting regression model (XGBoost) emerged as the ML algorithm that delivered the best results, obtaining a mean absolute error (MAE) of 7.01, a mean squared error (MSE) of 93.55, and an RMSE of 9.67. Sulaimon et al. (2022), on the other hand, achieved a performance improvement of at least 20 % and an error reduction of 18.97 % by including traffic flow-related data into the training of different ML models. They also observed varying behaviors and performances across the algorithms, with no single algorithm consistently yielding the best results overall. Their findings were primarily influenced by variations in the study areas and the combinations of datasets used.

Finally, according to Liu et al. (2023), the LightGBM model was employed to predict the atmospheric concentration of PM2.5 by optimizing its parameters. After analyzing monitoring data from various times and regions, the prediction curve of haze concentration indicated that the model exhibited strong learning capability, a high degree of fit, and significant improvement in both accuracy and stability. It has been demonstrated that optimization with this algorithm significantly enhances precision. The improved model is highly applicable in practice and is suitable for forecasting PM2.5 concentrations, offering valuable insights for predicting trends in air quality changes.

2.3 Data Processing Techniques

Several data processing techniques frequently used in the analyzed research can be identified. First, data cleaning involves detecting and correcting errors, outliers, or missing data to improve the quality and reliability of the dataset. It remains one of the most widely used techniques across various datasets (Gryech et al., 2020; Cordova et al., 2021; Gokul et al., 2023; Ayus et al., 2023). This technique includes several sub-techniques, such as outlier detection and data interpolation, among others.

On the other hand, data integration is a technique that involves combining multiple data sources into a single, coherent dataset. This approach provides a more holistic view of the data by combining different variables and features, and its application is considered essential for effective data processing (Gokul et al., 2023; Gryech et al., 2020; Sulaimon et al., 2022).

In the studies by Zhang et al. (2020), Liang et al. (2020), and Ayus et al. (2023), data normalization is commonly highlighted as a method used to scale variable values within a specific range, facilitating the comparison and analysis of variables within the same context.

Several key findings from the research on atmospheric pollution estimation stand out. The critical role of meteorological data —especially variables such as temperature, wind direction, wind speed, and precipitation in the formation and dispersion of pollutants— has been demonstrated. Additionally, a strong seasonal correlation between atmospheric pollutants and temperature has been observed, underscoring the significant impact of weather conditions on air quality. The studies have also highlighted the need to integrate multiple data sources to achieve more accurate and robust results. These findings underline the complexity and multidimensionality of the atmospheric pollution problem, providing a solid foundation for the implementation of advanced ML models in future research.

3. METHODOLOGY

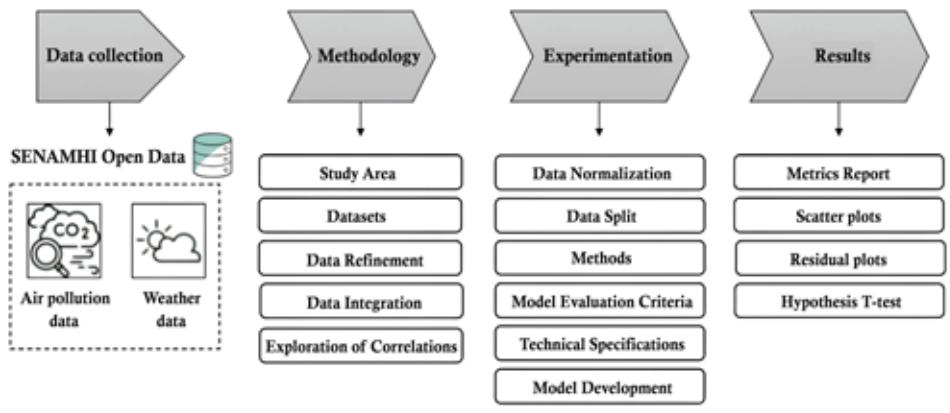
This section outlines the key procedures derived from a review of the aforementioned literature. Based on the research by Sulaimon et al. (2022), Gryech et al. (2020), Ayus et al. (2023), Yang et al. (2018), and Ameer et al. (2019), the methodology began with a systematic review of the state of the art in air quality prediction, followed by the data collection, which encompassed a wide range of pollutant concentrations and meteorological data. A comprehensive exploratory analysis was performed on these datasets, including a descriptive analysis, evaluation of missing data, and exploration of correlations between variables.

The data preprocessing phase included integrating, transforming, cleaning, and normalizing the extracted datasets, followed by the variable selection process. Finally,

prediction models were developed, and the performance of the trained models was evaluated through a comparative analysis.

Figure 1

Methodological Design



3.1 Study Area

The study area for training and testing the proposed approach is located in the province of Lima, Peru, specifically in the district of San Borja, as shown in Figure 2. Given its high traffic density and diverse lane distribution, San Borja is considered an ideal location for this study, as it also houses a variety of regulatory and commercial facilities. The air quality monitoring station is managed by SENAMHI, under the Dirección de Redes de Observación y Datos (Directorate of Observation Networks and Data). This station has been operational since May 26, 2010, and is located at coordinates 12.10859 latitude and 77.00769 longitude.

The data collected from the San Borja station spans from the start of its operation to the present date. It is important to note that this data is presented in its raw form, meaning it has not undergone any quality control process. Therefore, applying necessary corrections is deemed essential to minimize any negative impact on the results.

The dataset comprises 11 input variables, including temperature, relative humidity, wind direction, wind speed, precipitation, CO, nitrogen oxides (NOx), ozone (O3), sulfur dioxide (SO2), and particulate matter (PM2,5 and PM10).

Figure 2

Map Showing the Study Area and the Location of the Air Quality Monitoring Station in San Borja

3.2 Datasets

Two datasets were examined, both describing pollutant concentrations and meteorological conditions for the same period and area. The data was recorded through the San Borja monitoring station and provided by SENAMHI (2024), with public access via their web portal.

The first dataset includes hourly observations from January 1 to April 30, 2024, measuring atmospheric pollutant concentrations, expressed in micrograms per cubic meter ($\mu g/m^3$) for PM2.5, PM10, NO2, O3, and CO. It consists of a total of 2810 records.

The second dataset contains meteorological variables such as temperature (in degrees Celsius), precipitation (millimeters per hour), relative humidity (percentage), wind direction (degrees), and wind speed (meters per second). It includes 2841 records.

3.3 Data Refinement

Based on the work of Sulaimon et al. (2022), data refinement was applied, which involves handling missing data and cleaning outliers. In the pollutant dataset, four missing records were identified, specifically in the NO2 variable. Similarly, the meteorological dataset showed missing values for the temperature and humidity variables, with one missing record for each. Given the small number of missing records relative to the total dataset size, these records were removed to maintain data consistency.

Additionally, it was observed that 98.9 % of the records for the *precipitation* variable had a value of zero. This unusually high proportion suggests a possible anomaly in data capture, potentially indicating a malfunction of the measurement device during the observation period. Therefore, the *precipitation* column was removed to prevent misinterpretation or bias in further analyses.

For outlier detection, the multidimensional outlier detection method was employed, specifically the local outlier factor (LOF) method, which is widely recognized for its ability to identify anomalous points in multidimensional datasets. In this case, a contamination parameter of 0.1 was used.

3.4 Data Integration

According to Sulaimon et al. (2022), the use of multiple datasets requires a data integration process to ensure the information can be accessed through a unified repository. For this purpose, the datasets were processed and integrated. The common reference points were the date and time features. Below, Table 1 shows an initial fragment of the integrated dataset, while Table 2 provides a description of each feature.

Table 1

Header of the Integrated Dataset

N.º	PM10	PM2,5	NO2	O3	CO	Temperature	Humidity	Wind Direction	Wind Speed
0	31.0	14.4	15.3	11.6	332.4	21.9	85.0	168	2.1
1	32.0	14.9	11.8	12.6	226.8	21.5	87.0	184	1.3
2	25.8	10.8	9.6	15.3	242.7	22.0	83.0	183	0.8
3	23.7	11.8	9.2	35.9	181.7	26.2	67.0	276	1.0
4	25.5	12.7	11.2	39.1	188.6	26.6	68.0	210	0.7

Table 2

Data Dictionary

Variable	Unit	Data Type	Description
PM2.5	$\mu g/m^3$	Float	Contains numerical data represented as floating-point values with units of micrograms per cubic meter ($\mu g/m^3$). This feature measures the concentration of fine particles in the air with a diameter of 2.5 micrometers or less.
PM10	$\mu g/m^3$	Float	Contains numerical data represented as floating-point values with units of micrograms per cubic meter ($\mu g/m^3$). This feature measures the concentration of fine particles in the air with a diameter of 10 micrometers or less.
NO2	$\mu g/m^3$	Float	Contains numerical data represented as floating-point values with units of micrograms per cubic meter ($\mu g/m^3$). This feature measures the concentration of nitrogen dioxide in the air.
O3	$\mu g/m^3$	Float	Contains numerical data represented as floating-point values with units of micrograms per cubic meter ($\mu g/m^3$) or parts per million (ppm). It measures the concentration of ozone in the air.

(*continúa*)

(*continuación*)

Variable	Unit	Data Type	Description
CO	$\mu g/m^3$	Float	Contains numerical data represented as floating-point values with units of micrograms per cubic meter ($\mu g/m^3$) or parts per million (ppm). It measures the concentration of carbon monoxide in the air.
Temperature	Degrees Celsius	Float	Stores numerical data in floating-point format with units of degrees Celsius (°C). These values represent the ambient temperature at the time of measurement.
Humidity	Percentage	Float	Recorded as floating-point numerical values with units of percentage (%). It indicates the level of relative humidity in the air at the time of measurement.
Wind Direction	Degrees	Float	Stored as floating-point numerical data with units of degrees (°). This feature indicates from where the wind is coming at the time of measurement.
Wind Speed	m/s	Float	Contains numerical data in floating-point format with units of meters per second (m/s). It represents the wind speed at the time of measurement.

3.5 Exploration of Correlations

Based on the work of Bai et al. (2016), an analysis of correlations between pollutant variables and meteorological variables was conducted, revealing several significant relationships that provide valuable insights into the factors influencing air quality. First, PM10 and PM2.5 were highly correlated with each other (0.8), indicating a strong association in their concentration levels. This was expected, as both are particles of similar size and share common emission sources, such as fossil fuel combustion and industrial activities.

On the other hand, NO2 showed a relatively high correlation with CO at 0.69. This is because both substances are emitted during combustion processes and are influenced by similar factors, such as traffic density and meteorological conditions.

Wind direction and wind speed showed moderate correlations with other variables. This indicates that wind conditions can influence the dispersion of pollutants in the air. Stronger winds or favorable wind directions are likely to promote greater dispersion of pollutants, while weak winds may result in the accumulation of pollutants in a specific area.

Finally, a high negative correlation was found between temperature and humidity (-0.89). This inverse relationship is consistent with basic principles of physics, where the air's capacity to hold moisture decreases as temperature increases. This result underscores the important role of temperature and humidity in the formation and dispersion of atmospheric pollutants.

Table 3

Correlation Map Between Pollution and Meteorological Variables

Variables	PM10	PM2.5	NO2	O3	CO	T	H	WD	WS
PM10	1	0.79	0.60	-0.43	0.72	-0.03	-0.15	0.11	-0.15
PM2,5		1	0.49	-0.30	0.61	-0.07	-0.01	0.14	-0.27
NO2			1	-0.14	0.69	0.30	-0.39	0.18	0.03
O3				1	-0.28	0.58	-0.42	0.15	0.21
CO					1	0.17	-0.28	0.20	-0.17
T						1	-0.89	0.24	0.19
H							1	-0.24	-0.22
WD								1	-0.37
WS									1

Note. T represents temperature (°C), H represents humidity (%), WD represents wind direction (°), and WS represents wind speed (m/s).

Based on the analysis, two scenarios were created for the experimental phase of this study. In scenario 1 (S1), prediction models for PM2.5 and PM10 were trained using all available meteorological and pollution variables. In scenario 2 (S2), prediction models were developed separately for PM2.5, excluding the PM10 variable, and for PM10, excluding the PM2.5 variable. This approach aims to assess the impact that each variable has on the training of the other.

4. EXPERIMENTATION

Below is a detailed explanation of the workflow involved in the development of the prediction models.

- **Dataset:** The data was collected hourly at the San Borja meteorological monitoring station (SENAMHI) from January 1 to April 30, 2024. The integrated dataset included two categories —meteorological and pollution data— used as input variables. PM2.5 and PM10 were considered as the output variables.
- **Normalization:** As shown in Table 4, the dataset contains values with varying ranges. Therefore, the data is normalized. Normalization involves adjusting the scale of the data from its original range to a range between 0 and 1. Table 4 presents detailed statistics of the input and output variables used in this study. For this case, the standard scaling strategy was employed, with each value in the dataset normalized as follows:

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

where z represents the scaled value, x is the original value, μ is the mean, and σ represents the standard deviation of the dataset.

- **Data Split:** According to research by Shakya et al. (2023), Liu et al. (2023), Ayus et al. (2023), Zhang et al. (2023), and Liang et al. (2020), the dataset was divided into training and testing sets, with 80 % of data used for training and the remaining 20 % for testing, following the Pareto principle, as noted by one of the authors. Table 4 presents the dataset statistics used to predict PM2.5 concentrations.
- **Methods:** This study used conventional ML methods, such as XGBoost and LightGBM, to predict PM2.5 concentrations in San Borja. The results from these models were compared to assess their accuracy and effectiveness.
- **Model Evaluation Criteria:** Shakya et al. (2023), Martín-Baos et al. (2022), and Pan (2018) used different metrics to evaluate their models. In this study, performance metrics such as coefficient of determination (R^2), RMSE, relative root mean square error (RRMSE), and MAPE were applied. The descriptions, formulas, and ranges of performance metrics are shown in Table 5.
- **Technical Specifications:** Python was chosen for the experimental phase due to its extensive access to key libraries such as Pandas, NumPy, and Scikit-learn, as well as other libraries specific to the ML algorithms applied in this work.

Table 4

Data Statistics

Dataset	Variables	Minimum	Maximum	Mean	Standard Deviation	25 %	50 %	75 %
Target Data	PM2.5	6.70	26.10	15.11	3.6842	12.40	14.70	17.40
	PM10	17.50	109.20	51.45	19.2692	37.10	47.40	63.80
Pollutants	NO2	4.50	41.80	21.02	7.0858	15.80	20.85	25.90
	O3	4.00	45.40	16.37	9.6222	6.60	15.30	23.80
	CO	101.20	1360.50	631.95	240.6815	458.90	632.50	810.80
Meteoro-logical	Temperature	18.50	31.50	24.36	2.5980	22.40	23.90	26.30
	Humidity	48.00	98.00	77.03	10.7499	68.00	79.00	86.00
	Wind Direction	104.00	310.00	203.44	36.2904	179.00	191.00	227.25
	Wind Speed	0.00	2.70	1.02	0.6063	0.60	0.90	1.40

Based on Table 4, relevant information can be observed. First, the variability of PM2.5, with a standard deviation of 3.6842, suggests significant daily fluctuations in air quality. It can also be noted that the median (14.70) is close to the mean (15.11), indicating a relatively symmetrical distribution of the data.

The presence of other pollutants —such as PM10, NO2, O3, and CO— along with their respective variations and standard deviations provides a framework for understanding potential correlations between these pollutants and PM2.5. For example, the high standard deviation of CO (240.6815) indicates significant fluctuations that could have a substantial impact on PM2.5 levels. On the other hand, meteorological conditions are crucial factors affecting the dispersion and concentration of pollutants. For instance, the temperature with a mean of 24.36 and a standard deviation of 2.5980 suggests some stability, but fluctuations can still influence PM2.5 levels. Wind speed, although low on average (1.02 m/s), shows variations that can disperse pollutants.

Table 5

Formulas, Descriptions, and Ranges of Performance Metrics

Metric	Formula	Description	Range
R^2	$\left(\frac{\sum_{i=1}^{n}(a_i-\bar{a})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{n}(p_i-\bar{p})^2}}\right)^2$	It is a widely used metric that indicates how well the trends of the model simulation match the trends of the actual data.	$0 \leq R^2 \leq 1$ (Higher values indicate better performance)
MAE	$\frac{1}{n}\sum_{i=1}^{n}(\lvert p_i-a_i\rvert)$	Quantifies the dispersion between the actual values and the predicted values.	$0 \leq MAE \leq \infty$ (Lower values indicate better performance)
RRMSE	$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i-p_i)^2}}{\bar{a}}$	Calculated by dividing the RMSE value by the mean of the actual values.	$0 \leq RRMSE \leq \infty$ (Lower values indicate better performance)
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{a_i-p_i}{a_i}\right\rvert$	It is a statistical measure that provides a benchmark for the accuracy of a prediction model.	Lower values indicate better performance

Note. a_i represents the actual values, p_i represents the predicted values, $\bar{a}$ represents the mean of the actual values, p represents the mean of the predicted values and n represents the number of observations (Shakya et al., 2023).

4.1 Model Development

For the training process of each model, 100 iterations were performed to adjust the hyperparameters and obtain optimal results. This adjustment was carried out using RandomizedSearchCV, an efficient random search technique that systematically explores the hyperparameter space. Each hyperparameter is defined within a specific range,

determining the values it can take during the random search process. The ranges used are shown in Table 6.

Table 6

Hyperparameter Values

Algorithm	Hyperparameter	Range of Values
XGBoost and LightGBM	n_estimators	Random integer between 20 and 150. This hyperparameter controls the number of estimators (trees) used in the model.
XGBoost	max_depth	Random integer between 3 and 12. This parameter determines the maximum depth of each tree in the model, influencing its complexity and ability to fit the data.
XGBoost and LightGBM	learning_rate	Random continuous value between 0.05 and 0.35. This learning rate controls the magnitude of updates to the model weights in each iteration, affecting the speed and accuracy of training.
XGBoost and LightGBM	colsample_by-tree	Random continuous value between 0.5 and 1.0. This parameter determines the percentage of features to consider for each tree during training, helping to control overfitting.
XGBoost and LightGBM	subsample	Random continuous value between 0.5 and 1.0. This hyperparameter specifies the percentage of samples (instances) to consider for each tree during training, helping to control overfitting and improve model generalization.
XGBoost and LightGBM	reg_alpha	Random continuous value between 0.05 and 9.95. This parameter controls the strength of L1 regularization (Lasso regression) in the model, helping to prevent overfitting by penalizing large coefficients.
LightGBM	reg_lambda	Random continuous value between 0.05 and 9.95. Similar to reg_alpha, this parameter controls the strength of L2 regularization (Ridge regression) in the LightGBM model.
XGBoost	min_child_weight	Random continuous value between 1 and 19. This hyperparameter sets the minimum weight required to create a new partition in a tree node during the growth process, influencing the complexity and structure of the final tree.
LightGBM	min_child_samples	Random integer between 20 and 50. This hyperparameter sets the minimum number of samples required to create a new partition in a tree node during the growth process, affecting the structure and complexity of the final tree.
LightGBM	num_leaves	Random integer between 31 and 100. This parameter determines the maximum number of leaves allowed in each tree of the model, affecting its complexity and adaptability.

5. RESULTS

This section presents an analysis of the results from the models developed to predict PM2.5 and PM10 concentrations. First, the performance of the models is evaluated using the mentioned metrics and considering the proposed scenarios informed by the correlation analysis. Next, residual and scatter plots related to these results are included. Finally, a paired t-test is conducted to statistically assess potential differences between the algorithms under study.

Table 7

Performance of XGBoost and LightGBM Models for PM2.5 and PM10 Predictions: A Scenario Comparison

		XGBoost				LightGBM			
		R^2	MAE	RRMSE	MAPE	R^2	MAE	RRMSE	MAPE
S1	PM2.5	0.75	1.6376	0.1454	10.0227	0.75	1.654	0.1474	10.1317
	PM10	0.84	6.6538	0.1595	12.2376	0.83	6.7814	0.1632	12.4313
S2	PM2.5	0.54	2.1142	0.1810	13.6036	0.55	2.1367	0.1810	13.7412
	PM10	0.69	8.9187	0.2156	17.1318	0.68	8.9814	0.2190	17.0944

Table 7 compares the performance of the XGBoost and LightGBM models in predicting PM2.5 and PM10 concentrations across the proposed scenarios. In S1, the results for PM2.5 prediction are quite similar, with both models achieving an R^2 of 0.75. However, XGBoost shows a slight advantage in some metrics, such as MAE, where it records a slightly lower value (1.6376) compared to LightGBM (1.654). Similarly, XGBoost exhibits slightly lower RRMSE and MAPE values, at 0.1454 and 10.0227, respectively, compared to 0.1474 and 10.1317 for LightGBM. On the other hand, for PM10 prediction, XGBoost demonstrates a slightly superior performance, with an R^2 value of 0.84 compared to 0.83 for LightGBM, Additionally, XGBoost records lower MAE (6.6538) and RRMSE (0.1595) than LightGBM, as well as a slightly better MAPE (12.2376) compared to LightGBM (12.4313).

In S2, a similar trend is noted. For PM2.5 prediction, LightGBM shows a slight improvement in the R^2 value (0.55) compared to XGBoost (0.54). The MAE and RRMSE values are almost identical for both models, with values around 2.1142 and 0.1810 for XGBoost, and 2.1367 and 0.1810 for LightGBM, respectively. The MAPE values follow a similar trend, at 13.6036 for XGBoost and 13.7412 for LightGBM. For PM10 prediction in this scenario, XGBoost has a slightly higher R^2 (0.69). The MAE, RRMSE, and MAPE values are very similar, with slight differences favoring XGBoost in MAE (8.9187) and RRMSE (0.2156).

At a macro level, both models demonstrate a marked decrease in predictive performance when excluding the PM10 variable for PM25 predictions, and vice versa. This suggests that including both variables is crucial for achieving more accurate predictions. The residual and scatter plots for each target variable and scenario are presented below.

Figure 3

S1 (PM2.5): Scatter and Residual Plots of the XGBoost and LightGBM Models

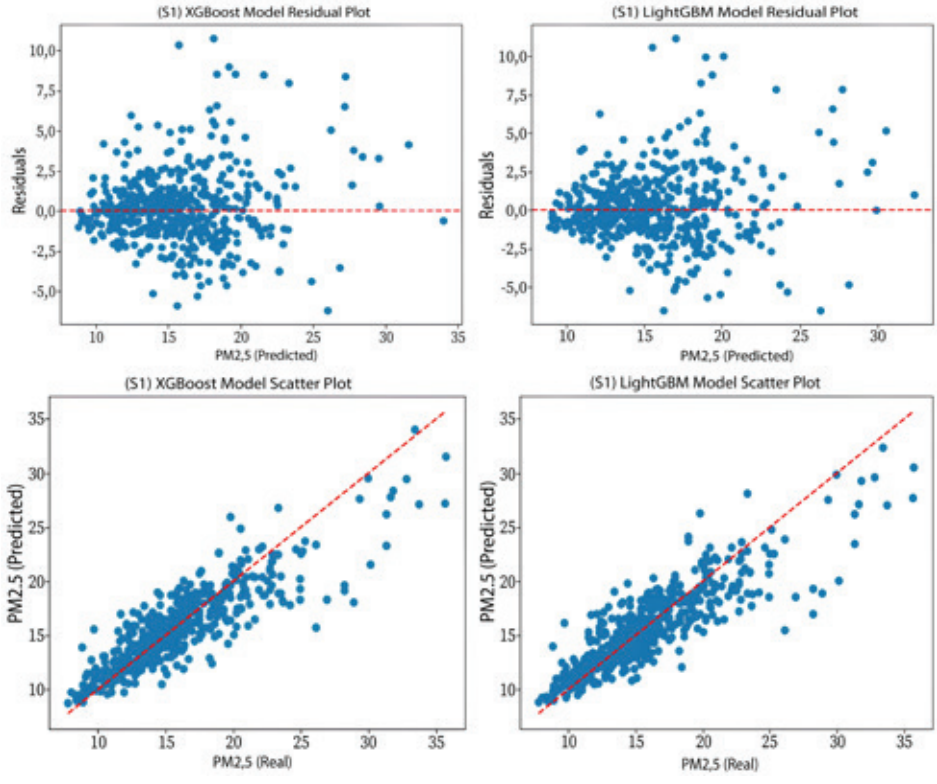


Figure 4

S1 (PM10): Scatter and Residual Plots of the XGBoost and LightGBM Models

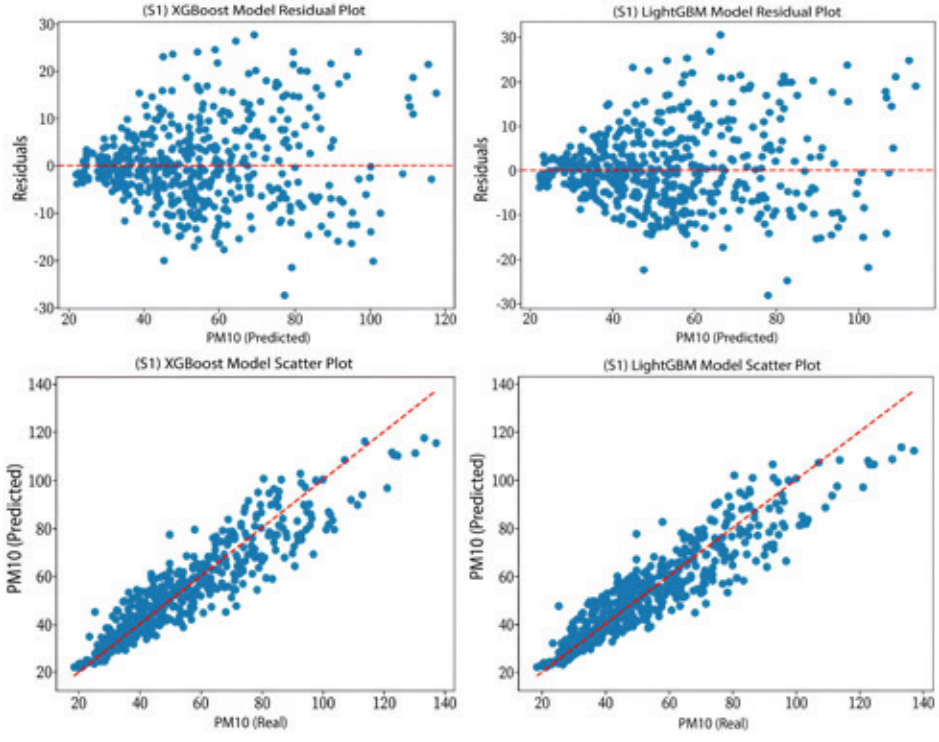


Figure 5

S2 (PM2.5): Scatter and Residual Plots of the XGBoost and LightGBM Models

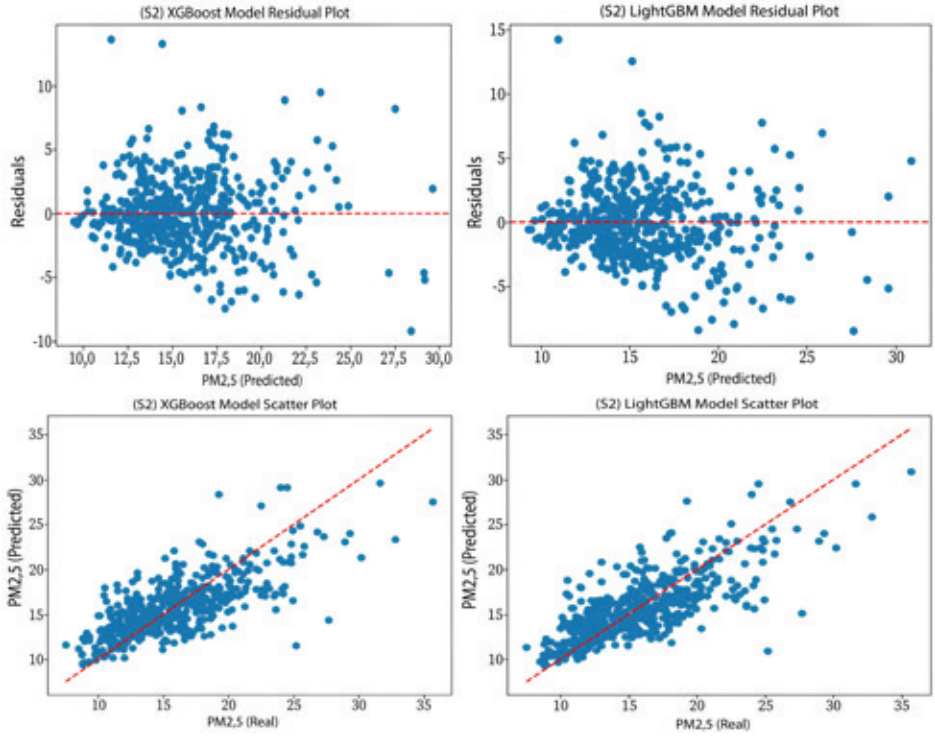


Figure 6

S2 (PM10): Scatter and Residual Plots of the XGBoost and LightGBM Models

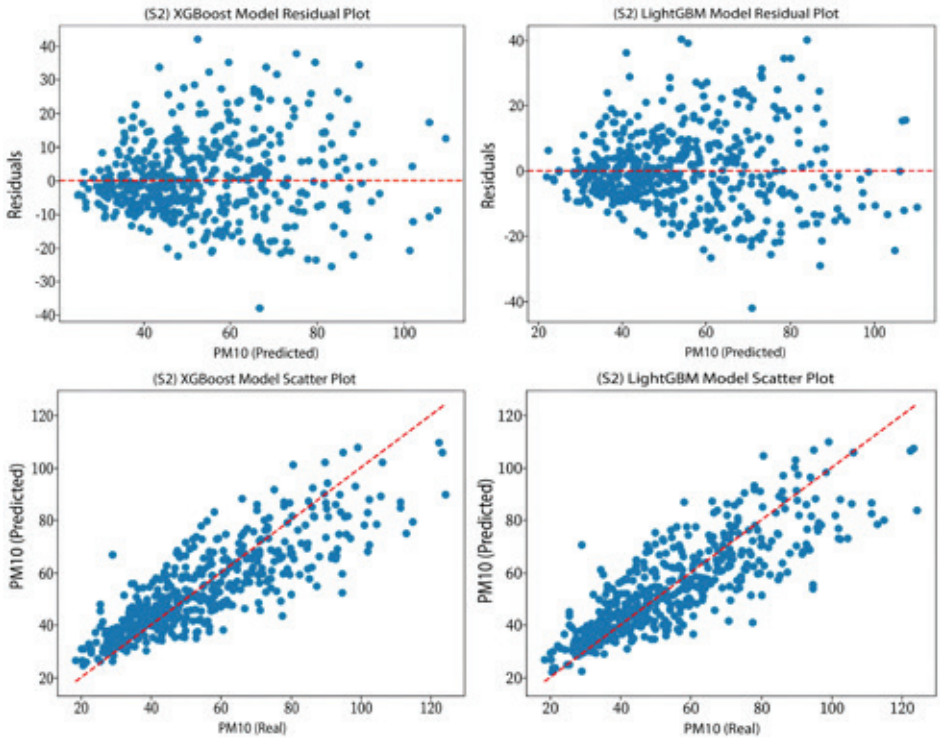


Table 8

Hypothesis T-Test

		T-Statistic	P-Value
S1	PM2,5	-0.1357	0.8921
	PM10	1.4859	0.1380
	PM2,5	-0.4743	0.6355
S2	PM10	-3.042	0.0025

Table 8 shows the paired t-test results comparing the performance of the XGBoost and LightGBM models across both scenarios for PM2.5 and PM10 predictions. In S1, for PM2.5 prediction, the t-statistic was -0.1357, with a p-value of 0.8921. Since this p-value is significantly higher than the 0.05 significance level, there is no statistically significant evidence to reject the null hypothesis, indicating no difference in performance between the XGBoost and LightGBM models for PM2.5.

For PM10 prediction in S1, the t-statistic was 1.4859, with a p-value of 0.1380. Similarly, this p-value exceeds the 0.05 threshold, suggesting insufficient evidence to conclude that the performance differences between the XGBoost and LightGBM models for PM10 are statistically significant. Therefore, the observed differences in performance metrics for PM10 in this scenario are also not statistically significant.

In S2, for PM2.5 prediction, the t-statistic was -0.4743, with a p-value of 0.6355. This result indicates that, as in S1, the observed differences in performance metrics between the two models are not statistically significant for PM2.5 in this scenario.

Finally, for PM10 prediction in S2, the t-statistic was -3.042, with a p-value of 0.0025. This result suggests that the observed differences in performance metrics for PM10 in this scenario are statistically significant, highlighting a superior performance of one model over the other.

6. DISCUSSION

The results provide a comprehensive assessment of the effectiveness of the XGBoost and LightGBM algorithms in predicting PM2.5 and PM10 concentrations in Lima, Peru. In S1, both algorithms exhibited similar performance in terms of R^2. Although XGBoost showed a slight improvement in some metrics, such as MAE and RRMSE, especially for PM10 prediction, the paired t-test revealed no statistically significant differences.

In contrast, a significant decrease in predictive performance was observed S2. This highlights the importance of including both variables to achieve more accurate predictions. The strong correlation between PM2,5 and PM10 appears to be critical, and excluding one undermines the predictive capability of the models.

These findings underscore the value of using advanced ML models like XGBoost and LightGBM in addressing complex air pollution problems. The slight performance advantage of XGBoost in certain scenarios could be leveraged to optimize early warning systems and mitigation policies.

7. CONCLUSIONS

One of the main benefits of this research lies in the ability of predictive models to accurately estimate PM2.5 and PM10 concentrations without the need for specialized equipment and machinery for measuring pollutants and climatic conditions. This approach offers significant cost savings, as it avoids the need for expensive equipment with limited spatial coverage. Prediction techniques enable the integration of multiple datasets, offering a broader and more accurate perspective than traditional measurement methods. These capabilities support more effective air quality management and better public health policy planning.

Despite the promising results, the study has some limitations. The analysis was based on data collected from a single monitoring station, which may not fully capture the spatial variability of air pollution across Lima. Additionally, important factors such as extreme weather conditions, changes in emission sources, precipitation, solar radiation, traffic congestion, and natural events were not considered due to limitation in the available datasets.

8. FUTURE RESEARCH

Future research should consider expanding the geographic scope by incorporating data from multiple monitoring stations distributed across Lima. This would enable a more accurate representation of the spatial variability of air pollution, providing a broader understanding of environmental conditions in different urban and suburban areas. Moreover, extending the data collection period to capture seasonal variations and long-term trends is recommended. Analyzing data over several years could offer deeper insights into how factors such as seasonal climate changes and fluctuations in emissions affect PM2.5 and PM10 concentrations.

Another promising area for future research is the development of new variants of ML algorithms specifically adapted for air pollution prediction. This could involve modifying existing algorithms or creating new hybrid approaches that combine the strengths of multiple algorithms. Moreover, it is important to include additional factors that may influence air quality, such as extreme weather conditions, changes in emission sources, precipitation, solar radiation, traffic congestion, and natural events. Incorporating these factors into predictive models could further enhance their accuracy and relevance for air quality management decision-making.

REFERENCES

Ameer, S., Shah, M, A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities, IEEE Access, 7, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

Amuthadevi, C., Vijayan, D. S. & Ramachandran, V. (2021). Development of air quality monitoring (AQM) models using different machine learning approaches, Journal of Ambient Intelligence and Humanized Computing, 13(1), 33. https://doi.org/10.1007/s12652-020-02724-2

Ayus, I., Natarajan, N. & Gupta, D. (2023). Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China, Asian Journal of Atmospheric Environment, 17, Article 4. https://doi.org/10.1007/s44273-023-00005-w

Bai, Y., Li, Y., Wang, X., Xie, J., & Li, C. (2016). Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions, Atmospheric Pollution Research, 7(3), 557–566. https://doi.org/10,1016/j.apr.2016.01.004

Cordova, C. H., Portocarrero, M. N. L., Salas, R., Torres, R., Canas, P., & López-Gonzales J. L. (2021). Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru, Scientific Reports, 11, Article 24232. https://doi.org/10.1038/s41598-021-03650-9

Gokul, P. R., Mathew, A., Bhosale, A., & Nair, A. T. (2023). Spatio-temporal air quality analysis and PM2,5 prediction over Hyderabad City, India using artificial intelligence techniques, Ecological Informatics, 76, Article 102067. https://doi.org/10.1016/j.ecoinf.2023.102067

Gryech, I., Ghogho, M., Elhammouti, H., Sbihi, N., & Kobbane, A. (2020). Machine learning for air quality prediction using meteorological and traffic related features, Journal of Ambient Intelligence and Smart Environments, 12(5), 379–391. https://doi.org/10.3233/AIS-200572

Liang, Y-C., Maimury, Y., Chen, A. H-L., & Juarez, J. R. C. (2020). Machine learning-based prediction of air quality, Applied Sciences, 10(24), Article 9151. https://doi.org/10.3390/app10249151

Liu, X., Zhao, K., Liu, Z., & Wang, L. (2023). PM2,5 Concentration Prediction Based on LightGBM Optimized by Adaptive Multi-Strategy Enhanced Sparrow Search Algorithm, Atmosphere, 14(11), Article 1612. https://doi.org/10.3390/atmos14111612

Martín-Baos, J. Á., Rodriguez-Benitez, L., García-Ródenas, R., & Liu, J. (2022). IoT based monitoring of air quality and traffic using regression analysis, Applied Soft Computing, 115, Article 108282. https://doi.org/10.1016/j.asoc.2021.108282

Pan, B. (2018). Application of XGBoost algorithm in hourly PM2,5 concentration prediction, IOP Conference Series: Earth and Environmental Science, 113, Article 012127. https://doi.org/10.1088/1755-1315/113/1/012127

Servicio Nacional de Meteorología e Hidrología del Perú. (2024). Monitoreo de la Calidad de Aire, para Lima Metropolitana. https://www.senamhi.gob.pe/?p=calidad-del-aire-estacion&e=112194

Shakya, D., Deshpande, V., Goyal, M. K., & Agarwal, M. (2023). PM2,5 air pollution prediction through deep learning using meteorological, vehicular, and emission data: A case study of New Delhi, India, Journal of Cleaner Production, 427, Article 139278. https://doi.org/10.1016/j.jclepro.2023.139278

Sulaimon, I. A., Alaka, H., Olu-Ajayi, R., Ahmad, M., Ajayi, S. & Hye, A. (2022). Effect of traffic data set on various machine-learning algorithms when forecasting air quality, Journal of Engineering, Design and Technology, 22(3), 1030–1056. https://doi.org/10.1108/JEDT-10-2021-0554

Wang, Z., Chen, P., Wang, R., An, Z., & Qiu, L. (2023). Estimation of PM2,5 concentrations with high spatiotemporal resolution in Beijing using the ERA5 dataset and machine learning models, Advances in Space Research, 71(8), 3150–3165. https://doi.org/10.1016/j.asr.2022.12.016

World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM2,5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. https://apps.who.int/iris/handle/10665/345329

World Health Organization. (2022). Air pollution. https://www.who.int/health-topics/air-pollution#tab=tab_1

Yang, W., Deng, M., Xu, F., & Wang, H. (2018). Prediction of hourly PM2,5 using a space-time support vector regression model, Atmospheric Environment, 181, 12–19. https://doi.org/10.1016/j.atmosenv.2018.03.015

Zhang, D., & Woo, S. S. (2020). Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network, IEEE Access, 8, 89584–89594. https://doi.org/10.1109/ACCESS.2020.2993547

Zhang, K., Yang, X., Cao, H., Thé, J., Tan, Z., & Yu, H. (2023). Multi-step forecast of PM2,5 and PM10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning, Environment International, 171, Article 107691. https://doi.org/10.1016/j.envint.2022.107691