Global Air Quality Index Prediction Using Machine Learning on Major Pollutants

Richard Santoso[1,*], Karto Iskandar[2,]

[1]*Computer Science Department, Master of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia*

[2]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia*

(Received: September 5, 2025; Revised: November 1, 2025; Accepted: January 15, 2026; Available online: February 17, 2026)

Abstract

Air pollution remains a major global concern due to its significant impact on public health and environmental sustainability. This study aims to develop a reliable global Air Quality Index (AQI) prediction model by evaluating five regression-based machine learning algorithms, including Linear Regression, Support Vector Regression, Random Forest, XGBoost, and LightGBM. The dataset contains over twenty thousand pollutant concentration records from multiple countries. Since the dataset consists of independent pollutant observations without timestamps or temporal sequences, this research employs supervised regression techniques rather than time-series forecasting methods to ensure methodological consistency with the non-temporal structure of the data. The methodology includes data preprocessing, validation of geocoded country information for missing values, transformations to address skewed pollutant distributions, and feature selection based on established environmental standards. Sample weights were applied to account for uneven regional representation, and systematic hyperparameter tuning with cross-validation was conducted to optimize model parameters and reduce potential overfitting. Evaluation metrics are supported by correlation analysis to quantify relationships between pollutants and AQI. The results show that XGBoost delivers the highest and most stable performance, with a MAE of 0.0216, MSE of 0.0010, RMSE of 0.0318, R^2 of 0.9971, and MAPE of 0.5664. Feature importance analysis highlights PM2.5 as the most influential pollutant, followed by ozone, nitrogen dioxide, and carbon monoxide. The predicted AQI values closely align with observed measurements, demonstrating strong generalizability across regions. An interactive dashboard was developed to visualize AQI predictions and pollutant contributions across countries, improving practical usability for environmental monitoring. Overall, this study provides a comprehensive framework for global AQI prediction and demonstrates the potential of machine learning to support decision-making in environmental management and public health planning.

Keywords: Air Quality Index (AQI); Air Quality; Prediction; Air Pollution; Machine Learning

1. Introduction

Air is one of the most important natural resources for the survival of all living creatures on Earth. Air can offer health to humans and animals as well as high-quality soil for plants [1]. Air pollution is the most hazardous sort of pollution and must be addressed immediately since we depend on clean air to breathe [2]. Air pollution has a direct impact on human health and can increase the risk of respiratory disorders such as asthma. In the long run, it increases the frequency of more severe weather events. It also raises pollution levels, which endangers vulnerable people, as well as preventive and health services [3]. The burning of fossil fuels, transportation, industrial activity, and agriculture are the main sources of air pollution and greenhouse gas emissions. On the other hand, the meteorological condition can increase the concentration of pollutants in air [4], [5]. Hazardous pollutants commonly found include fine particulate matter ($PM_{2.5}$); nitrogen dioxide (NO_2); sulfur dioxide (SO_2) and ozone (O_3). Of all these pollutants, $PM_{2.5}$ is the most dangerous as it can penetrate through the respiratory system and reach the lungs where it has a huge impact on health [6]. The Lancet Commission on Pollution and Health (2017), pollution caused around nine million deaths, or 16% of total global deaths, in 2015, with air pollution the leading contributor. According to the 2019 Global Burden of Disease Study, the number of deaths from air pollution is rising. This proves that air pollution is still a serious problem [7].

Several regions in South Asia and other developing areas consistently record annual $PM_{2.5}$ concentrations far exceeding the World Health Organization guideline value of 5 $\mu g/m^3$, placing them among the regions with the poorest air quality

*Corresponding author: Richard Santoso (richard.santoso@binus.ac.id)

DOI: https://doi.org/10.47738/jads.v7i1.1112

globally [8]. Persistent exposure to elevated $PM_{2.5}$ levels in these regions not only indicates insufficient emission reduction efforts but also underscores the urgent need for targeted mitigation strategies to improve air quality and public health [9].

Epidemiological evidence suggests that air pollution can cause many cases of heart disease, lung disease, reproductive disorders, and cancer. Humans create most of the dangerous substances around us. We're talking about the man-made molecules that come from activities such as transport, industry, construction, and agriculture. About 10% of the hazardous products occur naturally [10]. In worldwide rankings of mortality risk factors, $PM_{2.5}$ is the fifth ranked of global risk factors for mortality which is responsible for 7.6% of deaths. There is a lot of study going on the air pollution levels and the air quality prediction to make useful use pollution control. $PM_{2.5}$ is a type of fine particulate matter that consists of a combination of harmful gases and particles when it is emitted into the air [11]. Thus, there is need for more far-reaching and global prediction models to assist environmental and public health policies.

The dataset used in this study consists of 23,463 air quality records, including overall AQI values and specific AQI for major pollutants, namely carbon monoxide (CO), ozone (O_3), and nitrogen dioxide (NO_2). The data was collected from various cities and countries worldwide without specific regional boundaries, so the prediction results are global and can be used to support public health policies and more effective environmental management. Furthermore, the dataset does not contain temporal information or sequential timestamps; each record represents an independent observation of pollutant concentrations and AQI values. Therefore, this study employs supervised regression models rather than time-series forecasting methods to ensure methodological consistency with the non-temporal structure of the data.

Several studies have tried to model air quality with different machine learning techniques. However, most research remains focused on specific regions or pollutants, with small datasets that limit generalizability.Several machine learning models have been built to predict air quality, however the majority of research remains limited to specific regions or nations and focuses on only one or two types of pollutants. As a result, a broader and more complete methodology is required to develop predictive models that can be used globally. This study employed many machine learning approaches, including Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM). This study used a supervised learning technique to calculate the Air Quality Index (AQI) based on the concentrations of important pollutants, including CO, NO_2, O_3, and $PM_{2.5}$, which are recognized as air pollution indicators according to global standards. Because the data is tabular and does not have a temporal dimension, this work created a predictive regression model that bases AQI forecasts on the concentrations of the detected contaminants, allowing for direct predictions to support air quality monitoring.

The aim of this study is to predict global air quality using a variety of machine learning algorithms. Furthermore, this study examines model accuracy, defines the optimal strategy, and assesses the impact of important contaminants on air quality. The proposed approach can be used in nations or regions that lack sufficient air quality monitoring systems, notably in low- and middle-income countries. It can also help environmental authorities, health groups, and governments identify high-risk locations, create legislation, and provide early warnings.

2. Literature Review

2.1. Air Quality

Air quality is one of the most intuitively perceived environmental elements, directly affecting human health and overall well-being. Low air quality, has a direct impact on public health [12]. The Air Quality Index (AQI) measures air pollution and divided into six levels on the scale of 0 to 500 [13]. Table 1 displays the AQI categories.

Table 1. Air Quality Index Category

AQI Value	AQI Index Category
0 – 50	Good
51 – 100	Moderate

101 – 150	Unhealthy for Sensitive Groups
151 – 200	Unhealthy
201 – 300	Very Unhealthy
301 – 500	Hazardous

2.2. Air Pollution

Air pollution is the contamination of air with hazardous substances caused by combustion, cars, factories, and forest fires. Air pollution is a primary source of health issues and deaths around the world. Outdoor air pollution causes almost 4 million deaths per year, whereas indoor air pollution causes 2.3 million deaths. Its consequences include cardiovascular, pulmonary, and neurological disorders, as well as lower productivity, growing social disparity, and diminished cognitive ability [14].

2.3. Related Works

Several previous studies evaluated the effectiveness of various machine learning techniques in predicting air quality and pollutant concentrations. Random Forest (RF), XGBoost, Neural Network (ANN), and ensemble models have been shown to be the most effective, although accuracy varies by study context [2], [15], [16], [17], [18].

Madan et al. [2] found that neural networks and boosting models produced more accurate predictions than other methods. Liang et al. [15] analyzed RF, SVM, AdaBoost, and Stacking in Taiwan, demonstrating that ensemble models outperformed individual models for long-term forecasts and locations with high pollution levels. Lei et al. (2022) [16] used RF, Gradient Boosting, SVR, and MLR to predict PM_{10} and $PM_{2.5}$ in Macao, with RF outperform the other methods, especially under extreme conditions. Lei et al. (2023) [17] predicted $PM_{2.5}$, PM_{10}, and CO for the next 24-48 hours using ANN, RF, XGBoost, SVM, and MLR; the results showed that RF and SVM were the most effective, with the selection of meteorological features and prior pollutant data playing a key role. Kumar and Pande [18] evaluated data from 23 Indian cities using GNB, SVM, RF, XGBoost, and CatBoost, and discovered that XGBoost had the highest accuracy. These findings support the use of ensemble-based models and neural networks to increase air quality prediction accuracy, but they also demonstrate that model performance is heavily influenced by region, prediction period, and feature selection.

2.4. Gap Analysis

The existing literature can be grouped into three main themes. The first theme focuses on machine learning models developed for air quality prediction, comparing the performance of various algorithms. The second theme explores the influence of individual pollutants on prediction accuracy and air quality behavior. The third theme analyzes air quality trends within specific regions or cities, showing that most studies rely on localized datasets with limited geographical coverage. Organizing the literature into these themes provides clearer context for positioning the present study within the broader research landscape. Several previous studies have demonstrated the effectiveness of machine learning methods for predicting air quality and pollutants. Each study has different advantages and limitations, which are summarized in table 2.

Table 2. Gap Analysis

Research	Method	advantages	disadvantages
Cordova et al., 2021 [19]	MLP, LSTM, BNCV	LSTM achieved high accuracy for PM_{10}, effective in handling extreme data	Focused only on PM_{10} prediction, not tested in other regions
Kothandaraman et al., 2022 [11]	LR, RF, KNN, RL, XGBoost, AdaBoost	XGBoost and RF showed strong performance for $PM_{2.5}$ prediction	Requires complete meteorological data, performance highly data-dependent
Gladkova & Saychenko, 2022 [6]	ARIMA, Prophet, LSTM	Provided monthly pollutant forecasts useful for pollution control	Russian data is limited, long-term accuracy affected

Kumar & Pande, 2023 [18]	KNN, GNB, SVM, RF, XGBoost	XGBoost achieved high accuracy in both training and testing phases	Based on data from only 23 cities in India, not nationally representative
Lei et al., 2023 [17]	ANN, RF, XGBoost, SVM, MLR	Effective for 24–48 hours predictions, strong performance of RF and SVM	Limited to Macau, R^2 for 48-hour predictions was only 0.55
Suri et al., 2023 [20]	ANN, GPR	GPR highly accurate for 10 pollutants ($R^2 = 0.9843$)	Based on only 6 Indian cities, limited global relevance
Ravindiran et al., 2023 [21]	LightGBM, RF, CatBoost, AdaBoost, XGBoost	CatBoost achieved very high accuracy ($R^2 = 0.9998$), successfully processed 12 pollutants and 10 meteorological parameters over 5 years	Focus on Visakhapatnam city, generalization of other regions is limited
Liu et al., 2024 [22]	RF, LightGBM, LSTM	LSTM and LightGBM showed strong accuracy, considering weather and seasonal factors	Based only on Jinan data, pollutant-specific contributions not explained
Essamlali et al., 2024 [23]	LSTM, RF, ANN, SVR	Effective ML method predicts PM, NOx, CO, O_3	Based only on literature review, without direct experimental validation
Adiwidya et al., 2024 [24]	LightGBM, PCA	PCA-LightGBM combination accurately predicted $PM_{2.5}$ and CO_2	Short-term predictions (3 days/8 hours) with limited datasets

Machine Learning (ML) techniques are widely used to forecast air quality [25]. Most studies mainly emphasize accuracy performance while overlooking interpretability and operational feasibility, which are essential for real-world deployment. In addition, many previous studies remain limited to specific regions or analyze only one or two pollutants, resulting in a fragmented understanding of global air quality patterns. Therefore, this study seeks to fill this gap by analyzing several key pollutants simultaneously on a global scale.

3. Research Methodology

At this stage, the algorithms and methods employed in this study are explained. The methodological flow is shown in figure 1.

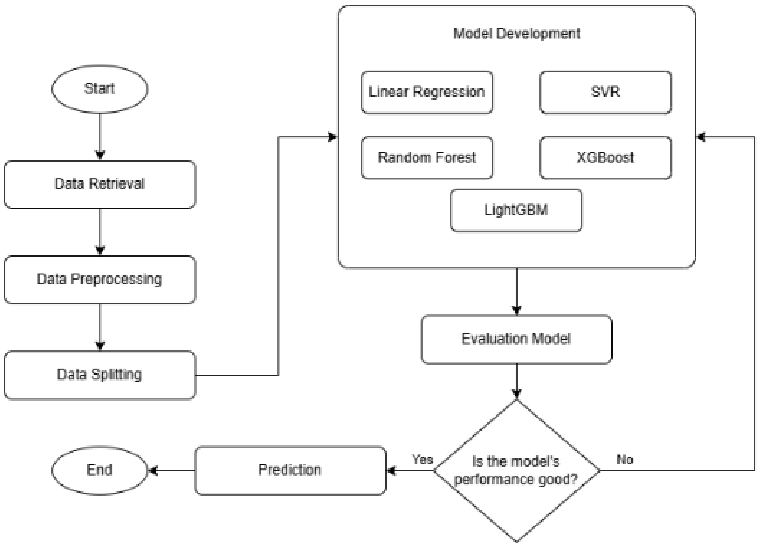


Figure 1. Flowchart research steps

3.2. Data Collection

The dataset used in this study was obtained from an online public platform; it is available in the Data Availability section. It contains 23,463 records with 175 countries and 23,462 cities as of March 11, 2022. The data was obtained through web scraping from elichens.com and underwent a simple feature engineering process by the provider. The main features available include Country, City, AQI Value, AQI Category, and the AQI value and category for each major pollutant (CO, O_3, NO_2, $PM_{2.5}$). This information allows for the analysis of the contribution of each pollutant to air quality in various cities worldwide, as shown in table 3.

Table 3. Features in the dataset

No	Feature	Description
1	Country	Name of the country
2	City	Name of the city
3	AQI Value	Overall Air Quality Index (AQI) value
4	AQI Category	Overall Air Quality Index (AQI) category
6	CO AQI Category	AQI category for Carbon Monoxide (CO)
7	Ozone AQI Value	AQI value for Ozone (O_3)
8	Ozone AQI Category	AQI category for Ozone (O_3)
9	NO2 AQI Value	AQI value for Nitrogen Dioxide (NO_2)
10	NO2 AQI Category	AQI category for Nitrogen Dioxide (NO_2)
11	PM2.5 AQI Value	AQI value for fine particulate matter ($PM_{2.5}$, ≤ 2.5 micrometers)
12	PM2.5 AQI Category	AQI category for fine particulate matter ($PM_{2.5}$, ≤ 2.5 micrometers)

The dataset was chosen because of its global coverage and comprehensive features, making it suitable for building air quality prediction models. To illustrate the data distribution, figure 2 displays the top twenty countries based on the number of cities in the dataset.

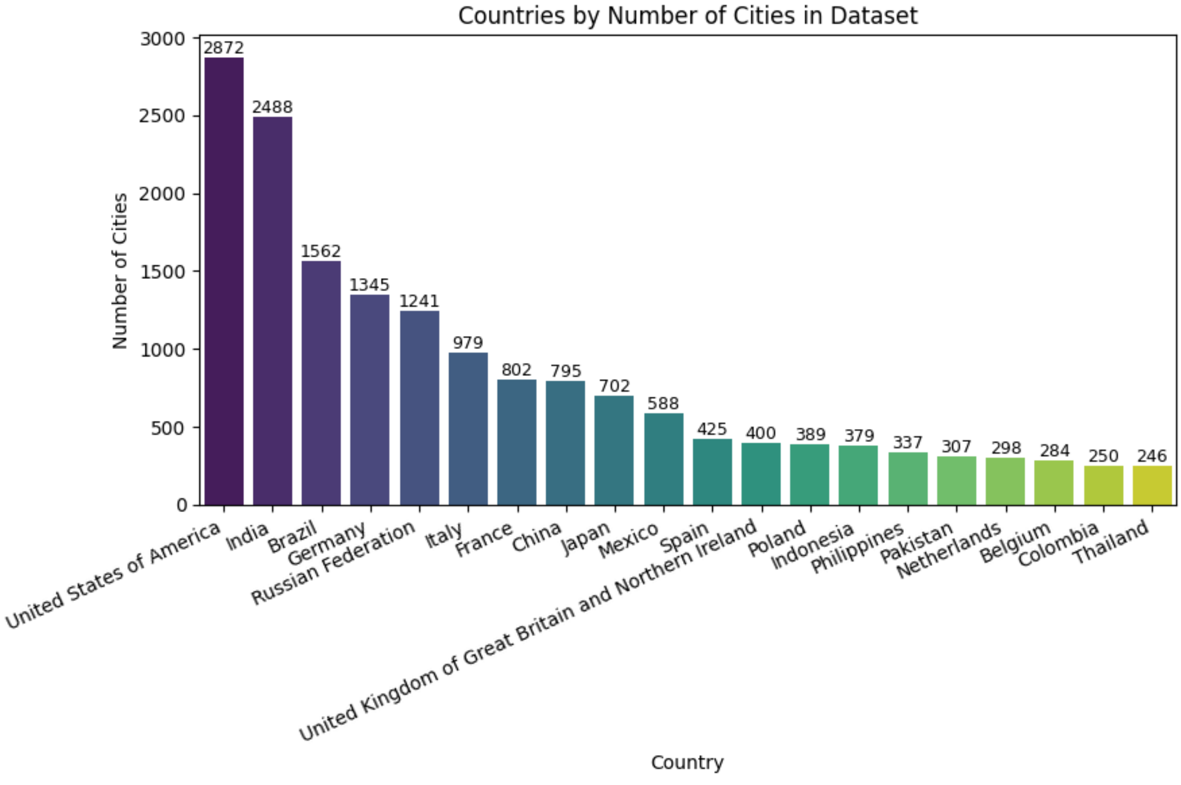


Figure 2. Twenty countries based on the number of cities in the dataset

3.3. Data Processing

The data processing step ensured the dataset's quality and consistency. The basic dataset contained 23,463 rows and 12 columns. Data inspection included checking dimensions, data types, missing values, and duplicates. The initial inspection found missing values in the Country (427 rows) and City (1 row) columns. Missing values in the City were removed and missing information in the Country were filled in using Geopy-Nominatim geocoding techniques. However, geocoding accuracy may vary due to ambiguous city names or incomplete location metadata. To reduce potential misclassification, fallback checks and manual verification were applied where inconsistencies were detected. The rest of the blank rows have been removed. To ensure compatibility with the ISO 3166 standard, country names were standardized using the pycountry library. The distribution of AQI values has a rightward skew (positive skewness), as shown in figure 3.

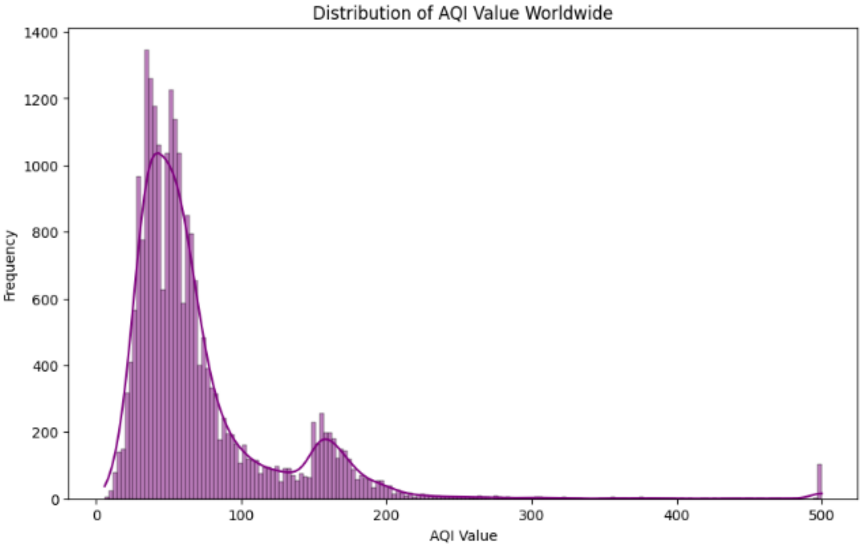


Figure 3. Distribution of AQI values worldwide

To reduce the imbalance in the distribution, a logarithmic transformation was applied to make it more symmetrical, as shown in figure 4. The log-transformed AQI values were used as the target variable in the predictive models.

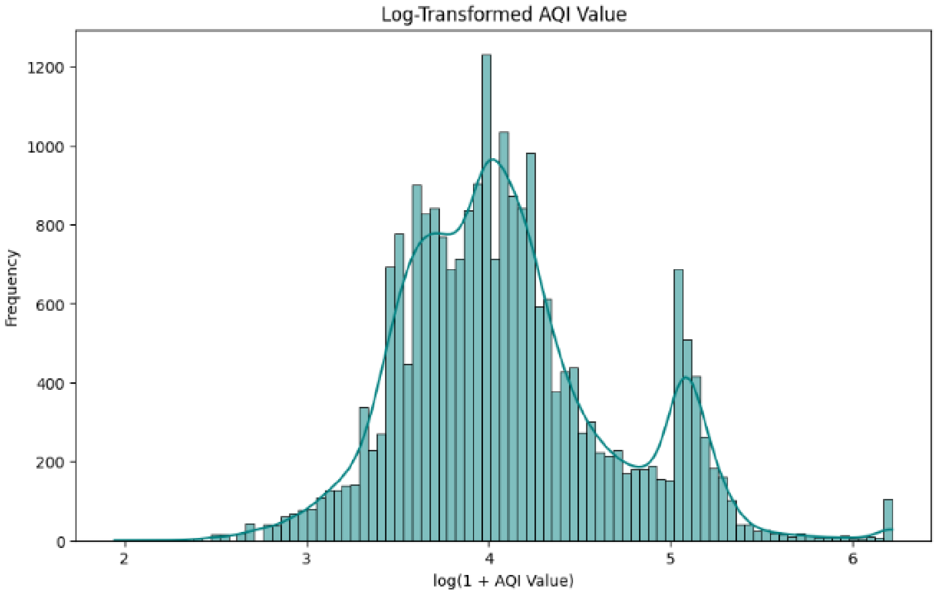


Figure 4. Distribution of AQI values worldwide on a logarithmic scale

The distribution of AQI categories was shown as a beginning overview, however the model is designed to predict numerical values. Next, a correlation analysis was performed to investigate the correlations between variables and ensure that the data were correctly prepared for the modeling stage. The correlation matrix was generated between the Air Quality Index (AQI) and four primary pollutants ($PM_{2.5}$, O_3, CO, and NO_2) to examine the strength of linear connections. Figure 5 depicts the correlation matrix, which demonstrates the degree of interdependence between various variables.

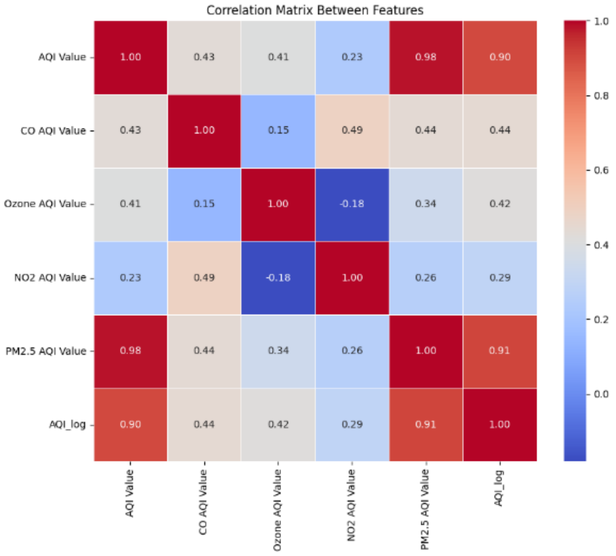


Figure 5. Correlation matrix of numeric variables in the air quality dataset

The correlation matrix indicated strong associations between pollutants and AQI, including $PM_{2.5}$ (r ≈ 0.91), NO_2 (r ≈ 0.29), O_3 (r ≈ 0.42), and CO (r ≈ 0.44). Reporting these coefficients provides clearer evidence of pollutant influence. Overall, the analysis found that $PM_{2.5}$ had the highest link with the AQI when compared to other pollutants, showing that $PM_{2.5}$ had a stronger impact on overall air quality levels. This finding provides preliminary insight into the data patterns before to the modeling procedure.

The dataset was then split into training and test data at an 80:20 ratio using a train-test split. StandardScaler was used to standardize the features, ensuring that all variables were on the same scale. Sample weights were used during model training to balance the contribution of data points, accounting for uneven regional representation. Given that this study employs a regression model and the target AQI values were log-transformed to reduce skewness, stratified splitting was not applied. Outliers were then examined, but they were kept to ensure that the model accurately mirrored real-world conditions and that its forecasts were valid..

3.4. Variable and Model Selection

The research dataset contains features such as country, city, overall AQI value, AQI category, and values and categories for many key pollutants (CO, O_3, NO_2, and $PM_{2.5}$). This study applies supervised learning with a regression model.

The AQI Value is the target variable to be predicted, with input features containing CO, O_3, NO_2, and $PM_{2.5}$ AQI values. The four pollutants were chosen based on global air quality regulations, which make them essential pollution indicators, as well as correlation analysis results that show a significant impact to the AQI. Other pollutants such as SO_2 and PM_{10} were excluded due to inconsistent availability across countries and incomplete data in the source dataset. The model focuses on numerical prediction of the AQI, with conversion to AQI categories occurring only during the evaluation and interpretation phases.

Furthermore, the dataset does not contain temporal information or sequential timestamps. Each record represents an independent observation of pollutant concentrations and AQI values, without any chronological order. Due to the absence of time-based features, this study does not employ time series forecasting methods. Instead, supervised regression models are used to directly predict AQI values from pollutant inputs. This approach aligns with the structure of the data and ensures methodological consistency with the non-temporal nature of the dataset.

The prediction model was built using five regression algorithms, including Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting

Machine (LightGBM). These five algorithms were chosen for their ability to handle complex data, nonlinear relationships, noisy data, and outliers.

3.5. Model Development

In this stage, an air quality prediction model was built using five machine learning algorithms such as LR, SVR, RF, XGBoost, and LightGBM. LR served as the baseline, while RF, XGBoost, and LightGBM were chosen for their capacity to handle non-linear relationships, noisy data, and outliers. SVR with an RBF kernel was used as an alternative nonlinear model. To ensure consistent and reliable evaluation, the model training process used a fixed 80:20 train-test split. Systematic hyperparameter tuning was conducted using GridSearchCV for LR, SVR, and RF, and RandomizedSearchCV for XGBoost and LightGBM, each with 3-fold cross-validation. Sample weights were applied to account for uneven regional representation. These strategies help mitigate overfitting and ensure that the models generalize well to unseen data.

Hyperparameter tuning was conducted using GridSearchCV for LR, SVR, and RF, and RandomizedSearchCV for XGBoost and LightGBM, with 3-fold cross-validation. Sample weights were applied to account for uneven regional representation. These strategies help mitigate overfitting and ensure that the models generalize well to unseen data. The models were trained and evaluated with an 80:20 data split. The evaluation metrics used included MAE, MSE, RMSE, R^2, and MAPE. The formulas for each metric are shown in Equations (1)-(5).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

where $y_i$is the actual value, $\hat{y}_i$is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the total number of test samples. MAE is the average absolute difference between predicted and actual values, a lower MAE indicates better prediction accuracy. MSE is the average of the squared differences between predicted and actual values, with lower values indicating less model error. RMSE is the square root of MSE, which expresses prediction error on the same scale as the data. R^2 measures how closely predictions match actual data, with values closer to 1 indicating higher model performance.

4. Results and Discussion

4.1. Evaluation Metrics

This study compared five regression algorithms, namely LR, SVR, RF, XGBoost, and LightGBM with the objective to predict AQI based on CO, O_3, NO_2, and $PM_{2.5}$ pollutants. Model performance was evaluated using MAE, MSE, RMSE, R^2, and MAPE, as shown in table 4.

Table 4. Model Performance Metrics

Model	MAE	MSE	RMSE	R-Squared	MAPE
Linear Regression	0.1993	0.0707	0.2659	0.7951	5.2344

SVR	0.0498	0.0037	0.0605	0.9894	1.2511
Random Forest	0.0499	0.0050	0.0707	0.9855	1.2818
XGBoost	**0.0216**	**0.0010**	**0.0318**	**0.9971**	**0.5664**
LightGBM	0.0261	0.0017	0.0417	0.9950	0.6619

Among the tested models, the XGBoost model outperformed the other models by achieving the lowest MAE, MSE, RMSE, and MAPE values, as well as the greatest R², indicating accurate predictions and an excellent match to the real AQI data. Linear Regression scored poorly due to its linear assumptions, while ensemble models like Random Forest and LightGBM gave results close to XGBoost.

4.2. Feature Importance

The feature importance analysis shown in figure 6 indicates that $PM_{2.5}$ had the greatest impact on AQI predictions across most of the models, while O_3 had the most impact in LightGBM. CO and NO_2 made small but significant impacts to AQI variation. These findings are consistent with previous studies identifying $PM_{2.5}$ as the major pollutant affecting air quality.

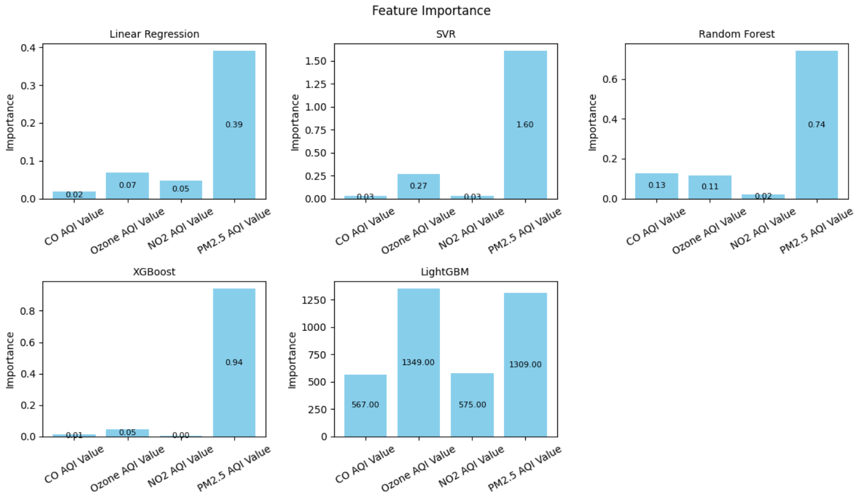


Figure 6. Feature importance of pollutants across five models

4.3. Actual vs Predicted

To validate the model's accuracy, figure 7 shows a comparison of actual and predicted AQI values using the XGBoost model on the test dataset. Actual values are illustrated as red dashed lines, while predictions are shown by blue dots. The projected values closely match the observed values, showing that the model has good predictive accuracy.

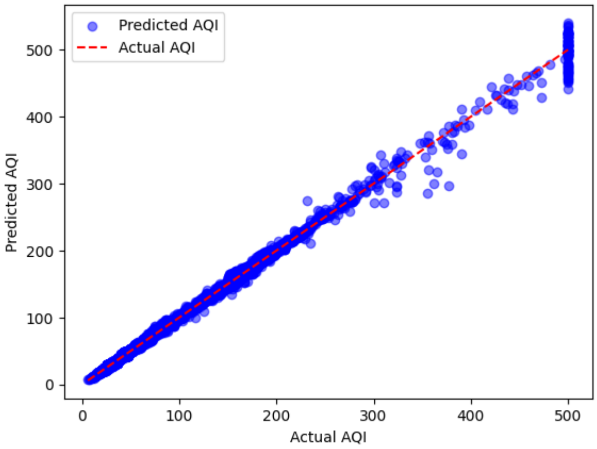
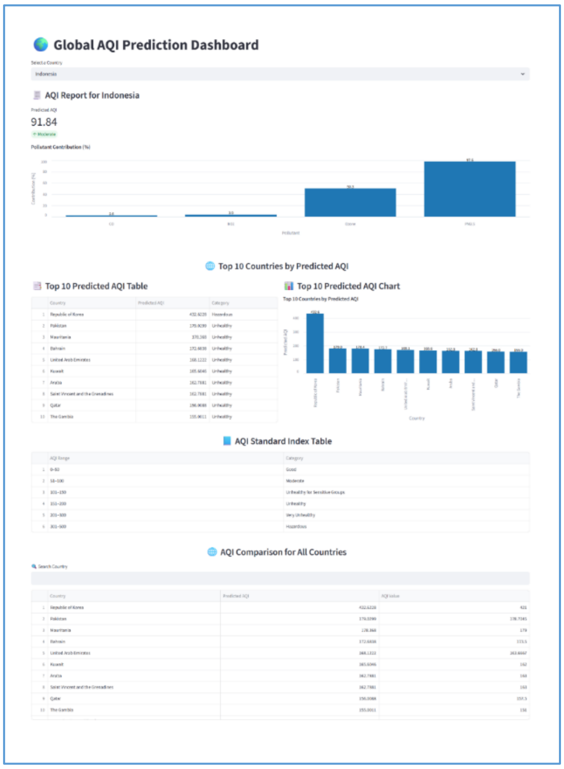


Figure 7. Comparison between predicted and actual AQI values using XGBoost

4.4. Dashboard

An interactive dashboard was developed using Streamlit and Altair to visualize predicted AQI values, pollutant contributions, and country rankings. The dashboard features country selection, top 10 predicted AQI tables and charts, an AQI standard index table, and a search-enabled comparison table for all countries, as shown in figure 8. Users can examine predicted AQI values alongside the contribution of each pollutant for selected countries. The dashboard also highlights the top ten countries with the highest predicted AQI and provides a comparison between predicted and actual values, facilitating intuitive insights into air quality patterns and supporting data-driven environmental management decisions.



Global AQI Prediction Dashboard

Select a Country

Indonesia

AQI Report for Indonesia

Predicted AQI

91.84

↑ Moderate

Pollutant Contribution (%)

Top 10 Countries by Predicted AQI

Top 10 Predicted AQI Table

	Country	Predicted AQI	Category
1	Republic of Korea	432.6228	Hazardous
2	Pakistan	179.0399	Unhealthy
3	Mauritania	178.368	Unhealthy
4	Bahrain	172.6838	Unhealthy
5	United Arab Emirates	168.1222	Unhealthy
6	Kuwait	165.6046	Unhealthy
7	Aruba	162.7881	Unhealthy
8	Saint Vincent and the Grenadines	162.7881	Unhealthy
9	Qatar	156.0088	Unhealthy
10	The Gambia	155.0011	Unhealthy

Top 10 Predicted AQI Chart

Top 10 Countries by Predicted AQI

AQI Standard Index Table

	AQI Range	Category
1	0–50	Good
2	51–100	Moderate
3	101–150	Unhealthy for Sensitive Groups
4	151–200	Unhealthy
5	201–300	Very Unhealthy
6	301–500	Hazardous

AQI Comparison for All Countries

Search Country

	Country	Predicted AQI	AQI Value
1	Republic of Korea	432.6228	421
2	Pakistan	179.0399	178.7045
3	Mauritania	178.368	179
4	Bahrain	172.6838	173.5
5	United Arab Emirates	168.1222	163.6667
6	Kuwait	165.6046	162
7	Aruba	162.7881	163
8	Saint Vincent and the Grenadines	162.7881	163
9	Qatar	156.0088	157.5
10	The Gambia	155.0011	156

Figure 8. Snapshot of the interactive Global AQI Prediction Dashboard

Although the dataset is described as having no specific regional boundaries, the distribution of cities across countries is uneven. This imbalance introduces potential regional and reporting bias, which may cause the model to learn more from highly represented regions while underperforming in areas with limited data coverage. As a result, the generalizability of the predictions may be reduced, and future studies should consider more balanced datasets or bias-mitigation approaches.

In the end, the combined evaluation of performance metrics, feature importance, and pollutant contribution indicates that XGBoost is the most effective model for predicting AQI, outperforming the other algorithms in both accuracy and fit. $PM_{2.5}$ was identified as the main factor affecting air quality, with O_3, CO, and NO_2 contributing to varying degrees according to the model. The predicted values showed close to the actual AQI, proving that the model was trustworthy. Also, the interactive dashboard serves as a useful tool for visualizing predicted AQI and pollutant contributions around the world, enabling data-driven decision-making for air quality management and pollution reduction.

5. Conclusion

This study compared five regression models for predicting the AQI, which were LR, SVR, RF, XGBoost, and LightGBM. The results of the evaluation showed that XGBoost performed the best, achieving MAE of 0.0216, MSE of 0.0010, RMSE of 0.0318, R^2 of 0.9971, and MAPE of 0.5664. This demonstrated the lowest prediction error and a strong ability to explain AQI variability. LightGBM ranked second with reasonably high performance, while SVR and RF showed moderate accuracy. Linear Regression had the lowest accuracy, making it unsuitable for reliable AQI prediction. These findings confirm that XGBoost is the most effective model for global AQI analysis. Feature importance analysis revealed that $PM_{2.5}$ had the greatest impact on AQI predictions across most of the models, while O3 had the highest influence in LightGBM. CO and NO2 contributed smaller yet meaningful effects. These findings provide a comprehensive understanding of the primary pollutants that shape air quality and can support evidence-based environmental regulation.

Furthermore, the interactive dashboard developed in this study provides an informative overview of predicted AQI values worldwide. Users can view predicted AQI values and the contribution of each pollutant for selected countries. The dashboard also displays the top ten countries with the highest predicted air quality index, as well as a comparison of predicted and actual AQI values. This supports informed decision-making for effective air quality monitoring and management. Overall, this study demonstrates the potential of machine learning models for global AQI prediction, with XGBoost providing the most accurate and stable performance. The findings can serve as a reference for future research on air quality prediction and pollutant impact assessment.

Future research could extend this work by incorporating additional environmental and meteorological variables, such as temperature, humidity, and rainfall, to improve the robustness of AQI prediction. Using datasets with temporal information would also enable time series modelling, allowing the analysis of trends and fluctuations in air quality over time.

6. Declarations

6.1. Author Contributions

Conceptualization: R.S. and K.I.; Methodology: R.S. and K.I.; Software: R.S.; Validation: R.S. and K.I.; Formal Analysis: R.S. and K.I.; Investigation: R.S.; Resources: K.I.; Data Curation: K.I.; Writing Original Draft Preparation: R.S. and K.I.; Writing Review and Editing: R.S. and K.I.; Visualization: R.S.; all authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Mitreska Jovanovska, V. Batz, P. Lameski, E. Zdravevski, M. A. Herzog, and V. Trajkovik, "Methods for urban air pollution measurement and forecasting: Challenges, opportunities, and solutions," *Atmosphere*, vol. 14, no. 9, pp. 14-41, 2023, doi: 10.3390/atmos14091441.

[2] T. Madan, S. Sagar, and D. Virmani, "Air quality prediction using machine learning algorithms – A review," *in Proc. 2nd Int. Conf. Advances in Computing, Communication Control and Networking (ICACCCN)*, vol. 1, no. 1, pp. 140–145, Dec. 2020, doi: 10.1109/ICACCCN51052.2020.9362912.

[3] H. M. Tran, F.-J. Tsai, Y.-L. Lee, J.-H. Chang, L.-T. Chang, T.-Y. Chang, K. F. Chung, H.-P. Kuo, K.-Y. Lee, K.-J. Chuang, and H.-C. Chuang, "The impact of air pollution on respiratory diseases in an era of climate change: A review of the current evidence," *Sci. Total Environ,* vol. 898, no.1, pp. 1-20, Nov. 2023, doi: 10.1016/j.scitotenv.2023.166340.

[4] R. Munsif, M. Zubair, A. Aziz, and M. N. Zafar, "Industrial air emission pollution: Potential sources and sustainable mitigation," *in Environmental Emissions, R. Viskup, Ed. London: IntechOpen*, 2021, ch. 4, doi: 10.5772/intechopen.93104.

[5] G. E. Edo, L. O. Itoje-Akpokiniovo, P. Obasohan, V. O. Ikpekoro, P. O. Samuel, A. N. Jikah, L. C. Nosu, H. A. Ekokotu, U. Ugbune, E. E. A. Oghroro, O. L. Emakpor, I. E. Ainyanbhor, W. A.-S. Mohammed, P. O. Akpoghelie, J. O. Owheruo, and J. J. Agbo, "Impact of environmental pollution from human activities on water, air quality and climate change," *Ecological Frontiers,* vol. 44, no. 5, pp. 874–889, 2024, doi: 10.1016/j.ecofro.2024.02.014.

[6] E. Gladkova and L. Saychenko, "Applying machine learning techniques in air quality prediction," *Transportation Research Procedia*, vol. 63, no.1, pp. 1999-2006, 2022, doi: 10.1016/j.trpro.2022.06.222.

[7] R. Fuller, P. J. Landrigan, K. Balakrishnan, G. Bathan, S. Bose-O'Reilly, M. Brauer, et al., "Pollution and health: a progress update," *The Lancet Planetary Health*, vol. 6, no.1, pp.1-12, 2022, doi: 10.1016/S2542-5196(22)00090-0.

[8] P. Baharvand, P. Amoatey, Y. Omidi Khaniabadi, P. Sicard, H. Naqvi, and R. Rashidi, "Short-term exposure to PM2.5 pollution in Iran and related burden diseases," International Journal of Environmental Health Research, vol. 35, no. 1, pp. 1-13, 2025, doi: 10.1080/09603123.2025.2449969.

[9] M. Olczak, A. Piebalgs, and P. Balcombe, "A global review of methane policies reveals that only 13% of emissions are covered with unclear effectiveness," *One Earth*, vol. 6, no. 5, pp. 519–535, 2023. doi: 10.1016/j.oneear.2023.04.009.

[10] G. Markozannes, K. Pantavou, E. C. Rizos, O. A. Sindosi, C. Tagkas, M. Seyfried, I. J. Saldanha, N. Hatzianastassiou, G. K. Nikolopoulos, and E. Ntzani, "Outdoor air quality and human health: An overview of reviews of observational studies," *Environmental Pollution*, vol. 306, no.1, pp. 1-19, 2022, doi: 10.1016/j.envpol.2022.119309.

[11] D. Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, D. Dhabliya, S. Satla, and W. Abera, "Intelligent forecasting of air quality and pollution prediction using machine learning," *Adsorption Science & Technology,* vol. 2022, no. 1, pp. 1-12, 2022, Article ID 5086622, doi: 10.1155/2022/5086622.

[12] X. Zhang, L. Han, H. Wei, X. Tan, W. Zhou, W. Li, and Y. Qian, "Linking urbanization and air quality together: A review and a perspective on the future sustainable urban development," *J. Clean. Prod*, vol. 346, no. 1, pp. 1-18, Apr. 2022, doi: 10.1016/j.jclepro.2022.130988.

[13] M.-J. Chen, Y. L. Guo, P. Lin, H.-C. Chiang, P.-C. Chen, and Y.-C. Chen, "Air quality health index (AQHI) based on multiple air pollutants and mortality risks in Taiwan: Construction and validation," *Environmental Research*, vol. 231, no. 1, pp. 1-14, 2023, doi: 10.1016/j.envres.2023.116214.

[14] J. Rentschler and N. Leonova, "Global air pollution exposure and poverty," *Nature Communications*, vol. 14, no.1, pp. 1-32, 2023, doi: 10.1038/s41467-023-39797-4.

[15] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," *Applied Sciences*, vol. 10, no. 24, pp. 1-21, 2020, doi: 10.3390/app10249151.

[16] T. M. T. Lei, S. W. I. Siu, J. Monjardino, L. Mendes, and F. Ferreira, "Using machine learning methods to forecast air quality: A case study in Macao," *Atmosphere*, vol. 13, no. 9, pp. 1412-1426, 2022, doi: 10.3390/atmos13091412.

[17] T. M. T. Lei, S. C. W. Ng, and S. W. I. Siu, "Application of ANN, XGBoost, and other ML methods to forecast air quality in Macau," *Sustainability*, vol. 15, no. 6, pp. 1-21, 2023, doi: 10.3390/su15065341.

[18] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: A case study of Indian cities," *International Journal of Environmental Science and Technology,* vol. 20, no. 6, pp. 5333–5348, 2023, doi: 10.1007/s13762-022-04241-5.

[19] C. H. Cordova, M. N. L. Portocarrero, R. Salas, R. Torres, P. C. Rodrigues, and J. L. López-Gonzales, "Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru," *Scientific Reports*, vol. 11, no. 1, pp. 24-32, 2021, doi: 10.1038/s41598-021-03650-9.

[20] R. Suri, A. Jain, N. Kapoor, A. Kumar, H. Arora, K. Kumar, and H. Jahangir, "Air quality prediction - A study using neural network based approach," *Journal of Soft Computing in Civil Engineering,* vol. 7, no. 1, pp. 99–113, 2023, doi: 10.22115/SCCE.2022.352017.1488.

[21] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, no. 1, pp. 1-18, 2023, doi: 10.1016/j.chemosphere.2023.139518.

[22] Q. Liu, B. Cui, and Z. Liu, "Air quality class prediction using machine learning methods based on monitoring data and secondary modeling," *Atmosphere*, vol. 15, no. 5, pp. 553-568, 2024, doi: 10.3390/atmos15050553.

[23] I. Essamlali, H. Nhaila, and M. El Khaili, "Supervised machine learning approaches for predicting key pollutants and for the sustainable enhancement of urban air quality: A systematic review," *Sustainability*, vol. 16, no. 3, pp. 976-985, 2024, doi: 10.3390/su16030976.

[24] A. S. Adiwidya, A. Romadhony, I. Chandra, A. N. D. Sukmawati, H. M. Sholihah, D. U. Islamiah, and A. Rinaldi, "Prediction of PM2.5 and CO2 concentrations using the PCA-LightGBM method in Bandung, Indonesia," *Journal of Physics: Conference Series*, vol. 2942, no. 1, pp. 1-14, 2024, doi: 10.1088/1742-6596/2942/1/012004.

[25] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artificial Intelligence Review*, vol. 56, no.1, pp. 10031–10066, 2023, doi: 10.1007/s10462-023-10424-4.