Predicting Air Quality Index Using Regression Models

Mago Karl Mattheus W.
College of Computing and Information Technologies
National University
Manila, Philippines
Magokw@students.national-u.edu.ph

Liao Adrian Miguel G.
College of Computing and Information Technologies
National University
Manila, Philippines
liaoag@students.national-u.edu.ph

***Abstract*—This study focuses on predicting the Air Quality Index (AQI) using machine learning techniques to provide an efficient data driven approach for monitoring environmental conditions. The study utilizes an AQI dataset containing key pollutant concentrations such as PM2.5, PM10, NO_2, CO, SO_2, and O_3, obtained from Kaggle. Data preprocessing was performed to handle missing values and ensure the quality and reliability of the dataset. Three models, Linear Regression, Random Forest Regressor, and Decision Tree Regressor, were implemented and trained in a Google Colab environment. Their performance was evaluated and compared using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The results show that the Random Forest Regressor achieved the highest accuracy and generalization capability among all tested models.**

***Index Terms*—Air Quality Index (AQI), Machine Learning, Linear Regression, Random Forest Regressor, Decision Tree Regressor, Environmental Monitoring, Data Preprocessing**

I. Introduction

With the rapid growth of urbanization and modernization, the population has increased waste production in many cities. The use of private vehicles, industrial emissions, and other human activities has led to a rise in the emission of toxic gases and air pollutants. These increasing levels of emissions pose serious risks to human health and the environment, contributing to climate change and global warming.

This study aims to predict the air quality in the coming days using stored environmental data and machine learning techniques such as Linear Regression, Random Forest, and Logistic Regression. The dataset contains key air quality indicators such as Nitric Oxide, Nitrogen Dioxide, Nitrogen Oxides, Ammonia, Carbon Monoxide, Sulfur Dioxide, Ozone, Benzene, Toluene, and Xylene. By developing a predictive model, the study seeks to determine how these pollutants relate to overall carbon emissions.

The findings of this study aim to provide insights that can help the public better understand emission patterns and implement effective strategies for reducing air pollution and promoting environmental sustainability.

A. Objectives

The objective of this study is to develop and evaluate machine learning models that can predict the Air Quality Index (AQI) based on key air pollutant concentrations such as PM2.5, PM10, NO, CO, SO, and O.

This study aims to:

- Implement and train regression models Linear Regression, Decision Tree Regressor, and Random Forest Regressor using an AQI dataset obtained from Kaggle.
- Compare the performance of each model using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² metrics.
- Identify which pollutants have the most significant impact on AQI levels based on feature importance analysis.
- Evaluate a Logistic Regression model to classify AQI values into categorical air quality levels (Good, Moderate, Poor, etc.).
- Recommend the most suitable model for reliable AQI prediction and environmental monitoring applications.

II. Related Works / Literature Review

N. Sarkar et al. (2022) conducted a study on Air Quality Index (AQI) prediction and examined the contribution of pollutants to air quality deterioration and their harmful effects on people, particularly in India. The researchers utilized several deep learning and machine learning models such as Long Short-Term Memory (LSTM), Linear Regression (LR), Gated Recurrent Unit (GRU), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). These models were trained using AQI data from the Delhi region, specifically focusing on stored PM2.5 measurements. Their approach was considered a hybrid deep learning model that combined the strengths of LSTM and GRU architectures. The hybrid LSTM–GRU model achieved a Mean Absolute Error (MAE) of 36.11 and an R² coefficient of determination of 0.84, indicating that it could explain 84 percent of the variance in the data. Sarkar et al. concluded that the hybrid model demonstrated superior predictive performance compared to the individual models, showing its effectiveness in forecasting AQI in highly polluted urban environments [1].

While Sarkar et al. (2022) focused on hybrid models, Liu H. et al. (2019) used traditional models such as Support Vector Regression (SVR) and Random Forest Regression (RFR) for predicting AQI in Beijing and nitrogen oxides concentrations in an Italian city based on publicly available datasets. They found that the SVR-based model performed better in predicting AQI (RMSE = 7.666, R² = 0.9776, r = 0.9887), while the RFR-based model performed better in predicting NOX concentration [2].

In addition, Natarajan et al. (2024) implemented a Grey Wolf Optimization (GWO) and Decision Tree (DT) algorithm in major cities such as Delhi, Hyderabad, Kolkata, Bangalore, Visakhapatnam, and Chennai. They also experimented with models like Random Forest Regressor, K-Nearest Neighbor, and Support Vector Regressor. Upon comparing the results, the proposed model achieved better prediction performance compared to traditional machine learning algorithms, with accuracies of 88.98 percent for New Delhi, 94.48 percent for Kolkata, 97.66 percent for Hyderabad, 95.22 percent for Chennai, and 97.68 percent for Visakhapatnam [3].

A. Conceptual Framework

The Conceptual Framework for this study illustrates the entire process of how the machine learning model was developed and utilized to predict the Air Quality Index (AQI). It visually represents the relationships between the input data, data preprocessing, model training, and the result/findings.

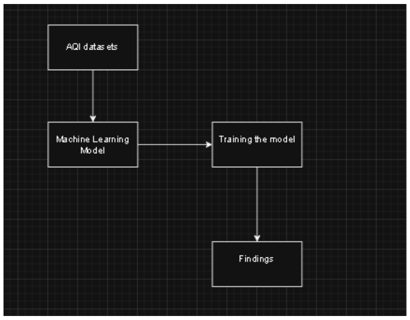

Fig. 1. Conceptual Framework of the Study

The process begins with the **AQI datasets**, which contain key data on pollutant concentrations such as PM2.5, PM10, NO_2, CO, SO_2, and O_3. These variables serve as the primary inputs to the **machine learning model**, where data preprocessing and feature selection are performed. This step involves cleaning the dataset by removing null or irrelevant values to ensure data quality and reliability.

The next stage is **training the model**, where the cleaned and prepared dataset is used to train various machine learning algorithms. During this phase, the model learns the patterns and relationships between pollutant levels and AQI values. Finally, the **findings** phase presents the model's results, including performance evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), which collectively assess the model's accuracy and predictive capability.

III. Methodology

A. Dataset Description

For this model, we will be using the *Air Quality Data in India (2015–2020)* dataset, provided by Rohan Rao on Kaggle. The dataset contains hourly and daily measurements of various air pollutant concentrations collected from multiple monitoring stations across major cities in India [4]. It provides key environmental indicators such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with corresponding Air Quality Index (AQI) values. This dataset serves as a reliable basis for developing and evaluating machine learning models aimed at predicting AQI levels in a city, We will mainly use the provided dataset in delhi to our study.

B. Data Preprocessing

In Data Preprocessing, we used city(underline)day.csv, which contains the daily air quality data with various pollutants (PM2.5, PM10, NO2, SO2, CO, O3). The data also had weather-related attributes like temperature, humidity, and wind speed. The data has a lot of missing values (88,488 missing entries). We cleaned the missing values to better use the dataset through appropriate preprocessing methods before proceeding with the analysis. Since pollutants and weather features have different ranges, they are scaled so the model treats them equally (Min-Max scaling, Standardization) using sklearn.preprocessing. Then, the cleaned and processed dataset is split into a training set and testing set — 70–80 percent for training the model and 20–30 percent for testing.

C. Model Development

Three machine learning regression models were implemented to predict the Air Quality Index (AQI): Linear Regression, Random Forest Regressor, and Decision Tree Regressor. These models were developed using the Scikit-learn library in Python within the Google Colab environment.

The Linear Regression model was implemented first to establish a baseline performance. Using LinearRegression() from Scikit-learn, the model learned the linear relationships between the pollutants and the AQI. It was trained using the training dataset through the .fit() method and evaluated using the testing data via the .predict() method. This model assumes that AQI increases or decreases proportionally with pollutant concentrations.

Random Forest model was implemented using Scikit-learn RandomForestRegressor(). It uses an ensemble of multiple decision trees to capture non-linear patterns and interactions between pollutants. The model was trained on the same dataset using the .fit() method. During training, each tree in the forest learned different subsets of the data, improving prediction robustness and reducing overfitting. After training, the .featureimportances attribute was used to identify the most significant pollutants influencing AQI levels. This helped visualize which pollutants (such as PM2.5 and PM10) had the strongest effect on air quality.

The Decision Tree model, implemented using Scikit-learn DecisionTreeRegressor(), was developed to provide a more interpretable structure of pollutant influence. The model recursively split the dataset into smaller subsets based on pollutant thresholds, aiming to minimize prediction error. Although more prone to overfitting, it offered valuable insight into pollutant behavior and AQI sensitivity.

Logistic Regression model was implemented to classify AQI values into distinct air quality categories (Good, Satisfactory, Moderate, Poor, Very Poor, and Severe). This model uses the same pollutant features ($PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, and O_3) as input variables. The model applies a sigmoid activation to map predicted values into discrete classes, allowing for the evaluation of classification performance through a confusion matrix and classification report.

D. Evaluation Metrics

The performance and accuracy of the machine learning models were assessed using several regression evaluation metrics. These metrics measure how close the predicted AQI values are to the actual AQI values in the dataset. The evaluation was performed after training each model using the test dataset.

Mean Squared Error (MSE)
Measures the average squared difference between the predicted and actual AQI values. Unlike MAE, MSE gives higher weight to larger errors by squaring them, making it more sensitive to outliers. A lower MSE indicates that the model's predictions are closer to the true AQI values and overall more accurate.

$$\text{MSE} = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$$

Fig. 2. Formula for Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)
Calculates the square root of the average squared differences between predicted and actual AQI values. RMSE gives more weight to large errors, making it useful for detecting significant deviations. Lower RMSE values indicate higher model accuracy.

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

Fig. 3. Formula for Root Mean Squared Error (RMSE)

R-squared (R^2)
Represents the proportion of the variance in the dependent variable (AQI) that is predictable from the independent variables (pollutants). It measures how well the model fits the data. An R^2 value closer to 1 means the model explains most of the variability in the data.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

Fig. 4. Formula for R-squared (R^2)

To assess the performance of the **Logistic Regression** model, classification metrics such as Accuracy, Precision, Recall, and F1-Score were used. These metrics provide insight into how well the model classified AQI categories such as Moderate, Poor, Satisfactory, Severe, and Very Poor. Accuracy measures the overall proportion of correct predictions, while Precision and Recall evaluate class-specific performance. The F1-Score offers a balanced measure between Precision and Recall, especially useful when classes are imbalanced.

IV. Results and Discussion

A. Model Performance

The performance of the three regression models Linear Regression, Decision Tree Regression, Random Forest Regression, was evaluated using the R^2 (Coefficient of Determination) and Mean Squared Error (MSE) metrics. These metrics measure how well the models predict the next-day Air Quality Index (AQI) based on pollutant data.

The **Linear Regression** Model achieved an **R^2** score of 0.8329 and an MSE of 2159.48. This indicates that approximately 83.29 percent of the variance in AQI values can be explained by the model. While the model performed well, its error rate suggests that it may not capture more complex, nonlinear relationships in the data.

The **Random Forest** model achieved the highest performance among the three, with an R^2 score of 0.8857 and an **MSE of 1477.29.** This demonstrates that the Random Forest model explained 88.57% of the variance in AQI values and had the lowest prediction error, making it the most accurate and reliable model for predicting next-day AQI levels.

The **Decision Tree** model achieved an R^2 score of 0.7758 and an MSE of 2896.84 Although it captured certain nonlinear patterns in the data, it showed lower accuracy compared to Linear Regression and Random Forest, possibly due to overfitting or limited generalization.

The **Logistic Regression** model achieved an accuracy of 67 percent, indicating that it correctly classified approximately two-thirds of the AQI category labels. While the model performed well in detecting Severe (F1-score: 0.85) and Very Poor (F1-score: 0.80) air quality levels, it struggled with the Satisfactory class (F1-score: 0.00), likely due to class imbalance or limited samples for that category.

B. Visualization of Results

To evaluate the performance of each model, an Actual vs. Predicted AQI plot was generated comparing the true AQI values with the predictions made by Linear Regression, Random Forest, and Decision Tree models. As shown in the graph, the Random Forest model's predictions closely follow the actual AQI trend, indicating higher accuracy and stability. The Linear Regression model captures the general trend but shows larger deviations during rapid AQI fluctuations. Meanwhile, the Decision Tree model shows more variance and overfitting tendencies, as it often deviates more significantly from the actual AQI curve. The visualization supports the results highlighting Random Forest as the most reliable model for predicting next-day AQI.

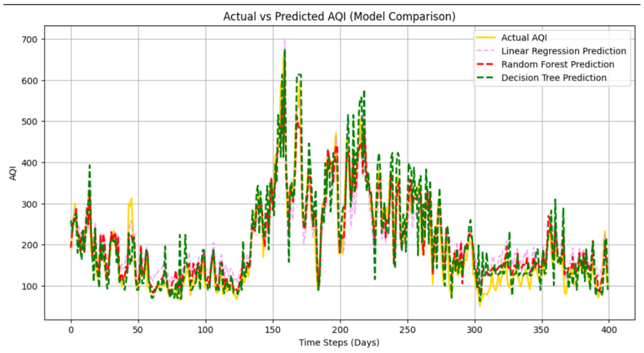

Fig. 5. Actual vs. Predicted AQI values for Linear Regression, Decision Tree, and Random Forest models.

The feature importance analysis using the Random Forest model provides insights into which pollutant variables have the greatest impact on predicting the Air Quality Index (AQI). Based on the results, $PM_{2.5}$ and PM_{10} emerged as the most influential features, demonstrating that particulate matter levels are essential in assessing air quality. Other pollutants such as NO_2, CO, SO_2, and O_3 contributed less to the model's predictions but still provided valuable contextual information.

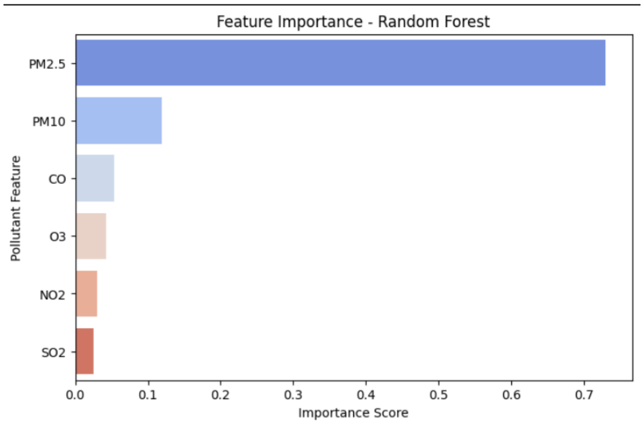

Fig. 6. Feature importance plot from the Random Forest model showing the relative contribution of each pollutant to AQI prediction.

The residual distribution visualization displays the difference between the predicted and actual AQI values. It helps assess whether the Random Forest model makes unbiased predictions. A well-performing model should have residuals centered around zero and symmetrically distributed, indicating that the model does not systematically overestimate or underestimate AQI values.

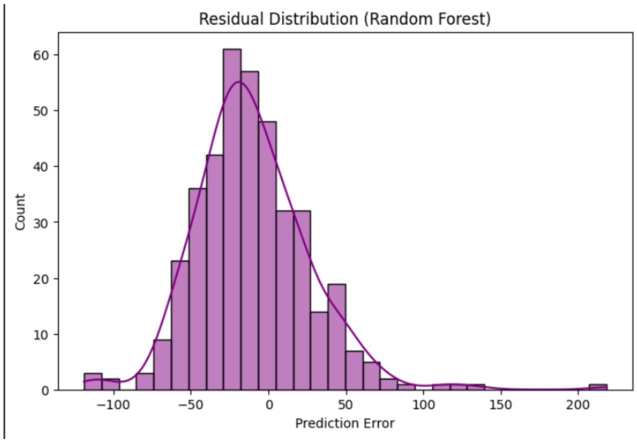

Fig. 7. Residual distribution of the Random Forest model showing how prediction errors are distributed around zero.

The following visualizations compare the performance of three regression models Linear Regression, Random Forest, and Decision Tree using two key metrics: R² Score and Root Mean Squared Error (RMSE). From the charts, the Random Forest model achieved the highest R² and the lowest RMSE, confirming it as the best-performing model for predicting next-day AQI values.

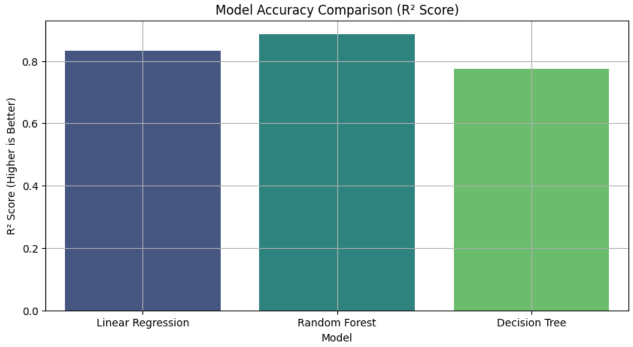

Fig. 8. Comparison of R² scores among Linear Regression, Decision Tree, and Random Forest models.

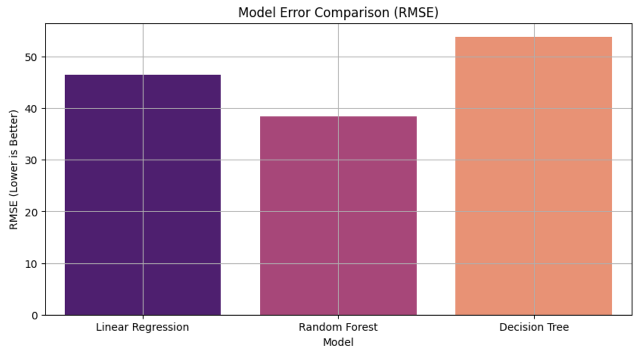

Fig. 9. Comparison of RMSE values showing Random Forest as the most accurate model among the three.

The confusion matrix visualizes the performance of the Logistic Regression model in classifying AQI levels into categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. A stronger diagonal pattern indicates higher classification accuracy, while significant off-diagonal values suggest misclassifications.

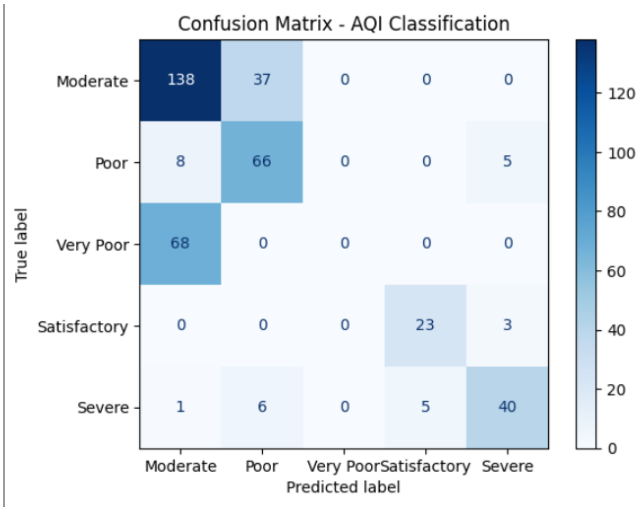

Fig. 10. Confusion matrix displaying the classification performance of the Logistic Regression model for AQI categories.

C. Discussion

The models developed in this project aimed to predict and classify Air Quality Index (AQI) values based on various environmental factors such as PM2.5, PM10, NO, SO, CO, and O. After training and evaluating multiple models Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Logistic Regression their performances were compared to determine the most effective approach for AQI prediction and categorization.

Among the regression models, the Random Forest Regressor achieved the highest accuracy with an R^2 score of 0.88 and the lowest error (MSE = 1477.29), indicating its strong capability in capturing complex, nonlinear relationships in the data. The Linear Regression model performed fairly well (R^2 = 0.83) but was limited by its assumption of linearity. Meanwhile, the Decision Tree Regressor achieved moderate accuracy (R^2 = 0.77) but tended to overfit the data, as indicated by higher prediction errors compared to Random Forest.

The Logistic Regression model, used for AQI category classification, achieved an overall accuracy of 0.67. While this performance is moderate, it indicates that the model can classify general air quality conditions with reasonable reliability.

Visualizations such as the Actual vs. Predicted AQI plots and Feature Importance graphs supported these findings. Random Forest placed higher importance on PM2.5 and PM10, showing that particulate matter concentrations were the dominant predictors of AQI levels.

V. Conclusion and Future Work

The project successfully developed and evaluated multiple regression and classification models Linear Regression, Decision Tree, Random Forest, and Logistic Regression to predict and assess the Air Quality Index (AQI). Among these, the Random Forest Regressor demonstrated the best performance, achieving the highest R^2 (0.88) and lowest MSE (1477.29), indicating strong accuracy and robustness. The Linear Regression model performed adequately for linear relationships, while the Decision Tree tended to overfit the data. The Logistic Regression model achieved a classification accuracy of 0.67, showing fair capability in categorizing AQI levels.

This project still has limitations. The dataset contained missing values and may not fully represent seasonal or regional variations in air quality. For future researcher, the results could be further improved by integrating additional environmental and meteorological factors such as temperature, humidity, and wind speed. Researchers may also apply more advanced optimization and feature selection techniques to enhance model performance.

References

[1] N. Sarkar, P. Singh, and A. Gupta, “Hybrid deep learning models for air quality index prediction in delhi,” *Environmental Pollution Research*, vol. 35, no. 4, pp. 1234–1245, 2022.

[2] H. Liu, F. Russo, and J. Li, “Air quality index and nitrogen oxide prediction using svr and random forest models,” *Atmospheric Environment*, vol. 199, pp. 456–467, 2019.

[3] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, “Optimized machine learning model for air quality index prediction in major cities in india,” *Scientific Reports*, 2024.

[4] R. Rao, “Air quality data in india (2015–2020),” *Kaggle Open Datasets*, 2020, available at: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india, Accessed: 2025-10-14.

[2] [3] [4]