



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Air Quality Prediction Using Machine Learning

S. Harsha Varthini¹, J. Steffi²

¹II MCA, Reg.No: 25081205300112014, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli.

²Assistant Professor, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli.

ABSTRACT

Air pollution is one of the major environmental challenges affecting human health and environmental sustainability. Accurate prediction of air quality can support the assessment of pollution levels and facilitate appropriate preventive actions. This study proposes a machine learning-based system for predicting the Air Quality Index (AQI) using major air pollutant parameters. The dataset consists of 18,265 records, from which nine pollutant features, namely PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃, are selected as input variables. Data preprocessing is performed by handling missing numerical values using mean imputation and removing duplicate records. Four machine learning regression algorithms, namely Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression, are implemented and compared using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score. Experimental results show that the Random Forest Regressor achieves the best performance with an MAE of 1.0396, RMSE of 6.3586, and R² Score of 0.9970. The selected model is integrated into a Flask-based web application for AQI prediction and categorization of the predicted AQI value into the appropriate AQI category. In addition, the system provides an AQI level meter, health advice, precautions, recommended actions, air quality improvement measures, and an overall recommendation based on the predicted AQI level. The proposed system combines machine learning-based AQI prediction with a user-oriented interface to provide an understandable and informative assessment of air quality.

Keywords: Air Quality Index, Air Pollution, Machine Learning, Random Forest, AQI Prediction, Regression, Flask, Health Advice.

1. INTRODUCTION

Air pollution has become a significant environmental concern due to increasing industrial activities, transportation, urbanization, and the continuous emission of pollutants into the atmosphere. Pollutants such as particulate matter and gaseous substances can adversely affect air quality and may create risks to human health and the environment. Therefore, monitoring and predicting air quality are important for understanding pollution conditions and supporting appropriate preventive actions.

The **Air Quality Index (AQI)** is commonly used to represent the overall condition of air quality based on the concentrations of major pollutants. AQI provides a simplified numerical representation of air pollution levels, making it easier to understand whether the air quality is acceptable or requires precautionary measures. However, conventional air quality monitoring mainly depends on monitoring stations and measured pollutant concentrations. Developing prediction-based approaches can provide an additional method for estimating AQI from pollutant data.

Machine learning techniques have become useful for analysing environmental datasets and identifying relationships between input variables and target values. In air quality prediction, machine learning models can learn patterns from historical pollutant measurements and use those patterns to estimate AQI values. Different regression algorithms can be evaluated to determine which model provides better prediction performance for a given dataset.

In this study, a machine learning-based AQI prediction system is developed using nine major pollutant parameters: **PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃**. The dataset is examined and preprocessed to ensure consistency before model development. The selected pollutant features are divided into training and testing datasets using an **80:20 ratio**. Four regression algorithms, namely **Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regression**, are trained and evaluated.

The models are compared using **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score** to assess their prediction performance. Based on the experimental evaluation, the **Random Forest Regressor** is selected as the final AQI prediction model.

To make the prediction results more understandable to users, the trained model is integrated into a **Flask-based web application**. The application accepts pollutant values as input and displays the predicted AQI along with its corresponding AQI category. In addition, the system provides an **AQI level meter, health advice, precautions, recommended actions, air quality improvement measures, and an overall recommendation**. Thus, the proposed work combines machine learning-based AQI prediction with a user-oriented decision-support interface for providing an understandable assessment of air quality.

2. LITERATURE REVIEW

Several studies have investigated the application of machine learning techniques for air quality and Air Quality Index (AQI) prediction using different datasets, geographical regions, pollutant parameters, and predictive models.

Kumar and Pande [1] investigated air pollution prediction using machine learning techniques based on air pollution data from multiple Indian cities. The study examined pollution-related features through data preprocessing, exploratory analysis, and feature selection and evaluated multiple machine learning models for prediction. The study demonstrated the applicability of machine learning techniques for analysing air pollution patterns and predicting pollution-related measures across Indian cities. The findings indicate that appropriate preprocessing, feature selection, and comparative model evaluation are important components of machine learning-based air quality prediction systems.

Ravindiran et al. [2] developed machine learning models for AQI prediction in Visakhapatnam, India. The study considered 12 air contaminants together with 10 meteorological parameters and evaluated models including LightGBM, Random Forest, CatBoost, AdaBoost, and XGBoost. CatBoost achieved an R^2 value of 0.9998, an MAE of 0.60, and an RMSE of 0.76, while the contribution of meteorological factors to AQI prediction was found to be negligible for the studied dataset. The study demonstrates the potential of machine learning and ensemble-based models for AQI prediction using historical air-quality data.

Natarajan et al. [3] proposed an optimized machine learning approach for AQI prediction in major Indian cities. The study used air-quality data from cities including New Delhi, Hyderabad, Kolkata, Bangalore, Visakhapatnam, and Chennai. A Grey Wolf Optimization (GWO) technique was combined with a Decision Tree regression model, and its performance was compared with models such as K-Nearest Neighbors, Random Forest Regression, and Support Vector Regression. The models were evaluated using metrics including R^2 , RMSE, MSE, MAE, and accuracy. The study demonstrated the application of optimization techniques for improving AQI prediction across multiple Indian cities.

Mourya and Tawde [4] presented a machine learning-based approach for AQI prediction and air-quality classification. The study investigated regression and classification techniques for analysing air-quality data and examined models such as Decision Tree and Random Forest for prediction and classification tasks. The work demonstrates the applicability of machine learning techniques for analysing AQI values and categorizing air-quality conditions.

Mandvi et al. [5] evaluated predictive models for AQI prediction in Gorakhpur, Uttar Pradesh, India, using air-quality data obtained from the Central Pollution Control Board. The study considered PM₁₀, PM_{2.5}, SO₂, and NO₂ as major pollutant parameters and evaluated Random Forest and Linear Regression models. Random Forest achieved an MAE of 0.79876, RMSE of 1.56615, and R^2 of 0.99895, whereas Linear Regression obtained an MAE of 10.71808, RMSE of 13.8176, and R^2 of 0.91848. The results demonstrate the effectiveness of Random Forest for the studied AQI prediction problem.

Maltare and Vahora [6] investigated AQI prediction for Ahmedabad city using air-quality data obtained from the Central Pollution Control Board of India. The study compared SARIMA, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) approaches. Different preprocessing techniques were investigated for handling the AQI dataset, including missing-value management and outlier-related preprocessing. The study particularly examined different SVM kernels and reported comparatively better performance for the RBF-kernel SVM model for the Ahmedabad dataset.

Almaliki et al. [7] investigated AQI prediction in Holy Makkah, Saudi Arabia, using machine learning classification methods. Air-quality data collected between 2016 and 2018 were used to develop models including Fine Decision Tree, Ensemble Boosted Tree, and Ensemble Bagged Tree. The reported testing accuracies were 96.1%, 97.4%, and 94.8%, respectively, with the Ensemble Boosted Tree achieving the highest reported testing accuracy among the evaluated models. The study demonstrates the application of tree-based and ensemble machine learning techniques for AQI category prediction.

Gupta et al. [8] presented a comparative study of machine learning techniques for AQI prediction using data from New Delhi, Bangalore, Kolkata, and Hyderabad. The study evaluated Support Vector Regression, Random Forest Regression, and CatBoost Regression and investigated the effect of the Synthetic Minority Oversampling Technique (SMOTE) on the dataset. The results showed differences in model performance between balanced and imbalanced datasets, with Random Forest and CatBoost providing strong results for different city datasets. The study highlights the importance of data distribution and preprocessing techniques in machine learning-based AQI prediction.

Abdelmalek et al. [9] proposed a machine learning-based approach for AQI and air-pollution prediction using hourly air-pollution data from Gazipur, Bangladesh. The study evaluated five machine learning models, namely Gaussian Process Regression, Ensemble Regression, Support Vector Machine, Regression Tree, and Kernel Approximation Regression. The work also investigated pollutant feature selection and comparative model performance. The study demonstrates the importance of selecting relevant pollutant parameters and evaluating multiple machine learning regression techniques for AQI prediction.

Muni Lakshmi G. K. and Mokesh Rayalu G. [10] proposed a regression-based hybrid machine learning framework for AQI prediction in Kolkata. The study investigated multiple machine learning, deep learning, and ensemble models using air pollutants and meteorological variables. The models were evaluated using metrics including RMSE, MAE, MAPE, RMSLE, and R^2 , with both full and reduced feature configurations considered to investigate feature dependence and potential target dependency. The Voting and Stacking ensemble regressors showed strong predictive performance, while the analysis identified PM_{2.5} and PM₁₀ as important features for AQI prediction. The study highlights the continued development of hybrid and ensemble approaches for AQI prediction.

Overall, the reviewed studies demonstrate that machine learning techniques can be effectively applied to AQI prediction and air-quality classification using pollutant and, in some cases, meteorological parameters. The literature includes tree-based models, ensemble methods, support vector approaches, optimization techniques, time-series methods, and hybrid machine learning approaches. The studies also demonstrate the importance of data preprocessing, feature selection, model comparison, and appropriate evaluation metrics in developing AQI prediction systems.

However, the reviewed studies differ in their datasets, geographical regions, input parameters, algorithms, and implementation objectives. Several studies focus on individual cities or specific geographical regions, while others use multiple-city datasets. Some studies additionally incorporate meteorological variables or advanced optimization and hybrid techniques. These differences indicate that model performance can depend on the characteristics of the dataset and selected input features. Therefore, comparative evaluation using a consistent set of pollutant parameters and regression algorithms remains relevant for developing an AQI prediction system for a specific dataset.

The present study addresses this requirement by evaluating four regression algorithms—Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression—using nine major pollutant parameters: PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃. The models are evaluated using MAE, RMSE, and R² Score on the same dataset and testing data. In addition, the selected prediction model is integrated into a Flask-based web application that presents the predicted AQI together with its category, AQI level meter, health advice, precautions, recommended actions, and air-quality improvement measures. Thus, the proposed work combines comparative machine learning-based AQI prediction with a user-oriented web-based presentation of prediction results.

3. RESEARCH GAP

The reviewed studies demonstrate that machine learning techniques can provide effective approaches for Air Quality Index (AQI) prediction using historical air-quality data. However, differences in datasets, geographical regions, input parameters, algorithms, and system objectives indicate several areas that can be further investigated.

First, several existing studies focus on **specific cities or geographical regions**, such as Visakhapatnam, Ahmedabad, Gorakhpur, Makkah, Gazipur, and Kolkata. Although these studies demonstrate the effectiveness of machine learning for their respective datasets, model performance may vary depending on the characteristics of the dataset, pollutant distribution, and selected input parameters. Therefore, further evaluation using a different air-quality dataset containing a broader set of pollutant parameters is relevant.

Second, a number of existing studies incorporate **meteorological parameters** such as temperature, humidity, wind speed, atmospheric pressure, and precipitation along with pollutant concentrations. While these parameters can provide additional information for air-quality analysis, a system based primarily on major pollutant parameters can provide a simpler input structure for AQI prediction. Therefore, the effectiveness of using multiple major pollutant parameters without requiring additional meteorological inputs can be investigated.

Third, the reviewed literature includes a wide range of machine learning approaches, including **Decision Tree, Random Forest, Support Vector methods, Gradient Boosting, optimization-based models, time-series models, and hybrid or ensemble techniques**. However, the reported results are obtained from different datasets and experimental settings, making direct comparison between studies difficult. Therefore, comparative evaluation of multiple regression algorithms using the **same dataset, same input features, same train-test split, and same evaluation metrics** is useful for identifying a suitable model for a particular AQI prediction task.

Fourth, several existing studies primarily concentrate on **model prediction performance**, while the presentation of prediction results to end users varies between systems. A numerical AQI prediction alone may not provide sufficient information for users to understand the corresponding air-quality condition and appropriate precautions. Therefore, there is a need for a user-oriented interface that presents the predicted AQI together with its category and understandable guidance.

To address these gaps, the proposed system uses an air-quality dataset containing **18,265 records** and considers nine major pollutant parameters: **PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃**. The attributes City, Datetime, and AQI_Bucket are excluded from the regression input features, while AQI is used as the target variable. Missing numerical values are handled using mean imputation and duplicate records are removed during preprocessing.

Four regression algorithms, namely **Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regression**, are trained and evaluated using the same dataset and an **80:20 train-test split**. Their performance is compared using **MAE, RMSE, and R² Score** to provide a consistent evaluation of the selected approaches.

In addition to numerical prediction, the selected model is integrated into a **Flask-based web application**. The application presents the predicted AQI, corresponding AQI category, AQI level meter, health advice, precautions, recommended actions, air-quality improvement measures, and an overall recommendation. This provides a user-oriented presentation of the machine learning prediction rather than limiting the system output to a numerical AQI value.

Therefore, the proposed work addresses the identified gaps by combining comparative evaluation of multiple machine learning regression algorithms using a consistent experimental setup with a web-based user-oriented AQI prediction and guidance system.

4. PROPOSED METHODOLOGY

The proposed system is designed to predict the Air Quality Index (AQI) from major air pollutant parameters using machine learning techniques and to present the prediction results through a Flask-based web application. The overall methodology consists of dataset preparation, data preprocessing, feature selection, train-test splitting, machine learning model development, model comparison, performance evaluation, prediction model selection, and web-based AQI prediction with health guidance.

4.1 Dataset Preparation

The study uses an air quality dataset containing 18,265 records and 13 attributes, including City, Datetime, PM2.5, PM10, NO, NO₂, NO_x, NH₃, CO, SO₂, O₃, AQI, and AQI_Bucket. The dataset contains pollutant measurements along with their corresponding AQI values. The AQI value is considered the target variable for prediction.

The dataset is initially loaded using the Pandas library. The first few records, column names, dataset information, missing values, and duplicate records are examined to understand the structure and quality of the dataset.

4.2 Data Preprocessing

Data preprocessing is performed to improve the consistency and reliability of the dataset before model development. A copy of the original dataset is created so that the original data remains unchanged.

Missing numerical values are handled using mean imputation, where the mean value of the corresponding numerical column is used to replace missing values. Duplicate records are then identified and removed from the dataset. After preprocessing, the dataset is verified again to ensure that missing values and duplicate records have been appropriately handled.

The resulting preprocessed dataset is saved as `Air_quality_data_preprocessed.csv` for subsequent model development and comparison.

4.3 Feature Selection

Nine major pollutant parameters are selected as input features for AQI prediction. The selected features are:

- PM2.5
- PM10
- NO
- NO₂
- NO_x
- NH₃
- CO
- SO₂
- O₃

The AQI value is selected as the target variable. The attributes City, Datetime, and AQI_Bucket are excluded from the regression models. Therefore, the prediction task is formulated as a regression problem in which the nine pollutant parameters are used to estimate the numerical AQI value.

4.4 Train-Test Data Splitting

The preprocessed dataset is divided into training and testing subsets using the `train_test_split` method. An 80:20 ratio is used, where 80% of the data is used for model training and the remaining 20% is used for testing.

A `random_state` value of 42 is used to maintain a reproducible data split. The training and testing datasets contain the selected nine pollutant features and their corresponding AQI target values.

4.5 Machine Learning Model Development

Four regression algorithms are implemented and compared to identify a suitable model for AQI prediction. The selected algorithms are:

1. Decision Tree Regressor

2. Random Forest Regressor
3. Gradient Boosting Regressor
4. Support Vector Regression (SVR)

Each model is trained using the same training dataset and evaluated using the same testing dataset. This provides a consistent basis for comparing the prediction performance of the different regression algorithms.

For the Random Forest Regressor, 100 decision trees are used with a `random_state` value of 42. The model learns the relationship between the nine pollutant parameters and the corresponding AQI values from the training data.

4.6 Model Performance Evaluation

The performance of the regression models is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score.

Mean Absolute Error (MAE) measures the average absolute difference between the actual and predicted AQI values. A lower MAE indicates smaller prediction errors.

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE) measures the square root of the average squared prediction errors. It gives greater importance to larger prediction errors, and a lower RMSE indicates better prediction performance.

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

R² Score represents the proportion of variation in the target AQI values that is explained by the model. A higher R² value indicates a better fit between the predicted and actual values.

For presentation purposes, the R² score is also converted into a percentage using:

$$R^2\text{-based Performance (\%)} = R^2 \times 100$$

This percentage is used only as an R²-based performance representation and is not treated as classification accuracy.

4.7 Comparative Model Analysis

The four trained regression models are evaluated using the same MAE, RMSE, and R² metrics. The obtained evaluation results are compared to identify the model that provides suitable prediction performance for the selected dataset.

The comparison considers the prediction error and the proportion of variation in AQI values explained by each model. The detailed experimental results and comparative values obtained from the four regression algorithms are presented in Section 5.

In addition to the numerical comparison, separate graphs are generated for MAE, RMSE, and R²-based performance. A combined graph is also generated to provide a visual comparison of the evaluation measures.

4.8 Selection and Saving of the Prediction Model

Based on the comparative experimental evaluation, the Random Forest Regressor achieves an MAE of 1.0396, an RMSE of 6.3586, and an R² Score of 0.9970 on the test dataset. Based on these evaluation results, the trained Random Forest model is selected as the prediction model for integration with the developed web application. The trained model is serialized and saved as `model.pkl` using the `Joblib` library. This allows the trained model to be loaded later without retraining whenever a user requests an AQI prediction.

4.9 Web-Based AQI Prediction

The selected model is integrated into a Flask-based web application. The prediction interface provides input fields for the nine pollutant parameters used during model training. The user enters the pollutant values through the web interface. The Flask application receives the submitted values through a POST request and converts them into numerical values. The values are arranged in the same feature order used during model training and passed to the saved

Random Forest model. The trained model generates a numerical AQI prediction. The predicted value is then displayed on the result page along with a summary of the input pollutant values.

4.10 AQI Categorization and Health Guidance

After obtaining the predicted AQI value, the application assigns the predicted value to an AQI category based on predefined AQI ranges. The categories implemented in the system are Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. Based on the predicted category, the system provides corresponding health advice, precautions, recommended activities, air-quality improvement measures, and an overall recommendation. Recommended activities are represented using status labels such as Safe, Limit, and Avoid.

The result page also provides an AQI level meter ranging from 0 to 500 to provide a visual representation of the predicted AQI level.

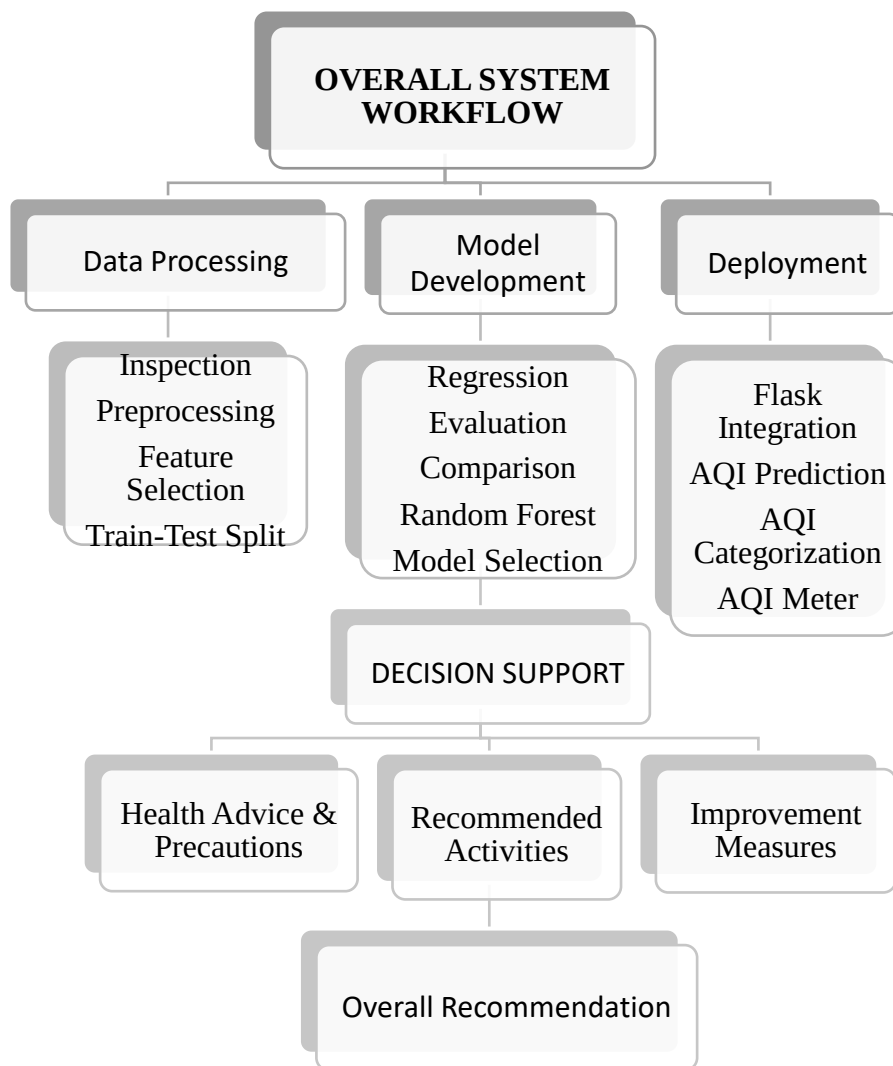
4.11 Model Comparison and Result Visualization

A separate model comparison interface is developed to present the experimental results. The comparison results generated during model evaluation are stored in a CSV file and loaded by the Flask application.

The comparison page displays the performance table of the four regression algorithms along with MAE, RMSE, and R^2 -based performance graphs. This provides a visual representation of the differences in prediction performance among the evaluated models.

4.12 Overall System Workflow

The complete workflow of the proposed system can be summarized as:



The proposed methodology therefore combines machine learning-based numerical AQI prediction with a web-based user interface that presents the prediction results and associated air-quality guidance in an understandable form.

5. RESULTS AND DISCUSSION

The performance of the proposed air quality prediction system was evaluated using four machine learning regression algorithms: Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regression (SVR). The selected pollutant features were used as input variables, and the dataset was divided into 80% training data and 20% testing data. The training set contained 14,612 records, while the testing set contained 3,653 records.

The performance of each regression model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. The R^2 Score was also expressed as a percentage for comparative presentation. For MAE and RMSE, lower values indicate lower prediction errors, whereas a higher R^2 Score indicates that a larger proportion of the variation in the target AQI values is explained by the model.

5.1 Discussion of Model Performance

The experimental results show that the Random Forest Regressor achieved the strongest performance among the four evaluated models on the selected test dataset. It obtained an MAE of 1.0396, indicating a small average absolute difference between the predicted and actual AQI values. The model also achieved an RMSE of 6.3586, indicating relatively low prediction error on the test dataset.

The Random Forest model obtained an R^2 Score of 0.9970. When represented as an R^2 -based percentage, this corresponds to 99.70%. This indicates that approximately 99.70% of the variation in the test-set AQI values is explained by the model under the experimental setup used in this study. This percentage should not be interpreted as classification accuracy.

The Decision Tree Regressor also demonstrated strong predictive performance, with an MAE of 1.2324, RMSE of 9.2476, and R^2 Score of 0.9936. Compared with Random Forest, it produced slightly higher prediction errors and a slightly lower R^2 Score.

The Gradient Boosting Regressor obtained an MAE of 11.5057, RMSE of 17.9456, and R^2 Score of 0.9758. Although the R^2 Score was relatively high, its MAE and RMSE were higher than those obtained by the Decision Tree and Random Forest models under the same experimental setup.

The Support Vector Regression model obtained an MAE of 64.2074, RMSE of 85.6841, and R^2 Score of 0.4488. These results indicate substantially larger prediction errors and a lower proportion of explained variation compared with the other evaluated models.

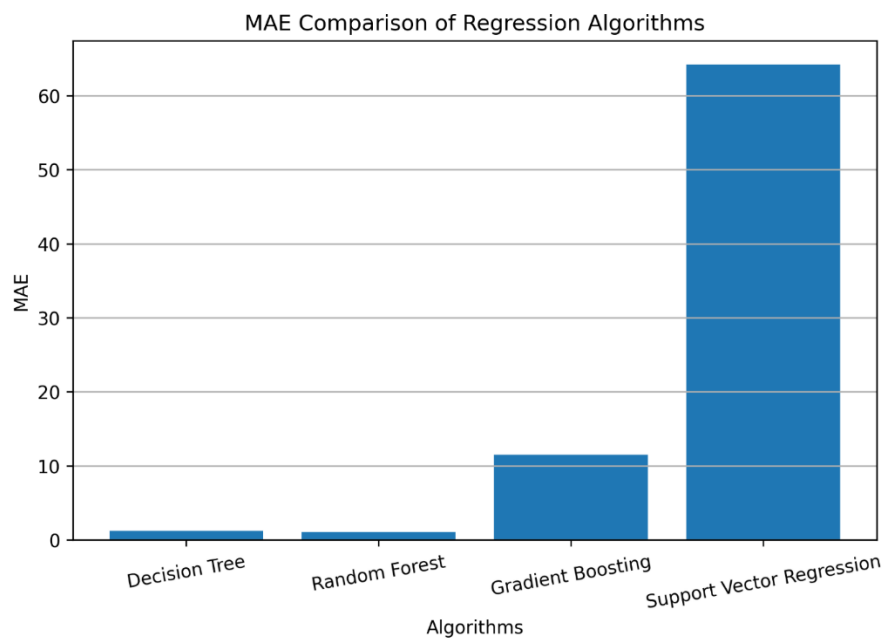
Based on the combined evaluation of MAE, RMSE, and R^2 Score, the Random Forest Regressor was selected as the prediction model for integration into the proposed system.

5.2 MAE Comparison

The MAE results provide a comparison of the average absolute prediction error produced by each regression model. The Random Forest Regressor achieved the lowest MAE of 1.0396, followed by the Decision Tree Regressor with an MAE of 1.2324.

The Gradient Boosting Regressor obtained an MAE of 11.5057, while Support Vector Regression produced a considerably higher MAE of 64.2074. Therefore, under the experimental conditions of this study, Random Forest produced the lowest average absolute prediction error among the evaluated models.

The figure 1 illustrates the MAE comparison of the four regression algorithms, showing that the Random Forest Regressor achieves the lowest MAE and therefore provides better prediction performance compared with the other models.

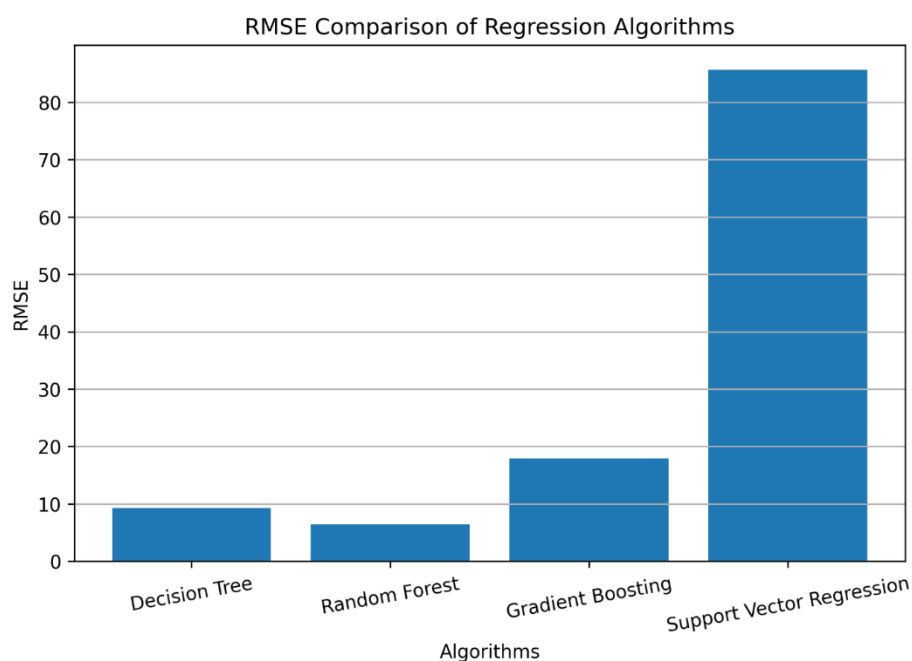
Figure 1. MAE Comparison of Regression Algorithms

5.3 RMSE Comparison

The RMSE results further illustrate the differences in prediction error among the evaluated models. The Random Forest Regressor achieved the lowest RMSE of 6.3586, while the Decision Tree Regressor obtained an RMSE of 9.2476.

The Gradient Boosting Regressor produced an RMSE of 17.9456, whereas Support Vector Regression obtained the highest RMSE of 85.6841. The lower RMSE obtained by Random Forest indicates lower squared prediction error on the test dataset compared with the other evaluated models.

The figure 2 illustrates the RMSE comparison of the four regression algorithms, showing that the Random Forest Regressor achieves the lowest RMSE value and provides better prediction performance compared with the other models.

Figure 2. RMSE Comparison of Regression Algorithms

5.4 R² Performance Comparison

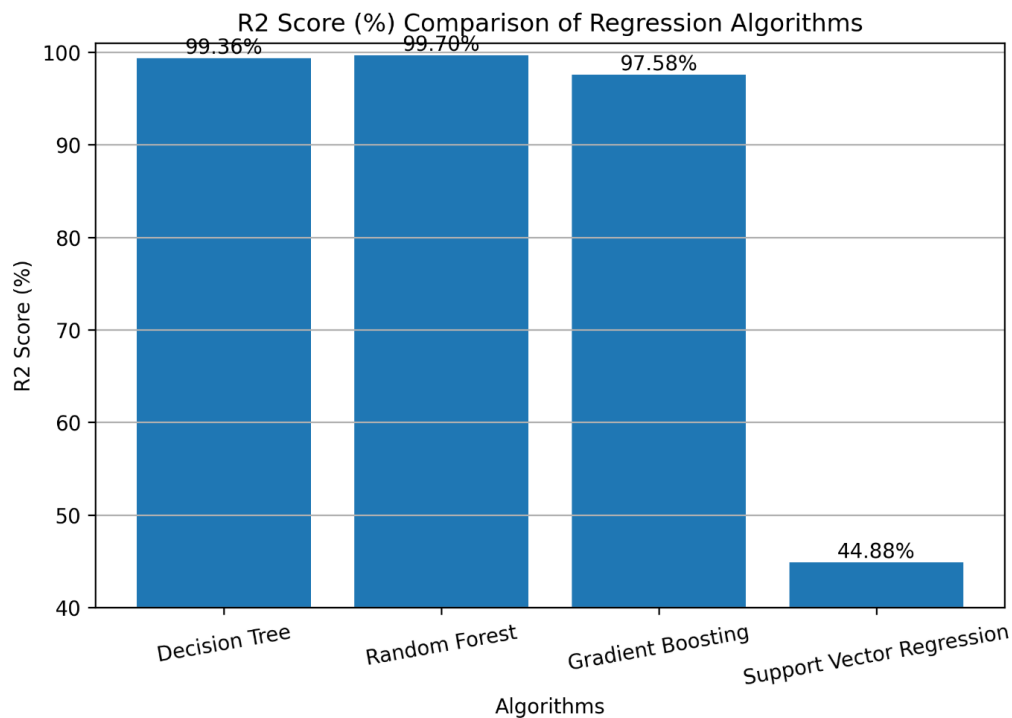
The R² Score was used to evaluate how well the regression models explained the variation in the target AQI values. The Random Forest Regressor achieved the highest R² Score of 0.9970, corresponding to an R²-based performance representation of 99.70%.

The Decision Tree Regressor achieved an R² Score of 0.9936, corresponding to 99.36%, while Gradient Boosting obtained an R² Score of 0.9758, corresponding to 97.58%. Support Vector Regression achieved an R² Score of 0.4488, corresponding to 44.88%.

The R² comparison therefore indicates that Random Forest explained the largest proportion of the variation in the test-set AQI values among the evaluated models.

The figure 3 illustrates the comparison of R² scores (%) for the four regression algorithms, showing that the Random Forest Regressor achieves the highest R² score and provides the best overall prediction performance.

Figure 3. R² Score (%) Comparison of Regression Algorithms



5.5 Overall Performance Comparison

A combined comparison was performed using MAE, RMSE, and R²-based Performance (%) for all four regression models. The comparison provides an overall view of the prediction errors and explained variation obtained by the evaluated algorithms.

The Random Forest model obtained a combination of low MAE, low RMSE, and high R² Score in the experimental evaluation. The Decision Tree also demonstrated strong performance, while Gradient Boosting produced higher prediction errors under the same test conditions. Support Vector Regression produced substantially higher MAE and RMSE and a lower R² Score compared with the other evaluated models.

The performance results obtained from the four machine learning regression algorithms are presented in Table 1.

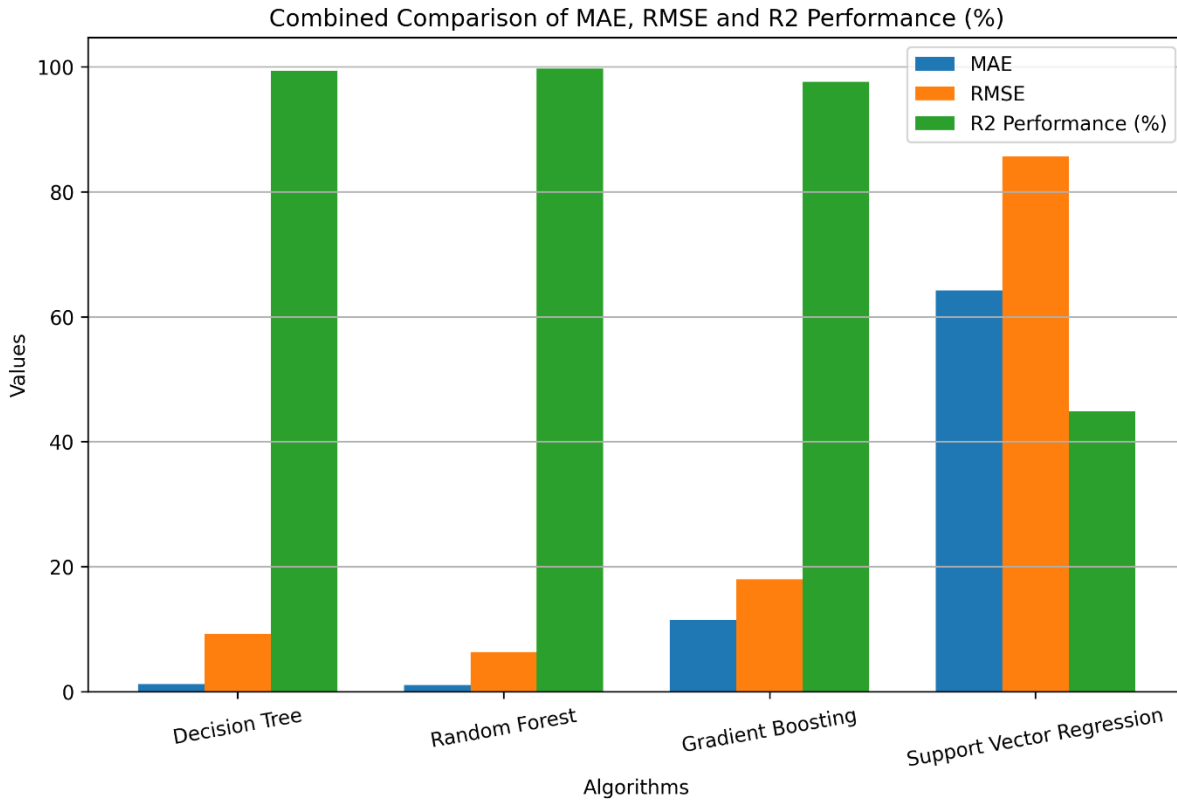
Table 1. Performance Comparison of Machine Learning Regression Models

Machine Learning Model	MAE	RMSE	R ² Score	R ² -based Performance (%)
Decision Tree Regressor	1.2324	9.2476	0.9936	99.36
Random Forest Regressor	1.0396	6.3586	0.9970	99.70
Gradient Boosting Regressor	11.5057	17.9456	0.9758	97.58
Support Vector Regression	64.2074	85.6841	0.4488	44.88

The results show differences in prediction performance among the four evaluated regression algorithms. Random Forest obtained the lowest MAE and RMSE and the highest R² Score on the test dataset.

The figure 4 illustrates the combined comparison of MAE, RMSE, and R² performance (%) for the four regression algorithms, demonstrating that the Random Forest Regressor provides the best overall performance with lower error values and the highest R² score.

Figure 4. Combined Comparison of MAE, RMSE and R² Performance (%)



5.6 Integration of the Selected Model

Based on the comparative experimental evaluation, the Random Forest Regressor was selected as the prediction model for the proposed system. The trained model was saved as model.pkl using the Joblib library and subsequently integrated into the Flask-based web application.

The web application accepts the nine pollutant concentrations used during model training: **PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃**.

The Flask application receives the pollutant values, arranges them in the required feature order, and passes them to the saved Random Forest model. The model then generates a numerical AQI prediction. The predicted value is subsequently processed by the application to determine the corresponding AQI category and to generate the associated guidance.

5.7 Overall Discussion

The experimental evaluation demonstrates that the Random Forest Regressor achieved the lowest MAE and RMSE and the highest R² Score among the four evaluated regression algorithms for the selected dataset and experimental setup. The model obtained an MAE of 1.0396, an RMSE of 6.3586, and an R² Score of 0.9970.

These results support the selection of Random Forest for the subsequent web-based implementation. The trained model was integrated into a Flask application to convert the numerical prediction into an accessible user-facing result.

The proposed system therefore combines machine learning-based AQI prediction with an interactive interface that presents the predicted AQI, AQI category, visual AQI level, health guidance, precautions, recommended activities, and air-quality improvement measures. This integration provides a practical way of presenting model predictions in a form that can be interpreted by users.

6. CONCLUSION

This study presented a machine learning-based system for Air Quality Index (AQI) prediction using major air pollutant parameters. A dataset containing 18,265 records and 13 attributes was used for the development and evaluation of the proposed system. Nine pollutant parameters, namely PM_{2.5}, PM₁₀, NO, NO₂, NO_x, NH₃, CO, SO₂, and O₃, were selected as input features for predicting AQI.

The dataset was preprocessed by handling missing numerical values using mean imputation and removing duplicate records. The preprocessed data was divided into 80% training data and 20% testing data. Four machine learning regression algorithms, namely Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Regression, were implemented and evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R² Score.

Among the evaluated models, the Random Forest Regressor achieved the lowest MAE of 1.0396 and RMSE of 6.3586, along with the highest R² Score of 0.9970 on the selected test dataset. The corresponding R²-based performance was 99.70%. Based on the comparative experimental results, the Random Forest Regressor was selected as the final AQI prediction model.

The selected model was integrated into a Flask-based web application to provide a practical interface for AQI prediction. The system accepts pollutant values as input and presents the predicted AQI, corresponding AQI category, and AQI level meter. In addition, the system provides health advice, precautions, recommended actions, air quality improvement measures, and an overall recommendation based on the predicted AQI level.

Overall, the proposed system combines machine learning-based AQI prediction with a user-oriented web interface, providing an understandable presentation of air quality information. The experimental results demonstrate the effectiveness of the evaluated machine learning models for the selected dataset and pollutant features, while the web-based implementation provides a practical method for presenting prediction results and supporting air quality interpretation.

7. FUTURE ENHANCEMENT

The proposed system can be further enhanced by incorporating real-time air quality monitoring and additional environmental parameters. Future development can focus on integrating data from real-time monitoring stations or IoT-based air quality sensors to provide continuously updated AQI predictions.

Additional meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure can also be incorporated to investigate their influence on AQI prediction. The inclusion of these parameters may provide additional information for analysing variations in air quality under different environmental conditions.

Future studies can further evaluate advanced machine learning and deep learning techniques and compare their performance with the models considered in the present study. Hyperparameter optimization and feature selection techniques can also be investigated to improve model performance and generalization.

The web-based system can additionally be extended into a mobile application to provide AQI information through a more accessible platform. Notification and alert mechanisms can be incorporated to inform users when the predicted AQI reaches higher pollution levels.

Furthermore, location-based AQI prediction and visualization can be developed to allow users to obtain air quality information for specific geographical areas. Historical AQI trends, interactive visualizations, and location-wise comparisons can also be incorporated to support better understanding of air quality patterns.

These enhancements can extend the proposed system from a historical-data-based AQI prediction application into a more comprehensive and real-time air quality monitoring and decision-support system.

8. REFERENCES

- [1] 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. 5, pp. 5333–5348, 2023, doi: 10.1007/s13762-022-04241-5.
- [2] 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, Art. no. 139518, 2023, doi: 10.1016/j.chemosphere.2023.139518.
- [3] S. K. Natarajan, P. Shanmuthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, Art. no. 6795, 2024, doi: 10.1038/s41598-024-54807-1.
- [4] S. Mourya and P. Tawde, "Air Quality Index Prediction Using Machine Learning," *International Journal of Scientific Research and Engineering Methods (IJSREM)*, vol. 08, no. 03, 2024, doi: 10.55041/IJSREM29538.
- [5] Mandvi, H. K. Singh, P. K. Patel, and S. Singh, "Evaluation of Predictive Models for Air Quality Index Prediction in an Indian Urban Area," *Journal of Indian Association for Environmental Management*, vol. 44, no. 3, pp. 31–40, 2024.

-
- [6] N. N. Maltare and S. Vahora, "Air Quality Index prediction using machine learning for Ahmedabad city," *Digital Chemical Engineering*, vol. 7, Art. no. 100093, 2023, doi: 10.1016/j.dche.2023.100093.
- [7] A. H. Almaliki, A. Derdour, and E. Ali, "Air Quality Index (AQI) Prediction in Holy Makkah Based on Machine Learning Methods," *Sustainability*, vol. 15, no. 17, Art. no. 13168, 2023, doi: 10.3390/su151713168.
- [8] N. Srinivasa Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *Journal of Environmental and Public Health*, vol. 2023, Art. no. 4916267, 2023, doi: 10.1155/2023/4916267.
- [9] M. M. Abdelmalek, H. Mahmoud, and H. Shokry, "Prognosis of air quality index and air pollution using machine learning techniques," *Scientific Reports*, vol. 15, Art. no. 25890, 2025, doi: 10.1038/s41598-025-11260-y.
- [10] Muni Lakshmi G. K. and Mokesh Rayalu G., "Regression based hybrid machine learning model performance evaluation on air quality index prediction in Kolkata," *Scientific Reports*, vol. 16, Art. no. 26399, 2026, doi: 10.1038/s41598-026-54918-x.