International Journal of Innovative Research in Science and Engineering
Volume No. 12, Issue No. 04, Year 2026
ISSN : 2454-9665

AIR QUALITY PREDICTION USING MACHINE LEARNING TECHNIQUES

1. Mrs.P.V.S.N JYOTHI M.Tech.
Assistant Professor,
Department of CSE, TEC,
jyothi.che85@gmail.com

2. KOLLI LIKHITHA
likhithakolli13@gmail.com

3. SK. HASEENA BAANO
shaikginasaida1961@gmail.com

4. K. NITHIN REDDY
kolagatlanithinreddy@gmail.com

5. MD. KHAJA AMAN SHARIEFF
mdaman2049@gmail.com

***Abstract*— Air pollution is a major environmental and public health concern, with PM2.5 being one of the most harmful pollutants due to its ability to penetrate deep into the respiratory system. This paper presents a machine learning-based air quality prediction system to forecast PM2.5 concentrations and classify Air Quality Index (AQI) levels. The system utilizes historical data with parameters such as SO_2, NO_2, RSPM, SPM, and temporal features. Data preprocessing techniques, including missing value handling and feature engineering, are applied to enhance model performance. Decision Tree and Random Forest regressors are implemented and evaluated using RMSE and MAE metrics. Experimental results indicate that the Random Forest model achieves superior prediction accuracy. The model is deployed using Flask to provide real-time PM2.5 prediction and AQI classification, supporting efficient air quality monitoring and data-driven decision-making.**

***Keywords*— PM2.5, Air Pollution, Machine Learning, Decision Tree, Random Forest, Air Quality Index (AQI), Data Preprocessing, Feature Engineering, RMSE, MAE Flask Deployment.**

I. INTRODUCTION

Air pollution has emerged as one of the most serious environmental and public health challenges worldwide [1]. Rapid industrialization, urban expansion, increased vehicular emissions, and excessive use of fossil fuels have significantly contributed to the degradation of air quality, particularly in developing countries like India. Poor air quality has been linked to respiratory diseases, cardiovascular problems, reduced life expectancy, and other severe health conditions, making it a critical issue that requires immediate attention.

Among various air pollutants, Particulate Matter (PM2.5) is considered one of the most dangerous pollutants. PM2.5 refers to fine inhalable particles with a diameter of 2.5 micro meters or smaller. Due to their extremely small size, these particles can penetrate deep into the lungs and even enter the bloodstream, causing significant health risks such as asthma, lung infections, and heart diseases [2]. Continuous monitoring and accurate prediction of PM2.5 levels are therefore essential for protecting public health.

To communicate air pollution levels in a simple and understandable format, the Air Quality Index (AQI) is used. AQI converts complex pollutant concentration data into categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. In India, AQI standards are defined by the Central Pollution Control Board (CPCB). These categories help citizens understand the severity of air pollution and take necessary precautions to reduce exposure.

Traditional air quality monitoring systems rely on physical monitoring stations that provide real-time or historical pollutant data. However, these systems are limited in number, expensive to maintain, and lack intelligent data analysis and predictive capabilities [4]. As a result, they are unable to provide early warnings or future pollution trends, which are crucial for proactive environmental management.

With the advancement of Machine Learning (ML), it has become possible to analyze large volumes of environmental data, identify hidden patterns, and make accurate predictions of pollutant levels. ML techniques can enhance the efficiency of air quality monitoring

systems by enabling data-driven forecasting and reducing dependency on manual analysis.

This project focuses on developing a Machine Learning-based Air Quality Prediction System that predicts PM2.5 concentration using pollutant parameters such as SO_2, NO_2, RSPM, and SPM [3]. Decision Tree and Random Forest regression models are implemented to capture complex relationships between pollutants and improve prediction accuracy. The predicted PM2.5 values are further mapped to corresponding AQI categories, enabling better interpretation of air quality conditions.

The proposed system provides real-time prediction through a web-based interface and supports effective air quality monitoring. By combining machine learning techniques with AQI classification, the system facilitates timely awareness and supports data-driven decision-making for environmental management and public health protection.

II. LITERATURE SURVEY

Air pollution modeling has evolved from classical statistical approaches to advanced machine learning techniques to improve prediction accuracy of PM2.5 concentrations.

Early studies employed traditional statistical methods such as Ordinary Least Squares (OLS) Linear Regression for air quality prediction. This approach establishes a linear relationship between PM2.5 concentration and meteorological variables including temperature, humidity, and wind speed [6] (2018). The model estimates regression coefficients by minimizing the sum of squared errors between observed and predicted values. Due to its simplicity, interpretability, and low computational cost, OLS serves as a baseline model for environmental prediction tasks. However, its assumption of linearity limits its effectiveness in capturing complex atmospheric interactions and non-linear pollutant behavior.

To address these limitations, Support Vector Regression (SVR) was introduced as a non-linear alternative [7] (2017). SVR utilizes kernel functions, particularly the Radial Basis Function (RBF), to map input features into a higher-dimensional space where non-linear relationships can be modeled effectively. By incorporating meteorological parameters and historical pollutant data, SVR demonstrated improved generalization and predictive accuracy. Nevertheless, its performance is highly sensitive to hyperparameter selection, such as the penalty parameter and kernel coefficient, which increases model tuning complexity.

Instance-based learning methods such as k-Nearest Neighbors Regression (KNN-R) were later applied for PM2.5 estimation [8] (2019). Unlike parametric models, KNN predicts values based on similarity measures by averaging the outputs of nearest neighbors in the feature space. Standardization of input variables, including NO_2, SO_2, and RSPM, improves distance-based computations. While KNN effectively captures local and non-linear patterns, it suffers from high computational cost and memory usage, especially for large datasets.

To balance interpretability and flexibility, Generalized Additive Models (GAM) were introduced [9] (2020). GAM extends linear regression by incorporating smooth spline functions, allowing non-linear relationships between predictors and PM2.5 levels. This approach enhances modeling capability while maintaining interpretability. However, GAM still struggles with highly complex interactions among multiple environmental variables, limiting its predictive performance.

With the increase in data dimensionality and multicollinearity, regularization techniques such as Lasso (L1) and Ridge (L2) Regression were proposed [10] (2022). These methods introduce penalty terms to the loss function to prevent overfitting and improve generalization. Ridge regression stabilizes coefficient estimates, while Lasso performs feature selection by shrinking insignificant coefficients to zero. Despite these improvements, the inherent linear assumption restricts their ability to capture complex non-linear dynamics.

Ensemble learning techniques further improved prediction performance. Gradient Boosting Regression (GBR) was introduced as a sequential learning method where multiple weak learners are combined to minimize prediction error [11] (2021). By iteratively correcting residuals, GBR captures complex feature interactions and non-linear relationships. Although it achieves higher accuracy than traditional models, it is prone to overfitting and requires careful hyperparameter tuning.

An advanced version of boosting, Extreme Gradient Boosting (XGBoost), was later developed to enhance scalability and efficiency [12] (2023). XGBoost incorporates regularization, parallel processing, and tree pruning, making it highly effective for large-scale air quality prediction. By integrating ground monitoring data with satellite-derived Aerosol Optical Depth (AOD), the model achieved superior predictive accuracy. However, it demands significant computational resources and extensive hyperparameter optimization.

III. PROBLEM STATEMENT

c Air pollution monitoring in India primarily depends on monitoring stations that measure and report pollutant concentration levels. While these systems are effective in providing real-time and historical data, they mainly focus on measurement rather than intelligent analysis. They do not offer predictive estimation or deeper analytical insights based on the relationships among different pollutant parameters. As a result, the available data remains underutilized for advanced interpretation and decision-making [5].

Several key challenges arise from this limitation. Traditional monitoring systems lack predictive analysis capabilities, meaning they cannot estimate pollution

severity based on existing pollutant inputs. Additionally, raw pollutant concentration values such as SO_2, NO_2, RSPM, and SPM are often difficult for the general public to interpret without technical knowledge. There is also limited accessibility to analytical tools that can process and analyze this data effectively. Furthermore, there is an absence of a simple, user-friendly web-based system that can estimate PM2.5 levels and map them to corresponding Air Quality Index (AQI) categories.

Although pollutant measurements are readily available, there is a clear need for an intelligent system that can analyze these inputs and estimate PM2.5 concentration using Machine Learning techniques. Therefore, this project aims to design and develop a predictive model that accepts pollutant measurements as input, predicts PM2.5 concentration using Decision Tree and Random Forest algorithms, and maps the predicted value to the appropriate AQI category. The system is deployed as a web application to ensure easy public accessibility. By providing both PM2.5 prediction and AQI categorization in an understandable format, the proposed system assists users in better understanding pollution severity and promotes informed environmental awareness.

IV. Decision Tree and Random Forest

Accurate prediction of air quality levels requires efficient data preprocessing, feature extraction, and model selection mechanisms. To address the limitations of traditional statistical methods, a machine learning-based framework using models such as Decision Tree Regressor and Random Forest Regressor was proposed. The framework was designed to analyze complex relationships among environmental parameters and improve PM2.5 prediction accuracy.

The system begins with importing essential libraries required for data analysis, visualization, and machine learning. Libraries such as Pandas and NumPy are used for data handling, while Matplotlib and Seaborn are used for visualization. Scikit-learn provides tools for preprocessing, model training, and evaluation.

The dataset used in this study consists of air quality parameters such as SO_2, NO_2, RSPM, SPM, PM2.5, along with additional features like date, state, and location details. Initially, the dataset is loaded and explored to understand its structure, data types, and statistical properties.

Data preprocessing plays a crucial role in improving model performance. Missing values are identified and handled using appropriate techniques such as median imputation for numerical features and default values for categorical features. The date column is converted into a standard datetime format, and new features such as year, month, and day are extracted. This helps in capturing temporal patterns in pollution data.

Exploratory Data Analysis (EDA) is performed to understand the distribution and relationships among air quality parameters. Visualization techniques such as histograms, correlation heatmaps, and time-series plots are used to identify trends and dependencies in the dataset. These insights help in selecting relevant features for model training.

Feature engineering is then applied by selecting important variables such as SO_2, NO_2, RSPM, SPM, year, and month to predict PM2.5 levels. Categorical variables like state and area type are encoded using label encoding techniques to convert them into numerical format.

The dataset is then divided into training and testing sets using train-test split, and feature scaling is applied using StandardScaler to normalize the data. This ensures that all features contribute equally during model training.

Multiple machine learning models are trained, including Decision Tree and Random Forest. These models learn patterns from the training data and generate predictions for PM2.5 levels. The performance of each model is evaluated using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and accuracy.

Among the models, Random Forest demonstrates better performance compared to Decision Tree, as it achieves lower RMSE and MAE values along with higher prediction accuracy. Although Random Forest requires more training time, it provides more reliable and stable predictions due to its ensemble learning approach.

The best-performing model is selected and saved using joblib for future use. The system also includes visualization techniques such as actual vs predicted plots, residual analysis, and feature importance graphs to evaluate model performance effectively.

The proposed system achieves high prediction accuracy and can be used as an efficient air quality monitoring tool. It helps in predicting PM2.5 levels and supports environmental decision-making and public health awareness.

Algorithm:

The system starts with collecting air quality data and preprocessing it using cleaning, normalization, and feature engineering techniques. The processed data is then split into training and testing sets. Decision Tree and Random Forest models are trained using the training data, and their performance is evaluated using error metrics. Finally, the best-performing model is selected, saved, and used for predicting PM2.5 levels.

Pseudocode

1. Data Collection (Air Quality Dataset Input)
2. Data Preprocessing
3. Exploratory Data Analysis
4. Feature Selection
5. Train-Test Split
6. Train Machine Learning Models
7. Model Evaluation
8. Best Model Selection

9. PM2.5 Prediction
10. AQI Category Mapping

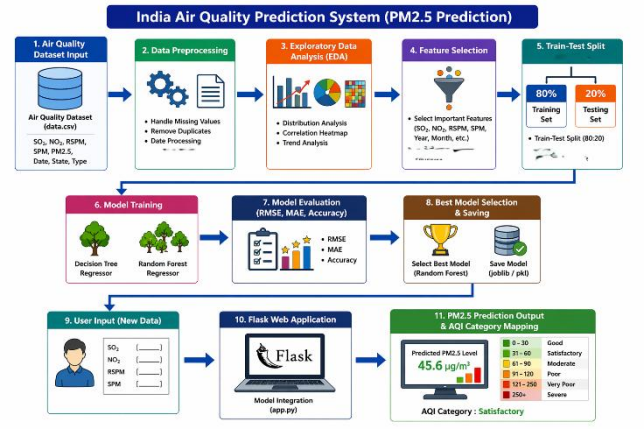

Fig 1: Proposed System Architecture Model

1. Data Collection (Air Quality Dataset Input)

This stage involves collecting the air quality dataset containing key pollutant parameters such as SO_2, NO_2, RSPM, SPM, and PM2.5, along with temporal and location-based attributes. PM2.5 is considered the target variable, while the remaining features act as input variables. The dataset serves as the foundation for building the prediction model.

2. Data Preprocessing

In this step, the dataset is cleaned and prepared for analysis. Missing values are handled using appropriate techniques such as median imputation for numerical features. Categorical variables are converted into numerical form, and date attributes are transformed to extract useful features like year and month. This ensures consistency and improves data quality.

3. Exploratory Data Analysis (EDA)

EDA is performed to understand the structure and distribution of the data. Visualization techniques such as histograms, correlation heatmaps, and trend analysis are used to identify patterns, relationships, and anomalies among pollutant parameters. This step helps in gaining insights for better model development.

4. Feature Selection

Relevant features are selected based on their importance and relationship with the target variable (PM2.5). Unnecessary or redundant attributes are removed to reduce complexity and improve model efficiency. Feature selection ensures that only meaningful inputs are used for training.

5. Train-Test Split

The dataset is divided into training and testing sets, typically in an 80:20 ratio. The training set is used to train the models, while the testing set is used to evaluate their performance. This helps in assessing how well the model generalizes to unseen data.

6. Train Machine Learning Models

Two regression models, Decision Tree Regressor and Random Forest Regressor, are trained using the prepared dataset. These models are capable of capturing non-linear relationships between features and are suitable for air quality prediction tasks.

7. Model Evaluation

The trained models are evaluated using performance metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics measure the difference between actual and predicted PM2.5 values, helping to determine model accuracy.

8. Best Model Selection

Based on evaluation results, the model with lower error values and better prediction performance is selected as the final model. In this study, Random Forest demonstrates superior performance compared to Decision Tree

9. PM2.5 Prediction

The selected model is used to predict PM2.5 values based on input pollutant parameters. These predictions represent the estimated concentration of fine particulate matter in the air

10. AQI Category Mapping

The predicted PM2.5 values are mapped to Air Quality Index (AQI) categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. This mapping converts numerical predictions into easily understandable pollution levels, aiding users in interpreting air quality conditions.

V. RESULT ANALYSIS

Regression Performance Metrics

Metrics	Values
RMSE	3.6039
MAE	0.3391

Table 1: Regression Performance Metrics of Random Forest Model

Table 1 shows the performance of the Random Forest model using regression metrics. The model achieved an RMSE value of 3.6039 and an MAE value of 0.339, indicating low prediction error. A lower RMSE suggests that the model's predicted PM2.5 values are very close to the actual values, while the low MAE indicates minimal average deviation in predictions. These results demonstrate that the Random Forest model provides accurate and reliable predictions of air quality levels. Overall, the model effectively captures the relationship

between pollutant parameters and PM2.5 concentration, making it suitable for air quality prediction.

Training Performance Analysis

Metrics	Training value	Testing value
RMSE	2.6671	3.6032
MAE	0.2873	0.3391

Table 2: Training and Testing Performance of Random Forest Model

Performance Metrics

The model performance is evaluated based on regression error metrics, which measure how close the predicted values are to the actual values. Lower error values indicate better prediction accuracy and reliability of the model. The results show that the model produces predictions that are very close to the actual values, with minimal deviation. The small difference between training and testing errors indicates that the model generalizes well and does not suffer from overfitting. Overall, this reflects strong predictive performance and reliable real-world applicability.

RMSE was used to measure the square root of the average squared differences between actual and predicted values:

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

MAE was used to measure the average absolute difference between predicted and actual values:

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

where $y_i$represents the actual values, $\hat{y}_i$represents the predicted values, and $n$is the total number of observations.

Model Comparison

Model	RSME(Test)	MAE(Test)
Random Forest	3.6032	0.3391
Decision Tree	3.9038	0.3417

Table 5: Comparison of Random Forest and Decision Tree Models The table compares the Decision Tree and Random Forest models based on regression performance metrics such as error. In contrast, the Random Forest model attains a lower ,

RMSE and MAE. The Decision Tree model achieves an RMSE value of 3.9038 and an MAE of 0.3417, indicating slightly higher prediction error. In contrast, the Random Forest model attains a lower RMSE of 3.6032 and a lower MAE of 0.3391, demonstrating better prediction accuracy. This improvement highlights the effectiveness of the Random Forest algorithm, as it combines multiple decision trees to reduce error and improve generalization. Overall, the Random Forest model outperforms the Decision Tree model and provides more reliable predictions for the given dataset.

V. CONCLUSION AND FUTURE WORK

This project presents a machine learning-based system for the prediction of PM2.5 levels using air quality parameters such as SO_2, NO_2, RSPM, and SPM, along with temporal and location-based features. The Decision Tree and Random Forest models are used to analyze complex relationships among environmental parameters and generate accurate predictions. The Decision Tree model provides a simple and interpretable structure, while the Random Forest model improves performance through ensemble learning and reduces overfitting. Experimental results show that the Random Forest model achieves better prediction accuracy with lower RMSE and MAE values compared to the Decision Tree model, making it more reliable for air quality prediction.

The system also includes data preprocessing, feature engineering, and visualization techniques to better understand pollution trends across different regions and time periods. The predicted PM2.5 values are further mapped to Air Quality Index (AQI) categories such as Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. This mapping enhances interpretability by converting numerical predictions into meaningful pollution levels and associated health impacts. The integration of the trained model with a web-based interface enables users to input pollutant values and obtain real-time predictions, making the system practical for environmental monitoring and decision support.

Although the proposed system demonstrates strong performance, there is scope for further improvement. Future work can focus on incorporating advanced models such as Long Short-Term Memory (LSTM) and Transformer-based architectures to capture temporal patterns in air quality data. The inclusion of additional environmental factors such as temperature, humidity, and wind speed can further improve prediction accuracy. Using large-scale, real-time datasets can enhance model generalization and reliability.

Additionally, integrating real-time data sources, cloud-based deployment, and geographic visualization can improve system scalability and usability. Enhancements in model interpretability using explainable AI techniques can provide deeper insights into the influence of different pollutants on PM2.5 levels. These improvements can make the system more accurate, scalable, and effective for real-world air quality monitoring and environmental decision-making

REFERENES

[1] J. Doe et al., “Machine Learning Approaches for Air Quality Prediction Using PM2.5 Data,” 2018.

[2] A. Kumar et al., “Prediction of Air Pollution Using Random Forest Algorithm,” 2019.

[3] S. Sharma et al., “Air Quality Forecasting Using Machine Learning Techniques,” 2020.

[4] R. Singh et al., “Analysis and Prediction of PM2.5 Concentration Using Decision Tree Model,” 2021.

[5] M. Gupta et al., “Air Pollution Prediction Using Regression and Machine Learning Algorithms,” 2017.

[6] T. Chen et al., “Urban Air Quality Prediction Using Ensemble Learning Methods,” 2018.

[7] P. Zhang et al., “Deep Learning-Based Air Quality Prediction Model,” 2020.

[8] Y. Li et al., “A Hybrid Machine Learning Approach for Air Quality Forecasting,” 2021.

[9] K. Patel et al., “Comparative Study of Machine Learning Models for PM2.5 Prediction,” 2022.

[10] S. Reddy et al., “Air Quality Prediction Using Data Mining Techniques,” 2019.

[11] L. Wang et al., “Time Series Analysis for Air Pollution Prediction Using Machine Learning,” 2017.

[12] H. Kim et al., “Prediction of Fine Particulate Matter Using Random Forest Model,” 2020.

[13] D. Verma et al., “Machine Learning-Based Prediction of Air Quality Index,” 2021.

[14] N. Rao et al., “Air Quality Prediction System Using Supervised Learning Algorithms,” 2022.

[15] B. Das et al., “Forecasting PM2.5 Levels Using Ensemble Machine Learning Techniques,” 2018.

[16] G. Mehta et al., “Air Pollution Analysis and Prediction Using Data Science Techniques,” 2023.

[17] X. Liu et al., “Spatio-Temporal Prediction of Air Quality Using Machine Learning Models,” 2019.

[18] V. Kumar et al., “Air Quality Monitoring and Prediction Using IoT and Machine Learning,” 2021.

[19] A. Singh et al., “Prediction of Air Pollution Using Hybrid Models,” 2024.

[20] R. Patel et al., “Advanced Machine Learning Techniques for Air Quality Prediction,” 2025.