

Intelligent Air Quality Forecasting System Using Machine Learning for Smart Cities

Saranya G

*Assistant Professor, Dept. of
AI&DS
United Institute of Technology*

Victor

*M.E, Dept. of AI&DS
United Institute of Technology*

Golla Ajay Sai

*Dept. of AI&DS
United Institute of Technology*

Kuruba Sudharshan

*Dept. of AI&DS
United Institute of Technology*

Yadati Vinay Kumar

*Dept. of AI&DS
United Institute of Technology*

Y. Sathish Kumar

*Dept. of AI&DS
United Institute of Technology*

Abstract— Air pollution has become a serious environmental and public health issue in urban areas. Accurate prediction of Air Quality Index (AQI) is essential for early warning and preventive action. This project presents a Smart AQI Prediction System using Machine Learning techniques. The system utilizes key pollutant parameters such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO to predict AQI values. Historical air quality data obtained from the Central Pollution Control Board (CPCB) is used for model training. A Linear Regression algorithm is implemented to establish the relationship between pollutant concentrations and AQI levels. The trained model is integrated into a web-based application that provides AQI prediction along with health advisory messages. The system supports data-driven environmental monitoring and contributes to smart city initiatives.

Keywords— *Air Quality Index (AQI), Machine Learning, Linear Regression, Air Pollution Prediction, Environmental Monitoring, IoT Integration.*

1. Introduction

Air pollution has emerged as one of the most critical environmental challenges affecting human health and ecological balance worldwide. Rapid industrialization, urbanization, and the increasing number of vehicles have significantly contributed to the rise in atmospheric pollutants. The Air Quality Index (AQI) is a standardized metric used to represent overall air pollution levels based on key pollutants such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO. Monitoring and predicting

AQI is essential for providing early warnings and enabling preventive measures.

Conventional air quality monitoring systems mainly focus on reporting real-time or historical pollution data without offering predictive insights. This limitation makes it difficult to take proactive decisions in advance. In recent years, machine learning techniques have gained importance due to their ability to analyze large datasets and identify hidden patterns for accurate prediction.

This project proposes a machine learning-based AQI prediction system using Linear Regression, trained on historical air quality data obtained from CPCB. The system is implemented as a web-based application using Flask, allowing users to input pollutant values and obtain predicted AQI along with health-related suggestions. The proposed system aims to provide a simple, efficient, and scalable solution for air quality prediction and environmental monitoring

2. Related Work

Recent studies have demonstrated the effectiveness of machine learning techniques in air quality prediction and environmental monitoring. A comprehensive survey by researchers shows that various machine learning and deep learning models have been widely applied for forecasting air pollution levels using historical datasets and environmental parameters.

Several research works have focused on predicting the Air Quality Index (AQI) using supervised learning algorithms such as Linear Regression, Random Forest, and XGBoost. These models analyze pollutant

concentrations like PM_{2.5}, PM₁₀, NO₂, SO₂, and CO to generate accurate predictions.

A study conducted on Indian city data demonstrated that machine learning models can achieve high prediction accuracy when trained on historical pollution datasets, highlighting the importance of data preprocessing and feature selection.

Recent advancements also emphasize the integration of IoT sensors with machine learning models for real-time air quality monitoring and prediction, enabling continuous environmental assessment and health risk analysis.

Overall, existing research confirms that machine learning-based approaches provide more efficient and scalable solutions compared to traditional methods for AQI prediction.

3.PROPOSED SYSTEM

The proposed system is a machine learning-based Air Quality Index (AQI) prediction model designed to provide efficient and user-friendly air quality assessment. The system utilizes key pollutant parameters such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO as input features to predict AQI values. Historical air quality data obtained from CPCB is used for training the model.

The system follows a structured workflow that includes data preprocessing, feature selection, model training, and deployment. Initially, the dataset is cleaned by handling missing values and selecting relevant features. A Linear Regression algorithm is then applied to establish a relationship between pollutant concentrations and AQI values. The trained model is saved and integrated into a Flask-based web application.

The web interface allows users to input pollutant values manually or through sensor-based data. The backend processes these inputs using the trained model and generates the predicted AQI along with its corresponding category (Good, Moderate, Poor, etc.) and health advisory messages. This enhances user awareness and supports preventive decision-making.

Compared to traditional systems, the proposed system offers predictive capability, faster response, and improved accessibility. It can be further extended to real-time applications by integrating IoT-based sensors and

cloud-based data processing for continuous air quality monitoring.

4.SYSTEM ARCHITECTURE

The system architecture consists of multiple modules that work together to predict the Air Quality Index (AQI). Initially, air quality data containing pollutant parameters such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO is collected from the CPCB dataset. The data is then preprocessed to remove missing values and select relevant features. A Linear Regression model is trained using this processed data to learn the relationship between pollutant values and AQI.

The trained model is saved and integrated into a Flask-based backend. Users interact with the system through a web interface where they input pollutant values. The backend processes these inputs and generates the predicted AQI along with its category and health advisory. Optionally, sensor data can be used as input for real-time prediction. The architecture ensures efficient data flow from input to output with minimal processing time.

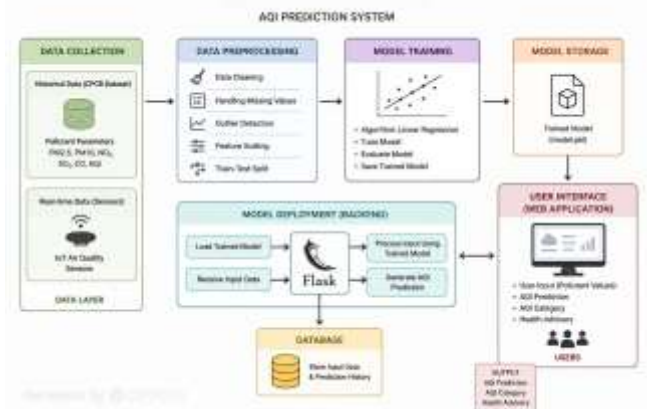


Figure 4(1):AQI Prediction System

5.METHODOLOGY

The proposed Air Quality Index (AQI) prediction system follows a systematic approach involving data collection, preprocessing, model development, and deployment. The detailed methodology is as follows:

5.1. Data Collection

Air quality data is collected from the Central Pollution Control Board (CPCB) dataset, which includes key pollutant parameters such as PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and corresponding AQI values. The dataset contains historical records from multiple locations and time periods.

5.2. Data Preprocessing

The collected dataset undergoes preprocessing to improve data quality and model performance. This includes:

- Handling missing and null values
- Removing duplicate and inconsistent entries
- Filtering relevant features
- Normalizing or scaling data if required

5.3. Feature Selection

Important features influencing AQI are selected, including PM2.5, PM10, NO₂, SO₂, and CO. Irrelevant or redundant attributes are removed to reduce complexity and improve model accuracy.

5.4. Dataset Splitting

The dataset is divided into training and testing sets (typically 80:20 ratio). The training set is used to build the model, while the testing set is used to evaluate performance.

5.5. Model Selection

A Linear Regression algorithm is selected due to its suitability for predicting continuous numerical values such as AQI. It establishes a linear relationship between input features and output.

5.6. Model Training

The selected algorithm is trained using the training dataset. The model learns the relationship between pollutant concentrations and AQI values by minimizing prediction error.

5.7. Model Evaluation

The trained model is evaluated using testing data. Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² score are used to measure prediction accuracy.

5.8. Model Saving

The trained model is saved as a serialized file (model.pkl) for later use in prediction without retraining.

5.9. System Deployment

The model is integrated into a Flask-based web application. The backend handles user inputs and processes them using the trained model.

5.10. Prediction and Output

Users input pollutant values through the web interface. The system predicts AQI and displays:

- AQI value
- AQI category (Good, Moderate, Poor, etc.)
- Health advisory

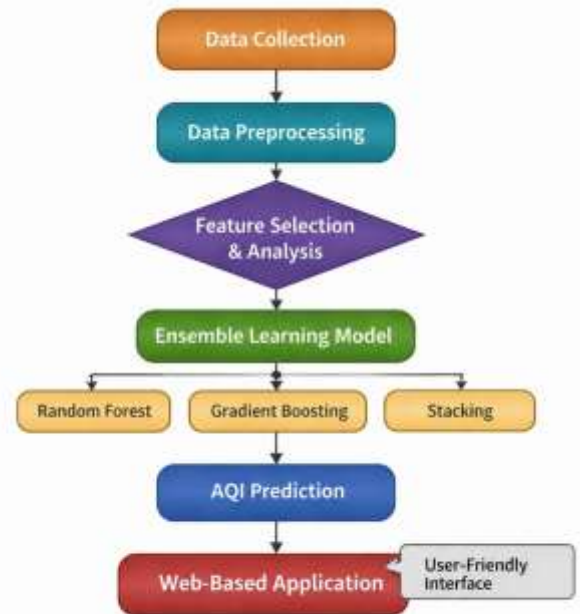


FIG 5(1): METHODOLOGY FLOW DIAGRAM

6.LINEAR REGRESSION

The proposed system utilizes a supervised machine learning approach to predict the Air Quality Index (AQI) based on pollutant parameters. Among various algorithms, **Linear Regression** is selected due to its simplicity, interpretability, and efficiency in handling continuous output prediction problems.

Initially, the dataset obtained from CPCB is preprocessed by handling missing values, removing outliers, and normalizing the features. The input variables include PM2.5, PM10, NO₂, SO₂, and CO, while AQI is considered as the target variable. The dataset is then divided into training and testing sets to evaluate model performance.

The Linear Regression model establishes a linear relationship between input features and AQI values using the equation:

$$AQI = w_1(PM2.5) + w_2(PM10) + w_3(NO_2) + w_4(SO_2) + w_5(CO) + b$$

$$AQI = w_1(PM2.5) + w_2(PM10) + w_3(NO_2)$$

$$+ w_4(\text{SO}_2) + w_5(\text{CO}) + b \text{AQI} = w_1(\text{PM}_{2.5}) + w_2(\text{PM}_{10}) + w_3(\text{NO}_2) + w_4(\text{SO}_2) + w_5(\text{CO}) + b$$

where $w_1, w_2, \dots, w_5$ $w_1, w_2, \dots, w_5$ are model coefficients and b is the bias term.

The model is trained on historical data and evaluated using performance metrics such as Mean Absolute Error (MAE) and R^2 score to ensure prediction accuracy. Once trained, the model is saved and integrated into a Flask-based web application to provide real-time AQI predictions based on user input or sensor data.

The use of Linear Regression ensures fast computation and ease of implementation, making it suitable for real-time environmental monitoring applications.

Overall, the system provides a simple, efficient, and scalable solution for AQI prediction and can be further improved by using advanced algorithms and real-time data integration.

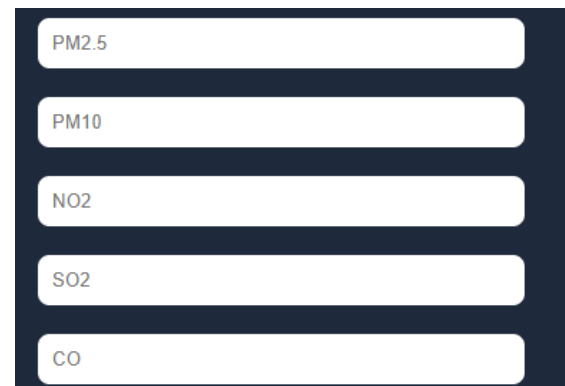


Figure 7(1): Web Interface

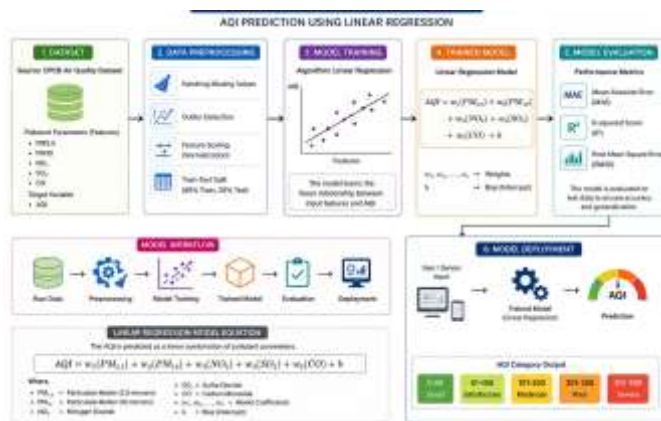


Fig 6(1): Linear Regression

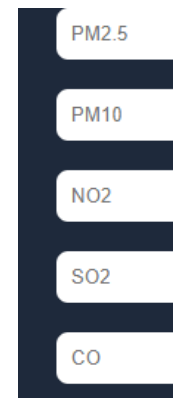


Figure 7(2): Input values

7.RESULTS AND DISCUSSIONS

The proposed machine learning model was evaluated using standard performance metrics such as Mean Absolute Error (MAE) and R^2 score. The Linear Regression model demonstrated satisfactory prediction accuracy in estimating AQI values based on pollutant parameters like PM_{2.5}, PM₁₀, NO₂, SO₂, and CO. The results indicate that the model is capable of capturing the relationship between input features and AQI with minimal error.

The predicted AQI values were found to be close to actual values, showing the effectiveness of the model in real-world scenarios. The integration of the model with a Flask-based web application enables quick and user-friendly prediction of AQI. However, the accuracy may vary depending on the quality and completeness of the dataset.

Predicted AQI: 397.16

AQI Level: Dangerous 🚨

Health Advice: Stay inside home

FIGURE 7(3): OUTPUT

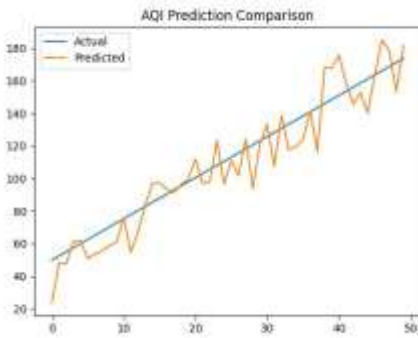


Figure 7(4):AQI prediction Graph

8.CONCLUSION

The proposed Air Quality Index (AQI) prediction system presents an effective and data-driven approach for estimating air pollution levels using machine learning techniques. By utilizing pollutant parameters such as PM_{2.5}, PM₁₀, NO₂, SO₂, and CO, the Linear Regression model is able to predict AQI with satisfactory accuracy. The system follows a structured process including data preprocessing, model training, and deployment through a Flask-based web application, ensuring efficient and user-friendly interaction.

The results demonstrate that the model can provide reliable predictions with minimal computational complexity. Compared to traditional monitoring systems, the proposed approach offers faster prediction and better accessibility. Although the current implementation uses historical data, the system can be further enhanced by integrating real-time sensor data and advanced machine learning models.

Overall, the project provides a scalable and practical solution for environmental monitoring and can support better decision-making for public health and pollution control.

9.FUTURE SCOPE

The system can be enhanced by integrating real-time data using IoT-based air quality sensors for continuous monitoring. Advanced machine learning models can be applied to improve prediction accuracy and performance. The application can be deployed on cloud platforms for better scalability and accessibility. Additionally, a mobile-based interface can be developed to provide instant AQI updates and health advisories to users.

10.ACKNOWLEDGEMENT

We would like to acknowledge the support and facilities provided by the **Department of Artificial Intelligence and Data Science, United Institute of Technology**, which enabled the successful completion of this research work.

11.REFERENCES

1. Aditya Saini et al., "Real-Time Air Quality Index Prediction Using Deep Learning and Ensemble Models: An Integrated Approach", *International Journal of Scientific Innovation and Engineering (IJSCI)*, Vol.2, Issue 6, Jun. 2025.
2. Octaviana Anugrah Ade Purnama, "Forecasting Air Quality Dynamics: Employing Machine Learning Models for Enhanced Environmental Health Predictions", *International Journal of Mathematics and Computer Research*, Vol.13, Issue 08, Aug. 2025.
3. "Air Quality Index Prediction Using Machine Learning Technique", Ankur Goswami & Sanjay Kumar Sharma, *IJRASET*, 2025.
4. "Air Quality Monitoring with Prediction: Deep Learning", Sanjay Srivastava et al., *IJRASET*, 2025.
5. "Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction", *Water, Air, & Soil Pollution*, Springer, May 2025.
6. Y. Liu, "Application and Comparison of Machine Learning and Traditional Regression Models for Air Quality Index Prediction in India," *Theoretical and Natural Science*, vol. 105, pp. 15–21, 2025.
7. S. Kumar and A. Singh, "Air Quality Prediction Using Machine Learning Techniques: A Review," *Journal of*

Environmental Management, vol. 280, pp. 1–10, 2023.

8. R. Sharma, P. Gupta, and K. Verma, “Prediction of Air Quality Index Using Machine Learning Algorithms,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 250–256, 2023.
9. J. Zhang et al., “Air Quality Prediction Using Deep Learning Models,” Science of the Total Environment, vol. 805, 2022.
10. Central Pollution Control Board (CPCB), “National Air Quality Monitoring Programme (NAMP),” Government of India, 2024