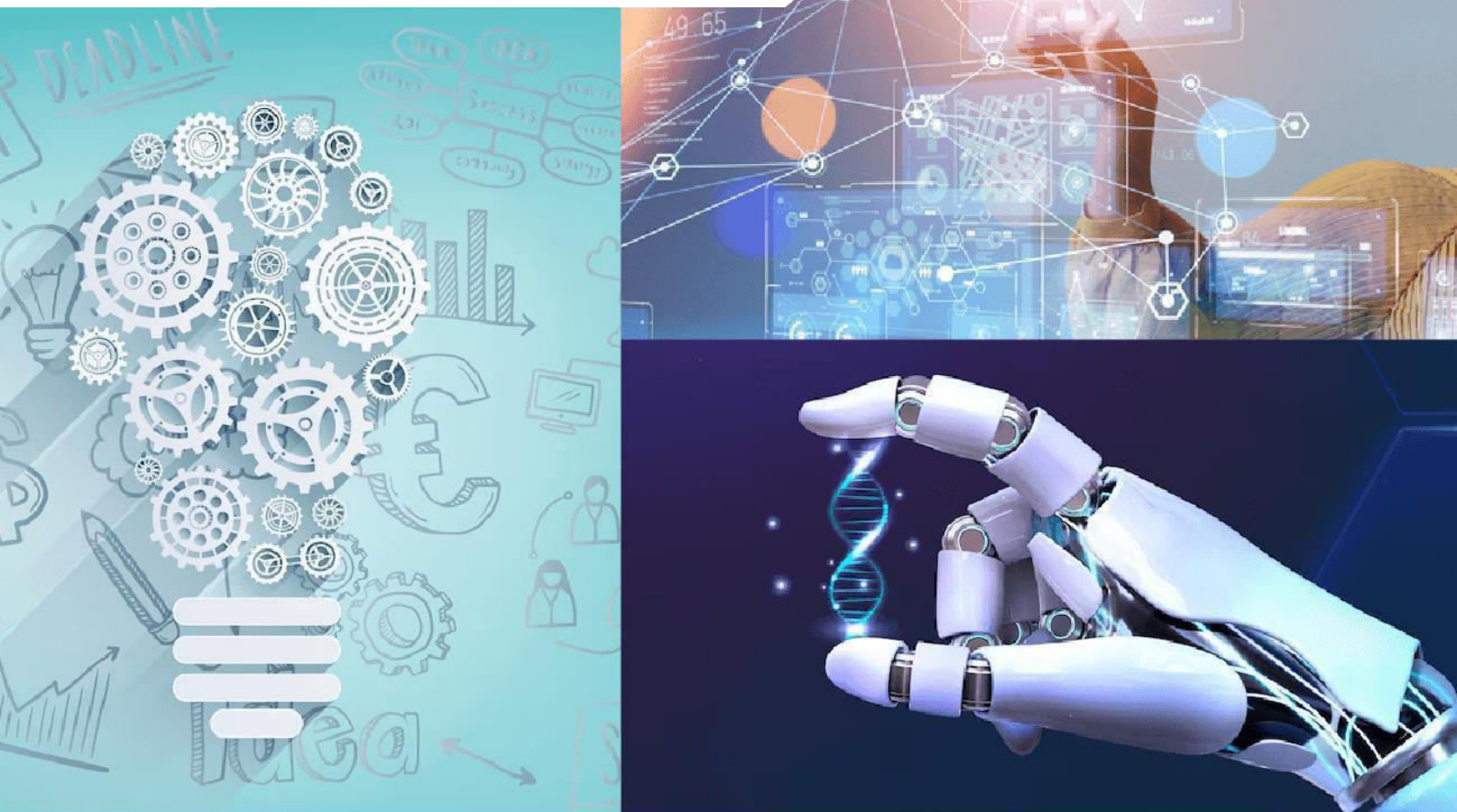




International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 13, Issue 2, March-April 2026

Impact Factor: 8.152



Air Quality Index Prediction and Analysis System Using Machine Learning

Mathur Akanksha

Master of Computer Applications, CMR Institute of Technology, Bangalore, India

ABSTRACT: Air pollution one of the major pressing environmental issues affecting modern society, particularly in urban and industrial regions. The continuous rise in population, increase in vehicular traffic, industrial emissions, and construction activities have led to a significant decline in air quality. Poor air quality has a direct impact on human health, causing respiratory illnesses, cardiovascular problems, and other long-term health complications. Although air quality data is made available by environmental agencies, it is often presented in technical ways that are difficult for common users to understand and interpret. This project presents an Air Quality Index (AQI) Prediction and Analytics System designed to monitor and analyse air quality conditions in real time. This project monitors air quality in real-time by pulling in live pollution data from reliable sources. The system tracks essential pollutants which include PM2.5 and PM10 and carbon monoxide and nitrogen dioxide and sulfur dioxide and ozone and ammonia to compute an Air Quality Index (AQI) score which shows current air quality. The system assesses air quality changes throughout the day by performing hourly data analysis which enables you to determine times of highest pollution and plan your outdoor activities. The system provides customized health recommendations to users who belong to high-risk categories which include children and senior citizens and people facing breathing issues. The data is displayed in easy-to-understand charts, graphs, and maps, so you don't need to be a scientist to understand what's going on. The goal is to make people more aware of air quality and help them make smarter choices. It's a good example of how data is used to tackle environmental problems

KEYWORDS: AQI, Machine Learning, XGBoost, Prediction, Environmental Monitoring

I. INTRODUCTION

Air pollution is a major environmental and community health concern across the world. Urban growth combined with industrial expansion and vehicle emissions from increased road traffic is leading to escalating air pollution problems. The environmental threat exists as a hidden danger that harms both nature and human health. Long-term exposure to pollution results in multiple health conditions which include breathing difficulties and heart disease and lung capacity reduction. The air quality in cities experiences continuous variation throughout the day because of traffic patterns and industrial operations and meteorological conditions. The problem is that most people don't realize how much it changes from hour to hour. Data is available through monitoring stations and websites, but the information becomes difficult for general audiences to understand because it contains complex technical specifications. The Air Quality Index (AQI) provides a solution to this problem. The system transforms complex information into one simple value which users can easily comprehend. Numerous systems display only numerical value. The system fails to provide you with essential trend analysis and practical guidance needed to interpret the data. The system requires development into a more intelligent and straightforward system which needs to monitor air quality and create valuable insights that assist people in making secure daily choices about their health and wellbeing. Air Quality Index (AQI) Prediction and Analytics System is developed with this objective in mind. This project focuses on designing a web-based application that presents real-time AQI information in an understandable format. By combining data acquisition, analysis, and visualization, the system aims to improve awareness about air quality and encourage preventive health measures among users.

II. LITERATURE REVIEW

Air quality monitoring and prediction have gained significant importance due to increasing pollution levels in urban areas. Earlier studies mainly focused on real-time AQI monitoring systems using API-based data sources [1]. These systems provided live updates but lacked predictive capabilities and deeper analysis. Traditional machine learning models such as Linear Regression and Decision Trees were initially used for AQI prediction [2]. These models were simple and easy to implement but could not handle complex relationships between environmental factors like temperature, humidity, and pollutant concentration. Later, advanced models such as Random Forest and Gradient Boosting were introduced to improve prediction accuracy [3]. Among these, XGBoost has shown superior performance due to its ability to handle structured data and reduce overfitting [4].

2.1 Role of Real-Time Data in AQI Systems

Real-time data plays a crucial role in AQI monitoring systems. APIs such as OpenWeatherMap provide pollutant and weather data including PM2.5, PM10, temperature, and humidity [1]. This data ensures that users receive updated information about air quality. However, real-time data alone is not sufficient for prediction and must be combined with historical datasets and machine learning models [2].

2.2 Analysis of Existing AQI Systems

Most existing AQI systems focus only on displaying real-time data without prediction features [1]. Some systems include basic visualization such as graphs and charts but lack personalized recommendations. Recent systems have started integrating machine learning models for prediction, but they often lack user-friendly interfaces and analytical features [3].

2.3 Machine Learning Models for AQI Prediction

Machine learning models are widely used for AQI prediction. Algorithms such as Decision Trees, Random Forest, and XGBoost have been applied in previous research [2], [3]. XGBoost provides higher accuracy due to its gradient boosting approach and ability to handle large datasets efficiently [4].

2.4 Integration of Prediction and Analysis Systems

Modern AQI systems integrate prediction with analytical features such as trends, comparison, and recommendations [3]. This integration improves usability and helps users make informed decisions based on air quality data.

III. METHODOLOGY

The proposed system is designed to provide a complete solution for monitoring, analysing, and predicting air quality using a combination of real-time data and machine learning techniques. The system follows a structured pipeline where each component performs a specific task, ensuring smooth data flow and accurate output generation. The overall architecture of the system consists of multiple modules, including data collection, preprocessing, model training, prediction, and visualization. These modules work together to transform raw environmental data into meaningful insights that can be easily understood by users.

3.1 Data Collection

The first step in the system is the collection of real-time environmental data. This is achieved using the OpenWeatherMap API, which provides reliable and up-to-date information about air pollution and weather conditions [1]. The system retrieves key pollutant parameters such as PM2.5, PM10, carbon monoxide (CO), nitrogen dioxide (NO₂), sulphur dioxide (SO₂), ozone (O₃), and ammonia (NH₃). In addition to pollutant data, weather-related parameters such as temperature, humidity, and wind speed are also collected. These parameters play a significant role in determining air quality, as weather conditions directly influence the dispersion and concentration of pollutants in the atmosphere. The data is fetched dynamically based on the user-selected city or current location.

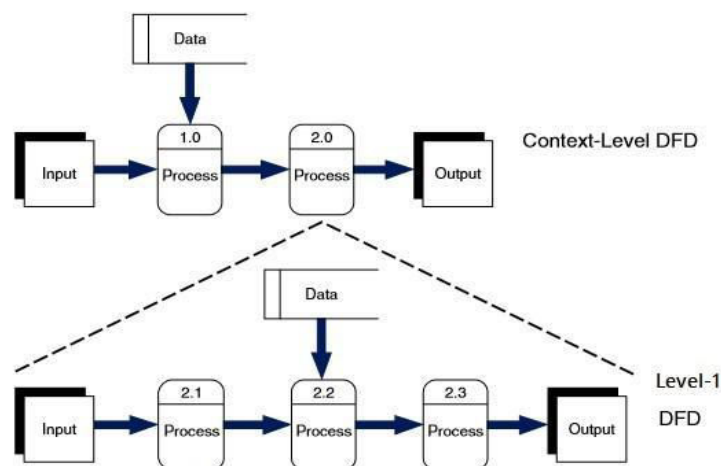


Figure 1: Data Flow Diagram of AQI System

3.2 Data Preprocessing

The raw data obtained from APIs is not always suitable for direct use in machine learning models. Therefore, preprocessing is required to ensure data consistency and quality. This step involves handling missing values, removing inconsistencies, and converting the data into a structured format. Feature selection is performed to identify the most relevant parameters required for AQI prediction. Normalization and scaling techniques are also applied to ensure that all features are within a similar range, which improves model performance and stability [2].

3.3 Feature Engineering

Feature engineering is an important step in improving the accuracy of the prediction model. In this process, the collected environmental parameters are transformed into meaningful input features. For example, pollutant concentrations are combined with weather parameters to understand their combined effect on air quality. This helps the model learn complex relationships between different variables. The final feature set includes PM2.5, PM10, temperature, humidity, and other relevant parameters, which are used as input for the machine learning model.

3.4 Model Training (XGBoost)

The prediction model is trained using historical AQI datasets obtained from publicly available sources such as Kaggle [5]. These datasets contain past records of pollutant levels and corresponding AQI values. The XGBoost algorithm is used for training due to its high efficiency and accuracy in handling structured data [4]. It works by combining multiple weak learners to form a strong predictive model. The dataset is divided into training and testing sets, typically in an 80:20 ratio. The model learns patterns from the training data and is evaluated using the testing data to ensure its performance.

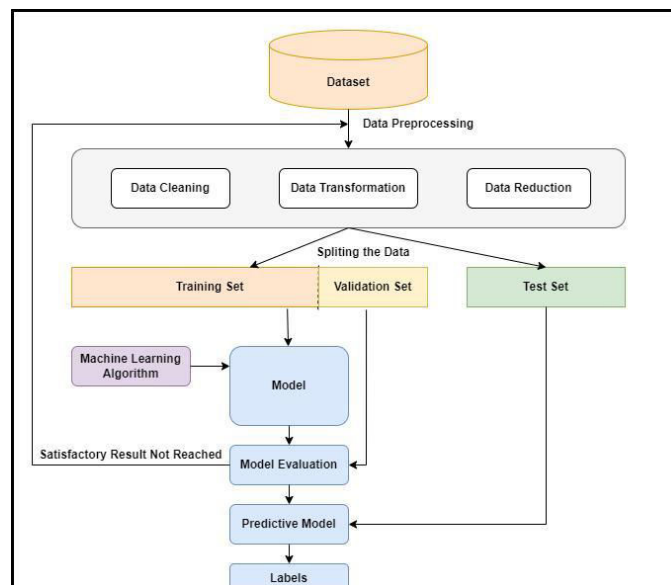


Figure 2: Machine Learning Model Workflow

3.5 Prediction Module

Once the model is trained, it is integrated into the backend system. During execution, real-time data collected from APIs is passed to the trained model. The model processes the input features and generates a predicted AQI value. This value is then sent to the frontend, where it is displayed to the user. The prediction module ensures that users not only see the current AQI but also get an estimated AQI based on environmental conditions.

3.6 System Architecture

The system follows a layered architecture consisting of frontend, backend, API integration, and machine learning components.

1. **User Layer** – Allows users to select a city or use location-based services.
2. **Frontend Layer** – Displays AQI data, predictions, and visualizations using an interactive interface.
3. **Backend Layer** – Handles API requests, data processing, and model integration.

4. **Processing Layer** – Performs preprocessing, feature engineering, and prediction.
5. **Data Layer** – Stores historical data and processed results.

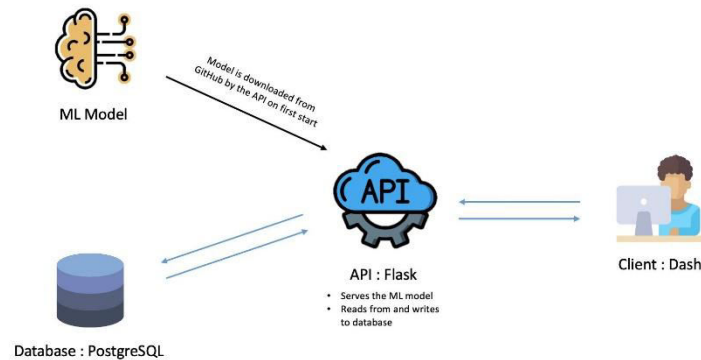


Figure 3: System Architecture of AQI Prediction System

3.7 System Workflow

The workflow of the system can be summarized as follows: User selects location → Data fetched from API → Data preprocessing → Model prediction → Results displayed. This workflow ensures that the system provides real-time as well as predictive insights in an efficient manner.

IV. RESULT AND OUTCOME

The proposed system was tested using multiple cities and environmental conditions to evaluate its performance and reliability.

4.1 Dataset Description

The dataset used for training consists of historical AQI values along with environmental parameters such as PM2.5, PM10, temperature, and humidity. A sufficient number of records were used to ensure that the model learns accurate patterns and relationships between variables.

TABLE I: Dataset Summary

Parameter	Description
Dataset Source	Kaggle AQI Dataset
Features Used	PM2.5, PM10, Temperature, Humidity
Data Type	Structured
Training Split	80%
Testing Split	20%

4.2 Experimental Setup

The system was implemented using Python for backend processing and machine learning. Libraries such as NumPy, Pandas, Scikit-learn, and XGBoost were used. The frontend was developed using modern web technologies to provide an interactive user interface.

The preprocessing steps included:- Handling missing values, Feature scaling, Data normalization

4.3 Performance Metrics

The performance of the model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Accuracy is calculated as: $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$. These metrics help in understanding how well the model predicts AQI values.

4.4 Results Analysis

The results indicate that the model performs well in predicting AQI values with good accuracy and consistency.

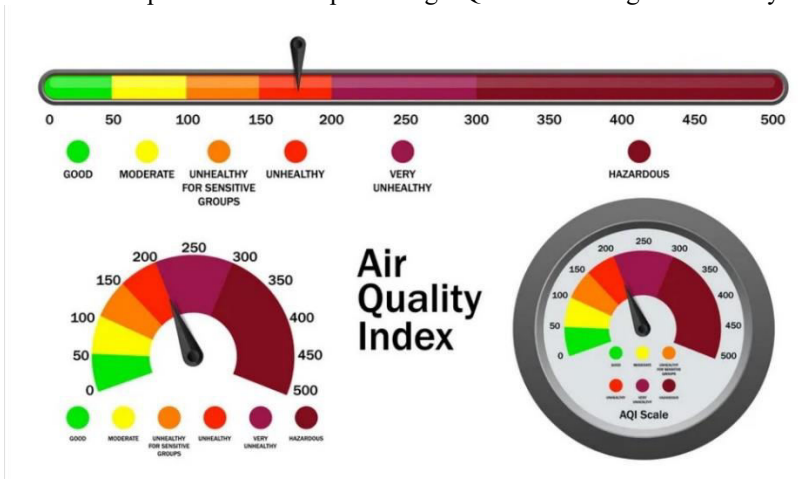


Figure 4: AQI Analysis and Visualization Dashboard

4.5 Discussion

The system successfully integrates real-time data with machine learning prediction. The predicted AQI values are close to actual values, indicating reliable performance. The analytical features such as trends, comparison, and best time recommendations enhance user understanding of air quality. However, the system performance depends on the quality of data and API reliability. Any inconsistency in real-time data may affect prediction accuracy.

V. CONCLUSION

The Air Quality Index (AQI) Prediction and Analysis System developed in this work provides an effective and user-friendly approach to understanding air quality conditions. The system successfully integrates real-time environmental data with machine learning techniques to deliver both current AQI information and predicted values. By combining data collection, preprocessing, and model prediction, the system is able to generate reliable insights based on environmental parameters. The use of the XGBoost algorithm improves prediction performance and ensures better accuracy compared to traditional methods. In addition to prediction, the system includes analytical features such as AQI trends, city comparison, best time suggestions, and health recommendations. These features make the system more practical and useful for everyday users. Overall, the proposed system not only simplifies complex air quality data but also helps users make informed decisions regarding their health and daily activities. It demonstrates how machine learning and data analytics can be effectively applied to solve real-world environmental problems.

REFERENCES

- [1] OpenWeatherMap, Air Pollution API Documentation, Available: <https://openweathermap.org/api>
- [2] Pedregosa F., Varoquaux G., Gramfort A., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011
- [3] Breiman L., "Random Forests," Machine Learning Journal, 2001
- [4] Chen T., Guestrin C., "XGBoost: A Scalable Tree Boosting System," Proceedings of the ACM SIGKDD Conference, 2016
- [5] Kaggle, Air Quality Index Dataset (India), Available: <https://www.kaggle.com>
- [6] World Health Organization (WHO), Air Pollution and Health, Available: <https://www.who.int>
- [7] United States Environmental Protection Agency (EPA), Air Quality Index (AQI) Basics, Available: <https://www.epa.gov/aqi>
- [8] Zhang Y., et al., "Air Quality Prediction Using Machine Learning Methods," Environmental Science and Pollution Research, 2019
- [9] Jiang D., et al., "A Review on Air Quality Prediction Using Machine Learning Algorithms," Atmospheric Environment, 2020

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152