IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Air Quality Monitoring and Health Risk Prediction System Using Machine Learning

[1]Tanmay Kamble, [2]Aqab Madre, [3]Ayan Munshi, [4]Abdullah Khan, [5]Prof. Rashmi More

Information Technology,
Finolex Academy of Management and Technology, Ratnagiri, India

Abstract: Air pollution has become a serious issue affecting human health and the environment across the world, and in this project, we developed an intelligent air quality monitoring and health risk prediction system by combining Internet of Things (IoT) and Machine Learning (ML) techniques. The system uses sensors such as PM2.5, MQ135, MQ7, and temperature sensors connected to an ESP32 microcontroller to collect environmental data in real time, and a dataset from Kaggle was also used to train the machine learning models. Two models, Linear Regression and XGBoost, were implemented and compared to evaluate their performance, and based on the results, XGBoost performed better with a higher R^2 score, indicating more accurate predictions. The system was implemented using Python, along with a web-based interface developed using HTML, CSS, and JavaScript to display real-time data and prediction results, while MySQL was used for efficient data storage and management. Overall, the system helps users monitor air quality and understand potential health risks, enabling them to take preventive actions.

Keywords

Air Quality Monitoring, Machine Learning, XGBoost, Linear Regression, Internet of Things (IoT), ESP32, Health Risk Prediction, PM2.5, Environmental Monitoring, Smart System

I. Introduction

Air pollution is one of the major environmental problems today and has a direct impact on human health, contributing to respiratory diseases, heart-related issues, and even premature deaths, while traditional air quality monitoring systems are often expensive, not easily accessible to common people, and do not always provide real-time updates, limiting their usefulness. With increasing urbanization and industrial activities, pollutants such as PM2.5, CO, and NOx have risen significantly, making continuous monitoring essential. Using modern technologies like IoT and machine learning, it is now possible to develop smart and cost-effective systems that can monitor and analyze environmental conditions in real time. In this project, we developed a system that uses IoT sensors to collect air quality data and machine learning models to analyze and predict AQI values along with possible health risks, where sensors such as PM2.5, MQ135, MQ7, and temperature sensors are connected to an ESP32 microcontroller for data collection. We implemented Linear Regression and XGBoost models to improve prediction accuracy, and a web-based interface is provided so that users can easily view real-time data and results. The main aim of this work is to create an efficient and user-friendly system that improves awareness and supports better decision-making regarding health and the environment.

II. Related Work

Air quality monitoring has been studied by many researchers because of its importance in maintaining public health and environmental safety. Earlier systems, such as those developed by Yi et al. [1], used IoT-based wireless sensor networks to collect environmental data in real time. These systems were helpful in reducing cost and improving accessibility, but they mainly focused on data collection and did not include prediction features.

Later, Kumar and Goyal [2] developed low-cost monitoring systems using MQ sensors. These systems made air quality monitoring more affordable but still lacked intelligent analysis. To improve prediction, machine learning techniques were introduced. Zhang et al. [3] used models like Linear Regression and Support Vector Machines, but these models were not very effective for complex environmental data. Singh and Verma [4] also highlighted that traditional models give lower accuracy for large datasets.

The idea of combining multiple models was introduced by Breiman [5], which improved prediction accuracy. Based on this, Chen and Guestrin [6] developed XGBoost, which is more efficient and accurate for handling complex data. However, many existing systems still do not use such advanced models in real-time applications. Reports from the World Health Organization [7] also show that air pollution is a major cause of serious health problems, which increases the need for better monitoring systems.

Even after all these developments, most systems focus either on monitoring or prediction, but not both together. Therefore, there is a need for a system that combines real-time monitoring, accurate prediction, and health risk analysis. Our proposed system aims to solve this problem by integrating IoT, machine learning, and a user-friendly interface.

III. System Architecture:

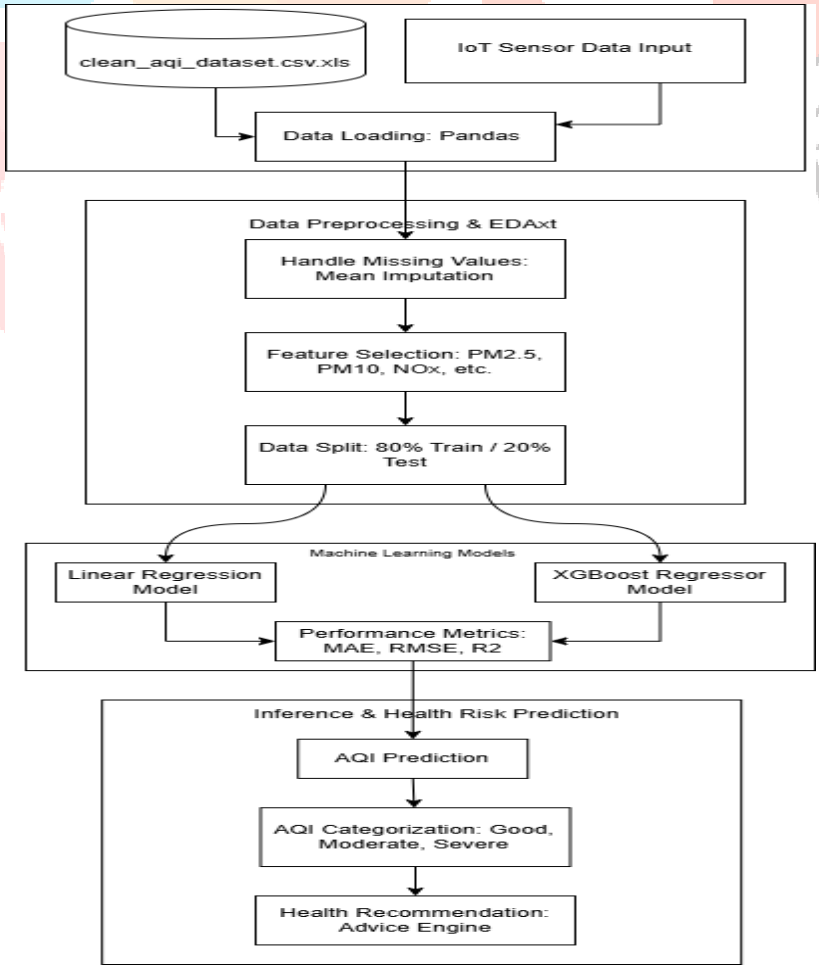

Fig. 3.1 Architecture Diagram

IV. Methodology

The proposed system aims to monitor air quality in real time and predict potential health risks using IoT and machine learning techniques. The methodology consists of multiple stages including data collection, preprocessing, model training, prediction, and visualization. The system ensures accurate air quality monitoring by integrating sensor data with advanced machine learning models.

4.1 Data Collection and Dataset Preparation

The data used in this system is obtained from two sources: real-time sensor data and a publicly available dataset from Kaggle. The hardware setup includes sensors such as PM2.5, MQ135, MQ7, and temperature sensors connected to an ESP32 microcontroller. These sensors continuously collect environmental parameters such as particulate matter, gas concentration, and temperature.

In addition to real-time data, a Kaggle air quality dataset was used to train the machine learning models. The dataset contains various air quality parameters and corresponding pollution levels. The data was divided into training and testing sets to evaluate model performance effectively.

4.2 Data Preprocessing

Before training the models, several preprocessing steps were performed to improve data quality and ensure consistency. Missing values in the dataset were handled using appropriate techniques such as mean substitution or removal of incomplete records.

Feature scaling and normalization were applied to bring all input parameters to a similar range, improving model efficiency. Relevant features such as PM2.5, gas concentrations, and temperature were selected for model training. Data cleaning ensured removal of noise and outliers to enhance prediction accuracy.

4.3 Machine Learning Models

Two machine learning algorithms were implemented in this study: Linear Regression and XGBoost. Linear Regression was used as a baseline model to establish a relationship between input features and air quality levels.

XGBoost, an advanced ensemble learning algorithm, was used to improve prediction accuracy by capturing complex patterns in the data. It uses gradient boosting techniques to combine multiple weak learners into a strong predictive model, making it suitable for handling structured environmental datasets.

4.4 Model Training and Evaluation

The models were trained using the preprocessed dataset. The dataset was split into training and testing sets to evaluate performance. Linear Regression and XGBoost models were trained using Python libraries such as Scikit-learn and XGBoost.

The performance of both models was evaluated using the R^2 (coefficient of determination) metric. The results showed that XGBoost achieved a higher R^2 score compared to Linear Regression, indicating better prediction accuracy. Therefore, XGBoost was selected as the final model for prediction.

4.5 Real-Time Data Processing and Prediction

For real-time operation, sensor data collected through the ESP32 is transmitted to the server and stored in a MySQL database. The trained machine learning model processes this incoming data to generate real-time predictions of air quality.

The system continuously analyzes environmental conditions and updates predictions dynamically, ensuring accurate and up-to-date information for users.

4.6 Health Risk Prediction

Based on the predicted air quality levels, the system estimates potential health risks associated with pollution exposure. Different air quality ranges are mapped to health risk categories such as low, moderate, and high risk.

For example:

- Good Air Quality: Minimal or no health risk
- Moderate Air Quality: May affect sensitive individuals
- Poor Air Quality: Harmful to general public

This feature helps users understand the impact of air pollution on their health and take preventive actions.

4.7 Visualization and User Interface

The system includes a web-based interface developed using HTML, CSS, and JavaScript. The interface displays real-time sensor data, predicted air quality levels, and health risk indicators in an interactive format.

Users can easily monitor environmental conditions through graphs, values, and alerts. The integration of frontend technologies ensures a user-friendly and responsive experience.

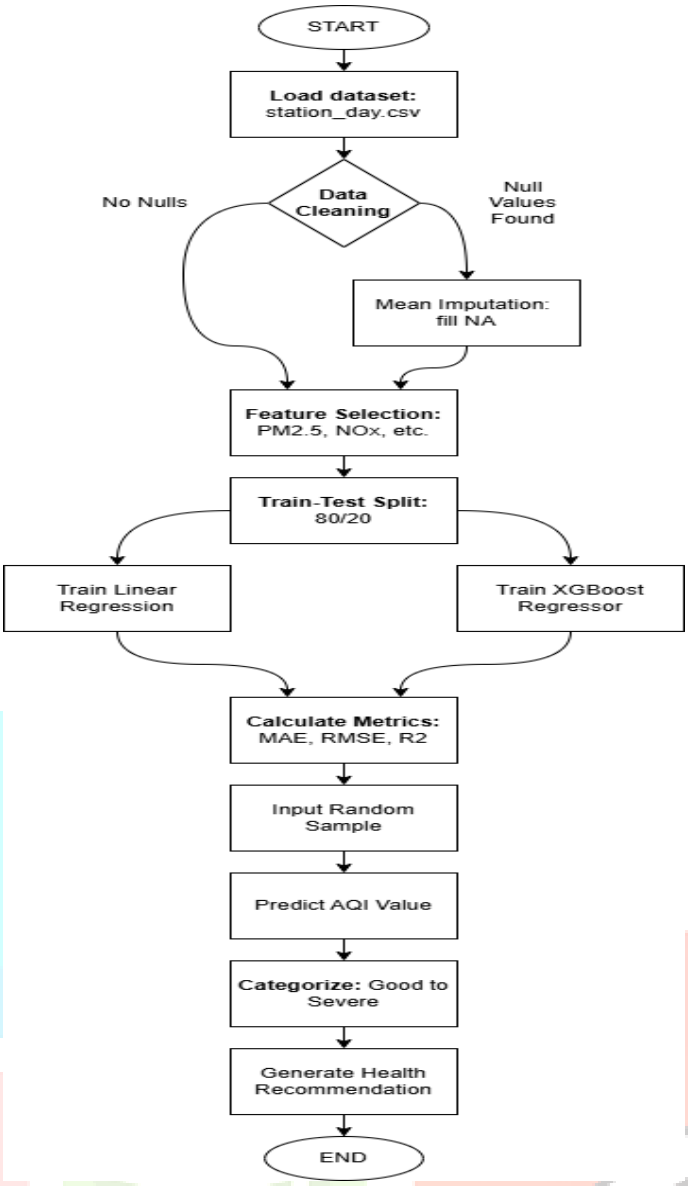

Fig. 4.1 Flowchart

The flowchart illustrates the overall working process of the proposed air quality monitoring and prediction system. The process begins with loading the dataset (station_day.csv), followed by data cleaning to handle missing values. If null values are found, mean imputation is applied; otherwise, the data proceeds directly to the next step. After cleaning, feature selection is performed by choosing important parameters such as PM2.5 and NOx. The dataset is then divided into training and testing sets using an 80:20 ratio. Two models, Linear Regression and XGBoost, are trained and evaluated using performance metrics such as MAE, RMSE, and R^2. Once the models are trained, a random input sample is provided to predict the AQI value. The predicted AQI is then categorized into levels (e.g., Good to Severe), and based on this, appropriate health recommendations are generated. Finally, the process ends after displaying the results.

V. Results

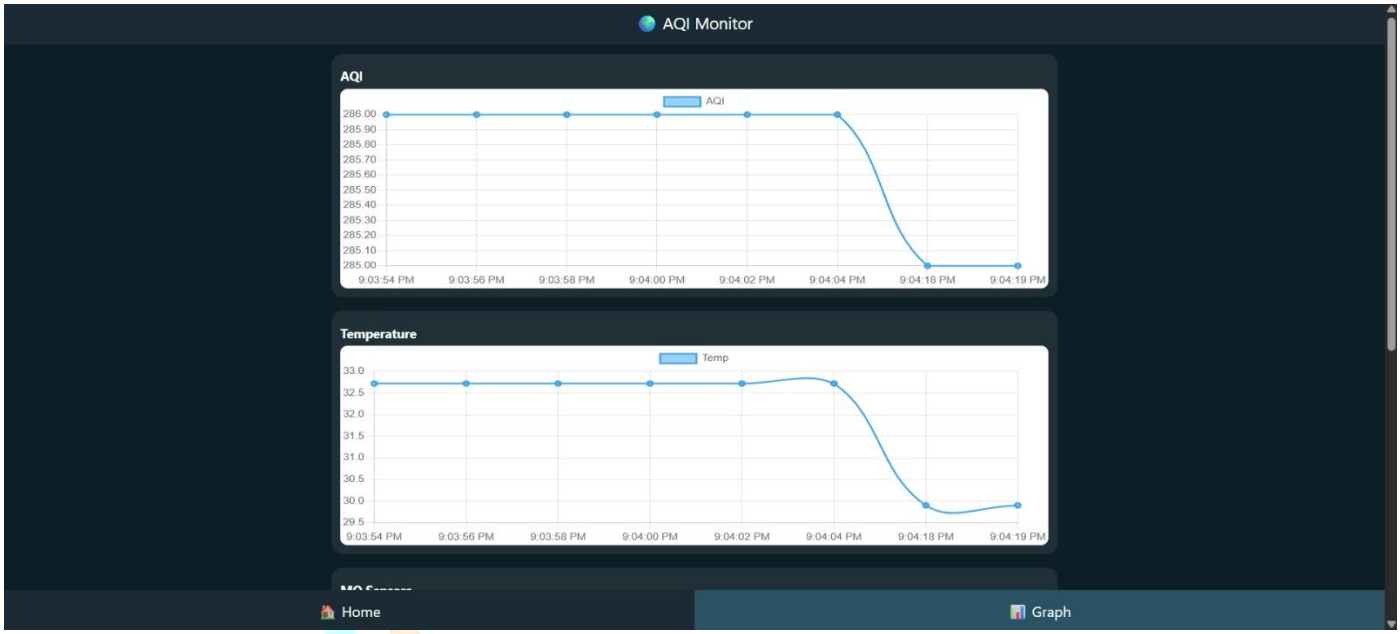

Fig 5.1 AQI Monitoring

Fig 5.2 Live Sensors Reading

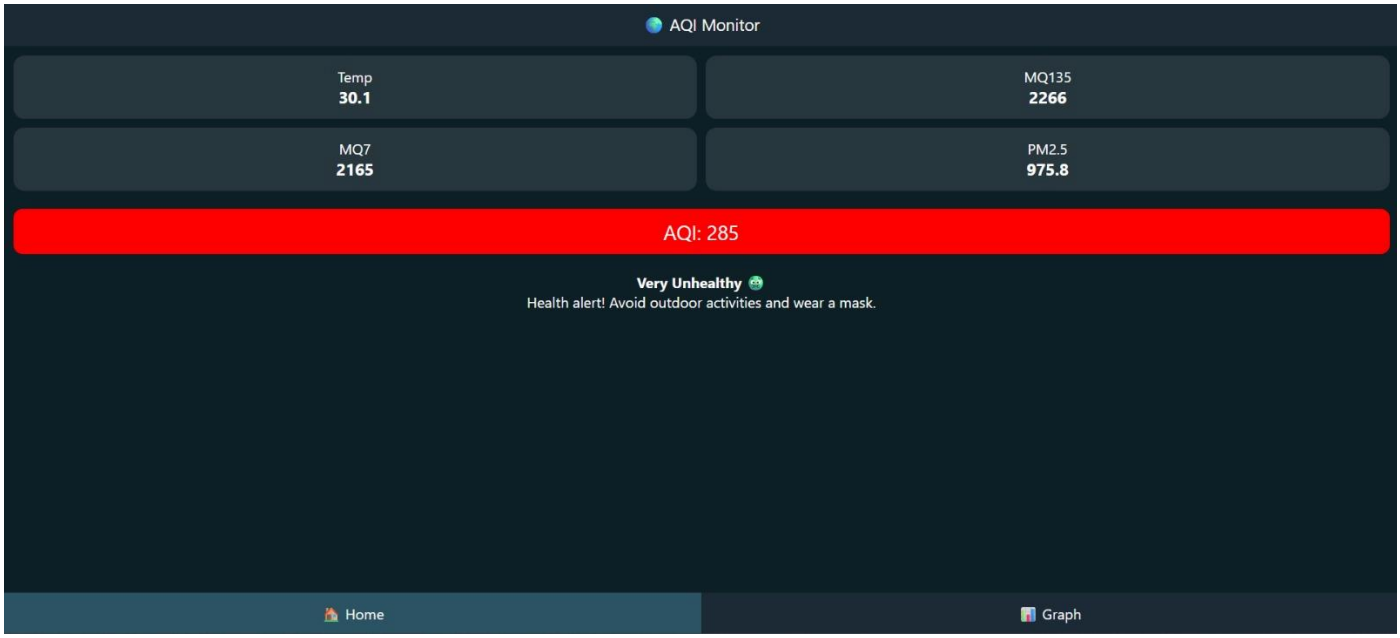

Fig. 5.3 AQI Value & Risk Prediction

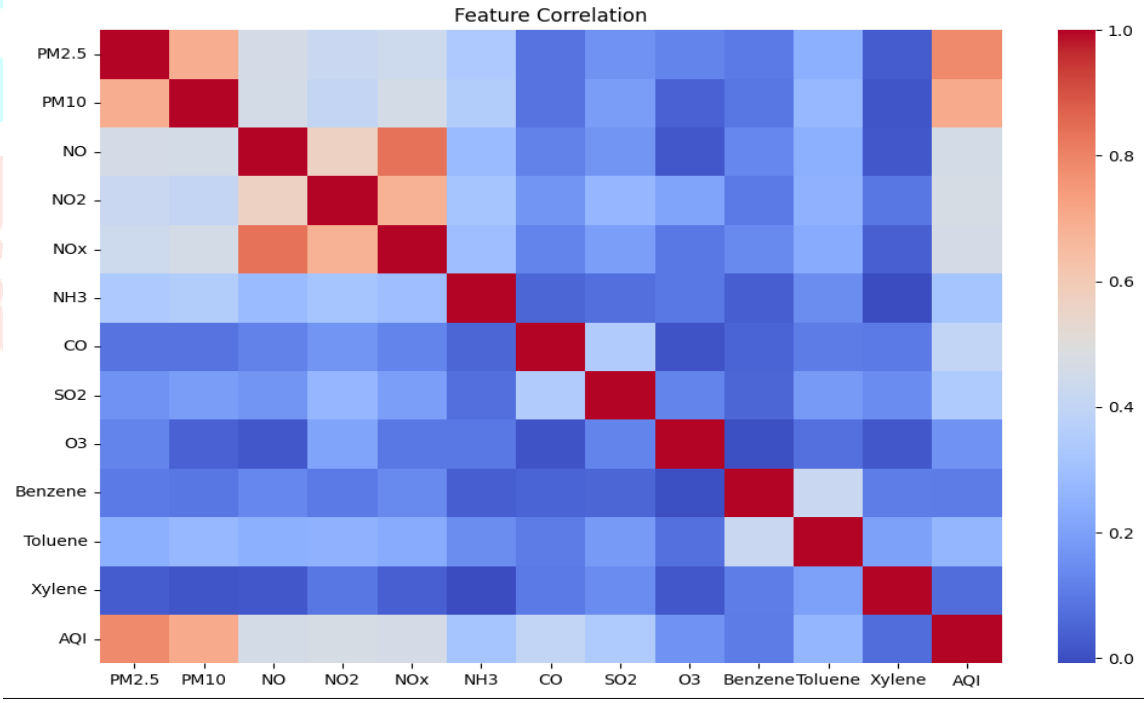

Fig 5.4 Correlation Matrix

The above figure represents the correlation matrix of different air quality parameters. It shows the relationship between various pollutants and AQI using correlation coefficients. From the matrix, PM2.5 and PM10 exhibit a strong positive correlation with AQI, indicating that they significantly influence air quality levels. Similarly, NO and NOx also show moderate correlation among themselves, suggesting interdependency between nitrogen-based pollutants.

On the other hand, gases like CO, SO2, and O3 show weaker correlations with AQI, indicating a comparatively lower direct impact. The heatmap helps in identifying the most important features, which improves model performance by focusing on highly correlated variables.

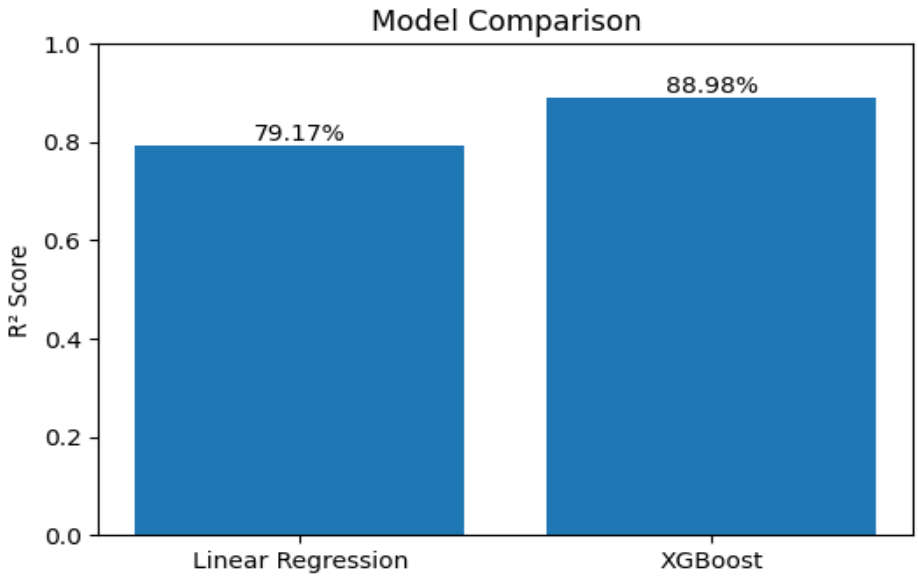

Fig. 5.5 Model Comparison

The above graph presents the comparison between Linear Regression and XGBoost models based on the R^2 score. It is observed that Linear Regression achieved an R^2 score of 79.17%, while XGBoost achieved a higher score of 88.98%.

This indicates that XGBoost performs better in predicting air quality as it can capture complex and non-linear relationships in the data more effectively than Linear Regression. The significant improvement in R^2 score demonstrates the superiority of XGBoost for this application.

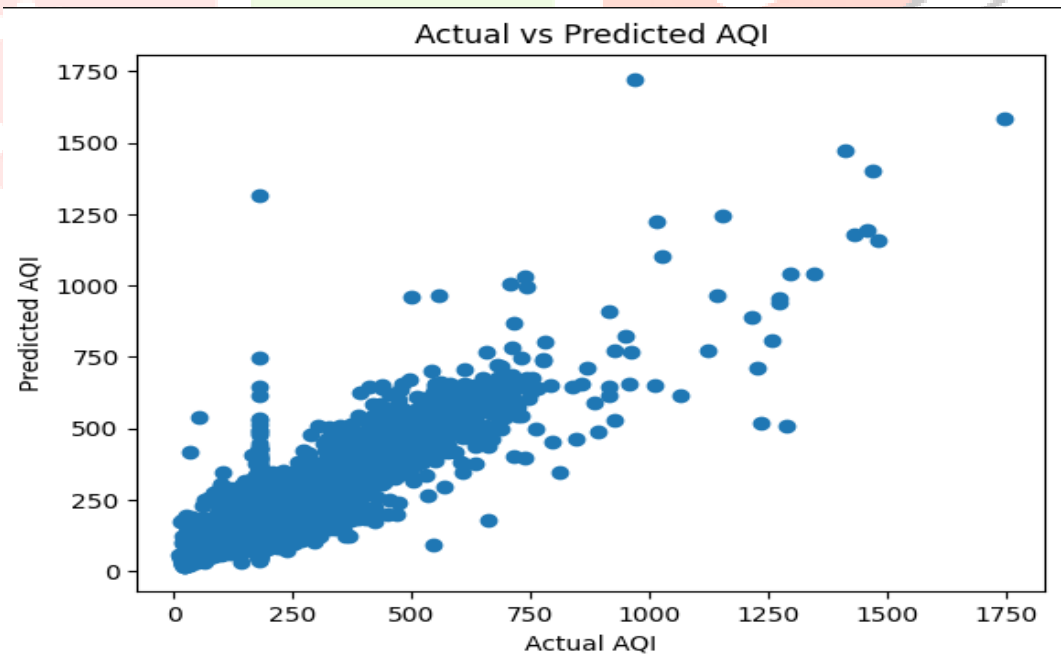

Fig. 5.6 Actual vs Predicted AQI Analysis

The scatter plot shows the relationship between actual and predicted AQI values. Most points lie close to the diagonal, indicating good prediction accuracy and that the model has learned the data patterns effectively. However, slight deviations at higher AQI values suggest minor prediction errors due to data complexity. Overall, the model performs well, supporting the high R^2 score, especially for XGBoost.

VIII. Conclusion

In this project, we developed a real-time air quality monitoring and health risk prediction system using IoT and machine learning techniques, where environmental data is collected using sensors connected to an ESP32 and processed using machine learning models to predict AQI values and associated health risks. We implemented and compared Linear Regression and XGBoost models, and based on the results, XGBoost performed better due to its ability to handle complex data and provide more accurate predictions. The system also includes a web-based interface that allows users to view real-time air quality data and understand possible health effects, making it easy to use and suitable for real-world applications. Overall, the developed system is cost-effective, scalable, and helps in improving awareness about air pollution while enabling users to take preventive actions, and it can be further applied in smart cities and environmental monitoring systems.

References

[1] Y. Yi, Z. Wang, and Y. Li, "Air Quality Monitoring System Based on Wireless Sensor Networks," *International Journal of Smart Home*, vol. 9, no. 6, pp. 143–150, 2015.

[2] A. Kumar and P. Goyal, "Low-Cost Air Quality Monitoring System Using MQ Sensors," *International Journal of Engineering Research & Technology (IJERT)*, vol. 5, no. 6, pp. 234–238, 2016.

[3] Y. Zhang, Q. Chen, and J. Wang, "Air Quality Prediction Using Machine Learning Techniques," *IEEE Access*, vol. 5, pp. 123–130, 2017.

[4] R. Singh and P. Verma, "Air Quality Prediction Using Linear Regression," *International Journal of Scientific Research in Computer Science*, vol. 9, no. 3, pp. 45–50, 2021.

[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] World Health Organization, "Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease," WHO Press, 2018.