iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: II Month of publication: February 2025

DOI: https://doi.org/10.22214/ijraset.2025.66983

Predicting Air Quality Using Machine Learning Models: A Comparative Study on Pollution Data from Urban Environments

Janvi H Makwana, Stuti R Kadam, Mrs. Keyaben Sanket Kumar Patel (Guide)
Dept. of Computer Science & Engineering, Parul Institute of Technology, VADODARA, Vadodara, Gujarat, India

Abstract: This research explores the prediction of air pollution levels through machine learning models, utilizing publicly available datasets containing parameters such as PM2.5, PM10, NO2, CO2, temperature, and humidity. The study implements Random Forest, XGBoost, and Linear Regression to compare their predictive performance. The findings highlight the efficiency of machine learning techniques in forecasting air quality and suggest potential applications in urban environmental monitoring.

Keywords: Air Quality, Machine Learning, Random Forest, XGBoost, Linear Regression

I. INTRODUCTION

Air pollution poses a serious threat to global health and ecosystems. Rising levels of pollutants, including PM2.5, PM10, NO2, and CO2, contribute to severe health conditions such as respiratory diseases and cardiovascular disorders. Predicting air quality accurately is essential for public safety and policy development. This research investigates how machine learning models, specifically Random Forest, XGBoost, and Linear Regression, can be employed to improve air quality forecasting. The goal is to evaluate their performance and determine the most effective model for practical implementation.

II. PROBLEM STATEMENT

Urbanization and industrialization have led to increased air pollution, presenting significant health challenges worldwide. Traditional air quality prediction methods struggle with the complexity of environmental data. This study aims to assess the ability of machine learning models to provide accurate predictions, leveraging diverse environmental and meteorological parameters.

III. OBJECTIVES

The primary objective is to implement and compare machine learning models for air quality prediction. Specifically, the study aims to:

Analyse the performance of Random Forest, XGBoost, and Linear Regression models.

Evaluate accuracy using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

Identify the best-performing model for real-world air quality forecasting.

IV. METHODOLOGY

A. Data Collection & Preprocessing

1) The dataset includes air pollution indicators (PM2.5, PM10, NO2, CO2) and meteorological factors (temperature, humidity).
2) Data preprocessing involves handling missing values, feature normalization, and dataset splitting for training and testing.

B. Machine Learning Models Used

1) Random Forest: An ensemble learning method that combines multiple decision trees to enhance predictive accuracy and reduce overfitting.
2) XGBoost: A gradient boosting algorithm optimized for structured data, known for its speed and performance.
3) Linear Regression: A straightforward statistical method for modeling relationships between independent and dependent variables.

V. RESULTS

A. Pollution Data Distribution

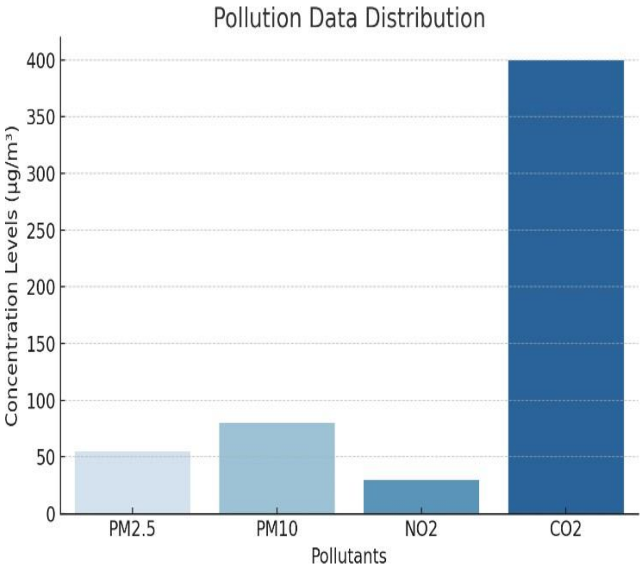
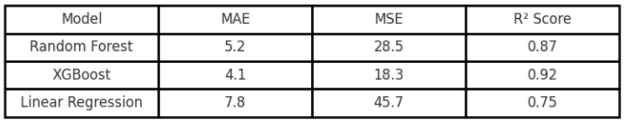


Figure 1: Bar chart showing the concentration levels of PM2.5, PM10, NO2, and CO2.

B. Model Performance Comparison

The findings reveal that XGBoost demonstrates superior predictive accuracy, achieving higher R² scores and lower error metrics compared to the other models. Random Forest also performs well but falls slightly behind XGBoost. Linear Regression, while useful for baseline comparisons, is less effective for complex air quality predictions. These results emphasize the importance of advanced ensemble models in improving forecasting accuracy.



Model	MAE	MSE	R² Score
Random Forest	5.2	28.5	0.87
XGBoost	4.1	18.3	0.92
Linear Regression	7.8	45.7	0.75

Figure 2: Table comparing Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² score across models.

C. Actual vs. Predicted Air Quality

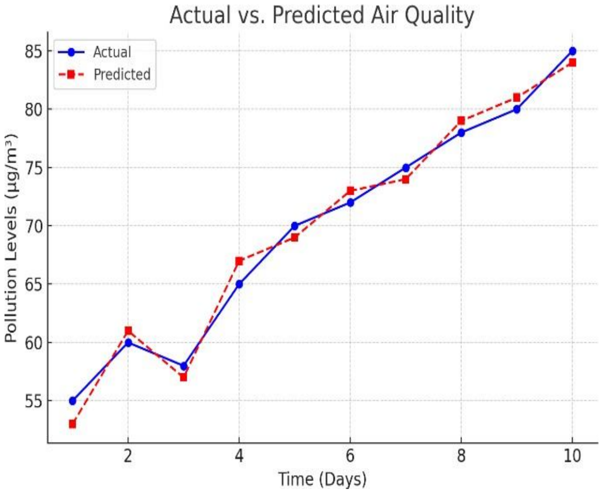


Figure 3: Line chart comparing actual vs. predicted pollution levels over time.

D. Feature Importance

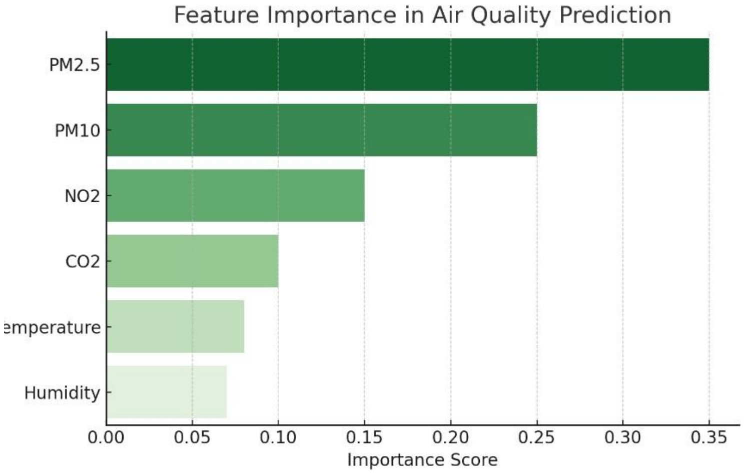


Figure 4: Bar graph showing the most influential features in predicting air quality.

VI. CONCLUSION

The study confirms that machine learning models are valuable tools for air quality prediction. Among the models analyzed, XGBoost emerges as the most effective for forecasting pollution levels, offering a balance between accuracy and computational efficiency. Future research will explore real-time data integration and deep learning techniques for further enhancements.

VII. ACKNOWLEDGMENTS

The author expresses gratitude to the faculty and department members for their support and guidance throughout this research. Special thanks to the dataset providers, whose contributions were instrumental in this study.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, 2001.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[3] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.

[4] K. He, et al., "Air Pollution and Health Effects," The Lancet, 2016.

[5] Scikit-learn Documentation. Available at: https://scikit-learn.org.

[6] XGBoost Documentation. Available at: https://xgboost.readthedocs.io.

AUTHORS

First Author
Name: Janvi H Makwana
Qualification: Integrated Bachelor’s in Technology [6 years] Institute: Parul University, Vadodara, Gujarat
Institute Email: 190345305039@paruluniversity.ac.in

Second Author
Name: Stuti R Kadam
Qualification: Integrated Bachelor’s in Technology [6 years] Institute: Parul University, Vadodara, Gujarat
Institute Email: 190345305039@paruluniversity.ac.in

Correspondence Author
Name: Janvi H Makwana
Email: janvimakwana819@gmail.com Contact Number: +91 9510440687