

Air Quality Prediction using Machine Learning: A Data-Driven Study of Ambient Pollutants in Delhi

Lakshmi K, Niveditha GC, Narendra KN

Department of CSE

Sri Sairam College of Engineering, Bengaluru, India

<https://orcid.org/0009-0006-7302-4435>

<https://orcid.org/0009-0005-7003-3769>

<https://orcid.org/0009-0004-2409-4336>



Publication History

Manuscript Reference No: IJIRIS/RS/Vol.11/Issue11/NVISX110096

Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue11/NVISX110096 Received: 28, October 2025, Revised: 05, November 2025, Accepted: 12, November 2025, Published Online: 21, November 2025.

<https://www.ijiris.com/volumes/Vol11/iss-11/17.NVISX110096.pdf>

Citation: Lakshmi, Niveditha, Narendra (2025), Air Quality Prediction using Machine Learning: A Data-Driven Study of Ambient Pollutants in Delhi, IJIRIS: International Journal of Innovative Research in Information Security, Volume 11, Issue 11 of 2025 pages 810-811 **Doi:** <https://doi.org/10.26562/ijiris.2025.v1111.17>

BibTeX Key: Lakshmi@Air

IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security, AM Publications, India 2025, ISSN 2349-7017, <https://doi.org/10.26562/ijiris.2025.v1111.17>) The journal's official abbreviation is IJIRIS.

Orcid: <https://orcid.org/0009-0004-9398-7488>

Copyright © 2025 copyright by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Air pollution is one of the major threats to the environment and health worldwide. This research analyzes an air quality dataset consisting of major pollutants like NO_2 , SO_2 , CO , O_3 , NH_3 , $\text{PM}_{2.5}$, PM_{10} in the capital city of India-Delhi. Using Python-based exploratory data analysis (EDA) and a comparison and evaluation of various machine learning models like Linear Regression, Random Forest, XGBoost, and LightGBM were performed on the data to understand pollutant behavior, inter-correlations, and their influence on Air Quality Index (AQI). The performance of each model was evaluated based on the results like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). It is observed that ensemble-based models gave better performance than linear models. LightGBM gave the best result with the lowest RMSE and the highest R^2 , followed by XGBoost with a very close result. Random Forest gave moderate results, whereas Linear Regression failed to give good results due to non-linear pollutant interactions. It is evident from the results that the ensemble and boosting algorithms work best for the prediction of AQI.

Keywords: Air Quality Index (AQI), Machine Learning, Exploratory Data Analysis (EDA), Linear Regression, Random Forest, XGBoost, LightGBM

I. INTRODUCTION

Air pollution is a significant issue affecting both developing and developed nations worldwide, causing environmental and public health challenges. Regular exposure to major pollutants, such as $\text{PM}_{2.5}$, PM_{10} , NO_2 , O_3 and SO_2 leads to cardiovascular, respiratory, and neurological disorders. Forecasting AQI helps concerned authorities take preventive measures, such as managing emission sources, implementing environmental policies, and warning the public. Air pollution in Delhi is increasing at an alarming rate, making it unfit for living. The public is affected by various respiratory problems and low visibility. Traditional AQI predictions struggle to tackle the non-linear interactions among pollutants as they depend on statistical or atmospheric dispersion models. This study compares four machine learning models to determine the one with the best performance for AQI prediction: Linear Regression, Random Forest, XGBoost and LightGBM.

II. RELATED WORK

Air quality prediction has been carried out for several years. Various machine learning models have been used over the years for air quality forecasting. Random Forest outperformed other traditional models for pollution estimation due to its ability to model non-linear relationships [8]. Recent studies involve gradient-boosting models like XGBoost and LightGBM, which performed better for AQI prediction [6], [7]. This study analyses the relative performance of these models on the dataset containing pollutant levels in Delhi.

III. DATASET DESCRIPTION

The dataset includes daily pollutant concentrations in the atmosphere of Delhi, such as:

- $\text{PM}_{2.5}$
- PM_{10}
- NO_2

- SO₂
- CO
- NH₃
- AQI (target variable)

Basic exploratory data analysis (EDA) concluded:

- No missing values in the dataset after preprocessing
- The strongest and dangerous contributors to AQI are PM_{2.5} and PM₁₀.
- Correlation analysis showed multi-collinearity among pollutants. As a result, non-linear ML models performed better on this dataset.

IV. METHODOLOGY

The study was conducted in Python using libraries Pandas and NumPy for data preprocessing, Seaborn and Matplotlib for visualization, and Scikit-learn for ML modeling. Data preprocessing included finding shape, missing values, data types, duplicates, and summary statistics of the dataset. Exploratory data analysis was performed on the dataset to arrive at conclusions about the dataset using distribution plots of pollutants, a correlation heatmap, and trend analysis using line charts. The data was split into training and testing as 80/20. It was then modeled using machine learning models such as Linear Regression, Random Forest Regressor, XGBoost Regressor, and LightGBM Regressor for air quality forecasting. Conclusions were made from the models regarding their performance based on the evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score.

V. RESULTS AND DISCUSSION

Ensemble models LightGBM and XGBoost gave the best accuracy after tuning with R^2 score 0.9. Among both, LightGBM outperformed XGBoost with a slight margin. Random Forest gave a moderate result with R^2 score 0.88. Linear regression failed to fit the model with R^2 score 0.60 due to non-linearity in the data. Particulate matter concentrations showed higher variance than gaseous pollutants, indicating severe pollution levels in Delhi. The results highlight the severe influence of particulate matter on air quality, making PM_{2.5} and PM₁₀ the dominant predictors. This aligns with public health studies, as there is a high chance of respiratory and cardiovascular diseases. The strong performance of the ensemble models demonstrates the accuracy of air quality predictions, which can be used to take preventive measures. The models achieved high accuracy due to fine-tuning, and since the dataset had no missing values.

VI. CONCLUSION

This research studies a comprehensive comparison of four machine learning models, such as Linear Regression, Random Forest, XGBoost, and LightGBM, for AQI prediction in Delhi. The results show that the ensemble models dominate other models due to their ability to capture multi-collinearity interactions among pollutants. Linear regression failed to fit the data due to non-linearity in the data.

VII. FUTURE SCOPE

This research can be further improved by integrating deep learning models such as LSTM networks for spatiotemporal AQI forecasting, integrating with real-time environmental sensors, developing an AQI prediction app, and including health impact analysis.

REFERENCES

1. World Health Organization (WHO). Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. World Health Organization, 2016.
2. Pope, C. A., & Dockery, D. W. "Health Effects of Fine Particulate Air Pollution: Lines that Connect." *Journal of the Air & Waste Management Association*, vol. 56, no. 6, 2006, pp. 709–742.
3. Central Pollution Control Board (CPCB). National Air Quality Index: Technical Report. Government of India, 2014.
4. Gulia, S., Shiva Nagendra, S. M., Khare, M., & Khanna, I. "Urban Air Quality Management—A Review." *Atmospheric Pollution Research*, vol. 6, no. 2, 2015, pp. 286–304.
5. Basu, S., & Srinivasan, S. "Air Quality Prediction Using Machine Learning Algorithms." *International Journal of Environmental Science and Technology*, 2021.
6. Vu, T. V., et al. "Deep Learning Models for Forecasting Air Pollution." *Atmospheric Environment*, vol. 218, 2019.
7. Patra, S., et al. "Prediction of Air Quality Index Using Random Forest Regression." *International Journal of Computer Applications*, 2018.
8. Breiman, L. "Random Forests." *Machine Learning*, vol. 45, 2001, pp. 5–32.
9. Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning*. Springer, 2009.
10. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
11. G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. <https://doi.org/10.48550/arXiv.1711.07363>