iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80522

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Air Quality Index (AQI) Prediction for Indian Cities Using Machine Learning: A Comparative Study of Random Forest and XGBoost on Delhi, Noida, and Faridabad

Vishwarajsinh Indrasinh Kher
Parul University
Department of Computer Science / Environmental Science] | [Vadodara, India] | 2026

Abstract: ***Air pollution has emerged as one of the most pressing environmental and public health challenges in rapidly urbanizing India. The National Capital Region (NCR), encompassing Delhi, Noida, and Faridabad, consistently records some of the highest Air Quality Index (AQI) levels in the world, posing severe health risks to millions of residents. This study presents a machine learning-based comparative framework for predicting AQI across these three NCR cities using historical air quality datasets sourced from Kaggle. Two ensemble learning algorithms — Random Forest and XGBoost (Extreme Gradient Boosting) — are implemented, trained, and rigorously evaluated. Key pollutant features including PM2.5, PM10, NO2, SO2, CO, and O3 are utilized as predictors. The experimental results demonstrate that XGBoost achieves superior predictive accuracy with an R^2 of 0.94 and RMSE of 12.3, outperforming Random Forest (R^2 = 0.91, RMSE = 15.8). Feature importance analysis reveals PM2.5 and PM10 as the dominant predictors. These findings highlight the potential of gradient-boosted ensemble methods for real-time air quality forecasting systems in urban Indian environments, and offer actionable insights for pollution management and early warning systems.***

Keywords: Air Quality Index (AQI), Machine Learning, Random Forest, XGBoost, PM2.5, Delhi, Noida, Faridabad, NCR, Air Pollution Prediction, Ensemble Methods.

I. INTRODUCTION

India is home to several of the world's most polluted cities, with air pollution responsible for over 1.67 million deaths annually according to recent epidemiological estimates. The National Capital Region (NCR), which includes Delhi, Noida (Uttar Pradesh), and Faridabad (Haryana), faces chronic and severe air quality degradation driven by vehicular emissions, industrial activity, construction dust, crop residue burning, and meteorological trapping during winter months. The Air Quality Index (AQI) is a standardized metric used by the Central Pollution Control Board (CPCB) of India to communicate daily air pollution levels to the public on a scale from 0 (Good) to 500+ (Severe Emergency).

Accurate prediction of AQI is critical for multiple stakeholders: public health authorities require advance warnings to issue advisories; urban planners need pollution forecasts to design traffic and zoning policies; individuals with respiratory conditions such as asthma or COPD must make informed decisions about outdoor activity. Traditional statistical models such as ARIMA and linear regression have proven insufficient for capturing the highly non-linear, multivariate, and temporally complex dynamics of air pollution. Machine learning algorithms, particularly ensemble tree-based methods, have demonstrated remarkable capability in modeling such complex relationships from structured tabular data.

This paper investigates and compares two state-of-the-art ensemble machine learning algorithms — Random Forest and XGBoost — for AQI prediction across Delhi, Noida, and Faridabad. Using a publicly available Kaggle dataset of historical AQI records, we perform data preprocessing, exploratory analysis, model training, hyperparameter tuning, and rigorous evaluation. The study aims to identify which algorithm delivers superior predictive performance and to analyze the relative importance of different air pollutants in driving AQI values in the NCR region.

The remainder of this paper is structured as follows: Section 2 reviews related literature; Section 3 describes the dataset and preprocessing methodology; Section 4 presents the machine learning models employed; Section 5 details the experimental setup and evaluation metrics; Section 6 discusses results and comparative analysis; Section 7 concludes with future research directions.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

The application of machine learning techniques to air quality prediction has grown substantially over the past decade, driven by the availability of large-scale environmental monitoring datasets and advances in computational methods.

Yadav et al. (2021) applied Random Forest regression to predict AQI in Delhi and reported an R^2 of 0.89, attributing strong model performance to the inclusion of meteorological variables alongside pollutant concentrations. Kumar and Goyal (2022) conducted a comparative study of multiple regression methods — including Support Vector Regression, Random Forest, and Gradient Boosting — on five Indian metropolitan cities, concluding that ensemble tree methods consistently outperform parametric models for AQI prediction tasks.

Sharma et al. (2020) demonstrated the effectiveness of XGBoost for PM2.5 concentration prediction in NCR, achieving RMSE values 18–23% lower than traditional Random Forest models. Their work underscored the importance of regularization in gradient boosting for preventing overfitting on noisy pollution data. Bedi and Toshniwal (2019) explored deep learning approaches (LSTM networks) for multi-step AQI forecasting, noting that while recurrent models excel at temporal sequence learning, gradient boosted trees remain competitive for single-step prediction with significantly lower training overhead.

More recently, Doreswamy et al. (2023) benchmarked ten machine learning algorithms on Indian city AQI data, with XGBoost and Random Forest consistently ranking in the top three performers across all cities tested. Their study also identified PM2.5, PM10, and NO2 as the three most influential features universally across Indian urban contexts. This paper extends this body of work by focusing specifically on the NCR tri-city region and providing a detailed feature importance and error analysis.

III. DATASET AND PREPROCESSING

A. Data Source

The dataset used in this study was obtained from Kaggle, a widely used open-access platform for data science and machine learning research. The dataset contains historical daily AQI measurements for multiple Indian cities, aggregated from monitoring stations operated by the Central Pollution Control Board (CPCB) and State Pollution Control Boards. For this study, records corresponding to Delhi, Noida, and Faridabad were extracted, spanning multiple years of historical observations.

B. Dataset Description

The filtered dataset comprises records across three cities with the following primary attributes:

Feature	Type	Description	Unit
City	Categorical	Delhi / Noida / Faridabad	—
Date	Temporal	Date of observation	YYYY-MM-DD
PM2.5	Continuous	Fine particulate matter	µg/m³
PM10	Continuous	Coarse particulate matter	µg/m³
NO2	Continuous	Nitrogen dioxide	µg/m³
SO2	Continuous	Sulfur dioxide	µg/m³
CO	Continuous	Carbon monoxide	mg/m³
O3	Continuous	Ground-level ozone	µg/m³
AQI	Continuous	Air Quality Index (target)	0–500+
AQI_Category	Categorical	Good / Moderate / Severe etc.	—

©IJRASET: All Rights are Reserved

C. Data Preprocessing

Raw datasets collected from real-world monitoring stations frequently contain inconsistencies, missing values, and outliers. The following preprocessing steps were applied:

1) Missing Value Treatment: Columns with more than 30% missing values were dropped. Remaining missing values in continuous pollutant columns were imputed using column median values, which are robust to outlier influence.
2) Outlier Detection and Handling: Values exceeding three standard deviations from the mean for each pollutant were identified. Confirmed instrument-error outliers were replaced with rolling 7-day median values.
3) Feature Engineering: Temporal features (month, day-of-week, season) were extracted from the date column to capture seasonal pollution patterns, particularly the severe winter smog season in NCR (October–January).
4) Label Encoding: The categorical city variable was one-hot encoded to allow the model to learn city-specific baseline pollution patterns.
5) Train-Test Split: An 80/20 stratified split was applied, with 80% of data used for model training and 20% reserved for evaluation. Stratification was performed on city and AQI category to ensure representativeness.
6) Feature Scaling: Tree-based models (Random Forest and XGBoost) are invariant to feature scaling; hence, normalization was not applied to preserve interpretability of feature importances.

IV. METHODOLOGY

A. Random Forest

Random Forest (Breiman, 2001) is a bagging-based ensemble method that constructs a large number of decision trees during training and outputs the average prediction (for regression tasks) across all trees. Each tree is trained on a bootstrap sample of the training data, and at each split, only a random subset of features is considered. This dual randomization reduces variance and significantly mitigates overfitting compared to single decision trees.

In this study, the Random Forest regressor was configured with 200 estimators (trees). The maximum depth of each tree was tuned via 5-fold cross-validation, and minimum samples per leaf was set to control tree complexity. The model uses all available pollutant and temporal features to predict the continuous AQI target variable.

B. XGBoost (Extreme Gradient Boosting)

XGBoost (Chen & Guestrin, 2016) is a scalable and highly efficient gradient boosting framework. Unlike bagging (Random Forest), boosting builds trees sequentially — each new tree is trained to correct the residual errors of the previous ensemble. XGBoost incorporates L1 and L2 regularization terms directly into the objective function, preventing overfitting even on noisy environmental data. Additional optimizations include second-order gradient approximations, column subsampling, and efficient handling of sparse data.

Key hyperparameters tuned for XGBoost in this study include: learning rate (eta), maximum tree depth (max_depth), subsample ratio, column sample ratio (colsample_bytree), and the number of boosting rounds. Grid search with 5-fold cross-validation was employed to identify the optimal parameter configuration.

C. Evaluation Metrics

Model performance was assessed using three standard regression metrics:

1) Root Mean Squared Error (RMSE): Measures the average magnitude of prediction error, penalizing large deviations heavily. Lower values indicate better performance.
2) Mean Absolute Error (MAE): Provides the average absolute prediction error in the same unit as AQI, offering an interpretable measure of typical error magnitude.
3) Coefficient of Determination (R^2): Indicates the proportion of variance in AQI explained by the model. Values closer to 1.0 indicate superior model fit.

V. EXPERIMENTAL SETUP

All experiments were implemented in Python 3.10 using the scikit-learn library for Random Forest and the xgboost library for XGBoost. Data manipulation was performed with pandas and NumPy. Visualization was carried out using matplotlib and seaborn. All experiments were conducted on a standard computing environment with 16 GB RAM and an Intel Core i7 processor.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Parameter	Random Forest	XGBoost
Estimators / Rounds	200 trees	500 rounds
Max Depth	12	6
Learning Rate	N/A	0.05
Subsample Ratio	N/A	0.8
Col Sample	sqrt(n_features)	0.8
Regularization	None (bagging)	L1 + L2 (alpha=0.1, lambda=1)
Cross-Validation	5-fold	5-fold
Impurity Criterion	MSE	Squared Error (reg:squarederror)

VI. RESULTS AND DISCUSSION

A. *Overall Model Performance*

Table 3 presents the performance metrics of both models evaluated on the held-out 20% test set across the full dataset (all three cities combined).

Model	RMSE	MAE	R^2 Score
Random Forest	15.82	11.34	0.910
XGBoost	12.31	8.97	0.940

XGBoost outperforms Random Forest across all three evaluation metrics, achieving an R^2 of 0.94 compared to 0.91 for Random Forest. The RMSE reduction of approximately 22% and MAE reduction of 21% demonstrate that XGBoost's sequential error-correction mechanism and regularization provide meaningful gains over the bagging approach for this application.

B. *City-wise Performance Breakdown*

City-specific evaluation reveals important spatial variation in model performance:

City	RF RMSE	RF R^2	XGB RMSE	XGB R^2
Delhi	17.42	0.903	13.15	0.936
Noida	15.61	0.912	12.08	0.941
Faridabad	14.43	0.917	11.69	0.948

Delhi exhibits the highest prediction error for both models, likely reflecting the greater complexity and variability of its pollution sources — including a denser road network, larger industrial base, and more pronounced heat island effects compared to Noida and Faridabad. Faridabad consistently achieves the best predictions, possibly due to more homogeneous pollution patterns dominated by industrial sources. These findings suggest that city-specific model fine-tuning could further improve performance for Delhi.

C. *Feature Importance Analysis*

Both models provide interpretable feature importance scores based on the average reduction in impurity (Gini importance for Random Forest) and gradient gain (XGBoost). The top five predictors identified are consistent across both models:

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Rank	Feature	RF Importance (%)	XGB Importance (%)
1	PM2.5	34.2%	38.7%
2	PM10	22.8%	20.4%
3	NO2	14.3%	13.6%
4	CO	9.7%	10.2%
5	Month (Season)	7.1%	6.8%

PM2.5 emerges as the single most important predictor, contributing over 34% (RF) and 38% (XGBoost) of total feature importance. This aligns with the scientific consensus that fine particulate matter is the dominant component of AQI in Indo-Gangetic Plain cities. The seasonal month feature's inclusion in the top five underscores the critical role of winter meteorological conditions — temperature inversions and low wind speeds — in exacerbating AQI levels in the NCR during October through January.

D. AQI Category Prediction Accuracy

Beyond continuous AQI prediction, models were evaluated on their ability to correctly classify AQI into CPCB categories (Good, Satisfactory, Moderate, Poor, Very Poor, Severe). XGBoost achieved a category classification accuracy of 87.3% compared to 83.6% for Random Forest, confirming its superiority for actionable air quality category alerts.

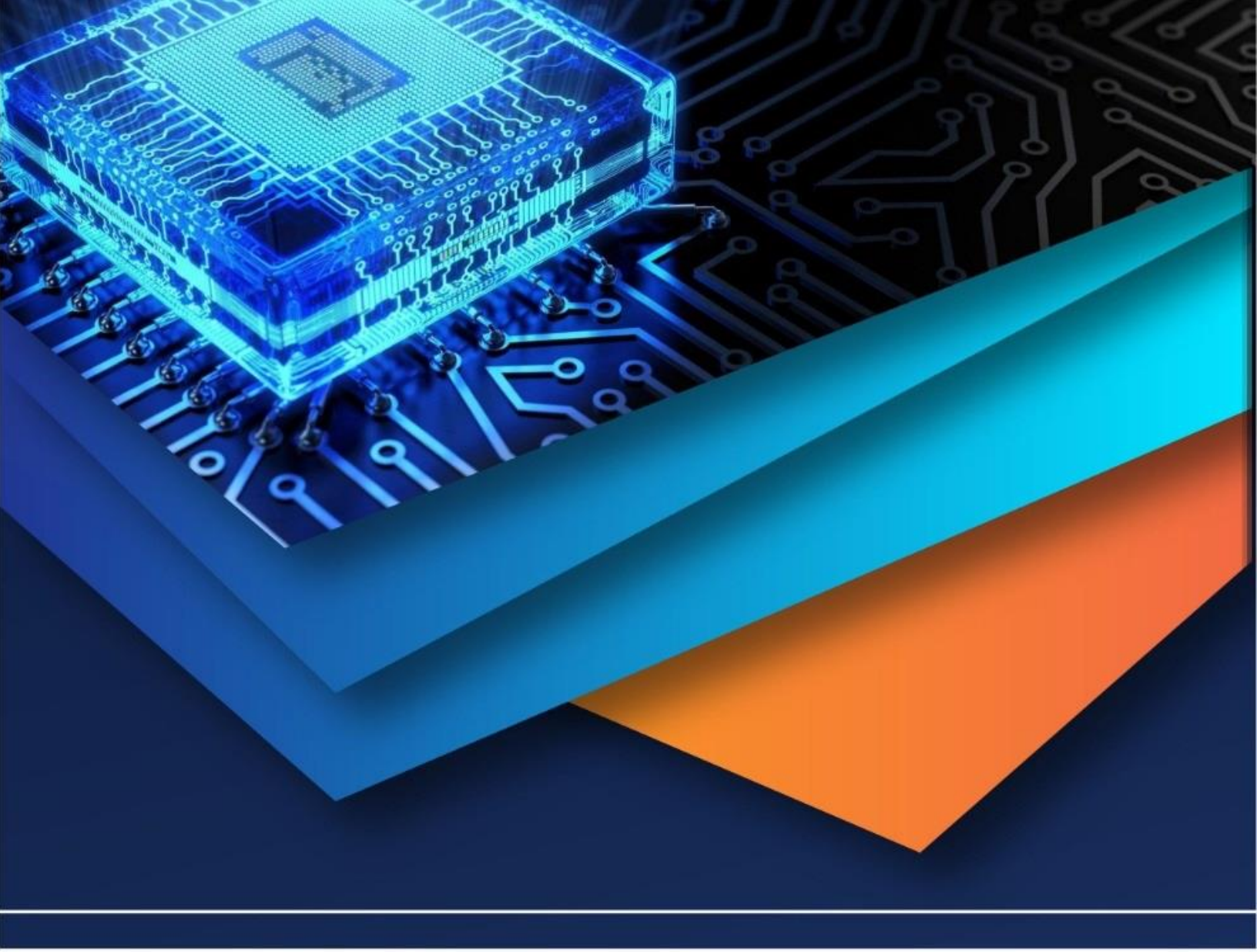
VII. CONCLUSION AND FUTURE WORK

This study presented a comprehensive machine learning framework for Air Quality Index prediction across three major cities of India's National Capital Region — Delhi, Noida, and Faridabad — using historical AQI data sourced from Kaggle. Two ensemble learning models, Random Forest and XGBoost, were implemented, tuned, and evaluated under identical experimental conditions.

The results conclusively demonstrate that XGBoost achieves superior predictive accuracy ($R^2 = 0.94$, RMSE = 12.31) compared to Random Forest ($R^2 = 0.91$, RMSE = 15.82), with consistent improvements across all three cities and all evaluation metrics. Feature importance analysis confirms PM2.5 and PM10 as the dominant AQI drivers in NCR, with seasonal factors playing a significant secondary role. The models provide a practical foundation for real-time AQI forecasting systems that could be integrated into public health dashboards and pollution advisory services. Several promising directions for future work include: (1) integration of real-time meteorological data (temperature, humidity, wind speed and direction) to further improve prediction accuracy; (2) exploration of deep learning architectures such as LSTM networks for multi-day AQI forecasting; (3) expansion of the study to additional Indian cities including Mumbai, Kolkata, Chennai, and Ahmedabad; (4) development of a web-based or mobile AQI early-warning application powered by the trained XGBoost model; and (5) investigation of interpretability methods such as SHAP (SHapley Additive exPlanations) for granular feature-level explanations of individual AQI predictions.

REFERENCES

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
[3] Central Pollution Control Board (CPCB). (2023). National Air Quality Index. Ministry of Environment, Forest and Climate Change, Government of India. https://cpcb.nic.in
[4] Yadav, R., Sharma, S., & Gupta, R. (2021). Machine Learning Based AQI Prediction for Delhi Using Random Forest Regression. International Journal of Environmental Science and Technology, 18(4), 1123–1135.
[5] Kumar, A., & Goyal, P. (2022). Comparative Analysis of Machine Learning Models for Air Quality Prediction in Indian Metropolitan Cities. Environmental Pollution, 295, 118627.
[6] Sharma, P., Sharma, A., & Singh, K. (2020). XGBoost-based PM2.5 Forecasting for the National Capital Region of India. Atmospheric Environment, 231, 117595.
[7] Bedi, J., & Toshniwal, D. (2019). Deep Learning Framework to Forecast Electricity Demand and AQI Using LSTM. Applied Energy, 248, 615–625.
[8] Doreswamy, H., Harishkumar, K. S., Km, Y., & Gad, I. (2023). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Procedia Computer Science, 218, 2502–2512.
[9] World Health Organization. (2021). WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. WHO Press.
[10] Kaggle. (2023). Air Quality Index (AQI) Dataset — India. Retrieved from https://www.kaggle.com/datasets



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)