EPJ Web of Conferences **379**, 05004 (2026) https://doi.org/10.1051/epjconf/202637905004
*ICS2MT-2026*

# Comparative Analysis of Machine Learning Models for Air Quality Index Prediction in Hyderabad and Delhi Cities

*Arun Kumar* Katkoori [1*], *Nayani* Sateesh[2], and *Sathya Prakash Racharla*[3], *V D S* Krishna[4]

[1]Department of ECE, CVR College of Engineering, Hyderabad, Telangana, India
[2,3,4]Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India

**Abstract.** According to the WHO, air pollution contributes to around 7 million premature deaths each year. This study aims to analyze the prediction of air quality using machine learning techniques. It utilizes data collected from India's Hyderabad and Delhi. The classification and prediction accuracy of various key pollutants, such as Particular Matrix (PM10), Nitrogen Dioxide (NO2,) and Sulfur Dioxide (SO2,) are evaluated. The XGBoost model performed well in Delhi, with a Coefficient of Determination ($R^2$) of 0.9998 and a 99.5% accuracy rate. In Hyderabad on the other hand the model found to be very satisfactory with $R^2$ of 0.786 and accuracy of 86%. The findings from the study suggest that the XGBoost model is able to predict the air quality in an urban area accurately. It has the potential to help improve the management of health risks and reduce pollution levels.

## 1 Introduction

Air pollution releases harmful gases and particulates like $SO_2$, NOx, CO and PM, affecting health, ecosystems, and agriculture [1-2]. The Air Quality Index (AQI) is a standard classification based on the level of air pollution to which you will be exposed to various categories of pollution levels, which are classified into various health-relevant levels (Good, Satisfactory, Moderate, Poor, Very Poor and Severe) [3].

Monitoring of AQI is important in India because of urbanization and industrialization and the risk of pollution in cities like Delhi and Hyderabad is very high and has severe health impacts [4-5].

The conventional forecasting methods have proven to be less precise, and consequently, ML methods have been proven to be promising alternatives. The present study compares the ML models used for the prediction of AQI for cities of Delhi and Hyderabad with particular emphasis on regression and classification problems.

* Corresponding author: dr.arunkatkoori@gmail.com

## 2 Literature Survey

The usefulness of ML in forecasting AQI has been illustrated in previous studies. Better predictive accuracy has been obtained by using feature selection and preprocessing techniques. Manikandan [6] introduced optimized GPR and SARIMA, and Liu [7] analyzed the prediction of AQI and pollutant concentrations using ML. Decision Trees and Random Forests demonstrated high accuracy in Indian context [4]. XGBoost is a highly accurate ensemble method for AQI forecasting. For Macau, [8] used the ANN and XGBoost models, and [9] proved XGBoost was effective in predicting $PM_{2.5}$ in Shanghai. Subrahmanyam [5] also conducted a comparative analysis of various ML models for AQI prediction in India, highlighting the effectiveness of ensemble models. Moreover, the robustness of ML for smart city air quality management [8] was also confirmed through comparative review and large-scale analysis [10-11].

## 3 Proposed System

The following stages were included in the proposed system process:

### 3.1 Dataset Description:

The data for air quality in the study is obtained from the Central Pollution Control Board (CPCB) and Telangana Pollution Control Board (TPCB) for the period 2017-2023 in the cities of Hyderabad and Delhi of India. The data provide monitoring information on nine major pollutants ($NO_2$, $O_3$, $PM_{10}$, $PM_{2.5}$, $SO_2$, CO, Benzene, $NH_3$, AQI) to monitor regional changes in pollution. There are 2,090 rows and 9 features in the Hyderabad data and 18,777 rows and 8 features in the Delhi data. The pollutant concentration measurements give a detailed analysis of the air quality developments in the two large cities.

### 3.2 Data Pre-Processing:

To control noise, inconsistencies and missing values, data pretreatment is required for data analysis. This also involves replacing missing pollutant parameter values with mean values to retain important information. Exploratory Data Analysis (EDA) is used to provide information about the characteristics of the data set using summary descriptive statistics such as the number of data points, mean, standard deviation, and Interquartile Ranges, Q1 (25th percentile), Q2 (50th percentile), Q3 (75th percentile). It offers data distribution & relationship information to uncover anomalies & to create training sets of data.

#### 3.2.1 Calculation of AQI and Sub-Index of Each Air Pollutant:

The Air Quality Index (AQI) is a common index ranging from 0 to 500 to determine the impact of air pollution on health and the environment. It is based on important pollutants like $SO_2$, $PM_{10}$, $PM_{2.5}$, CO, and $NO_2$, which are then turned into sub-indices that range from "good" to "hazardous." The overall AQI is determined by the highest sub-index and the weights of pollutants are derived from health effects to provide a comprehensive picture of air quality in the cities.

$$I_P=\frac{I_{HI}\text{-}I_{LO}}{BP_{HI}\text{-}BP_{LO}}(C_P\text{-}BP_{LO})+I_{LO} \tag{1}$$

For pollutant p, I serves as the index; Cp represents the pollutant p's truncated concentration; BPH is the concentration breakpoint at or above Cp; BPL is the concentration breakpoint at or below C; IHI corresponds to the AQI value for BPHI;; ILO corresponds to the AQI value for BPLO; The proposed air quality index prediction block diagram is shown in figure 1.

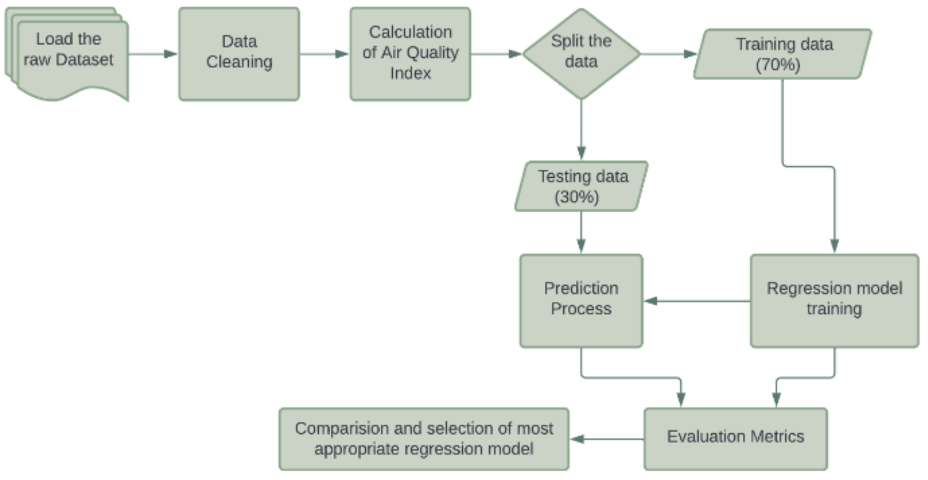



**Fig.1.** AQI Prediction Block Diagram

$$AQI=max(AQI_{PM2.5}, AQI_{PM10}, AQI_{SO_2}, AQI_{NO_X}, AQI_{NH_3}AQI_{O_3}, AQI_{CO} \tag{2}$$

### 3.3 Feature Selection:

After preprocessing, correlation analysis was conducted to ascertain the pollutant characteristics most pertinent to AQI. The correlation matrix showed that all pollutants were positively correlated with AQI. This means that higher levels of pollutants make the air quality worse. Since all the coefficients were greater than zero, no feature was thrown out. This shows that AQI depends on all the pollutants that were looked at.
Feature importance was analyzed using the correlation matrix to identify the most influential variables for AQI prediction. Key observations:

- Delhi: PM2.5 and PM10 showed the highest correlation with AQI, indicating their dominant role in pollution levels.
- Hyderabad: $NO_2$ and $SO_2$ were more significant because of the local sources such as industrial emissions.

Split Data: The data set was partitioned into a training set (70%) and a testing set (30%) to balance the learning and testing. This separation allowed the model train to take in more of the data, and leave some of the data aside for testing. This strategy ensured adequate training and testing of the model for its generalization capability.

### 3.4 Model Building

In this step, classification and regression methods like Linear Regression, Random Forest, Logistic Regression, XGBoost, Naïve Bayes, and KNN are used to make models that could predict the future. The metrics adopted for regression were Mean Absolute Error (MAE), $R^2$ score and Root Mean Squared Error (RMSE) [12] while the metrics for classification were accuracy, log loss, Cohen's Kappa, and Matthew's correlation coefficient. These measures enabled a detailed comparison to determine the most suitable model to predict AQI.

## 4 RESULTS

### 4.1 Regression Models

Table 1 shows the regression result of the model for predicting AQI for the city of Hyderabad with Linear Regression, Decision Tree, Random Forest and XGBoost. Random Forest ($R^2$ = 0.9687 train, 0.7749 test) and XGBoost ($R^2$ = 0.9972 train, 0.7859 test) performed well in terms of predicting the values of the target variable.Random Forest ($R^2$ = 0.9687 train, 0.7749 test) and XGBoost ($R^2$ = 0.9972 train, 0.7859 test) produced the highest accuracy along with the lowest MAE and RMSE, indicating good prediction ability and generalization. Linear Regression ($R^2$ = 0.7352), Decision Tree ($R^2$ = 0.6024) had lower performance with higher error rates. Add to this, the two best models for accurate AQI prediction were Random Forest and XGBoost.

**Table 1**. Comparison of Performance Metrics of Training and Testing for Hyderabad City

| | **Linear Regression** | **Decision Tree** | **Random Forest** | **XGBoost** |
|---|---|---|---|---|
| **MAE (Train)** | 16.143597 | 0.019757 | 4.563567 | 1.805243 |
| **MAE (Test)** | 15.681730 | 17.499443 | 12.733510 | 13.481968 |
| **RMSE (Train)** | 26.810731 | 0.388004 | 8.933714 | 2.659204 |
| **RMSE (Test)** | 27.159831 | 33.279986 | 25.035417 | 24.415521 |
| **$R^2$ (Train)** | 0.718373 | 0.999941 | 0.968730 | 0.997229 |
| **$R^2$ (Test)** | 0.735171 | 0.602371 | 0.774980 | 0.785985 |

The results in Table 2 show that the Random Forest model performed best in terms of MAE and RMSE values for Delhi, demonstrating its high accuracy and robustness. Decision Tree was also good with a testing $R^2$ of 0.9998, and MAE of 0.5971, trailing only Random Forest. Linear Regression was also less generalizable while XGBoost was effective, but had a higher error rate than the two best models.

Overall, the best models for AQI prediction in Delhi were random forest and decision tree.

**Table 2.** Comparison of Performance Metrics of Training and Testing for Delhi City

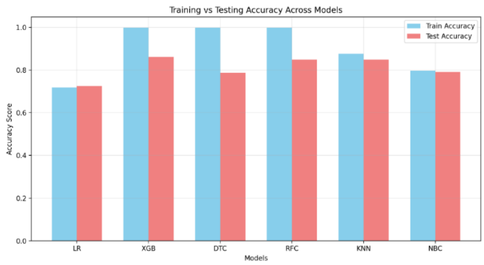
| | Linear Regression | Decision Tree | Random Forest | XGBoost |
|---|---|---|---|---|
| **MAE (Train)** | 2.789294 | 1.374725 | 1.634999 | 1.622991 |
| **MAE (Test)** | 1.374725 | 0.597109 | 0.475948 | 2.984849 |
| **RMSE(Train)** | 3.583013 | 9.288621 | 1.092457 | 3.757832 |
| **RMSE (Test)** | 36.387017 | 3.215786 | 3.113929 | 10.728646 |
| **$R^2$ (Train)** | 0.992788 | 1.000000 | 0.9999984 | 0.999811 |
| **$R^2$ (Test)** | 0.982152 | 0.999861 | 0.999869 | 0.998448 |

### 4.2 Classification Models

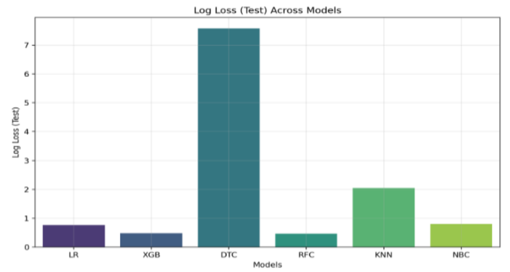
Table 3 shows the evaluation of six classification models (LR, XGB, DTC, RFC, KNN, NBC) for predicting AQI in Hyderabad. XGBoost (XGB) and Random Forest Classifier (RFC) showed outstanding results with testing accuracy of 86.09% and 84.93% respectively, log loss of 0.4839 and 0.4690 respectively and Kappa score of 0.7554 and 0.7291 respectively, reflecting strong agreement and predictability. KNN is also performing good, with almost an equal accuracy with RFC, but with a slightly lower Kappa score. Logistic Regression and Naive Bayes Classifier have a moderate performance with accuracy scores below 80% and Decision Tree Classifier has poor testing accuracy and high training accuracy, suggesting overfitting. In general, XGBoost and RFC are the best models. The comparison of classification performance metric of Hyderabad city is shown in Figure 2.

**Table 3**. Benchmark metrics calculation for Hyderabad City

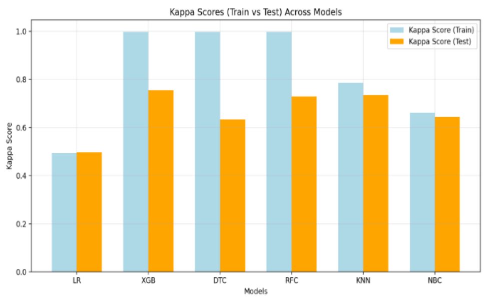
| | LR | XGB | DTC | RFC | KNN | NBC |
|---|---|---|---|---|---|---|
| **Accuracy (Train)** | 0.71837 | 0.99857 | 0.99857 | 0.99857 | 0.87634 | 0.79699 |
| **Accuracy (Test)** | 0.72463 | 0.86086 | 0.78695 | 0.84927 | 0.84927 | 0.79130 |
| **Log Loss (Test)** | 0.76190 | 0.48387 | 7.5757 | 0.46903 | 2.05 | 0.8033 |
| **Kappa Score (Train)** | 0.49408 | 0.99754 | 0.99754 | 0.99754 | 0.78576 | 0.66162 |
| **Kappa Score (Test)** | 0.49687 | 0.75542 | 0.63373 | 0.72911 | 0.73466 | 0.64427 |
| **MCC** | 0.49957 | 0.75605 | 0.63421 | 0.73132 | 0.73499 | 0.64590 |



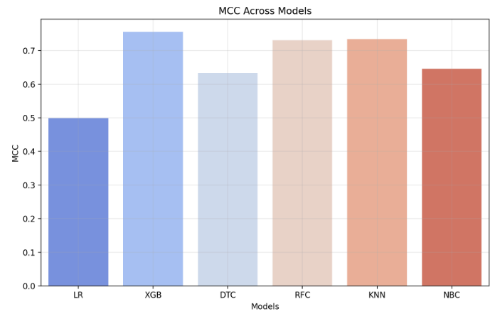

(a)




(b)




(c)

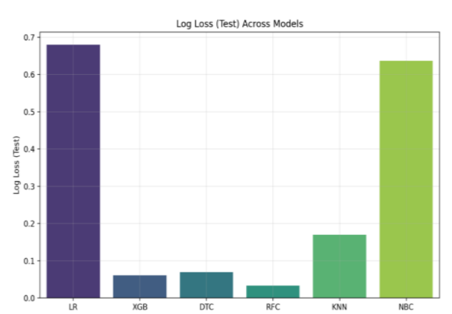



(d)

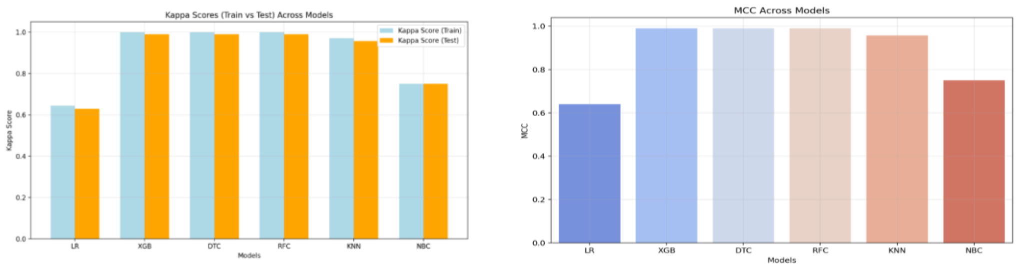
**Fig.2.** Comparison of various metrics for classification models for Hyderabad city

Table 4 presents the classification results of the models trained on the dataset for the city of Delhi; XGBoost, Decision Tree and Random Forest models show the highest testing accuracy of >99%, lowest log loss (<0.01), and the highest Kappa and MCC at 0.99, with good performance. KNN also had a good accuracy of 96% with slightly worse agreement measures, whereas Logistic Regression (73.1%) and Naïve Bayes (81.3%) had lower predictive power. Overall, XGBoost, DTC, and RFC showed the best performance in terms of the accuracy of AQI prediction using the tree-based models. The classification performance metrics for Delhi city are compared as shown in Figure 2.


(a)




(b)




(c) (d)

**Fig.3.** Comparison of various metrics for classification models for Delhi city

**Table 4**. Benchmark metrics calculation for Delhi city

| | LR | XGB | DTC | RFC | KNN | NBC |
|---|---|---|---|---|---|---|
| **Accuracy (Train)** | 0.74171 | 1.0 | 1.0 | 1.0 | 0.97 | 0.8118 |
| **Accuracy (Test)** | 0.73099 | 0.995 | 0.99822 | 0.99693 | 0.96 | 0.81297 |
| **Log Loss (Test)** | 0.67978 | 0.0612 | 0.06979 | 0.0330 | 0.17 | 0.63650 |
| **Kappa Score (Train)** | 0.64473 | 1.0 | 1.0 | 1.0 | 0.97 | 0.75004 |
| **Kappa Score (Test)** | 0.63 | 0.99 | 0.99 | 0.99 | 0.957 | 0.75 |
| **MCC** | 0.64 | 0.99 | 0.99 | 0.99 | 0.957 | 0.75 |

## 5 CONCLUSIONS

This study is a comparison of regression model and classification model for the prediction of AQI for Delhi and Hyderabad. XGBoost consistently achieves excellent results, like almost perfect accuracy (99.5%) for classification and $R^2 = 0.9998$ for regression, in Delhi. XGBoost's adaptability is showcased with a high $R^2$ of 0.786 and classification accuracy rate of 86% for the Hyderabad dataset. Based on these data, XGBoost is able to capture complex interactions between pollutants and achieve accurate prediction of AQI. XGBoost in AQI measurement reduces air pollution health hazards and aids environmental management decision-making.

## References

1. I. Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and health impacts of air pollution: a review. *Front. Public Health* **8**, 14 (2020)
2. J. Dutta, P.P. Singh, Air pollutants and acid precipitation: impact on ecology and human health, in *Multidimensional Approaches to Impacts of Changing Environment on Human Health* (CRC Press, Boca Raton, 2021), pp. 3–25

3. P. Kumar, A critical evaluation of air quality index models (1960–2021). *Environ. Monit. Assess.* **194**(5), 1–45 (2022)
4. A. Pant, S. Sharma, K. Pant, Evaluation of machine learning algorithms for air quality index (AQI) prediction. *J. Reliab. Stat. Stud.* (2023).
5. M. Subrahmanyam, P. Nagendrudu, T. Ramana, Comparison of effective machine learning technique for air quality forecast, in *Proceedings of the International Conference on Smart Computing and Communication* (Springer, Singapore, 2023).
6. S.V. Manikandan, Y. Abilash, S.H. Prasanth, Optimized feature selection for air quality index forecasting using GPR and SARIMA models, in *Proceedings of the 2023 International Conference on Data Science, Machine Learning, and Applications (ICDSMLA)* (2023)
7. H. Liu, Q. Li, D. Yu, G. Yu, Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Appl. Sci.* **9**, 4069 (2019).
8. T.M.T. Lei, S.C.W. Ng, S.W.I. Siu, Application of ANN, XGBoost, and other ML methods to forecast air quality in Macau. *Sustainability* **15**(6), 5341 (2023).
9. J. Ma, Z. Yu, Y. Qu, J. Xu, Y. Cao, Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai. *Aerosol Air Qual. Res.* **20** (2020).
10. J.K. Sethi, M. Mittal, A new feature selection method based on machine learning technique for air quality dataset. *J. Stat. Manag. Syst.* **22**(4), 456–472 (2019).
11. S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, Niveshitha, Comparative analysis study for air quality prediction in smart cities using regression techniques. *IEEE Access* **11** (2023).
12. S.A. Aram, E.A. Nketiah, B.M. Saalidong, H. Wang, A.R. Afitiri, A.B. Akoto, P.O. Lartey, Machine learning-based prediction of air quality index and air quality grade: a comparative analysis. *Int. J. Environ. Sci. Technol*. 21(2), 1345–1360 (2024)