EQA - International Journal of Environmental Quality
ISSN 2281-4485 - Vol. 71 (2026): 19-28
Journal homepage: https://eqa.unibo.it/

published by

Leveraging machine learning to analyze and forecast air quality trends in Kota City, India

Monika Sharma, Mahendra Pratap Choudhary*, Anil K. Mathur

Department of Civil Engineering, Rajasthan Technical University, Kota (Rajasthan), India

*Corresponding author E.mail: mpchoudhary@rtu.ac.in

Article info
Received 25/5/2025; received in revised form 21/8/2025; accepted 10/9/2025
DOI: 10.60923/issn.2281-4485/21975

Abstract
Air quality is a critical indicator of environmental health, directly impacting human well-being and ecological stability. Rapid urbanization and industrialization have recently exacerbated air pollution, necessitating robust monitoring and predictive frameworks. This study investigates air quality trends in Kota city of Rajasthan, India and using data from 2017 to 2023. Machine learning models, including linear regression (LR), random forest (RF), decision tree (DT), support vector regressor (SVR), and K-nearest neighbors (KNN), were employed to analyze predict air quality index (AQI) values based on key pollutants such as $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, SO_2, CO, Ozone, Benzene, Ethyl-Benzene, *m* & *p*-Xylene considering the effects of meteorological factors like relative humidity (RH), wind speed (WS), wind directions (WD), and barometric pressure (BP). Among these, the decision tree regressor shows almost perfect fit on the training set (R^2 score 0.9999) and excellent test performance (R^2 score 0.9991), suggesting a very accurate prediction model. However, it exhibits potential overfitting, limiting its generalization capabilities. On the other hand, the random forest regressor provides a balance of accuracy and robustness, achieving an R^2 score of 0.9831, making it the preferred model for reliable predictions. The study delves into pollutant contributions, evaluates model performances, and explores actionable insights for policymakers. By leveraging machine learning approaches, the study aims to provide a comprehensive framework for analyzing air quality trends and supporting decision-making processes.

Keywords: *Air quality index, Machine learning models, Exploratory data analysis, NCAP*

Introduction

Air pollution remains one of the most pressing global challenges, with profound implications for human health, ecosystems, and the climate. According to the World Health Organization, ambient air pollution is responsible for approximately 4.2 million premature deaths annually (WHO, 2018). This stark statistic underscores the urgent need for effective air quality management strategies across the globe. In India, the introduction of the National Clean Air Programme (NCAP) in 2019 marked a significant step toward tackling urban air pollution, setting ambitious goals to reduce particulate matter (PM_{10} and $PM_{2.5}$) concentrations by 20-30% by 2024 compared to 2017 levels (NCAP, 2019; Sharma et al., 2024). The city of Kota is included in this program, where rapid urbanization, industrial expansion, and vehicular emissions have led to concerning air quality trends. Kota, a growing industrial and educational hub in Rajasthan, epitomizes the challenges faced by rapidly urbanizing cities in India. Industrial activities, construction projects, and a surge in vehicular traffic have contributed to elevated levels of pollutants such as PM_{10}, $PM_{2.5}$, and nitrogen dioxide (NO_2). As the city grapples with these challenges, leveraging advanced technologies, particularly machine learning (ML) and artificial intelligence (AI), offers promising solutions for monitoring and mitigating air pollution. Recent studies have highlighted the transformative potential of ML and AI in air quality monitoring and prediction. These technologies enable the analysis of lar-

EQA 71 (2026): 19 - 28

DOI: 10.60923/issn.2281-4485/21975

ge datasets, including pollution metrics and meteorological variables, to provide accurate forecasts and actionable insights. A variety of ML and deep learning (DL) models have been employed, each offering unique strengths and applications. The use of ML techniques was recently explored in a study including AdaBoost, support vector regression (SVR), random forest (RF), and K-nearest neighbors (KNN), alongside DL models like the multi-layer perceptron (MLP) regressor and long short-term memory (LSTM) networks. The study focused on optimizing features to enhance the prediction of pollutants such as $PM_{2.5}$, PM_{10}, and ozone (O_3). Among these, LSTM demonstrated exceptional accuracy, achieving R^2 values as high as 0.998, thereby outperforming other models (Neo et al., 2023). AQNet, a multimodal AI model was introduced that integrates satellite data from the European Space Agency's Copernicus project with ground-level pollution measurements. This approach improved the prediction of NO_2, O_3, and PM_{10} concentrations, emphasizing the importance of urban and traffic features in influencing pollution levels (Rowley & Karakuş, 2023).A hybrid SVR-GWO (Gray Wolf Optimizer) model was developed to predict aerosol optical depth (AOD) in Pakistan, demonstrating improved accuracy over standalone models. Similarly, Gupta et al. (2023) compared ML algorithms like RF regression, CatBoost, and SVR for predicting the Air Quality Index (AQI) in Indian cities, identifying RF and CatBoost as the most reliable models for AQI prediction (Zaheer et al., 2023). Other studies have showcased the effectiveness of time-series models like LSTM and GRU in forecasting pollutant concentrations. These models captured temporal patterns in pollution data, offering valuable tools for proactive air quality management (Hsieh et al., 2022; Cican et al., 2023). Globally, researchers have employed ML and AI to address region-specific air quality challenges. In Visakhapatnam, it was found that the CatBoost model excelled in AQI prediction, achieving an impressive R^2 of 0.9998 (Ravindiran et al., 2023). Similarly, studies in Jaipur identified PM_{10} as the major pollutant post-COVID-19 lockdown, underscoring the persistent air quality challenges in urban areas (Ruhela et al., 2022). Remote sensing data was combined with RF models to analyze air pollution in Egypt, highlighting the impact of road proximity and temperature on air quality (Abu El-Magd et al., 2023). In China, a boosted regression tree (BRT) model was used to explore the relationship between $PM_{2.5}$ levels and land use patterns, revealing significant seasonal variations (Liang et al., 2020). Inspired by these advancements, the present research focuses on applying ML techniques to analyze air quality trends in Kota city. By integrating statistical and ML methods, this study aims to bridge the gap between traditional monitoring approaches and modern predictive frameworks. The integration of ML and AI into air quality monitoring systems represents a paradigm shift in environmental management. By providing accurate, real-time predictions, these technologies empower stakeholders to make informed decisions, mitigating the adverse effects of air pollution on health and the environment. For cities like Kota, where traditional monitoring methods may fall short, embracing innovative approaches is not just an opportunity, it is a necessity.

Research Methodology

The research methodology for AQI prediction followed a structured, multi-phase approach as depicted in Figure 1, ensuring a systematic workflow from data preparation to model evaluation.

Data Source

The dataset for this study was sourced from the Rajasthan State Pollution Control Board (RSPCB). It includes comprehensive records of pollutant concentrations such as $PM_{2.5}$, PM_{10}, NO, NO_2, NOx, NH_3, SO_2, CO, ozone, benzene, ethyl-benzene, and *m* & *p*-xylene, as well as meteorological parameters like relative humidity (RH), wind speed (WS), wind direction (WD), and

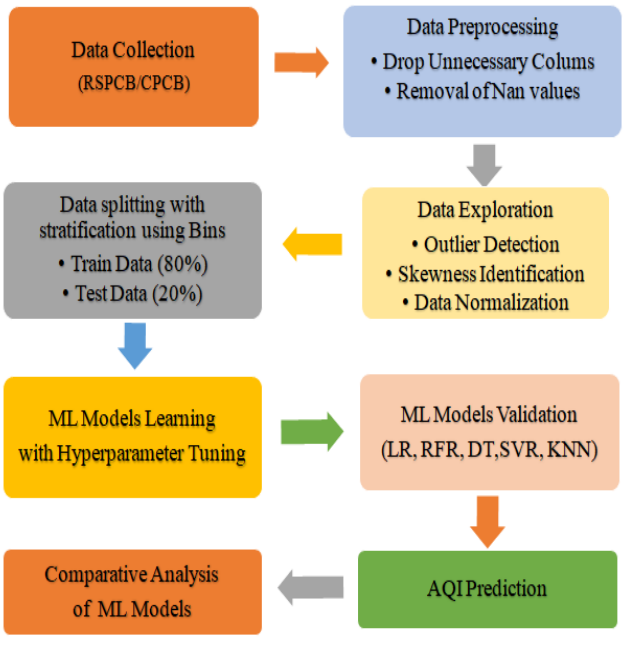

Figure 1. *Flowchart showing research methodology*

DOI: 10.60923/issn.2281-4485/21975

barometric pressure (BP). These records, spanning from 2017 to 2023, provide a robust temporal dataset for analyzing air quality trends in Kota city, Rajasthan.

Data Pre-processing

To ensure the quality and consistency of the dataset, a structured preprocessing workflow was employed:

Cleaning and Formatting, Unnecessary columns were removed, and missing values were handled appropriately. Depending on the nature of the missing data, strategies such as mean imputation or interpolation were applied to preserve the dataset's integrity.

Exploratory Data Analysis (EDA): Exploratory data analysis included the following steps.

Outlier Detection. Statistical methods and visual tools like box plots were used to identify and handle anomalies in the data.

Skewness Identification. The skewness of data distributions was analyzed to understand the need for transformation or normalization.

Normalization. To standardize feature values, normalization techniques were applied, ensuring that all variables contributed equally to model performance.

Data Splitting with Stratification

The dataset was stratified into bins based on AQI levels to maintain a balanced distribution of different AQI categories. This stratification ensured that both training and test datasets adequately represented the range of air quality conditions. The data was then split into training (80%) and testing (20%) subsets.

Model Training

Five machine learning models were selected to capture diverse data patterns and relationships.

Linear regression (LR). A baseline model to understand linear dependencies in the data.

Random forest regressor (RF). A robust ensemble model leveraging decision tree to capture complex patterns.

Decision tree regressor *(DT).* A tree-based model that splits data iteratively for prediction.

Support vector regressor (SVR). A model designed to minimize prediction error by finding the optimal hyperplane in feature space.

K-nearest neighbors (KNN). A non-parametric model that predicts based on the closest training samples in feature space.

Model validation

The trained models were evaluated using the testing dataset to assess their ability to generalize predictions. Validation metrics ensured that the models performed reliably under unseen conditions.

Performance evaluation and comparison

The performance of the machine learning models was assessed using the following metrics:

Mean absolute error (MAE). Measures the average magnitude of errors in predictions, without considering their direction.

Mean squared error (MSE). Highlights larger errors by squaring the differences between predicted and actual values.

Root mean squared error (RMSE). The square root of MSE, providing a scale-sensitive measure of prediction error.

R^2 Score. Indicates the proportion of variance in the dependent variable explained by the model.

These metrics provided a comprehensive understanding of each model's accuracy and reliability in predicting AQI values.

Results and Discussion

The exploratory data analysis (EDA), which allowed us to visually inspect the distribution and outliers for each feature in the dataset in a structured manner, is shown in Figure 2. It consists of a series of "violin plots" for multiple variables related to air pollution and meteorological parameters. The violin plot combines a box plot and a kernel density plot. The "red dots" represent the individual data points (scatter plot overlay) and the "blue shapes" represent the distribution of the data (density plot), showing how the data is spread across the range of values. The "box plot inside" the violin (horizontal lines) shows key statistics and the central line is the "median". The edges of the box represent the "inter-quartile range (IQR)", which captures the middle 50% of the data. The whiskers show the range of data excluding outliers.

Explanation of variables and graphs

Air pollutants

$PM_{2.5}$ and PM_{10} ($\mu g/m^3$). These plots show the distribution of particulate matter concentrations. The spread indicates some extreme outliers, possibly during pollution events like festivals or crop burning.

NO, NO_2, NOx ($\mu g/m^3$ or ppb).These represent

EQA 71 (2026): 19 - 28

DOI: 10.60923/issn.2281-4485/21975

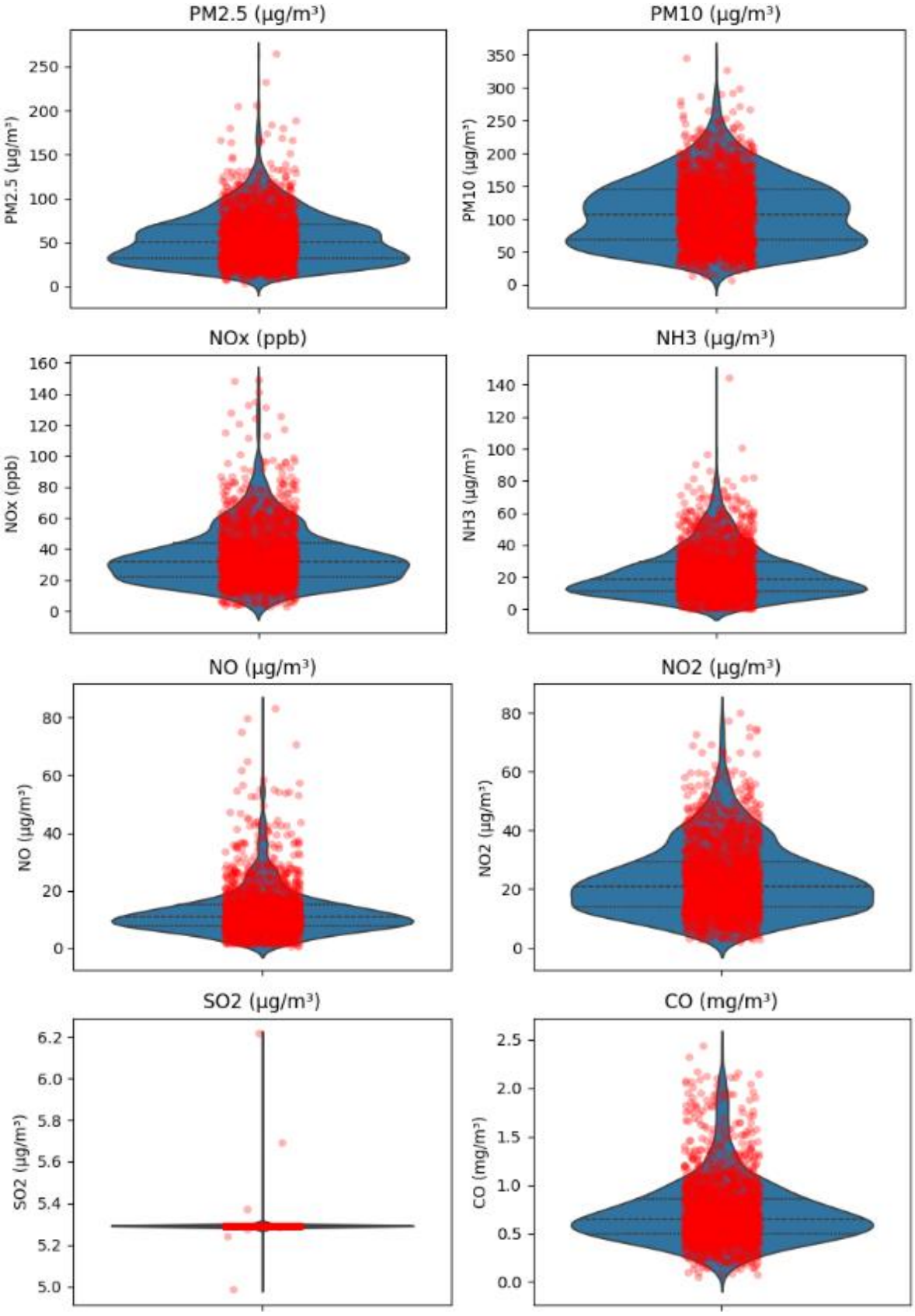

Figure 2. *Exploratory data analysis of air pollution and meteorological variables*

EQA 71 (2026): 19 - 28

DOI: 10.60923/issn.2281-4485/21975

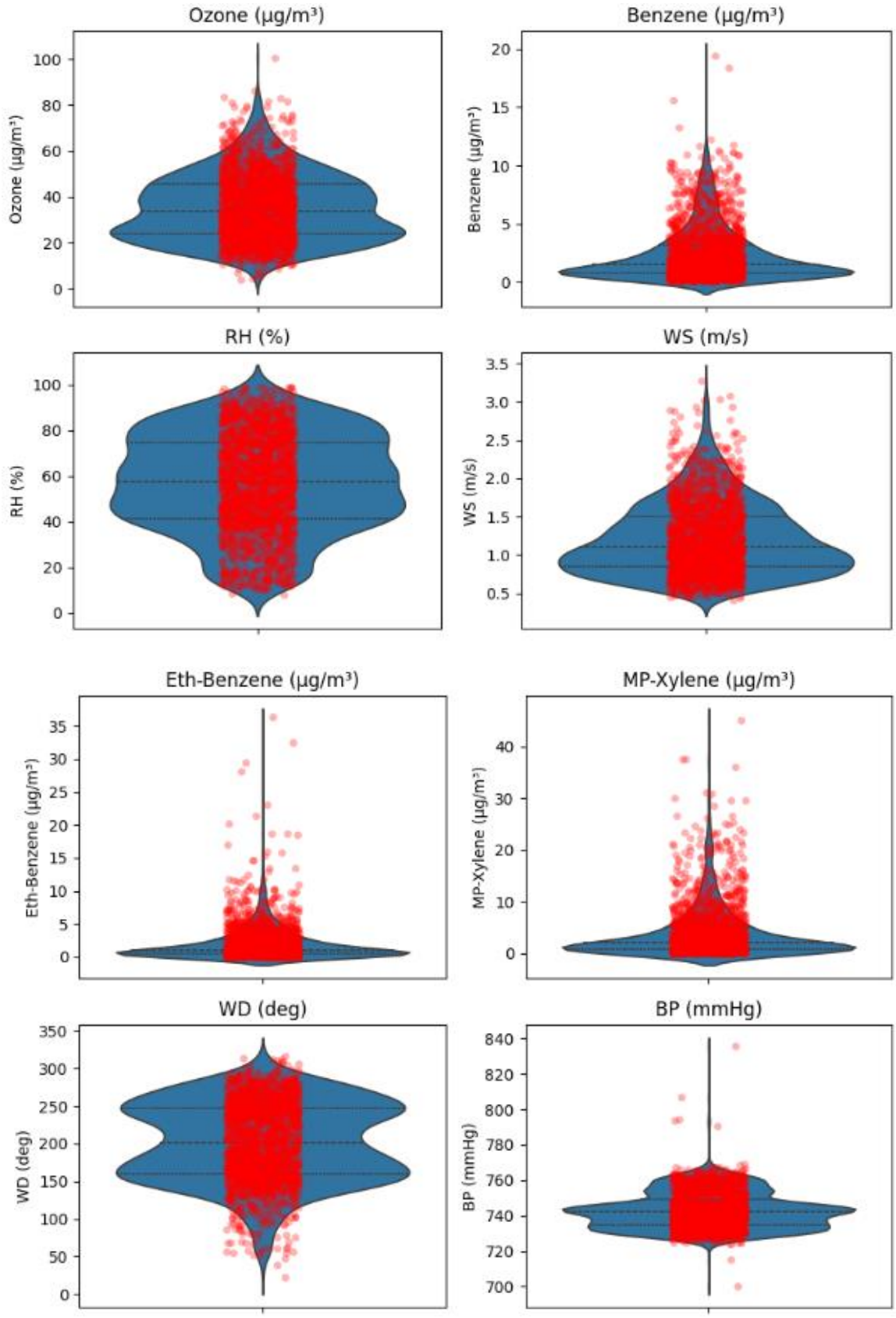

Figure 2 (segue). *Exploratory data analysis of air pollution and meteorological variables*

DOI: 10.60923/issn.2281-4485/21975

nitrogen oxides. Their distributions are skewed, with most values concentrated at lower ranges but some outliers at high concentrations.

NH₃ (µg/m³). Shows ammonia concentrations. The distribution appears slightly skewed with outliers at higher values.

SO₂ (µg/m³). Sulfur dioxide levels are highly concentrated in a very narrow range, with almost no significant variability.

CO (mg/m³). The distribution of carbon monoxide shows a spread with some prominent outliers.

Ozone (µg/m³). The distribution of ozone levels shows moderate spread and no extreme outliers.

Benzene, ethyl-benzene, m & p-xylene (µg/m³). These volatile organic compounds show distributions with notable skewness and some outliers.

Meteorological parameters

Relative humidity (RH, %). The data shows a relatively uniform distribution with no significant outliers.

Wind speed (WS, m/s). Wind speed distribution is skewed, with most values concentrated at lower ranges.

Wind direction (WD, degrees). Wind direction is evenly distributed, as expected in many datasets with varying wind patterns.

Barometric pressure (BP, mmHg). Barometric pressure data shows a tight distribution, indicating low variability over time.

Based on the Figure 2, it can be inferred that many pollutants (e.g., $PM_{2.5}$, PM_{10}, NOx, CO, and benzene) exhibit significant outliers, suggesting episodic high pollution events. SO_2 shows almost no variability, which may imply a consistent source or limited emission variability in the area. Meteorological parameters like wind direction and pressure are more stable, while relative humidity and wind speed show moderate variability.

Performance of machine learning models

Linear regression. The linear regression model demonstrates strong training performance, with predicted values closely matching actual ones. Metrics, including MAE of 17.98 and R^2 of 0.87, indicate effective learning of patterns and a good fit for the training dataset. The testing data shows a similar trend but with greater dispersion at higher values, reflecting prediction errors. Metrics (MAE 18.82, R^2 0.86) indicate good generalization but slightly reduced performance compared to the training dataset**.** Figures 3(a) and 3(b) show the linear regression results on training v/s testing datasets.

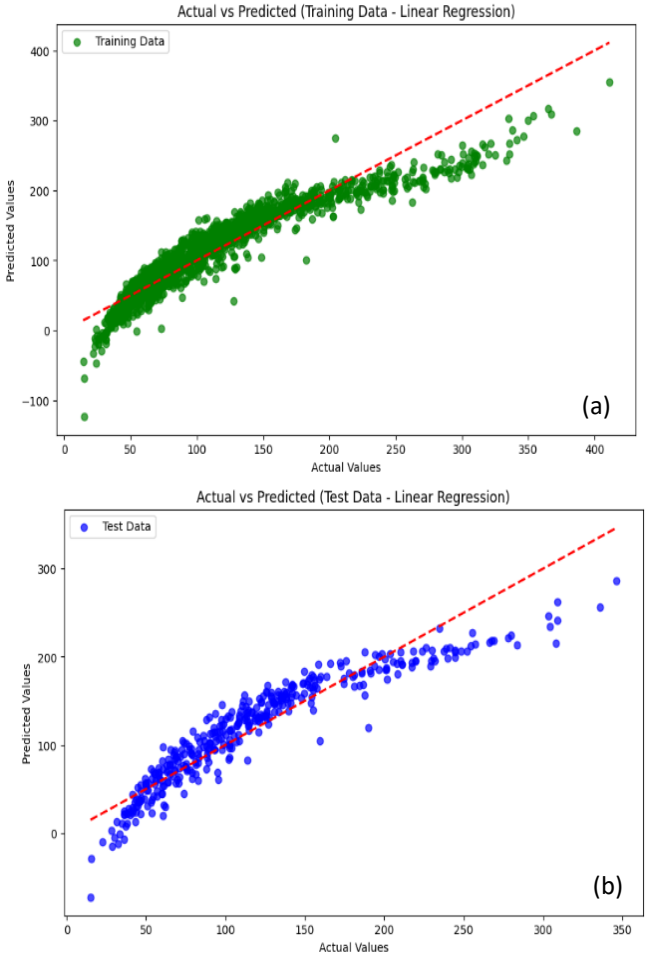

Figure 3. *Linear regression results of actual v/s predicted value*

In conclusion, the linear regression model effectively learns patterns in the training dataset, achieving a strong fit. While it generalizes well to unseen data, slight performance drops highlight opportunities for refinement to improve prediction accuracy on higher values.

Random forest regressor. The random forest regressor demonstrates exceptional accuracy and generalization, with closely aligned training and testing metrics. On the training set, it achieves an MAE of 6.66, MSE of 79.95, RMSE of 8.94, and an R^2 score of 0.98, reflecting minimal errors and a strong fit. On the testing set, it performs slightly better with an MAE of 6.53, MSE of 69.90, RMSE of 8.36, and an R^2 score of 0.98, indicating effective generalization to unseen data. Figures 4(a) and 4(b) show the random forest regression results on training v/s testing datasets. The minimal difference in error metrics and consistently high R^2 scores underscore this model's robustness, reliability, and capability to capture patterns while avoiding overfitting.

EQA 71 (2026): 19 - 28

DOI: 10.60923/issn.2281-4485/21975

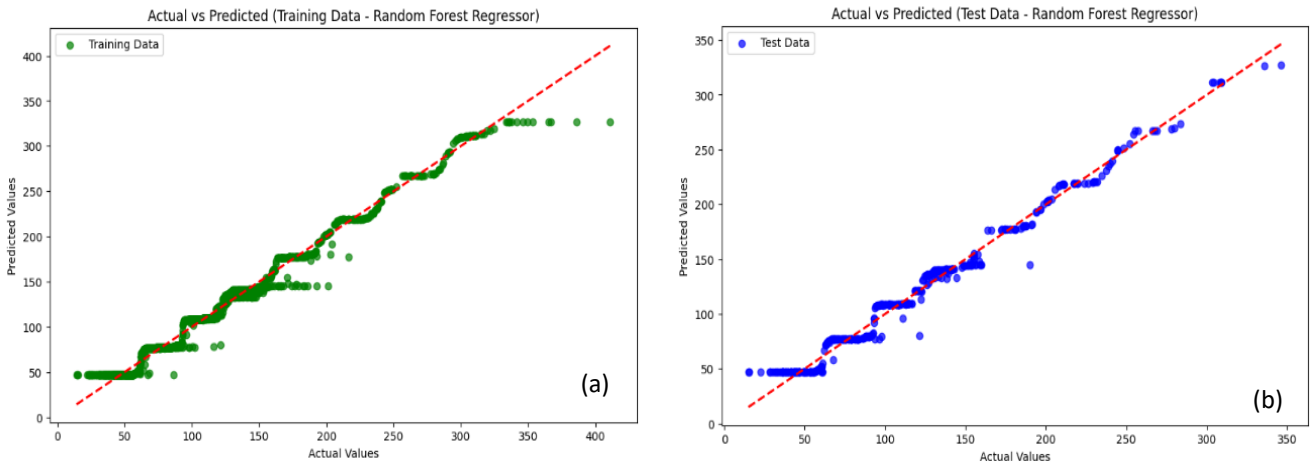

Figure 4. *Random forest regression results of actual v/s predicted values*

Decision tree regressor. The decision tree regressor demonstrates exceptional performance on the training dataset, with an MAE of 0.2464, MSE of 0.5811, RMSE of 0.7623, and an R^2 score of 0.9999, indicating near-perfect fit. On the test dataset, performance slightly declines but remains highly accurate, with an MAE of 0.7070, MSE of 3.5985, RMSE of 1.8970, and an R^2 score of 0.9991. Figures 5(a) and 5(b) show the decision tree results on training v/s testing datasets The increase in errors on unseen data suggests mild overfitting, as the model performs exceptionally well on the training set but experiences a slight drop in generalization. Overall, it is a highly accurate and effective predictive tool.

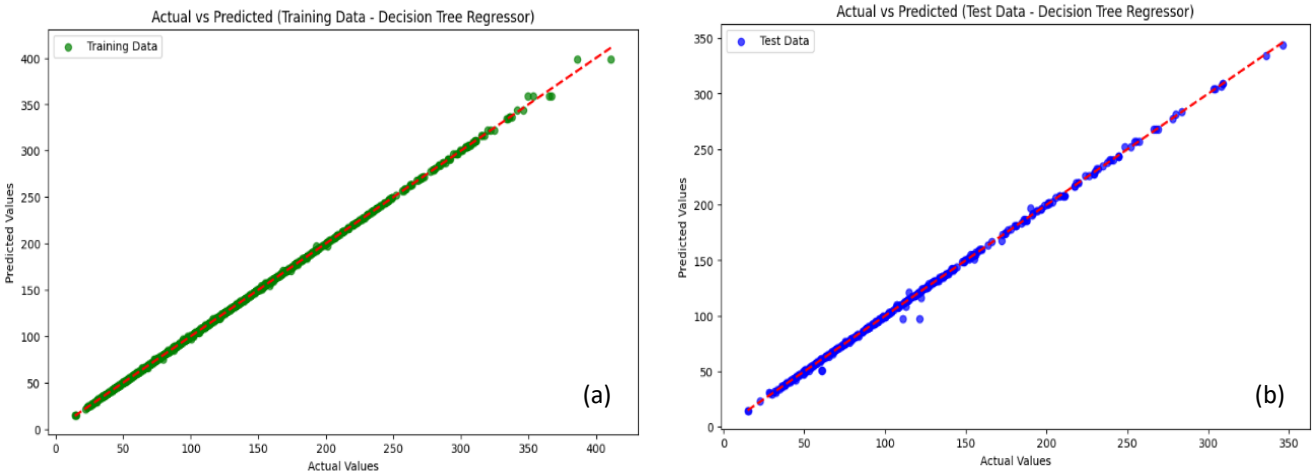

Figure 5. *Decision tree regression results of actual v/s predicted values*

Support vector regressor. The support vector regressor (SVR) performs well on the training data, with an MAE of 2.4066, MSE of 36.9748, RMSE of 6.0807, and an R^2 score of 0.9916, indicating a strong fit. However, on the test data, performance declines, with an MAE of 6.9799, MSE of 163.4713, RMSE of 12.7856, and an R^2 score of 0.9605, suggesting a loss in accuracy and generalization. Figure 6(a) and 6(b) show the support vector regressor results on training v/s testing datasets. The increase in errors and drop in R^2 suggest potential overfitting, as the model struggles to generalize to new data despite high accuracy on the training set.

K-nearest neighbors. The KNN model shows strong performance on the training dataset, with an MAE of 10.7026, MSE of 279.7702, RMSE of 16.7263, and an R^2 score of 0.9361. On the test dataset, the MAE increases to 15.9931, MSE to 603.7686, RMSE to 24.5717, and R^2 drops to 0.8542, indicating slightly reduced accuracy and higher errors. This performance gap suggests potential over fitting, as the model fits the

EQA 71 (2026): 19 - 28

DOI: 10.60923/issn.2281-4485/21975

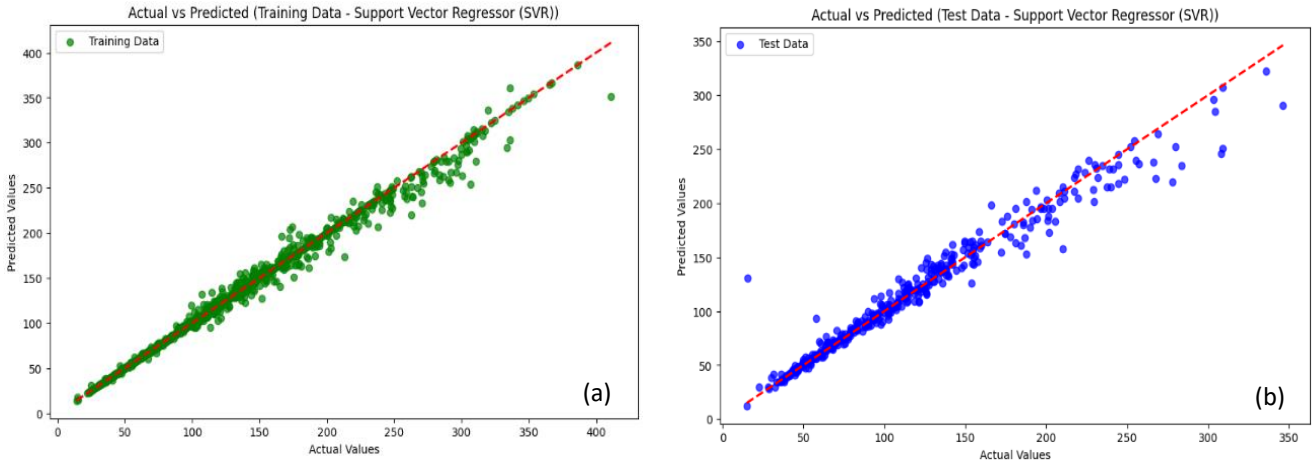

Figure 6. *Support vector regressor results of actual v/s predicted values*

training data better than the test data. Despite this, the relatively high R^2 score for both datasets shows the model maintains good overall predictive ability. Figures 7(a) and 7(b) show the K-nearest neighbors results on training v/s testing datasets.

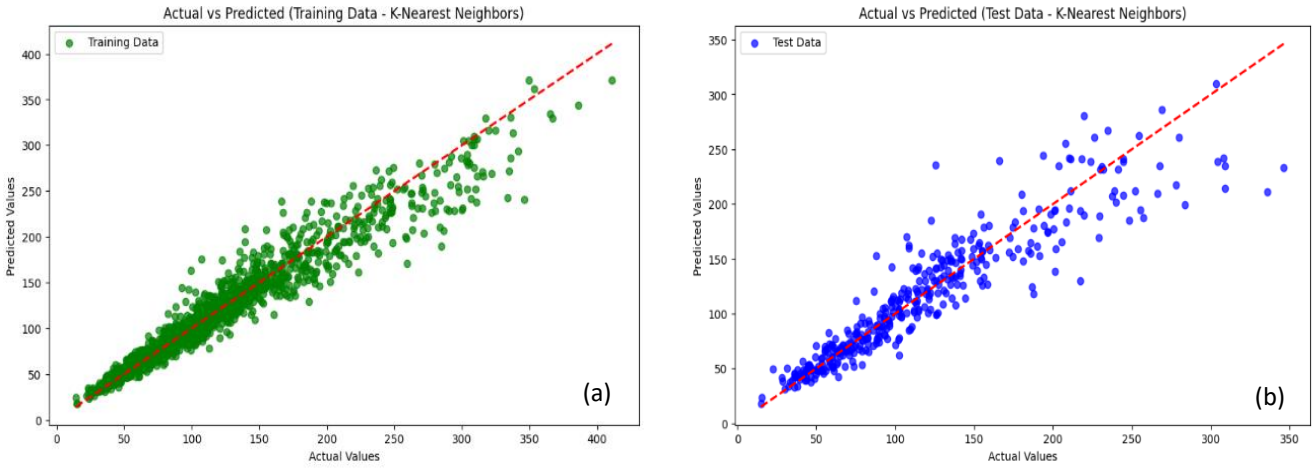

Figure 7. *K-nearest neighbors results of actual v/s predicted values*

The overall performance of each model evaluated on the training and test datasets is represented in Table 1. The Table 1 shows that the decision tree regressor offers near-perfect accuracy but could be overfitting due to its complexity. Random forest regressor is the most balanced, with consistently high accuracy across

Model	Dataset	MAE	MSE	RMSE	R^2 Score
Linear regression (LR)	Training	17.9764	560.6084	23.6772	0.8719
	Test	18.8163	567.5179	23.8226	0.8629
Random forest regressor (RF)	Training	6.6575	79.9481	8.9414	0.9817
	Test	6.5289	69.9036	8.3608	0.9831
Decision tree regressor (DT)	Training	0.2464	0.5811	0.7623	0.9999
	Test	0.7070	3.5985	1.8970	0.9991
Support vector regressor (SVR)	Training	2.4066	36.9748	6.0807	0.9916
	Test	6.9799	163.4713	12.7856	0.9605
K-nearest neighbors (KNN)	Training	10.7026	279.7702	16.7263	0.9361
	Test	15.9931	603.7686	24.5717	0.8542

Table 1
Evaluation metrics for each model on training and test datasets

DOI: 10.60923/issn.2281-4485/21975

both training and test datasets. While the SVR performs well on the training set, its generalization to the test set isn't as strong. KNN and linear regression are less accurate compared to other models. Random forest regressor is likely the best choice if we are optimizing for generalization and consistent performance. However, if near-perfect accuracy is critical, decision tree regressor might be considered, keeping in mind potential overfitting risks.

Discussion on variable influence

The predictive modeling results underscore the relative importance of air pollutants and meteorological parameters in determining AQI levels. Across all models, pollutant concentrations such as $PM_{2.5}$, PM_{10}, NO_2, and benzene showed higher influence on AQI predicttion compared to meteorological factors. The decision tree and random forest models, which offer insight into variable importance, consistently ranked $PM_{2.5}$ and PM_{10} among the top predictors highlighting the dominant role of particulate matter in air quality degradation. Nitrogen-based pollutants (NO, NO_2, NOx) and benzene also emerged as strong contributors, likely reflecting vehicular and industrial emissions in Kota. In contrast, meteorological variables like wind direction and barometric pressure exhibited lower importance, although wind speed and relative humidity showed moderate influence, potentially by modulating pollutant dispersion and concentration levels. Notably, the random forest model's robustness allowed a clearer interpretation of variable ranking, with $PM_{2.5}$ often emerging as the most influential single factor. This suggests that air quality management strategies should prioritize particulate emission controls. The findings affirm that while meteorological conditions shape pollutant behavior, the core drivers of AQI variability in Kota are anthropogenic emissions, primarily from transport, construction, and industry.

Conclusions

This study successfully applied machine learning models to analyze and predict air quality trends in Kota city, Rajasthan, using pollutant and meteorological data spanning from 2017 to 2023. Five models namely linear regression, random forest regressor, decision tree regressor, support vector regressor, and K-nearest neighbors were evaluated for their predictive capabilities based on metrics such as MAE, MSE, RMSE, and R^2 score. Among these models, the decision tree regressor (DTR) achieved the best performance, with the lowest MAE, MSE, and RMSE with an R^2 score of 0.9981, showcasing its ability to accurately capture the relationships between features. The only issue with DTR could be overfitting due to its complexity. On the other hand, random forest regressor (RFR) is the most balanced, with consistently high accuracy across both training and test datasets. By integrating pollutants like $PM_{2.5}$, PM_{10}, NO_2, and benzene with meteoro-logical parameters such as wind speed, wind direction, and humidity, the models provided actionable insights into pollutant contributions and their effects on air quality. These findings offer significant value to policy-makers, enabling data-driven decision-making and the assessment of initiatives like the national clean air programme (NCAP).

Acknowledgement

The authors are thankful to the Central Pollution Control Board, New Delhi and Rajasthan State Pollution Control Board for providing the necessary data on their websites.

References

ABU EL-MAGD S., SOLIMAN G., MORSY M., KHARBISH S. (2023) Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing. International Journal of Environmental Science and Technology, 20(6): 6103–6116. https://doi.org/10.1007/s13762-022-04367-6

CICAN G., BUTURACHE A. N., MIREA R. (2023). Applying Machine Learning Techniques in Air Quality Prediction—A Bucharest City Case Study. Sustainability (Switzerland), 15(11). https://doi.org/10.3390/su15118445

GUPTA N. S., MOHTA Y., HEDA K., ARMAAN R., VALARMATHI B., ARULKUMARAN G. (2023) Prediction of air quality index using machine learning techniques: A comparative analysis. Journal of Environmental and Public Health, 1–26. https://doi.org/10.1155/2023/4916267

HSIEH H. P., WU S., KO C. C., SHEI C., YAO Z. T., CHEN Y. W. (2022) Forecasting fine-grained air quality for locations without monitoring stations based on a hybrid predictor with spatial-temporal attention based network. Applied Sciences (Switzerland), 12(9). https://doi.org/10.3390/app12094 LIANG Y. C., MAIMURY Y., CHEN A. H. L., JUAREZ J. R. C. (2020) Machine learning-based prediction of air quality. Applied Sciences (Switzerland), 10(24): 1–17.

https://doi.org/10.3390/app10249151

DOI: 10.60923/issn.2281-4485/21975

NCAP. (2019) National clean air programme. Ministry of Environment, Forest and Climate Change, Govt. of India.https://prana.cpcb.gov.in/ncapDashboard/download_public_portal_file/NCAP_Report.pdf

NEO E. X., HASIKIN K., LAI K. W., MOKHTAR M. I., AZIZAN M. M., HIZADDIN H. F., RAZAK S. A., YANTO. (2023) Artificial intelligence-assisted air quality monitoring for smart city management. PeerJ Computer Science, 9. https://doi.org/10.7717/peerj-cs.1306

RAVINDIRAN G., HAYDER G., KANAGARATHINAM K., ALAGUMALAI A., SONNE C. (2023) Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam. Chemosphere, 338. https://doi.org/10.1016/j.chemosphere.2023.139518

ROWLEY A., KARAKUŞ O. (2023) Predicting air quality via multimodal AI and satellite imagery. Remote Sensing of Environment, 293.
https://doi.org/10.1016/j.rse.2023.113609

RUHELA M., MAHESHWARI V., AHAMAD F., KAMBOJ V. (2022) Air quality assessment of Jaipur city Rajasthan after the COVID-19 lockdown. Spatial Information Research, 30(5):597–605.
https://doi.org/10.1007/s41324-022-00456-3

SHARMA M., CHOUDHARY M. P., MATHUR ANIL K. (2024) Mitigating air pollution and protecting public health: analyzing the impact of national clean air programme in Kota, Rajasthan. Current World Environment. 19(3). http://dx.doi.org/10.12944/CWE.19.3.12

WHO. (2018) Burden of disease from ambient air pollution for 2016. Version 2 May 2018. Summary of results. Geneva: World Health Organization; 2018.
https://cdn.who.int/media/docs/default-source/air-pollution-documents/air-quality-and-health/aap_bod_results_may2018_final.pdf

ZAHEER K., SAEED S., TARIQ S.(2023) Prediction of aerosol optical depth over Pakistan using novel hybrid machine learning model. Acta Geophys. 71:2009–2029.
https://doi.org/10.1007/s11600-023-01072-x

ZIMMERMAN N., PRESTO A.A., KUMAR S.P.N., GU J., HAURYLIUK A., ROBINSON E.S., ROBINSON A.L., SUBRAMANIAN R. (2018) A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmospheric Measurement Techniques, 11(1):291–313.
https://doi.org/10.5194/amt-11-291-2018