DOI: doi.org/10.64388/IREV9I3-1710528-450

Predicting Air Quality Index (AQI) in Western Uttar Pradesh Cities Using Machine Learning Models: A Comparative Analysis

SALIL KUMAR GUPTA[1], PRAVEEN KUMAR YADAV[2]
[1, 2]Department of Civil Engineering, Institute of Engineering and Technology, Lucknow, Uttar Pradesh, India

Abstract- ***The objective of this project is to use a variety of machine learning techniques to create a reliable predictive model for predicting the Air Quality Index (AQI) in Indian cities in Western Uttar Pradesh. The data came from the Central Pollution Control Board (CPCB) and covered the years January 2024 to December 2024. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R2) were the primary evaluation metrics used to compare the performance of the four machine learning models: Linear Regression, K-Nearest Neighbors (KNN), Decision Tree Regressor, and Random Forest Regressor. In terms of predicted accuracy, XGBoost outperformed the other algorithms among these models, exhibiting an impressive R2 value of 0.9967. Together with lower MSE and RMSE values, the XGBoost model demonstrated a significant decrease in MAE (3.5282) when compared to the other models, whose MAEs ranged from 18.148. These results suggested that the predictions were more accurate and had less volatility. Agra, Meerut, Ghaziabad, Bareilly, and Vrindavan were the five cities in which the study also examined AQI trends. According to the findings, the highest AQI recorded in Agra was 321.89 in November, while the highest AQI recorded in Meerut was 427.09 in the same month. The highest AQI in Bareilly was recorded in February (118.73), while the highest in Ghaziabad was recorded in January (413.57). October saw the highest AQI in Vrindavan (309.41). These cities' varying average AQIs were a reflection of seasonal variations in air pollution levels.***

Indexed Terms- ***Linear Regression, Random Forest, Decision Tree Regressor, K-Nearest Neighbors (KNN), Western Uttar Pradesh***

I. INTRODUCTION

Air pollution is one of the main concerns in emerging nations like India, where rapid growth and environmental problems can create a number of difficulties. Two types of environmental pollution that seriously endanger both human health and the sensitive ecosystem's balance are air pollution and other types of pollution (Samad et al., 2023). Furthermore, breathing in contaminated air can exacerbate respiratory conditions and raise the risk of lung ailments, putting human health at risk. Therefore, it is necessary to ensure that the next generation has access to significant natural resources in order to protect human health, as well as to maintain ecosystem integrity and reduce air pollution (Ayus et al., 2023a). The primary cause of the increase in global temperatures, which poses a threat to the environment and human life, is climate change. The research conducted by Dewan and Lakhani (2022), mentioned that the pollutants such as ozone and particulate matter have sophisticated and interconnected relationships with air quality and climate change which adversely impacts the earth's energy balance by getting in the way of short-wave and long-wave radiations (Isaev et al., 2022). The frequency and intensity of air stagnation events, anticyclonic conditions, heat waves and meteorological phenomena will all increase as a result of this (Mondal et al., 2024).

The research on an ML method such as linear regression, decision tree, XGBoost, and random forests, based on research, it was inferred that the random forest method is the most accurate among all four methods mentioned by Sri Eshwar College of Engineering and Institute of Electrical and Electronics Engineers (Wang et al., 2023).

Some of study's research gaps are as follows: First off, large datasets spanning five years have often been used in prior studies to train machine learning models. Nevertheless, the dataset used in this research only spans two and a half years (Janarthanan et al., 2021). It's unclear how successful machine learning models are, particularly when working with smaller datasets. Second, several research has evaluated the performance of different machine-learning models. However, there isn't a comprehensive comparison of how well the random forests and XGBoost regression models perform (Natarajan et al., 2024).

In order to better understand the relationship between air quality and climate change at the local level, we can apply the developed model to a larger dataset from multiple cities to validate the results. We can also use the model to develop effective air quality control plans, ensure effective air quality, and improve the accuracy of AQI prediction by incorporating other pollutants and meteorological parameters (Liang et al., 2020).

Sources of Air Pollution

The presence of gases, liquids, or solids in the atmosphere in concentrations high enough to endanger people, other living things, or materials is known as air pollution. They can be characterized in Fig.1 as a variety of airborne circumstances when specific compounds are present in such high quantities that they may have unfavorable impacts on people and other substances (Gupta et al., 2023).

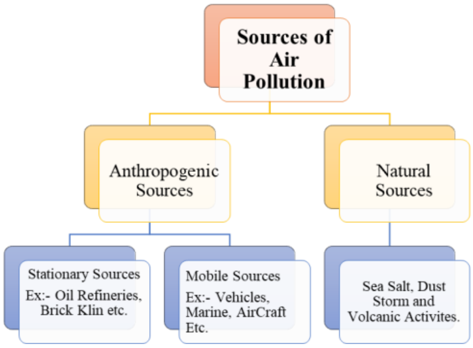

Sources of Air Pollution
Anthropogenic Sources
Natural Sources
Stationary Sources
Ex:- Oil Refineries, Brick Klin etc.
Mobile Sources
Ex:- Vehicles, Marine, AirCraft Etc.
Sea Salt, Dust Storm and Volcanic Activites.

Fig.1. Sources of Air Pollution

Air Quality Index (AQI)

A system for measuring and communicating the air quality in a particular location is called the Air Quality Index (AQI). It offers details on the air's cleanliness and pollution levels as well as any related health risks that the general public may find concerning (K. Kumar and B. P., 2022).

The AQI is calculated based on the concentration levels of several key air pollutants, which include:

1. Ground-level ozone (O_3)
2. Particulate matter (PM2.5 and PM10)
3. Carbon monoxide (CO)
4. Sulphur dioxide (SO_2)
5. Nitrogen dioxide (NO_2)

Particulate matter, which includes PM2.5 and PM10, are tiny airborne particles that can cause respiratory issues if they are inhaled. Ozone at Ground Level: Ground-level ozone is a dangerous pollutant that can lead to respiratory problems and other health problems. One gas created by combustion activities that might irritate the respiratory system is nitrogen dioxide (NO_2).

Sulphur Dioxide (SO_2): A gas that can cause respiratory problems and help cause acid rain. Carbon Monoxide (CO): An odourless, colourless gas that disrupts the body's oxygen transport system.

Machine Learning

A branch of artificial intelligence (AI) called machine learning (ML) allows computers and other machines to learn similarly to people, carry out activities on their own, and enhance their performance and accuracy through experience and exposure to fresh data (Sekeroglu et al., 2022).

1. Decision Process: Machine learning techniques are typically employed for prediction or categorization.

2. An error function evaluates model predictions. If there are known examples, an error function can compare them to determine the model's accuracy.

3. Model Optimization: Adjust weights to improve model fit to training data points and reduce disparity between known and estimated results.

Machine Learning Workflow

- Data Collection: Gathering relevant data for the problem.

- Data Preprocessing: Cleaning and transforming data into a usable format (e.g., handling missing values, normalization, feature engineering).
- Model Training: Using a machine learning algorithm to learn from the data.
- Evaluation: Assessing the model's performance using evaluation metrics such as accuracy, precision, recall, F1-score, etc.
- Deployment: Deploying the model in real-world applications.
- Monitoring & Maintenance: Continuously monitoring model performance and retraining it as necessary.

For this study data on air pollution has been made available on request from CPCB, India for Western UP cities. India, the dataset was collected between Jan 2024 to Dec 2024, for the prediction of AQI, we create Four Machine Learning Models; Linear Regression, Random Forest, Decision Tree & XG Boost (Liu et al., 2019). To assess the performance of both the machine learning models, the comparative study was done by comparing value of evaluation metrics such as MAE, RMSE & R^2. XG Boost model is one of the most widely accepted Ml models for its prediction accuracy. XG Boost Handles large data set.

This study also aims to conduct a comparative analysis of the used model based on their performance metrics including MAE, RMSE & R^2.

II. MATERIAL AND METHODS

Study Area

For this research, the proposed study area is Western Uttar Pradesh Cities shows in Fig.2 , Agra, Bagphat, Vrindavan, Meerut, Ghaziabad, Kasganj, Barielly. Agra is a city of banks of the Yamuna River. Thus, city is densely populated which increases its susceptibility to air pollution and more industrialization and urbanization add even more to this (Mahesh et al., 2022). The main sources of air pollution in the cities are vehicular emissions, industrial emissions etc. which produces PM_{10}, $PM_{2.5}$, SO_2, NO_2, NH_3, and O_3 in a very high concentration (Rahman et al., 2024).

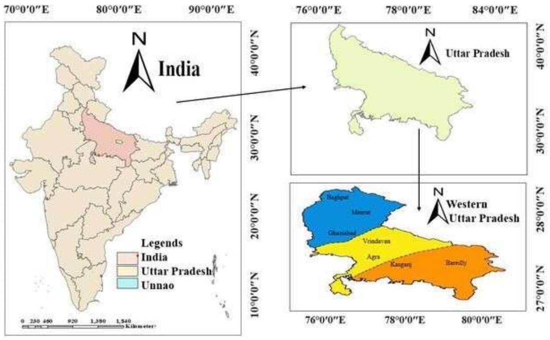


Fig.2. Western Uttar Pradesh Cities

III. METHODOLOGY

The data collection procedure shown in Fig.3 for Western Uttar Pradesh Cities has been completed. The first stage in data preprocessing was to properly inspect the dataset to verify there were no null values. Following this, features were chosen, and correlation was assessed using exploratory data analysis (Kumar & Pande, 2023). The dataset was then divided into training and testing datasets, with 70% of the data assigned to the training set to train the model and the remaining 30% to the testing set. This step consists of model selection, hyperparameter tuning, model training, and cross-validation (Ameer et al., 2019). After training a model on a training dataset, it was applied to the test dataset to make predictions (Méndez et al., 2023).

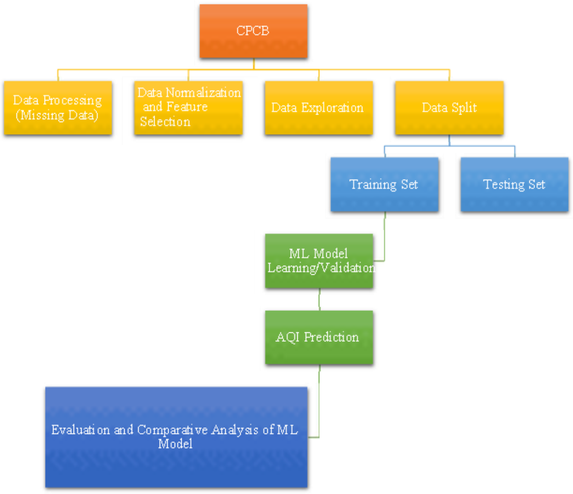

CPCB
Data Processing (Missing Data)
Data Normalization and Feature Selection
Data Exploration
Data Split
Training Set
Testing Set
ML Model Learning/Validation
AQI Prediction
Evaluation and Comparative Analysis of ML Model

Fig.3. Flowchart of Methodology

Parameters Involved in the Study

The CPCB website provided the basic raw data for this investigation. The dataset contains important meteorological metrics like temperature, humidity, wind speed, and precipitation ("Air Pollution and Disasters," 2016). The dataset includes air pollutants like PM10 (particles suspended in the environment with a size of 10um or less), PM2.5 (ultrafine particles and liquid droplets suspended in the atmosphere with a size of 2.5 µm or less), SO2 (emitted by automobiles and chemical industries), and NO2 (emitted in the atmosphere primarily by vehicles). In this investigation, these features were employed to calculate the AQI (Ma et al., 2020a). At the very first we looked for ant null and missing values and addressed them to maintain the dataset integrity (Bekkar et al., 2021). For this study, we determined the AQI for a specified period using the daily mean values for selected pollutants provided by CPCB.

Air quality Index

The state of the air has been assessed using the Air Quality Index (AQI). AQI levels normally fall between 0 to 500, according to the criteria set by the Central Pollution Control Board (CPCB) in India. Serious air pollutants, which can have catastrophic impacts on the environment and human health, are indicated by the highest index value (Ma et al., 2020b). On the other hand, the cleanest air is indicated by the lowest AQI number. The concentration of different air pollutants in the atmosphere, staying within the designated limits for each pollutant, is shown by the minimum AQI values. The study's data on air quality was gathered throughout the course of a typical day (Doreswamy et al., 2020).

The CPCB provides information on the health effects associated with each pollutant alongside its AQI readings. AQI values ranging from 0 to 50 suggest minimal health impacts. Individuals who are sensitive may experience mild respiratory issues when AQI values are between 51 and 100. Those suffering from lung diseases and respiratory problems may find an AQI of 101 to 200 to be difficult (Wu & Lin, 2019). Short-term exposure in the range of 201 to 300 might be uncomfortable for individuals with heart conditions. Prolonged exposure at levels between 301 and 400 could lead to respiratory illnesses. The impact may be more pronounced in individuals with pre-existing lung and heart issues. Even healthy people can face significant health risks when the AQI reaches the severe levels of 401 to 500, particularly those with heart and lung conditions (Lee et al., 2020). The relevance of the AQI levels is displayed in Table 1.

Table 1 AQI Value with its Significance

AQI Range	Significance
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very poor
401-500	Severe

Evaluation and Comparison Criteria

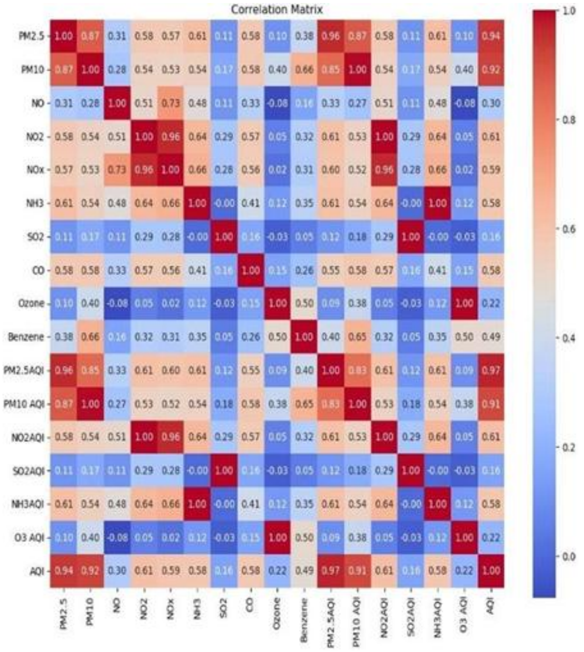
Employing evaluation metrics such as the coefficient of determination (R2 value), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), the final outcomes were analyzed. MSE places more emphasis on significant errors by squaring them before calculating the average (Ma et al., 2019). Consequently, researchers will find it easier to investigate errors in the databases. However, recurring errors have a considerable effect on MSE values, leading to a deterioration in these values due to the repeated mistakes. The MAE was also utilized as a metric to evaluate the accuracy and performance of machine learning regression models. It solely examines the magnitude of differences between the actual and predicted outputs, disregarding the direction of the errors (Bellinger et al., 2017).

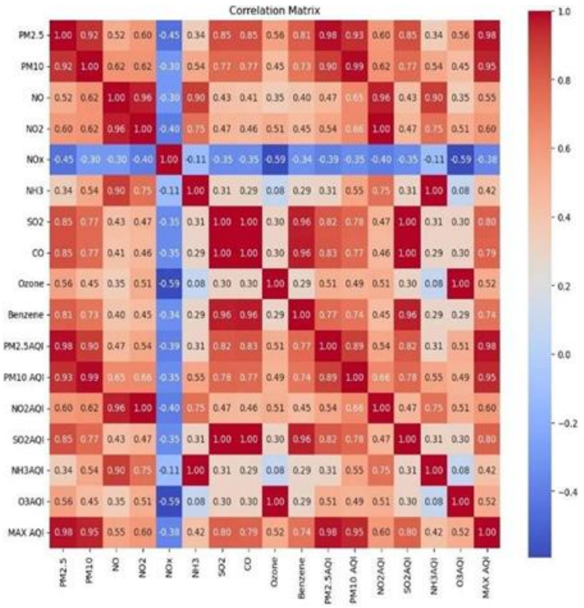
The predictive accuracy of regression models was assessed using the RMSE metric. This metric computes the average of the differences between predicted and actual values, similar to MSE. Furthermore, in contrast to MSE, which squares the errors, RMSE determines the square root of the average of the squared errors. RMSE was deemed beneficial in relation to the actual data, as it offers an estimation of deviation and imposes greater penalties on larger errors. Additionally, it is more straightforward to interpret (C R et al., 2018). The R^2 value closely resembles the MSE, assessing how effectively the outcomes in the dataset are achieved. The R^2 value measures the relationship between the predicted values and the actual results. This provides a basis for evaluating the models, allowing researchers to conduct a more thorough comparative analysis (Zhu et al., 2018).

Machine Learning Algorithms

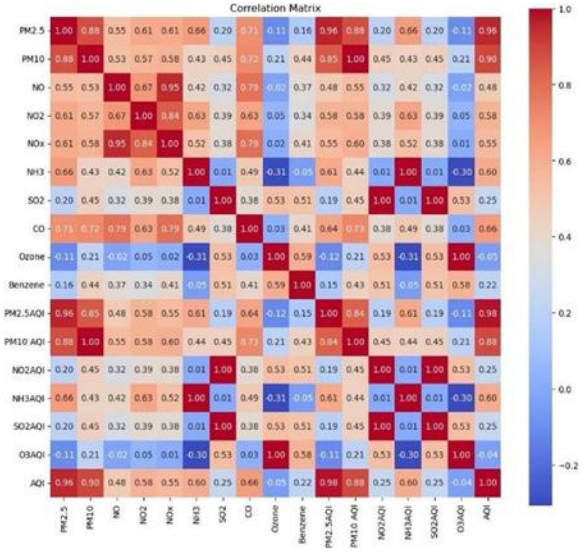
For a small dataset, traditional methods can be more efficient and comprehensible, despite the accuracy and usefulness of ML models in numerous applications (Ayus et al. 2023). In this study, we ensured the completeness of the dataset by loading it from an Excel file and removing any row with missing values. This laid the groundwork for predictive modeling, which also involves separating the dataset into target variables (Y) and independent features (X). To gain insights into the pairwise relationships among variables and to identify the underlying patterns present in the data, correlation heat maps and pair plots were created (Gul & Das, 2023). A bar graph was produced to highlight the most important features that significantly influence air quality. The train_test_split function was utilized to divide the dataset into training and testing subsets. The dataset was partitioned with a ratio of 70:30 for evaluating the model. This approach ensured that the model could be applied practically while also facilitating an understanding of its performance on the test data (Tien et al., 2022).

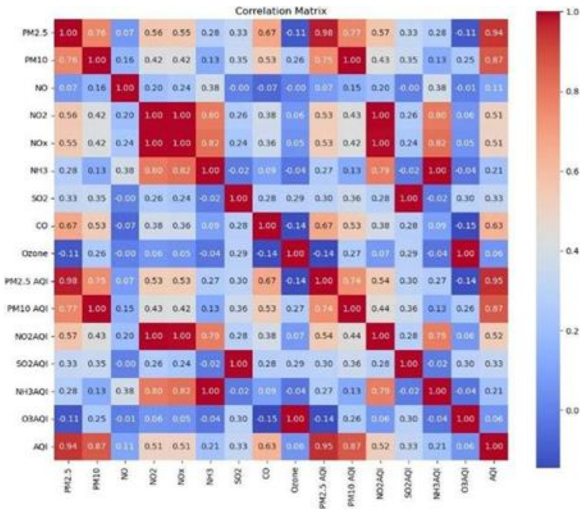
Dataset split and Standardization
There are several reasons for choosing this ratio: to begin with, the 70:30 ratio strikes a balance between model training and evaluation. Achieving this balance is crucial for creating a strong model that can generalize effectively to new, unseen data. Next, a standardization technique was applied to the numerical parameters. This technique transforms the features to yield a mean value of 0 and a standard deviation of 1. This step is vital as it ensures that no particular feature disproportionately impacts the model's fitting procedure. Additionally, it guarantees that all features used in this analysis are on the same scale (Ayus et al., 2023b).

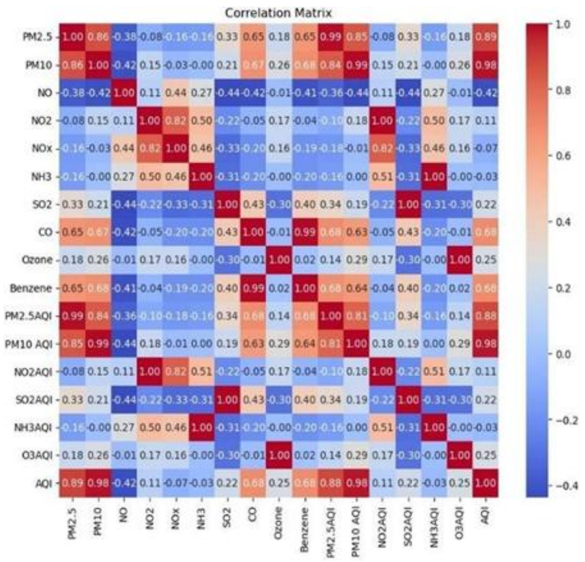




Correlation Matrix
A correlation matrix is a table that shows the correlation coefficients between variables. Each table cell shows the correlation between two variables. A correlation matrix can be used to summarize data and can also be used as input into more complex analysis and as a diagnostic tool. Western rising cities are depicted in the correlation matrix (Kumar Patel et al., 2020).







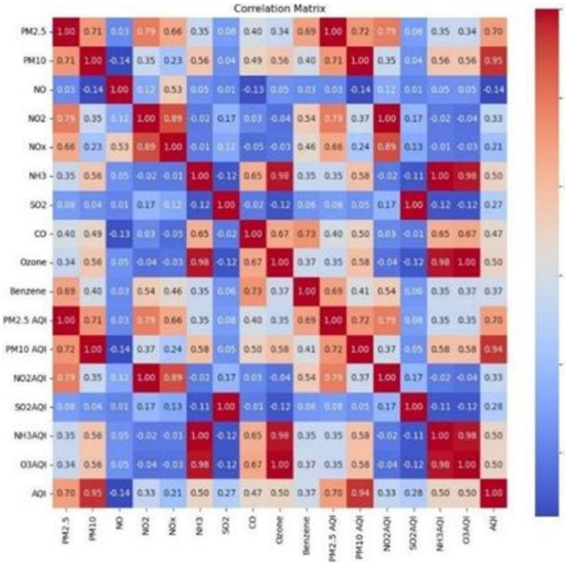


Fig.4. Correlation matrix for all stations (Agra, Ghaziabad, Meerut, Bareilly)

In Fig.4 the dataset was divided, and subsequently, a standardized model was developed. To create a productive and precise model, various training and testing sets were utilized to forecast AQI. The dataset was carefully prepared to guarantee that the model could effectively learn from the training data and accurately apply the knowledge to new, unseen datasets. To ensure the reliability of the predictive model, a precise data split was essential, as an improper data division in a machine learning model could lead to overfitting, underfitting, or inaccurate predictions. Thus, it is vital to thoughtfully choose how to split the dataset. The effectiveness of the machine learning model on a new dataset is assessed to confirm its ability to efficiently and accurately predict outcomes for previously unseen data. In this research, AQI values were designated as the target variable, represented by "y," while all other features were labeled as exploratory variables, denoted by "x." For partitioning the dataset into training and testing sets, we employed the train_test_split () function from the sklearn model selection module. We allocated 70% of the total dataset for training purposes, 30% to the testing set.

We trained the model using one dataset while reserving another for evaluation purposes. In the following phase, the model underwent training with a dataset that allowed it to identify trends and patterns. Due to effective training on the initial dataset, the model's capability to discern the underlying trends and relationships within the data improved. Since the training and testing datasets were entirely distinct, we conducted an analysis of the model's performance to assess its accuracy on an unseen dataset. This strategy improved the potential for reliability and reproducibility of the predictive model.

Evaluation Criteria

The accuracy and dependability of the machine learning models used to predict the AQI were assessed in this work using a variety of performance evaluators. Mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R2) are among the evaluator metrics (Asadi et al., 2023; Moharana et al., 2022; Sahoo et al., 2024).

Mean Absolute Error (MAE)

MAE quantifies the average size of errors between predicted and actual values. It is determined by averaging the absolute differences between these values, providing a clear interpretation, as it reflects the average error in the same units as the original data.

$$MAE = \frac{1}{n}\sum_{i\equiv1}^{n}[yi - \hat{y}_i]$$

Mean Squared Error (MSE)

MSE determines the average of the squared discrepancies between the predicted values and the actual values. Due to its emphasis on larger errors, MSE is influenced more by outliers than MAE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}i)^2$$

Root Mean Squared Error (RMSE)

RMSE is obtained by taking the square root of the MSE. It expresses the error in the same units as the original data, making it easier to understand. RMSE is useful for assessing and comparing errors across various models or datasets.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}i)^2}$$

R-squared (R^2)

R-squared (R^2) is a statistical metric that shows the proportion of variance in the dependent variable explained by the independent variables. It illustrates how well the model corresponds to the data. An R^2 value of 1 signifies a flawless prediction, while a value of 0 means the model does not account for any variance.

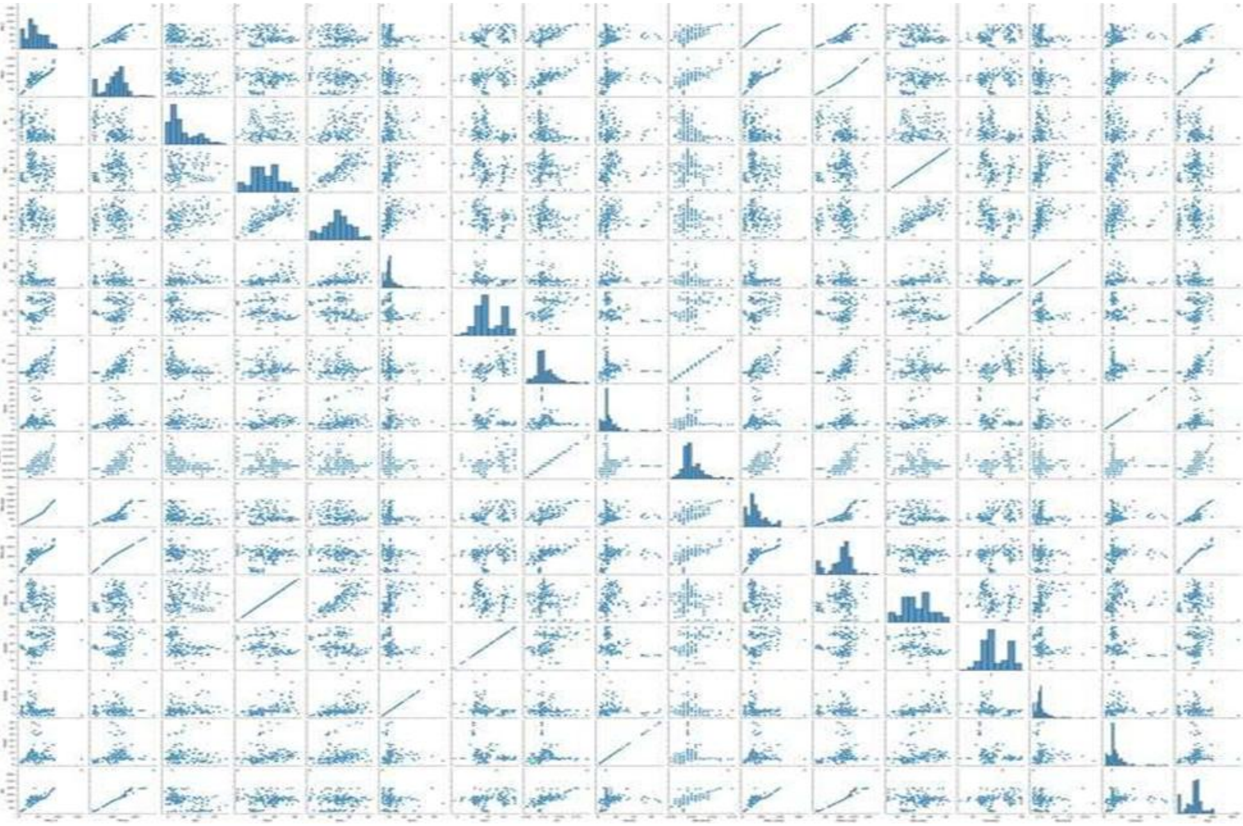
$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}i)^2}$$

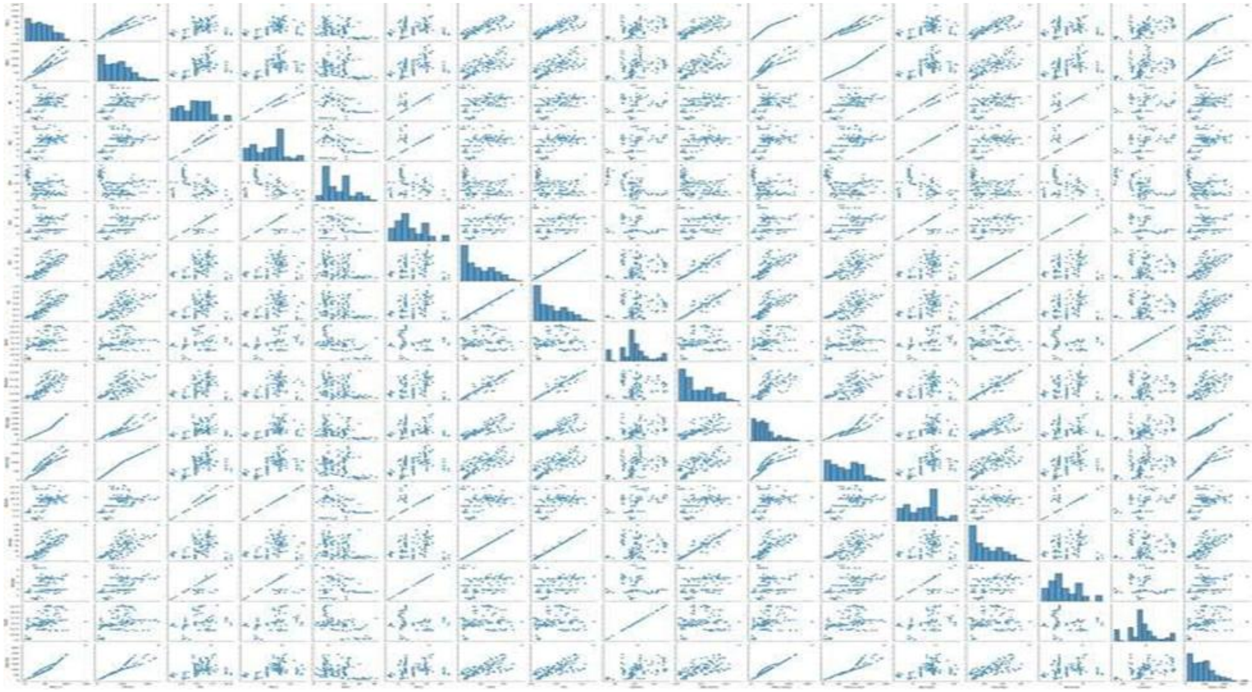
IV. RESULT AND DISCUSSION

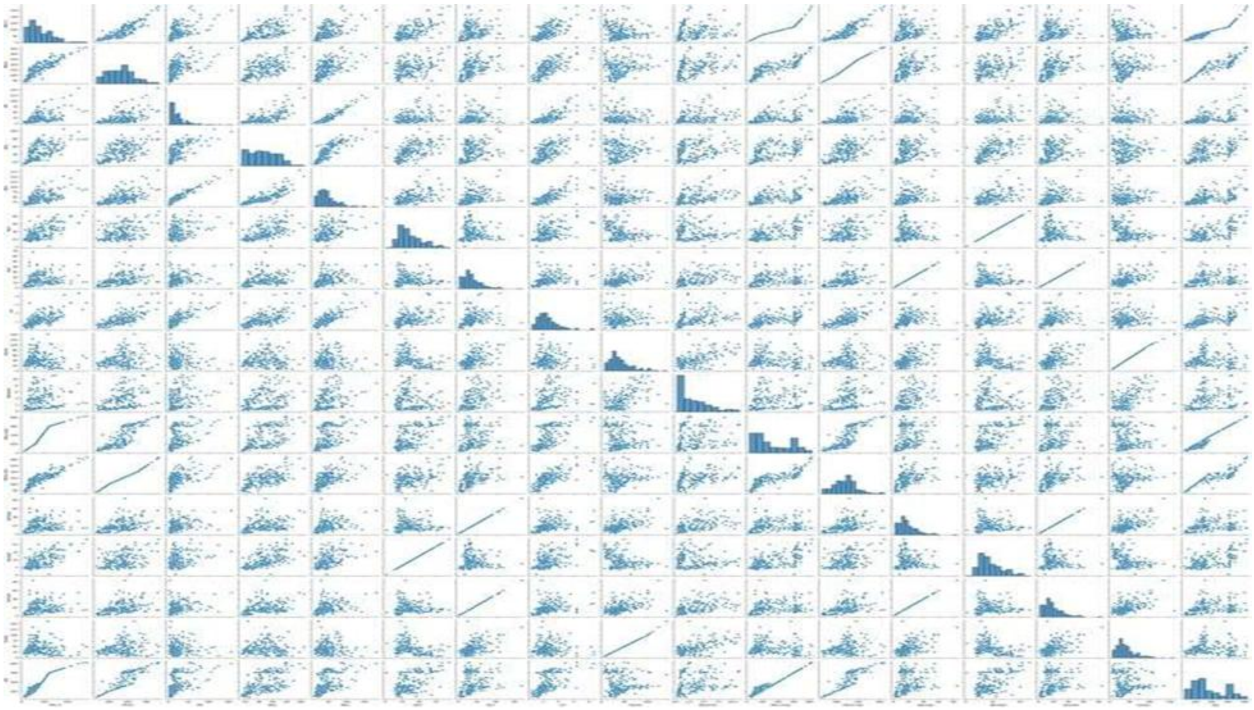
Data exploration and visualization

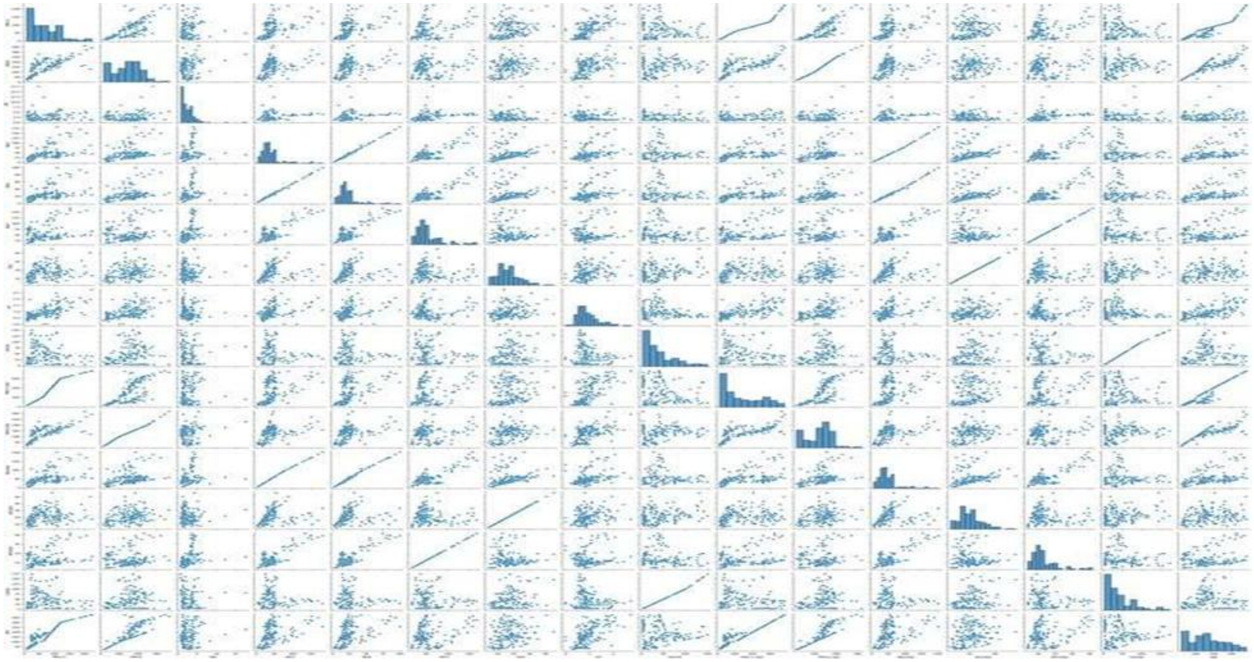
Data exploration and visualization techniques are essential for creating an accurate and dependable model for AQI prediction and for offering valuable insights into the variables influencing air quality. The Python seaborn library's Pairplot feature provides a thorough display of scatterplots and histograms for every possible combination of variables in the dataset. A pairplot's main purpose is to compare one variable to the others.

A kind of representation Fig.5 that aids in examining the connections between several numerical variables in a dataset is a pair plot, often known as a pairplot. It is frequently utilised to find patterns, correlations, and variable distributions in exploratory data analysis (EDA). The scatterplots in the upper and lower triangles show the relationships between pairs of variables, while the histograms show the distribution of single variables along the grid's diagonal. Analysing trends, patterns, and connections between several parameters is made easier by pairplots. The correlation heatmap improves visualization by showing weaker correlations with lighter colours and stronger correlations with deeper colours. correlations between variables, whether positive or negative. Clusters have been identified (helpful for classification issues) identifying outliers or anomalies.









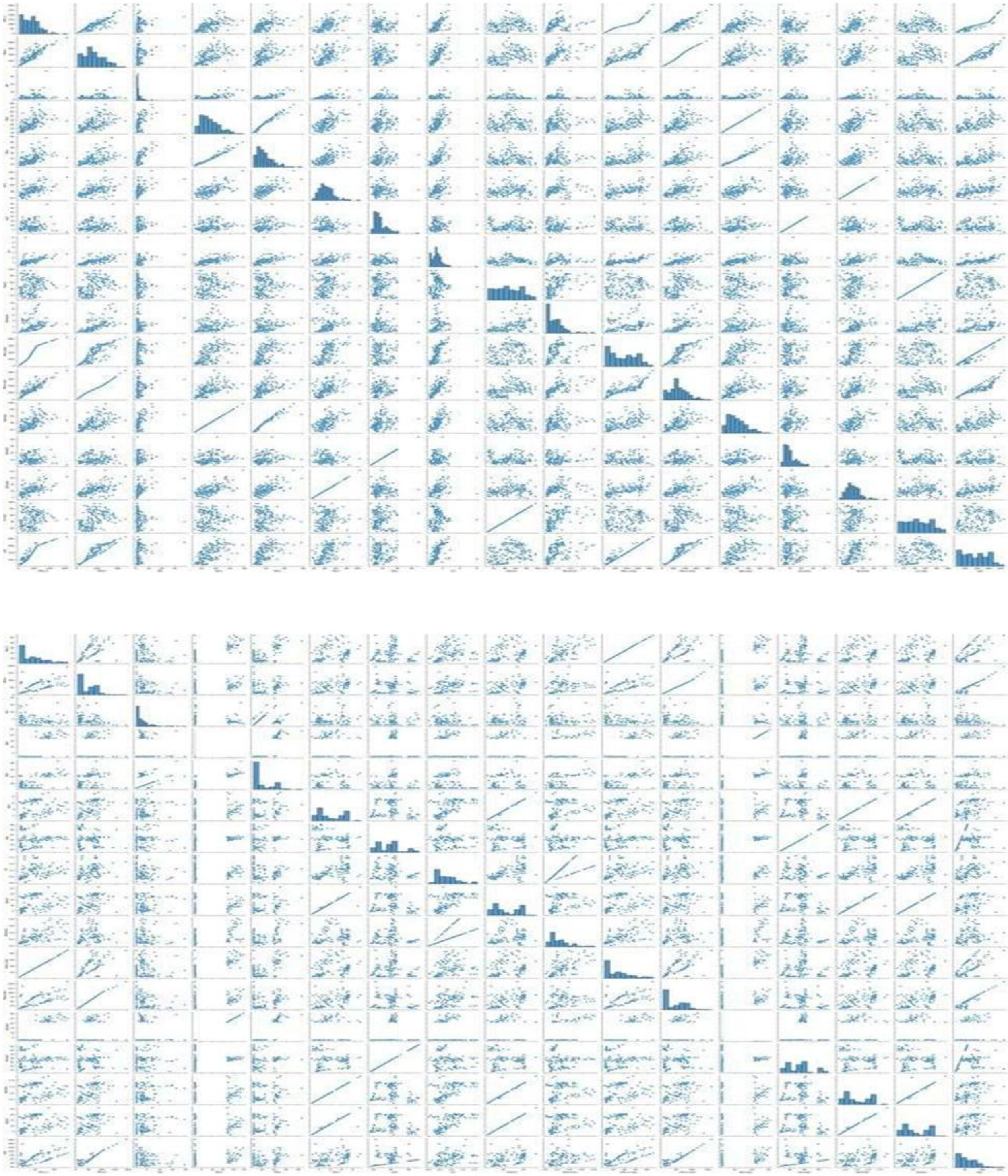


Fig.5. Pairplot for all stations (Agra, Ghaziabad, Meerut, Bareilly)

One of the very important stages in evaluating a regression model's performance is residual assessment, additionally referred to as residual analysis in Fig.6. the discrepancies between actual and predicted values generated by the models are known as residuals. To identify whether the regression algorithm satisfies the underlying assumption and to determine any patterns or trends in the model's mistakes, the residual analysis includes exploring these residuals.

A tree model is a kind of prediction model that bases its judgements on a hierarchical structure in Fig.7. It is frequently employed for jobs involving both regression and classification. Decision trees, random forests, and gradient boosting trees are the most widely used tree-based models. Every internal node stands for a feature-based decision. Every branch symbolises a decision's result. Every leaf node denotes a final forecast.

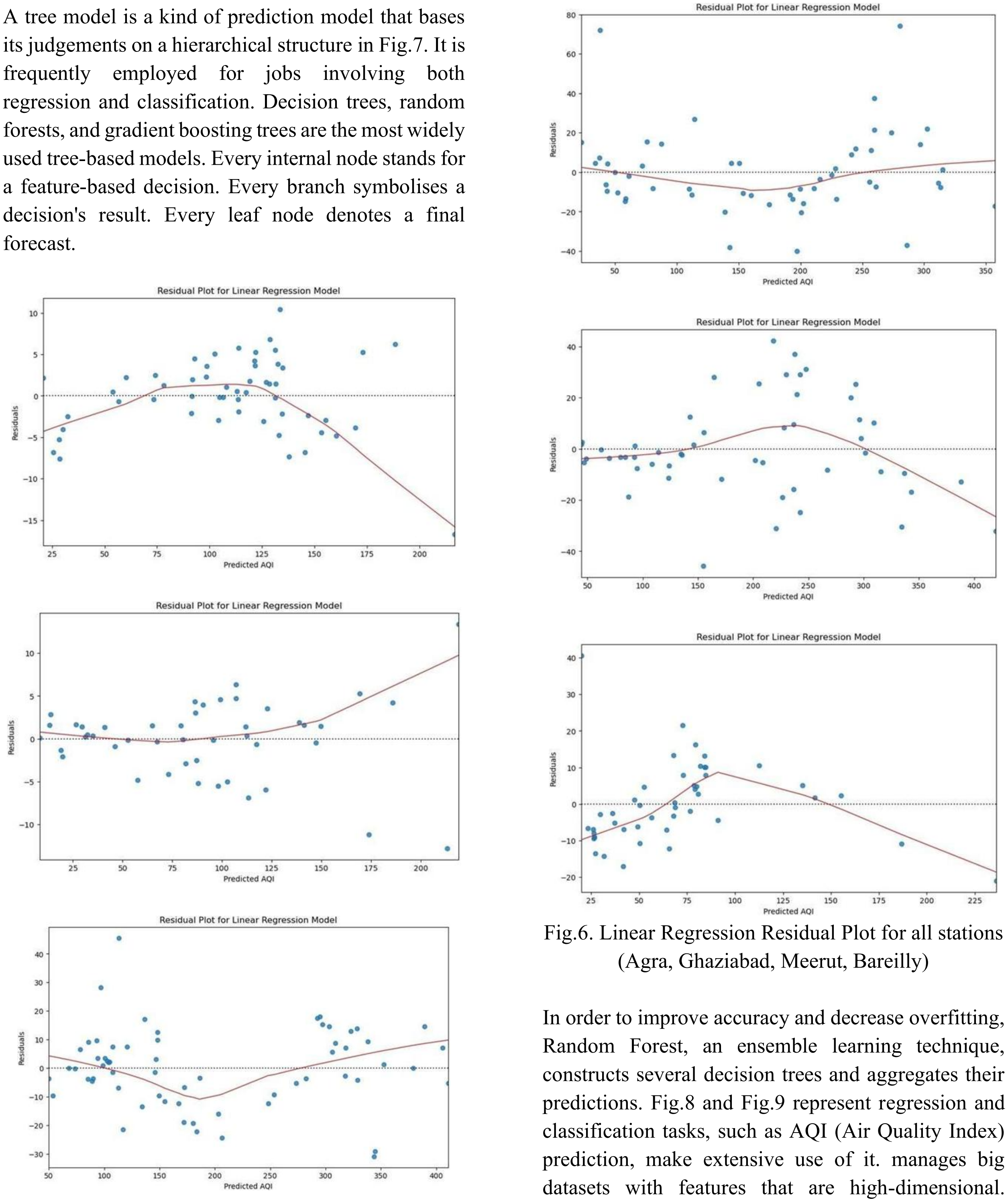
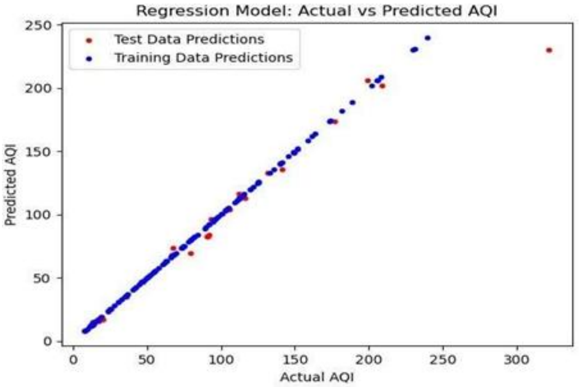


Fig.6. Linear Regression Residual Plot for all stations (Agra, Ghaziabad, Meerut, Bareilly)

In order to improve accuracy and decrease overfitting, Random Forest, an ensemble learning technique, constructs several decision trees and aggregates their predictions. Fig.8 and Fig.9 represent regression and classification tasks, such as AQI (Air Quality Index) prediction, make extensive use of it. manages big datasets with features that are high-dimensional. lessens overfitting if compared to a decision tree alone. perform well with categorical variables and missing data.

A sophisticated boosting method called XGBoost (Extreme Gradient Boosting) creates trees one after the other, enhancing predictions at each stage. Fig. 10 and Fig.11 show contrast to Random Forest, which trains trees separately, XGBoost trains trees sequentially, with each new tree fixing the mistakes of the one before it. Gradient Boosting: Iteratively, trees are constructed with an emphasis on minimising prior errors. Regularisation: Uses L1 (Lasso) and L2 (Ridge) penalties to stop overfitting. Compared to other boosting techniques, parallel processing is faster. Managing Missing Values: Learns the best values for missing data automatically.



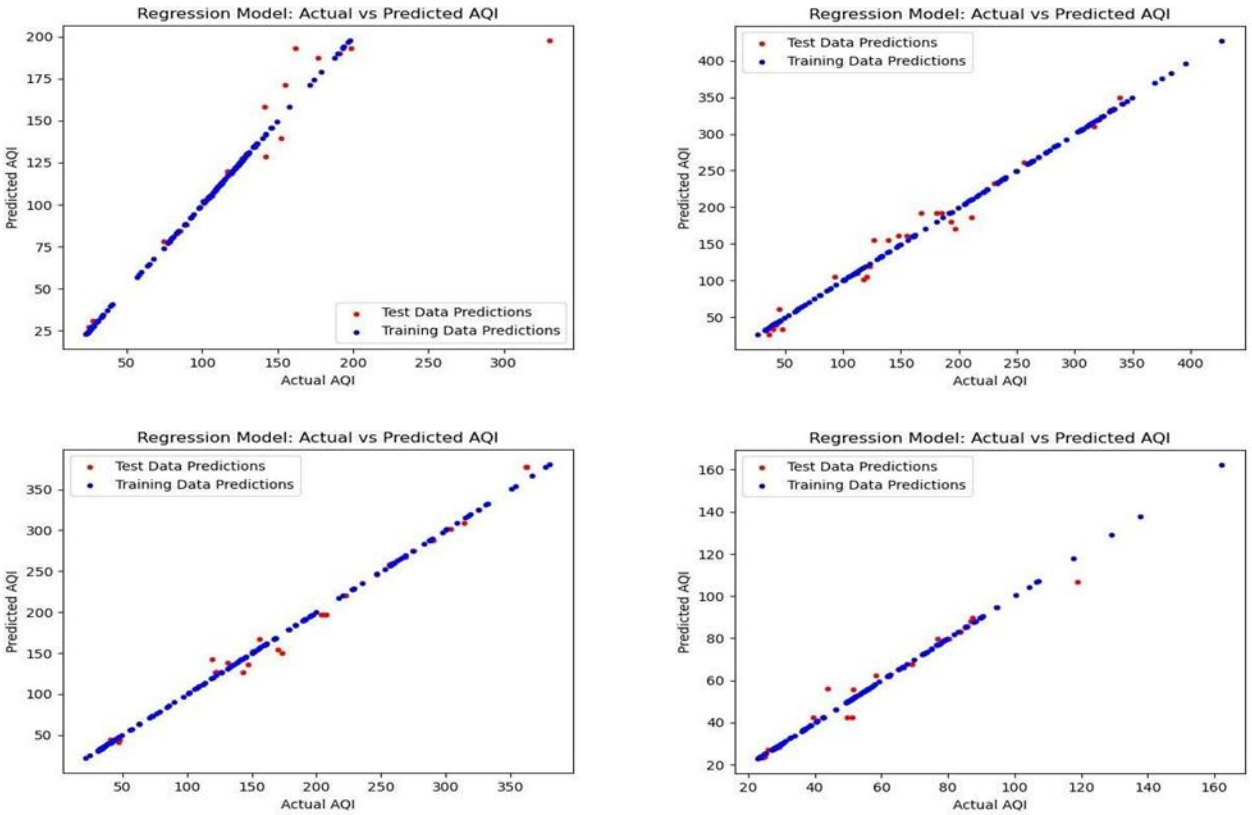
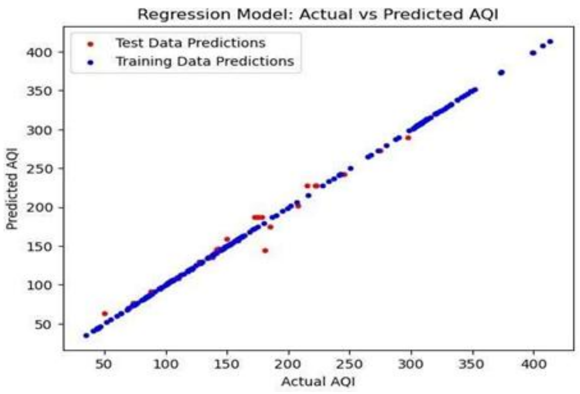
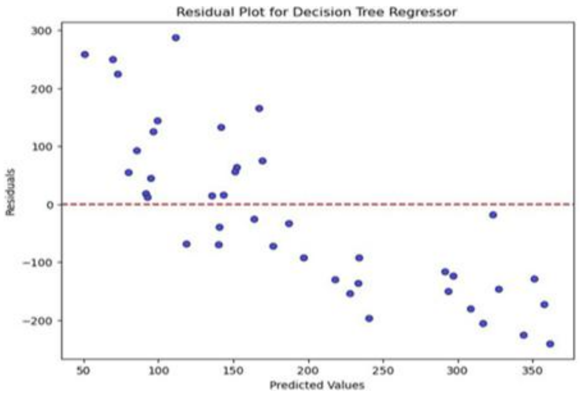
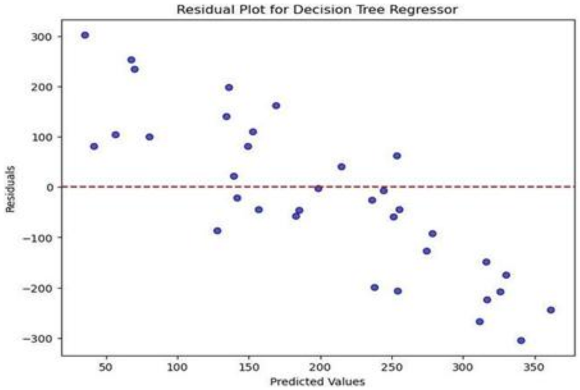
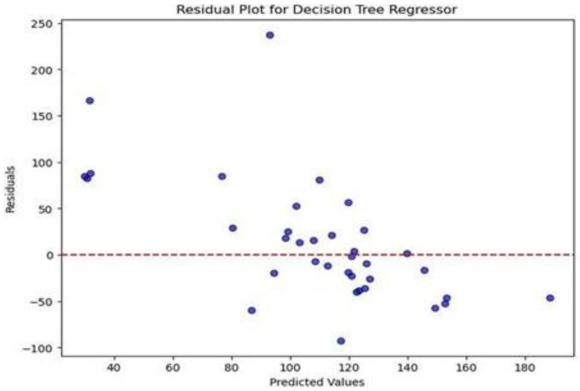


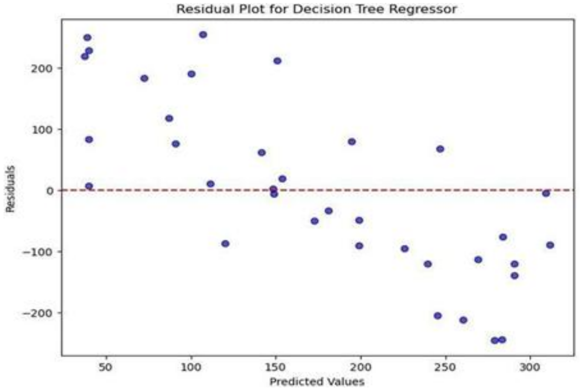
Fig.7. Decision Tree model for all stations (Agra, Ghaziabad, Meerut, Bareilly)

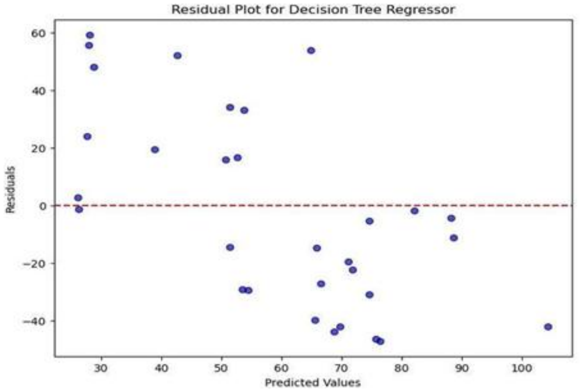












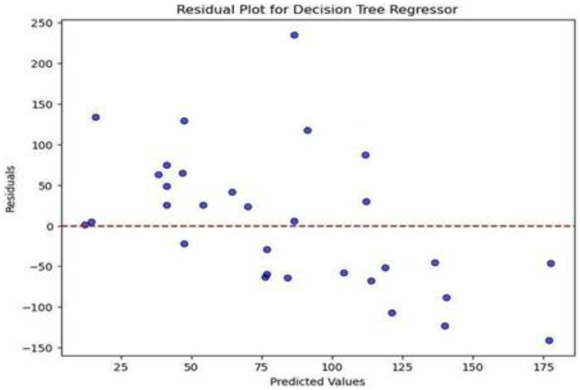
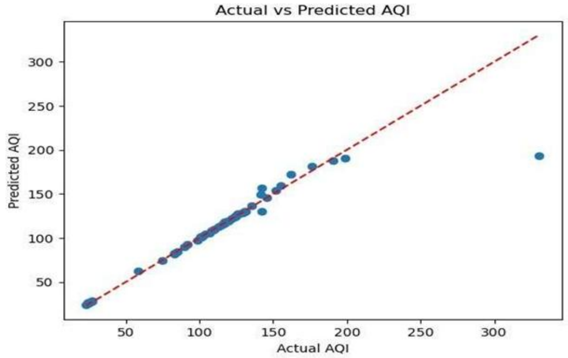
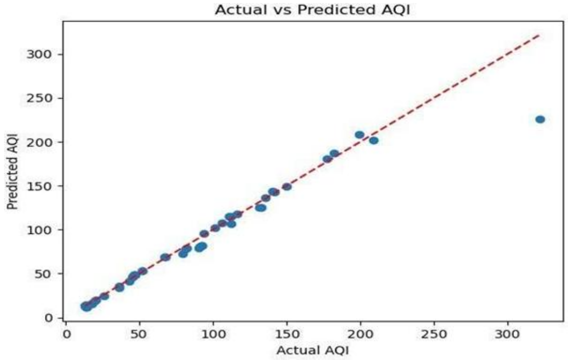
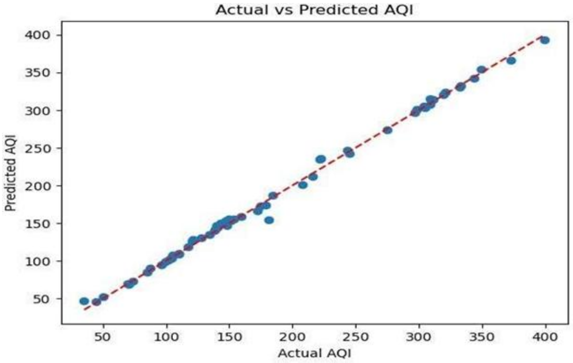
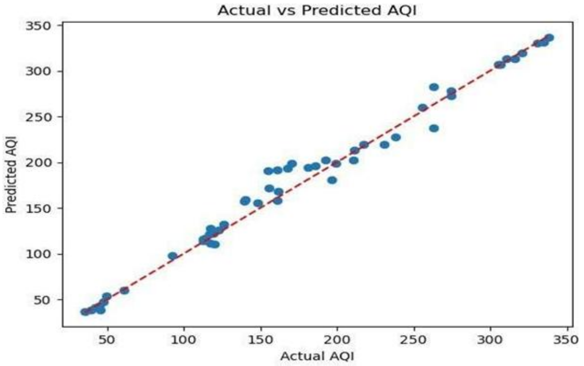


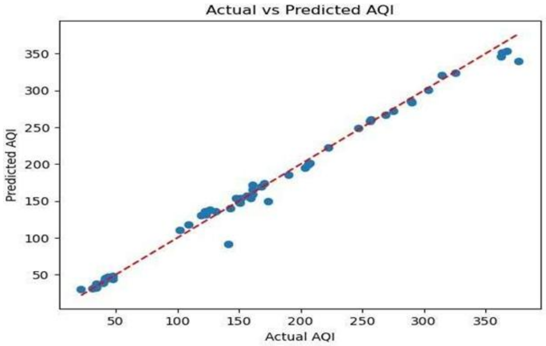
Fig.8. Decision Tree Residual Plot for all stations (Agra, Ghaziabad, Meerut, Bareilly)











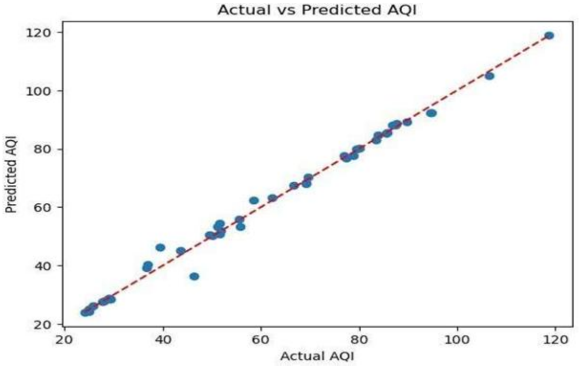
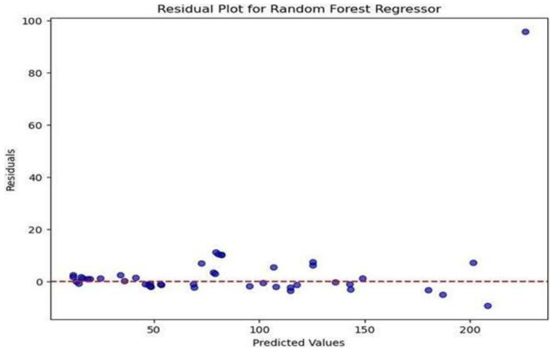
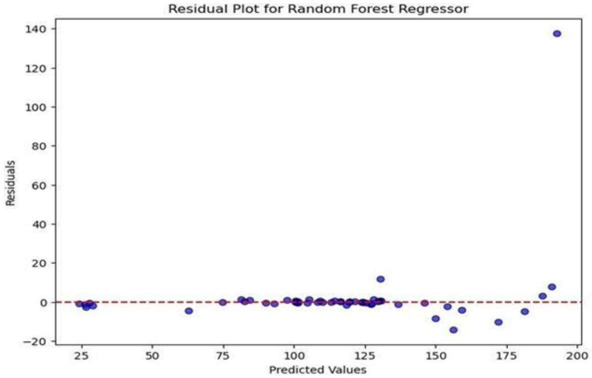
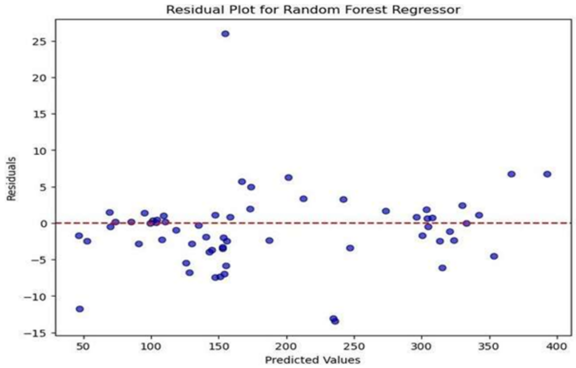
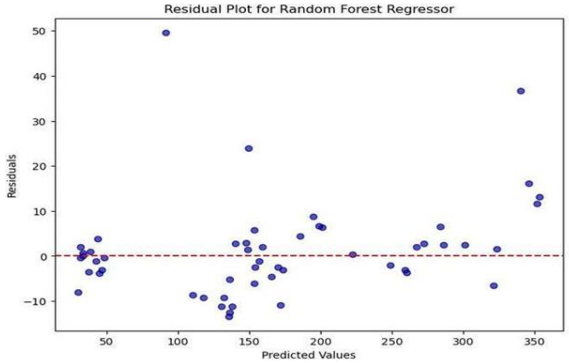


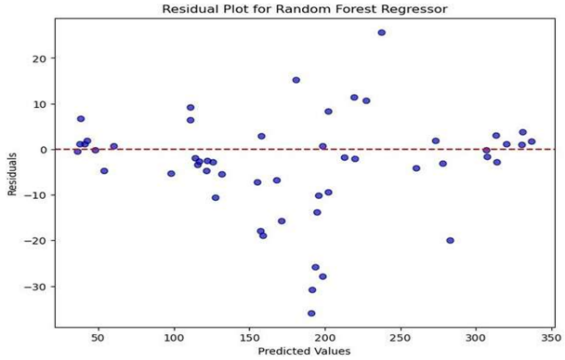
Fig.9. Random Forest for all stations (Agra, Ghaziabad, Meerut, Bareilly)











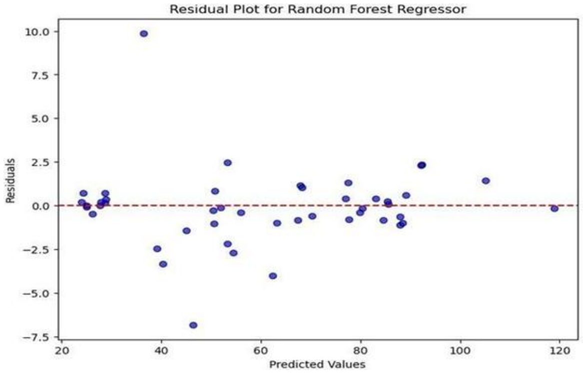
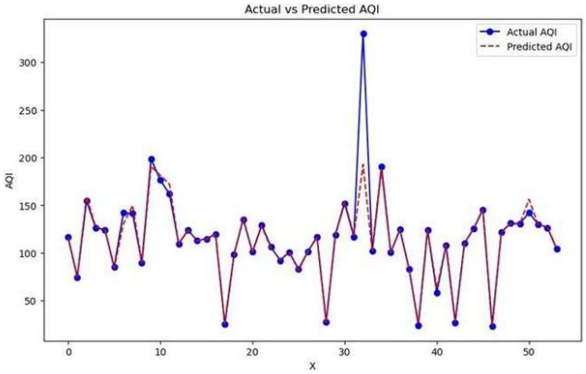
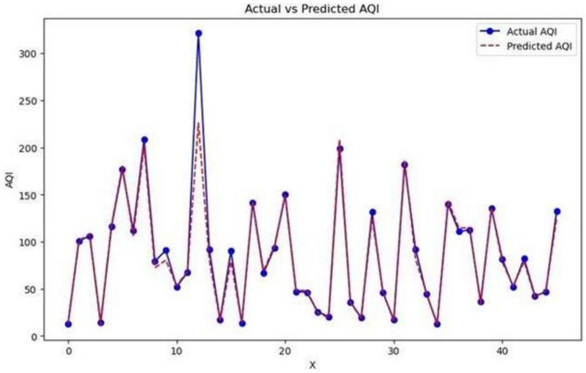
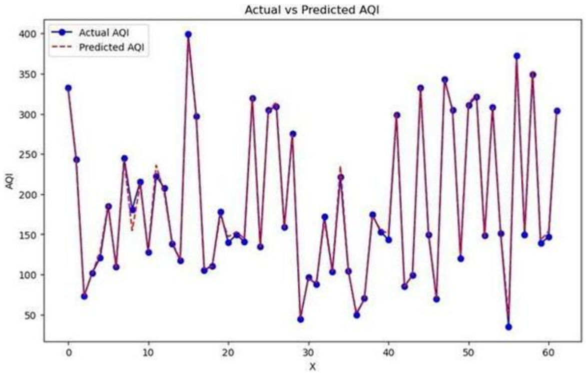
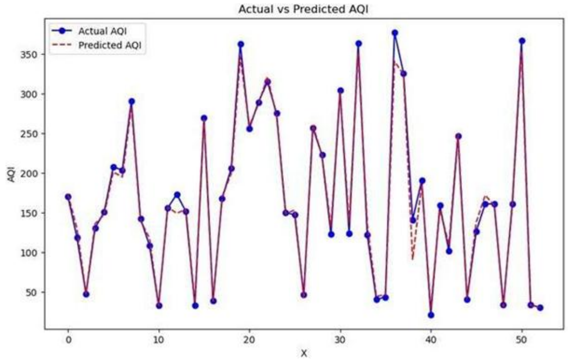


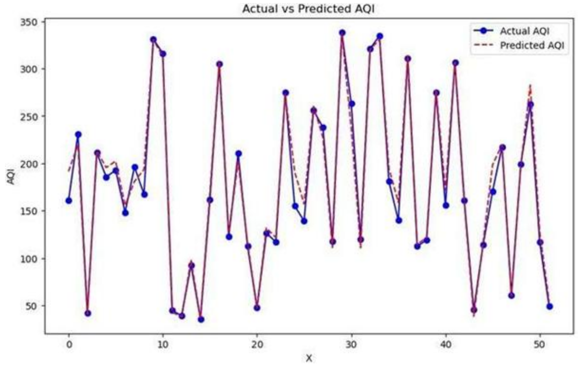
Fig.10. Random Forest on Residual Plot for all stations (Agra, Ghaziabad, Meerut, Bareilly)











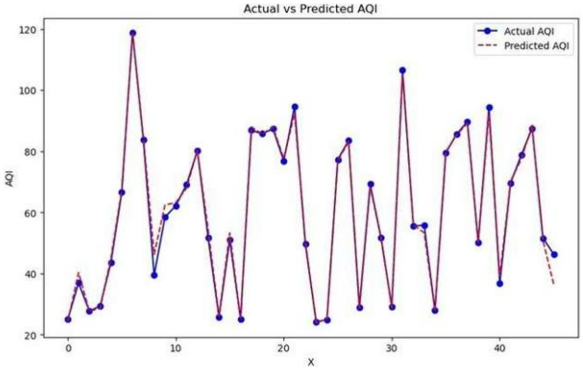


Fig.11. XG Boost for all stations (Agra, Ghaziabad, Meerut, Bareilly)

Table 2. Evaluations of Stations

Monitoring Stations	Linear Regression Model				K Nearest N Model			
	M.S. E	*M.A. E*	*R.M.S. E*	*R^2*	*M.S. E*	*M.A. E*	*R.M.S. E*	*R^2*
Agra (Shastri)	*26.6*	*3.6*	*5.15*	*0.983*	*213.51*	*10.64*	*14.61*	*0.876*
Agra (Rohta)	*20.45*	*3.13*	*4.522*	*0.992*	*104.06*	*8.19*	*10.2*	*0.949*
Ghaziabad	*188.95*	*10.29*	*13.74*	*0.982*	*197.51*	*9.37*	*12.2*	*0.991*
Meerut (Ganga)	*470.55*	*14.98*	*21.69*	*0.951*	*797.59*	*20.6*	*28.24*	*0.921*
Meerut (JaiBhim)	*343.92*	*13.91*	*18.54*	*0.966*	*558.65*	*19.43*	*23.63*	*0.943*
Bareilly	*300.14*	*9.62*	*17.32*	*0.712*	*220.54*	*7.62*	*14.84*	*0.708*

Monitoring Stations	*Decision Tree Model*				*Random Forest Model*				*XG Boost Model*			
	M.S.E	*M.A. E*	*R.M.S. E*	*R^2*	*M.S.E*	*M.A. E*	*R.M.S. E*	*R^2*	*M.S.E*	*M.A. E*	*R.M.S. E*	*R^2*
Agra (Shastri)	*440.02*	*9.13*	*12.12*	*0.782*	*363.36*	*4.35*	*14.02*	*0.841*	*363.37*	*4.35*	*19.06*	*0.841*
Agra (Rohta)	*690.7*	*8.21*	*11.1*	*0.939*	*218.89*	*5.2*	*10.21*	*0.944*	*218.89*	*5.2*	*14.79*	*0.944*
Ghaziabad	*616.96*	*18.45*	*24.82*	*0.945*	*29.93*	*3.52*	*14.61*	*0.996*	*29.93*	*3.52*	*14.61*	*0.996*
Meerut (Ganga)	*199.27*	*13.02*	*22.34*	*0.991*	*122.81*	*6.89*	*11.01*	*0.987*	*122.81*	*6.89*	*11.01*	*0.974*
Meerut (Jai Bhim)	*242.07*	*16.03*	*11.06*	*0.982*	*138.6*	*7.99*	*11.77*	*0.982*	*139*	*7.95*	*11.01*	*0.982*
Bareilly	*117.61*	*14.75*	*20.76*	*0.975*	*4.89*	*1.29*	*2.21*	*0.992*	*5.02*	*1.29*	*2.23*	*0.992*

In table 2 shows XGBoost machine learning model is used along with other machine learning models. The MAE, MSE and RMSE (29.93, 3.52, and 14.61) in our study of Ghaziabad are lower than Agra (Shastri Station) MAE, MSE, and RMSE (26.6, 3.6, and 5.15). But Linear regression shows.

CONCLUSION

This paper presents a complete investigation of machine learning methods and their potential to generate models for urban settlements. This comparative analysis yielded the following key findings:

(i) In this study, the XGBoost algorithm depicted a much higher value of the coefficient of determine (R^2) as compared to the value obtained from the Linear Regression, KNN, Decision Tree Regressor and Random Forest Regressor.

(ii) When both the models were compared, the R^2 value of 0.9967 for the XGBoost.

(iii) The evaluation measures show that the XGBoost model performs noticeably better than the other models. In contrast to the other model's MAE of 18.148, the XGBoost model significantly decreased the MAE of 3.5282. Furthermore, when compared to the other models, the XGBoost exhibits much lower MSE and RMSE, indicating higher prediction precision and less variation.

(iv) These AQI value has follow:

a) The maximum AQI of Agra city was found in the month of Nov i.e 321.89 and the minimum AQI was found in the month of May i.e 8.07, The average AQI through the year was found to be 106.40.

b) The maximum AQI of Meerut City was found in the month of Nov i.e 427.09 and the minimum AQI was found in the month of Aug i.e 32.78, the average AQI through the year was found to be 187.30.

c) The maximum AQI of Ghaziabad City was found in the month of Jan i.e 413.57 and the minimum

AQI was found in the month of Sep i.e 35.27, then average AQI through the year was found to be 184.54.

d) The maximum AQI of Barielly was found in the month of Feb i.e 118.73 and the minimum AQI was found in the month of July i.e 22.93, then average AQI through the year was found to be 56.724.
e) The maximum AQI of Vrindavan was found in the month of Oct i.e 309.41 and the minimum AQI was found in the month of Oct i.e 30.22, then average AQI through the year was found to be 133.44.

By offering a robust predictive model for the air quality index that can be used in practical applications, the study's conclusions address the issues mentioned in the introduction. Therefore, we can say that the XGBoost machine learning model performed the best in terms of prediction for this dataset.

Acknowledgment
The authors would like to thank the Civil Engineering Department, Institute of Engineering and Technology, Lucknow, and CPCB, India for providing guidance, data, and support required to conduct the study.

REFERENCES

[1] Air pollution and disasters. (2016). Environmental Science and Engineering (Subseries: Environmental Science), 143, 325–343. https://doi.org/10.1007/978-3-319-21596-9_8

[2] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. IEEE Access, 7, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

[3] Ayus, I., Natarajan, N., & Gupta, D. (2023a). Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China. Asian Journal of Atmospheric Environment, 17(1). https://doi.org/10.1007/s44273-023-00005-w

[4] Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00548-1

[5] Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. In BMC Public Health (Vol. 17, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12889-017-4914-3

[6] C R, A., Deshmukh, C. R., D K, N., Gandhi, P., & astu, V. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology, 59(4), 204–207. https://doi.org/10.14445/22315381/IJETT-V59P238

[7] Doreswamy, Harishkumar, K. S., Km, Y., & Gad, I. (2020). Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Procedia Computer Science, 171, 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[8] Gul, H., & Das, B. K. (2023). The Impacts of Air Pollution on Human Health and Well-Being: A Comprehensive Review. Journal of Environmental Impact and Management Policy, 36, 1–11. https://doi.org/10.55529/jeimp.36.1.11

[9] 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, 2023, 1–26. https://doi.org/10.1155/2023/4916267

[10] Isaev, E., Ajikeev, B., Shamyrkanov, U., Kalnur, K. U., Maisalbek, K., & Sidle, R. C. (2022). Impact of Climate Change and Air Pollution Forecasting Using Machine Learning Techniques in Bishkek. Aerosol and Air Quality Research, 22(3). https://doi.org/10.4209/aaqr.210336

[11] Janarthanan, R., Partheeban, P., Somasundaram, K., & Navin Elamparithi, P. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67. https://doi.org/10.1016/j.scs.2021.102720

[12] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333–5348. https://doi.org/10.1007/s13762-022-04241-5

[13] Kumar Patel, P., Kumar Singh, H., & Hrishikesh Kumar Singh, E. (2260). A MACHINE LEARNING-BASED APPROACH TO PREDICT GORAKHPUR CITY'S AQI: A CRITICAL STUDY *Corresponding Author, 14.

[14] Lee, M., Lin, L., Chen, C. Y., Tsao, Y., Yao, T. H., Fei, M. H., & Fang, S. H. (2020). Forecasting Air Quality in Taiwan by Using Machine Learning. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-61151-7

[15] 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

[16] Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences (Switzerland), 9(19). https://doi.org/10.3390/app9194069

[17] Ma, J., Cheng, J. C. P., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmospheric Environment, 214. https://doi.org/10.1016/j.atmosenv.2019.116885

[18] Ma, J., Ding, Y., Cheng, J. C. P., Jiang, F., Tan, Y., Gan, V. J. L., & Wan, Z. (2020a). Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. Journal of Cleaner Production, 244. https://doi.org/10.1016/j.jclepro.2019.118955

[19] Mahesh, T. R., Vinoth Kumar, V., Muthukumaran, V., Shashikala, H. K., Swapna, B., & Guluwadi, S. (2022). Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. Journal of Sensors, 2022. https://doi.org/10.1155/2022/4649510

[20] Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 56(9), 10031–10066. https://doi.org/10.1007/s10462-023-10424-4

[21] Mondal, S., Adhikary, A. S., Dutta, A., Bhardwaj, R., & Dey, S. (2024). Utilizing Machine Learning for air pollution prediction, comprehensive impact assessment, and effective solutions in Kolkata, India. Results in Earth Sciences, 2, 100030. https://doi.org/10.1016/j.rines.2024.100030

[22] Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-54807-1

[23] Rahman, M. M., Nayeem, M. E. H., Ahmed, M. S., Tanha, K. A., Sakib, M. S. A., Uddin, K. M. M., & Babu, H. M. H. (2024). AirNet: predictive machine learning model for air quality forecasting using web interface. Environmental Systems Research, 13(1). https://doi.org/10.1186/s40068-024-00378-z

[24] Ravindiran, G., Rajamanickam, S., Kanagarathinam, K., Hayder, G., Janardhan, G., Arunkumar, P., Arunachalam, S., AlObaid, A. A., Warad, I., & Muniasamy, S. K. (2023). Impact of air pollutants on climate change and prediction of air quality index using machine learning models. Environmental Research, 239. https://doi.org/10.1016/j.envres.2023.117354

[25] Samad, A., Garuda, S., Vogt, U., & Yang, B. (2023). Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations. Atmospheric Environment, 310. https://doi.org/10.1016/j.atmosenv.2023.119987

[26] Sekeroglu, B., Ever, Y. K., Dimililer, K., & Al-Turjman, F. (2022). Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems. Data Intelligence, 4(3), 620–652. https://doi.org/10.1162/dint_a_00155

[27] Tien, P. W., Wei, S., Darkwa, J., Wood, C., & Calautit, J. K. (2022). Machine Learning and Deep Learning Methods for Enhancing Building Energy Efficiency and Indoor Environmental Quality – A Review. In Energy and AI (Vol. 10). Elsevier B.V. https://doi.org/10.1016/j.egyai.2022.100198

[28] Wang, Y., Huang, L., Huang, C., Hu, J., & Wang, M. (2023). High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city. Environment International, 172. https://doi.org/10.1016/j.envint.2023.107752

[29] Wu, Q., & Lin, H. (2019). A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Science of the Total Environment, 683, 808–821. https://doi.org/10.1016/j.scitotenv.2019.05.288

[30] Zhu, D., Cai, C., Yang, T., & Zhou, X. (2018). A machine learning approach for air quality prediction: Model regularization and optimization. Big Data and Cognitive Computing, 2(1), 1–15. https://doi.org/10.3390/bdcc2010005