Journal of Indian Association for Environmental Management

Journal homepage: www.http://op.niscair.res.in/index/php/JIAEM/index

e-ISSN No.: 2582-4228

Evaluation of Predictive Models for Air Quality Index Prediction in an Indian Urban Area

Mandvi[a], Hrishikesh Kumar Singh*[b], Prabhat Kumar Patel[c], Shivam Singh[d]

[a,b,c,d] Department of Civil Engineering, Institute of Engineering and Technology, Lucknow,226021, Uttar Pradesh, India
*Corresponding Er. Hrishikesh Kumar Singh, Email: hrishikeshsingh93@gmail.com

Submitted: 29 July 2024 Revised: 05 September 2024 Accepted: 07 September 2024

***Abstract*: With rapid urbanization, the air quality standards for cities have deteriorated due to increased emissions. This increased addition of pollutants to the atmosphere severely affects city life. To identify ambient air quality in a city, an air quality index (AQI) number is provided by CAAQMS situated in those cities. This study delves deeper into predicting AQI using machine learning-based models. In this study, the primary data was collected from CPCB for Gorakhpur City Uttar Pradesh, India. Particulate matter (PM_{10} and $PM_{2.5}$), SO_2, and NO_2 were considered as the primary AQI pollutant parameters. This study develops a statistical comparative analysis of two machine learning models vis a vis linear regression and Random Forest for predicting this AQI. The evaluation metrics used for validating and evaluating the prediction accuracies of models were MAE, MSE, RMSE, and R^2. Also, it includes statistical metrics such as T-statistics, 95% Confidence intervals, and p-values, for determining the significant difference between the models developed. The value of the R^2 matrix for Random Forest (0.99895) was significantly more than the R^2 value for Linear Regression (0.91848), indicating high accuracy and low variance of Random Forest in predicting AQI. Also, the Random Forest displayed a higher degree of accuracy than the Linear Regression, as indicated by the higher values of MAE, MSE, and RMSE for the latter. Statistically significant differences between Random Forest and Linear Regression were demonstrated by the t-statistics, p-values, and confidence intervals calculated for MAE, MSE, RMSE, and R^2. 95% confidence intervals calculated for all evaluation metrics indicate the higher performance of Random Forest over the Linear Regression model.**

***Keywords*: Air Pollution; Machine Learning models; Air Quality Index; Random Forest; Linear Regression; Air Quality Prediction; New technologies.**

I. INTRODUCTION

One of the primary challenges is air pollution, especially in nations like India where environmental issues constantly coexist with fast development. Just consider for a moment that the daily activities we perform are threatening the air we breathe, which is essential for every life on earth. Air quality has been degraded by rapid population growth, industrial activity, vehicular emissions, and unsustainable agricultural practices. Pollutant emissions from this degradation include $PM_{2.5}$ and PM_{10}, CO_2, SO_2, NO_2, CO, and O_3. Directly generated from sources like industrial facilities and vehicles, these air pollutants severely threaten the health of those who are exposed to them either any individuals or the environment, and worsen the quality of the air (Méndez et al. 2023). Particulate matter and gaseous pollutants are the two most important airborne contaminants.

A fast-growing nation like India in present time confronted with the problems of outdoor as well as indoor air pollution. The research conducted by (Balakrishnan et al. 2019) highlighted that in country like India addressing air pollution is more important and also a big challenge to the government and other stakeholders. In the study conducted by (Bodor et al. 2020; Villanueva et al. 2016), the increase in pollutant concentration in the atmosphere is influenced by various meteorological variables which consist of precipitation, humidity, wind speed, and temperature. Global urbanization and industrialization without a doubt have accelerated to an extreme level which continuously leads to the increase in pollutant concentration in the atmosphere. About 9 million accidental deaths globally are associated with air pollution, thereby constituting as greatest concern to public health (Manisalidis et al. 2020)

The complex relationship between air quality, global warming, and climate change was studied by (Dewan and Lakhani 2022)

along with the air pollutants most likely particulate matter and ozone. These contaminants interfere with incoming solar radiation and increase the greenhouse effect by impacting both short-wave and long-wave radiations. This will also lead to fluctuation in the intensity and frequency of air stagnation events, heat waves, and meteorological phenomena, and their overall incidence. The interconnection between atmospheric pollution and global warming forms a never-ending cycle, where one is the cause and the effect of the other (Ma et al. 2019; Von Schneidemesser et al. 2015).

Monitoring air pollution is very crucial and the air quality index (AQI) is proven to be an important tool for this task and it also assesses the associated health risk. Accuracy in the prediction is the need of the hour for efficient air pollution management it also provides awareness among people and policymakers to accordingly take measures to reduce exposure to the potentially hazardous pollutants.

Methods such as statistical forecasting and potential forecasting techniques are involved in the prediction of air quality. Both statistical and numerical forecasts are used by potential forecast methods to generate a pollution potential index that predicts future air quality. On the other hand, the statistical forecast technique employs conventional statistical models including linear regression, which involves methods like Neural network models and Gaussian process regression approaches (Bai et al. 2018).

Machine learning approaches are the best options for the researcher to obtain accuracy as they provide the most accurate and reliable results in AQI prediction. Commonly used Machine Learning (ML) for air quality prediction are Linear Regression, Decision tree regression, Random Forest XGBoost, etc. The random forest and XGBoost proved to be the best model in the prediction of AQI (Tien et al. 2022). The existing comparison studies fail to conduct a thorough analysis of the model's ability to accurately estimate pollution levels using unprocessed and raw datasets. This research will also explore the potential advantages of machine learning (ML) techniques, including Artificial intelligence (AI) Therefore, to construct an error-free ML model, XG-Boost, Random Forest, linear Regressor, and Lasso Regressor are examined for this study using the Python platform. Most of the research shows that the, ML models created for cities are built for big dataset sizes (more than five years' worth of data are taken into account). Before using a machine learning tool to handle the dataset, it is additionally refined to remove errors, null data, overfitting, etc. (Ameer et al. 2019; Natarajan et al. 2024).

To estimate future air quality, ML algorithms, on the other hand, use previous data and assess the trend, they don't require any chemical or physical mechanism to be used (Janarthanan et al. 2021). Two ML models have been used in this study to estimate air quality since they have provided a useful and increasingly well-liked method for AQI prediction (Liu et al. 2019; Taylan et al. 2021).
For this study data on air pollution has been made available on request from CPCB, India for Gorakhpur City in Uttar Pradesh, India. the dataset was collected between July 2021 to December 2023. for the prediction of AQI, we create two machine learning models: Random Forest and Linear Regression. To assess the performance of both the machine learning models, the comparative study was done by comparing the values of Evaluation metrics such as MAE, MSE, RMSE, and, R^2. The Random Forest model is one of the most widely accepted machine learning models for its prediction accuracy. Random forest is capable of handling large datasets. It integrates numerous decision trees to build a strong and precise ensemble machine learning model to predict AQI accurately. Developing robust and accurate machine learning models to predict the air quality index is the aim of this study. This study also aims to conduct a comparative analysis of the used models based on their performance metrics including mean absolute error, mean squared error, root mean squared error, and coefficient of determination.

II. Materials and Methods

Study Area

For this research, the proposed study area is Gorakhpur city. It is located in the state of Uttar Pradesh on the bank of river Rapti. Geographically Gorakhpur is located about 26.76˚N and 83.37˚E. The study area has a unique terrain that encompasses a metropolitan area of about 3483.8 sq. km. The city is densely populated which increases its susceptibility to air pollution and rapid industrialization and urbanization add even more to this. The main sources of air pollution in the city are vehicular emissions, burning of agricultural waste, etc. which produces PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO_2, and CO in a very high concentration. The CPCB monitoring station is located at the MMMUT, Deoria Road, Singhariya, Kunraghat, Gorakhpur with a latitude and longitude of 26.730136˚N, and 83.433859˚E as shown in Fig. 1.

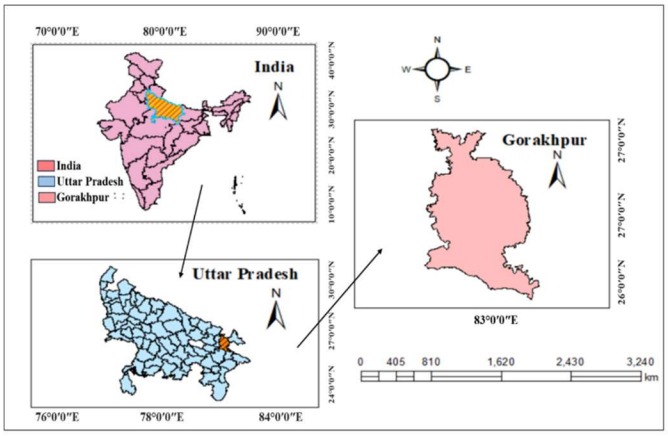

Fig. 1 Map of the Study Area

Methodology

The data collection procedure was finished for Gorakhpur city. The first step in data preprocessing was thoroughly examining the dataset to ensure that there were no null values. Following this, features were selected and correlation evaluation was conducted by performing exploratory data

analysis. The next step was splitting of dataset into training and testing datasets by assigning 80% of the data to the training set to train the model and the remaining 20% of the dataset to the testing set. This stage involves model selection, hyperparameter tuning, model training, and cross-validation. After training a model on a training dataset the model was applied to the test dataset to generate predictions. In the final stage, the model's performance was evaluated as mentioned in Fig. 2.

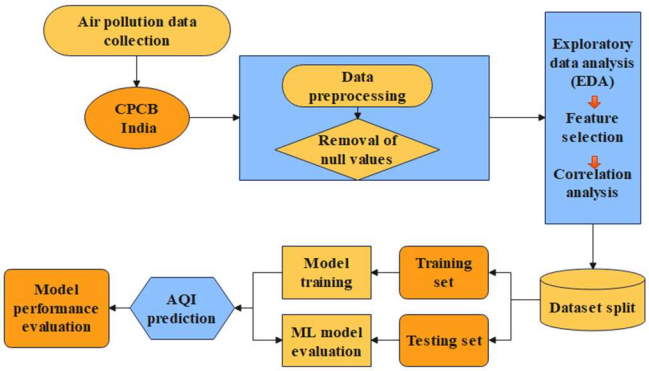

Fig. 2 Flowchart of Methodology

Parameters Involved in the Study

The CPCB website was the source of primary raw data for this study. The dataset includes key meteorological parameters such as temperature, humidity, wind speed, and precipitation. The dataset also contains air pollutants such as PM_{10} (particles suspended in the environment with a size of 10µm or less defined as PM_{10}), $PM_{2.5}$ (it is defined as an ultrafine particle which also includes liquid droplets suspended in the atmosphere having a size of 2.5 µm or less), SO_2 (automobiles and chemical industries are the sources of sulphur dioxide), and NO_2 (nitrogen dioxide is generally emitted in the atmosphere through vehicular emission). In this study, these were the features used for AQI calculation. At the very first we looked for ant null and missing values and addressed them to maintain the dataset integrity. For this study, we determined the AQI for a specified period using the daily mean values for selected pollutants provided by CPCB. The dataset comprises 9 variables with 906 instances. The variables used in this study were PM_{10}, $PM_{2.5}$, SO_2, NO_2, wind speed, temperature, humidity, and precipitation.

Air quality Index

The AQI has been employed to determine the status of air quality. As per the CPCB, India's guideline the AQI values typically fall between 0 and 500. The highest possible index value signifies the presence of acute air pollutants, which means for its devastating effect on both environmental and human health. Similarly, the purest air is defined through the lowest AQI index value. The lowest AQI values represent the concentration of various air contaminants present in the atmosphere within the specified limits set for each pollutant. The air quality data for this study was collected on average over 24 hours.

Along with AQI values, CPCB also provides the health impact associated with each pollutant. AQI values 0-50 show minimum health impact. 51-100 may cause slight breathing issues in sensitive people. 101-200 could make it difficult for people with breathing issues and lung disease. 201-300 may cause discomfort for individuals with heart disease with short-term exposure. 301-400 prolonged exposers may result in respiratory disease. People already suffering from heart and lung disease may experience a more pronounced effect. 401-500 is the severe level of AQI and may have major health impacts on those with lung and cardiovascular diseases even in healthy people. Table 1 shows the significance of the AQI values.

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

Data Preprocessing

The most significant prerequisite for efficient visualization and the creation of ML models is ensuring the data quality. Data preprocessing is necessary for reducing noise in the data which enhances the processing speed and ML methods' capability to generalize (Kumar and Pande 2023a). The raw dataset contained a total of 914 instances in the raw dataset. After removing the null or missing values only 906 instances were left with 9 characteristics. The consistency of the dataset was ensured after removing the missing values. Data integrity ensures that all instances and parameters give reliable values.

Evaluation and Comparison Criteria

The final results have been compared by utilizing the evaluation metrics which were the coefficient of determination (R^2 value), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). MSE implies squaring the errors before averaging them, providing more importance to the significant mistakes. Thereby this makes it simpler for researchers to investigate mistakes in the datasets. Despite this, the repetition of errors has a significant influence on MSE values, with recurring mistakes leading to worsening the MSE values. The MAE was another matrix employed to evaluate the accuracy and efficiency of ML regression models. MAE excludes the direction of the mistakes and only analyses the magnitude of the differences between the actual and predicted outputs. It is believed that more stable and consistent results can be produced with MAE. (Sekeroglu et al. 2022).

RMSE matrix was employed to assess the prediction accuracy of regression models. Similar to MSE, it averages the magnitude of the errors between predicted and actual values. Also, RMSE determines the square root of the average squared errors, unlike MSE, which squares the mistakes. RMSE was proven helpful as it penalizes greater errors with greater severity and provides an estimation of deviation, proving simple to interpret compared to the actual data. The calculation of these metrics has been showed in annexure 1.

The R^2 value is very similar to the MSE. It evaluates the extent to which the results obtained in the dataset perform well. R^2 value gauges the correlation between predicted and actual values. This furnished scale assessment findings for the models, which enables scholars to perform a more robust comparative analysis.

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 established the foundation for predictive modeling, which also includes the division of the dataset into target variables (Y) and independent features (X). For a better understanding of the pairwise relationship between variables and to recognize the underlying patterns in the data, correlation heatmap and pair plots visualization were generated. We implemented the Extra Trees Regressor, which formulates a robust model that can evaluate feature importances, to determine the most important in the prediction of AQI. A bar graph plot has been generated to show the most critical features that have a significant impact on air quality. The train_test_split function was used for splitting the dataset into training and testing sets. The dataset was divided into a ratio of 80:20 for model evaluation. This strategy ensured the model's practical applicability while simultaneously making it easier to understand how well it performed on test data. Two regression models were used for prediction: Random Forest, and Linear Regression.

Random Forest Regression

A supervised learning method known as Random Forest (RF) involves building ensemble decision trees and combining their predictions to generate the outcome (Wang et al. 2009). individual decision trees have been built using randomly selected variables and training data in RF regression (Ravindiran et al. 2023). Generally, the implementation of RF regression requires importing the important modules from the scikit-learn library like RandomForestRegressor used to generate the random forest module. GridSearchCV and KFold classes have been used for hyperparameter optimization and cross-validation, and the best model was executed based on its performance. Hyperparameters have been defined as a grid. Hyperparameters governed the learning process in RF. The different values for parameters are specified through grid sets. These involve the various trees (n_estimators), the max_depth which was the maximum depth of trees, and the minimum number of samples needed to split an internal node that was min_sample_split, among others. For prediction, the test set was used by utilizing the learned RF model. The performance of the model was then assessed using multiple metrics including MAE, MSE, RMSE, and, R^2 by comparing predicted values with the targeted values. The pickle module was used to serialize the trained RF model and stored on disk for use afterward

Linear Regression

A linear regression model has been put into practice for predictive analysis. The dataset was pre-processed to deal with the missing or null values. X train and Y train, the training dataset, were used to produce and train instances of the LinearRegression class from scikit-learn. Cross-validation was done to evaluate the model performance. For visualization purposes, the distribution and scatter plots of actual against predicted value were generated.

Implementation of Machine Learning Models

Random Forest and Linear Regression Implementation

Random Forest is an ensemble learning technique and efficient machine-learning model that excels in regression as well as in classification tasks and Linear regression is one of the simplest methods and the most specific machine-learning models. Fig. 3, Fig. 4, Fig. 5, Fig. 6, and Fig. 7 represents the implementation of machine learning model.

Following is the explanation of its implementation:

- Data preprocessing: The very first stage was data preprocessing. This includes checking data for missing values and splitting the dataset into testing and training sets.
- Model initialization: The Random Forest and Linear regression models were initiated by utilizing the 'RandomForestRegressor' from the 'sklearn.ensemble' library and the 'LinearRegression' class from the 'sklearn.linear_model' module.

```
from sklearn.linear_model import LinearRegression
```

```
[18]:
regressor=LinearRegression()
regressor.fit(X_train,y_train)
```

```
[18]:
▾ LinearRegression
LinearRegression()
```

Fig. 3 Linear Regression Model Initialization

```
from sklearn.ensemble import RandomForestRegressor
```

```
# Define the parameter grid for RandomizedSearchCV
random_grid = {
    'n_estimators': [int(x) for x in np.linspace(start=100, stop=1200, num=12)],
    'max_features': ['auto', 'sqrt'],
    'max_depth': [int(x) for x in np.linspace(5, 30, num=6)],
    'min_samples_split': [2, 5, 10, 15, 100],
    'min_samples_leaf': [1, 2, 5, 10]
}
```

```
[17]:
```

```
# Create RandomForestRegressor
rf = RandomForestRegressor()
```

Fig. 4 Random Forest Model Initialization

- Model training: The model was trained on a training dataset using the 'fit' function.

```
regressor=LinearRegression()
regressor.fit(X_train,y_train)
```

```
rf_random.fit(X_train, y_train)
```

Fig. 5 Model Training

- Prediction: After training the model prediction on the dataset

```
prediction=regressor.predict(X_test)
```

```
# Make predictions
predictions = rf_random.predict(X_test)
```

Fig. 6 Model Prediction

- Evaluation: The model performance was evaluated using MAE, MSE, and RMSE metrics.

```
# Evaluate the model
mae = metrics.mean_absolute_error(y_test, predictions)
mse = metrics.mean_squared_error(y_test, predictions)
rmse = np.sqrt(mse)
```

```
# Evaluation metrics
print('MAE:', metrics.mean_absolute_error(y_test, prediction))
print('MSE:', metrics.mean_squared_error(y_test, prediction))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, prediction)))
```

Fig. 7 Model Evaluation

Exploratory Data Analysis (EDA)

An initial evaluation is performed using exploratory data analysis before the application of ML techniques. Machine learning-based AQI prediction improves visual representation and highlights the relationship between various air contaminants (Kumar and Pande 2023b). EDA was employed to find hidden connections and patterns. Stronger pairwise relationship visualization was made possible by pairplot and correlation heatmaps, which also help in the identification of possible correlations that enhance data understanding. Fig. 13 depicts a correlation heatmap for the input data and the values generally vary from positive 1 to negative 1. The association between the input variables is accurately represented in a heatmap that was generated using the correlation matrix technique. The correlation was negative for certain factors and positive for some. It was determined that the positive correlation is important to the AQI prediction after studying the negative and positive correlations. Fig. 8, Fig. 9, Fig. 10, Fig. 11, and Fig. 12 shows the steps involved in the creation of pairplots and heatmaps.

Steps involved in the creation of heatmap and pair plots:

1. Import necessary libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn import metrics
import pickle
```

Fig. 8 Input Libraries

3. Load and prepare the data

```
df = pd.read_excel("C:\\Users\\hp\\Desktop\\paper2.xlsx")
```

Fig. 9 Data Loading

4. Calculate the correlation matrix

```
df.corr()
```

Fig. 10 Calculation of Correlation Matrix

5. Create the heatmap and the pair plots

```
import seaborn as sns
#get correlations of each features in dataset
corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))
#plot heat map
g=sns.heatmap(df[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

Fig. 11 Creation of Heatmap

```
sns.pairplot(df)
```

Fig. 12 Creation of Pairplot

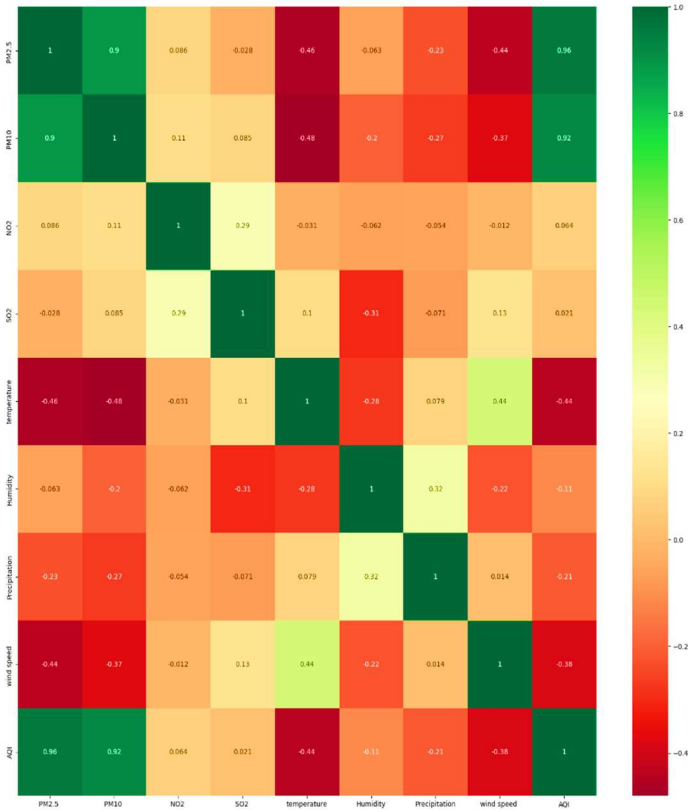

Fig. 13 Metrix of the Correlation Coefficient

III. Results and Discussion

In this research work, AQI of the Gorakhpur city was our regression target. For this study, two regression models were used for predicting air quality. In the first phase, the model was split into training and testing sets with a ratio of 80:20. The model was trained using the first 80% of the data, and for testing the remaining 20% of the data was used. For data processing purposes python libraries like Scikit-learn, NumPy, Pandas, and Seaborn, are utilized. After this, the dataset was explored to determine the overall AQI value for the pollutant concentration which is significantly responsible for raising AQI value. The correlation between the variables is depicted in Fig. 13 which visualizes the pollutants that show the relationship between AQI values and different meteorological parameters and air pollutants. For model evaluation, the metrics including MAE, MSE, RMSE, and, R^2 were utilized.

Data Exploration and Visualization

Data exploration and visualization approach helps in building a robust model for AQI prediction by providing crucial information regarding the factors influencing air quality. scatterplot, histogram display, and pairplot for every possible combination are provided by the Python Seaborn library as depicted in Fig. 14. The main function of a pairplot is to plot each parameter against the other's value whereas scatterplots illustrate the association between the combination of variables in the upper and lower triangles of grids. The function of the histogram is to show the distribution of each variable along the diagonal of the grids. Analysis of correlation, trend, and, pattern among variables is facilitated through pairplot visualization. The correlation heatmap as shown in Fig. 13 provides easy visualization by facilitating graphical representation.

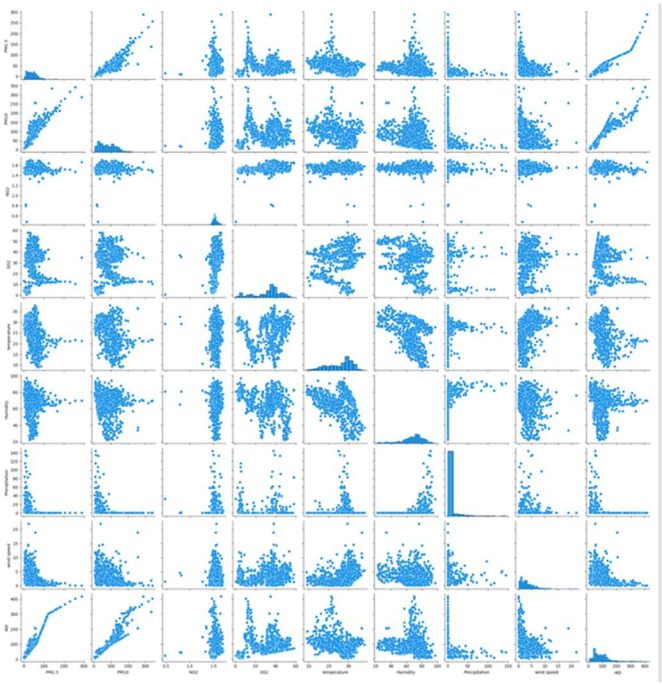

Fig. 14 Pairplot representation of each parameter

AQI Prediction by Machine Learning Models

In the optimization phase, to identify the model's optimum performance across several data subsets, cross-validation is used. Training and testing the model on several folds of data ensures a more precise assessment of the model's capacity. Across all the folds model performs effectively, confirming that the model is not overfitting for certain data points. For model evaluation, the dataset was divided into a ratio of 80:20 which means the model is trained over 80% of the dataset and tested over 20% of the dataset. For assessing the effectiveness and accuracy of the developed ML model metrics like RMSE. MSE, MAE, and, R^2 are important (Bao and Zhang,2020). In a very frequent manner, the R^2 is utilized to compare the best fit, however, to find the well-fit model, it additionally becomes necessary to compare different model errors.

Random Forest Results

The MAE calculates the average magnitude of errors in an entire set of predictions without taking into consideration their direction. An MAE of 0.79876 for the Random Forest demonstrates that the model's prediction generally varies by 0.8 units from the actual values. The lower value of MAE indicates that the model's prediction is quite accurate. The MSE indicates the average of squares of the errors and determines the larger errors. With an MSE of 2.45283 in between observed and actual values, Random Forest indicates the average squared difference of about 2.45. The low MSE values represent that the model is well-trained and rarely makes larger mistakes. The error measures generated by the RMSE are in equivalent units as the target variable, which is the squared root of the MSE value. The Random Forest model's prediction error is about 1.57 units with an RMSE value of 1.56615 comparatively lower error, indicating higher prediction performance. The R^2 value shows the percentage of the variance of the dependent variables from the independent variables. With the R^2 value of 0.99895 Random Forest model highlights its outstanding fit and greater accuracy of 99.89%. Fig. 15, Fig. 16, and Fig. 17 show the scatter plot of actual values against predicted value, the residual plot for Random Forest, and the histogram for the model respectively.

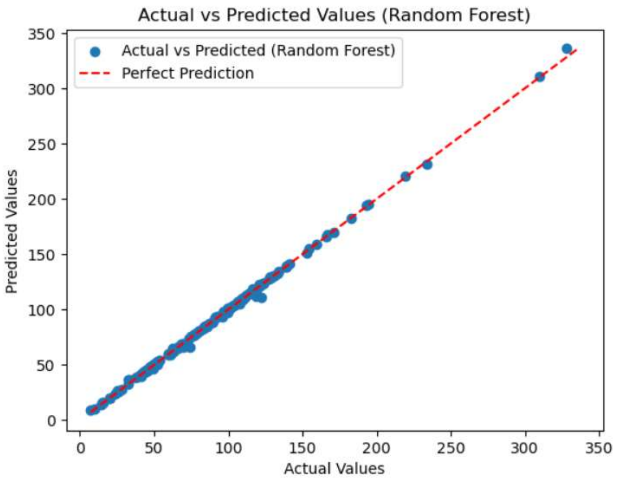

Fig. 15 Scatter plot for Random Forest model

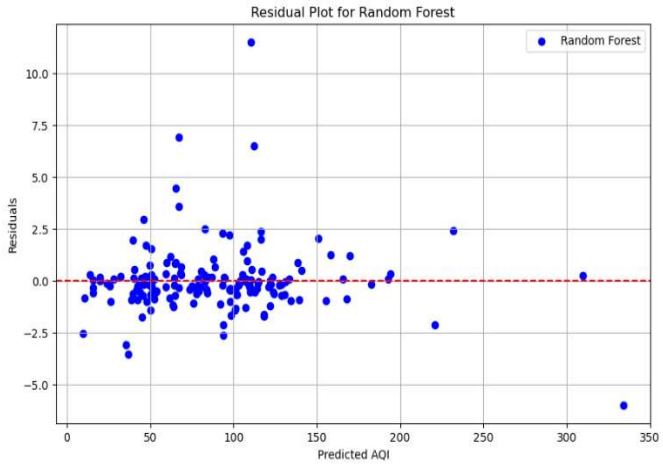

Fig. 16 The residual plot for the Random Forest model

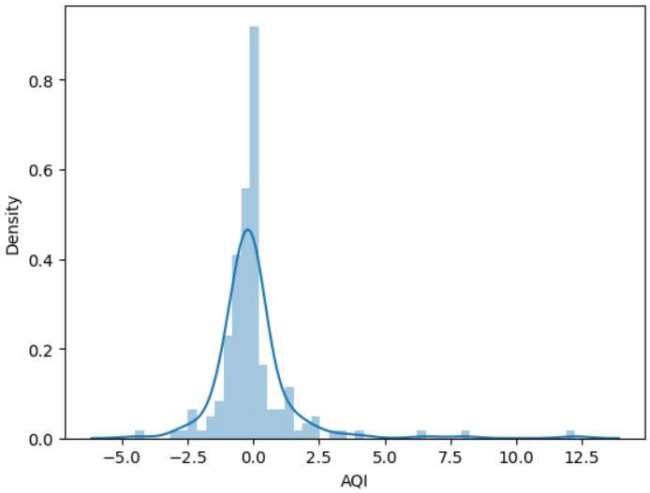

Fig. 17 Histogram for Random Forest model

Linear Regression Results

For Linear Regression an MAE value of 10.71808 shows that the model's prediction fluctuated by 10.72 units from the actual values. The higher value of MAE indicates a lower precision in the model's prediction. This model has a higher value of MSE of 190.9275 which indicates the difference between the predicted and actual values. The greater value of MSE demonstrates that the model is susceptible to significant errors and inaccuracies. The RMSE value for Linear Regression is 13.8176 which indicates that the typical prediction error is about 13.82 units. The R^2 value of 0.91848 shows that the model is susceptible to approximately 91.8% of variance in the data. Though it is less than that of Random Forest's R^2 value but still it shows a comparatively good fit. Fig. 18 shows the scatter plot of actual values against predicted value, Fig. 19 depicts the residual plot for Linear Regression, and Fig. 20 shows the histogram for the Linear Regression model.

Accuracy Comparison

The prediction accuracy of both machine learning models is shown in Table 2. Fig. 21 and Fig. 22 show the value of evaluation metrics for the Random Forest and Linear Regression model for better visualization. These results show that Random Forest performed superiorly compared to the Linear regression model, with the lowest error evaluation metrics and highest R^2. The greater error metrics and lower R^2 value Linear Regression model show its failure to understand or learn over complex datasets.

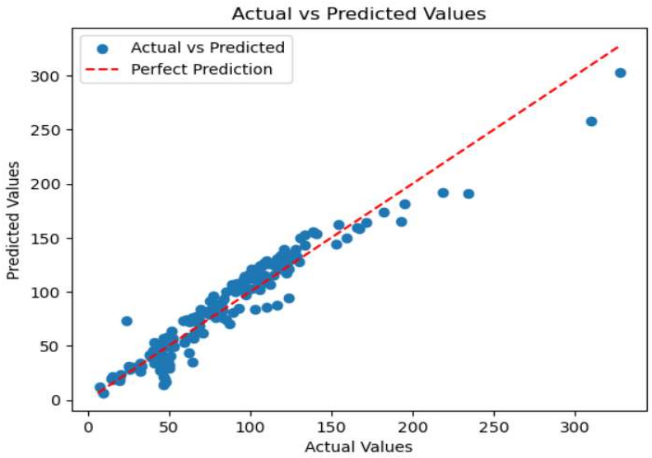

Fig. 18 Scatter plot for Linear Regression model

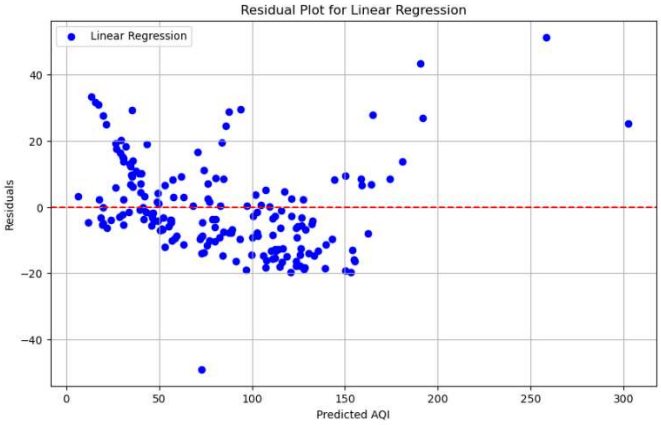

Fig. 19 The residual plot for the Linear Regression model

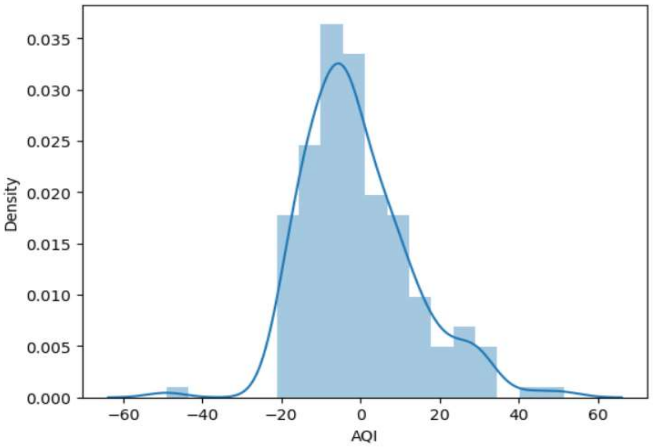

Fig. 20 Histogram for Linear Regression model

TABLE 2
Model Performance Metrics

Model	MAE	MSE	RMSE	R^2
Random Forest	0.79876	2.45283	1.56615	0.99895
Linear Regression	10.71808	190.9275	13.8176	0.91848

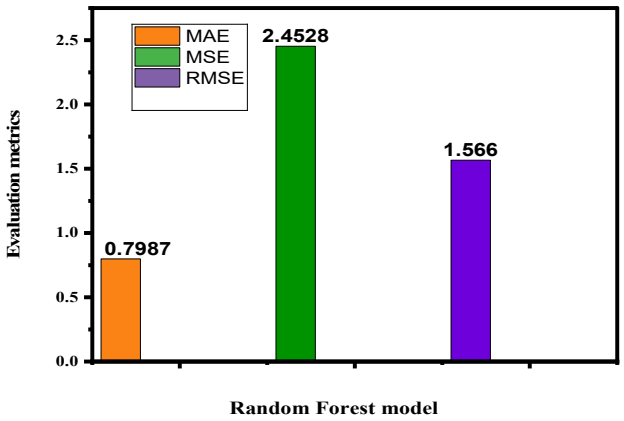

Fig. 21 Evaluation Metrics for Random Forest

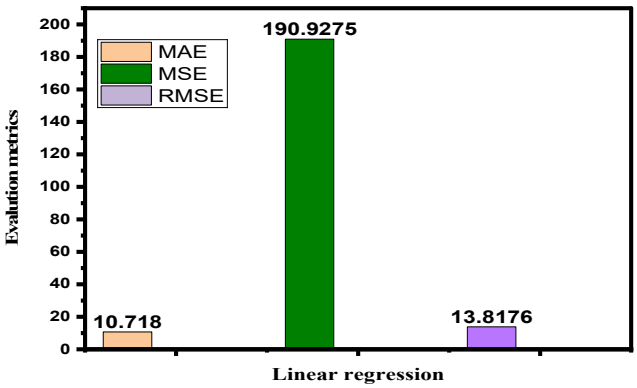

Fig. 22 Evaluation Metrics for Linear Regression model

Statistical Analysis

The statistical measures used for this study were R^2, mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). The T-statistic test was also performed on the models with associated P-values for the models. The coefficient of determination (R^2) calculated for both models measured the variance produced in the dependent variable which was predicted using the independent variable. The Random Forest model shows a higher R^2 which is significantly more than that of the Linear Regression. The mean R^2 value for linear regression was 0.9972 with Std. deviation of 0.0022, which is more significant than that of Linear Regression (0.9351). the differences are statistically significant as shown by p-values of 3.09433e-08 and t-statistic of 17.3968 as generated through T-test for R^2. For the Random Forest, a 95% confidence interval (0.0541, 0.0702) indicates its superior performance.

The mean MAE value of the Random Forest was 1.1093 which was less than the mean MAE value of the Linear Regression (11.3382). A lower MAE value reflects good prediction and better accuracy, indicating with a substantially lower MAE value the Random Forest model's prediction is significantly closer to the actual values. The MAE T-test showed a t-statistic of -2208789 and a p-value of 2.76476e-09 which indicates a significant difference between both the models. The better performance of Random Forest was supported by the 94% confidence interval (-11.2403, -9.2175) for the MAE difference. The random Forest model shows a lower mean MSE value of 9.6423 compared to the Linear Regression model (247.0357), indicating that Random Forest produces fewer mistakes than that of Linear Regression. The t-statistic (-14.2035) and p-value (1.81026e-07) obtained from the MSE T-test also show statistically significant differences. This result was also backed by a 95% confidence interval for MSE (-275.2025, -199.5844). For better accuracy lower RMSE value is required so as shown by the Random Forest model, meaning that the predicted values are significantly closer to the actual values. For the T-test for RMSE, the t-statistic value of -19.32038 and p-value of 1.23138e-08 was obtained show that the difference is exceptionally significant. The 95% confidence interval for the RMSE (-14.230159, -11.24711373) proved that the Random Forest performs much better than the Linear Regression model.

This result explained the higher variance for Random Forest when compared to the Linear Regression. The observed T-test and corresponding p-value for both models also indicate that the difference in R^2 between the Random Forest and Linear Regression is statistically greater. This observation provides a conclusion that Random Forest is significantly outperforming as compared to Linear Regression, in terms of calculating variance in data. The results are shown in Table 3, Table 4, and Table 5.

TABLE 3
Performance Metrics mean and std. Deviation

Performance Metrics	Model	mean	Std. deviation
R^2	Random Forest	0.997244	0.002201
	Linear Regression	0.935078	0.012021
MAE	Random Forest	1.109311	0.199533
	Linear Regression	11.33820	1.251270
MSE	Random Forest	9.64227	6.598139
	Linear Regression	247.03570	48.67721
RMSE	Random Forest	2.912425	1.077057
	Linear Regression	15.65106	1.442208

TABLE 4
T-statistics and p-value

Metrics	T-statistics	p-value
R^2	17.39682	3.094339524148819e-08
MAE	-22.87892	2.7647649286815677e-09
MSE	-14.20350	1.8102695137983046e-07
RMSE	-19.32038	1.2313833435517461e-08

TABLE 5
Confidence Interval of Difference

Metrics	95% confidence interval of the difference
R^2	0.0540819311516525, 0.07024903211963895
MAE	-11.24027251100685, -9.217506821864749
MSE	-275.2024917405684, -199.58437698244967
RMSE	-14.230159611920834, -11.247113762129173

IV. Conclusion

This research paper comprehensively analyzes machine learning models and their application for building models for urban cities. This comparative assessment concluded with the following key results.

- The Random Forest model demonstrates exceptional performance across all the metrics compared to the Linear Regression model, showing that the Linear Regression has less ability to handle non-linear datasets.
- The R^2 value for the Random Forest model (0.99895) is very close to 1 showing that the Random Forest model almost perfectly explains the degree of variance in the dataset. In contrast, the Linear Regression model's R^2 (0.91848) is less than the random forest's R^2 explaining that it accounts for less variance.
- The Random Forest model with a much lower MAE (0.79876) and MSE (10.71808) produces superior prediction than the Linear Regression model which shows significantly higher MAE and MSE values (10.71808, and 190.9275)
- Furthermore, the statistical assessment proves that the Random Forest model performs superiorly when compared to the Linear Regression. By using t-statistics and p-values for different evaluation metrics like MAE, MSE, RMSE, and R^2 it is confirmed that there is a significant difference in performance. The Confidence intervals of the differences with negative values which means the predicted values are much closer to the actual values, further validate its higher performance.

Overall, the Random Forest model compared to the Linear Regression model performs considerably more efficiently both in terms of accuracy and reliability for this research work.

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. The authors would also like to mention the USEPA website and **preventionweb.net** for delivering the information useful for this study.

V. References

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

Ayus, I., Natarajan, N., & Gupta, D. (2023). 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

Bai, L., Wang, J., Ma, X., & Lu, H. (2018, April 17). Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*. MDPI AG. https://doi.org/10.3390/ijerph15040780

Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., et al. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *The Lancet Planetary Health*, *3*(1), e26–e39. https://doi.org/10.1016/S2542-5196(18)30261-4

Bodor, Z., Bodor, K., Keresztesi, Á., & Szép, R. (2020). Major air pollutants seasonal variation analysis and long-range transport of PM10 in an urban environment with specific climate condition in Transylvania (Romania). *Environmental Science and Pollution Research*, *27*(30), 38181–38199. https://doi.org/10.1007/s11356-020-09838-2

Dewan, S., & Lakhani, A. (2022, December 15). Tropospheric ozone and its natural precursors impacted by climatic changes in emission and dynamics. *Frontiers in Environmental Science*. Frontiers Media S.A. https://doi.org/10.3389/fenvs.2022.1007942

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

Kumar, K., & Pande, B. P. (2023a). 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

Kumar, K., & Pande, B. P. (2023b). 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

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

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

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020, February 20). Environmental and Health Impacts of Air Pollution: A Review. *Frontiers in Public Health*. Frontiers Media S.A. https://doi.org/10.3389/fpubh.2020.00014

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

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

Ravindiran, G., Rajamanickam, S., Kanagarathinam, K., Hayder, G., Janardhan, G., Arunkumar, P., et al. (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

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

Taylan, O., Alkabaa, A. S., Alamoudi, M., Basahel, A., Balubaid, M., Andejany, M., & Alidrisi, H. (2021). Air quality modeling for sustainable clean environment using anfis and machine learning approaches. *Atmosphere*, *12*(6). https://doi.org/10.3390/atmos12060713

Tien, P. W., Wei, S., Darkwa, J., Wood, C., & Calautit, J. K. (2022, November 1). Machine Learning and Deep Learning Methods for Enhancing Building Energy Efficiency and Indoor Environmental Quality – A Review. *Energy and AI*. Elsevier B.V. https://doi.org/10.1016/j.egyai.2022.100198

Villanueva, F., Notario, A., Tapia, A., Albaladejo, J., Cabañas, B., & Martínez, E. (2016). Ambient levels of volatile organic compounds and criteria pollutants in the most industrialized area of central Iberian Peninsula: intercomparison with an urban site. *Environmental Technology (United Kingdom)*, *37*(8), 983–996. https://doi.org/10.1080/09593330.2015.1096309

Von Schneidemesser, E., Monks, P. S., Allan, J. D., Bruhwiler, L., Forster, P., Fowler, D., et al. (2015, May 27). Chemistry and the Linkages between Air Quality and Climate Change. *Chemical Reviews*. American Chemical Society. https://doi.org/10.1021/acs.chemrev.5b00089

Wang, S., Feng, X., Zeng, X., Ma, Y., & Shang, K. (2009). A study on variations of concentrations of particulate matter with different sizes in Lanzhou, China. *Atmospheric Environment*, *43*(17), 2823–2828. https://doi.org/10.1016/j.atmosenv.2009.02.021