Predicting Air Quality Index (AQI) Using Machine Learning in Urban Indian Cities

Raj Prakashchandra Chauhan
Department of Data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India
Guide: Prof. Shubhangi Tidke
Co-Guide: Prof. Prashant Kulkarni

Abstract- This study investigates the use of machine learning models like Random Forest and XGBoost to predict Air Quality Index (AQI) in Indian urban cities. By analyzing major pollutants such as PM2.5, PM10, NO2, and CO, the research aims to support early warning systems and data-driven environmental policy decisions.

Index Terms—AQI, Machine Learning, Random Forest, XGBoost, Pollution Prediction.

I. INTRODUCTION

Air pollution in the present world become a major environmental as well as public health concern in many Indian cities, mainly in densely populated urban areas. The aim of this study is to improve early warning systems and sustainable living in the urban areas. As cities such as Delhi, Mumbai, and Kolkata face dangerous levels of air pollution, there is a subsequent requirement to have intelligent models that would facilitate proactive health planning and policy-making. The analysis makes use of indicators of pollutants, including PM2.5, PM10, NO2, SO2, CO, and others, combined with a machine learning method based on regression. Random Forest and XGBoost models are elaborated with the aim of finding the most accurate predictors of AQI.

II. PROBLEM STATEMENT

The short-term AQI levels in Indian urban cities are not easily predicted because of the dynamic nature of various pollutants and environmental conditions. The prediction model of machine learning is required to enhance real-time insight to forestall the chances of projecting the AQI (Pande *et al.* 2025). The emergency interventions in the city health sectors and informed policy formulations to deal with the air quality of the urban environment can occur.

III. OBJECTIVES

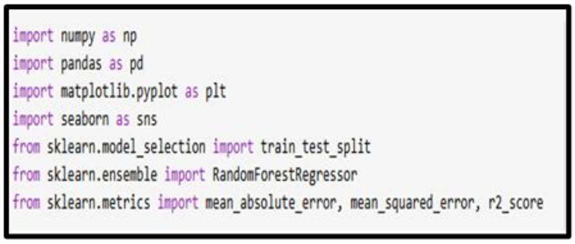
- To gather and pre-process the past air quality and other pollutant levels, such as PM2.5, PM10, NO_2, CO, SO_2 and other important indicators of the environment in various Indian urban cities.
- To compare and apply machine learning regressions, including Random Forest and XGBoost, to predict air factors correctly.
- To assess the effectiveness of the model using suitable error measures such as MAE, RMSE, and R^2 to identify the effectiveness of the prediction.
- To present observations that enhance early warning and data-based environmental policy programming.

IV. METHODOLOGY

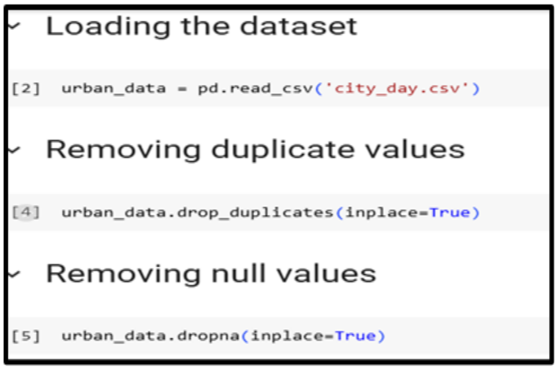
The processing of the results of the work is based on the methodology of collecting and preprocessing the historical information on the air quality of the major Indian cities. The main characteristic of the pollutants, including PM2.5, PM10, NO_2, CO_2, and so on, is chosen to be analyzed (Natarajan *et al.* 2024). The data is loaded and cleaned by eliminating duplicates and nulls using pandas or other Python libraries. Linear Regression, Random Forest, and XGBoost regression models are applied in predicting AQI, which reflects the pollution level (Binbusayyis *et al.* 2024). The data recorded is separated into a training and a testing set. The performance of the model is measured using the values of MAE, RMSE, and R^2 to achieve accuracy and strength.

V. OUTCOME

The research, a quality machine learning model is to be obtained to help predict short-term AQI levels and assist in early health warnings and environmental decisions based on data.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```



Loading the dataset

```
[2]  urban_data = pd.read_csv('city_day.csv')
```

Removing duplicate values

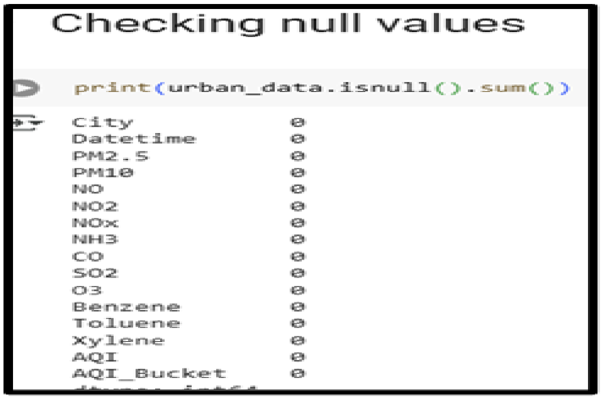
```
[4]  urban_data.drop_duplicates(inplace=True)
```

Removing null values

```
[5]  urban_data.dropna(inplace=True)
```

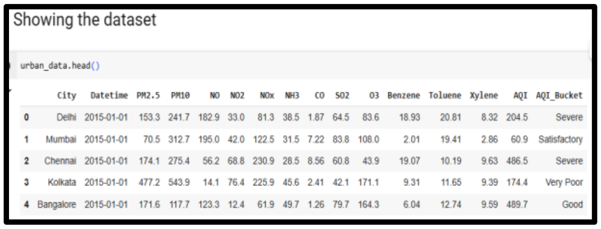
Figure 1: Importing libraries and removing duplicate and null values

The Python code loads the dataset, which contains the air quality data in the file city_day.csv, into a DataFrame entitled urban_data. It drops duplicate rows to maintain the integrity of data and then drops the rows with missing data to leave the dataset, preparing it to be analyzed or modelled.



Checking null values

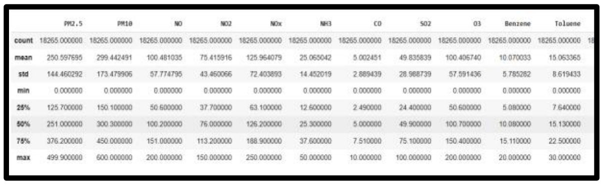
```
print(urban_data.isnull().sum())

City           0
Datetime       0
PM2.5          0
PM10           0
NO             0
NO2            0
NOx            0
NH3            0
CO             0
SO2            0
O3             0
Benzene        0
Toluene        0
Xylene         0
AQI            0
AQI_Bucket     0
dtype: int64
```



Showing the dataset

```
urban_data.head()
```

```
        City    Datetime  PM2.5   PM10     NO   NO2    NOx   NH3    CO   SO2     O3  Benzene  Toluene  Xylene    AQI    AQI_Bucket
0      Delhi  2015-01-01  153.3  241.7  182.9  33.0   81.3  38.5  1.87  64.5   83.6    18.93    20.81    8.32  204.5        Severe
1     Mumbai  2015-01-01   70.5  312.7  195.0  42.0  122.5  31.5  7.22  83.8  108.0     2.01    19.41    2.86   60.9  Satisfactory
2    Chennai  2015-01-01  174.1  275.4   56.2  68.8  230.9  28.5  8.56  60.8   43.9    19.07    10.19    9.63  486.5        Severe
3    Kolkata  2015-01-01  477.2  543.9   14.1  76.4  225.9  45.6  2.41  42.1  171.1     9.31    11.65    9.39  174.4     Very Poor
4  Bangalore  2015-01-01  171.6  117.7  123.3  12.4   61.9  49.7  1.26  79.7  164.3     6.04    12.74    9.59  489.7          Good
```

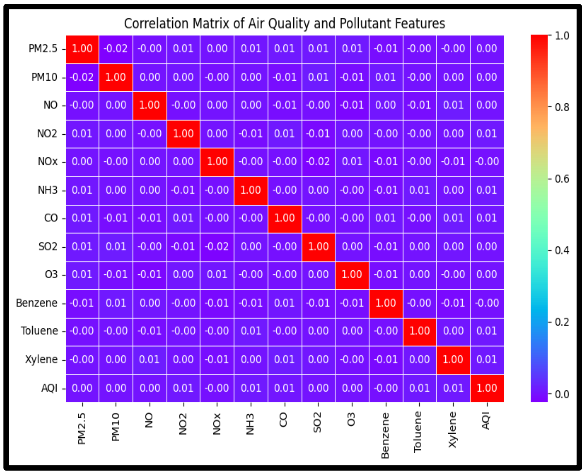


```
              PM2.5          PM10            NO           NO2           NOx           NH3            CO           SO2            O3       Benzene       Toluene
count  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18265.000000  18
mean     250.597695    299.442491    100.481035     75.415916    125.964079     25.065042      5.002451     49.835839    100.406740     10.070033     15.063365
std      144.460292    173.479906     57.774795     43.460066     72.403893     14.452019      2.889439     28.988739     57.591436      5.785282      8.619433
min        0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000
25%      125.700000    150.100000     50.600000     37.700000     63.100000     12.600000      2.490000     24.400000     50.600000      5.080000      7.640000
50%      251.000000    300.300000    100.200000     76.000000    126.200000     25.300000      5.000000     49.900000    100.700000     10.080000     15.130000
75%      376.200000    450.000000    151.000000    113.200000    188.900000     37.600000      7.510000     75.100000    150.400000     15.110000     22.500000
max      499.900000    600.000000    200.000000    150.000000    250.000000     50.000000     10.000000    100.000000    200.000000     20.000000     30.000000
```

Figure 2: Checking for null values and showing the dataset and Descriptive Statistics

The output has no missing values or null values in any of the columns of the urban_data DataFrame. The figure presents the first five records of the data, consisting of the level of various air pollutants such as PM2.5, PM10, NO^2, CO, and SO^2 of various cities and their AQI as well as AQI category.

The descriptive statistics present the presentation of pollutant concentrations throughout 18,265 records. Every pollutant displays the count, mean, standard deviation, minimum, and quartiles of the pollutant. The dispersions and averages of PM10 and PM2.5 demonstrate extreme values, showing large variations in the level of pollution across urban areas in India.



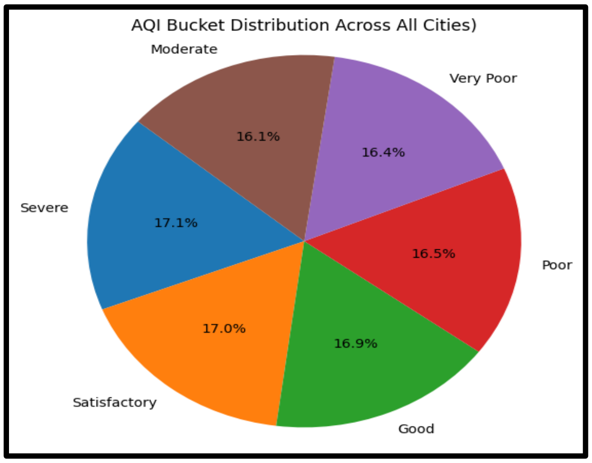


Figure 3: Correlation matrix and AQI Bucket Distribution across All (Cities)

The heat-map demonstrates low associations between the majorities of pollutants. There is a low positive correlation between PM2.5 and PM10, whereas AQI does not strongly correlate with a particular pollutant. The pie chart demonstrates that the distribution of AQIs is somewhat even, with a slight dominance of Severe and Satisfactory. This indicates the changing levels of air quality among the cities of India that highlights the variability, and inconsistency in pollution control performance.

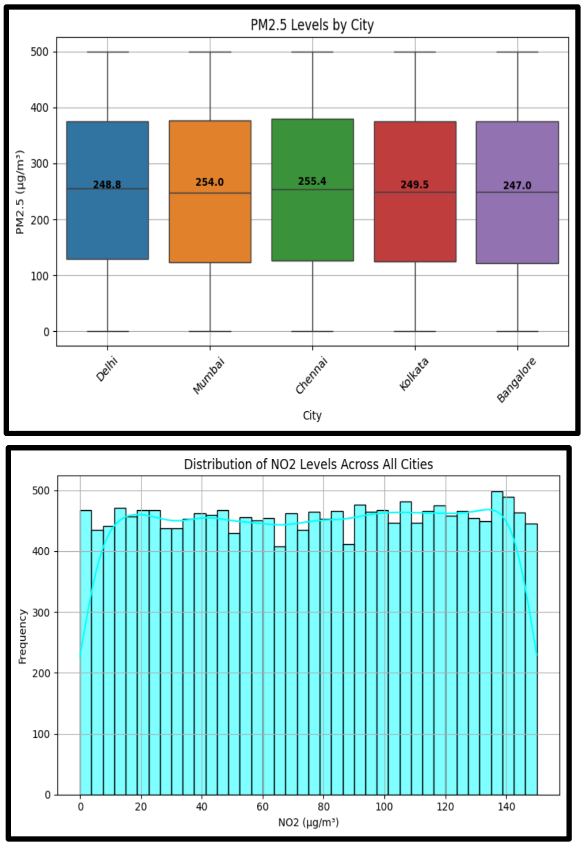


Figure 4: PM2.5 distribution by city and Distribution of NO2 across all cities

Boxplot comparison interprets that Chennai and Mumbai are the highest in terms of the median of PM2.5. Wide variability is observed in all cities, which indicates that PM2.5 is a high pollutant in all urban centers of India probably caused by traffic, construction, or industrial activities.

Histogram interprets that NO_2 have a distribution across different cities except that there are slight peaks showing that there is regular exposure. This regularity is an indication of regular sources of pollution such as emissions by vehicles, but the slight variation is a sign that there are local causes.

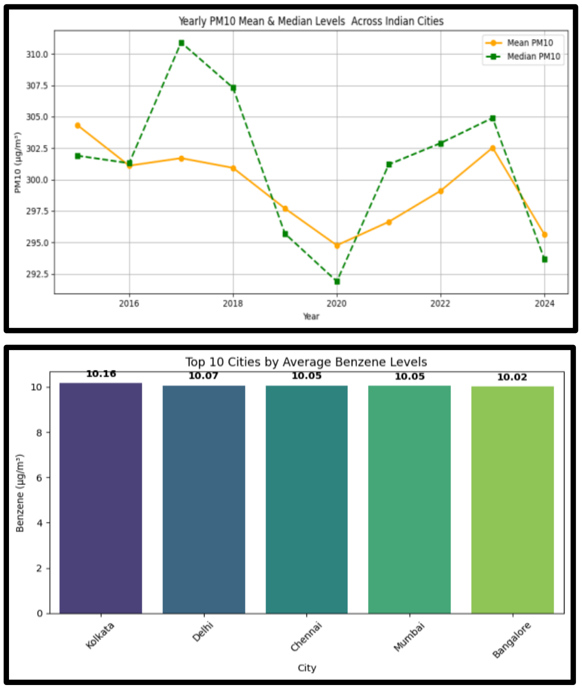


Figure 5: Average and median PM10 over time, all cities and Top 10 cities by average Benzene Levels

The decrease in 2020 implies a decline in activity linked to the pandemic. This is because the mean and median values are very close, which means they are fairly distributed without large outliers in most of the years. The plot of the time series of PM10 measurements made in Google colab by the help of Python. The average levels of benzene are highest in Kolkata, closely followed by Delhi and Chennai. Benzene is a carcinogenic chemical, as this is an indication of a health issue among citizens exposed to industrial pollutants and fuel-burning discharges. The small variation in value indicates high prevalence of benzene in the urban space and not a single region.

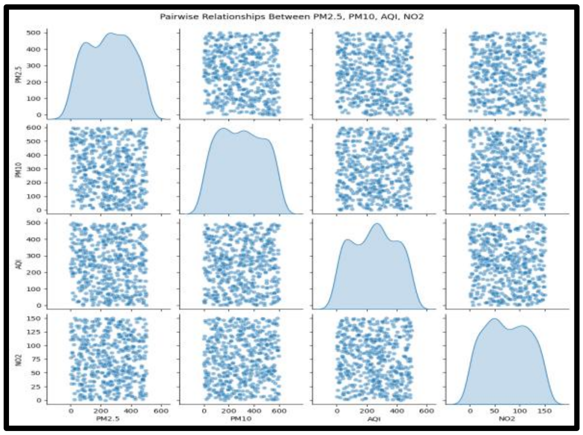
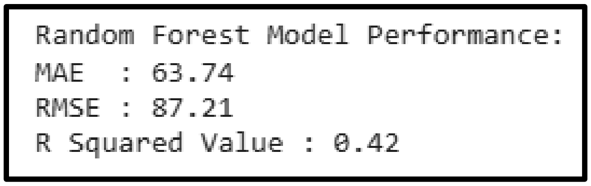


Figure 6: Pairwise relationships between PM2.5, PM10, AQI, NO2

The scatterplots indicate that there is no significant clustering between PM2.5 and AQI, though one can observe weak clustering. The right-skew nature of the pollution levels on the diagonals confirms the occurrence of extremely high levels of pollution in some cases despite the majority of the available data points being moderate. The weak linearity trends highlight the reasons behind the use of a non-linear ML model.



```
Random Forest Model Performance:
MAE  : 63.74
RMSE : 87.21
R Squared Value : 0.42
```

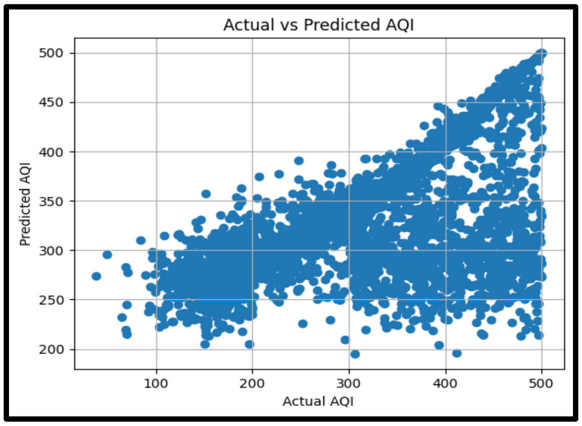
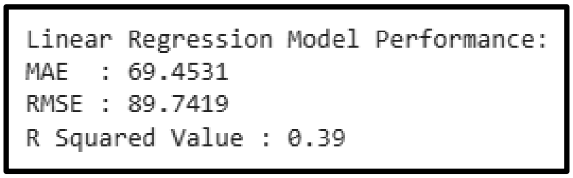


Figure 7: Random Forest Model Performance

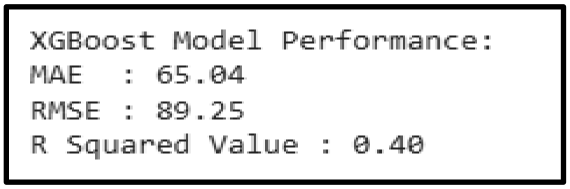
The Random Forest model exhibits average results with one of the values being the MAE, which is 63.74, and the RMSE of 87.21, which suggests that values in the predictions are moderately accurate. The amount of predictability, R^2 value, implies that the model has a 42 percent chance of explaining the variance in AQI and therefore has a partial predictive ability.



```
Linear Regression Model Performance:
MAE  : 69.4531
RMSE : 89.7419
R Squared Value : 0.39
```

Figure 8: Linear Regression Model Performance

The Linear Regression model shows satisfactory results as its MAE is 69.45 and RMSE is 89.74, and it comes to the predictive performance.. The R^2 of 0.39 also means that it explains 39 percent of the variance of AQI with limited but reasonable accuracy.



```
XGBoost Model Performance:
MAE  : 65.04
RMSE : 89.25
R Squared Value : 0.40
```

Figure 9: XGBoost Model and LSTM Model Performance

The XGBoost model has an MAE at the level of 65.04 and RMSE 89.25, which is moderate prediction accuracy. The model exhibits moderately better results compared to those of the linear regression, as the R^2 is 0.40, with 40 percent of AQI variance being explained.

VI. SCOPE & LIMITATIONS

This study focuses on predicting short-term AQI levels in Indian urban cities using historical air quality data and machine learning models. The scope covers the analysis of major pollutants such as PM2.5, PM10, NO2, and CO within several cities and the implementation of models as Random Forest and XGBoost to forecast accurately (Swamynathan *et al.* 2024). These models are effective and work with complex trends of data; their precision is always prone to what is within the data. The limitations arise due to the quality and magnitude of data available. Publicly sourced data is utilised, and this data can be non-representative of pollutant hotspots or real-time changes (Barthwal and Goel, 2024.). The models are trained using past data, and thus, they are not sensitive to sudden environmental changes. Data for only urban centers is taken into consideration, and no data about the rural air quality movements or about the posts of the personal sensors can be found there (SK and Ravindiran, 2024). The model effectively interprets AQI by using important pollutants such as PM2.5 and NO2.

VII. CONCLUSION

The current research shows that machine learning models including Random Forest, Linear Regression, and XGBoost, can be used to successfully forecast the short-term levels of AQI in major cities across India. The models assist in timely health warnings and policy formulation using pollutant data but the predictions are subject to the accuracy of the data and cannot reflect instantaneous events.

REFERENCES

[1] Barthwal, A. and Goel, A.K., 2024. Advancing air quality prediction models in urban India: A deep learning approach integrating DCNN and LSTM architectures for AQI time-series classification. *Modeling Earth Systems and Environment*, *10*(2), pp.2935-2955.

[2] Binbusayyis, A., Khan, M.A., Ahmed A, M.M. and Emmanuel, W.S., 2024. A deep learning approach for prediction of air quality index in smart city. *Discover Sustainability*, *5*(1), p.89.

[3] Natarajan, S.K., Shanmurthy, P., Arockiam, D., Balusamy, B. and Selvarajan, S., 2024. Optimized machine learning model for air quality index prediction in major cities in India. *Scientific Reports, 14(1),* p.6795.

[4] Pande, C.B., Radwan, N., Heddam, S., Ahmed, K.O., Alshehri, F., Pal, S.C. and Pramanik, M., 2025. Forecasting of monthly air quality index and understanding the air pollution in the urban city, India based on machine learning models and cross-validation. *Journal of Atmospheric Chemistry*, *82*(1), p.1.

[5] SK, A. and Ravindiran, G., 2024. Integrating machine learning techniques for Air Quality Index forecasting and insights from pollutant-meteorological dynamics in sustainable urban environments. *Earth Science Informatics*, *17*(4), pp.3733-3748.

[6] Swamynathan, S., Sneha, N., Ramesh, S.P., Niranjana, R., Ponkumar, D.D.N. and Saravanakumar, R., 2024, December. A Machine Learning Approach for Predicting Air Quality Index in Smart Cities. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* (pp. 1609-1615). IEEE.