Machine Learning Approach to Analyze Sensor Data of Air Pollutants for Sustainable Smart Cities

[1]Radhika V. Kulkarni
[1]Associate Professor
Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, Maharashtra, India -411037.

Email: radhikavikaskulkarni@gmail.com
ORCID ID:- https://orcid.org/0000-0001-6614-6297

[2]Rakhi J. Bharadwaj
[2]Assistant Professor
Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, Maharashtra, India -411037.
rakhi.bharadwaj@vit.edu

[3]Kaustubh V. Sakhare
[3]Senior Data Scientist II
John Deere India.
Email: kaustubhsakhare@gmail.com
ORCID ID:- https://orcid.org/0000-0001-8108-8590

***Abstract*—** Rapid urbanization demands solutions for sustainable healthy smart cities. It aims to make cities safe, environmentally friendly, equitable, and resilient. One of the critical problems in cities is air pollution. Monitoring air quality is essential for ensuring public health. The established remote air quality monitoring stations in smart cities produce a large volume of sensor data on air pollutants. Intelligent systems to analyze air pollution to build a sustainable solution for smart cities are of the utmost importance. The current study reports the applicability of several machine learning (ML) models for air pollutants analysis in major smart cities in India. The work focuses on three major tasks: 1) multiclass categorization of air quality using a variety of classifiers, 2) Air Quality Index (AQI) prediction by employing different regression models, and 3) comparative study of empirical results of different classifiers and regressors used in air quality analysis of smart cities. The research employs five classifiers and five regressors to analyze real-time sensor data of air pollutants from major five Indian metro cities from 2021 to 2023. The performance of these learning models is evaluated on a variety of metrics.

Keywords: Air Quality Index (AQI) Prediction, Air Pollutants Sensor Data Analysis, Classification, Machine Learning (ML), Regression, Smart Cities.

I. INTRODUCTION

The worldwide urban population is predicted to rise by 2.4 billion by 2050 [1]. The speedy growth in urbanization leads to many livability challenges like pollution, waste management, scarcity of resources, climate change, and economic and social diversity. Developing sustainable smart cities is the solution to these problems [2]. A smart city is livable, environmentally friendly, and has a flourishing economy that gives its residents many possibilities to follow their interests [3], [4]. Thus, sustainability in smart cities is looked at through the lenses of environmental, social, and economic aspects.

Globally, air pollution is a serious hazard to society's health. There are several diseases and medical disorders linked to disability, morbidity, and mortality that are among the adverse consequences of air pollution on human health [5], [6]. The major cause of air pollution is the growth and modernization of the metropolitan area. Worldwide the main contributors to severe air pollution are air-polluting factors such as Particulate Matter having an aerodynamic diameter of 2.5 and 10 µm or smaller ($PM_{2.5}$ and PM_{10}), Carbon monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), etc. Exposure to these air pollutants increases the risk of stroke, heart disease, and chronic diseases like diabetes [2]. The amount of air polluting factors present in ambient air are used to determine the Air Quality Index (AQI) [7].

The sensor infrastructure to capture the data of air pollutants and Information and Communication Technology (ICT) to analyze it for developing environmentally sustainable solutions encompasses the "smartness" of cities. The applicability of machine learning (ML) models to analyze sensor data of air pollutants is a paramount use case of digital technology for sustainable smart cities. It is a trigger to the research work presented through this correspondence.

Air pollutants sensor data for the period of 2021 to 2023 of major Indian smart cities, namely, Delhi, Mumbai, Bengaluru, Chennai, and Kolkata are examined in this research to evaluate different ML models for air quality analysis. The objectives of this study are to:

- Compile the real-time sensor dataset of air pollutants collected from major Indian smart cities.
- Categorize air quality into six classes – 1) good, 2) satisfactory, 3) moderately polluted, 4) poor, 5) very poor, and 6) severe using a variety of classifiers.
- Predict AQI by employing different regression models.
- Do a comparative empirical analysis of different classifiers and regressors used in air quality analysis.

The current communication is further organized into four more sections. Section II throws light on the related research work. Section III describes the research method. Section IV reports the empirical results and their analysis. The communication concludes in section V.

II. RELATED WORK

An intelligent system for the correct analysis of air quality is essential. The majority of nations in the world have designated a certain set of laws and regulations to gauge air quality. Many researchers contributed to building intelligent systems for air quality prediction using ML.

Abu El-Magd S. et al. [8] apply a Random Forest (RF) algorithm to predict the distribution of the air pollutant PM_{10} in the North Ras Gharib region of Egypt. The study in [9] applies various ML

models, namely, Decision Tree (DT), Expectation Maximization, Artificial Neural Network (ANN), Logistic Regression, and Support Vector Machine (SVM) to classify air pollution levels as low, medium, and high by referring to $PM_{2.5}$, PM_{10}, SO_2, and NO_2 pollutants data in Visakhapatnam city, Andhra Pradesh, India. The work in [10] examines various polluting factors for AQI prediction using five ML models namely, K-Nearest Neighbor (KNN), SVM, Gaussian Naive Bayes (GNB), RF, and Extreme Gradient Boosting (XGBoost) on a dataset from Indian cities. Ketu S. [11] proposes a Recursive Feature Elimination with a Random Forest Regression (RFERF) model for predicting AQI and pollutant NO_x in data taken from the Central Pollution Control Board of India (CPCB)[1]. The study in [12] reports the applicability of various feature selection and regression methods to estimate $PM_{2.5}$ in five Chinese smart cities. In [13] AQI estimation in data collected in Taiwan is done using adaptive boosting (AdaBoost), SVM, ANN, RF, and stacking ensemble. PM_{10} forecasting in a couple of cities in Poland by using linear regression, ANN and RF are available in [14].

Researchers proposed deep learning (DL) models for the prediction of air pollutants. Liu X. et al. [15] present a combination of Graphic Convolutional Networks (GCN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Q-learning to predict the distribution of $PM_{2.5}$ in data collected from Changsha, China, whereas Kristiani E. et al. [16] propose LSTM model to predict $PM_{2.5}$ concentration in air pollutants data collected from substations in Taiwan. In [17] air pollution data by the Turkey Ministry of Environment and Urbanization is analyzed using LSTM, Recurrent Neural Networks (RNN), and GRU to estimate $PM_{2.5}$. The study in [18] incorporates Exponential Adaptive Gradients (EAG) optimization with one-dimensional CNN to forecast the air pollution index in Klang, Malaysia. A deep learning model based on wavelet-packet transform is available in [19] for $PM_{2.5}$ forecasting in Qingdao, China. In [20] RF, XGBoost, and Deep Neural Net (DNN) are employed to predict $PM_{2.5}$ in Tehran, Iran.

Several studies are being conducted on how to calculate the Air Quality Index employing ML models. Measuring the Air Index Quality properly is a vital first step in the mitigation of air pollution. Accurately assessing the air quality index depends heavily on ML techniques. The majority of literature on air pollution analysis focuses on predicting only specific pollutants $PM_{2.5}$ and PM_{10}. This research focuses on overall air quality analysis considering various polluting factors. It prefers various machine learning models over the computationally costlier deep learning models for classifying air quality levels and forecasting the overall AQI of popular Indian smart cities.

III. RESEARCH METHODS

This section presents the proposed methodology and machine learning models used in this research work.

A. Proposed methodology

The proposed work on air quality analysis uses a case study of Indian smart cities. The proposed multiclassification module categorizes the air quality of those cities into six classes. However, the regression modules forecast overall AQI based on the concentration of various polluting factors. Figure 1 depicts the workflow of the proposed research work. Indian smart cities have air pollution monitoring substations at different locations. Air sensors at those substations record different polluting factors in the air. The Central Pollution Control Board of India (CPCB) maintains these records from different sources. The dataset preparation phase applies cleaning and pre-processing techniques to raw data from CPCB to obtain the experimentation dataset. Exploratory data analysis finds different features, their

[1] https://www.cpcb.nic.in/

correlation, and statistical parameters of the dataset and then selects appropriate features for building ML models. Splitting experimentation datasets into training and testing datasets and employing different classifiers and regressors training and testing phases are carried out. Referring to the experimentation results, the performance of ML models is analyzed.

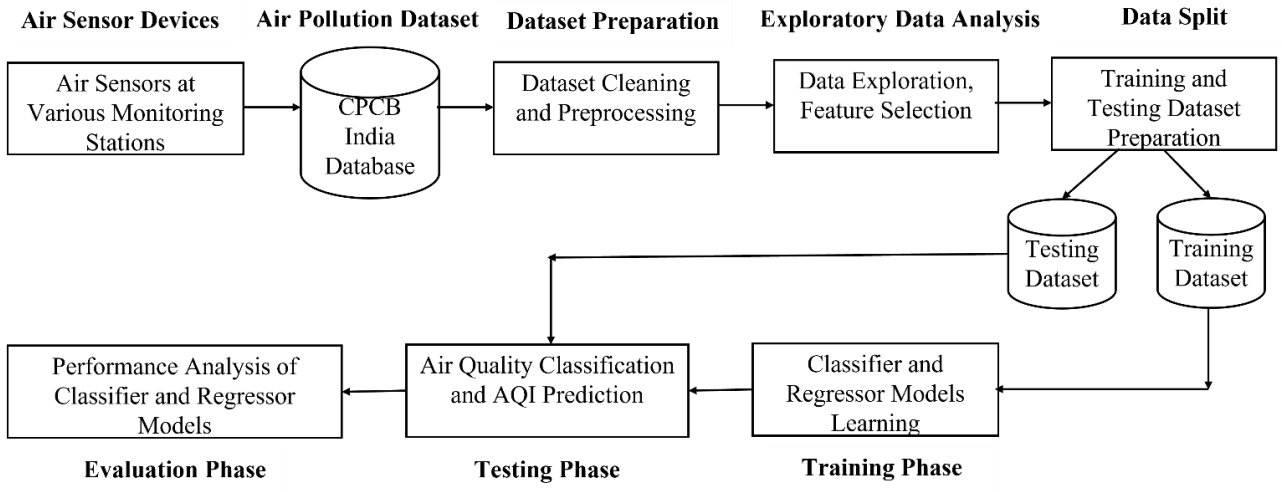

Figure 1. The proposed workflow

Table 1 presents the AQI category, polluting factors, and health breakpoints as per the National Ambient Air Quality Standards. It categorizes air quality into six classes such as, good, satisfactory, moderate, poor, very poor, and severe. Considering health breakpoints for different levels and different concentrations of polluting factors Pollutant Sub-Indices (PSIs) of pollutants are calculated. Equation 1 formulates it. Overall AQI is dependent on these PSIs.

Table 1: AQI category, polluting factors and health breakpoints

AQI Category (Range) and Health Impact	**PM_{10} $\mu g/m^3$ (24hr)**	**$PM_{2.5}$ $\mu g/m^3$ (24hr)**	**NO_2 $\mu g/m^3$ (24hr)**	**O_3 $\mu g/m^3$ (8hr)**	**CO mg/m^3 (8hr)**	**SO_2 $\mu g/m^3$ (24hr)**	**NH_3 $\mu g/m^3$ (24hr)**	**Pb $\mu g/m^3$ (24hr)**
Good (0 – 50)	0 – 50	0 – 30	0 – 40	0 – 50	0 – 1.0	0 – 40	0 – 200	0 – 0.5
Health Impact	Minimal impact on health.							
Satisfactory (51 – 100)	51 – 100	31 – 60	41 – 80	51 – 100	1.1 – 2.0	41 – 80	201 – 400	0.5 – 1.0
Health Impact	Minor breathing discomfort to sensitive people.							
Moderately polluted (101 – 200)	101 – 250	61 – 90	81 – 180	101 – 168	2.1 – 10	81 – 380	401 – 800	1.1 – 2.0
Health Impact	Breathing discomfort to children, older adults, and people with lung, and heart disease.							
Poor (201 – 300)	251 – 350	91 – 120	181 – 280	169 – 208	10 – 17	381 – 800	801 – 1200	2.1 – 3.0
Health Impact	Breathing discomfort to people on prolonged exposure.							
Very poor (301 – 400)	351 – 430	121 – 250	281 – 400	209 – 748*	17 – 34	801 – 1600	1200 – 1800	3.1 – 3.5
Health Impact	Respiratory illness in people on prolonged exposure.							

Severe (401 – 500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+
Health Impact	Respiratory effects even on healthy people.							

$$PSI_j = \left(PC_j - PLB_j\right) \times \left(\frac{AUB_i - ALB_i}{PUB_j - PLB_j}\right) + ALB_i \qquad \text{----------} \qquad 1$$

where,

j = Pollutants; $j \in \{PM_{2.5}, PM_{10}, NO_2, NH_3, SO_2, CO, O_3\}$

i = Air quality category based on AQI value; $i \in$ {Good, Satisfactory, Moderately polluted, Poor, Very poor, Severe}

PSI_j = Pollutant sub-index of j

PC_j = Pollutant concentration value of j

PLB_j = Pollutant concentration lower breakpoint i.e. $\leq PC_j$

PUB_j = Pollutant concentration upper breakpoint i.e. $>PC_j$

ALB_j = AQI lower breakpoint corresponding to PLB_j

AUB_j = AQI upper breakpoint corresponding to PUB_j

AQI is calculated by selecting the maximum value among the pollutant sub-indices PSI_j provided that there must be at least three non-zero values of pollutant sub-indices PSI_j are available and one of them must be either pollutant sub-index of $PM_{2.5}$ or PM_{10}.

B. *Machine learning models*

The proposed work employs both classification and regression models for air quality analysis in smart Indian cities. It applies classifiers AdaBoost, DT, RF, XGBoost, extremely randomized trees (ExtraTrees), and the corresponding regressors for this work.

1) *AdaBoost:* Adaptive boosting ensemble [21] starts by fitting a classification model to the initial dataset.It then fits subsequent copies of the classifier in an ensemble to the same set of data instances but with more weightage to instances that were incorrectly classified so that later classifiers would concentrate more on challenging data instances.

2) *Decision Trees (DT):* A non-parametric simple supervised learning model for classification and regression is called a decision tree (DT) [22]. It aims to learn straightforward decision rules resulting from the data characteristics in order to build a learning model that estimates the value of a target variable. DT tree is a piecewise constant approximation. From the learning data instances, decision trees are developed in a deterministic way and are trimmed using the cost-complexity pruning method with error estimates.

3) *Random Forests (RF):* Breieman L. [23] proposes an ensemble model RF to improve tree bagging. It integrates a bootstrap copy of the learning sample, the Random Subspace method [24], and the Classification and Regression Trees (CART) algorithm (without pruning) [22]. The optimum split is determined at each test node by examining a size K random subset of candidate attributes (chosen from the candidate attributes without replacement).

4) *Extreme Gradient Boosting (XGBoost):* Extreme Gradient Boosting is a robust and scalable ensemble based on a gradient boosting technique put out by Chen T. and Guestrin C. [25]. In contrast to earlier methods, it provides stronger control against overfitting by virtue of the more regularized model formalization. Learning in XGBoost is accelerated by parallel and distributed computing, which facilitates model exploration more quickly. It uses out-of-core processing to process billions of instances on a desktop for data scientists.

5) *Extremely randomized trees (ExtraTrees):* It is a meta-estimator that fits multiple randomized decision trees on different sub-samples of the dataset [26]. It employs averaging to increase prediction accuracy and reduce overfitting. It involves dividing a tree node while substantially randomly choosing attribute and cut point. In the most extreme scenario, it creates completely randomized trees with structures independent of the learning data instances' output values. It tunes the power of the randomization to the particulars of the situation by selecting the right parameter.

IV. EXPERIMENTAL FRAMEWORK

The presented study performs experimentation on the real data set of air pollutants using various classifiers and regressors to analyze air quality in Smart Indian cities. This section describes details of the dataset, the setup for experimentation carried out, and different performance metrics.

A. Datasets

CPCB collects data on air pollutants from numerous air quality sensing stations set up at different locations in Indian cities. This research prepares a real dataset of air pollutants recorded during the period of January 2021 to March 2023 in major five smart cities namely, Delhi, Mumbai, Bengaluru, Chennai, and Kolkata. The values of air pollutants are recorded by taking eight hourly averages of air pollutants at each sensing station. The dataset has eight attributes covering AQI and corresponding significant seven air pollutants, namely, $PM_{2.5}$, PM_{10}, NH_3, NO_2, CO, SO_2, and O_3. It is a multi-class dataset with six class labels such as good, satisfactory, moderately polluted, poor, very poor, and severe based on AQI. Table 2 presents the details of the air pollution real dataset used in this research. Table 3 gives the statistics of different air pollutants and AQI in the city wise datasets.

Table 2: Details of used datasets

Smart City Dataset	No. of Sensing Stations	No. of Samples	No. of Features	No. of Classes
Delhi	32	76611	8	6
Mumbai	18	40958	8	6
Bengaluru	10	17959	8	6
Chennai	8	16526	8	6
Kolkata	7	16872	8	6

Table 3: Statistics of different air pollutants and AQI in the city wise datasets

City →		Delhi		Mumbai		Bengaluru		Chennai		Kolkata	
Attr.	Unit	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
$PM_{2.5}$	$\mu g/m^3$	107.54	86.94	53.09	43.63	34.65	33.29	28.73	26.21	55.27	43.04
PM_{10}	$\mu g/m^3$	220.63	132.15	123.99	87.07	74.02	41.38	64.98	43.73	109.07	76.20
NO_2	$\mu g/m^3$	41.43	31.80	27.92	34.74	23.84	16.68	13.94	9.95	25.40	22.14
NH_3	$\mu g/m^3$	45.03	30.29	26.32	40.16	14.21	13.88	22.25	17.86	23.57	21.61
SO_2	$\mu g/m^3$	11.20	9.42	14.64	16.53	5.81	3.35	9.98	10.45	11.09	8.76
CO	mg/m^3	2.69	9.41	1.47	2.74	0.66	0.54	0.82	0.52	0.68	0.54

City →		Delhi		Mumbai		Bengaluru		Chennai		Kolkata	
O_3	$\mu g/m^3$	28.35	26.89	26.04	25.37	25.79	16.03	23.34	21.73	36.22	25.23
AQI	--	240.25	145.51	141.14	94.44	80.75	51.85	77.18	49.12	123.66	91.90

Data exploration gives the analysis of the distribution of pollutants in each city during the period 2021 to 2023. It is depicted in Figure 2. It shows that PM_{10} is the most significant pollutant in studied Indian cities. Overall pollutants concentration in different Indian cities from 2021 to 2023 is available in Figure 3. $PM_{2.5}$, PM_{10}, NO_2, NH_3, and CO have the highest distribution in Delhi and secondly in Mumbai. However, the distribution of pollutants SO_2 and O_3 is balanced in mentioned cities - Delhi, Mumbai, Bengaluru, Chennai, and Kolkata.

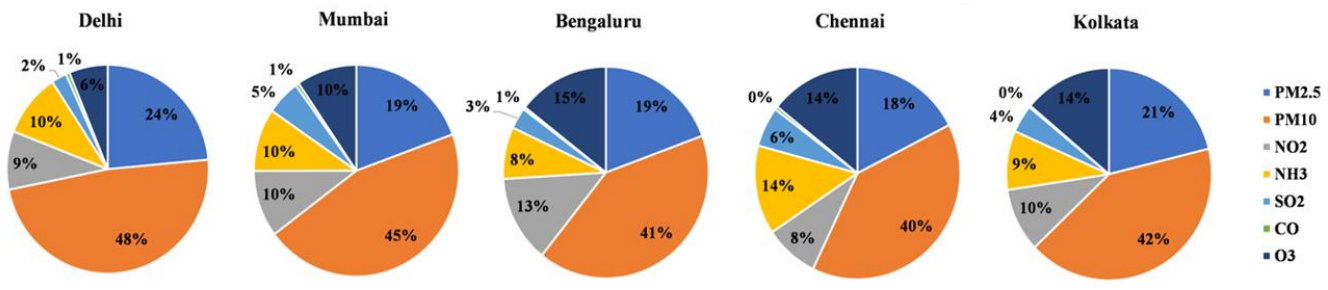

Figure 2. Air pollutants distribution in different Indian cities from 2021-2023

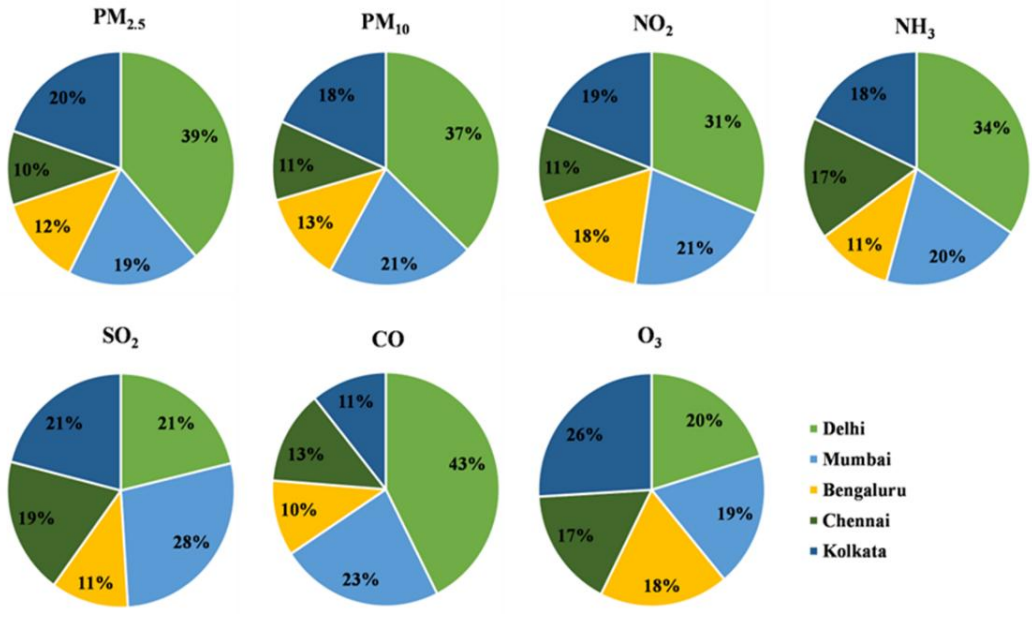

Figure 3. Overall pollutants concentration in different Indian cities from 2021-2023

B. Experimentation setup

The real dataset of air pollutants of each of the five Indian smart cities is split into training and testing datasets with a ratio of 70:30. Classifiers AdaBoost, DT, RF, XGBoost, and ExtraTrees

categorize air quality of the city into six classes, however, their corresponding regressors AdaBoostR, DTR, RFR, XGBoostR, and ExtraTreesR, respectively predict the AQI value of each city based on the distribution of air pollutants in ambient air. The implementation in this study is carried out using the Scikit-learn library [27] in Python.

C. *Evaluation metrics*

Using a variety of metrics this study compares the performance of employed ML models for air quality analysis.

1) *Classification:* This research performs the classification of multi-class datasets to categorize air quality. To measure the classification performance it uses metrics like accuracy ($(TP+TN)/(TP+TN+FP+FN)$), recall ($TP/(TP+FN)$), precision ($TP/(TP+FP)$), and F1-measure ($2(Recall \cdot Precision)/(Recall+Precision)$) where *TP* is true positive samples, *TN* is true negative samples, *FP* is false positive samples and *FN* is false negative samples in the binary classification confusion matrix. However, to evaluate the results of multilabel classification it extends the binary classification metrics by viewing the data as a set of binary problems, one for each class, and then computes the average of binary classification metrics across the set of classes. This research applies macro-averaging of binary metrics by simply computing their mean value while giving each class the same weight.

2) *Regression:* This study performs regression to forecast the value of AQI based on historic data. To evaluate the prediction performance of regressors it uses metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Absolute Error (RMSE), and coefficient of determination (R^2). Equations 2, 3, 4, and 5 formulate MAE, MSE, RMSE, and R^2 respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|A_i - P_i| \quad \text{----------} \quad 2$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2 \quad \text{----------} \quad 3$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2} \quad \text{----------} \quad 4$$

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(A_i-\bar{A})(P_i-\bar{P})]}{\sigma_A\sigma_P}\right]^2 \quad \text{----------} \quad 5$$

where,
N is the number of data instances
A_i is the actual value of i^{th} instance, $i = 1, 2, \ldots, N$
P_i is the predicted value of i^{th} instance, $i = 1, 2, \ldots, N$
$\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A_i$ i.e., mean of actual values of N instances
$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$ i.e., mean of predicted values of N instances
σ_A is the standard deviation of actual values of N instances
σ_P is the standard deviation of predicted values of N instances.

V. RESULTS AND DISCUSSION

The experimentation results are available for two tasks carried out for air quality analysis: 1) air quality classification and 2) AQI prediction. This section presents the empirical analysis of the work.

A. *Air quality classification*

The study investigates the multilabel classification performance of AdaBoost, DT, RF, XGBoost,

and ExtraTrees using metrics accuracy, recall, precision, and F1-measure on real datasets of five Indian metro cities. Tables 4, 5, 6, and 7 respectively present the empirical results of accuracy, recall, precision, and F1-measure of five classifiers on different datasets. They also mention the ranks ranging from 1 to 5 of each classifier based on the average value of each metric. Rank 1 indicates the best performer while rank 5 indicates the worst performer among the five classifiers.

Table 4: Accuracy of different classifiers

Dataset	**AdaBoost**	**DT**	**RF**	**XGBoost**	**ExtraTrees**
Delhi	0.8221	0.9997	0.9997	0.9998	0.9965
Mumbai	0.9135	0.9997	0.9993	0.9996	0.9941
Bengaluru	0.7706	0.9983	0.9991	0.9987	0.9915
Chennai	0.9899	0.9994	0.9994	0.9994	0.9921
Kolkata	0.9872	0.9976	0.9974	0.9980	0.9929
Mean	**0.8966**	**0.9990**	**0.9990**	**0.9991**	**0.9934**
Rank	**5**	**3**	**2**	**1**	**4**

Table 5: Recall of different classifiers

Dataset	**AdaBoost**	**DT**	**RF**	**XGBoost**	**ExtraTrees**
Delhi	0.6650	0.9994	0.9993	0.9996	0.9948
Mumbai	0.6627	0.9981	0.9978	0.9997	0.9852
Bengaluru	0.6564	0.9926	0.9722	0.9972	0.9723
Chennai	0.9946	0.9997	0.9847	0.9997	0.9512
Kolkata	0.8368	0.9667	0.9772	0.9981	0.9731
Mean	**0.7631**	**0.9913**	**0.9862**	**0.9989**	**0.9753**
Rank	**5**	**2**	**3**	**1**	**4**

Table 6: Precision of different classifiers

Dataset	**AdaBoost**	**DT**	**RF**	**XGBoost**	**ExtraTrees**
Delhi	0.6177	0.9995	0.9993	0.9996	0.9952
Mumbai	0.6010	0.9996	0.9993	0.9996	0.9939
Bengaluru	0.5745	0.9990	0.9923	0.9992	0.9860
Chennai	0.9947	0.9997	0.9928	0.9995	0.9799
Kolkata	0.9885	0.9858	0.9979	0.9890	0.9940
Mean	**0.7553**	**0.9967**	**0.9963**	**0.9974**	**0.9898**
Rank	**5**	**2**	**3**	**1**	**4**

Table 7: F1-measure of different classifiers

Dataset	**AdaBoost**	**DT**	**RF**	**XGBoost**	**ExtraTrees**
Delhi	0.6345	0.9995	0.9993	0.9996	0.9950
Mumbai	0.6253	0.9988	0.9985	0.9996	0.9894
Bengaluru	0.6087	0.9957	0.9814	0.9982	0.9788

Dataset	AdaBoost	DT	RF	XGBoost	ExtraTrees
Chennai	0.9946	0.9997	0.9883	0.9996	0.9636
Kolkata	0.8436	0.9756	0.9868	0.9934	0.9829
Mean	**0.7413**	**0.9939**	**0.9909**	**0.9981**	**0.9819**
Rank	**5**	**2**	**3**	**1**	**4**

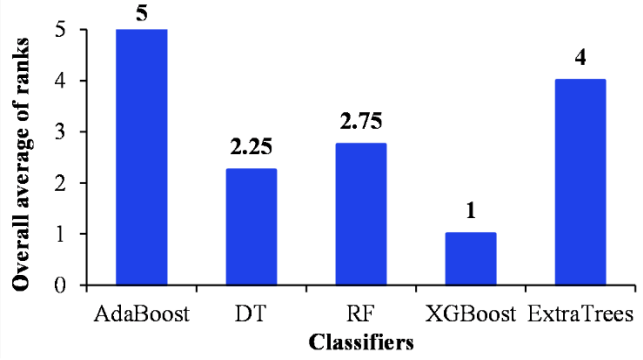

Figure 4. Overall average ranks of all classifiers

The study computes the overall average rank of five classifiers by taking the mean of their performance ranks based on all metrics. Figure 4 depicts the overall average ranks of all classifiers. XGBoost ensemble performs the best among the five ML models in classifying the air quality of cities based on air quality sensor data.

B. Air Quality Index Prediction

The paper evaluates the prediction performance of regressors AdaBoostR, DTR, RFR, XGBoostR, and ExtraTreesR by measuring MAE, MSE, RMSE, and R^2. Tables 8, 9, 10, and 11, respectively show the empirical findings of metrics MAE, MSE, RMSE, and R^2 for five regressors on various datasets. They also indicate the ranks of regressors which range from 1 to 5, according to the average value of each statistic.

Table 8: Mean Absolute Error (MAE) of different regressors

Dataset	AdaBoostR	DTR	RFR	XGBoostR	ExtraTreesR
Delhi	61.429	0.491	0.389	1.720	0.407
Mumbai	44.261	0.760	0.575	1.772	0.562
Bengaluru	44.837	0.726	0.561	1.107	0.546
Chennai	26.843	0.651	0.514	1.291	0.518
Kolkata	16.563	0.668	0.523	1.070	0.594
Mean	**38.786**	**0.659**	**0.512**	**1.392**	**0.525**
Rank	**5**	**3**	**1**	**4**	**2**

Table 9: Mean Squared Error (MSE) of different regressors

Dataset	AdaBoostR	DTR	RFR	XGBoostR	ExtraTreesR
Delhi	4506.436	29.461	20.261	23.587	7.752
Mumbai	2517.264	31.772	20.641	12.532	11.614

Dataset	AdaBoostR	DTR	RFR	XGBoostR	ExtraTreesR
Bengaluru	2241.561	40.342	25.925	7.346	11.414
Chennai	908.950	19.105	18.354	9.829	9.910
Kolkata	404.054	46.240	29.411	19.616	30.694
Mean	**2115.653**	**33.384**	**22.918**	**14.582**	**14.277**
Rank	**5**	**4**	**3**	**2**	**1**

Table 10: Root Mean Squared Error (RMSE) of different regressors

Dataset	AdaBoostR	DTR	RFR	XGBoostR	ExtraTreesR
Delhi	67.130	5.428	4.501	4.857	2.784
Mumbai	50.172	5.637	4.543	3.540	3.408
Bengaluru	47.345	6.352	5.092	2.710	3.378
Chennai	30.149	4.371	4.284	3.135	3.148
Kolkata	20.101	6.800	5.423	4.429	5.540
Mean	**42.979**	**5.717**	**4.769**	**3.734**	**3.652**
Rank	**5**	**4**	**3**	**2**	**1**

Table 11: R^2 of different regressors

Dataset	AdaBoostR	DTR	RFR	XGBoostR	ExtraTreesR
Delhi	0.781	0.999	0.999	0.999	1.000
Mumbai	0.710	0.996	0.998	0.999	0.999
Bengaluru	0.042	0.983	0.989	0.997	0.995
Chennai	0.598	0.992	0.992	0.996	0.996
Kolkata	0.954	0.995	0.997	0.998	0.996
Mean	**0.617**	**0.993**	**0.995**	**0.998**	**0.997**
Rank	**5**	**4**	**3**	**1**	**2**

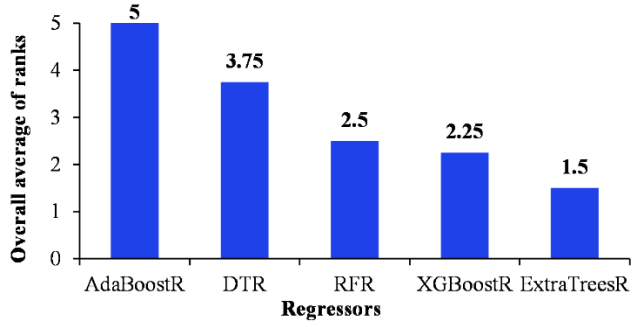

Figure 5. Overall average ranks of all regressors

Figure 5 shows the overall average ranks of five regressors by computing the average of their performance rankings based on MAE, MSE, RMSE, and R^2. The ExtraTreesR regression model

outperforms the other five regressors in predicting the AQI of cities on the basis of historical data on air pollution.

VI. CONCLUSION

For a better life, environmental sustainability is one of the significant aspects of a smart city. In order to mitigate the negative effects of air pollution on public health, social planning, and management require the ability to predict the AQI. This investigation presents the applicability of ML models for analyzing air pollutants sensor data collected from five major Indian smart cities during 2021-2023. It applies multilabel classification to categorize the air quality of a city by focusing on AQI and pollutants $PM_{2.5}$, PM_{10}, NH_3, NO_2, CO, SO_2, and O_3. Also, the work employs regression to forecast the AQI value of metropolises. The presented research work on air quality analysis reports a comparative study of the learning performances of various classifiers and regressors on a variety of metrics. The classifier XGBoost and regressor ExtraTreesR perform the best in this investigation.

REFERENCES

1. M. Swilling et al., "The weight of cities resource requirements," International Resource Panel (IRP) Report, United Nations Environment Programme, 2018.
2. "The sustainable development goals report 2022," 2022. [Online]. Available: https://unstats.un.org/sdgs/report/2022/
3. A. M. Toli and N. Murtagh, "The concept of sustainability in smart city definitions," Front. Built Environ., 6, Jun. 2020, doi: 10.3389/FBUIL.2020.00077/BIBTEX.
4. S. E. Bibri and J. Krogstie, "Smart sustainable cities of the future: An extensive interdisciplinary literature review," Sustain. Cities Soc., 31, May 2017, pp. 183–212, doi: 10.1016/J.SCS.2017.02.016.
5. L. Pimpin et al., "Estimating the costs of air pollution to the National Health Service and social care: An assessment and forecast up to 2035," PLoS Med., 15(7), Jul. 2018, doi: 10.1371/JOURNAL.PMED.1002602.
6. P. Orellano, J. Reynoso, and N. Quaranta, "Effects of air pollution on restricted activity days: systematic review and meta-analysis," Environ. Heal., 22(1), Dec. 2023, pp. 31, doi: 10.1186/S12940-023-00979-8.
7. "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," 2021. https://www.who.int/publications/i/item/9789240034228 (accessed Jun. 08, 2023).
8. S. Abu El-Magd, G. Soliman, M. Morsy, and S. Kharbish, "Environmental hazard assessment and monitoring for air pollution using machine learning and remote sensing," Int. J. Environ. Sci. Technol., 20, 2023, pp. 6103–6116, doi: 10.1007/S13762-022-04367-6.
9. A. Sanapala, B. J. Lakshmi, and K. B. Kundra, K. SandhyaRani Madhuri, "Air pollution detection and control system using ML techniques," Int. J. Recent Innov. Trends Comput. Commun., 11(4), pp. 220–225, 2023, Accessed: May 10, 2023. [Online]. Available: https://ijritcc.org/index.php/ijritcc/article/view/6442/5827
10. K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," Int. J. Environ. Sci. Technol., 20, May 2023, pp. 5333–5348, doi: 10.1007/S13762-022-04241-5/TABLES/7.
11. S. Ketu, "Spatial air quality index and air pollutant concentration prediction using linear regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India," Nat. Hazards, 114(2), Nov. 2022, pp. 2109–2138, doi: 10.1007/S11069-022-05463-Z.
12. A. Banga, R. Ahuja, and S. C. Sharma, "Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities," Int. J. Syst. Assur. Eng. Manag., 2021, doi: 10.1007/S13198-020-01049-9.
13. Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, "Machine learning-based prediction of air quality," Appl Sci, 10(9151), Dec. 2020, pp. 1–17, doi: 10.3390/app10249151.

14. K. Karatzas, N. Katsifarakis, C. Orlowski, and A. Sarzyński, "Revisiting urban air quality forecasting: a regression approach," Vietnam J. Comput. Sci., 5(2), May 2018, pp. 177–184, doi: 10.1007/S40595-018-0113-0.
15. X. Liu, M. Qin, Y. He, X. Mi, and C. Yu, "A new multi-data-driven spatiotemporal PM2.5 forecasting model based on an ensemble graph reinforcement learning convolutional network," Atmos. Pollut. Res., 12(10), Oct. 2021, pp. 101197, doi: 10.1016/J.APR.2021.101197.
16. E. Kristiani et al., "Short-term prediction of PM2.5 using LSTM deep learning methods," Sustainability, 14(4), Feb. 2022, pp. 2068, doi: 10.3390/SU14042068.
17. Y. A. Ayturan et al., "Short-term prediction of PM2.5 pollution with deep learning methods," Glob. NEST J., 22(1), 2020, pp. 126–131.
18. M. G. Ragab et al., "A novel one-dimensional CNN with exponential adaptive gradients for air pollution index prediction," Sustainability, 12(23), Dec. 2020, pp. 10090, doi: 10.3390/SU122310090.
19. Q. Zheng et al., "Application of wavelet-packet transform driven deep learning method in PM2.5 concentration prediction: A case study of Qingdao, China," Sustain. Cities Soc., 92, May 2023, pp. 104486, doi: 10.1016/J.SCS.2023.104486.
20. M. Z. Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2.5 prediction based on Random Forest, XGBoost, and deep learning using multisource remote sensing data," Atmosphere (Basel)., 10(7), pp. 373, Jul. 2019, doi: 10.3390/ATMOS10070373.
21. Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., 55(1), 1997, pp. 119–139, doi: 10.1006/jcss.1997.1504.
22. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees. Taylor & Francis, 1984. Accessed: Apr. 14, 2023. [Online]. Available: https://doi.org/10.1201/9781315139470
23. L. Breiman, "Random forests," Mach. Learn., 45(1), Oct. 2001, pp. 5–32, doi: 10.1023/A:1010933404324/METRICS.
24. T. K. Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell., 20(8), 1998, pp. 832–844, doi: 10.1109/34.709601.
25. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.
26. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., 63(1), Apr. 2006, pp. 3–42, doi: 10.1007/S10994-006-6226-1/METRICS.
27. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., 12(85) , 2011, pp. 2825–2830, Accessed: Jun. 22, 2023. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html