IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Aqi Prediction Using Classification Techniques: A Case Study Of Delhi And Kolkata

Jasleen Bhatia, Amanpreet Kaur*

Assistant Professor, Assistant Professor

Information Technology, Computer Science

Guru Tegh Bahadur Institute of Technology, New Delhi, India

Abstract: The rapid development in commercial and social areas of modern society has affected the air quality adversely. The pollutant concentration due to industries and transport continue to grow and have aneffect on human life. Thus, tracking air quality index is essential in developing countries so that measures to curb air pollution could be tackled. In this research paper the primary aim is to anticipate the levels of Air Quality Index using classification techniques namely Decision Tree, Random Forest, Linear Regression and Support Vector Machines. For the experimental work the dataset of Delhi: Ashok Vihar and 2 stations of Kolkata: Ballygunge and Victoria has been gathered from Central Pollution Control Board Website. It has been concluded that Random Forest depicts maximum performance than the other classification techniques based on accuracy, precision and recall.

***Index Terms*:** Classification , Support Vector Machine, Decision Tree, , Linear Regression, Random Forest, Air Quality Index

1. INTRODUCTION

Air pollution refers to presence of harmful and excessive substances called pollutants leading to contamination of air. Natural composition of air is changed due to the presence of pollutants[1]. Variety of human activities that contribute to poor air quality are: growth withinside the variety of motorized vehicles, emissions from the combustion of fossil fuels, industrial plants, automobiles, environmental problems like global warming, acid rain, decreased visibility, smog, weather change, are liable for multiplied greenhouse gases, environmental degradation[2]. These dangerous pollutants are hazardous to human health leading to respiratory and cardiovascular problems, premature mortality, impaired lung function, cancer, neurological effects, reproductive and developmental effects. The Air Quality Index (AQI) serves as a communication device to inform the public about air quality conditions. It simplifies complicated facts concerning the awareness of various air pollutants right into a unattached value. This manual is crucial for alerting the common population concerning the harmful effects associated with elevated AQI levels[3].The AQI tool condenses estimation of various impurities (such as, $PM_{2.5}$, NH_3, PM_{10}, Ozone(03), SO_2, NO, CO and NO_2) into a single value for assessing air quality. The AQI is derived by initially calculating sub-indices for individual pollutants, which are then combined using weighted additive methods. In India, the AQI is categorized into six levels: Severe (401–500), Very Poor (301–400), Poor (201–300), Moderately Polluted (101–200), Satisfactory (51–100), and Good (0–50) [4,5].The Air Quality Index (AQI) is a critical measure directly impacting public health, with higher levels indicating increased risk due to pollutant exposure.

Researchers are actively working to forecast AQI levels, particularly challenging in urban areas with growing industrial and motor activities. Ravindiran et al. models for applied machine learning, such as CatBoost, AdaBoost, LightGBM, XGBoost and Random Forest to predict AQI in Visakhapatnam,Andhra Pradesh, India from July 2017 to September 2022 aiming at10 meteorological and 12 contaminants parameters. Among these models, CatBoost outperformed others while AdaBoost showed the least effective prediction[11].Samad et al. adopted 5 machine learning techniques specifically: random forest,ridge regressor, extreme gradient boosting, , support vector regressor and extra trees regressor at two locations in Stuttgart (Am Neckartor and Marienplatz) for predicting air pollutant concentrations (NO_2, $PM_{10,}$ $PM_{2.5}$), data incorporated from monitoring stations including meteorological parameters, information on traffic and data on pollutant ,from January 01, 2018, to March 31, 2022. It was concluded that the data notably enhance prediction accuracy [1].Verma et al. analysed trends in air quality between 2015-2020 in Tamil Nadu's industrial cluster (Thoothukudi, Cuddalore, Manali) and important Indian urban areas (Delhi,Mumbai, Chennai, Kolkata), where coastal industrial areas depicted less seasonal variation as compared to non-coastal areas. Delhi exhibit themaximum NO_2 , PM_{10}, $PM_{2.5}$ levels, with the highest SO_2 levels found in Mumbai[7].Veljanovska & Dimoski made comparability between three various machine learning algorithms including:k-nearest neighbour ,decision treeand neural network. Datasetwas collected from Macedonia in 2016.It was found that neural network algorithm is most effective with 92.3% accuracy and decision tree algorithm has lowest accuracy of 81%[8].Sethi & Mittal has foreseen different classification techniques named Linear Model, SVM and Decision Trees, and three different ensemble techniques named Stacking ensemble, Random Forest and Voting ensemble and four distinct methods for regression analysis named Quantile Regression, Stepwise Regression, Linear Regression and SVR for predicting AQI from Faridabad and Haryana.It was noticed that F1 score has least error rate and Stacking ensemble has maximum accuracy and ensemble techniques performed best in the ensemble category. SVR has maximum min max accuracy and minimum MAPE for regression techniques. When comparing decision trees to other classification techniques, they offer the highest accuracy and lowest error rate [4].Hussain et al. applieddecision tree (DT), supervised machine-learning (ML) technique, for calculating AQ. A study was conducted to analyse air quality (AQ) and its relationship with climate variables (wind speed, relative humidity, rainfall and temperature), using MATLAB 2018b. Daily conclusions spanning seven years (2014 to 2020) revealed consistently toxic AQ levels (65% to 75% per year) across the chosen urbanite areas. Using a Decision Tree (DT) as a anticipating model, data contents from Dhaka were used for validation (20%) and fortraining (80%)**,** achieving an impressive accuracy of 98.8%. The DT model was then utilized to predict recurrent air quality for Chittagong**,** demonstrating a consistency of ≥97%.Additionally, AQ predictions for eight cities in the year 2020 were made with a resulting accuracy of 96%[9].Liu et al. carried out prediction using Support vector regression (SVR) technique in three cities namely, Tianjin, Beijing and Shijiazhuang from the Jingjinji Region in China. The study demonstrated that the results incorporated from multiple cities and meteorological parameters showed more reliable forecast accuracy in prediction results[10].

Despite extensive research on air quality in developing nations where pollutants like $PM_{2.5}$ are concentrated, there's a notable gap in analysis for countries like India. Recognizing this, there's an urgent need for tailored analysis and prediction of AQI in the Indian context [6]. In this paper, the focus is to predict the AQI using classification techniques for two cities of India. The section that follows describes theMethods and Materials used in this research paper. In Section 3, the methods of machine learning used to forecast the AQI are outlined. The AQI prediction methodologyis presented in Section 4. Section 5 has a discussion of the performance evaluation results for the different classifications. Section 6 provides a precise interpretation of the paper.

2. MATERIALS AND METHODS USED

2.1 Input Parameters

$PM_{2.5}$or Particulate Matterare minisculeparticles with a diameter less than 2.5 micrometre. Due to their small size, these particles can evade the body's bloodstream, causing health issues likereduced lung function, heart problems, worsened asthma and early death. They are constantly released during the burning of fossil fuels in various settings such as vehicles, power plants, industries, and home heating systems[25].PM_{10} particles can originate from both human activities and natural events, having significant health and environmental impacts. They can irritate eyes, nose, and throat, worsen health conditions, and reduce visibility. Sources of PM_{10} include manufacturing, construction, mining, traffic, and agricultural operations, as well as natural events such as wildfires, volcanic eruptions, and windblown dust [12]. Nitric oxide (NO) is a colourless, odourless

gas composed of nitrogen and oxygen molecules. It is a naturally occurring component of the atmosphere and is produced by various natural processes, including lightning and microbial activity in soils. In the presence of sunlight and other pollutants, NO can react with other compounds to form nitrogen dioxide (NO_2) and ozone (O_3), both of which are harmful to human health and contribute to air pollution [12,13].NH3, or ammonia, is naturally produced through biological processes like animal waste decomposition and soil microbial activity. Elevated atmospheric NH3 levels, can contribute to acid rain, harming ecosystems and human health.Sulphur dioxide (SO_2) significantly impacts AQI readings due to its effects on air quality and health. Inhalation can cause respiratory issues, exacerbate asthma, and increase susceptibility to infections. SO_2 directly contributes to pollution levels, influences particulate matter formation, and interacts with other pollutants, all affecting AQI.Carbon monoxide (CO) binds to haemoglobin, reducing oxygen transport and causing hypoxia. Short-term exposure leads to symptoms like headaches and nausea, while prolonged exposure can cause severe health issues, including cardiovascular problems and death. Monitoring and controlling CO emissions are crucial for safeguarding public health.Ozone (O_3) is a pale blue gas with a sharp smell, found in both the upper atmosphere and the ground.It protects Earth from damaging UV radiation in the upper atmosphere, preventing skin cancer and other health issues. However, at ground level, ozone contributes to smog formation and can also damage crops, forests, and other vegetation. While upper atmospheric ozone is beneficial, ground-level ozone is a pollutant. Efforts to reduce ground-level ozone typically involve reducing release of its initial contaminants, for instance volatile organic compounds and nitrogen oxides, through regulations and technological advancements[23].Benzene is a colourless, flammable liquid with a sweet odour, made up of of six ringed carbon atomswith a hydrogen atom tacked to each one. It's a natural component of crude oil and widely used in producing plastics, rubber, detergents, and pharmaceuticals, as well as present in gasoline. Long-term exposure is linked to cancer, especially leukemia, while short-term exposure causes dizziness, headaches, drowsiness, and even unconsciousness. Regulations aim to limit exposure in workplaces and the environment due to its health risks.Metrological Parameters- Meteorological parameters are variables that describe the state of the atmosphere and its behaviour. These parameters are measured to understand weather patterns, atmospheric conditions, and climate.The amount of water vapor in the air in relation to its maximum capacity at a specific temperature and pressure is measured by relative humidity, or RH. High RH promotes particle formation, increasing PM levels like $PM_{2.5}$, affecting the Air Quality Index (AQI). RH also influences chemical reactions, leading to secondary pollutant formation, potentially raising AQI. Moderate RH aids in dispersing pollutants, lowering AQI, while low RH enhances evaporation. Additionally, RH impacts visibility by altering light absorption, affecting AQI readings and public perception of air quality [13].Wind Speed (WS) measures the horizontal movement of air past a fixed point and influences weather, climate, and human activities. While not directly factored into Air Quality Index (AQI) calculations, WS indirectly impacts AQI by affecting how pollutants disperse. Higher WS can disperse pollutants, potentiallyreducing ground-level concentrations and lowering AQI levels. Conversely, low WS can lead to stagnant air, allowing pollutants to accumulate and elevate AQI levels. Despite its indirect role in AQI, WS is crucial in determining air quality by influencing pollutant movement in the atmosphere. AQI primarily considers pollutant concentrations alongside factors like, ozone, nitrogen, carbon monoxide, PM or particulate matter and sulphur dioxide [14].

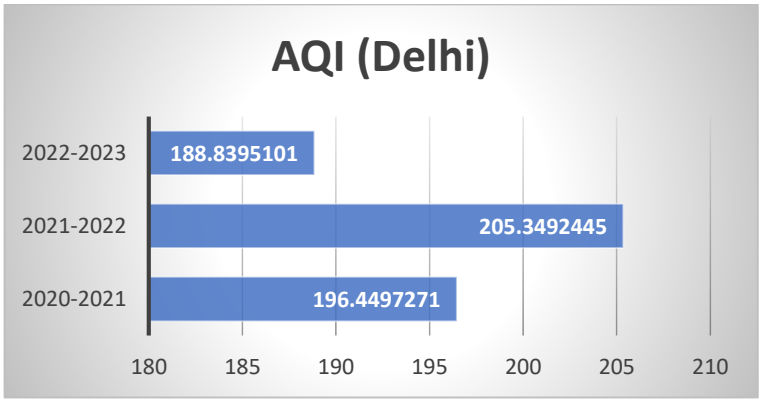

Figure 1: Bar chart of Delhi with their average AQI values from 2020 to 2023

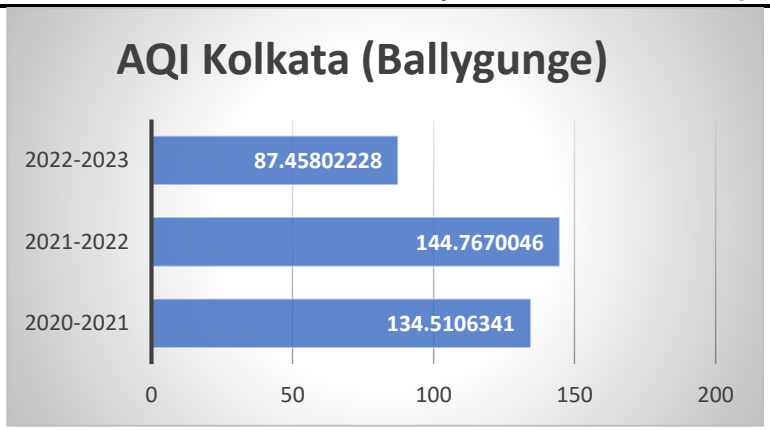

Figure 2: Bar chart of Kolkata (Ballygunge) with their average AQI values from 2020 to 2023

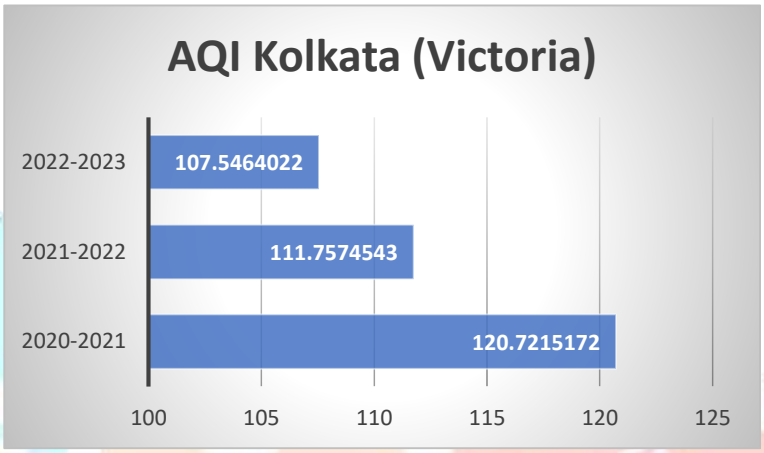

Figure 3: Bar chart of Kolkata (Victoria)with their average AQI values from 2020 to 2023

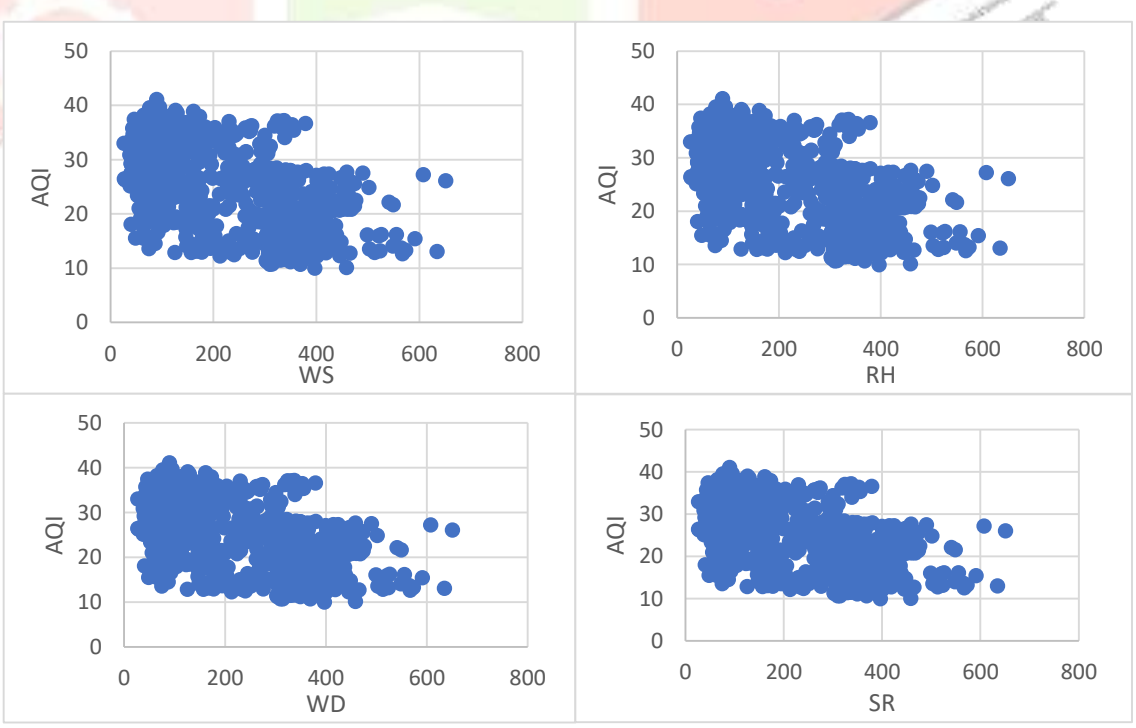

Figure 4: Scatter plot between AQI and Metrological Parameters at station Delhi

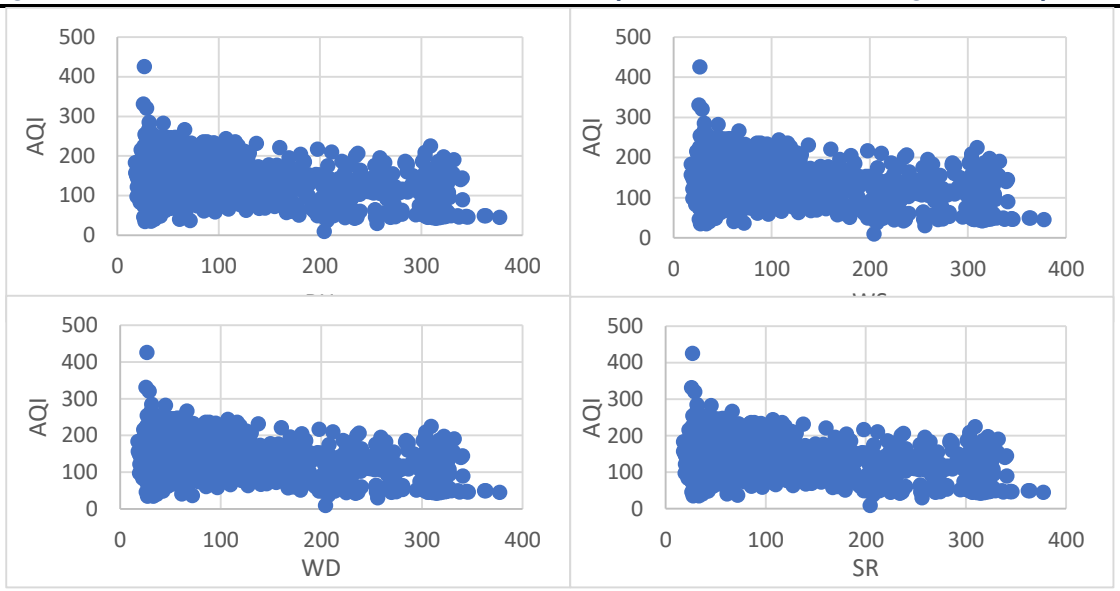

Figure 5: Scatter plot between AQI and Metrological Parameters at station Kolkata (Ballygunge)

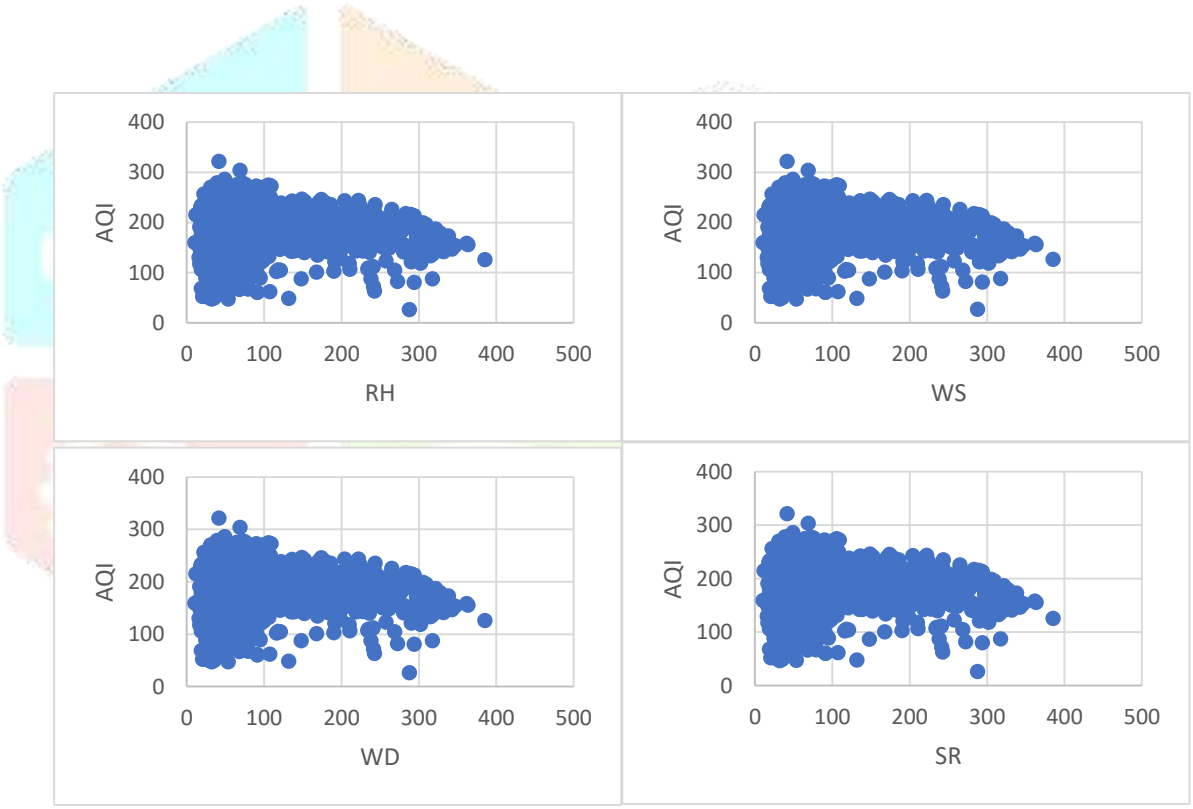

Figure 6: Scatter plot between AQI and Metrological Parameters at station Kolkata (Victoria)

3. MACHINE LEARNING TECHNIQUES TO ESTIMATE AIR QUALITY INDEX

3.1 Decision Tree:

Decision Trees are classification tools that utilize nodes to denote parameters and branches to indicate parameter values, allowing for hierarchical division of data. The construction of these classifiers depends on the tree's architecture, selected feature subsets, and the decision rules applied at each node. Design considerations encompass reducing error rates, managing the number of nodes, and optimizing information gain. Methods such as branch and bound and greedy algorithms help in selecting feature subsets. Decision rules, typically based on entropy and information, direct the partitioning of data. Overall, Decision Tree

design involves balancing structure, feature selection, and decision rules to efficiently partition data for accurate predictions while managing complexity [16].

$$IG(A_P, g) = I(A_p) - \left(\frac{X_{left}}{X_p} I(A_{left}) + \frac{X_{right}}{X_p} I(A_{right}) \right) \quad (1)$$

Here, Xp is the total number of samples at the parent node, Xleft and Xright are the number of samples in the child nodes, I is the impurity measure, and the feature to carry out the split is called g. The datasets of the parent and child nodes are Ap, Aleft, and Aright.[15,16].

3.2 Linear Regression:

A model with a linear relationship between one dependent output variable (y) and all independent input variables (x) is called a linear regression. One input variable (x) is used in simple linear regression, whereas multiple input variables (x) are used in multiple linear regression to predict a single output variable (y).

$$\hat{q} = m_0 + m_1 Y_1 + m_2 Y_2 + \ldots + m_p Y_p \quad (2)$$

Here, the dependent variable that requires prediction is q and Y1, Y2, Y3, Y4, ..., Yp are the independent variables or features. The coefficients m0, m1, ..., mp represent the regression coefficients [15,17].

3.3 Support Vector Machines:

To predict Air Quality Index (AQI) using Support Vector Machines (SVM), you gather historical data on factors like pollutant concentrations, weather conditions, and geographic features. After cleaning and preprocessing the data, select relevant features and engineer new ones if needed[24]. SVM are a family of supervised learning methods utilized for regression and classificationtasks. By avoiding overfitting, they improve predictive accuracy by utilizing machine learning principles. Originally designed for classification, SVMs have expanded to tackle regression problems as well. Their versatility has led to widespread adoption across various domains, including handwriting and facial analysis. SVMs excel in pattern classification and regression-based applications, offering robust solutions for diverse predictive tasks[15,18].

Using a fixed (non-linear) mapping, the input A is first mapped onto an n-dimensional feature space in SVM regression. A linear model is then built in this feature space. The linear model (in the feature space)p(A, α) is given by using mathematical notation.

$$p(A, \alpha) = \sum_{i=1}^{n} \alpha_i g_i \ (A) + b \quad (3)$$

3.4 Random Forest:

An ensemble technique for classification and regression problems is called Random Forest. It's applied to detect air pollutants such asnitrogen dioxide, carbon monoxide, sulphur dioxide, carbon dioxide, ozone, $PM_{2.5}$, and PM_{10}. Additionally, to evaluate the quality of the air, it uses a variety of datasets, including road, traffic, point-of-interest, and meteorological data [16].

Using data samples, the supervised classification algorithm Random Forest creates decision trees. It employs an ensemble approach, where each tree contributes to predictions through voting or bagging. Bagging involves dividing the training data into several random subsets, which are then used to train different decision trees[15,19].

$$p(a) = Y0(a) + Y1(x) + Y2(x) \ldots \quad (4)$$

Using this method, the sum of the simple base models, Y(i), equals the final model, p. Every base classifier consists of a basic decision tree. The term "model ensemble" describes this general strategy of combining multiple models to improve prediction accuracy[16].

4 METHODOLOGY USED:

The primary objective of this study is to estimate the pollution levels at particular locations. Following data collection, an initial examination of the pollutants has been conducted to detect outliers and missing values. Subsequently, an in-depth investigation of the relevant features and how they relate to pollutant concentrations has been observed. Next, the data was divided into testing and training sets. Machine learning models have then been trained with selected hyperparameter settings. The following Fig. 1 shows the methodological steps of the chosen process.

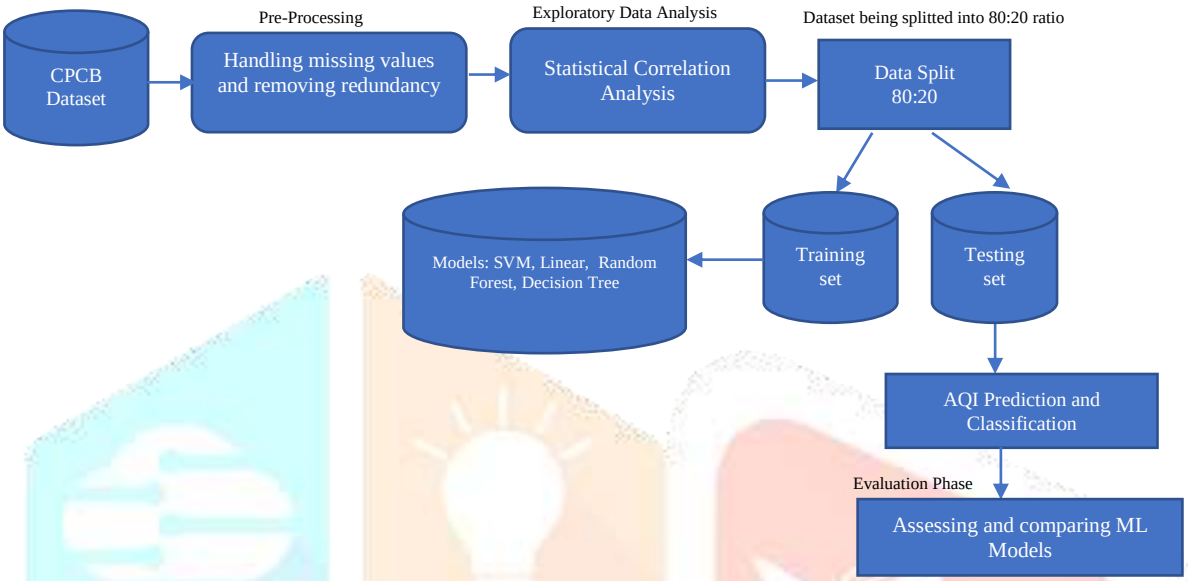

Figure 7: Flowchart of the proposed model

4.1 Study Area

The current study examines data on air pollution from the Central Pollution Control Board (CPCB) in India.According to the global pollution database maintained by the World Health Organization (WHO), India has the greatest number of polluted cities globally. One of the cities mentioned earlier in the database is Delhi, which is located 28.7 degrees N, 77 degrees East with an area of 1,483 km^2. The city of Kolkata, located in West Bengal, India, at latitude 22°33′36.00 North and longitude 88°25′12.00 East, is also included in the study area. The maps for the study area has been shown in Fig. 2 [21,22].

Figure 8: Study area map of station Delhi (Ashok Vihar)

Figure 9: Study area map of station Kolkata (Ballygunge and Victoria)

4.2 Data Pre-processing:

The durationof the research was spanned from November 2020 to November 2023 and it includes 17 features and 1010, 917 and 1009 instances from 2 different cities namely Delhi and Kolkata. The dataset has been collected from one station of Delhi: Ashok Vihar and 2 stations of Kolkata: Ballygunge and Victoriarespectively. The independent variables in each dataset include several pollutants likePM_{10}, $PM_{2.5}$ SO_2,NO_2,Ozone etc. and metrological parameters such as relative humidity, wind speed, temperature etc. The dependent variable in each case is AQI which is computed from these pollutants.After gathering data, the next step is to pre-process it, which is essential for evaluating its quality. The pre-processing steps reduce noise in the data, thus improving the processing speed and the generalization performance of machine learning algorithms. Outliers and missing data are the most frequent issues encountered in data extraction and monitoring applications. To pre-process the dataset, the missing data has been removed [1,2,15].

4.3 ***Technique Used****:*

Further, four machine learning techniques namely Support Vector Machines (SVM), Decision Tree, Random Forest, Linear Regression has been applied on the dataset of Delhi and Kolkata. These models performance has been assessed using a number of metrics, including accuracy, recall, and precision.

4.3.1 Precision: A useful metric for assessing a model's accuracy in making positive predictions is precision, which computes the ratio of correctly predicted positive outcomes to all instances predicted as positive.

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \qquad \textbf{(5)}$$

Where:

- TP (True Positives): The number of positive samples that were accurately predicted to be favourable. They are also known as Real Positives.
- FP (False Positives): The quantity of negative samples that are mistakenly reported as favourable.

*4.3.2 Recall:*By dividing the total number of correctly identified positive instances by the total number of actual positive instances, recall assesses the completeness of positive predictions made by a model.

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \qquad \textbf{(6)}$$

Where:

- FN (False Negative): The quantity of positive samples that were mistakenly reported as negative.

*4.3.3 Accuracy:*Accuracy is the ratio of all observations—true positives and true negatives—to correctly predicted observations.

$$\boldsymbol{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}} \qquad \textbf{(7)}$$

Where**:**

- TN(True Negatives): The quantity of negative samples that were accurately predicted to be negative.

5 RESULTS:

Four classification techniques are used to forecast the Air Quality Index (AQI)— Support Vector Machine (SVM), Decision Tree, Linear Regression and Random Forest have been utilized. The effectiveness of these classification has been assessed using various metrics namely Precision, Recall and Accuracy. The results of precision and recall for every class label, across all classification techniques of Delhi (Ashok Vihar) and Kolkata (Ballygunge and Victoria), are presented in Table 1to Table 7 respectively.

Table 1: Precision Results of Kolkata (Victoria)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**0.9911**	**0.9776**	**0.9847**	**1**	**1**	**1**
SVM	**0.9735**	**0.8882**	**0.832**	**0.8615**	**0.8732**	**1**
Random Forest	**0.9911**	**0.9944**	**1**	**0.9897**	**0.9859**	**1**
Linear Regression	**0.9382**	**0.8938**	**0.9771**	**0.8615**	**0.9859**	**1**

Table 2: Precision Results of Kolkata (Ballygunge)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**0.9896**	**0.9685**	**0.992**	**0.992**	**1**	**1**
SVM	**0.927**	**0.8451**	**0.8583**	**0.9264**	**0.8543**	**1**
Random Forest	**0.9965**	**1**	**0.9666**	**0.988**	**1**	**1**
Linear Regression	**0.9166**	**0.9806**	**0.975**	**0.9243**	**0.9708**	**1**

Table 3: Precision Results of Delhi (Ashok Vihar)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**0.8421**	**0.8484**	**0.8333**	**0.9615**	**1**	**1**
SVM	**0.375**	**0.6428**	**0.3**	**0.96**	**0.6666**	**0.8857**
Random Forest	**1**	**1**	**1**	**1**	**0.8333**	**0.9714**
Linear Regression	**1**	**0.9285**	**0.8**	**0.9318**	**0.75**	**0.9428**

Table 4: Recall Results of Kolkata (Victoria)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**1**	**0.9943**	**1**	**0.9763**	**0.9861**	**1**
SVM	**0.9169**	**0.8502**	**0.8515**	**0.9222**	**0.9687**	**1**
Random Forest	**0.9911**	**1**	**0.9849**	**0.9897**	**1**	**1**
Linear Regression	**0.9246**	**0.9142**	**0.9624**	**0.8736**	**0.9722**	**1**

Table 5: Recall Results of Kolkata (Ballygunge)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**0.9896**	**0.9935**	**1**	**0.9726**	**0.9903**	**1**
SVM	**0.9888**	**0.8239**	**0.7984**	**0.8594**	**0.9777**	**1**
Random Forest	**0.9896**	**0.981**	**1**	**0.992**	**1**	**1**
Linear Regression	**0.9462**	**0.962**	**0.959**	**0.8992**	**1**	**1**

Table 6: Recall Results of Kolkata of Delhi (Ashok Vihar)

Classes						
Technique	**Good**	**Satisfactory**	**Moderately Polluted**	**Poor**	**Very Poor**	**Severe**
Decision Trees	**0.8421**	**0.9655**	**1**	**0.8771**	**0.923**	**1**
SVM	**1**	**0.75**	**0.375**	**0.7272**	**1**	**0.7948**
Random Forest	**1**	**1**	**0.909**	**1**	**1**	**0.9444**
Linear Regression	**1**	**0.8666**	**0.8**	**0.9534**	**1**	**0.8918**

Table 7: Accuracy

Classes			
Technique	**Kolkata (Victoria)**	**Kolkata (Ballygunge)**	**Delhi (Ashok Vihar)**
Decision Trees	**0.991**	**0.987**	**0.9062**
SVM	**0.9**	**0.8952**	**0.7549**
Random Forest	**0.9921**	**0.9912**	**0.9801**
Linear Regression	**0.9168**	**0.9432**	**0.9172**

From the above tables, it has been concluded that the highest values are found in random forest and SVM has the lowest values of precision and recall out of all the classification techniques.These outcomes have been condensedin Table 1 to Table 6. Further it can be noted that Random forest depicts the maximum accuracy of 0.9921,0.9912 and 0.9801 and SVM depicts the minimum accuracy of 0.9, 0.8952 and 0.7549 for Kolkata (Victoria), Kolkata (Ballygunge) and Delhi (Ashok Vihar) respectively.These results have been represented in Table 7.

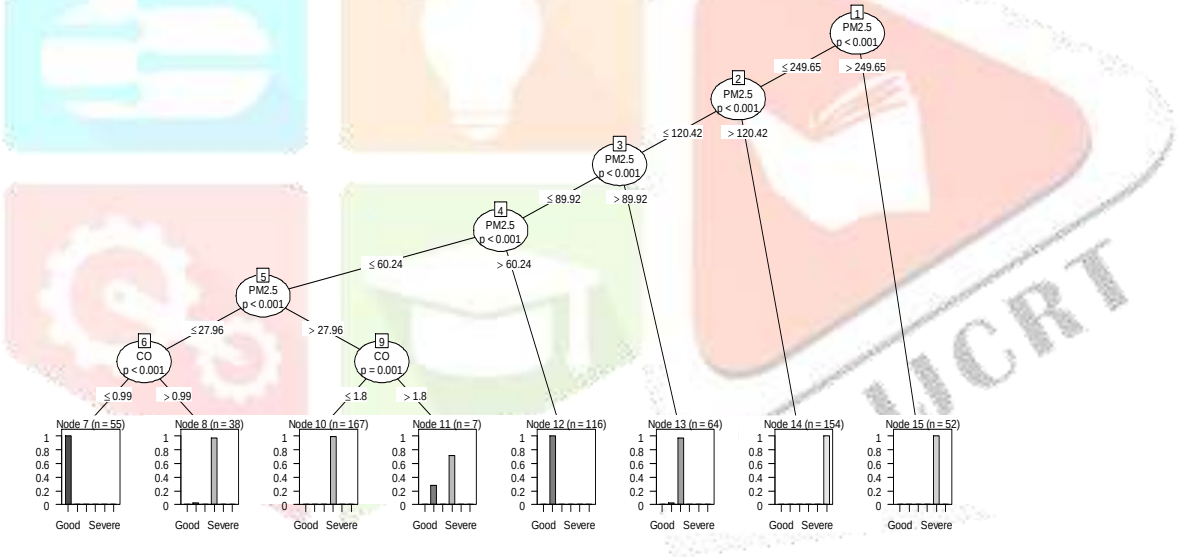

Figure 10: Decision Tree of Delhi (Ashok Vihar)

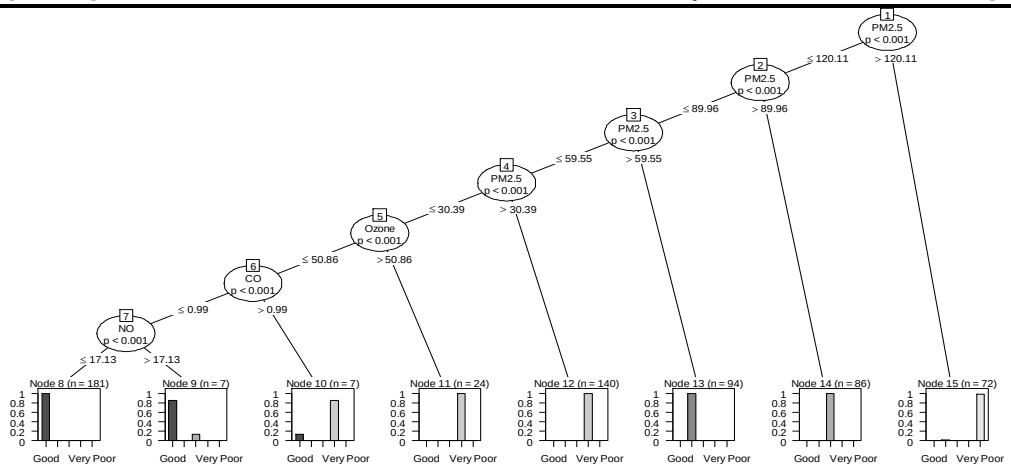

Figure 11: Decision Tree of Kolkata (Ballygunge)

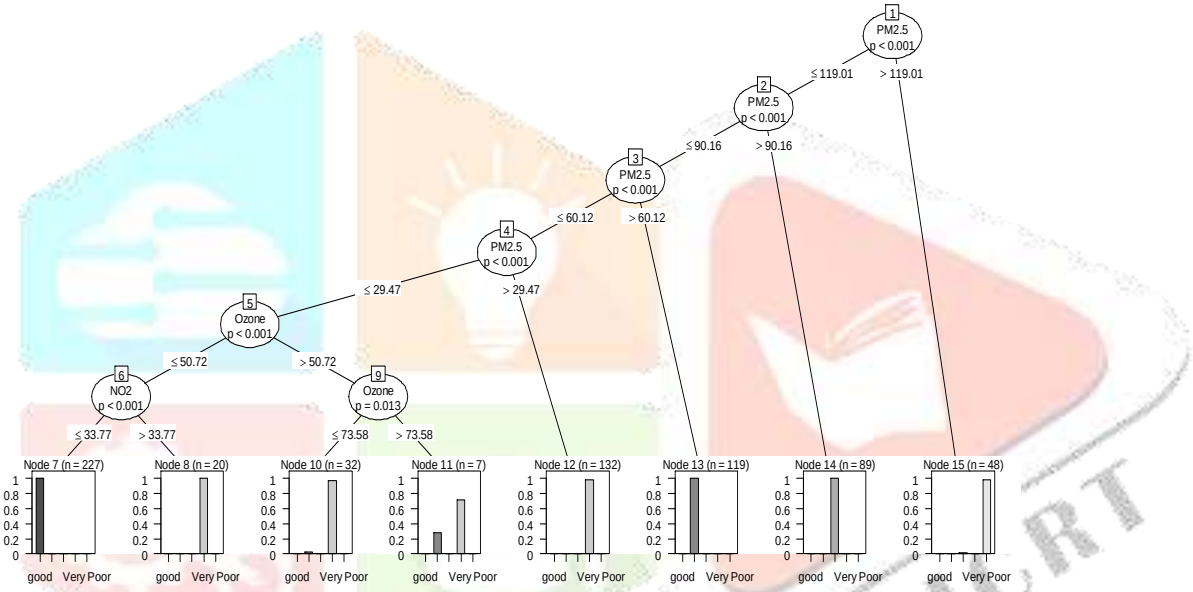

Figure 12: Decision Tree of Kolkata (Victoria)

6 CONCLUSION:

Monitoring air quality is crucial, particularly in developing nations like India where air pollution has serious negative effects on public health. In this study, four classification techniques were used to predict the air quality index for two stations in Kolkata and one station in Delhi: SVM, Random Forest, Decision Trees, and Linear Regression. In each case, the performance evaluation has been done based on metrics like precision, recall and accuracy. From the experimental work it has been concluded that out of all techniques, Random Forest has the highest values and SVM has the lowest values of precision and recall. Moreover, For Kolkata (Victoria), Kolkata (Ballygunge), and Delhi (Ashok Vihar), respectively, it has been discovered that Random Forest exhibits the highest prediction accuracy of 0.9921, 0.9912, and 0.9801 and SVM exhibits the lowest prediction accuracy of 0.9, 0.8952, and 0.7549.

REFERENCES

[1] Samad, A., Garuda, S., Vogt, U., & Yang, B. (2023). Air pollution prediction using machine learning techniques–an approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, *310*, 119987.

[2] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. *International Journal of Environmental Science and Technology*, *20*(5), 5333-5348.

[3] Wang D, Wei S, Luo H, Yue C, Grunder O (2017) A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci Total Environ 580:719–733

[4] Sethi J, Mittal M. Ambient air quality estimation using supervised learning techniques. EAI Endors Trans Scal Inform Syst. 2019;6(22), doi: 10.4108/ eai.13-7-2018.159406.

[5] Sethi JK, Mittal M. Analysis of air quality using univariate and multivari- ate time series models. In 2020 IEEE 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). 2020;epub, 823-827. doi: 10.1109/Confluence47617.2020.9058303

[6] Rybarczyk Y, Zalakeviciute R (2021) Assessing the COVID-19 impact on air quality: a machine learning approach. Geophys Res Lett. https://doi.org/10.1029/2020GL091202

[7] Verma, R. L., Gunawardhana, L., Kamyotra, J. S., Ambade, B., & Kurwadkar, S. (2023). Air quality trends in coastal industrial clusters of Tamil Nadu, India: A comparison with major Indian cities. *Environmental Advances*, *13*, 100412.

[8] Veljanovska, K., & Dimoski, A. (2017). Machine Learning Algorithms in Air Quality Index Prediction. *International Journal of Science and Engineering Investigations (IJSEI)*, *6*(71), 123-127.

[9] Hussain, M., Afrin, S., Irin, A., & Park, S. K. (2021, December). Applying Decision Tree Algorithm for Air Quality Prediction in Bangladesh. In *2021 5th International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

[10] Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K., & Tsao, C. C. (2017). Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. *PloS one*, *12*(7), e0179763.

[11] Ravindiran, G., Hayder, G., Kanagarathinam, K., Alagumalai, A., & Sonne, C. (2023). Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*, *338*, 139518.

[12]https://cpcb.nic.in/upload/NAAQS_2019.pdf

[13] Delavar, M. R., Gholami, A., Shiran, G. R., Rashidi, Y., Nakhaeizadeh, G. R., Fedra, K., & Hatefi Afshar, S. (2019). A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS International Journal of Geo-Information, 8(2), 99.

[14]Alpan, K., & Sekeroglu, B. (2020). Prediction of pollutant concentrations by meteorological data using machine learning algorithms. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *44*, 21-27

[15]Sethi, J. K., & Mittal, M. (2021). Prediction of air quality index using hybrid machine learning algorithm. In *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019* (pp. 439-449). Springer Singapore.

[16] Safavian SR, Landgrebe D. A survey of decision tree classifier methodol- ogy. IEEE Trans Syst Man Cybern. 1991;21(3):660-674. doi: 10.1109/21. 97458.)

[17] Halsana, S. (2020). Air quality prediction model using supervised machine learning algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *8*, 190-201.

[18] Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*(2.5), 3.

[19] Singh, S. (2021). Prediction of Air Pollution Using Random Forest. *Annals of the Romanian Society for Cell Biology*, 19314-19322

[20] Haq, M. A. (2022). SMOTEDNN: A novel model for air pollution forecasting and AQI classification. *Computers, Materials & Continua*, *71*(1).

[21] https://www.mapsofindia.com/maps/delhi/

[22] https://www.mapsofindia.com/maps/westbengal/districts/kolkata.html

[23] Cican, G., Buturache, A. N., & Mirea, R. (2023). Applying Machine Learning Techniques in Air Quality Prediction—A Bucharest City Case Study. *Sustainability*, *15*(11), 8445.

[24] Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied sciences*, *9*(19), 4069.

[25] Benchrif, A., Wheida, A., Tahri, M., Shubbar, R. M., & Biswas, B. (2021). Air quality during three covid-19 lockdown phases: AQI, PM2. 5 and NO2 assessment in cities with more than 1 million inhabitants. *Sustainable Cities and Society*, *74*, 103170.