SEMARAK ILMU PUBLISHING
Journal of Advanced Research in Applied Sciences and Engineering Technology
Journal homepage: https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Air Quality Prediction using Ensemble Classifiers and Single Decision Tree

Tanisha Madan[1,*], Shrddha Sagar[1], Tien Anh Tran[2], Deepali Virmani[3], Deependra Rastogi[4]

[1] School of Computer Science and Engineering, Galgotias University, Greater Noida, India
[2] Department of Marine Engineering, Vietnam Maritime University, Haiphong, Vietnam
[3] School of Engineering and Technology, VIPS-TC, Delhi, India
[4] School of Computer Science and Engineering, IILM University, Greater Noida, India

ARTICLE INFO

Article history:
Received 5 December 2023
Received in revised form 4 February 2024
Accepted 28 August 2024
Available online 1 October 2024

ABSTRACT

Air pollution is a major concern nowadays that needs immediate action. Various acts have already been initiated by the government in controlling it. The harmful gases present in the air are Nitrogen dioxide, Sulphur dioxide, Ozone, and Carbon monoxide that is causing air pollution and are becoming the major cause of harmful diseases. It does not affect human beings only but the entire environment. According to World Health Organization, air pollution is killing seven million people around the world. Planting trees and avoiding using plastics is the possible solution to control air pollution. The air quality index is the measure through which we categorize the air pollution of various cities. In this paper, we will discuss the bagging and boosting method, the extra trees method of machine learning in predicting air quality index, and its comparison with Decision trees. The dataset is collected from the Open Weather Application Program Interface. In this, the prediction of the Air Quality Index is done based on previous data. Time i.e., hourly prediction and space series prediction are done on the dataset. After implementing each algorithm confusion matrix is obtained and from this precision and recall are calculated for each category of the Air quality index ranging from one to five with one being good and five being very unhealthy. It has been observed that Gradient Boosting achieves the highest accuracy with 98.89%.

Keywords:
Air Pollution; Air Quality Index; Bagging; Boosting; Decision trees; Extra trees

1. Introduction

Pollution is the contamination of toxic gases in the environment. When these gases are present in the air, it is called air pollution. There are other types of pollution also like soil pollution and water pollution. Various types of pollutants present in the air are carbon monoxide, nitrogen dioxide, sulphur dioxide, particulate matter. Various factors causing air pollution such as

i. Burning of fossil fuels
ii. Vehicles
iii. Factories
iv. Burning of crackers

* *Corresponding author.*
E-mail address: tanishamadan@gmail.com

https://doi.org/10.37934/araset.52.1.5667

It is causing various diseases like bronchitis, asthma, lung, and cancer. Many people die globally due to disease caused by pollution. Pollution needs to be controlled immediately. It is necessary to create awareness among human beings so that they make limited use of chemicals causing air pollution. World pollution prevention day is also observed every year describing its importance and details. Planting trees, avoiding plastics, and recycling and reusing are various solutions to prevent pollution.

The air quality index (AQI) is the measure through which we categorize air pollution in various cities among 5 categories. In today's time, various cities are categorized as having poor AQI. Wind speed, humidity air pressure, and air temperature also affect air quality.

In earlier times AQI is calculated manually and it is not accurate. So, various machine learning algorithms like decision trees, bagging and boosting models, and Naïve Bayes, K-nearest neighbour (KNN) are there to determine air quality index. In this paper bagging and boosting, extra trees models are compared with decision trees in terms of precision and recall.

The probability of $PM_{2.5}$ and PM_{10} can be calculated with the help of decision tree [1]. The hybrid method of ARIMA and Prophet method can also be used for air quality prediction [2]. It has been observed that ensemble methods produce better results [3].

Our aim is to determine the AQI of various cities on the basis on historical data of pollutants with a delay of one hour. Time and space prediction is followed.

Generally, AQI 1 means good and AQI 5 means very unhealthy for the population as shown in Table 1. Particulate matter 10 (PM_{10}), particulate matter 2.5 ($PM_{2.5}$), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO) variables or indicators are used to calculate AQI. AQI value is calculated as in Eq. (1).

$$I = I_{high} - I_{low}/C_{high} - C_{low}(C - C_{low}) + I_{low} \tag{1}$$

where I = Air quality index
I_{high} = Index breakpoint corresponding to C_{high}
I_{low} = Index breakpoint corresponding to C_{low}
C =pollutant concentration
C_{high} = Concentration breakpoint >= C
C_{low} = Concentration breakpoint <= C

For example, for AQI 1, pollutant concentration values given in dataset by Openweathermap [4].

Table 1
AQI categories

AQI Category	Meaning	AQI Range
1	Good	0-25
2	Fair	25-50
3	Moderate	50-75
4	Poor	75-100
5	Very Poor	>100

2. Literature Survey

Previous study conducted by Su [5] had predicted the particulate matter 2.5 ($PM_{2.5}$) value with the light and extreme gradient boosting method and it was concluded that the light gradient boosting method performs better in terms of accuracy. Liu *et al.,* [6] use leveraging bagging method in predicting air quality and is compared with all algorithms. Zheng *et al.,* [7] compare the autoregressive integrated moving average model, random forest, logistic regression, boosting and it was predicted that the ensemble model performs better in case of forecasting air quality index value. Chang *et al.,* [8] predicted AQI and $PM_{2.5}$ values using data mining algorithms and proposes a framework called Extract-Transform and load on a cloud platform. Wenjing Wang and Shengquan Yang [9] propose a Neural network on a big data platform and six pollutant concentration is taken for predicting AQI. Through self-learning characteristics of neural networks, it has higher prediction accuracy. Li *et al.,* [10] propose a spatiotemporal autoencoder neural network that achieves higher accuracy of 87.2% than a support vector machine (SVM). Mahalingam *et al.,* [11] predicted the air quality of Delhi city, data collected from the central pollution control board and support vector machine achieves higher accuracy than neural networks. Six functions of the support vector machine are used, and medium Gaussian support vector machine achieves higher accuracy of 97.3%. Ma *et al.,* [12] predicted particulate matter 2.5 ($PM_{2.5}$) with bidirectional Long short-term memory and inverse distance weighted and achieves higher accuracy than other models. Chiang *et al.,* [13] proposes a gated recurrent unit and long short-term memory for the prediction of particulate matter 2.5 on the basis of the hidden number of neurons from 30 to 100 and the training time and Root means square error are calculated. Decoder transfer learning can also be used for predicting personalized air quality [14]. Zhang *et al.,* [15] predicted the $PM_{2.5}$ pollutant level using light GBM model for high dimensional data over the next 24 hours of Beijing. Shaban *et al.,* [16] compared the neural network, M5P model trees, SVM in predicting SO_2, NO_2 and O_3 and M5P algorithm performs better. Univariate and multivariate modelling are done. SO_2 using ANN in univariate modelling produces the worst result. Murugan and Palanichamy [17] compares the multilayer perceptron and random forest, and it has been seen that random forest performs better in the prediction of $PM_{2.5}$ in Malaysia air quality dataset. Kothandaraman *et al.,* [18] predicted $PM_{2.5}$ in polluted cities using linear regression, random forest, KNN, AdaBoost, XGBoost and compare the results in terms of MAE and RMSE. ArunaKumari *et al.,* [19] compares the SVM and neural network in predicting air quality of Delhi. Van *et al.,* [20] compares algorithms decision trees, random forest and XGBoost, and compare algorithms by MAE, RMSE, and R^2 and XGBoost outperforms other models. Popa *et al.,* [21] collected dataset using pollution sensor data from six atmospheric air quality stations and various machine learning algorithms are compared and they have predicted the air pollution of the crowded area of Bucharest, Romania. Sensors are used for data collection as the dataset is huge. Castelli *et al.,* [22] has used SVM with radial basis function for accurate prediction of pollutants like CO, SO_2, NO_2, O_3, and $PM_{2.5}$ on an hourly basis in the California area.

It has been concluded that machine learning plays a vital role in predicting and forecasting air quality index and pollutants value. Various Ensembling methods like Random Forest, Gradient Boosting method outperforms other methods. But still, various challenges are faced by researchers like data availability, outliers in data, sudden climatic change, and slow training in neural networks, fixed to a particular region i.e. if the region is changing model does not show a better result. Most researchers followed a Time Series prediction with a delay of particular hours i.e. 24 hours,5 hours,1 hour. Precision, Recall, F1 Score are some parameters on which results are calculated.

3. Methodology

The dataset is collected from Open Weather Application Program Interface. The dataset contains CO, NO, NO_2, O_3, SO_2, NH_3, PM_{10}, $PM_{2.5}$, and AQI values as Input and Output parameters.

It is being tested on one dataset. Cities taken are Chandigarh and Visakhapatnam. The total no of rows in dataset are 5,712.

First Dataset pre-processing is performed i.e. removing all rows containing null values, then normalization of input parameters is done using the Z score method. The dataset is split into training (70%) and 30%). After splitting bagging, boosting, decision tree, and extra trees algorithms are applied, and a confusion matrix is created for comparison.

NH_3, NO, O_3 contain a lot of noisy data. In this, we have chosen input and output parameters as follows: input parameters considered are PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO and output parameter is AQI.

Bagging model: In the bagging model data points are selected more than once to be passed to each estimator i.e., weak classifier. Then the weak classifier is trained independently and in parallel, and the majority voting rule is applied in case of classification and the mean is taken in case of regression to obtain the final prediction [23]. The decision tree is chosen as a weak classifier in our approach.

Boosting model: In boosting, models are trained sequentially instead of parallel. There are various types of boosting models like AdaBoost, gradient boosting, and XGBoost [24].

Decision tree: It is also used for classification. Beginning with the root node that has the full dataset. The internal node represents the features of the dataset, and the leaf node contains the outcome. After beginning with the root node containing the entire dataset, the best attribute is selected using information gain. Then divide the dataset into subsets using the best attribute and the decision tree is created. This continues with the sub-dataset until you cannot classify the nodes and the leaf node represents the outcome. In the decision tree for calculating information gain, we have included Gini Impurity. Gini Index is calculated as in Eq. (2).

$$1-\sum_{i=1}^{n}(P_i)^2 \tag{2}$$

Extra trees: It combines the prediction of various decision trees. Extra trees choose the split randomly and give unique samples to each decision tree.

Algorithm: -

I. Reading SQL Table containing cities.
II. Removing rows containing NaN values.
III. Defining the input column and target column to be used for prediction.
IV. Normalizing the data using the Z-Score method

$$x-\mu/\sigma \tag{3}$$

where, x = input parameters, μ = mean, and σ = standard deviation

V. Splitting the dataset into training (70%) and testing (30%)
VI. Defining various ensembling models like bagging, extra trees, gradient boosting, Adaboost, and single classifier decision trees.
VII. Plotting confusion matrix for each model.

The methodology is illustrated in Figure 1.

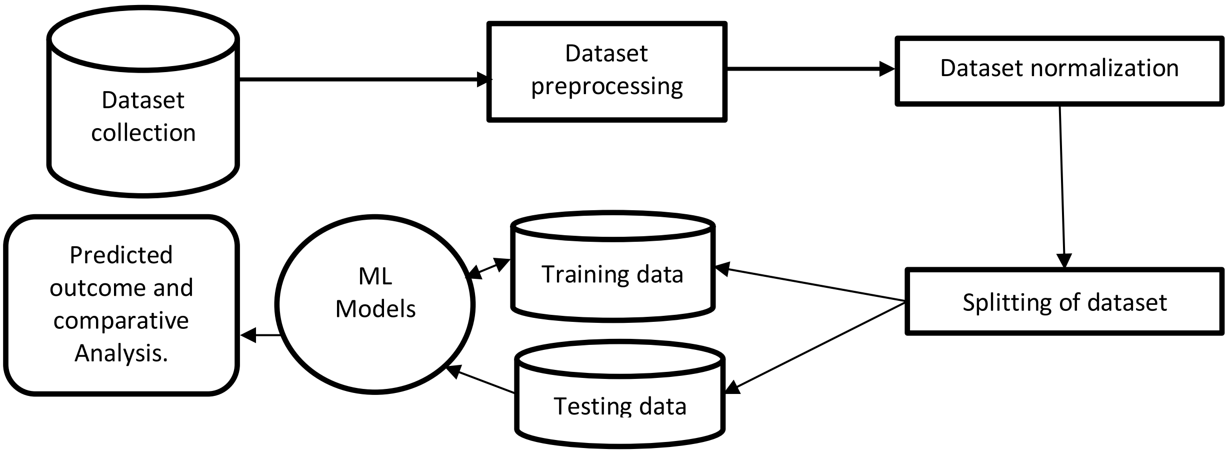

Fig. 1. Methodology

The parameters chosen are explained in result section and confusion matrix obtained are explained in result section.

4. Implementation

Implementation is done on Jupyter Notebook.

I. Bagging Classifier: In the bagging classifier we have used a decision tree classifier and the max samples and max features taken are 0.5 as shown in Figure 2 and Table 2.

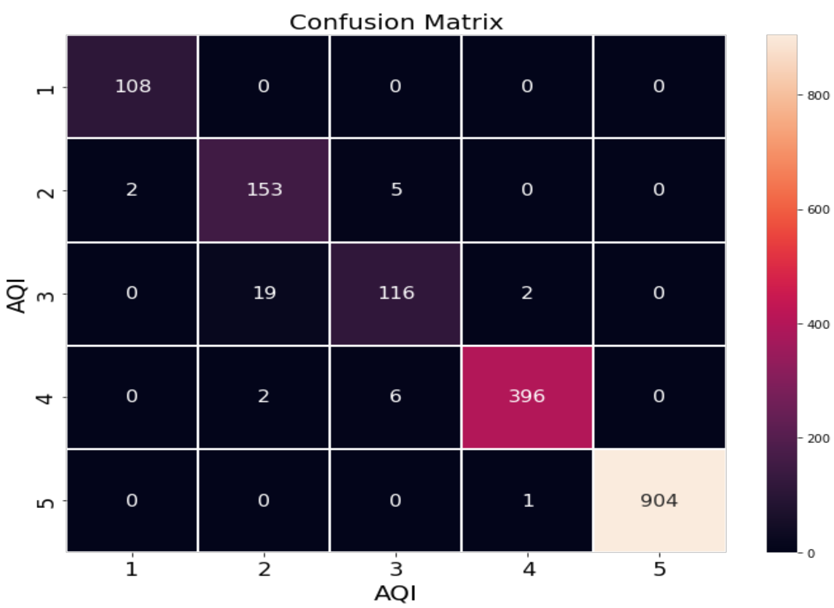

Fig. 2. Confusion matrix of bagging

Table 2
Bagging performance

	Precision	Recall	F1-score	Support
1	0.9818	1.0000	0.9908	108
2	0.8793	0.9563	0.9162	160
3	0.9134	0.8467	0.8788	137
4	0.9925	0.9802	0.9863	404
5	1.0000	0.9989	0.9994	905
Accuracy			0.9784	1714
Macro avg	0.9534	0.9564	0.9543	1714
Weighted avg	0.9789	0.9784	0.9784	1714
			Accuracy	97.84

II. AdaBoost classifier: In the AdaBoost classifier, we have taken estimators as 100 and the random state is 0 as shown in Figure 3 and Table 3.

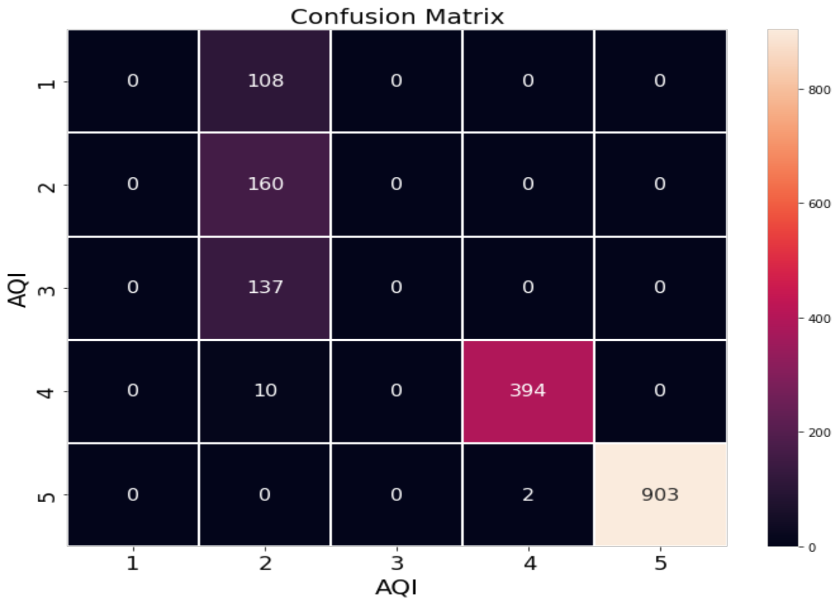

Fig. 3. Confusion matrix of AdaBoost

Table 3
AdaBoost performance

	Precision	Recall	F1-score	Support
1	0.0000	0.0000	0.0000	108
2	0.3855	1.0000	0.5565	160
3	0.0000	0.0000	0.0000	137
4	0.9949	0.9752	0.9850	404
5	1.0000	0.9978	0.9989	905
Accuracy			0.8501	1714
Macro avg	0.4761	0.5946	0.5081	1714
Weighted avg	0.7985	0.8501	0.8115	1714
			Accuracy	85.01

III. Gradient boosting classifier: In gradient boosting classifier estimators are 100, the learning rate is 0.1, the max depth is 1 and the random state is 0 as depicted in Figure 4 and Table 4.

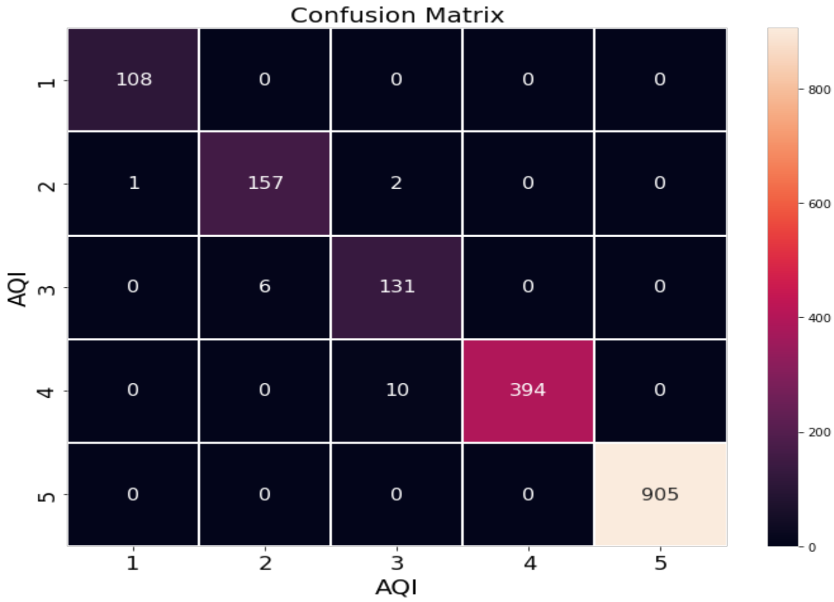

Fig. 4. Confusion matrix of gradient boosting

Table 4
Gradient boosting performance

	Precision	Recall	F1-score	Support
1	0.9908	1.0000	0.9954	108
2	0.9632	0.9812	0.9721	160
3	0.9161	0.9562	0.9357	137
4	1.0000	0.9752	0.9875	404
5	1.0000	1.0000	1.0000	905
Accuracy			0.9889	1714
Macro avg	0.9740	0.9825	0.9781	1714
Weighted avg	0.9893	0.9889	0.9890	1714
			Accuracy	98.89

IV. Extra trees: In extra trees, total estimators are 100 and random state is 0 as illustrated in Figure 5 and Table 5.

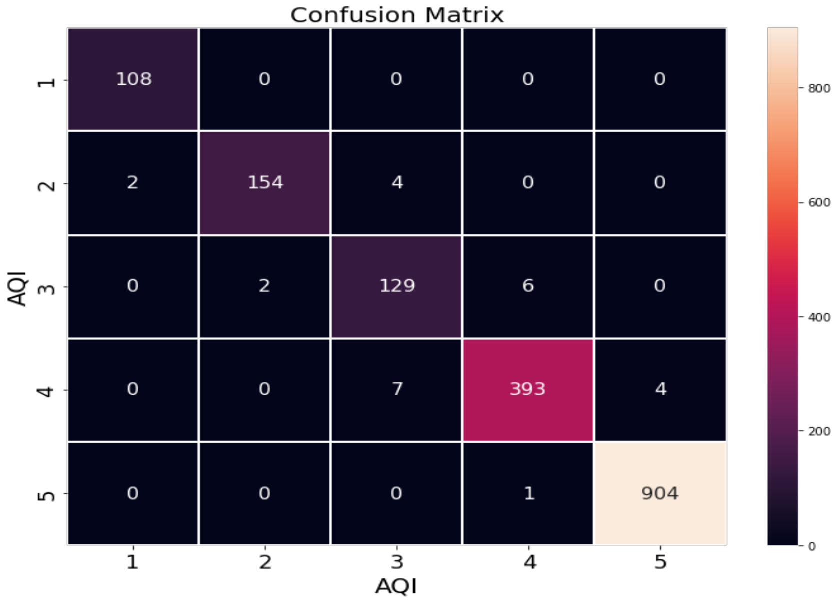

Fig. 5. Confusion matrix of extra trees

Table 5
Extra trees performance

	Precision	Recall	F1-score	Support
1	0.9818	1.0000	0.9908	108
2	0.9872	0.9625	0.9747	160
3	0.9214	0.9416	0.9314	137
4	0.9825	0.9728	0.9776	404
5	0.9956	0.9989	0.9972	905
Accuracy			0.9848	1714
Macro avg	0.9737	0.9752	0.9744	1714
Weighted avg	0.9849	0.9848	0.9848	1714
			Accuracy	98.48

V. Decision trees: In decision trees, we have used Gini impurity as best attribute selection measure as shown in Figure 6 and Table 6.

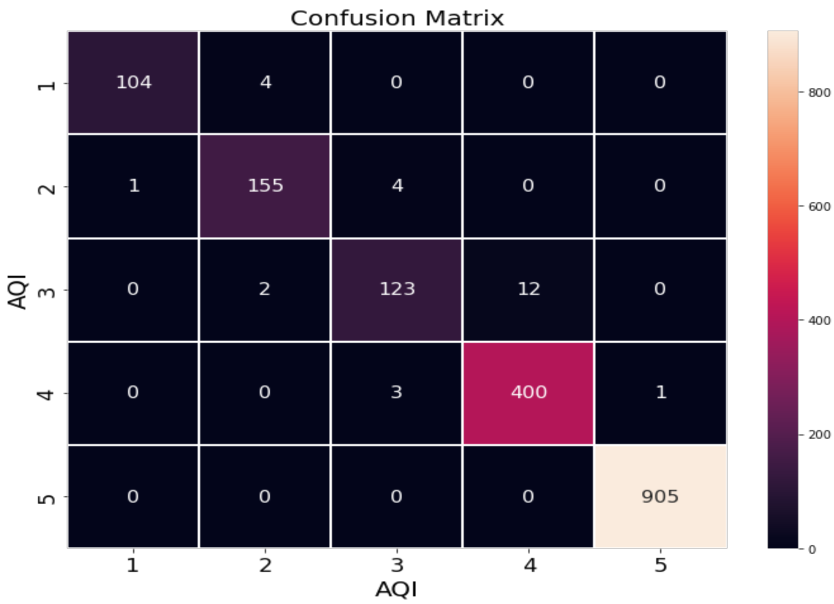

Fig 6. Confusion matrix of Decision trees

Table 6
Decision trees performance

	Precision	Recall	F1-score	Support
1	0.9905	0.9630	0.9765	108
2	0.9627	0.9688	0.9657	160
3	0.9462	0.8978	0.9213	137
4	0.9709	0.9901	0.9804	404
5	0.9989	1.0000	0.9994	905
Accuracy			0.9842	1714
Macro avg	0.9738	0.9639	0.9687	1714
Weighted avg	0.9842	0.9842	0.9841	1714
			Accuracy	98.42

Figure 2 – Figure 6 are the confusion matrix obtained for each model. From the Confusion matrix, we can test our model and how it is performing. Here X-axis represents AQI predicted, and the Y-axis represents AQI actual. AQI Ranges are described in Table 1 as shown before.

Table 2 – Table 6 are the calculated precision, recall, f1 score, weighted average, and macro average, and formulas for calculating are described in the result section.

5. Results and Discussion

This paper focuses on predicting AQI using the ensemble classifier and single decision tree. In boosting AdaBoost model and Gradient Boosting models are taken. It is observed that the Gradient Boosting model performs better than the bagging and AdaBoost models. Ada Boost model exhibits the least accuracy. Gradient boosting is also compared with decision trees and extra trees that exhibit almost similar accuracy, but the gradient boosting model performs better than all models. The result and calculation are shown below [25].

i. True positive (TP) is defined as when the predicted outcome has been correctly classified as positive class.

ii. False positive (FP) is defined as when the predicted outcome is a positive class but actually belongs to another class.
iii. False negative (FN) is defined as when actually it belongs to a positive class but is predicted as a negative class.
iv. True negative (TN) is defined as when the predicted outcome has been correctly classified as negative class.
v. Support is the total no of samples of each class in which there are true responses.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

$$Accuracy = \frac{Correctly\ classified\ samples}{Total\ no\ of\ samples} \quad (7)$$

The macro average is the average of all categories. The weighted average for precision is calculated as in Eq. (8).

$$\frac{S(1) * P(1) + S(2) * P(2) + S(3) * P(3) + S(4) * P(4) + S(5) * P(5)}{S(1) + S(2) + S(3) + S(4) + S(5)} \quad (8)$$

where $S(1)$ is the support sample of class 1, $P(1)$ is the precision for class 1.

Similarly for recall and F1 Score, weighted average can be calculated. The result for all algorithms is shown in Table 7.

Table 7
Results for all algorithms

Algorithms	Accuracy	Precision	Recall	F1 Score
Bagging method	97.84%	0.9534	0.9564	0.9543
Gradient Boosting method	98.89%	0.9740	0.9825	0.9781
AdaBoost Method	85.01%	0.4761	0.5946	0.5081
Decision trees	98.42%	0.9738	0.9639	0.9687
Extra trees	98.48%	0.9737	0.9752	0.9744

Based on the results in Table 7, it is concluded that gradient boosting method perform better than other algorithms.

Various Single models suffer from bias-variance trade i.e. they have high bias and high variance. To achieve low bias and variance we use ensembling methods. AdaBoost exhibit least accuracy than Gradient Boosting because Gradient Boosting learns from the previous classifier residuals and the final prediction depends on the maximum vote of weak learners i.e. it learns from gradient whereas AdaBoost learns from high-weight data points as it put more weight on misclassified samples. Bagging learns independently and follow averaging process and work parallelly and produces a model with less variance. Extra trees performs better than decision tree as it randomly selects the node for splitting and uses the entire dataset to build decision tree.

6. Conclusion

It is concluded that the gradient boosting outperforms all other models. Gradient boosting, decision trees, and extra trees exhibit almost similar accuracy and AdaBoost is having least accuracy because Adaboost learns from misclassified samples. In the future, various deep learning algorithms and other datasets, cities, and pollutants could be taken for validation of our results. Meteorological factors like wind speed, direction, and humidity can also be taken into account in the calculation of AQI.

Acknowledgement
This research was not funded by any grant.

References

[1] Sharma, Ritik, Gaurav Shilimkar, and Shivam Pisal. "Air Quality Prediction by Machine Learning." *Int. J. Sci. Res. Sci. Technol* 8 (2021): 486-492. https://doi.org/10.32628/IJSRST218396
[2] Ye, Ziyuan. "Air pollutants prediction in shenzhen based on arima and prophet method." In *E3S Web of Conferences*, vol. 136, p. 05001. EDP Sciences, 2019. https://doi.org/10.1051/e3sconf/201913605001
[3] Liang, Yun-Chia, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez. "Machine learning-based prediction of air quality." *applied sciences* 10, no. 24 (2020): 9151. https://doi.org/10.3390/app10249151
[4] Openweathermap.org "Weather API - OpenWeatherMap: Data Set".
[5] Su, Yuelai. "Prediction of air quality based on Gradient Boosting Machine Method." In *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, pp. 395-397. IEEE, 2020. https://doi.org/10.1109/ICBDIE50010.2020.00099
[6] Liu, Weike, Hang Zhang, and Qingbao Liu. "An air quality grade forecasting approach based on ensemble learning." In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pp. 87-91. IEEE, 2019. https://doi.org/10.1109/AIAM48774.2019.00024
[7] Zheng, Hong, Yunhui Cheng, and Haibin Li. "Investigation of model ensemble for fine-grained air quality prediction." *China Communications* 17, no. 7 (2020): 207-223. https://doi.org/10.23919/J.CC.2020.07.015
[8] Chang, Yue Shan, Kuan-Ming Lin, Yi-Ting Tsai, Yu-Ren Zeng, and Cheng-Xiang Hung. "Big data platform for air quality analysis and prediction." In *2018 27th Wireless and Optical Communication Conference (WOCC)*, pp. 1-3. IEEE, 2018. https://doi.org/10.1109/WOCC.2018.8372743
[9] Wang, Wenjing, and Shengquan Yang. "Research on Air Quality Forecasting Based on Big Data and Neural Network." In *2020 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, pp. 180-184. IEEE, 2020. https://doi.org/10.1109/ICCNEA50255.2020.00045
[10] Li, Yan, Xiajiong Shen, Daojun Han, Jun Sun, and Yatian Shen. "Spatio-temporal-aware sparse denoising autoencoder neural network for air quality prediction." In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 96-100. IEEE, 2018. https://doi.org/10.1109/CCIS.2018.8691167
[11] Mahalingam, Usha, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam. "A machine learning model for air quality prediction for smart cities." In *2019 International conference on wireless communications signal processing and networking (WiSPNET)*, pp. 452-457. IEEE, 2019. https://doi.org/10.1109/WiSPNET45539.2019.9032734
[12] Ma, Jun, Yuexiong Ding, Vincent JL Gan, Changqing Lin, and Zhiwei Wan. "Spatiotemporal prediction of PM2. 5 concentrations at different time granularities using IDW-BLSTM." *Ieee Access* 7 (2019): 107897-107907. https://doi.org/10.1109/ACCESS.2019.2932445
[13] Chiang, Yu-Lun, Chao-Liang Hsieh, Hsiang-Yu Huang, Jen-Cheng Wang, Cheng-Ying Chou, Chih-Hong Sun, Tzai-Hung Wen, Jehn-Yih Juang, and Joe-Air Jiang. "Urban Area PM 2.5 Prediction with Machine Methods: An On-Board Monitoring System." In *2018 12th International Conference on Sensing Technology (ICST)*, pp. 25-30. IEEE, 2018. https://doi.org/10.1109/ICSensT.2018.8603564

[14] Zhao, Peijiang, and Koji Zettsu. "Decoder transfer learning for predicting personal exposure to air pollution." In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5620-5629. IEEE, 2019. https://doi.org/10.1109/BigData47090.2019.9006604

[15] Zhang, Ying, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang, and Linyan Huang. "A predictive data feature exploration-based air quality prediction approach." *IEEE Access* 7 (2019): 30732-30743. https://doi.org/10.1109/ACCESS.2019.2897754

[16] Shaban, Khaled Bashir, Abdullah Kadri, and Eman Rezk. "Urban air pollution monitoring system with forecasting models." *IEEE Sensors Journal* 16, no. 8 (2016): 2598-2606. https://doi.org/10.1109/JSEN.2016.2514378

[17] Murugan, Rishanti, and Naveen Palanichamy. "Smart city air quality prediction using machine learning." In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1048-1054. IEEE, 2021. https://doi.org/10.1109/ICICCS51141.2021.9432074

[18] Kothandaraman, D., N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla, and Worku Abera. "Intelligent forecasting of air quality and pollution prediction using machine learning." *Adsorption Science & Technology* 2022 (2022). https://doi.org/10.1155/2022/5086622

[19] ArunaKumari, P., Y. Vijayalata, G. Susmitha Valli, and Y. Lakshmi Prasanna. "Air Contamination Prediction and Comparison Using Machine Learning Algorithms." In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pp. 661-677. Singapore: Springer Nature Singapore, 2023. https://doi.org/10.1007/978-981-19-2358-6_60

[20] Van, N. H., P. Van Thanh, D. N. Tran, and D-T. Tran. "A new model of air quality prediction using lightweight machine learning." *International Journal of Environmental Science and Technology* 20, no. 3 (2023): 2983-2994. https://doi.org/10.1007/s13762-022-04185-w

[21] Popa, Cicerone Laurentiu, Tiberiu Gabriel Dobrescu, Catalin-Ionut Silvestru, Alexandru-Cristian Firulescu, Constantin Adrian Popescu, and Costel Emil Cotet. "Pollution and weather reports: Using machine learning for combating pollution in big cities." *Sensors* 21, no. 21 (2021): 7329. https://doi.org/10.3390/s21217329

[22] Castelli, Mauro, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. "A machine learning approach to predict air quality in California." *Complexity* 2020 (2020). https://doi.org/10.1155/2020/8049504

[23] IBM. "What is Bagging." (2021).

[24] Pedamkar, P. "Ensemble Learning: Bagging vs Boosting." (2023)

[25] Precision | Definition, Precision Vs Accuracy, Recall, Formula.