Performance Analysis of Machine Learning Models for Air Pollution Prediction

Harshit Srivastava
Dept. of ECE, NIT Rourkela
Odisha, India
harshit_srivastava@nitrkl.ac.in

Goutam Kumar Sahoo
Dept. of ECE, NIT Rourkela
Odisha, India
goutamkr_sahoo@nitrkl.ac.in

Santos Kumar Das
Dept. of ECE, NIT Rourkela
Odisha, India
dassk@nitrkl.ac.in

Poonam Singh
Dept. of ECE, NIT Rourkela
Odisha, India
psingh@nitrkl.ac.in

***Abstract*—Air pollution includes contamination of air due to harmful gases, residues, fumes, etc. Contaminated air gives rise to important issues for the solid endurance of plants, organisms and individuals, including natural life. This paper focuses on predicting air pollutants using machine learning (ML) techniques and its performance analysis. Various regression and classification models like Support Vector Machine (SVM), Random Forest Classifier, Logistic Regression, Linear Regression and Random Forest Regression are used to optimize the air pollutants for better accuracy in forecasting. The performance of ML models is evaluated using State Pollution Control Board (SPCB) dataset, Odisha. The performance of Regression models is evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). It prevails in Random Forest Regression having RMSE and MAE as 2.63 and 3.32 respectively. For classification models, Random Forest Classifier precede with an accuracy of 93.5%. The efficient performance of the model in predicting air pollutants can help in alerting the public to safer living.**

***Index Terms*—Air pollution, Air quality index (AQI), Machine learning (ML), Logistic regression, Linear regression, Random forest, Support vector machine (SVM)**

I. Introduction

Air pollution is considered one of the fundamental detriments to people's well-being. The most prevalent air pollutants are greenhouse gases which includes carbon dioxide (CO2), methane (CH4), nitrogen oxides, sulphur dioxide, and particulate matter [1]. Aerosols, which are solid and liquid particles and certain gases, end up in the air, generating pollution. Most likely, people with heart or lung diseases, infants, children, and older adults experience health effects caused by particle pollution. PM2.5 and PM10 are two types of particle pollution. The former has minute air particles with a diameter of 2.5 micrometres or less while the latter has a diameter of fewer than 10 micrometres [2] . These air pollutants include meteorological parameters like temperature, humidity, dew point, solar radiation, rainfall, wind speed, direction, etc. Wind speed and direction are monitored to determine the source or area of emission. Chemical reactions in the atmosphere due to air pollution are influenced by temperature and sun radiation. The air quality index (AQI) is a comprehensive scheme for converting the weighted values of various air contaminants into a single number. Primarily sub-indices of each air pollutant (ppm) are calculated, which is directly proportional to the concentration of the air pollutant. Secondly, aggregation of sub-indices gives overall AQI and is categorized in Table I.

TABLE I
AQI Category

AQI Value	Category
0-50	Good
51-100	Satisfactory
101-200	Moderately Polluted
201-300	Poor
301-400	Highly Poor
401-500	Severe

Recent technological advances in Artificial Intelligence (AI) enable the benefits of using various cloud computing platforms, machine learning analytics, and AI technologies in real-world scenarios. IoT applications use machine learning techniques to analyze enormous volumes of data collected by cloud-based sensors [3] . Many literature studies reported accurate and precise information regarding air pollution. Shahriar *et al.* [4] collected air pollutant and meteorological data from four air monitoring stations in Bangladesh. The assessment metrics RMSE, R2, and MAE, were calculated using Gaussian Process Regression and fared best in predicting PM2.5 and PM10 concentrations in Dhaka. In contrast, ANN performed best in two of the stations for predicting PM2.5. Similarly, Kingsy Grace R *et al.* [5] investigated air pollution data employing the Enhanced K-Means clustering model. A comparative analysis was done with the possibilistic Fuzzy C-Means (PCFM) clustering algorithm.

Ayele *et al.* [6] worked on an IoT-enabled pollution monitoring system that monitors air pollutants in a given area, analyses it, and forecasts air quality. The data from the sensors were stored in the cloud. The recurrent neural network (RNN)machine learning method was used to anticipate the pollution rate. Kiruthika R *et al.* [7] have proposed a prototype that alters the user through a website. An embedded system was utilized to connect several sensors in their proposed IoT-based system, and the collected data was saved in a MySQL database. A prediction algorithm is constructed based on the obtained data to analyze and report it to the users. The exterior parameters are measured using temperature, humidity, moisture, and MQ5 gas sensors. Dan zhang *et al.* [8] focused on establishing a model that can be used both stationary and mobile. IoT sensors are installed on patrol cars around the region to anticipate immediate air quality. Different machine learning models, such as SVM, RF, and GBM, were employed to forecast air quality.

Work by Gore *et al.* [9] used the Decision Tree and Naive Bayes J48 techniques to classify the air quality index and its impact on health. The best performance of 91.9978% was achieved for the decision tree algorithm. However, using a small dataset and overfitting problems with decision tree approaches might perform poorly for continuous variables. Kingsy *et al.* [10] presented the K-means technique for categorizing the air quality index. However, the small dataset and trouble predicting future values put limitations on the study. Ishak *et al.* [11] examined the amount of ozone in Tunisia. The Support Vector Regression and Random Forests techniques were used to make future predictions while measuring the ozone content at three monitoring stations. Random Forests better estimate ozone forecasting. However, only one variable was considered for future forecasts due to the small amount of data from the three stations.

The study of various techniques and limitations creates scope to overcome the shortcomings. This work aims to use different machine learning regression and classification techniques for performance improvement. It will be helpful for the prediction of air pollutants for the benefit of public healthcare. The main contributions of the proposed work are as follows.

- Data collection from SPCB and pre-processing
- Prediction using various Regression models like Linear Regression and Random Forest Regression.
- Prediction using different Classification models like Logistic Regression, Support Vector Machine, Random Forest Classifier.
- Performance evaluation and comparative analysis of machine learning models.

The paper contributes towards the air pollution prediction for providing public care. The rest of the paper is as follows. Section II explains the the methodology, Section III explains the results and discussion, Section IV provides the conclusion with future scope.

II. METHODOLOGY

In this research work, Logistic Regression, Support Vector Machine, Random Forest Regression and Classification, Linear Regression models were implemented. Predictions of Air Quality Index have been made using these models. The dataset used is taken from State Pollution Control Board (SPCB) . PM10, PM2.5, O3 pollutants were considered for calculation and prediction of AQI.

A. Linear Regression:

The subject of regression analysis is the relationship between variables. Specifically the linear relationship between a dependent variable, y, and an independent variable, x. A supervised machine learning method with a continuous and constant slope projected output is linear regression. It is employed to forecast values across a broad range. Simple regression and multivariable regression are the two main forms. Simple linear regression employs the conventional slope-intercept form, where (m) and (b) are the variables that the algorithm will attempt to learn to produce the most accurate predictions. (x) and (y) denotes input data and prediction respectively. Multivariable Regression, on the other hand, accepts more than one independent variable and is a little more complicated.

B. Random Forest Regression:

The ensemble learning method is used in the regression supervised learning technique known as Random Forest Regression. By combining predictions from various machine learning algorithms, the ensemble learning technique produces forecasts that are more accurate than those from a single model. Random forest is a bagging technique rather than a boosting technique. The process involves creating a lot of decision trees, training them, and then calculating the class that represents the mean prediction (regression) of all the trees.

C. Logistic regression:

It is one of the machine learning mechanism. In this, probabilities must be transformed into binary numbers in order to make a prediction. This is due to the logistic function, often known as the sigmoid function. The S-shaped curve is the Sigmoid function. The major function of this is to trnsform real- valued numbers in between 0 and 1, but never exactly. A threshold classifier will then turn this value between 0 and 1 into either 0 or 1. Using Logistic Regression classifier AQI value above 100 is classified as class1 and AQI value below 100 is classified as 0.

D. Support Vector Machine Classifier:

A linear model for solving classification and regression problems. It can handle both linear and nonlinear issues. In this the data is being divided into classes. This division of data is done by drawing a line or hyperplane. The SVM Algorithm finds the nearest points in both classes to the line. The points are termed as Support vectors. It is now time to determine how far apart the line and the support vectors are from one another. This space is frequently referred to as a margin. Increasing the profit margin is our goal. The optimum hyperplane is that with the largest margin. As a result, SVM aims to create a decision boundary that provides as much separation between the two classes (that street) as possible.

E. Random Forest Classifier:

Random Forest works for prediction of both continuous and discrete values. It implements multiple decision trees. For regression, regression tree which predicts value is used.

F. Air Qulaity Index:

The principle for calculating AQI [12] given by CPCB (Central Pollution Control Board) is based on transforming the weighted value of each air pollutant into a single number or set of numbers. Thus, it first forms the sub-indices for each pollutant parameter and then the aggregate of sub-indices using the weighted additive form to calculate overall AQI.

$$I_i = \frac{I_{hi} - I_{lo}}{B_{hi} - B_{lo}} (C_p - B_{lo}) + I_{lo} \tag{1}$$

$$AQI = \sum_{k=1}^{n} W_i I_i \tag{2}$$

where, $\sum_{k=1}^{n} W_i = 1$, B_{hi}, and B_{lo} are the break-point concentrations greater and smaller to given concentration respectively, C_p is the pollutant actual concentration, I_{hi} and I_{lo} are AQI values corresponding to B_{hi} and B_{lo} respectively and W is the weight assigned to each pollutant.

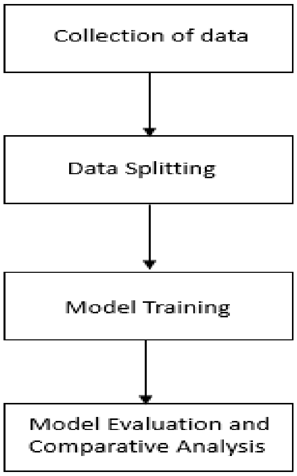

Fig. 1. Workflow of the mechanism

The complete workflow is being depicted in the Fig. 1. The pollutant values gets collected from SPCB after which Data splitting is done into training set and testing set . The data is the trained for the prediction by implementing various machine learning models and a comparison analysis has been done.

III. Results and Discussion

Machine learning Algorithms were modelled to predict AQI. Dataset is split into two sets: 80% as training set and 20% as testing set. The training set is used for training the model and the test data is used for making predictions. Regression models Linear Regression, and Random Forest Regression are implemented. Classification models Logistic Regression, Support Vector Machine and Random Forest Classifier are implemented for the prediction of status of AQI value. Evaluation RMSE, MAE is calculated for evaluating the performance of the regression models. Whereas for evaluation of classification models, accuracy is calculated.

A. *Linear Regression:*

In this work, Linear Regression was implemented using machine learning modules in python. AQI was predicted with respect to Particulate Matter (PM2.5). Root Mean Square Error (RMSE) was obtained as evaluation metrics for future purpose. The RMSE and MAE obtained was 32.358 and 20.02 respectively as shown in Fig 2.

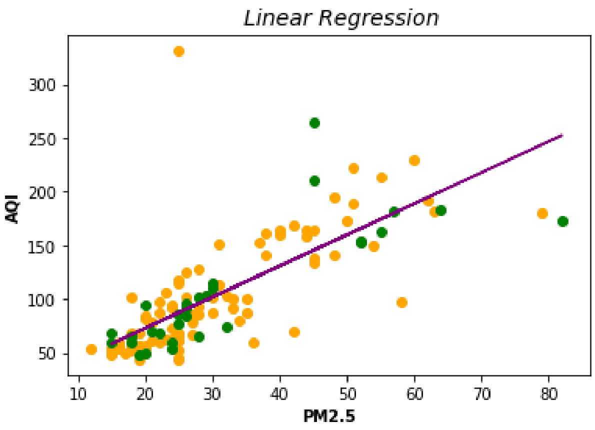

Fig. 2. Predicted value (yellow) vs Real Value (Green) for PM2.5

B. *Random Forest Regression:*

The following plots in Fig. 2-5, are inidvidual AQI value prediction for each pollutant. In this work, PM10, PM2.5, O3 are considered for prediction of AQI.

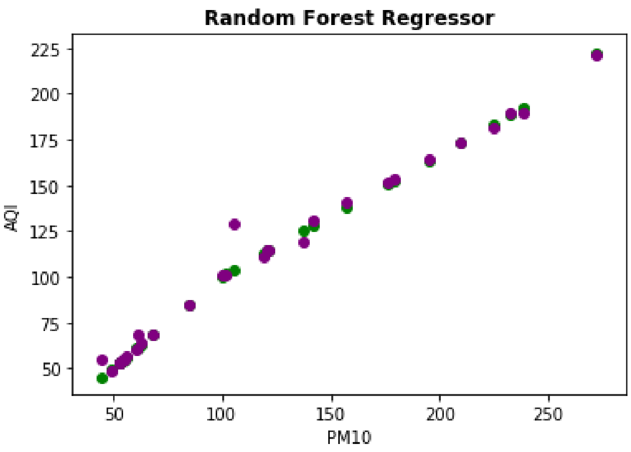

Fig. 3. Predicted value (purple) vs Real Value (Green) for PM10

C. *Logistic regression:*

Here, the prediction is done using logistic regression and PM10, PM2.5, O3 are considered for prediction of AQI as shown in Fig. 6-8. The accuracy of this model before cross validation was 83.5% and after cross validation the accuracy was 93.3%.

D. *Support Vector Machine Classifier:*

Here, the prediction is done using logistic regression and PM10, PM2.5, O3 are considered for prediction of AQI as depicted in Fig. 9-11. The accuracy of this model before cross validation was 90.0% and after cross validation the accuracy was 92.8%.

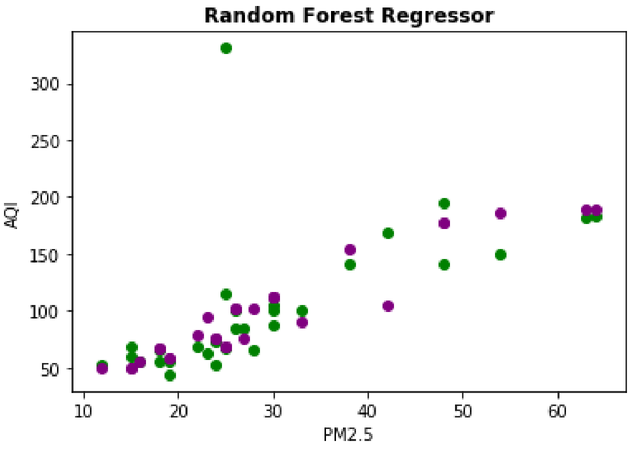

Fig. 4. Predicted value (purple) vs Real Value (Green) for PM2.5

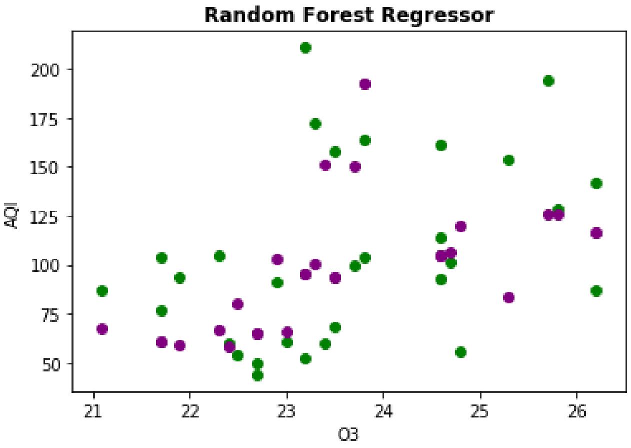

Fig. 5. Predicted value (purple) vs Real Value (Green) for O3

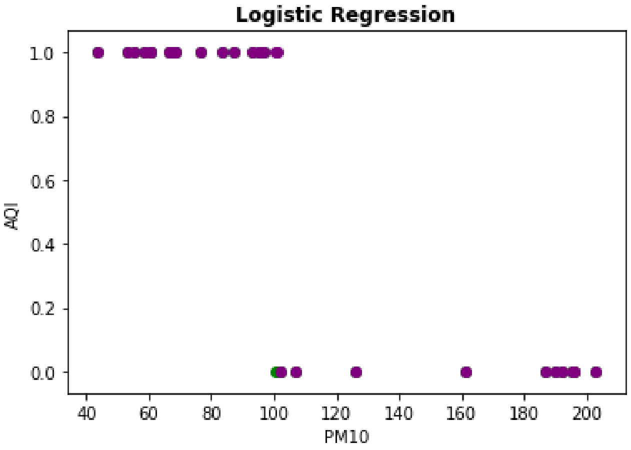

Fig. 6. Predicted value (purple) vs Real Value (Green) for PM10

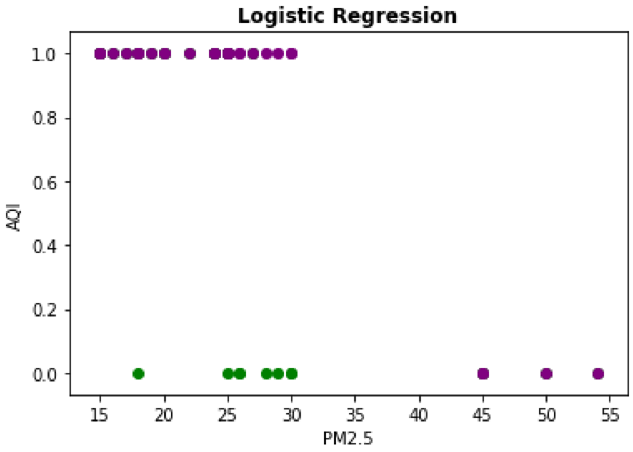

Fig. 7. Predicted value (purple) vs Real Value (Green) for PM2.5

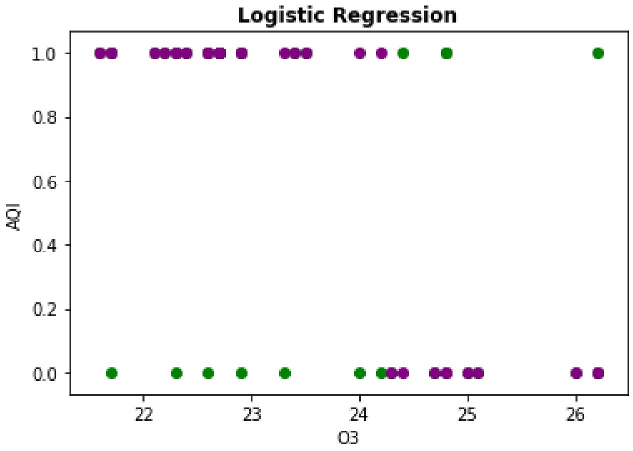

Fig. 8. Predicted value (purple) vs Real Value (Green) for O3

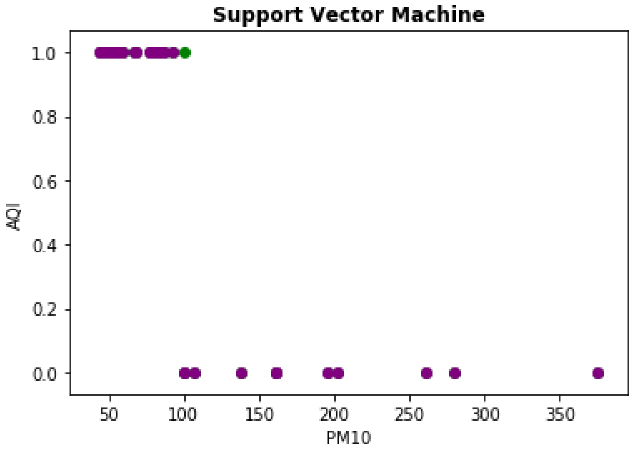

Fig. 9. Predicted value (purple) vs Real Value (Green) for PM10

E. Random Forest Classifier:

Here, the prediction is done using logistic regression and PM10, PM2.5, O3 are considered for prediction of AQI as illustrated in Fig. 12-14. The accuracy of this model before cross validation was 90.0% and after cross validation the accuracy was 93.5%.

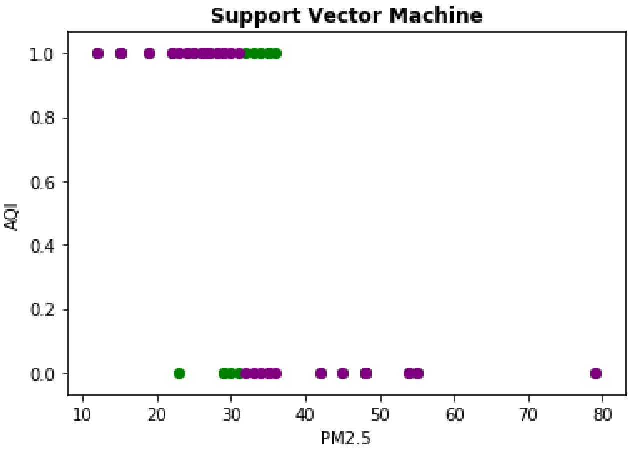

Fig. 10. Predicted value (purple) vs Real Value (Green) for PM2.5

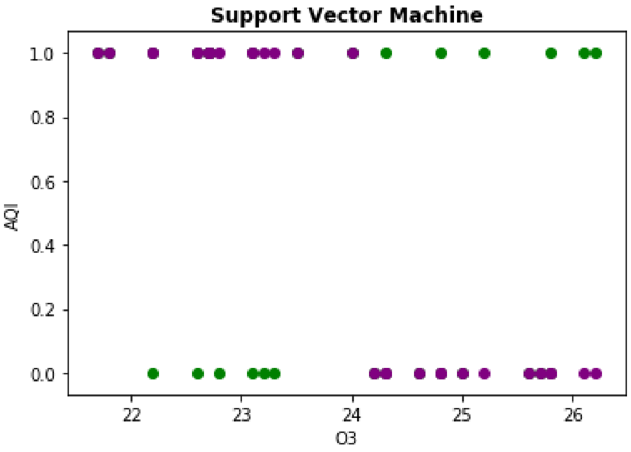

Fig. 11. Predicted value (purple) vs Real Value (Green) for O3

For classification of labels or class, classification tree is used.

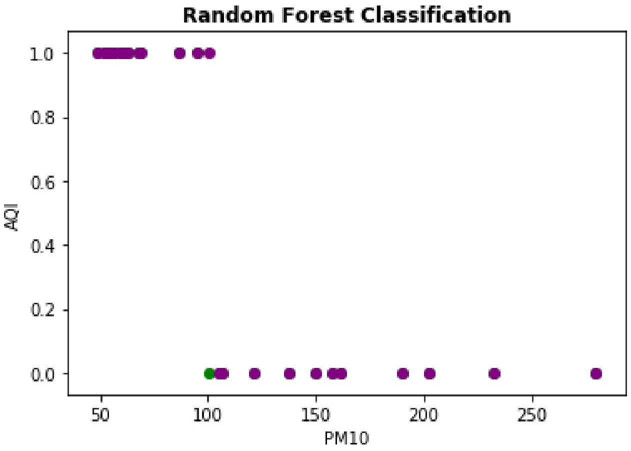

Fig. 12. Predicted value (purple) vs Real Value (Green) for PM10

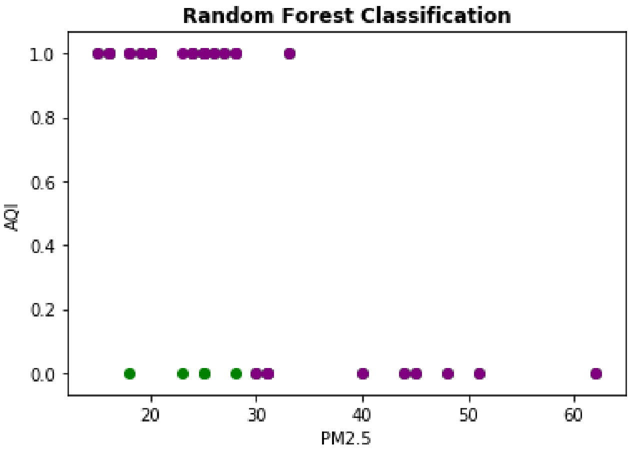

Fig. 13. Predicted value (purple) vs Real Value (Green) for PM2.5

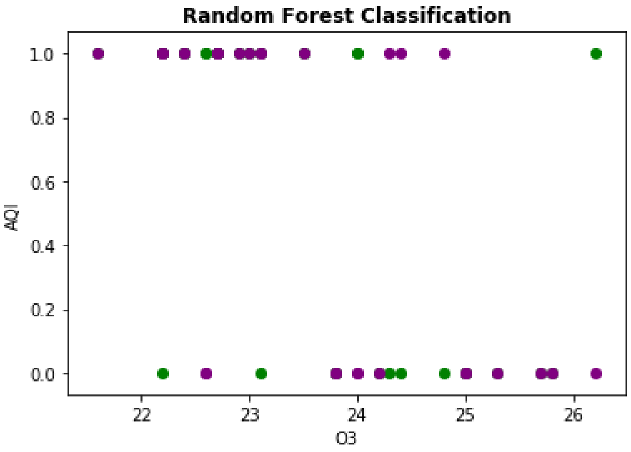

Fig. 14. Predicted value (purple) vs Real Value (Green) for O3

When evaluating the performance of regression models, RMSE and MSE are used. Accuracy is calculated for evaluating the performance of the classification models. To improvise the classification models cross validation is implemented. As the RMSE value is high for Linear Regression, optimization Technique is being implemented to reduce the error in the prediction. The comparison analysis is being shown in Table II and Table III for regression models and for different classification models respectively.

TABLE II
COMPARISON ANALYSIS OF REGRESSION MODELS

Regression Model	**Evaluation Metrices Calculated**
Linear Regression	RMSE: 32.358 MAE : 20.02
Random Forest Regression	RMSE: 2.63 MAE: 3.32

IV. CONCLUSION AND FUTURE WORK

Real time air monitoring is very much needed in today's Life. To obtain such monitoring we need fast and efficient

TABLE III
COMPARISON ANALYSIS OF CLASSIFICATION MODELS

Classification Model	**Metrics Before Cross Validation**	**Metrics After Cross Validation**
Logistic Regression	Accuracy: 83.5%	Accuracy: 93.3%
Support Vector Machine	Accuracy: 90.0%	Accuracy: 92.8%
Random Forest Classifier	Accuracy: 90.0%	Accuracy: 93.5%

prediction such that the user gets information in advance. It has been found out that in regression models, random forest regression is best in our case having RMSE and MAE as 2.63 and 3.32 respectively, and in classification models, the accuracy of random forest classifier is best and is found to be 93.5%. Keeping this in view, the main objective is to predict the air pollutants or the pollution level in the environment,lassifications models have been implemented to classify whether the pollution level is in good range. This work will be extended to App Development or Webpage Development which can be a small contribution to Air Pollution Monitoring and Prediction. Extended Kalman Filter algorithm will be implemented in the future work for the estimation of AQI.

V. ACKNOWLEDGMENT

Under the grant [Grant No. - 7794/2016], the Ministry of Human Resource Development (MHRD) and Urban Development (UD) are funding this research project.

REFERENCES

[1] Singh, Jayant Kumar, and Amit Kumar Goel. ”Prediction of Air Pollution by using Machine Learning Algorithm.” In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 1345-1349. IEEE, 2021.

[2] Sunori, Sandeep Kumar, Pushpa Bhakuni Negi, Kapil Ghai, Amit Mittal, Manoj Chandra Lohani, Mehul Manu, and Pradeep Juneja. ”Prediction of Air Pollutant PM 10 using Various SVM Models.” In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 1541-1546. IEEE, 2022.

[3] Srivastava, Harshit, Shashidhar Mishra, Santos Kumar Das, and Santanu Sarkar. ”An IoT-Based Pollution Monitoring System Using Data Analytics Approach.” In Electronic Systems and Intelligent Computing, pp. 187-198. Springer, Singapore, 2020.

[4] Shahriar, Shihab Ahmad, Imrul Kayes, Kamrul Hasan, Mohammed Abdus Salam, and Shawan Chowdhury. ”Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh.” Air Quality, Atmosphere Health 13, no. 10 (2020): 1247-1256.

[5] Kingsy, Grace R., R. Manimegalai, Devasena MS Geetha, S. Rajathi, K. Usha, and Baseria N. Raabiathul. ”Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data.” In 2016 IEEE Region 10 Conference (TENCON), pp. 1945-1949. IEEE, 2016.

[6] Ayele, Temesegan Walelign, and Rutvik Mehta. ”Air pollution monitoring and prediction using IoT.” In 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1741-1745. IEEE, 2018.

[7] Kiruthika, R., and A. Umamakeswari. ”Low cost pollution control and air quality monitoring system using Raspberry Pi for Internet of Things.” In 2017 International conference on energy, communication, data analytics and soft computing (ICECDS), pp. 2319-2326. IEEE, 2017.

[8] Zhang, Dan, and Simon S. Woo. ”Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network.” IEEE Access 8 (2020): 89584-89594.

[9] Gore, Ranjana Waman, and Deepa S. Deshpande. ”An approach for classification of health risks based on air quality levels.” In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), pp. 58-61. IEEE, 2017.

[10] Kingsy, Grace R., R. Manimegalai, Devasena MS Geetha, S. Rajathi, K. Usha, and Baseria N. Raabiathul. ”Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data.” In 2016 IEEE Region 10 Conference (TENCON), pp. 1945-1949. IEEE, 2016.

[11] Ishak, A. Ben, M. Ben Daoud, and A. Trabelsi. ”Ozone concentration forecasting using statistical learning approaches.” J. Mater. Environ. Sci 8, no. 12 (2017): 4532-4543.

[12] National Air Quality Index, Central Pollution Control Board (CPCB), http://www.indiaenvironmentportal.org.in/files/file/Air%20Quality%20Index.pdf