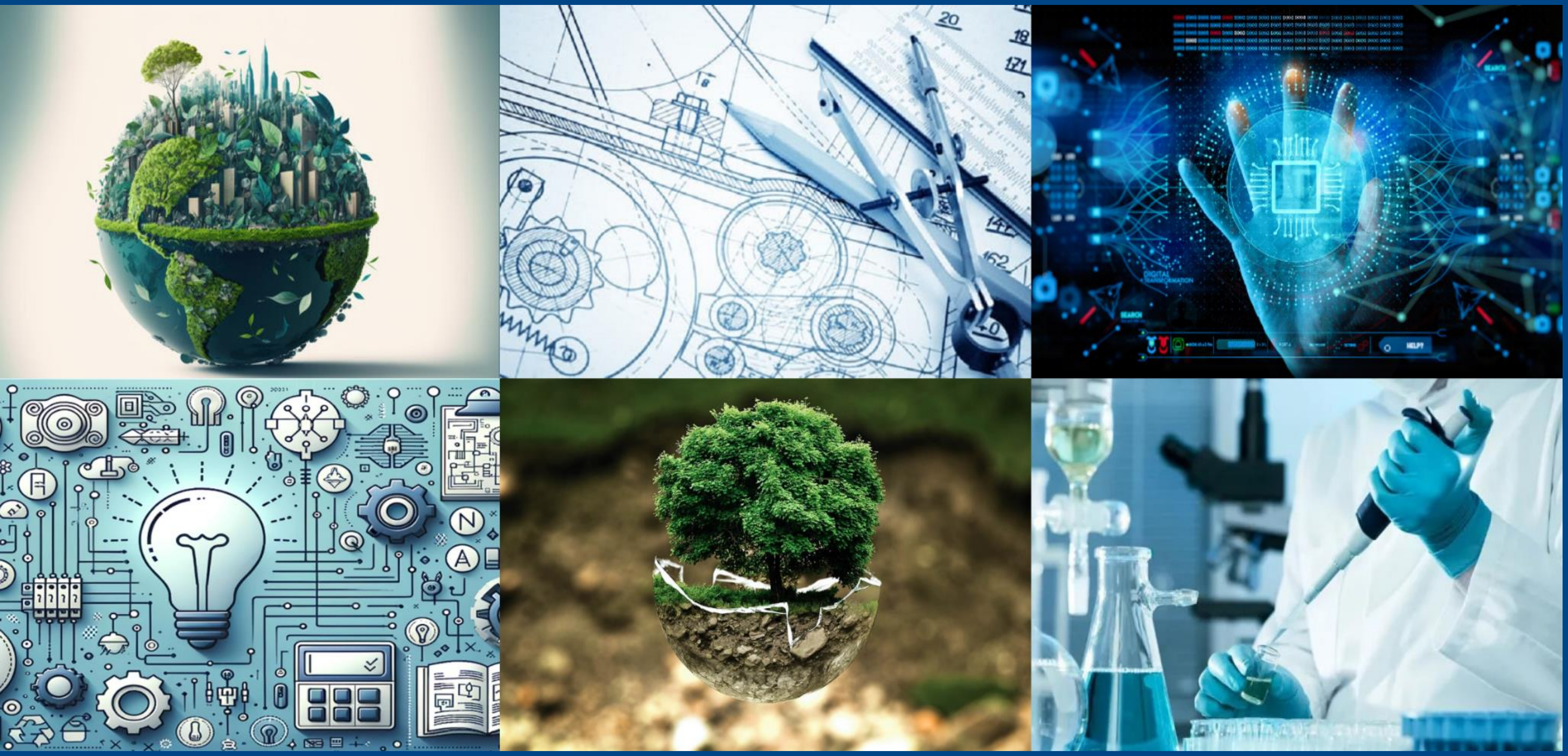
ISSN: 2582-7219

# International Journal of Multidisciplinary Research in Science, Engineering and Technology

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*



Impact Factor: 8.206

Volume 8, Issue 4, April 2025

# Air Quality Prediction by Machine Learning

**T.R.Anand, Sreenaveen S S L**

Assistant Professor, Department of Computer Technology, Dr. N.G.P. Arts and Science College, Coimbatore, India

B.Sc., Computer Technology, Dr. N.G.P. Arts and Science College, Coimbatore, India

**ABSTRACT:** As we all know that pollution in our country is increasing day by day as a result of which the death toll rate around the world is rising vigorously, but among these one cause which is affecting us in a vigorous manner is “air pollution”. Air pollution is now becoming a major issue to human life as well as to the other living organisms. The air quality in our country is decreasing or we can say is being affected day by day which is a matter of concern to the health department. Due to the increase in the pollution level, the air quality index is increasing day by day. Air quality index of India is a high degree statistical factor which is used for getting an analysis of the pollutant present in the atmosphere NO2, Respirable Suspended Particulate Matter, SO2, Suspended particulate matter etc. levels over a period of time. Our idea is basically to do a better analysis of the air quality index by applying various algorithms of machine learning.

**KEYWORDS: Air Quality Index, Machine Learning, Linear Regression, Random Forest, Naïve Bayes**

## I. INTRODUCTION

S. Li, X. Deng et al., [10] provides a novel methodology framework for the spatio-temporal predictions of air pollutants at various time granularities. The BLSTM captures the long-term temporal mechanism of air pollution quite well. The IDW layer, on the other hand, can take into account air pollution's spatial correlation and interpolate its distribution. To verify the efficiency of the suggested methodology, a case study is done. The concentration of PM2.5 in Guangdong, China, is expected to rise. The LSTM network's prediction performance is provided at various time intervals. The projected PM2.5 concentrations and their spatial distribution, as well as the prediction errors, are investigated.The authors compared the suggested technique to ElasticNet, Support Vector Regression, Autoregressive Integrated MA, GBDT, ANN, andRNN. IDW-BLSTM, CNN-LSTM, BLSTM, traditional LSTM, and RNN all perform better than the others[11].A. Masih [13]has systematically reviewed this paper and highlighted the underlying principles of machine learning techniques (LR: Linear Regression, NN: Neural Network and, SVM: Support Vector Machine or Ensemble learning algorithms). During the previous six years, 38 of the most significant papers in the area of engineering have used machine learning(ML) techniques. Author has divided the whole work into two main classes, namely estimation and forecasting of air pollutant in air quality. This paper indicates that linear regression is suited for pollution estimation and the remaining other algorithms like SVM based approaches are suitable for prediction of air quality index.[14]V. Athira et. al., have used $PM_{10}$ as a pollutant in his work. Authors have used RNN (Recurrent Neural Network) and LSTM(LongShort-Term Memory).The goal of this study is to look at a range of big data and (machine learning)ML-based air quality prediction methods. This paper summarises the findings of earlier research on air quality assessment that included artificial intelligence, decision trees, and deep learning, among other methodologies[15].Kalapanidas et al. suggested an ANN ML model for predicting photochemical pollutant concentrations in operational conditions of air quality monitoring. The data source for Athens,Greece, is used in this work [17].

## II. ALGORITHM USED AND EXPERIMENTAL SETUP

In this analysis work three major machine learning(ML) algorithms a) Random Forest(RF), b) Linear regression and, c) Gaussian Naïve Bayes are used (Fig 1).

a. Random Forest(RF):

RF (Random Forest) is a set of trees that may be used for both classification and regression. It reduces the error correlation across various classifiers, raises the correlation, and reduces the error in forest by raising the power of the specific tree. When dealing with large amounts of data, and if the data is equally dispersed, the rate of error is less and efficient. A random forest's two most essential factors are the number of trees utilised in the forest and the number of random variables used in each tree. Based on the Mean Decrease Accuracy (MDA) and, Mean Decrease Gini, it may be used to priorities the relevance of variables in a regression or classification task (MDG). Over-fitting is not achieved by running as many trees as possible in a RF[12].

Linear Regression:

Linear regression (LR); one of the machine learning algorithm is used for the successful prediction of air quality index for the current dataset. A simplified linear regression mathematically

represented in (1) as: I

$Y= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots.. + \alpha_n X_n$ (1) c. Gaussian Naïve Bayes:

The Naive Bayes classifier is a probabilistic machine learning approach for classification tasks. It is based on the Bayes theorem, as shown in (2). The probability :

$P(X/Y) = \underline{(XP/(X)P)(Y)}.(2)$

P(X) is the predictor's prior probability, and P(Y) is the predictor's prior class probability. When a class output label is supplied, P (X |Y) is the likelihood probability of the input feature (X). The probability that the input feature (X) is the label is given by P (Y|X) (Y) [12].

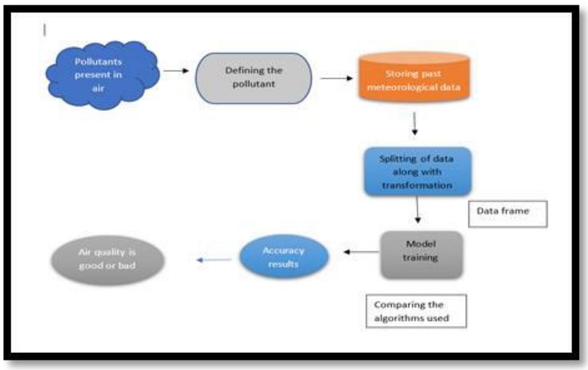

Pollutants present in air
Defining the pollutant
Storing past meteorological data
Splitting of data along with transformation
Data frame
Model training
Comparing the algorithms used
Accuracy results
Air quality is good or bad

Fig 1: Workflow for each method

AQI data set was taken from the Kaggle data repository[16]. The used dataset contains the various parameters related to AQI of Ahmedabad of 2015. It has 16 attributes like city, day, PM2.5, PM10, NO, NO2, NH3, CO, SO2 etc. The value of AQI is depending on the values of these features in the dataset (namely NO2, PM2.5, NH3 etc.), python is used for preprocessing the data and implementation of all the algorithms used in this paper is done in the all the algorithms used in this paper is done in the

| S. No. | Method | Accuracy in percentage |
|---|---|---|
| 1 | Random Forest | 91.77 |
| 2 | Linear Regression | 85.84 |
| 3 | Gaussian Naive Bayes | 79.21 |

Jupyter Notebook' along with several libraries like; sklearn and others are used. Since few entries in the dataset were missing, they get imputed, using the mean values of the rest of the entries in the dataset. Consequently, a data

visualization technique heat map was also observed for the correlation among the features in the dataset. With the help of a heat map, few insignificant features like xylene, benzene, AQI Bucket etc. were discarded.
We have used the total size of data 18314 x10. In order to perform above we broke the dataset in two parts i.e. 80% for training and 20% testing. Training dataset size was 14651 x10, while the test dataset size was 3663 x10. Then normalization was also done to make the dataset in the range o to 1.
After applying the above preprocessing steps, AQI values was predicted using Gaussian Naïve Bayes, Linear Regression and Random Forest. First training was performed and on the basis of training a model was formed, using which testing was performed, and compared the results of the above models (drawn from different algorithms used) with the given actual AQI in the dataset (Table 1).

## III. RESULT AND DISCUSSION

While comparing the actual and predicted values of AQI, it was observed that the comparative values of predicted and actual values of AQI were different, and that too depends upon the algorithm used in this work. On the basis of this we may find the accuracy of the predicted algorithm on the used dataset.
Accuracy = *CorrectlyTotal predicted row _Irow in AQI*
The result obtained has helped us in making a comparative analysis of the used algorithms and hence drawing a better conclusion out of it. While computing the predicted AQI values from these models, we may observe the difference with the actual values of same and that is plotted using scatter plot in Fig. 2, 4, and 6. Out of the three algorithms used "Random Forest and Regression "algorithm has proved its performance to the upmost level (Table 1).
Another visualization technique displot was also used (Fig. 3, 5 and 7) for better understanding in the variance between the actual and predicted values of AQI.

Table 1: Predicted accuracy with each method

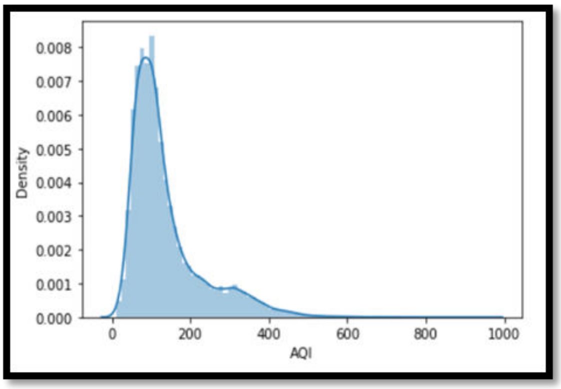


Fig 1: Scatter Plot for Random Forest

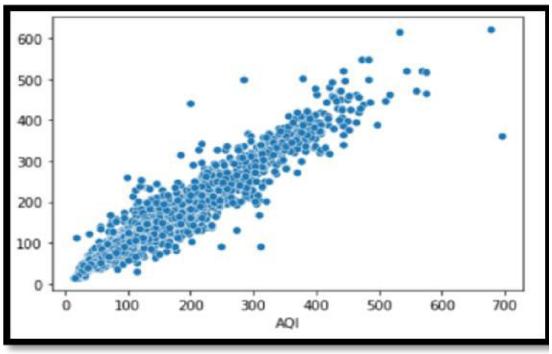


Fig 2: Displot for Random Forest

Fig 3: Scatter Plot for Linear regression

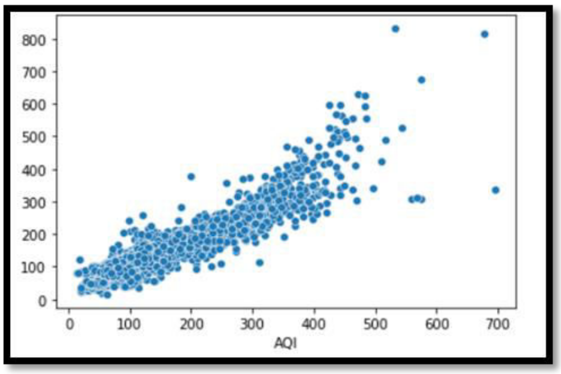
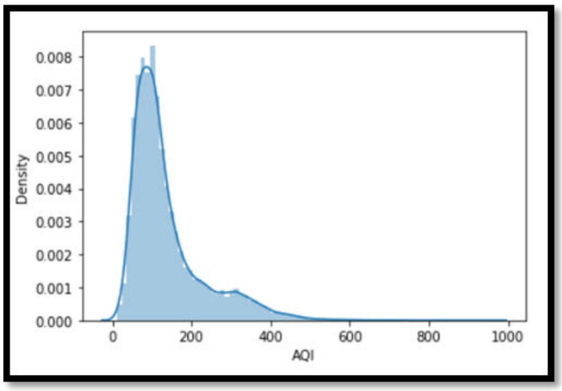


Fig 4: Displot for Linear regression



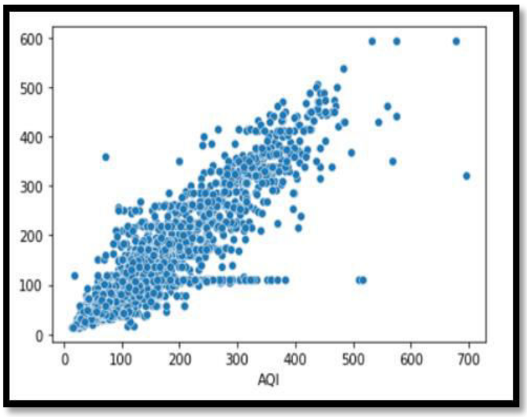


Fig 5: Scatter Plot for Gaussian Naive Bayes

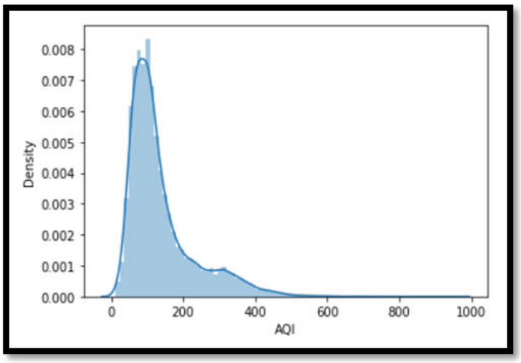
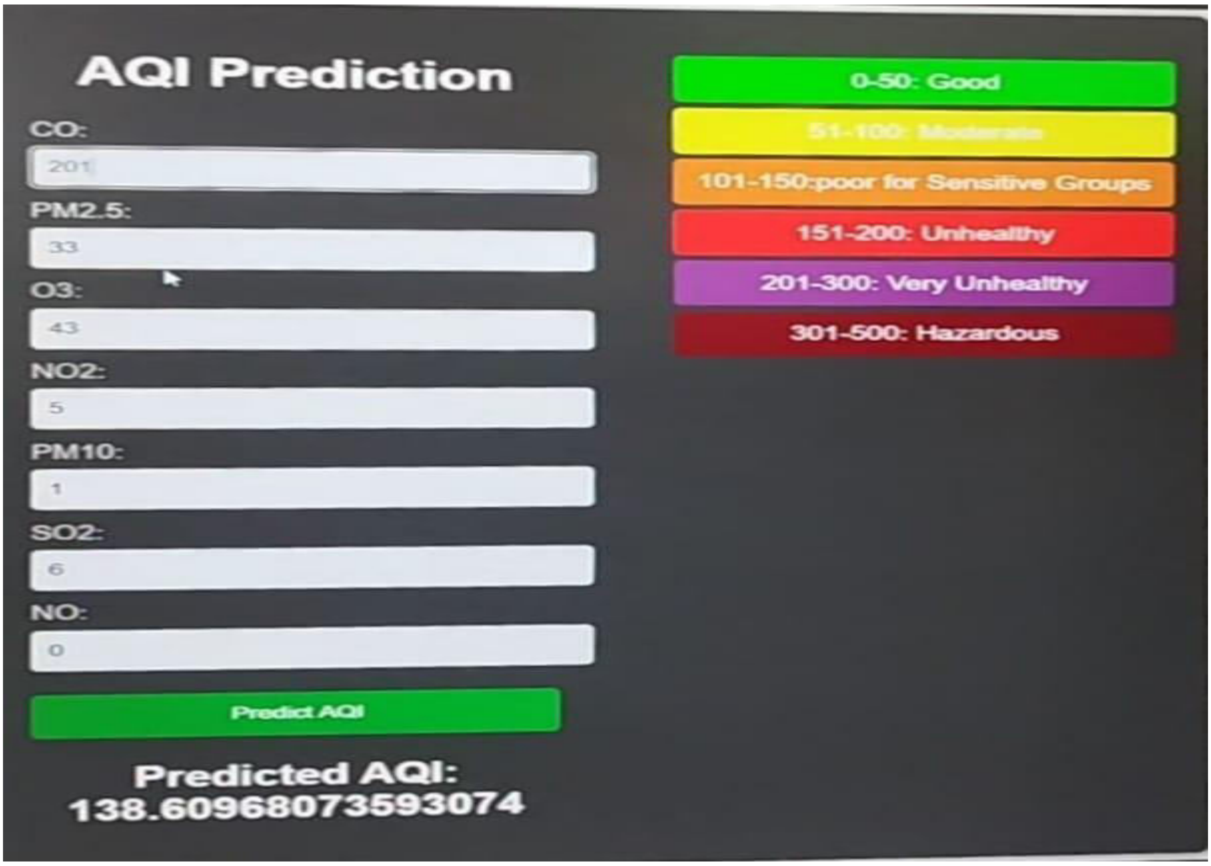


Fig 6: Displot for  Gaussian Naive Bayes

## IV. OUTPUT DESIGN



AQI Prediction

CO: 201

PM2.5: 33

O3: 43

NO2: 5

PM10: 1

SO2: 6

NO: 0

Predicted AQI: 138.6096807359307

0-50: Good

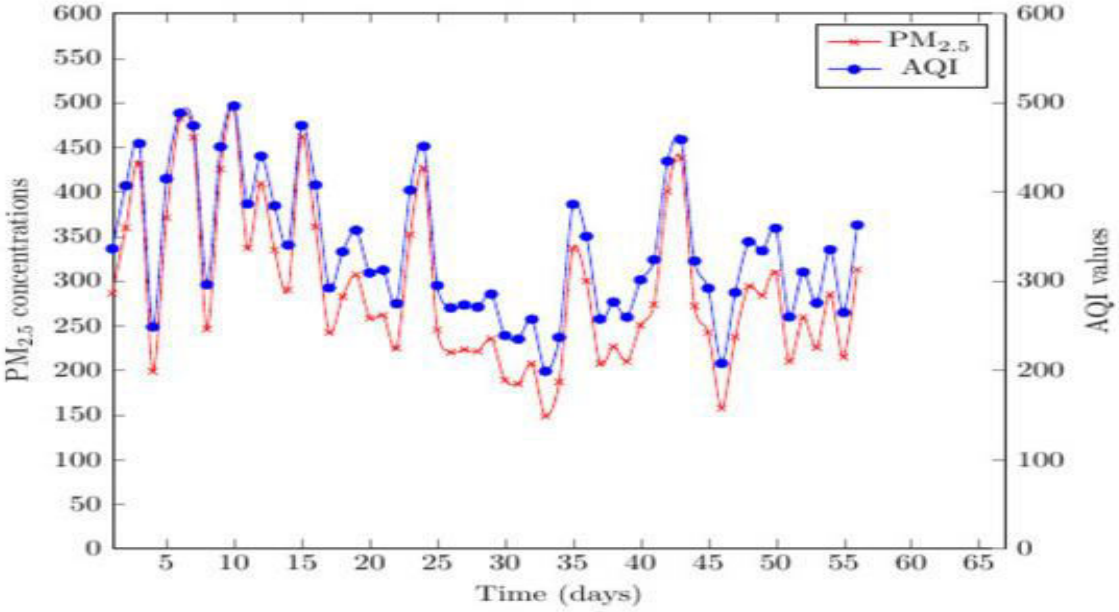
51-100: Moderate

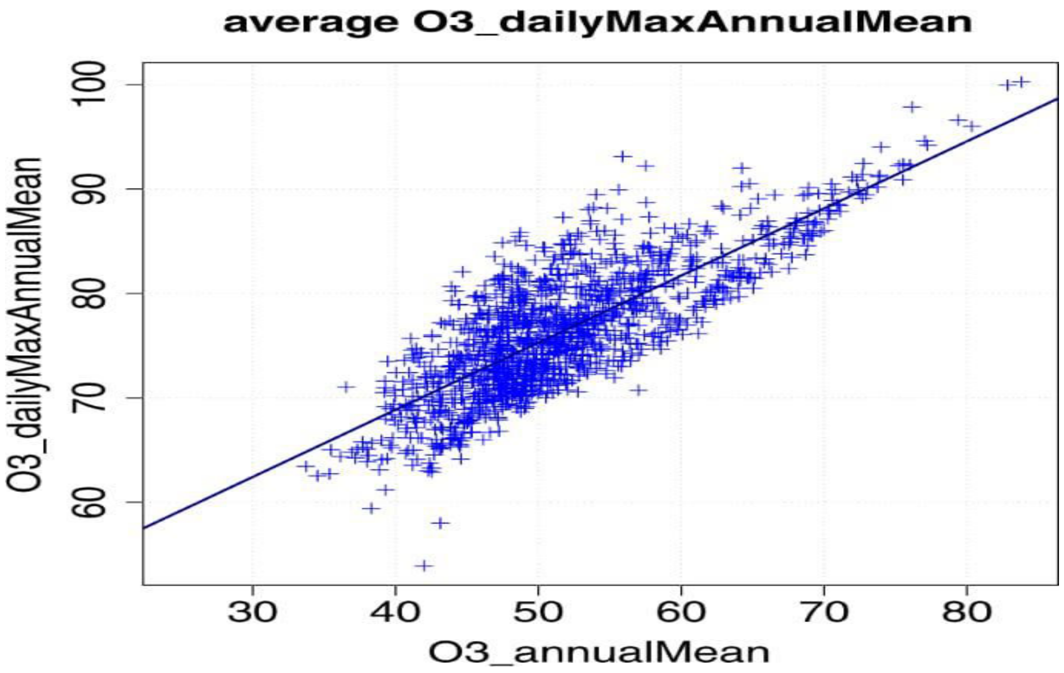
101-150:poor for Sensitive Groups

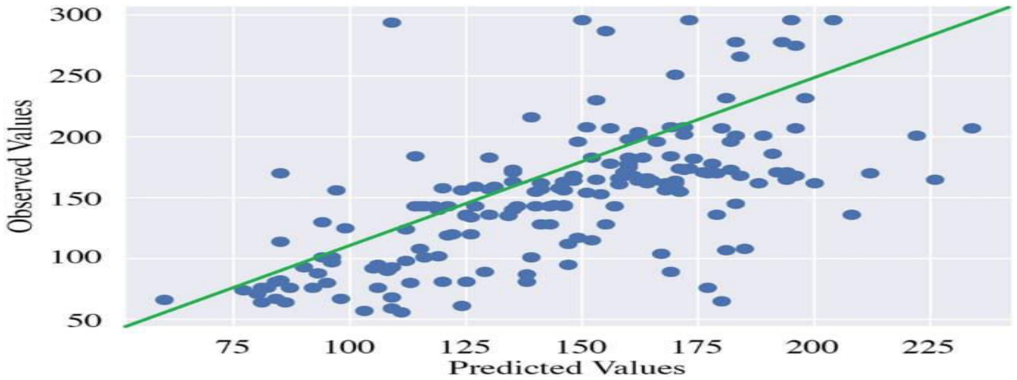
151-200: Unhealthy

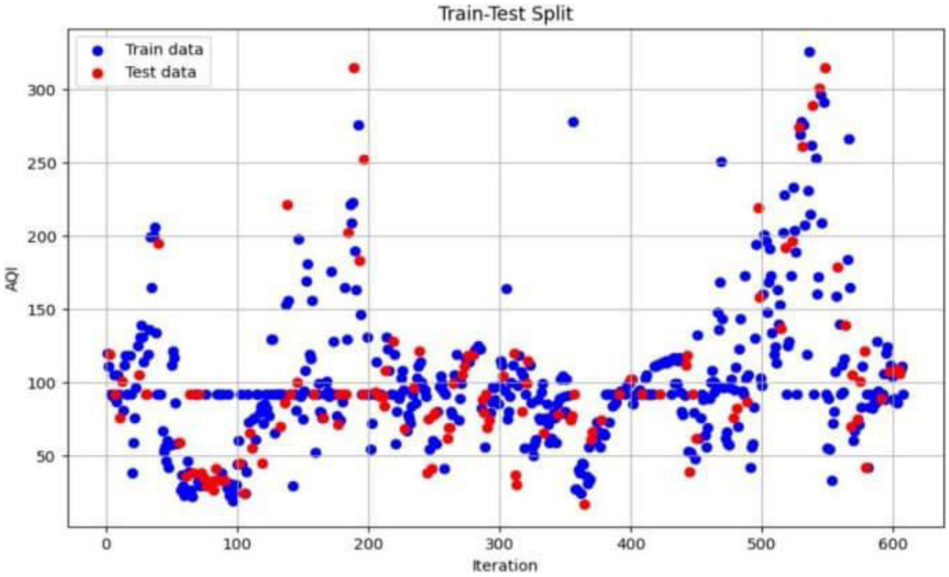
201-300: Very Unhealthy

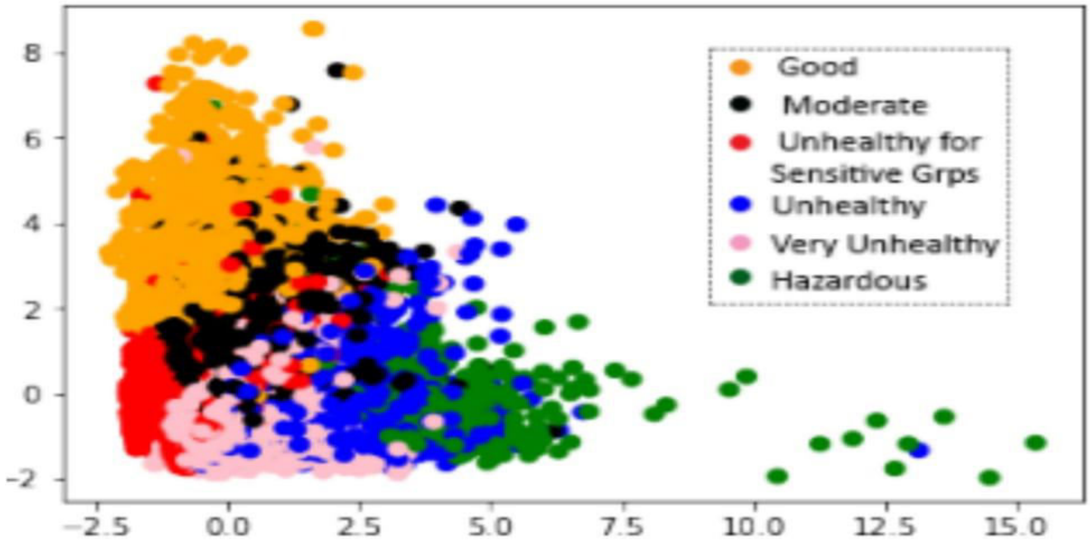
301-500: Hazardous


PM$_{2.5}$
AQI
PM$_{2.5}$ concentrations
AQI values
Time (days)
0
50
100
150
200
250
300
350
400
450
500
550
600
5
10
15
20
25
30
35
40
45
50
55
60
65


average O3_dailyMaxAnnualMean
O3_dailyMaxAnnualMean
O3_annualMean
60
70
80
90
100
30
40
50
60
70
80


Observed Values
Predicted Values
300
250
200
150
100
50
75
100
125
150
175
200
225


Train-Test Split
Train data
Test data
AQI
Iteration
300
250
200
150
100
50
0
100
200
300
400
500
600


Good
Moderate
Unhealthy for Sensitive Grps
Unhealthy
Very Unhealthy
Hazardous
8
6
4
2
0
-2
-2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0

SS

## V. CONCLUSION

From the results of accuracy, scatter plot and displot, we may draw the conclusion that Random Forest gives the best accuracy on this dataset. Therefore, this algorithm may be used for such dataset or similar one. Since matter of pollution is going to be a sensitive issue in future, so the integrity of the dataset must be high. In this technique, various preprocessing technique like imputation, heat map, scatter plot and displot was used for making analysis robust.

## REFERENCES

1. I. Mokhtari, W. Bechkit, H. Rivano, and M.R. Yaici, "Uncertainty-Aware Deep Learning Architectures for Highly Dynamic AirQuality Prediction," IEEE Access, vol. 9, pp. 14765–14778, 2021, doi:
2. 10.1109/ACCESS.2021.3052429.
3. U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, "A machine learning model for air quality prediction for smart cities," 2019 Int. Conf.Wirel. Commun. Signal Process. Networking, WiSPNET 2019, pp. 452–457,2019,doi:10.1109/WiSPNET45539.2019.9032734.
4. Y. Zhang et al., "A Predictive Data Feature Exploration-Based Air Quality Prediction Approach," IEEE Access, vol. 7, pp.30732–30743,2019,doi:10.1109/ACCESS.2019.2897754.
5. D.Zhang and S. S. Woo, "Real TimeLocalized Air Quality Monitoring and Prediction through Mobile and Fixed IoT Sensing Network," IEEE Access, vol. 8, pp.89584594,2020,oi:10.1109/ACCESS.2020.2993547.
   Polytechnique, "Prediction of daily PM 10 concentration using machine learning," 2020.
6. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," Int. J.Environ. Sci. Dev., vol. 9, no. 1, pp. 8–16, 2018, doi: 10.18178/ijesd.2018.9.1.1066.
7. T. R. Patil, "Analysis of Air Quality Estimation based on Air PollutantsParameters," vol. IV, no. 2350.
8. L. Ma, Y. Gao, and C. Zhao, "Research on machine learning prediction of air quality index based on SPSS," Proc. - 2020 Int. Conf. Comput. Network, Electron. Autom. ICCNEA 2020, pp. 1–5,2020,doi:10.1109/ICCNEA50255.2020.00011.
9. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction:Modelregularizationand optimization," Big Data Cogn. Comput., vol.2,no.1,pp.1–15,2018,doi:10.3390/bdcc20100053
10. S.Li,X.Deng,andB. Tang, "Using0DFKLQH Machine / HDUQLQJ Learning 0HWKRGVMethodsIRU for
11. UHGLFWLRQ Prediction RI of $ LU Air 4XDOLW \ Quality LQ in : XOLQJ Mountain $ UHD Area LQ in & KLQD," pp. 426–430, 2021.
12. J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan, "Spatiotemporal Prediction of PM2.5

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com