

# AQI PREDICTION USING MACHINE LEARNING

## MINI PROJECT REPORT

*Submitted in partial fulfilment of the requirements for the award of the degree*

*Of*

## BACHELOR OF TECHNOLOGY

### *in* INFORMATION TECHNOLOGY

*by*

Sayyam Jain

Rohan Khilnani

Vasu Bansal

02311503119

00411503119

35311503119

*Guided by*

**Mr. Puneet Singh Lamba**  
**Assistant Professor**



## DEPARTMENT OF INFORMATION TECHNOLOGY

**BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING  
(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY)  
DELHI – 110089  
MAY 2022**

## **CANDIDATE'S DECLARATION**

---

---

It is hereby certified that the work which is being presented in the B. Tech Mini Project Report entitled "**AQI PREDICTION USING MACHINE LEARNING**" in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology and submitted in the Department of Information Technology of **BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi)** is an authentic record of our own work carried out during a period from **January 2022 to May 2022** under the guidance of **Mr. Puneet Singh Lamba**.

The matter presented in the B. Tech Mini Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

**Sayyam Jain**

**02311503119**

**Rohan Khilnani**

**00411503119**

**Vasu Bansal**

**35311503119**

**This is to certify that the above statement made by the candidate is correct to the best of my knowledge. They are permitted to appear in the External Minor Project Examination**

**Mr. Puneet Singh Lamba**

**Dr. Prakhar Priyadarshi**

The B. Tech Mini Project Viva-Voce Examination of **Sayyam Jain (Enrollment No: 02311503119), Rohan Khilnani (Enrollment No: 00411503119), Vasu Bansal (Enrollment No: 35311503119)**, has been held on .....

**Mr. Ravi Arora**

**Ms. Anshu Khurana**

## **ABSTRACT**

---

---

We forecast the air quality of India by using machine learning to predict the air quality index of a given area. Air quality index of India is a standard measure used to indicate the pollutant (SO<sub>2</sub>, NO<sub>2</sub>, O<sub>3</sub>, PM. etc.) levels over a period. We developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. we improve the efficiency of the model by applying hyper parameter optimization in Random Forest Regressor.

Our model will be capable for successfully predicting the air quality index. In our model by implementing the proposed parameter-reducing formulations, we achieved better performance than the standard regression models. our model has 98.52% accuracy on predicting the current available dataset on predicting the air quality index of Delhi.

The Mini pollutants Such as (NO<sub>2</sub>, SO<sub>2</sub>, O<sub>3</sub>) indexes AQI is acquired, with this individual AQI, the data can be categorised based on the limits. We collected the data from the Indian Central Pollution Control Board, which contains pollutant concentration occurring at various places across India.

We start by calculating the individual index of the pollutant for every available datapoints and find their respective AQI for the region. We have designed a model to predict the air quality index of every available data point in the dataset, our model is capable of forecasting the air quality of India in any given area. By predicting the air quality index, we can backtrack the Mini pollution causing pollutant and the location affected seriously by the pollutant across India. This gives more information and knowledge about the cause and seniority of the pollutants.

## **ACKNOWLEDGEMENT**

---

---

We express our deep gratitude to **Mr. Puneet Singh Lamba** Department of Information Technology for his valuable guidance and suggestion throughout our project work.

We are thankful to **Mr. Ravi Arora, Ms. Anshu Khurana**, for their valuable guidance.

We would like to extend our sincere thanks to **Head of the Department, Dr. Prakhar Priyadarshi** for his time-to-time suggestions to complete our project work. We are also thankful to **Prof. Dharmender Saini** for providing us the facilities to carry out our project work.

**Sayyam Jain**

**02311503119**

**Rohan Khilnani**

**00411503119**

**Vasu Bansal**

**35311503119**

## TABLE OF CONTENTS

---

---

CANDIDATE'S DECLARATION .....	II
ABSTRACT .....	III
ACKNOWLEDGEMENT .....	IV
TABLE OF CONTENTS .....	V
LIST OF FIGURES .....	VI
LIST OF ABBREVIATION .....	VII
CHAPTER 1: INTRODUCTION .....	1
1.1 INTRODUCTION .....	1
1.2 MOTIVATION .....	3
1.3 OBJECTIVE .....	4
1.4 SUMMARY OF THE REPORT .....	4
CHAPTER 2: DESCRIPTION OF PROJECT .....	5
2.1 ABOUT THE PROJECT .....	5
2.2 AIR QUALITY INDEX PREDICTION MODEL .....	6
2.3 DATA SOURCES .....	6
2.4 PRE-PROCESSING THE DATA .....	7
2.5 MODELLING .....	7
2.5.1 RANDOM FOREST REGRESSION .....	7
2.5.2 ACCURACY MEASURE .....	8
CHAPTER 3: RESULTS AND DISCUSSION .....	10
CHAPTER 4: CONCLUSION .....	12

## **LIST OF FIGURES**

---

---

Figure 1: Primary Air Pollutants.....	2
Figure 2: AQI Value, Color Code and Remarks.....	3
Figure 3: AQI Data (Delhi - April 2022 Vs April 2020).....	5
Figure 4: Yearly Air Quality Index for India .....	6
Figure 5: Random Forest Regression Tree Flow.....	7
Figure 6: Best Parameters For Predicting AQI .....	8
Figure 7: Distplot for Prediction .....	10

## **LIST OF ABBREVIATION**

---

---

PM 2.5: Particulate Matter 2.5

AT: Absolute Temperature

RH- Relative Humidity

WS- Wind Speed

CO- Carbon Monoxide

NO<sub>2</sub>-Nitrogen Dioxide

SO<sub>2</sub>- Sulphur Dioxide

AQI- Air Quality Index

# **CHAPTER 1: INTRODUCTION**

---

---

## **1.1 INTRODUCTION**

The Nature is nothing but everything that encircles us. The environment is getting polluted due to human activities and one of the most common is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. The Major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO<sub>2</sub> etc. Carbon Monoxide is produced due to the deficient Oxidation of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel: Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometre or less than that affects more to human health.

The traffic and the energy consumption continue to increase due to the urbanization and industrialization in developing countries. Number of contaminants like CO, CO<sub>2</sub>, SO<sub>2</sub>, NO<sub>2</sub> are released into atmosphere that led to severe air pollution. Air Quality Index (AQI) is an assessment of the air quality that is closely related to human health. Therefore, prediction of Air Quality plays a vital role in management of air pollution. In literature, there are various Air Quality Prediction models: Deterministic Models, Statistical Models, Physical Models, Photochemical models and Machine Learning. Although several models have been proposed, most of the models have high operational and storage overhead. Machine learning approach overcomes these drawbacks and is an excellent tool for solving air pollution problems. In India, the air quality index is calculated based on the concentration of various pollutants.

According to Liu et al, PM2.5 is one of the most harmful air pollutants and has the highest influence on AQI out of all the pollutants. Fine particulate matter (PM2.5) consists of microscopic particles with diameter 2.5 am or smaller. They can penetrate deeply into the lungs and cause number of health problems. A plethora of research has been carried out for the prediction of AQI. Nieto et al applied Autoregressive Integrated Moving Average (ARIMA) timeseries model for the short-term prediction of AQI concentration for Beijing, China using data from multiple sources. Deters et al used algorithms like Boosted Trees (BT), Linear Support Vector Machines (L-SVM), Random Boosted Tree (RBT) to perform classification for the concentration of PM2.5 for two stations in Quito, Ecuador. Regression analysis is performed using, L-SVM, Neural network and Convolutional Generalised Model (CGM). Using regression analysis, CGM gave the best results for the prediction of PM2.5 concentration. Hourly spot concentration forecasting for Ozone, PM2.5 and NO<sub>2</sub> for six cities of Canada was performed using algorithms like Multiple Linear Regression (MLR), Online Sequential Multiple Linear Regression (OSMLR), Multilayer Perceptron Neural Network (MLPNN) and

Online Sequential Extreme Learning Machine (OSEL) by Peng et al. also outperformed other techniques in the prediction of all three pollutants. Urbanczyk et al [6] performed the prediction of PM 2.5 for two stations of the city of Quito, Ecuador.

To perform classification of PM2.5, several algorithms like J48, Zero, Naïve Bayes were used. Out of all these algorithms the best accuracy for classification was given by J48 algorithm. Air Quality Index (AQI) is used to measure the quality of air. Earlier it was difficult and imprecise to know the quality of air and was based on predictions but now due to advancement of technology, it is easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the pollutants needs vigorous analysis. Convolution Neural networks, Recursive Neural networks, Deep Learning, Machine learning algorithms assure in accomplishing the prediction of future AQI so that measures can be taken appropriately. Machine learning which comes under artificial intelligence has three kinds overlearning algorithms, they are the Supervised Learning, Unsupervised learning, Reinforcement learning. In the proposed work we will use these all algorithms to bring more accuracy like supervised learning algorithms such as Linear Regression, Random Forest.

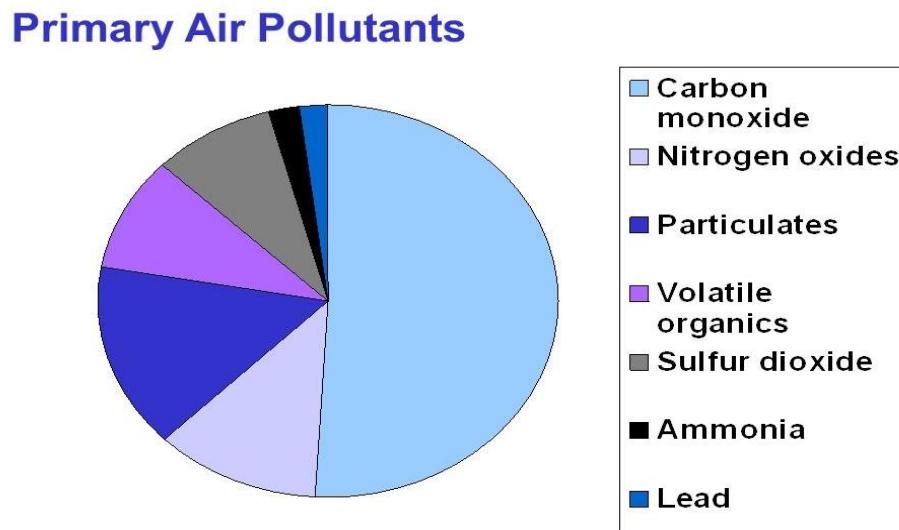


Figure 1: Primary Air Pollutants

AQI	Remark	Color Code
0-50	Good	
51-100	Satisfactory	
101-200	Moderate	
201-300	Poor	
301-400	Very Poor	
401-500	Severe	

Figure 2: AQI Value, Color Code and Remarks

## 1.2 MOTIVATION

Several studies, experiments, and research have been approved over the years to get precise and accurate results for prediction of air quality index using Machine Learning algorithms, here are the details of some foremost research papers, thoroughly examined. Ishan et.al [1] described the benefits of the Bidirectional Long - Short Memory [Belts] method to forecast the severity of air pollution. The proposed technique achieved better prediction which models the long term, short term, and critical consequence of PM2.5 severity levels. In the proposed method predictions made at 6h, 12h, 24h. The results obtained for 12h is consistent, but the result obtained for 6h, and 24h are not consistent. Chao Zhang et.al [2] proposed web service methodology to predict air quality. They provided service to the mobile device, the user to send photos of air pollution. The proposed method includes 2 modules:

- a) GPS location data to retrieve the assessment of the quality of the air from nearby air quality stations.
- b) they have applied dictionary learning and convolution neural network on the photos uploaded by the user to predict the air quality. The proposed methodology has less error rate compared to other algorithms such as PAPLE, DL, PCALL but this method has a disadvantage in learning stability due to this the results are less accurate. Ruijin Yang et.al [3] used the Bias network to find out the air quality and formed DAG from the data set of the town called as shanghai.

The dataset is divided for the training and testing model. The disadvantage of this approach is they have not considered geographical and social environment characteristics, so the results may vary based on these factors. Madhuri VM et .al [4] The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Air Quality Index (AQI) is used to measure

the quality of air. The proposed work is a supervised learning approach using different algorithms such as LR, SVM, DT and RF. The result show that Map Reduction obtained through RF are promising which are analysed with results.

### **1.3 OBJECTIVE**

- Objective of the project is to measure air quality index accurately, by predicting the air quality index, we can backtrack the Mini pollution causing pollutant and the location affected seriously by the pollutant.
- With the Machine Learning based forecasting model, various knowledge about the data is extracted using various techniques to obtain heavily affected regions on a particular region(cluster).
- The air quality index is needed to provide a metric for warning citizens about the dangers of air pollution at varying levels of intensity.
- AQI informs the public about environmental conditions. It is especially useful for people suffering from illnesses aggravated or caused by air pollution. If AQI is not taken seriously it can cause damage to nature and human life. Some of the Risk associated with High AQI to humans is lung diseases, such as asthma, chronic bronchitis, and emphysema.

### **1.4 SUMMARY OF THE REPORT**

In this study, the analysis of air quality dataset has been carried out using various machine learning techniques. The analysis has been performed for the whole dataset. The performance of the models based on all the algorithms have been evaluated based on the Mean Squared Error, Root Mean Squared Error, absolute mean error (AME). It has been observed that Random Forest has the highest accuracy in comparison with other techniques. It has been further observed that the experimental results improve when feature selection methods are used, and all the four models produce good results.

Data collection was done, and suitable data analysed to make the best model for better prediction. Earlier the collected data was trained using Linear Regression, and Lasso Regression. Where, Lasso worked better. Now the data was further trained using Decision Tree and Random Forest Regressor among all Random Forest showed the best accuracy.

## CHAPTER 2: DESCRIPTION OF PROJECT

---

---

### 2.1 ABOUT THE PROJECT

As the largest growing industrial nation, India is producing record number of pollutants specifically Co2, pm2.5 etc and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicate the levels of Mini pollutants on the atmosphere. There are various atmospheric gases which causes pollution on our environment.

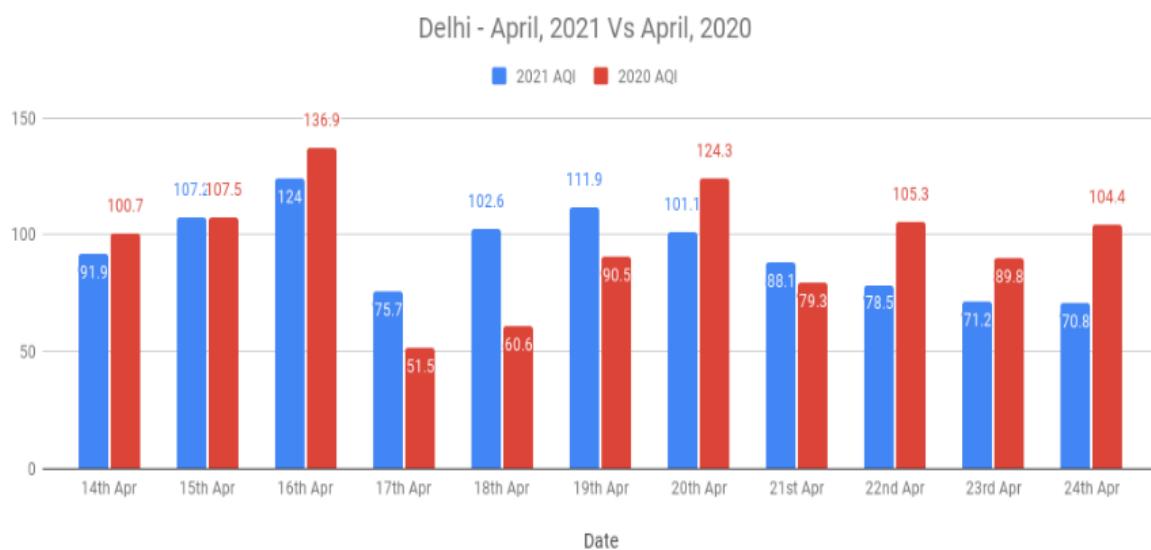


Figure 3: AQI Data (Delhi - April 2022 Vs April 2020)

Each pollution has individual index and scales at different levels. The Mini pollutants Such as (NO2, SO2, O3) indexes AQI is acquired, with this individual AQI, the data can be categorized based on the limits. We collected the data from the Indian government database, which contains pollutant concentration occurring at various places across India. We start by calculating the individual index of the pollutant for every available datapoints and find their respective AQI for the region.

We have designed a model to predict the air quality index of every available data point in the dataset, our model can forecast the air quality of India in any given area. By predicting the air quality index, we can backtrack the Mini pollution causing pollutant and the location affected seriously by the pollutant across India. With this forecasting model, various knowledge about

the data is extracted using various techniques to obtain heavily affected regions on a particular region(cluster). This gives more information and knowledge about the cause and seniority of the pollutants.

## 2.2 AIR QUALITY INDEX PREDICTION MODEL

Fine material (PM2.5) could be an important one as a result of it's a giant concern to people's health once its level within the air is comparatively high. PM2.5 refers to little particles within the air that scale back visibility and cause the air to look hazy once levels are elevated. But in the proposed system we calculate the air quality index of all the pollutants using the AQI formulae to know the air quality level in a particular city using gradient descent and Box-Plot analysis. In the proposed system the air quality index of the upcoming years can be predicted using the present AQI values.



Figure 4: Yearly Air Quality Index for India

## 2.3 DATA SOURCES

To predict the air quality index of a particular region, we need the pollutant concentration of all the gases which will be available in the cpcb.nic.in website, which holds all the data that pollutes the cities every year. The AQI formulae will be applied to calculate the AQI by using the linear regression algorithm for a particular year. Several datasets will be imported inside the directory and null values will be dropped.

## 2.4 PRE-PROCESSING THE DATA

In this dataset the outliers are mainly of faulty sensor or transmission errors, these errors have huge variation than the normal valid results. We know the standard range of pollutants occurs on a particular area so to remove the outliers from the data we use boundary value analysis. By using BVA we found the upper quartile range and lower quartile range of a given data.

## 2.5 MODELLING

The Data was trained and tested using various machine learning algorithms Like Linear Regression, Lasso Regression, Extra Tree Regressor, Random Forest and we got our best results with Random Forest Regression.

### 2.5.1 RANDOM FOREST REGRESSION

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

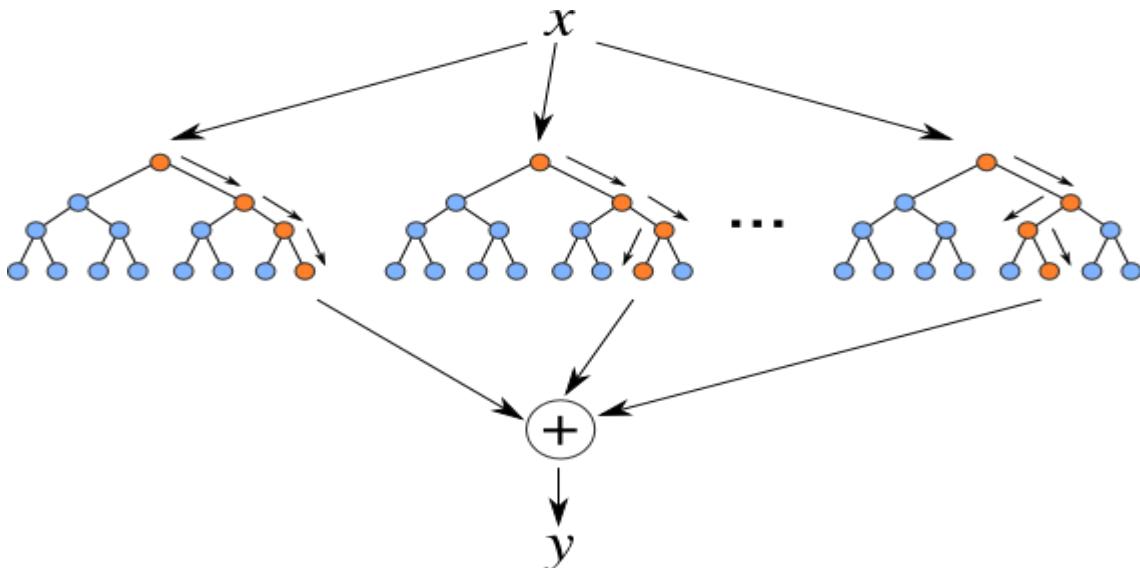
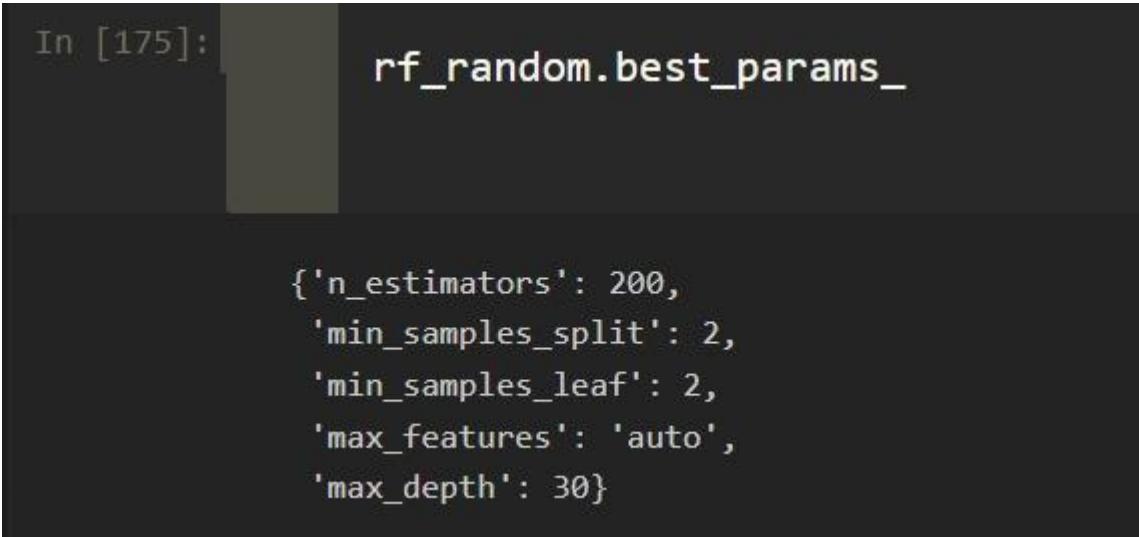


Figure 5: Random Forest Regression Tree Flow

We will use the `sklearn` module for training our random forest regression model, specifically the `RandomForestRegressor` function. The `RandomForestRegressor` documentation shows many different parameters we can select for our model. Some of the important parameters are highlighted below:

- `n_estimators` — the number of decision trees you will be running in the model.

- **criterion** — this variable allows you to select the criterion (loss function) used to determine model outcomes. We can select from loss functions such as mean squared error (MSE) and mean absolute error (MAE). The default value is MSE.
- **max\_depth** — this sets the maximum possible depth of each tree
- **max\_features** — the maximum number of features the model will consider when determining a split
- **bootstrap** — the default value for this is True, meaning the model follows bootstrapping principles (defined earlier)
- **max\_samples** — This parameter assumes bootstrapping is set to True, if not, this parameter doesn't apply. In the case of True, this value sets the largest size of each sample for each tree.



```
In [175]: rf_random.best_params_
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 2,
 'max_features': 'auto',
 'max_depth': 30}
```

Figure 6: Best Parameters For Predicting AQI

## 2.5.2 ACCURACY MEASURE

For measuring the accuracy of the model built we have used three metrics.

1. Root Mean Squared Error: Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. Root mean square error can be expressed as.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N \|y(i) - \hat{y}(i)\|^2}{N}},$$

where N is the number of data points,  $y(i)$  is the i-th measurement, and  $\hat{y}(i)$  is its corresponding prediction.

2. Absolute Mean Error: MAE measures the average magnitude of the errors in a set

of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

3. Mean Squared Error: Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

## CHAPTER 3: RESULTS AND DISCUSSION

---

---

Various machine learning algorithms have been applied such as linear regression, lasso regression, extra tree regressor, random forest and random forest regression with hyperparameter optimization.

The accuracy for our Model is 98.52%. The best accuracy has been achieved using random forest regression along with applying the hyperparameter optimization.

The analysis of air quality dataset has been carried out using various machine learning techniques. The analysis has been performed for the whole dataset and for the subset of features extracted. The performance of the models based on all the algorithms have been evaluated based on the mean absolute error, mean squared error, root mean squared error and accuracy (ACC).

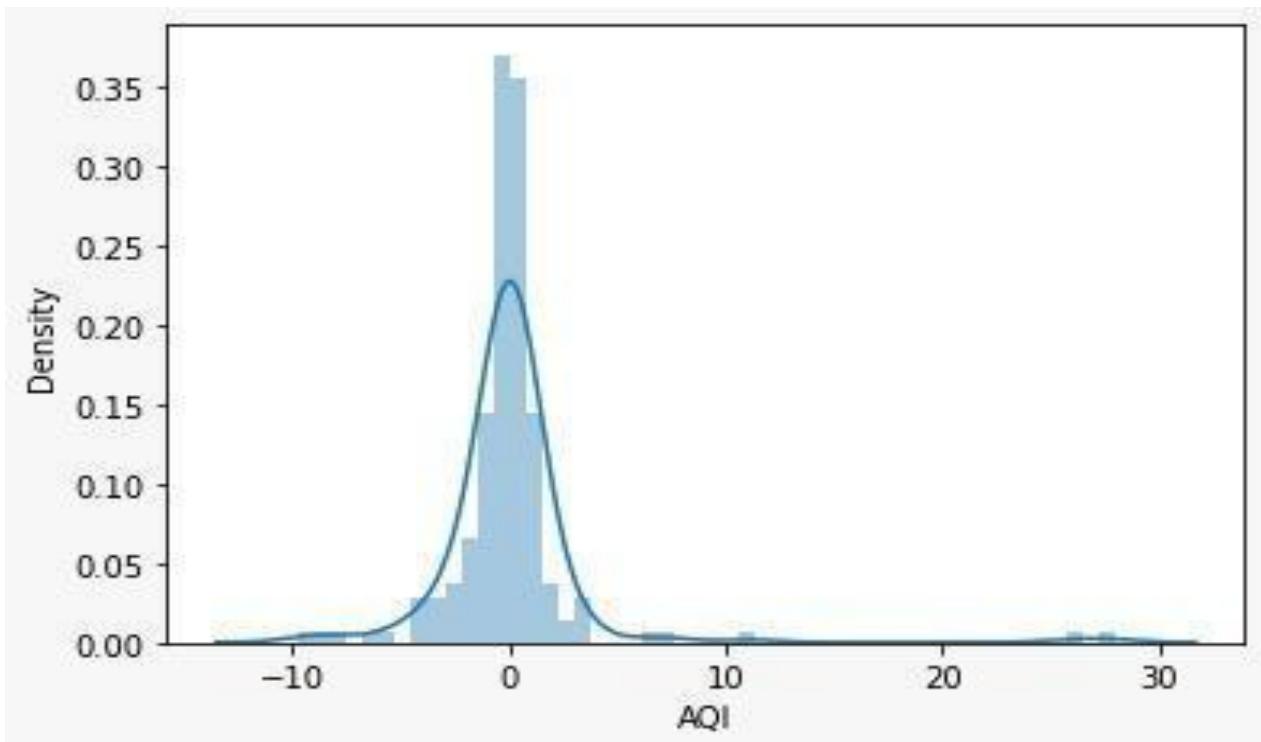


Figure 7: Distplot for Prediction

Three Different Scores were calculated for the determination of accuracy of our machine learning model and the following were:

- Mean Absolute Error (MAE) – **1.48**
- Mean Squared Error (MSE) – **12.48**
- Root Mean Squared Error (RMSE) – **3.53**

This way we achieve to our best possible result using the random forest regression using the hyperparameter optimization technique various other estimators were also applied we vary the number of trees to be used during the training model ranging from 200 to 1200 trees.

Furthermore, we can now work on using the non – supervised learning algorithms where we can group different behaviours of AQI in different seasons in different regions.

## **CHAPTER 4: CONCLUSION**

---

---

Since our model can predict the current data with 98.52% accuracy it will successfully predict the upcoming air quality index of any data within a given region. With this model we can forecast the AQI and alert the respected region of the country also it a progressive learning model it is capable of tracing back to the location needed attention provided the time series data of every possible region needed attention. The air quality information utilized in this paper originates from the China air quality checking and investigation stage, and incorporates the normal every day fine particulate issue (PM2.5), inhalable particulate issue (PM10), ozone (O<sub>3</sub>), CO, SO<sub>2</sub>, NO<sub>2</sub> fixation and air quality record(AQI).The essential perspectives that should be viewed as with regards to gauging of the poison focus are its different sources alongside the components that impact its fixation. Furthermore, we can now work on using the non – supervised learning algorithms where we can group different behaviours of AQI in different seasons in different regions.

## REFERENCES

---

---

- [1] Jasleen Kaur Sethi, Mamta Mittal (2018). A Study of Various Air Quality Prediction Models. Circulation in Computer Science, ICIC 2017, 128-131.
- [2] Wang, D., Wei, S., Luo, H., Yue, C., & Grinder,(2017). A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Science of The Total Environment, 580, 719-733.
- [3] Verma, Ishan, Rahul Ahuja, Hardik Meshier, and LipikaDey. "Air pollutant severity reduction using Bi-directional LSTM Network." In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 651-654. IEEE, 2018.
- [4] Figures Zhang, Chao, Baoxian Liu, Junchi Yan, Jinghai Yan, Lingjun Li,Dawei Zhang, Xiaoguang Rui, and RongfangBie. "Hybrid Measurement of Air Quality as a 5 Fig. 8. RH w.r.t tin oxide Fig. 9. RH w.r.t C6H6 Mobile Service: An Image Based Approach." In 2017 IEEE International Conference on Web Services (ICWS), pp. 853- 856. IEEE,2017.
- [5] Yang, Ruijun, Feng Yan, and Nan Zhao. "Urban air quality based on Bayesian network." In 2017 IEEE 9th Fig. 10. RH w.r.t NO Fig. 11. RH w.r.t NO<sub>2</sub> International Conference on Communication Software and Networks (ICCSN), pp. 1003-1006. IEEE,2017.
- [6] Madhuri VM, Samaya Gunjal GH, Savitha Kamalapurkar "Air Pollution Prediction Using Machine Learning Supervised Learning Approach" In 2020 international journal of scientific and technology research volume 9 pp. ISSN 2277-8616
- [7] Kumar, Anikender and P. Goyal, "Forecasting of daily air quality index in Delhi", Science of the Total Environment 409, no. 24(2011): 5517- 5523.
- [8] Kumar, Anikender and P. Goyal, "Forecasting of daily air quality index in Delhi", Science of the Total Environment 409, no. 24(2011): 5517-5523.
- [9] Singh Kunwar P., et al. "Linear and nonlinear modelling approaches for urban air quality prediction, "Science of the Total Environment 426(2012):244-255.
- [10] Sivakumar R, et al, "Air pollution modelling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001): 471-477
- [11] Gokhale sharad and Namita Raokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008): 9-24.

- [12] Bhanarkar, A. D., et al, "Assessment of contribution of SO<sub>2</sub> and NO<sub>2</sub> from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745-India." Atmospheric Environment 39.40 (2005):7745-7760.
- [13] Challa Venkara Srinivas et al," Data Assimilation and performance of Wrf for Air Quality Modelling in Mississippi Gulf Coastal Region "
- [14] Hutchison Keith D., Solar Smith and Shazia J. Farooqui, "Correlating MODIS aerosol optical thickness data with ground-based PM2.5 observations across Texas for use in a real time air quality prediction system, "Atmospheric Environment 39.37(2005) :7190 – 7203.
- [15] Wang Z et al, "A nested air quality prediction modelling system for urban and regional scales: Application for high high-ozone episode in Taiwan " Water, Air and Soil Pollution 130.1-4(2001):391-396
- [16] Nallakaruppan, M. K., and U. Senthil Kumaran. "Quick fix for obstacles emerging in management recruitment measure using IOT- based candidate selection." *Service Oriented Computing and Applications* 12.3-4 (2018): 275- 284.
- [17] Nallakaruppan, M. K., and Harun Surejllango. "Location Aware Climate Sensing and Real Time Data Analysis." *Computing and Communication Technologies (WCCCT), 2017 World Congress on*. IEEE, 2017.
- [18] Mittal, M., Goyal, L. M., Sethi, J. K., & Hemanth, D.J. (2019). Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning. Computational Economics, 53(4), 1467-1485.
- [19] Mittal, M., Goyal, L. M., Hemanth, D. J., & Sethi, J.K. Clustering approaches for high-dimensional databases: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1300.
- [20] Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental journal of computer science & technology, 8(1), 13-19.
- [21] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31, 249-268.
- [22] Sánchez, A. S., Nieto, P. G., Fernández, P. R., del Coz Díaz, J. J., & Iglesias-Rodríguez, F. J. (2011). Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area(Spain). Mathematical and Computer Modelling, 54(5-6),1453-1466.
- [23] Koenker, R., & Hallock, K. F. (2001). Quantity regression. *Journal of economic perspectives*, 15(4), 143-156.

- [24] Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1), 74-89.
- [25] Lewis, M. (2007). Stepwise versus Hierarchical Regression: Pros and Cons. Online Submission.
- [26] Denil, M., Matheson, D., & De Freitas, N. (2014, January). Narrowing the gap: Random forests in theory and in practice. In International conference on machine learning (pp. 665-673).
- [27] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar), 1289-1305.
- [28] Gunawardena, A., & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(Dec), 2935-2962.
- [29] Hemanth, D. J., Anitha, J., & Mittal, M. (2018). Diabetic retinopathy diagnosis from retinal images using modified hopfield neural network. *Journal of medical systems*, 42(12), 247.
- [30] Dholakia, H. H., Purohit, P., Rao, S., & Garg, A. (2013). Impact of current policies on future air quality and health outcomes in Delhi, India. *Atmospheric environment*, 75, 241-248.
- [31] Li, H., Wang, Y., Wang, H., & Zhou, B. (2017). Multi-window-based ensemble learning for classification of imbalanced streaming data. *World Wide Web*, 20(6), 1507-1525.
- [32] Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., & Zhang, X. (2017). A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web*, 20(2), 325-350.