Machine learning's applicability in modelling atmospheric particle pollution in India

Prateek Parashar[1], Abhijeet Ghldiyal[2], and Atul Kumar Srivastava[3]

[1, 2, 3] *School of Computing, DIT University, Dehradun, India*

E-mail address: parasharprateek6423@gmail.com, abhijeet.ghildiyal928@gmail.com, atulbhuphd@gmail.com

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: Air pollution has been increasing from the past century at a very fast pace. Its adverse effects are clearly visible in almost every country, especially in India. There are many factors responsible but major factors are particulate matter in air (mainly PM2.5 and PM10). High enough concentration of these in air might cause breathing problems and might damage the lung permanently if inhaled for long time. By monitoring and predicting their concentration in air, we can use preventive measures to tackle their concentration and mitigate the problem. We will predict their concentration using various machine learning models like Decision Tree Regression, Random Forrest Regression (RFR) and Gradient Boosting Regression. NO, NO2, NOx, NH3 and CO, along with meteorological variables from AP001 (Amravati) for the period of 2015 to 2019, were utilized as exploratory variables.

Keywords: Air Pollution, Machine Learning, Particulate Matter, Regression Analysis.

1. INTRODUCTION

Particulate matter is a threat to public health [1]. There are various health hazards associated with concentration of PM 2.5 and PM 10, such as respiratory diseases, cancer, and cardiovascular disease [2] A meta-analysis study Kim et al. 2015 revealed that 3% of cardiorespiratory and 5% of lung cancer deaths are related to PM exposure. This study also showed that PM particles are far more dangerous than any other particles present in the atmosphere. Furthermore, a study showed that an increase of 1gm-3 in PM2.5 accelerates the mortality rate of the Covid-19 disease by 15%. These PM particles have this effect because of their size and composition of particles. PM10 has a size of 10 micrometres and PM2.5 is generally 2.5 micrometres in diameter. Both PM2.5 and PM10 particles are constituted by other classes of pollutants such as sulphates, nitrates, and other volatile organics. These are emitted from natural and artificial sources and as these particles coz health hazards it is essential to implement proper mechanisms to regulate them. Pollution modelling can act as the first step of controlling mechanisms.

This modelling process or APM (Atmospheric Pollution Modelling) is the tool that illustrates the casual relationship between various emissions, atmospheric concentrations, and depositions.

There have been many studies and each study used a different kind of model for their approach, but each of these studies demonstrate the regional relationships between PM2.5 and PM10 concentration with meteorological parameters and other pollutants for e.g., NOx, SO2, CO, and ozone(O3).

2. LITERATURE REVIEW

In a lot of recent studies modelling, analysis and forecast of air pollutants is being studied. Masood Adil et al. [2] developed ANN models forecasting applications in semi humid climatic conditions, their results exhibited better performance for PM2.5 concentrations using ANN-based models rather than SVM models. In a study in 2017 Biancofore F, et al. used the air quality information of urban areas for building PM10 concentrations [3]. During this study M-Linear Regression and Neural Networks performed poorly. In another study Qin S et al (2014) proposed a forecasting algorithm based on a trajectory-based search algorithm, Ensemble Empirical Mode Decomposition and Back propagation with atmospheric parameters as inputs [4]. Most of the recent works have involved only the use of either Linear Regression or SVR. Few works have used these atmospheric variables using Decision Tree Regressor, Random Forest, and Gradient Boosting. These studies are the motivation behind our comprehensive study to develop an accurate model to forecast outdoor pollution.

3. MODELS

There are various models has been used to analyze the Air pollution.

A. Decision Tree Regression

Although Linear Regression is easier to execute, it can't be used to study non-linear relationships [5]. This model

contains nodes with decisions by listing the decision consequences of it and to choose decision node in each step according to some rules for the outcome will be optimal as shown in figure 1.

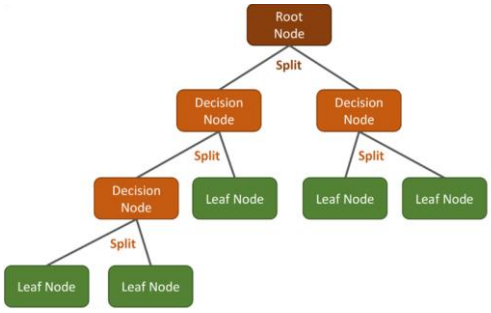

Figure 1. Decision tree model

In this model we choose a variable, to control the complexity of the model. We choose various parameters to build trees and to evaluate training and test error. After evaluation on several variables MSE measurements are drawn. This is a huge improvement when compared to the errors in linear regression

B. *Random Forest Regressor*

Random Forests are more suitable for predicting when compared to ANN or SVR as they don't have parameters and don't need any statistical information about the variables [6]. The main feature of this model is that it can create sub features from the given features. These can be applied to both classification and prediction, it is widely used to predict stock prices, in recommender systems and even identifying various compositions for any manufacturing industry.

In this model a test is carried out at each node on an attribute, each branch of the decision tree then carries the outcome and terminals hold the feature label. Using many splitting variables, the data set is broken down into leaves. This algorithm calculates the information gain and predicted entropy.

- Take 'x' number of trees.
- S [1 to x] samples are taken from a dataset; all these samples contain various predictors (randomly selected).
- Using the sample, a tree is created 'T'. All the parameters and mid points among the x samples are selected by increasing the information gain to break down each leaf into a pair of smaller leaves.
- Final decision is taken by computing the predictions of the 'x' trees as shown in Fig 2.

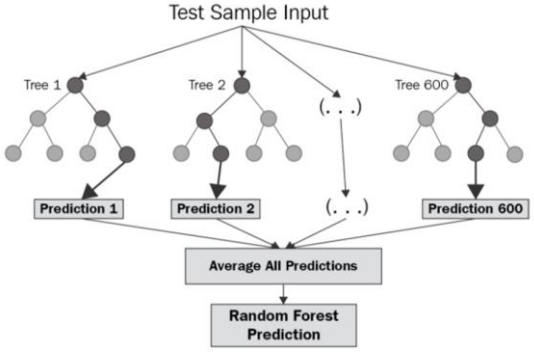

Figure 2. Random forest regressor model

C. *Gradient Boosting Regressor*

GBR is the idea that a weak learner can be changed to become better [7]. This model works on three major elements: Optimizable function, Learner to make predictions and an additive model to increase the number of learners to decrease the Optimizable function. This algorithm can fit any training dataset easily. The model is mainly used for forecasting and image recognition.

In this model the method used to increase the contribution of every tree by some amount is called Shrinkage. If the rate at which the model learns is reduced, the model will in turn have more trees, increasing the efficiency.

4. EXPERIMENT AND RESULT ANALYSIS

Dataset has been download from the Kaggle of various states of India for the research analysis purpose. Data Pre-processing for this dataset included dropping unwanted columns and features and printing out the relevant features of the dataset, after filtering out the required characteristics we removed redundant and null values from all the desired features. For the missing values we filled in the gaps with the mean of the column. Now before exploring the machine learning models on the dataset, we performed exploratory data analysis to gain knowledge about it.

Table 1. Literature analysis of various researchers

Author	Year	Study	Models	Parameters	Accuracy
Doreswamy, Harishkumar KS, Yogesh KM, Ibrahim Gad [1]	2020	Air Pollution forecasting using ML models	RF, Gradient Boosting, Decision Tree Regression, MLP Regression, Lasso, Ridge	Ozone, CO, NO_2, SO_2 , time, date, PM values	0.8891
Adil Masood, Kafeel Ahmad [2]	2020	PM predictions in New Delhi using ML models	ANN, SVM	$PM_{2.5}$, CO, SO_2, NO, C_7H_8, NO_2, Wind Speed and direction	0.8567
Anurag Barthwal, Debopam Acharya, Divya Lohani [5]	2021	Analysis of PM concentration using ML techniques	Multiple Linear Regression, SVR. RF, Gradient Boosting	CO, Ozone, SO_2 AQI, $PM_{2.5}$, PM_{10}	0.95
A. Suleiman, M.R. Tight, A.D. Quinn [6]	2019	Traffic related PM concentration using ML methods	ANN, SVM, Boosted Regression	Year, NO_2, Vehicle Emissions, Temperature, Wind Direction, Wind Speed, Humidity	0.95
Nurul Amalin Fatihah Kamarul Zaman, Kasturi Devi Kanniah, Dimitris G. Kaskaoutis, Mohd Talib Latif [7]	2021	PM2.5 concentration in Malaysia using ML techniques	Support Vector Regression, Random Forest	Ozone, CO, NO_2, SO_2, time, date, PM values	0.69
Limei Ma, Chen Zhao, Yijun Gao [8]	2020	Air Quality Index using SPSS Machine learning techniques	Multivariate Linear Regression Model	AQI, Date, $PM_{2.5}$, PM_{10}, NO_2, Ozone, Temperature	0.897
Fabiana Franceschi, Martha Cobo, Manuel Figuerdo [9]	2018	PM2.5 and PM10 concentrations in Bogota and Colombia	Multi layered Perceptron model, K means Clustering	Year, Month, DOW, Wind Speed, meridional component, Temperature, Humidity	0.72
Xinzhi Lin [10]	2021	Predicting PM concentration using ML models	RF, linear regression and NN	Dew point, Temperature, Pressure, Wind Speed, Conc. Of SO_2, NO_2, CO, $PM_{2.5}$ and PM_{10}	0.742
Chavi Srivastava, Amit Prakash Singh, Shyamli Singh [11]	2018	Air pollution in delhi using ML models	ANN, SVM	RH, Temperature, Wind Speed, AQI, Humidity	0.812

Figures 3 and 4 answer our curiosity about the pattern the of PM2.5 and PM10 concentration follow on an hourly basis.

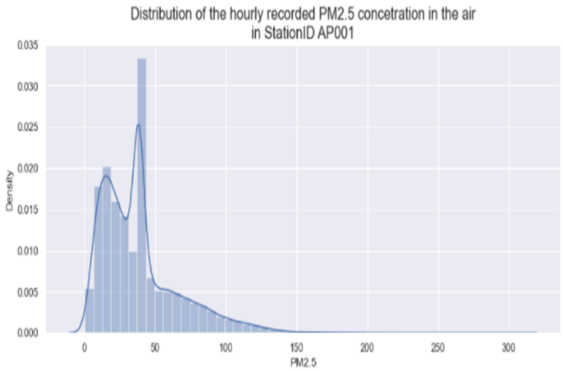

Figure 3. Hourly distribution of PM2.5 in AP001

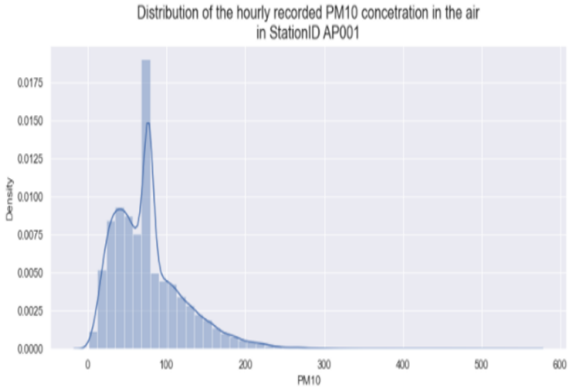

Figure 4. Hourly distribution of PM10 in AP001

Table 1 shows the description of all the variables in the dataset and the relationship between the variables is shown using the heatmaps in figure 5 and figure 6. Figure 5 shows the relationship between PM2.5 and other variables, whereas Figure 6 shows the relationships for PM10. Apart from the meteorological parameters, PM was highly dependent with other air pollutants (as seen in the heatmaps).PM was significantly related with NO2 and NOx scaling up to 56-66%.

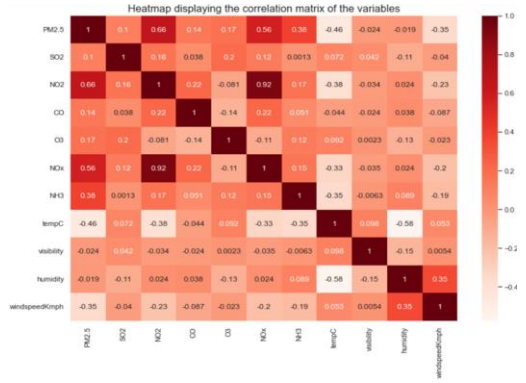

Figure 5. Heatmap PM2.5

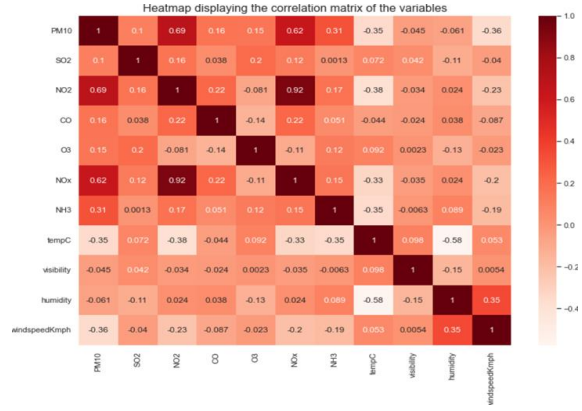

Figure 6. Heatmap PM10

The statistics from Table 1 for air pollution show that the mean concentration for PM10 were observed to be 77.0079 μg m−3 which is above average when it comes to PM10 levels in the world. Whereas the PM2.5 concentration levels were found to be 3.38 μg m−3. These concentrations in AP001 are 1.5 times that of the standard value of annual PM10 is 50 μg m−3 but the value of PM2.5 in AP001 is below the average value that is 15 μg m−3.

Among the three models: Decision Tree Regression, Random Forest Regression and Gradient Forest Regression the best results were shown by the Gradient Boosting Model (as shown in Table 2 and Table 3) scaling up to 88 and 89% accuracy for PM 2.5 and PM10 respectively.

From the below given graphs (Figure 7 and 8) we can observe that the PM10 and PM2.5 concentrations was at its peak during the winter of 2019. After 2019 the concentration levels in AP001 Amravati have started to decrease and are half of what they were back in 2019.

Table 2. Description of Dataset

	PM2.5	PM10	NO2	NH3	CO	SO2	O3	humidity	tempC	visibility	windspeed
count	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.000000	22776.0
mean	3.384328	77.007920	21.938381	12.084864	0.618724	14.096526	38.142658	45.337329	27.967422	10.047945	12.8
std	0.784808	46.122278	21.340178	6.135896	0.481677	11.789017	25.126509	19.671439	5.441298	1.436572	6.4
min	-1.386294	1.000000	0.100000	0.100000	0.000000	0.030000	0.600000	4.000000	12.000000	0.000000	0.0
25%	2.890372	43.500000	9.030000	7.995000	0.400000	8.500000	21.270000	29.000000	24.000000	10.000000	8.0
50%	3.569533	76.000000	17.050000	12.084864	0.618724	13.350000	37.250000	45.000000	28.000000	10.000000	11.0
75%	3.860730	96.500000	22.070000	15.300000	0.690000	15.350000	45.800000	61.000000	31.000000	10.000000	16.0
max	5.732532	559.250000	198.050000	197.970000	9.920000	195.000000	199.920000	100.000000	46.000000	20.000000	37.0

Table 3. Analysis of PM2.5

MODEL NAME	**R^2** Score	**RMSE**	**MAE**
Decision Tree Regression	0.83	11.27	7.43
Random Forest Regression	0.85	10.57	6.42
Gradient Boosting Regression	0.89	8.86	5.34

Table 4. Analysis of PM210

MODEL NAME	**R^2** Score	**RMSE**	**MAE**
Decision Tree Regression	0.82	18.68	12.78
Random Forest Regression	0.85	17.33	11.22
Gradient Boosting Regression	0.88	15.46	9.82

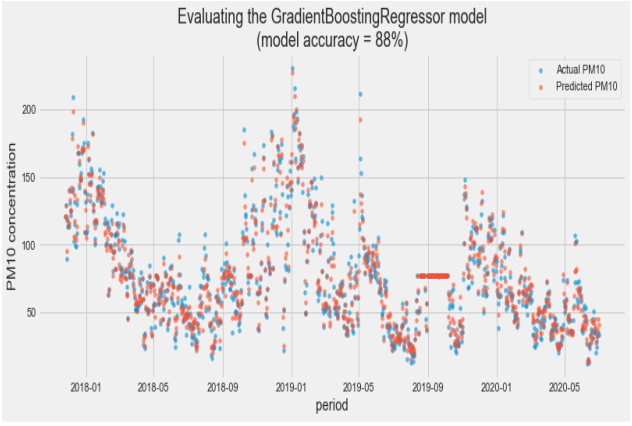

Figure 7. Gradient Boosting Regressor for PM10

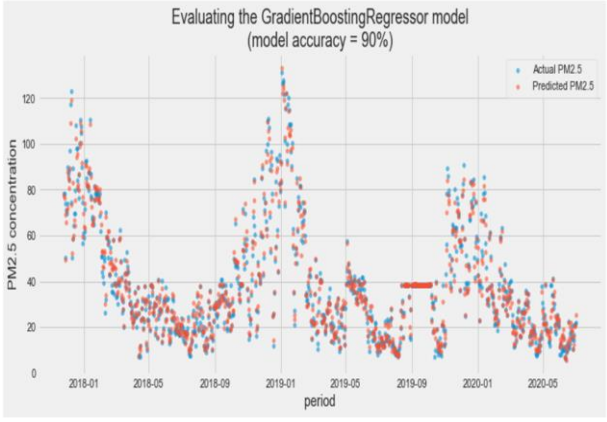

Figure 8. Gradient Boosting Regressor for PM2.5

5. CONCLUSION

The machine learning models provided us with excellent forecasting of air pollution, the research showed us the application of various ML models such as Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression. Among these, the Gradient Boosting Regression model showed the best accuracy predicting the PM2.5 and PM10 in the AP001 station. The study, also showed that the PM particles were in high concentrations during the Winter-Spring Cycle which is the December to

February period, this is shown in figure 7 for PM10 and figure 8 for PM2.5. After all these results, we can state that machine learning can offer us reliable information that the state can use to issue alerts of air pollution incidents and accordingly regulate it to protect the residents from harmful diseases caused by them.

REFERENCES

[1] Doreswamy, ; KS, Harishkumar; KM, Yogesh; Gad, Ibrahim et al (2020) Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models https://sci-hub.hkvisa.net/10.1016/j.procs.2020.04.221

[2] Adil Masood, Kafeel Ahmad et al (2020) A model for particulate matter (PM2.5) prediction for Delhi based on machine learning pproach https://www.sciencedirect.com/science/article/pii/S1877050920307249

[3] Biancofore F, et al. (2017) Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmos Pollut Res 8(4): 652- 659. https://doi.org/10.1016/j.apr.2016.12.014

[4] Qin S et al (2014) Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. Atmosp Environ 98:665–675. https://doi.org/10.1016/j.atmosenv.2014.09.046

[5] Anurag Barthwal, Debopam Acharya, Divya Lohani et al (2021) Prediction and analysis of particulate matter (PM2.5 and PM10) concentrations using machine learning techniques https://link.springer.com/article/10.1007/s12652-021-03051-w

[6] Suleiman A, M.R Quinn et al (2018) Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5) https://www.sciencedirect.com/science/article/abs/pii/S1309104218301272

[7] Nurul Amalin Fatihah Kamarul Zaman, Kasturi Devi kanniah, Dimitris G. Kaskoaoutis and Mohn Talib Latif et al (2021) Evaluation of Machine Learning Models for Estimating PM2.5 Concentrations across Malaysia https://www.mdpi.com/2076-3417/11/16/7326/htm

[8] Limei Ma, Yijun Gao, Chen Zhao et al (2020) Research on Machine Learning Prediction of Air Quality Index Based on SPSS https://ieeexplore.ieee.org/abstract/document/9239825

[9] Fabiana Franceschi, Martha Cobo, Manuel Figueredo et al (2018) Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering ttps://www.sciencedirect.com/science/article/abs/pii/S1309104217305494

[10] Xinzhi Lin et al (2021) The Application of Machine Learning Models in the Prediction of PM2.5/PM10 Concentration https://dl.acm.org/doi/abs/10.1145/3450588.3450605

[11] Chavi Srivastava, Shyamli Singh, Amit Prakash Singh et al (2018) Estimation of Air Pollution in Delhi Using Machine Learning Techniques https://ieeexplore.ieee.org/abstract/document/8675022

Mr. Prateek Parashar is a graduate student of School of Computing, DIT University. Mr. Parashar is having interest in Machine learning, Python Programming.

Mr. Abhijeet Ghldiyal is a graduate student of School of Computing, DIT University. Mr. Ghldiyal is having interest in Data Science, Image processing.

Dr. Atul Kumar Srivastava is an Assistant Professor in Computer Science at DIT University. His primary research interests are in Social network analysis, Data Sciene, and Machine learning. Dr. Srivastava received Masters's and Doctoral degree in Computer Science from the Banaras Hindu University, Varanasi, India during 2011 and 2018, respectively. He has served as a Faculty member at Amity University, Patna, India (2018 to 2019) Lucknow University, India (2019-2020).