IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Estimating Particulate Matter using Machine Learning Technique

Jash J Patel[1], Prof. Neha R Patel[2], Prof. Hetal Gaudani[3]

[1:] M. Tech. Environmental Engineering Student, Department of Civil Engineering, BVM Engineering College, Vallabh Vidyanagar, Gujarat, India.

[2:] Assistant Professor, Department of Civil Engineering, BVM Engineering College, Vallabh Vidyanagar, Gujarat, India.

[3:] Assistant Professor, Department of Computer Engineering, GCET Engineering College, Vallabh Vidhyanagar, Gujarat, India

Abstract: Urbanization and Industrial growth have been identified as the primary sources of air pollution in India's metropolitan cities. One of the most important challenges in urban areas has been highlighted as air pollution. Particulate matter is still one of the most prominent forms of air pollution in cities, with serious health repercussions for both acute and chronic exposures. The presence of particulate matter is mostly determined by meteorological conditions. Citizens and governments around the world have become increasingly concerned about the effects of air pollution on human health, and have proposed sustainable development solutions to address air pollution challenges caused by current industrialization, which include liquid droplets, solid particles, and gas molecules dispersed throughout the atmosphere. The high concentration of particulate matter of sizes PM10 and PM2.5 has a major negative impact on human health. The determination of particulate matter concentration in atmospheric air for the betterment of human being well of primary importance. So, in this paper, a machine learning model were used for predicting particulate matter concentration in atmospheric air is investigated on Vapi Air Quality Monitoring data sets, which were obtained from 2017 to 2021. These models were compared on their performance in particulate matter prediction and the best model was found for predicting particulate matter.

Keywords: **Particulate Matter, Meteorological factors, Random Forest, Decision Tree, and Prediction.**

1. Introduction

One of the most important components of every living entity on the planet is air. Urbanization, industrialization, vehicles, power plants, and chemical activity, as well as other natural activities like as volcanic eruptions, agricultural burning, and wildfires, have all contributed to an increase in pollution during the previous 50 years. All of these activities increase pollution, and particulate matter (PM) is one of the major causes of air pollution (Jung, 2017). Several causes contribute to pollution, including stubble burning and harmful particles such as PM10 and PM2.5 (Zanobetti et al. 2009). These particulate materials are generally made up of solids and liquids floating in the air, with a wide range of chemical compositions, including certain organic molecules, sulphur dioxide, and so on (Davidson, 2005). PM2.5 particles, as their name indicates, are tiny atmospheric particulate matter with a diameter less than 2.5 m", or roughly 3% of the diameter of a human hair. These particles are particularly dangerous to one's health since they may readily enter deep into the lungs, irritate and erode the alveolar wall. As a result of all of this, the lungs are gravely harmed. PM2.5 has various detrimental impacts, including asthma, lung inflammation, and cardiovascular illness, as well as the possibility of cancer (Valavanidis et al. 2006). If these small patches are injected into the lungs, they may be able to overcome COVID-19 infection's inflexibility, as the new coronavirus also targets the respiratory system (Kumar et al. 2020). If these pollutant patches receive a lot of attention in the atmosphere, it has a negative impact on human health and can cause life-threatening difficulties in a short time (Graff, 2007). Particulate pollution has been shown to have an inheritable effect on mortal health in studies.

1.1 Particulate Matter

Particulate air pollution is a combination of solid, liquid, or solid and liquid particles suspended in the air. These scattered particles vary in size, nature, and origin. Aerodynamic properties are important for particle classification because they impact particle transit and removal from the air, as well as particle deposition inside the respiratory system, and they are connected to particle chemical composition and origins. Particulate matter (PM) is a common indoor and outdoor air pollutant with particle sizes ranging from a few nanometres to tens of micrometres. Natural sources, human sources, and atmospheric change all contribute to PM in the ambient air. Penetration from the outside air, cooking, and resuspension from household dust are the primary sources of indoor particulate matter (Pope et al. 2020Indoor air chemistry might be a substantial contribution to indoor PM in particular circumstances. Despite being regulated as a single chemical, PM can contain hundreds of inorganic and organic species. Depending on the source, the size and chemical composition of PM varies substantially. Coarse PM (particles with a diameter of 2.5–10 mm) is mostly produced by mechanical activities such as resuspended road dust, abrasive mechanical operations in industry and agriculture, and certain bioaerosols. $PM_{2.5}$ particles have a size range of 0.1 to 2.5 mm, whereas $PM_{2.5}$ particles have a size range of 10mm. (Z. Fan and colleagues, 2008). Particulate matter is the most common contaminant in today's air, and its concentration is a key criterion for assessing air quality. Particulate matter not only lowers vision and sunlight in the atmosphere, but it also generates more hazy days and is harmful to human health. As a result, precise prediction of Particulate Matter Concentration can support the management in understanding the present state and trend of air quality, allowing management to develop appropriate preventative or control measures.

1.2 Machine Learning

Machine learning is the study of computer algorithms that can learn and grow based on their own experience and data. Artificial intelligence is considered to be involved. Without the need for explicit programming, machine learning algorithms generate a model from training data and use it to make predictions or judgements. The following are the stages of analysing a machine learning model: Data Understanding- Before developing the different ways of dealing with data, it is required to first examine the original data. Merging data, imputing missing values, eliminating variables with too many missing values, sorting data, and so on are all examples of data preparation. The act of putting models through their paces and assessing the outcomes is known as model training. The outcomes evaluation is a crucial stage in understanding the findings since it allows the models and the initial research approach to be changed. Furthermore, explaining and displaying the several models that can be used: In supervised learning, a collection of independent parameters is matched to one or more dependent variables. Regression and classification are two examples of these sorts of problems. Unsupervised learning, on the other hand, does not require any prior "correct" data, and its goal is to discover the data's underlying patterns. Optimization techniques are ways for determining the optimal set of parameters for minimising a cost function. (Wilcox and colleagues, 2013).

1.2.1 Decision Tree

A decision tree is a graph that represents choices and their outcomes as a tree. The graph's nodes represent events or choices, while the graph's edges reflect decision rules or conditions. Nodes and branches constitute each tree. Each branch indicates a value that the node can take, and each node represents attributes in a group that needs to be categorized. Decision Tree is a supervised learning approach that may be used for classification and regression issues; however, it is most commonly employed to solve classification problems. Internal nodes contain dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Leaf nodes are the result of those decisions and do not include any more branches, whereas Choice nodes are used to make any decision and have several branches. The judgments or tests are based on the characteristics of the provided dataset. It's a graphical depiction for obtaining all feasible answers to a problem/decision depending on certain parameters. It's termed a decision tree because, like a tree, it starts with a root node and grows into a tree-like structure with additional branches. We utilize the cart algorithm, which stands for Classification and Regression Tree algorithm, to form a tree. A decision tree simply asks a question and divides the tree into subtrees based on the answer (Yes/No).

1.2.2 Random Forest

Random Forest Regression is a regression supervised learning method that use the ensemble learning method. To produce a more accurate forecast than a single model, the ensemble learning technique integrates predictions from numerous machine learning algorithms. The Random Forest Regression model is both strong and exact. It works effectively in a variety of circumstances, including those with non-linear interactions. There are several disadvantages, however: there is no interpretability, overfitting is possible, and we must pick the number of trees to include in the model. These two methods are employed. 1. Boosting method: The word "boosting" refers to a set of algorithms that assist poor learners in improving their performance. In ensemble learning, boosting is a bias and variance reduction approach. A weak learner is defined as a classifier that is arbitrarily well-correlated with the actual classification, whereas a strong learner is defined as a classifier that is arbitrarily well-correlated with the real classification. 2. Bagging technique: Bagging or bootstrap aggregating is used to increase the accuracy and stability of a machine learning system. Bagging also helps to minimise overfitting by reducing variance.

1.2.3 Gradient Boosting

Gradient Boosting is a sequential ensemble learning approach in which the model's performance increases with time. The model is built in stages using this manner. It infers the model by allowing an absolute differentiable loss function to be optimized. A new model is constructed as each weak learner is added, giving a more exact assessment of the response variable. To work, the gradient boosting technique requires the following components: 1. Loss function: We must optimize the loss function to decrease prediction errors. In contrast to AdaBoost, the wrong result in gradient boosting is not given a larger weighting. By averaging the outputs from weak learners, it attempts to lower the loss function. 2. Weak learner: We need weak learners to generate predictions in gradient boosting. We employ regression trees to retrieve true numbers as output. We generate trees in a greedy approach to find the best split point; as a result, the model overfits the dataset. 3. Additive model: We aim to minimize the loss in gradient boosting by adding decision trees. We can also reduce the mistake rate by reducing the parameters. As a result, we construct the model in this situation so that adding a tree does not modify the existing tree.

1.2.4 K Nearest Neighbour

K-Nearest Neighbour is a Supervised Learning-based Machine Learning algorithm that is one of the most basic. The K-NN method assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories. The K-NN method maintains all of the available data and classifies a new data point based on its resemblance to the existing cheval. This implies that fresh data may be quickly sorted into a suitable category using the K-NN method. The K-NN algorithm may be used for both regression and classification, however, it is more commonly utilized for classification tasks. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the data. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and then takes an action on it during classification. The KNN method simply saves the information during the training phase, and when it receives new data, it classifies it into a category that is quite similar to the new data.

1.3 Aim

- To find the best machine learning prediction model like Random Forest, Linear Regression, Decision Tree and Support Vector Machines.

1.4 Objectives

- To predict Particulate Matter by machine learning models in the Vapi region.
- To determine the contribution of the explanatory variable (Wind speed, Wind direction, Temperature, and Relative humidity in the development of the prediction model for PM2.5 and PM10.
- To select the best Machine Learning Prediction model.

1.5 Scope

- In this study the PM2.5 and PM10 are predicted with the meteorological parameters (Wind speed, Wind direction, Temperature, and Relative humidity by using the various machine learning models like Random Forest, Linear Regression, Decision Tree and K-Nearest Neighbour. The Vapi region is selected to for this study. The secondary data were collected from the Vapi GPCB office and from this data the particulate matter was predicted and to get the predicted values as nearer to the measured valu e so, from the data the null values were removed and after that the data were trained for the prediction and get the prediction as nearer to the measured values and in that the K-Nearest Neighbour predicted as nearer to the measured values as compared to other models.

2. Methodology

2.1 Study area

The site selected as per the daily 24-hour average data availability is the Continuous Ambient Air Quality Monitoring Station installed at Vapi in the state of Gujarat, India.

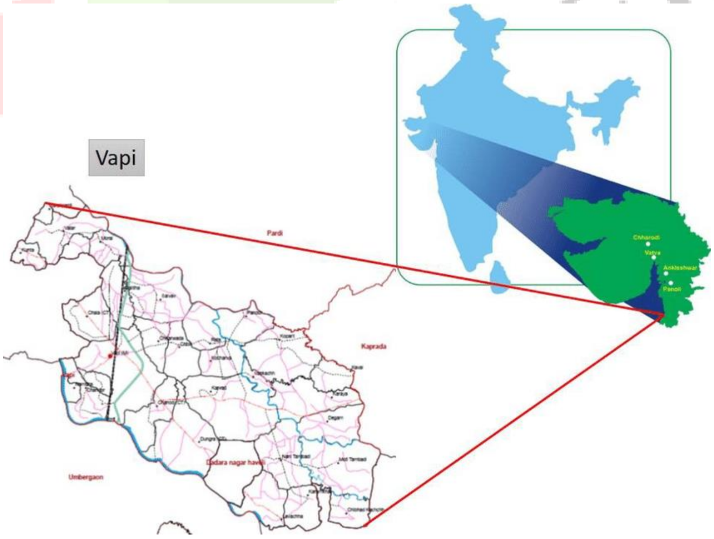


Figure 1: Location of study area.

2.2 Data Observation

In the present study, the secondary data set was obtained from the GPCB monitoring station at Vapi, Gujarat for a period of five years (2017 - 2021). The daily concentration data of pollutants, such as PM2.5 (particulate matter with an aerodynamic diameter less than 2.5μm), PM10 (particulate matter with an aerodynamic diameter less than 10 μm), and daily meteorological data including wind speed, wind direction, temperature, relative humidity and for the same period have been utilized as primary variables to develop the Machine Learning models. To avoid any bias with a particular missing value in the dataset that may affect the Predicting model results, the null values were removed.

2.3 Data Pre-Processing

In the pre-processing step, data gaps in Air Quality data were replaced with null values. And afterwards the null values were removed from the data so that the prediction for the particulate matter can be reached as nearer to the measured values. Moreover, the general statistics of the experimental observations considered in this work from 2017 to 2021 has been shown in Table 1.

Table 1. General statistics of the experimental data considered in this study

Parameters	Unit	Mean	Standard Deviation	Minimum	Maximum
PM2.5	µg/m^3	61.9	52.2	0	408.7
PM10	µg/m^3	120.6	77.8	2.8	712.1
Wind Speed (WS)	m/s^1	0.54	0.35	0.1	0.6
Wind Direction (WD)	degrees	164.7	131.1	0.1	9.7
Temperature	°C	26.2	4.9	18	26.7
Relative Humidity (RH)	%	62.9	15.6	25	64

3. Data Visualization

I. PM10 and PM2.5 parameters secondary data collected for the duration 01-01-2017 to 31-12-2021 for Vapi station.

As, it can be seen from the figure that at the time between Monsoon season there is a drop in the Particulate matter concentration due to the rain as after the monsoon season there is a rise of Particulate matter due to the winter season and Diwali festival where firecrackers are burn so there is an increase of pollution at that period.

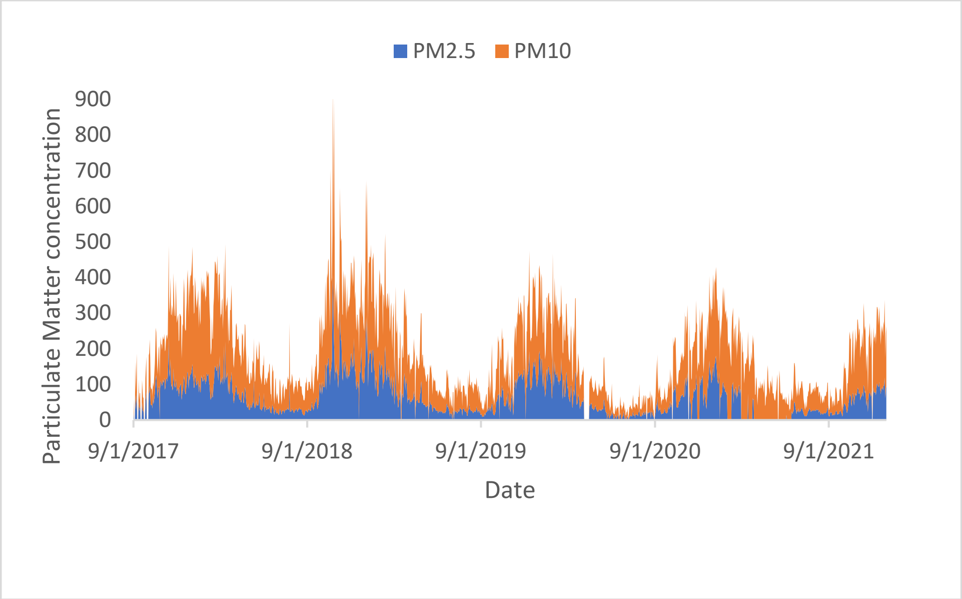
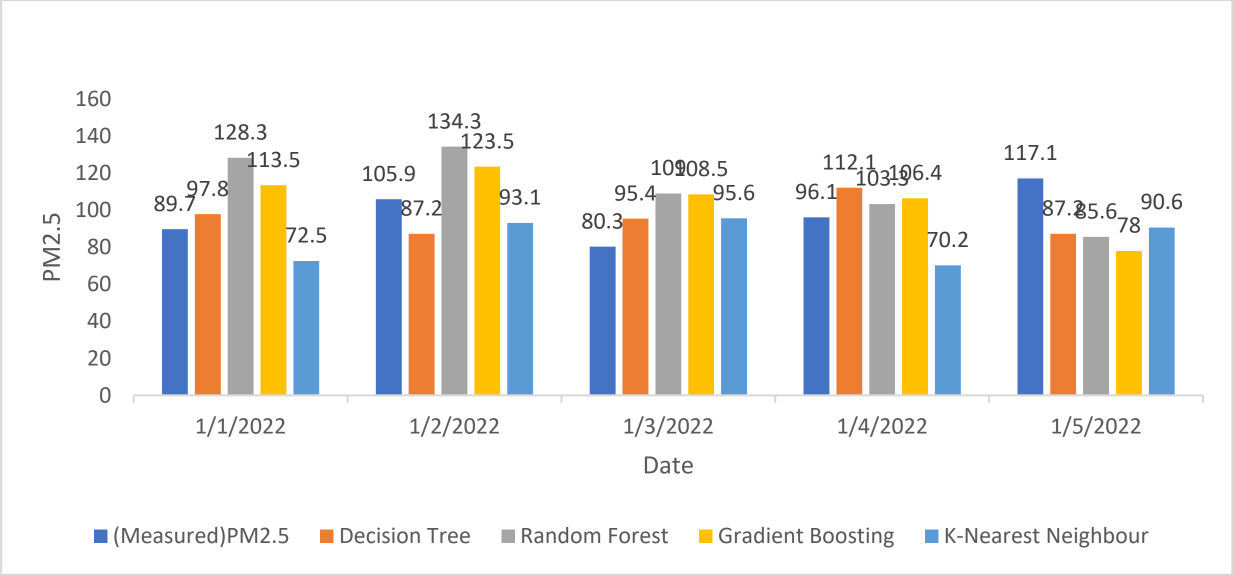


Figure 1 Overtime representation of PM10 and PM2.5 parameters from 2017-2021

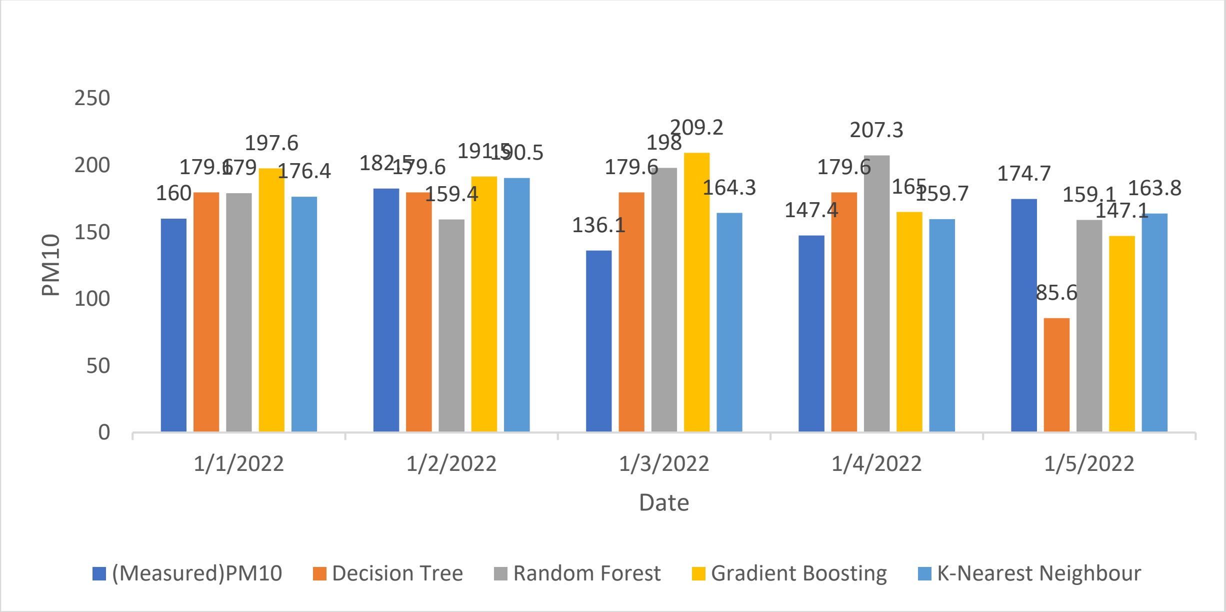
I. Prediction Models for PM2.5 from 2017-2021 for 70:30 ratio.

As from the figure it can be said that when the models were trained for the 4 years data with 70:30 ratio for PM2.5 it can be seen that prediction values when compared with the measured values were not nearer. K-Nearest Neighbour predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.



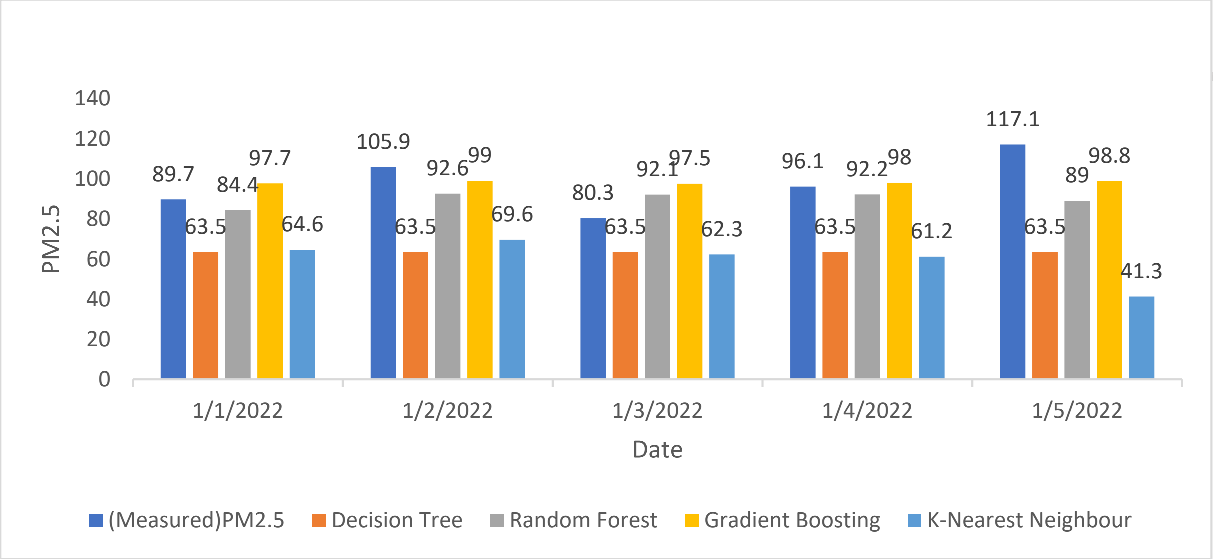
II. Prediction Models for PM10 from 2017-2021 for 70:30 ratio.

As from the figure it can be said that when the models were trained for the 4 years data with 70:30 ratio for PM10 it can be seen that prediction values when compared with the measured values were not nearer. K-Nearest Neighbour predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.



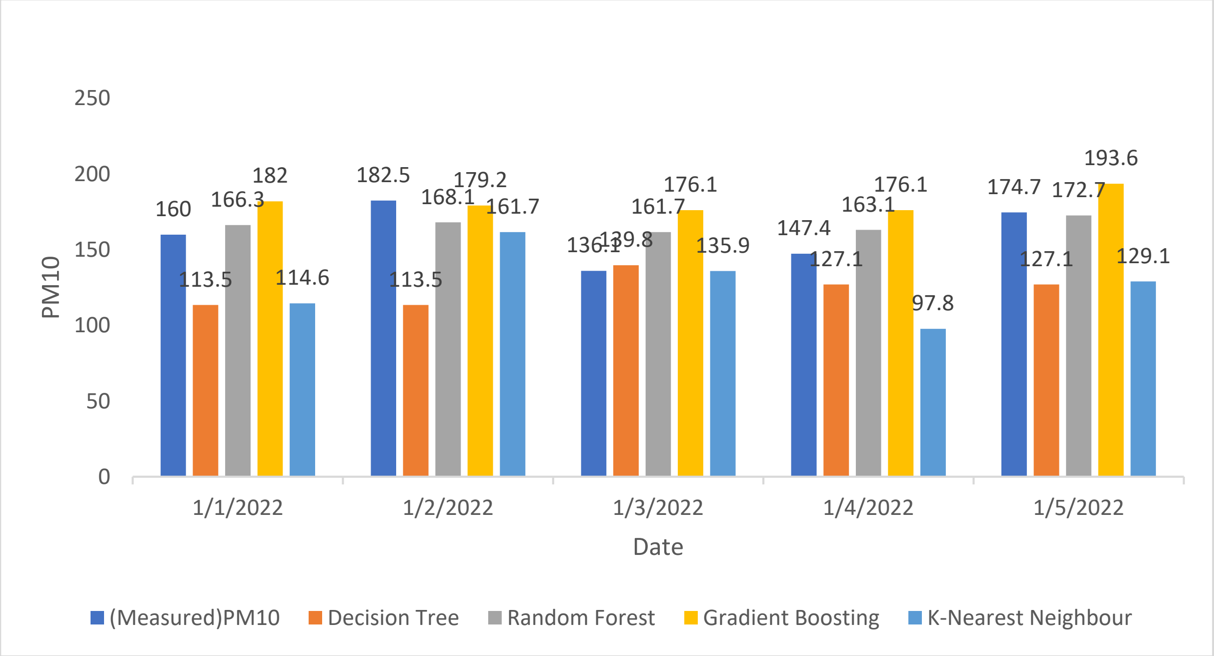
III. Prediction Models for PM2.5 from 2021 for 70:30 ratio.

As from the figure it can be said that when the models were trained for the one-year data with 70:30 ratio for PM10 it can be seen that prediction values when compared with the measured values were not nearer. Random Forest predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.



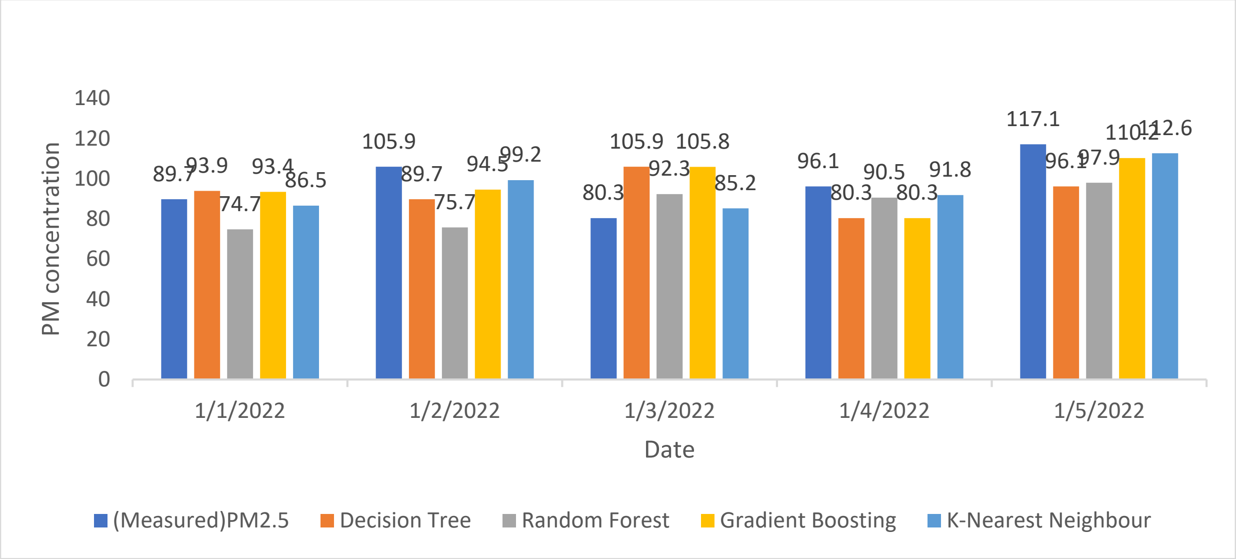
IV. Prediction Models for PM10 from 2021 for 70:30 ratio.

As from the figure it can be said that when the models were trained for the one-year data with 70:30 ratio for PM10 it can be seen that prediction values when compared with the measured values were not nearer. Random Forest predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.



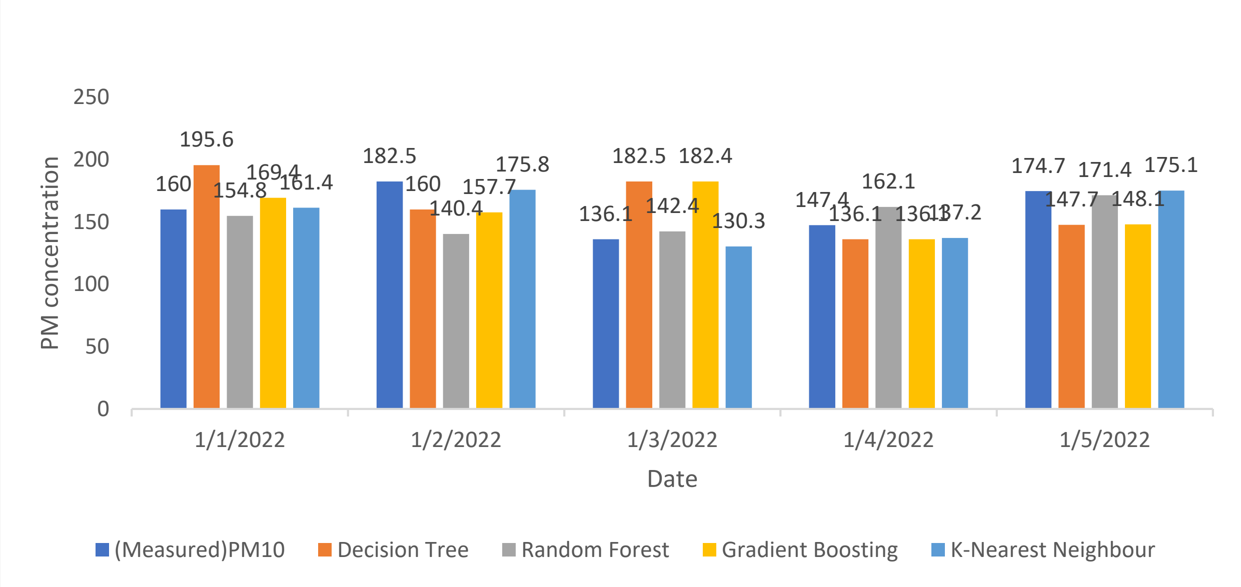
V. Prediction Models for PM2.5 from Last four days.

As from the figure it can be said that when the models were trained for the past four days for PM2.5 it can be seen that predicted values when compared with the measured values they were nearer to the measured value. K-Nearest Neighbour predicted the values were nearer to the measured value as per required.



VI. Prediction Models for PM10 from Last four days.

As from the figure it can be said that when the models were trained for the past four days for PM2.5 it can be seen that predicted values when compared with the measured values they were nearer to the measured value. K-Nearest Neighbour predicted the values were nearer to the measured value as per required.



4. Results

70% of the data was used for training and the remaining 30% of the data was used as testing data. So, as a results when the models were trained. Results of comparison of predicted values with measured values are shown in table.

Table 2: Comparison of predicted values using four years data for 80:20 ratio.

The duration of 1st Jan 2017-31st Dec 2021 for PM2.5 with 80:20 ratio. It can be seen that K-Nearest Neighbour predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.

Date	(Measured) PM2.5	Decision Tree	Random Forest	Gradient Boosting	K-Nearest Neighbour
01-01-2022	89.7	79.2	124.4	114.9	77.5
02-01-2022	105.9	79.2	130.5	115.3	88.8
03-01-2022	80.3	85.2	109.2	109.6	89.5
04-01-2022	96.1	103.2	103.9	108.3	90.5
05-01-2022	117.1	27.7	87.5	77.9	109.6

The duration of 1st Jan 2017-31st Dec 2021 for PM10 with 80:20 ratio. It can be seen that K-Nearest Neighbour predicted the values nearer to the measured value but was not as nearer as it should be to the measured values.

Date	(Measured) PM10	Decision Tree	Random Forest	Gradient Boosting	K-Nearest Neighbour
01-01-2022	160	183.1	191.9	197.8	148.1
02-01-2022	182.5	183.1	193.3	190.3	139.5
03-01-2022	136.1	183.1	186.6	202.8	190.5
04-01-2022	147.4	183.1	182.5	196.8	194.2
05-01-2022	174.7	154	158	114.6	122.3

Table 6: Comparison of predicted values using past four days data for Particulate matter

Past four days of the data was used as testing data. So, as a results when the models were trained. Results of comparison of predicted values with measured values are shown in table

It can be seen that K-Nearest Neighbour the values nearer to the measured value as compared to the other models was predicted the vales as nearer to the measured values for PM2.5.

Date	(Measured)PM2.5	Decision Tree	Random Forest	Gradient Boosting	K-Nearest Neighbour
01-01-2022	89.7	93.9	74.7	93.4	86.5
02-01-2022	105.9	89.7	75.7	94.5	99.2
03-01-2022	80.3	105.9	92.3	105.8	85.2
04-01-2022	96.1	80.3	90.5	80.3	91.8
05-01-2022	117.1	96.1	97.9	110.2	112.6

It can be seen that K-Nearest Neighbour the values nearer to the measured value as compared to the other models was predicted the vales as nearer to the measured values for PM10.

Date	(Measured)PM10	Decision Tree	Random Forest	Gradient Boosting	K-Nearest Neighbour
01-01-2022	160	195.6	154.8	169.4	161.4
02-01-2022	182.5	160	140.4	157.7	175.8
03-01-2022	136.1	182.5	142.4	182.4	130.3
04-01-2022	147.4	136.1	162.1	136.1	137.2
05-01-2022	174.7	147.7	171.4	148.1	175.1

5. Conclusion

In this data analysis phase as it can been seen from the results when the prediction of PM10 and PM2.5 were carried out and in that when the comparison of predicted values with the measured values was carried out so, it can be easily seen that the K-Nearest Neighbour model has got to the nearest value to the measured values as compared with the other models like Decision Tree, Random Forest and Gradient Boosting. So, for this Study K-Nearest Neighbour was best prediction model.

6. References

1. A. Masood, K. Ahmad, “A model for particulate matter PM2.5 prediction for Delhi based on machine learning approaches”, Elsevier, 2020.
2. B. Pan, “Application of XGBoost algorithm in hourly PM2.5 concentration prediction”, IEEE 2017.
3. C. Feng, W. Wang, Y. Tian, X. Que, and X. Gong, “Estimate Air Quality Based on Mobile Crowd Sensing and Big Data”, IEEE,2017.
4. D Petelin, A Grancharova, J Kocijan, “Evolving Gaussian process models for prediction of ozone concentration in the air”. Elsevier, 2013.
5. J. Ma, Z. Yu, Y. Qu, J. Xu, Y. Cao, “Application of the XGBoost machine learning method in PM2.5 prediction: a case study of Shanghai”, Aerosol and Air Quality Research. 2020.
6. J.K. Deters, R. Zalakeviciute, M. Gonzalez, Y. Rybarczyk, “Modeling PM2.5 Urban pollution using machine learning and selected meteorological parameters”, Journal of Electrical and Computer Engineering. 2017.
7. Jung, C Ren, B Hwang, and W Chen. “Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2.5 concentrations in Taiwan from 2005 to 2015.” Elsevier, 2017.
8. K. B. Shaban, A. Kadri, and E. Rezk,” Urban Air Pollution Monitoring System with Forecasting Models”, IEEE,2016.

9. K. Hu, V. Sivaraman, H. Bhrugubanda, S. Kang, A. Rahman, "SVR Based Dense Air Pollution Estimation Model Using Static and Wireless Sensor Network", IEEE, 2016.
10. KS Harishkumar, KM Yogesh, I. Gad "Forecasting air pollution particulate matter PM2.5 using machine learning regression model", Elsevier, 2019.
11. Kumar A, Kamal, D, Maji, Jyoti, Deshpande, Ashok. "Disability-adjusted life years and economic cost assessment of the health effects related to $PM_{2.5}$ and PM10 pollution in Mumbai and Delhi, in India from 1991 to 2015." Springer, 2017
12. Kumar S, Mishra S, Singh S, "A machine learning-based model to estimate $PM_{2.5}$concentration levels in Delhi's atmosphere". Journal of Heliyon, 2020.
13. M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, S. Talebiesfandarani, "PM 2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data", atmosphere, Multidisciplinary Digital Publishing Institute, 2019.
14. Nevin, G., Zur, G. "The regional prediction model of PM_{10} concentrations for Turkey" Journal of Physics: Conference Series, 2016.
15. O. Kisi, K. S. Parmar, K. Soni, and V. Demir, "Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, andM5 model tree models", springer,2017
16. P. Wang, H. Zhang, Z. Qin, G. Zhang, "A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting", Atmospheric Pollution Research 2017.
17. Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, MB. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Y. Wang, LJ. Mickley, J. Schwartz, "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution", Environment International. ELSEVIER, 2019.
18. R. Waman Gore and D. S. Deshpande," An Approach for Classification of Health Risks Based on Air Quality Levels", IEEE, 2017.
19. Rubal, D Kumar, "Evolving Differential evolution method with random forest for prediction of Air Pollution". Elsevier, 2018.
20. S Yarragunta, M Nabi, Jeyanthi, Revathy, "Prediction of Air Pollutants Using Supervised Machine Learning". IEEE, 2021.