# Machine Learning Based Air Quality Prediction:

# New Delhi Okhla Phase II Area Case Study

##### **Sahil Negi**

Computer Science & Engineering Department

Galgotias University

***Abstract*-** Air pollution has been one of the most important problems in human evolution over the last century because of its detrimental impact on the ecosystem of humans. The research presented in this article focused on using machine learning to predict behavior related to air quality.

Central Pollution Control Board: CPCB is a source from where data is collected to validate the model. In this paper, the Okhla Phase II area in New Delhi was selected as the study target, which suffers from severe air pollution. Based on CPCB data on major air pollutants from February 2018 to November 2022. The paper initially included monthly air quality assessments. The findings show that air quality in the Okhla Phase II region generally follows the same trend with respect to historical assessments during the study period. According to this study, a significant proportion of Okhla phase II air pollution is attributed to the pollutants PM2.5, PM10, NO2, and CO. Therefore, the study was conducted to determine air quality based on Particulate Matter 2.5, Particulate Matter 10, Nitrogen Dioxide, and Carbon Monoxide pollution concentrations in Okhla phase II. Air quality prediction used data based on CPCB from February 2018 to November 2022 for key air pollutants and machine learning techniques were used. These methods include linear regression, decision tree regressor, random forest regressor, support vector regressors, and the K Nearest Neighbour method for air quality prediction. The Study found that Random Forest Regressor was the most reliable algorithm for predicting air pollution, with a result of 99.3%.

***Index Terms***- Air Pollution; Prediction; Machine Learning; Linear regression; DTR; SVR; KNN; New Delhi Okhla Phase II Area Case Study;

1. Introduction

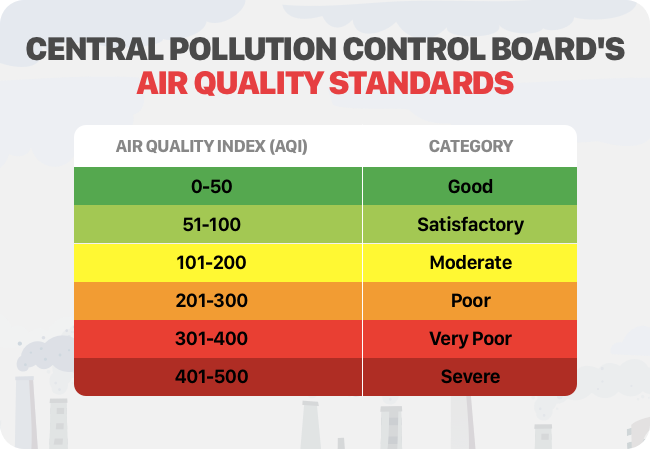
A

**ir** pollution consists of one or more pollutants present in the air which can cause **serious** environmentalharm to **people, plants,** or **animals, or interacting** unexpectedly with property in developed and developing countries.

Air pollution distribution is a complex process that depends on many factors. Indeed, predicting air pollution that exhibits nonlinear dynamics is a challenging task, it requires accurate knowledge of pollutants mixed with air in the atmosphere, which is prohibitively expensive. In some cases, air pollution in megacities has exceeded standards, raising concerns. Pollution, the most critical issue of the last decade, it is a constant debate in the media, government conferences, and environmental activism. In recent decades, controlling air pollution seemed impossible. Machine learning (ML) technology can be used to predict future increases in air pollution in urban areas. Air pollution monitoring implies the existence of scales of air quality defined in table 1. AQI is measured by airborne droplets and solid particles, primarily composed of Nitrogen dioxide (NO2), Ozone (O3), Sulfur Dioxide (SO2), and Carbon Monoxide (CO). Particulate matter (PM2.5 and PM10) is the main contaminant used for AQI calculations.with air quality level classes ranging from 0 to 500 (good, fair, moderate, poor, extremely poor, and severity). Each AQI category has particular implications. Different air pollutants have a different impacts. If an area has the highest levels of pollutant particles, that means fossil fuels are burning there, for example, the elevated levels of nitrogen dioxide may be due to heavy traffic in areas. The Air Quality Index (API) or Pollutant Standard Index (PSI), gives a complete picture of each air pollutant quality. For example, PM2.5 is determined by adjusting the arithmetic mean of hourly values ​​recorded over the last 24 hours. PM2.5 Index has concentrations measured in µg/m3. PM2.5 pollutants include particles produced by combustion or secondary particle formation and compression. PM10 particles include particles less than 10 microns in diameter that can penetrate the first line of defense (nose and throat) and cause damage and deposition in the lungs. According to Studies exposure to polluted air is directly connected with your health which can cause cardiovascular and respiratory disease. According to the World Health Organization if the average concentration of PM2.5 get reduced from 55.8μg/m3 to 40μg/m3annually and the WHO standard (5μg/m3) respectively, then the deaths related to air pollution will be reduced by 15% and Indians stand to gain 1.6 and 5 years of life. In fact, exposure to high concentrations of particulate matter is associated with increased daily and annual mortality, and reducing concentrations of these pollutants without changing other factors reduces the associated mortality. These particles are small and pose a high hazard to human health. This study uses PM2.5, PM10, NO2, and CO as pollutants to predict air pollution.

Details of the AQI concentrations are shown in Table 1.

Table 1Specifies the AQI index.



Source: Central Pollution Control Board (CPCB)

**Sources of pollution**

We know that there are 5 key sources of pollution:

a. Vehicles – There is an increasing number of highly polluting vehicles such as trucks and diesel vehicles, and vehicles that negate the impact of clean fuel and emission technologies.

b. Combustion in power plants and industries using dirty fuels such as petroleum coke, FO (and its variants), coal, biomass

c. Waste incineration, landfill and other collection, treatment.

d. Dust management on roads, construction sites, etc. which causes fine dust pollution.

e. Farmers burn crop residues as they have no choice but to use straws.

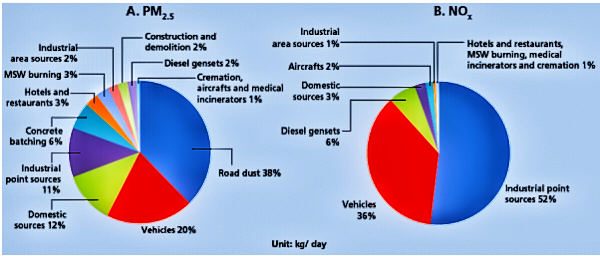


Fig 1:Emissions inventory - PM 2.5 and NOx

Source: 2016, Comprehensive Study on Air Pollution and Green House Gases(GHGs) in Delhi, IIT Kanpur

One of the biggest issues affecting individuals in metropolitan areas is air pollution. The issue is caused by a huge number of motor vehicles, industrial output emissions, and the combustion of petroleum products for power generation and transportation.

Over the past decades, two general approaches of have been used to predict air pollution: deterministic and probabilistic. One of the deterministic techniques created in diverse places to analyze and monitor air pollution is diffusion modeling. These models' results are influenced by their input data, and their use requires access to data on the distribution and diffusion of pollutants in the atmosphere Therefore, it is sufficient to use these models. Since the data collecting necessary for diffusion models is challenging and impractical on a broad scale, researchers have moved to more effective techniques like statistical modeling. Statistical approaches have more uses for forecasting air pollution than deterministic methods do.

1. Machine learning

Computers can now learn without being explicitly programmed thanks to the discipline of machine learning.One of the most intriguing technologies ever is machine learning. As the name suggests, it gives computers what makes them more human: the ability to learn. Contrary to popular belief, there are now more applications for machine learning.Deep learning may be divided into four primary categories: reinforcement learning, semi-supervised learning, unsupervised learning, and supervised learning.

**Supervised learning:** Learning which contains labeled dataset this learning already knows the output, regression and classification are its type. The method that builds a mathematical representation of a dataset with both intended inputs and outputs. A collection of training samples make up the so-called training data. Each training example contains a desired output, also known as a monitor signal, and one or more inputs. The algorithm can accurately estimate outputs for inputs that are not included in the training data when it performs at its best. Algorithms are considered to have learnt to do a job when they gradually increase the accuracy of their outputs or predictions.

**Unsupervised learning:** An algorithm that learn from unlabeled, unclassified, or uncategorized test data. Unsupervised learning algorithm detect patterns hidden in the data and for each new data it respond depending on such patterns. Unsupervised learning has several uses, including determining probability density functions in the field of density estimation in statistics. However, unsupervised learning also encompasses other areas that summarize and describe properties of data. Although many machine learning researchers have shown that mixing unlabeled data with sparse quantities of labeled data considerably enhances learning accuracy, even when some training examples lack training labels.

**Semi-supervised learning:** The learning which lies between Supervised and Unsupervised learning is Semi Supervised Learning. Although many machine learning researchers have shown that mixing unlabeled data with sparse quantities of labeled data considerably enhances learning accuracy, even when some training examples lack training labels.

**Reinforcement learning:** the area of machine learning that examines how software agents need to act in various contexts to maximise the idea of cumulative reward. In machine learning, the environment is usually represented as a Markov Decision Process (MDP). Dynamic programming is a common component in reinforcement learning techniques. When a precise model of the MDP cannot be established, reinforcement learning techniques are applied. These algorithms do not need knowledge of the exact mathematical model of the MDP. Reinforcement learning algorithms are used when learning to play games against human opponents and self-driving cars.

1. Machine Learning Role In Air Quality Prediction

## In the study of air quality prediction, many models and techniques have been used those are Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, K Nearest Neighbor, and Support Vector Classifier. A brief description of each model and its studies is provided.

## **Linear Regression**

One of the simplest, and Most used Machine Learning Algorithm.

This method is a statistical technique used for predictive analytics. This method is used for data with continuous or numerical variables lets understand this algorithm with example, predicting employee Salary with respect to his Years of Experience. Linear regression produces predictions on the basis of linear connection between a dependent (y) variable and one or more independent (x) variables. Hence it is called linear regression. It shows how X and Y are related. Linear Regression looks at how X and Y values ​are changing.

Linear regression model provides a straight line that represents the relationship between variables.

Linear regression algorithm types:

**Simple Linear Regression**

When just one X is used to predict the value of Y, where X is independent variable and Y is dependent variable.

Example Years of Experience use to predict Salary of Employee.

**Multiple Linear Regression**

When more than one X is used to predict the value of Y, where X is independent variable and Y is dependent variable.

Example Engine size and number of cylinders use to predict Emission of Carbon Dioxide

## **Decision Tree Regression:**

Decision Trees are a sort of supervised machine learning algorithm using true or false responses to a given question, that are trained using AutoML tools to classify or regress data. The resulting structure looks like a tree with different types of nodes like root node, inner node, and leaf node. The decision tree begins at the root node and branches out into internal and leaf nodes. The final categories or actual values are found in the leaf nodes.

**Random Forest Regression:**

Random Forest is a well-known machine learning algorithm that uses supervised learning methods. Both classification and regression issues may be solved with it. It is built on the idea of ensemble learning, which mixes many classifiers to address complicated issues and enhance model functionality.

**Support Vector Regression:**

Support vector algorithm is used for both Classification and Regression. Where Classification is used for Categorical data and Regression is used to define relationship between X and Y variable. Support Vector Regression and SVM have same principle. The goal of Support Vector machine is to find maximum marginal plane between two different data points. The hyperplane is the best line fitting in two different data points, The Data Point are the support vectors. Example Detecting face in image.

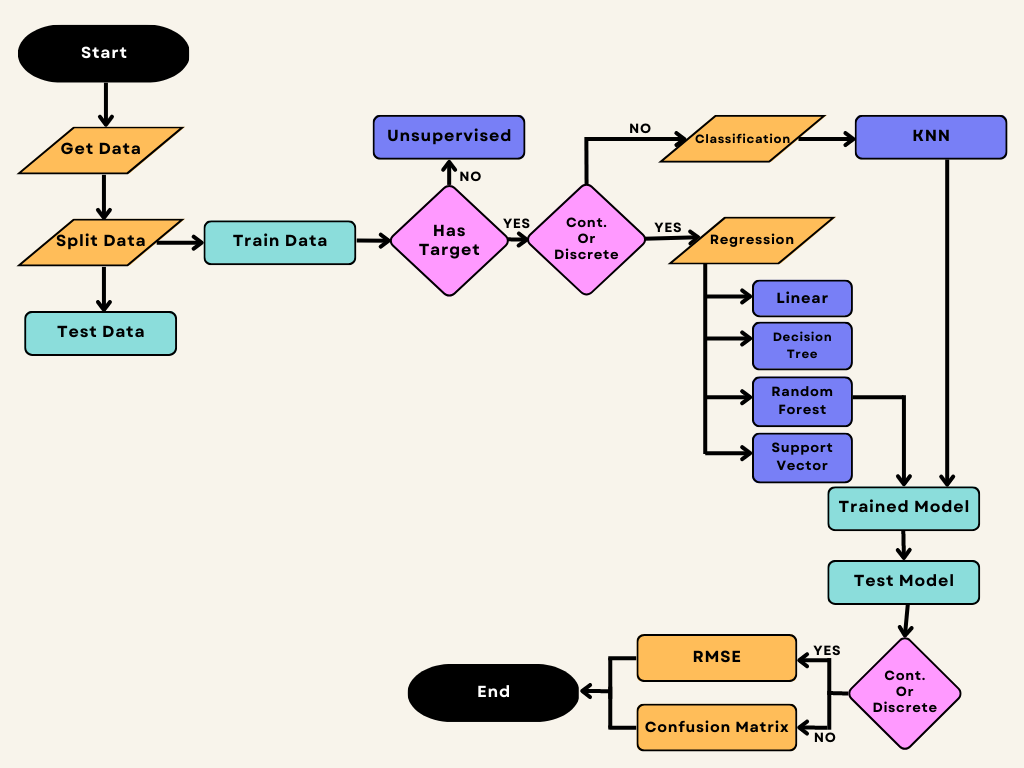
# KNN (K Nearest Neighbor):

KNN is K Nearest Neighbor, is a nonparametric supervised learning classifier that use proximity to classify or predict groups of data points. It is a classification technique that finds the close related points.

# Support Vector Classifier:

The supervised machine learning technique known as the SVC, or Support Vector Classifier, is often used for classification problems. SVC separates the data into two classes by mapping the data points into a high-dimensional space and locating the ideal hyperplane.

1. Materials and Methods



1. Materials and Methods

Material used in this study is data which is taken from CPCB Indian Government website for Okhla Phase II stations in New Delhi

You can Download it from Here - <https://github.com/sahilnegi1/Aqi_data.git> with name [AQI\_2018\_2022.xlsx](https://github.com/sahilnegi1/Aqi_data/blob/main/AQI_2018_2022.xlsx)

Methods used in this study are linear regression, decision tree regressor, random forest regressor, support vector regressor, and the K Nearest Neighbour method for air quality prediction which is programmed in Python Language. In this study, different machine learning algorithms are used to predict air quality. The predicted Air quality Value should act as an incentive to make necessary changes to purify the air, and encourage the population to help reduce air pollution like switching off your vehicle engine at traffic signal.

Formula for Calculating AQI over PM10, PM2.5, NO2, CO

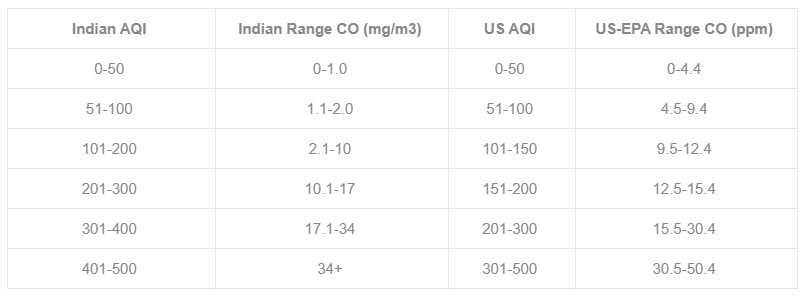
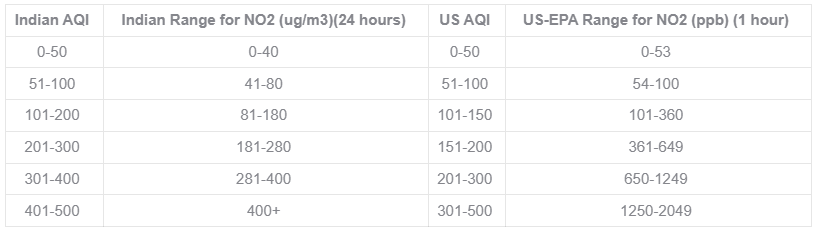
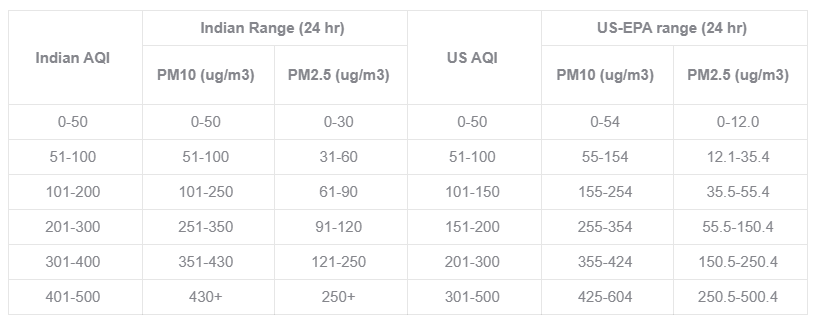


Image Credit : Prana Air

The first step is to collect the data once you collected then prepare it in a format for desired output like our case study data in table 2 which consists of 1756 entries from February 2018 to November 2022. The data is divided into five columns: "Date", "PM2.5", "PM10", "NO2" and "CO".

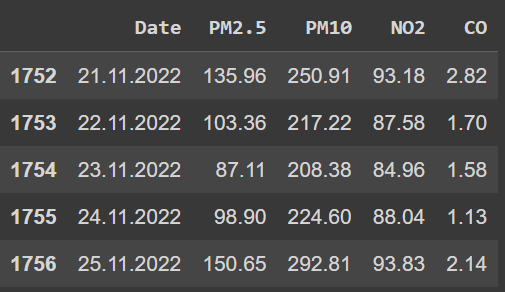


Table 2

Next, calculating Air quality value with respect to NO2, PM2.5, PM10, CO from table 2.

Table 3 Specifies the Air Quality value with respect to NO2, PM2.5, PM10, CO

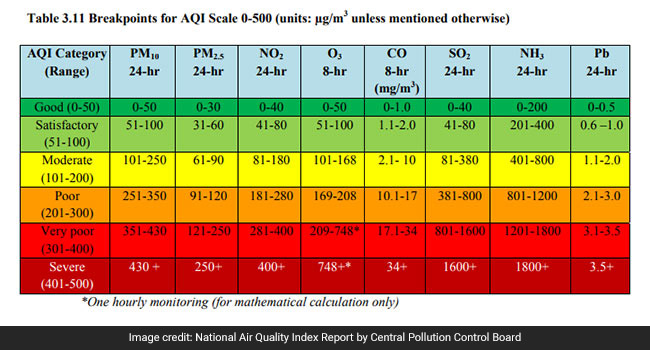


Table 3

Using table 3 Air Quality value with respect to NO2, PM2.5, PM10, CO was calculated and table 4 was created

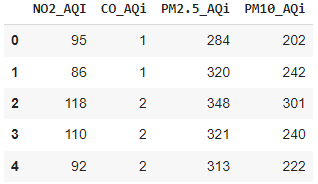


Table 4 Air quality value for different pollutants

AQI value column is created using formula Max(NO2, PM2.5, PM10, CO) from air quality value, AQi in table 5.

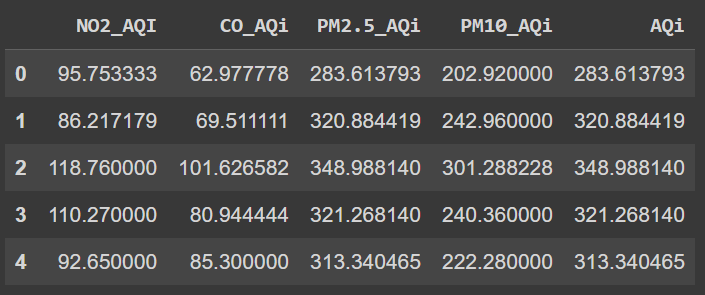
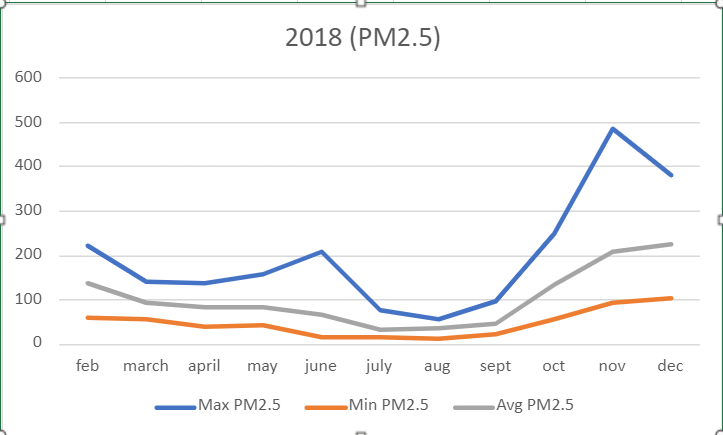


Table 5 Calculated Air quality value.

After that, the data were divided into 80% of training data and 20% for testing data.

Observing Monthly PM2.5 pollutant Conc. FY 2018,2019,2020,2021,2022.



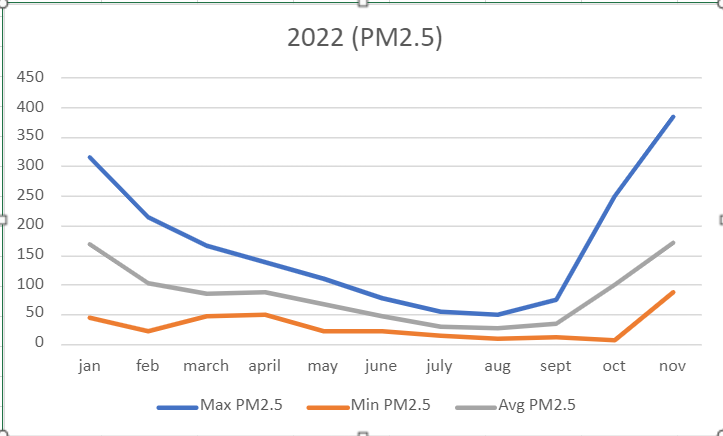
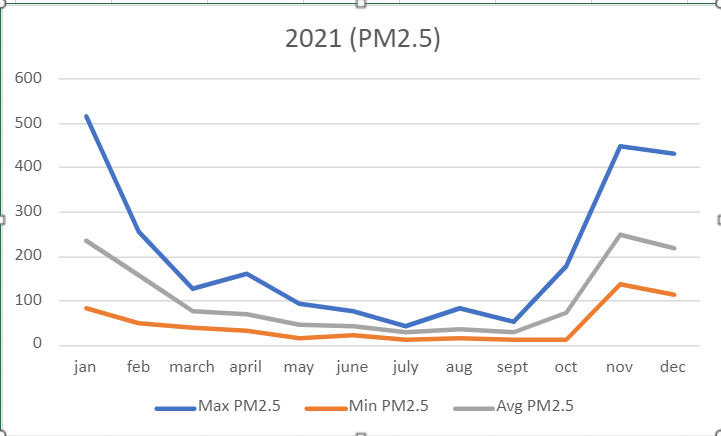
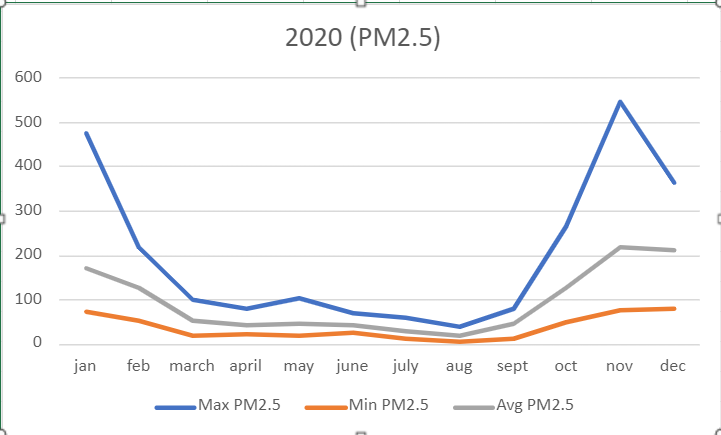
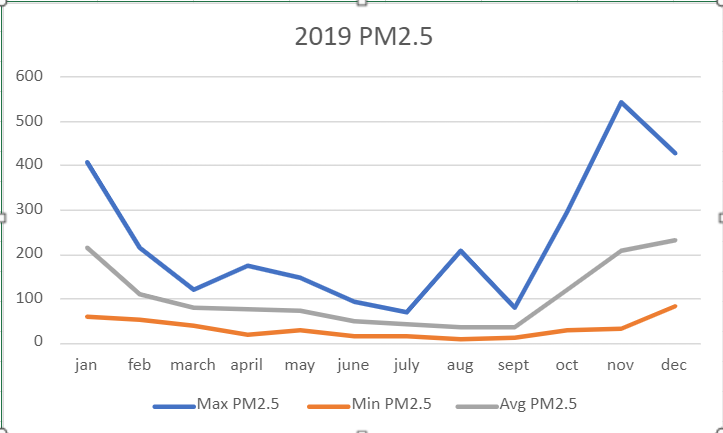


Fig 2: PM2.5 Conc. FY 2018,2019,2020,2021,2022

According to Figure 2, The months of June through September saw the lowest monthly amounts of PM2.5 pollution due to a decline in traffic and an improvement in meteorological conditions. The months of October through January have the greatest PM2.5 monthly values because of increased airflow, greater atmospheric stability, and temperature inversions that cause pollution to accumulate in the city.

Observing Monthly PM10 pollutant Conc. FY 2018,2019,2020,2021,2022.

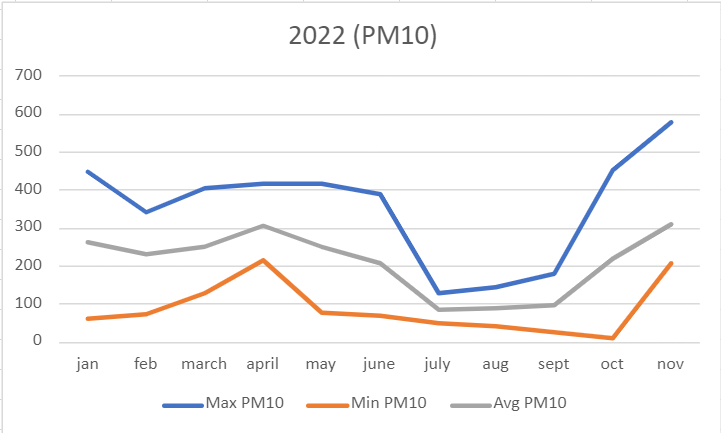
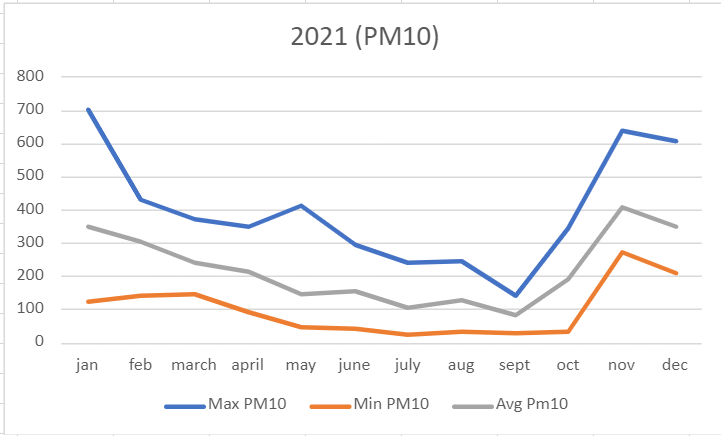
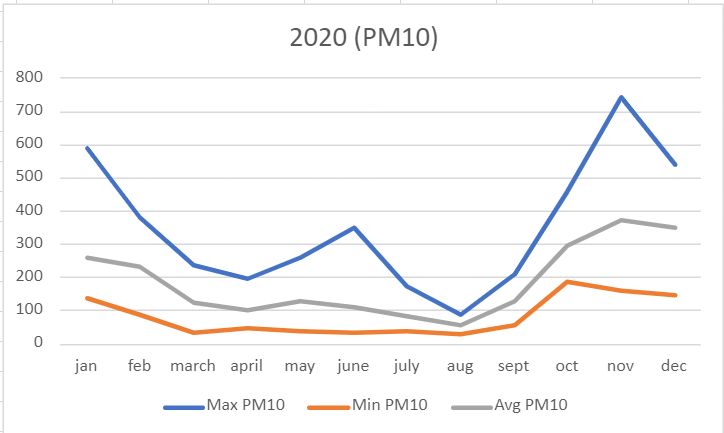
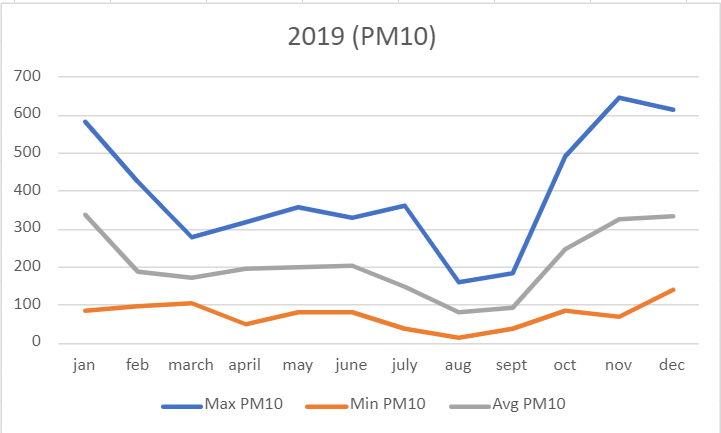
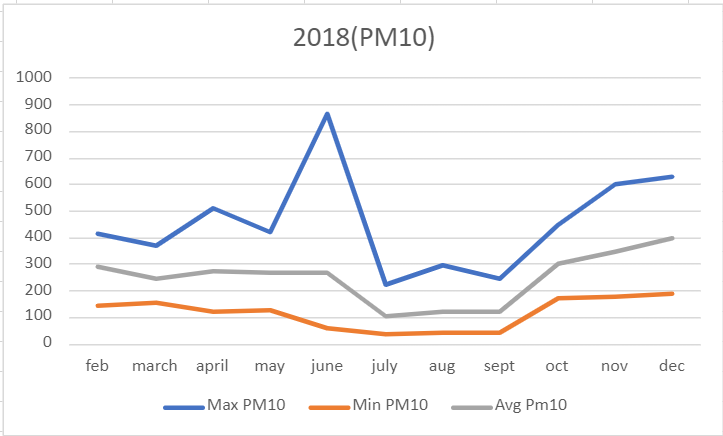


Fig 3: PM10 Conc. FY 2018,2019,2020,2021,2022

According to Figure 3, The months of June through September saw the lowest monthly amounts of PM10 pollution due to a decline in traffic and an improvement in meteorological conditions. The months of October through January have the greatest PM10 monthly values because of increased airflow, greater atmospheric stability, and temperature inversions that cause pollution to accumulate in the city.

Observing Monthly NO2 pollutant Conc. FY 2018,2019,2020,2021,2022.

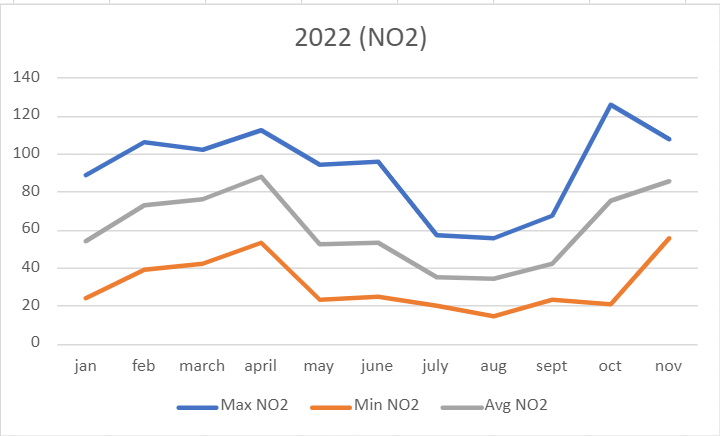
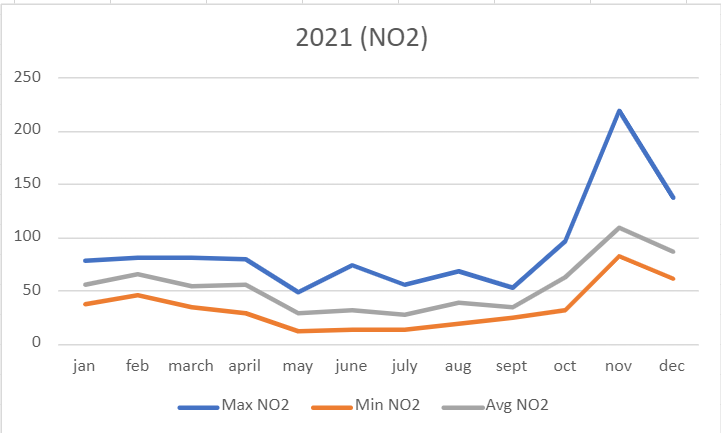
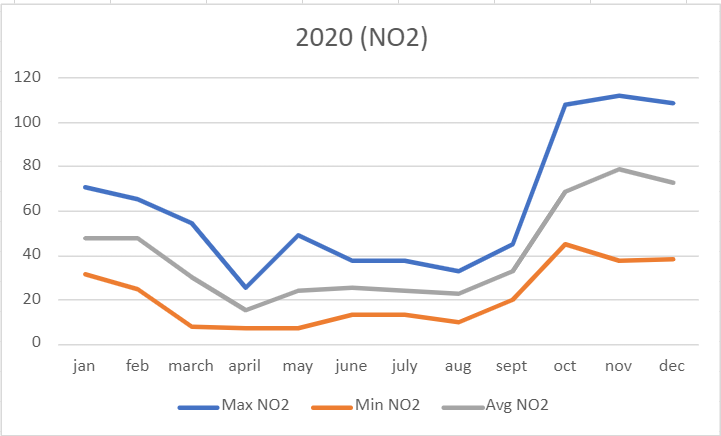
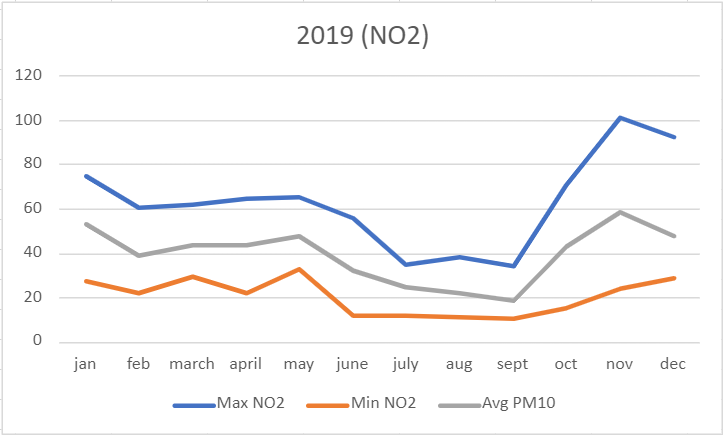
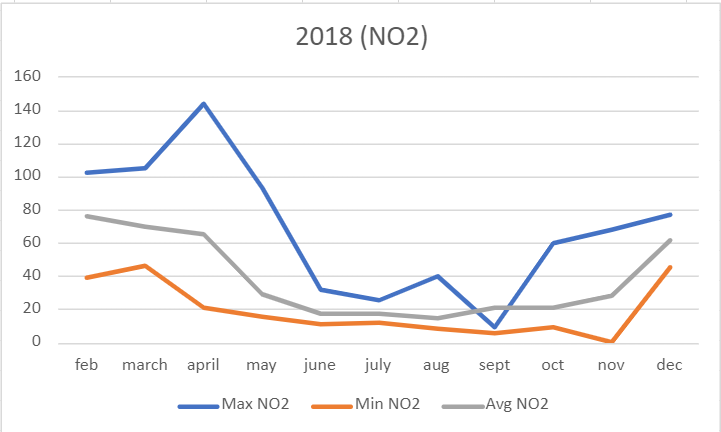


Fig 4: NO2 Conc. FY 2018,2019,2020,2021,2022

According to Figure 4, The months of June through August saw the lowest monthly quantities of NO2 pollution due to a decline in traffic and an improvement in meteorological conditions. Due to enhanced atmospheric stability, temperature inversions, and higher airflow, the months with the greatest NO2 monthly concentrations are recorded to be October through January, correspondingly.

Observing Monthly CO pollutant Conc. FY 2018,2019,2020,2021,2022.

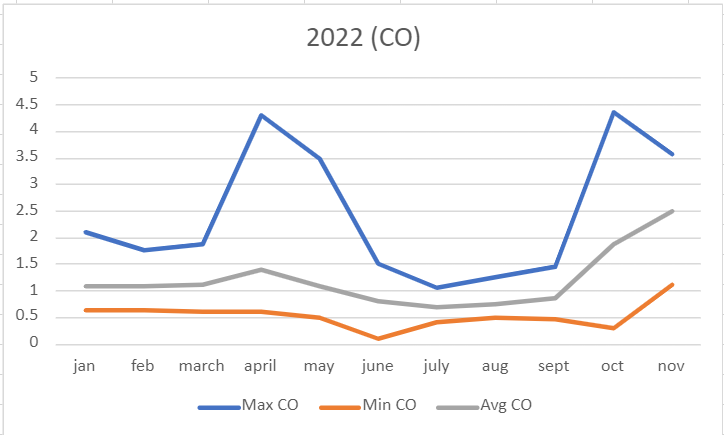
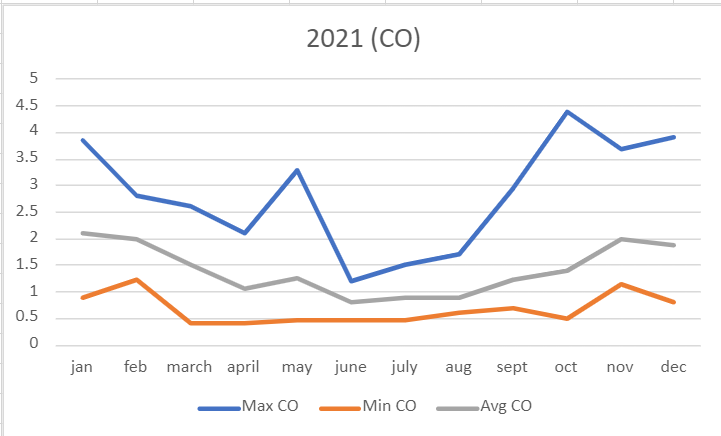
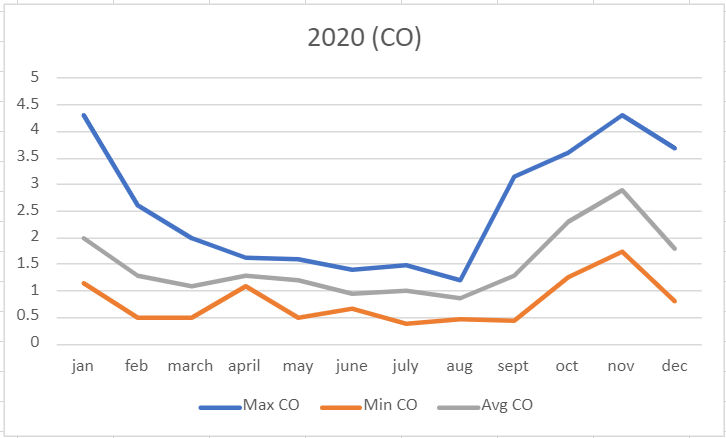
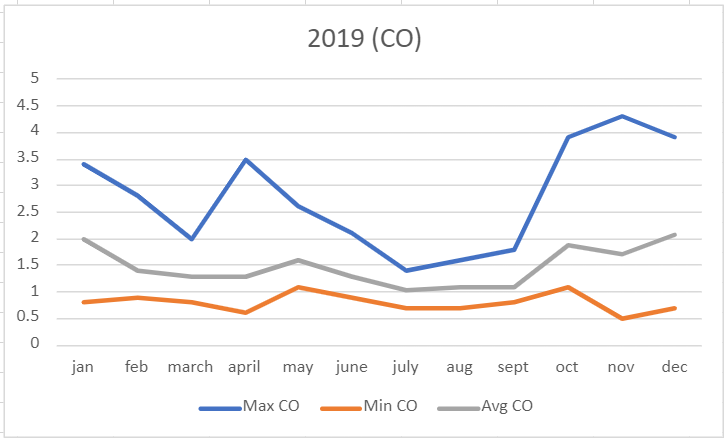
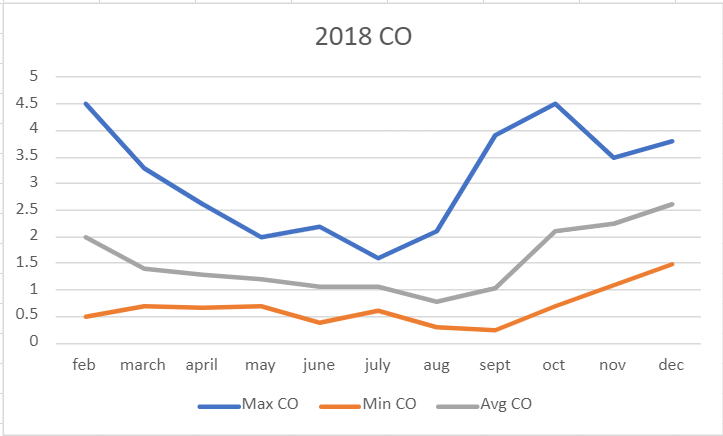


Fig 5: PM10 Conc. FY 2018,2019,2020,2021,2022

According to Figure 5, The months of June through September saw the lowest monthly CO pollution concentrations due to less traffic and improved atmospheric conditions. Due to airflow, enhanced atmospheric stability, and temperature inversions, which allow pollutants to accumulate in the city, the largest quantities of CO monthly concentrations are seen from October to January, respectively.

You can see all the parameters in Figure 6.

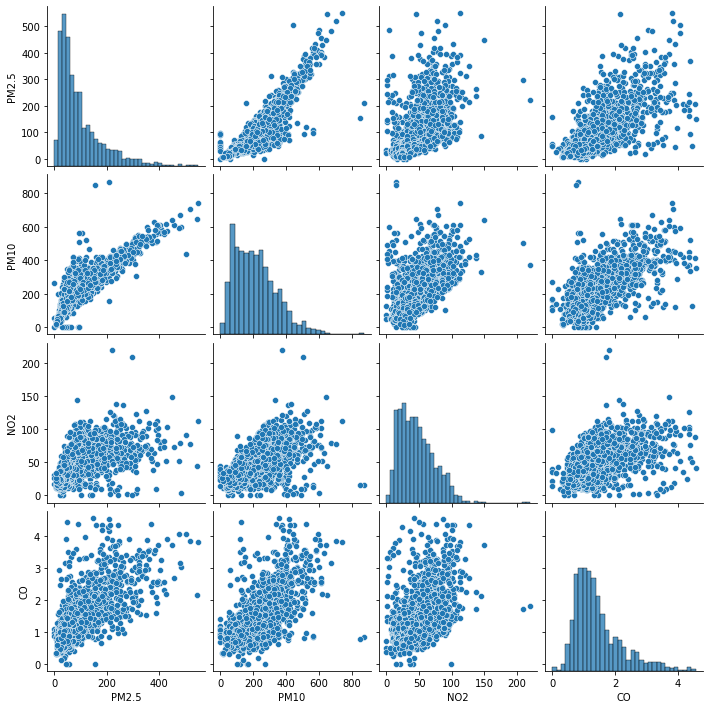
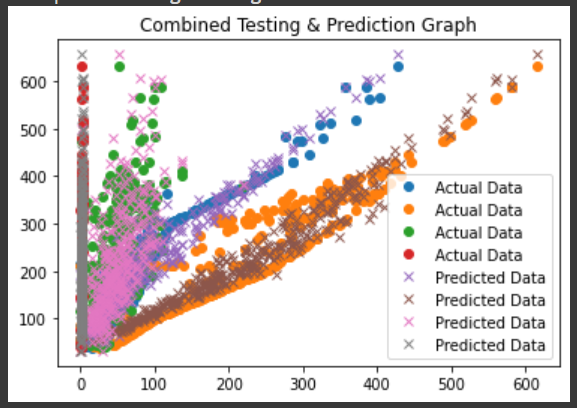


Fig 6: Pairplot of (PM2.5,PM10,NO2 and CO)

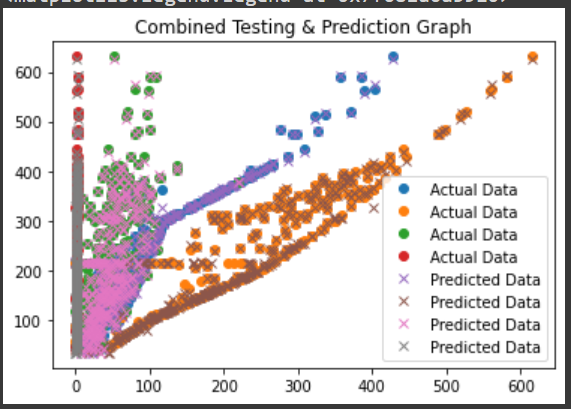
1. Results and Discussion

Based on the air quality evaluated Graph from February 2018 to November 2022 as displayed in Fig: 2,3,4,5,6 this paper has noticed The research period's evaluation outcomes and the previous air quality evaluation results across this region showed consistent quality assessment results. The following factors contribute to this conformity: first, the physical positions of the city are quite similar, the city has a shared external environment, and the air pollution management strategies have similar characteristics.

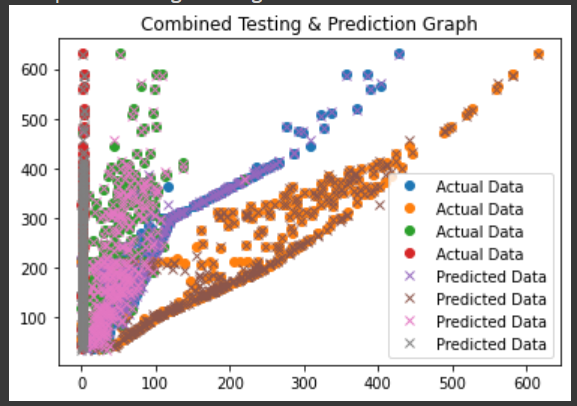
We can Understand the Results from Graph 1, 2, 3, 4.



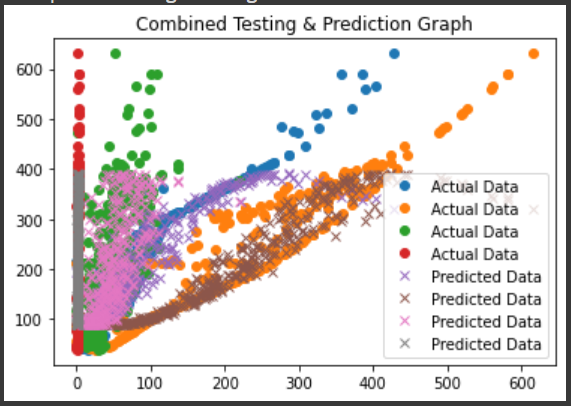
Gaph 1 Prediction by Linear Regression



Garph 2 Prediction by Decision Tree Regression



Graph 3 Prediction by Random Forest Regressor



Graph 4 Prediction by Support Vector Regressor

Observing Graphs 1, 2, 3, and 4 gives the study a clearly perspective that Graph 3 Prediction by Random Forest Regressor has made the most accurate prediction.

According to a recent research, Delhi may only be responsible for 26% of its PM2.5 (microscopic, respirable particles) pollution. In the winter, this number rises to 36%. The upwind areas outside of the city are where the majority of the pollution in this category originates. The research found that air pollution level goes high every winter from the end October to mid February and goes low in monsoon from July to Mid-September.

1. CONCLUSION

After reviewing certain papers applying machine learning techniques to predict air pollution. This study is carried out which is currently in its infancy. air quality prediction using machine learning based on Okhla Phase II CPCB data for the New Delhi region from February 2018 to November 2022, this paper uses different machine learning algorithm to evaluate Daily air quality to demonstrate past effectiveness and future projections. The RMSE & R2 was highlighted in Fig: 11,12,13,14,15 these are different models predictions accuracy score that make the Study found that Random Forest Regressor was the most reliable algorithm for predicting air pollution, with a result of 99.3%.

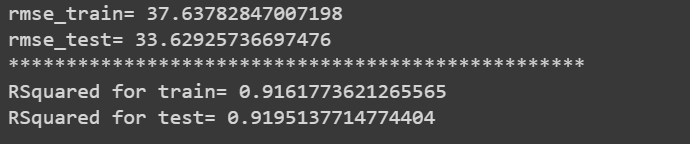


Fig 11: Linear Regression Accuracy

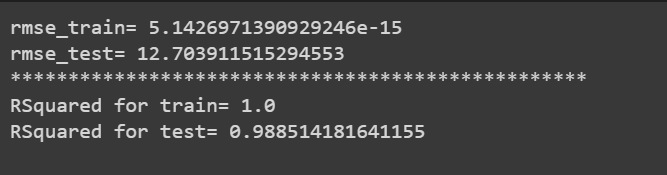


Fig 12: Decision Tree Regression Accuracy

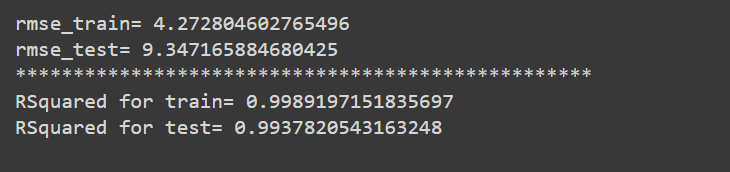


Fig 13: Random Forest Regression Accuracy

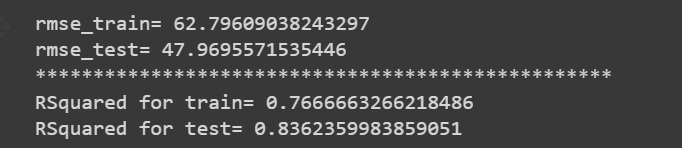


Fig 14: Support Vector Regression Accuracy



Fig 15: KNN Accuracy

References

1. Ghufran Isam Drewil , Dr. Riyadh Jabbar Al-Bahadili. Forecast Air Pollution in Smart City Using Deep Learning Techniques: A Review
2. Cicerone Laurentiu Popa , Tiberiu Gabriel Dobrescu, Catalin-Ionut Silvestru, Alexandru-Cristian Firulescu,

Constantin Adrian Popescu and Costel Emil Cotet. Pollution and Weather Reports: Using Machine Learning for Combating Pollution in Big Cities

[3] Hao Gao 1,†, Weixin Yang 1,\* ,† , Jiawei Wang 1,† and Xiaoyun Zheng. Analysis of the Effectiveness of Air Pollution Control Policies Based on Historical Evaluation and Deep Learning Forecast: A Case Study of Chengdu-Chongqing Region in China

[4] Mahmoud Reza Delavar 1,\* , Amin Gholami 2 , Gholam Reza Shiran 3 , Yousef Rashidi 4 , Gholam Reza Nakhaeizadeh 5 , Kurt Fedra 6 and Smaeil Hatefi Afshar 2. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran.

[5] <https://www.delhitourism.gov.in/delhitourism/aboutus/seasons_of_delhi.jsp>

[6] <https://epca.org.in/state-air-winter-2017-18.pdf>

[7] https://timesofindia.indiatimes.com/city/delhi/36-of-delhis-winter-bad-air-is-its-own-doing-study/articleshow/65431266.cms

[8] https://thewire.in/health/air-pollution-india-life-expectancy-five-years#:~:text=India's%20average%20PM2.,5%20%CE%BCg%2Fm3)%20respectively.

[9] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[10] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[11] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[12] https://www.geeksforgeeks.org/machine-learning/

[13] https://timesofindia.indiatimes.com/city/delhi/delhi-air-takes-severe-hit-worse-yet-to-come/articleshow/66467293.cms

[14] https://timesofindia.indiatimes.com/auto/miscellaneous/reducing-air-pollution-biggest-reason-to-buy-electric-vehicles-not-rising-fuel-cost-survey/articleshow/65709873.cms

[15] <https://timesofindia.indiatimes.com/city/delhi/why-no2-may-be-your-next-big-headache/articleshow/62842535.cms>

[16] <https://timesofindia.indiatimes.com/city/delhi/study-says-people-didnt-feel-air-better-this-winter/articleshow/63177870.cms>

[17] <https://airquality.cpcb.gov.in/AQI_India/>

[18] [https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-comparison-data](https://airquality.cpcb.gov.in/ccr/)

[19] https://swachhindia.ndtv.com/air-pollution-what-is-air-quality-index-how-is-it-measured-and-its-health-impact-40387/

[20] Google Scholar

[21] [www.mdpi.com](http://www.mdpi.com/)

[22] www.ncbi.nlm.nih.gov

[23] mrcet.com

[24] www.ijates.com

[25] mpdi-res.com

[26] https://www.pranaair.com/blog/what-is-air-quality-index-aqi-and-its-calculation/

Author

Sahil Negi, Galgotias University, Negisahil1960@gmail.com.