EasyChair Preprint

№ 9202

Predictive Analysis of Air Polluting Factors Using Support Vector Machines Algorithm

Adhikari Durga Venkata Madhav, Addanki Gargeya, Amanchi Sravan Kumar and T. Dhiliphan Rajkumar

October 31, 2022

Predictive Analysis of Air Polluting factors Using Support Vector Machines Algorithm

Abstract — Today Growing trends in Air pollution is possessing threat to environment. Various Researchers have extended their work in predicting air polluting using various predictive analytics. and in the recent paper the accuracy is too low and they have did using ANN algorithm . so In this paper, we are implementing a predictive model for monitoring air polluting factors level in different cities of India and publishing it as a web service .The algorithm being used is support vector machines . A comparison has also been carried out in different predictive analytics mainly using Machine Learning techniques support vector machine.

Keywords: SVM, ANN, accuracy, predictive model , machine learning

INTRODUCTION

With the ever growing development, air pollution is becoming a serious threat to environment. Government and Environment bodies are taking necessary steps to monitor the air pollution factors and their levels and take precautionary measures. Some Indian cities fall in the array of the most polluted cities in the world, and the threat of air pollution is being raised day by day. Poor air quality in India is now considered a signifcant health challenge and a major obstacle to economic growth. Various Researchers have applied predictive modelling techniques to develop air pollution models which can be used to monitor air polluting factor levels in different regions. In this paper also, we had tried to develop a prediction model for air polluting factor . The model is published as a web service, so that it can be used by Third party vendors or government agencies to use air polluting predictive data for monitoring and regulating the levels of these factors . For this project, we are capturing data form different cities in India and monitoring parameters like Nitrogen Dioxide, Oxides of Nitrogen, Sulfur Dioxide, Carbon Monoxide, Ozone, pollution released by vehicles and industries . Dataset has been obtained from the web repository of Central board of Air pollution, India. The methodology which is being applied to develop predictive model had been developed by using support vector machines. SVR is an analytical machine learning approach which is used to explore the relationship between one or more predictor variables and a real valued dependent variable. It is an approach in which the model learns the importance of distinguishing the co-relation between the input and output. It can applied to regression problems by introducing a loss funvtion. The basics of SVR is to map a input data to a random high dimensional feature space and then to obtain and solve the regression problem in that particular feature space.

Literature survey:

We are aware that in our planet, we use things like gasoline, soil, water, and so forth on a daily basis. We are polluting them, which causes air pollution and global warming, both of which result in dangerous air pollution.

Firstly, we've identified keywords like air quality index, forecasting, prediction, supervised machine learning algorithms, machine learning which are mainly related to our paper

Using the identified keywords, we searched for the previously done researches related to prediction of Air pollution in google scholar, IEEE, etc.

Important research works were filtered by considering inclusion criteria and exclusion criteria.

Inclusion criteria:

1. Include studies that are related to both prediction of air pollution and machine learning.

2.peer review of the articles related to air quality forecasting.

3. Include the articles written in english.

4. Only published papers should be included. Exclusion criteria:

1. The gathered works which are not scientific are excluded.

2. The works which don't follow proper guidelines like without having abstract are excluded.

Some research works related to air pollution prediction were gathered from the search.

In the past, there have been many research papers on this topic, but their accuracy was low due to a lack of data and they did not use any other techniques. In our paper, accuracy is somewhat improved over previous papers, and we are using the machine learning algorithm support vector machines, which is a well-known machine learning algorithm and has a high accuracy.

In support vector machines, the graph's data is divided into positive and negative segments, which are then separated by a hyperplane, which can be thought of as the centre line.

The dataset is crucial when conducting research on prediction, and if we provide more historical data, it will be easier to predict. According to our research, we are providing a 5 year dataset of various cities, including Delhi, Ahmedabad, etc. Several websites, like cpcb, google cloud data, etc., may provide the data set.

In our dataset, we take into account factors like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzen, Toluene, and Xylene. In previous research papers, they also took into consideration some other factors, but they did not provide a proper dataset, such as an unfiltered dataset, and in this, we provide a proper dataset that aids in accurate prediction and may also affect accuracy.

Data is gathered from a source, however during collection, there are duplicates, irrelevant data, missing values, and other issues with the data that we don't want to include.

As a result, some procedures must be taken to remove the undesired data, such as data processing, data cleaning, removing duplicates, handling missing values, clearing format errors, converting data types, data reduction,etc.

Data cleaning: Data cleaning is the process of removing unwanted data like incorrect data, duplicate data, unformatted data from the data set. By cleaning the data, we can improve the accuracy of the result.

Convert data types: Numbers may occasionally be entered as text. Then the numbers data type will be string. We are unable to do mathematical operations on them because they were strings. Therefore, in order to conduct operations on the data types, we must convert them.

Clear formatting: Machine learning models cannot process data that has a high level of formatting. It will be unclear if our data uses different formats.

Handling missing values: We can deal with missing data by either eliminating the entire tuple or by adding the missing values. In the blank field, we can provide a rough figure. The tuple data with missing values can be removed if the data set is too big.

Data reduction: Analysis becomes more difficult when dealing with large amounts of data. We employ data reduction techniques to deal with this. Data reduction aims to increase storage effectiveness while lowering the cost of data storage and processing. Data cube aggregation, attribute subset selection, numerosity reduction, and dimensionality reduction were some of the processes in the data reduction process.

After analysing the data, we need to train the machine learning model, then test it. To do this, we use the Python compiler and support vector machines for the machine learning algorithm.

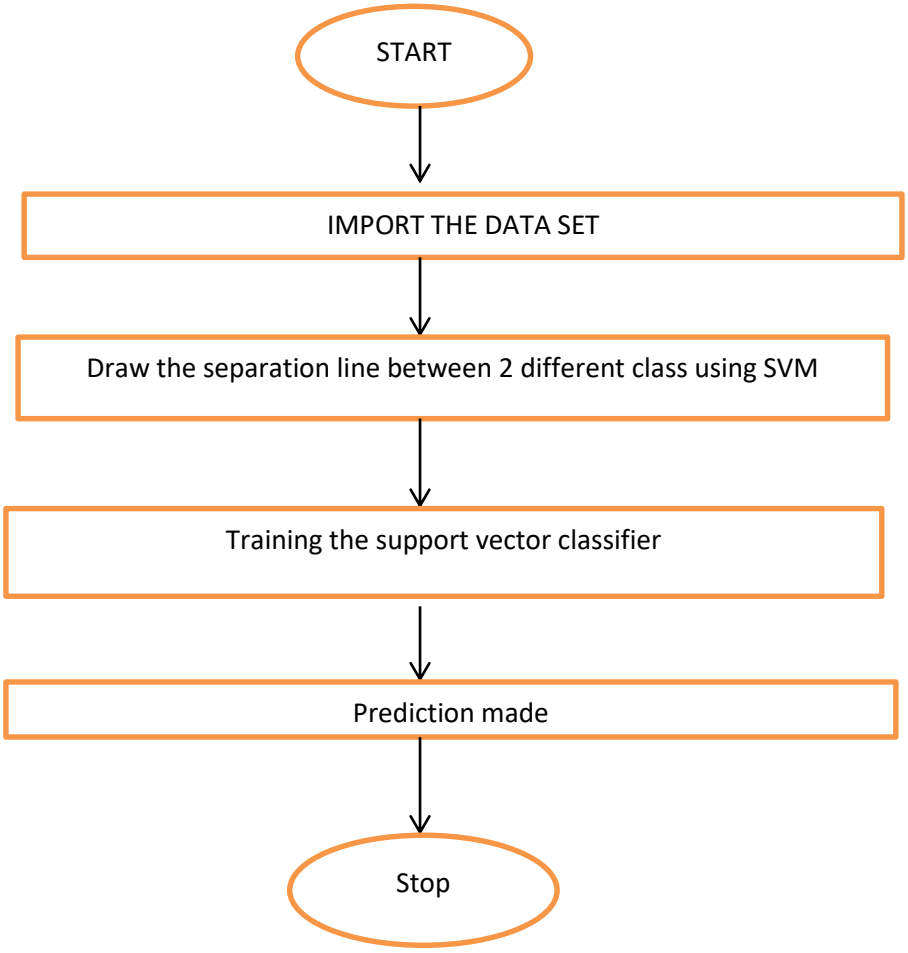
The output of a machine learning algorithm is often a performance matrix, from which we may calculate various metrics such as accuracy, prediction, f1 score, recall, confusion matrix, etc. are determined by the performance matrix.

RESEARCH METHODOLOGY:

Dataset: For the current Research work, the dataset provided by central pollution control board has been exploited, which concentrates on the scientifically accepted parameters responsible for air pollution. The parameter which are being monitored for research work are Nitrogen ,Sulfur Dioxide ,Carbon Monoxide ,Ozone,PM2.5,Relative Humidity, Wind Speed and Wind Direction and some other parameters .

ALGORITHM:

SVM is a Supervised Learning algorithm in which it divides the plane into 2 parts by drawing a line between the 2 different classes. The line which separates the plane into different parts is called hyperplane. It always gives a perpendicular distance from the data point to the line of separation. It can do both linear and nonlinear classification. It is mainly used to do the classification and regression.



Flow chart about Support Vector Machines

Input Data : we collect the data from central pollution of control board the dataset contains various attributes like PM2.5,PM10,NO,N02,NOx,NH3,CO,SO2,O3,Benzene,Toluene,xylene . we considered these

12 values and collected dataset containing 5years of pollution data . we summarized the attributes and corresponding values.

Data preprocessing : The first and most crucial need for effective visualisation and the development of effective ML models is data quality. The preprocessing procedures aid in decreasing the noise in the data, which ultimately speeds up processing.intially the data set contains noisy, inconsistent data and missing values. The data has to be preprocessed to remove the unwanted data and to make the data useful. Data preprocessing helps to transform data into useful format we have to follow certain steps

1.Data Cleaning

2.Data reduction

3.Data transformation

To Download Data set : https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data

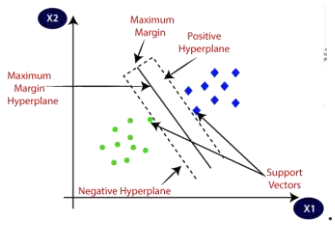
Apply Machine Learning Techqniues : once the data is ready for modelling, we apply the most popular machine learning algorithm support vector machines to predict the pollution factors

About Support Vector Machines:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



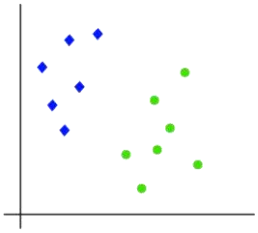
Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

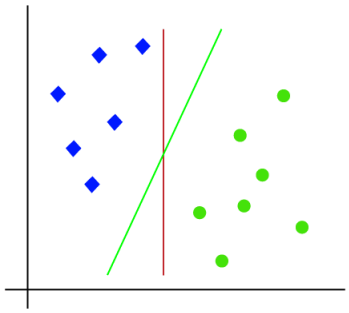
How does SVM works?

Linear SVM:

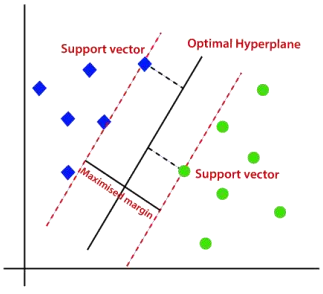
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:

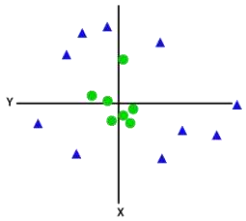


Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



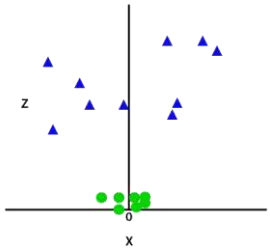
Non-Linear SVM:

If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as:

$$z=x^2+y^2$$



Performance Metrics :

To analyze the performance of a machine learning model we need some metrics. These metrics are statistical criteria that can be used to measure and monitor the performance of a model.

The performance evaluation metrics used in this experiments are listed below

- If TP belongs to true positive rate and FP belongs to false positive rate then according to the formal definition of precision is Precision = (TP) / (TP + FP).
- Recall is defined as below where FN represents the false Negative rate.

  Recall = (TP) / (TP + FN)
- Accuracy = (TP + FN) / (TP + TN + FP + FN)

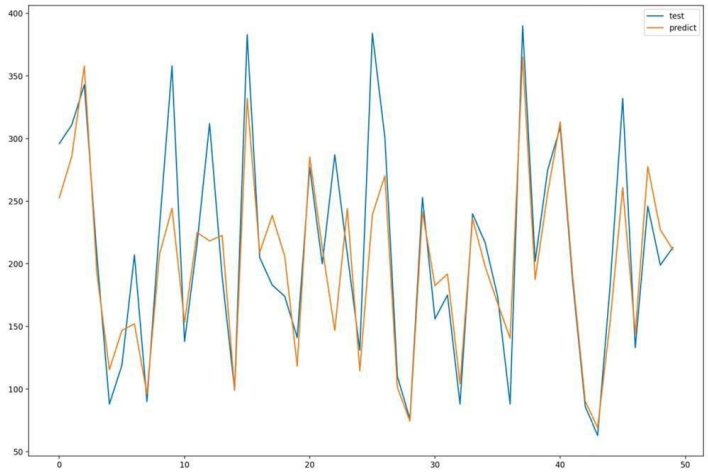
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       388
           1       1.00      1.00      1.00      2397
           2       1.00      1.00      1.00      2608
           3       1.00      1.00      1.00       805
           4       1.00      1.00      1.00       861
           5       1.00      1.00      1.00       324

    accuracy                           1.00      7383
   macro avg       1.00      1.00      1.00      7383
weighted avg       1.00      1.00      1.00      7383

[[ 388    0    0    0    0    0]
 [   0 2397    0    0    0    0]
 [   0    0 2608    0    0    0]
 [   0    0    0  805    0    0]
 [   0    0    0    0  861    0]
 [   0    0    0    0    0  324]]
```

Result Analysis:

The CPCB dataset under study involves a specific parameter viz, AQI and government agencies use this parameter to alert people about the quality of the air and also practice forecasting it. According to the National Ambient Air Quality Standards, there are six AQI categories: good (0–50), satisfactory (51–100), moderate (101–200), poor (201–300), very poor (301–400), and severe (401–500). The table also illustrates the effects of pollution.and the current pollution levels, if they were severe, would result in a red colour and a prediction system that would signal pollution.

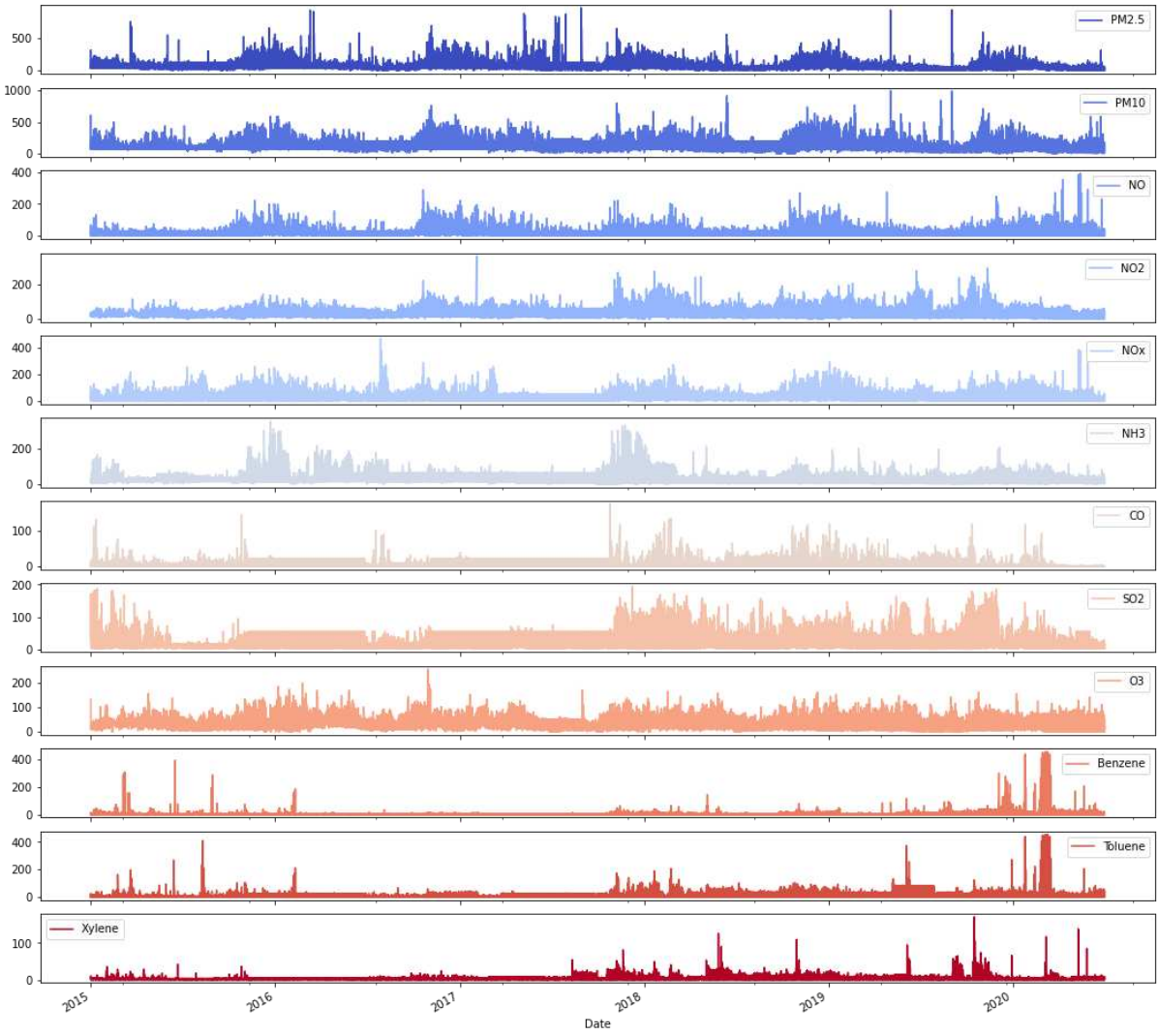
AQI Category	Associated Health Impact
Good (0 to 50)	Minimal impact
Satisfactory (51 to 100)	May cause minor breathing discomfort to sensitive people
Moderately Polluted (101 to 200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201 to 300)	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease
Very Poor (301 to 400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401 to 500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

According to the data set, some pollutants are higher in some cities, such as Ahmedabad and Bhopal, and the northen states are having highest pollution than the southeren states, and most of these are due to the factories are more in the north states. The below graph shows the average AQI in the last five years in different cities, and mostly Ahmedabad is having highest pollution and delhi is having the second postion and the pollutants like PM2.5,CO are mostly present in this city

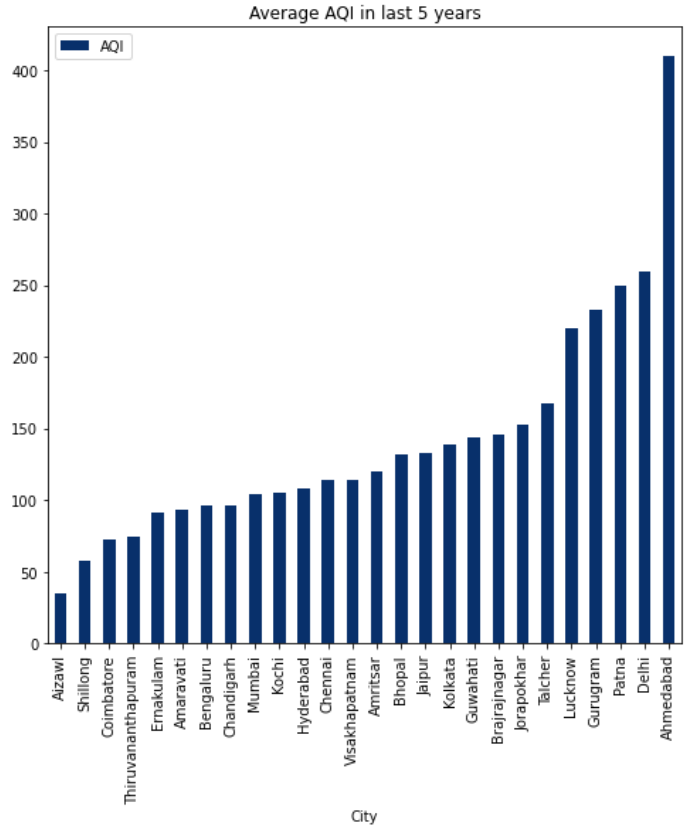


Prediction of SVM and Real values

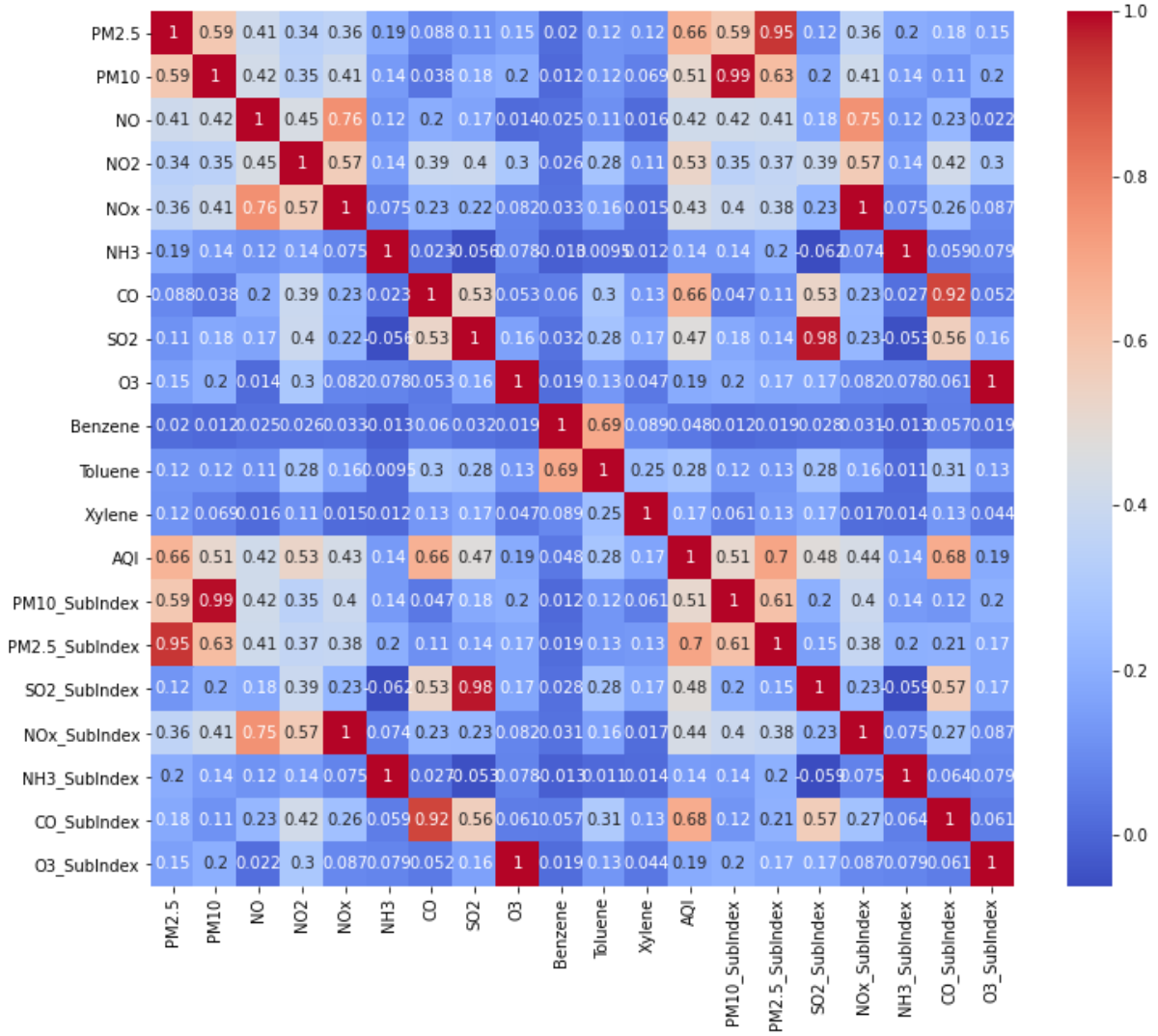
The accuracy is higher here since the predicted values and test values are somewhat similar.



The graph up above the concentrations of PM2.5, PM10, NO, NO2, NOX, NH3, CO, Benzene, TOULENE, and XYLENE. and these are shown in this graph with various colours according to their factors, taking into account five years' worth of data. Most commonly, factors like PM2.5 and PM10 have the highest levels of pollution, and they are also subject to some limitations, such as when they can be taken. And the graph is mean while increasing from 2015 to 2020 and these are due to rising levels of pollutants. When we compare the graph to other factors, such as xylene and benzene, which had low percentages in 2015 but are now rising, and pm2.5 and pm10, which are still rising, we have compared these factors and come to the conclusion that the factors are rising every year.



Scholars in the realm suggest that reducing input variables lowers the computational cost of modeling and enhances prediction performance. A correlation-based feature selection method has been exploited in the present work to determine the optimal number of input variables (pollutants) when developing a predictive model. Statistical correlation-based feature selection algorithms compute correlations between every pair of the input variable and the target variable. The variables possessing the strongest correlation with the target variable are then filtered for further study. Since many ML algorithms are sensitive to outliers, any feature in the input dataset which does not follow the general trend of that data must be found. For the present dataset, a correlation-based statistical outliers detection method has been applied to identify the outliers. To select significant features, the correlation analysis of the AQI feature has been exercised with features of other pollutants. shown below clearly reveals that pollutants PM10,PM2.5, CO, NO2,SO2,NOX, and NO are generally responsible for the AQI to attain higher values.



Conclusion:

Prediction of air quality is a challenging task because of the dynamic environment, unpredictability, and variability in space and time of pollutants.The concentration of air pollutants in ambient air is governed by the meterological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Air Quality Index(AQI), is used to measure the quality of air.

Reference:

1.Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar, Air Pollution Prediction Using Machine Learning Supervised Learning, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 04, APRIL 2020

2.A. Gnana Soundari , J. Gnana Jeslin, Akshaya A, INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING , Volume 14, Number 11, 2019

3.Samayan Bhattacharya, Sk Shahnawaz , Using Machine Learning to Predict Air Quality Index in New Delhi

4.Andreas Jacobsen Lepperød, Air Quality Prediction with Machine Learning, June 2019

5.Yun-Chia Liang , Yona Maimury , Angela Hsiang-Ling Chen ,and Josue Rodolfo Cuevas Juarez , Machine Learning-Based Prediction of Air Quality, 21 December 2020

6.Miss Ruchita Nehete, D. D. Patil , AIR QUALITY PREDICTION USING MACHINE LEARNING , 2021 IJCRT | Volume 9, Issue 6 June 2021

7.Rohit Adke, Suyog Bachhav, Akash Bambale, Bhushan Wawre, Air Pollution Prediction using Machine Learning , Volume: 06 Issue: 10 | Oct 2019

8.JohnH.Ludwvig,Sc.D.,F.A.P.H.A.,andB.J.Steigerwald,Ph.D.,RESEARCHINAIRPOLLUTION:CURRENTTRENDS,VOL.55.NO.7.A.J.P.H.

9.Anthony Heyes Matthew Neidell Soodeh Saberian, THE EFFECT OF AIR POLLUTION ON INVESTOR BEHAVIOR: EVIDENCE FROM THE S&P500,http://www.nber.org/papers/w22753,MA 02138 October 2016

10.SVM Example Dan Ventura March 12, 2009

11.Andrew ng, CS229 Lecture notes

12.A. Suárez Sáncheza, P.J. García Nietob, P. Riesgo Fernándeza, J.J. del Coz Díazc, F.J.Iglesias-Rodríguez, Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain), Mathematical and Computer Modelling 54 (2011) 1453–1466

13.Mrs. Monia , Ms. Akanksha Gupta ,Ms. Sameiksha Sharma , Predictive Analysis of Air Pollution Using Collaborative Filtering Prediction Algorithm , 2018

14.Jayant Kumar Singh , Amit Kumar Goel, prediction of Air pollution by using machine learning algorithm, 2020,7th ICACCS

15.Avan Chowdary Gogineni Vamsi Sri Naga Manikanta Murukonda, Prediction of Air Quality Index Using Supervised Machine Learning, May 2022

16.K. Kumar,B. P. Pande, Air pollution prediction with machine learning: a case study of Indian cities, https://doi.org/10.1007/s13762-022-04241-5, 19 April 2022

17.Arpan Chatterji , Air Pollution in Delhi: Filling the Policy Gaps, December 2020

18.Sapan Bali,Ms Nidhi Sengar , INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING , vol 11 Issue 5 may 2020,ISSN NO:03777-9254

19.shiv das meena, POLLUTION CONTROL ACTS, RULES & NOTIFICATIONS ISSUED THEREUNDER, CENTRAL POLLUTION CONTROL BOARD , Delhi 08April,2021

20.Dan Wei , Predicting air pollution level in a specific city

21.Shuyue Zhang, Minfeng Lin, Xiuguo Zou, Steven Su, Wentian Zhang, Xuhui Zhang and Zijie Guo, LSTM-based Air Quality Predicted Model for Large Cities in China , 02-07-2019, Vol. 19,p-ISSN: 0972-6268 e-ISSN: 2395-3454,

22.K Srinivasa Rao , G. Lavanya Devi , N. Ramesh , Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks , 08 February 2019 ,http://www.mecs-press.org

23.Muthukumar Neelamegam, Karin Anna Hummel, March, 2022,PREDICTING AIR QUALITY USING WEATHER FORECASTING AND MACHINELEARNING

24.Sweta, Machine Learning Approaches to Ambient Air Quality Prediction , Volume 11 Issue 4, April 2022

25.Aleksandar Trenchevski, Marija Kalendar, Hristijan Gjoreski and Danijela Efnusheva , Prediction of Air Pollution Concentration Using Weather Data and Regression Models , March 2020 ,8th International Conference on Applied Innovations in IT, (ICAIIT)

26.Anurag Sinha, Shubham Singh, A Review on Dynamic Forecasting of Air Pollution in Delhi Zone , Volume 7 ~ Issue 4 ,www.questjournals.org,10 May, 2021