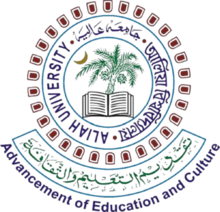
****

**ALIAH UNIVERSITY**

**NEWTOWN CAMPUS, KOLKATA-156**

**WEST BENGAL**

**Prediction of Air Quality Using**

**Regression Models**

*A Project*

*Submitted in partial fulfillment for the*

*Award of the degree of*

**Master of Computer Applications**

***Submitted by***

|  |
| --- |
| **Julfikar Sk (MCA173010)** |
| **Shahajahan Ahamed (MCA173020)** |

*Under the Guidance of*

**Dr. Zeenat Rehena**

**DECLARATION**

We hereby declare that this submission is our own work to the best of our knowledge and belief, it contains no material previously or written by any other person nor material which to a substantial extent has been accepted for the award of any degree of the university or other institute of higher learning, except where the acknowledgment has been made in the text. We have taken care in all respect to honor the intellectual property right and have acknowledged the contribution of others for using them in academic purpose and further declare that in case of any violation of intellectual property right or copyright we, as a candidate, will be fully responsible for the same. Our supervisor should not be held responsible for full or partial violation of copyright or intellectual property rights.

**Name: Julfikar Sk Name: Shahajahan Ahamed**

**Roll No: MCA173010 Roll No: MCA173020**

Date: 03-07-2020

Place: Kolkata

**CERTIFICATE**

Certified that the Project entitled **Prediction of Air Quality Using Regression Models** submitted in partial fulfillment of the degree of Master of Computer Applications (**MCA**) in Aliah University done by **Julfikar Sk**, Roll No MCA173010, is a record of the student’s own work carried out under my supervision and guidance. The matter embodied in this project has not been submitted earlier for award of any degree to the best of my knowledge and belief. It is further understood that by his certificate the undersigned does not endorse or approve of any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose for which it is submitted. Project guide Head of the department.

|  |
| --- |
| Project guide :  Dr. Zeenat Rehena |
|  |
|  |
| Head of the department  Dr. Obaidullah Sk |

**CERTIFICATE**

Certified that the Project entitled **Prediction of Air Quality Using Regression Models** submitted in partial fulfillment of the degree of Master of Computer Applications (**MCA**) in Allah University done by **Shahajahan Ahamed**, Roll No MCA173020, is a record of the student’s own work carried out under my supervision and guidance. The matter embodied in this project has not been submitted earlier for award of any degree to the best of my knowledge and belief. It is further understood that by his certificate the undersigned does not endorse or approve of any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose for which it is submitted. Project guide Head of the department.

|  |
| --- |
| Project guide :  Dr. Zeenat Rehena |
|  |
|  |
| Head of the department  Dr. Obaidullah Sk |

**ACKNOWLEDGEMENT**

First and foremost I bow to the almighty for giving me the opportunity to undertake the writing of this project and the strength and capacity to complete it. Then I take this opportunity to express my sincere and deep sense of gratitude to **Prediction of Air Quality Using Regression Models,** My humble and heartfelt acknowledgements are also to my esteemed guide **Dr. Zeenat Rehena** for her guidance and support, without her help this task would not have been accomplished. I would like to thank my friend **Shahajahan Ahamed** if any for his constant and timely help, moral support and valuable suggestions. In addition I thank one and all who have been instrumental in helping me complete this project. I am extremely grateful and indebted to my parents and my siblings for being pillars of strength, for their unfailing moral support, and encouragement. I treasure their blessings and good wishes and dedicate this study to them.

**Name: Julfikar Sk**

**Roll No: MCA173010**

**ACKNOWLEDGEMENT**

First and foremost I bow to the almighty for giving me the opportunity to undertake the writing of this project and the strength and capacity to complete it. Then I take this opportunity to express my sincere and deep sense of gratitude to **Prediction of Air Quality Using Regression Models** , My humble and heartfelt acknowledgements are also to my esteemed guide **Dr. Zeenat Rehena** for her guidance and support without her help this task would not have been accomplished. I would like to thank my friends, who have helped me during this study. In addition I thank one and all who have been instrumental in helping me complete this project. I am extremely grateful and indebted to my parents and my siblings for being pillars of strength, for their unfailing moral support, and encouragement. I treasure their blessings and good wishes and dedicate this study to them.

**Name: Shahajahan Ahamed**

**Roll No: MCA173020**

**ABSTRACT**

Air pollution is defined as a phenomenon harmful to the ecological system and the normal conditions of human existence and development when some substances in the atmosphere exceed a certain concentration. The quality of air is adversely affected due to various forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases including carbon dioxide, nitrous oxide, methane, ozone, and mainly Particulate Matter (PM10 & PM2.5) etc. is creating a serious threat for the quality of life in all over the world. In this situation, we need to implement efficient air quality monitoring systems which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation, monitoring, and prediction has become an important research area and these things cannot be possible without Machine Learning. In the past, many environmental researchers have dedicated their research efforts on this subject using many approaches. The aim of this research paper is to collect the concentration of the pollutants and train some regression model like Multiple Linear Regression, Support Vector Regression and Decision Tree Regression to find out the best model for prediction, which can be further used for developing the Air Quality Monitoring System.

Index

|  |  |
| --- | --- |
| **Table of Contents** | **Page No** |
| Declaration | i |
| Certificate(2) | ii- iii |
| Acknowledgement (2) | iv- v |
| Abstract | vi |
| 1. Introduction | 1-2 |
| 1.1 Background | 2-4 |
| 1.2 Research Objectives | 4 |
| 2. Literature Review | 5 |
| 2.1 Introduction | 5 |
| 2.1.1 About National Air Quality Index | 5 |
| 2.2 Machine Learning Techniques | 6 |
| 2.3 Regression Models | 6 |
| 2.3.1 What is Regression Model? | 6 |
| 2.3.2 Why do we use Regression Model? | 6-7 |
| 2.2.3 How many types of regression techniques do we have?  A. Multiple Linear Regression  B. Support Vector Regression  C. Decision Tree Regression | 7  8  8-9  9-10 |
| 3. Data | 10-11 |
| 3.1 Data Collection | 11 |
| 3.2 Creating Data Set | 12 |
| 3.2.1 Calculate sub index   * PM2.5 * PM10 * NO2 * NH3 * SO2 * CO * Ozone | 12-20  13-14  14-15  15-16  16  17  18  19 |
| 3. 3 Calculating AQI | 20 |
| 3.3.1 How to calculate the | 20-21 |
| 3.4 Data Correlation | 22-23 |
| 3.5 Training Testing Splitting | 23 |
| 4. Methods and Models Set up | 24 |
| 4.1 Multiple Linear Regression(MLR) | 24-25 |
| 4.2 Support Vector Regression | 25 |
| 4.3 Decision Tree Regression | 25-26 |
| 4.4 Pre-processing and Feature Selection | 26-27 |
| 4.5 Result | 27-28 |
| 4.6 Algorithm Comparison | 28-29 |
| 5. Conclusion | 30 |
| 6. References | 30 |

**1. Introduction**

In developing countries like India, Air pollution has been a common health concern not only for humans but also for animals, plants, oceans, and aquatic life worldwide.  In most countries air quality monitoring is done manually via centrally located stations. Air pollution has a direct impact on human health. There has been increased public awareness about the same in our country. Global warming, acid rains, increase in the number of asthma patients are some of the long-term consequences of air pollution.

According to World Health Organization Air pollution is accountable for the death of seven million people worldwide each year or one in eight premature deaths yearly. Almost 570,000 children under the age of five die every year from respiratory infection linked to indoor/outdoor pollution and second-hand smoke.

In the prediction of air pollution, several researchers worldwide have developed models to monitor many of the pollution gases such as Sulphur Dioxide (SO2), Carbon Monoxide (CO), Carbon Dioxide (CO2), Nitrogen Oxides (NO), PM10 and PM2.5 etc.

This paper focuses on building a model of air quality prediction systems. It discusses how the air quality can be predicted by collecting the concentration of pollutants from stations where sensors are available. The main objective of this paper is to find the best model which can predict the air quality very well by taking the concentration value of the pollutants as an input. This model can be used to inform people of how good or bad the air condition is in various places, so that the people can change their lifestyle and try to reduce the air pollution as much they can.

In these days, the whole world is badly facing and suffering from many types of pollution due to change in environment, and Air Pollution is one of the major causes of ailing the whole world. The air is getting polluted from many sources that may be by the emission coming out from Vehicles, factories and Industries. While doing the survey for this work, some research papers are studied which have evaluated the major causes of air pollution and solutions are also proposed in those works using various approaches.

In various measures had been reviewed, which had been used for the prediction of air pollution pollutants. Some of them are Deep Learning, Machine Learning, Feed Forward Neural Network. In the proposed approach utilizes the information pertaining to the unlabelled data to improve the performance of the interpolation and the prediction, and perform features selection. A mechanism has constructed for monitoring the air health in smart city using Context Aware Computing. One of a research article , worked with time series data using deep learning approaches and used Recurrent Neural Networks and Long Short-Term Memory unit as a framework for forecasting the air pollution in South Korea. A deep learning model has proposed to forecast the air pollution. Compared four unpretentious machine learning algorithms, linear regression, Naive Byes, support vector machine and random forest to predict air pollutants levels ahead of time. First introduced a method of integrating multi-source air quality data, for the data preparation of the artificial intelligence based smart urban services. Then, a system is set up with the deployment of air quality-aware healthcare applications. In one of an another article developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem.

The article has proposed a work to predict the Air Quality Index (AQI) using support vector regression (SVR) and random forest regression (RFR) in Beijing and the nitrogen oxides (NOX) concentration in an  
Italian city, based on two publicly available datasets. A investigated machine learning based techniques for air quality prediction and air quality is predicted using supervised machine learning techniques like Logistic Regression, Random Forest, K-Nearest Neighbors, Decision Tree and Support Vector Machines.

One another research article , Logistic regression is employed to simply detect whether a data sample is either polluted or not polluted. The proposed work compared four simple machine learning algorithms, neural network, k-nearest neighbor, support vector machines and decision tree to predict the Air Quality Index.

In this work, the ideas gained from the survey are used to construct a good model which will predict the quality of the air.

1.1 Background

Air pollution is the introduction of particulates, biological molecules, or other harmful materials into the Earth’s atmosphere, causing disease, death to humans, damage to other living organisms such as food crops, or damage to the natural or man-made environment. An air pollutant is a substance in the air that can have adverse effects on humans and the ecosystem. The substance can be solid particles, liquid droplets, or gases. Pollutants are classified as primary or secondary. Primary pollutants are usually produced from a process, such as ash from a volcanic eruption. Other examples include carbon monoxide gas from motor vehicle exhaust, or sulfur dioxide released from factories. Secondary pollutants are not emitted directly. Rather, they form in the air when primary pollutants react or interact. Ground level ozone is a prominent example of a secondary pollutant. The seven “criteria pollutants” are fine particulate matter (PM2.5), PM10, nitrogen dioxide (NO2), ammonia(NH3), sulfur dioxide (SO2), carbon monoxide (CO), ground level ozone (O3). PM2.5 and NO2 (main component of NOx) are the most widespread health threats.

Ground level O3, a gaseous secondary air pollutant formed by complex chemical reactions between NOx and volatile organic compounds (VOCs) in the atmosphere, can have significant negative impacts on human health. Prolonged exposure to O3 concentrations over a certain level may cause permanent lung damage, aggravated asthma, or other respiratory illnesses. Ground level O3 can also have detrimental effects on plants and ecosystems, including damage to plants, reductions of crop yield, and increase of vegetation vulnerability to disease .

Particle pollution (also called particulate matter or PM) is the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope. Fine particulate matter (PM2.5) consisting of particles with diameter 2.5 µm or smaller, is an important pollutant among the criteria pollutants. The microscopic particles in PM2.5 can penetrate deeply into the lungs and cause health problems, including the decrease of lung function, development of chronic bronchitis and nonfatal heart attacks. Fine particles can be carried over long distances by wind and then deposited on ground or water through dry or wet deposition. The wet deposition is often acidic, as fine particles containing sulfuric acid contribute to rain acidity, or acid rain. The effects of acid rain include changing the nutrient balance in water and soil, damaging sensitive forests and farm crops, and affecting the diversity of ecosystems. PM2.5 pollution is also the main cause of reduced visibility.

Nitrogen dioxide (NO2) is one of a group of highly reactive gases known as “nitrogen oxides” (NOx). US Environmental Protection Agency (EPA) Ambient Air Quality Standard uses NO2 as the indicator for the larger group of nitrogen oxides. NO2 forms quickly from emissions of automobiles, power plants, and off-road equipment. In addition to contributing to the formation of ground-level ozone, and fine particle pollution, current scientific evidence links short-term NO2 exposures, ranging from 30 minutes to 24 hours, with adverse respiratory effects including airway inflammation in healthy people and increased respiratory symptoms in people with asthma (EPA, 2005).

The Air Quality Health Index (AQHI) is a public information tool designed in Canada to help understand the impact of air quality on health. Basically, the AQHI is defined as an index or rating scale range from 1 to 10+ based on mortality study to indicate the level of health risk associated with local air quality (Chen and Copes, 2013). The higher the number, the greater the health risk and the need to take precautions. The formulation of Canadian national AQHI is based on three-hour average concentrations of ground-level ozone (O3), nitrogen dioxide (NO2), and fine particulate matter (PM2.5). The AQHI is calculated on a community basis, each community may have one or more monitoring stations and the average concentration of 3 substances is calculated at each station within a community for the 3 preceding hours. AQHI is a meaningful index protecting residents on a daily basis from the negative effects of air pollution. Our study gives direction to predicting individual pollutants of one hour average concentration instead of AQHI (or its maximum) as the formulation of AQHI is based on health related science and may evolve over time. Building a forecast system based on individual pollutants and one hour average concentration will make it more flexible to future changes in health indices. Our result can also be beneficial to external clients and meteorologists.

The concentration of air pollutants including ground level ozone, PM2.5 and NO2 varies depending on meteorological factors, the source of pollutants and the local topography. Among these three factors, the one which most strongly influences variations in the ambient concentration of air pollutants is meteorological factors. Meteorological factors experience complex interactions between various processes such as emissions, transportation and chemical transformation, as well as wet and dry depositions. In addition, the spatial and temporal behavior of wind fields are affected by the surface roughness and differences in the thermal conditions , which further influence the dispersion of pollutants. For example, Revlett and Wolff and Lioy found that ambient ozone concentration not only depended on the ratio and reactivity of precursor species, but also on the state of the atmosphere - the amount of sunlight, ambient air temperature, relative humidity, wind speed, and mixed layer (ML) depth, while Tai found that daily variations in meteorology as described by the multiple linear regression (MLR) including nine predictor variables (temperature, relative humidity, precipitation, cloud cover, 850-hPa geopotential height, sea-level pressure tendency, wind speed and wind direction) could explain up to 50% of the daily PM2.5 variability in the US. Hence, me3 teorological factors play an important role in air pollutant concentrations, also making them difficult to model.

Most current air quality forecasting uses straightforward approaches like box models, Gaussian models and linear statistical models. Those models are easy to implement and allow for the rapid calculation of forecasts. However, they usually do not describe the interactions and non-linear relationship that control the transport and behaviour of pollutants in the atmosphere with these challenges, machine learning methods originating from the field of artificial intelligence have become popular in air quality forecasting and other atmospheric problems . For instance, several neural network (NN) models have already been used for air quality forecast, in particular for forecasting hourly averages and daily maximum . Although NN have advantages over traditional statistical methods in air quality forecasting, NN-based models still need to improve in order to achieve good prediction performance as effectively and efficiently as possible. A number of difficulties associated with NN hamper their effectiveness in air quality

forecasting. These difficulties include computational expense, multiple local minima during optimization, over-fitting to noise in the data, etc. Furthermore, there are no general rules to determine the optimal size of network and learning parameters, which will greatly affect the prediction performance.

Another key consideration of forecast models is their updatability when doing realtime forecasting. For a forecast model, recently observed data should be used to refine the model. This generally follows a procedure that links the discrepancy between model forecasts and the corresponding latest observation to all or some of the parameters in model. Normally there are two ways for model updating: batch learning and online learning. Whenever new data are received, batch learning uses the past data together with the new data and performs a retraining of the model, whereas online learning only uses the new data to update the model. Batch learning can be computationally expensive in real-time forecasting as the procedure means repeatedly altering a representative set of parameters calibrated over a long historical record. Linear models are generally easy to update online , and even with batch learning, linear models are fast and easy to implement. As for non-linear methods, true online learning is difficult for many formulations such as the non-linear kernel method.

1.2 Research Objectives

The research goal of this study is to develop a non-linear updatable model for real-time air quality forecasting, to potentially replace the updatable linear regression models currently being used. The ultimate goal is to improve air pollution forecasting in Canada and in other countries.

**Literature Review**

2.1 Introduction

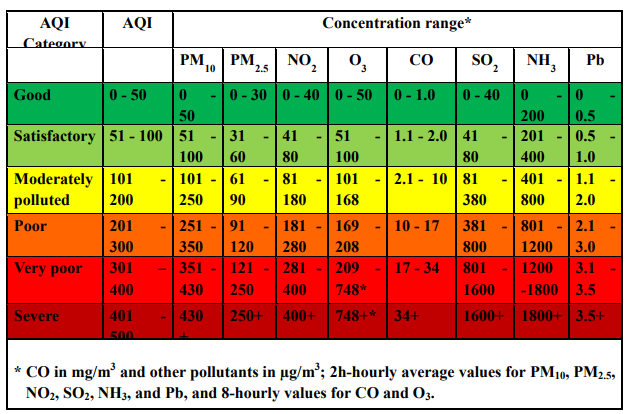
Air pollution is major threat to health and exerts a wide range of impacts on biological and economic systems. The purpose of this literature review is to justify the research objectives of this study in light of previous work by investigating past air quality prediction studies and determining where future research is needed. Literature related to air quality prediction and various types of machine learning methods used in this study are reviewed. Machine learning theory and past applications are examined to show why these methods are likely to perform well in air quality forecasting.

2.1.1 About National Air Quality Index

1. Air Quality Index is a tool for effective communication of air quality status to people in terms, which are easy to understand. It transforms complex air quality data of various pollutants into a single number (index value), nomenclature and colour.

2. There are six AQI categories, namely Good, Satisfactory, Moderately polluted, Poor, Very Poor, and Severe. Each of these categories is decided based on ambient concentration values of air pollutants and their likely health impacts (known as health breakpoints). AQ sub-index and health breakpoints are evolved for eight pollutants (PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb) for which short-term (upto 24-hours) National Ambient Air Quality Standards are prescribed.

3. Based on the measured ambient concentrations of a pollutant, sub-index is calculated, which is a linear function of concentration (e.g. the sub-index for PM2.5 will be 51 at concentration 31 µg/m3 , 100 at concentration 60 µg/m3 , and 75 at concentration of 45 µg/m3 ). The worst sub-index determines the overall AQI. AQI categories and health breakpoints for the eight pollutants are as follow:

  
 Fig.1. Air Quality Index

2.2 Machine Learning Techniques

Machine learning is a major sub-field in computational intelligence (also called artificial intelligence). Its main objective is to use computational methods to extract information from data. Machine learning has a wide spectrum of applications including handwriting and speech recognition, robotics and computer games, natural language processing, brain-machine interface and so on. In the environmental sciences, machine learning methods have been heavily used in data processing, model emulation, weather and climate prediction, air quality forecasting, oceanographic and hydrological forecasting.

2.3 Regression Models

## 2.3.1 What is Regression Model?

Regression analysis is a form of predictive modeling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

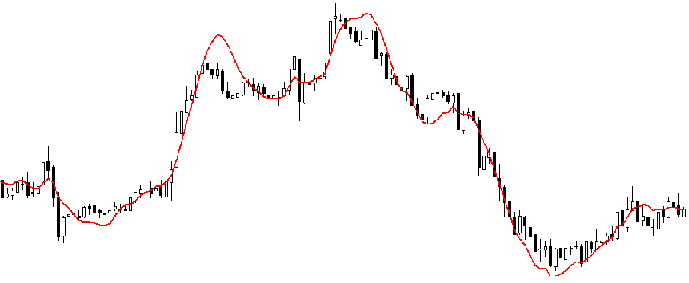


Fig.2.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. I’ll explain this in more details in coming sections.

## 2.3.2 Why do we use Regression Model?

As mentioned above, regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

Let’s say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

## 2.3.3 How many types of regression techniques do we have?

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We’ll discuss them in detail in the following sections.

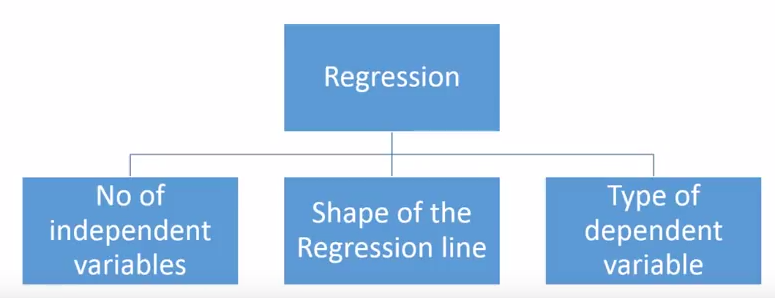


Fig.3.

For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven’t used before. But before you start that, let us understand the most commonly used regressions:

Basic Regression Models which are used in this project is discussed below-

A. Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

yi​=β0​+β1​xi1​+β2​xi2​+...+βn​xin​+ϵ

where, for i= *n* observations:

*yi*​=dependent variable

*xi*​=independent variables

*β*0​=y-intercept (constant term)

*βn*​=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals) ​

A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

* There is a linear relationship between the dependent variables and the independent variables.
* The independent variables are not too highly correlated with each other.
* yi observations are selected independently and randomly from the population.
* Residuals should be normally distributed with a mean of 0 and variance *σ.*

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

B. Support Vector Regression

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample.

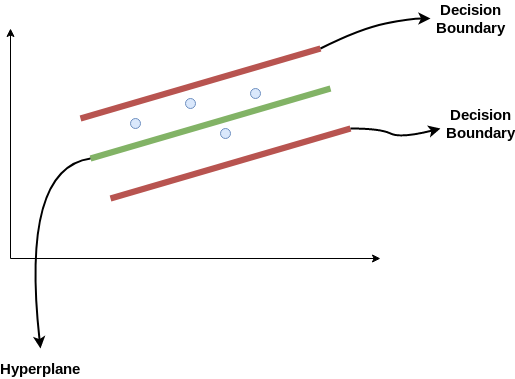
[](https://cdn.analyticsvidhya.com/wp-content/uploads/2020/03/SVR1.png)

Fig.4. Support Vector Regression Working Functionality

Consider these two red lines as the decision boundary and the green line as the hyper plane. Our objective, when we are moving on with SVR, is to basically consider the points that are within the decision boundary line. Our best fit line is the hyperplane that has a maximum number of points. The first thing that we’ll understand is what is the decision boundary (the danger red line above!). Consider these lines as being at any distance, say ‘a’, from the hyperplane. So, these are the lines that we draw at distance ‘+a’ and ‘-a’ from the hyper plane. This ‘a’ in the text is basically referred to as epsilon.

Assuming that the equation of the hyperplane is as follows:

Y = wx + b (equation of hyper plane)

Then the equations of decision boundary become:

wx + b = +a

wx + b = -a

Thus, any hyperplane that satisfies our SVR should satisfy:

-a < Y- wx+b < +a

Hence, we are going to take only those points that are within the decision boundary and have the least error rate, or are within the Margin of Tolerance. This gives us a better fitting model

C. Decision Tree Regression

The decision tree is the simplest, yet the most powerful algorithm in machine learning. Decision tree uses a flow chart like tree structure to predict the output on the basis of input or situation described by a set of properties. It falls under the category of supervised learning in machine learning and works for:

* Categorical output problem
* Continuous output problems.

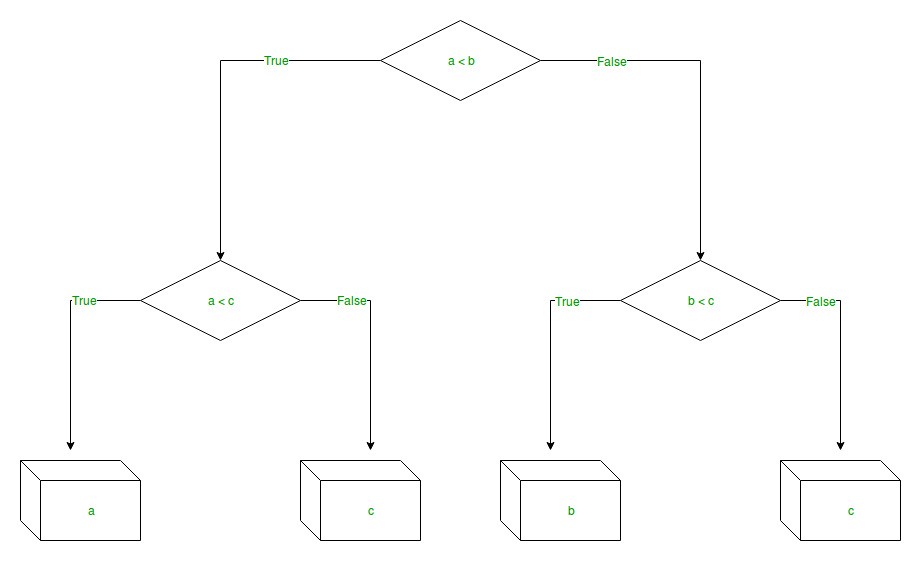


Fig.5. Working process of DTR

In a decision tree we have:

* + Nodes, which represent a condition. On the basis of the condition we go to either left(true) or right(false) child of the node as shown in Fig. 3.
  + Leaves or end nodes, which represents the output of the algorithm.

focusing on decision tree regression, So, decision tree regression is used for the continuous output problem. Continuous output means the output of the result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values. Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

3. Data

The concentrations of several criteria pollutants can be included in the air pollution index (API), which is easy to be understood by the public. API and air quality index (AQI) have been used to obtain and convey information regarding the urban air quality in a number of countries and zones worldwide.

The World Air Quality Index system will take care of regularly checking the data from the feed, and each time an update is available, it will be processed, converted to the US EPA scale AQI values, and publish on the World Air Quality Index website within minutes.

Also, although only PM2.5, PM10, NO2, NH3,SO2, CO and Ozone. Air Quality data is published, the system does collect more pollutants for forecasting purpose: Benzen, Toluene, Ethyl benzene, NOx, THC, NMHC, PM1, Formaldehyde, Mercury, Ammonia, Methane, Hydrogen sulfide, Nitrous acid, Phenol, Naphthalene, paraxylene (p-Xylen), metaxylene (m-Xylen), etc..

It is also possible to publish meteorological data: Temperature, Atmospheric Pressure, Humidity, Precipitation, Wind Speed, Wind Direction, Solar Radiation and UVI.

Now at first we will create a CSV file using python with our chosen parameter and check how many null values there. We have to remove null values and fill at this position mean values.

* 1. Data Collection

We can easily collect dataset from difference website. Here, We have to collected the dataset from Central Control Room for Air Quality Management and its website links is , <https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-data-availability>

|  |  |
| --- | --- |
| AQI Station Name | AQI Station Address |
| Anand Vihar | Anand Vihar delhi DPCC  (Chaudhary Charan Singh Marg, Isbt Anand Vihar, Anand Vihar, Delhi – 110092) |

We have chosen at Anand Vihar, Delhi DPCC Air Monitoring station for working our prediction.

.

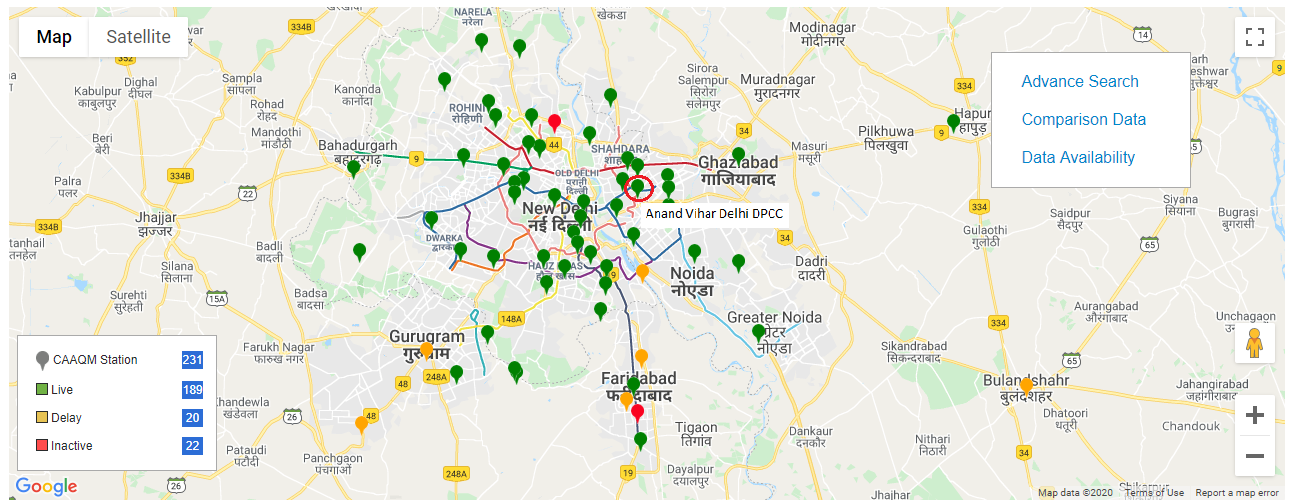


Fig.6. Air Pollution in Delhi: Real-time Air Quality Index Visual Map

3.2 Creating Data Set:

At first, the lists of the primary pollutants to be considered are decided. From the Central Control Room for Air Quality

Management the air pollution data are gathered for each hour from the following location:

 Station Name: Anand Vihar, Delhi – DPCC

 Address: Delhi, Delhi

 Latitude: 28.646835, Longitude: 77.316032

Only 8 pollutants have been chosen among the 19 pollutants available in the dataset provided by the Central Control

room for Air Quality Management, as these 8 pollutants are considered to calculate the AQI as mentioned in the India

Environment Portal. In India Environment Portal the breakpoints of the pollutants, within scale 0-500, are clearly

mentioned which are used to calculate the AQI.

The 8 pollutants or the attributes of the dataset are PM2.5, PM10, NO2, NH3, SO2, CO, O3, C6H6.

After creating the one year dataset in excel format, we have to import CVS file then its look like that . . . . . . .

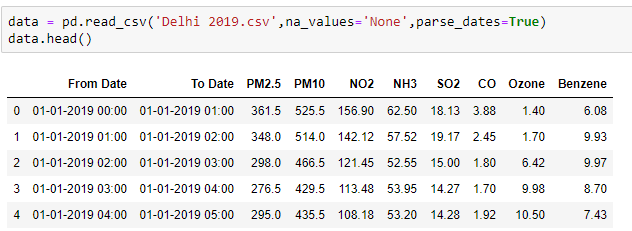


Fig.7. Dataset

3.2.1 Calculate sub index

The seven pollutants which are chosen for this project are Particulate Matter (PM2.5), Particulate Matter (PM10), Nitrogen Dioxide (NO2), Ammonia(NH3), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Ozone (O3). Calculating sub indexes of each and every pollutants individually by using the breakpoints of every pollutant.

After constructing the dataset, the sub index of every pollutant has been calculated by the equation

**Ip = [{(IHI - ILO) / (BHI -BLO)} \* (Cp-BLO)] + ILO**

where,

Ip= The Sub index of a specific pollutant concentration Cp.

BHI= Breakpoint concentration greater or equal to given concentration.

BLO= Breakpoint concentration smaller or equal to given concentration.

IHI =AQI value corresponding to BHI

ILO = AQI value corresponding to BLO

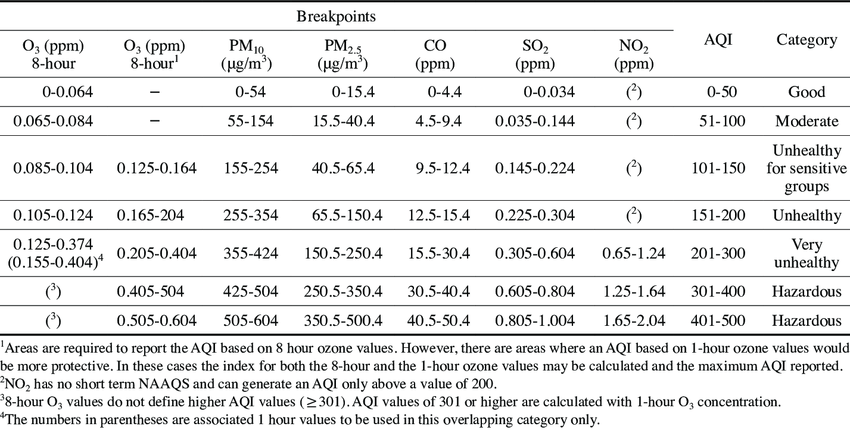
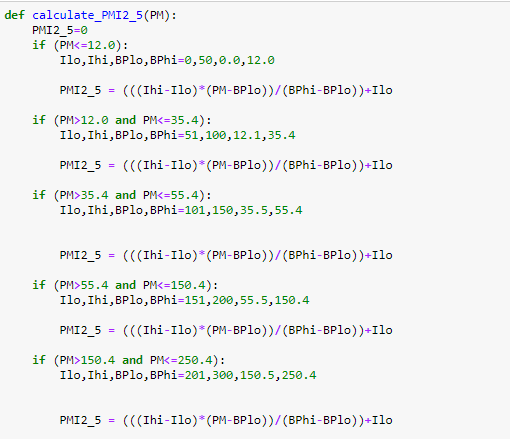


Fig.8. Breakpoints

* PMI2.5

Particulate Matter (PM) is a mixture of solid and liquid particles that are suspended in the air. These are categorized into coarse, fine and ultrafine. Coarse particles have a diameter of 2.5 micrometres to 10 micrometres (about 25 to 100 times thinner than a human hair), are relatively heavier and thus tend to settle. Dust, spores and pollen are some examples. PM2.5 refers to particles that have diameter less than 2.5 micrometres (more than 100 times thinner than a human hair) and remain suspended for longer.



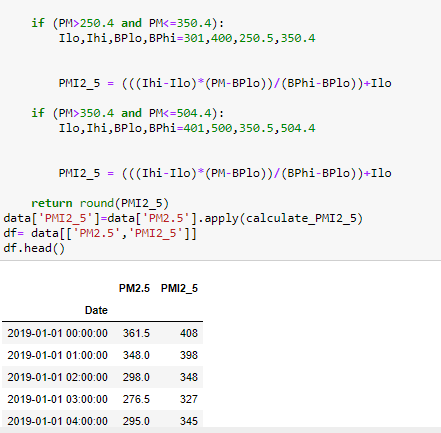
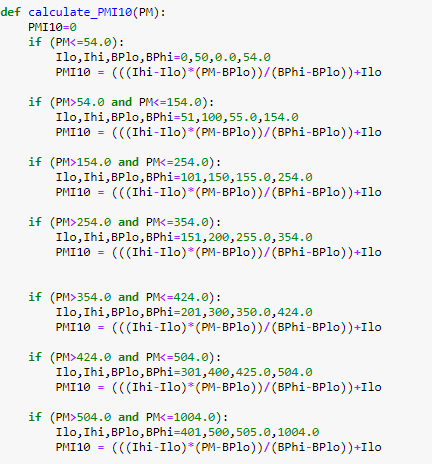


Fig.9. PMI 2.5

### PM10

PM10  is particulate matter 10 micrometers or less in diameter, PM2.5 is particulate matter 2.5 micrometers or less in diameter. ... By way of comparison, a human hair is about 100 micrometres, so roughly 40 fine particles could be placed on its width.



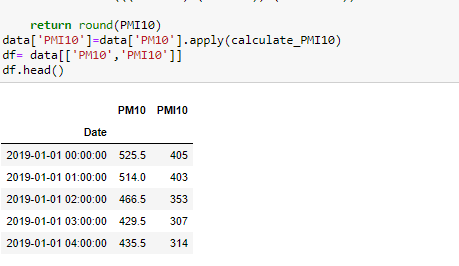
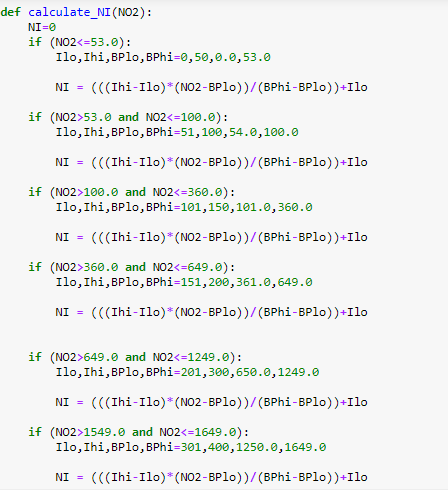


Fig.10. PM10

* NO2

Nitrogen Dioxide (NO2) is one of a group of highly reactive gases known as oxides of nitrogen or nitrogen oxides (NOx). Other nitrogen oxides include nitrous acid and nitric acid. NO2 is used as the indicator for the larger group of nitrogen oxides.

NO2 primarily gets in the air from the burning of fuel. NO2 forms from emissions from cars, trucks and buses, power plants, and off-road equipment.



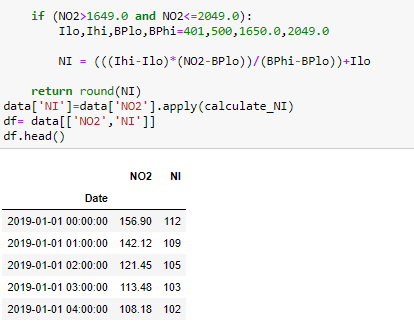


Fig. NO2

* NH3

**Ammonia (NH3)**, colorless, pungent gas composed of nitrogen and hydrogen. It is the simplest stable compound of these elements and serves as a starting material for the production of many commercially important nitrogen compounds.



Fig.12. NH3

* SO2

**Sulfur dioxide** is a colorless, water-soluble gas that forms when sulfur burns. Sulfur dioxide is not flammable, although it does transition to sulfur trioxide when exposed to an oxygen environment. It's poisonous and has a strong odor. Sulfur dioxide occurs naturally during volcanic activity and is produced as a waste gas when sulfur-containing materials are heated, such as the burning of coal or oil, or the smelting of metal. It's a significant pollutant and currently considered an ecological threat.

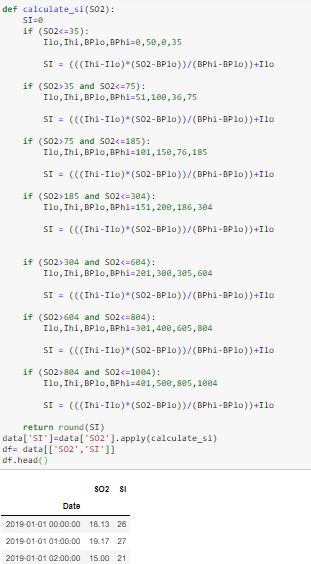


Fig.13. SO2

* CO

Carbon monoxide is a gas and is found in air. High levels of carbon monoxide are poisonous to humans and, unfortunately, it cannot be detected by humans as it has no taste or smell and cannot be seen.

The natural concentration of carbon monoxide in air is around 0.2 parts per million (ppm), and that amount is not harmful to humans. Natural sources of carbon monoxide include volcanoes and bushfires.

The main sources of additional carbon monoxide are motor vehicle exhaust and some industrial activities, such as making steel.

Tobacco smoke is one of the main indoor sources of carbon monoxide.

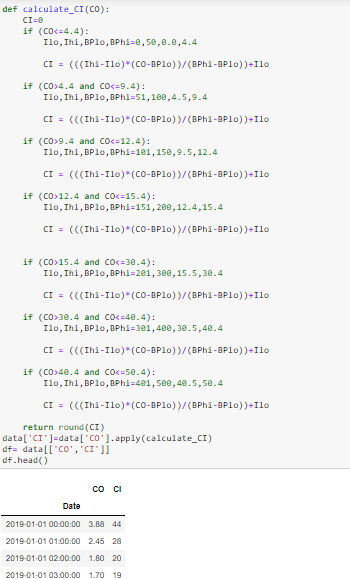


Fig.14. CO

* Ozone

Ozone is a gas that is formed when nitrogen oxides react with a group of air pollutants known as 'reactive organic substances' in the presence of sunlight. (See fact sheet on air toxics.) The chemicals that react to form ozone come from sources such as: motor vehicle exhaust, oil refining, printing, petrochemicals, lawn mowing, aviation, bushfires and burning off. Motor vehicle exhaust fumes produce as much as 70% of the nitrogen oxides and 50% of the organic chemicals that form ozone.

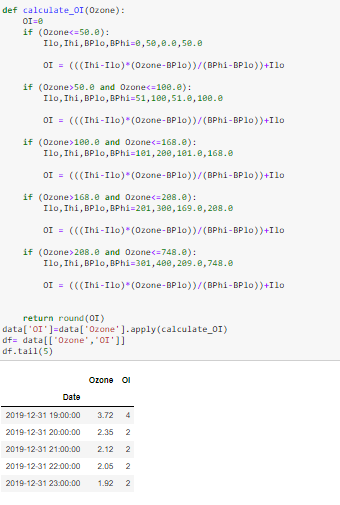


Fig.15. Ozone

* Now the full dataset with all the sub-indices looks like-

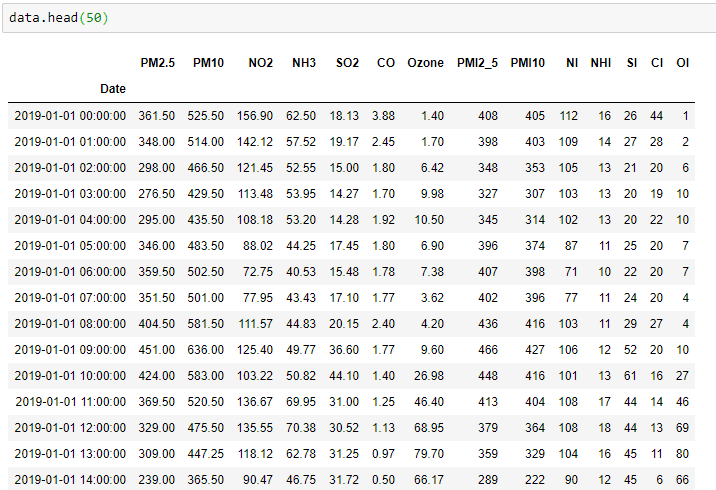


Fig.16. Full Dataset with all sub-indices

3.3 Calculating AQI

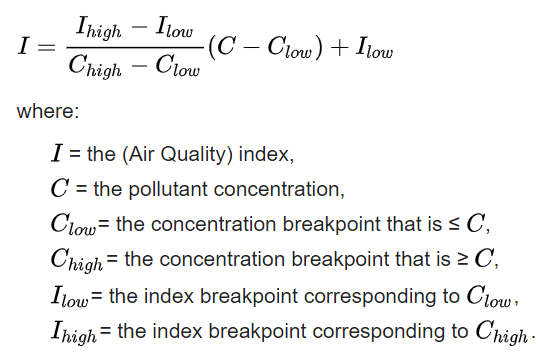
An individual score (Individual Air Quality Index, IAQI) is assigned to each pollutant and the final AQI is *the highest* of these six scores. The final AQI value can be calculated either per hour or per 24 hours. The concentrations of pollutants can be measured quite differently.

**3.3.1 How to calculate the AQI:**

The air quality index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI this equation is used:

{\displaystyle I={\frac {I\_{high}-I\_{low}}{C\_{high}-C\_{low}}}(C-C\_{low})+I\_{low}}

Where:



* After calculating all the sub-indices , now the final dataset can be created using these sub-indices of all the pollutants.

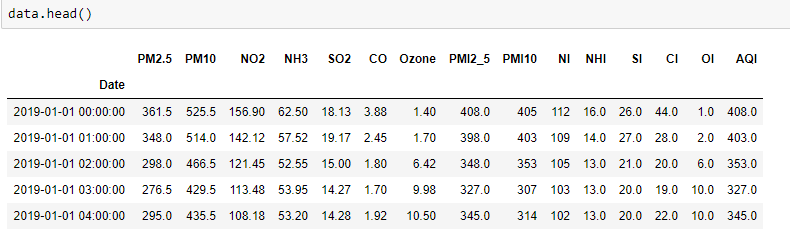


Fig.17. Dataset after calculate the sub-indices

{\displaystyle I}

* Once the sub-indices are formed, AQI can be calculates as AQI= Max (Ip) as shown in below figure:

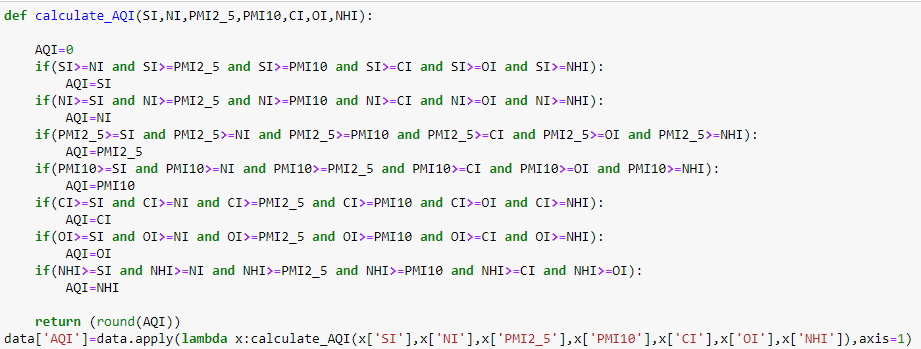


Fig.18. Calculating AQI with the with sub-indices

3.4 Data Correlation

**Correlation Matrix**is basically a covariance matrix. Also known as the auto-covariance matrix, dispersion matrix, variance matrix, or variance-covariance matrix. It is a matrix in which i-j position defines the correlation between the ith and jth parameter of the given data-set.

When the data points follow a roughly straight-line trend, the variables are said to have an approximately linear relationship. In some cases, the data points fall close to a straight line, but more often there is quite a bit of variability of the points around the straight-line trend. A summary measure called the correlation describes the strength of the linear association. Correlation summarizes the strength and direction of the linear (straight-line) association between two quantitative variables. Denoted by r, it takes values between -1 and +1. A positive value for r indicates a positive association, and a negative value for r indicates a negative association.  
The closer r is to 1 the closer the data points fall to a straight line, thus, the linear association is stronger. The closer r is to 0, making the linear association weaker.

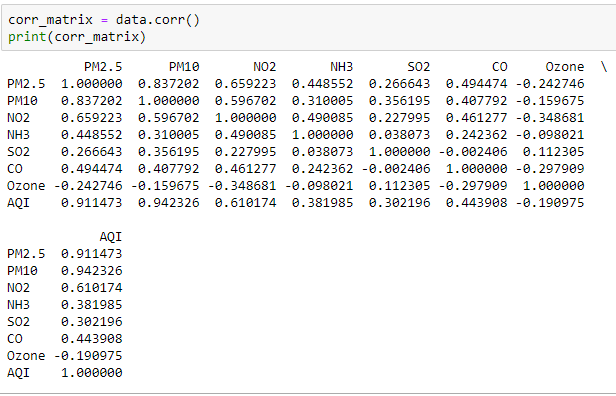


Fig.19. Data Correlation

* **In this dataset total numbers of rows are 8276.**
* **The following plot (Fig. 20.) shows that all the features that are considered for the prediction are correlated and thus can be considered to train the model.**

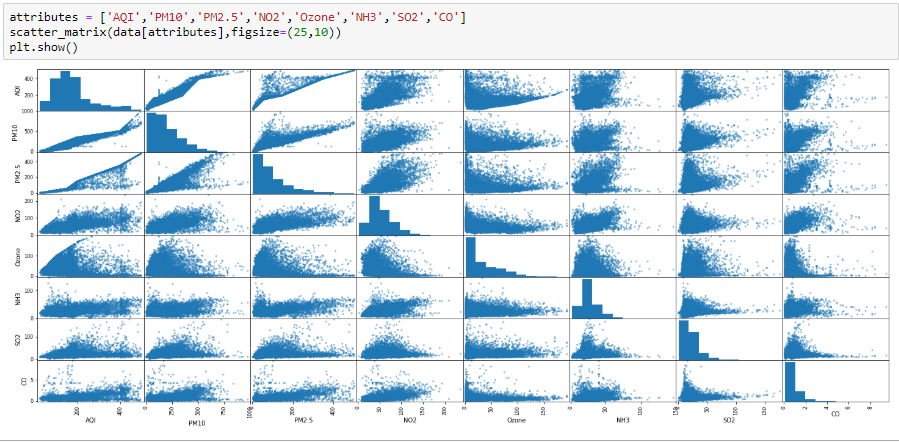


Fig.20. Relation among all attributes

* **Data splitting was done as 80% for training and 20% for testing in this work.**

3.5 Training Testing Splitting

Train/Test is a method to measure the accuracy of your model.

It is called Train/Test because you split the the data set into two sets: a training set and a testing set.

80% for training and 20% for testing.

We train the model using training set.

We test the model using testing set.

Train the model means create the model.

Test the model means test the accuracy of the model.

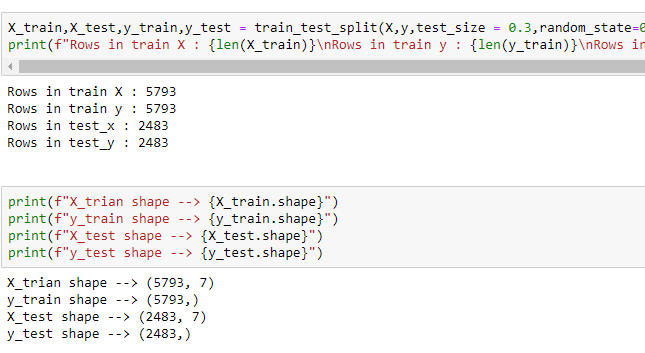


Fig.21. Training Test/Splitting

4. Methods and Models Set up

In this Session, several different methods and models are introduced. Each method produces a different model per station per forecast hour per pollutant. The model development can be separated into two phase: 1) training and 2) testing and updating. Models are first trained and validated using 1-year data sets (01/01/2019-28/12/2019). Model updating is conducted by either batch learning algorithm or an online-sequential learning algorithm from the newly arrived data.

When data become available, batch learning performs a complete retraining of the model using all past data plus the new data. It can be used to update the multiple linear regression (MLR), multi-layer perceptron neural network (MLP NN) and extreme learning machine (ELM) methods. Depending on computation resources, batch updating can be applied daily, monthly or seasonally. Batch learning can be computationally intensive for nonlinear models as it may involve many iterations through the training data. There are many applications where online-sequential learning algorithms are preferred over batch learning algorithms as sequential learning algorithms do not require retraining with the full dataset whenever new data arrive .

A versatile online-sequential learning algorithm means the data for training are sequentially presented to the learning algorithm. At any time, only the newly arrived data are needed to update the model. The new data, once learned by the model, can be discarded. The machine learning algorithm has no prior knowledge as to how many training dataset will be presented.

4.1 Multiple Linear Regression (MLR)

Continuous output Multiple linear regression(MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression(MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

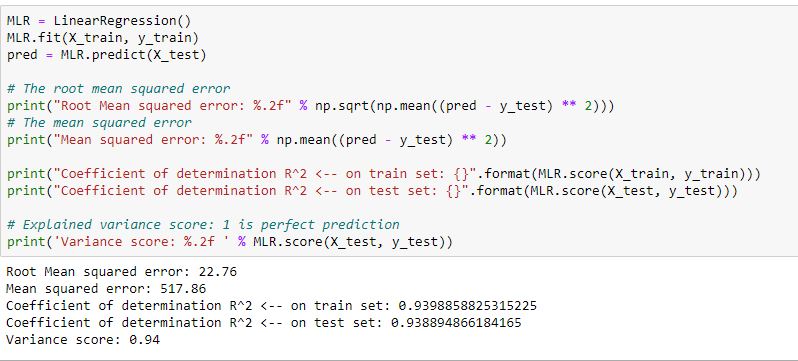


Fig.22. Multiple Linear Regression Code

4.2 Support Vector Regression

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample.

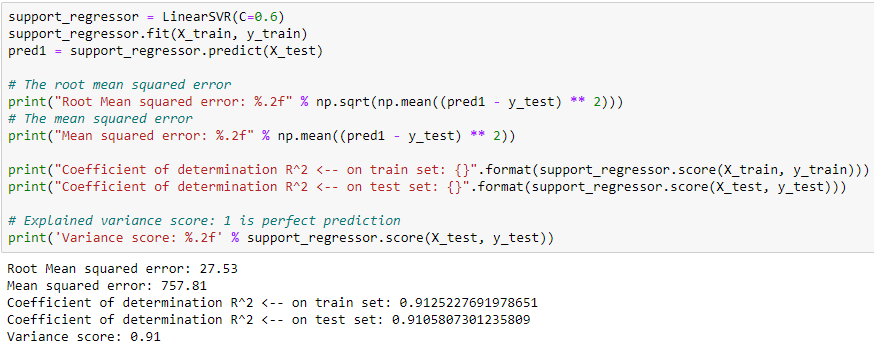


Fig.23. Support Vector Regression Code

4.3 Decision Tree Regression

The decision tree is the simplest, yet the most powerful algorithm in machine learning. Decision tree uses a flow chart like tree structure to predict the output on the basis of input or situation described by a set of properties.

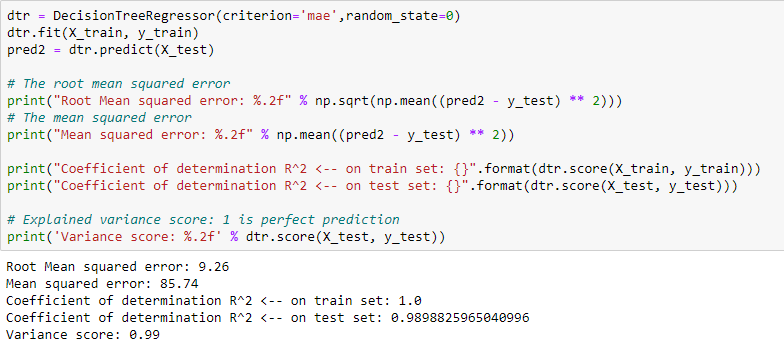
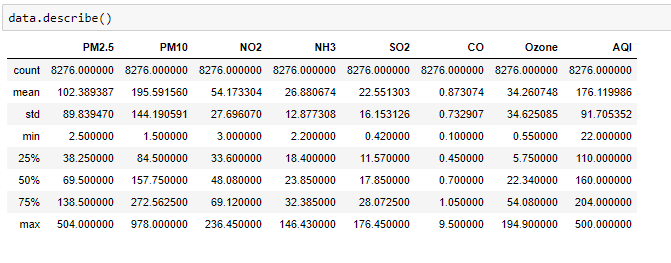


Fig.24. Decision Tree Regression Code

* **After doing all this steps, final dataset has been created using 7 pollutants as shown in Fig. 25.**

 Fig.25. Final Dataset

4.4 Pre-processing and Feature Selection

Here only studied and applied algorithms on the data of the particular station of Delhi. Data pre-processing has done (example-removing null values, filling null values, etc.) and has used Standard scalar as Feature Scaling, as shown in Fig.26.

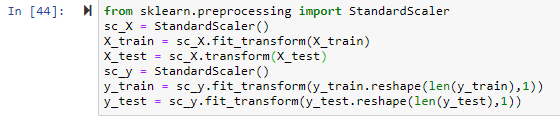


Fig.26. Snippets for feature selection

Standardize features by removing the mean and scaling to unit variance

The standard score of a sample x is calculated as:

z = (x - u) / s

where u is the mean of the training samples or zero if with mean=False, and s is the standard deviation of the training samples or one if with std=False.

Below are the following table of each Model prediction details-

4.5 Result:

1. ***AQI prediction using Multiple Linear Regression***

|  |  |  |
| --- | --- | --- |
| RMSE | MSE | Score |
| 22.76 | 517.86 | .94 |

***2. AQI prediction using Support Vector Regression***

|  |  |  |
| --- | --- | --- |
| RMSE | MSE | Score |
| 30.61 | 936.69 | .89 |

***3. AQI prediction using Decision Tree Regression***

|  |  |  |
| --- | --- | --- |
| RMSE | MSE | Score |
| 9.26 | 85.74 | .99 |

4.6 Algorithm Comparison

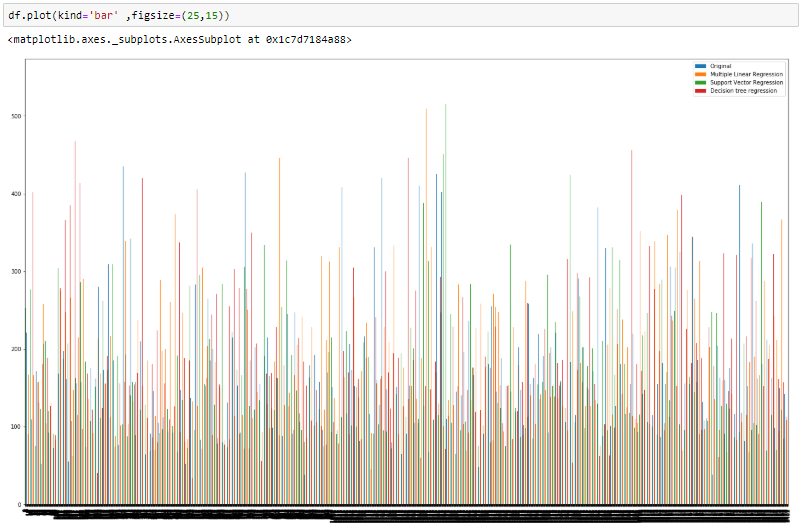
****

Fig.27. Comparison of all models using histogram

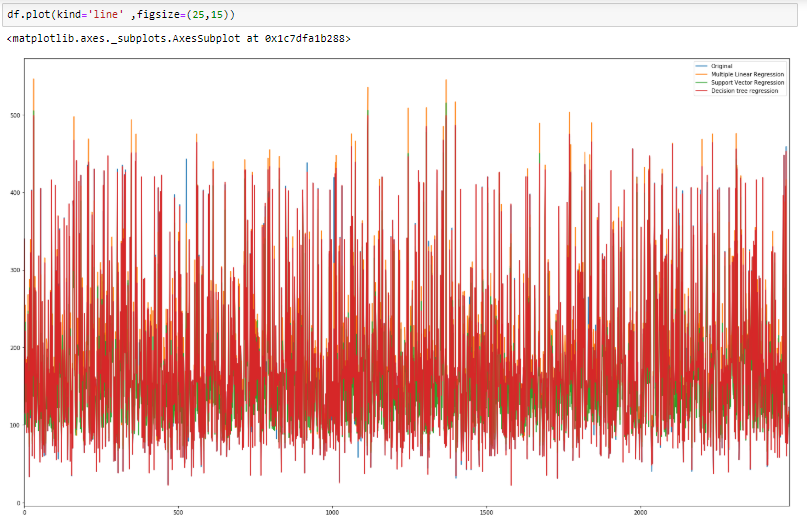
****

Fig.28. of all models using Line plot

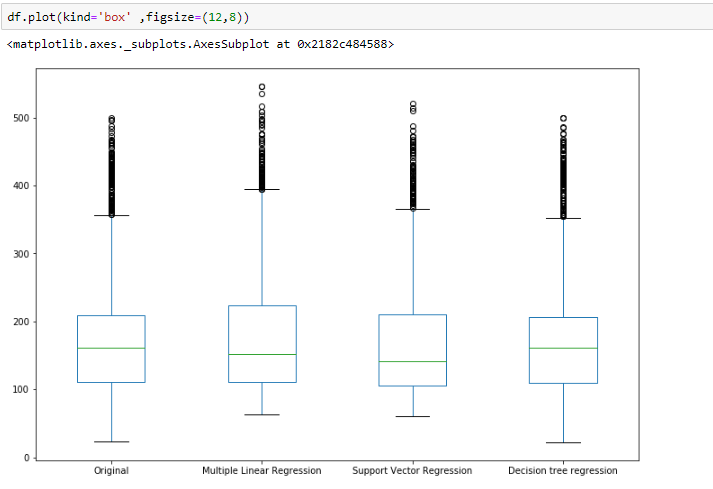


Fig.29. Comparison of all models using box plot

As shown in the Fig.8., when compared to other machine learning models applied on the data set Decision Tree Regression (DTR) suits the best for this system with the mean accuracy and standard deviation accuracy to be .99 (99%). It also can be seen in the above figure that the original value and the predicted values of DTR are very close. Hence, Decision Tree Regression can be used to predict the AQI value generated based on the given sample of Air Quality.

5. CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it. The results show that machine learning models (Decision Tree Regression) can be efficiently used to detect the quality of air and predict the level of AQI in the future.

The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

6. References

[1] Air Pollution, WHO, URL https://www.who.int/health-topics/air-pollution#tab=tab\_1

[2] CCR,2020, Central Control Room for Air Quality Management. URL: https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing

[3] Central Pollution Control Board, October 2014. Air Quality Index. URL: http://www.indiaenvironmentportal.org.in/files/file/Air%20Quality%20Index.pdf

[4] Z. Qi, T. Wang, G. Song, W. Hu, Xi Li, Z. Zhang, 2018, Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality

[5]T. C. Bui, V. Duc. Le, S. K, Cha,. A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM

[6] M. Chen, J. Yang, L. Hu, M. S. Hossain, G. Muhammad, 2018, Urban Healthcare Big Data System based on Crowdsourced and Cloud based Air Quality Indicators, IEEE COMMUNICATIONS MAGAZINE

[7]H. Liu, Q. Li, D. Yu, Yu Gu , 2019, Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms, MDPI Journal[8] P. P. Raj, S. Dewangan, S. Harne, 2019, Prediction and Optimization of Air Pollution-A Review Paper, International Journal for Research in Applied Science & Engineering Technology (IJRASET)

[9] S. Nemade, C. Mankar, 2019,  A Survey on Different Machine Learning Techniques for Air Quality Forecasting for Urban  
Air Pollution, IJRASET

[10] K. Mahesh Babu, J. Rene Beulah, 2019, Air Quality Prediction based on Supervised Machine Learning Methods, IJITEE

[11]A.G. Soundari,  J. G. Jeslin, Akshaya A.C, 2019, Indian Air Quality Prediction and Analysis using Machine Learning, IJAER

[12] K. Veljanovska, A. Dimoski, 2018, Air Quality Index Prediction Using Simple Machine Learning Algorithms, IJETTCS

[13] Aditya C. R, C. R Deshmukh, Nayana D K, P. G. Vidyavastu, 2018, Detection and Prediction of Air Pollution using Machine Learning Models, IJETT.

[14] B. Ghose, Z. Rehena, 2019, A Mechanism for Air Health Monitoring in Smart City using Context Aware Computing,. CoCoNet’19.

[15] B. Ghose, Z. Rehena, A Deep Learning Approach for Predicting Air Pollution  in Smart Cities.