**Prediction of Air Quality Using Regression Models**

*A Project*

*Submitted in partial fulfillment for the*

*Award of the degree of*

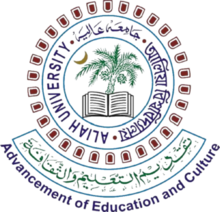
**Master of Computer Applications**

***Submitted by***

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

*Under the Guidance of*

**Dr. Zeenat Rehena**

****

**ALIAH UNIVERSITY,**

**NEWTOWN CAMPUS, KOLKATA-156,**

**WEST BENGAL**

**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 right.

|  |
| --- |
| NAME : Julfikar Sk |
| Roll No : MCA173010 |
| NAME : Shahajahan Ahamed |
| Roll No : MCA173020 |
|  |
| DATE : 03-07-2020 |
| PLACE : KOLTATA |

**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**). 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**). Aliah 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 teacher guide **Dr. Zeenat Rehena** for her guidance and support without which this task would not have been accomplished. I would like to thank my co-guide **Shahajahan Ahamed** if any for his constant and timely help, moral support and valuable suggestions. I also 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 : 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 teacher guide **Dr. Zeenat Rehena** for her guidance and support without which 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 &amp; 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 system 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 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.

|  |
| --- |
| Table of Contents |
|  |
| Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ii |
| Preface . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iii |
| Table of Contents . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iv |
| List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vi |
| List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vii |
| List of Abbreviations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xi |
| 1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1 1.1 |
| Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1 1.2 Research Objectives . . . . . . . . . . . . . . . . . . . . . . . . . . 1.3 Organization of Thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . |
| 2 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2.1 |
| 2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2.2 |
| 2.2 Machine Learning Techniques . . . . . . . . . . . . . . . 2.2.1 |
| 2.3 Neural Network . . . . . . . . . . . . . . . . . . . . . . . . . . . 2.2.2 |
| Extreme Learning Machine (ELM) . . . . . . . . . . . . 2.3 |
| Air Quality Forecasting Models . . . . . . . . . . . . . . 2.3.1 |
| Dispersion Models . . . . . . . . . . . . . . . . . . . . . . . . 2.3.2 |
| Photochemical Models . . . . . . . . . . . . . . . . . . . . . 2.3.3 |
| Regression Models . . . . . . . . . . . . . . . . . . . . . . . 2.3.4 |
| Neural Network Models . . . . . . . . . . . . . . . . . . . . 2.4 |
| Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17 iv  3 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 |
| 3.1 Study Area . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 |
| 3.2 Data Set . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18 4 |
|  |
| 4 Methods and Models Set up . . . . . . . . . . . . . . . . . . . . . . . . . 22 |
| 4.1 Input Data Preprocessing . . . . . . . . . . . . . . . . . . . . . 23 4.2 Multiple Linear Regression (MLR) . . . . . . . . . . . . . 23 4.3 Online-Sequential Multiple Linear Regression (OS-MLR) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .24 |
| 4.4 Multi-layer Perceptron Neural Network (MLP NN) . 26 4.5 Extreme Learning Machine (ELM) . . . . . . . . . . . . . . 27 4.6 Online-Sequential Extreme Learning Machine (OS-ELM) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30 |
| 4.7 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . 31 |
| 4.7.1 Pearson Correlation Coefficient (r ) . . . . . . 31 |
| 4.7.2 Mean Absolute Error (MAE) . . . . . . . . . . . . 31 4.7.3 MAE/MAD . . . . . . . . . . . . . . . . . . . . . . . . . . 32 4.7.4 Root Mean Square Error (RMSE) . . . . . . . . . 32 4.7.5 Skill Score (SS) . . . . . . . . . . . . . . . . . . . . . . . 32 |
| 4.8 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .33 |
|  |
| 5 Results and Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34 |
| 5.1 Ozone . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34 5.2 PM2.5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48 5.3 NO2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61 5.4 Model Results with Antecedent Predictors . . . . . . . 74 |
|  |
| 6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84 |
|  |
| 7 Reference . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . |

Chapter 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 [4] 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. In [14], A mechanism has constructed for monitoring the air health in smart city using Context Aware Computing. In a research article [5], 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. In a research paper [15], a deep learning model has proposed to forecast the air pollution. In the article [9], compared four unpretentious machine learning algorithms, linear regression, Naive Bayes, support vector machine and random forest to predict air pollutants levels ahead of time. In [6], 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 another article [11], 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 [7], 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. In [10], 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 [13], Logistic regression is employed to simply detect whether a data sample is either polluted or not polluted. In [12], 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 (Chen et al., 2007; Brauer and Brook, 1997). 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 (EPA, 2005).

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

* 1. Background

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 (haze) (EPA, 2005).

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 (Dominick et al., 2012). Among these

1.1 Background

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

three factors, the one which most strongly influences variations in the ambient concentration of air pollutants is meteorological factors (Banerjee and Srivastava, 2009). Meteorological factors experience complex interactions between various processes such as emissions, transportation and chemical transformation, as well as wet and dry depositions (Seinfeld and Pandis, 1997; Demuzere et al., 2009). In addition, the spatial and temporal behavior of wind fields are affected by the surface roughness and differences in the thermal conditions (Oke et al., 1989; Roth, 2000), which further influence the dispersion of pollutants. For example, Revlett (1978) and Wolff and Lioy (1978) 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 (2012) 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 (Luecken et al., 2006). With these challenges, machine learning methods originating from the field of artificial intelligence have become popular in air quality forecasting and other atmospheric problems (Comrie, 1997; Hadjiiski and Hopke, 2000; Reich et al., 1999; Roadknight et al., 1997; Song and Hopke, 1996). For instance, several neural network (NN) models have already been used for air quality forecast, in particular for forecasting hourly averages (Kolehmainen et al., 2001; Perez et al., 2000) and daily maximum (Perez, 2001). 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 (Wang et al., 2003). A number of difficulties associated with NN hamper their effectiveness in air quality

1.2 Research Objectives \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 (Wilson and Vall´ee, 2002), 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.

Chapter 2

**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:

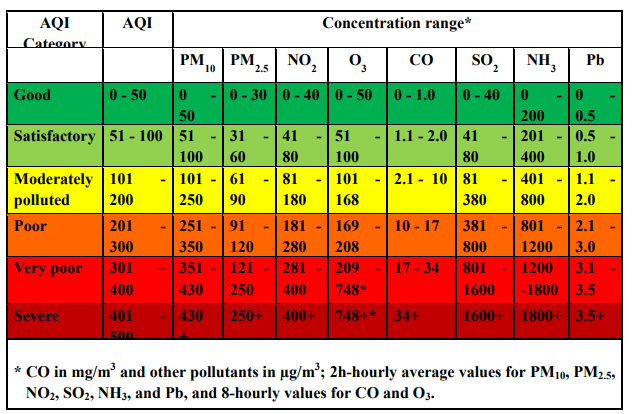


Fig: Air Quality Index

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. (Hsieh, 2009).

2.3 Air Quality Forecasting Models

An air quality model is a numerical tool used to describe the causal relationship between emissions, meteorology, atmospheric concentration,

2.3. Air Quality Forecasting Models

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

deposition and other factors. It can give a complete deterministic description of the air quality problem (Nguyen, 2014). The most commonly used air

quality models include dispersion models, photochemical models and regression models. Various neural network models, as non-linear regression models, have also been shown to be effective in air quality forecasting. In this section, different models and their applications will be introduced.

2.3.1 Regression Models

Both linear regression and non-linear regression models have been employed for air quality forecasting. The general purpose of a linear regression model

is to learn about the linear relationship between several independent variables (predictors) and a dependent variable (predictand).

Prybutok et al. (2000) built a simple linear regression model for forecasting the daily peak O3 concentration in Houston. The final model used four meteorological and O3 precursor parameters: O3 concentration at 9:00 a.m., maximum daily temperature, average NO2 concentration between 6:00 a.m. and 9:00 a.m. and average surface wind speed between 6:00 a.m. and 9:00 a.m. The correlation coefficient r of this model was 0.47. Chaloulakou et al. (1999) proposed a multiple regression model to forecast the next day’s hourly maximum O3 concentration in Athens, Greece. The set of input variables consisted of eight meteorological parameters and three persistence variables, which were the hourly maximum O3 concentrations of the previous three days. Testing this linear regression model on four separate test data sets, the mean absolute error (MAE) ranged from 19.4% to 33.0% of the corresponding average O3 concentrations.

Non-linear regression models are superior to simple linear regression models because they capture the non-linear relationships between air pollutant and meteorological parameters. Bloomfield et al. (1996) described a non-linear regression model to explain the effects of meteorology on O3 in the Chicago area. The model input variables consisted of a seasonal term, a linear annual trend term, and twelve meteorological variables. The observed ozone and meteorological data in 1981-1991 were divided into subsets for

2.3. Air Quality Forecasting Models

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

model development and validation. The model error were within ±5 ppb about half the time, and within ±16 ppb about 95% of the time. Bloomfield

et al. (1996) demonstrated that the meteorological data accounted for at least 50% of the ozone concentration variance. As the reference model in this thesis, the updatable model output statistics - air quality (UMOS-AQ) system applies multiple linear regression (MLR) to forecast air quality predictands. UMOS-AQ is a statistical post-processing system for air quality forecasting in Canada. The current Environment Canada (EC) operational AQ forecast model is the GEM-MACH15 (global environmental multi-scale model - modeling air quality and chemistry with 15-km grid spacing). GEM-MACH15 runs twice daily at 00 and 12 UTC to give 48-hour AQ forecasts (Anselmo et al., 2010). UMOS-AQ is based on post-possessing the GEM-MACH15 forecasts. The UMOS post-processing package has been used by EC to forecast meteorological predictands such as surface temperature and probability of precipitation since 1995 (Wilson and Vall´ee, 2002, 2003).

UMOS-AQ uses the existing UMOS framework and became operational in July 2010. Three predictands are currently considered by UMOS-AQ: O3, PM2.5 and NO2. Possible MLR predictors include O3, PM2.5 and NO2 hourly concentrations at a station for each hour of the previous day (i.e., persistence) plus 84 other chemical, meteorological, and physical predictors (e.g., solar flux, sine of scaled Julian day). Two seasons (summer and winter) are considered with a transitional period of 6 weeks. A minimum of 250 observation-model pairs per season are needed to generate robust MLR equations and the equations are regenerated with the latest model data every week (Moran et al., 2014). One of UMOS-AQ’s main advantages is its ability to adapt to the model changes, as its equations are updated four times per month. However, UMOS-AQ can only be constructed for locations where historical AQ measurements are available (Wilson and Vall´ee, 2002; Moran et al., 2014). This becomes a limitation because most AQ stations are not co-located with public weather forecast stations. A solution is to blend the UMOS-AQ point forecasts with GEM-MACH15 gridded forecast fields. This is now being done by optimal interpolation (OI) using MIST (Moteur d’Interpolation STatistique), an EC statistical interpolation package that uses the OI algorithm described by Mahfouf et al. (2007).

2.4 Summary

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.4 Summary

Studies from the fields of machine learning and air quality models show that much effort has been put into air quality forecasting, including the use of various machine learning methods. Machine learning methods have been widely used in environmental science problems and the applications of the MLP NN tend to provide some advantages over linear methods based on the results of the previous studies. In air quality forecasting, machine learning methods are promising when compared with the linear regression model and the photochemical dispersion model. The ELM method has been introduced to overcome some of the drawbacks in the popular MLP NN model, e.g. in computing time and the local minima problem.

Chapter 3

Data

3.1 Study Area

The updatable model output statistics - air quality (UMOS-AQ) model uses observations from more than 36 stations across Delhi . The stations belong to the National Air Pollution Surveillance Network (NAPS), where each station measures all or a combination of the concentrations of fine particulates (PM2.5) and (PM10) and nitrogen dioxide (NO2), ammonia(NH3), sulfur dioxide(SO2), Carbon Monoxide(CO) and ozone (O3) . 36 stations across delhi are used for model testing: As per **the** NAMP of CPCB, manual **air pollution monitoring** is carried out at Sarojini Nagar, Chandni Chowk, Mayapuri Industrial Area, Pritampura, Shahadra, Shahzada Bagh, Nizamuddin, Janakpuri, Siri Fort, and at ITO as traffic

intersection **station** across **the Delhi**. These ten stations include the largest AQI monitoring station of Delhi. They all have different topography, weather conditions and major pollution sources. We have been mentioned real time air quality index[Fig: 3.1] .

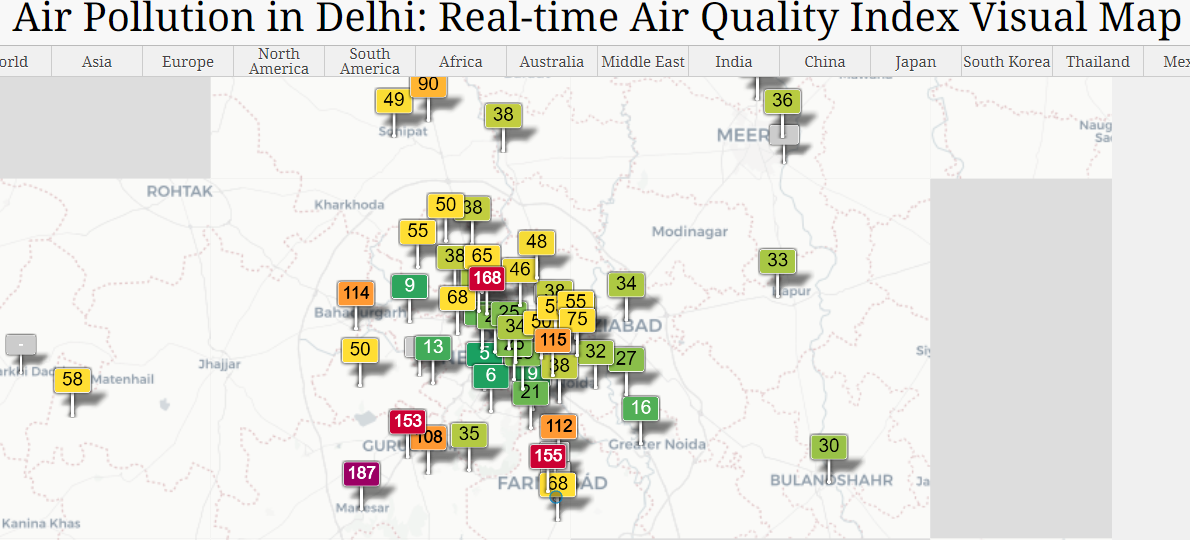
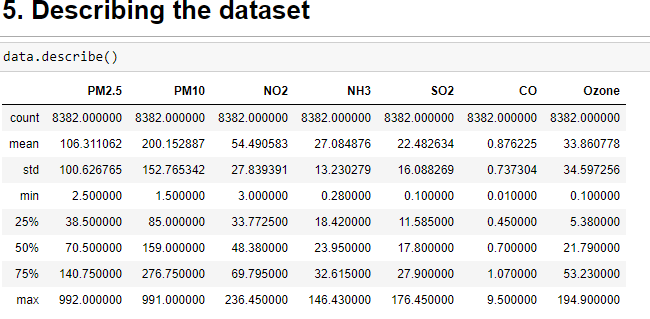


Fig 3.1: Air Pollution in Delhi: Real-time Air Quality Index Visual Map

3.2 Data Set

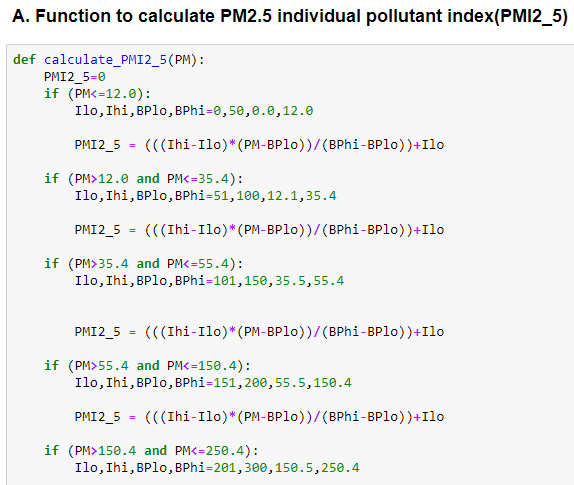
The data set used in this study covers the period on 01/01/2019 and was provided by UMOS-AQ model of Environment Delhi. In this years of data were for model training and validation and the finally we had been used for model testing as well as model updating. As mentioned before, UMOS-AQ is a post possessing system that combines multiple sources of information: AQ forecasts, meteorological forecasts, AQ measurements and physical variables. Hence, the input data sets consisted of observational and numerical data. The observational air pollutant data were from automated near-real-time (NRT) hourly reports of local PM2.5,PM10 and NO2 , NH3, SO2, CO,C3 concentrations from around 36 urban and rural AQ measurement stations located across Delhi. See [Fig: 3.2] the data set after removing null values and fill the mean values

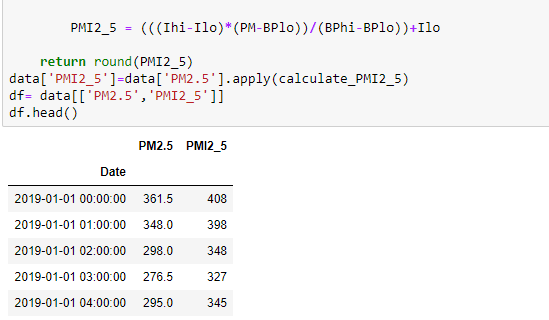


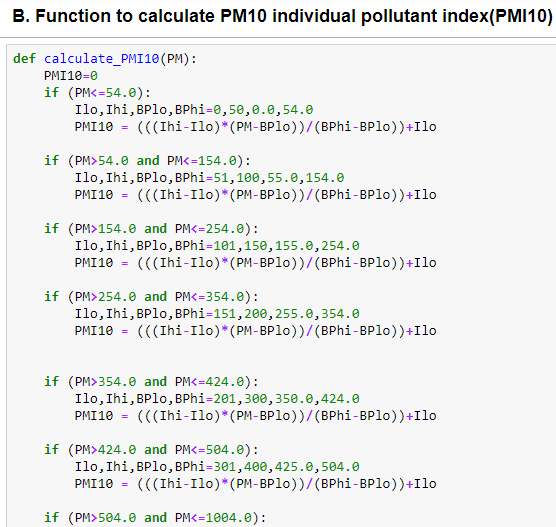
Fig[3.2]: Dataset

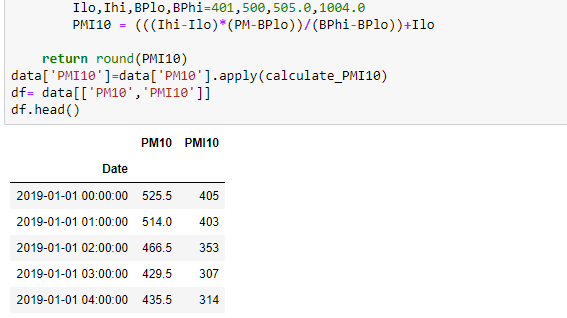
Creating The Dataset by Calculating AQI:

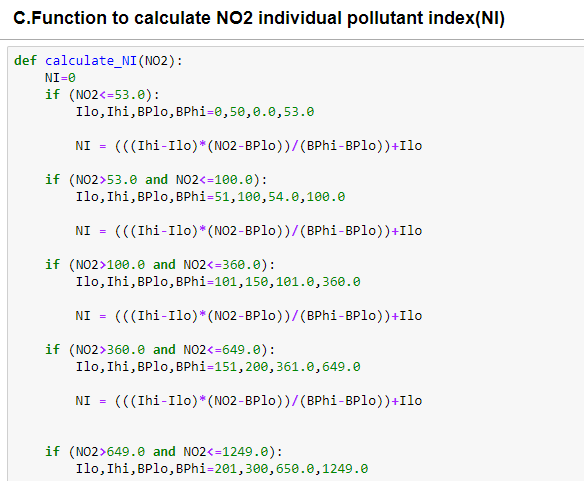
The seven “criteria pollutants” are fine particulate matter (PM2.5), PM10, nitrogen dioxide (NO2), ammonia(NH3), sulfur dioxide (SO2), carbon monoxide (CO), ground level ozone (O3). We have been working at these parameter. Calculating sub index of each and every pollutants and now we will calculate all the pollutants indivisually.

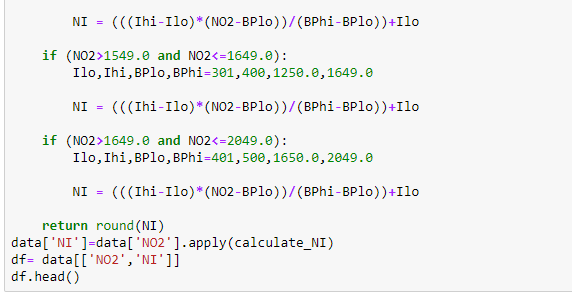


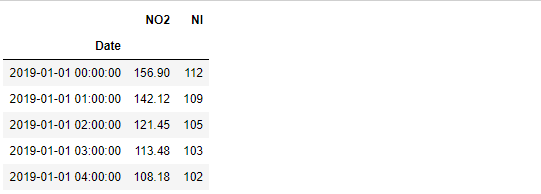


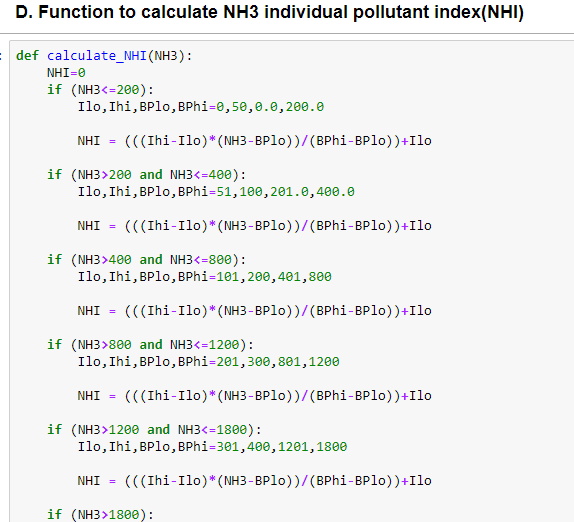


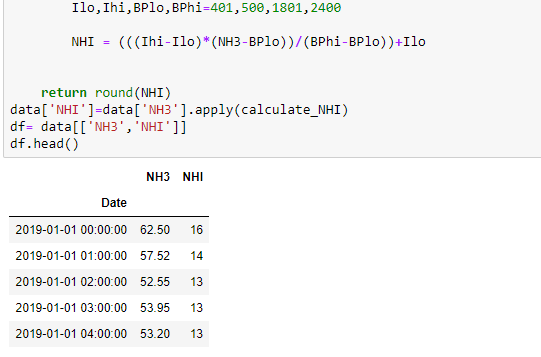


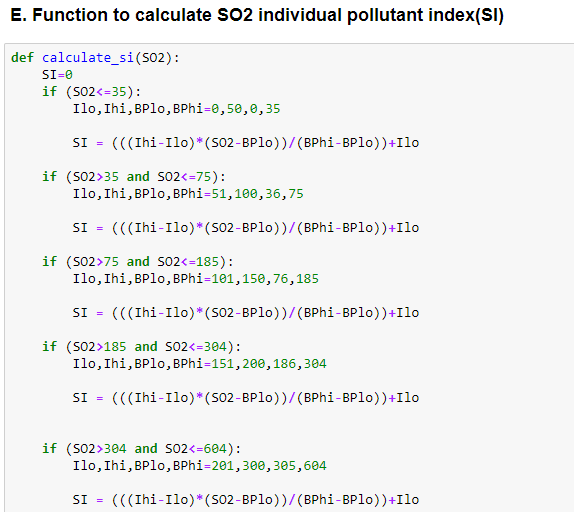


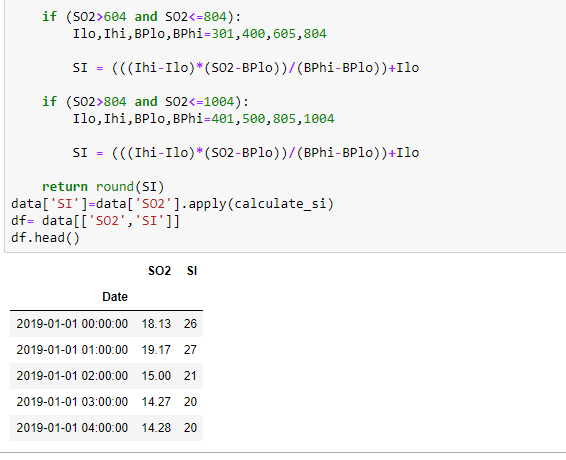


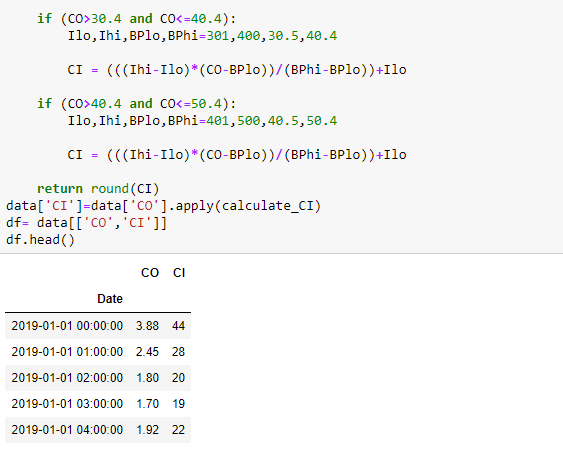
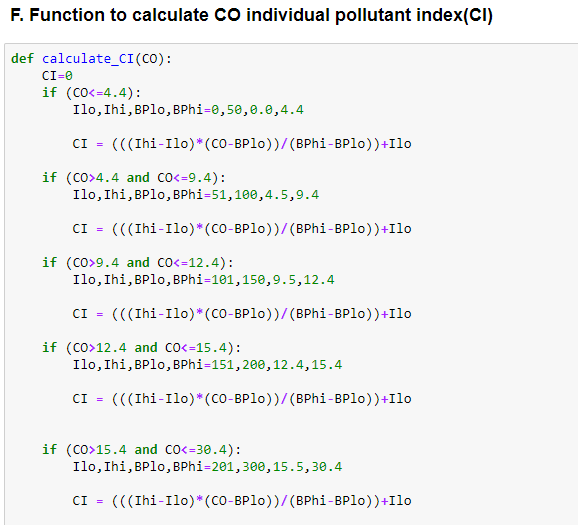


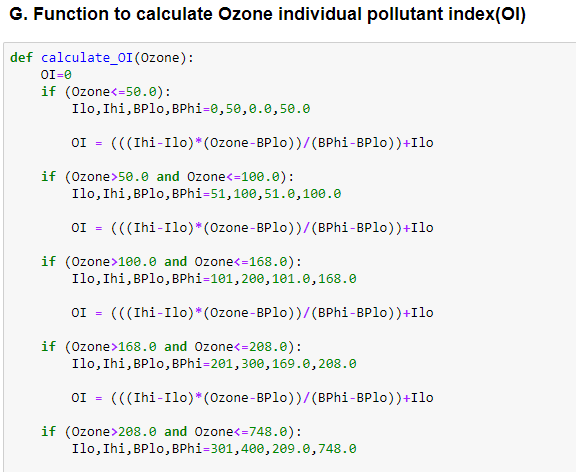


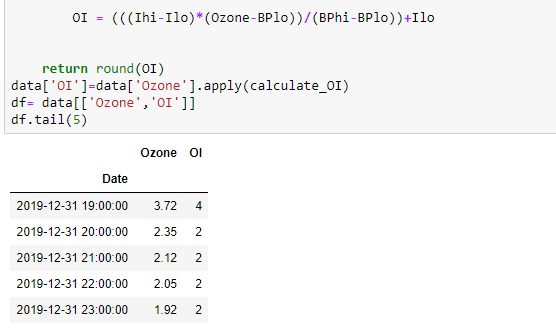




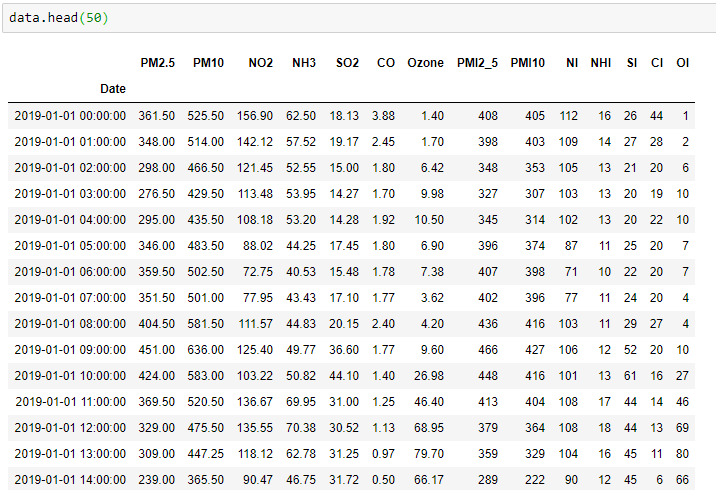






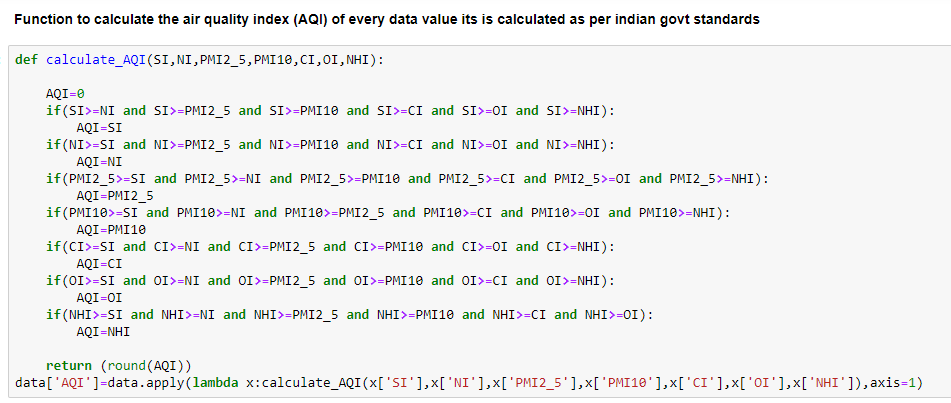


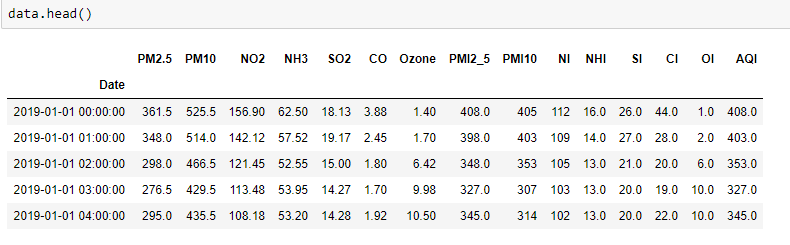
Calculating all the sub index , the full Dataset along with all the sub index,



Calculating AQI

How to calculate the AQI :





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 2-year data sets (2009/07-2011/07) and after the initialization phase, the models are used to predict the air pollutant concentration with newly arrived single datum or a chunk of data during 2011/08-2014/07. Model updating is conducted by either batch learning algorithm or an online-sequential learning algorithm from the newly arrived data.

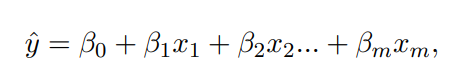
In this chapter, several different methods and models are introduced. Each method produces a different model per station per forecast hour per pollutant. The model development can be separated into two phase: 1) training and 2) testing and updating. Models are first trained and validated using one-year data sets (01/01/2019 – 28/12/2019) and after the initialization phase, the models are used to predict the air pollutant concentration with newly arrived single datum or a chunk of data during 2011/08-2014/07. 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 (Liang et al., 2006). A versatile online-sequential learning algorithm means the data for training are sequentially presented (singly or as a chunk of data) to the learning algorithm. At any time, only the newly arrived data (instead of all past data) are needed to update the model. The new data, once learned by the model, can be discarded (Liang et al., 2006). The learning algorithm has no prior knowledge as to how many training dataset will be presented. A comparison will be made between the online-sequential extreme learning machine (OS-ELM), the online-sequential multiple linear regression (OS-MLR) and the reference model, UMOS which is also online-sequential.

4.1 Input Data Preprocessing

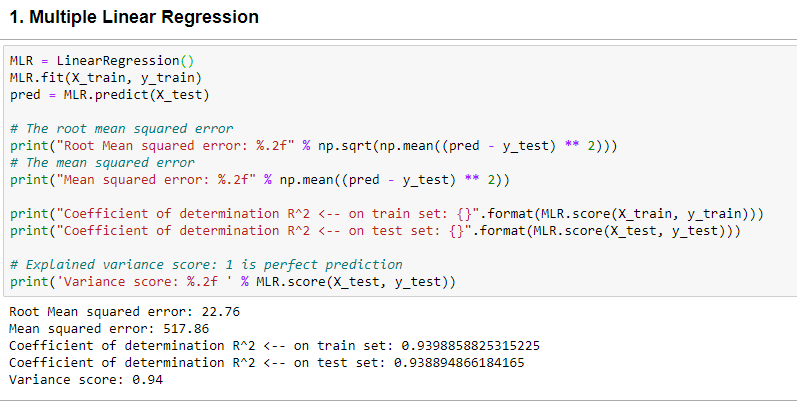
Without properly transforming or scaling the input data, machine learning methods may be trained inefficiently and the resulting model may perform poorly. If the input variables in the training dataset vary greatly in magnitude, the model weights have to adapt to the differences. The resulting weights will also have a large spread in magnitude, rendering the training algorithm inefficient (Rasouli et al., 2012). Input data preprocessing/scaling is an efficient way to solve the problem. The commonly used scaling methods include: (i) linear transformation, (ii) statistical standardization, and (iii) nonlinear transformation (e.g. the logarithmic transformation). Input data in this study is standardized, i.e. data have the mean value subtracted, then divided by the standard deviation, yielding variables with zero mean and unit standard deviation. As separate forecast models are developed for the different hours of the day, there is no need to remove the diurnal cycle from the input data. Control filters with minimum concentration, maximum concentration and rate of change criteria were applied here to remove unrealistic low/high observations and to ensure reasonable rates of changes in the measurements.

4.2 Multiple Linear Regression (MLR)

Multiple linear regression models were developed in the free R software (R Development Core Team, 2011) environment for statistical computing with the package “stats”. MLR is a statistical technique for finding the linear relation between the independent variables (predictors) and the dependent or response variable (Kumar and Goyal, 2013). The general MLR model is built from N observations of the multiple predictor variables xk (k = 1, . . . , m) and the observed target data y. The MLR output variable ˆy can be written in terms of the input predictor variables as

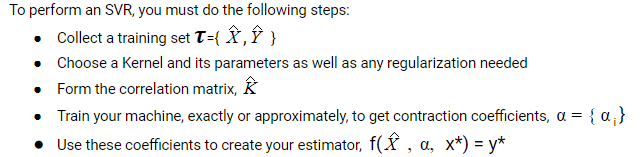
 (4.1)

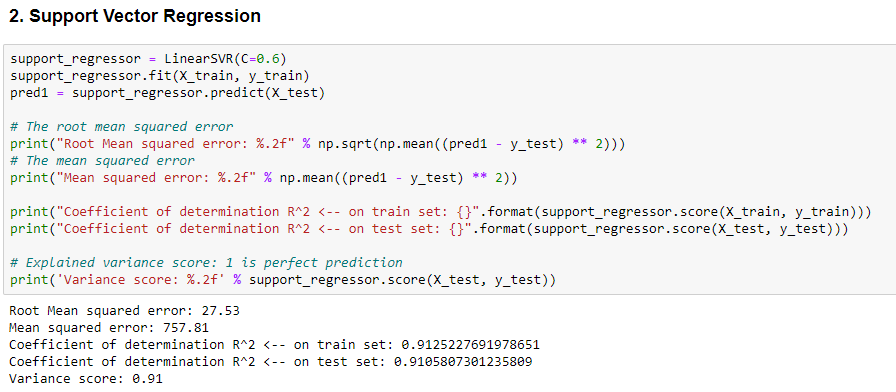
where βj (j = 0, . . . , m) are the regression coefficients or parameters determined by minimizing the MSE between the model output and the target data using a linear least squares algorithm. Stepwise regression is applied here using the R software to choose relevant predictor variables by an automatic procedure (going both forward and backward). The model was trained and testing and also updating with one years of data (01/01/2019. Predictors were re-selected and the linear regression was recalculated during each model update.



4.3 Support Vector Regression

Support Vector Regression(SVR) is quite different than other Regression models. It uses the Support Vector Machine(SVM, a classification algorithm) algorithm to predict a continuous variable. While other linear regression models try to minimize the error between the predicted and the actual value, Support Vector Regression tries to fit the best line within a predefined or threshold error value. What SVR does in this sense, it tries to classify all the prediction lines in two types, ones that pass through the error boundary( space separated by two parallel lines) and ones that don’t. Those lines which do not pass the error boundary are not considered as the difference between the predicted value and the actual value has exceeded the error threshold, 𝞮(epsilon). The lines that pass, are considered for a potential support vector to predict the value of an unknown. The following illustration will help you to grab this concept.

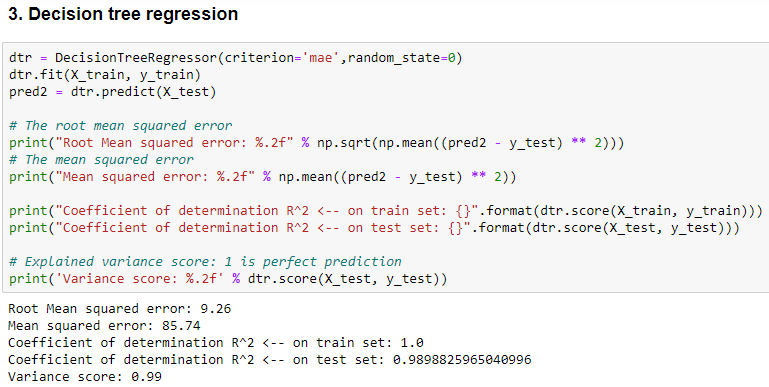




4.4 Decision Tree Regression

The decision tree is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

We can see that if the maximum depth of the tree (controlled by the max\_depth parameter) is set too high, the decision trees learn too fine details of the training data and learn from the noise, i.e. they overfit.



**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.

**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.