**MARRI LAXMAN REDDY**



**INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

(Approved by AICTE -NEW Delhi, Accredited by NAAC with `A`& Affiliated to JNTU)

DUNDIGAL, Quthbullapur (M) , HYD-500043



WE ARE FROM ***COMPUTER SCIENCE AND ENGINEERING DEPARTMENT***

**PRESENTED BY**

***1.PRAVEEN KUMAR***

***2.VARUN KANIKARAM***

***3.SAI GOPINADH***

***4.SURENDHRA***

***5.ROHITH***

***Air Quality Prediction:-***

#### **1.1.Introduction**

***Air is one of the most essential natural resources for theexistence and survival of the entire life on this planet. Allforms of life including plants and animals depend on air for their basic survival. Thus, all living organisms need gooquality of air which is free of harmful gases to continue theirlife. According to the world&#39;s worst polluted places byBlacksmith Institute in 2008 [1], two of the worst pollutionproblems in the world are urban air quality and indoor airpollution. The increasing population, its automobiles andindustries are polluting all the air at an alarming rate. Airpollution can cause long-term and short-term health effects.It&#39;s found that the elderly and young children are moreaffected by air pollution. Short-term health effects includeeye, nose, and throat irritation, headaches, allergic reactions,and upper respiratory infections. Some long-term health effects are lung cancer, brain damage, liver damage, kidneydamage etc. and we are explaining about air quality predictionin Delhi According to data set there are 5 columns they are Ozone, Solar radiation, Air, Temp, Month, Day etc.5***

##### ***1.2 PROBLEM STATEMENT:-***

***The purpose of this project is to identify the effect that surface modifications have on the urban area in Delhi phenomenon and related ozone problem in the metropolitan area of Delhi, IL. The basic hypothesis is that urban, summertime temperatures can be significantly lowered by increasing the vegetative landscape cover and enhancing the solar reflectivity of paved and roofed surfaces within an urban area. It is proposed that in addition to a decrease in temperature, the modification of an urban surface to include more vegetative cover and lighter, lower albedo surfaces will also reduce energy consumption, ozone exceedances, and detrimental environmental and human health effects associated with high levels of ozone.***

***The analysis is divided into three main parts. The first section of this report introduces the causes of ground level ozone and its effects in urban areas. It explains both the chemistry and transport associated with ozone exceedances. The second section is a compilation of the most viable mitigation strategies of urban heat islands: increasing vegetative cover and increasing proportions of light to dark surfaces. The effects, implementation strategies, and specific strengths and weaknesses associated with each approach are described, including a comparison of asphalt and concrete pavements systems using a life cycle analysis approach.***

***The final section provides a case study of the Delhi. This study entailed an examination of the land use, development of an urban fabric analysis in which total vegetative, paved, and roofed surfaces are investigated and quantified, and discussion on the effectiveness of possible mitigation strategies in the Delhi. In general, the associated findings of my research are located within this final section. Delhi is the one of the most polluted area in India. Based on Delhi Air Quality Dataset we are finding the research.***

***1.3.*** ***Related Work***

Many previous works have been proposed to apply machine learning algorithms to air quality predictions. Some researchers have aimed to predict targets into discretized levels. Kalapanidas et al. [32] elaborated effects on air pollution only from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity and classified air pollution into different levels (low, med, high, and alarm) by using a lazy learning approach, the case-based reasoning (CBR) system. Athanasiadis et al. [45] employed the σ-fuzzy lattice neurocomputing classifier to predict and categorize O 3 concentrations into three levels (low, mid, and high) on the basis of meteorological features and other pollutants such as SO2, NO, NO2, and so on. Kurt and Oktay [33] modeled geographic connections into a neural network model and predicted daily concentration levels of SO2, CO, and PM10 3 days in advance. However, the process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently is inaccurate. Other researchers have worked on predicting concentrations of pollutants. Corani [46] worked on training neural network models to predict hourly O3 and PM10 concentrations on the basis of data from the previous day. Mainly compared were the performances of feed-forward neural networks (FFNNs) and pruned neural networks (PNNs). Further efforts have been made on FFNNs: Fu et al. [47] applied a rolling mechanism and gray model to improve traditional FFNN models. Jiang et al. [48] explored multiple models (physical and chemical model, regression model, and multiple layer perceptron) on the air pollutant prediction task, and their results show that statistical models are competitive with the classical physical and chemical models. Ni, X. Y. et al. [49] compared multiple statistical models on the basis of PM2.5 data around Beijing, and their results implied that linear regression models can in some cases be better than the other models.

***1.4****.* ***preprocessing:-***

We paired the collected meteorological data and air pollutant data on the basis of time to obtain the required data format for applying the machine learning methods. In particular, for each variable, we formed one value for each hour. However, the original data may have contained multiple records or missing values at some hours. To preprocess the data, we calculated the hourly mean value of each numeric variable if there were multiple observed records within an hour and chose the category with the highest frequency per hour for each categorical variable if there were multiple values. Missing Big Data Cogn. Comput. 2018, 2, 5 5 of 15 values existed for some variables, which was not tolerable for applying the machine learning methods used in this study

***1.5.OBJECTIVES OF RESEARCH:-***

***Solar radiation***

***Wind***

***Temperature***

***2.Review on literature:-***

***Air quality index***

Air Quality Index The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health affects you may experience within a few hours or days after breathing polluted air. AQI is value between 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. The AQI is an ”index” determined by calculating the degree of pollution in the city or at the monitoring point and includes five main pollutants - particulate matter, ground-level ozone, sulfur dioxide, carbon monoxide and nitrogen dioxide.

Each of these pollutants has an air quality standard which is used to calculate the overall AQI for the city. For better understanding and presentation, the AQI is broken down into six categories, each color coded with the number scale.( www.imd.gov.in/ ) The measurement scale is based on color system and a definite scale as per the

The AQI of each air pollutant is calculated using the Eq.2.3 Ip = [(Ihi − Ilow)/(BPhi − BPlow) ∗ (Cp − BPlow) + Ilow] (2.3)

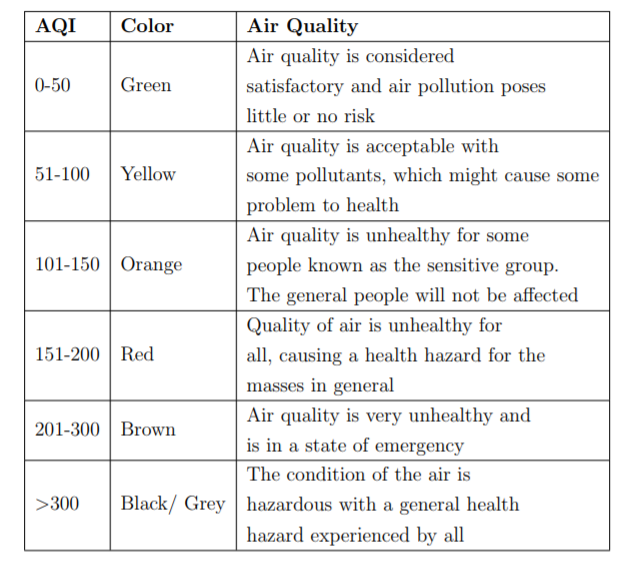
where Ip is the index of the pollutant; Cp is the rounded concentration of pollutant p;

BPhi is the breakpoint greater or equal to Cp;

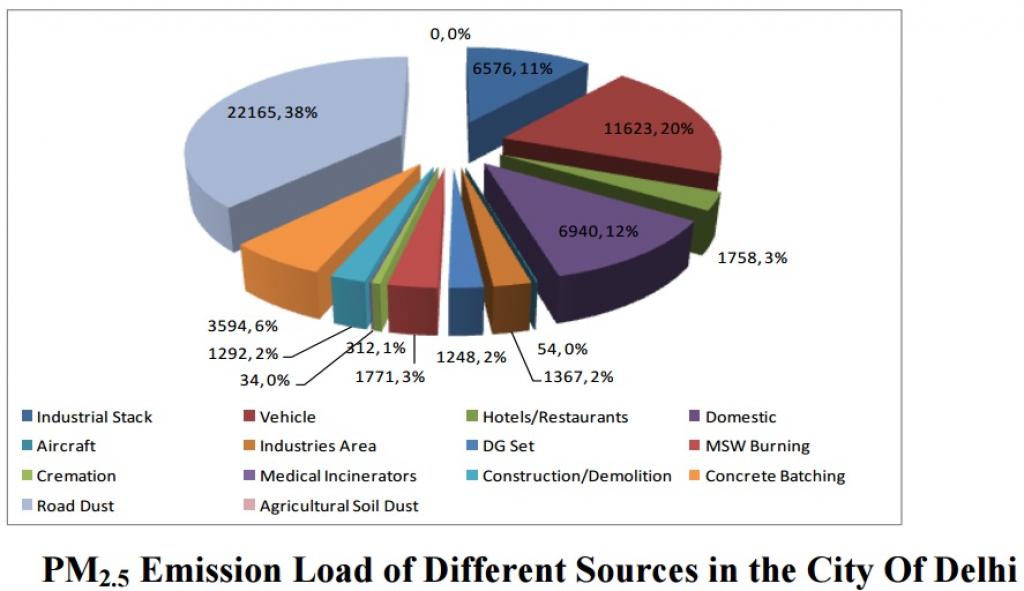
BPlow is the breakpoint less than or equal to Cp ;

37 Ihi is the AQI corresponding to BPhi;

Ilow is the AQI corresponding to BPlow;



### ***4. COLLECTION DATA:***



***In the above figure explain about the Air quality in Different sources in the city of Delhi. How much percentage is there about Vehicles, Industries Area, Road Dust, Agricultural Soil Dust, Industrial Stack etc.***

# *4.@@@:- METHODOLOGY OF AIR QUALITY PREDICTION:-*

# *The principal stages of the proposed methodology are:*

# *1.Multi Linear Regression*

# *1.In this method we are use Multi Linear Regression For to*

# *find Air Quality Prediction*

# *2. Generation of a mathematical formalization for this model.*

# *3. Creation of a multivariate hierarchical structural model*

# *based on system analysis.*

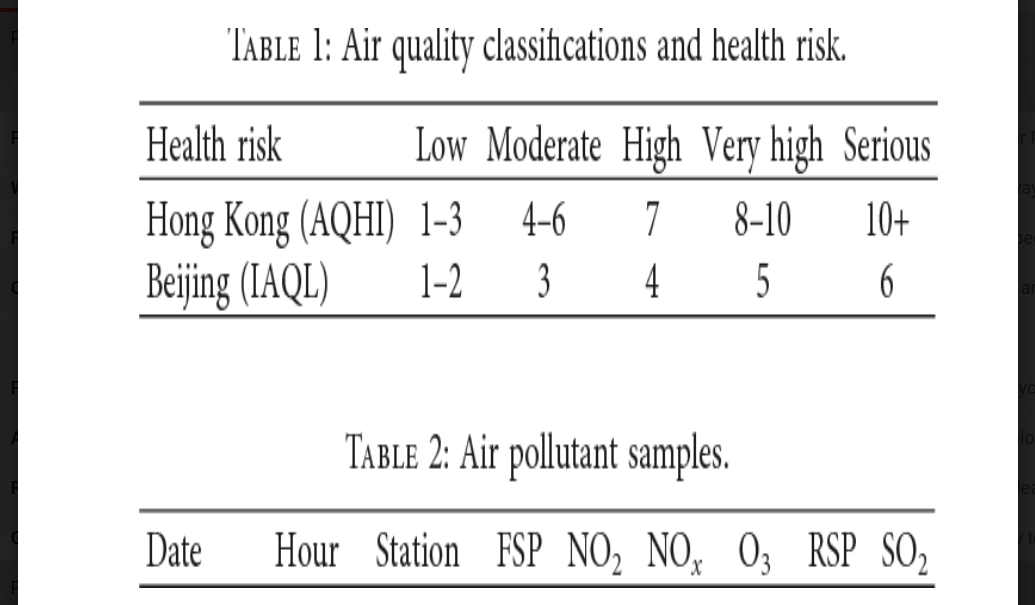
# *4. These coefficients allow an epidemiological meaningful*

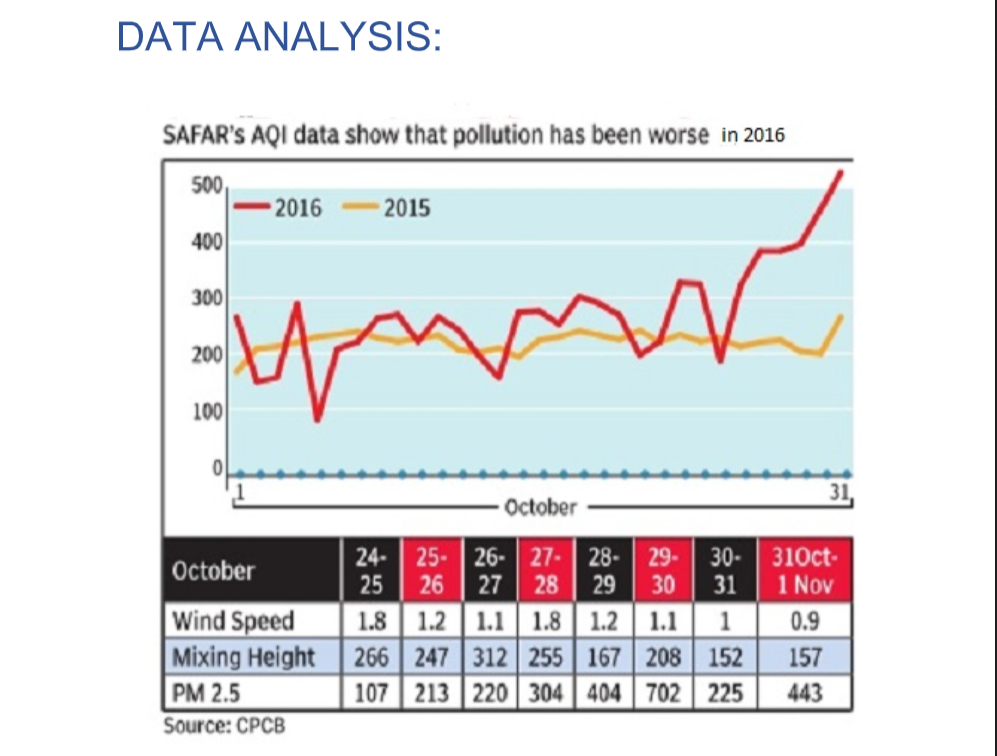
# *model interpretation. Thus,*

# *We proposed an integration air quality prediction framework in airports, which includes two main parts: air quality classification in airports and the supervised learning for AQI prediction. The air quality classification in airport is mainly composed of air quality index calculation method, airport emissions inventory evaluation, and airport emissions dispersion modeling. In the supervised learning for prediction part, all the AQI level results for the airport are used as training examples; the supervised learning method is used to analyze the training data and produce an inferred function.*

***Exploratory data analysis:-***

***TABLES AND FIGURES*:-**





***trigger a variety of health problems such as lung irritation andinflammation, asthma attacks, wheezing, coughing, andincreased susceptibility to respiratory illnesses.Particles matter (PM), or airborne particles, include dust, dirt,soot, and smoke. Some particles are directly emitted into the air, for example, cars, trucks, buses factories, constructionsites, and wood burning. Other particles are formed in the airwhen gases from burning fuels react with sunlight and water vapor. Such gases, from incomplete combustion in motor vehicles, at power plants and in other industrial processes,contribute indirectly to particulate pollution. PM can cause chronic bronchitis, asthma attacks, decreased lung function,coughing, painful breathing, as well as a variety of seriousenvironmental impacts such as acidification of lakes and streams and nutrient depletion in soils and water bodies.***

***STATISTICAL TECHNIQUES AND DATA***

***VISUALIZATION:-***

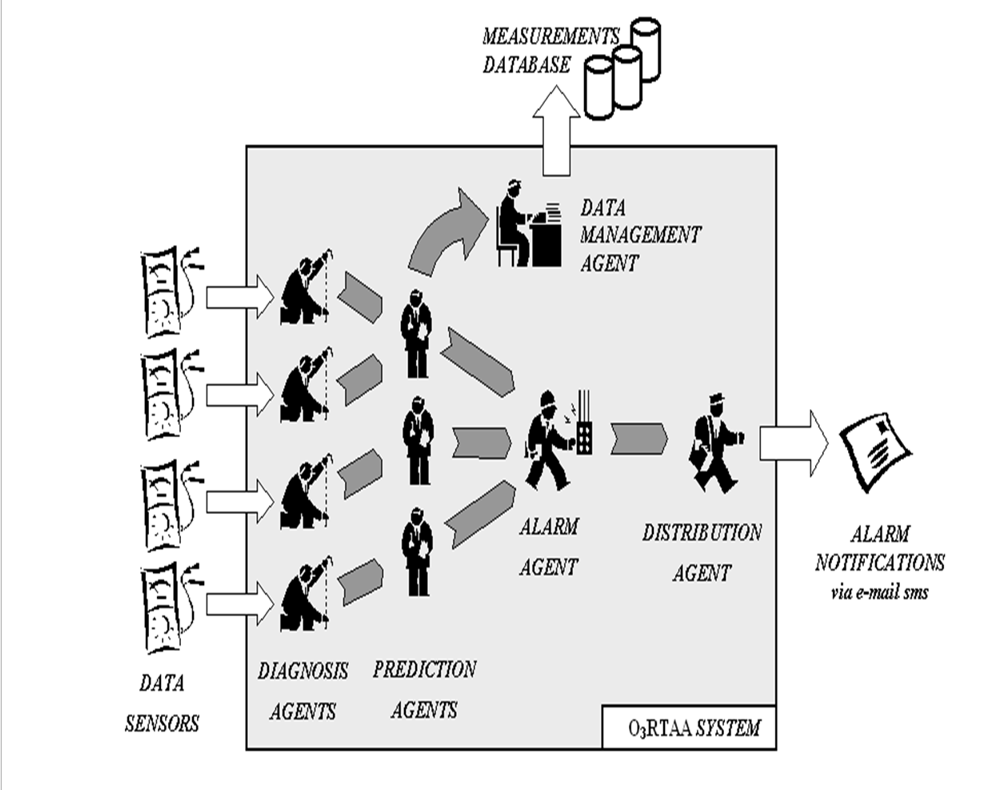
***Air quality simulation models have been the subject ofextensive evaluations to determine their performance under a variety of environmental and meteorological conditions.While much information has been gathered, no clearly defined methodology exists for comparing the performanceof two or more models. The purpose of this paper is to present a statistically oriented procedure to test if the performance of one model is superior to others using a composite performance index involving the bootstrap resampling technique The detection of NO 2 by its chemiluminescent reaction with luminol is a rapid and sensitive means of measuring atmospheric NO 2 . However, testing and field use of a commercial NO 2 monitor employing this detection scheme have shown that several corrections are necessary in order to obtain accurate NO 2 data at low concentrations. In use aboard aircraft, the NO 2 data must be corrected for zero offset, altitude (i.e. pressure), nonlinearity of response, and interferences from ozone and PAN. Detector response is dependent on the age of luminol reagent solution. This paper describes the tests performed to determine correction factors, the algorithms and order of precedence for applyingthe corrections, and other observations regarding detector performance.***

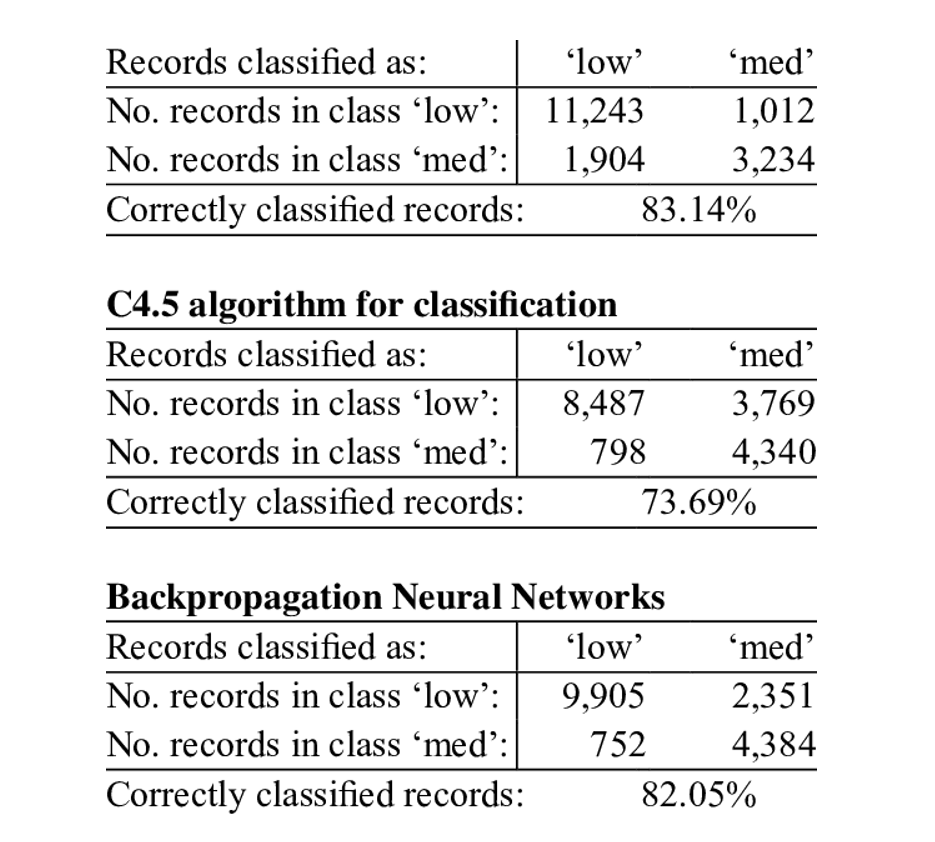
***Air Quality Forecasts:-***

***The District is attaining the national ambient air quality standards (NAAQS) for all pollutants except ground-level ozone.  Ground-level ozone, also known as smog, is created by a chemical reaction between precursor pollutants, primarily oxides of nitrogen (NOx) and volatile organic compounds (VOCs) in the presence of sunlight and high temperatures.***

***“Ozone season” lasts from May to September.  During ozone season, air quality forecasters rate the quality of the air on a daily basis and recommend actions when predictions indicate that air quality may be bad for public health.***

***DATA MODELING USING SUPERVISED ML TECHNIQUES:-***





##### ***FINDING AND SUGGESTIONS:-***

Raise awareness about **air pollution**. ...

* Run tree plantation drives. ..
* Conserve energy. ...
* Curb Unbridled Industrialization. ...
* Reduce use of personal vehicles. ...
* Let your vehicles stick to **pollution** control norms. ...
* Choose cleaner options. ...
* Ensure low waste production

In this example we would look at the task of predicting air quality. We would use the following dataset:

<https://archive.ics.uci.edu/ml/datasets/Air+Quality>

This dataset contains the responses of a gas multisensor device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data was recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) are provided by a co-located reference certified analyzer.

Attribute Information:

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
4. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
6. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
7. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. True hourly averaged NOx concentration in ppb (reference analyzer)
9. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
11. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
12. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13. Temperature in Â°C
14. Relative Humidity (%)
15. AH Absolute Humidity

Download the dataset from the link and save it in the same directory as your code. Next we import all the required modules:-

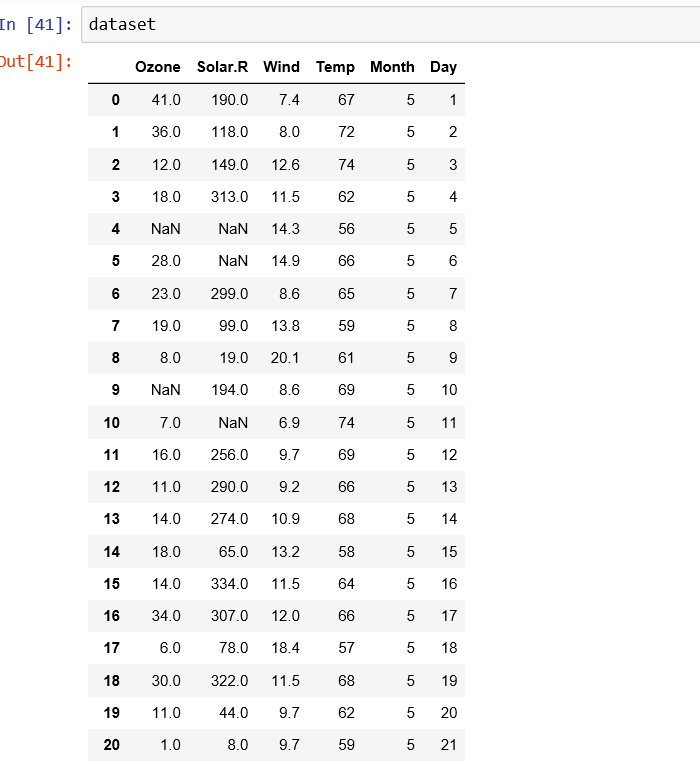


**

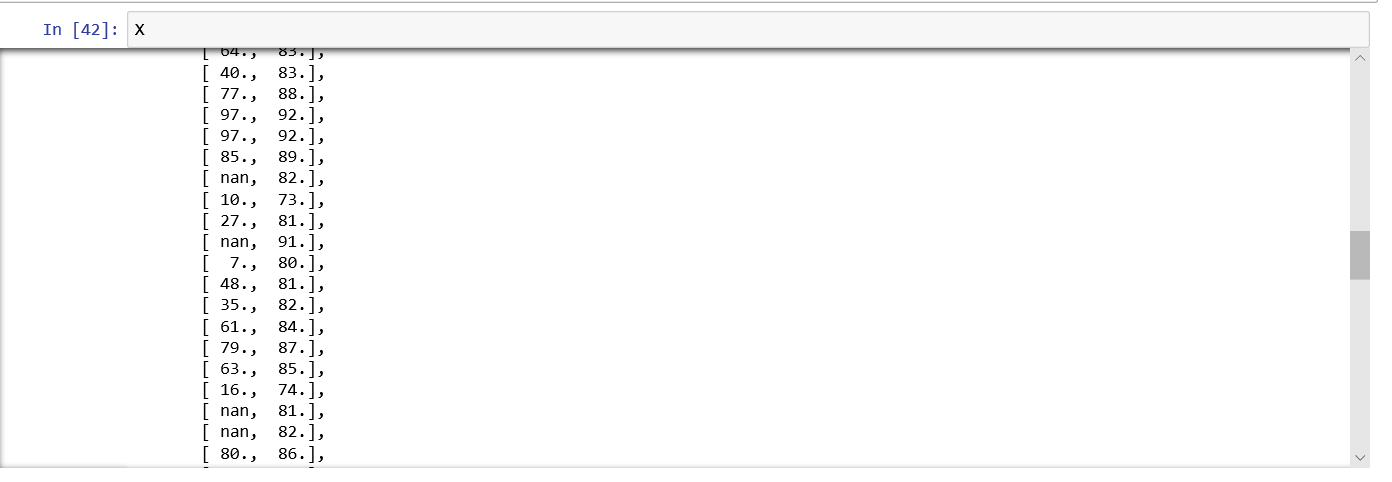
*We will use the AirQualityUCI.csv file as our dataset. It is a seperated file so we'll specify it as a parameter for the read\_csv function. We'll also use parse\_dates parameter so that pandas recognizes the 'Date' and 'Time' columns and format them accordingly.*

*This is our dataset:-*

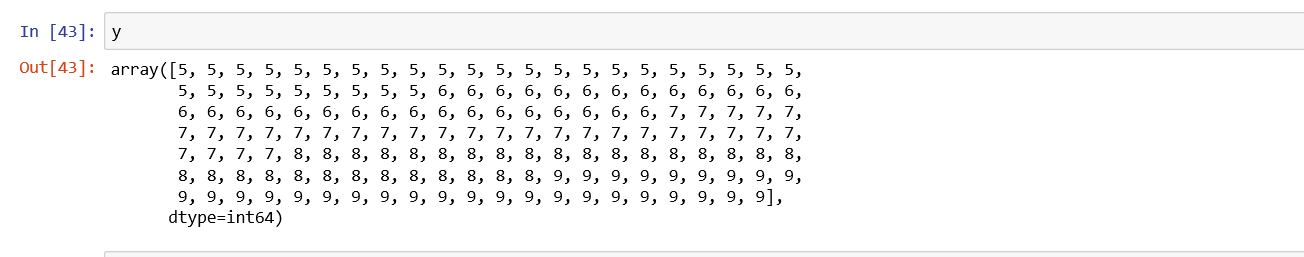
*IT INDICATES A SOLAR.R ,HOW MUCH WIND IS FLOWING ,HOW MUCH TEMP IS THERE ..,*



***And this our x vaeriables:-***



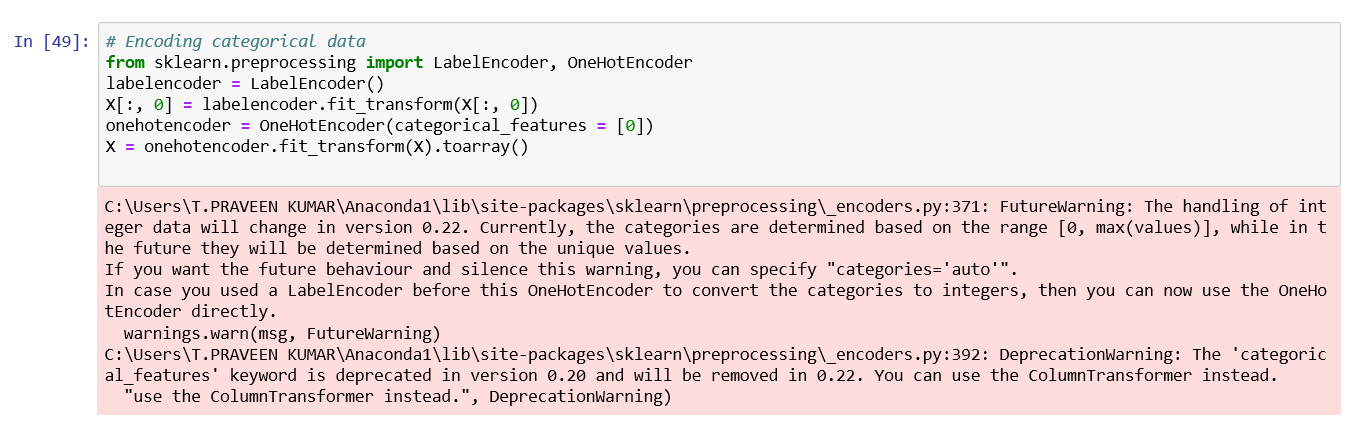
***And this y variable’s:-***



***The data contains null values. So we drop those rows and columns containing nulls.***

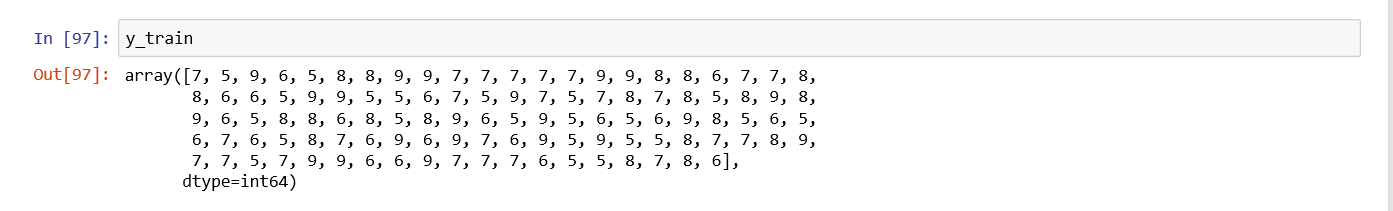
***AFTER THAT THAY CAN PERFORM A ENCODEING…!***

* Integer ***Encoding***. As a first step, each unique category value is assigned an integer value. For example, “red” is 1, “green” is 2, and “blue” is 3. ...



We will define our features and ignore those that might not be of help in our prediction. For example, date is not a very useful feature that can assist in predicting the future values.

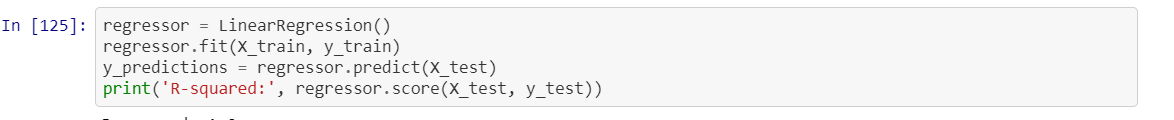




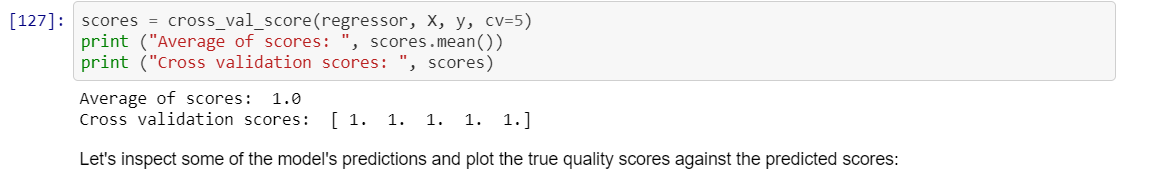
# Regression

Please see the previous examples for better explanations. We have already implemented Decision Tree Regression and Random Forest Regression to predict the Electrical Energy Output.

## Multi Linear Regression



The R-squared score of 1 indicates that 100 percent of the variance in the test set is explained by the model. The performance can change if a different set of 75 percent of the data is partitioned to the training set. Hence Cross-validation can be used to produce a better estimate of the estimator's performance. Each cross-validation round trains and tests different partitions of the data to reduce variability.



### ***Fitting models with gradient descent***

*Gradient descent is an optimization algorithm that can be used to estimate the local minimum of a function.*

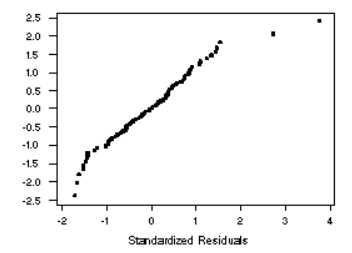
*We can use gradient descent to find the values of the model's parameters that minimize the value of the cost function. Gradient descent iteratively updates the values of the model's parameters by calculating the partial derivative of the cost function at each step. Although the calculus behind the cost function is not entirely required to implement it with scikit-learn, having an intuition for how gradient descent will always help to you use it effectively.*

# *And also use Multiple Linear Regression*

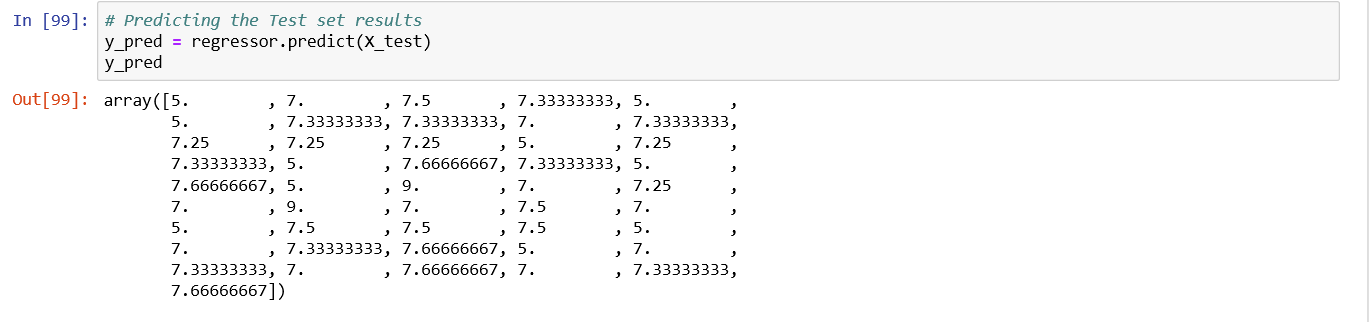
# *A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. 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*.

# 

***After fitting the regression line, it is important to investigate the residuals to determine whether or not they appear to fit the assumption of a normal distribution.,Despite two large values which may be outliers in the data, the residuals do not seem to deviate from a random sample from a normal distribution in any systematic manner.***



And we predicted a array values:-





***And finally we got 0.1974677214038908 this percentage***

##### ***CONCLUSION:***

We performed a series of analyses, and came to some rather interesting conclusions. Our methodology included collecting the Air Quality Index (AQI) data in four areas in Delhi in 2016 as well as 2015 and drawing spatial as well as temporal comparisons to give a clear picture of what was happening. We also collected data from 2014 but it was sparse, and yet it supported our findings definitively, if not conclusively. One of the most relevant is the fact that this year was truly anomalous, if only firecrackers are considered major pollutants during and after 30th October. Comparing the AQI this year to that observed in 2015, it was obvious that firecrackers could not have caused the shocking level of pollution seen this year.