CAPSTONE PROJECT

ON

AIR QUALITY PREDICTION OF RELATIVE HUMIDITY

EDA & MODELLING

Student Name:	Haritika Jolly
Student ID:	
Programme:	Capstone Project
Year:	2022
Professor:	**Ms. Savita Seharawat**
Submission Due Date:	27rd July 2022
Project Title:	Exploratory analysis of Air Quality Dataset and to check the quality of air using 'Air Quality Chemical Multisensor Device' by finding the R^2 score and coefficient of regression using different regression models and the best model is selected to evaluate the Air Quality.

Table of Contents

Abstract:

Examining and protecting air quality in this world has become one of the essential activities for every human in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution, we need to implement models that will record information about concentrations of air pollutants. The deposition of these harmful gases in the air is affecting the quality of people's lives by altering their health, especially in urban areas. In this paper, regression techniques are used to predict the concentration of Carbon monoxide in the environment. Carbon monoxide causes headaches, dizziness, vomiting, nausea, and heart diseases. The dataset is downloaded and imported to the project. It contains data on average hourly responses of major air pollutants for nearly one year. This dataset is used to predict the Quality of Air based on other parameters using regression analysis. It creates awareness among people about the air quality degradation, and it's health effects. Support environmentalists and government to frame air quality standards and regulations based on issues of toxic and pathogenic air exposure and health-related hazards for human welfare.

In comparison with all 3 classification algorithms used in this work,

Decision Tree Regression outperformed Linear Regression and Lasso Regression classification by achieving the highest accuracy among all of the three classification algorithms.

Classification algorithms used: Linear Regression, Lasso Regression and Decision Tree Regression.

Keywords: pollution, health, Carbon monoxide, time-series data, Regression analysis

1. Introduction:

Air pollution will endanger human health and life in big cities, especially to the elderly and children. This is not an individual problem of one person but a global problem. Therefore, many countries in the world made air pollution monitoring and control stations in many cities to observe air pollutants such as NO2, CO, SO2, PM2.5, and PM10 and to alert the citizens about pollution index which exceeds the quality threshold. Particulate Matter PM 2.5 is a fine atmospheric pollutant that has a diameter of fewer than 2.5 micrometers. Particulate Matter PM10 is a coarse particulate that is 10 micrometers or less in diameter. Carbon Monoxide CO is a product of combustion of fuel such as coal, wood, or natural gas. Vehicular emission contributes to the majority of carbon monoxide let into our atmosphere. Nitrogen dioxide or nitrogen oxide expelled from high-temperature combustion: sulfur dioxide SO2 and Sulphur Oxides SO produced by volcanoes and in industrial processes. Petroleum and Coal often contain sulfur compounds, and their combustion generates sulfur dioxide. Air pollution is caused by the presence of poison gases and substances; therefore, it is impacted by the meteorological factors of a particular place, such as temperature, humidity, rain, and wind. To clear out this statement, weather data including air temperature, relative humidity, precipitation, wind speed, wind direction, which was also collected in real-time by sensors and analyzed them along with the air pollution values. In this project, the regression analysis technique is used to evaluate the quality of Air by finding the R^2 score and coefficient of regression using different regression models and the best model is selected to evaluate the Air Quality.

. This project supports environmentalists and the government to frame air quality standards and regulations based on issues of toxic and pathogenic air exposure and health-related hazards for human welfare.

1.1 Research Questions & Objective

- **Find the Target Variable.**
- **Find the variable highly correlated with the target variable. Decide independent parameter to predict the dependent parameter for modelling.**
- **Find the best Model for Evaluating the Quality of Air with highest Accuracy Score.**
- **Summarize**

GitHub Repository Link :

https://github.com/haritikajolly/Air-Quality-Prediction

2. Literature Review Existing System

Aditya C R and et al., [1] performed two important tasks (i) Detected the levels of PM 2.5 based on given atmospheric values. (ii) Predicted the level of PM2.5 for a particular date. Logistic regression is used to detect whether a data sample was either polluted or not polluted. Autoregression was employed to predict future values of PM 2.5 based on the previous PM2.5 readings. This paper mainly predicts the air pollution level in the city with the ground data set.RuchiRaturi A and Dr. J.R Prasad [6] used the linear regression and artificial neural network (ANN) Protocol for prediction of the pollution of the next day. The system helped to predict next date pollution details based on basic parameters and analyzing pollution details and forecast future pollution. Time Series Analysis was also used for recognition of future data points and air pollution prediction.Zheng Y and et al. [7] tried to forecast the air pollution by reading an air quality monitoring station data over the next 24 hours, considers current meteorological data, weather forecasts, air quality data of the station within 100 km, and other stations within 150 km. They used machine learning and deep learning algorithms, including a linear regression-based temporal predictor along with a neural network-based spatial predictor. The prediction values were not good because PM2.5 value was increased, but they predicted decreasing. Baldasano J.M, Akita Y [2] has done modeling a framework based on the Bayesian Maximum Entropy method that integrated monitoring data and outputs from existing air quality models based on Regression and CTM (Chemical Transport Models). It was applied to estimate the yearly average of NO2 concentrations in Spain. It gave the output of Care Transition Measure CTM and the interurban scale variability through regression model output.McCollister G.M and Wilson K.R [4] described the development of an application to predict the peak ozone levels with the help of meteorological and air quality prediction variables Athens area. For this purpose, a number of regression models were considered, while the selection of the final model was based on extensive analysis and on literature. The model adapted includes variables that are available on a daily basis, so as daily operational maximum ozone concentration level forecast can be achieved.Rao S.T. and Zurbenko I.G [5] presented a statistical method for filtering or moderating the influence of meteorological fluctuations on ozone layer concentrations. The use of this statistical technique in examining trends in ambient ozone air quality is demonstrated with ozone data from a monitoring location in Newyork. The results indicate that it can detect changes in the ozone layer due to changes in emissions in the presence of meteorological fluctuations.GnanaSoundari.A and et al. [3] developed a model to predict the air quality index based on historical data of previous years. It made the prediction using a multivariable regression model. It improved the efficiency of the model by applying cost Estimation for predictive Problems. This model had only 46% accuracy in predicting the available dataset on predicting the air quality index of India.

2.1 Proposed System

In the proposed system, the air quality dataset is downloaded, which is available in CSV format. The comma separated value data format can easily be processed and analyzed fast using a computer and the data utilized for various purposes. It is imported to the project by using a panda package available in anaconda software. The dataset contains 15 important attributes that help in air quality prediction. Initially, the dataset is preprocessed with suitable techniques to remove the inconsistent and missing valued data, and the needed features from the dataset are selected for better results. Then the dataset is split off into training and test dataset in order to evaluate the performance of the model. The processed data sets are analyzed through different regression analysis techniques for accurate results. Regression analysis is the form of a predictive modeling technique that investigates the relationship between a dependent and independent variable. This technique is used for forecasting or predicting, time series modeling, and finding the causal effect relationship between the variables. Regression analysis is a method of analyzing and modeling data. There are different kinds of regression techniques available to make predictions namely Linear regression, Decision tree regression and Lasso regression.

3. Research Methodology

The methodology to be used for this research follows the cross-industry standard process for data mining (CRISP-DM). It provides a systematic approach to planning a data mining project. This approach is reliable and well-proven due to its step by step process and its general applicability. This CRISP-DM includes five phases which are hierarchical and will be implemented during a data mining project.

These are shown in the diagram below;

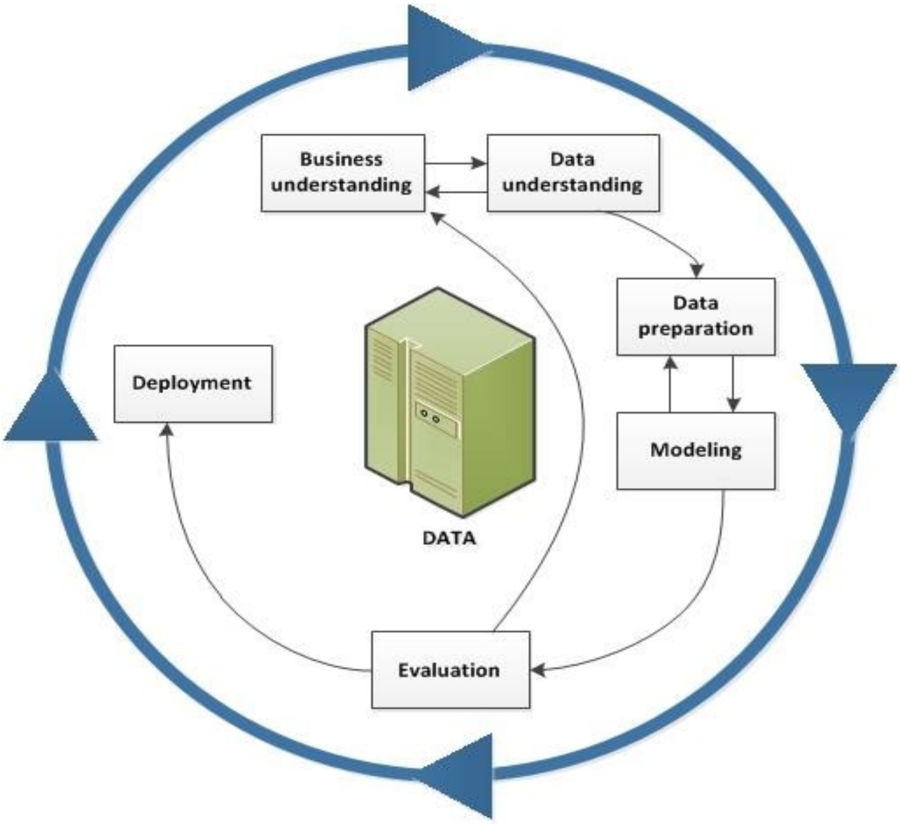


Figure 1 CRISP-DM Approach

Figure 2 represents the flow diagram of the system. The diagram represents the step-by-step process, from data preprocessing to air quality prediction.

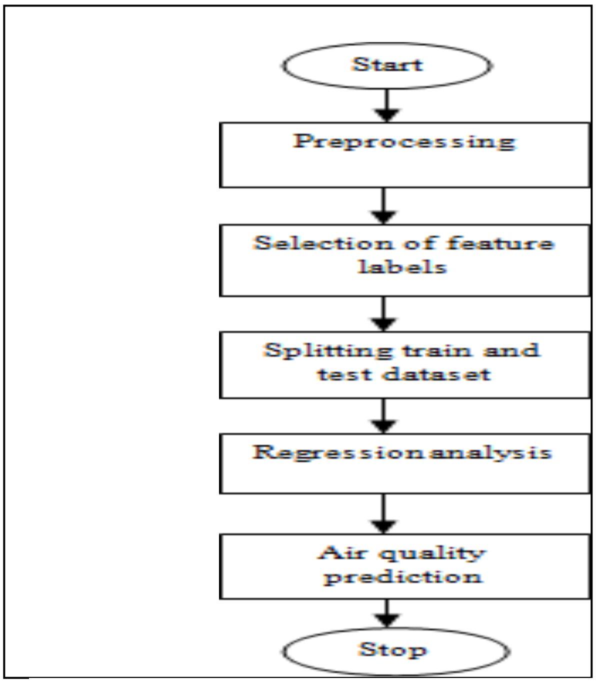


Figure 2 Process - Data pre-processing to Air Q Prediction

3.1 Business understanding

The first phase of the approach chosen for this data mining project is market segment awareness. For the main objective of this project to be achieved, an appropriate dataset needs to be put into consideration. **The historical dataset used for this project was derived from the UCI . Containing 8991 rows with 15 columns.**

3.2 Data preparation

The data for this analysis will be designed finally after having understood the data. The data is expected to be used for their potential modelling for this analysis. this analysis will start by gathering data from the UCI. The data will be obtained by downloading the Air Quality dataset which will be followed by the selected data exploration. The exploration will help to provide a clearer understanding of the data characteristics, size, and structure. Exploration will also assist in choosing the variables to be used while keeping in mind the study's question and purpose Better analysis of the data for this research will expose data quality problems and observations. It will eventually help in defining the type of data mining technique. It will ultimately help to decide what type of data mining technique to use for the research to better achieve a reasonable outcome.

3.3 Data Modelling

The data will Present its independent variables and target variable after the data preparation has been completed to a certain level which will be further divided by a certain percentage for validation into test and training data. A certain percentage will be allocated to the training, while its percentage will be allocated to the test though smaller. The training is assigned higher so that the classifier can learn from the training larger part of the Dataset. This breaking will make it possible to apply the chosen modeling technique. **This research will be limited to three chosen modeling techniques** from other modeling approaches used for a data mining problem. **Linear regression, Lasso Regression, Decision Tree** are the models to be used for this research work.

3.4 Model evaluation

These model(s) shall be evaluated for every analytical function. The models to be implemented for this analysis will be tested, as this will shape as necessary an important part of the research process. the model with the highest R^2 score on both training and testing datasets will be concluded as the best model for evaluating the Quality of Air.

4. Implementation:

The figure below shows the mechanism owing by which the execution of this research is directed. the python programming language has been used to perform all the experiments.
The flow diagram above shows how the experiment on the dataset downloaded from UCI will be carried out using the python code.

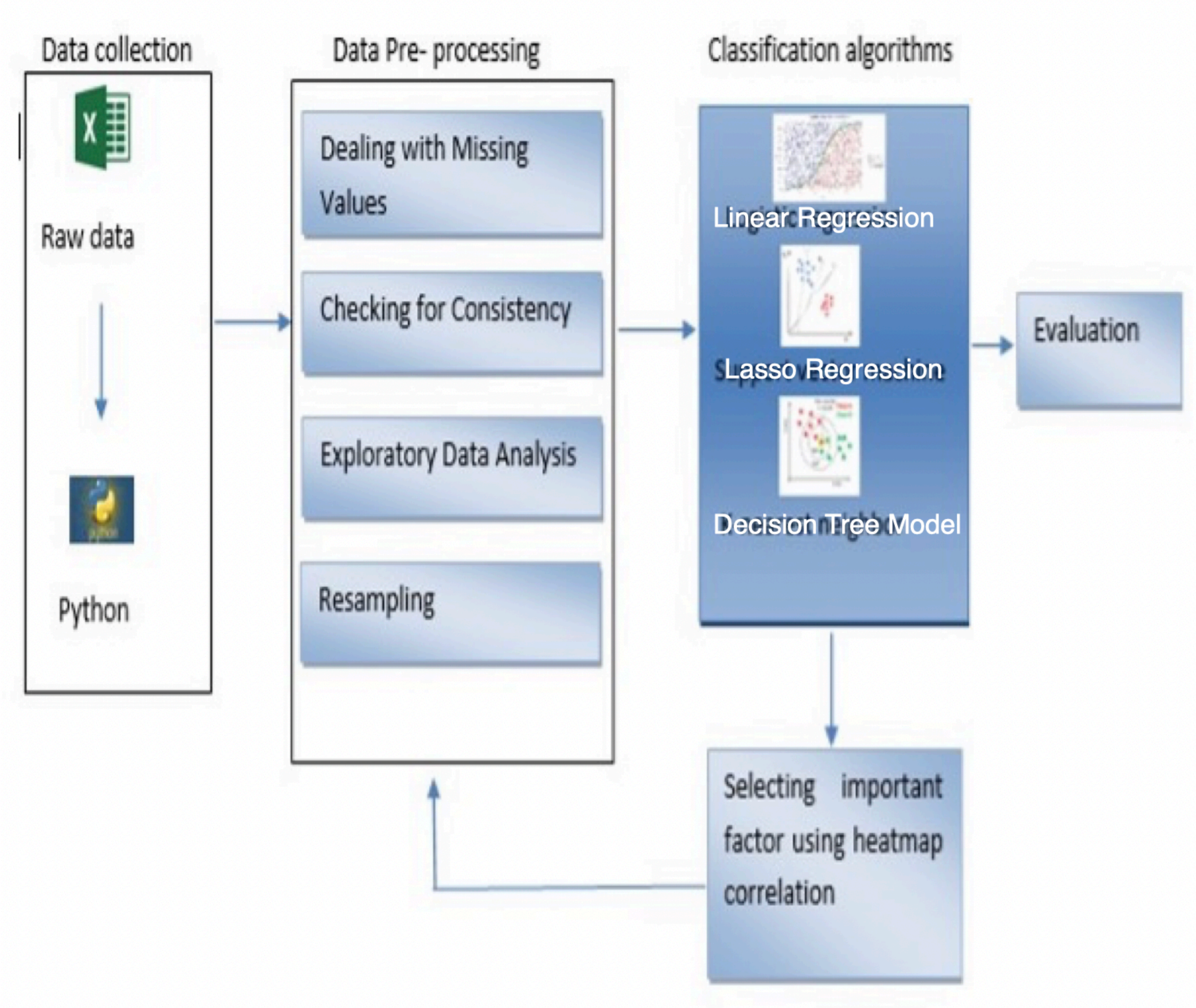


Figure 3 Implementation Flow Diagram

5. Data Detail:

For this research work, the Air Quality dataset was downloaded from the UCI Machine Learning Repository.

The dataset contains data of average hourly responses of different elements in the air for nearly one year from March 2018 to April 2019.

Dataset consists of 9357 rows and 15 columns. The following tables 1 display the attributes used in the dataset and their standards in unpolluted air.

13 Numeric Columns & 2 Categoric Columns.

Table 1- Table including all the attributes present in the dataset:

S.No	Attribute name
0	Date (DD/MM/YYYY)
1	Time (HH.MM.SS)
2	True hourly average concentration CO in mg/m^3
3	PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4	True hourly averaged overall Non-Metanic Hydro Carbons concentration in microgram/m^3 (reference analyzer)
5	True hourly averaged Benzene(C6H6) concentration in microgram/m^3 (reference analyzer)
6	PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7	True hourly averaged NOx concentration in ppb (reference analyzer)
8	PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9	True hourly averaged NO2 concentration in microgram/m^3 (reference analyzer)
10	PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11	PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12	The temperature in Â°C
13	Relative Humidity (%)
14	AH Absolute Humidity

5.1 Data pre-processing

After the data was extracted to achieve the specified goals, the dataset was read into a jupyter notebook using python code because it is flexible in handling large datasets. A python feature was being called to replace strings that have space with an underscore, the **duplicated().sum()** function was utilized to search for **duplicate values** in the imported data. The result shows that the **data has no duplicate values**. Many Null, missing were found, with the use of boxplot, outliers were detected from the dataset and this will be properly addressed in the modeling phase.

Also, the **NHHC_GT** column was dropped as it had 90% missing data.

5.2. Splitting training and test dataset

Separating dataset into training and testing datasets is an important part of evaluating data mining models. Typically, while separating a data set into a training dataset and testing dataset, most of the data is used for the training process, and a smaller portion of the data is used for testing. After a model has been made by using this training set, test the model by making predictions against the test Set. By using the same data for the training and testing process will minimize the data discrepancies effects of data and helps in a better understanding of the characteristics of the model. Table 3 shows the splitting of testing and training dataset for the air quality prediction.

Whole dataset	Data from March 2018 to April 2019
Training dataset	Data from March 2018 to December 2018
Test set	Data from January 2019 to April 2019

Figure 4 Splitting testing and training dataset

5.3 Summary of Numeric Attributes:

Thirteen numeric Attributes are present in the dataset:

CO_GT PT08_S1_CO NMHC_GT C6H6_GT PT08_S2_NMHC NOX_GT PT08_S3_NOX AH

NO2_GT PT08_S4_NO2 PT08_S5_O3 T RH

The describe() feature was called to find the **descriptive statistics.** Descriptive statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or median.

Find Max, min, mean and standard deviation of attributes:

Table 2 Max, min, mean and standard deviation:

	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	NOX_GT	PT08_S3_NOX	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	7344.000000	8991.000000	887.00000	8991.000000	8991.000000	7396.000000	8991.000000	7393.000000	8991.000000	8991.000000	8991.000000	8991.000000	8991.000000
mean	2.129711	1099.833166	218.607666	10.083105	939.153376	242.189292	835.493605	112.145137	1456.264598	1022.906128	18.317829	49.234201	1.025530
std	1.436472	217.080037	206.615130	7.449820	266.831429	206.312007	256.817320	47.629141	346.206794	398.484288	8.832116	17.316892	0.403813
min	0.100000	647.000000	7.000000	0.100000	383.000000	2.000000	322.000000	2.000000	551.000000	221.000000	-1.900000	9.200000	0.184700
25%	1.100000	937.000000	66.000000	4.400000	734.500000	97.000000	658.000000	77.000000	1227.000000	731.500000	11.800000	35.800000	0.736800

	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	NOX_GT	PT08_S3_NOX	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
50%	1.800000	1063.000000	145.000000	8.200000	909.000000	178.000000	806.000000	109.000000	1463.000000	963.000000	17.800000	49.600000	0.995400
75%	2.800000	1231.000000	297.000000	14.000000	1116.000000	321.000000	969.500000	140.000000	1674.000000	1273.500000	24.400000	62.500000	1.313700
max	11.900000	2040.000000	1189.000000	63.700000	2214.000000	1479.000000	2683.000000	333.000000	2775.000000	2523.000000	44.600000	88.700000	2.231000

5.4 Missing Values:

Using **isnull().sum()** feature, we were able to find the missing values in the dataset attributes.

```
Count of missing values:

CO_GT           1647
NMHC_GT         8104
NOX_GT          1595
NO2_GT          1598
```

- As we can see **three columns (CO_GT, NHMC_GT, NOX_GT, NO2_GT) contain missing values**.
- **Fill missing value strategy :**

- **CO_GT, NOX_GT, NO2_GT will be filled by monthly average of that particular hour.**
- **NHHC_GT will be dropped as it has 90% missing data.**

5.5 Criteria for Cleaning Missing Values:

If missing values are present in the data-

- We First do **Outlier Treatment** and then **Missing Data Imputation** because the outliers will also influence the missing data algorithms in a negative manner.

5.5 (a) Handling Outlier :

There are two type of outliers -

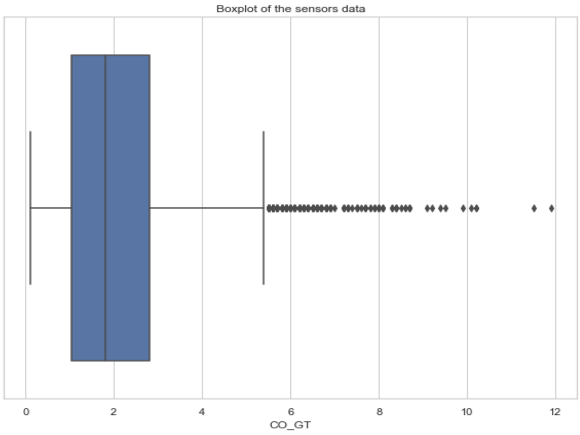
Univariate Outliers: Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

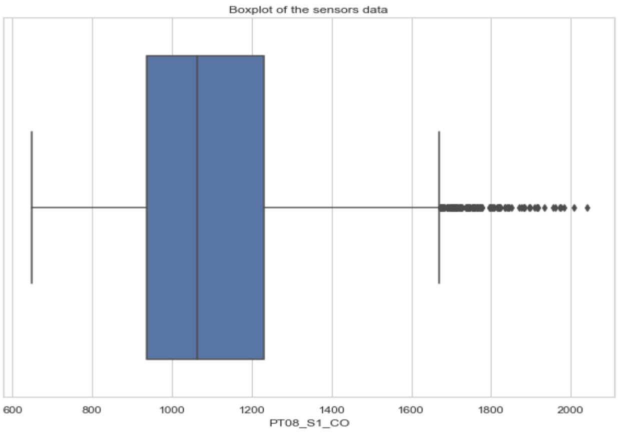
Multivariate Outliers: While plotting data, some values of one variable may not the beyond the expected range, but when you plot the data with some other variable, these values may be lie far from the expected value.

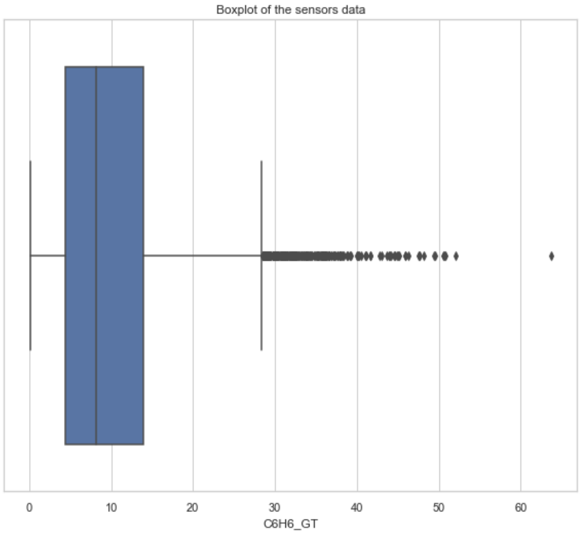
We **checked for outlier, treated them and then Missing data imputation** was performed in the end.

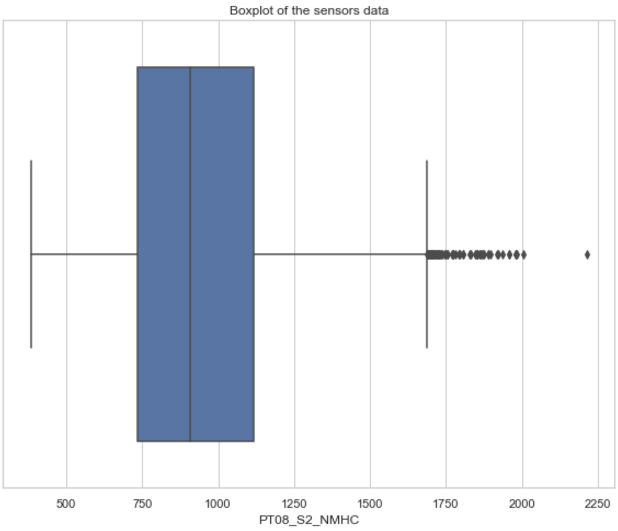
Outliers Detection and Imputation:

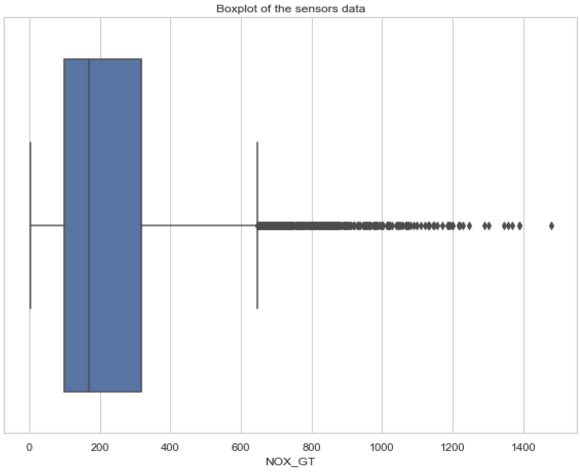
- Using Boxplot we detected Outliers in most of the variables (sensors data).

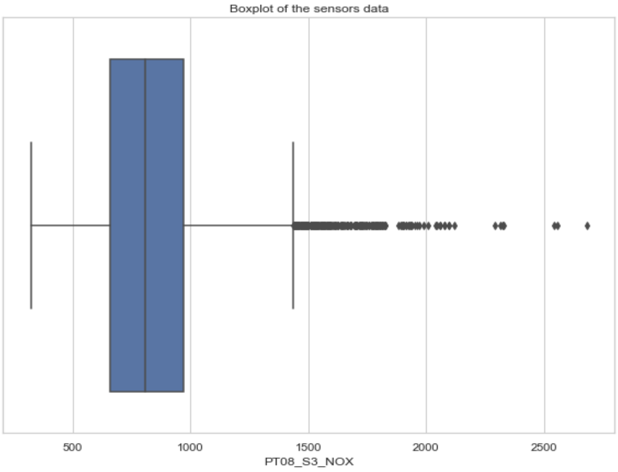


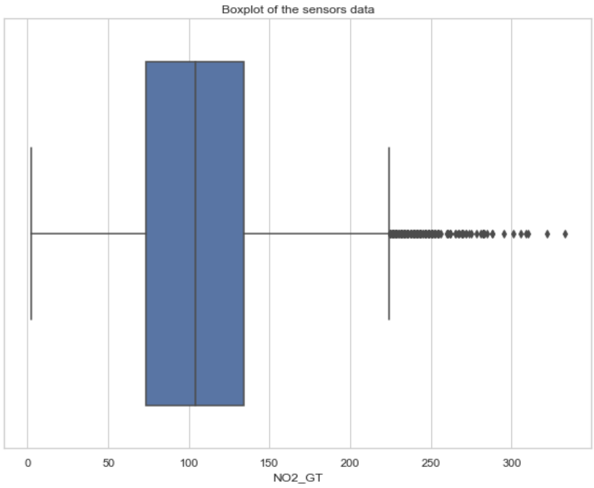


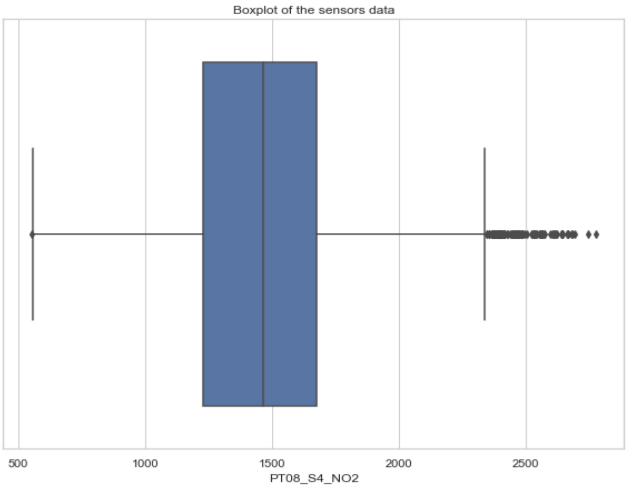

Boxplot of the sensors data
0
10
20
30
40
50
60
C6H6_GT


Boxplot of the sensors data
500
750
1000
1250
1500
1750
2000
2250
PT08_S2_NMHC


Boxplot of the sensors data
0
200
400
600
800
1000
1200
1400
NOX_GT


Boxplot of the sensors data
500
1000
1500
2000
2500
PT08_S3_NOX


Boxplot of the sensors data
0
50
100
150
200
250
300
NO2_GT


Boxplot of the sensors data
500
1000
1500
2000
2500
PT08_S4_NO2

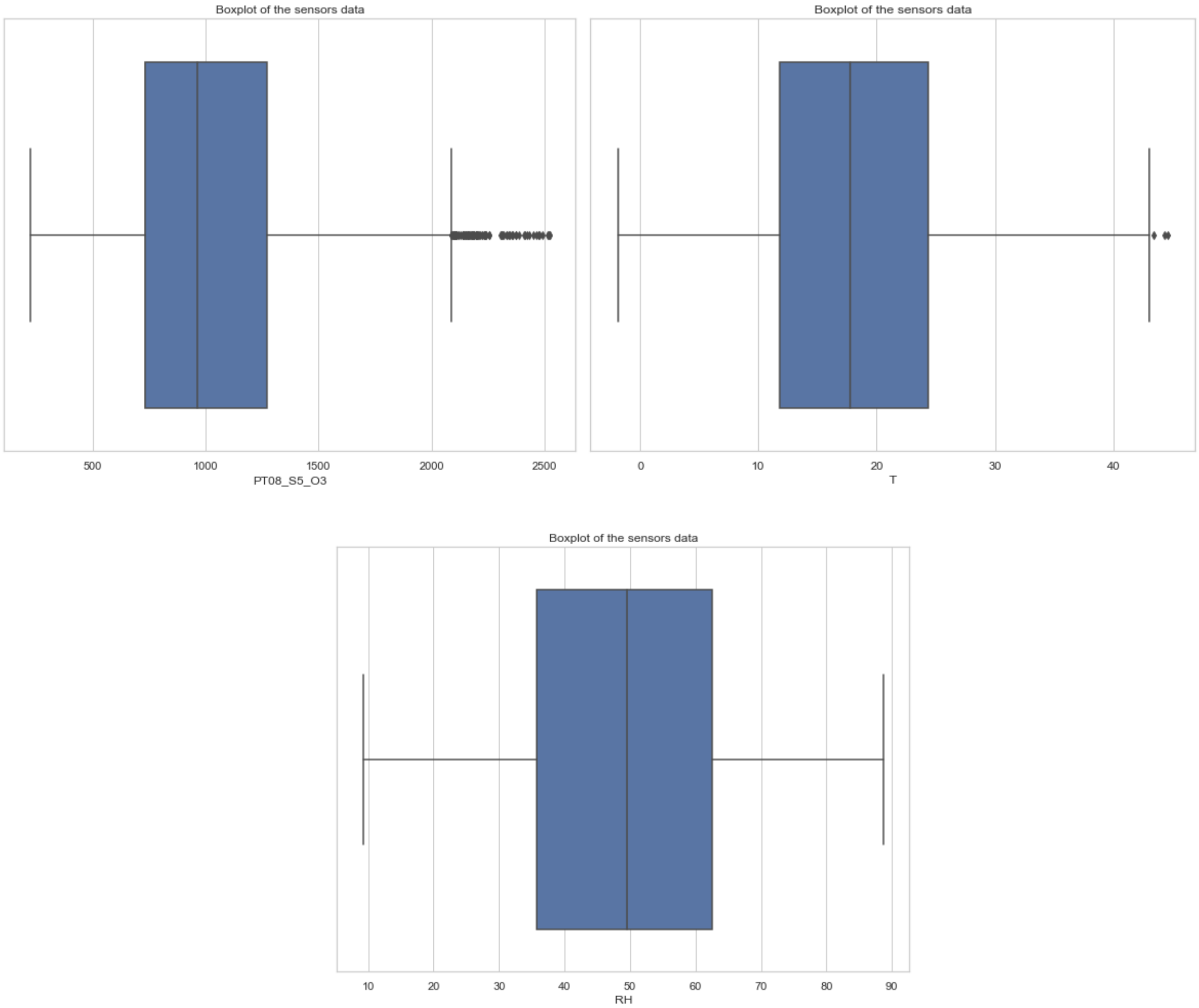


Figure 5 Outlier Detection of all variables (sensors data)

Outlier Treatment:

- We have used Interquartile Range Method (IQR) to treat/ remove Outliers from the above variables.

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has *2n / 2n+1* data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.
IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

Example:
Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

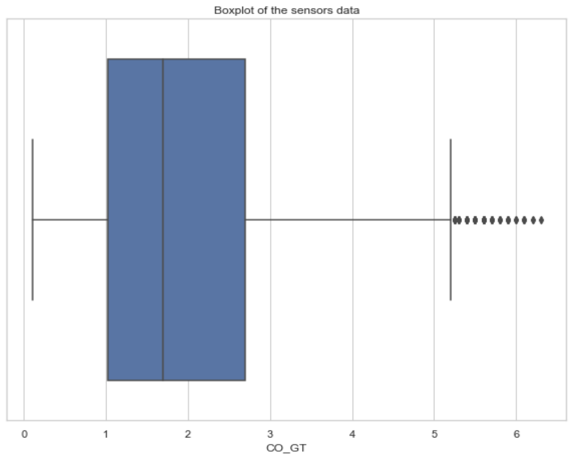
In our case,

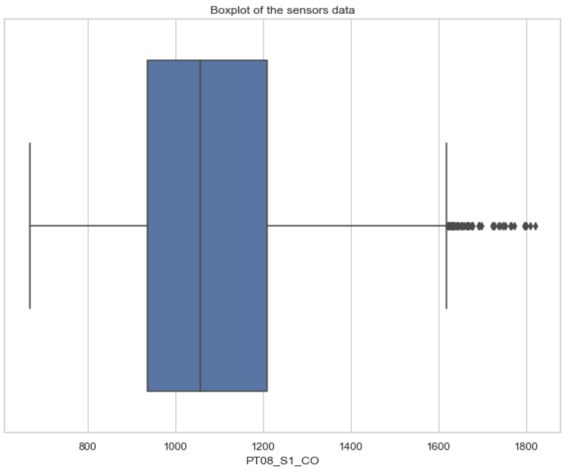
- We divided the data in **Q1** (first 25% of the data) and **Q3** (first 75% of the data).

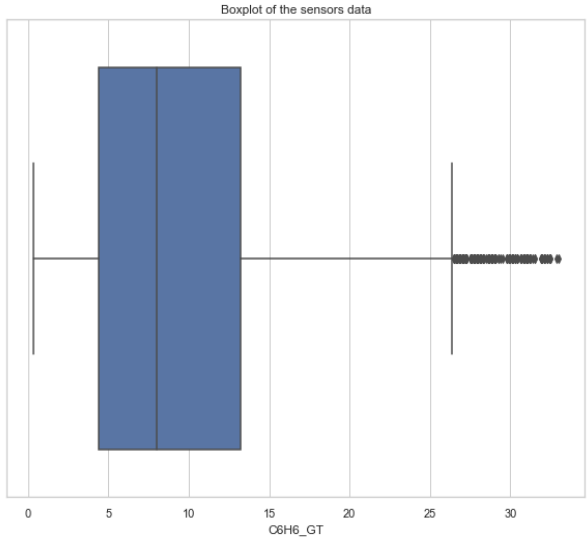
- Then we found IQR (Q3-Q1) and created new Data Frame without the outliers.

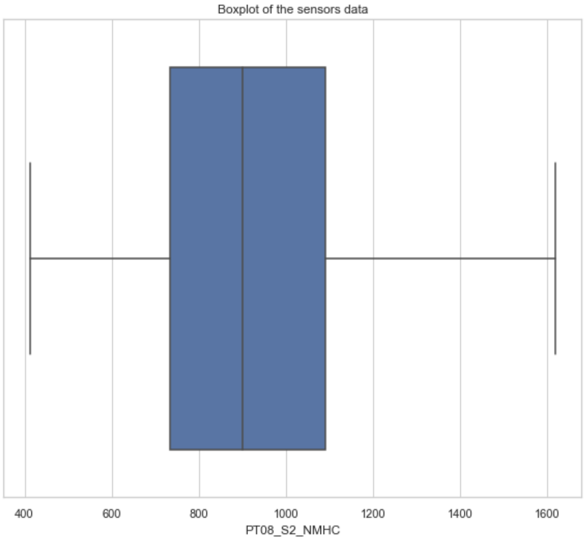
After Removing Outliers :

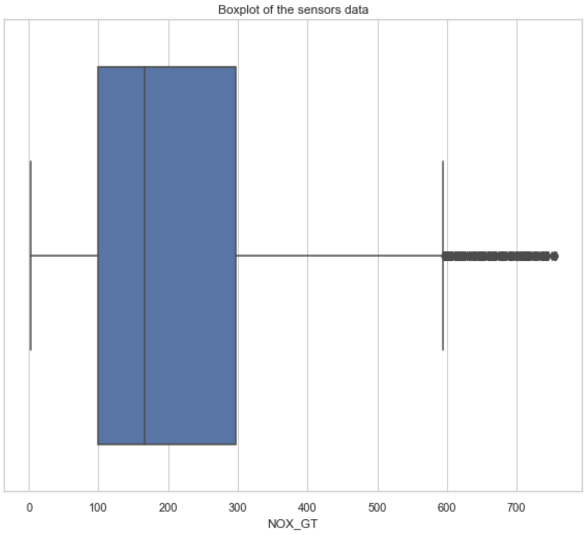
Boxplots Visualisation after **Outlier Treatment** and removing **NAN & Missing Values -**

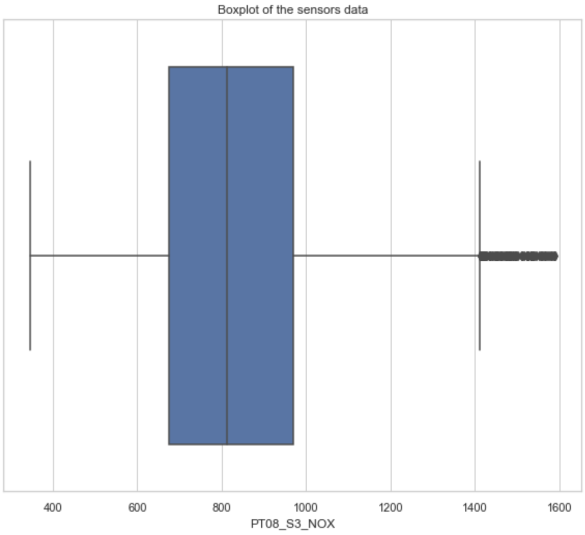


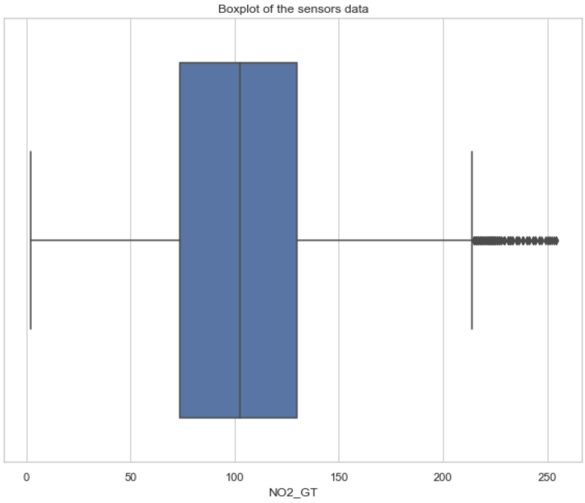


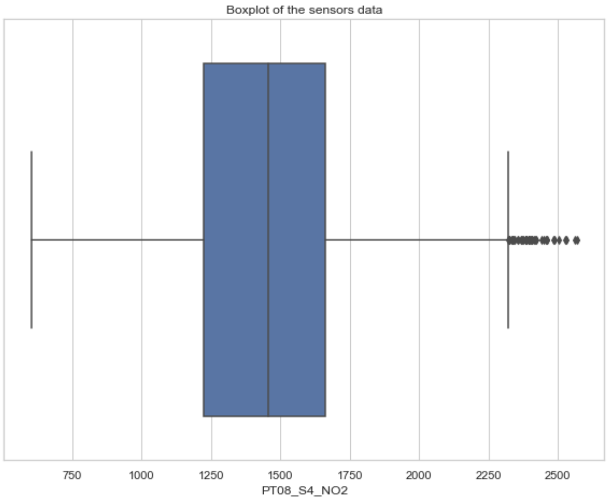


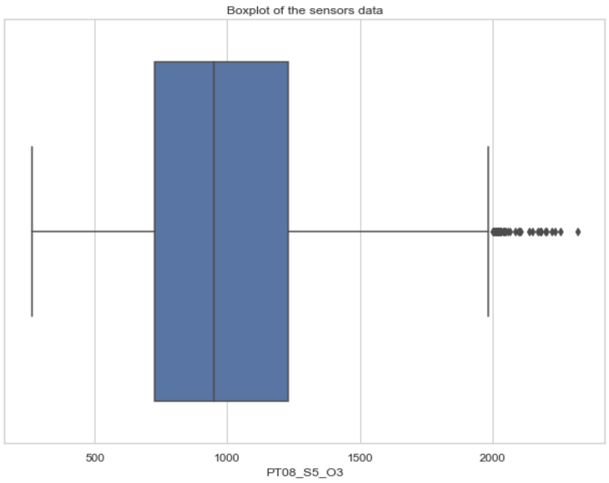


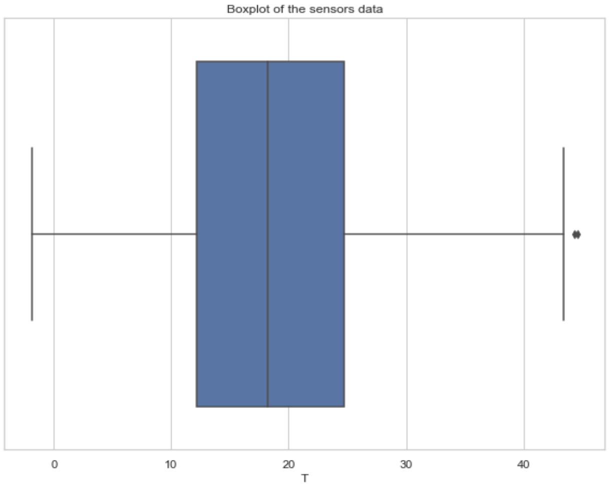

Boxplot of the sensors data
0
100
200
300
400
500
600
700
NOX_GT


Boxplot of the sensors data
400
600
800
1000
1200
1400
1600
PT08_S3_NOX


Boxplot of the sensors data
0
50
100
150
200
250
NO2_GT


Boxplot of the sensors data
750
1000
1250
1500
1750
2000
2250
2500
PT08_S4_NO2


Boxplot of the sensors data
500
1000
1500
2000
PT08_S5_O3


Boxplot of the sensors data
0
10
20
30
40
T

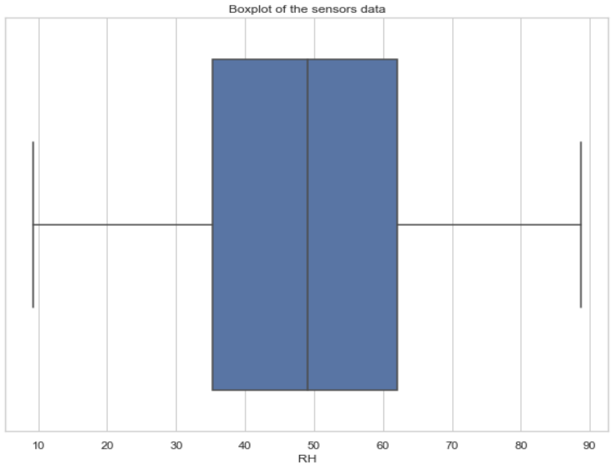


Figure 6 After Outlier Treatment - Boxplots with No outliers.

- With the help of visualisation of the boxplots given above we can say that, we have removed all **outliers, missing & Nan values** from the dataset successfully.

- So, **after imputation**, we see we get **0 missing values**. This is how we have removed all the missing values and Outliers.

6. Exploratory Data analysis

Exploratory Data Analysis on "Air Quality" dataset to understand the importance of EDA and classification model in the process of Machine Learning.

6.1 Graphs with explanation below:

Analyse the distribution of numeric attributes (normal or other). Plot histograms for attributes of concern and analyse whether they have any influence on the class attribute.

Histogram: It means a statistical view of our dataset. **The histogram is a pictorial representation of a dataset distribution with which we could easily analyse which factor has a higher amount of data and the least data.**

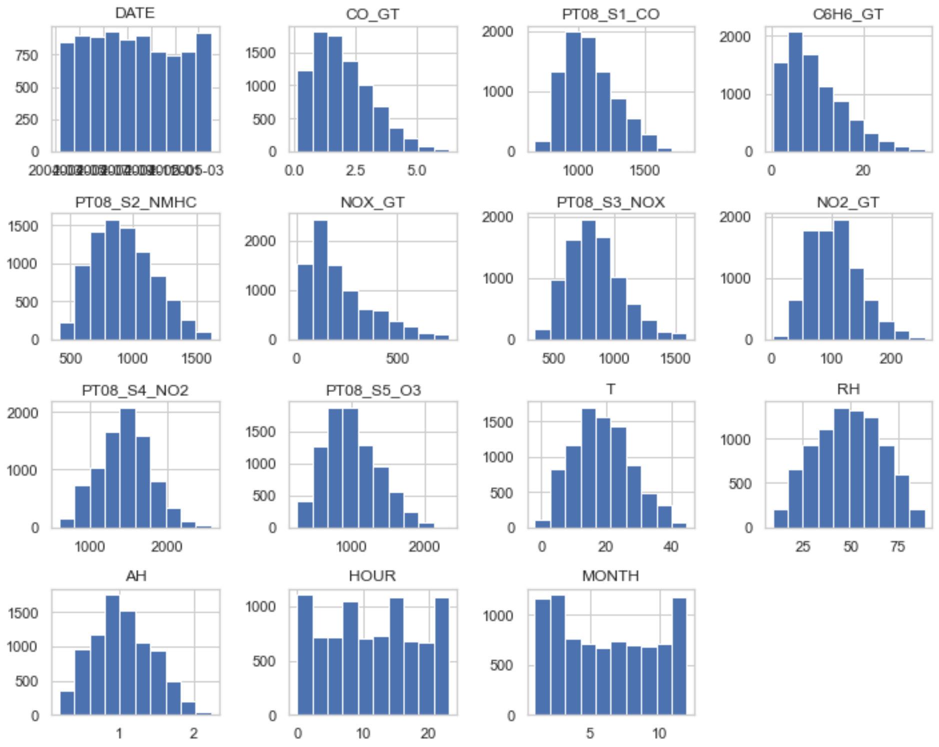


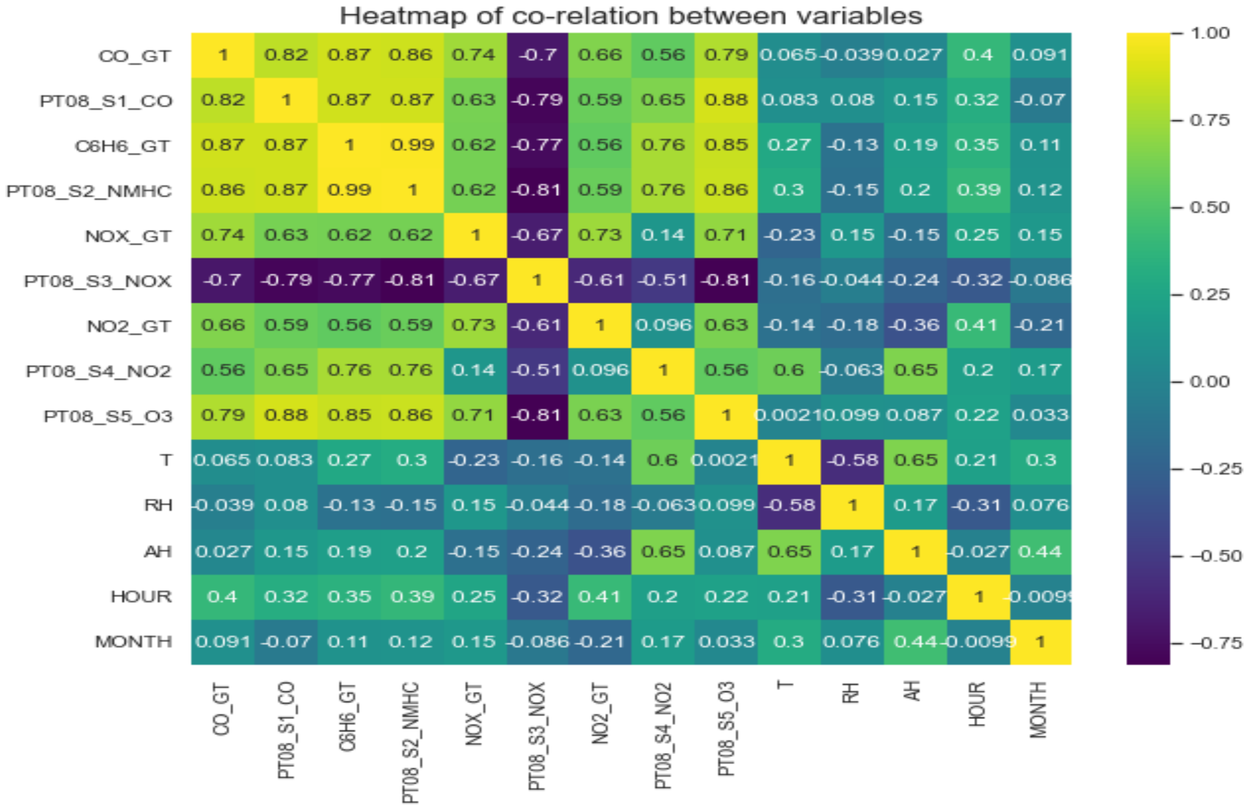
Figure 7 Histogram Plot of every Attribute

6.2 Correlation Matrix:

# Which attributes seem to be correlated? Which attributes seem to be most linked to the class attribute?

- A heat map is a two-dimensional representation of data in which values are represented by colours which provides a visual summary of information.

Heat map work with only numerical values.



	CO_GT	PT08_S1_CO	C6H6_GT	PT08_S2_NMHC	NOX_GT	PT08_S3_NOX	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH	HOUR	MONTH
CO_GT	1	0.82	0.87	0.86	0.74	-0.7	0.66	0.56	0.79	0.065	-0.039	0.027	0.4	0.091
PT08_S1_CO	0.82	1	0.87	0.87	0.63	-0.79	0.59	0.65	0.88	0.083	0.08	0.15	0.32	-0.07
C6H6_GT	0.87	0.87	1	0.99	0.62	-0.77	0.56	0.76	0.85	0.27	-0.13	0.19	0.35	0.11
PT08_S2_NMHC	0.86	0.87	0.99	1	0.62	-0.81	0.59	0.76	0.86	0.3	-0.15	0.2	0.39	0.12
NOX_GT	0.74	0.63	0.62	0.62	1	-0.67	0.73	0.14	0.71	-0.23	0.15	-0.15	0.25	0.15
PT08_S3_NOX	-0.7	-0.79	-0.77	-0.81	-0.67	1	-0.61	-0.51	-0.81	-0.16	-0.044	-0.24	-0.32	-0.086
NO2_GT	0.66	0.59	0.56	0.59	0.73	-0.61	1	0.096	0.63	-0.14	-0.18	-0.36	0.41	-0.21
PT08_S4_NO2	0.56	0.65	0.76	0.76	0.14	-0.51	0.096	1	0.56	0.6	-0.063	0.65	0.2	0.17
PT08_S5_O3	0.79	0.88	0.85	0.86	0.71	-0.81	0.63	0.56	1	0.0021	0.099	0.087	0.22	0.033
T	0.065	0.083	0.27	0.3	-0.23	-0.16	-0.14	0.6	0.0021	1	-0.58	0.65	0.21	0.3
RH	-0.039	0.08	-0.13	-0.15	0.15	-0.044	-0.18	-0.063	0.099	-0.58	1	0.17	-0.31	0.076
AH	0.027	0.15	0.19	0.2	-0.15	-0.24	-0.36	0.65	0.087	0.65	0.17	1	-0.027	0.44
HOUR	0.4	0.32	0.35	0.39	0.25	-0.32	0.41	0.2	0.22	0.21	-0.31	-0.027	1	-0.009
MONTH	0.091	-0.07	0.11	0.12	0.15	-0.086	-0.21	0.17	0.033	0.3	0.076	0.44	-0.0099	1

Figure 8 Correlation matrix to find correlated attributes.

- By seeing these correlation we conclude that **Target variable : C6H6_GT** is **highly correlated** to **PTO8.S2_NMHC**, so we take **PTO8.S2_NMHC as independent parameter** to **predict the dependent parameter i.e., C6H6_GT**

7.0 Classification Models & Machine learning techniques :

Before applying machine learning techniques, we prepared the dataset:

1. **With the help of the Correlation Matrix we took PTO8.S2_NMHC as independent parameter as X And C6H6_GT as taking dependent parameter as Y.**
2. **We splitted the train (X_train, Y_train) and test (X_test, Y_test) data.**
3. **Giving 10 samples for training and 90 samples for testing this is very quiet interesting to do because we only have 10 sample out of 100 for training and 90 samples for testing.**

For designing the model for predicting Quality of Air:

We applied **Linear Regression**, **Decision Tree**, **Lasso Regression**.

We tested on test to find MSE, RMSE and R^2 scores from different algorithms.

Air quality prediction Regression Evaluation Metrics :

Air quality is predicted using the **R squared** value. R square determines the proportion of variance in the dependent variable of the system that can be explained by the independent variable. It is a statistical measure in a regression model. It is also called a coefficient of determination. The predicted **R square** values indicate how well a regression model predicts responses for the given observations. R square value generally lies between -1 to +1. In this project, R square value for training and test dataset is calculated using four different regression models. Here in this project R square value of the training dataset is always greater than the test data. If the R square value is near to 1, then the regression model is better for than dataset.

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

Root Mean Square Error (RMSE) -

Root Mean Square Error is the standard deviation (SD) of the prediction errors. Residuals are the measure of how far from the regression line data points are; RMSE tells you how the data is concentrated around the best fit line or a measure of how the residuals are spread out. It is commonly used in forecasting, climatology, and regression analysis to verify experimental results.

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

Where f = forecasts (unknown results) and o = observed values (known results).

7.1 Experimental Designs:

The research would include a qualitative comparison of three machine learning models which will be considered as the strategies to be used in this research for the proper implementation of the models outlined for this report. The machine learning models can also be referred to as machine learning classifiers or classification approaches in the study for greater understanding and clarification since the study deals with a classification problem. The algorithms are;

Linear regression

Linear regression is a basic and best-used type of predictive analysis. Linear regression is used to examine two things; namely, it checks whether a set of predictor variables is doing a good job in predicting an outcome (dependent) variable And checks which variables, in particular, are the significant predictors of the outcome variable.

$$Y = c + b*x$$

Decision Tree

Decision tree regression is a supervised data mining model used to predict a target by learning decision rules from features. A decision tree is constructed using recursive partitioning starting from the parent or root node. Each node can be split into the left and right child nodes. These nodes can be further split themselves to become parent nodes of their resulting children nodes.

$$S(T,X) = \sum_{c \in X} P(c)S(c)$$

Lasso Regression

Lasso regression is similar to linear regression but uses shrinkage. Shrinkage is where data values are shrunk toward a point as they mean. The lasso procedure encourages simple, sparse models. This regression is well suited for models showing high levels of multicollinearity or if you want to automate certain parts of model selection, like variable selection/parameter elimination.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

7.2 Model Implementation

Applying Various Models –

- **Results after applying all three classification:**

1) Linear regression:

Evaluating Metrics –

```
MSE value for LinearRegression model is 1.1097
RMSE value for LinearRegression model is 1.0534
R^2 value for LinearRegression model is 97.2476
```

2) Lasso Regression Model:

Evaluating Metrics:

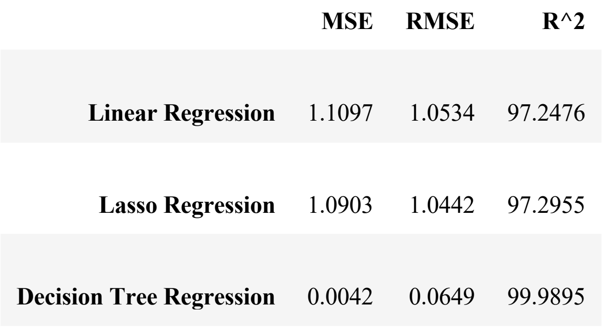
```
MSE value for LassoRegression model is 1.0903
RMSE value for LassoRegression model is 1.0442
R^2 value for LassoRegression model is 97.2955
```

3) Decision Tree Model:

Evaluating Metrics:

```
MSE value for Decision Tree Regression model is 0.0042
RMSE value for Decision Tree Regression model is 0.0649
R^2 value for Decision Tree Regression model is 99.9895
```

7.3 Final Accuracy Table:



	MSE	RMSE	R^2
Linear Regression	1.1097	1.0534	97.2476
Lasso Regression	1.0903	1.0442	97.2955
Decision Tree Regression	0.0042	0.0649	99.9895

Result: By seeing the table we conclude that Decision Tree Regression model is giving More accuracy out of all these model so it will be considered as the best model for evaluating the Quality of Air.

8.0 Conclusion from Modelling & Research answers:

Q: Find the classification model that can be used in the future as a predictive model, that will result in higher accuracy?

In comparison with all 3 classification algorithms **Decision Tree Regression** outperformed **Linear Regression** and **Lasso Regression** classification by achieving the highest accuracy among all of the three classification algorithms.

So Decision Tree Regression model will be considered as the best predictive model for evaluating the Quality of Air.

Q: Find the variable highly correlated with the target variable. Decide independent parameter to predict the dependent parameter for modelling **?**

By seeing the correlation we conclude that **Target variable : C6H6_GT** is **highly correlated** to **PTO8.S2_NMHC**, so we take **PTO8.S2_NMHC as independent parameter** to **predict the dependent parameter i.e., C6H6_GT**

Github Repository Link:

https://github.com/haritikajolly/Air-Quality-Prediction

9.0 References:

1 [1] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu "Detection and Prediction of Air Pollution using Machine Learning Models." (IJETT) – Volume 59 Issue 4 – May 2018

[2] Baldasano J.M, Akita Y "Large scale Air pollution estimation method combining land use regression modeling in geostatistical framework," Environmental Science & Technology, vol. 48, no. 8, 2014

[3] GnanaSoundari.A Mtech, (Ph.D.), Mrs. J.GnanaJeslin M.E, (Ph.D.), Akshaya A.C. "Indian Air Quality Prediction And Analysis Using Machine Learning." ISSN 0973-4562 Volume 14, Number 11, 2017

[4] McCollister G.M. and Wilson K.R. (2008), "Linear regression model for forecasting daily maxima and hourly concentrations of air pollutants," Atmospheric Environment.

[5] Rao, S.T., and Zurbenko, I.G.(2014)."Detecting and Tracking Changes in Air Quality using regression analysis". J. Air Waste Manage. Assoc. 44: 1089–1092.

[6] RuchiRaturi, Dr. J.R. Prasad "Recognition Of Future Air Quality Index Using Regression and Artificial Neural Network" IRJET .e-ISSN: 2395- 0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[8] https://en.wikipedia.org/wiki/Airpollution https://www.activesustainability.com/environment/effects-air-pollution-human-health/ https://www.geeksforgeeks.org/working-csv-filespython/

- Stackoverflow
- GeeksforGeeks
- Machinelearningmastery
- Stackexchange

Scikit-learn Documentation

The 5 Classification Evaluation metrics every Data Scientist must know

The Python Graph Gallery - Grouped Bar Plot