**Predicting Ozone Layer Concentration using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree**

**CSE3001 -- SOFTWARE ENGINEERING**

**J-Component Slot : F2**

**FACULTY : Prof . SWARNALATHA P**

**School of Computer Science and Engineering**



**Group Members**

|  |  |
| --- | --- |
| **Y.Amartya** | **16BCE0033** |
| **K.Hemanth** | **16BCE0126** |
| **W. Arrshi** | **16BCE0146** |

# CERTIFICATE

This is to certify that the project work entitled “Predicting Ozone Layer Concentration using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree” that is being submitted by us for “Software Engineering (CSE3001) is a record of bonafide work done. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted for any other CAL course. Thankyou.

Place: Vellore

Date: 03-04-19

Signature of Faculty Advisor

(Dr. Swarnalatha P)

# ACKNOWLEDGEMENTS

We would like to acknowledge our mentor Prof. Dr. Swarnalatha P for providing us the opportunity to perform this experiment as part of the J component. We would like to thank her to give us a chance to apply our knowledge on practical footings making the study topics all the more interesting and easy to grasp. Also we would like to thank the Dean and the University Management to provide us with a platform that enabled this. We are grateful to the Computer lab assistants to allow us to do the project under their experienced supervision.

**ABSTRACT**

Air pollution is one of the major environmental worries in recent time. Abrupt increase in the concentration of any gas leads to air pollution. The cities are mostly affected due to the abundance of population there. One of the worst gaseous pollutants is OZONE (O3). In this project, we propose three predictive models for estimation of concentration of ozone gases in the air which are Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree. Evaluation of the prediction models indicates that the Multivariate Adaptive Regression Splines model describes the dataset better and has achieved significantly better prediction accuracy as compared to the Random Forest and Classification And Regression Tree. A detailed comparative study has been carried out on the performances of Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree. MARS gives the result by considering less variables as compared to other two. Moreover, Random Forest takes a little more time for building the tree as the elapsed time was calculated to 45 second in this case. In addition variable importance for each model has been predicted. Observing all the graphs Multivariate Adaptive Regression Splines gives the closest curve of both train and test set when compared. It can be concluded that multivariate adaptive regression splines can be a valuable tool in predicting ozone for future.

**Keywords** - Ozone, Multivariable Adaptive Regression Spline, Random Forest, Classification And Regression Tree.

**INTRODUCTION**

Urban atmospheric pollutants are increasing day by day. They are considered as one of the main causes of increased incidence of respiratory illness in citizens. It is now irrefutable that air pollution is being caused by large amount of Total Suspended Particulates (TSP) and respiratory particulate of Particulate Matter less than 10µm in aerodynamic diameter that has numerous undesired consequences on human health . Air pollutants in an area with good airflow quickly mixes with air and disperses however when trapped in an area, pollutant concentration can increase rapidly which ultimately leads to degradation of air quality. In order to measure how polluted the air is Air Quality Index is examined while for properties of air we see Qualities of Air . All these factors affect the ozone layer which is Earth’s protective layer. It is a belt of naturally occurring ozone gas that sits 9.3-18.6 miles above Earth and serves as shield from harmful ultraviolet B radiation emitted by sun . Several steps like Montreal Protocol which declines emission of ODS (ozone depleting substances) have been taken. It is expected to result in a near complete recovery of ozone layer near the middle of 21st century. By 2012, the total combined abundance of anthropogenic ODS in the troposphere has decreased by nearly 10% from its peak value in 1994 . In the present era, there is a wide spread concern for ozone layer depletion due to the release of pollution. As particulate matter causes several kind of respiratory and cardiovascular disease, it also leads to ozone depletion which attracts more and more attention for air quality information. This shows the need for integration of different information system in a similar way as done by Birgersson et al. (2016) for data integration using Machine Learning . Prediction of air quality has thus become a necessary need to save the future. Machine learning has been applied in various fields . Medical and other fields have also been covered by various classification techniques . Just as application of rough set technique was done for data investigation by Roy et al. (2013), in this project application of Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree techniques has been applied for predicting the concentration of ozone . Chuanting Zhang and Dongfeng Yuan (2015) worked on Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark . Previously existing methods could not meet the demand of real time analysis so a distributed random forest algorithm is implemented using Spark on the basis of resilient distributed dataset and shared variable. Parallelized random forest is also used as prediction model. Estimation of benzene by on field calibration of an electronic nose has been carried by Vito et al. (2007) in which gas multi-sensor played an important role which helps to raise the density of the monitoring network. But their concentration estimation capabilities are seriously limited by the known stability and selectivity issues of solid-state sensors they often rely on. Sensor fusion algorithm used in regression need to be properly tuned via supervised learning. But this training was revealed to be unsuccessful . Forecasting and prediction of things has become an essential part for future life. Roy et al. (2015) worked on prediction of Stock Market Forecasting using Lasso Linear regression model . Vito et al. (2009) worked on CO, NO2 & NOx urban pollution monitoring. Some authors have used gas multisensor devices as a tool for densening the urban pollution monitoring mesh due to the significantly low cost per unit . But the drawback is that these sensors are not reliable for long term and selectivity issues. In this project we concentrate on regression technique MARS (Multiple Adaptive Regression Spline) for Air Quality dataset. Hui et al. (2013) used this regression model for prediction of emission of CO2 in ASEAN countries. A comparative study of multiple regression (MR) and multiple adaptive regression splines (MARS) was carried for statistical modelling of CO2 over period of 1980-2007. MARS model was concluded as more feasible and with better predictive ability.This project shows the comparison of regression techniques like Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree on Air Quality data showing the prediction using Salford Predictive Modeller. This project is organised as follows. Section 2 overviews proposed techniques of Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree. Section III gives the experimental setup and the steps involved in performing the regression techniques on the given dataset. Section IV displays the results and discussion. Section V concludes the project.

**PROPOSED METHODOLOGY**

To work with Salford modeller it is necessary to know the working of the regression techniques that are going to be used. All are type of machine learning like a computer program is said to learn from experience ‘E’ with respect to some class of tasks ‘T’ and performance measure ‘P’ if its performance at tasks in ‘T’ as measured by ‘P’ improves with experience ‘E’. All these have been used for prediction of ozone concentration by extracting knowledge from dataset.

**Random Forest algorithm :**

It is a tree-based ensemble learning method involving the combination of several models to solve a single prediction problem. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method. It may also be said as a collection of many CART trees that are not influenced by each other when it is constructed . It works as a large collection of decor related decision trees. It comes under bagging technique (average noisy and unbiased models to create a model with low variance). Tress are based on random selection of data as well as variable. This develops lots of decision trees based on random selection of data and random selection of variables. After all the tree are built the data get fed in the tree and proximities are calculated for each pair of cases. If any two cases occupy the same terminal node, their proximity is changed and incremented by one. At last, the proximities get normalized by dividing it by the number of trees. Proximities can be used in replacing missing data, locating outliers, and producing illuminating low dimensional views of the data. It serves as one of the most useful tools in random forests. The proximities originally form an NxN matrix. After a tree is built, both training and test data are pulled down the tree. At the end, the proximities get normalized by dividing by the number of trees. Since the large data set could not fit an NxN matrix into fast memory, a modification reduced the required memory size to NxT where T stands for the number of trees in the forest. In order to speed up the computation-intensive scaling and iterative missing value replacement, the user is provided with the option of retaining only the nrnn largest proximities for each case. When the dataset is presented, the proximities of each case in the test set with each case in the training set can also be computed and compared. The amount of additional computing is moderate. The dataset contains thousands of data from which concentration of ozone is predicted. Thus Random Forest is useful in handling thousands of input variables without variable deletion. Hence Random Forest gives variable importance to each and every variable involved.

**Classification And Regression Tree algorithm :**

Classification Regression Tree was introduced by Breiman et al. (1984) for classification or regression predictive modelling problems. It is often referred as ‘Decision Tree’ but now named as CART in modern software. It provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees. It is a binary tree that splits a node into two child nodes repeatedly beginning with the root that contains whole learning sample. Say for x being a nominal categorical variable of I categories, there are 2I-1-1 possible splits for the predictor. If X is an ordinal categorical or continuous variable with K different values there are K-1 different splits on X. At any node say t, the best split s is chosen to maximize a splitting criterion ∆i(s, t) . There are 3 splitting criteria available.

Gini Criterion

The impurity measure at a node t is defined as

i(t)= ∑\_(i,j)▒〖c(i│j)p(i│t)p(j|t)〗 (1)

It is the decrease of impurity given by

∆i(s,t)=i(t)-pL i(tL) - pR i(tR) (2)

Where pL and pR are probabilities of sending case to left child node and right child node where

pL=p(tL)/p(t) (3)

And

pR=p(tR)/p(t) (4)

CART does not require any special data preparation other than a good representation of the problem.

**Multivariate Adaptive Regression Splines algorithm**

It is a form of regression analysis developed by Friedman in 1991 with the aim to predict dependent variable from set of independent variable. It is simpler than other models like random forest and neural network. It is seen as an extension of linear models that automatically models non linearity interaction between in variables. Mars is not affected by outliners. It produces model that can be written as an equation. It models both classification and regression tasks. It accepts a large number of predictor and chooses important predictor variable. The extensive use of MARS model can be done for prediction as it has been done for concentration of ozone in this project. It is used for prediction and classification problems (Islam et al.2015, Zhang et al.2015). The details of the MARS model can be observed through a website by Salford Systems. Also, this regression is influenced by the recursive partitioning method for which any criteria can be chosen for the selection of basis function of multivariate spline. One of the advantage of the mars model is that MARS can reduce the outliers. The proposed MARS forms the model with the use of two sided truncated functions of the predictor x . Equation works as a basis function for linear and non-linear functions. Also, this equation works to approximate any function f(x).In MARS, let us assume that the dependent variable (output) is y and number of terms is M.

In the equation MARS works over M term. The terms β\_0 , β\_m are the parameter. Hinge function i.e the H can be written as the following equation,

In the above equation (8) the product of kth of the mth term is given as, x\_v(k,m)

The value of K=1 and K=2 gives additive and pairwise interaction respectively. For this work, the opted value of K is 2.

**PERFORMANCE ANALYSIS**

The experiment to predict concentration of ozone is carried out by using software named ‘Salford Predictor Modeller 7.0’ founded in 1983 which supports the three techniques Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree.

1. **DATASET**

The dataset is of East zone of Uttar Pradesh , which includes cities like Kanpur , Murtaza nagar , akhbarpur .

Over 9000 values of various gases have been recorded over the span of 1 year . The starting date of data is 10-03-2016 and the final date of data recording is 04-04-2017 The gases recorded are - CO , PT08.S1( tin oxide) , NMHC ( non metanic hydrocarbon ) , C6H6 , PT08.S2(Titania), NO , PT08.S3 (Tungsten oxide) , NO2 , PT08.S4(tungsten oxide ) , PT08.S5 (Indium oxide), TEMP , Relative humidity absolute humidity.

1. **Air Quality Prediction Steps**

This describes the various steps taken for prediction by Salford Modeller.

Step 1: The database is opened in the software as it supports all type of file.

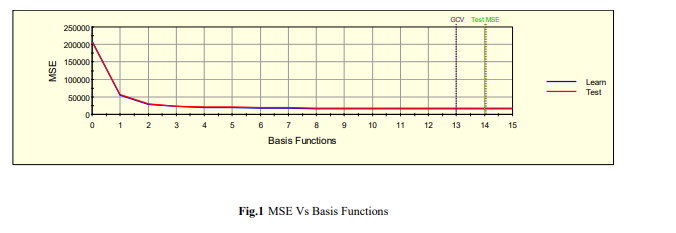
Step 2: The model is designed by selecting the predictor. A total of 12 predictors are selected for MARS/Random Forest/CART for this dataset. Date and time are not chosen as they don’t have any effect. PT08\_S5\_O3\_ is set as the target in all the cases.

Step 3: The analysis method is selected as MARS/Random Forest/CART with analysis type being ‘regression’ in all three cases.

Step 4: It’s time to separate the dataset as learning set and test set. This is done by selecting Fraction of cases selected at random for testing by assigning any value. Remember the values are in terms of percent. Here we put the test set as 0.30. Step 5: Now the model is started and resulting graph pops up showing the information required for future prediction of target variable. It also provides summary for all other details which is discussed in the next section.

**RESULT AND DISCUSSION**

In this section we compare the results given for the target variable PT08\_S5\_O3\_ through Random Forest, Multivariate Adaptive Regression Splines and Classification And Regression Tree by Salford Predictive Modeller. Out of 15 attributes, 12 are being used for used as predictors while 1 is selected as the targeted variable and the targeted variable being PT08.S5 (O3). 30% of the 9358 instances are selected for test case while rest go in for the learn case. This project contains the graphical representation of the learn and test value, summary of important terms, list of variable importance by the three models used. On applying Multivariate Adaptive Regression Splines, we get Fig 1. which shows the graph shows result of learn and test case where the Y-axis represents MSE with an interval of 50000 and X-axis representing Basis functions which was taken as 15 initially. From the graph conclusion can be made that there are least error as both the learn and test cases are same. Initially the MSE value starts from 200000 drops till 5000 and gradually becomes constant.

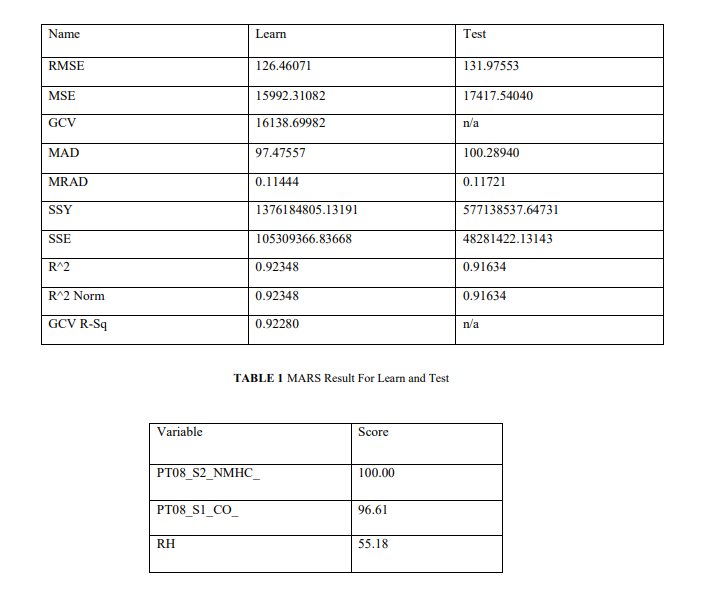


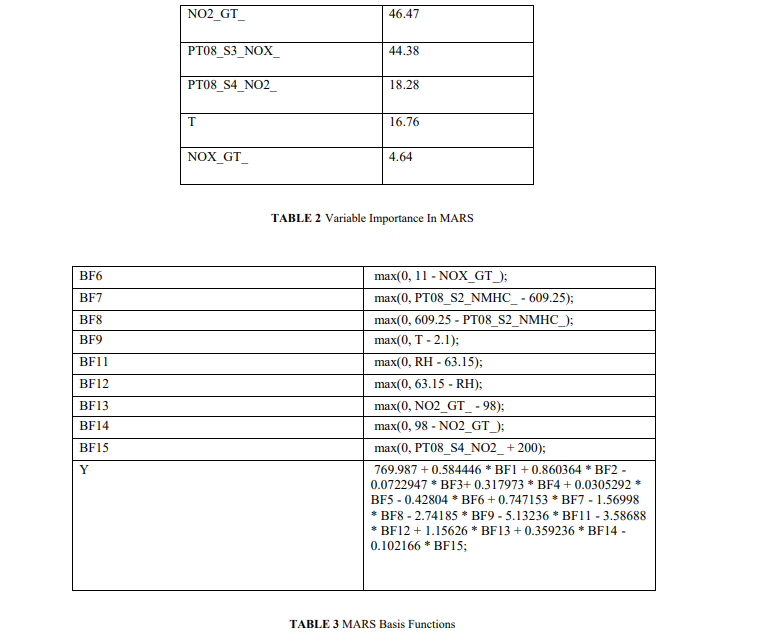
There are several important parameters that give the model error measure. These important parameters have been listed in table 1 showing their value for both learn and test. The variables include RMSE, MSE, GCV, MAD, MRAD, SSY, SSE, R^2, R^2 Norm, GCV R-Sr. Out of 15 attributes, 12 were set as predictors but after the regression model was prepared it was deducted that only 8 variables were important for prediction of PT08.S5 (O3). The most important variable was found to be PT08\_S2\_NMHC. The scores of all the variables are given in decreasing order of their importance in predicting the predictor in table 2. The number of basis function was set as 15 initially. The model assigns special variables to make a new equation to cover all points of nonlinearity. These variables are termed as basis variables. The model is a weighted sum of basis function. Each basis function takes one of the following forms

1) Constant. Only one term i.e. the intercept

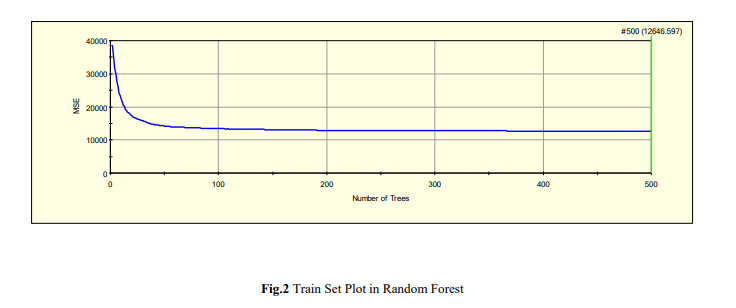
2) Hinge function

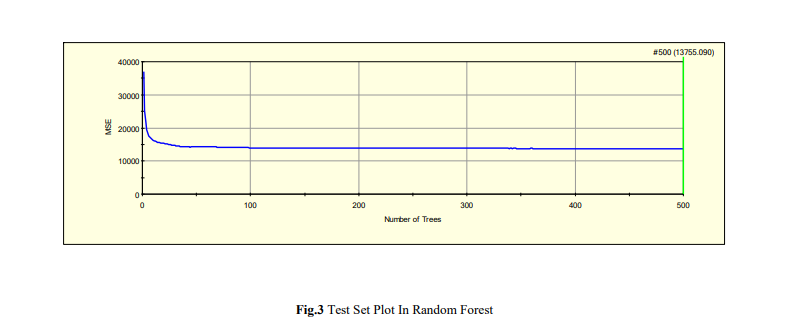
3) Product of 2 or more Hinge function. Table 3consists all the basis function and their combination to give the final equation of Y.



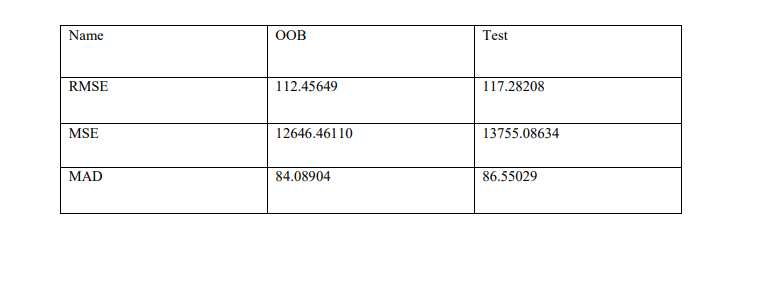


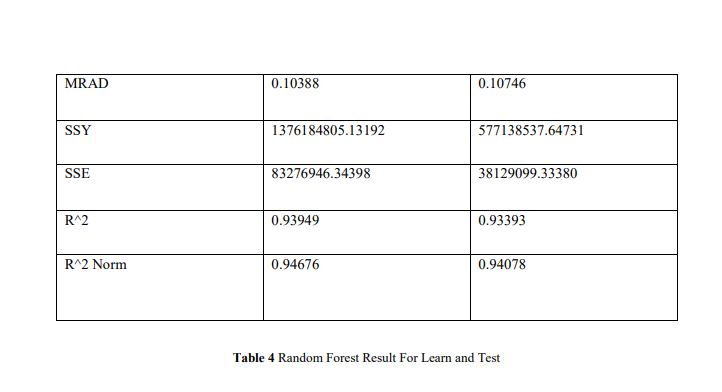
Random Forest was started by setting the number of trees to be built as 500 and number of predictors for each node as 3. The frequency of the report was set 10 along with parent minimum case as 2. The elapsed time was nearly 45 second for creating the trees. Separate graphs were obtained showing the comparison of train as in fig.2 and test cases as in fig.3 with 500 trees having maximum terminal node of 3293. Observing the curve in both the cases, a clear difference can be seen in the curve. Both start with MSE 40000 but the train set is less steep when compared with test set. Train set shows a turn at 16th tree (18502.625) while test set shows a turn at 8th tree (16911.168). Both the graph for train and test are given in fig.2 and fig.3 respectively.

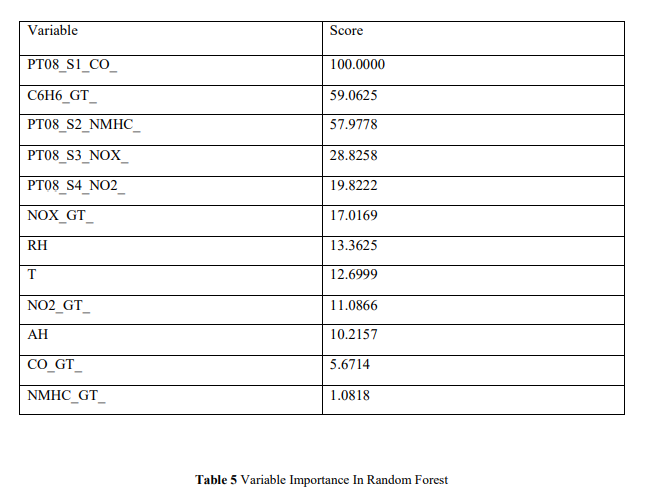




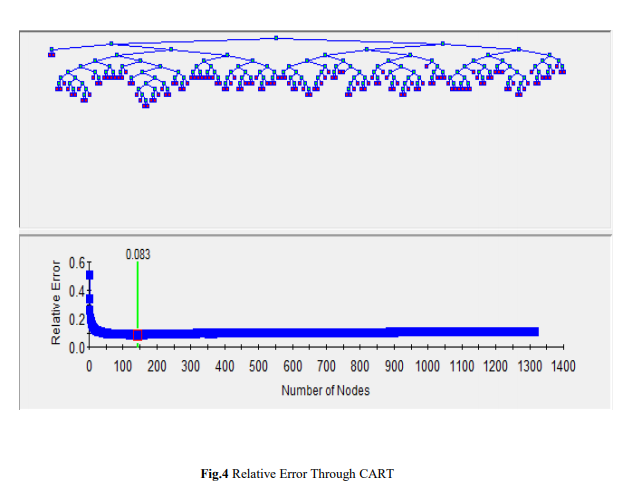
There are several important parameters that give the model error measure. These important parameters have been listed in table 4 showing their value for both learn and test. The variables include RMSE, MSE, MAD, MRAD, SSY, SSE, R^2, R^2 Norm. Unlike MARS in random forest all 12 variables have their own importance and show their contribution for building trees. Importance of each variable has been shown in table 5.



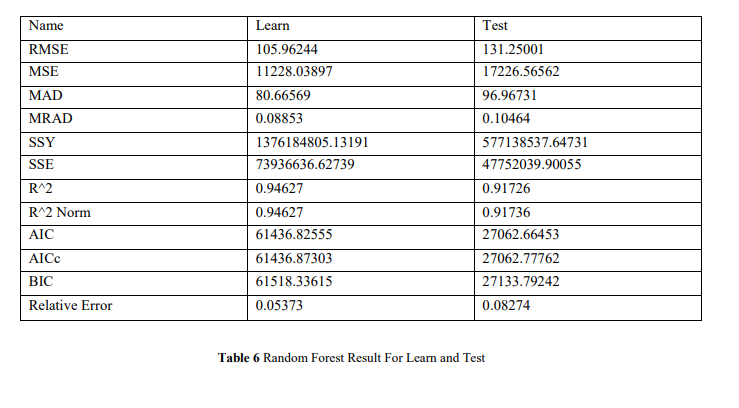


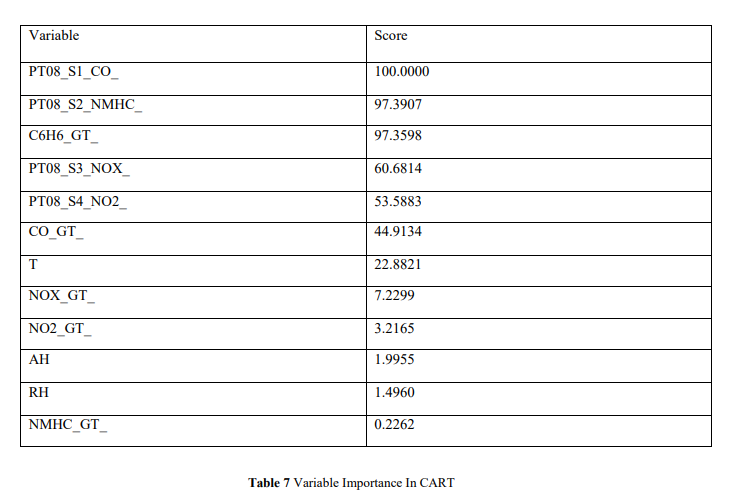


Classification And Regression Tree model also leads in building up of trees. It gives a graph where the Y-axis shows Relative Error while X-axis shows the number of nodes. The graph value shows the relative error value as 0.083 at 150th node. So by examining the graph directly we can get the relative error of test from the train set as shown in fig.4



There are several important parameters that give the model error measure. These important parameters have been listed in table 6 showing their value for both learn and test. The variables include RMSE, MSE, MAD, MRAD, SSY, SSE, R^2, R^2 Norm, AIC, AICc, BIC, Relative Error. CART also requires the use all 12 predictor variables just as Random Forest does. Table 7 lists the variable according to their importance.





**Software Test Case Specification**

**Assumptions/Constraints/Risks**

**Assumptions**

The software manly works in the windows environment so windows os has to be used , the functionality of the software can be modified by changing the attribute of the Salford modeler.

**Constraints**

If the data set is corrupted or incorrect then wrong equation will be generated by the algorithms and the prediction will be wrong . The product should have the Salford predictive modeler because if it is not present then the graphs could not be generated and comparison cannot be done .

The assumption done in this project is that the data set collected of the gases in correct and not corrupted , the corruption of this will result in the wrong prediction of the gases

**Test Case Summary**

Test Case Summary Test Case/Script Identifier Test Case/Script Title Execution Priority System Overview system testing

HIGH

User Overview User testing Medium

Data Overview Data testing HIGH

**Test Case Details**

**System Overview**

The system has to be checked to use the software , whether the correct build of OS is being used or not , whether the software installed is correct or not .

**Test Objective**

The objective of the test is to verify whether the system installed is correct or not .

**Inter-Case Dependencies**

The software installed is bought by the user and windows OS should be used

**Test Items**

The requirement and design document is benchmark for it , according to the document the testing is done for each part

**Input Specifications**

Os : windows Software : Salford Modeler Algorithm : Data Mining

**Expected Test Results**

According to the test result the specified objects has to be same . Pass/Fail Criteria If the OS is different or the software installed is different it is considered to be fail criteria .

**Assumptions and Constraints**

It is assumed that the hardware and the system is build as given in the requirement and the design document .

**Data Overview**

**Test Objective**

To confirm whether the chosen data is correct or wrong and the data set is valid or invalid.

**Inter-Case Dependencies**

The data has to be collected from a valid source for it to be correct , the data should be in numbers and not string for it to be tested .

**Test Items**

The requirement and design document is benchmark for it , according to the document the testing is done for each part

**Prerequisite Conditions**

The prerequisite condition is that the data should be in numbers and not string and the data should be related to air quality index.

**Input Specifications**

The data should be taken from a reliable source for its originality and environmental websites should be double checked before using .

**Expected Test Results**

The result should be that the dataset should be correct and valid .

**Pass/Fail Criteria**

If the data is in form of strings it should be discarded and if the data is numbers then it should be valid

**Test Procedure**

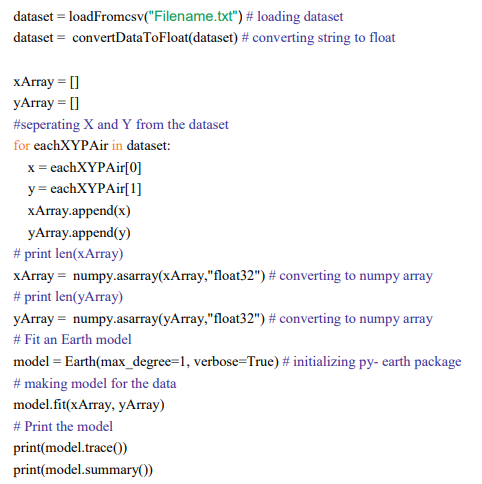
The data would be placed in the software and the generated result would be analysed .

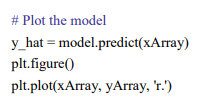
**Functionality/Module Implementation - Interfaces, Software Code**

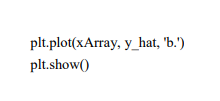
MARS :

In statistics, multivariate adaptive regression splines (MARS) is a form of regression analysis introduced by Jerome H. Friedman in 1991. It is a nonparametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables.

MARS ALGORITHM CODE :



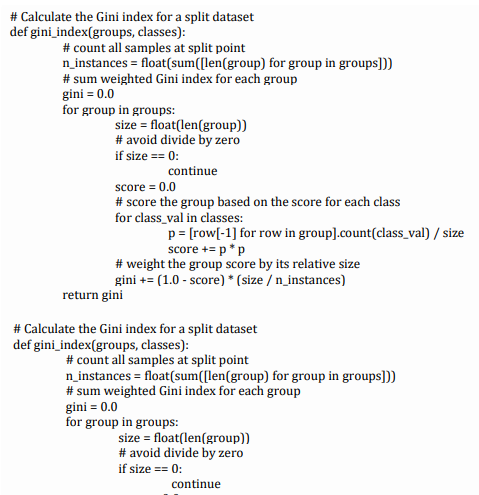


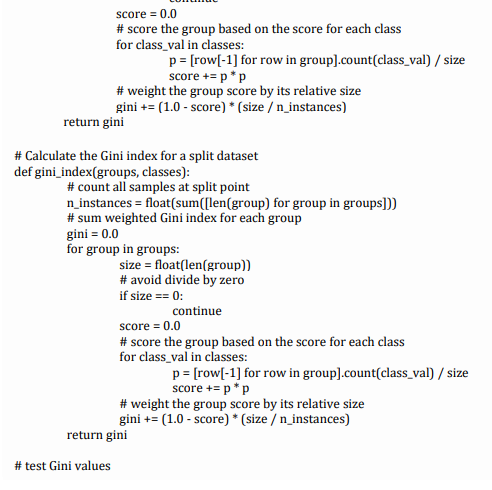


CART :

CART (Classification and Regression Trees) is a modern, c.1984, flavor of data mining that employs decision trees and can be used for a variety of business and scientific applications. Its advantages include quick insight into database patterns and significant relationships using simple tools such as graphs, charts and reports. Both

CART ALGORITHM CODE :



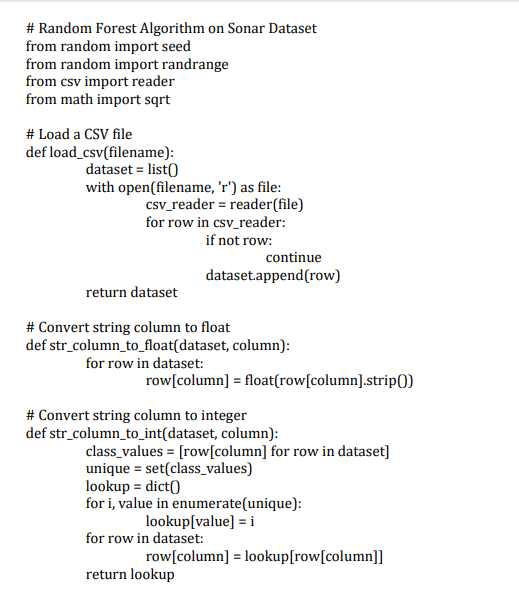


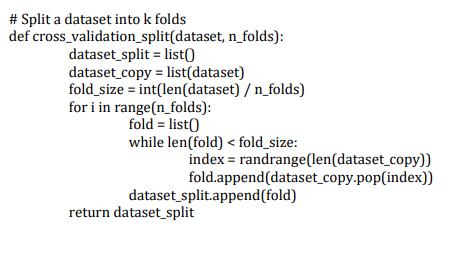


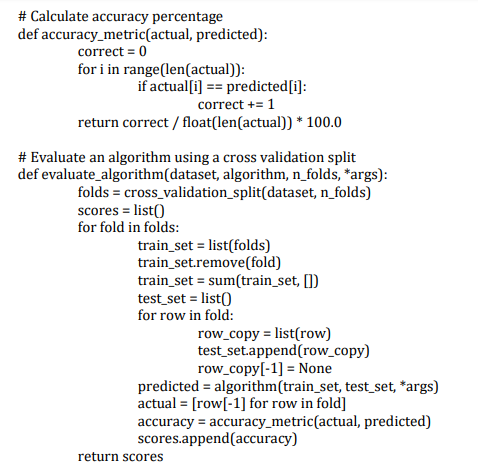
RANDOM FOREST :

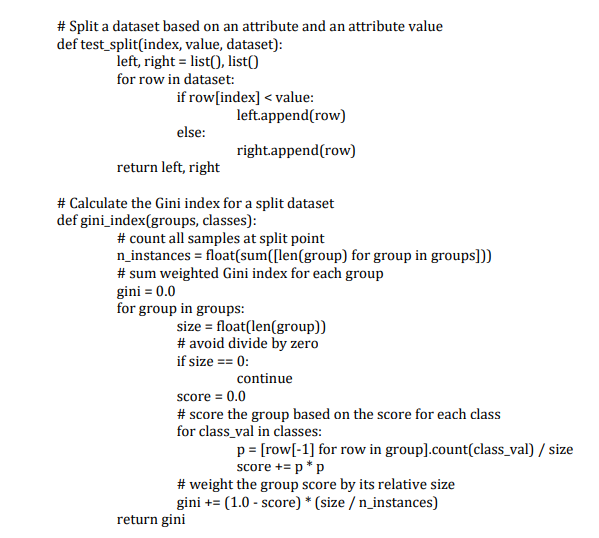
Random forest (or random forests) is a trademark term for an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random forests are collections of trees, all slightly different. It randomize the algorithm, not the training data.

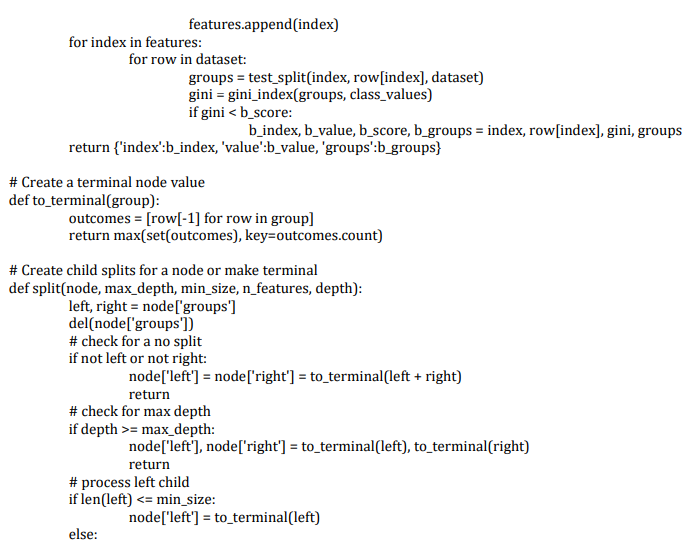
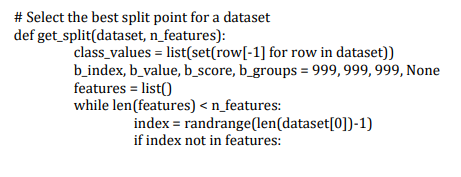
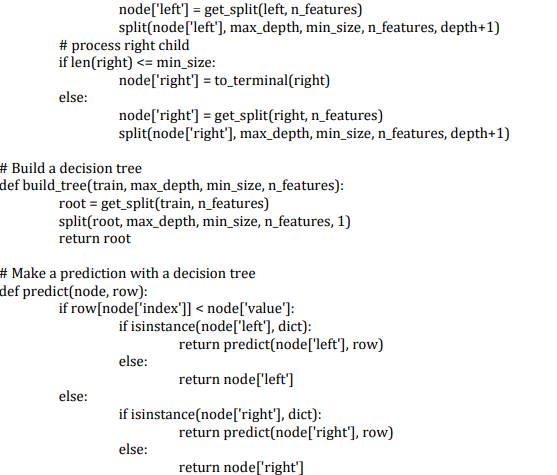
RANDOM FOREST ALGORITHM CODE :

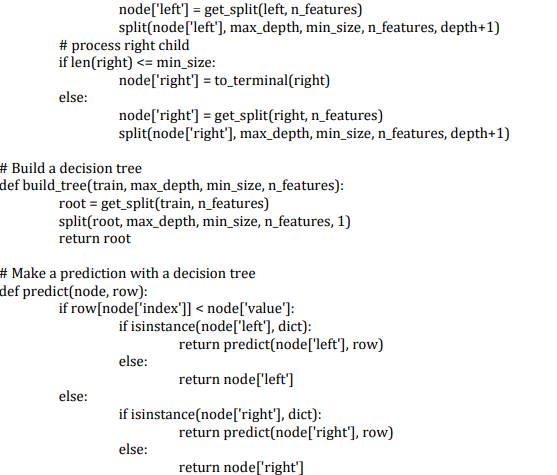


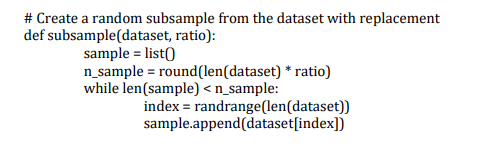


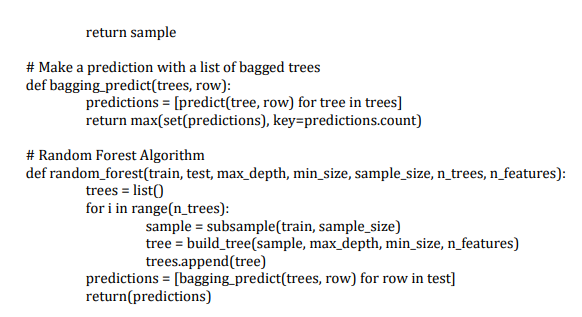


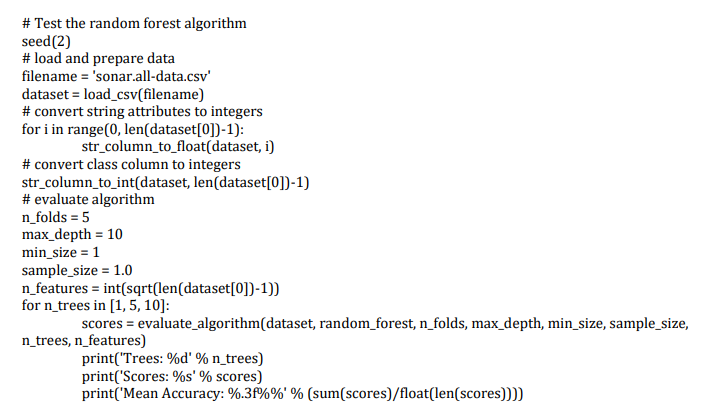


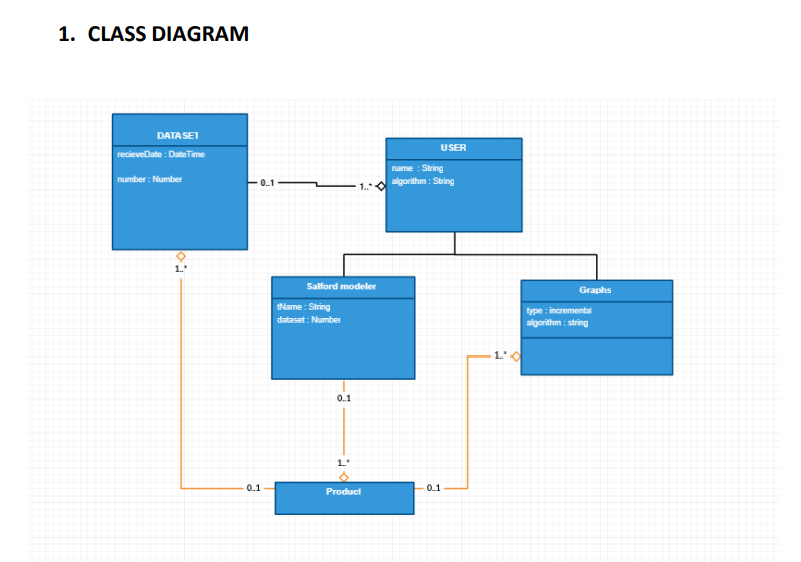
 



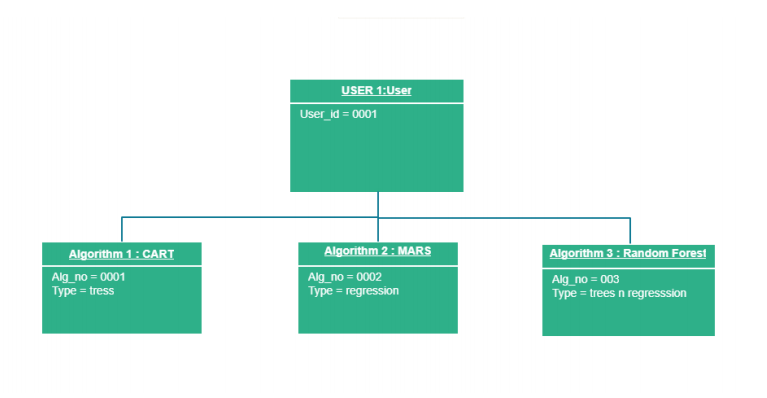


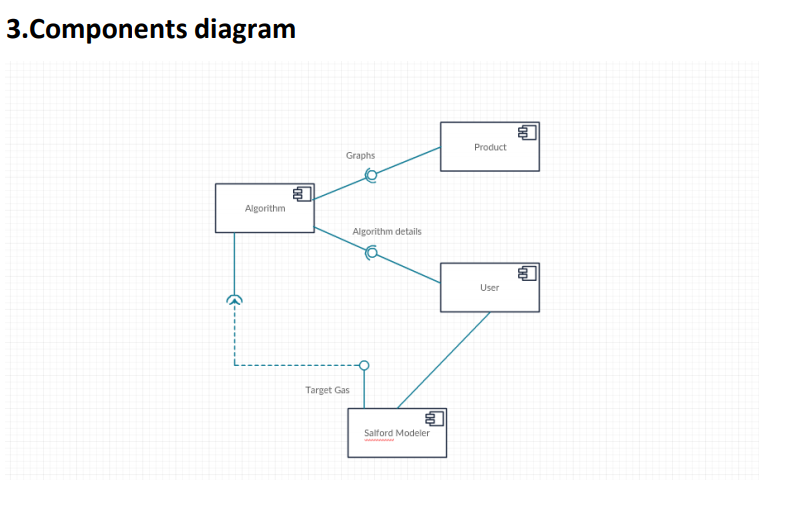




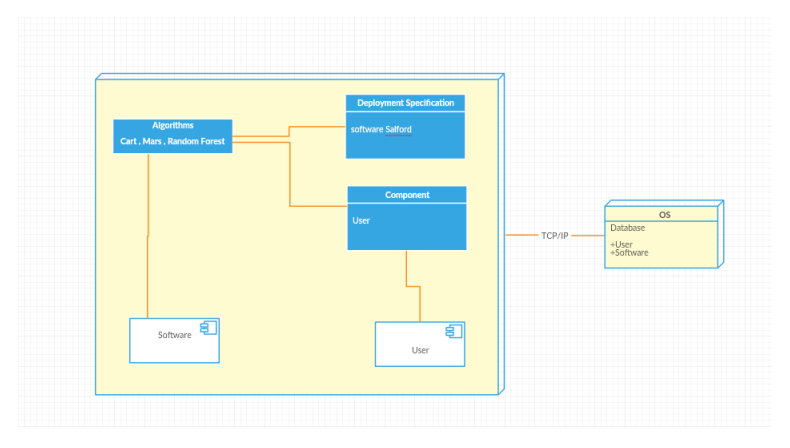


**2. OBJECT DIAGRAM**

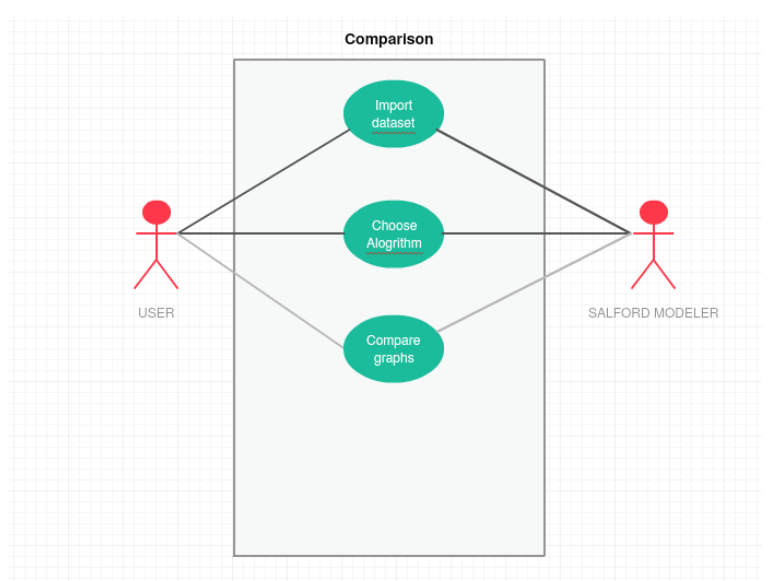




**4. DEPLOYEMENT DIAGRAM**



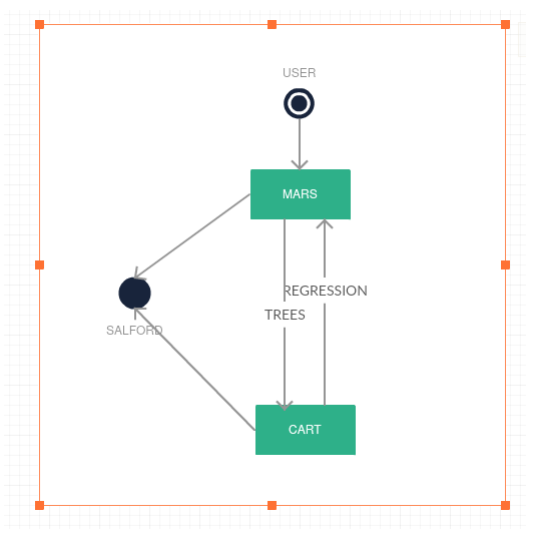
**5. Use Case Diagram**

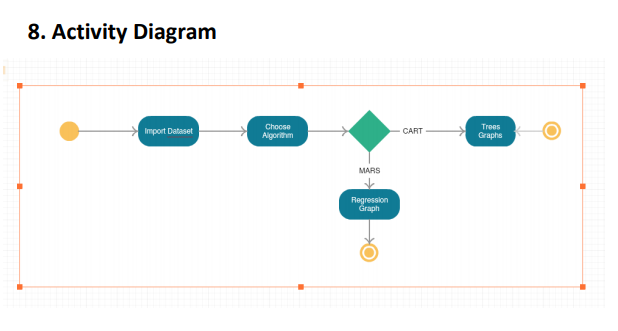


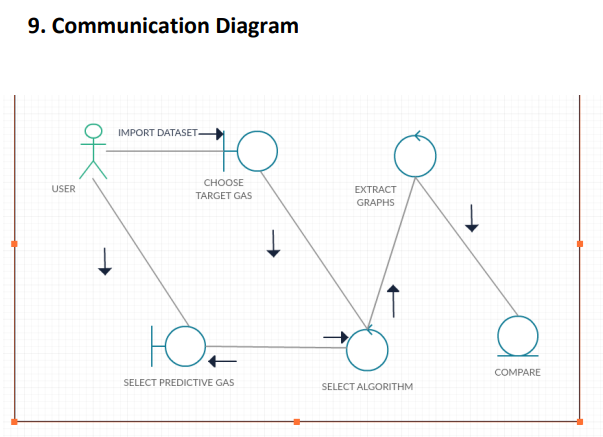
**6. Interaction Diagram**



**7. State Chart**







**CONCLUSION**

In this project we have proposed to show the prediction of ozone concentration by using three regression model. By keeping the train and test in the ratio 7:3 we compare the result from all three cases. Evaluation of the prediction models indicates that the Multivariate Adaptive Regression Splines model describes the dataset better and has achieved significantly better prediction accuracy as compared to the Random Forest and Classification And Regression Tree. Multivariate Adaptive Regression Splines gives the result by considering less variables as compared to other two. It evaluates on basis of 8 variables while other two require all variables. Moreover, Random Forest takes a little more time for building the tree as the elapsed time was calculated to 45 second in this case. PT08\_S2\_NMHC\_ is the most important variable as given by Multivariate Adaptive Regression Splines while PT08\_S1\_CO\_ is most important variable as given by Random Forest and Classification And Regression Tree. Observing all the graphs Multivariate Adaptive Regression Splines gives the closest curve of both train and test set when compared. It can be concluded that multivariate adaptive regression splines can be a valuable tool in predicting ozone for future.

**REFERENCES :**

1. Pierre Ozer, Mohamed Béchir Ould Mohamed Laghdaf, Sidi Ould Mohamed Lemine, Jean Gassan “Estimation of air quality degradation due to Saharan dust at Nouakchott, Mauritania, from horizontal visibility data”,Water, Air, and Soil Pollution, 2007, Volume 178, Number 1-4, Page 79, 2007.
2. 2. W. Y. Zhang, T. T. Han, Z. B. Zhao, J. Zhang and Y. F. Wang, "The Prediction of Surface Layer Ozone Concentration Using an Improved AR Model," 2011 International Conference of Information Technology, Computer Engineering and Management Sciences,Nanjing, Jiangsu, 2011, pp. 72-75.
3. Assessment for Decision-Makers: Scientific Assessment of Ozone Depletion: 2014, World Meteorological Organization, Global Ozone Research and Monitoring Project— Report No. 56, Geneva, Switzerland, 2014.
4. C. Zhang and D. Yuan, "Fast Fine-Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark," 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), Beijing, 2015, pp. 929-934.
5. Y. Wang and W. Yue, "Target data association in communication constrained environment using CART: Compressed adaptive reference topology," 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), Beijing, China, 2016, pp. 333-338.
6. M. Birgersson, G. Hansson and U. Franke, "Data Integration Using Machine Learning," 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW), Vienna, Austria, 2016, pp. 1-10.
7. S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, “On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario,” Sensors and Actuators B: Chemical Volume129, Issue 2, 22 February 2008, Pages 750–757.
8. Saverio De Vito, Marco Piga, Luca Martinotto, Girolamo Di Francia, CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical, Volume 143, Issue 1, 4 December 2009, Pages 182-191, ISSN 0925-4005
9. T. S. Hui, S. A. Rahman and J. Labadin, "Comparison between multiple regression and multivariate adaptive regression splines for predicting CO2 emissions in ASEAN countries," Information Technology in Asia (CITA), 2013 8th International Conference on, Kota Samarahan, 2013, pp. 1-5
10. Friedman, J. H. (1991). Multivariate adaptive regression splines. The annals of statistics, 1-67.