Amrita School of Engineering, Amrita Viswa Vidyapeetham, Amritapuri
21MAT117, Mathematics for Intelligent Systems

Air Quality prediction using Linear Regression

By Group No – 13, Members -

S HARI SANKAR	AM.EN. U4AIE21056
K SUBHASH	AM.EN. U4AIE21036
SAI MANASA SOUMYA	AM.EN. U4AIE21057
CHERISHMA AKSHAYA	AM.EN. U4AIE21051
R RAKESH	AM.EN. U4AIE21052

I. INTRODUCTION

Air Quality security has gotten one of the foremost fundamental exercises for the administration in numerous mechanical and concrete zones in today's world. The meteorological and traffic factors, consuming crude oil derivatives and mechanical parameters perform critical jobs in air contamination which make an adverse effect on living beings. With this expanding pollution on the earth, we also had different executing models that can record data about centralizations of air pollutants (SO2, NO2, etc.). The affidavit of those unsafe gases is noticeable all around; is influencing the character of individuals' lives, particularly in urban territories. Of late, numerous specialists began to study about this concern and mentioned several measures to manage these conditions with the assistance of the presidency and native people.

Data Analytics is a leading approach as it includes natural detecting systems and sensor information accessible. During this paper, Machine Learning strategies are utilized to predict the ratio with relation to other components present in the earth’s atmosphere. Mainly 14 factors which are present in an Air Quality Dataset and a few other atmospheric components encompass an adverse effect on the ratio. Human skin and mucous layers of the eyes supports different ratio in the numerous atmospheric conditions. Thus, it's very necessary to grasp how various factors will make an effect on the relative humidity.

Problem:

Humans are very sensitive to humidity, as the skin relies on the air to get rid of moisture. The process of sweating is your body's attempt to keep cool and maintain its current temperature. If the air is at 100-percent relative humidity, sweat will not evaporate into the air. As a result, we feel much hotter than the actual temperature when the relative humidity is high. If the relative humidity is low, we can feel much cooler than the actual temperature because our sweat evaporates easily, cooling us off. For example, if the air temperature e is 75 degrees Fahrenheit (24 degrees Celsius) and the relative humidity is zero percent, the air temperature feels like 69 degrees Fahrenheit (21 C) to our bodies. If the air temperature is 75 degrees Fahrenheit (24 C) and the relative humidity is 100 percent, we feel like it's 80 degrees (27 C) out.

Objective

To predict the relative humidity at a given point of time using all the attributes affecting relative humidity.

II. DATA SET

II.1. Data description

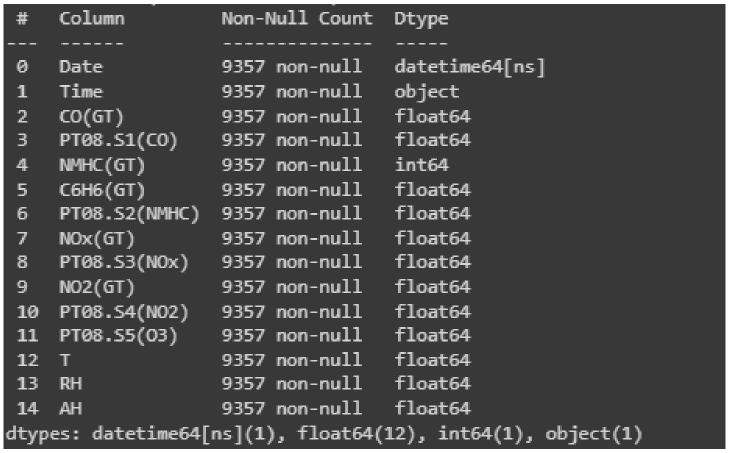
The air quality dataset for this project is collected from the UCI repository. The data set is in the UCLX format. The dataset contains data of average hourly responses of different elements in the air for nearly one year from March 2018 to April 2019. Dataset consists of 9357 rows and 15 columns. The table shows the details of various 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 average sensor response
4	NMHC(GT) True hourly arranged overall Non-Metanic Hydro Carbons concentration in microgram/m^3
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

Total no of rows of data in dataset – 9357.

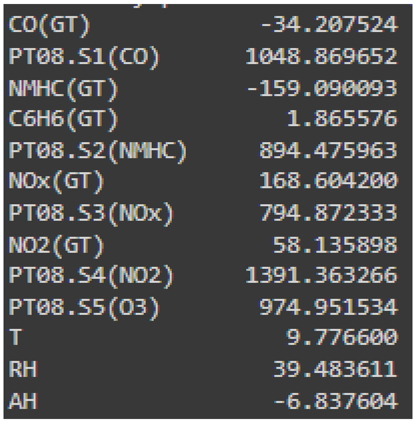
The count and data type of every attribute is given below –



```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Date           9357 non-null   datetime64[ns]
 1   Time           9357 non-null   object
 2   CO(GT)         9357 non-null   float64
 3   PT08.S1(CO)    9357 non-null   float64
 4   NMHC(GT)       9357 non-null   int64
 5   C6H6(GT)       9357 non-null   float64
 6   PT08.S2(NMHC)  9357 non-null   float64
 7   NOx(GT)        9357 non-null   float64
 8   PT08.S3(NOx)   9357 non-null   float64
 9   NO2(GT)        9357 non-null   float64
 10  PT08.S4(NO2)   9357 non-null   float64
 11  PT08.S5(O3)    9357 non-null   float64
 12  T              9357 non-null   float64
 13  RH             9357 non-null   float64
 14  AH             9357 non-null   float64
dtypes: datetime64[ns](1), float64(12), int64(1), object(1)
```

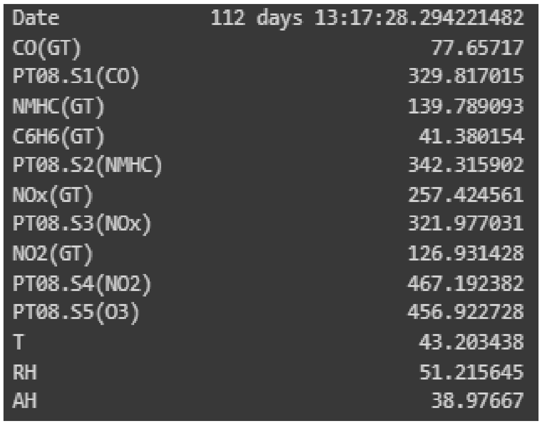
II.2. Data Set analysis

The mean fig (a) and standard deviation fig(b) of all the attributes presents in the data set in given below –



```
CO(GT)            -34.207524
PT08.S1(CO)      1048.869652
NMHC(GT)         -159.090093
C6H6(GT)            1.865576
PT08.S2(NMHC)     894.475963
NOx(GT)           168.604200
PT08.S3(NOx)      794.872333
NO2(GT)            58.135898
PT08.S4(NO2)     1391.363266
PT08.S5(O3)       974.951534
T                   9.776600
RH                 39.483611
AH                 -6.837604
```

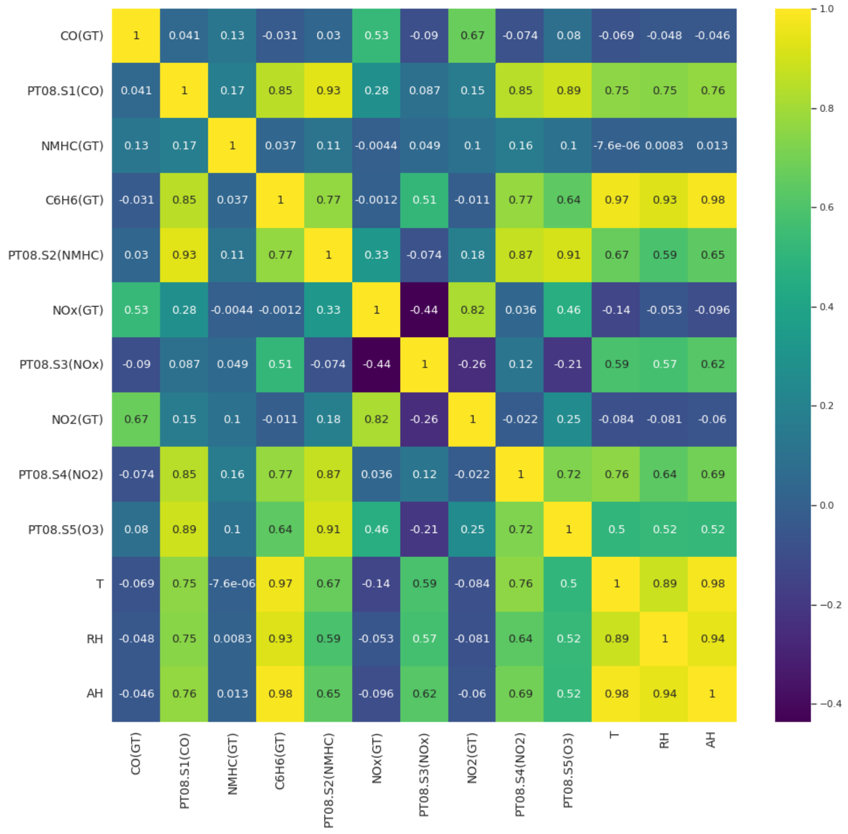
(a)



```
Date           112 days 13:17:28.294221482
CO(GT)                            77.65717
PT08.S1(CO)                     329.817015
NMHC(GT)                        139.789093
C6H6(GT)                         41.380154
PT08.S2(NMHC)                   342.315902
NOx(GT)                         257.424561
PT08.S3(NOx)                    321.977031
NO2(GT)                         126.931428
PT08.S4(NO2)                    467.192382
PT08.S5(O3)                     456.922728
T                                43.203438
RH                               51.215645
AH                                38.97667
```

(b)

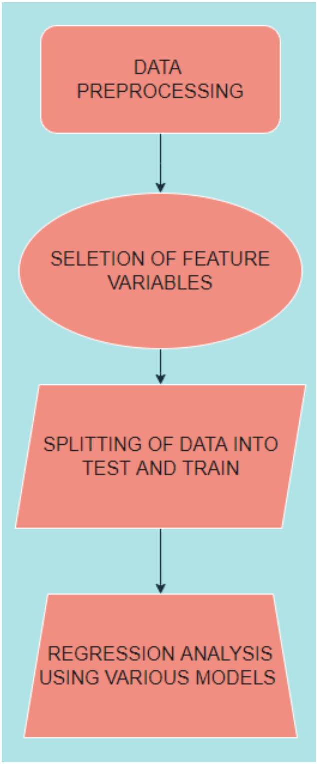
A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.



	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
CO(GT)	1	0.041	0.13	-0.031	0.03	0.53	-0.09	0.67	-0.074	0.08	-0.069	-0.048	-0.046
PT08.S1(CO)	0.041	1	0.17	0.85	0.93	0.28	0.087	0.15	0.85	0.89	0.75	0.75	0.76
NMHC(GT)	0.13	0.17	1	0.037	0.11	-0.0044	0.049	0.1	0.16	0.1	-7.6e-06	0.0083	0.013
C6H6(GT)	-0.031	0.85	0.037	1	0.77	-0.0012	0.51	-0.011	0.77	0.64	0.97	0.93	0.98
PT08.S2(NMHC)	0.03	0.93	0.11	0.77	1	0.33	-0.074	0.18	0.87	0.91	0.67	0.59	0.65
NOx(GT)	0.53	0.28	-0.0044	-0.0012	0.33	1	-0.44	0.82	0.036	0.46	-0.14	-0.053	-0.096
PT08.S3(NOx)	-0.09	0.087	0.049	0.51	-0.074	-0.44	1	-0.26	0.12	-0.21	0.59	0.57	0.62
NO2(GT)	0.67	0.15	0.1	-0.011	0.18	0.82	-0.26	1	-0.022	0.25	-0.084	-0.081	-0.06
PT08.S4(NO2)	-0.074	0.85	0.16	0.77	0.87	0.036	0.12	-0.022	1	0.72	0.76	0.64	0.69
PT08.S5(O3)	0.08	0.89	0.1	0.64	0.91	0.46	-0.21	0.25	0.72	1	0.5	0.52	0.52
T	-0.069	0.75	-7.6e-06	0.97	0.67	-0.14	0.59	-0.084	0.76	0.5	1	0.89	0.98
RH	-0.048	0.75	0.0083	0.93	0.59	-0.053	0.57	-0.081	0.64	0.52	0.89	1	0.94
AH	-0.046	0.76	0.013	0.98	0.65	-0.096	0.62	-0.06	0.69	0.52	0.98	0.94	1

III. METHODOLOGY

The below flowchart represents the steps involved in the prediction process-



DATA PREPROCESSING
SELETION OF FEATURE VARIABLES
SPLITTING OF DATA INTO TEST AND TRAIN
REGRESSION ANALYSIS USING VARIOUS MODELS

III.1. Data Preprocessing and Feature Selection

It is a technique used in data mining that involves transforming raw data into an understandable format. The data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. As it contains some missing value, the dataset is cleaned, and decimal values are converted into proper float values.

As a part of feature selection, we have dropped the attribute time and date from our data set using python. This attribute will not be used in any of the prediction models which we are developing

III.2. Splitting of data into test and train

After the data is processed, it is split into two categories training and testing. This is done to train the model which is being used and to measure parameters like RMSE, Accuracy etc. 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.

In the prediction models developed, the data set has been splitted in such a way that the test size consists of 33% of the whole dataset. The rest of the data set is being used up for training the models that are being developed. The random state for splitting the data set is 42.

III.3. Models used for the prediction

The features used for prediction is CO(GT), C6H6(GT), PT08.S2(NMHC), NOx (GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3), T, AH

- In the initial 10 models, we have taken each attribute individually for predicting the dependent variable i.e., relative humidity. The accuracy MSE and RMSE for each individual independent variable are recorded. The accuracy among different attributes is compared and the best model is selected

- In the next 9 models, we have taken set of 2 attributes at a time from the independent variables and fitted the model. The best model out of all the model is selected.

- In the next 8 models, we have taken set of 3 attributes at a time from the independent variables and fitted the model. The best model out of each is taken by comparing the accuracy

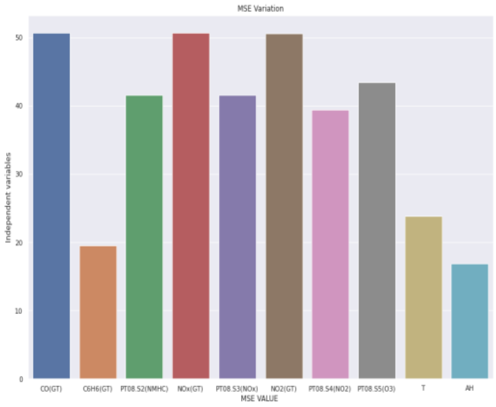
- A model in which we are using every independent variables of the data sheet is also being used. The accuracy RMSE and MSE of the model is also recorded.

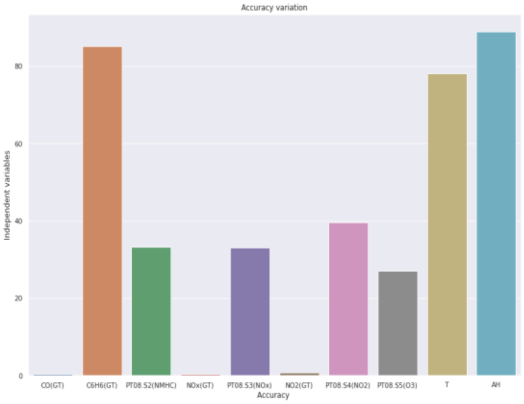
- A model based on polynomial regression of degree 2 is also used for prediction of relative humidity. The RMSE, MSE and accuracy of the model is recorded.

- The total number of models used for the project = 28

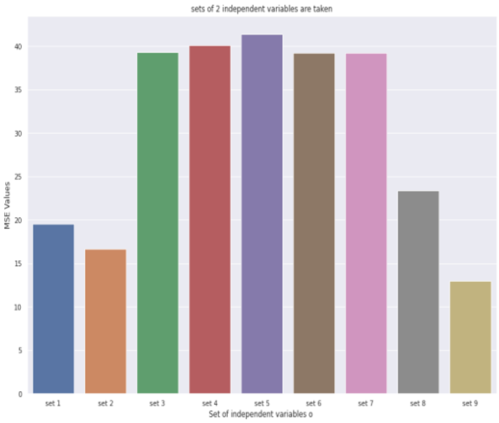
IV. RESULTS

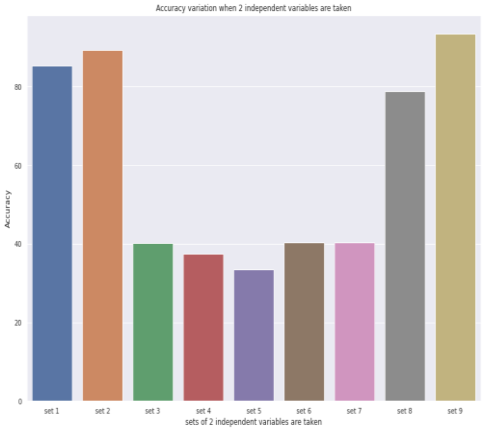
- A bar graph showing the Accuracy and MSE for various independent variables among the first 10 models are shown below



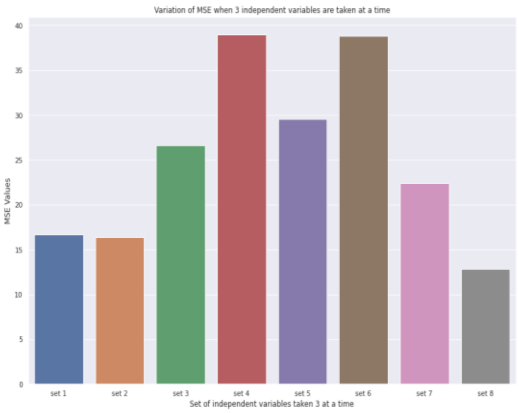


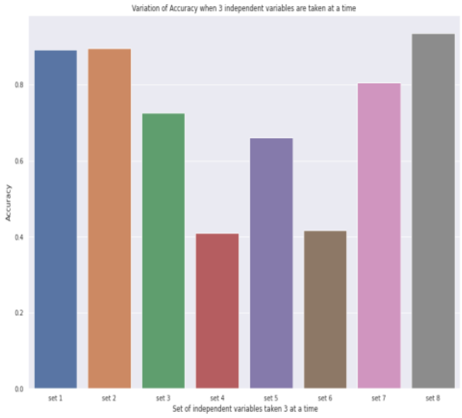
- A bar graph showing the variation of accuracy and MSE while taking set of 2 independent variables are shown below –



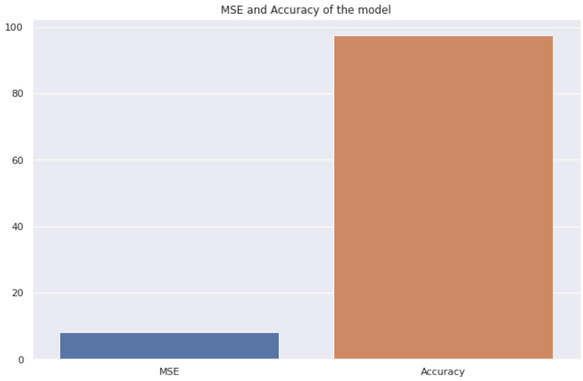


- A bar graph showing the variation of accuracy and RMSE while taking set of 2 independent variables are shown below –

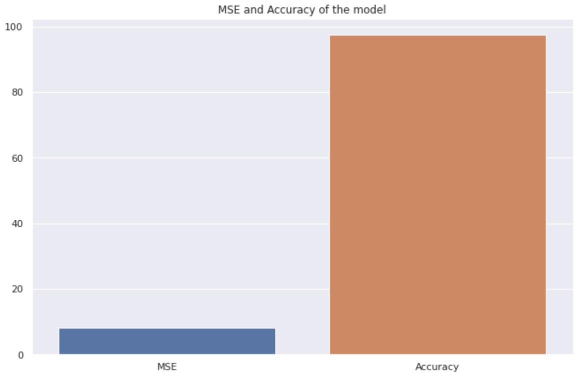




- A bar graph showing the accuracy and MSE of the model in which every independent variable is used is shown below.



- A bar graph showing the accuracy and MSE of polynomial regression model with degree 2 is also shown.



V. CONCLUSIONS AND DISCUSSIONS

- From the initial 10 models where we have used each independent variable individually, we can conclude that the best model when each individual feature is taken into consideration is that when the independent feature is Absolute humidity.

  The accuracy of the above model comes out to be `88.98238321814209%` and the RMSE value comes out to be `16.855962514000264`

- From the models where set of two independent variables are used, the most accurate model comes out to be the one where the independent variables are Temperature and Absolute humidity

  The accuracy of the above model comes out to be `93.48680051703874%` and the RMSE value comes out to be `12.960057111832572`

- From the models where set of three independent variables are used, the most accurate model comes out to be the one where the independent variables are PT08.S5, Temperature and Absolute humidity.

  The accuracy of the above model comes out to be `93.62408972903646%` and the RMSE value comes out to be `12.822739627512387`

- Out of all the models discussed above the one with the high accuracy is the one where we have taken every independent variable together as it presents with an accuracy of `97.47618144392487%`

VI. REFFERENCES

- https://github.com/MananJethwani/air_quality_prediction-linear-regression-

- http://www.ijstr.org/final-print/mar2020/Air-Quality-Prediction-Through-Regression-Model.pdf

- GitHub - ishanag9/air-quality-prediction: Air Quality Prediction of Relative Humidity - Regression

- Air Quality prediction of Relative Humidity | Kaggle

DATASET LINK

- https://archive.ics.uci.edu/ml/datasets/air+quality

PYTHON CODE

- https://colab.research.google.com/drive/1FJrIIfJV5c0puhjr3nMg0XJxYfOqoY12?usp=sharing