

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Technology Degree in

Information Technology

By

MOHAMMAD HABIB MEMON
(Reg. No.37120046)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

DECEMBER 2020



SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(Established under Section 3 of UGC Act, 1956)
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai -
600119 www.sathyabama.ac.in



SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **MOHAMMAD HABIB MEMON (Reg.no.37120046)** who carried out the project entitled “AIR QUALITY INDEX PREDICTION” under our supervision from July 2020 to Dec 2020.

Internal Guide

Dr. R. SUBHASHINI M.E., Ph.D.,

Head of the Department

Dr. R. SUBHASHINI M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **MOHAMMAD HABIB MEMON(Reg.no.37120046)** hereby declare that the Project Report entitled “**AIR QUALITY INDEX PREDICTION**” done by me under the guidance of **Dr. R. SUBHASHINI** at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree in INFORMATION TECHNOLOGY

DATE: 20-01-2021

PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. R. SUBHASHINI M.E., Ph.D.**, Head of the Department, Dept. of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. SUBHASHINI** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **INFORMATION TECHNOLOGY** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

AKON ONLINE SERVICES PRIVATE LIMITED

Quality Never Goes out of style....
GST No 27AATCA650481ZS
UDYAM-MH-33-0007682
ceo@akonzon.com

Bungalow no 6, Shrushti
Complex Hiranandani
Estate GB Road Patlipada
Thane West (400607)
Maharashtra India

Date : 26-09-2020

TO WHOM IT MAY CONCERN

This is to certify that **MR. MOHAMMED HABIB MEMON** has successfully completed his internship during his tenure from 25-07-2020 to 25-09-2020 in our organisation as **DATA RESEARCH INTERN**.

During his tenure we found him to be punctual, hardworking, inquisitive and dedicated to the project allotted to him. He has duly completed the project and assignments assign to him to our satisfaction.

We wish him all the best for his future endeavours.



AKON ONLINE SERVICES PRIVATE LIMITED

CEO / Managing Director
Founder and CEO



ABSTRACT

Air pollution is one of the most severe problems of the current time. It is growing day by day because of the vast level of industrialization and urbanization, causing massive damage to flora and fauna of the planet. Every moment, we are breathing air that is full of pollutants, going to our lungs, impregnating our blood and then the whole body, causing uncountable health problems. Both state and central governments have put in many efforts to keep air pollution under control. Air quality index(AQI) is a measure of air quality. It is used to inform citizens about the associated health impacts of air pollution exposure. This project is one scientific contribution towards this challenge. We compare three simple machine learning algorithms, linear regression, ridge-lasso and random-forest. The results are promising and it was proven that implementation of these algorithms could be very efficient in predicting air quality index.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	I
	LIST OF FIGURES	ii
	LIST OF TABLES	iii
	LIST OF ABBREVIATIONS	iv
1	INTRODUCTION	
	1.1 What is AQI?	2
2.	SYSTEM REQUIREMENTS	
	2.1 Software Requirements	3
	2.2 Hardware Requirements	3
3.	IMPLEMENTATIONS	
4.	EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED	
	4.1 Linear Regression	7
	4.2 Random Forest Regression	8
5.	RESULT AND SREENSHOT	
6.	SUMMARY AND CONCLUSIONS	
	REFERENCES	14
	APPENDIX	15

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
5.1	Dataset	9
5.2	Source Code	10
5.3	x_train and x_test	10
5.4	y_train and y_test	11
5.5	pred and rpred	11
5.6	Result	12

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO.
-----------	------------	----------

1.1	AQI Level Classification	2
2.1	Hardware Requirements	3

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
AQI	Air Quality Index
HTML	Hyper Text Markup Language
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

CHAPTER 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, PM_{2.5}, 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 relationship between these factors and predict carbon monoxide C.O. based on other parameters. 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 WHAT IS AQI?

The air quality index is an index for reporting air quality on a daily basis. In other words, it is a measure of how air pollution affects one's health within a short time period. The AQI is calculated based on the average concentration of a particular pollutant measured over a standard time interval. Generally, the time interval is 24 hours for most pollutants, 8 hours for carbon monoxide and ozone.

The purpose of the AQI is to help people know how the local air quality impacts their health. Awareness of daily levels of air pollution is important, especially for those suffering from illnesses caused by exposure to air pollution.

Table 1.1 AQI level classification

POLLUTION LEVEL	INDEX VALUE	INDEX COLOR
Good	0-50	Green
Moderate	51-100	Yellow
Unhealthy for sensitive group	101-150	Orange
Unhealthy	151-200	Red
Very unhealthy	201-300	Purple
Hazardout	301-500	Maroon

CHAPTER 2

SYSTEM REQUIREMENTS

2.1 SOFTWARE REQUIREMENTS:

- Operating System Windows 7 onwards
- Database Any Database

2.2 HARDWARE REQUIREMENTS:

A brief description of the hardware requirement to implement the “AIR QUALITY INDEX PREDICTION” is given below:

Table 2.1 Hardware Requirements

SERIAL NUMBR	FEATURE	MINIMUM	RECOMMENDED
01	Processor	P II	P III
02	RAM	2 GB RAM	4 GB RAM
03	Hard Disk Size	4.3 GB	40 GB for Server
04	Monitor(Color/SVGA)	Any Standard Monitor	Any Standard Monitor
05	Operating System	Win-7,8,10	Win-NT / 2K

		Networking with office 2010	Server
06	Network/ Multi-user	Win-7,8,10 Networking	Win-NT / 2K Server
07	System Printer	Any Good Printer	Any Good Printer
08	Network Interface Card	NE-2000 Ethernet	Card or Similar
09	Scanner	TWAIN compatible OCR software	Flat-bed Scanner

CHAPTER 3

IMPLEMENTATION

Steps for implementations:

STEP 1 WEB SCRAPING:

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

STEP 2 DATA PREPROCESSING:

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.

STEP 3 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. Split can be made 70/30 or 80/20 or 85/15 or 75/25

STEP 4 FEATURE SELECTION:

The data features that used to train machine learning models have a huge influence on the performance of the model. Irrelevant or partially relevant features can negatively impact model performance. In this project, the attributes such as date, time, C6H6 (benzene), PTO8.S4 tungsten oxide in the dataset are the selected features for better results.

STEP 5:

Fitting a Predictive Model (here random forest)

Since Random forest is ensemble model (made of many trees) from sklearn.ensemble, import RandomForestClassifier class

With 501 tree or “n_estimators” and criterion as ‘entropy’

Fit the model via .fit() method with attributes X_train and y_train

STEP 6 AIR QUALITY PREDICTION:

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.

$R^2 = 1 - (\text{Explained Variation} / \text{Total Variation})$

CHAPTER 4

ALGORITHMS

4.1 LINEAR REGRESSION:-

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog)

Simple Regression

Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to “learn” to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$Y = mx + c$$

Multivariable regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x, y, z) = w_1x + w_2y + w_3z$$

The variables x, y, z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$\text{Sales} = w_1\text{Radio} + w_2\text{TV} + w_3\text{News}$$

4.2 RANDOM FOREST

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART models with different samples and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

Algorithm 1. Random Forest

Input: B = Number of Trees, N = Training Data, F = Total-Features, f = Subset of Features

Output: Bagged class label for the input data.

1. For each tree in Forest B :

a) Select a bootstrap sample S of size N from training data.

b) Create the tree T_b by recursively repeating the following steps for each internal node of the tree.

i. Choose f at random from the F .

ii. Select the best among f .

iii. Split the node.

2. Once B Trees are created, Test instance will be passed to each tree and class label will be assigned based on majority of votes.

CHAPTER 5

RESULT AND SCREENSHOTS

df - DataFrame

Index	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
0	7.4	9.8	4.8	1017.6	93	0.5	4.3	9.4	219.721
1	7.8	12.7	4.4	1018.5	87	0.6	4.4	11.1	182.188
2	6.7	13.4	2.4	1019.4	82	0.6	4.8	11.1	154.038
3	8.6	15.5	3.3	1018.7	72	0.8	8.1	20.6	223.208
4	12.4	20.9	4.4	1017.3	61	1.3	8.7	22.2	200.646
5	16	25.2	10	1013.2	79	0.6	4.8	11.1	205.225
6	13.4	21	9.2	1015.1	87	0.5	1.5	7.6	236.825
7	14.3	22.6	6.6	1016.3	76	0.8	0.4	3.5	276.908
8	12.7	18.9	7.3	1021.2	76	1.8	6.1	16.5	108
9	11.9	18.9	6.3	1020.1	76	1.1	8.3	20.6	107.625
10	12.2	19.9	6.2	1021	69	2.1	4.4	18.3	125.892
11	13.4	21.1	5.6	1020	60	1.9	8	22.2	181.013
12	12.3	21.4	6.4	1019.6	70	1.9	6.7	14.8	152.554
13	13.7	22	7	1018.2	63	1.8	7.2	18.3	152.321
14	14.9	21.9	7.1	1018.3	73	0.6	2	7.6	319.738
15	15	24	7.6	1017.4	81	0.6	2.2	7.6	332.708
16	16.3	24	12	1017.2	80	0.8	1.3	9.4	279.6
17	18.7	26.3	9.5	1015.3	64	1.6	1.7	7.6	179.117
18	14.6	21.6	9.4	1018.4	72	2.7	9.4	18.3	54.7917
19	14.7	21.7	8.4	1017.1	71	1.8	10.9	22.2	93.375
20	14.8	22.6	8.5	1016.8	67	2.1	9.3	18.3	103
21	17.4	25.4	10.3	1015.8	63	1.9	6.7	14.8	132.208
22	18.1	26.7	10.2	1013.8	61	2.6	5	11.1	127.708
23	18.3	23.2	12.7	1011.2	75	1.9	10	22.2	109.333
24	16.4	22	11.2	1015.5	81	1.6	5.9	18.3	56.4583

Fig : 5.1 : Dataset

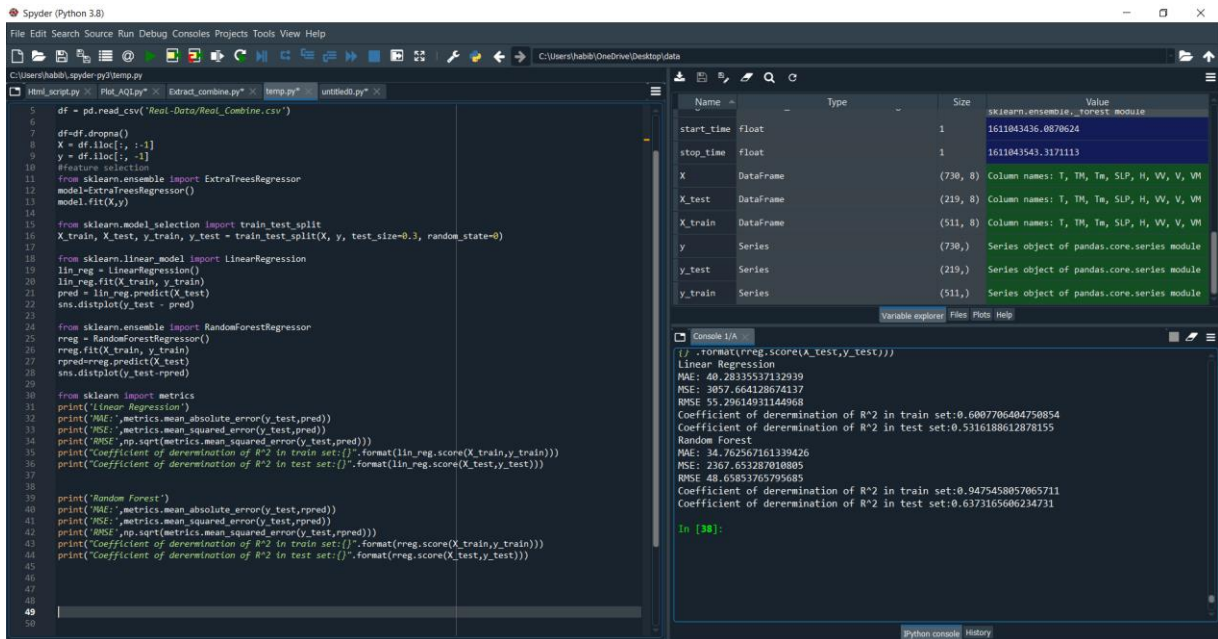


Fig.5.2 : Source Code

X_test - DataFrame									X_train - DataFrame								
Index	T	TM	Tm	SLP	H	VV	V	VM	Index	T	TM	Tm	SLP	H	VV	V	VM
197	14.3	20.3	9.9	1019.1	88	0.6	8.9	14.8	236	25.7	32.8	17.5	1010.3	41	2.6	12.8	25.9
188	11.9	20.7	4.4	1017.9	78	0.8	1.9	7.6	525	21.9	29	17.1	1016	75	3.1	1.9	5.4
14	14.9	21.9	7.1	1018.3	73	0.6	2	7.6	12	12.3	21.4	6.4	1019.6	70	3.9	6.7	14.8
31	19	26.8	11.7	1016.3	58	2.6	16.7	29.4	161	17.8	26.2	11	1014.3	68	3	3	11.1
391	19.3	28.6	10.6	1009.5	65	2.6	4.3	14.8	557	15.6	23.2	10.9	1015.6	82	0.6	4.3	9.4
320	30.4	36.2	25.2	1003.6	61	2.6	5.6	14.8	503	30.2	36.5	26.2	1002.3	70	3.9	5.7	14.8
300	30.5	36.2	25.4	997.3	65	2.3	10.5	29.4	474	30	36.6	27	998.0	76	2.6	16.9	25.9
703	23.9	32	14.5	1015.6	63	3	2	5.4	723	18	26.6	10.5	1017.1	61	3.9	8	14.8
463	31.4	39.8	27.6	999.7	54	3.9	13	25.9	502	29.1	34.8	27.2	1003.4	76	3.9	4.6	11.1
27	17.2	24.2	10.8	1018.3	74	2.6	4.6	14.8	159	18.9	27.1	14.5	1015.2	80	0.6	1.3	7.6
40	27.1	34.8	17.7	1009.7	52	3.6	6.1	11.1	364	9.5	17.5	5.6	1018.1	84	0.5	0.9	3.5
361	14.2	20	10.4	1019.2	82	1.6	1.9	9.4	410	24.9	31.7	15.8	1014.8	65	2.6	5.9	13
55	29.3	37.9	20.3	1002.9	35	2.3	4.4	11.1	230	24.2	30.8	17	1008.2	48	3.9	13.1	22.2
157	20.9	28.1	13.6	1016	59	0.8	6.3	18.3	78	32.4	38.5	26.6	1001	63	2.3	13.3	22.2
194	14.6	22.3	9	1021.2	86	3	4.4	9.4	92	31.9	39.2	27.8	995.2	62	3.9	8.9	14.8
170	18.1	28	10	1016.1	66	3.4	1.7	7.6	691	28.6	35	19.8	1000.2	49	3.7	11.5	22.2
619	36.5	43	27	1001.8	35	2.6	4.1	22.2	66	34.1	42	26.8	1000.8	32	3.9	7.8	14.8
680	30.4	35.6	25.7	1006.5	72	2.9	4.3	14.8	694	27.6	35	17.6	1007.6	53	3.1	5.9	13
383	13.8	21.5	7.2	1018	60	2.4	9.1	24.1	501	30.4	35.9	26	1003.6	78	3.6	2	7.6
79	34.1	40.3	28.8	999.7	59	3.9	10.9	16.5	436	28.3	38.7	23.8	1006.5	65	2.4	7.8	14.8
608	33.3	39.4	24.8	1000.7	28	3.1	8.3	18.3	303	33.3	38.4	28	1002.5	54	3.9	8.9	18.3
728	13.9	24.5	11.4	1015	95	0.6	8.7	14.8	669	32.6	36	27.6	999.6	70	2.7	10.6	20.6
675	30.7	34	25.5	1000.1	73	2.6	13.7	22.2	404	18.1	24.1	13.3	1021.1	60	2.7	12.2	20.6
162	17.6	25.6	11.2	1015.2	68	3.1	3.5	14.8	396	23.2	28.6	18.4	1008.5	65	3.4	6.5	13
387	15.7	25.6	7.4	1016.9	68	3.4	1.9	7.6	565	17.4	27.2	12.2	1016.4	79	3.1	7.8	18.3

Fig: 5.3: x_train and x_test

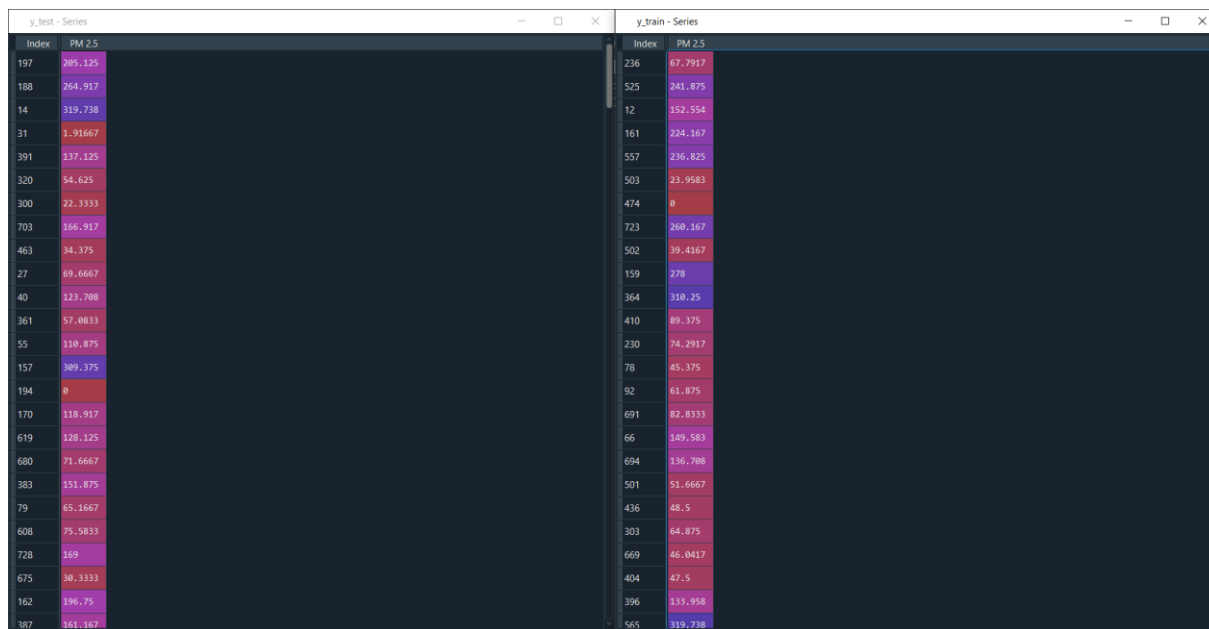


Fig : 5.4 : y_train and y_test

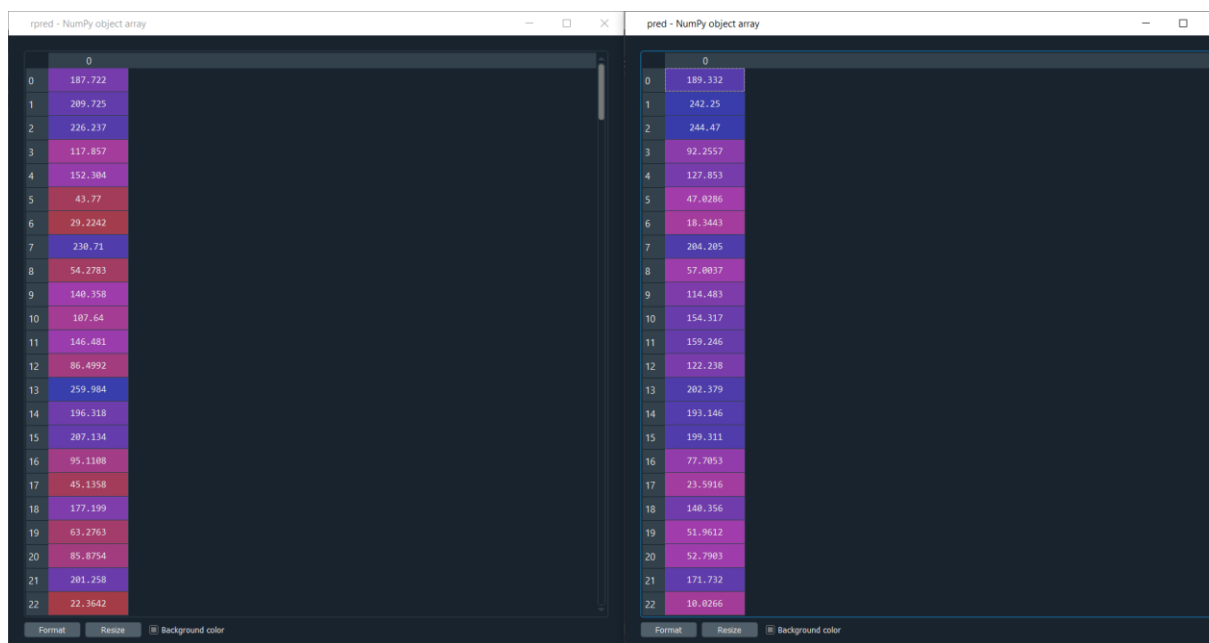


Fig:5.5: pred and rpred

```
Console 1/A X
Linear Regression
MAE: 40.28335537132939
MSE: 3057.664128674137
RMSE 55.29614931144968
Coefficient of derermination of R^2 in train set:0.6007706404750854
Coefficient of derermination of R^2 in test set:0.5316188612878155
Random Forest
MAE: 34.762567161339426
MSE: 2367.653287010805
RMSE 48.65853765795685
Coefficient of derermination of R^2 in train set:0.9475458057065711
Coefficient of derermination of R^2 in test set:0.6373165606234731

In [38]:

IPython console History
```

Fig: 5.6 : Result

CHAPTER 6

CONCLUSION

Regression analysis techniques are used to predict the results. The model will successfully predict the upcoming air quality index of any particular data within a given region. With this model we can forecast the AQI and alert the respected region of the country. The result for Linear Regression MAE: 40.28335537132939, MSE: 3057.664128674137 and RMSE 55.29614931144968 and the result for Random Forest Regression is MAE: 34.762567161339426, MSE: 2367.653287010805 and RMSE 48.65853765795685.

As future work, I intend to improve and investigate the usage of SVR to forecast air quality and last but not least, I intend to compare the results obtained by SVR to the ones achieved by other machine learning algorithms of a different nature, like artificial neural networks, Bayesian networks, decision trees, random forests, and genetic programming.

REFERENCES

1.The reference for the project has been taken from

- Sathyabama institute of science and technology
- www.geeksforgeeks.org
- Udemy(Machine learning A-Z course)
- youtube

APPENDIX

SOURCE CODE

```
#importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#importing datasets
df = pd.read_csv('Real-Data/Real_Combine.csv')

df=df.dropna()
X = df.iloc[:, :-1]
y = df.iloc[:, -1]
#feature selection
from sklearn.ensemble import ExtraTreesRegressor
model=ExtraTreesRegressor()
model.fit(X,y)

#splitting dataset into training and test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

```
#fitting Linear Regression to Training set
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
```

```
#predicting the test set results
pred = lin_reg.predict(X_test)
sns.distplot(y_test - pred)
```

```
#fitting Random Forest Regression to Training set
from sklearn.ensemble import RandomForestRegressor
rreg = RandomForestRegressor()
rreg.fit(X_train, y_train)
```

```
#predicting the test set results
rpred=rreg.predict(X_test)
sns.distplot(y_test-rpred)
```

```
from sklearn import metrics
```

```
print('Linear Regression')
```

```
print('MAE:',metrics.mean_absolute_error(y_test,pred))
print('MSE:',metrics.mean_squared_error(y_test,pred))
print('RMSE',np.sqrt(metrics.mean_squared_error(y_test,pred)))
print("Coefficient of derermination of R^2 in train
set:{}".format(lin_reg.score(X_train,y_train)))
print("Coefficient of derermination of R^2 in test
```

```
set:{}".format(lin_reg.score(X_test,y_test)))
```

```
print('Random Forest')
```

```
print('MAE:',metrics.mean_absolute_error(y_test,rpred))
```

```
print('MSE:',metrics.mean_squared_error(y_test,rpred))
```

```
print('RMSE',np.sqrt(metrics.mean_squared_error(y_test,rpred)))
```

```
print("Coefficient of derermination of R^2 in train
```

```
set:{}".format(rreg.score(X_train,y_train)))
```

```
print("Coefficient of derermination of R^2 in test
```

```
set:{}".format(rreg.score(X_test,y_test)))
```