Air Quality Index Prediction

Pavithra Avvari[1*], *Preethi* Nacham[1], *Snehitha* Sasanapuri[1], *Sirija Reddy* Mankena[1], *Phanisree* Kudipudi[1] and *Aishwarya* Madapati[1]

[1]Department of Information Technology, GRIET, India

Abstract Falling back past few years rapid progress in Air pollution has become a life-threatening concern in many nations throughout the world due to human activity, industrialisation, and urbanisation.. As a result of these activities, sulphur oxides, carbon dioxide (CO2), nitrogen oxides, carbon monoxide (CO), chlorofluorocarbons (CFC), lead, mercury, and other pollutants be emitted into atmosphere. Simultaneously, estimating quality of air is a tough undertaking because of evolution, variability, also unreasonable unpredictability over pollution and particle region and time. In this project we compare the two Algorithms of machine learning in predicting Index of Air Quality and its predominant. Support vector machine (SVM) exists as prominent machine learning method beneficial to forecasting pollutant plus particle levels and predicting the air quality index (AQI), and Random Forest Regression is another. We'll be working with data from India's Open Government Data Platform. This website displays Air Quality Index readings from around India, including Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), and Particulate Matter (PM) are examples of contaminants (PM10 and PM2.5), Carbon Monoxide (CO), and others. The output of the project is the predict of Air Quality index using two different algorithms and the comparison of models using various error metrics.

Keywords: Machine Learning, Air Quality Index, SVM, Random Forest Regression, Error Metrics

1 Introduction

Air pollutants alludes to the difficulty of pollutants within the air which can be dangerous upon humane along with the complete world. It may be defined as among the maximum risky threats humanity has ever faced. Affecting human's respiration and cardiovascular system, they are motive for extended mortality and extended chance for sicknesses for the population. To save you this trouble there's a want to are expecting air pleasant from pollution the use of device gaining knowledge of strategies. Subsequently, air quality index prediction and evaluation have became a universal studies sector. The purpose is to research device gaining knowledge of primarily based totally strategies for air pleasant prediction. Proper or accurate

[*] Corresponding Author: pavithra.griet@gmail.com

prediction or forecast of Air Quality or the degree of awareness in regard to numerous kinds ambient contaminants in the air, such as Ozone, Nitrogen Dioxide and Sulphur Dioxide, Carbon Monoxide, Particulate Matter less than 10 microns in diameter, Particulate Matter less than 2.5 microns in diameter, etc, may be very vital due to the fact the effect of those elements on human fitness will become severe.

2 Literature Review

Multiple Linear Regression and Principal Component Regression (PCR)Techniques have been accustomed to analyse previous AQI as well as meteorological data. They use prior records to predict day to day AQI of the current year. Then, using the Multiple Linear Regression Technique, this predicted value is compared to the observed value of AQI for the current year's seasons of summer, monsoon, post-monsoon, and winter. To discover collinearity among the absolute attributes, where the Principal Component Analysis is utilised .Principal components were used in Multiple Linear Regression to eliminate collinearity among predictor variables and reduce the number of predictors. In comparison to other seasons, the Principal Component Regression performs better in forecasting the AQI in the winter Only meteorological parameters were analysed or employed in this study for forecasting future AQI, but no ambient air pollutants that may cause harmful health consequences were considered.

Bayesian network model has been utilized to forecast air quality index in previous years. The model's evaluation factors are SO2, NO2, O3, CO, PM2.5, and PM10, and the model's output is the AQI value, after which the Bayesian network model is constructed. Finally, the model is utilised to forecast air quality and compare the predicted value to the actual value. The results show that air quality prediction accuracy is over 80%, and the forecast value is close to the real value in most cases, specifying the Bayesian network model posesses some practical use like air quality detection method. Greater the air pollution , higher would be the harm to human health, because of higher values in these six categories : PM2.5, PM10, SO2,NO2,CO, and O3 are all factors that influence the air quality index.

The precision of items for single and many steps ahead of concentrations of floor-degree ozone(O3), nitrogen dioxide(NO2), and other pollutants.and sulphur dioxide (SO2) is investigated using three machine learning (ML) techniques (SO2). Machine learning methods include support vector machines, M5P model bushes, and artificial neural networks (ANN).There are two types of simulations used: 1) uni-variate and 2) multivariate. The measures used to evaluate the performance include prediction pattern accuracy and root mean square error (RMSE).The results show that the M5P algorithm gives reliable estimates when using unique attributes in multivariate simulation. Air quality is a significant issue that directly affects human health. Monitoring motes wirelessly collect amazing air knowledge. These data are processed and used to forecast pollution awareness values utilising a sophisticated desktop-to-computer technology. The platform employs machine learning-based algorithms to create forecasting items based on the data collected.

3 Method

3.1 Implementation

3.1.1 Random Forest Model:

Random Forest (RF) is a combined learning technique for clustering and decision trees. The purpose of the Random Forest algorithm is to swap different decision trees for random selection of all dataset features and subsamples. All trees are averaged together to increase efficiency and prevent overfitting. Thus, this strategy minimizes high variance while slightly increasing the bias. This tradeoff usually makes the model more robust when predicting the input model.

Random Forest is a popular choice due to its fast training time, no input data for normalization, and hyperparameters that require little tuning. The uses randomly selected features to distribute individual trees, while random frequency selection is used to generate a subset of information for each decision tree. Deviations from the random number of attributes are checked for separation between each decision tree. If the target behavior is categorical, Random Forest will choose the most frequent based on its prediction. If it is a numerical problem, the mean of all estimates will be chosen. All measured data points were run over each decision tree in the forest to make predictions. Trees then vote on the results and predictions are made based on the majority vote in the sample, getting stronger per student. The predicted mean will be like the base true (distribution) or the true value, allowing the random forest to overcome the variation (regression) in the prediction that each decision tree has. Scikit-learn Random Forest classifier class for generating RF models. This parameter estimation results in several different decision trees among different models of each dataset.

3.1.2 Support Vector Model:

Support Vector Machines, a supervised learning system for classification, regression, and anomaly detection so that the output can be determined later. SVM was implemented in two different ways. Support vectors are data points located on the edge of the region closest to the hyperplane in the distribution problem in figure a. The edge of the course is somewhere between these two areas. The number of classes in the dataset will be determined using the hyperplane, and the output for missing data will be estimated based on which classes are similar to the new data.

In the regression problem, linear regression is used to construct this hyperplane approximation to the nonlinear function of the maximum margin. Therefore, -insensitive loss was added as an additional parameter to allow some time variation in the tube region. Section (assuming a straight line with equation. For products manufactured outside of the greater than or less than I range, the SVR uses the concept of a penalty denoted by the C index (appropriate value). However, data points at the border will be avoided. Since the support vectors represent the points of these boundary lines, the number of support vectors decreases as you get to the plane; otherwise, the number of support vectors increases as the plane moves upward.

3.2 Model Evaluation Methods:

Several statistical scores were used to evaluate the performance of O3, NO2 and PM2.5 model, including the Coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE) and Root Mean Squared Logarithmic Error (RMSLE).

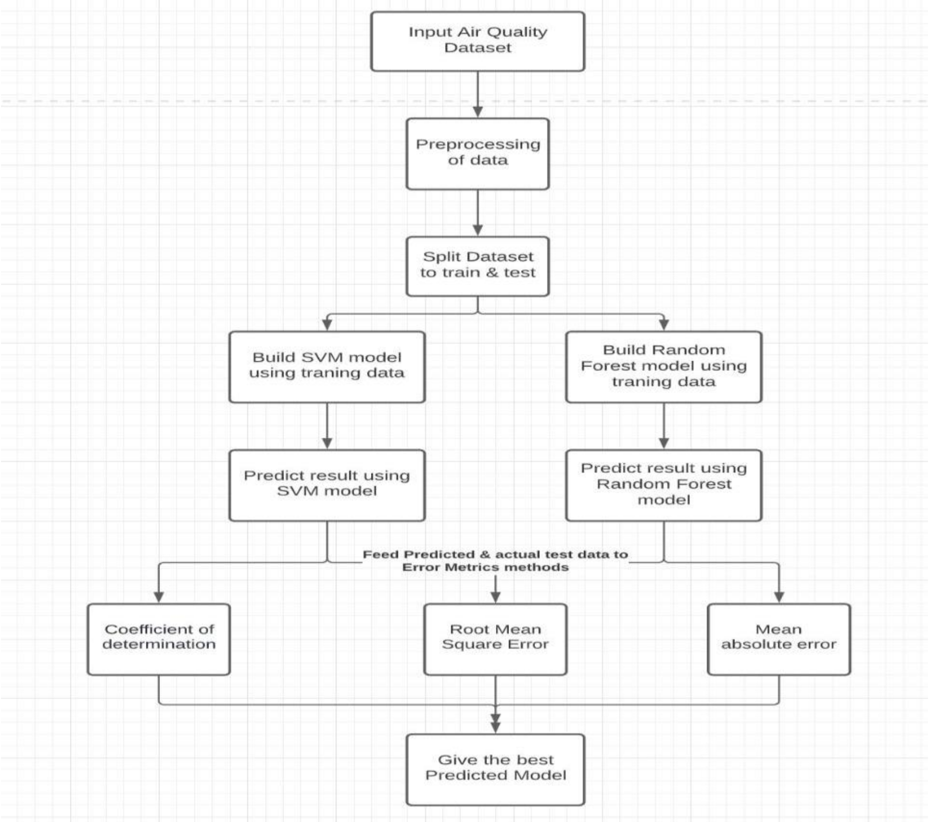
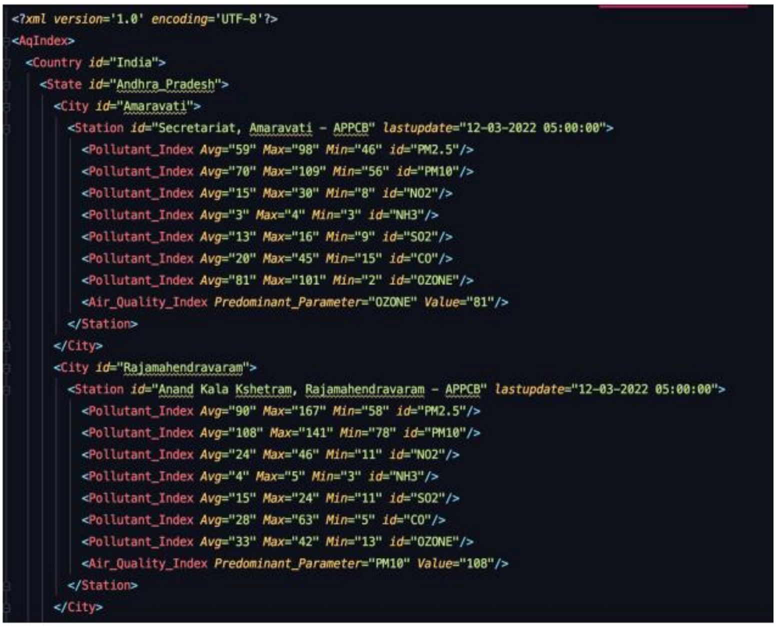

Input Air Quality Dataset
Preprocessing of data
Split Dataset to train & test
Build SVM model using traning data
Build Random Forest model using traning data
Predict result using SVM model
Predict result using Random Forest model
Feed Predicted & actual test data to Error Metrics methods
Coefficient of determination
Root Mean Square Error
Mean absolute error
Give the best Predicted Model

Fig 1: Architecture of the proposed model



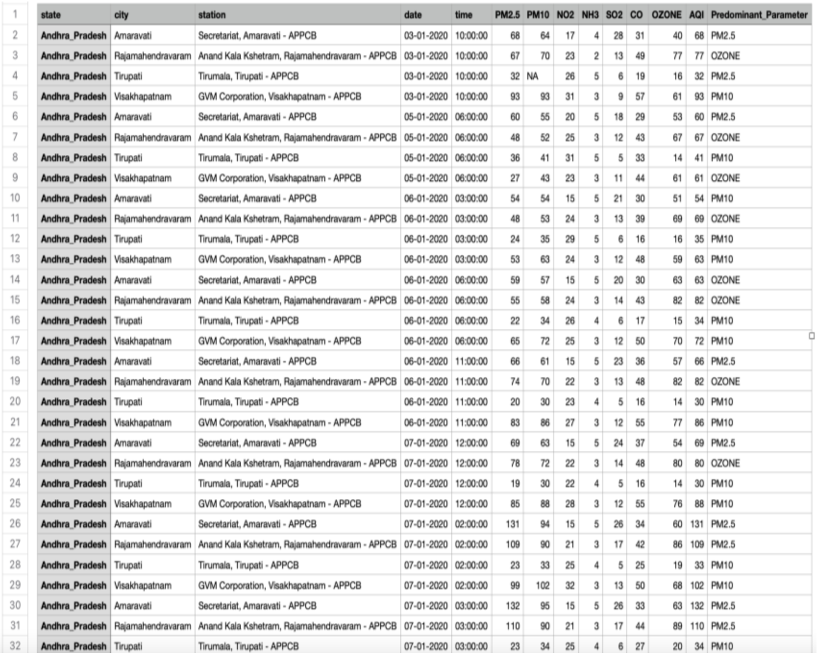
```
<?xml version='1.0' encoding='UTF-8'?>
<AqIndex>
  <Country id="India">
    <State id="Andhra_Pradesh">
      <City id="Amaravati">
        <Station id="Secretariat, Amaravati - APPCB" lastupdate="12-03-2022 05:00:00">
          <Pollutant_Index Avg="59" Max="98" Min="46" id="PM2.5"/>
          <Pollutant_Index Avg="70" Max="109" Min="56" id="PM10"/>
          <Pollutant_Index Avg="15" Max="30" Min="8" id="NO2"/>
          <Pollutant_Index Avg="3" Max="4" Min="3" id="NH3"/>
          <Pollutant_Index Avg="13" Max="16" Min="9" id="SO2"/>
          <Pollutant_Index Avg="20" Max="45" Min="15" id="CO"/>
          <Pollutant_Index Avg="81" Max="101" Min="2" id="OZONE"/>
          <Air_Quality_Index Predominant_Parameter="OZONE" Value="81"/>
        </Station>
      </City>
      <City id="Rajamahendravaram">
        <Station id="Anand Kala Kshetram, Rajamahendravaram - APPCB" lastupdate="12-03-2022 05:00:00">
          <Pollutant_Index Avg="90" Max="167" Min="58" id="PM2.5"/>
          <Pollutant_Index Avg="108" Max="141" Min="78" id="PM10"/>
          <Pollutant_Index Avg="24" Max="46" Min="11" id="NO2"/>
          <Pollutant_Index Avg="4" Max="5" Min="3" id="NH3"/>
          <Pollutant_Index Avg="15" Max="24" Min="11" id="SO2"/>
          <Pollutant_Index Avg="28" Max="63" Min="5" id="CO"/>
          <Pollutant_Index Avg="33" Max="42" Min="13" id="OZONE"/>
          <Air_Quality_Index Predominant_Parameter="PM10" Value="108"/>
        </Station>
      </City>
```

Fig 2: XML data set

3.3 Dataset Description

In this paper, the data had been acquired from the archive statistics furnished with the aid of using the Open Government Data (OGD) Platform India. The statistics from the OGD includes day by day concentrations of particulate count of sizes much less than or same to 2.5 microns (PM2.5) and particulate counts of less than or equal to 10 microns in diameter (PM10). They derive the Air Quality Index primarily based totally on those values. The statistics obtained is absolutely in XML format.

The raw data received for this evaluation is in XML format as shown in Figure,which is initially converted to .csv and executed diverse data preprocessing steps like getting ready extra correct and whole datasets via way of means of imputing lacking data, making sure data is uniformly disbursed via way of means of normalization and standardization of data, developing a smaller and compact dataset via way of means of extraction and choice of features



1	state	city	station	date	time	PM2.5	PM10	NO2	NH3	SO2	CO	OZONE	AQI	Predominant_Parameter
2	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	03-01-2020	10:00:00	68	64	17	4	28	31	40	68	PM2.5
3	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	03-01-2020	10:00:00	67	70	23	2	13	49	77	77	OZONE
4	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	03-01-2020	10:00:00	32	NA	26	5	6	19	16	32	PM2.5
5	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	03-01-2020	10:00:00	93	93	31	3	9	57	61	93	PM10
6	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	05-01-2020	06:00:00	60	55	20	5	18	29	53	60	PM2.5
7	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	05-01-2020	06:00:00	48	52	25	3	12	43	67	67	OZONE
8	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	05-01-2020	06:00:00	36	41	31	5	5	33	14	41	PM10
9	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	05-01-2020	06:00:00	27	43	23	3	11	44	61	61	OZONE
10	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	06-01-2020	03:00:00	54	54	15	5	21	30	51	54	PM10
11	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	06-01-2020	03:00:00	48	53	24	3	13	39	69	69	OZONE
12	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	06-01-2020	03:00:00	24	35	29	5	6	16	16	35	PM10
13	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	06-01-2020	03:00:00	53	63	24	3	12	48	59	63	PM10
14	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	06-01-2020	06:00:00	59	57	15	5	20	30	63	63	OZONE
15	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	06-01-2020	06:00:00	55	58	24	3	14	43	82	82	OZONE
16	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	06-01-2020	06:00:00	22	34	26	4	6	17	15	34	PM10
17	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	06-01-2020	06:00:00	65	72	25	3	12	50	70	72	PM10
18	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	06-01-2020	11:00:00	66	61	15	5	23	36	57	66	PM2.5
19	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	06-01-2020	11:00:00	74	70	22	3	13	48	82	82	OZONE
20	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	06-01-2020	11:00:00	20	30	23	4	5	16	14	30	PM10
21	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	06-01-2020	11:00:00	83	86	27	3	12	55	77	86	PM10
22	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	07-01-2020	12:00:00	69	63	15	5	24	37	54	69	PM2.5
23	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	07-01-2020	12:00:00	78	72	22	3	14	48	80	80	OZONE
24	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	07-01-2020	12:00:00	19	30	22	4	5	16	14	30	PM10
25	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	07-01-2020	12:00:00	85	88	28	3	12	55	76	88	PM10
26	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	07-01-2020	02:00:00	131	94	15	5	26	34	60	131	PM2.5
27	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	07-01-2020	02:00:00	109	90	21	3	17	42	86	109	PM2.5
28	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	07-01-2020	02:00:00	23	33	25	4	5	25	19	33	PM10
29	Andhra_Pradesh	Visakhapatnam	GVM Corporation, Visakhapatnam - APPCB	07-01-2020	02:00:00	99	102	32	3	13	50	68	102	PM10
30	Andhra_Pradesh	Amaravati	Secretariat, Amaravati - APPCB	07-01-2020	03:00:00	132	95	15	5	26	33	63	132	PM2.5
31	Andhra_Pradesh	Rajamahendravaram	Anand Kala Kshetram, Rajamahendravaram - APPCB	07-01-2020	03:00:00	110	90	21	3	17	44	89	110	PM2.5
32	Andhra_Pradesh	Tirupati	Tirumala, Tirupati - APPCB	07-01-2020	03:00:00	23	34	25	4	6	27	20	34	PM10

Fig 3: .csv file

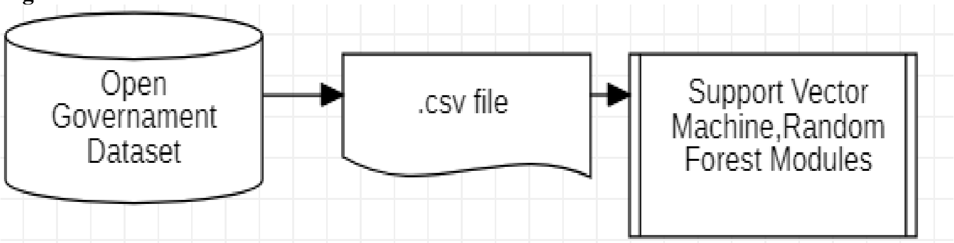


Fig 4: Input flow chart

3.3.1 Output Data Flowchart:

After all the data is processed ,it is fed to Support Vector Machine and Random Forest Regression models. The predicted results are compared with error metrics and finally best accurate model is visualized.

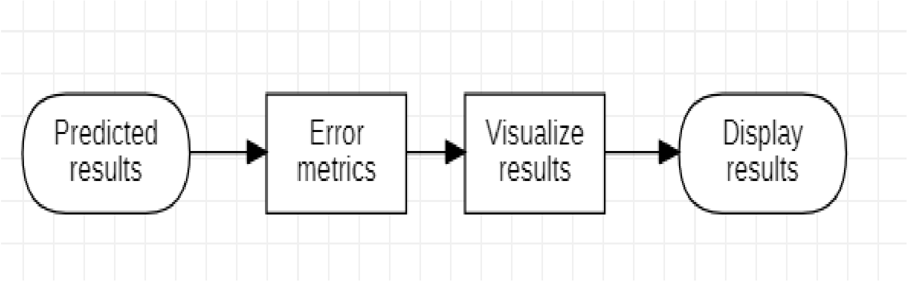


Fig 5: Output Data Flowchart

4 Experiment and Results:

4.1 Procedure:

- Keras is a High-level Neural Networks API for building network models.
- In this paper, keras is a front-end interface which is used for running the model.
- After that enter the code in Visual Studio Code and save the file as aqi.py
- Compile and run through the options present in VS code app.
- Successful completion of code opens a pop of graph of the models.
- Tensorflow is an open-source software library for machine learning applications such as Neural Networks.
- Here, Tensorflow acts as a backend.

4.2 Results

The output of this model gives the best accurate model by comparing both the outputs of SVM and Random Forest by using Error Metrics.

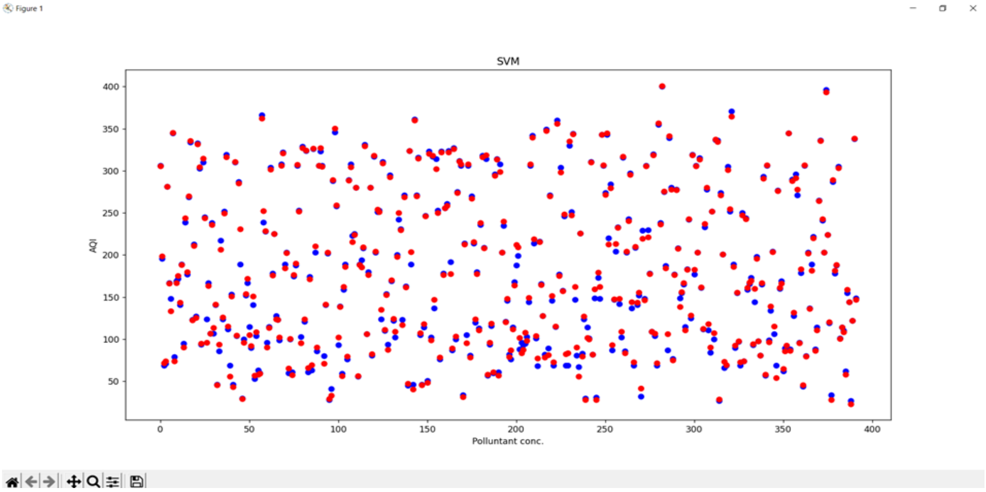


Fig 6: Output of SVM model

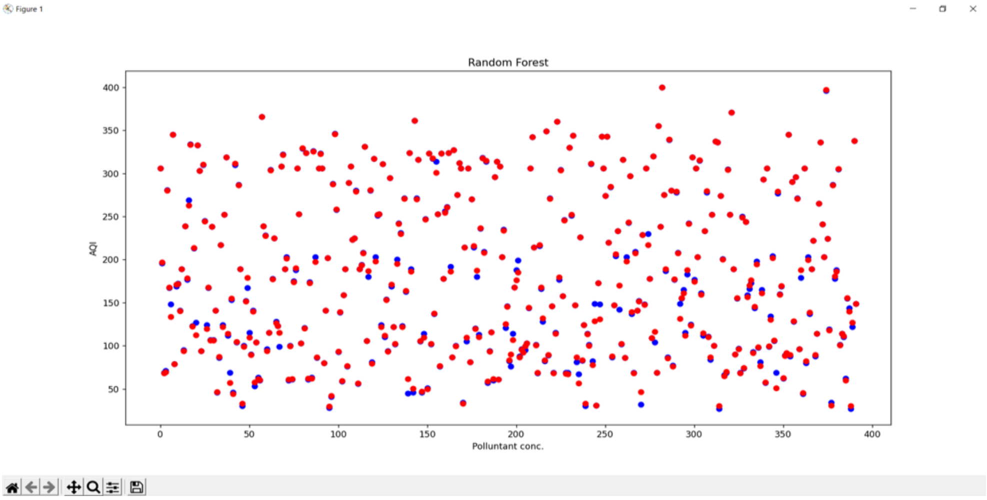
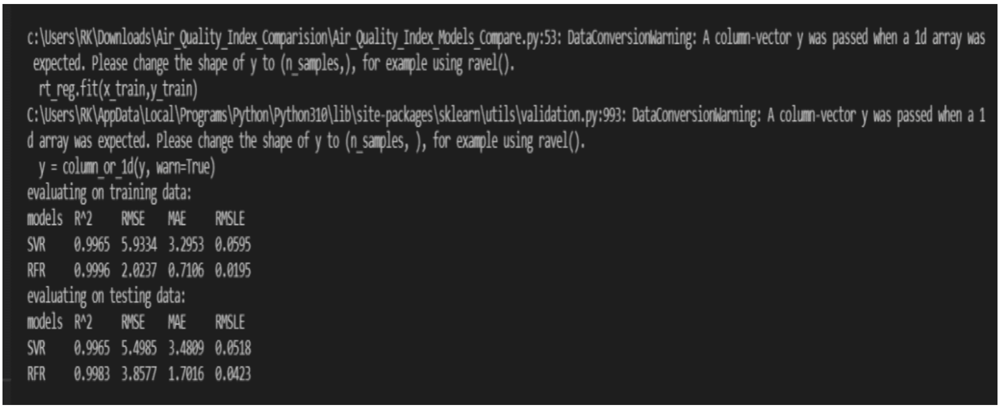


Fig 7: Output of Random Forest model

Error metrics that are used in this model for evaluation are Root Mean square error, Mean absolute error, Root mean square logarithmic error and coefficient of determination. These metrics takes the value of AQI and gives the results for comparison.



```
c:\Users\RK\Downloads\Air_Quality_Index_Comparision\Air_Quality_Index_Models_Compare.py:53: DataConversionWarning: A column-vector y was passed when a 1d array was
 expected. Please change the shape of y to (n_samples,), for example using ravel().
  rt_reg.fit(x_train,y_train)
C:\Users\RK\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\utils\validation.py:993: DataConversionWarning: A column-vector y was passed when a 1
d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
evaluating on training data:
models  R^2     RMSE    MAE     RMSLE
SVR     0.9965  5.9334  3.2953  0.0595
RFR     0.9996  2.0237  0.7106  0.0195
evaluating on testing data:
models  R^2     RMSE    MAE     RMSLE
SVR     0.9965  5.4985  3.4809  0.0518
RFR     0.9983  3.8577  1.7016  0.0423
```

Fig 8: Error metrics of the model

5 Advantages and Disadvantages

5.1 Advantages:

- It helps people in identifying the purity of air this impacts their health and reduces the chances of occurring any health issues by maintaining a moderate ambiance or as required.
- It is done by using only software which is used to predict the Air Quality Index and no hardware is required which helps in reducing the cost.
- It helps people monitor the presence of pollutants, resulting in better environmental conditions for humans to reside.
- It is used to determine the quality of air.

5.2 Disadvantages:

- The dataset may not work well in order to give accurate predictions if dataset is too complex.
- It doesn’t support sensor-based prediction.

6 Conclusion

In this project, we have created a comparative model where we compare the best accurate model using various error metrics methods. Based on error metrics methods and graph we conclude that the random forest model is better than the support vector model.

It is essential to recognise approximately AQI due to the fact except and till the humans recognise the worst influences or risks of air pollutants they'll now no longer end up that a whole lot privy to the air pollutants and attempt to lessen it. As in line with this evaluation maximum of the researchers worked on AQI and pollution attention degree forecasting in order to provide the real concept approximately AQI. Our method started out from data cleansing and processing, incomplete records, specified assessment and ultimately version building and assessment of those models to decide the fine ML set of rules that could provide us the fine end result in the given dataset.

References

[1] Zhongjie Fu, Haiping Lin, Bingqiang Huang and Jiana Yao, *Research on air quality prediction method in Hangzhou based on machine learning*, (2020).

[2] Anikender Kumar, PramilaGoyal, *Research on Machine Learning Prediction of Air Quality Index*, (2011)

[3] Mauro Castelli,[1] Fabiana Martins Clemente,[1] Aleš Popovič,[1,2] Sara Silva,[3] and Leonardo Vanneschi[1,3] *A Machine Learning Approach to Predict Air Quality* in California, **Volume 2020**, Published04 Aug (2020)

[4] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu*, Detection and Prediction of Air Pollution using Machine Learning Model*, International Journal of Engineering Trends and Technology, **volume 59**, Issue 4 May (2018).

[5] Pan B. *Application of xgboost algorithm in hourly pm2. 5 concentration prediction.* In: IOP conference series: earth and environmental science, **vol. 113**. IOP publishing; 2018, p. 012127(2018).