

Air Quality Index Prediction using Machine Learning Techniques

Bharathi P.T¹ and Malatesh S Jangannavar²

¹Assistant Professor, Department of Master of Computer Applications, Siddaganga Institute of Technology, Tumakuru, Karnataka, India

Email: bharathi_pt@sit.ac.in

²Post Graduate Student, Department of Master of Computer Applications, Siddaganga Institute of Technology, Tumakuru, Karnataka, India

Email: malateshsj19@gmail.com

Abstract— Pollution is the most vital and disturbing issues faced in today's world. Above 2000 people die due to diseases whose root cause is pollution. Pollution can be of various forms and each of these types can have different effects on different people. Increased pollution levels are capable of causing mass destruction to the earth as well as to the species residing in it. One of the prevalent environmental challenges right now is air pollution. Air pollution has been noticeable as one of the most important problems of metropolitan regions around the globe, exclusively in Delhi, Beijing and Tehran where its inhabitants and administrators have long been struggling with air pollution impairment such as the health issues of its citizens. With the rapid development in the availability of data and computational technologies, various machine learning techniques have been proposed for predicting air pollution. Air Quality Index can be predicted using both Classification and Regression models. Decision tree, Linear Regression, Support Vector Regression and Random Forest Techniques are implemented in this proposed research work. From the above models, it is evident that, Random Forest Regression based model out performs better in the air quality index prediction with an accuracy of 98%.

Index Terms— Air Pollution, Machine Learning Techniques, Air Quality Index (AQI), Decision tree, Random Forest, Linear Regression, Support Vector Regression.

I. INTRODUCTION

Pollution has been an increasing world issue since the American Industrial revolution. Nowadays, pollution is one of the biggest universal killers. Above 2000 people die every day due to pollution related issues. The major categories of pollution are: Air pollution, Light pollution, Noise pollution, Radioactive pollution, Soil pollution, Thermal pollution and Water pollution. The environment and people are effected by each type of pollution in various ways. Pollution levels are as high as they are due to human consumption and their communication with the world initiating destruction to the world and all species living on the earth.

Our planet has been plagued by bad air quality since the beginning of the industrial era. Especially in recent years, major metropolitan cities like Beijing, New Delhi and Tehran have had extremely high air pollution levels which led to the development of major health problems among the people living in these areas.

Air pollution can be recognized as one of the main detriments to human health. According to a survey led by

“WHO” (World Health Organization) over the recent years, air pollution or air contamination has turned out to be a concerning hazard, being accountable for over 7 million deaths annually. Air pollution may be defined as the insufficient variation in the physical, chemical or biological physiognomies of the air which impedes the lifetime in addition leading to possible well-being difficulties. Air contamination stands as a vital risk aspect for health such as difficulty in inhalation and coughing, irritation of the eyes, nose and throat, bronchitis, lung cancer, cardiovascular problems, contagions in the skin.

The air we inhale can comprehend an assortment of contaminants radiated into the outdoors of the atmosphere (which is also known as the ambient air) in addition to the air indoors. Air pollution can comprise a combination of solid particles, liquid droplets and gases from a variety of sources such as industry, motor vehicles, heating appliances and tobacco smoke. Air pollution can also be generated by natural events such as bush fires and can contain windblown dust, pollen and mould spores. The composition of air pollution can vary greatly depending on the season, the weather and the types and number of sources.

A. Air Quality Index (AQI) and Calculation of AQI:

Air pollution and its anticipation are facing continuous scientific challenges. Affecting human’s respiratory system and cardiovascular system, they are a reason for augmented mortality and augmented risk for diseases. To ensure better living conditions, air quality has to be constantly monitored and assessed. To standardize the air quality, U.S. Environmental Protection Agency (EPA) uses the Air Quality Index (AQI).

The Air Quality Index (AQI) is an index for recording the day-to-day air quality. AQI tells the public exactly how polluted or clean the air is, and what are the allied health effects might be a concern for them. The AQI emphasizes on health effects that public may experience within a couple of hours or days after breathing the contaminated air. The AQI is premeditated by EPA for five main pollutants controlled by the Clean Air Act: Ground-Level Ozone (O₃), Particulate Matter (PM₁₀ or PM_{2.5}), Carbon Monoxide (CO), Sulphur Dioxide (SO₂) and Nitrogen Dioxide (NO₂). To protect the public health, national air quality standards have been recognized by EPA for each of these contaminants. Air Quality Index is a measure which runs from 0 to 500. The higher the AQI esteem, the more noteworthy the degree of air contamination and the more noteworthy the wellbeing concern. For instance, AQI estimation of 50 or beneath signifies to Good air quality, while AQI esteem more than 300 signifies to Hazardous air quality.

For every contamination an AQI estimation of 100 usually relates to an ambient air concentration which matches the level of the short-term national ambient air quality standard for prediction of public health. The AQI values at or underneath 100 are usually believed to be Satisfactory air quality. When the AQI values are over 100, then it is thought of as Unhealthy air quality. The AQI is alienated into six groups. Each group corresponds to a diverse level of health concern. Each group has a particular color. The color makes it simple for individuals to rapidly decide if the air quality is arriving at unhealthy levels in their societies.

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Fig 1: AQI values with colour symbols and the levels of Health Concern

Fig 1 describes the specified colour for each category, the levels of health concern and the corresponding Air Quality Index values along with the description of Air Quality.

Fig 2 and Fig 3 describes the AQI values and the corresponding ambient concentrations of the pollutants i.e., health breakpoints as well as the associated health impacts for the major pollutants.

Fig 4 describes the EPA’s table of breakpoints for the major pollutants along with the AQI category.

AQI Category, Pollutants and Health Breakpoints								
AQI Category (Range)	PM ₁₀ (24hr)	PM _{2.5} (24hr)	NO ₂ (24hr)	O ₃ (8hr)	CO (8hr)	SO ₂ (24hr)	NH ₃ (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Fig 2: AQI Category, Pollutants and Health Breakpoints

AQI	Associated Health Impacts
Good (0–50)	Minimal impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people.
Moderately polluted (101–200)	May cause breathing discomfort to people with lung disease such as asthma, and discomfort to people with heart disease, children and older adults.
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease.
Very poor (301–400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases.
Severe (401–500)	May cause respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

Fig 3: Associated Health Impacts

O ₃ (ppb)	O ₃ (ppb)	PM _{2.5} (µg/m ³)	PM ₁₀ (µg/m ³)	CO (ppm)	SO ₂ (ppb)	NO ₂ (ppb)	AQI	AQI
<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>C_{low} - C_{high}</i> (avg)	<i>I_{low} - I_{high}</i>	Category
0-54 (8-hr)	-	0.0-12.0 (24-hr)	0-54 (24-hr)	0.0-4.4 (8-hr)	0-35 (1-hr)	0-53 (1-hr)	0-50	Good
55-70 (8-hr)	-	12.1-35.4 (24-hr)	55-154 (24-hr)	4.5-9.4 (8-hr)	36-75 (1-hr)	54-100 (1-hr)	51-100	Moderate
71-85 (8-hr)	125-164 (1-hr)	35.5-55.4 (24-hr)	155-254 (24-hr)	9.5-12.4 (8-hr)	76-185 (1-hr)	101-360 (1-hr)	101-150	Unhealthy for Sensitive Groups
86-105 (8-hr)	165-204 (1-hr)	55.5-150.4 (24-hr)	255-354 (24-hr)	12.5-15.4 (8-hr)	186-304 (1-hr)	361-649 (1-hr)	151-200	Unhealthy
106-200 (8-hr)	205-404 (1-hr)	150.5-250.4 (24-hr)	355-424 (24-hr)	15.5-30.4 (8-hr)	305-604 (24-hr)	650-1249 (1-hr)	201-300	Very Unhealthy
-	405-504 (1-hr)	250.5-350.4 (24-hr)	425-504 (24-hr)	30.5-40.4 (8-hr)	605-804 (24-hr)	1250-1649 (1-hr)	301-400	Hazardous
-	505-604 (1-hr)	350.5-500.4 (24-hr)	505-604 (24-hr)	40.5-50.4 (8-hr)	805-1004 (24-hr)	1650-2049 (1-hr)	401-500	

Fig 4: EPA's Table of Breakpoints

The Air Quality Index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is discontinuous jump of one AQI unit. The formula to calculate the Individual Air Quality Index (IAQI) given the concentration of the pollutant is as follows:

$$I = \frac{I_{\text{high}} - I_{\text{low}}}{C_{\text{high}} - C_{\text{low}}} (C - C_{\text{low}}) + I_{\text{low}}$$

Where,

I = Individual Air Quality Index
 C = Pollutant concentration
 C_{low} = Concentration breakpoints that is < C
 C_{high} = Concentration breakpoint that is >= C
 I_{low} = Index breakpoint corresponding to C_{low}
 I_{high} = Index breakpoint corresponding to C_{high}

The formula to find the Air Quality Index given the Individual Air Quality Index of various pollutants is as follows:

$$AQI = \max(IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n)$$

Where,

AQI = Air Quality Index
 IAQI = Individual Air Quality Index
 Max = Maximum

The present work is organized as follows: Section 2 describes literature survey of the problem statement. Section 3 describes methodology of the work. Section 4 elaborates on the proposed methodology of the problem statement and section 5 discusses on implementation and experimental results obtained with different machine learning techniques. Finally, section 6 concludes with some final remarks and all the references been made for completion of this work.

II. LITERATURE SURVEY

A literature survey is a summary of published research work which are relevant to the topic under consideration for research work. The purpose is to create familiarity with the research on air quality index prediction.

In [1] Air pollution is the major concern since it can cause variable health effects on human beings. This causes a need to utilize the available information gathered on the environment for the advancement of Smart Cities so as to improve the air quality which thus can improve the quality of life for a city. Here, the main focus is on the analysis of the air which is based on the data collected of various pollutants such as Ground-Level Ozone (O₃), Carbon Monoxide (CO), Sulphur Dioxide (SO₂) and Nitrogen Dioxide (NO₂). The proposed solution is to use the classification approach for data analysis and prediction of air quality. The models derived are used to decide the data classes and also to predict the future values.

The algorithms implemented include Naïve Bayes and Decision Tree. In Naïve Bayes algorithm, a classifier is constructed given a set of training samples and by using the classifier, testing sample is assigned with the corresponding class labels. In Decision Tree, J48 classifier is used which is the simple for classification and hence the decision tree is constructed which is then applied to the testing set to assign the class labels. The Naïve Bayes classifier gives the accuracy of 86.663% and 13.337% instances are classified as incorrect which is the error rate. The Decision Tree J48 classifier gives the accuracy of 91.9978% and 8.0022% instances are classified as incorrect which is the error rate. Hence Decision Tree with J48 algorithm gives the more accurate result compared to the Naive Bayes algorithm. Therefore, AQI values for a region are provided as an input to the J48 classifier, where the class labels are assigned based on the corresponding health concern.

In [2] the major concern is to forecast the air quality in Delhi as prediction of urban air pollution turns out to be a vital substitute to control its harmful outcomes. To forecast and to predict the Air Quality Index of major pollutants like Ground-Level Ozone (O₃), Particulate Matter (PM₁₀ and PM_{2.5}), Carbon Monoxide (CO), Sulphur Dioxide (SO₂) and Nitrogen Dioxide (NO₂) at the three monitoring sites located in the different districts of New Delhi, the proposed solution makes use of different regression and classification techniques such as Adaptive Boosting regression, Decision Tree Regression, Gradient Boosting regression, Linear Tree Regression, Random Forest regression, Stochastic Descent Gradient Regression, Support Vector Regression and Artificial Neural Networks.

The pollutant data was collected from three monitoring stations namely, Anand Vihar, Punjabi Bagh and R. K. Puram. The pollutant concentrations were collected from an "Air and Noise Pollution Monitoring System". The machine learning techniques are evaluated using Mean Square Error (MSE), Mean Absolute Error (MAE) and Coefficient of Determination (R²) for the pollutant concentrations data collected from three monitoring stations. When compared the results of all the techniques, it says that Artificial Neural Networks and Support Vector Regression are the best suited machine learning techniques for predicting the air quality in Delhi.

In [3] the foremost environmental challenge in smart city environments is dealing with air pollution. Monitoring of real-time pollution data permits local authorities to study the present traffic condition of the city and thus make the verdicts accordingly. Here the air pollution prediction is performed on the multiple dataset's of four cities namely, Beijing City, Shanghai City, Shenyang City and Guangzhou City by using four of the advanced regression techniques like Decision Tree Regression, Gradient Boosting Regression, Random Forest Regression and ANN Multilayer-Perceptron Regression. A relative analysis is presented to define the best model for predicting the air quality with reference to data size, error rate and pre-processing time. The evaluation metrics used for the comparison of these regression techniques are Mean Absolute Error (MAE) and Random Mean Square Error (RMSE). As per the results got from simulation, the best regression technique for air pollution prediction is the Random Forest Regression for multiple datasets having varying location and size and also different characteristics.

In [4], in the recent decades, air pollution has turned out to be a significant environmental problem. Air quality forecasting plays a vital role in warning general public about the air pollution as well as regulating the air pollution. In this paper, the prediction of Air Quality Index in Beijing city and the prediction of Nitrogen Oxides (NOx) concentration in an Italian city are carried out by implementing the regression techniques like Random Forest Regression and Support Vector Regression. The metrics used to evaluate the performance of the regression models include the Correlation Coefficient (r), Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE).

The following norms can be used as references to find a good regression model: (i) high r , (ii) high R^2 and (iii) low RMSE. Finally according to the experimental results, the Support Vector Regression model performed better than the Random Forest Regression model in the prediction of Air Quality Index ($r = 0.9987$, $R^2 = 0.9776$ and $RMSE = 7.666$). The Random Forest Regression model performed better than the Support Vector Regression model in the prediction of Nitrogen Oxides (NOx) concentration ($r = 0.9180$, $R^2 = 0.8401$ and $RMSE = 83.6716$).

In [8], four simple machine learning algorithms like neural network, k-nearest neighbour, support vector machines and decision tree are been implemented. In this research work, air pollution database are taken by measurements stations in capital city of the Republic of Macedonia which contains data of each day in the year 2017. The results are promising and it was proven that implementation of these algorithms could be very efficient in predicting air quality index. Among the four simple machine learning algorithms like neural network, k-nearest neighbour, support vector machines and decision tree, it is concluded that the most accurate algorithm for classification is neural network with maximum accuracy of 92.3%, while KNN algorithm has an accuracy of 80.0%, DT algorithm has an accuracy of 78.0% and SVM algorithm has produced an accuracy of 80.0%.

III. METHODOLOGY

A Machine Learning

Machine Learning is described as "a study area that enables computers to learn without explicit programming". Machine Learning (ML) can be described in a very layman way as automating and enhancing the computer's learning process based on their experiences without ever being programmed, i.e. without any human assistance. The process starts with feeding good quality data and then training our machines (computers) using the data and various algorithms to create machine learning models. Choosing algorithms depends on what type of data we have and what type of function we attempt to automate.

B Programming Language

Python is a high level interpreted programming language. Python language has some emphasizing reliability of code and it uses the whitespaces. For large scales and small scales the python language enables the clear programming structures. Python coding language has some features such as

- Automatic multiple memory management.
- It will change automatically.
- Some of the challenging and standard library.
- It's Time efficient
- It's Efficient and Functional
- It's Imperative and procedure.

Python is an open-source environment and it's capable of interpreting for many operating systems. In the industry type it's oriented for many community based models. Here we are using python language for coding because,

- It has Coding guidelines for good programming
- Fast execution of Results.
- Time efficiency.
- Good Analytical model can be done by using python.
- Efficient Testing Result.
- Metrics can be done by using the python language.
- Commenting section for coding
- The above are the some of the features of the python coding language.

C Anaconda Distribution

Anaconda is open software centre for the Python and r distribution. It's free software for the field of data science and machine learning. The applications related to the anaconda packages are used in the field of Predictive analytics, scientific computing and Data processing. Anaconda is helpful for the efficient Deployment and package management. Its inherently manages the different versions of the package. Over 6 million users are efficiently using the anaconda packages with respect to the different operating system such as Windows, Linux and Mac operating system.

Anaconda distribution includes a Anaconda Navigator which is version of Graphical user interface (GUI) which is desktop based versions which allows to launch different anaconda packages and environments. Anaconda has a local repository which can be further installed in the environment. Some of the python coding background can be seen such as Spyder, Jupyter etc are some of the python programming tool for the efficient and classic output based results.

IV. PROPOSED METHODOLOGY

This section describes the methodology used for designing the research work development. The methodology or the process adopted in this research work is divided into following set of subsections as shown in Fig 4.1.

The first step in this proposed research work is to collect the data required that is relevant to start the project. The data collected is analysed and then for the data collected pre-processing is performed by dropping the unwanted fields and observations. The target data i.e., the independent variable (y) will be the Air Quality Index which is calculated by considering the Individual Air Quality Index of four pollutants, where the highest IAQI value becomes the final AQI value for that particular row. Likewise AQI is calculated for all the observations. Therefore the actual dataset is obtained. Next, the actual dataset is split into input dataset and target dataset. Further both the input dataset and the target dataset is split into training dataset and testing dataset.

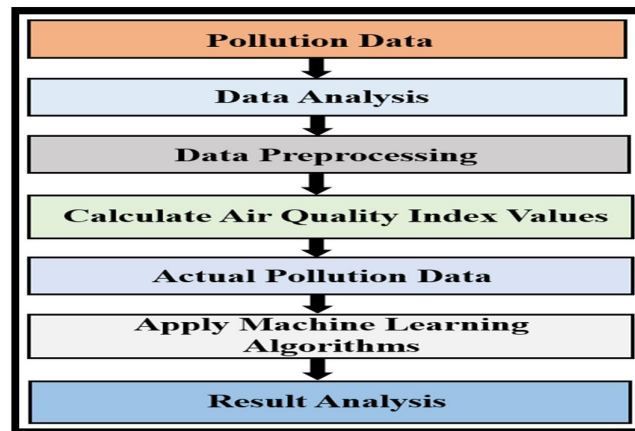


Fig 4.1: Detailed working process of the Proposed System

Furthermore, train the training dataset by applying Decision Tree Regression to predict the result. Next predict the result with the train dataset which is not trained and find out the accuracy of Air Quality Index Prediction by using regression metrics like R2, MSE, RMSE and MAE. Likewise, predict the results of train dataset and test

dataset by using the other four regression techniques and also find out the accuracy by using the four regression metrics. Finally, compare all the four machine learning techniques used to predict the Air Quality Index and analyse the results as to which Technique is the best for prediction.

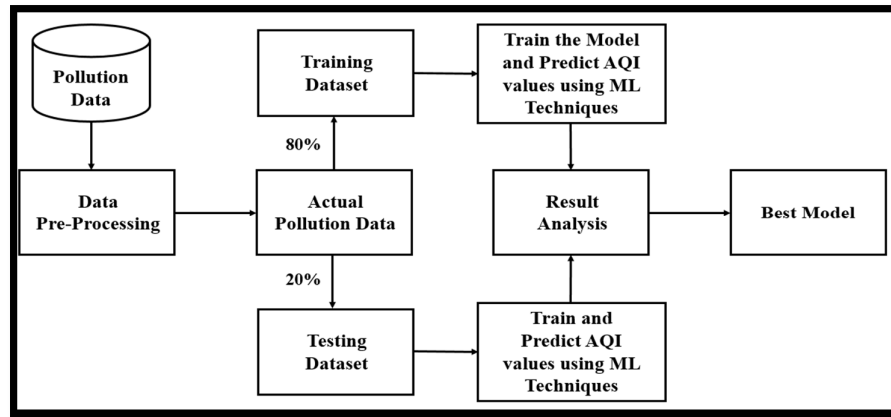


Fig 4.2: System Architecture for AQI Prediction

This section gives architecture of the proposed system. It is the conceptual model that defines structural and behavioural representation of the proposed system. It is an outcome of the design process. It is represented by block diagram as shown in Fig 4.2 and Fig 4.3 describes the step by step procedure for the entire project.

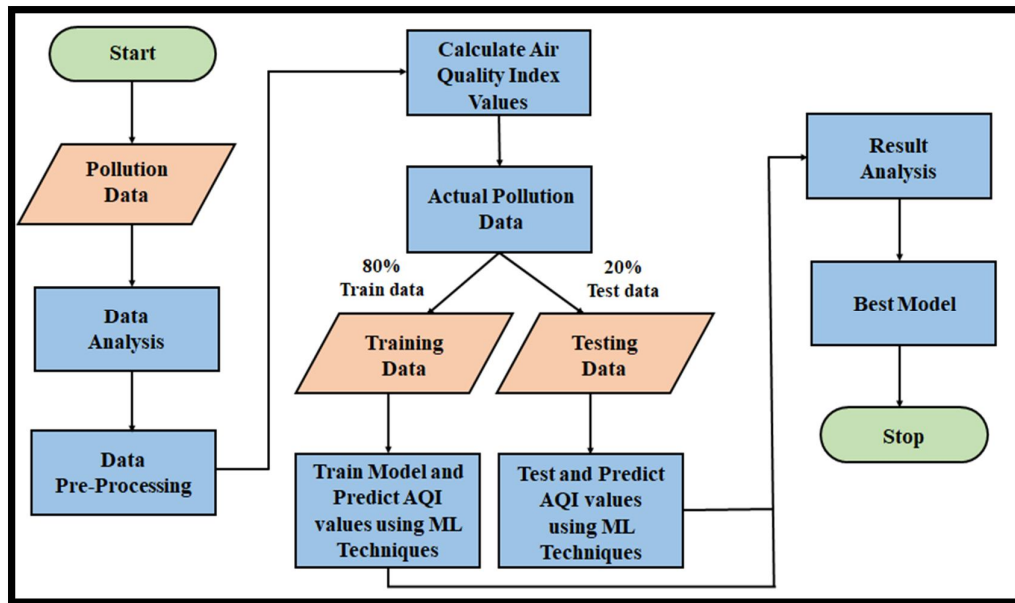


Fig 4.3: Flowchart of the proposed system

V. RESULTS AND DISCUSSION

A. Decision Tree Regression

Decision Tree is a type of Supervised Machine Learning where the data is continuously divided according to some parameter. Decision trees are called regression trees where the target variable will take the continuous values. Decision Regression Analysis of the Regression Tree is used where the expected outcome can be considered as a real number. Decision tree regression asserts an object's characteristics and trains a model in a tree's structure to anticipate information to generate meaningful continuous output in the future.

The training dataset is trained by applying Decision Tree Regression to predict the result. The result is predicted with the test dataset which is not trained and the accuracy of AQI prediction is analysed by using regression metrics like R-Square (R²), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

- *R-Square (R²):*

R-Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R-Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean. R-Square value will be in the range 0 to 1. Higher the value indicates a better fit between prediction and actual value.

- *Mean Square Error (MSE)*

While R Square is a relative measure of how well the model fits on dependent variables, MSE is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted result deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model [5, 9].

- *Root Mean Square Error (RMSE)*

RMSE is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation [10]

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

- *Mean Absolute Error (MAE)*

Mean Absolute Error is similar to Mean Square Error. However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Comparing MSE and RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalization to big prediction error by squaring the errors while MAE treats all errors the same.

	Actual	Predicted
14327	47.0	41.672635
241910	36.0	30.464749
212836	36.0	41.672635
85307	35.0	30.464749
15539	33.0	30.464749
...	...	...
240382	30.0	30.464749
9657	64.0	60.429963
174992	47.0	41.672635
285284	67.0	60.429963
310636	41.0	45.086536
20000 rows × 2 columns		

Fig 5.1: Actual and Predicted AQI Values of the testing dataset for Decision Tree Technique

Fig 5.1, describes the actual and predicted AQI values of testing data for decision tree regression technique. From figure it can be seen that when the AQI values are predicted little variance is observed between actual and predicted values.

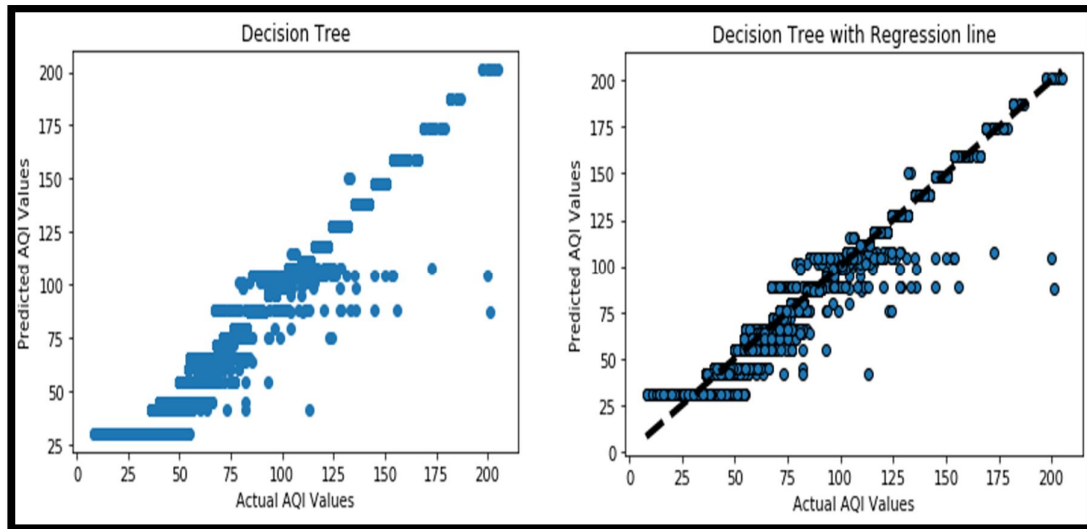


Fig 5.2: Scatter Plot of Actual v/s Predicted AQI Values for Decision Tree Technique

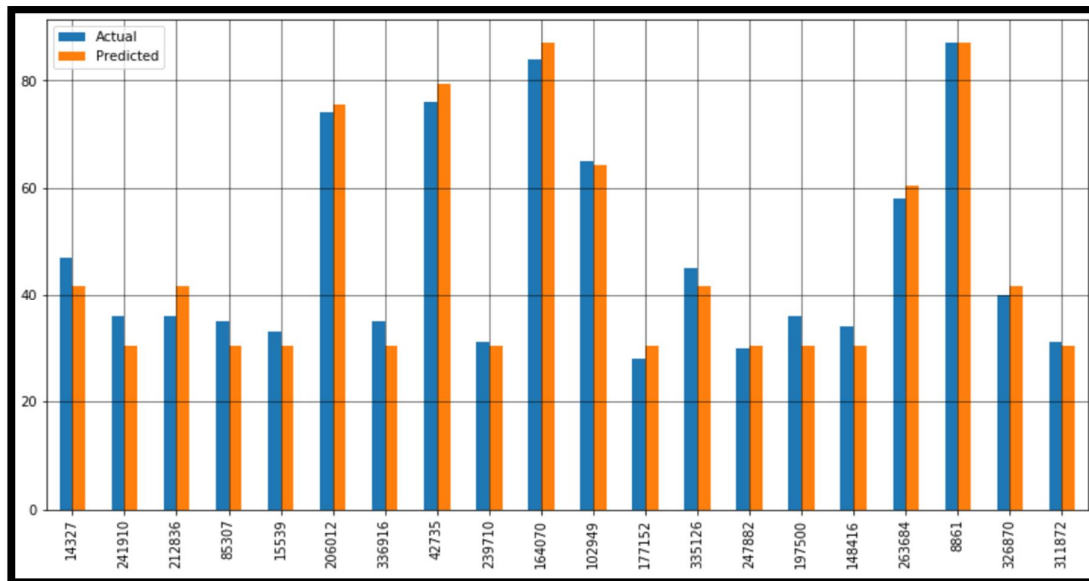


Fig 5.3: Bar Graph of Actual v/s Predicted AQI Values for Decision Tree Technique

Fig 5.2 describes the scatter plot of actual and predicted AQI values for Decision Tree Regression along with the regression line and Fig 5.3 describe the bar plot for actual and predicted AQI values. From both the figures we can see that there is a slight difference between the actual and predicted AQI values.

B. Random Forest Regression

Random Forest is a type of supervised Technique of machine learning based on learning the ensemble. Ensemble learning is a type of learning where various types of Techniques are joined together or the same technique is joined so many times to form a more powerful model of prediction. The random forest technique combines multiple Technique of the same type i.e. multiple decision trees, resulting in a forest of trees.

	Actual	Predicted
14327	47.0	47.0
241910	36.0	36.0
212836	36.0	36.0
85307	35.0	35.0
15539	33.0	33.0
...	...	...
240382	30.0	30.0
9657	64.0	64.0
174992	47.0	47.0
285284	67.0	67.0
310636	41.0	41.0

20000 rows × 2 columns

Fig 5.4: Actual and Predicted AQI Values of the testing dataset for Random Forest Technique

Fig 5.4 describes the actual and predicted AQI values of testing data for Random Forest Regression algorithm. From figure it can be seen that when the AQI values are predicted for testing data, actual and predicted values are the same (no variance). Hence this model predicts the values accurately.

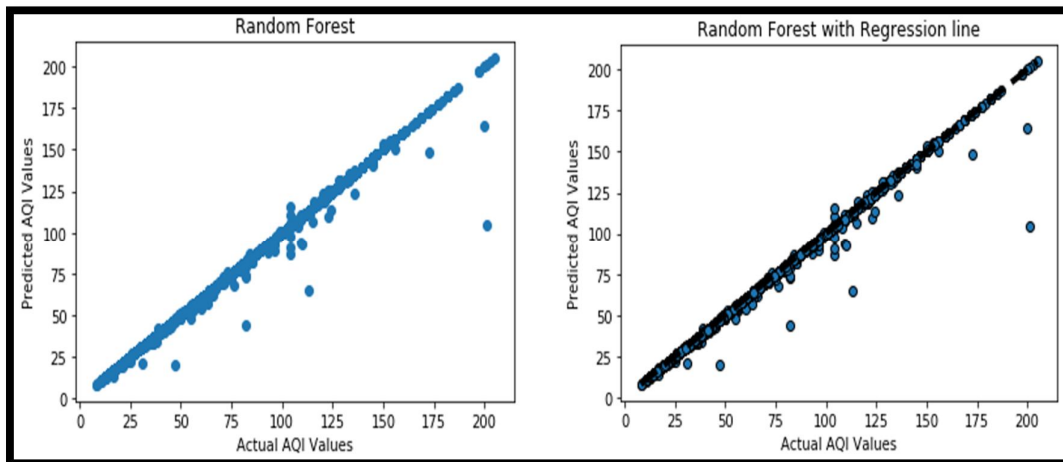


Fig 5.5: Scatter Plot of Actual v/s Predicted AQI Values for Random Forest Technique

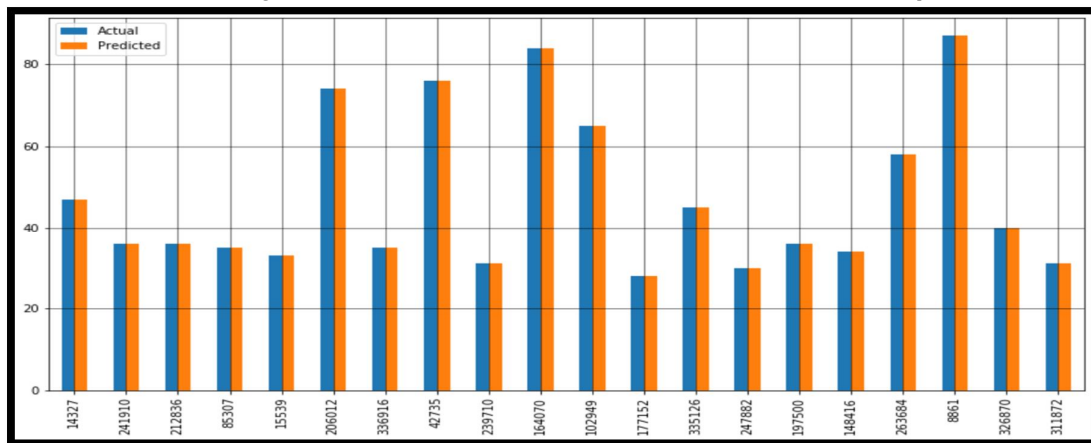


Fig 5.6: Bar Graph of Actual v/s Predicted AQI Values for Random Forest Technique

Fig 5.5 describes the scatter plot of actual and predicted AQI values for Random Forest Regression along with the regression line and Fig 5.6 describe the bar plot for actual and predicted AQI values. From both the figures, it is clear that there is no difference between the actual and predicted AQI values.

C. Linear Regression

Simple Linear Regression is a statistical method of summarizing and analysing relationships between two continuous (quantitative) variables. Linear regression is a linear model in which a model assumes a linear relationship between the input (x) and the single output (y) variables. Here, a linear combination of the input variables (x) can calculate the y. The approach is called a simple linear regression when there is one single input variable (x). The procedure is called multiple linear regressions when there are multiple input variables.

	Actual	Predicted
14327	47.0	39.049873
241910	36.0	33.720295
212836	36.0	39.546070
85307	35.0	30.204193
15539	33.0	39.821888
...	...	...
240382	30.0	29.596982
9657	64.0	58.492959
174992	47.0	46.505528
285284	67.0	58.848031
310636	41.0	32.508926
20000 rows × 2 columns		

Fig 5.7: Actual and Predicted AQI Values of the testing dataset for Linear Regression Technique

Fig 5.7 describes the actual and predicted AQI values of testing data for Linear Regression algorithm. From figure it can be seen that when the AQI values are predicted for testing data, little variance is observed between actual and predicted values.

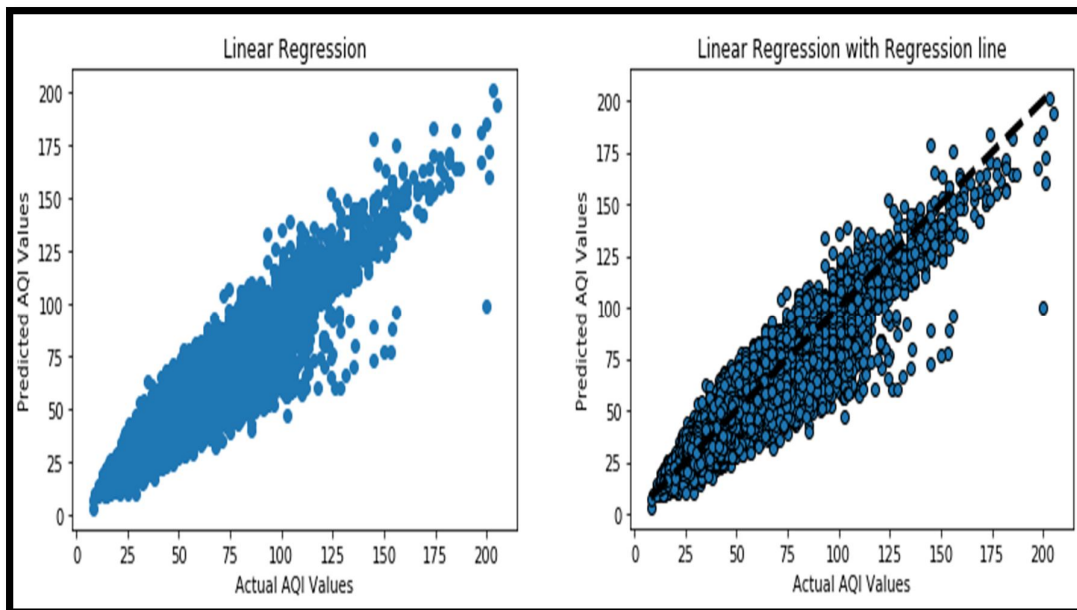


Fig 5.8: Scatter Plot of Actual v/s Predicted AQI Values for Linear Regression Technique

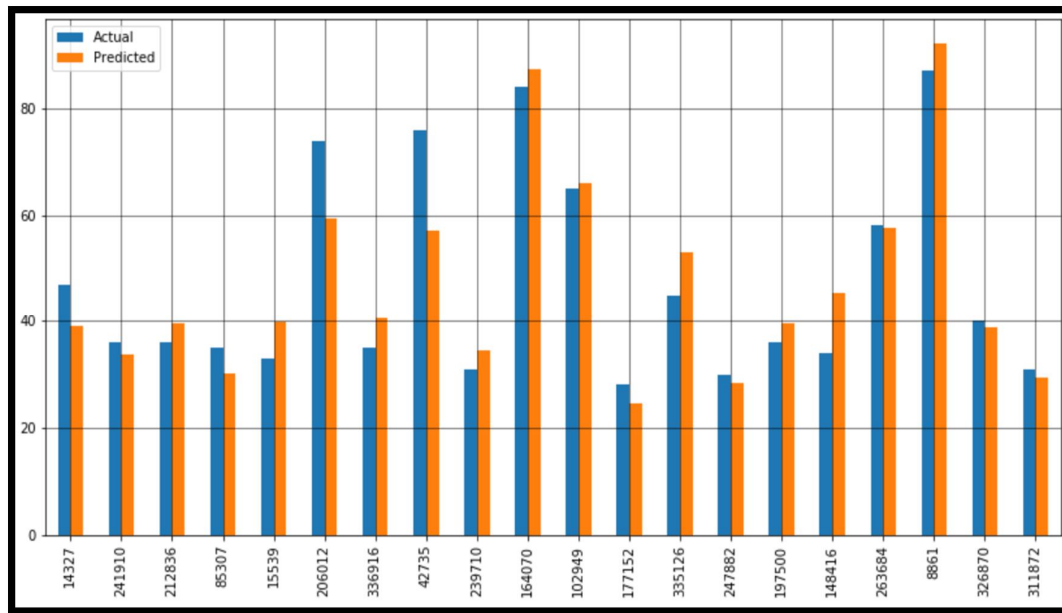


Fig 5.9: Bar Graph of Actual v/s Predicted AQI Values for Linear Regression Technique

Fig 5.8 describes the scatter plot of actual and predicted AQI values for Linear Regression along with the regression line and Fig 5.9 describe the bar plot for actual and predicted AQI values. From both the figures we can see that there is a lot of difference between the actual and predicted AQI values.

D. Support Vector Regression

Supporting vector regression (SVR) is quite distinct from other types of Regression. To predict a continuous variable, it uses the Support Vector Machine (SVM, classification Technique). While other linear regression models are attempting to minimize the error between the predicted and actual value, Support Vector Regression is attempting to fit the best line within a predefined or threshold error [8].

Support Vector Machine has different types of kernel functions. In this proposed work, the author has tried implementing radial basis function, polynomial and linear models.

When Kernel is rbf

	Actual	Predicted
14327	47.0	47.495076
241910	36.0	36.131901
212836	36.0	35.371557
85307	35.0	34.469408
15539	33.0	35.547525
...	...	...
240382	30.0	29.879817
9657	64.0	64.422496
174992	47.0	45.961614
285284	67.0	67.375203
310636	41.0	40.936150
20000 rows × 2 columns		

Fig 5.10: Actual and Predicted AQI Values of the testing dataset for SVR with (rbf) Technique

Fig 5.10 describes the actual and predicted AQI values of testing data for Support Vector Regression algorithm when the kernel is rbf. From figure it can be seen that when the AQI values are predicted for testing data, little variance is observed between actual and predicted values. The predicted values are almost near to the actual values.

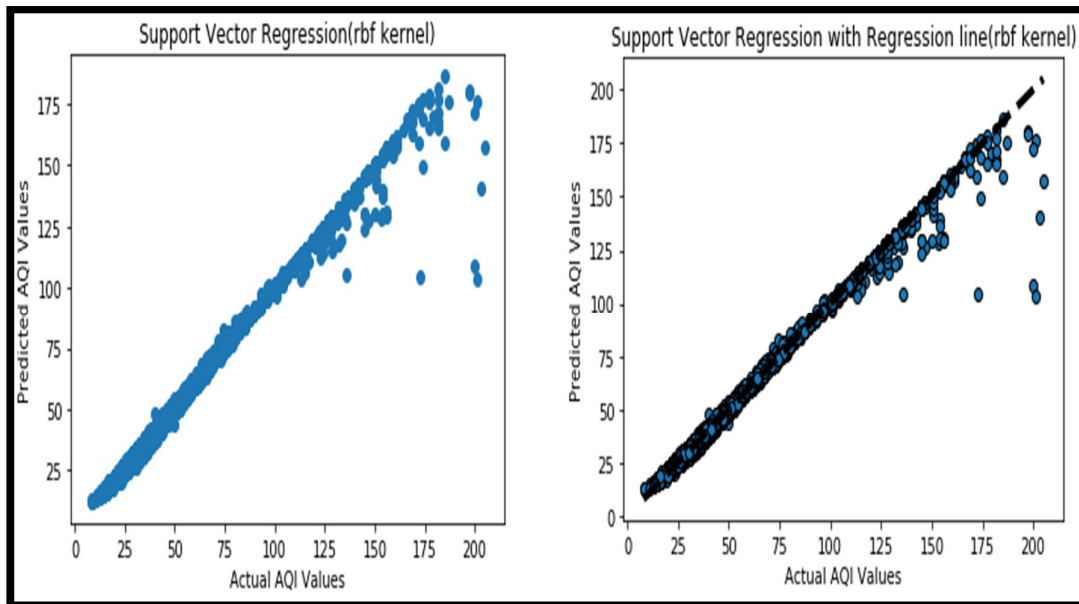


Fig 5.11: Scatter Plot of Actual v/s Predicted AQI Values for SVR with (rbf) Technique

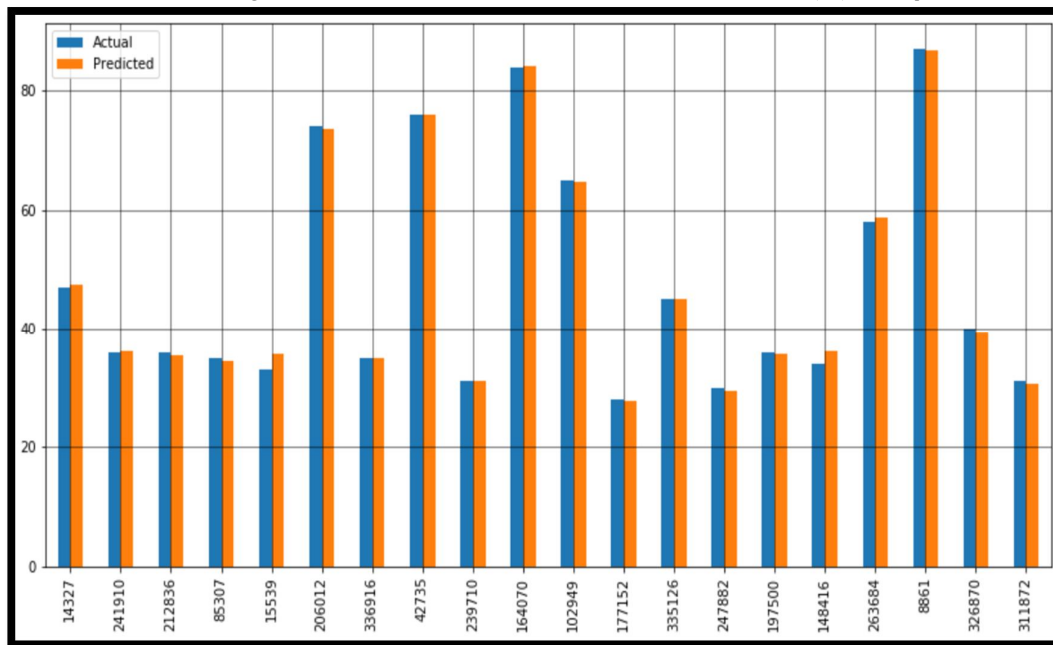


Fig 5.12: Bar Graph of Actual v/s Predicted AQI Values for SVR with (rbf) Technique

Fig 5.11 describes the scatter plot of actual and predicted AQI values for Support Vector Regression when the kernel is rbf along with the regression line and Fig 5.12 describe the bar plot for actual and predicted AQI values. From both the figures we can see that there is a slight difference between the actual and predicted AQI values.

When Kernel is polynomial

Fig 5.13 describes the actual and predicted AQI values of testing data for Support Vector Regression algorithm when the kernel is polynomial. From figure it can be seen that when the AQI values are predicted for testing data, little variance is observed between actual and predicted values.

	Actual	Predicted
14327	47.0	41.668367
241910	36.0	34.826384
212836	36.0	37.581209
85307	35.0	36.336993
15539	33.0	37.576744
...	...	...
240382	30.0	34.069196
9657	64.0	45.946835
174992	47.0	43.640001
285284	67.0	47.747196
310636	41.0	35.111407
20000 rows × 2 columns		

Fig 5.13: Actual and Predicted AQI Values of the testing dataset for SVR with (polynomial) Technique

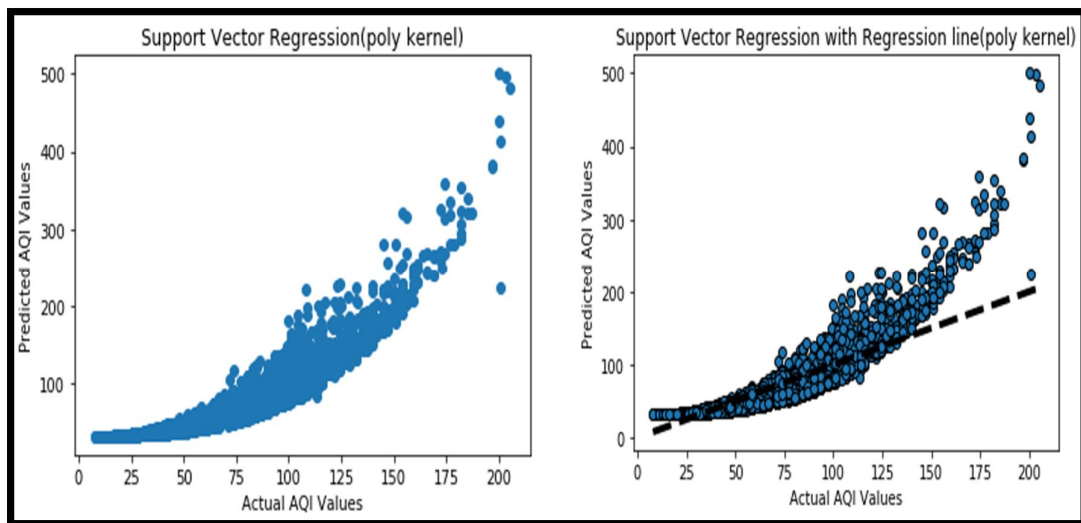


Fig 5.14: Scatter Plot of Actual v/s Predicted AQI Values SVR with (polynomial) Technique

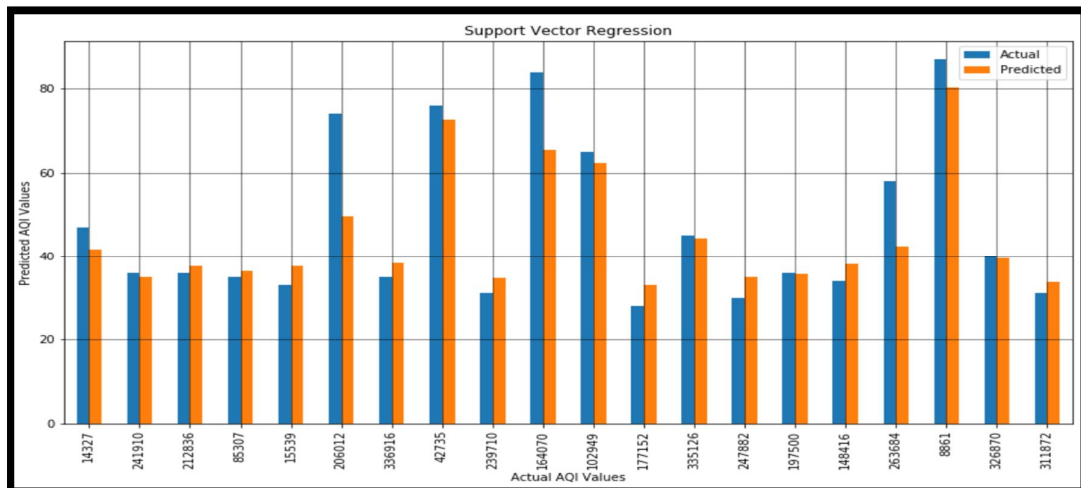


Fig 5.15: Bar Graph of Actual v/s Predicted AQI Values for SVR with (polynomial) Technique

Fig 5.14 describes the scatter plot of actual and predicted AQI values for Support Vector Regression when the kernel is polynomial along with the regression line and Fig 5.15 describe the bar plot for actual and predicted

AQI values. From both the figures we can see that there is a slight difference between the actual and predicted AQI values.

In this research work, air quality index values are calculated for the U.S. Pollution dataset and then regression models are built to predict air quality index based on traditional machine learning techniques.

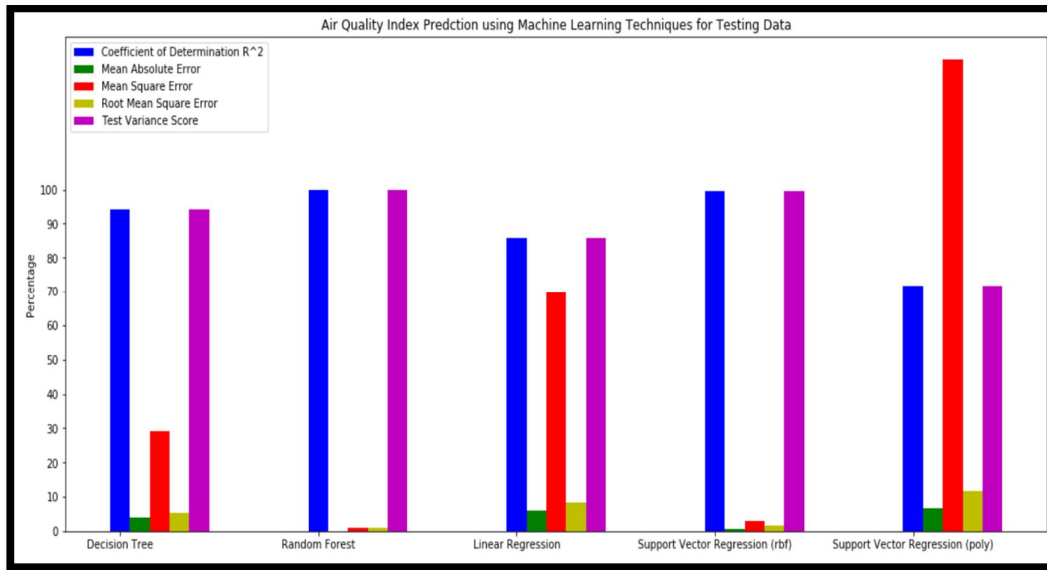


Fig 5.16: Results of Prediction of Air Quality Index using Machine Learning Techniques

Fig 5.16 shows the results of the prediction of air quality index for Decision Tree Regression, Random Forest Regression, Linear Regression, Support Vector Regression when the kernel is rbf and Support Vector Regression when the kernel is polynomial. The results have been calculated using regression metrics like Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). Thus, when the results are compared and analysed, Random Forest Regression gives the best accuracy ($R^2 = 0.9981$, $MAE = 0.0386$, $MSE = 0.9201$, $RMSE = 0.9592$).

E. Results Analysis

Accuracy of Different Machine Learning Algorithms																
	Decision Tree Regression				Random Forest Regression				Linear Regression				Support Vector Regression			
	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE
Training Accuracy	0.9473	3.8971	27.434	5.2378	0.9998	0.0112	0.0602	0.2453	0.8684	6.0460	68.519	8.2776	0.9948	0.7012	2.6567	1.6299
Testing Accuracy	0.9404	3.9298	28.951	5.3807	0.9981	0.0386	0.9201	0.9592	0.8562	6.0805	69.832	8.3566	0.9940	0.6958	2.9043	1.7042

Fig 5.17: Accuracy of AQI Prediction for different Machine Learning Techniques

Fig 5.17 shows both the training and testing accuracy results of different machine learning Techniques. Finally, when the accuracy results are got by calculating Coefficient of Determination or R-Squared (R²), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Test Variance score for Decision Tree Regression, Random Forest Regression, Linear Regression and Support Vector Regression are compared, we find that Random Forest Regression gives the best results for Air Quality Index prediction.

IV. CONCLUSION

As a serious concern, air pollution in the 21st century has received a great attention in recent years, and also various air pollution forecasting methodologies and approaches have been advanced in these few decades. This research work reviewed on the methods of air pollution forecasting. At the beginning of this work, the current research status of air pollution from the perspective of pollution emission inventories, health effects, and air pollution assessment to air pollution control efficiency and air pollution early warning systems has been reviewed with the literature study. The methodology has been proposed for the prediction of Air Quality Index with accurate air quality forecasting has the important of theoretical and practical value for the public. Without it, neither the government nor the public can effectively avoid the health damage caused by air pollution or improve the emergency response capability of heavy pollution days. From the experimental result, it shows that all the four models, Decision Tree Regression based model, Linear Regression based model and Random Forest Regression based model and Support Vector Regression based model can achieve with a satisfactory results. Random Forest Regression based model out performs better in the air quality index prediction with an accuracy of 98%.

REFERENCES

- [1] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif ul Islam, Muhammad Nabeel Asghar, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities", IEEE Access, Vol. 7, June 26, 2019, pp. 128325-128338.
- [2] Ranjana Waman Gore, Deepa S. Deshpande, "An Approach for Classification of Health Risks Based on Air Quality Levels", IEEE, 2017, pp. 58-61.
- [3] Chavi Srivastava, Shyamli Singh and Amit Prakash Singh, "Estimation of Air Pollution in Delhi Using Machine Learning Techniques", 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India, Sep 28-29, 2018, pp. 304-309.
- [4] Huixiang Liu, Qing Li, Dongbing Yu and Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences 2019, 29 September 2019, pp. 1-9.
- [5] Lu Bai, Jianzhou Wang, Xuejiao Ma, Haiyan Lu, "Air Pollution Forecasts: An Overview", International Journal of Environmental Research and Public Health", April 2018, 1-44.
- [6] MS. Varsha Hable-Khandekar, Dr. Pravin Srikanth, "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring", Third International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017.
- [7] Rekha GS, Indrajeet Kumar, DJ Vineeth, Kushal Pandita, Abhishek M Patil, "Deep Air Learning Prediction and Feature Analysis of Air Quality", International Journal of Latest Trends in Engineering and Technology (IJLTET), Volume 13, Issue 2, pp. 067-069.
- [8] Kostandina Veljanovska, Angel Dimoski, "Air Quality Index Prediction using Simple Machine Learning Algorithms", International Journal of Emerging Trends and Technology in Computer Science (IJETTCS), January-February 2018, Volume 7, Issue 1, pp. 25-30.
- [9] Bharathi P. T and P. Subashini, "Optimal Feature Subset Selection Using Differential Evolution and Extreme Learning Machine", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 3 Issue 7, pp. 1898 – 1905, July 2014
- [10] Alexei Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology", Interdisciplinary Journal of Information, Knowledge, and Management, pp. 45-79, 2018