International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206427

Comparative Analysis of Machine Learning Regression Algorithms on Air Pollution Dataset

Sumit Upadhyay

Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

ABSTRACT

Article Info

Volume 6, Issue 4

Page Number: 125-136

Publication Issue :

July-August-2020

Air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs. Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. Air pollutants, such as carbon monoxide (CO), sulfur dioxide (SO(2)), nitrogen oxides (NOx), volatile organic compounds (VOCs), ozone (O(3)), heavy metals, and respirable particulate matter (PM2.5 and PM10), differ in their chemical composition, reaction properties, emission, time of disintegration and ability to diffuse in long or short distances. The main objective of this paper to build a model for predicting Air Quality Index(AQI) of the specific cities using various types of machine learning algorithms namely Multiple Linear Regression, K Nearest Neighbours(KNN), Support Vector Machine(SVM) and Decision Tree. And also evaluate and compare the performance of every algorithm based on their accuracy score and errors. Air Pollution dataset is publicly available on different government sites. The implementation phase dataset is divided as 80% for the training of different models and the rest of the dataset is used for testing the model.

Article History

Accepted : 10 July 2020

Published : 20 July 2020

Keywords : Machine Learning, Regression, Prediction, Air Quality Index, So2, PM2.5, Accuracy score, Support Vector Regression.

I. INTRODUCTION

Air pollution refers to the release of pollutants into the air that are detrimental to human health and the planet as a whole.Major outdoor air pollutants in cities include ozone (O3), particle matter (PM), sulfur dioxide (SO2), carbon monoxide (CO), nitrogen oxides (NOx), volatile organic compounds (VOCs), pesticides, and metals, among others . Increased mortality and morbidity rates have been found in association with increased air pollutants (such as O3, PM and SO2) concentrations.Particle size is critical in determining the particle deposition location in the human respiratory system . PM2.5, referring to particles with a diameter less than or equal to 2.5 μm, has been an increasing concern, as these particles can be deposited into the lung gas-exchange region, the alveoli [1] . As per a study supported 2016 data, a minimum of 140 million people in India breathe air that's 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the very best annual levels of pollution are in India.The 51% of pollution is caused by the economic pollution, 27% by vehicles, 17% by crop burning and 5% by diwali fireworks. Air pollution contributes to the premature deaths of two million Indians per annum . Emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm.Along with harming human health, air pollution can cause a variety of environmental effects such as acid rain,eutrophication,haze,effects on wildlife,ozone depletion,crop and forest damage,global climate change.Global warming is an environmental phenomenon caused by natural and anthropogenic air pollution. It refers to rising air and ocean temperatures round the world. This temperature rise is a minimum of partially caused by a rise within the amount of greenhouse gases within the atmosphere. Greenhouse gases trap heat within the Earths atmosphere (Usually, more of Earths heat escapes into space).

Carbon dioxide may be a greenhouse emission that has had the most important effect on heating . Carbon dioxide is emitted into the atmosphere by burning fossil fuels (coal, gasoline, and natural gas). Humans have come to believe fossil fuels to power cars and planes, heat homes, and run factories. Doing this stuff pollutes the air with CO2.

Other greenhouse gases emitted by natural and artificial sources also include methane, laughing gas , and fluorinated gases. Methane may be a major emission from coal plants and agricultural processes. Nitrous oxide may be a common emission from industrial factories, agriculture, and therefore the burning of fossil fuels in cars. Fluorinated gases, like hydrofluorocarbons, are emitted by industry. Fluorinated gases are often used rather than gases like chlorofluorocarbons (CFCs). CFCs are outlawed in many places because they deplete the ozonosphere .

In 2006 the planet Health Organization issued new Air Quality Guidelines. The WHOs guidelines are tougher than most individual countries existing guidelines. The WHO guidelines aim to scale back air pollution-related deaths by 15 percent a year.

“The less gasoline we burn, the higher we’re doing to scale back pollution and harmful effects of global climate change ,” Walke says.“Make good choices about transportation. When you can, walk, ride a motorcycle , or take public transportation. For driving, choose cars that recover miles per gallon of gas or choose an electrical car.”You also can investigate your power provider options—you could also be ready to request that your electricity be supplied by wind or solar. Buying your food locally cuts down on the fossil fuels burned in trucking or flying food in from across the country. And perhaps most vital , “Support leaders who push for clean air and water and responsible steps on global climate change ,”Walke says.[2]

Currently, three major approaches are wont to forecast PM2.5 concentrations: statistical models, chemical transport, and machine learning. Statistical models, which are mainly supported single variable linear regression , have shown a negative correlation between different meteorological parameters (wind, precipitation, and temperature) and PM concentrations (PM10, PM2.5, and PM1.0).

Chemical transport and Atmospheric Dispersion Modeling are numerical methods, and therefore the most advanced ones are WRF-Chem and CMAQ. These models are often wont to predict atmospheric pollution, but their accuracy relies on an updated source list that's very difficult to supply . In addition, complex geophysical characteristics of locations with complex terrain complicate the implementation of these models of weather and pollution forecast mostly because of the complexity of the air flows (wind speed and direction) around the topographic features . Unlike a pure statistical procedure , a machine learning approach can consider several parameters during a single model. The most popular classifiers to forecast pollution from meteorological data are artificial Neural Networks . Other successful studies use hybrid or mixed models that combine several artificial intelligence algorithms, such as fuzzy logic and Neural Network , or Principal Component Analysis and Support Vector Machine , or numerical methods and machine learning .[3]

Recent studies show that the machine learning approach seems to beat the opposite two methods for forecasting pollution . This is the rationale why it's increasingly wont to predict air quality . However, the data mining does not only differ from one study to another, in terms of classification algorithms, but also regarding the used features. Some of them consider a quite exhaustive list of meteorological factors , whereas others proceed with a careful selection,or don't even use climatic parameters in the least.

The proposed system is capable of predicting air quality index by analyzing different types of air pollutants present in air using different machine learning algorithms.

II. RELATED WORK

In this research paper the students have forecasted the air quality of India by using machine learning algorithms to predict the air quality index(AQI) of a given area. Air quality Index is a standard measure to determine the quality of air. Concentrartion of Gases such asso2, no2,co2, rspm, spm. etc. are recorded by the agencies . These students have developed a model to predict the air quality index based on historical data of previous years and predicting over a particular upcoming year as a Gradient decent boosted multivariable regression problem. They improved the efficiency of the model by applying cost Estimation for predictive Problem. They say that this model is capable of successfully predicting the air quality index of a total county or any state or any bounded region provided with the historical data of pollutant concentration.

This paper presents an integrated model using Artificial Neural Networks and Kriging to predict the level of air pollutants at various locations in Mumbai and Navi Mumbai using past data available from meteorological department and Pollution Control Board. The proposed model is implemented and tested using MATLAB for ANN an R for Kriging and the results are presented.

This system has used the Linear regression and Multilayer Perceptron (ANN)Protocol for prediction of the pollution of next day. The system helps to predict next date pollution details based on basic parameters and analyzing pollution details and forecast future pollution. Time Series Analysis was also used for recognition of future data points and air pollution prediction.

This proposed system does two important tasks (i). Detects the levels of PM2.5 based on given atmospheric values. (ii) Predicts the level of PM2.5 for a particular date. Logistic regression is used to detect whether a data sample is either polluted or not polluted. Autoregression is employed to predict future values of PM2.5 based on the previous PM2.5

readings. The primary goal is to predict air pollution level in City with the ground data set.[4]

The major objective of our paper is to comapre various types of regression algoritms namely Multiple Linnear regression,K Nearest Neighbours,Decision Tree,Support Vector Machine and evaluate performance based on their accuracy and errors while predicting AQI.

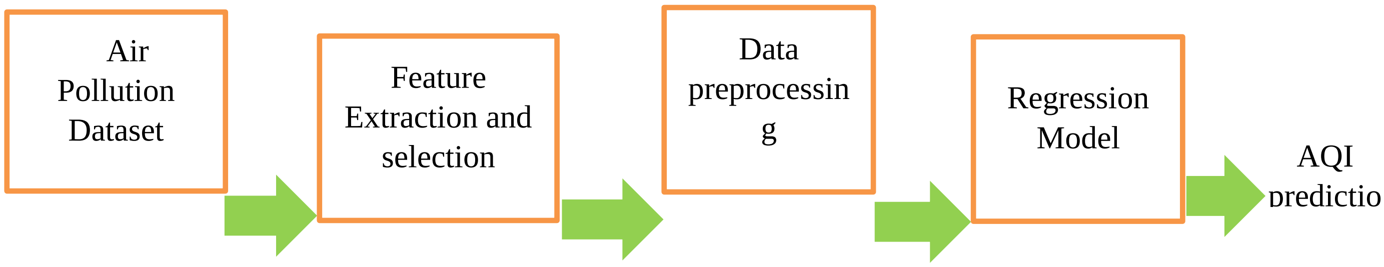
III. LITERATURE REVIEW

Sr No.	Paper Title	Author	Objective	Methodology	Conclusion
1	Regression Model to predict Air Pollution from Affordable Data Collection.[5]	Yves Rybarczyk And Rasa Zalakeviciute	The aim of this study was to design accurate regression model and predict the Pm 2.5.	The rectllinear model and modifying it based on its previous performance.	The model was able to produce ultimately following results r = 0.49 And RMSE = 7.65
2	Detection and Prediction of Air Pollution using Machine Learning Models[6]	Aditya C, Chandana R Deshmukh, Nay ana Dand Praveen Gandhi Vidyavas tu	The aim was to find best result by using different classification machine learning algorithms.	Different classification algorithms such as Logistic regression etc.	Had achieved best accuracy using Logistic regression which suits best for the data with mean accuracy and standard deviation accuracy to be 0.998859.
3	A thorough Survey on prediction of Airpollution.[7]	Mushtak Sayyed, Akshay Sarode, Adesh Salunke and Swaraj Desai	This paper mainly deals with the fact of different methodologies that deals with the Air pollution prediction technique.	They have survey on air pollution by studying different aspects from different resources.	This paper after studying all these facts putforwadrs an idea of using Fuzzy C means clustering on Air pollution index data along with the

					Fuzzy Artificial neural network to improve the prediction rate,
4	INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING[8]	Mrs. A. Gnana Sound ari,Mrs. J. Gnana and Ak shaya A.C	The study aims to find best regression model able to predict accurate result on air pollution	They have use Linear regression and Gradient Boosting ML algorithms.	This paper finds best results by using boosting algorithm.

IV. Machine Learning Algorithms

There are so many algorithms to predict data such as linear regression,KNN,SVM,Decision tree etc.We have tried these algorithms to achieve maximum accuracy.All algorithms have different working procedure.Figure 1 shows the air pollution regression model architecture.



A. Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called **multiple linear regression**. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.The linear regression equation is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

A. K Nearest Neighbours(KNN)

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

B. Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**.The core algorithm for building decision trees called **ID3** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with *Standard Deviation Reduction*.

C. Support Vector Machine

In machine learning, **support-vector machines** (**SVMs**, also **support-vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems.In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

V. DATASET

5.1 Description

The dataset consist of several hundred weather and various pollutant records,each of which contains the value of a set of characterstics. The dataset having following attributes:

1. Temperature
2. Humidity
3. Wind speed
4. Visibility
5. Pressure
6. So2
7. No2
8. Rainfall
9. PM 10
10. PM 2.5
11. AQI

Temperature is the factor that effects air and its pollutants so it represents the temperature of the enviornment at the time of prediction.Humidity the represents the moisture in the air so that AQI can be predicted more precisely.Wind speed represents the direction of air at the time of prediction.Visibility is used to descibe how much one is able to see that can effect AQI and able to predict result more precisely.Rainnfall can shuttle down the pollutant of air which can cause air pollution.And So2,No2,PM10,PM2.5 are the pollutant that cause air pollution, these are the element in the air which cabn produce result more accurately.

5.2 Splitting for testing

Data splitting was done as 70% of training and 30% for testing.

5.3 Preprocessing and Feature Selection

In this step we clean the data and after cleaning the data the data is grouped together. Then we have to change its format if needed. After that we check for missing the values and remove or replace according to our need and then we check for correlated features and remove highly correlated features.

After preprocessing the data we select the feature which will use for training and prediction of the data i.e splitting of dependent and independent features.

VI. Observation

In this study we will observe the results which we were obtained during the experiment and what errors and accuracies we had drawn from that,the process observing the results during the practical time is called observation.

6.1 Analysis

Figure 2 shows how does so2 pollutant effects the air quality as well as air quality index. The following graph shows how effected by emmision of so2.

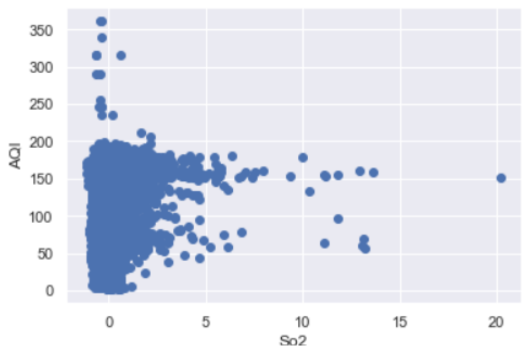


Figure 2. so2 vs. AQI

Figure 3 shows how No2 effects the Air Quality Index. Following scatter plot shows emmission of NO2 against AQI.

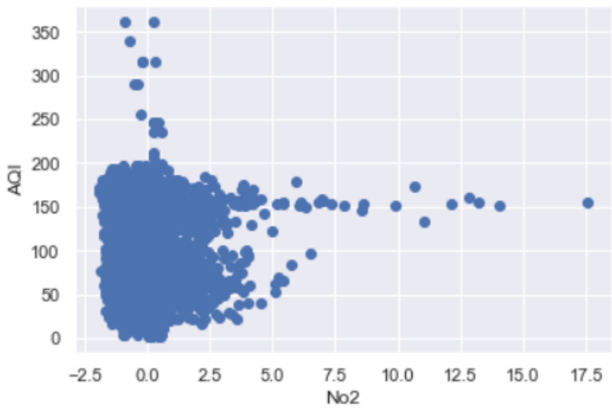


Figure 3. no2 vs. AQI

Figure 3 and Figure 4 shows particulate matter (PM) 10 and PM 25 emission against AQI. These particles are very highly cause the air pollution as this graph also shows.

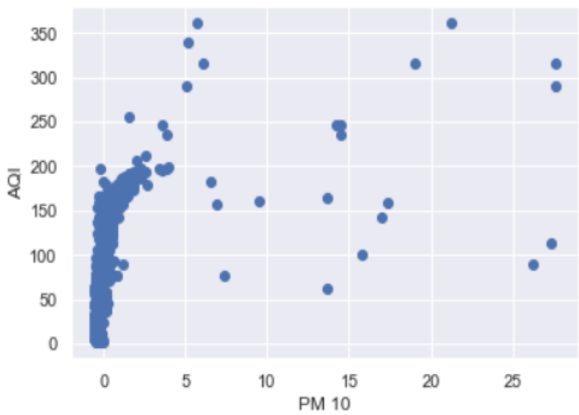


Figure 3. PM 10 vs. AQI

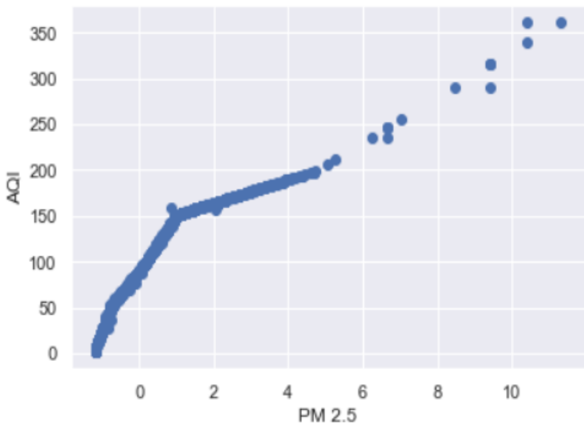


Figure 4. PM 25 vs. AQI

6.2 Accuracy, prediction and errors

a. Linear Regression

Table 1: AQI Prediction with Linear Regression

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	PM25	Predicted_AQI	Original_AQI
6213	0.017727	1.309815	0.554879	1.558790	0.133046	-0.123471	0.039591	-0.991779	-0.234271	-0.178459	78.968189	80
7168	-0.932841	0.820826	1.734667	0.086796	0.207987	-0.485819	-0.049133	0.694777	0.596350	0.843286	114.386474	140
2699	-0.467960	1.092487	1.935327	0.106796	0.108243	-0.281014	0.010016	1.266208	-0.319028	-0.785874	49.780969	45
3533	-1.285366	0.820826	-1.466232	-1.693196	0.166140	-0.627608	-0.574081	0.655942	-0.268174	-0.524404	59.850404	62
1560	0.417216	0.331837	0.571803	-1.905195	0.233653	0.475193	0.283581	0.633750	-0.081708	-0.190527	76.472905	79
...	...	...	...	...	...	...	...	...	...	...	...	...
3655	-0.853181	0.875158	-1.543595	0.062797	0.168859	-0.627608	-1.416956	-1.219242	-0.183417	-0.258911	73.408964	76
1563	1.142478	-0.754805	0.068942	-2.293193	0.237433	0.475193	0.209644	0.556080	0.545496	0.795015	117.179437	137
3374	-1.171227	0.820826	-0.327544	0.062797	0.236505	-0.879677	0.024803	0.167728	-0.352931	-0.604857	57.333601	58
3746	-1.957125	-0.048487	0.213998	1.342791	0.147968	-0.312522	0.357517	0.278686	-0.132562	-0.198572	73.174387	79
397	0.737639	-1.569786	-0.859415	0.106796	0.157850	4.004155	-0.714561	-1.058353	0.257321	0.690427	119.465808	131

Accuracy Score on training data: 0.91
Accuracy Score on testing data: 0.900
Mean squared error: 181.715
Mean absolute error: 8.098
r2 score: 0.9000

b. K Nearest Neighbours

Table 2: AQI Prediction with KNN

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	PM25	Predicted_AQI	Original_AQI
6213	0.017727	1.309815	0.554879	1.558790	0.133046	-0.123471	0.039591	-0.991779	-0.234271	-0.178459	86.333333	80
7168	-0.932841	0.820826	1.734667	0.086796	0.207987	-0.485819	-0.049133	0.694777	0.596350	0.843286	138.666667	140
2699	-0.467960	1.092487	1.935327	0.106796	0.108243	-0.281014	0.010016	1.266208	-0.319028	-0.785874	48.666667	45
3533	-1.285366	0.820826	-1.466232	-1.693196	0.166140	-0.627608	-0.574081	0.655942	-0.268174	-0.524404	65.666667	62
1560	0.417216	0.331837	0.571803	-1.905195	0.233653	0.475193	0.283581	0.633750	-0.081708	-0.190527	77.000000	79
...	...	...	...	...	...	...	...	...	...	...	...	...
3655	-0.853181	0.875158	-1.543595	0.062797	0.168859	-0.627608	-1.416956	-1.219242	-0.183417	-0.258911	75.666667	76
1563	1.142478	-0.754805	0.068942	-2.293193	0.237433	0.475193	0.209644	0.556080	0.545496	0.795015	131.333333	137
3374	-1.171227	0.820826	-0.327544	0.062797	0.236505	-0.879677	0.024803	0.167728	-0.352931	-0.604857	62.333333	58
3746	-1.957125	-0.048487	0.213998	1.342791	0.147968	-0.312522	0.357517	0.278686	-0.132562	-0.198572	69.666667	79
397	0.737639	-1.569786	-0.859415	0.106796	0.157850	4.004155	-0.714561	-1.058353	0.257321	0.690427	153.000000	131

Accuracy Score on training data: 0.97
Accuracy Score on testing data: 0.93
Mean squared error: 110.810
Mean absolute error: 7.399
r2 score: 0.93

c. Decision Tree

Table 3: AQI Prediction with Decision Tree

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	PM25	Predicted_AQI	Original_AQI
6213	0.017727	1.309815	0.554879	1.558790	0.133046	-0.123471	0.039591	-0.991779	-0.234271	-0.178459	80.0	80
7168	-0.932841	0.820826	1.734667	0.086796	0.207987	-0.485819	-0.049133	0.694777	0.596350	0.843286	140.0	140
2699	-0.467960	1.092487	1.935327	0.106796	0.108243	-0.281014	0.010016	1.266208	-0.319028	-0.785874	45.0	45
3533	-1.285366	0.820826	-1.466232	-1.693196	0.166140	-0.627608	-0.574081	0.655942	-0.268174	-0.524404	62.0	62
1560	0.417216	0.331837	0.571803	-1.905195	0.233653	0.475193	0.283581	0.633750	-0.081708	-0.190527	79.0	79
...	...	...	...	...	...	...	...	...	...	...	...	...
3655	-0.853181	0.875158	-1.543595	0.062797	0.168859	-0.627608	-1.416956	-1.219242	-0.183417	-0.258911	76.0	76
1563	1.142478	-0.754805	0.068942	-2.293193	0.237433	0.475193	0.209644	0.556080	0.545496	0.795015	138.0	137
3374	-1.171227	0.820826	-0.327544	0.062797	0.236505	-0.879677	0.024803	0.167728	-0.352931	-0.604857	58.0	58
3746	-1.957125	-0.048487	0.213998	1.342791	0.147968	-0.312522	0.357517	0.278686	-0.132562	-0.198572	79.0	79
397	0.737639	-1.569786	-0.859415	0.106796	0.157850	4.004155	-0.714561	-1.058353	0.257321	0.690427	131.0	131

Accuracy Score on training data: 1.0
Accuracy Score on testing data: 0.998
Mean squared error: 2.037
Mean absolute error: 0.196
r2 score: 0.998

d. SVR(Support Vector Regression)

Table 4: AQI Prediction with SVR

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	PM25	Predicted_AQI	Original_AQI
6213	0.017727	1.309815	0.554879	1.558790	0.133046	-0.123471	0.039591	-0.991779	-0.234271	-0.178459	79.067250	80
7168	-0.932841	0.820826	1.734667	0.086796	0.207987	-0.485819	-0.049133	0.694777	0.596350	0.843286	134.260507	140
2699	-0.467960	1.092487	1.935327	0.106796	0.108243	-0.281014	0.010016	1.266208	-0.319028	-0.785874	51.299227	45
3533	-1.285366	0.820826	-1.466232	-1.693196	0.166140	-0.627608	-0.574081	0.655942	-0.268174	-0.524404	63.479616	62
1560	0.417216	0.331837	0.571803	-1.905195	0.233653	0.475193	0.283581	0.633750	-0.081708	-0.190527	79.333438	79
...	...	...	...	...	...	...	...	...	...	...	...	...
3655	-0.853181	0.875158	-1.543595	0.062797	0.168859	-0.627608	-1.416956	-1.219242	-0.183417	-0.258911	76.205416	76
1563	1.142478	-0.754805	0.068942	-2.293193	0.237433	0.475193	0.209644	0.556080	0.545496	0.795015	126.800206	137
3374	-1.171227	0.820826	-0.327544	0.062797	0.236505	-0.879677	0.024803	0.167728	-0.352931	-0.604857	55.138289	58
3746	-1.957125	-0.048487	0.213998	1.342791	0.147968	-0.312522	0.357517	0.278686	-0.132562	-0.198572	78.198121	79
397	0.737639	-1.569786	-0.859415	0.106796	0.157850	4.004155	-0.714561	-1.058353	0.257321	0.690427	117.374992	131

Accuracy Score on training data: 0.889
Accuracy Score on testing data: 0.870
Mean squared error: 235.402
Mean absolute error: 6.585
r2 score: 0.870

6.3 Accuracy and Error Comparison

Table 5: Algorithms Comparison

Algorithms	Accuracy	Mean squared error	Mean absolute e rror	r2 score
1.Linear Regression	0.90	181.715	8.098	0.900
2.K nearest Neighbour	0.93	110.810	7.399	0.93
3.Decision Tree	0.99	2.037	0.196	0.998
4.Support Vector Regression	0.87	235.402	6.585	0.870

6.4 Visualization

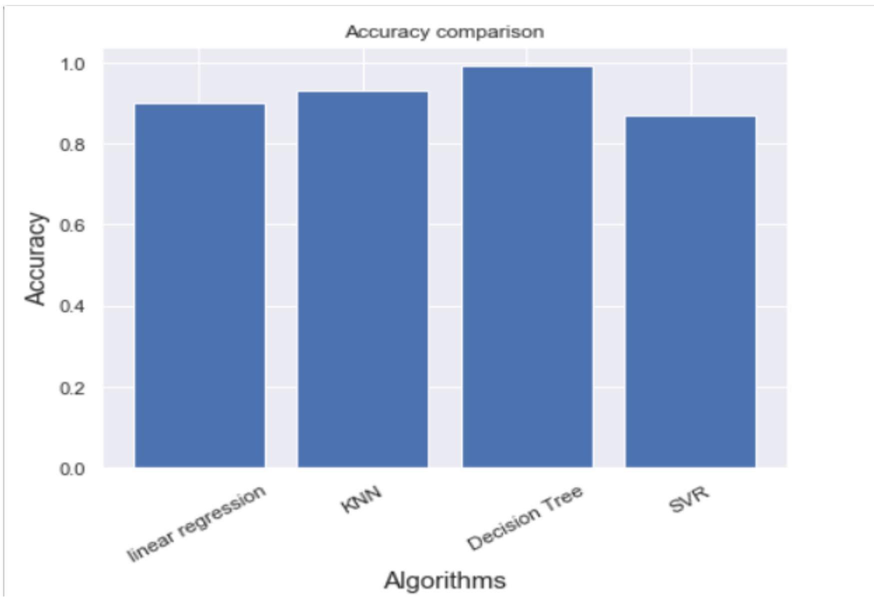


Figure 5. Accuracy Comparison of various algorithms

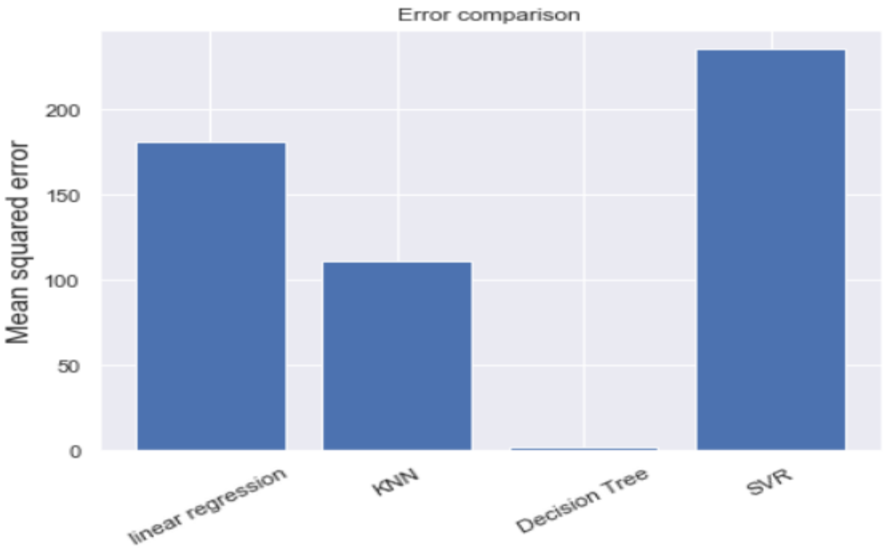


Figure 6. Mean Squared Error Comparison

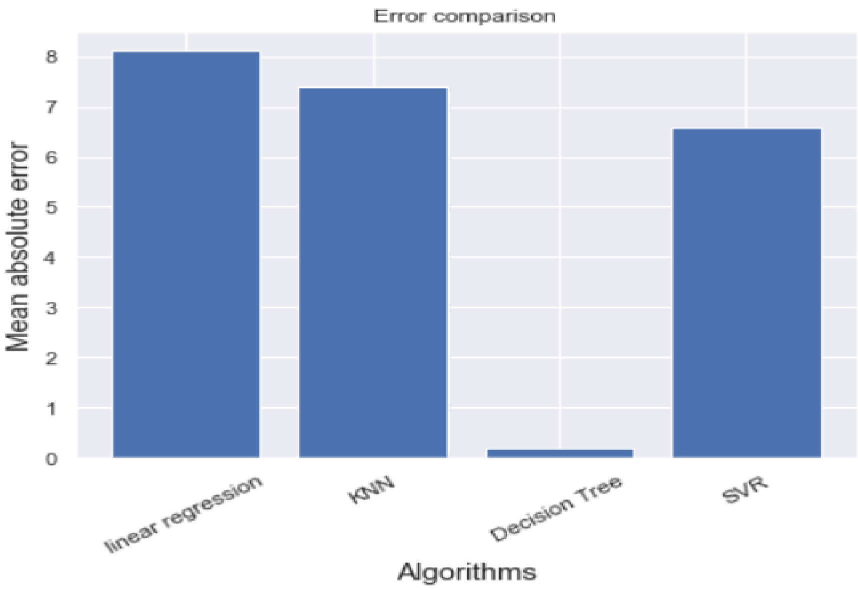


Figure 7. Mean absolute error comparison

VII. RESULTS AND DISCUSSION

After successful training of the dataset having so2, no2,PM 10, PM25 like pollutants data, more than satisfactory accuracy is achieved by Decision tree algorithm for the AQI prediction . After the successful prediction, predicted value for each algorithms has been compared indivisually in Table 1,2,3,4 and it compares the predicted value with original value.Before the prediction value the graph has been shown in analysis section which conclude effect of different pollutant on the air quality index.The accuracy has been shown on training and testing both to conclude that there is no over fitting.Table 5 shows comparison of different algorithms on the basis of their accuracy and errors. The graphical representation of accuracy and error has been shown for accuracy and errors for better understanding of results as we can see in fig 5,6 7.After all the examination we can see from the results that the decision tree is come with most efficient result as its accuracy and errors are better to any of algorithms that we have applied on this dataset.As a result we can say after comparing all the algoritms using machine learning decision tree was able to produce the best result.

VIII. CONCLUSION

In this paper, the data (PM10,PM25,so2,no2 etc.) is analysed and preprocessed before prediction. Linear Regression,KNN,Decision tree and svr ML algorithms are used to predict and forecast the air quality prediction of the specific city. The matplot library is used to plot the predicted and the forecasted data. So, at last the following conclusion can be drawn – A model for forecasting AQI for specific city is trained and tested with different input variables like PM10, PM25, nitrous oxide, sulphur dioxide by mentioned algorithms. Some graphs are plotted as a graphical interface for the predicted and forecasted data for all the inputs with the help of matplot library.

IX. Limitation and Future Work

There is some limitation in this study or this project. This model is predicted only on the static data of air pollution of any city. It does not give result on the live data. The data set is used for this model,the data which has been produced past few monthes.The prediction will be more useful if the live data will be given to the model. In future the prediction would be better by using the live dataset. The data set can be generate live using web scraping then the result will be more valuble . The data can be predicted in the current time as well as it could be predicted for any reigon or country, which will make this analysis more useful and reasonable for all people to follow.

X. REFERENCES

[1]. Dixian Zhu 1, Changjie Cai 2 , Tianbao Yang 1 and Xun Zhou 3 “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization”.

[2]. https://www.nrdc.org/stories/air-pollution-everything-you-need-know “Air Pollution : Everything You Need to Know”.

[3]. an Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez and Yves Rybarczyk 18 Jun 2017 “Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters”.

[4]. Pooja Bhalgat , Sejal Pitale and Sachin Bhoite “Air Quality Prediction using Machine Learning Algorithms”.

[5]. Yves Rybarczyk and Rasa Zalakeviciute December 20th 2017 “Regression Model to predict Air Pollution from Affordable Data Collection”.

[6]. Aditya C, Chandana R Deshmukh, Nayana D, Praveen Gandhi Vidyavastu 4 may 2018 “Detection and Prediction of Air Pollution using Machine Learning Models”.

[7]. Mushtak Sayyed, Akshay Sarode, Adesh Salunke and Swaraj Desai march 2020 “A thorough Survey on prediction of Airpollution”.

[8]. Mrs. A. Gnana Soundari, Mrs. J. Gnana and Akshaya A.C 2019 “INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING”.

About Author

Sumit Upadhyay is B.Tech 3rd year student in Department of Computer Science and Engineering in IMS Engineering College, Ghaziabad, UP, India. His Area of interest is programming in Python, Machine Learning and Deep Learning.

Cite this article as :

Sumit Upadhyay, "Comparative Analysis of Machine Learning Regression Algorithms on Air Pollution Dataset", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 4, pp.125-136, July-August-2020. Available at doi : https://doi.org/10.32628/CSEIT206427

Journal URL : http://ijsrcseit.com/CSEIT206427