International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206435

Air Quality Prediction Model using Supervised Machine Learning Algorithms

Suprateek Halsana

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

Article Info

Volume 6, Issue 4

Page Number: 190-201

Publication Issue :

July-August-2020

Article History

Accepted : 20 July 2020

Published : 25 July 2020

ABSTRACT

Air pollution is the "world's largest environmental health threat"[1], causing 7 million deaths[1] worldwide every year. Its major constituents are $PM_{2.5}$, PM_{10} and the harmful green house gases $S0_2$, $N0_2$, C0 and other effluents from vehicles and factories affecting not only humans but also other living organisms both on land and sea. The only effective solution to this global issue is to implement machine learning algorithms to predict the AQI (Air Quality Index) that can make the people aware of the condition of the air of a certain region such that certain actions could be issued by the government for the improvement of the air quality in the future. The prime objective behind this project is to predict the AQI based on the concentration of $PM_{2.5}$, PM_{10},$S0_2$, $N0_2$, C0 as well as weather conditions like temperature, pressure and humidity[2].Hence the data set is combined from various web sources like cpcb.nic.in and uci repository in order to bring accuracy in the prediction and to justify whether the Quality of air is suitable or not. This prediction will be brought about with the help of some supervised machine learning algorithms and the observation and the result will state which algorithm is giving better accuracy in prediction of AQI and which one is giving less error.

Keywords : AQI, PM2.5, PM10, Machine Learning Algorithms, Dataset, Preprocessing, Regression, Training, Testing, Standardization, Normalization, Outlier, Correlation.

I. INTRODUCTION

In this sophisticated era with the rapid growth of population and their demand, world has advanced in technology as well as industrialization.However, the dark side of this advancement is overlooked i.e. the unregulated emission of the harmful gases from vehicles, burning of fossil fuels as well as effluents from industries. This has lead to the global issue causing degradation of air quality called air pollution. Air Pollution refers to the release of pollutants into the air that are detrimental to human as well as other

organisms. “According to the research by WHO[1] (World Health Organization) approximately 7 million people die”[1] worldwide due to this global crisis of air pollution. Even though a lot of technologies are still working for it but still this crisis is affecting us globally.

The major constituent of such harmful gases are PM 2.5 (Particulate matter < 2.5 microns) and PM 10. They are the most hazardous ones for our health. The PM 2.5 [2] are the particles of size less than 2.5 microns which enter deeper in our lungs and cause various problems and issues like heart attacks, strokes, asthma etc and the PM 10 are the particles less than 10 microns (>2.5 microns) that affect the upper respiratory tract and cause nasal complications.

The Air Quality Prediction model intends to work on the concentration of various pollutants such as PM2.5, PM10,S02, N02, C0 and also on the weather conditions that also affects the AQI i.e. the Air Quality Index of a region scaling them to a range and defining whether it is healthy, satisfactory, moderate or unhealthy for the region. The various machine learning algorithms are applied after preprocessing the data and scaling the data properly.

Air Quality Index is the numerical data that may range from 0 to as high as 300+ but these values even after prediction are not making any sense to a lay man. Hence after the prediction of AQI, the AQI will be converted into 5 categories using a Quality Check function. They are specifically ‘Healthy, moderate, ‘Unhealthy’,’Very Unhealthy’ and ‘Hazardous’ for the extreme case.

II. Literature Survey

S. No	PAPER TITLE	AUTHOR	OBJECTIVE	METHODOLGY	CONCLUSION
1	INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING [3]	Mrs. A. Gnana Soundari, Mrs. J. Gnana Jeslin and Akshaya A.C	The aim of the Project is to predict the AQI and produce better prediction than standard regression algorithms.	The Algorithms that are used are Linear Regression and the Gradient Boost Algorithm. The data Preprocessing is first done.	The Algorithm Gradient Boost Algorithm shows better accuracy of 96 % and the Naive Forecast approach is used.
2	Detection and Prediction of Air Pollution using Machine Learning Models [4]	Aditya C R, Chandana R Deshmukh, Nayana D K and Praveen Gandhi Vidyavastu	The aim of the project is to predict the PM2.5 on the basis of the features like Temp,Wind speed etc.	The Algorithms that are used are Logistic Regression and Auto Regressiion.	The model has been successfully in predicting the PM2.5 values upto 99.88% accuracy and error of 0.0006 by Logistic

					Regression.
3	Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters [5]	Jan Kleine Deters, Rasa Zalakeviciute, Mario Gonzalez and Yves Rybarczyk	The aim of the project is to predict the PM2.5 on the basis of the selected meterological parameters.	The Algorithms that have been used are the Decision tree and the SVM i.e. Support vector machine	The model was able to predict the PM2.5 to an accuracy of nearly 89 % by Decision Trees.
4	A thorough Survey on prediction of Air pollution [6]	Mushtak Sayyed, Akshay Sarode, Adesh Salunke, Swaraj Desai	The aim is to provide a survey on the various researches done on prediction of the Air Quality.	The various research papers including the techniques like SVM, Decision Tree and effective Algorithms.	Its conclusion is that there is no specific factors on which Quality of air depend, hence the prediction may vary in real scenarios.
5	Air Quality Prediction using Machine Learning Algorithms [7]	Pooja Bhalgat, Sejal Pitale and Sachin Bhoite	The aim is to predict the AQI based on the Features like PM2.5, PM10 and other gaseseous concentration.	The Algorithms that are used are ARIMA model, Auto Regression and also Linear Regression.	The Algorithms are able to provide the good prediction however the error is high which they are qorking to overcome in near future.

III. MACHINE LEARNING ALGORITHMS

Linear Regression

Linear regression [8] is a linear model, I.e. a model that has a linear relationship between the independent input variables (x) and the single dependent output variable (y) such that y can be evaluated or predicted from the linear combination of the independent input variables (x). The linear regression is called 'Simple Linear Regression' if there is only a single input variable or independent variable

(x). However in this research paper we are working on Multiple Linear Regression as the independent variables (x) are more and the dependent variable (y) is only one i.e. AQI in my case.The formula for multiple linear regression is :

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

Figure 1 : Formula for Multiple Linear Regression

The where y is the dependent variable to be predicted and the x1, x2, x3, x4….xp are the dependent variables or features. The b0, b1, bp are the regression coefficients.

Support Vector Regression (SVR)

Support Vector Machine (SVM) is a supervised algorithm i.e. used for applying regression as well as regression. In this research, I have tried to best utilize the 'Support Vector Regression' (SVR) model that is quite different from other regression techniques as it tries to fit the best possible line within the predefined or threshold variance or error. However, the prime idea is always to minimize the error, individualizing the hyper plane that will maximize the margin, and keeping in mind that part of the error in most of the cases is tolerated.

In SVM regression, the input $\mathbf{x}$ is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) $f(\mathbf{x}, \varpi)$ is given by

$$f(\mathbf{x}, \varpi) = \sum_{j=1}^{M} \varpi_j g_j(\mathbf{x}) + b$$

Figure 2 : Formula for SVM (Support Vector Machine)

Decision Tree Regression

The Decision Tree [9] is an algorithm which tends to fit a sine curve with addition to noisy observation. This in fact helps it to learn local linear regressions approximating the sine curve. The max_depth parameter in the decision tree model describes the maximum depth of the tree. If this parameter is set too high then the decision tree learns too fine details of training data and also the noise in depth i.e they overfit.

The main objective behind the Decision tree is to maximize the Information gain at each split as the decision tree keep splitting in depth.

$$IG(D_p, f) = I(D_p) - \left(\frac{N_{left}}{N_p} I(D_{left}) + \frac{N_{right}}{N_p} I(D_{right})\right)$$

Figure 3 : Maximizing Information Gain Formula

Here, f is the feature to perform the split, Dp, Dleft, and Dright are the datasets of the parent and child nodes, I is the impurity measure, Np is the total number of samples at the parent node, and Nleft and Nright are the number of samples in the child nodes.

Random Forest Regression

A Random Forest [10] is an ensemble technique which capable performs both regression and classification tasks with the utilization of multiple decision trees and a technique called Bootstrap and Aggregation which is referred to as bagging. The random forest provides better results as it doesn't rely on single decision tree rather it works on multiple decision trees to determine better prediction.

The random forest model is a kind of additive model which makes the predictions by combining the decisions from a sequence of the base models. More precisely and correctly we can write the class of models as:

$$g(x)=f0(x)+f1(x)+f2(x)+...$$

Where, g as the final model makes the summation of the simple base models f(i). Here, each of the base classifier is a simple decision tree. This broad technique of using the multiple models an approach to obtain a better predictive performance is called the model ensembling technique.

IV. DATASET

The dataset has been referred from cpcb.nic.in as well as from the uci repository.The dataset plays a major role in accurate prediction also as the right features make the prediction more realistic and with least variance. Hence, the dataset used in my research has the following features as discussed below :

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	AQI	PM25
0	14.033333	0.93	14.1197	15.8263	1015.13	4.8	17.4	50.7	87	168	89.1
1	15.055556	0.93	14.2646	15.8263	1015.63	3.1	7.0	52.1	122	177	105.5
2	15.916662	0.89	3.9284	14.9569	1015.94	6.2	28.5	53.8	95	174	100.2
3	16.094444	0.93	14.1036	15.8263	1016.41	6.3	14.7	53.7	79	169	89.6
4	16.094444	0.94	11.0446	15.8263	1016.51	4.7	7.5	54.5	63	162	76.3
...	...	...	...	...	...	...	...	...	...	...	...
7283	5.105556	0.82	13.1054	9.9015	1021.20	4.7	15.2	58.1	59	141	51.9
7284	4.450000	0.83	15.1984	10.9480	1020.64	4.7	15.0	56.0	51	139	51.0
7285	3.888889	0.86	11.1090	11.1251	1019.83	4.7	15.0	62.0	50	138	50.7
7286	3.438889	0.88	8.4364	11.1251	1019.74	5.0	15.2	60.4	48	136	49.8
7287	3.077778	0.90	11.3988	10.5938	1020.01	4.9	15.0	59.0	44	132	48.0

7288 rows × 11 columns

Figure 4 : Dataset Sample

The above F**igure (4)** shows the dataset has **7288 rows x 11 columns** which is divided into two parts i.e. dependent variable and independent variable.The 11 columns basically are the features of the dataset and the 7288 rows determine the values of the features which may help in training of the model to successfully predict the AQI.

The Independent variable (x) contains : **Temperature, Humidity, Wind speed (km/h), Visibility, Pressure, So2, No2, Rainfall, PM10, PM2.5.**

The Dependent variable or the the variable to be predicted is **AQI**.

Data Splitting and Testing :

The Dataset was splitted in such a way that 80% was training data and rest 20 % was testing data. It was done so that the model could be first trained and then could be tested on the testing data such that the accuracy score, precision and the error in prediction could be checked and proper results could be marked.

V. FLOW OF WORK

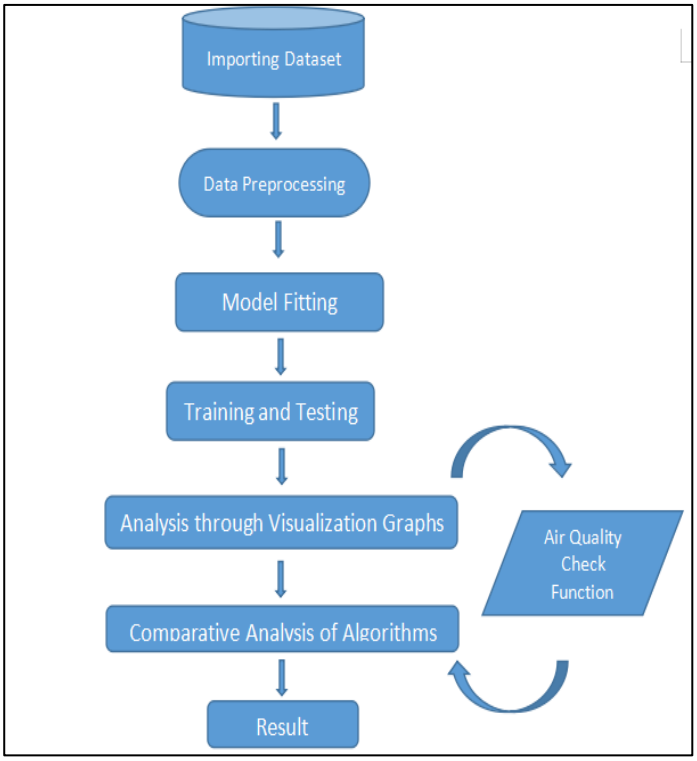

Figure 5 : Description of the Work Flow

The Above flow chart describes the working of the model where the Air Quality Check Function works to convert the AQI into Air Quality Describing Labels like '**Healthy**' , '**UnHealthy**' , '**Very UnHealthy**' and '**Moderate**' and '**Hazardous**' for extreme cases.

VI. Observation

In this Air Prediction model, the primary work is the 'Data preprocessing' and 'feature engineering' as these play a major role in shaping our dataset.The statistical data can be observed as below for good understanding of the preprocessing steps that are needed.

```
#Statisctical Data Description
data.describe()
```

	Temperature	Humidity	Wind.Speed..km.h.	Visibility	Pressure	so2	no2	Rainfall	PM10	AQI	PM25
count	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000	7288.000000
mean	12.034334	0.758924	10.279578	9.552143	995.668551	7.983727	28.164531	75.876715	32.820115	84.537870	30.336389
std	9.345925	0.184066	6.659961	4.025294	150.795632	6.347910	13.526067	18.026142	58.996060	42.370068	24.861132
min	-10.133333	0.230000	0.000000	0.000000	0.000000	0.900000	2.600000	-99.000000	1.000000	1.000000	1.000000
25%	4.050000	0.640000	5.119800	7.132300	1012.460000	4.800000	19.200000	64.900000	14.000000	57.000000	14.700000
50%	11.650000	0.810000	9.370200	9.982000	1017.200000	5.700000	29.800000	80.300000	22.000000	73.000000	22.600000
75%	19.136111	0.910000	13.781600	11.270000	1022.620000	9.800000	33.000000	87.200000	38.000000	104.000000	36.700000
max	34.811111	1.000000	45.933300	16.100000	1045.140000	136.100000	265.200000	103.700000	1660.000000	362.000000	311.900000

Figure 6 : Statistical Data Description

6.1 Preprocessing

Preprocessing [11] is a data mining technique to transform our raw data into a standardized data, which would provide better result. We have often heard about "Garbage in Garbage Out ", the concept explains that the quality of the output and the prediction we expect depends largely on the quality of the input.

The various preprocessing steps include :

Check for Missing Value or Null Value

The most crucial aspect of the preprocessing steps is the check for null values and their necessary replacement or removal as required according to the dataset. Similarly in this dataset we checked for the null values and after removal of the very few null values we get a proper numerical dataset with no null values.

```
#checking null values
data.isnull().sum()
```

```
Temperature          0
Humidity             0
Wind.Speed..km.h.    0
Visibility           0
Pressure             0
so2                  0
no2                  0
Rainfall             0
PM10                 0
AQI                  0
PM25                 0
dtype: int64
```

Figure 7 : Check for Null after their removal

Check for Most Relevant Features

Now we are ready for the rest of the preprocessing steps. This step basically aim towards finding the features that affect the target majorly. The aim is to find the correlation coefficient value between the individual feature and the target.

Check for most relevant Feature

```
data.corr()['AQI']
```

```
Temperature          0.167722
Humidity            -0.143279
Wind.Speed..km.h.   -0.031410
Visibility           0.130885
Pressure            -0.000402
so2                  0.247686
no2                  0.024833
Rainfall            -0.026996
PM10                 0.503657
AQI                  1.000000
PM25                 0.948092
Name: AQI, dtype: float64
```

So as the PM2.5 is the most relevant we would check for the presence of outliers and remove them if present

Figure 8 : Check for Most Relevant Feature

In above Figure 7, the most relevant feature is the PM2.5 and PM10 across the target i.e 'AQI '

Outlier Detection and Removal

In a dataset, an outlier [12] is a data point that differs significantly from other observations. It is really important to check for the outliers in the dataset primarily in most significant ones like $PM_{2.5}$. In the PM2.5 we find a lot of outliers causing distortion in shape as well as the distribution. The removal of outliers make the dataset normally distributed at

times which is a good attrribute for a dataset towards better prediction.

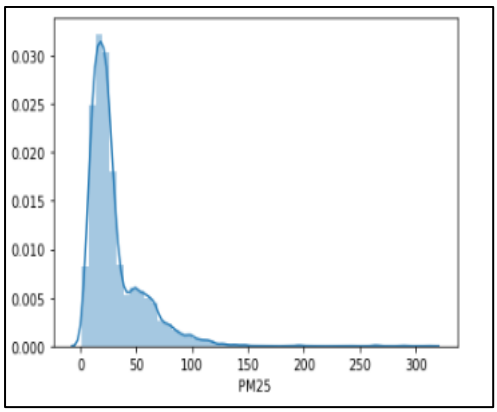

Figure 9 : PM2.5 with Outliers

After Outlier Removal

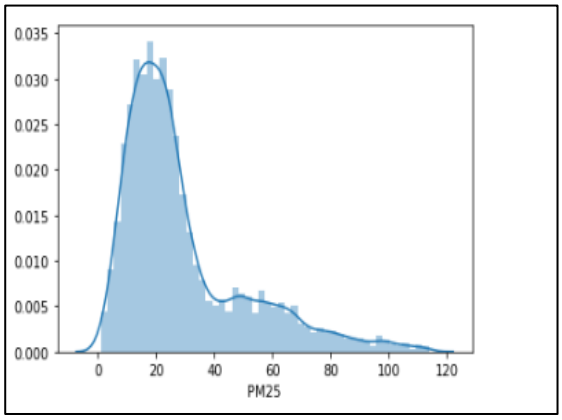

Figure 10 : PM2.5 without Outliers

The above Outlier removal is obtained by the Quantile method that tries to remove the 0.01 part of the data that act as Outlier.

Data Standardization or Data Scaling

So far we have removed the outliers now we need to work on data tranformation like data standardization or data scaling as the dataset has data varying from 0 to 10^3 in one feature and while some features work on data upto maximum 10^2. Hence to scale them in one range, there comes the need of standardization or scaling. In most cases we use standardization but at times data scaling give far better results. Hence for this dataset, data scaling is used.

Data Normalization (if needed)

If the data is not normally distributed then you can use **preprocessing. normalize (data)** but it is important to make sure whether it is needed or not. At times using the above may alter the results and may cause affect in correlation among the features.

Correlation Check

The correlation Check is also an important factor that counts in preprocessing as it check for the features which are correlated highly among themselves leading to a high redundancy and may cause poor prediction. The correlation could be checked and evaluated using the 'heatmap' easily. As the heatmap visualization makes the correlation among features clear.

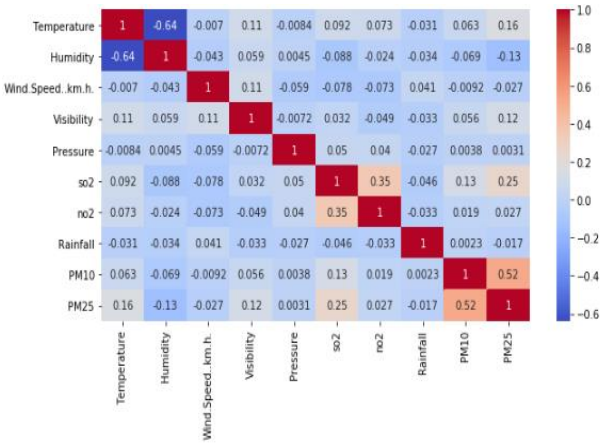

Figure 11 : Visualization of Correlation using Heatmap

The above result shows that we have none of the features correlated highly as the none of the correlation is greator than 0.70. The maximum correlation of PM_{10} could be seen with $PM_{2.5}$ i.e. 0.52 which is less.

6.2 Understanding Visualization graphs

For proper understanding of the relation between the 'PM10 concentration vs AQI ' and 'PM25 concentration vs AQI ' we have drown the scatter plot which clearly shows that we have linear graph for the PM2.5 whereas, the PM10 has a little steep distribution i.e. the gaussian distribution.

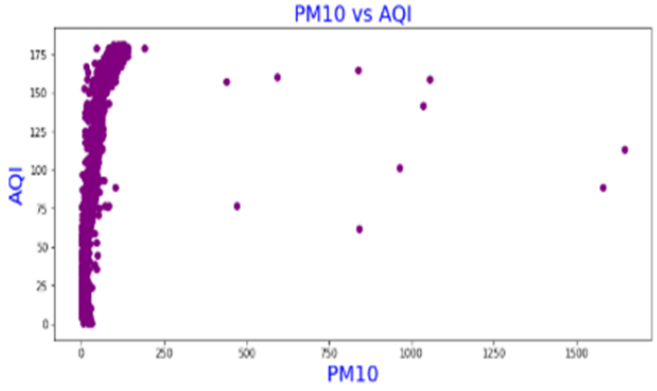

Figure 12 : Scatter plot for PM10 vs AQI

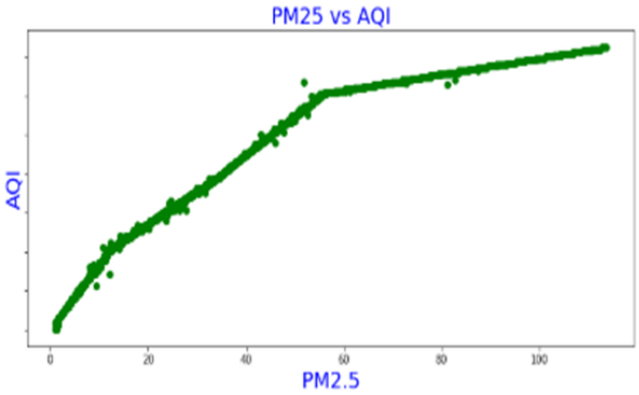

Figure 13 : Scatter plot for both PM2.5 vs AQI

The next important Observation through visualization graphs is really clear and understandable. As we have such a high accuracy in prediction that the frequency of healthy, Unhealthy, very Unhealthy, Moderate Air Quality levels are same in both 'Test Air Quality' data and 'Predicted Air Quality' data.

The final self explanatory observation is the following Dataframe that I have created showing the accuracy in prediction of AQI through different algorithmic models and their test Quality and predicted Quality helps the people to better understand the level of prediction . The below is the random sample of DataFrame

	ytest	pred of Linear R	pred of SVR	pred of DTree	pred of RF	test Quality	predictedQuality
417	145	129.427862	138.365435	145.0	144.65	Unhealthy	Unhealthy
2113	34	42.380492	41.501737	34.0	34.00	Healthy	Healthy
5789	71	67.936626	68.791423	71.0	71.00	Moderate	Moderate
5054	59	60.726339	58.065177	59.0	59.00	Healthy	Healthy
4497	67	68.484002	65.672150	67.0	67.00	Moderate	Moderate
...	...	...	...	...	...	...	...
3816	68	64.710115	73.056818	68.0	68.22	Moderate	Moderate
5620	60	58.660186	59.426245	60.0	60.00	Healthy	Healthy
7035	102	94.694277	107.741288	102.0	102.02	Unhealthy	Unhealthy
32	173	212.027750	163.599377	173.0	173.02	Very Unhealthy	Very Unhealthy
994	90	86.357053	92.803504	90.0	90.00	Moderate	Moderate

1443 rows × 7 columns

Figure 14 : Sample of Data Frame showing accuracy in prediction

VII. RESULT

Comparative Analysis of various Algorithms

After the Preprocessing and transforming the raw data into useful data, we can now split the dataset into two parts. One is the **training part** (comprising **80%** of dataset) and the **testing part** (comprising of **20%** of dataset). The splitting is done with the help of '**train_test_split**' method from the **sklearn. model_selection** module.

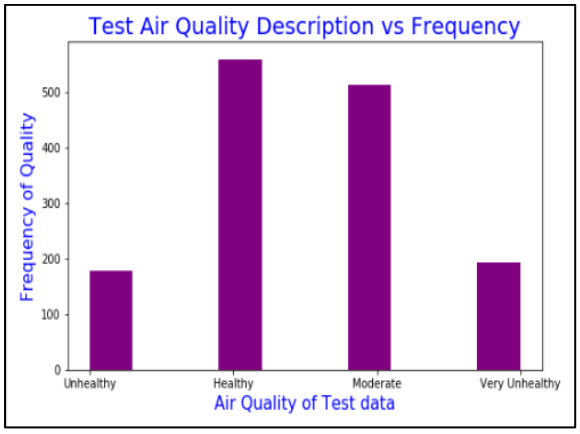

Figure 15 : Test Air Quality Vs Frequency of Quality

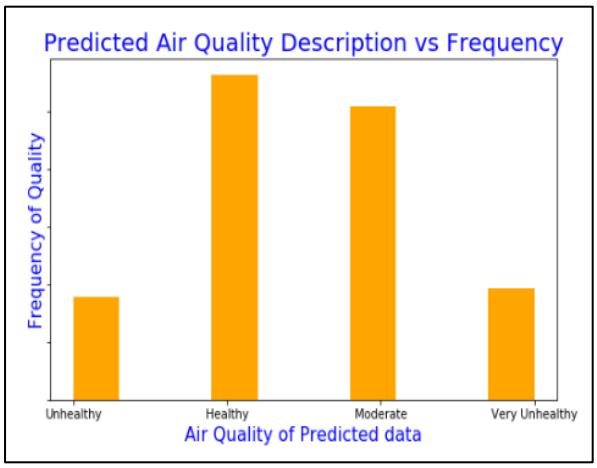

Figure 16 : Predicted Air Quality Vs Frequency of Quality

The Figure 15 & 16 shows the Similarity in the Frequency of Quality.

Prediction Score

The prediction score is found using r2_score method for the test data and model.score for training data.

The various Algorithms and their Prediction score is as shown below :

Algorithms	Prediction Score	
	Training dataset	Testing dataset
Multiple Linear Regression	0.9405	0.9379
Support Vector Regression	0.9937	0.9869
Decision Tree Regression	1.0000	0.9997
Random Forest Regression	0.99997	0.99985

Table 1 : Comparative Prediction score of both Training and Testing Dataset

The best accuracy in prediction is shown by Decision tree on the training dataset and Random Forest Regression for testing Dataset. The accuracy on Training dataset determines how well the model is trained. Whereas, the accuracy shown on testing dataset is the real prediction score to be taken into account. Hence best accuracy is shown by the **Random Forest Regression** Algorithm.

Error in Prediction

Along with the Prediction score we need to also consider the error in prediction also. As the algorithm with minimum error stands to be better than others with greator error. We have Mean Squared Error that calculates the mean of the squared differences between actual value and predicted value. Whereas Mean Absolute Error is the mean of the absolute difference between actual value and predicted value.

Algorithms	Error	
	Mean Squared Error	Mean Absolute Error
Multiple Linear Regression	0.05986	0.15332
Support Vector Regression	0.01257	0.07127
Decision Tree Regression	0.00020	0.00378
Random Forest Regression	0.00013	0.00373

Table 2 : Comparative analysis of Mean Squared Error and Mean Absolute Error

Along with the accuracy, the error is also of prime importance. From the above data of table 2, we can state that the least 'Mean Squared Error' and the least 'Mean Absolute Error' is obtained by the Random Forest Regression Algorithm.

VIII. Conclusion

On the basis of all the observations, the Visualization graphs & Comparative Analysis of the Prediction and the error, I can conclude safely that out of these four algorithms, the best algorithm suited well for this prediction of air pollution is Random Forest Regression Algorithm as it provides an accuracy of 0.99985 on the testing data with the least ' Mean Squared Error ' of 0.00013 and ' Mean Absolute Error ' of 0.00373. The above stated data is hereby true and experimented by the best of my knowledge.

IX. Future Works

In this research paper I have stated the observations and statistical information about how I worked on the dataset obtained from the sources like uci repository that is a good resource but are static. The prediction Quality has been improved a lot and the error is even minimized. However this Globally crucial issue calls for betterment and advancement in every approach we find, hence my aim is to work on the real time live dataset that could be directly extracted from online database and it would be done with the help of the technique called Web Scraping. I would even try to build a comparative analysis of the air Quality prediction among the various regions of the country. This will not only help people to understand their region's failure in keeping the quality of the air well but will also help them to improve their quality of air in comparison to the regions which are at better condition then them. It is really prime time to realize this global issue which I personally understand and will work as soon as possible in near future.

X. Acknowledgement

I have completed this work under the guidance of Dr. Pankaj Agarwal (Professor and head) & Ms. Sapna Yadav (Assistant Professor), Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad. I have been doing Summer Internship of Machine Learning under their mentorship and worked with various supervised and unsupervised machine learning algorithms. This work has been assigned by my instructors to showcase my efforts and learning that I have obtained. They beleived on me and showed me this opportunity to write a paper on my project on Air Pollution Prediction.

I express my heartfelt thanks to my mentors who have encouraged me, guided me throughout the project work. I really appreciate their cooperative support and moral boosting that they continuously provided me. I even thank my parents for their support and encouragement. My mentors as well as my parents both believed on my potential to successfully complete these paper. I even show my gratitude towards ‘ IMS Engineering College ’ for this wonderful opportunity. Any omission in this brief acknowledgement doesn't show my lack of gratitude.

XI. REFERENCES

[1]. Campbell-Lendrum, D., & Prüss-Ustün, A. (2018). Climate change, air pollution and noncommunicable diseases. Bulletin Of The World Health Organization, 97(2), 160-161. https://doi.org/10.2471/blt.18.224295

[2]. Li, J., Li, X., & Wang, K. (2019). Atmospheric PM2.5Concentration Prediction Based on Time Series and Interactive Multiple Model Approach. Advances In Meteorology, 2019, 1-11. https://doi.org/10.1155/2019/1279565

[3]. Soundari, M., Jeslin, M., & A.C, A. (2019). INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING. International Journal Of Applied Engineering Research, 14(0973-4562), 1-6. Retrieved 22 July 2020, from https://www.ripublication.com/ijaerspl2019/ijaerv14n11spl_34.pdf.

[4]. C R, A., Deshmukh, C., D K, N., Gandhi, P., & astu, V. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. International Journal Of Engineering Trends And Technology, 59(4), 204-207. https://doi.org/10.14445/22315381/ijett-v59p238

[5]. Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Journal Of Electrical And Computer Engineering, 2017, 1-14. https://doi.org/10.1155/2017/5106045

[6]. Sayyed, M., Sarode, A., Salunke, A., & Desai, S. (2020). A thorough Survey on prediction of Airpollution. Journal Of Emerging Technologies And Innovative Research, 7(3), 1-3. Retrieved 22 July 2020, from http://www.jetir.org/papers/JETIR2003302.pdf.

[7]. Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air Quality Prediction using Machine Learning Algorithms. International Journal Of Computer Applications Technology And Research, 8(9), 367-370. https://doi.org/10.7753/ijcatr0809.1006

[8]. Brownlee, J. (2020). Linear Regression for Machine Learning. Machine Learning Mastery. Retrieved 22 July 2020, from https://machinelearningmastery.com/linear-regression-for-machine-learning/.

[9]. Decision Tree Regression — scikit-learn 0.23.1 documentation. Scikit-learn.org. (2020). Retrieved 22 July 2020, from http://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html.

[10]. Random Forest Regression in Python - GeeksforGeeks. GeeksforGeeks. (2020). Retrieved 22 July 2020, from https://www.geeksforgeeks.org/random-forest-regression-in-

ython/#:~:text=A%20Random%20Forest%20is%20an,Aggregation%2C%20commonly%20known%20as%20bagging.&text=Random%20Forest%20has%20multiple%20decision%20trees%20as%20base%20learning%20models.

[11]. Data pre-processing. En.wikipedia.org. (2020). Retrieved 22 July 2020, from https://en.wikipedia.org/wiki/Data_pre-processing.

[12]. Outlier. En.wikipedia.org. (2020). Retrieved 22 July 2020, from https://en.wikipedia.org/wiki/Outlier#:~:text=In%20statistics%2C%20an%20outlier%20is,serious%20problems%20in%20statistical%20analyses.

Author

Suprateek Halsana is B.Tech 3rd year student in the department of Computer Science & Engineering at IMS Engineering College , Ghaziabad , UP , India . His areas of interest is Programming in Python , Data Analysis & Machine Learning .

Cite this article as :

Suprateek Halsana, "Air Quality Prediction Model using Supervised Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 190-200, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206435
Journal URL : http://ijsrcseit.com/CSEIT206435