Predicting Air Quality Index Using Machine Learning Techniques

Report By: Yash Agrawal

TEAM:

Roll No.	Name	Department	Email ID
18410	Yash Agrawal	EECS	yasha18@iiserb.ac.in
18265	Siddharth Sethi	EECS	siddharth18@iiserb.ac.in

CONTENTS

ABSTRACT

In the modern world, air pollution is one of the growing environmental issue. It poses major threat to health and climate. Air quality in cities is deteriorating day by day. Urban air quality monitoring has been a constant challenge with the advent of industrialization. Pollutants in air also impacts both water and land. This report presents a literature review of predicting air quality using Machine learning techniques and forecast the air pollution levels so that preventive measures can be taken by the people in order to minimize the air pollution. The air pollution databases were extracted from the US Embassy and Consulates in India. The proposed Machine Learning (ML) model is promising in prediction context for the

Delhi AQI. Now, these analysed results are very much helpful in improvement of accuracy for the predictions, hence they can be applied to other cities as well.

INTRODUCTION

Air pollution is the introduction of particulates, biological molecules or other harmful materials into the Earth's atmosphere. It causes disease, death to humans, damage to other living things or damage to the environment. Monitoring and preserving air quality is one of the most essential activities in many industrial and urban areas today. An air pollutant is a substance in the air that can create lot of adverse effects on humans and the ecosystem. The substance can be solid particles, liquid droplets or gases. The quality of air is affected due to many forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases including carbon dioxide, nitrous oxide, methane etc. is creating a serious threat for the quality of life in smart cities. Naive Bayes (NB) classifier has been shown to perform surprisingly well with very small amounts of training data. Though the NB is the fastest learning algorithm examining all its training inputs, Artificial Neural Networks' (ANNs') complexity makes them capable of handling huge amounts of data. Different algorithms and tools such as Neural Networks, fuzzy a system, Support Vector Machine for regression, fuzzy logic, Decision Trees has been used previously for different purposes. Many researches in various fields have targeted on air pollution seeking to solve the problem with different aspects but no research work has been published regarding Delhi Air Quality. This detailed report addresses Air Quality Prediction using various Machine Learning models. This data is nothing but previously existing data collected and stored for long period. Data Mining techniques is used for the collection of the data. Machine learning system is trained based on this training data; the systems accuracy increases with the increase of training data.

IMPORTED DATA INFORMATION AND VISUALIZATION

a. Data Source and Its Classification-

The following analysis and prediction is carried out using the dataset collected from the official website of 'US EMBASSY and CONSULATES in India' where the following data was selected:

1. Dataset of Delhi
2. Real Time AQI values
3. Records from 2016-2021
4. All places in Delhi city AQI value is integrated of nearly eight pollutants – Particulate Matter 2.5

Now, the AQI rage is classified over six categories shown as:

VALUE	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)
1	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
1 if >=151	Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
0	Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

The AQI(PM2.5 in the data) ranges are classified as **0 and 1** where **'1'** indicates that AQI is said to be **'Good'** and **'0'** indicates **'Poor'** AQI. If the value is '1' then it is said to be 'Good', in case of 2 it is 'satisfactory', likewise 3 it is 'Moderately polluted', 4 means 'poor', 5 represents 'very poor', 6 denotes 'severe' that is a dangerous situation which is denoted for the value '0'. Now, the reason to select Delhi for our study is due to its rising pollution level recently. For this research, we use the database collected based on Air Pollution in Delhi, which is the most polluted state in India. The value '1' directly implies PM 2.5 >= 151 ppm which is very serious level for breathing and may cause various breathing problems.

b. Data Description and Analysis-

The dataset provided for this study contains two separate files-

1. 'Complete' (Train Data)—which consists of all years' (2016-2020) data

2. 'weather-2021' (Test Data)—consisting of the weather data of year 2021 that has to be predicted.

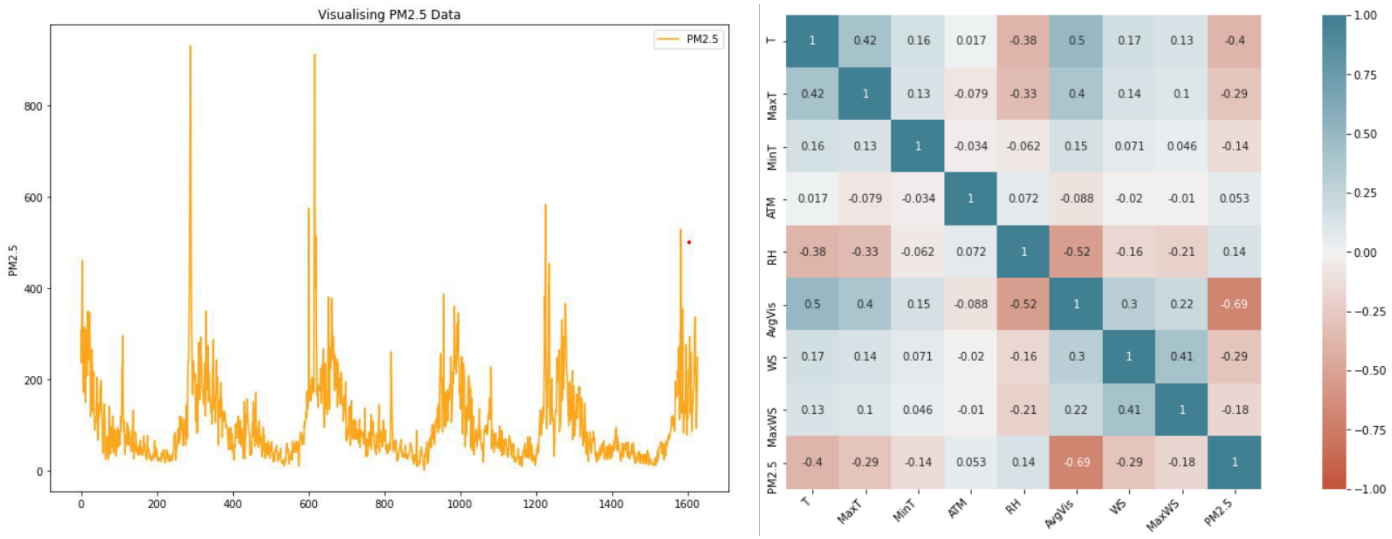
All the sets contain seven variables; Temperature, Max Temp, Min Temp, ATM, Relative Humidity, Average Visibility, Wind Speed, Maximum Wind Speed and PM 2.5, where the Pm 2.5 is the response variable indicated with 1 and 0 (good/poor).

The data description is the most vital part for analysing the nature of data and through this we can know what type of parameters to be used and eliminated. In this research work, PYTHON is used to carry out the descriptive statistics. Also, various libraries such as NUMPY and PANDAS provide a proper summarization of the collected data and hence very convenient for data analysis. The `.info()`and `.describe()` function gave the perfect statistics of the imported data shown as below:

Full form	Particulars	Count	Mean	STD	Min	25%	50%	75%	Max
Temperature	T	1627	26.352981	10.562436	6.9	19.8	28.5	31.9	311.0
Maximum Temperature	MaxT	1627	33.130854	9.33453	4.8	28	35.0	37.4	175.0
Minimum Temperature	MinT	1627	20.138844	26.186411	0.7	12.4	21.5	26.2	1017.8
Atmospheric Pressure	ATM	1627	1056.931346	679.770346	-18.2	1000.7	1007.2	1014.5	10178.8
Relative Humidity	RH	1627	2.293362	17.781513	0.0	53.0	65.0	75.0	102.0
Average Visibility	AvgVis	1627	6.636017	0.080788	0.2	1.2	2.4	2.9	5.8
Wind Speed	WS	1627	15.466318	4.624195	0.2	3.9	6.1	8.7	111.0
Maximum Wind speed	MaxWS	1627	10.684457	10.684457	3.5	9.4	14.8	18.3	222.0
PM2.5	PM2.5	1627	101.553910	90.819974	1.5	41.708333	72.25	134.520833	930.608696

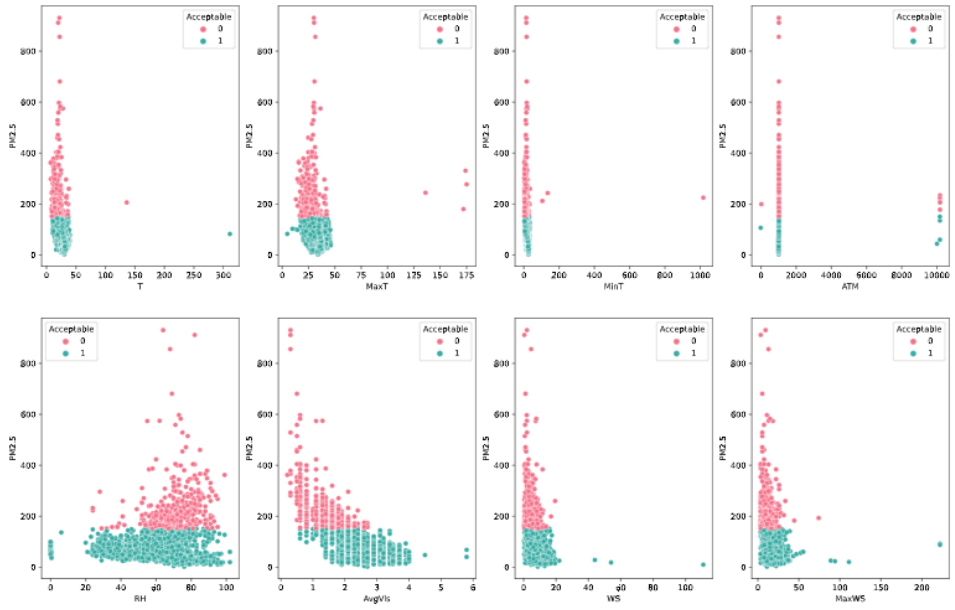
For visualizing the data, the graph for variation of AQI (PM 2.5) w.r.t. the days of year was generated as shown below.

This plot represents the AQI levels from date 1 Jan 2016 to 31 Dec 2020. In this dataset features like Temperature, Wind Speed and Average Visibility would be very essential for the prediction.



The correlation matrix clearly suggested to remove 'ATM' as a parameter for analysis. Moreover, this also showed the importance of Temperature and Wind Speed for the analysis of the data.

Now, as scatterplots are also used to determine the potential root of causes and effects, I also plotted scatterplots to determine which parameter is the most reliable ones.



The factor of 'ATM' is clearly discarded from the picture since we can see the behaviour of the plot w.r.t. PM 2.5 levels. Still, we haven't took that feature out from any of our model consideration. This is so because, there was no effect in R^2 score and in root mean square error, even if we removed ATM as a parameter. This showed that removing the parameters was not at all an option in this project.

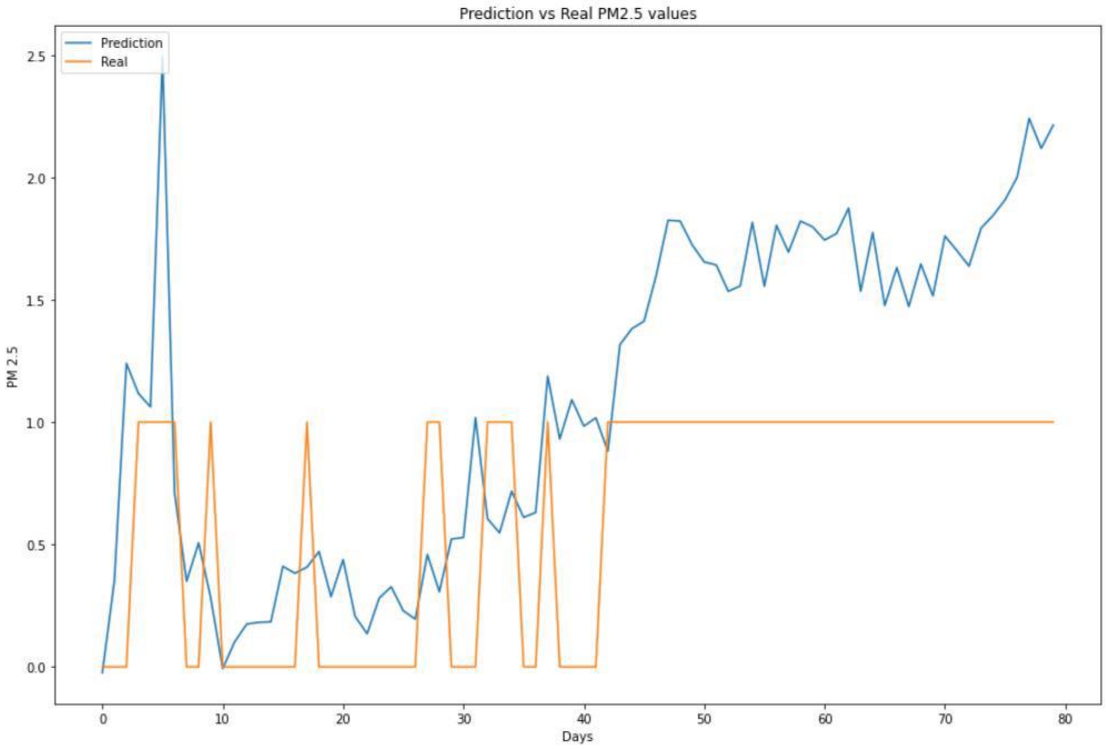
METHODOLOGY:

For this project, I am using Machine Learning techniques, i.e. Naive Bayes, Linear Regression, Logistic Regression, SVM, Random Forest, and ANN, to predict the level of PM 2.5 in Delhi for 3 months of data in 2021(i.e. Jan, Feb and March). Hence, we will discuss each machine learning algorithms used and analyse our results.

a. LINEAR REGRESSION:

Linear regression is perhaps one of the most well-known and well understood algorithms in statistics and machine learning. It is a statistical method that is used for predictive analysis. Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). Linear regression algorithm shows a linear relationship between a dependent (x) and one or more independent (y) variables.

It fits a linear model with coefficients *w = (w1,….., wp)* to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.



After applying the linear regression technique on the dataset, I found the relationship between the various parameters given and the given PM 2.5 levels more significantly. Although, it was observed from the curve that in order to fit the curve, the predictions were going out of range (i.e. 1). Hence , on running the code, we got the **R^2 score as 0.457** with a **root mean squared error as 0.661.** Hence, we can infer that this model clearly failed to determine the predictions for the upcoming year data. As we can see there were multiple parameters at a time, hence the technique used was **multiple linear regression.**

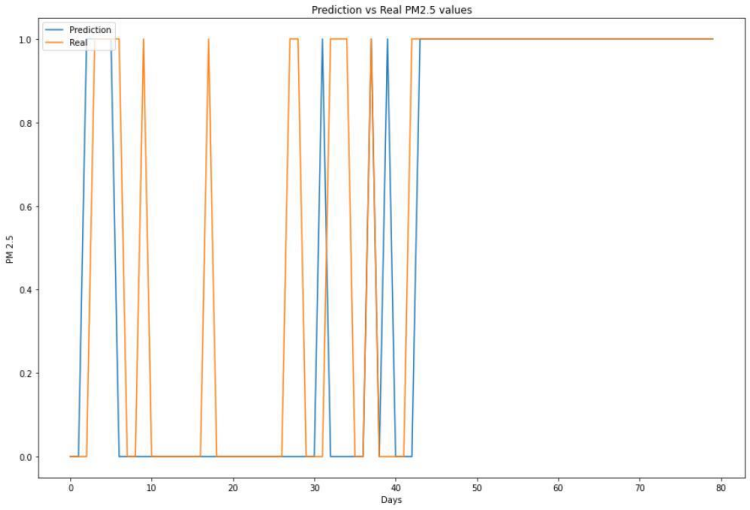
Now, this might have happened as we see that linear regression is only limited to linear relationships between dependent and independent variables. That is, it assumes a straight line relationship between them. Also, it is more sensitive to Outliers. Hence, clearly, this might would have been proposed a problem in our dataset, due to which satisfying results were not shown by this model. This linear regression model will fail if any random and unpredictable event occurs.

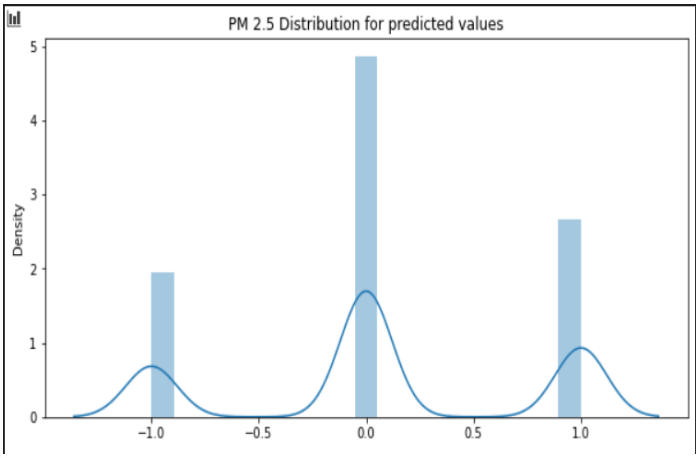
b. LOGISTIC REGRESSION:

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic Regression is a supervised learning algorithm that uses the logit (or logistic or sigmoid) function to calculate the probabilities of dependent categorial events and one or more independent variables. The logistic regression model computes a weighted sum of the input variables similar to the linear regression.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.

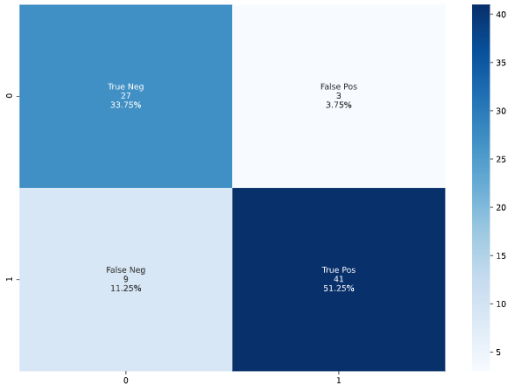
Now, I applied Logistic Regression to predict the AQI indices a 0 or 1. This model predicts the PM 2.5 levels as Good- '1' and poor- '0'. Hence, I used the '*liblinear*' solver to predict the test data.





Here, we observed that after applying the logistic regression model, the **R^2 score to come out as 0.85** with a **root mean squared error of 0.387**. This shows the Logistic Regression gives promising results towards prediction of AQI data. This model accurately predicts the value and can be beneficial for the prediction of the air quality index. We can also see the distribution of the predicted PM 2.5 values by this algorithm as shown in fig to the right.

As this model can interpret model coefficients as indicators of feature importance. Also, It makes no assumptions about distributions of classes in feature space. Hence, this might be the reason that this model has predicted the upcoming year's AQI value so accurately. Also the **confusion matrix** is as shown here:



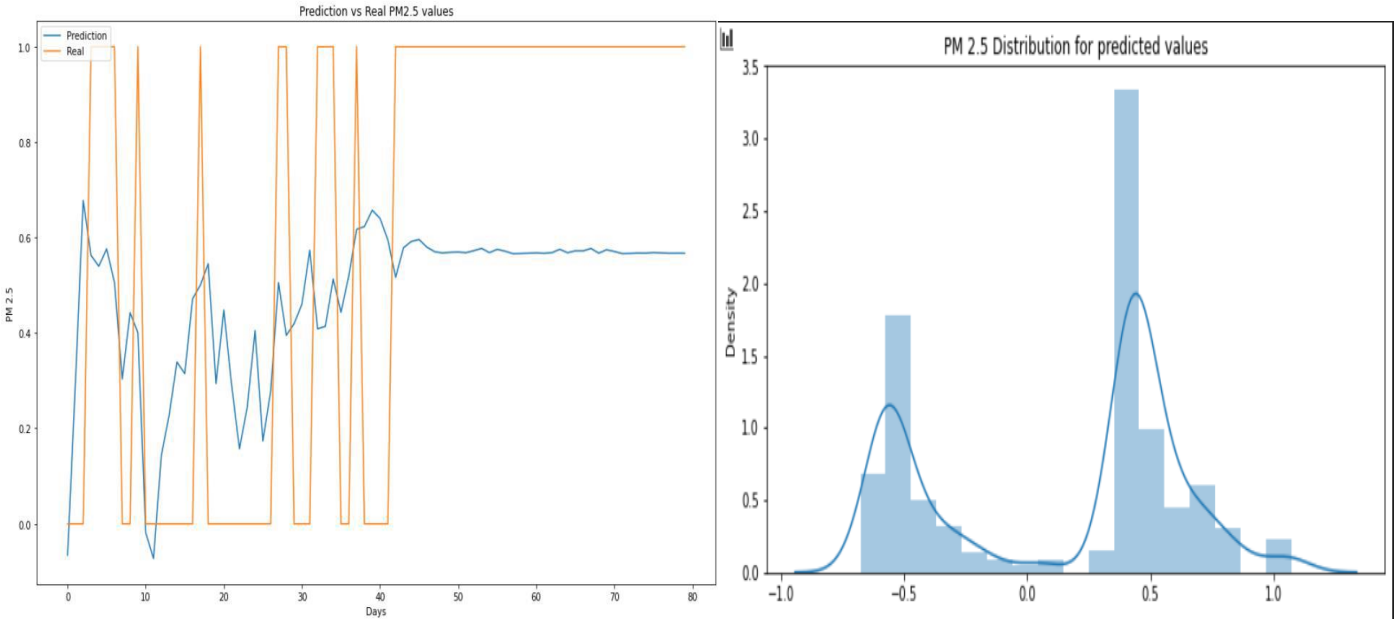
c. SUPPORT VECTOR MACHINE:

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. It can do both linear and nonlinear classification. It is mainly used to do the classification and regression. Kernel methods are a class of machine learning techniques that have become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection.

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function.Hence, the cost function looks like:

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Here, on using SVM regression model in the prediction, we see that the regression model provides us with the flexibility to define how much error to be accepted by the algorithm for fitting the data. We use **radial basis function kernel (RBF)** , a popular kernel used in various kernelized learning algos.

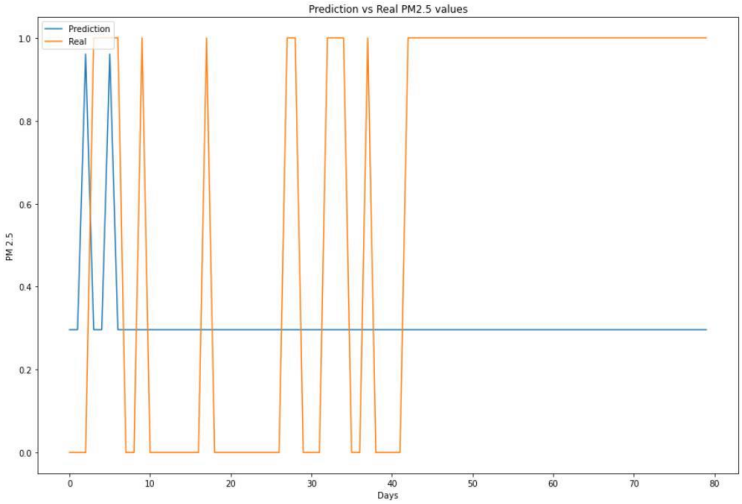


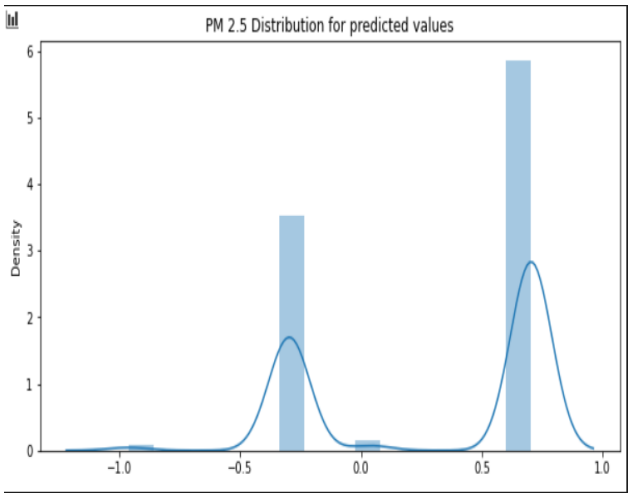
Now, for tuning the Regularization parameter (C) and epsilon value, we inferred that the model gave its best results on having **C = 1 and epsilon = 0.01(default values).** Hence, after applying this model we observed that the **R^2 value is 0.601** with **root mean squared error of 0.437** which shows that the model has predicted the values precisely.

d. DECISION TREES:

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It is a tree-structured classifier with three types of nodes. The ***Root Node*** is the initial node which represents the entire sample and may get split further into further nodes. The ***Interior Nodes*** represent the features of a data set and the branches represent the decision rules. Finally, the ***Leaf Nodes*** represent the outcome. This algorithm is very useful for solving decision-related problems. In real-time, as the number of variables increases tree grows larger and the algorithm becomes complex. In the Decision tree, we have two types, they are classification and regression trees. A classification tree is used to classify the dataset so that it is easy to analyse the data. But using this algorithm we cannot make a prediction. The Regression tree is a tree mainly used to predict continuous values.

Using this algorithm, we inferred although this model didn't give much of an R^2 score which is 0.**509 and a root mean squared error of 0.588** , still it showed a decent prediction of PM 2.5 levels and the model behaved fine and gives quite promising results.



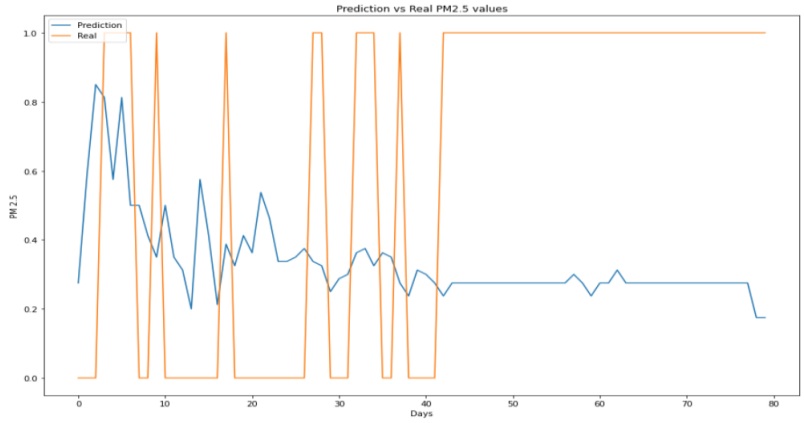


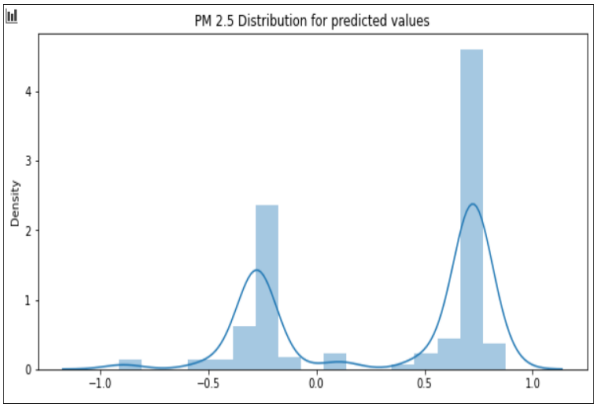
Decision trees have an advantage that it is easy to understand, lesser data cleaning is required, non-linearity does not affect the model's performance and the number of hyper-parameters to be tuned is almost null. However, it may have an over-fitting problem, which can be resolved using the ***Random Forest*** algorithm.

e. RANDOM FOREST REGRESSION:

Random Forest Regression is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It is defined as a set of decision trees to do regression and classification. Regression is used to calculate the mean value. This algorithm is more accurate, robust, and can handle a variety of data such as binary data, categorical data, and continuous data. Random Forest is nothing but multiple decision trees. It usually performs great on

many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.





When we applied this model for prediction of our air quality, we see that it gives us the score which is **R2 value to 0.951** and **root mean squared error of 0.603** which gives most accurate prediction for air quality index and outperforms other machine learning algorithms discussed so far.

Now, as various datasets require one hot encoding before implementing the Random Forest Algorithm. But, we felt that there's no need of that here as there were no categorical variables to turn to numerical form. Hence, we can infer that this model fits the test data completely and is the best model to predict the AQI values.

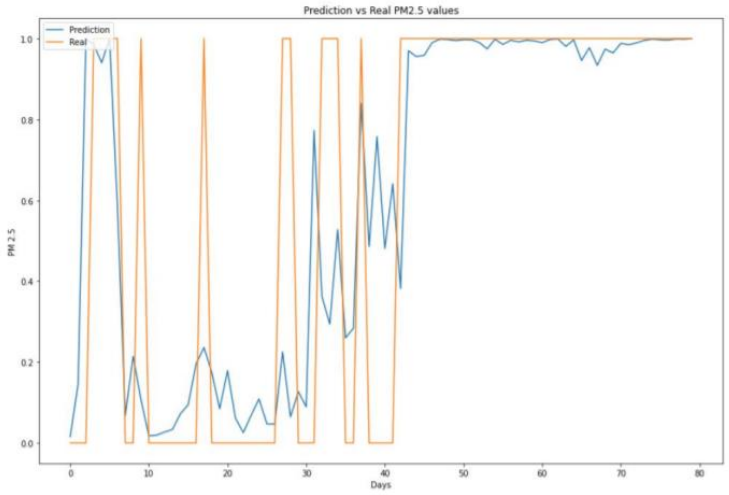
f. ARTIFICIAL NEURAL NETWORKS:

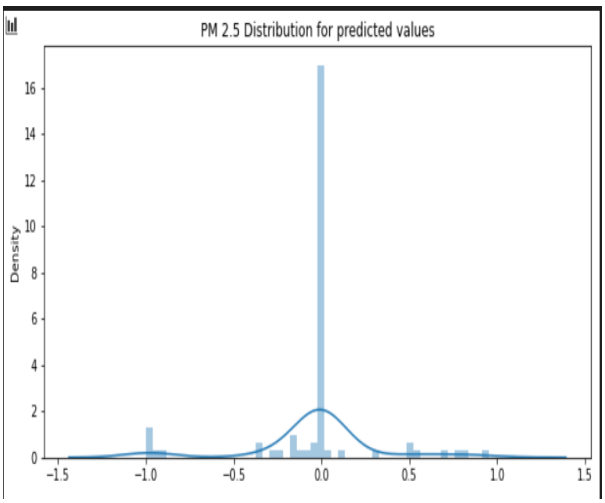
Artificial Neural Networks are a special type of machine learning algorithms that are modelled after the human brain. ANNs are nonlinear statistical models which display a complex relationship between the inputs and outputs to discover a new pattern. A variety of tasks such as image recognition, speech recognition, machine translation as well as medical diagnosis makes use of these artificial neural networks. A neural network may contain the following 3 layers:

- Input layer – The activity of the input units represents the raw information that can feed into the network.
- Hidden layer – To determine the activity of each hidden unit. The activities of the input units and the weights on the connections between the input and the hidden units. There may be one or more hidden layers.
- Output layer – The behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

Higher the weight of the artificial neuron, stronger is the input with which it is multiplied. The process of adjusting the weights is called learning or training the model. By feeding the network with a data set of desired inputs and set of desired outputs, the network can be made to repeatedly adjust the weights of each interior link to model more accurately and determine the correct output for a given input. By exploiting the behaviour of input and output of network, it can be trained with known data to predict the outcomes for new data. The network thus trained can be used to predict the outcome of new independent input data.

On Applying the model to our problem, we have used **"relu" activation function** for **input** and hidden layer and **"sigmoid" activation function** for the **output**. While compiling the optimiser was "**adam**" and the loss was "binary cross entropy" which enables the model neither to underfit nor overfit. Also, the batch size is 10 and epochs used is 100. Also, we used **L2 regularization** for the accuracy purposes. Briefly, L2 regularization works by adding a term to the error function used by the training algorithm. The additional term penalizes large weight values.
We see that the accuracy is **86.25%** which predicts the air quality index most accurately and can yield promising results.





g. PRINCIPAL COMPONENT REGRESSION:

In statistics, **principal component regression** (**PCR**) is a regression analysis technique that is based on principal component analysis (PCA). More specifically, PCR is used for estimating the unknown regression coefficients in a standard linear regression model. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors. It is hoped that the net effect will be to give more reliable estimates.

Hence, PCR is simply a 2-step process, which are described below:

1. Running PCA on the data to decompose the independent variables into the 'principal components', thereby removing correlated components.
2. Selecting a subset of the principal components and running a regression against the calibration values.

Here we have applied PCA and used several regression methods and all the scores and error are listed in the table shown below:

Legend: Performance reduced , No improvement, Improvement

MODELS	MEAN ABSOLUTE ERROR	MEAN SQUARED ERROR	ROOT MEAN SQUARED ERROR	R^2 SCORE
Linear Regression	0.578	0.436	0.660	0.457
Logistic Regression	0.150	0.150	0.387	0.889
Support Vector Regression	0.432	0.217	0.466	0.657
Decision Trees	0.404	0.335	0.579	0.291
Random Forest	0.578	0.369	0.608	0.952

After applying PCR, we see that SVR shows a good improvement and becomes a more stable model indicating better predictions than the one observed in the previous section of SVR. Random Forest is the most stable model where the score isn't changed much.

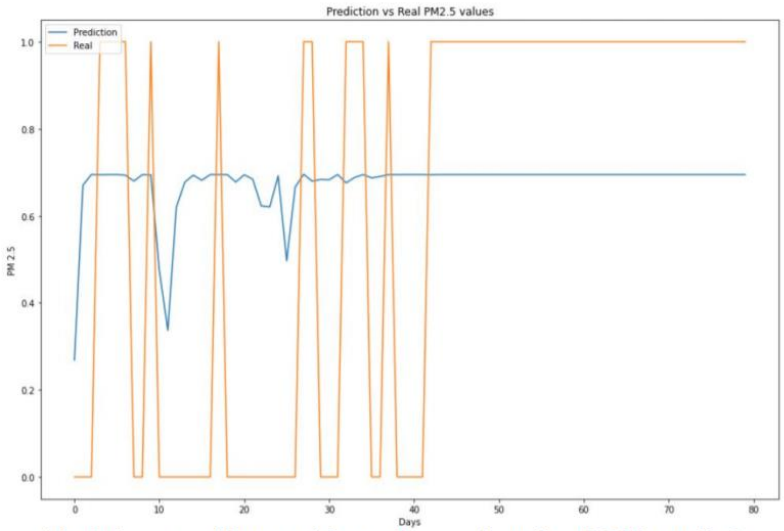


Fig 8: *Improved Support Vector Regression when PCA is applied*

Now, the advantages of using PCR are:

- PCR reduces the number of features of the model
- PCR is particularly useful on datasets facing the problem of multi-collinearity
- On datasets with highly correlated features, or even collinear features, PCR is quite useful
- PCR reduces the problem of overfitting
- Regression may take significantly lesser time because of lesser number of features
- The only decision we need to make while performing PCR is the number of input features to keep.

h. K-MEANS CLUSTERING:

K- means clustering is one of the most simple and popular unsupervised machine learning algorithm. It is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. A cluster refers to a collection of data points aggregated together because of certain similarities. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

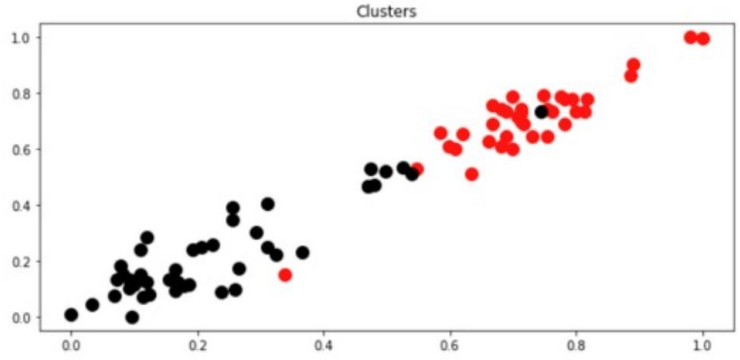
The k-means clustering algorithm mainly performs two tasks:

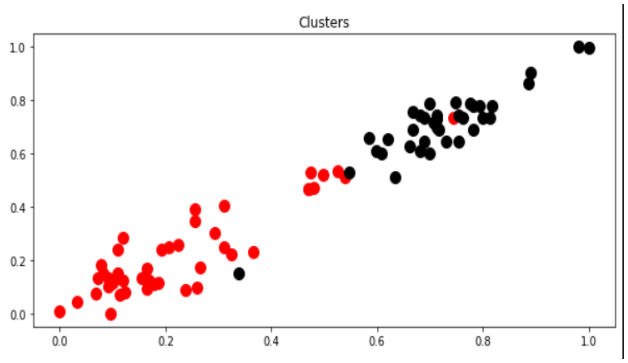
1 Determines the best value for K center points or centroids by an iterative process.

2 Assigns each data point to its closest k-center. Those data points which are near to the particular k center, create a cluster.

Now, on applying this model to the dataset, we observed the model is behaving very poorly, with an accuracy of 0.16 which clearly suggested that K-means clustering is not very convenient for the prediction on regression methods.

Note: For different compiling platforms, we got different accuracies with different plots





	precision	recall	f1-score	support
0	0.00	0.00	0.00	30
1	0.30	0.26	0.28	50
accuracy			0.16	80
macro avg	0.15	0.13	0.14	80
weighted avg	0.19	0.16	0.17	80

For Google Colab

	precision	recall	f1-score	support
0	0.70	1.00	0.82	30
1	1.00	0.74	0.85	50
accuracy			0.84	80
macro avg	0.85	0.87	0.84	80
weighted avg	0.89	0.84	0.84	80

For VS Code

CONCLUSION (With Comparative Analysis):

From the above survey, it can be concluded that from Supervised Machine Learning, Random Forest Regression performed the best out of all and also, Logistic Regression also produced promising results. Linear Regression was not so much convenient to the dataset. Moreover, after doing the PCR over the dataset, we found out that there was no significant change in the performance of the other models besides Logistic Regression whose score was somewhat improved. Also, PCR resulted in decreasing the performance of Decision Tree algorithm.

Talking about K-means clustering, we can say that the algorithm didn't give satisfying results and hence was discarded from the picture of prediction.

So, each model's score is given below:

MODELS	MEAN ABSOLUTE ERROR	MEAN SQUARED ERROR	ROOT MEAN SQUARED ERROR	R^2 SCORE
Linear Regression	0.579	0.436	0.660	0.457
Logistic Regression	0.150	0.150	0.387	0.886
Support Vector Regression	0.419	0.191	0.437	0.601
Decision Trees	0.551	0.347	0.588	0.509
Random Forest	0.574	0.371	0.609	0.951

Various factors like Temperature, Wind Speed, Atmospheric Pressure were used in the analysis of PM 2.5 levels for New Delhi's weather. We saw that various models gave promising results while quite a few didn't give great prediction results. Air pollutants concentration like SO_2, NO_2, O_3 etc. can also be considered for the prediction as input variables. . Furthermore, the model can be interfaced with the web applications for the user that could benefit from the work and take precautions in order to minimise the air pollution.

ACKNOWLEDGEMENT:

I would like to thank Dr. Kushal Kumar Shah, for their able to guidance and support in completing my project. Through the course "ECS 308: Data Science and Machine Learning" I would able to learn and apply various machine learning algorithms and learnt to use in real world application. This project and course would help me in pursuing a new field of research which can be coupled with other research fields and widen our knowledge.

REFERENCES:

1. Dataset: https://en.tutiempo.net/climate/2016/ws-421820.html

US Embassy and Consulates: https://www.airnow.gov/international/us-embassies-and-consulates/

2. LINEAR REGRESSION:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

3. LOGISTIC REGRESSION:

https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

4. SUPPOR VECTOR REGRESSION:

https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

5. DECISION TREE:

https://scikit-learn.org/stable/modules/tree.html

6. RANDOM FOREST REGRESSION:

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

https://www.geeksforgeeks.org/random-forest-regression-in-python/

7. ARTIFICIAL NEURAL NETWORK:

https://medium.com/@robertjohn_15390/simple-housing-price-prediction-using-neural-networks-with-tensorflow-8b486d3db3ca

8. PRINCIPAL COMPONENT REGRESSION:

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://nirpyresearch.com/principal-component-regression-python/

9. K-MEANS CLUSTERING:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

10. Introduction to Mathematical Statistics- Paul G. Hoel

11. An Introduction to Probability and Statistics- VIJAY K. ROHATGI, A. K. Md. EHSANES SALEH