Linear Regression Based Air Quality Data Analysis and Prediction using Python

Sonu S B
Department of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.p2ebs19024@cb.students.amrita.edu

A. Suyampulingam
Department of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
a_suyampulingam@cb.amrita.edu

2021 IEEE Madras Section Conference (MASCON) | 978-1-6654-0405-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/MASCON51689.2021.9563432

***Abstract*— Air pollution is a major scenario in the urban areas. The need of analyzing air quality is becoming an important requirement over past years. Atmosphere contains various levels of pollutants which needs to be monitored constantly. The analysis of air quality data becomes seamless using advanced software techniques. In this paper, the analysis of air quality is carried out by the means of python platform where the historical air quality data analysis of Coimbatore city is implemented and the prediction of the air quality data using linear regression model. For the air quality prediction part, dataset collection, data cleaning, pre-processing and prediction are implemented. The air quality data is processed using inbuilt libraries for analyzing the data which are available in python environment. The work consist of two parts where the air quality data analysis is implemented along with the prediction of air quality data which is the Air Quality Index(AQI) class from the previous records available.**

***Index Terms*— Air quality, Machine learning, data analysis, linear regression, python.**

I. INTRODUCTION

Air pollution is one of the greatest concerns across the world and in India. Many of the Indian cities today are suffering from the problem of air pollution and the tolerate standards are getting exceeded. In response to such grave situation there are automated air quality monitoring stations being operated across the country in various cities. These stations are providing real time data and historical data to the regulators and the citizens in order to understand the scale of pollution for taking control actions. The challenge lies in processing such big data using efficient statistical tools and to visualize the results for providing a meaningful interpretation. Unfortunately programs which coach the regulators, citizens and industries are missing to fill the gap of interpreting the air quality data [1].

Air quality is the measure of fresh air that human beings are in taking. Specifically the pollutants have direct health effects on living and non-living things. These air pollutants are generally emitted from traffic and also from natural sources like forest fire. The requirement for air quality analysis is to combine the data collection, analysis and visualizations of air quality data. In order to analyze the air quality data, an efficient software tool is required that should be easily accessible and understandable [2]. As a part of analysis and prediction of air quality tool, two experiments are performed. First experiment is the air quality analysis and the comparison of two time frames.

The dataset is in the .csv (comma separated value) form and is read into the data frame where air quality information is extracted for the comparison purpose Second experiment includes the prediction of Air Quality Index(AQI) class using regression technique. The first experiment includes the historical air quality data analysis of Coimbatore city to find the amount of atmospheric pollutants and visualized for the user reference. There are many atmospheric pollutants such as PM2.5, PM10, NO2, NO, CO Etc. PM2.5 is a major air pollutant present which is referred as Particulate Matter. PM2.5 is considered to be the major pollutant since it is emitted from industrial units. The amount of pm2.5 which is ejected over two time periods is compared using python tools. Studies prove that the particulate matter otherwise known as PM2.5 and PM10 causes more serious health hazards to the surroundings. The 2.5 and 10 are the diameter or the size of pollutant. It is vital to understand the amount of major pollutants which is been ejected over a time period. These results are evaluated for the same city on basis of two aspects; amount of major pollutant share on atmosphere, comparison of major pollutant across different time periods. Second experiments include the prediction of air quality of the same city using linear regression models. The Air Quality Index (AQI) is termed as the measure of air quality which is considered as a standard parameter [5]. In the second experiment, the dataset is taken from an open source for the Coimbatore city and extracted for the specific time period in order to feed the model. Coimbatore city is the second biggest after Chennai in Tamilnadu state, India is suffering from air pollution with its rates alarming every moment. Plethora of greeneries has been replaced by roads carrying vehicles vomiting large volumes of carbon dioxide and other green house gases.

The rate of emission has been rapidly increased due to urbanization and industrialization. The python provides a good platform to analyze the input data with its own inbuilt libraries. The experiments are carried out for monitoring and predicting purpose that can be useful for the common people to understand the havoc of the situation in many cities.

II. RELATED WORKS

The air quality analysis is an important study for understanding the pollution levels of various cities. The increase in the urbanization causes a demand for efficient monitoring about the atmospheric conditions.

Conventional studies about air quality had been exposed the fact regarding the development of software tools for predicting and analyzing the air quality data

Large amount of raw input data is processed using statistical tools to solve the problem of interest. Data is inspected and cleaned based on the requirements to derive insights that are valuable to the organization and to test theories. As like any other software, python also does user understandable techniques to handle the large amount of data [3].

The raw data requires reading of data, data processing and cleaning, summarizing data, visualization and deriving insights from data. The python libraries are already geared towards possibilities of data analytics and bring interesting results. Visualizations for the variation of pollutant level is required for Graphical User Interface and sophisticated machine learning algorithms are needed to process the air quality data which are made available in the python platform [4]. The big data is processed for machine learning model to learn and predict future data from previous records [6].

The prediction of Air Quality Index (AQI) for daily basis has been implemented using Principal Component Regression (PCR) in this study done [8]. The method performed the prediction of AQI for the year 2006 using previousyear's data. After that actual data is compared with predicted data over different seasons using Multiple Linear Regression Technique. The co linearity between independent variables is found out using Principal Component Analysis [8].

The next important study regarding air quality analysis is carried out by Huixiang Liu [9] where they had taken two cities for the study purpose. By the help of different datasets for the same city, they had forecasted the AQI values by taking the concentration of NOx into consideration. The first dataset contains the information such as hourly averaged AQI and the concentration of PM2.5, O3, SO2, PM10 and NO2 in Beijing City. The second dataset is collected from Italian city which contains hourly average concentrations of CO, hydrocarbons, Benzene, NOx, NO2 [9]. They concentrated mainly on NOx concentration as it is considered to be an important predictor for air quality evaluation.

This paper [10] uses machine learning algorithms to predict the PM2.5 concentration. Air beam, a mobile device is used to extract an unofficial data to measure PM2.5 value and in addition PM2.5 data had collected from the official site of Environment Protection Agency (EPA) for Melbourne city. Various machine learning algorithms were used to predict the PM2.5 concentration like Artificial Neural Network (ANN), Linear Regression (LR) and Long Short Term Memory (LSTM). Out of those LSTM gave best and accurate results compared to others [10].

The work done by Aditya C R [11] describes about deploying machine learning algorithms to forecast the levels of PM2.5 concentrations from the dataset available for that city. They initially made a classification based on the whether the air is polluted or not using Logistic Regression Algorithm and predicted the levels of PM2.5 by considering the previous records.

The air pollutants rapid increase in the urban areas is studied by Nidhi Sharma [12] in which critical observation on air quality data has been implemented. The future trend of air pollutants such as Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Suspended Particulate matter (PM), Ozone (O3) and Carbon Monoxide (CO). Data analytics based on the method of Time Series Regression Forecasting is used for the future prediction of pollutants. According to this study, there is a drastic increase in PM10 levels, NO2 and PM2.5 across Delhi City. The proportion of various pollutants with respect to time of day and day of the week has been studied in the paper [13].The effect of environmental factors such as wind speed, temperature and humidity is determined by the help of WEKA tool.

Based on the ZeroR algorithm used in WEKA tool, study shows that the pollutant levels are increasing at the peak hours of traffic and decreasing at weekends or holidays. K –means clustering algorithm is used to determine the dependencies between environmental factors like temperature, humidity and pollutant variables across Karnataka City. In this work [14] AirQ software is used which is proposed by World Health Organization (WHO) and the pollutant information of highly polluted city Tehran is taken where various health impacts are taken into consideration like lung disorders, cardio vascular problems. The work showed that the major air pollutant PM2.5 is responsible for the high mortality rate in the urban area of Tehran City.

The studies triggered the need of serious solutions for lowering the range of PM2.5 to avoid the health hazards faced by the people. All though a single machine learning algorithm is used for air quality analysis, there came a requirement for the comparative study of more than one algorithm. This paper [15] shows the study conducted by comparing more than one regression techniques. Multiple datasets have used to understand the variation of parameters like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Pollution prediction model is the best model that requires a large dataset and has less processing time [7]. Hence the proposed model is implemented for obtaining good accuracy in prediction which intakes less processing time.

III. METHODOLOGY

The system is divided into two parts, in first part a historical analysis is implemented using the dataset containing information of air quality of Coimbatore city and in the second part regression technique is incorporated for the prediction of air quality data from the previous records. Both experiments are implemented in the python platform. Various inbuilt libraries in python platform are used for the data analysis and prediction part.

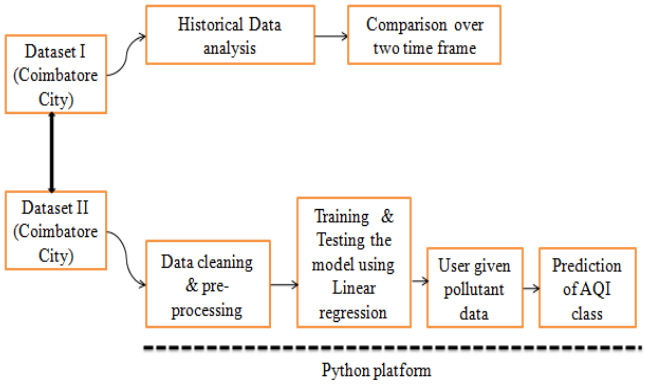

Fig.1. System overview

Fig.1 depicts the overall system and the methodology used for both the experiments to determine the air quality data. The first experiment contains the historical analysis where dataset I is used for analyzing the recorded data.

a. Historical data analysis :

The air quality information of Coimbatore city such as the date, pollutant levels are present in the dataset I. Air quality analysis is done based on the PM2.5 concentration in the atmosphere.

b. Comparison over time frame:

The analysis is implemented for two time frames. The months of December 2020 and January 2021 have been selected based on the lockdown period. . The two time periods are selected because both are observed as lockdown and non-lockdown period so that the variation can be clearly observed.

The second experiment contains the prediction of AQI class/bucket which determines the status of air quality as a measuring parameter.

a. Data cleaning and Pre-processing :

The dataset II is used for prediction for AQI class from the air pollutant information. The dataset contains the pollutant levels like PM2.5, NO2, and NOx etc. Major air pollutant is identified as PM2.5 in the first experiment. The dataset contains the raw data which is cleaned by using python libraries and data manipulation. The data cleaning is otherwise known as data modeling. The raw input data is shaped into easy processing format to feed into the regression model.

b. Testing and Training the model:

After the data pre –processing stage, the processed data is fed to the model in this stage. The machine learning algorithm uses a standard procedure to train and test the regression model. Linear regression technique is used to train and predict the future data.80% of dataset is used for training where 20% is used for testing the model. Basedon the model used, corresponding error probabilities are calculated and data is made to fit in the same model.

c. Prediction of AQI class:

After the training and testing stage, then the model is further checked for the accurate prediction for air quality class/bucket. The AQI class is divided into three divisions of air quality status which are user defined in this stage. The model takes a random data from the user to find out the array value aka AQI value. Based on threshold setting of the parameter, the user data is classified into correct class. The AQI determines the status of atmosphere quality of that particular test city.

IV. EXPERIMENT I: EXTRACTING AIR QUALITY INFORMATION AND COMPARISON

The first experiment is the historical data analysis compared over two months. The dataset shown in Fig.2. contains the air quality information of individual pollutants such as PM2.5, PM10, O3, NO2, SO2 and CO values over a time period. The pollutant values are extracted from the dataset initially. The timeframe used for comparison is in the months of December 2020 and January 2021.

	date	pm25	pm10	o3	no2	so2	co
0	2021/2/1	137					
1	2021/1/4	80	49	4	9	7	5
2	2021/1/5	109		3	9	7	4
3	2021/1/6	77		1	8	7	5
4	2021/1/7	64		2	9	8	4
..	...	...	...	...	...	...	...
584	2019/8/23		35	11	3	5	13
585	2019/8/30		29	10	4	4	13
586	2019/9/15		27	11	1	4	14
587	2019/6/14		33	11	2	5	9
588	2019/8/29			11	3	4	13

Fig.2. Dataset I for time frame comparison of PM2.5

Both time frames are monitored with lockdown and without. The specific timeframe is extracted and major share of pollutant is found out i.e. PM2.5. The test period is from 01-12-2020 to 29-01-2021.

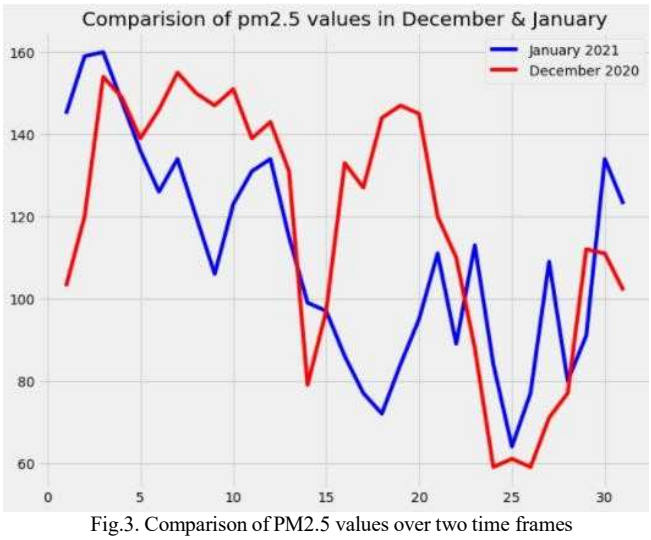

Fig.3. Comparison of PM2.5 values over two time frames

The individual share of PM2.5 shows a huge spike across the dates compared to other pollutants and can be inferred that PM2.5 is the major pollutant in Coimbatore city.

From Fig.3. it is observed that the comparison between two time frames shows a small variation of PM2.5. The two time frames are taken because both times are with lockdown and without lockdown. Hence the road transport and industrial units which contribute more pollutants will be closed therefore getting a closer observation on the variation of pollutant graphs.

V. EXPERIMENT II: PREDICTION OF AQI CLASS USING LINEAR REGRESSION ALGORITHM

One of the important machine learning algorithm for the prediction of future data is the Linear Regression(LR) method. This method is used mostly for the predictive modeling which is used for understanding the relationship between input and output variables.

The model is represented by the simple linear equation which represents a set of input values x and y with two co- efficient where the first co-efficient represents the scaling factor and the second one represents the degree of freedom on the classifier line often referred as intercept co-efficient. LR method is mainly used for predictive analysis applications like air quality data which helps to get the average hourly responses of air pollutants easily.

The Dataset II is collected for Coimbatore city over different time periods. The experiment aims to predict the air quality from the recorded data in the dataset. Based on the new values of each pollutant, AQI class is predicted. AQI class or bucket is used to group the AQI values into categories.

Dataset collection: The dataset II contains the air quality information for Coimbatore city which includes pollutant information. Dataset contains information between July 2019 to July 2020. The time period is taken as nearly one year to extract the overall amount of air quality data, so as to predict the AQI class more accurately. If the dataset size is large, then the model is trained and predicted with a good accuracy. Fig.4 shows dataset II for the AQI prediction where the information such as date of recorded data along with individual pollutant values of PM2.5, PM10, NO, NO2, NOx, CO, SO2, O3, AQI and AQI bucket.

	Date	PM2.5	PM10	NO	NO2	NOx	CO	SO2	O3	AQI	AQI_Bucket
0	2019-07-01	25.21	39.66	7.14	7.22	14.36	0.89	9.53	21.47	46.0	Good
1	2019-07-02	32.92	46.75	9.74	8.87	18.61	1.13	12.05	21.03	61.0	Satisfactory
2	2019-07-03	32.66	44.95	9.48	8.66	18.14	1.14	10.79	16.94	64.0	Satisfactory
3	2019-07-04	40.89	49.26	9.23	9.57	18.79	1.11	10.44	16.05	64.0	Satisfactory
4	2019-07-05	31.52	39.65	9.43	7.96	17.39	1.19	10.58	17.62	70.0	Satisfactory
...	...	...	...	...	...	...	...	...	...	...	...

Fig.4. Dataset II for AQI class prediction

Data cleaning and pre- processing: Data cleaning is the stage where the missing values are removed and data modeling are done. Data modeling is the process of changing the structure of bulk dataset to obtain an easy processing form.

The dataset contains many null values/missing values for each pollutant value; hence it should be identified and eliminated in the data cleaning stage itself. The raw data should be cleaned so that the null values must be filled and the columns having more missing values should be dropped for feeding to the regression model. These missing values may lead to uncertainty in the prediction of results. For identifying the number of missing values for each pollutant a heat map is used as the visualizing tool.

```
City           0
Date           0
PM2.5          7
PM10           6
NO            54
NO2           11
NOx            4
NH3          122
CO             3
SO2            3
O3             3
AQI           37
AQI_Bucket    37
dtype: int64
```

Fig.5. Null values present for each pollutant

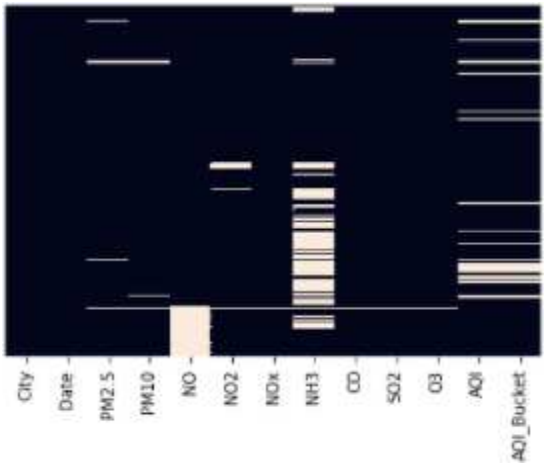

Fig.6. Plotting the missing values via heat map

Fig.6. shows the heat map for determining the missing values of pollutants in the dataset. Heat maps are the data visualization tool represented as color coding in order for the user to understand those columns need to be removed from the dataset. Fig.4. clearly shows columns of NH3, NO, AQI, AQI Bucket contributes more missing patches and hence those corresponding columns of pollutants are dropped at this stage. The specific columns which contains more missing values are eliminated since the loss of 10% dataset causes no big difference within the prediction results.

Checking normal distribution of dependent variable: Generally scatter plots are used to find the normal distribution between variables and the linking of pollutant variables (columns).

The linear regression model is based on the linear equation which represents the relation between input and output variables. The equation checks the relation between dependent and independent variables in the dataset is linear or not. The linear equation is as;

$$y = mx + c$$

Here y is representing the dependent variable i.e. the AQI which is plotted along Y axis and independent variables i.e. the pollutant values from PM2.5 to O3 is assigned to the term x which is plotted along X axis. This linear equation will be applied to the observed data.

The linear equation is to find the variables that are showing skewness otherwise known as the property of showing a deviation from the linearbehavior. The scatter plots helps to check the distribution of columns and to check whether the data is normallydistributed by obeying a linear relation.

Fig.7. Scatter plot for checking linear relation

Fig.7. shows the scatter plot between the dependent variable AQI and independent variables variables i.e. the pollutant values from PM2.5 to O3 to imply the strength of relationship between variables. From the scatter plot it can be inferred that the data is distributed normally and the plot appears to support the linear relation.

Checking the correlation: Correlation is the quantity of degree in which two variables are associated. Correlation of the data of all variables is checked using correlation matrix which is represented inFig.8. Correlation matrix is the table showing the correlation co- efficient between variables. A single cell in the correlation matrix depicts the association between two variables. The technique is used to find the linear association between two quantitative variables.

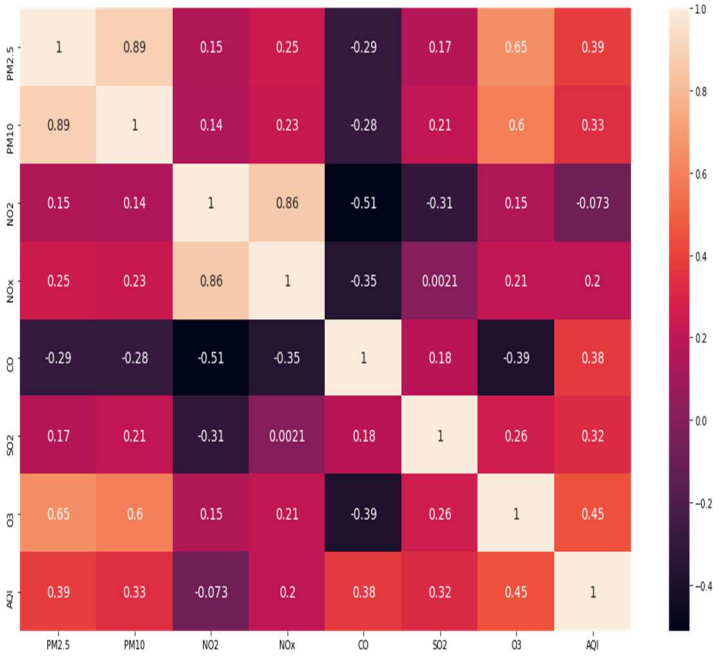

Fig.8.Covariance matrix

From the Fig.8, it is observed that the pollutant variables PM2.5 and PM10 are correlated and NOx and NO2 are positively correlated. The correlation between pollutant variables is depicted by the color coding of the corresponding cell at the intersection of both X and Y axis values. There are positive and negative correlation between independent and dependent variables Hence it can be removed since it may cause multi-co linearity.

Determining Inter Quantile Range (IQR) and Variation Inflation Factor (VIF): Quantile is the set of data points that divide the whole dataset into equal size and also used as a parameter for checking the normal distribution.

The Quantile range is calculated as the difference between 75th and 20th quantiles of the data. The Quantile values are named as Q1, Q2 and Q3. Q2 is the median where Q1 is the 25% of the normally distributed data and Q2 is the 75% of the data**.** Qauntile range is calculated as Q3-Q2. IQR is found out for determining the amount of outliers. Outliers are the extreme data points that lie far from the normal data pointswhich cause the non linearity.

In other hand, VIF determines the multi-co linearity in the regression analysis technique. It is calculated using software for determining the amount of inflation of regression co-efficient which causes the multi co linearity.

```
PM2.5    15.97
PM10     22.74
NO2      28.80
NOx      21.57
CO        0.83
SO2       3.80
O3       20.94
AQI      22.00
dtype: float64
```

	VIF	features
4	4.353593	CO
6	10.578851	O3
5	12.645131	SO2
2	16.305868	NO2
3	22.655574	NOx
1	36.488310	PM10
0	41.371458	PM2.5

Fig.9. IQR and VIF values

Fig.9. shows the IQR and VIF values calculated for each pollutant.VIF describes how the data is varied across the range of pollutant values across the dataset. Each pollutant posses an index that are directly passed to the python inbuilt functions to find the resultant VIF and the values are stored in the data frame.

V. IMPLEMENTATION OF TRAINING MODEL

As the data pre-processing stage is completed, the libraries areimported and the independent variables such as PM 2.5 and O3 are assigned to X and dependent variable AQI to Y. For training and testing the regression model for air quality analyzer tool, theregression method is imputed and the result is used to test the model shown in Fig.10. 80% of data is used for training the model and 20% is used to test the model.

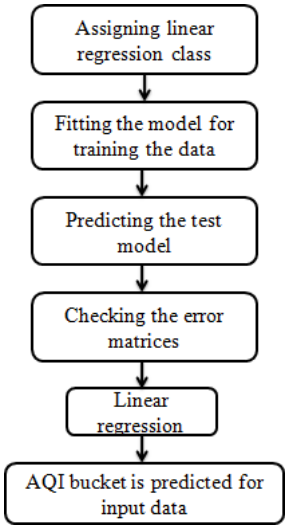

Fig.10. Methodology for AQI prediction

Error matrices are the error probabilities which are required to determine whether the regression model is satisfying the input data. Certain corrections have to make for the model to predict accurately and the regression assumptions must be satisfied.

i. Error matrices are determined by calculating Mean Absolute Error (MAE); Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
ii. MSE gives the square of the difference between calculated and true value. MAE gives the difference between true and predicted value. RMSE is used to predict the accuracy of model.
iii. R2 score is calculated that provides the information on how the regression model fits for observed data and theperformance of the model. The prediction of AQI is implemented using the pollutant values in this stage. R2 score is the statistical parameter to find out the performance of model.
iv. The correctness of R2 values depends on the model andthe findings; probably it should be greater than 50 %. The calculated R2 score for the regression model is 50.8 %.

VI. RESULTS

The historical analysis is done using the pre-recorded datasetof Coimbatore city. The comparison over two time frames i.e. over lockdown and non-lockdown period are implemented and the variation of major air pollutant PM2.5 is observed.

In the second experiment the prediction of AQI is done using the linear regression technique by extracting the significant pollutant data. The parameter values for the errorprobabilities are obtained as MAE=8.89, MSE=141.77, RMSE=11.9.

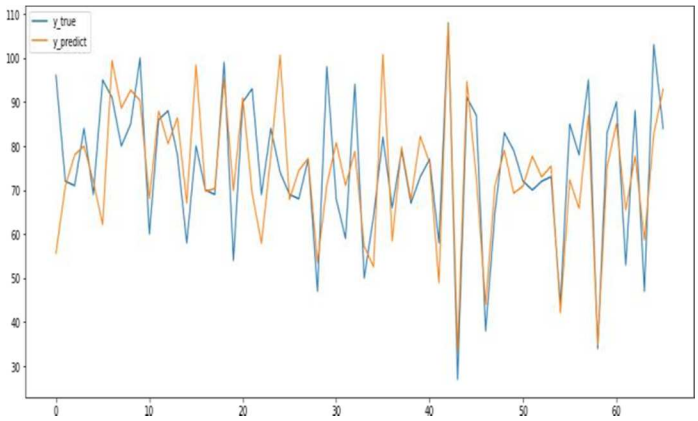

Fig.11. Plotted graph of test data and predicted data

In Fig.11. the blue line shows the variation of test data and the orange line shows the variation of predicted data along with test data. The graph depicts the variation of predicted data from the true data is small. In order to test the model, theuser provides random values for each pollutant to the regression model created.

	PM2.5	PM10	NO2	NOx	CO	SO2	O3
0	25.21	39.66	7.22	14.36	0.89	9.53	21.47

Fig.12. User given data for prediction

```
[62] if y_hat <= 50:
         print("Good")

     elif y_hat <= 100:
         print("Satisfactory")

     else:
         print("Moderate")

     Satisfactory
```

Fig.13. Predicted result for new data

Fig.12. shows the user data gives as input for the model to predict the AQI value. The model is taking these input pollutant values and calculating the AQI value which is storedin an array format. The data is inputted for the values of pollutants PM2.5, PM10, NO2, NOx, CO, SO2, O3.

The array value is referred as the AQI value and these AQI values are classified into different groups based on the standard pollution levels. Here for example, the AQI values below a threshold of 50 belongs to "*good"* class, above 50 belongs to "*satisfactory"* class and above 100 belongs to *"Moderate"* class. The obtained value for the user given data is 59.8 and it is predicted as *"satisfactory"* class.

VII. CONCLUSION AND FUTURE SCOPE

The air quality analysis and prediction of Coimbatore city is performed using linear regression algorithm by the help of python platform. The analysis of air quality is a crucial factor in deciding the prediction of future air quality data. The historical data of air quality is analyzed and the share of major pollutant PM2.5 is calculated and visualized. Using the regression model, the dataset is used to train and test the model in which data cleaning, pre processing, training and testing are executed. The AQI class is predicted from the user data with good accuracy. More regression techniques like Lasso, Ridge along with Linear Regression can be incorporated to tune the model further in the air quality analysis applications in future.

REFERENCES

[1] Yadav, S. K. & Sahay, M. "A Study on Automobile Industry growth in India and Its Impact on Air Pollution." Retrieved from:https://www.amrita.edu/sites/default/files/a-study-on-automobileindustry-growth-in-india-and-its-impact-on-air-pollution.pdf.

[2] D. K. Niranjan, N. Rakesh, "Real Time Analysis of Air Pollution Prediction using IoT," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 904-909, doi: 10.1109/ICIRCA48905.2020.9183251.

[3] S. Harikumar, V. Mannam, C. Mahanta, M. Smitha and S. Zaman, "Interactive Map Using Data Visualization and Machine Learning," 2020 6th IEEE Congress on Information Science and Technology (CiSt), 2020, pp. 104-109, doi: 10.1109/CiSt49399.2021.9357237.

[4] R. Ramachandran, G. Ravichandran and A. Raveendran, "Evaluation of Dimensionality Reduction Techniques for Big data," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 226-231, doi: 10.1109/ICCMC48092.2020.ICCMC-00043.

[5] P. Wang, H. Feng, G. Zhang, and D. Yu, "A Period-Aware Hybrid Model Applied for Forecasting AQI Time Series," Sustainability, vol. 12, no. 11, p. 4730, Jun. 2020

[6] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Feb. 2018.

[7] Wenjing Mao, Weilin Wang, Limin Jiao, Suli Zhao, Anbao Liu,Modeling air quality prediction using a deep learning approach: Method optimization and evaluation,Sustainable Cities and Society,Volume 65,2021,102567,ISSN 2210-6707,https://doi.org/10.1016/j.scs.2020.102567.

[8] Anikender Kumar, PramilaGoyal, "Forecasting of air quality in Delhi using principal component regression technique", Atmospheric Pollution Research, 2 (2011) 436-444.

[9] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences, ISSN 2076-3417; CODEN: ASPCC7, 2019, 9, 4069; doi:10.3390/app9194069

[10] Ziyue Guan and Richard O. Sinnot, "Prediction of Air Pollution through Machine Learning on the cloud", IEEE/ACM5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 978-1- 5386-5502-3/18/$31.00 ©2018 IEEE DOI 10.1109/BDCAT.2018.00015.

[11] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018 [11]. https://www.iqair.com/us/india.

[12] Nidhi Sharma , ShwetaTaneja , VaishaliSagar , Arshita Bhatt, "Forecasting air pollution load in Delhi using data analysis tools", ScienceDirect, 132 (2018) 1077– 1085.

[13] Mohamed Shakir, N. Rakesh, "Investigation on Air Pollutant Data Sets using Data Mining Tool", IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1- 5386-1442-6.

[14] KazemNaddafi, Mohammad SadeghHassanvand, MasudYunesian, , FatemehMomeniha, RaminNabizadeh, SasanFaridi, Akbar Gholampour, "Health impact assessment of air pollution in megacity of Tehran, Iran", IRANIAN JOURNAL OF ENVIRONMENTAL HEALTH SCIENCE & ENGINEERING, 2012, 9:28.

[15] S. Ameer et al., "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities," in IEEE Access, vol. 7, pp. 128325-128338, 2019, doi: 10.1109/ACCESS.2019.2925082.