Air Quality Index Prediction Using Machine Learning

Bhavadhaarani.R[1], Preetha.S[2], Sasirekha.R[3], Reena.R[4]
[1, 2] Dept of Computer Science and Engineering
[3]Assistant Professor, Dept of Computer Science and Engineering
[4]Associate Professor, Dept of Computer Science and Engineering
[1, 2, 3, 4] Prince Shri Venkateshwara Padmavathy Engineering College, Ponmar, Chennai

Abstract- *For human life in earth, Air plays an important role, but in recent times the standard of air has been compromised by economic activities, geographical factors, industrial parameters and prolonged consumption of non-renewable sources of energy. Due to poor air quality, many lung-related diseases, premature deaths are recorded in step with the World Health Organization (WHO).Several researches has been dedicated to predict the poor air quality, but most of the studies have failed because of insufficient longitudinal data, making it difficult to depict the seasonal and other factors. Therefore, the air quality must be monitored consistently. The air quality is predicted with the assistance of Air Quality Index (AQI)[1]. AQI includes averaging values of several factors like Carbon monoxide(CO), Nitrogen dioxide(NO2), Sulphur dioxide(S02),Ozone(O3),PM 2.5,PM 10.These values are collected from Central Control Pollution Board are fed as input and also the optimized AQI obtained from sensor's output are set as a target to train the regression model.*

Keywords- Air quality index, Air quality monitoring, Linear Regression, Root Mean Square Error, Data processing.

I. INTRODUCTION

According to the World Health Organization (WHO)[2],air pollution is responsible for around 1.3 million deaths worldwide. The pollutants from industries and vehicular emissions are mainly responsible for the aerial contaminants. Hence, the pollution is caused mainly from industrial countries. As the largest industrial country, India produces great deal of dioxide, Carbon monoxide and other harmful aerial contaminants. The recent studies in air quality shows that the pollutants present within the air are comparatively less during the COVID-19 lockdown

II. EXISTING APPLICATION

Many monitoring methods are applied for predicting and forecasting the air quality. However, it is extremely challenging to create accurate predictions on the standard of air. The prediction model is tougher to take care of accuracy and stability. The sufficient training data plays a serious role within the accuracy of prediction models. The lack of sufficient training data is that the common problem in prediction models. As we all know that the air quality monitoring in developing countries depends on the monitoring stations. These monitoring stations arranged by government are very sparse. Therefore, making accurate predictions and forecast based on insufficient training data may be a challenging and realistic problem.

III. PROPOSED SYSTEM

The air quality is predicted with the assistance of Air Quality Index (AQI). AQI includes averaging values of several factors like Carbon monoxide(CO), Nitrogen dioxide(NO2), Sulphur dioxide(S02),Ozone(O3),PM 2.5,PM 10.These values are collected from Central Control Pollution Board[4](CCPB),India are fed as input and also the optimized AQI obtained from sensor's output are set as a target to train the regression model.

IV. FLOW DIAGRAM

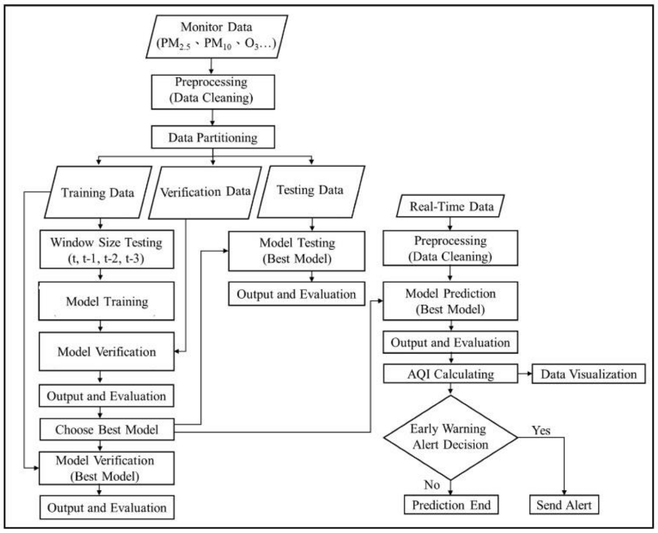


Figure 1: Flow Diagram

The Fig 1 consists of six components which are discussed below:

i.Dataset:

The pollutant concentration on all gases are needed to predict the air quality index of a part, which is obtained from Central Pollution Control Board website (cpcb.nic.in),which contains the info about all the pollutants that pollutes the cities every year. So as to calculate the AQI for a specific year, the Linear regression algorithm is employed. The null values are set to infinite to get rid of the outliers within the predicted and actual values the Box-plot analysis is employed.

ii.Pre-processing:

The outliers within the dataset are mainly of transmission errors has huge variation than the conventional valid results. To remove such outliers from the data we use Boundary Value Analysis(BVA).The Boundary Value Analysis is employed to seek out the upper quaritle range and lower quaritle range.

iii.Histogram:

Histogram equalization could be a method in image processing of contrast adjustment using the image's histogram.

iv.Prediction:

Using Linear regression algorithm, air quality is predicted.

v.Decisions:

The final decision results are given based on the Quality metrics.

V. ALGORITHM USED

Linear Regression[6](LR) could be a regression model that estimates the connection between one dependent variable and one independent variable by using a straight line. As referred in Fig 2, both the variables must be quantitative (measurable quantity). The prediction errors are identified using RMSE. Root Mean Square Error (RMSE)[5] is that the Standard Deviation (SD) of the errors that happens during the prediction of the dataset. It is used for determining the accuracy of the model. The prediction errors are usually called residuals.

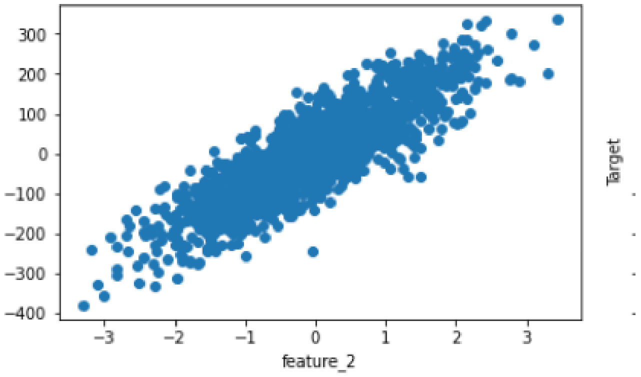
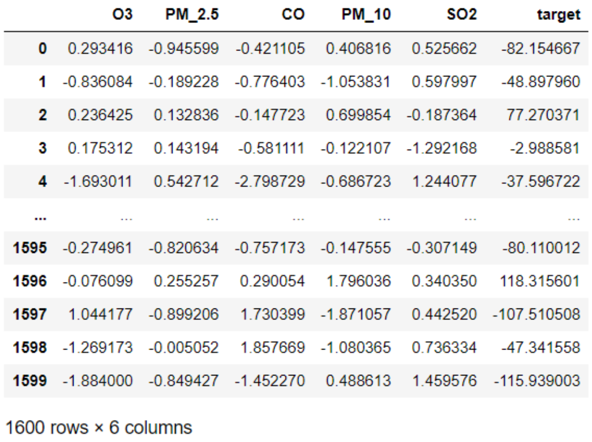


Figure 2: Linear Regression

VI. DATA COLLECTION AND PREPROCESSING

Information Collection is plays a significant role in building an AI model. It is the gathering of errand-related data depended with some focused factors to research and build some significant results. In any case, a little of the information could be uproarious, as an example, may contain mistaken qualities, inadequate qualities, or inaccurate qualities. Subsequently, it's necessary to handle the information before breaking it down and visiting the outcomes. Information pre-handling should be possible by information cleaning, information change, and information determination. So as to calculate the AQI for a selected year, the linear regression algorithm is employed.The null values are set to infinite to get rid of the outliers within the predicted and actual values the Box-plot analysis is employed. The Fig 3 consists of 1600 rows and 6 columns which makes 9600 values in total



	O3	PM_2.5	CO	PM_10	SO2	target
0	0.293416	-0.945599	-0.421105	0.406816	0.525662	-82.154667
1	-0.836084	-0.189228	-0.776403	-1.053831	0.597997	-48.897960
2	0.236425	0.132836	-0.147723	0.699854	-0.187364	77.270371
3	0.175312	0.143194	-0.581111	-0.122107	-1.292168	-2.988581
4	-1.693011	0.542712	-2.798729	-0.686723	1.244077	-37.596722
...	...	...	...	...	...	...
1595	-0.274961	-0.820634	-0.757173	-0.147555	-0.307149	-80.110012
1596	-0.076099	0.255257	0.290054	1.796036	0.340350	118.315601
1597	1.044177	-0.899206	1.730399	-1.871057	0.442520	-107.510508
1598	-1.269173	-0.005052	1.857669	-1.080365	0.736334	-47.341558
1599	-1.884000	-0.849427	-1.452270	0.488613	1.459576	-115.939003

1600 rows × 6 columns

Figure 3: Data Collection

VII. ANALYZING THROUGH HEAT MAP

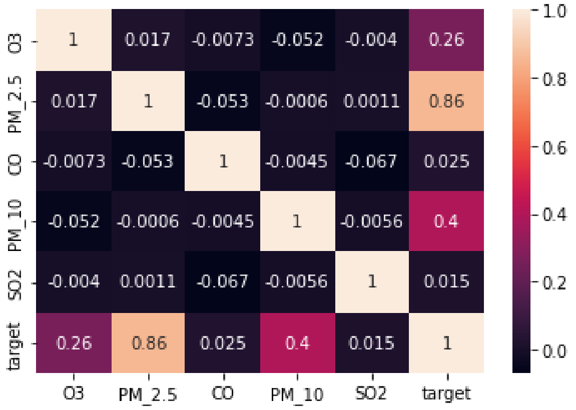


Figure 4: Heat Map

From the above heat map, we observe a strong correlation PM 2.5 and PM 10

It can be keenly observed that **feature 2** and **feature 4** are more important for predicting the target variable.

VIII. HISTOGRAM

A histogram is a graphical illustration that organizes a collection of data points into user-specific levels. Similar in look to a bar chart, the histogram at Fig 5, condenses data points into interpreted visible by taking many data points and grouping them into logical levels or bins

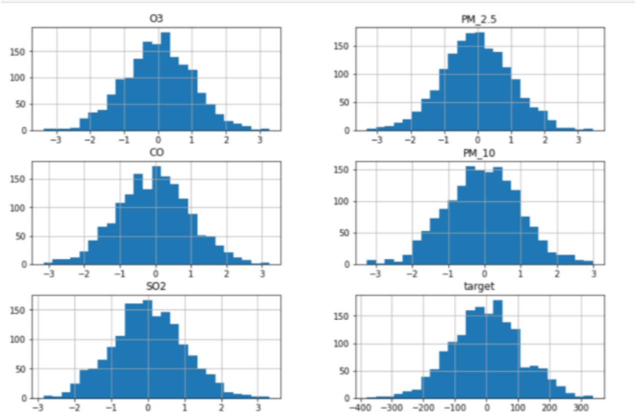
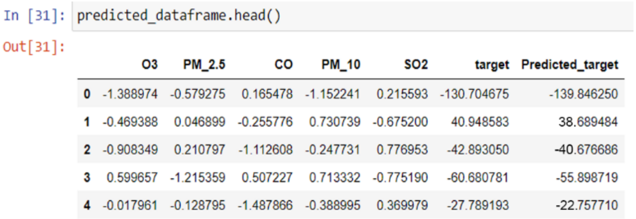


Figure 5: Histogram

IX. OUTPUT

We have predicted that the major pollutants PM 2.5 and PM 10 tries to form a linear graph. After the prediction of the testing set, we have evaluated our system with best accuracy and loss against the outliers, as mentioned in Fig 7.



```
In [31]: predicted_dataframe.head()
Out[31]:
         O3    PM_2.5        CO     PM_10       SO2       target  Predicted_target
0 -1.388974 -0.579275  0.165478 -1.152241  0.215593 -130.704675       -139.846250
1 -0.469388  0.046899 -0.255776  0.730739 -0.675200   40.948583         38.689484
2 -0.908349  0.210797 -1.112608 -0.247731  0.776953  -42.893050        -40.676686
3  0.599657 -1.215359  0.507227  0.713332 -0.775190  -60.680781        -55.898719
4 -0.017961 -0.128795 -1.487866 -0.388995  0.369979  -27.789193        -22.757710
```

Figure 6: Predicted Output

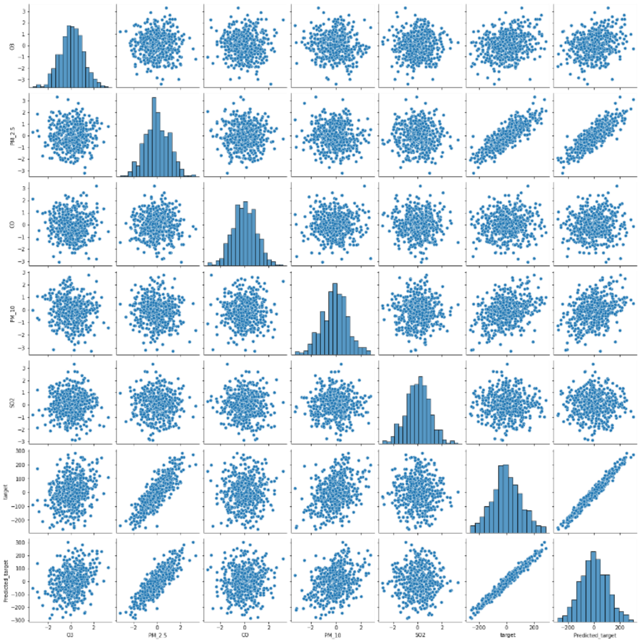


Figure 7: Visualisation of the Predicted Model

X. DECISION RESULTS

The decision results are given based on the standard quality metrics. Here, the analysis of air quality data in India for post and pre corona effects on pollution created by vehicles and industries. The final conclusion is that this pandemic has a decent effect on air pollution. Since the industries are mostly closed and most of them trying to adopt environment of work from home, which may be a reason we see there is a continuous downfall in pollution levels of the cities. Another reason behind air pollution is through vehicles, which additionally seems to be decreasing as most of the people self quaratined themselves in their houses in order to not get infected by this virus. Another reason would be that majority of the people is ignoring to travel long distance and moving out of their houses just to shop for essential stuff they need.

XI. CONCLUSION AND FUTURE WORK

Since our model is capable of predicting this data with 95% accuracy, as referred in Fig 6, it will successfully predict the upcoming air quality index of any particular data within a given region.With this model we will forecast the AQI and alert the respected region of the country also it a progressive learning model it's capable of tracing back to actual location needed attention provided the statistic data of each possible region needed attention. Firstly, the correlation between air pollutants was found, then the MLR model[8]was trained using the training data set and validated with the unseen test data.

Combining the proposed model with a real time application will help in real time prediction of the pollution bands. We can also improve our proposed model by working with a live AQI which keeps predicting the future AQI values automatically. Therefore, the proposed methodology is suggested to predict AQI and therefore the forecasting of AQI suits for real-time implementation within the future.

REFERENCES

[1] Corbitt, R. A., Standard Book of Environmental Engineering, McGraw-Hill, 1998.
[2] https://ieeexplore.ieee.org/document/8650344/
[3] USEPA, National Ambient Air Quality Standards (NAAQS)
[4] http://ieeexplore.ieee.org/document/8980024/
[5] http://airnow.gov/index.cfm?action=static.aqi
[6] https://ieeexplore.ieee.org/document/8300926/
[7] http://www.atsdr.cdc.gov/general/theair.html
[8] https://ieeexplore.ieee.org/document/8785602/