Chapter 5
Air Pollution Index Prediction: A Machine Learning Approach

Praveen Kumar Maduri, Preeti Dhiman, Chinmay Chaturvedi, and Abhishek Rai

Abstract Nowadays air pollution has become a major problem for our present generation. Many research and innovations are done in order to deliver fresh and clean air and the first and the most important step in this process is the prediction and monitoring of air pollution index. For accurate prediction of air pollution in surroundings many advanced models are developed taking from different machine learning models like logistic regression model, auto-regression model and many others to the several deep learning models. In this paper, we have proposed three different machine learning models LightGBM, GBM and random forest to compare them on the basis of 21 different physical parameters like humidity, temperature, traffic volume on roads for calculating Air quality index value and their dependency on determining the AQI value. The developed model was found to be more accurate as compared to previously developed models in predicting the level of contamination in surrounding air.

5.1 Introduction

With the increasing industrialization and urbanization, the area of forest-covered land is decreasing which is directly affecting our environmental condition and creating an imbalance in nature. Surrounding air is one of these influenced natural resources which is badly contaminated by different human activities. According to the world health organization, every nine out of ten people living on earth breathe contaminated air which directly or indirectly harms their body organs. The first and the foremost step in solving this issue is the accurate prediction of Air quality index of surrounding air and also determining the parameters affecting it most.

The proposed paper focuses on determining the Air Quality Index (AQI) value using three different machine learning models. For training the model a data set is imported which comprises 21 different parameters like traffic volume, temperature,

P. K. Maduri · P. Dhiman · C. Chaturvedi (✉) · A. Rai
Galgotias College of Engineering and Technology (AKTU), Greater Noida, India
e-mail: chinmaychaturvedi76@gmail.com

S. Yadav et al. (eds.), *Proceedings of Second International Conference in Mechanical and Energy Technology*, Smart Innovation, Systems and Technologies 290,
https://doi.org/10.1007/978-981-19-0108-9_5

wind direction, humidity, clouds, visibility, snow, month, year, etc. Based on the values from the available data set the three machine learning models are trained and compared by analyzing the error generated between actual value and predicted air quality index values. Out of these three models, one model is selected which produces least error. Once the models are compared and selection is done then the effect of the parameters on influencing the AQI value is calculated and plotted. After all these compilation processes the AQI values are predicted for every month in a year. The contribution of the given paper is given below:

- The proposed paper aims to determine the best decision tree-based model out of three different models which are Light GBM, GBM and Random Forest for value prediction purposes.
- The model also aims to predict the air quality index for surrounding air based on different physical parameters.

5.2 Literature Review

In the major developing countries, the monitoring and prediction of air pollution levels are becoming a difficult task and in order to solve this problem, many smart and advanced technologies are being developed and adopted by researchers like implementation of IoT (Internet of things) based smart real-time monitoring system and Machine learning or Artificial intelligence-based system for accurate and efficient prediction of air quality index.

Till now many machine learning models are being developed and deployed in different parts of the world for efficient forecasting of pollution levels in our surroundings. Particulate matter, which is the major component of smog, needs to be accurately predicted for which a combo of 1-D (one dimensional) CNN and multi-dimensional bi-directional LSTM model were developed for guessing the PM 2.5 level [1] in the air. Other than this a machine learning model based on regression and auto-regression models is developed for determining the level of small pollutants in air [2]. Adaptive neuro-fuzzy interface is another model which is deployed for predicting the PM particle level in the environment [3]. Other than particulate matter there are many other harmful gases that are responsible for contaminating surrounding air. Sulfur dioxide is a harmful gas that is generally emitted from power stations and by burning fossil fuels in large quantities. In order to monitor the Sulfur dioxide level, an orthogonal decomposition and machine learning model was proposed which is called parameterized non-intrusive reduced order model for reducing the pollutant transport equation [4]. Along with oxides of Sulfur, there are many other gases that are present in impure air like ground-level ozone, oxides of nitrogen, oxides of carbon such as carbon monoxide, carbon dioxide and others. For predicting these harmful components of polluted air, a classification algorithm was developed in which outputs from different sources were combined like chemical components, meteorological forecast output from different sources for training the model and predicting the air impurity level [5]. Real-time pollution monitoring is another big step in determining air contamination levels in the environment. For such different combinations of hardware and software were set up for efficient prediction of impurity. Image processing is one of the most efficient ways to predict pollution levels by analyzing the images and videos of gases emitting from chimneys of industries [6] and then using different classification programs for guessing the contamination level of air. Implementation of Internet of things with a machine learning model is another potential way of real-time monitoring the degree of impurity in air. Use of different microcontroller-based sensing devices like humidity and temperature sensor along with air pollution sensing modules for recording the real-time data of different polluting components found in air and analyzing it using a proper machine learning model [7] is another way of pollution monitoring of air. Another technique is using previously recorded data along with sensor data simultaneously and a machine learning model to depict the air quality value [8]. But after all these developments there is a need to deploy a proper framework that accurately predicts the AQI (Air quality index) for a large dataset and with a proper machine learning model. The proposed paper has implemented three different machine learning models to predict the AQI value and also has suggested the best model out of three models.

5.3 Flow Chart

See Fig. 5.1.

5.3.1 Methodology

In this paper, we have proposed a machine learning model for predicting the Air Quality index using three different frameworks which are Light GBM, GBM and Random Forest. For value prediction, a proper data set is imported consisting of 21 different parameters and based on these data the AQI value is determined. While after analyzing the data set few data variables are found to be of no use which is rain ph., snow ph. and weekend and thus are removed from the data set. The used dataset has different physical parameters which are recorded from the year 2012 to 2017 as shown in Fig. 5.2 for training the model.

The data set used in the paper is extracted and divided into two parts:

- Train dataset: The train data set compressed 80% of data from overall data set.
- Test dataset: the test data set consists of 20% of the total available data.
- The used data set consists of a balanced set of data which can be visualized from the graph shown in Fig. 5.3.

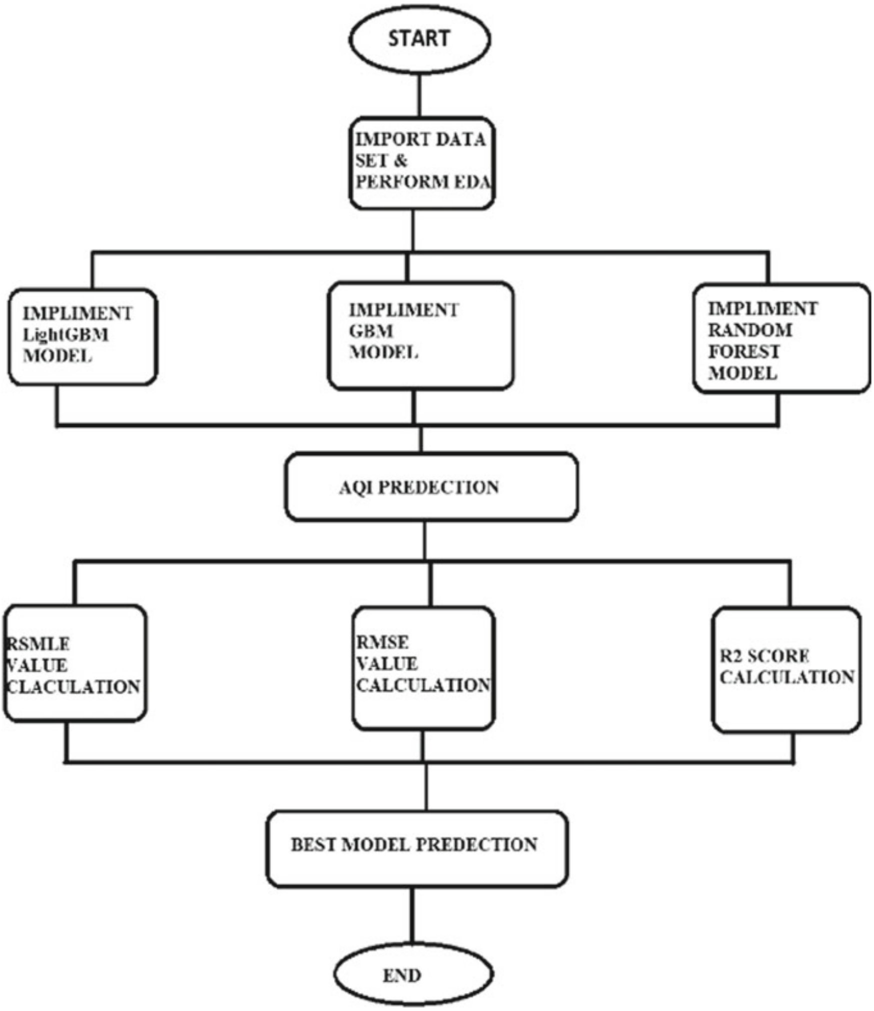
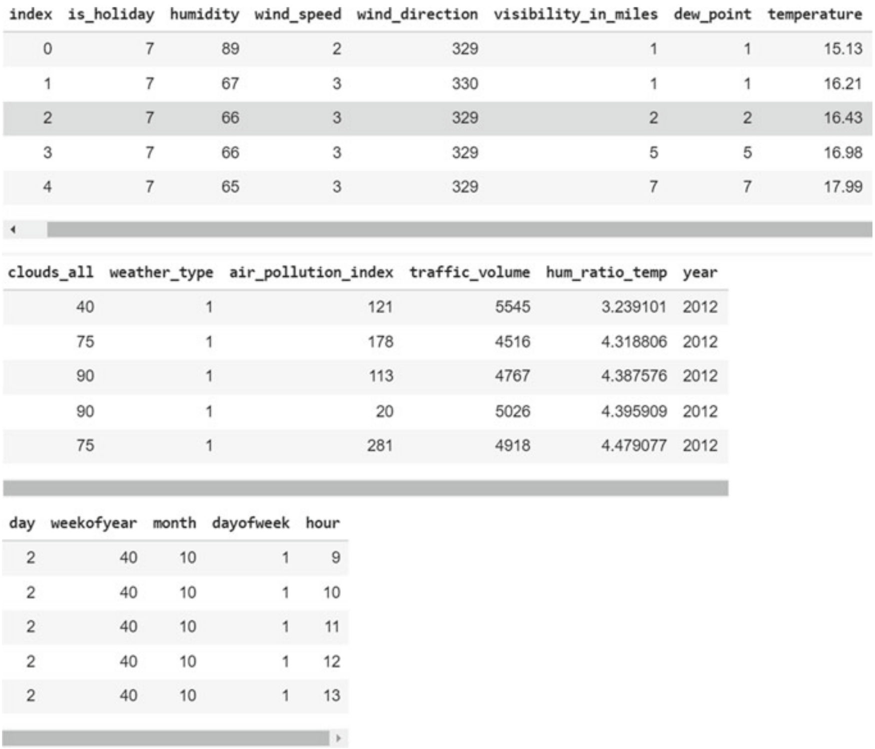


Fig. 5.1 Flow chart

5.3.2 Data Set

The used data set consists of total of 22 parameters out of which 21 are independent parameters which are date-time, holiday, humidity, wind speed, wind direction, visibility in miles, dew point, temperature, rain, snow, clouds, weather type, traffic volume, humidity and temperature ratio, year, day, week of the year, month, day of week, weekend and hour and one is dependent parameter which is Air quality index. The parameters in used data set consist of both integer and string values. The parameters like holiday consist array of different holidays falling in a year such as



index	is_holiday	humidity	wind_speed	wind_direction	visibility_in_miles	dew_point	temperature
0	7	89	2	329	1	1	15.13
1	7	67	3	330	1	1	16.21
2	7	66	3	329	2	2	16.43
3	7	66	3	329	5	5	16.98
4	7	65	3	329	7	7	17.99

clouds_all	weather_type	air_pollution_index	traffic_volume	hum_ratio_temp	year
40	1	121	5545	3.239101	2012
75	1	178	4516	4.318806	2012
90	1	113	4767	4.387576	2012
90	1	20	5026	4.395909	2012
75	1	281	4918	4.479077	2012

day	weekofyear	month	dayofweek	hour
2	40	10	1	9
2	40	10	1	10
2	40	10	1	11
2	40	10	1	12
2	40	10	1	13

Fig. 5.2 Used data set

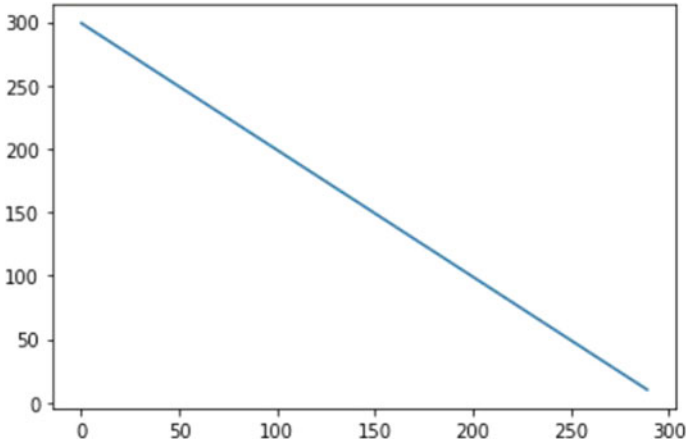


Fig. 5.3 Balanced dataset representation

Christmas Day, Independence Day, labor day, veteran's day, etc. while parameter weather type consists of cloudy, clear, rain, drizzle, fog, etc. Other than these two other parameters contain integer values depending on their type.

For more precise understanding of parameters in dataset different analysis is performed before implementing the models. Being a multivariate regression problem where one variable is depending on multiple independent variables a multivariate analysis is performed to understand how the variables are related to each and to get a view of how they behave with respect to each other.

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{17} x_{17} \tag{5.1}$$

In Eq. (5.1) the dependency between independent and dependent variables is shown where Y represents the dependent variable which is Air Quality Index and β_0, β_1 ..., β_{17} represent the coefficient of dependency while x_0, x_1 ..., x_{17} are independent parameters like humidity, temperature, etc.

To get insight into general distribution of dataset variables histogram distributional features plot is used to understand the data distribution of each used variable. The correlation between the variables is shown using heat map plot where the positive values represent that on changing the value of one feature the other feature will tend to follow the same order, i.e., on increasing the value of one variable the value of other will increase and on decreasing the value of one variable the value of other will also decrease. The negative value indicates the reverse of value, i.e., on increasing the value of one variable the value of other variables will increase and vice-versa.

5.3.3 *Used Frameworks*

The described machine learning model uses three different frameworks which are LightGBM, GBM and random forest.

1. **LightGBM**: This is a tree-based algorithm that uses gradient boosting framework for working. This is designed to provide faster training speed with low memory usage. Unlike other algorithms in place of growing trees depth-wise, it grows leaf wise which tends to achieve lesser loss as compared to previous ones.
2. **GBM**: Gradient Boosting machine constructs a forward level-wise additive model with the help of gradient descent in function space.
3. **Random Forest**: As the name suggests the random forest consists of a large number of decision trees that operate as a multiple learning algorithms to obtain better predictive performance (Figs. 5.4, 5.5 and 5.6).

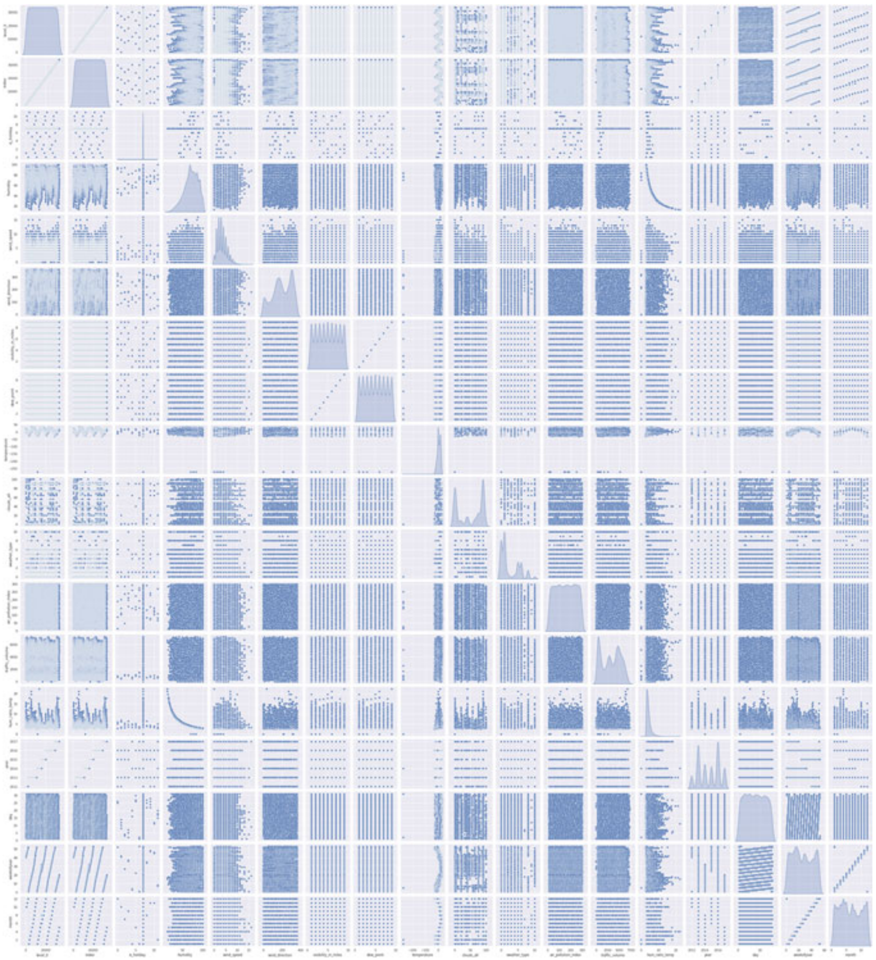


Fig. 5.4 Multivariate analysis chart

5.4 Working

The overall programming of the model is divided into two sections

a. Prediction of AQI value
b. Error calculation for actual value and predicted value of each model.

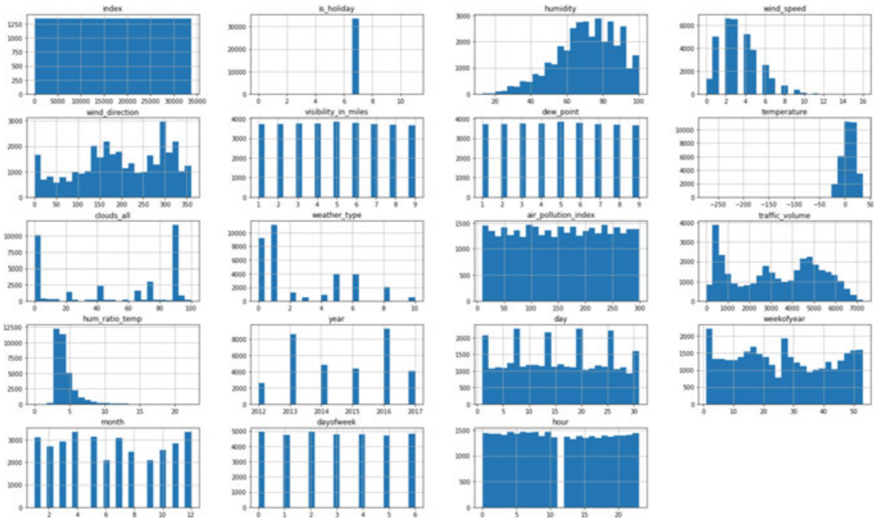


Fig. 5.5 Dataset variable distribution chart

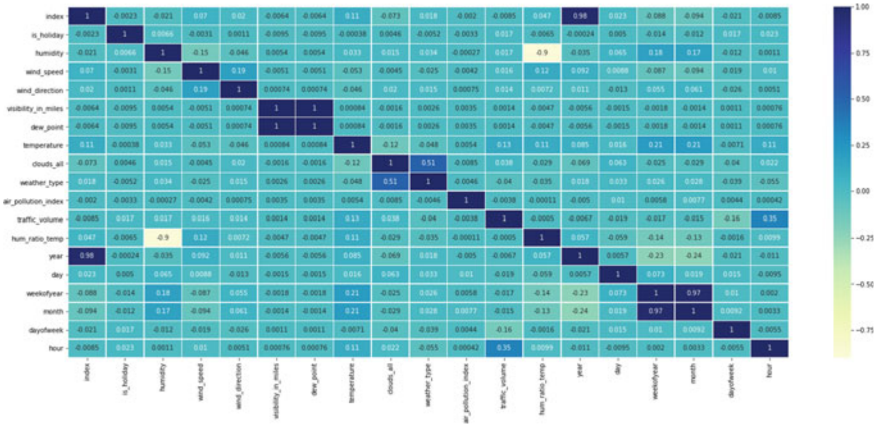
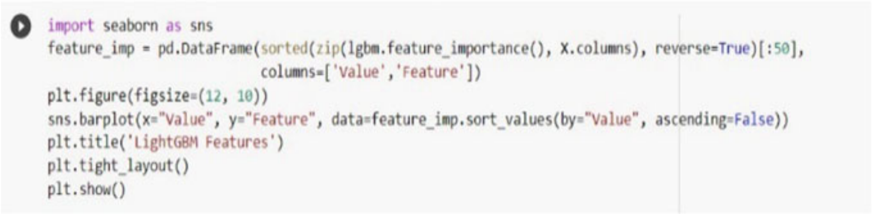


Fig. 5.6 Correlation heat map

5.4.1 Prediction of AQI Value

The model proposed here is a logistic regression problem that is solved using three different decision tree-based frameworks namely LightGBM, Gradient Boosting Machine (GBM) and random forest. The model is trained using 80% of the total data set. At the starting, the dataset is imported and Null values are calculated for data cleaning process and then after the string values present in the dataset are converted into integer values. Once the data processing is done the three frameworks



```
import seaborn as sns
feature_imp = pd.DataFrame(sorted(zip(lgbm.feature_importance(), X.columns), reverse=True)[:50],
                           columns=['Value','Feature'])
plt.figure(figsize=(12, 10))
sns.barplot(x="Value", y="Feature", data=feature_imp.sort_values(by="Value", ascending=False))
plt.title('LightGBM Features')
plt.tight_layout()
plt.show()
```

Fig. 5.7 Implementation of seaborn library for parameter dependency in Light GBM

are imported. Once the framework is imported then the feature dependency is calculated using seaborn library as shown in Figs. 5.7 and 5.9 and respective graph is also plotted as in Fig. 5.8 and fir10 for determining which parameter is largely responsible for determining the AQI value. After compiling all these steps then by using a test dataset the values of Air quality are predicted (Fig. 5.10).

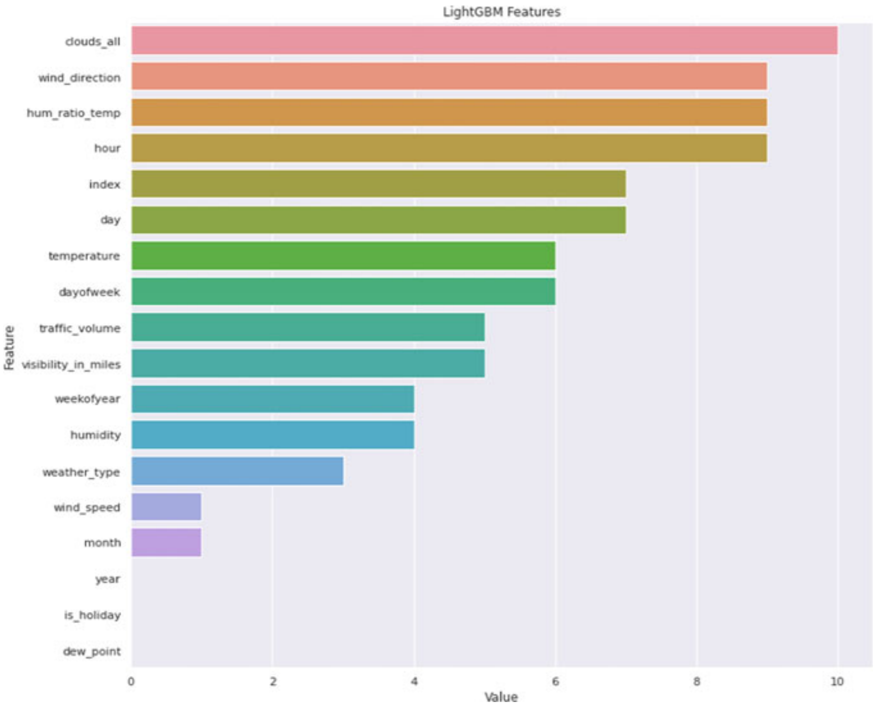
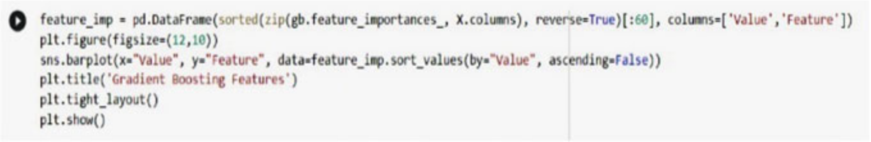


Fig. 5.8 Dependency of parameter for Light GBM model



```
feature_imp = pd.DataFrame(sorted(zip(gb.feature_importances_, X.columns), reverse=True)[:60], columns=['Value','Feature'])
plt.figure(figsize=(12,10))
sns.barplot(x="Value", y="Feature", data=feature_imp.sort_values(by="Value", ascending=False))
plt.title('Gradient Boosting Features')
plt.tight_layout()
plt.show()
```

Fig. 5.9 Implementation of seaborn library for parameter dependency in GBM model

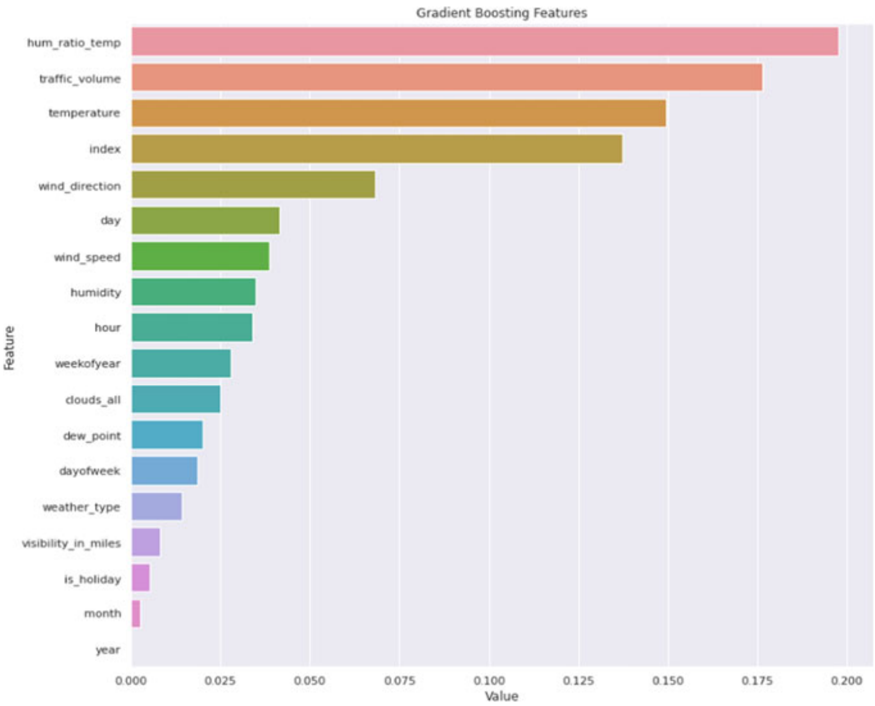


Fig. 5.10 Dependency of parameters for GBM model

5.4.2 *Error Calculation for Each Model*

After predicting the AQI values using three models the data comparison is done between predicted and actual value and three different matrices are used to calculate the performance of each used model which are root mean square error, root mean squared logarithmic error and R2 score.

$$\text{RSMLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(P_i + 1) - \log(A_i + 1))^2} \tag{5.2}$$

The formula written in Eq. 5.2 is used to calculate RMSLE where P_i is predicted value, A_i is actual value and n is total observation in dataset and $\log(y)$ is the natural log of y.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2} \tag{5.3}$$

The formula written in Eq. 5.3 is used to calculate root mean square error (RMSE) where P_i is predicted value, A_i is actual value and n is the total number of observations.

$$R2\,\text{score} = 1 - \frac{\text{Unexplained Variation}}{\text{total variation}} \tag{5.4}$$

The $R2$ score shown in Eq. 5.4 is used to calculate the measure of how well the observations are reproduced by the model.

For each model, the three matrices are calculated and out of all three models, the one which produced lower value is considered to be best model for AQI prediction.

5.5 Results

The proposed model is able to produce the promising results as the available dataset and is able to give the nearly accurate values accordingly. The proposed model is able to give the following results.

1. Prediction of Accurate AQI value.
2. RMSLE, RMSE, R 2 score calculation for each of three models.

5.5.1 Prediction of AQI Values

See Fig. 5.11.

5.5.2 Error Calculation

Once the predicted data and actual data is compared and matric error values are calculated using Eqs. 5.2, 5.3 and 5.4 and mean of error generated from each matric equation is calculated and plotted separately using bar chart as shown in Figs. 5.12, 5.13 and 5.14 and respective values are shown in Table 5.1. After plotting of graph of mean error value for each of three matrices the model which is able to generate

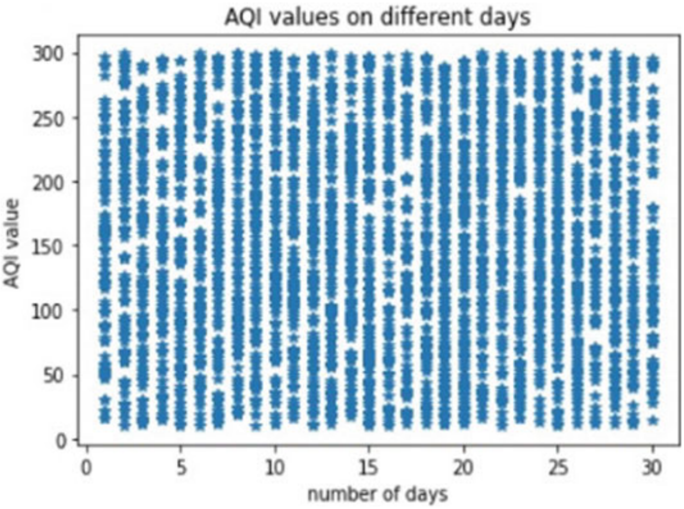


Fig. 5.11 AQI values for different days of a month

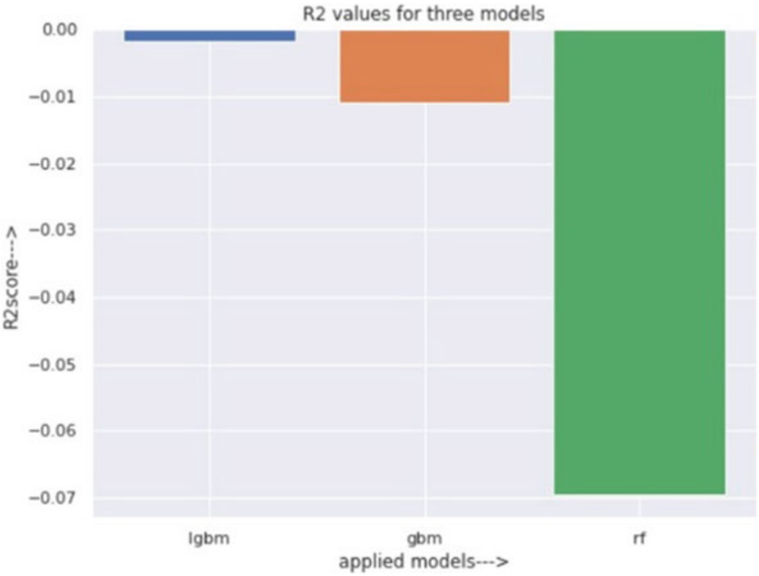


Fig. 5.12 Mean *R*2 score value representation

lesser error value is found to be the best model for the AQI prediction from the used data set.

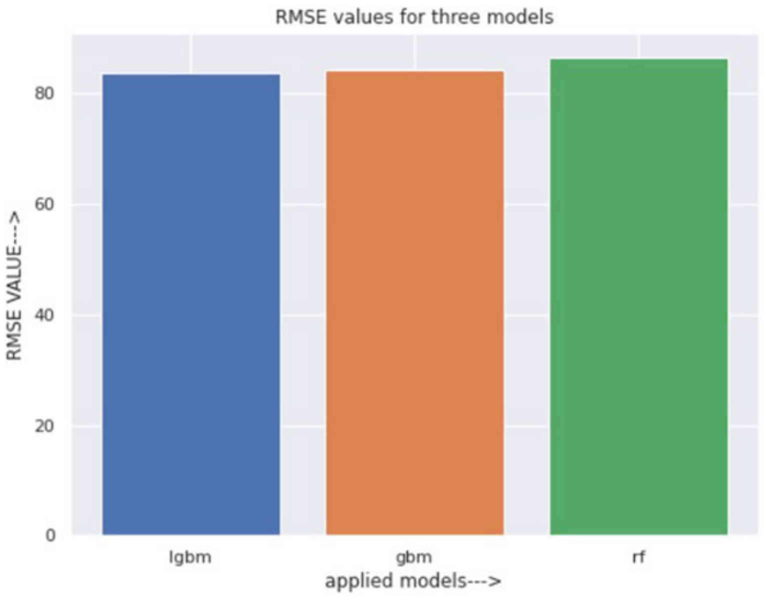


Fig. 5.13 Mean RMSE value representation

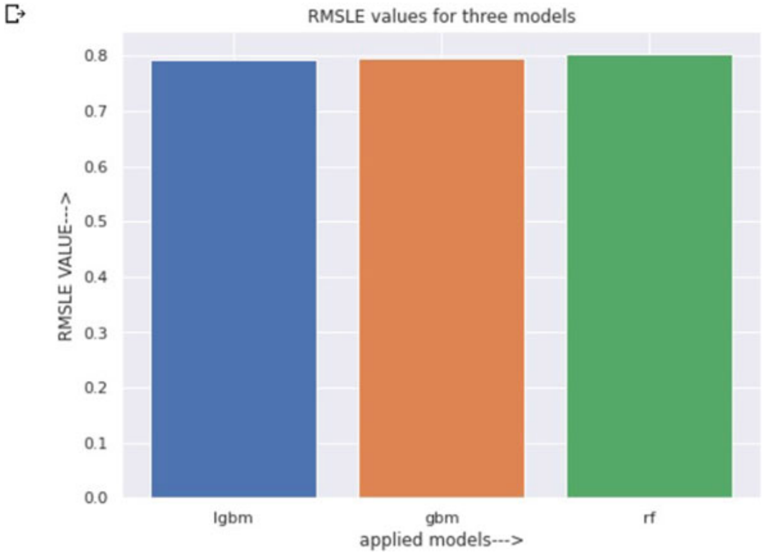


Fig. 5.14 Mean RMSE value for each applied model

Table 5.1 Model and their respective error metric value

Model/matrices	RMSLE	RMSE	*R*2 Score
Light GBM	0.7939	83.79	−0.0071
GBM	0.7953	84.19	−0.0110
Random Forest	0.8045	86.59	−0.6951

5.6 Conclusion

From the proposed machine learning model following conclusions can be drawn out.

1. The proposed model is able to predict the accurate air quality index value (AQI).
2. The best-suited framework for the used dataset is LightGBM model which in comparison to other two models GBM and Random Forest produces lesser error and is found to be most accurate framework for AQI value prediction for the used data set.

5.7 Future Scope

In order to develop more accurate machine learning models for AQI prediction, there is a need to develop a more powerful framework that is faster in the compilation process and also able to predict nearly exact value or produce very less error while undergoing the prediction process. Moreover, there is a need to implement the most accurate model in real-life scenarios which can predict the future values and alert the surrounding about the coming hazard of air pollution.

References

1. Du, S., Li, T., Yang, Y., Horng, S.-J.: Deep air quality forecasting using hybrid deep learning framework. IEEE Trans. Knowl. Data Eng. **33**(6), 2412–2424 (2021). https://doi.org/10.1109/TKDE.2019.2954510
2. Aditya, C.R., et al.: Detection and prediction of air pollution using machine learning models. Int. J. Eng. Trends Technol. (IJETT) **59**(4) (2018)
3. Mihalache, S.F., Popescu, M., Oprea, M.: Particulate matter prediction using ANFIS modelling techniques. In: 2015 19th International Conference on System Theory, Control and Computing (ICSTCC), pp. 895–900 (2015). https://doi.org/10.1109/ICSTCC.2015.7321408
4. Xiao, D., et al.: Machine learning-based rapid response tools for regional air pollution modelling. Atmos. Environ. **199**, 463–473 (2019)
5. Xi, X., et al.: A comprehensive evaluation of air pollution prediction improvement by a machine learning method. In: 2015 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). IEEE (2015)
6. Gacquer, D., et al.: Comparative study of supervised classification algorithms for the detection of atmospheric pollution. Eng. Appl. Artif. Intell. **24**(6), 1070–1083 (2011)

7. Ayele, T.W., Mehta, R.: Air pollution monitoring and prediction using IoT. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE (2018)
8. Srivastava, C., Singh, S., Singh, A.P.: Estimation of air pollution in Delhi using machine learning techniques. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE (2018)