Air Quality Prediction System using LightGBM

Narayani Sunil Pillai[1], Mythili S[2], Meera S Nair[3], Diliya M Khan[4]

[1,2,3]Computer Science and Engineering (Pursuing), Dept. of Computer Science and Engineering, LBS Institute of Technology for Women, Thiruvananthapuram, Kerala, India.
[4]Assistant Professor, Dept. of Computer Science and Engineering, LBS Institute of Technology for Women, Thiruvananthapuram, Kerala, India.

---***---

Abstract - *One major fundamental right is clean air which is integral to the idea of citizenship and it is without a doubt, the responsibility of each citizen to do his/her part to keep the air clean. Air quality forecasting has been looked into as the key solution of early warning and control management of air pollution. In this paper, we propose an air quality prediction system based on a machine learning framework called LightGBM model, to predict the air quality in Trivandrum, Kerala, 24 hours in advance. This model, trained using LightGBM classifier, takes weather forecasting data as one of the data sources for predicting the air quality thereby increasing the prediction accuracy by making full use of available spatial data. The existing air quality monitoring stations and satellite meteorological data provides real-time air quality monitoring information which is used to predict the trend of air pollutants in the future. The proposed system was found to give an accuracy of 98.38%.*

Key Words: **Air Quality Index, Air Quality Forecasting, LightGBM, Time Series Analysis, Machine Learning**

1. INTRODUCTION

Human actions reduce the quality of air on a big scale, traffic and industrial activities release huge amounts of air pollutants into the air. Continuous inhaling of these toxic matter leads to major respiratory diseases and even death. Therefore, air quality forecasting systems become indispensable. The air quality index (AQI) is an index for reporting the air quality of a locality on a daily basis, it can be considered to be a measure of how air pollution can affect a person's health within a short time period. By getting updates of AQI, people can understand how the local air quality can affect their health. The higher the value of AQI, the greater the risk factor with high level of air pollution and greater the health concerns. India follows the 500 point scale to report air quality where rating between 0 and 50 is considered good and between 300 and 500 is deemed hazardous. Generally, AQI values at or below 100 are thought of as satisfactory whereas when the AQI values are above 100, air quality is unhealthy.

In this paper, we propose an approach to predicting the quality of air in Trivandrum24 hours in advance based on the LightGBM model; a machine learning algorithm, by using hourly data of both historical air quality data and meteorological data within the time span of three years from the Central Pollution Control Board (CPCB). This model takes the weather forecasting data as one of the data sources for predicting the air quality thereby increasing the accuracy. The Classifier model upon testing, classifies the air quality into good or bad. The classifications are fairly close to the testing set. The results show that the proposed method improves prediction performance.

The rest of the paper is organized in the following way. Section II discusses the related works. Section III describes the methods and materials. Section IV describes the experimental results and Section V concludes the paper.

2. BACKGROUND AND RELATED WORK

A number of machine learning algorithms are used for predicting the quality of air worldwide. When it comes to air quality systems, integrating the correlation of temporal and spatial features and mining the local similarity and regional interaction is an important tactic which is discussed by by Guyu Zhao, Guoyan Huang, Hongdou He and Qian Wang [2019][1]. The major pollutants responsible for air pollution are sulphur dioxide (SO2), nitrogen dioxide (NO2), particulate matter (PM), carbon monoxide (CO) and ozone (O3). By using data mining techniques like linear regression and multilayer perceptron, Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi and Yash Navoria [2017][2] analysed the present trends in air pollution in Delhi and made prediction about the future. Out of these pollutants, PM2.5 is said to be the most hazardous factor. Qing Tao, Fang Liu, Yong Li and Denis Sidorov [2019][3] proposed a short-term forecasting model based on deep learning and the convolutional-based bidirectional gated recurrent unit (CBGRU) method to predict the concentration of PM2.5. Dongming Qin, Jian Yu, Guojian Zou, Ruihan Yong, Qin Zhao and Bo Zhang [2019][4] proposed a pollutant concentration prediction method by integrating big data by using two kinds of deep networks which predicts future particulate matter (PM2.5) concentrations as a time series. Shengdong Du, Tianrui Li, Yan Yang and Shi-Jinn Horng [2019][5] proposed a new deep learning model for air quality (mainly PM2.5) forecasting, which learns the spatial-temporal correlation features and interdependence of multivariate air quality related time series data by hybrid deep learning architecture. By testing it on two real-world air quality datasets, the experimental results indicated good forecasting ability. Pratyush Singh, Lakshmi Narasimhan T and Chandra Shekar Lakshminarayanan [2019][6] proposed a solution to predict the concentration of PM2.5 by considering the real time data obtained from CPCB at the Indira Gandhi

International Airport Area based on Long Short-Term Memory (LSTM) networks, which are deep neural networks that are known to perform well on sequential prediction problems. The RMSE achieved from this system was less than 1.2 for each of the next 23 hours of prediction from a given time instant. Different classification and regression techniques like Linear Regression, SDG Regression, Random Forest Regression, Decision Tree Regression, Support Vector Regression, Artificial Neural Networks, Gradient Boosting Regression and Adaptive Boosting Regression were implemented by Chavi Srivastava, Shyamli Singh and Amit Prakash Singh [2018][7] to predict the levels of pollutants like PM2.5, PM10, CO, NO2, SO2 and O3 at three monitoring sites located in different districts of New Delhi using meteorological features like wind speed (WS), vertical wind speed (VWS), wind direction (WD), temperature (Temp) and relative humidity (RH) as input parameters. Mean Square Error (MSE), Mean Absolute Error (MAE) and R2 were used for evaluating the regression techniques for building the model. Considering the overall performance, SVR and Neural Networks (MLP) were found to give the least errors in estimation and provided maximum accuracy with fair-low range of errors. A device is introduced by Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth and Hari Kiran Reddy [2020][8] that can predict the future data of pollutants by taking the past and present pollutant data into account. The sensed data is saved in an Excel sheet for further evaluation. The sensors are used on Arduino Uno platform to collect the pollutant data. Different machine learning algorithms like linear regression, Decision Tree and Random Forest were used to predict the air quality index by using linear regression algorithm which is used to perform the regression task. From the results, it was concluded that the Random Forest algorithm gives better prediction of air quality index. The air quality of India is predicted by A.Gnana Soundari, Akshaya A.C and J.Gnana Jeslin [2019][9] using a Gradient descent boosted multivariable regression problem. For this prediction, historical data of pollutant concentration is provided. With this model, some knowledge about the data are extracted using various techniques to obtain the heavily affected regions on a cluster. 96% accuracy was obtained on prediction of the air quality index of whole India. But in data training, after a series of processing, the dimension of the data would increase to more than one million. Hence, the training model must have a faster training rate, relatively lower memory cost, and higher accuracy, which is what was proposed by Ying Zhang, YanhaoWang, Minghe Gao, Qunfei Ma, Jing Zhao, Rongrong Zhang, Qingqing Wang and Linyan Huang [2019][10], a system using LightGBM framework for predicting the quality of air in future by making use of 3 real-time datasets. LightGBM is prefixed as 'Light' because of its high speed. This model was found to have high accuracy.

3. MATERIALS AND METHODS

3.1 Data Sources and Classification Standards

The entire experiment is carried out using the dataset collected from the Government official website named Central Pollution Control Board (CPCB).

Selection of Experimental dataset:

- Dataset of Trivandrum
- Real-time AQI values
- 24313 Records from 23.06.2017 to 31.03.2020

AQI value is integrated of nearly eight pollutants mainly Particulate Matter PM10, PM2.5, Nitrogen dioxide (NO2), NO, NOx, Sulphur dioxide (SO2), Carbon monoxide (CO), Ozone (O3). The AQI ranges are classified as six values (1 to 6). 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.

3.2 Dataset

A total of three datasets are used in this paper: air quality dataset, meteorological dataset and meteorological forecast dataset. The first two are the hourly data of pollutants and meteorological factors respectively, over the past three years in Trivandrum collected from the Central Pollution Control Board (CPCB) which are used for training the model. The pollutant parameters considered are PM10, PM2.5, NO, NO2, NOx, Ozone, SO2, CO and the meteorological factors include Atmospheric temperature (AT), Barometric pressure (BP), Wind direction (WD), Wind speed (WS) and Relative humidity (RH), with a total of 24,313 rows in each of these datasets.

The third dataset is the hourly meteorological forecast data of any random day of Trivandrum other than those used in training the model collected from the CPCB which is used in testing the model and not in the training process. The parameters considered in this dataset are same as those in the meteorological dataset with a total of 24 rows in it.

3.3 LightGBM

Light Gradient Boosting Method or LightGBM is a gradient boosting framework that utilizes tree based learning algorithm. LightGBM grows trees vertically while other algorithms grow trees horizontally. The leaf with max delta loss will be chosen to grow. Leaf-wise algorithm is capable of reducing more loss than a level-wise algorithm when growing the same leaf. LightGBM can handle the large data size and takes lower memory to run.

3.4 Proposed Work

The proposed system aims at predicting the air quality of Trivandrum using the LightGBM model. The model is trained with the statistical features of the historical air quality data and meteorological data collected over the past three years. By providing the weather forecast data of any particular day and pollutant data of any nearby day, we can predict the air quality of that day.

The modules of the system are as follows:-

1) Data pre-processing. The datasets are checked for any missing values and are replaced. This dataset is then visualised using matplotlib library to understand the trend of the data over time. Savitzky-Golay filter is applied for the purpose of smoothing the data. The datasets obtained after the application of the Savitzky-Golay filter are the pre-processed datasets.

2) Sliding window mechanism. The use of preceding time steps to predict the next time step is called the sliding window method. By this method we constructed high-dimensional temporal features for improving the prediction accuracy. We calculated summary statistics across the values in the sliding window and included these as features in our dataset. The most useful was the mean of the previous few values, also called the rolling mean. A rolling () function is provided by Pandas that creates a new data structure with the window of values at each time step. We then performed statistical functions on the window of values collected for each time step, such as calculating the mean. First, the series had to be shifted. Then the rolling dataset was created and the mean values calculated on each window of two values. We used the window size as 4 in our project.

3) Feature integration. Feature integration is to deeply mine the features that would significantly influence the predictors to provide the prediction model. Four characteristics were dealt with, i.e., the air quality feature, the meteorological feature, the predictive data feature, and the statistical feature.

a. Air quality feature. Different parameters of historical air quality data whose increased concentration in the atmosphere affects the air quality adversely were considered. The parameters are PM2.5, PM10, NO, NO2, NOx, Ozone, SO2 and CO.
b. Meteorological feature. Since the weather parameters such as atmospheric temperature, barometric pressure, wind speed, wind direction and relative humidity have an impact on air quality, they could have an influence over the variation of the pollutant concentration.
c. Predictive data feature. The weather forecast data of a particular day was used as the predictive data feature that includes the atmospheric temperature, barometric pressure, wind direction, wind speed and relative humidity of 24 hours. This feature is used for the testing purpose later on in the system.
d. Statistical feature. Based on the above features, we then constructed the statistical features such as mean and standard deviation for the air quality and the meteorological dataset and combine them into a single dataset for training and testing of the model. Labelling of data as 0 and 1 is also done on every observation after comparison with the ground truth values.

4) Dataset splitting. The new dataset is split into two, training set and testing set. The splitting is done in a 75-25 ratio. 75% of the dataset is taken as the Training Set which is used to train the model. The remaining 25% becomes the Test Set which is used to test the model, to analyse its accuracy. The testing set is never used for training, which could otherwise lead to overfitting the model.

5) Model training and testing. The LightGBM model is trained by fitting the training set to the LightGBM classifier model. The classifier model upon testing, classifies the air quality into good or bad. The classifications are fairly close to the testing set.

6) Prediction. In order to predict the air quality of a day, the meteorological forecast data of that day is provided. Air quality data of any nearby day is also provided. These data were pre-processed, sliding window is applied, statistical features of the two datasets are extracted and it is used to predict the air quality of that desired day.

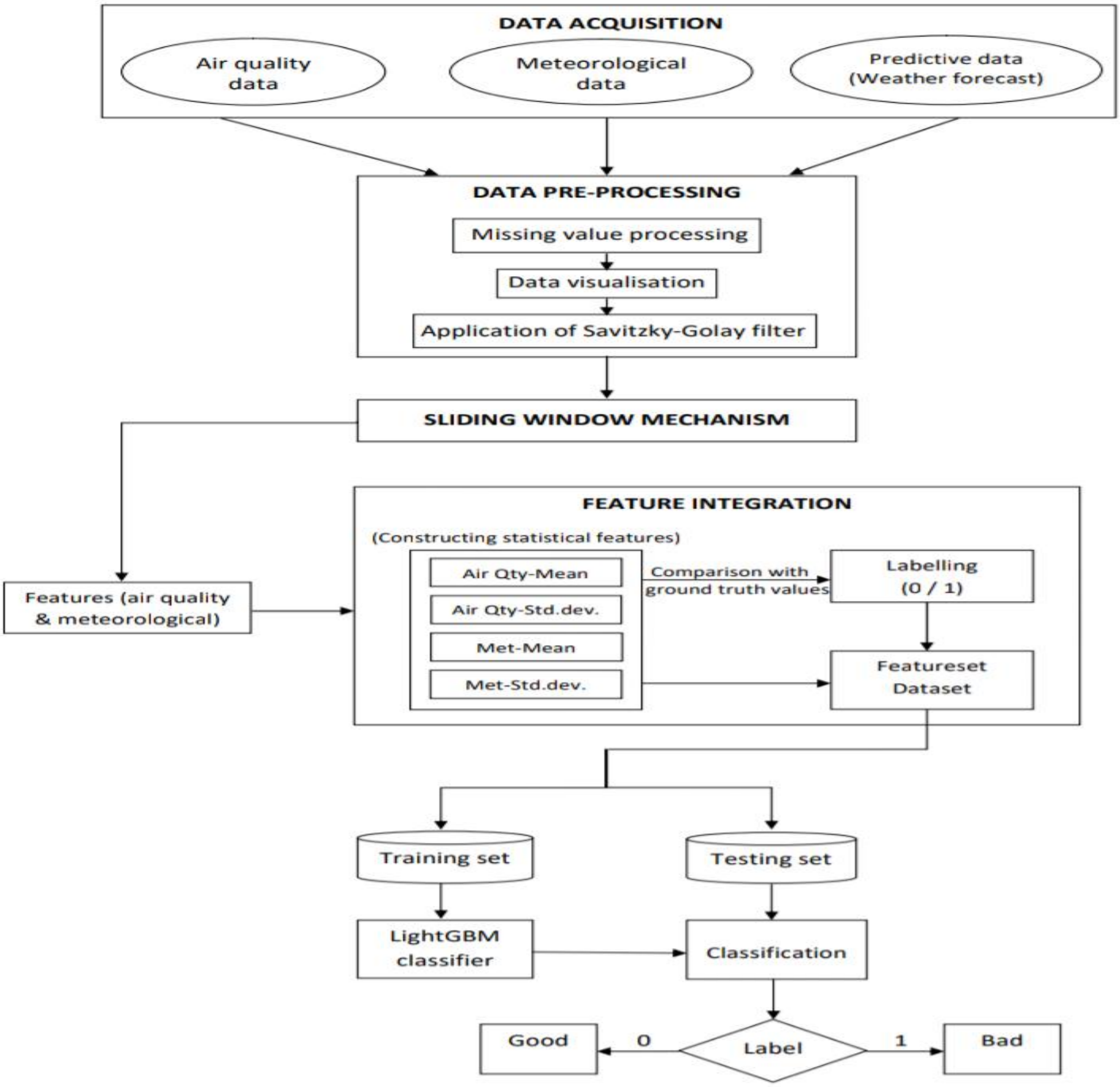

Fig -1: Dataflow of the model

4. RESULTS

Figure 2 shows the feature integration of all the required features into a dataset, which is then split in the 75-25 ratio to train and test the model. An accuracy of 98.38% was obtained. We used the Coefficient of Determination, denoted as r^2 to get the regression score. The r2 score varies between 0 and 100%. It is a statistic used in the conditions of statistical models for either the prediction of future outcomes or the testing of hypotheses, on the basis of other connected information. The two variables are perfectly related, that is, there is no variance at all if it is 100%.

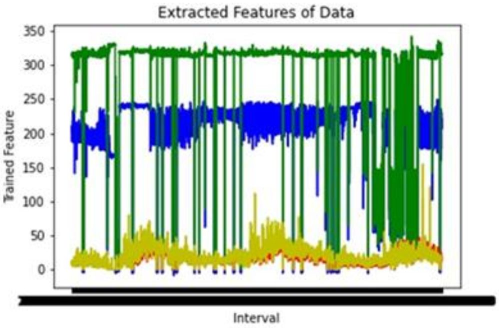

Fig -2: Extracted Features of Data

Figure 3 depicts the learning curves of the model. In this graph, we can see that that the plot of training loss and validation loss decreases to a point of stability. Also, there is only a very small gap between the training loss and validation loss plot. Hence, we can conclude that it shows a good fit.

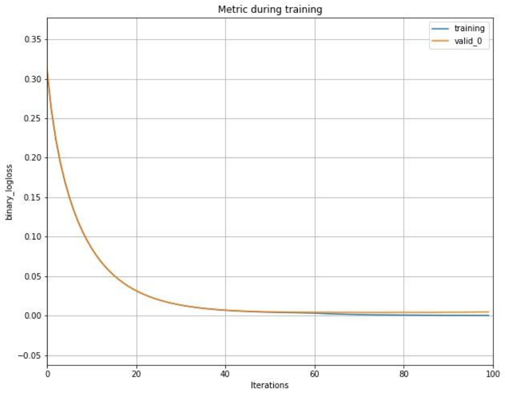

Fig -3: Learning Curves

Table 1 shows the Predicted Air Quality of the date 04-04-2020, as good or bad after giving the air quality data of 02-04-2020 and the meteorological forecast data of 04-04-2020 as input to the model, which thereby predicts the air quality of the locality of Trivandrum 24 hours in advance from the 2nd of April, 2020. The prediction plot of the trained model giving the air quality of the date 04-04-2020 is shown in Figure 4.

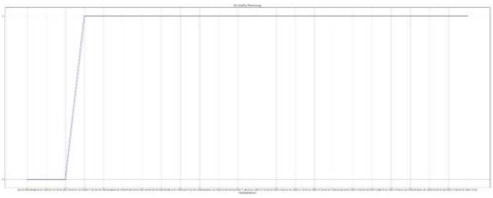

Fig -4: Dataflow of the model

From Date	To Date	Air quality
04.04.2020 00:00	04.04.2020 01:00	Bad
04.04.2020 01:00	04.04.2020 02:00	Bad
04.04.2020 02:00	04.04.2020 03:00	Bad
04.04.2020 03:00	04.04.2020 04:00	Good
04.04.2020 04:00	04.04.2020 05:00	Good
04.04.2020 05:00	04.04.2020 06:00	Good
04.04.2020 06:00	04.04.2020 07:00	Good
04.04.2020 07:00	04.04.2020 08:00	Good
04.04.2020 08:00	04.04.2020 09:00	Good
04.04.2020 09:00	04.04.2020 10:00	Good
04.04.2020 10:00	04.04.2020 11:00	Good
04.04.2020 11:00	04.04.2020 12:00	Good

Table -1: Predicted Air Quality

5. CONCLUSION AND FUTURE WORK

In this paper, we developed an air quality prediction system for the city of Trivandrum obtaining datasets from Pollution Control Board, India. This model can successfully predict the air quality 24 hours in advance with high accuracy. An accuracy of 98.38% was obtained upon testing the model. LightGBM is found to be a suitable framework for prediction of air quality with high accuracy, such that it is a better framework candidate for air quality prediction than the existing models. This model was developed overcoming the limitations of the existing systems, and adapted to the Indian scenario by considering the data from Trivandrum. It can be further developed to predict the ambient air quality of multiple regions together. In future, we can introduce more meteorological factors like precipitation, minimum and maximum temperature, solar radiation, vapor pressure etc. to increase the accuracy of the system. The unclear trend and wide fluctuations of air pollutants is also attributed to the emissions from pollution sources like transportation, industrial emissions etc. Those factors need to be considered as well.

REFERENCES

[1] Guyu Zhao, Guoyan Huang, Hongdou He, Qian Wang."Innovative Spatial-Temporal Network Modelling and Analysis Method of Air Quality"- IEEE Journal, Volume 7, 22 February 2019. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Shweta Taneja,Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria. "Predicting Trends in Air Pollution in Delhi using Data Mining"- IEEE Conference paper, 13 July 2017.

[3] Qing Tao, Fang Liu, Yong Li, Denis Sidorov. "Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU"- IEEE Journal, Volume 7, 7 June 2019.

[4] Dongming Qin, Jian Yu, Guojian Zou, Ruihan Yong, Qin Zhao, Bo Zhang. "A Novel Combined Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration"- IEEE Journal, Volume: 7, 1 February 2019.

[5] Shengdong Du, Tianrui Li, Senior Member, IEEE, Yan Yang, Member, IEEE, Shi-Jinn Horng. "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework"- IEEE, 20 November 2019.

[6] Pratyush Singh, Lakshmi Narasimhan T, Chandra Shekar Lakshminarayanan. "DeepAir: Air Quality Prediction using Deep Neural Network"- 2019 IEEE Region10 Conference (TENCON 2019).

[7] Chavi Srivastava, Shyamli Singh, Amit Prakash Singh. "Estimation of Air Pollution in Delhi Using Machine Learning Techniques". 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE (2018).

[8] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy. "Air Quality Prediction of Data Log by Machine Learning". 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE (2020).

[9] A.Gnana Soundari, Akshaya A.C, J.Gnana Jeslin. “Indian Air Quality Prediction and Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973- 4562 Volume 14, Number 11, 2019.

[10] Ying Zhang, Yanhao Wang, Minghe Gao, Qunfei Ma, Jing Zhao,Rongrong Zhang, Qingqing Wang,Linyan Huang. “A Predictive Data Feature Exploration-Based Air Quality Prediction Approach”-IEEE Journal, Volume 7,8 February 2019.