A Comparative Study on Air Quality Analysis and Prediction Using Machine Learning Techniques

Nandini K[1] , Fathima G[2]

[1] M.E Scholar, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

[2] Professor, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

Email: [1]nandinikannan44@gmail.com

[2]fathima.ace@gmail.com

***Abstract*— Air pollution is a complex mixture of toxic components with considerable impacts on humans. The increase in air pollution is of concern for many urban cities in India and other developing countries around the world. In this paper air pollution analysis and prediction is done using the machine learning techniques. Bengaluru is one of India's fastest growing metropolises and, although benefiting economically due to its rapid development, has a rapidly deteriorating environment. The data sets were collected from government repositories of Bengaluru region i.e Karnataka Pollution Control Board (KPCB). The collected datasets were pre-processed. Pre-processing includes, clustering of datasets using K-Means algorithm. The clustered datasets were further labelled. These labelled datasets were subjected to various machine learning algorithms such as Multinominal logistic regression, Decision tree and Random forest in order to analyze the air pollution. Comparisons were made using the three models to obtain the best fit model for analysis. To predict the next day air pollutant data, normalization is applied upon the datasets. Using the standard deviation and mean value the next day air pollutant data was obtained. The algorithms were compared based on the accuracy. Generally Random forest algorithm gives good result but in this case Decision tree and Multinominal logistic regression have given high accuracy.**

Keywords— ***Air pollution, Machine learning, K-Means, Multinominal logistic regression, Decision tree, Random forest.***

I. Introduction

Over the years, the profile of Bangalore has changed drastically and is currently better known as one of the country's major IT hub rather than as a "garden city". With economic development, there has been tremendous pressure on the environment. Deterioration of the air quality in Bangalore can be attributed to rapid increase in population and corresponding fuel combustion activities, which include transport, industrial, and domestic sectors. The population of Bangalore urban agglomerate increased to about 76 lakhs in 2007. The number of vehicles too increased rapidly to about 25 lakhs in 2007, majority of which are private vehicles such as two-wheelers and cars. In terms of contribution to the air pollution load (especially particulate matter), besides the transport sector emissions, the movement of vehicles over paved roads leads to resuspension of road dust that also contributes to the particulate matter emissions. Though there are no major highly polluting industries in Bangalore, however there do exist a number of industries located in some of the earmarked industrial areas in the city. These industries include Engineering, Metal, Textile, Wood, Printing, Rubber & Plastics, Chemicals, Glass, etc. Diesel Generator (DG) sets are additional source of pollution because of power cuts. Besides the industries, most of the commercial establishments and some households in Bangalore have DG sets. Domestic fuel combustion too has been proportionately rising with the rise in population. Other sources of air pollution in the city include restaurants, hotels, bakeries which burn fuel for cooking purposes. Construction activities across the city also add to the PM emission load.

Bangalore city (as per survey of India map, 2002) has been divided into grids of 2x2 km². Emission inventory has been prepared for the city as a whole as well as for the 2x2 km² zone of influence around the monitoring sites. Information has been collected from secondary sources to establish a baseline profile for the city.

Table 1 Total emission loads (T/d) in Bangalore

	PM_{10}	NO_x	SO_2
Transport	22.4	146.36	2.31
Road dust	10.9	0.00	0.00
Domestic	1.8	2.73	0.68
DG Set	3.6	50.96	3.35
Industry	7.8	17.19	8.21
Hotel	0.1	0.20	0.02
Construction	7.7	0.00	0.00
Total	54.4	217.4	14.6

The indicative sources based on Factor Analysis for the different sites are presented in Table 2

Table 2: Indicative sources based on factor analysis for the different sites

Site	Site description	Indicative sources
Silk Board	Traffic location	Motor vehicle exhaust, secondary particulate matter, construction activities, natural soil, road dust
Peenya	Industrial	Road dust, residual oil burning, crustal soil dust, industrial sources, metal industries,motor vehicle exhaust, construction activities

II. RELATED WORK

In the paper [1], a data science model for big data analytics of frequent patterns with Map Reduce is used. The model is evaluated by using social networks, which are good examples of big data. Evaluation results show the efficiency and practicality of data science model in mining and analyzing big data for the discovery of interesting frequent patterns from various real-life applications including social network analysis.

The authors in paper [2], knowledge discovery techniques have been applied for supporting the decision-making process involved in an operational assessment module reporting EMIS. While previous work has been concentrated in forecasting problems. In this paper, tackled issues related to online decision making and operational air quality assessment. The empirical approach followed yielded trustworthy decision making models, as analytically shown in the previous section. The results of this study have brought forth the potential value of data driven approaches for operational air quality assessment and decision making.

The authors in work [10], the requirements for the design of a genuine air quality information system. Urban regions of Howrah, With its cluster of foundries and other age old industries located within densely populated residential zone, is considered as a case study. Air quality data available for those zone is analyzed and compared with that of other Indian cities. The prospects of designing an information system to monitor air quality of this region were explored in the study.

In the paper [14] the author, proposes a data analysis engine, based on business intelligence methodologies and open technologies, to support different targeted analysis on air pollution data. To analyze the problem from different facets, air pollution measurements are enriched with additional information such as meteorological and traffic data, which are collected through sensor networks available in the smart city context. This integrated dataset is periodically analyzed to generate informative dashboards based on a selection of Key Performance Indicators (KPIs). The informative dashboards can provide useful insights about pollutant concentration at different time granularity levels in the urban areas and support a joint evaluation of pollutant concentrations with climate conditions and traffic flow. As a reference use case, open data on air pollution in the urban area of a major Italian city is analyzed to demonstrate the effectiveness of the proposed approach in a real smart city context.

III. PROPOSED WORK

This paper mainly contains the analysis of air pollution and prediction of air pollutants using machine learning algorithms. The proposed system is to build a best fit machine learning algorithm to analyze and predict the air pollutants. In the previous work, Multinominal Logistic Regression (MLR) and Decision Tree (DT) models were used for the analysis of air pollution. The result obtained was, MLR model was found to be the best fit model by providing the better accuracy when compared to DT model. Whereas, in the proposed system Random Forest (RF) algorithm is introduced for the analysis of air pollution and the results are compared with the MLR and DT models. Also, the normalization technique is used for the prediction of next day air pollutant data.

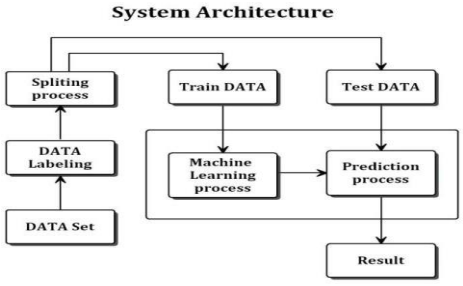

Fig.1: Architecture flow diagram

From the figure1, the data sets are collected from the government websites i.e KPCB. These collected data sets are found to be in a raw format, hence they are pre-processed and then labeled. The pre-processed data sets are clustered using K-Means algorithm. These clustered algorithms are labeled as low, moderate and unhealthy based upon the obtained centroid values after the iteration. These labeled data sets are split into train dataset and test dataset of 9:1 ratio for validation and testing purpose respectively. There are two primary phases in the system:

Training phase: The system is trained by using the datasets and fits a model based on the algorithm chosen accordingly.

Testing phase: The system is provided with the inputs and is tested for its working. The accuracy is checked.

And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to analyze and predict. Hence appropriate algorithms must be used to do the two different tasks. Different algorithms were compared for its accuracy.

IV. METHODOLOGY

Data preprocessing

The air pollution monitoring websites such as Central Pollution Control Board (CPCB) and Open Govt Data (OGD) provides the live data. The collected data sets will be found in the raw format hence the preprocessing of data is required. 60 days of data sets are obtained which contains few attributes such as NO, SO, HC, Humid, Sampling date, Location, Station code and station name. The data sets are obtained from three monitoring stations namely, Peenya, BTM Layout, Silkboard junction at Bangalore. The preprocessing of data includes data cleaning i.e the removal of null values using the normalization technique.

Implementation

K-means

K-means is one of the most popular and simplest unsupervised learning algorithm that solves the well known clustering problem [12]. The procedure, classifies a given dataset through a certain number of clusters. The main idea is to define *k* centroids, one for each cluster, but different locations cause different results. So, the better choice is to place them as much as possible far away from each other or randomly. The next step is to take each point belonging to a given dataset and associate it to the nearest centroid. Then, a new association has to be done in order to create new clusters. This procedure is repeated until *k* centroids do not change their location any more.

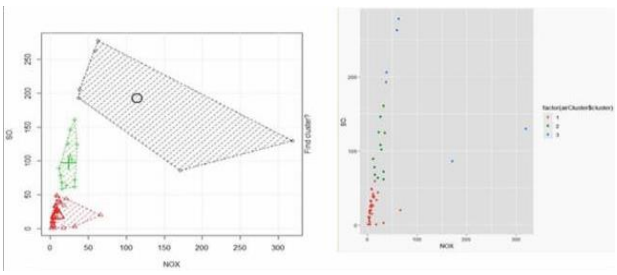

Fig.2: Clustering of datasets using K-Means algorithm

The figure 2 shows three clusters, each of 12, 42, 6 sizes. These clusters are labeled as low, medium and moderate based upon the obtained centroid values after the iteration.

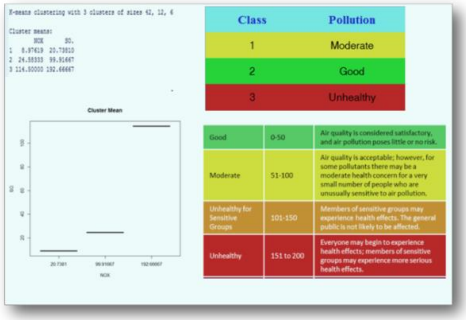

Fig.3: Labelling of clusters

From the above figure 3, the clustered algorithms are labeled as low, moderate and unhealthy based upon the obtained centroid values after the iteration.

Multinominal Logistic Regression

Multinomial Logistic Regression (MLR) is a form of linear regression analysis conducted when the dependent variable is nominal with more than two levels. It is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous level (interval or ratio scale) independent variables. The multinomial logistic regression estimates a separate binary logistic regression model for each dummy variable. The result is M-1 binary logistic regression models. Each model conveys the effect of predictors on the probability of success in that category, in comparison to the reference category.

Fig .4: Iterations values using MLR model

From figure 4, the iteration values are obtained using MLR model, the residual deviance value and AIC values are found to be 78.20799 and 98.20799 respectively.

Decision Tree

Decision tree supports both numerical and categorical data. It cuts down data set into small and smaller parts. The last result I s a tree with decision nodes and leaf odes. The

decision tree is read by starting at the root of the tree, then followed the line to the next left or right node, asking if the number of nodes positive is equal to or greater than 0. If not, then follow the line up the tree (through nodespos <= 0, which is true) to the next node, which asks if the tumor size is less than or equal to 20. If true, we reach as leaf where the outcome is 0.

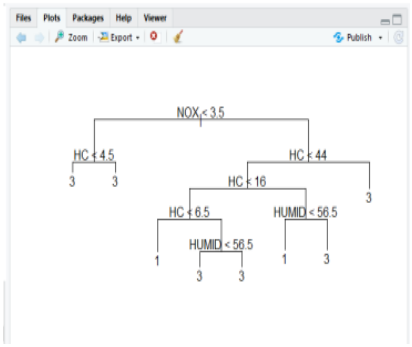

Fig.5: Decision Tree

From the figure 5, the decision tree model is constructed based upon the values obtained from the iterations. The values greater than 56.5 is found to be in the right leaf node and the values less than 56.5 are found in left leaf node.

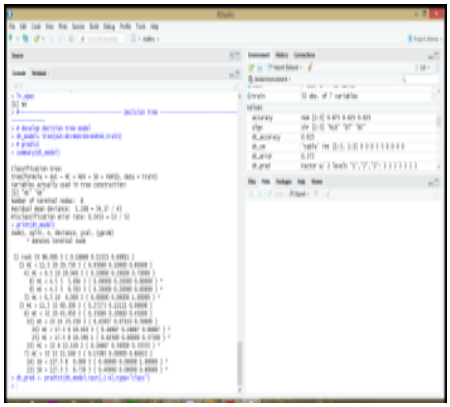

Fig.6: Iteration values obtained from Decision tree

The figure 6, shows the iteration values obtained from decision tree, the residual mean deviance and misclassification error rate is found to be 1.208 and 0.2453 respectively.

Random Forest

Random Forest is a ensemble method, but specifically designed for decision trees. Random Forest generates many decision trees by bootstrapping. The user can define many trees to construct, and the collection of trees is called the forest. The term "random" reflects the fact that each tree is built using a training set of random instances, unlike boosting, which creates new trees with a focus on hard to classify instances (Witten & Frank, 2005). Random Forest classifies a new instance by running it through each tree to make a prediction and then votes by taking the majority class predicted by all trees.

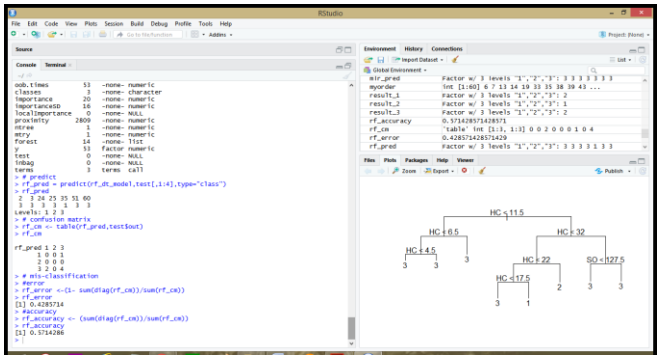

Fig.7: Random forest

From the figure 7, the error rate and accuracy is evaluated using the random forest algorithm. The error rate is found to be 0.42857 and accuracy is 0.57412.

V. RESULTS & ANALYSIS

Analyzing the different classifiers, the accuracy and error rate are recorded. The accuracy, it depends on the parameters chosen. It depends on the dependent and independent variables or predictor. For some attributes it will give high accuracy, for some it will give low accuracy.

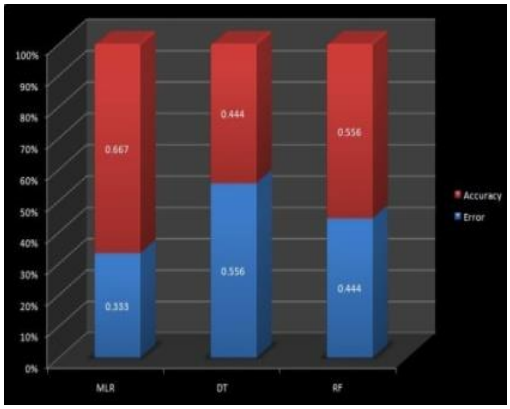

Fig.8: Comparison between accuracy and error rate

From the figure 8, the comparisons are made based upon the obtained accuracy and error rate. MLR model is found to be the best fit model providing the lower error rate and high accuracy when compared to the other models.

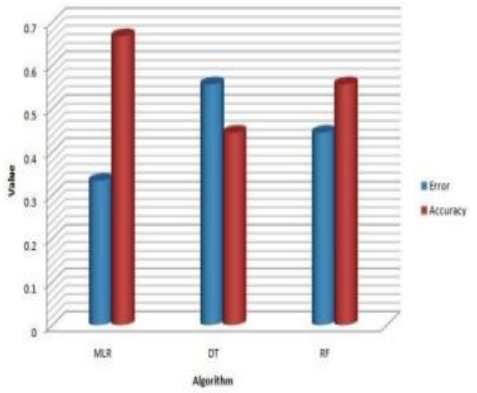

Fig.8: Comparison between MLR, DT and RF

From the figure 8, the MLR model provides the better accuracy of 66% when compared to DT and RF models. DT model provides 44% of accuracy and RF model provides 55% of accuracy. The error rates are found to be 33%, 55% and 44% for MLR, DT and RF models respectively.

Prediction of next day air pollutant

The prediction of next data day air pollutants is done by normalizing the datasets. Wherein, the Standard deviation value and the Mean value is been evaluated for the given data sets to predict the next day air pollutants.

Thereafter, the predicted air pollutant values are been tracked back to which class it belongs in order to identify whether the values belong to low, moderate or unhealthy category.

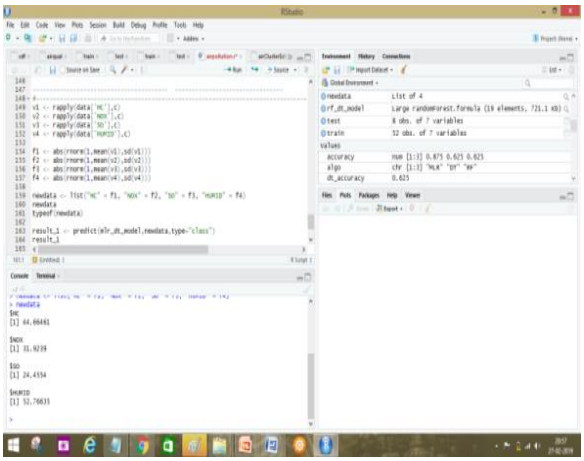

Fig.9: Predicted next day data

From the figure 9, shows the values predicted for HC, NOX, SO, Humid of the following values 44.664, 31.9239, 24.4554, 52.76635 respectively.

VI. CONCLUSIONS

In this work, the probabilistic models were used to reveal the values in cities in terms of different parameters. The various models were evaluated based upon the error rate and accuracy. The MLR model seems to be the best fit model since the error rate is found to be lesser compared to the DT and RF model error rate. Generally random forest gives good results but in this case decision tree and logistic regression have given high accuracy.

REFERENCES

[1] Carson K. Leung, Fan Jiang, Hao Zhang, and Adam G.M. Pazdor, "A Data Science Model for Big Data Analytics of Frequent Pat- terns" 978-1-5090-4065-0© 2016 IEEE.
[2] Mihaela Oprea, Hai-Ying Liu, "A knowledge-based approach for PM2.5 air pollution effects analysis" 978-1-4673-9910-4 ©2016 IEEE.
[3] Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner, Daniel Price," Air Quality Simulations using Big Data Programming Mod- els" 978-1-5090-2251-9/16 © 2016 IEEE.
[4] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, and Mohammad Reza Kavoosifar, "Analyzing air pollution on the ur- ban environment" MIPRO 2016, May 30 - June 3, 2016.
[5] Navjot Kaur Walia, Parul Kalra, Deepti Mehrotra, "Prediction of Carbon Stock Available in Forest using Naive Bayes Approach" 978-1-5090-0210-8/16 © 2016 IEEE.
[6] David G. Rickerby, Andreas N. Skouloudis, "Big data for innova- tive air-pollution assessments in the era of verifiable regulatory de- cisions" 978-1-5090-0058-6/16 ©2016 European Union.
[7] Jinsong Wu, Senior Member, IEEE, Song Guo, Senior Member, IEEE, Jie Li, Senior Member, IEEE, "Big Data application in Green Challenges" 1932-8184 © 2016 IEEE.
[8] James Manyika, Michael Chui, Brad Brown, "Big data: The next frontier for innovation, competition, and productivity" Report, McKinsey Global Institute, USA, May 2011.
[9] Nancy Agrawal and Arushi Baboota, "The Importance of Including Carcinogenic Benzene in Real-Time Ambient Air Quality Data in Delhi" COMSNETS 2016 - Net Health Workshop, 978-1-4673- 9622-6/16 ©2016 IEEE.
[10] Sreemoyee Roy and Abhik Mukherjee, "Information system analy- sis for monitoring of air quality in peri-urban Howrah" 2012 Third International Conference on Emerging Applications of Information Technology (EAIT), 978-1-4673-1827-3/12 ©2012 IEEE.
[11] Lily Bui, "Breathing Smarter: A critical look at Representation of air quality sensing data across platform and publics" 2015 IEEE.
[12] Wenjun Lv, Yu Kang, Zerui Li, Yunbo Zhao "Fusion Approach for Real-Time Mapping Street Atmospheric Pollution Concentration" 978-1-5090-1729-4/16©2016 IEEE.
[13] A. Cuzzocrea and C. K. Leung, "Computing theoretically-sound upper bounds to expected support for frequent pattern mining prob- lems over uncertain big data," Proc. IPMU 2016, Part II, pp. 379– 392.
[14] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, "Analyzing air pollution on the urban environment" MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia.
[15] Abhishek Pandey, Amit Sinha, "An Analytical Approach to Check the Development of any State in India" 2016 Second International Conference on Computational Intelligence & Communication Technology, 978-1-5090-0210-8/16 © 2016 IEEE.
[16] Valerio Persico, Antonio Montieri, Antonio Pescape, "On the Net- work Performance of Amazon S3 Cloud-storage Service" 2016 fifth IEEE International Conference on Cloud Networking, 978-1-5090- 5093-2/16 © 2016 IEEE.