Artificial & Computational Intelligence / Published online: Dec 2019
http://acors.org/ijacoi/VOL1_ISSUE1_05.pdf

Artificial & Computational Intelligence

Data Mining & Machine Learning Algorithms for Air Pollutant Prediction

Shaheen Usmani
Department of Computer Science
Shaheenusmani57@gmail.com
Madhav Institute of Technology & Science, Gwalior (M.P), India

Abstract
Air pollution can be defined as existence or initiation of a liquid, solid, and gases in the environment which have noxious effects on human health's and animals and environment. Now a day's fresh air of environment is getting polluted due to harmful substances like biological molecules, noxious substances and harmful gases. The focus of this paper is to study about the data mining and machine learning techniques used for prediction of air pollution and the mainly focus is on the prediction of PM2.5 on the basis of all the other air pollutants and temperature and humidity. For implementing such type of model different types of classification and prediction algorithms are used.

Keywords: Data Mining, Artificial Neural Network (ANN), Generalized Linear Model (GLM), Random Forest, Recursive Partitioning and regression tree (RPART), Air pollutants.

1. INTRODUCTION

Air pollution is defined as abominable contagion of gas, vapor, dirt, aroma, or synthetic particulates in the environment. In India advancement and progress in the field of industries and urbanization is increasing very rapidly and air pollution is also increasing to a high level which have noxious effects on human beings, animals and environment. Various types of sources are responsible for air pollution. The main sources of air pollution are burning of fossil fuels, emission of noxious gases and solid substances from vehicles and industries. Such materials are Sulphur oxides, nitrogen dioxides, particulate matter, carbon monoxide[1].Currently supervising and scrutinizing air quality is a very crucial issue to have a healthful life, and it also very important. By applying data mining techniques air pollution can be analyzed, so that apt actions can be taken for reduction of air pollution.

Data Mining is simply an approach for extracting intention-based knowledge from the raw data set. Data mining can also be used to explore huge data, the most frequent set of patterns in a dataset. The main aim behind the actual Data mining procedure is to mine the information data from a large collection of data and change it into an explainable framework for additional use.

Data mining can be used for Prediction, Identification, Classification, and Optimization. Data mining can be defined as the excerpting of hidden predictive knowledge from large database. It can be defined as a logical process that is used to

search through large amount of data with the objective of finding useful information. The main goal of this technique is to find patterns that were previously unknown as well as novel information. The terms knowledge discovery and data mining are distinct.

By applying data mining techniques air pollution analysis and prediction and forecasting of pollutants can be performed. And reason behind the air pollution can be identified.

For efficient result of classification and prediction optimization algorithms can be used for most appropriate feature selection.

2. LITERATURE REVIEW

Ranjana Waman Gore et.al.in [1] have proposed an approach in which Naïve Bayes and J48 classification algorithms are used for analyzing the air quality levels. The accuracy of dataset by using Naïve Bayes was 86.66% and the accuracy with J48 decision tree algorithm was 91.99%. And author also justify that J48 algorithm gives more accurate results than Naïve Bayes algorithm.

Sandhya P in [2] have proposed a method in which author aim is to predict the PM2.5 by using random forest, Naïve Bayes, and decision tree algorithm.

Bonny Paulose et.al. in [3] proposed mainly focused on analysis of air quality of Delhi and also find the reason behind the pollutants that cause air pollution by using K-means clustering algorithm. And the author showed that Anand Vihar, R k Puramand, Punjabi Bagh are one of the mostly polluted regions. Ranjana Gore et.al. in [4] proposed a method in which author used Random forest and multiclass classifier classification algorithms for analysis of air quality. The author also showed that multiclass classifier is superior than random forest. Mohamed Shakir et.al. in [5] proposed a model for investigation of air pollution of Karnataka state. The author used the ZeroR algorithm for air pollutants analysis. And author also shows the dependencies and relationship between pollutants.

Shweta Taneja et.al. in [6] have proposed an approach for predicting the air pollution in Delhi. The author used time series analysis techniques namely are Linear regression and multilayer perceptron for predicting air pollutants.

Kiymet Kaya et.al. in [11] have proposed a model for binary classification of PM10 levels. The author used the Extra tree classifier, Gradient boosting classifier and Random forest classifier for classification of PM10 levels. Author also justify that Random forest classifier gives more accurate results.

Kostandina Veljanovska et.al. in [15] have proposed an approach for predicting air quality index by using Machine learning approach. The algorithm used by an author are Neural Network, K- Nearest Neighbour, SVM and Decision tree. Author conclude that Neural network is more accurate in comparison of others.

Rubal et.al. in [10] have proposed an approach for prediction of air pollution. The author used hybrid technique for prediction i.e. combined approach of differential evolution method with random forest algorithm for obtaining precise results.

3. Data Mining Techniques

3.1 Decision Tree

Decision Tree is a flow chart like tree structure that represents sets of conditions and results and can generate rules for the classification of a data set [19]. Decision tree has different types of nodes namely are

- Root Node
- Internal Node
- Terminal Node

Basically, Decision tree algorithm procedure can be divided into two steps:

- Construction of Decision tree

- Pruning of tree

In data mining there are various types of decision tree algorithms available used for classification of data such as ID3 (Iterative Dichotomiser 3), J48, CART, C4.5, PUBLIC, CN2 etc.

3.2 Neural Network

An artificial neural network can be called as neural network. ANN is an efficient and most used approach in the field of the computing system. The idea behind the ANN is based on Biological Neurons in the human brain.

It comprised of an interrelated group of neurons and process information using connectionism method to computation [15]. Artificial neural network has an adaptive property because neural network changed their structure according to internal information or external information passes through this network.

Artificial neural network has various components are as input layer, hidden layer, output layer, weights, activation function, threshold.

3.3 Support Vector Machine (SVM)

SVM is a training algorithm for generating classification and regression rules from data [15]. In support vector machine classifier generates the support vectors which are data points which are lie at the boundary of an area, that distinguish one class from another. It is a supervised learning approach of classification.

3.4 Random Forest

Belongs to the family of supervised learning approaches, suitable for the classification and regression problems as well. Basic working ideas behind this approach are multiple collections of tree-structured classifiers. Random forest is an ensemble learning method. It is used when size of dataset is large and the very large number of input variables approximately hundreds or thousands [11].

3.5 Naïve Bayes

Naive Bayes produces a probabilistic model of the data. It is also called the probabilistic classifier, it works on the probability assumptions and generate strong independence rules among data.

These data mining techniques are used by various researchers for prediction and analysis of air pollution and author also used various forecasting techniques for prediction of a particular year based on past data.

4. PROPOSED METHODOLOGY

- **Data Collection**- In this research work the raw data of air pollution of Delhi is collected. The dataset contains the following attributes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Date	State	City	Location	SO2	NO2	PM10	PM2.5	O3	CO	NH3	C6H6	Temperat	Humidity	AQI	Class
2	1/11/2017	Delhi	Delhi	Nizamudd	14	25	189	45	19	2.3	35	2.8	24.6	67	159	3
3	2/11/2017	Delhi	Delhi	Nizamudd	14	24	156	111	30	1.5	32	2.9	24.5	68	270	4
4	3/11/2017	Delhi	Delhi	Nizamudd	16	26	200	89	56	1.9	35	3.9	25	67.7	196	3
5	4/11/2017	Delhi	Delhi	Nizamudd	14	24	156	59	31	0.5	41	2.6	23	66	137	3
6	5/11/2017	Delhi	Delhi	Nizamudd	17	23	189	45	35	6	38	4.3	24.6	59.3	159	4
7	6/11/2017	Delhi	Delhi	Nizamudd	16	40	156	56	20	0.7	68	4.1	24.7	63.8	137	3
8	7/11/2017	Delhi	Delhi	Nizamudd	21	35	178	101	23	0.8	118	5.9	22.3	75.5	236	4
9	8/11/2017	Delhi	Delhi	Nizamudd	20	22	200	101	30	1.9	128	7.1	22.4	69.4	236	4
10	9/11/2017	Delhi	Delhi	Nizamudd	14	38	135	54	35	3.5	141	6.1	22.6	64.6	123	3
11	10/11/2017	Delhi	Delhi	Nizamudd	17	50	150	101	20	4	207	6.2	23.3	57.9	236	4
12	11/11/2017	Delhi	Delhi	Nizamudd	20	45	100	56	27	5	186	7.5	22.3	63.7	137	3
13	12/11/2017	Delhi	Delhi	Nizamudd	14	29	200	180	24	6.7	32	6.7	21.9	63.2	346	5
14	13/11/2017	Delhi	Delhi	Nizamudd	17	42	200	120	25	11.7	127	4.8	22.2	61.9	300	4
15	14/11/2017	Delhi	Delhi	Nizamudd	20	33	300	129	21	2.4	48	3.2	21.8	55.1	306	5
16	15/11/2017	Delhi	Delhi	Nizamudd	18	57	123	56	37	1.5	48	3.1	21.8	60.8	115	3
17	16/11/2017	Delhi	Delhi	Nizamudd	16	50	101	78	18	1.2	45	3.6	20.4	62.1	160	3
18	17/11/2017	Delhi	Delhi	Nizamudd	13	53	150	121	18	0.8	28	2.7	20.8	61.9	300	4
19	18/11/2017	Delhi	Delhi	Nizamudd	15	34	316	124	26	0.7	26	2.6	20.4	56	303	5
20	19/11/2017	Delhi	Delhi	Nizamudd	15	42	235	113	24	0.7	27	1.8	19.4	45.2	276	4
21	20/11/2017	Delhi	Delhi	Nizamudd	13	46	249	101	24	0.8	31	2.1	18.7	44.3	236	4
22	21/11/2017	Delhi	Delhi	Nizamudd	17	56	214	110	25	0.9	36	3.2	19	45.2	266	4

Figure 1: Dataset of Delhi

- **Data Pre-processing-** This step makes the data ready to be processed. Processes include handling of noisy values, removal of redundant values. Selection of proper attributes, etc.
- **Model estimation & Building-**In this step a model is estimated and build for prediction of PM2.5 on the basis of NO2, SO2, CO,O3, C6H6, PM10, humidity and temperature. In this step machine learning algorithms can be

used for more accurate results.For training of the model certain amount of data is required. Similarly, for testing and validation the remaining amount of data is provided. Data Splitting includes the ratio in which training data and testing data is separated. Apply machine learning algorithms on training dataset for building the model. After that apply test data on the trained model.

- **Interpretation & Visualization** -In this step predicted crop production is interpreted and visualize by using different charts and graphs.

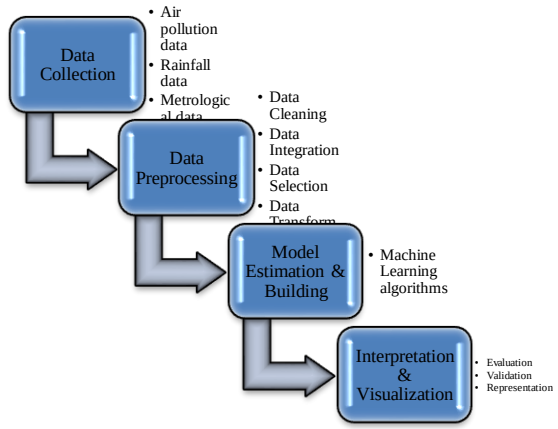


Figure 2: Proposed Methodology

The classification and predictive approaches tested for air pollution data are as follows:

a) RPART: Recursive partitioning is an analytical method for multivariable analysis purpose. Recursive partitioning generates a decision tree and also generates prediction rules by dividing data into subparts of data. [30].

Regression techniques consist a single response variable and one or more input variables. Input variable may consist categorical and continuous both types of variables. A regression tree is alternative of decision trees used for find out real valued functions. A regression tree is created by using the process of recursive partitioning [31].

b) Random Forest: Belongs to the family of supervised learning approaches, suitable for the classification and regression problems as well. Basic working ideas behind this approach are multiple collections of tree-structured classifiers. Random Forest is ensemble learning method. It is used when size of dataset is large and the very large number of input variables approximately hundreds or thousands [29].

c) Generalized Linear Model: Generalized linear model is a data mining classification tool. In Generalized Linear model various types of non-linear models may be tested according to the regression theory. It is supervised learning algorithm used for classification and regression [32].

d) Artificial Neural Network: An ANN often just called a neural network. ANN is an efficient and most used approach in the field of the computing system. The idea behind the ANN is based on Biological Neurons in the human brain [16].

It comprised of an interrelated group of neurons and process instruction using connectionist approach to computation. ANN has an adaptive property because neural network changed their structure according to internal information or external information passes through this network [17].

e) Support Vector Machine: SVM is a discriminatory classifier. SVM generates the hyperplane of data. SVM is a classification technique which come under the supervised learning. For generating the rules from the data, the classification and regression, SVM is a training algorithm [33]. In support vector machine classifier

generates the support vectors which are data points which are lie at the boundary of an area, that distinguish one class from another. If the data can be separate into two distinguish classes or linearly separable then a unique global minimum value exist. In support vector machine various kinds of kernels can be used [34].

5. RESULTS

Figure 2,3,4, and 5 shows the comparison between the Actual output and predicted output of PM2.5 here PM2.5 is predicted on the basis of all other air pollutants.

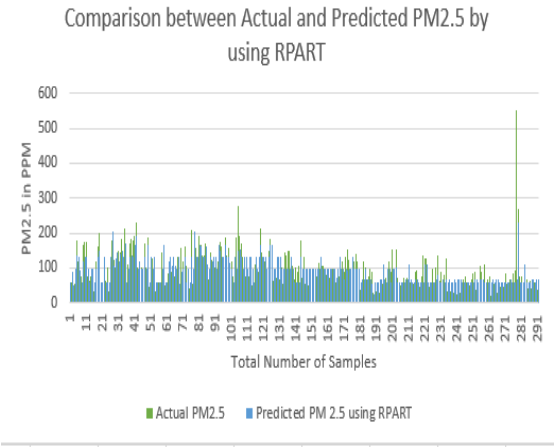


Fig:3. Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using RPART

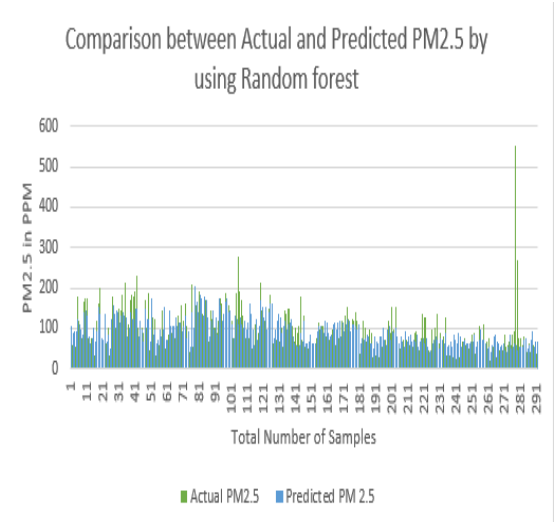


Figure 4: Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using Random Forest

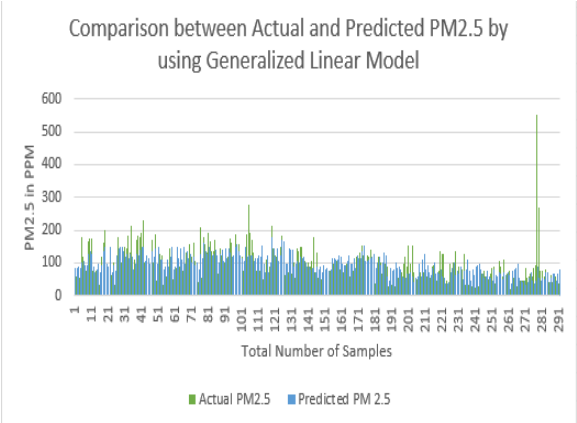


Figure 5: Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using GLM

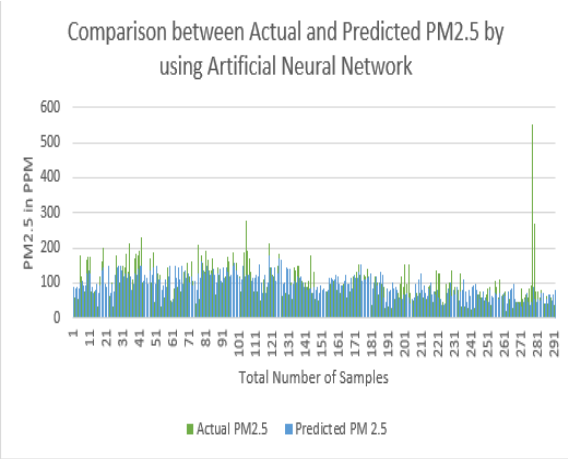


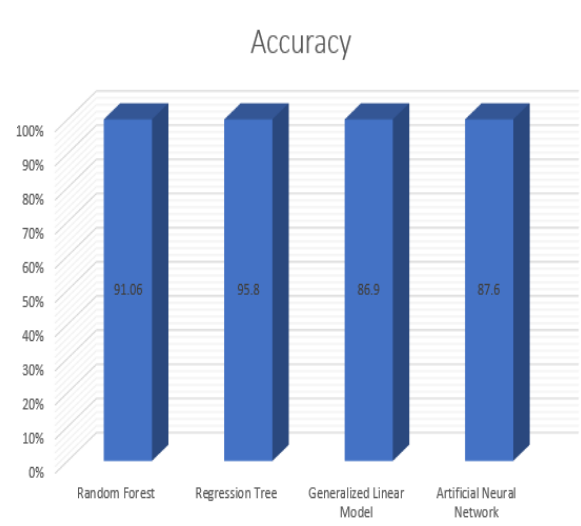
Figure 6: Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using ANN.

6. RESULT ANALYSIS

Figure no.6 shows the comparison among the accuracy of prediction of Random forest, RPART, GLM and ANN. The graph shows that Recursive Partitioning & Regression Tree is more accurate than others and its accuracy is 95.8%.

6. CONCLUSION

Now-a-days an enormous hazard to the environs and specially for human beings is air pollution. In



India air pollution is increasing very rapidly and which have noxious effects. That's why analysis

Fig:6. Accuracy Comparison

and prediction of air pollution is very imperative and necessary. In this research different data mining techniques are used for analysis purpose and Prediction of PM2.5 and among of them Recursive Partitioning & Regression Tree is more efficient and accurate.

7. FUTURE WORK

The efficiency of this model can be increased by integrating Ensemble learning algorithms in the future to get more accurate results.. For improving efficiency, existing optimization techniques can also be applied after slight parameter tuning.

REFERENCES

[1] R. W. Gore and D. S. Deshpande, "An approach for classification of health risks based on air quality levels," *1st International Conference on Intelligent Systems and Information Management (ICISIM)*, Aurangabad, 2017.

[2] Sandhya, P. "Ensemble learning on forecasting fine grained pollutant levels in air using random forest, naive bayes, decision tree algorithms," International Journal of Civil Engineering and Technology, 2018.

[3] Paulose, Bonny & Sabitha, Sai & Punhani, Ritu & Sahani, Ishaan, "Identification of Regions and Probable Health Risks Due to Air Pollution Using K-Mean Clustering Techniques," International Conference on "Computational Intelligence and Communication Technology", 2018.

[4] Ranjana Waman Gore, Deepa S. Deshpande, "Air Data Analysis for Predicting Health Risks," IJCSN - International Journal of Computer Science and Network, 2018.

[5] M. Shakir and N. Rakesh, "Investigation on Air Pollutant Data Sets Using Data Mining Tools," 2nd International Conference on I-SMAC, Palladam, India, 2018.

[6] S. Taneja, N. Sharma, K. Oberoi and Y. Navoria, "Predicting trends in air pollution in Delhi using data mining", *1st India International Conference on Information Processing (IICIP)*, Delhi, 2016.

[7] W. Wang, W. Shen, B. Chen, R. Zhu and Y. Sun, "Air Quality Index Forecasting Based on SVM and Moments," *5th International Conference on Systems and Informatics (ICSAI)*, Nanjing, 2018.

[8] 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), 2018.

[9] Kaur, Gaganjot& Gao, Jerry &Chiao, Sen & Lu, Shengqiang&Xie, Gang, "Air Quality Prediction: Big Data and Machine Learning Approaches," International Journal of Environmental Science and Development, 2018.

[10] Rubal & Kumar, Dinesh, "Evolving Differential evolution method with Random forest for Prediction of Air Pollution," Procedia Computer Science, 2018.

[11] K. Kaya and Ş. GündüzÖğüdücü, "A Binary Classification Model for PM10 Levels," *3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018.

[12] Doreswamy and K. S. Harishkumar, "Multidimensional Data Model for Air Pollution Data Analysis," *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Bangalore, 2018.

[13] S. Ali, S. S. Tirumala and A. Sarrafzadeh, "SVM aggregation modelling for spatio-temporal air pollution analysis," *17th IEEE International Multi Topic Conference*, Karachi, 2014.

[14] Krzysztof Siwek, Stanislaw Osowski, "Data Mining methods for prediction of Air Pollution," Int. J. Appl. Math. Comput. Sci., 2016.

[15] Kostandina Veljanovska, Angel Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2018.

[16] Mahmoud Reza Delavar, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, "A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran", International Journal of Geo-Information, 2019.

[17] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora and Naresh Dhami. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications 163(8):15-19, April 2017.

[18] Priyanka Gaur, "Neural Networks in Data Mining", International Journal of Electronics and Computer Science Engineering, ISSN- 2277-1956.

[19] Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods, 2009.

[20] Terry M. Therneau Elizabeth J. Atkinson Mayo Foundation, "An Introduction to Recursive Partitioning Using the RPART Routines", April 2019.

[21] Kolyshkina, Inna & Wong, Sylvia & Lim, Steven. Enhancing Generalised Linear Models with Data Mining, 2004.

[22] Nayak, Janmenjoy& Naik, Bighnaraj& Behera, Dr. H. A comprehensive survey on support vector machine in data mining tasks: Applications &

challenges, 2015.
[23] Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," *12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 2015.
[24] Han, J., & Kamber, M., Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers, 2001.
[25] Lei Shi, QiquoDuan, Xinming Ma and Mei Weng, "The Research of Support Vector Machine in Africultural Data Classification", IFIP, 2012.