IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 8, Issue 3, June - July, 2020
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

WEKA TOOL USED IN AIR POLLUTION STUDY

Monika Lambsonge[1], Girish Talmale[2], Poonam Prasad[3]
[1]M.Tech student, Computer science & Engineering Department, G H Raisoni College of Engineering, Nagpur, India.
[2] Associate Professor, Computer science & Engineering Department, G H Raisoni College of Engineering, Nagpur, India.
[3] Senior Scientist, Analytical Instrumentation Department, CSIR-National Environmental Engineering Research Institute (CSIR-NEERI), Nagpur,

ABSTRACT

Toxicity in the air is the largest problem in the whole world. The main reason of air pollution is manufacturing companies, vehicles, tree cutting, etc. many companies and environmental research centers are using ML tools for analyzing the dataset to take quick action. WEKA is software which is used to analyze real-time big dataset like industrial, medical, bank, shopping mall, school, colleges. In this paper WEKA tool used to analyze the air pollution dataset with different machine learning techniques such as multilayer perceptron and random forest tree for better solution.

KEYWORDS

Air pollution dataset, WEKA, MLP, RFT

INTRODUCTION

Today is the biggest challenge in front of the world is air pollution most of the countries try to overcome the hazardous condition of air pollution in their cities so, the government organized many environmental-related programs in school, colleges, etc. to motivate students about air pollution and help to increase greenery in there surrounding area And also motivate to their parents and families. In India Delhi government-organized even and odd rule for vehicles to decrease toxic air in the environment. Not only the vehicles but also the industries are also cause to increase toxic air in an environment like chemical industries, pulp and paper industries, etc. cause to emitting high toxic gases in the air and most of the time those people who are leaving in the industrial area they suffers from the lungs disease like asthma, respiratory, cardiovascular diseases and blood diseases. Nearby 30 million people including Children die due to asthma [10]. There are many social occasional programs cause to increase air pollution like Diwali [10].In this study WEKA tool is used to analyze the air pollution dataset for real-time monitoring with different data mining techniques such as multilayer perceptron and random forest tree methods. Two-month Mundka (Rohtak Road) dataset used here for analyzing the real-time monitoring of air pollution concentrations such as CO, CO2, NO2, NO, LEQ, LMIN, LMAX, AQI, PM 10, PM 2.5, humidity, temperature, wind speed, and wind direction.

Particulate matter 10 (PM 10) [11] is also known as aerosol particles present in the atmosphere.PM10 measuring ten micrometers. It causes lung disease, cancer, respiratory and cardiovascular diseases, headache etc. It is very small particles easily inhale through nose and effect on the heart and lungs [12].

PM 2.5[11] is a small particles stay longer in the atmosphere. It is very harmful for human beings and animals. Asthma patient's sufferer's very hazardous conditions when the level of PM 2.5 is high in the atmosphere. It is causing several respiratory diseases.

CO is a flammable, colourless and odourless gas which is produced during the fuel burned [11]. It is harmful to human beings and animals. a large amount of CO can directly effect on heart and brains. It causes several serious problems

headache, dizziness, vomiting, and nausea. It causes sudden death.

CO2 is a colourless, flammable gas. The main reason for the CO2 increase in the atmosphere is burned fossil fuel like coal, gas, oil. Its density high than the dry air, affects on humans health cause high blood pressure, high heartbeat, clumsiness.

Nitrogen monoxide (NO) has a long time been recognized as a poisonous environment. Nitrogen dioxide (NO2) is an inflammable gas. Its high concentration causes a toxic environment and the effect on human health like lungs, It can reduce the immunity power of lungs and the person suffers cold, cough, flu, etc. Its level increased in air during the industrial emittions as compare to vehicles [11].

Maximum level (Lmax) and Maximum level (Lmin) which used to denote the level of environment, etc. Lmax is a high root mean square level and Lmin is the lowest root mean square level. Leq is an equivalent level of sound.AQI stand for air quality index that can map the air quality in an atmosphere. AQI measured at least 8 hours to calculate real-time air quality.

ABOUT WEKA TOOL

WEKA stand for Waikato environment for knowledge analysis developed in 1993 at the University of Waikato in NZ.WEKA created in JAVA however easy to use in any platform. WEKA is free software easily available on the internet. WEKA has a graphically user interface with different machine learning and data mining algorithms like MLP, K-mans, Backpropagation, clustering, etc.It can be handled several algorithms with different data mining tasks by the use of proper algorithms. WEKA give chance to set your training dataset or data preprocessing according to your choice. It can easy to analyze large datasets like industries, banks, shopping malls, Cinemax, school, college, hospitals, etc. WEKA used only comma delimited or CSV file formate and converted it into the attribute related file formate (ARFF) which can be used to load the dataset and explore with the help of data mining techniques.

For example:

Steps of ARFF file format:-

Step 1: Dataset written in CSV format.

Step 2: Dataset converted into ARFF file format, set your dataset according to WEKA need.

Step 3: ARFF file loaded into the WEKA Explorer.

Step4: ones the file is loaded after that applies the algorithms according to your dataset.

Following is the ARFF file formate dataset of Air pollution. We used here **Mundka** (**Rohtak Road dataset**).

@RELATION TEST

%

%

@ATTRIBUTE PERIOD_DATE date "yyyy-MM-dd"

@ATTRIBUTE TIME NUMERIC

@ATTRIBUTE BATTERY NUMERIC

@ATTRIBUTE PM2.5 NUMERIC

@ATTRIBUTE NITROGENDIOXIDE NUMERIC

@ATTRIBUTE WINDDIR NUMERIC

@ATTRIBUTE PARTICULATEMATTER10 NUMERIC

@ATTRIBUTE HUMIDITY NUMERIC

@ATTRIBUTE WINDSPEED NUMERIC

@ATTRIBUTE TEMPRATURE NUMERIC

@ATTRIBUTE CARBONDIOXIDE NUMERIC

@ATTRIBUTE CARBONMONOXIDE NUMERIC

@ATTRIBUTE NITROGENOXIDE NUMERIC

@ATTRIBUTE leq NUMERIC

@ATTRIBUTE lmin NUMERIC

@ATTRIBUTE lmax NUMERIC

@ATTRIBUTE aqi NUMERIC

%

%

@data

AI & ML TECHNIQUES

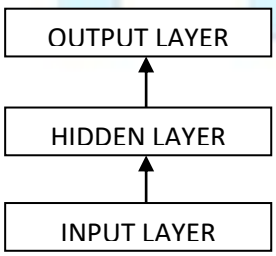
There are two techniques we used here, multilayer percptron and Random forest algorithms

Multilayer Perceptron (MLP)

In MLP supervised learning method for training dataset. ANN was developed in 1950 for the purpose of work like a human brain that is biological brain architecture which is made of several parallel distributed system called neurons. Nowadays ANN is one of the useful methods which act based on real-time air pollution dataset. ANN is a feed-forward [11] neural network that can be used to move forward information. MLP is beneficial to map the entire input for particulate output. MLP performs three layers such as the input layer, output layer, and hidden layer. Hidden layer plays an important role between calculation before forwarding the input and the output data. In our study multilayer perceptron take more time to analyze our air pollution dataset and system get slow down when we apply MLP separately on CO, CO2, NO2, NO, LEQ, LMIN, LMAX, AQI, PM 10, PM 2.5, humidity, temperature, wind speed, and wind direction according to result multilayer perceptron (MLP) take more time as compare to random forest tree (RFT).

Method of Multilayer perceptron (MLP):

Figure 1.1

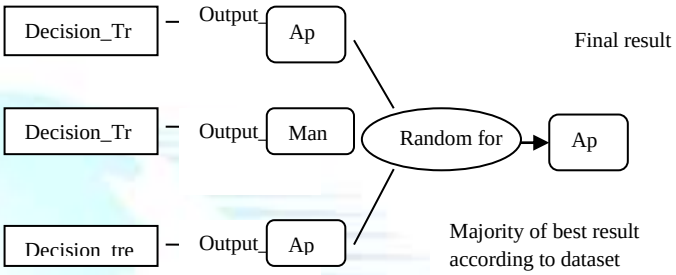


Random forest Tree (RFT):

Random forest algorithm is a classification and regression algorithm used a several decision tree on the entire training set and gets the final future selection result according to the best solution through voting. In this study, random forest tree get a better solution than the multilayer perceptron because with MLP system get slow during the analysis of real-time air pollution dataset and take more time to analyze data.

Method of Random forest tree:

For example: Figure 1.2



LITERATURE SURVEY on Air pollution and WEKA tool

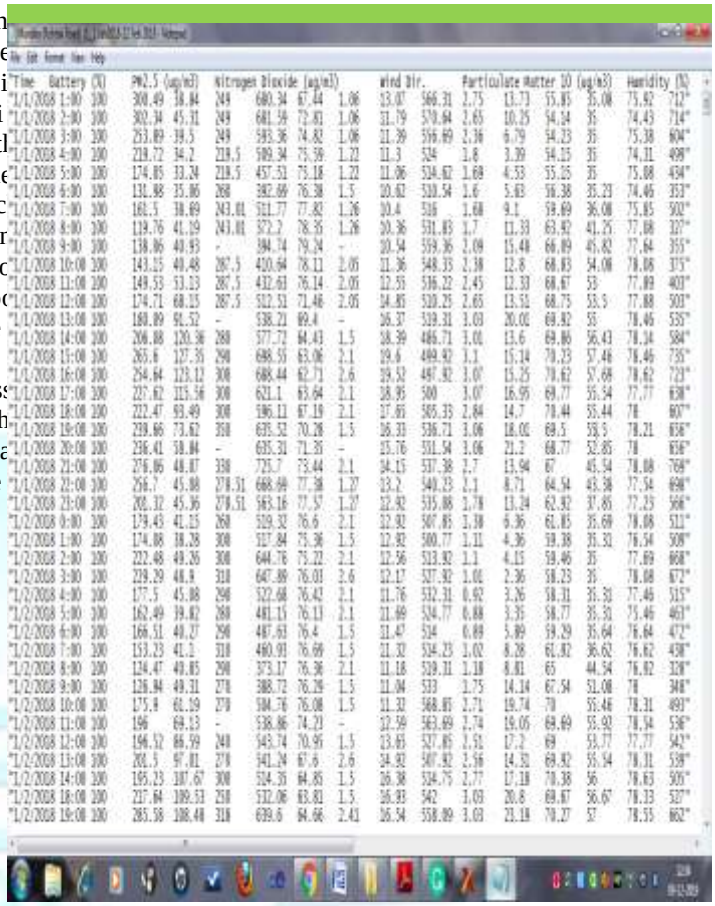
The author Ebrahim Sahafizadeh et al [1] having used data mining with prediction techniques to identify dusty days in Boushehr city, It has a main part of pollution increased in air by used the 53 years data. In this article author focused on dusty days of first three months such as Jan, Feb, and March and compare the relation between future months by the used of prediction method to classify each year data.
The author Min Huang et al [2] considered a new technology to detected pollution in the air by the use of data mining and the BP neural network algorithm. In this paper the author claim this algorithm based on the monitoring of air pollution dataset. This model has three steps, first to find out the air quality factors which were affected in air quality, second after collecting air quality factors used that data to train dataset and third evaluated the model. The author Lidia Contreras-achando et al [3] compared the four polluted concentrations in air like No, No2, SO2, and O3 along six cities through the IDW technique. The author Mirza Farhan Bin Tarek et al[4] used the new technique to analyze the big dataset of air pollution, In this paper author, used the clustering method for analyzing the large dataset with its polluted hot spot areas and time. In this paper author researched on polluted air PM (2.5 & 10), ozone in the UK from 2015-17 according to dataset particulate matter was increased in winter and ozone pollution had a downward trend except some areas. The author Yue Shan Chang et al [5] used the different data mining methods with a cloud platform for forecast the air quality such as PM (2.5 and 10). In this paper author established the extract transform load framework on cloud with computing and storage nodes. The author Chen Chen et al [6] studied a classification method of RF with thematic map for air pollution analysis. In this paper author used the real-time air pollution dataset author applied three

steps here, 1st obtain real-time air quality data in methods from the internet, 2nd step modulate the e pollution data with population, height, and i information. The author Nurul Ashikin Bte Mabahwi studied the air pollution effects on human health. In th author researched the association between humans he AQ, however, the author studied the ambient air suc (10, 2.5), NO2, O3 has more dangerous for humar cause many respiratory problems like asthma, heart pro The author Shweta Srivastava [8] studied the weka too features. In this paper, the author studied the algorithms such as classification, FS, clustering
The author Rohit Arora et al [9] studied the class algorithm in different data set using WEKA. In th author two algorithms for data analysis such as C4.5 a with different dataset however, In result MLP give result more time as compared to C4.5.

.

Dataset used: Mundka (**Rohtak Road dataset**).

Figure 1.3



RESULT table 1.1: According to our Air pollution **Mundka** (**Rohtak Road dataset**).

Table1.1:

Comparison between multilayer perceptron and Random forest tree.

RMSE: root mean square error

MAE: mean absolute error

RRSE: root relative squared error

TNOI: total number of instances

Attribute	Algorithm	CORRELA COEFICIE	MAE	RMSE	RAE	RRSE	TNOI	TIME TAKEN
CO	MLP	0.949	0.2015	0.2709	26.4075	32.1231	539	4.42
CO	**RFT**	**0.971**	**0.1579**	**0.2088**	**20.6947**	**24.7555**	**539**	**1.35**
CO2	MLP	0.8057	25.7755	33.1418	56.1158	61.8747	539	4.43
CO2	**RFT**	**0.8803**	**20.1995**	**26.1829**	**43.9764**	**48.8827**	**539**	**1.3**
NO	MLP	0.8882	3.7232	5.3468	42.1989	51.4666	539	3.98
NO	**RFT**	**0.9238**	**2.8137**	**4.0661**	**31.8899**	**39.1391**	**539**	**1.06**
NO2	MLP	0.8872	13.0893	17.0264	44.1069	48.5879	539	4.22
NO2	**RFT**	**0.9335**	**9.1233**	**12.7774**	**30.7428**	**36.4626**	**539**	**1.17**
PM10	MLP	0.9645	11.1816	48.0093	7.7044	26.7083	539	3.95
PM10	**RFT**	**0.9881**	**12.1132**	**28.1464**	**8.3463**	**15.6583**	**539**	**1.17**
PM2.5	MLP	0.9866	10.8968	14.9224	14.5506	16.295	539	4.48
PM2.5	**RFT**	**0.9798**	**10.7956**	**18.6787**	**14.4155**	**20.3966**	**539**	**1.11**
WIND DIRECTI	MLP	0.7837	58.5315	81.4566	57.7842	71.1098	539	4.44
WIND DIRECT	**RFT**	**0.9253**	**29.3204**	**44.8622**	**28.9461**	**39.1637**	**539**	**1.2**
HUMIDI	MLP	0.9556	2.7475	3.9885	26.0145	30.5307	539	4.03
HUMIDI	**RFT**	**0.9728**	**2.4314**	**3.1685**	**23.0214**	**24.2542**	**539**	**1.01**
WIND SI	MLP	0.5657	0.7612	1.6315	66.2883	84.981	539	3.73
WIND SI	**RFT**	**0.66**	**0.5585**	**1.4433**	**48.64**	**75.1769**	**539**	**0.86**
TEMPRA	MLP	0.9588	0.9121	1.3146	25.2554	30.0345	539	3.98
TEMPRA	**RFT**	**0.9725**	**0.8071**	**1.0538**	**22.3485**	**24.0764**	**539**	**1.2**
LEQ	MLP	0.9767	0.656	0.878	19.3352	22.198	539	4.41
LEQ	**RFT**	**0.9896**	**0.4252**	**0.5799**	**12.5331**	**14.6604**	**539**	**1**
LMIN	MLP	0.9838	1.2005	1.6282	14.8217	18.271	539	3.63
LMIN	**RFT**	**0.9828**	**1.0785**	**1.67**	**13.3151**	**18.7402**	**539**	**1.1**
LMAX	MLP	0.8515	0.4283	0.6676	48.8718	55.5088	539	4.37
LMAX	**RFT**	**0.8993**	**0.3727**	**0.5308**	**42.5247**	**44.1342**	**539**	**1.25**
AQI	MLP	0.9381	12.1277	73.7809	7.4262	36.5871	539	5.3
AQI	**RFT**	**0.9725**	**13.8687**	**46.9908**	**8.4922**	**23.3022**	**539**	**0.88**

Conclusion:

WEKA tool easily access several types of dataset with different ML techniques. In this paper, two artificial intelligence and machine learning algorithms that is MLP and RFT according to result RFT is better than the MLP because it take less time to analysis of dataset not only the time but also the all attributes MAE, RSME, RAE, RRSE get the better solution as compare to MLP. The main motto in this study to analyze air pollution real time dataset in less time with best result.

References:

[1] E. Sahafizadeh and E. Ahmadi, "Prediction of Air Pollution of Boushehr City Using Data Mining," *2009 Second International Conference on Environmental and Computer Science*, Dubai,2009,pp.33-36.doi: 10.1109/ICECS.2009.18

[2] Min Huang, Tao Zhang, Jingyang Wang and Likun Zhu," A New Air Quality Forecasting Model Using Data Mining and Artificial Neural Network",*Hebei University of Science and Technology, Shijiazhuang, 050018, China*,huangmin@hebust.edu.,978-1-4799-8353-7/15/2015IEEE

[3] Lidia Contreras-Ochando, C`esar Ferri,” airVIS: An application for visualizing wind-sensitive interpolation of urban air pollution forecasts” 2016 IEEE 16th International Conference on Data Mining Workshops, 2375-9259/16 $31.00 © 2016 IEEE, DOI 10.1109/ICDMW.2016.180

[4]Mirza F B T, Md Asaduzzaman and Mohammad Patwary,”Spatio-Temporal Analysis of Large Air Pollution Data”, *10th I C EC E,20-22 December, 2018,Bangladesh,*978-1-5386-7482-6/18/$31.00 c 2018 IEEE

[5] Yue Shan Chang, Kuan-Ming Lin, Yi-Ting Tsai, Yu-Ren Zeng, Cheng-Xiang Hung,” Big data platform for air quality analysis and prediction” Dept. of CSIE National Taipei University New Taipei City, Taiwan ,The 27th WOCC2018, IEEE

[6] Chen Chen, Qiang Liu, Pei-Wen Liu, Yao-Sen Huang, Wei-Qing Li, Hao Luo,” Integration of Real-time Air Pollution Analysis and Mapping Based on Multi -source Data”, 2017 International Conference on Network and Information Systems for Computers (ICNISC), 978-1-5386-1618-5/19/$31.00 ©2019 IEEE,DOI 10.1109/ICNISC.2017.00054

[7] Nurul A B M, Oliver L H L, Dasimah Omar,” Human Health and Wellbeing: Human health effect of air Pollution”, *Faculty of Architecture, Planning & Surveying, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia.2014,*AMER (ABRA malaysia).doi: 10.1016/j.sbspro.2014.10.056

[8]Shweta Srivastava,” Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining”, Assistant Professor (CSE Department)
ABES Engineering College, Ghaziabad, *IJCA (0975 - 8887) Volume 88 - No.10, February 2014.*

[9] Rohit Arora, Suman,” Comparative Analysis of Classification Algorithms on Different Datasets using WEKA”, CSE Deptt.Sonepat, Haryana, India, *I J C A (0975 - 8887) Volume 54- No.13, September 2012*

[10] V. Hable-Khandekar and P. Srinath, "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring," *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*,Pune, 2017, pp. 1-6.doi: 10.1109/ICCUBEA.2017.8463746

[11] A.Divya, Assistant professor, R.Kiruthika, Student (UG), D.Gayathri, Student (UG),
“Detecting and analysing the quality of air using low cost sensors to reduce air pollution in urban areas” Department of CSE,IFET College of Engineering, Villupuram, India” IEEE 978-1-7281-1524-5, PICSCAN 2019

[12]Kulwinder Kaur1 and Shivani Dhiman” Review of Data Mining with Weka Tool”,
IJCSE, 2016, india, volume 4, issue-8,S-ISSN 2347-2