Journal of Science and Technology
ISSN: 2456-5660 Volume 06,Special Issue 01, August 2021,

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2021.v6.i04.pp25-30*

Predicting Air Pollutant using Data Mining and Machine Learning Algorithms

Isha Jagtap[1],Prof. Nandini Babbar [2]

[1,2] *Department of Computer Science, NBN Sinhgad School of Engineering, Pune, India.*
[1]*ishajagtap21200@gmail.com*
[2]*Nandini.babbar.nbnssoe@sinhgad.edu*

To Cite this Article
Isha Jagtap,Prof. Nandini Babbar, "Predicting Air Pollutant using Data Mining and Machine Learning Algorithms", Journal of Science and Technology, Vol. 06, Special Issue 01, August 2021, pp25-30.

Article Info
Received: 15.07.2021 *Revised: 24.07.2021* *Accepted:10.08.2021* *Published: 16.08.2021*

Abstract: *Air pollution can be defined presence of harmful or hazardous substances in the air which deteriorate the quality of air. As we are moving ahead in future the environment is getting polluted day by day due to these biological molecules and harmful gases. These pollutant causes diseases, allergy, etc and death as well. The main aim of this article is to study data mining and machine learning algorithms for predicting air pollutants, especially PM2.5 .So as to control the emission of these harmful substances This is a scientific approach for predicting PM2.5 level in the air using a data set containing different attributes.*
Key Word*: Data Mining; Artificial neural network; Support Vector Machine; Decision Tree*

I. INTRODUCTION

Particulate matter can either be human made or can occur naturally. It is a combination of solid and liquid particles that are suspended in air and are hazardous for the life on earth. Its formed during combustion of solid and liquid fuels.

Fine particulate matter of size 2.5 (PM2.5) is an air pollutant that is of great concern for human health when its level in the air uprises.[4] . It reduces visibility as the air becomes cloudy at high values. Various machine learning algorithms are used to detect particulate matter and predict PM2.5 based on given input data. The primary reason to choose machine learning to predict air particulate matter , was its ability to collect the data from sensor and perform action to predict the output and adapt to the algorithms.

Artificial Intelligence and Machine Learning are new area of interest since last couple of years. In science of machine learning the system take its own decision unlike the traditional method of doing its as per the program written by programmer. And gradually it has influenced all aspect of our life. Machine learning [10] has become a key part starting from early stage start up companies to large platform of vendors.

For this we need useful data to fit machine learning algorithms. This is where Data Mining comes to our rescue. Data mining is a fundamental approach to extract important data or insights from the raw data set. Data mining can also be used to explore large data, the most frequent set of patterns in a dataset. The main goal of the data mining process can be used to extract useful information data from a huge collection of data and to transform it into an explainable framework for further use. Data mining [1] is also used for prediction, identification, classification and optimization. Data mining is also defined as the extraction of hidden predictive knowledge from a large data set. By applying data mining techniques to air pollution analysis and pollutant prediction are possible and the causes of air pollution can be identified

II. LITERATURE VIEW

Table no 1 : Shows literature survey

Sr No	Title	Authors	Methodology
1	Air Quality Index Prediction Using Simple Machine Learning Algorithms	KostandinaVeljanovska, Angel Dimoski	They have compared data mining techniques which include supervised machine algorithms such as k-nearest neighbor (k-NN), Support Vector Machines (SVM) and Decision Tree (DT) and one unsupervised algorithm Neural Network (NN).They also calculated the accuracy of each algorithm with different values for validation and testing and compared them
2	Detection and Prediction of Air Pollution using Machine Learning Models	Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu	They reviewed logistic regression in this research paper for modelling using .Logistic function is used to predict the air pollution. Based in the function the system detects whether the air is polluted (0) or not polluted (1). Also they compared the logistic regression with other model and it was found that logistic regression is best suited amongst all
3	A systematic review of data mining and machine learning for air pollution epidemiology	MohomedShazanMohomedJabbar, OsmarZaïane& Alvaro Osornio-Vargas	They have stated the exact review of data mining and machine learning algorithms and its paradigms as well as how to choose a perfect fitting model based on different aspect

DATA MINING TECHNIQUES

1 Decision Tree

Decision tree[8] is a decision support tool which has tree like structure of decision and their consequences. They are commonly used in operation research ,specially in decision analysis which decides which strategy is best suited to achieve the designated goal. Its also a popular Machine learning algorithm.

2 Neural Network

An Artificial Neural Network(ANN)[6]can also be called as Neural Network. ANN isthe most efficient and popular approach in computing field. It is a collection of units or neurons called artificial neurons which are inter-connected to forms the network like structure to get the desired output. A neurons receives the signal , process it and can signal the next neuron connected to it. They are characterized and aggregated into different layers. Different layer performs different processing . The signal travels from first layer to last layer.

3 Support Vector Machine

Support Vector Machine (SVM) [3] is a simple algorithm that every machine learning expert should know.Many prefer SVM because it produces significant precision with less processing power. The goal of the SVM algorithm is finding a hyperplane in n-dimensional space that obviously classifies the data point. Hyperplanes are decision boundaries that help us classify data points. Data points that are closest to the hyperplane which can influence its alignment are called support vectors. By using support vectors, we increase the scope of the classifier.

4 Random Forest

Random forests [7] is part of supervised learning approaches, used for the classification as well as regression problems . The objective of this approach are multiple collections of tree-structured classifiers. Random forest is a learning method. It is used when size of dataset is sizeable and the bulky input variables approximately in hundreds or thousands

III. METHODOLOGY

System has two phases:

1. Training :
   The model is trained by giving input data from dataset and fitting the model we choose accordingly

2. Testing :
   The model is given the inputs and is tested whether its working or not . Then the accuracy is determined.

Hence data given to the system for training and testing purpose must be appropriate[1] Appropriate Algorithms must be in each phase as we are designing the system to predict the PM2.5

1 Data Collection
In this step the raw data of air pollutant is collected from sensors[1] . The dataset contains many attributes.

2 Data Preprocessing
Data preprocessing is a crucial step in the data mining [1] . In this step we process and modify the data which by eliminating redundant values, selecting right attributes, etc.
It is a process of preparing the raw data and making it appropriate for a machine learning model.

3 Model Estimation and Building
In this step we actually build a model to prediction of the PM2.5 on the basis of NO2, SO2, CO,O3, C6H6, PM10, humidity and temperature [1] . Here we split the processed data into particular ratio into two parts i.e. training and test set.
Then we apply machine learning algorithms on training set for building the model. After that we apply test data on the trained model.

4 Interpretation and Visualisation
In this step we predicted air pollutant using the data provided and visualize by using different charts and graphs [1] .

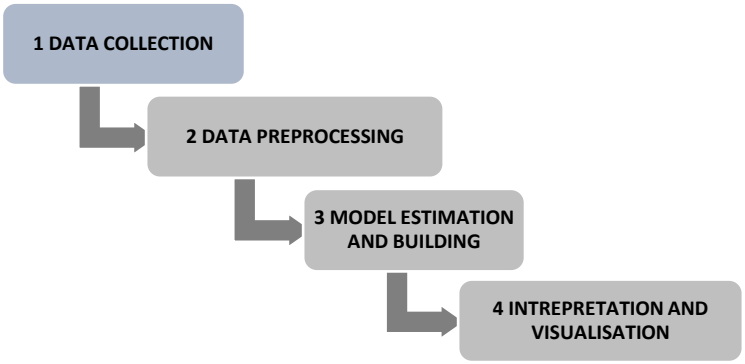

Figure 1 : Flow Chart

The algorithms used for predicting air particulate matter are as follows:
1: RANDOM FOREST
Algorithm :
Step 1 Importing dataset
Step 2 Data preprocessing
Step 3 Feature scaling
Step 4 Splitting data set into training and test set
Step 5 Training the Random Forest model on the Training set
Step 6 Predicting the Test set results
Step 7 Making the Confusion Matrix(to predict accuracy)
Step 8 Visualizing the Test set results

2: ARTIFICIAL NEURAL NETWORK (ANN)
Algorithm:
Step 1 Importing dataset

Step 2 Data preprocessing
Step 3 Feature scaling
Step 4 Splitting data set into training and test set
Step 5 Building ANN
Step 6 Adding the input layer and the first hidden layer
Step 7 Adding the second hidden layer
Step 8 Adding the output layer
Step 9 Training the ANN
Step 10 Predicting the Test set results
Step 11 Making the Confusion Matrix(to predict accuracy)
Step 10 Visualizing the Test set results

IV. RESULTS

Figures 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.

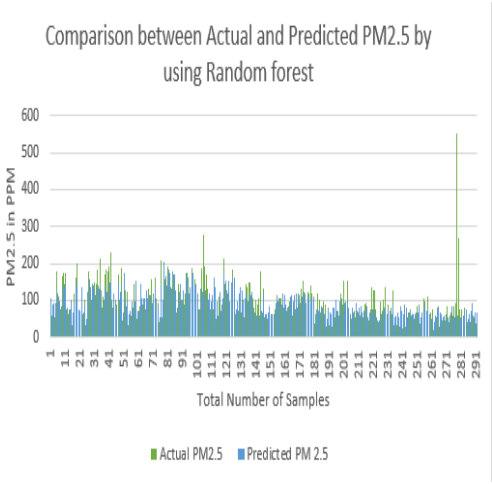

Figure 2: Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using Random Forest

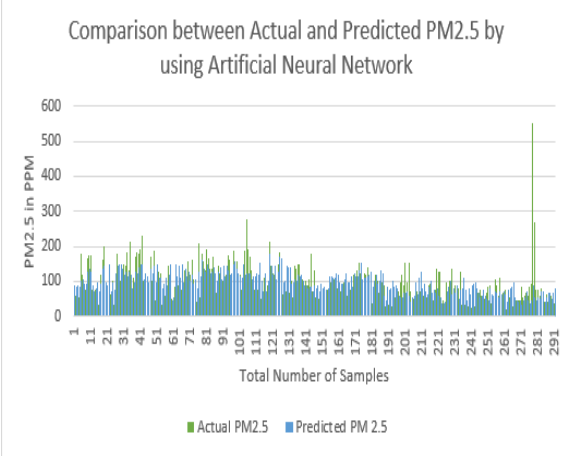

Figure 3: Comparison Graph between Actual PM2.5 and Predicted PM2.5 by using ANN

V. ANALYSIS

Random forest is has more accuracy of 91.06% than Artificial Neural Network which has accuracy of 87.5%. [7].

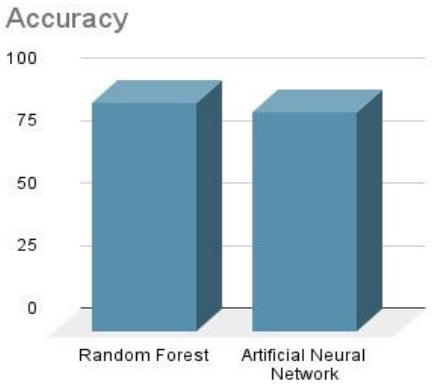

Figure 4 : Accuracy Histogram

VI. CONCLUSION

Now a days as we proceed into future due to increase in the industrialization the air quality is degrading day by day, which has worsen the effect of it on human being. Hence work air quality prediction is very important specially when it comes to India. In this research various basically random forest and ANN machine learning algorithms are used for prediction of PM2.5.Amongst which random forest was more accuracy rate.

VII. FUTURE WORK

The efficiency of the system can be increased by integrating it with upcoming advanced learning algorithms in future to increase the accuracy.

REFRENCES

[1]. Krzysztof Siwek, Stanislaw Osowski, "Data Mining methods for prediction of Air Pollution," Int. J. Appl. Math. Computer. Sci.,2016.
[2]. 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.
[3]. 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.
[4]. Marcazzan, G.M.; Vaccaro, S.; Valli, G. Characterisation of PM10 and PM2.5 particulate matter in the ambient air of Milan (Italy). Atmos. Environ. 2001, 35, 4639–4650.
[5]. Aggarwal, A.; Choudhary, T.; Kumar, P. A fuzzy interface system for determining Air Quality Index. In Proceedings of the 2017 International Conference on Infocom Technologies and Unmanned Systems, Dubai, UAE, 18–20 December 2017; pp. 786–790.
[6]. Azman Azid1,Hafizan Juahir1* MohdTalib Latif2,Sharifuddin Mohd Zain3, Mohamad Romizan Osman ,"Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia".Journal of Environmental Protection, 2013, 4, 1-10.
[7]. Altinçöp, H.; Oktay, A.B. Air Pollution Forecasting with Random Forest Time Series Analysis. In Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 28–29 September 2018.
[8]. N. Loya, et al., "Forecast of Air Quality Based on Ozone by Decision Trees and Neural Networks" Mexican International Conference on Artificial Intelligence (MICAI), pp 97-106, 2012
[9]. Makrand M Jadhav, Gajanan H. Chavan, and Altaf O. Mulani, "Machine Learning based Autonomous Fire Combat Turret", Turkish Journal of Computer and Mathematics Education, Vol.12 No.2 (2021), 2372-2381, https://doi.org/10.17762/turcomat.v12i2.2025