International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Air Pollution Prediction using Data Mining Technique

M.Gayathri[1], R. Shankar[2] and S. Duraisamy[3]

[1]M.Phil Research Scholar, Dept. of Computer Science, Chikkanna Government Arts College, Tirupur, Bharathiar University, Coimbatore, Tamilnadu, India.
[2&3]Assistant Professor, Dept. of Computer Science, Chikkanna Government Arts College, Tirupur, Bharathiar University, Coimbatore, Tamilnadu, India.

Abstract - *Air pollution is one of the major hazards among the environmental pollution. As each living organism needs fresh and good quality air for every second? None of the living things can survive without such air. But because of automobiles, agricultural activities, factories and industries, mining activities, burning of fossil fuels our air is getting polluted. These activities spread sulphur dioxide, nitrogen dioxide, carbon monoxide, particulate matter pollutants in our air which is harmful for all living organism. The air we breathe every moment causes several health issues. So we need a good system that predicts such pollutions and is helpful in better environment. So here we are predicting air pollution for our city using data mining technique. In our model we are using data mining technique c4.5decision tree algorithm. Our system takes past and current data and applies them to our model to predict air pollution. This model reduces the complexity and improves the effectiveness and practicability and can provide more reliable and accurate decision for environmental city.*

Key Words: **Air pollution prediction, Data mining, city, Time series, c4.5 decision tree, Complexity, Effectiveness, Practicable.**

1. INTRODUCTION

One out of every eight deaths in India can be attributed to air pollution, a study conducted by the Indian Council of Medical Research (ICMR) and the Union Health Ministry says. In 2019, 12.4 lakh people died due to air pollution, accounting for 12.5 per cent of total deaths in the country.

As IQAir Air Visual results(Fig-1) shows out of 1,2 most polluted cities are in India. So Air pollution is one of the major hazards among the environmental pollution. As each living organism needs fresh and good quality air for every second. None of the living things can survive without such air. But because of automobiles, agricultural activities, factories and industries, mining activities, burning of fossil fuels our air is getting polluted. These activities spread sulphur dioxide, nitrogen dioxide, Ozone, carbon monoxide, particulate matter pollutants in our air which is harmful for all living organism. Information about this pollutants given below.

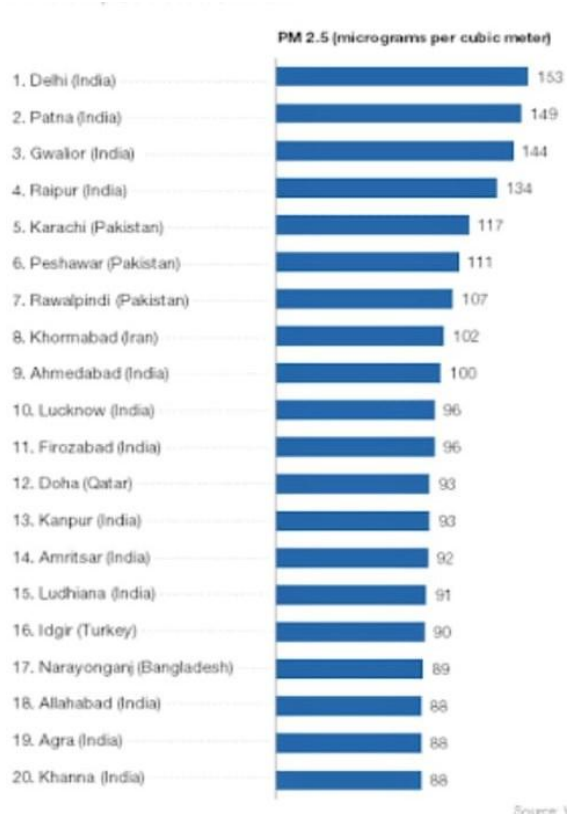

Fig-1: IQAir Air visual result of world's most polluted cities

1. Particulate Matter (PM2.5 and PM10): Particulate matter (PM) is a complex pollutant as it consists of a variety of components in different concentrations. The principle source of particulate matter in Delhi is road traffic emissions, particularly from diesel vehicles. It is also emitted from industrial combustion plants and power generation, Commercial and residential combustion, and some non-combustion processes. Particulate matter is further categorized on the basis of its size in micrometers. The particles under 10 micrometers, refers to PM10 sometimes called the 'coarse fraction'. The particles under 2.5 micrometers, refers to PM 2.5 sometimes called the 'fine fraction'. PM2.5 is considered to be more damaging to human

Health than PM10. The prominent health effects caused due to this are premature death, aggravation of respiratory and cardiovascular disease.

a. Nitrogen Dioxide (NO2): Nitrogen Dioxide is produced during high temperature burning of fuel from road vehicles, heaters and cookers. When this mixes with air, NOx is formed. NOx levels are highest in urban areas as it is related to traffic. It has harmful effects such as wide-range of respiratory problems in school children; cough, runny nose and sore throat etc.

b. Sulphur Dioxide (SO2): It is formed mostly by burning of fossil fuels particularly from power stations, converting wood pulp to paper, production of sulphuric acid, incineration of refused products and smelting. Volcanoes are natural source of emission of sulphur dioxide. This pollutant is the reason for acid rain and has adverse effects on lung functions.

c. Carbon Monoxide (CO): Carbon fuels when burned, either in the presence of too high temperature or too little oxygen, and then CO is formed. Vehicle deceleration and idling vehicle engines are one of its main causes.

d. Ozone (O3): It is formed when a chemical reaction of volatile organic compounds and nitrogen dioxide occurs in the presence of sunlight, so level of ozone is generally higher in the summer.

The air we breathe every moment causes several health issues. So we need a good system that predicts such pollutions and is helpful in better environment. It leads us to look for advance techniques for predicting the air pollution. So here we are predicting air pollution for our smart city using data mining technique. Our system takes past and current data and applies them to our model to predict air pollution. Also this attributes we use for the prediction.

1. Temperature

   Temperature affect air quality because of temperate inversion: the warm air above cooler air acts like a lid, suppressing vertical mixing and trapping the cooler air at the surface. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground.

2. Wind speed

   Wind speed plays a big role in diluting pollutants. Generally, strong winds disperse pollutants, whereas light winds generally result in stagnant conditions allowing pollutants to build up over an area.

3. Relative Humidity

   Humidity could affect the diffusion of contaminant.

4. Traffic index

   The large number of cars on the road cause high level of air pollution and traffic jam may increase the pollutants concentration from vehicles. The definition of traffic index is a index reflecting the smooth status of traffic. The index range is from 0 to 10. 0 represents smooth and 10 represents sever traffic jam.

5. Air quality of previous day

   The air pollution level is influenced by the condition of the previous day to some extent. If the air pollution level of the previous day is high, the pollutants may stay and affect the following day.

The predicting model improves the effectiveness and practicability and can provide more reliable and accurate decision for environmental protection departments for smart city. So here we are using Multivariate Multistep Time series prediction using Random Forest Algorithm. A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time series forecasting is the use of a model to predict future values based on previously observed values.

2. USAGE OF DATA MINING FOR PREDICTION

Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.

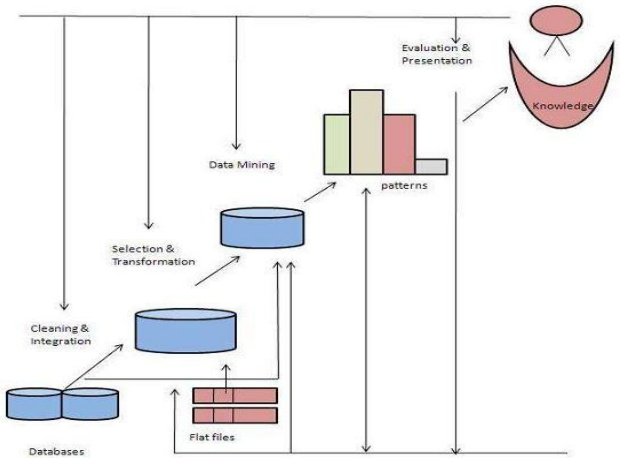

Fig-2: Basic flow diagram of Data Mining for knowledge Discovery

Here is the list of steps involved in knowledge discovery process:

- Data Cleaning - In this step the noise and inconsistent data is removed.
- Data Integration - In this step multiple data sources are combined.
- Data Selection - In this step relevant to the analysis task are retrieved from the database.
- Data Transformation - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining - In this step intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation - In this step, data patterns are evaluated.
- Knowledge Presentation - In this step, knowledge is represented.

Here are the main types of Data mining algorithms.

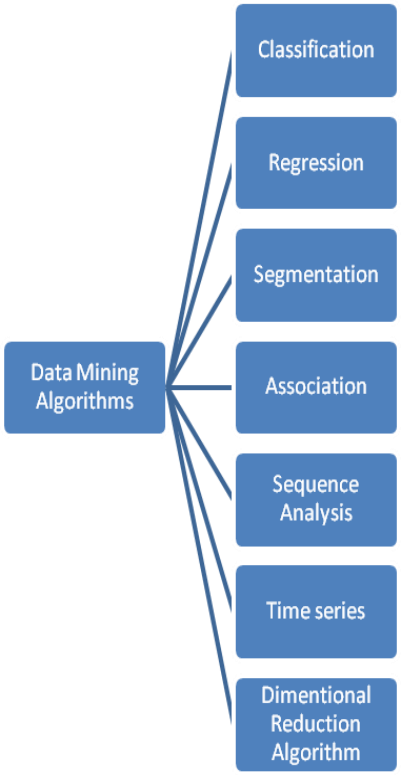

Fig-3: Basic Data Mining Algorithms

1. Classification: These algorithms put the existing data (or past data) into various 'classes' (hence classification) based on their attributes (properties) and use that classified data to make predictions.
2. Regression: These algorithms build a mathematical model based on existing data elements and use that model to predict one or more data elements are mostly used with numbers such as profit, cost, real estate values etc. The primary difference between classification algorithms and regression algorithms is the type of output in that regression algorithms predict numeric values whereas classification algorithms predict a 'class label'.
3. Segmentation or clustering: These algorithms divide data into groups, or clusters, of items that have similar properties.
4. Association: These algorithms find some relation (technically called correlation) between different attributes or properties in existing data and attempt to create 'association' rules to be used for predictions. The algorithms find items in data that frequently occur together.
5. Sequence analysis: These algorithms find frequent sequences in data (Ex: Series of clicks in a web site, or a series of log events preceding machine breakdown).
6. Time series: These algorithms are similar to regression algorithms in that they predict numerical values but time series is focused on forecasting future values of an ordered series and also incorporate seasonal cycles (ex: warehouse inventory management).
7. Dimensional Reduction Algorithms: Some datasets may contain many variables making it almost impossible to identify the important variables with an impact on prediction. Dimension reducing algorithms help identify the most important variables.

3. AIR POLLUTION IN DATA MINING

An important task in providing the proper quality of our life is protection of the environment from air pollution (Bhanu and Lin, 2003; Brunelli et al., 2007; Grivas, 2006; Perez and Trier, 2001). This problem is strictly associated with early prediction of air pollution, concerning the level of SO2, NO2, O3 and particulate matters of diameters up to 10 μm (PM10). Actually, PM is of special importance for a European policy (the new European Air Quality Directive EC/2008/50) defining restrictions for yearly and 24 h average PM10 concentrations.

To respect the short term limit values defined by these restrictions and diminish dangerous concentration levels, emission abatement actions have to be planned at least one day in advance. Moreover, according to EU directives, public information on the air quality status and on the predictable trend for the next day's should also be provided. Hence, one day ahead forecasting is needed.

The paper will discuss the numerical aspects of the air pollution prediction problem, Concentrating on the methods of data mining used for building the most accurate model of prediction.

There are two main tasks to be solved. The first one is generation of the best prognostic features influencing the prediction, and the second—building the structure of the predicting system which provides the most accurate forecast. There are a number of papers devoted to this problem (Bhanu and Lin, 2003; Brunelli et al., 2007; Grivas, 2006; Agirre-Basurko et al., 2006; Mesinet al., 2010).

However, most of them take into account only primary atmospheric variables (temperature, wind, humidity, etc.), on the basis of which the forecast is made. The derivatives of these variables, like the gradient, the estimated trend of their changes, the forecast made on the basis of such trends, etc., have not been used up to now, although their application might improve the quality of prediction. On the other hand, including all of them in the set of features increases the size of input attributes and may lead to decreasing the generalization ability of the prognostic system. Therefore, special methods of detection of the most important factors influencing the prognosis are necessary. This task is known as the feature selection problem (Guyon and Elisseeff, 2003; Tan et al., 2006). Various sets of potential features might be formed from the parameters measured by meteorological stations (temperature, wind, humidity, and insolation) at different hours of the day. The contents of these sets should be analyzed to detect the features which are most important from the prediction point of view (Siwek et al., 2011; Osowski et al., 2009).

In this paper, an analysis and comparison of two approaches to the feature selection will be presented. One applies a genetic algorithm (nonlinear approach) and the other—a linear method of stepwise fit. The former represents a global and the latter a local optimization method. Both the approaches determine the contents of the sets of input variables, treated as the most influential features in the prediction process. Because of different principles of operation the contents of both the sets are usually not the same.

The results of feature selection provide the input information to the system responsible for predicting the average level of air pollution on the next day. Two different systems of prediction will be studied here. In the first one, the features selected are applied to the random forest (RF) of decision trees, which performs two functions at the same time: regression (made by the individual decision trees) and integration (averaging the results of outputs of many decision trees). In the second approach, the features selected create the inputs to the individual predictors, built on the basis of neural networks: the multilayer perception (MLP), the radial basis function (RBF) network and the support vector machine (SVM) of the Gaussian kernel. The universal approximation ability of these networks (Haying, 2000; Scholkopf and Smola, 2002) will be exploited in this approach. All of them have the reputation of very good universal approximations. Their results are combined together in an ensemble providing the final prognosis of an increased accuracy. The numerical results of prediction of different air pollutants (PM10, SO2, NO2 and O3) will be presented and discussed.

4. RELATEDWORK

In this section, we discuss the different papers related to air pollution prediction using data mining technique. We take all the recent years papers.

Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria, proposed paper of Predicting Trends in Air Pollution in Delhi using Data Mining. In this Paper, They have used time series analysis method for analyzing the pollution trends in Delhi and predicting about the future. The time series method includes Multilayer Perceptron and Linear Regression[1].

In Elseviere(2018) paper, Forecasting air pollution load in Delhi using data analysis tools. In this paper, A systematic approach has been followed in this analysis. The approach starts with the collection of dataset from CPCB. Collected data has been pre processed to remove the redundancy. Pre processing of data includes steps like parsing of dates, noise removal, cleaning, training and scaling. Further, descriptive analysis has been carried out on two different platforms-Rstudio and Tableau for different stations. For observing the forecasted results, predictive analysis has been done[2].

KRZYSZTOF SIWEK, STANISŁAW OSOWSKI, Proposed paper for Data Mining methods for Prediction of Air Pollution. The paper will discuss the numerical aspects of the air pollution prediction problem, concentrating on the methods of data mining used for building the most accurate model of prediction. In this paper feature selection is done by using the genetic algorithm (GA). The application of several predictors and feature selection methods allowed integrating their results into one final forecast. The best results of integration were obtained in the direct application of selected features to the RF, performing at the same time the role of regression and integration[3].

In Springer (2019) Paper, Prediction of Air Quality Using Time Series Data Mining. Many of the modern databases are temporal, which makes the task of studying and developing time series data mining techniques an important and much needed task. Time series data mining identifies time-dependent features from time series databases. These features are used for building predictive models. This paper proposes an efficient algorithm to predict the concentration of the various air pollutants by using time series datamining techniques. The time series datamining algorithm CTSPD or

Continuous Target Sequence Pattern Discovery has been used for the prediction of air pollutants. The predictions made by the proposed solution are compared with the predictions made by SAFAR-India and found that the proposed solution provides more accurate results. By studying the obtained air quality patterns, it was found that the concentration of a pollutant need not depend on all the other pollutants[4].

In Springer (2018) Paper, Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi. In this work multi-variable linear regression model of ELM is used to predict air quality index for PM10, PM2.5, NO2, CO, O3. In the proposed model, the previous day air quality index of pollutants and meteorological conditions are used for prediction. Performance of the proposed model was compared with the prediction of an existing prediction system, SAFAR as well as with the actual values of next day. ELM-based prediction was found to have greater accuracy than the existing[5].

Khaled Bashir Shaban, Senior Member, IEEE, Abdullah Kadri, Member, IEEE, and Eman Rezk Proposed Paper of Urban Air Pollution Monitoring System With Forecasting Models. In this paper air quality data are collected wirelessly from monitoring motes that are equipped with an array of gaseous and meteorological sensors. These data are analyzed and used in forecasting concentration values of pollutants using intelligent machine to machine platform. The platform uses ML-based algorithms to build the forecasting models by learning from the collected data. These models predict 1, 8, 12,nd 24 hours ahead of concentration values. Based on extensive experiments, M5P outperforms other algorithms for all gases in all horizons in terms of NRMSE and PTA because of the tree structure efficiency and powerful generalization ability. On the other hand, ANN achieved the worst results because of its poor generalization ability when working on small dataset with many attributes that leads to a complex network that overfit the data, while having SVM better than ANN in our case due to its adaptability with high dimensional data[6].

PUBLICATION	TITLE	METHOD	LIMITATION
IEEE,	Predicting Trendsin	Linear regression,	Linear regression
2016	Air pollution inDelhi usingData Mining.	Multilayer perceptron, Time series analysis	only looks at linear relationships between dependent and independent variables. Sometimes thesis incorrect.
IEEE, 2016	Air Pollution Monitoring System With Forecasting Models.	SVM(Suppor t Vector Machine),A NN(Artificia l neural Network)	Neural Networks requires filling missing values andconverting categorical datainto numerical. We need to define the NN architecture.
AMCS,2016	Data mining methods for prediction of air pollution	SVM Regression RF_fusion	SVM algorithm is not suitable for large data sets. SVM does not perform very well, when the data set has more noise.
Springer, 2018	Pollution prediction using extreme learning machine: a case study on delhi.	ELM(Extre me Machine Learning)	ELM is much faster to train, but cannot encode more than 1 layer of abstraction, so it cannot be "deep".
Elsevier, 2018	Forecasting air pollution load in Delhi usingdata analysis tools.	Time series regression	Here we are using time series with regression.

Table -1: Comparison Table

5. PROPOSED WORK

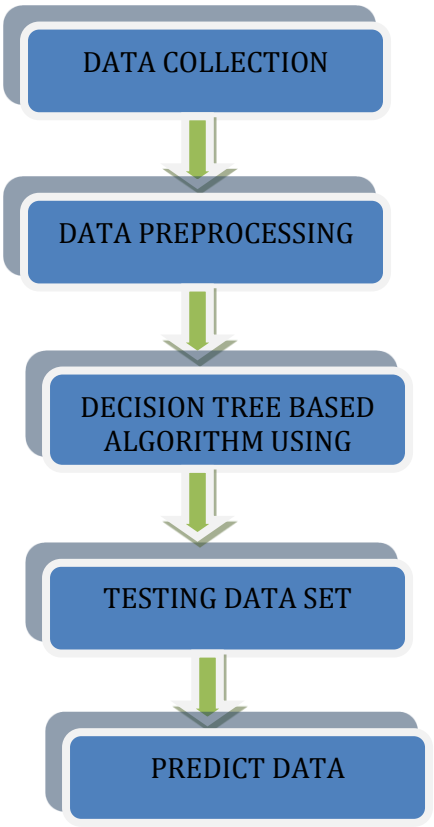

Fig-4: Workflow of proposed method for Air Pollution Prediction

As shown above the proposed model is divided in to five stages

Stage 1 Data Collection: Here we are collecting all the data of attribute which are affect the air pollution. There are many sensors available in smart cities which sense the pollutants.

Stage 2 Data Preprocessing: data are cleaned by removing noise and filling up the missing values.

Stage 3: Decision tree based c4.5 algorithm: Decision tree is the process of finding the most relevant inputs for predictive model. These techniques can be used to identify and remove unneeded, irrelevant and redundant features that do not contribute or decrease the accuracy of the predictive model.

Stage 4 Testing data: In this stage we are taking testing data and using decision tree algorithm we are predicting the air pollution.

Stage 5 Prediction: Here our system predicts the air pollution.

6. CONCLUSION

The proposed system will definitely help in improving the prediction of air pollution in our city. Prediction Using Decision tree based c4.5 algorithm technique improve the the performance and reduce the complexity of the air pollution prediction model. Also here we are using technique which makes our prediction even better.

REFERENCES

[1] Shweta Taneja, Dr. Nidhi Sharma, Kettun Oberoi, Yash Navoria ,"Predicting Trends in Air Pollution in Delhi using Data Mining", IEEE(2016)

[2] Nidhi Sharmaa, Shweta Tanejab*, Vaishali Sagarc, Arshita Bhattd, "Forecasting air pollution load in Delhi using data analysis tools.", Elseviere (ICCIDS 2018)

[3] KRZYSZTOF SIWEK, STANISŁAW OSOWSKI," Data mining methods for prediction of Air Pollution", amcs(2016)

[4] Mansi Yadav, Suruchi Jain and K. R. Seeja," Prediction of Air Quality Using Time Series Data Mining", Springer (2019)

[5] Manisha Bisht and K.R. Seeja," Air Pollution Prediction Using Extreme Learning Machine: A Case Study on Delhi.", Springer(2018)

[6] Khaled Bashir Shaban, Senior Member, IEEE, Abdullah Kadri, Member, IEEE, and Eman Rezk," Air Pollution Monitoring System With Forecasting Models.", Khaled Bashir Shaban, Abdullah Kadri, Eman Rezk, "Urban Air Pollution Monitoring System With Forecasting Models",IEEE SENSORS JOURNAL, VOL. 16, NO. 8, APRIL 15, 2016

[7] Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study Tikhe Shruti, Dr. Mrs. Khare , Dr. Londhe ,IOSR-JESTFT (Mar. - Apr. 2013)

[8] Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study Tikhe Shruti, Dr. Mrs. Khare , Dr. Londhe,IOSR-JESTFT (Mar. - Apr. 2013)

[9] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, hengqiang Lu, and Gang Xie," Air Quality Prediction: Big Data and Machine Learning Approaches" , International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

[10] Ebrahim Sahafizadeh, Esmail Ahmadi," Prediction of Air Pollution of Boushehr City Using Data Mining", 2009 Second International Conference on Environmental and Computer Science.