AIR POLLUTION ESTIMATION USING DATA MINING APPROACH

Suketha[1], Pooja N S[2], Vanishree B S[3]

[1] Department of CSE, SCEM, Karnataka, India
[2], Department of CSE, SCEM, Karnataka, India
[3] Department of CSE, SCEM, Karnataka, India

ABSTRACT

The main objective of this paper is to estimate air pollution using data mining approach. Today world is going behind industrialization which has made human life better as well as negatively affected the environment. Air pollution should be prevented as earlier as possible to avoid threat to human life. Air pollution includes excess levels of ozone, carbon-dioxide, nitrogen oxide etc. The Air Quality Index is an indicator of air quality standards, it is based on air pollutants that have bad effects on human health and the environment. The proposed method uses Random Forest Regression Algorithm and it selects the most important features and predictors. These allows increasing the forecasting accuracy of atmospheric pollution in a significant way.

Keyword : - *Random Forest Regression , Air Quality Index*

1. INTRODUCTION

Industrialization and traffic growth are the major causes of increase in air pollution. Air pollution is a serious concern worldwide which includes excess levels of pollutant particles like ozone, carbon-dioxide, nitrogen oxides and ultra small particles like $PM_{2.5}$. Among all these pollutants $PM_{2.5}$ is most dangerous for human health, as it could be inhaled directly into lungs and also it could be dissolved into blood causing high rate of lung cancer. The Air Quality Index (AQI) is a specific number which is used by governmental agencies and theses number helps characterize the quality of the air at a given location. AQI Scheme transforms the weighed values of individual air pollution related parameters into a single number or set of numbers. AQI is used for local and regional air quality management in many metropolitan cities of the world. The main objective of the present study is to forecast short–term daily AQI through previous day's AQI and meteorological variables using Random Forest regression technique.

Most of the people spends a significant amount of time either in the home, office or other types of buildings where gas, chemical and other pollutants causes headaches, eye irritation and allergies. Breathing quality of a indoor air is critical to maintain good health. Serious pollutants can cause various types of cancers and other long term health complications. Common indoor air pollutants include:

- Second hand smoke: A serious indoor air pollutant which can worsen symptoms for asthma sufferers, increase risks of ear infections in children and increase risks for Sudden Infant Death Syndrome.

- Radon: A dangerous gas pollutant which is identified as the second leading cause of lung cancer.

- Combustion Pollutants includes carbon monoxide and nitrogen dioxide: These gases come from burning materials or improperly vented fuel-burning appliances such as wood stoves, gas stoves, water heaters, dryers and fireplaces.

Carbon monoxide is also a main particle which causes air pollution and is a colorless and odourless gas. This particle is not easily detectable by human senses, and interferes with oxygen delivery throughout the body. Carbon monoxide causes dizziness, headaches, weakness, and toxic amounts can lead to death. Nitrogen dioxide causes shortness of breath, and increased risk for respiratory infections which is also a colorless and odourless gas. Because of all these air pollution prediction has become one of major concern.

2. RELATED WORK

Air pollution by harmful emissions from vehicles is one of basic problems of present time. Some of the harmful substances pollute the air, and the rest, due to rainfalls or natural sedimentation pollute soils. One of the most dangerous harmful substances in the structure of emissions is nitrogen dioxide, which has a significant effect when polluting the air, because there is a long stay in hanging state as well as in conjunction with water. The dynamic mathematical model of nitrogen dioxide distribution is build. The possibility of using the obtained model for pollution prediction in different points of city was confirmed [1] .

The proposed model will show that data mining provides a relatively high goodness of fit and sufficient space-time explanatory power, particularly air pollution frequency and affect areas. In the proposed model, a method using Dynamic Time Warping is proposed to analyse temporal similarity between stations. The proposed model can eliminate global effect on a single station through the performance of multiple stations [2] .

An advanced set of features, created on the basis of the atmospheric parameters, is proposed. Two methods of feature selection are compared. One applies a genetic algorithm , and the other—a linear method of stepwise fit . On the basis of such analysis, two sets of the most predictive features are selected. In the first one, the features selected are directly applied to the random forest (RF), which forms an ensemble of decision trees. In the second case, intermediate predictors built on the basis of neural networks are used. They create an ensemble integrated into the final prognosis [3].

Over the past years the development and urbanization in Delhi has led to increase in air pollution. This has led to study and research in this area. The data mining techniques used are linear regression and multilayer perceptron. We have seen the trends of various air pollutants like sulphur dioxide(SO2),nitrogen dioxide (NO2),particulate matter (PM),carbon monoxide (CO),ozone (O3) . By using the above techniques we have observed that there will be an increase in amount of PM10 by 45.9% in coming years. However amount of CO and NO2 may show slight increase due to increasing number of 2 wheelers on road. The other pollutants like SO2 may show decrease due to usage of non sulphur fuel and stringent pollution control measures [4].

3. METHODOLOGY

The architecture of the system is being emphasized in the architecture diagram that describes the structure and the behaviour of the system. It comprises of many different system components that will work together synchronously to implement the overall system. The architecture begins with menu which follows processing of data in a database. Once the processing is done,yearly statistics graphs are generated and then AQI is calculated. Using a data mining algorithm such as Random Forest Regression we predict PM2.5 value and generate a graph based on that.

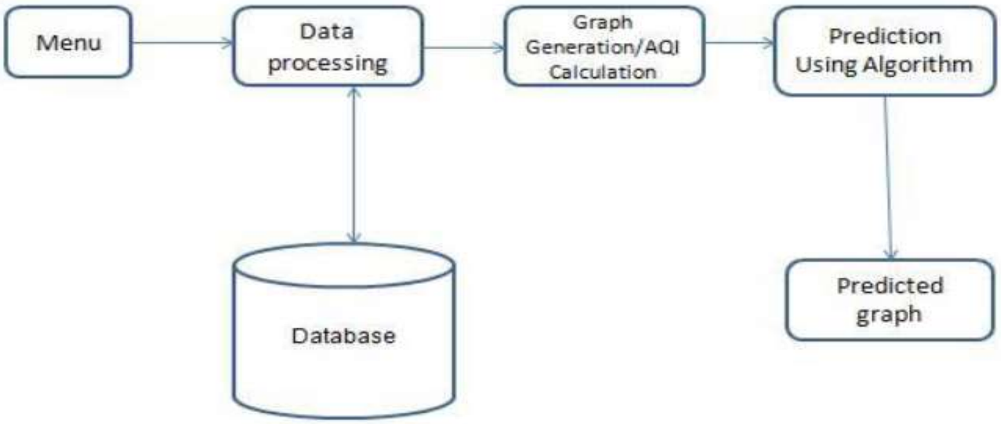

Fig -1: Architecture diagram of proposed method

Random forest regression is a tree based learning method which is popularly used in the fields such as pattern recognition, machine leaning and data mining. In this approach large number of decision tree is created. Prediction of various decision trees are combined together and the most common output is used as the final output. In random forest regression, the parameter n_estimators represents the number of tress to be constructed in the forest, default value of this parameter being 10. Many decision trees are created and to get the final predicted value it takes average of all the results obtained from individual trees in the forest hence resulting in regression. Decision tree maps the attribute or features with the target value forming a pattern between them. Decision tree begins by providing the data set at its root node which can be a table. From this root node the training samples are split into different spaces based on AQI values.

.

4. RESULT

In the Proposed method Air pollution is predicted using dataset of three years. The AQI generator can take various input factors like PM $_{10}$, PM $_{2.5}$, NO_2, SO_2. In this experiment Air pollution is predicted by using one of the dangerous pollutant factor $PM_{2.5}$ as input to the AQI generator. The graph shown below gives the result for the PM2.5 air pollutant particle for the dataset considered in a particular year.

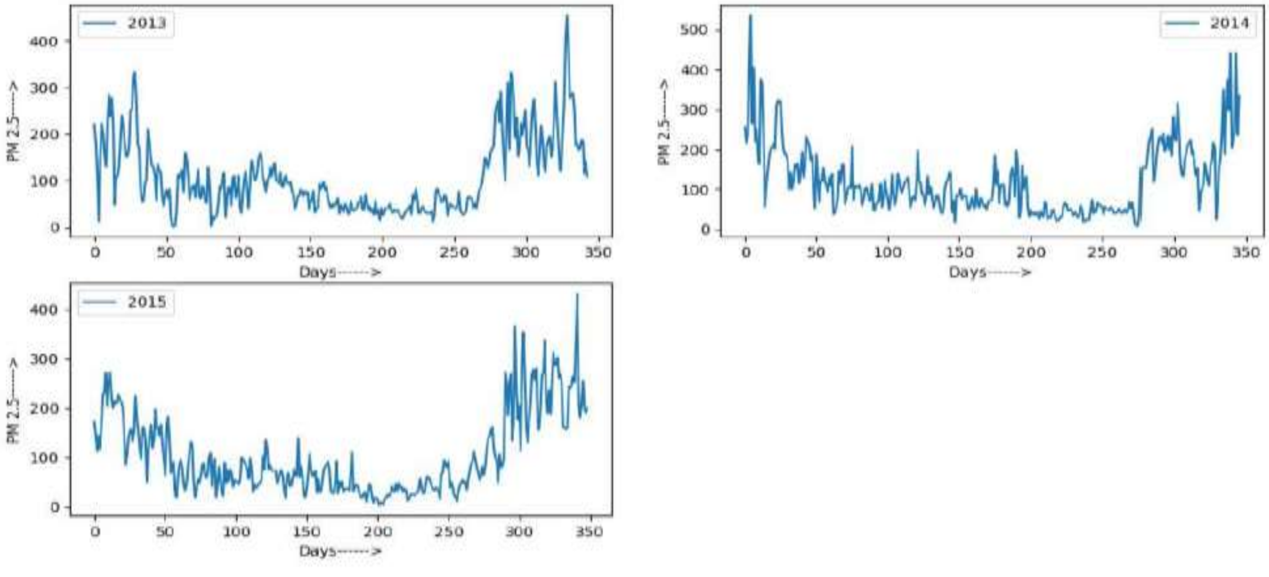

Fig -2: Graph showing PM 2.5 values for 2013, 2014, 2015.

5. CONCLUSION

In this paper, we proposed AQI level prediction method based on Random forest regression method to predict Air pollution in the coming years. In order to evaluate performance of proposed method we conducted experiments by collecting the dataset from the industry area where pollution causing factors present at high rate. Furthermore Random forest regression method applied to get the most common output as the final ouput. The result demonstrates the proposed method is highly suitable to apply any air polluted city.

6. REFERENCES

[1]. MykolaDyvak , IrynaVoytyuk , Natalia Porplytsya , AndriyPukas, “Modeling the process of air pollution by harmful emissions from vehicles”, Jan 2018.

[2]. Ping-Wei Soh , Kai Hsiang Chen , Jen-Wei Huang, Hone-Jay Chu, “ Spatial-Temporal pattern analysis and prediction of air quality in Taiwan” , Oct 2017

[3]. K. Siwek and S. Osowski, “Data mining methods for prediction of air pollution”, Feb 2016

[4]. Shweta Taneja,Dr. Nidhi Sharma, Kettun Oberoi, YashNavoria, “Predicting Trends in Air Pollution in Delhi using Data Mining”, Aug 2016.

[5]. Xia Xi ,Zhao Wei , RuiXiaoguang, “A comprehensive evaluation of air pollution prediction improvement by a machine learning method”, Nov 2015