Impact Factor Value 4.046 e-ISSN: 2456-3463
International Journal of Innovations in Engineering and Science, Vol. 4, No.6, 2019
www.ijies.net

Air Pollution Prediction Using Machine Learning

Maghvendra Singh[1],Harshit Saran[2], Mrs.Sapna Yadav[3]

[1,2]Students, [3] Assistant Professor
IMS Engineering College, Ghaziabad, India,

Abstract*: Monitoring and preserving air quality has become one of the most essential activities in many industrial and urban areas today. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation and prediction has become an important research area. In our study of making air pollution prediction system, the main focus will be in exploring the suitable machine learning techniques that will help in better forecasting of the pollution concentration. In our study, we are focused on predicting the air pollution level of a specific region i.e Ghaziabad by using certain parameters like PM10, PM 2.5, SO2, NO2, Benzene, CO and O3(ozone).*

The data analysis is done by using various machine learning techniques for the past 5 months (2017-2018).In our study, we also made a comparative analysis on the results predicted by the various algorithms like Multi linear Regression,SVM(support vector machines) and Random Forest. This will help in the prediction of air quality in areas of Ghaziabad and this could serve as an important reference for local government agencies in evaluating present and making future air pollution policies.

In predicting the air pollution we also found out those meteorological factors largely affects the air pollution. So, in our study, we also consider some meteorological factors to predict the pollution of air like Temperature, Minimum Temperature, Maximum Temperature, Wind speed and Relative Humidityy and several other features .Furthermore, our paper also throws light on some of the challenges and future research needs.

I- INTRODUCTION

Air is one of the most essential natural resources for the survival and existence of the eeach and every life on this planet. All forms of life including plants and animals depend on air for their basic survival and existence. Thus, all the living organisms need good quality of air which is free of harmful gases to continue their life. The increasing population, its automobiles and industries are polluting all the air at an alarming rate. Air pollution can cause various long-term and short-term health effects.

The development of air quality predictive models can be very useful as such models can provide early warnings of pollution levels increasing to unsatisfactory levels. Pollution concentrations in urban areas are primarily from vehicular exhaust, factories, and small scale industries. Air pollution can affect our health and environment in many ways. Thousands of premature deaths occur every year as a result of inhaling a high level of pollution concentrations like PM10, PM2.5, CO, Nitrogen Oxides (NO+NO2). In the past few years, the heavy environmental loading has led to the deterioration of air quality in urban and industrial areas in Ghaziabad.

This brought us to focus our study on air quality in Ghaziabad. The task of controlling and improving air quality level has attracted a great deal of national attention. The main focus of this paper is to explore the suitable machine learning techniques that will help in better forecasting of the air pollution concentration. The data is collected from CPCB (Central Pollution Control Board) online data and sensors over the target region. Then the distribution of suspended particles like PM10, PM2.5, SO2, and NO2 polluted environment air are identified. The data is analyzezed using data mining techniques for the past 5 months (2017-2018). In predicting the air pollution we also found out those meteorological factors largely affects the air pollution. So, in our study, we also consider some meteorological factors to predict the pollution of air like Temperature, Minimum Temperature, Maximum Temperature, Wind speed and Relative Humidity and several other features.

This will help in the prediction of air quality in urban and industrial areas of Ghaziabad and this could serve as an important reference for government agencies in evaluating present and devising future air pollution policies. Our study focuses on prediction of air pollution level of a particular or specific region. In air pollution prediction, model accuracy, efficiency and adaptability are key considerations.

II- RELATED WORK

A) Literature Review

Xiao Feng, Qi Li, Yajie Zhu et al., 2015 have proposed a novel hybrid model to predict daily average PM2.5 concentration. It is built by applying the trajectory based geographic model and wavelet transformation into the MLP type of neural network. Combined with meteorological forecasts and respective pollutant predictors the hybrid model is considered to be an effective tool to improve the forecasting accuracy of PM2.5.

M. S. Baawain and A. S. Al-Serihi et al., 2014 conducted a study which was carried out in the City of Sohar, Oman, where part of the data for the proposed research was gathered. The study proposed to develop models for daily predictions of CO, PM10, NO, NO2, NOx, SO2, H2S, and O3.The training of the prediction models was based on the Multi-layer Perceptron method with the Back-Propagation algorithm, and showed very high concurrence between the actual and predicted concentrations. In addition, the research investigated the MLP model's sensitivity to variation of epochs cycle (trial and error technique adopted to try different adjustments).

B) Implemented Work

a. Techniques used for predicting air pollution / constituents.

Obtaining a relationship of dependency among the concentrations of the seven pollutants measured is difficult. So we divide our target class into four categories and a analysis was carried out. The work and analysis done using regression are as follows:

1) ***Multi-Linear Regression Algorithm***: The multiple linear regression explains the relationship between one continuous dependent variable (y) and two or more independent variables (x1, x2, x3… etc).Here in our algorithm, we have output (PM2.5) as "Y" i.e the dependent variable and other features like Average Temperature('T'),Average Relative Humidity('H') and Average Wind Speed etc.

We use this algorithm to decide which feature among given will be most suitable. This action of omitting variables is part of stepwise regression. we can do this by 1) Forward Selection 2) Backward Elimination 3) Bidirectional Elimination. The various features which we take into account are 'T'(Average Temperature),'TM' (Maximum Temperature),'Tm'(Minimum Temperature) ,'H'(Relative Humidity),'V'(Average Wind Speed),' VM'(Maximum wind speed).

The pseudo-code for the implemented Multi Linear Regression algorithm in our model is as shown below:

Air Poll Predict Multi Linear Regression(Average Temperature(T),Maximum Temperature(TM),Minimum Temperature(Tm),Average Relative Humidity(H), Average Wind Speed(V), Maximum wind speed (VM), PM2.5)

{

X1 <-- contains training data set features of T,TM, Tm,H,V and VM

Y1 <-- contain only one training data set feature PM2.5

X2 <-- contains testing data set features of T, TM, Tm,H,V and VM

Y2 <-- contains only one testing data set feature PM2.5

Find Adjusted R-squared() method (Perform Feature selection by using this)

correlation matrix // check for collinearity

Applying Multi Regression model (from scalar import linear model)

fit(X1,Y1)

Calculate mean_absolute_error (Y2, lin. predict (X2))*100

The predict function will give the value of PM2.5

}

2) SVM (Support Vector Machine) : Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression.

In our prediction model, we use SVM(support vector machine) algorithm in order to predict the level of PM2.5 pollutant, which is assumed to be the major pollutant among several others. The various kernels including the linear kernel is used to predict the level of PM2.5.The various features which we take into account are 'T'(Average Temperature),'TM'(Maximum Temperature),'Tm'(Minimum Temperature),'H'(Relative Humidity),'V'(Average Wind Speed),'VM'(Maximum

wind speed).The classifier of the SVM algorithm constructs a hyper plane or set of hyper planes in a high-dimensional space, which can be used for classification, regression or other tasks.

In using this algorithm, we obtain a high accuracy as it provides a good separation by the hyper plane that has the largest distance to the nearest training data points of any class.

The pseudo-code for the implemented SVM algorithm in our model is as shown below:

Air Poll Predict SVM (Average Temperature (T),Maximum Temperature(TM),Minimum Temperature (Tm),Average Relative Humidity(H), Average Wind Speed(V),Maximum wind speed(VM), PM2.5)
{
X1 <-- contains training data set features of T,TM,Tm,H,V and VM
Y1 <-- contain only one training data set feature PM2.5
X2 <-- contains testing data set features of T,TM,Tm,H,V and VM
Y2 <-- contains only one testing data set feature PM2.5
applying SVM model from the sklearn.svm (applying the linear kernel approach)
fit(X1,Y1)
calculate mean_absolute_error(Y2, abc .predict(X2))
The predict function will give the value of PM2.5
}

3) ***Random Forest*:** Random Forest is a supervised learning algorithm. In our model, we have used this algorithm because it builds multiple decision trees and merges them together to get a more accurate and stable prediction.

In our study, we have used random forest algorithm in order to measure the relative importance of each feature on the prediction i.e relative importance of various meteorological feature like Average Temperature('T'),Average Relative Humidity('H') and Average Wind Speed on level of 'PM2.5' which is taken as the target output. Sklearn libraries provides a great tool that measures a feature's importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest.

The algorithm applied is advantageous to us in the following ways :

1) Increasing the Predictive Power: The "estimators" hyper parameter, which is just the number of trees the algorithm builds. More number of trees means more accuracy in prediction.

2) Increasing the Models Speed: The "n_jobs" hyper parameter tells the engine how many processors it is allowed to use. If it has a value of 1, it can only use one processor. A value of "-1" means that there is no limit.

The pseudo-code for the implemented Random Forest algorithm in our model is as shown below:

Air Poll Predict Random Forest (Average Temperature(T),Maximum Temperature(TM),Minimum Temperature(Tm),Average Relative Humidity(H),Average Wind Speed(V),Maximum wind speed(VM),PM2.5)
{
X1 <-- contains training data set features of T,TM, Tm,H,V and VM
Y1 <-- contain only one training data set feature PM2.5
X2 <-- contains testing data set features of T,TM, Tm,H,V and VM
Y2 <-- contains only one testing data set feature PM2.5
applying Random Forest Repressor (estimators =10)
fit(X1,Y1)
calculate mean _ absolute _ error (Y2, abc. predict(X2)) * 100
The predict function will give the value of PM2.5
}
The "n_estimators" hyper parameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking averages of predictions. In our algorithm, we have used 10 estimators which mean the algorithm will build 10 trees in order to predict the output value.

B. Use of Sensors: There are different types of sensors which are used for collection of the atmospheric data such as Temperature sensor, Humidity sensor, Rain sensor, Gas Sensor etc. Different types of gas sensors are also available to collect the different gases from the road traffic emission such as CO2 sensor,NO2 sensor,SO2 sensor etc. We are using sensors for measuring Temperature, Relative Humidity and Wind Speed in real time of Ghaziabad region specifically and can use this batch processing of data for prediction of the pollutant "PM2.5", (which is assumed to be the major pollutant in the pollution level in air) by applying several machine learning algorithms like SVM and Random Forest.

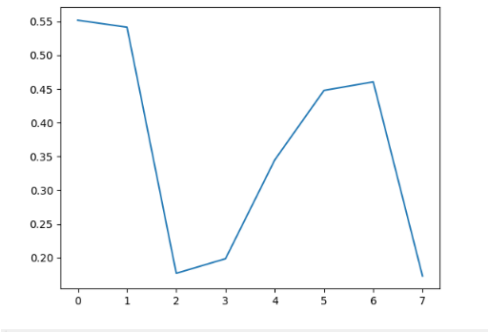
III- RESULT

Our air pollution prediction system predicts the pollution level of upcoming days by using Machine learning algorithm .The output is represented on the website and the users can view the digital value of air pollution and user can analyse it with a graph. It becomes very easy for us to rectify the pollution levels and air pollution around and plan for a healthy living and surrounding. The results of comparative study between different algorithm are also viewed in the form of a graph. Our system also predicts PM2.5 level based on some given meteorological factors.

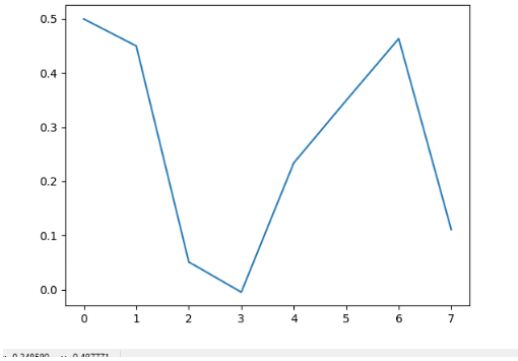
Our project is useful because it helps us to put a check on air pollution level by given meteorological features. The various machine learning algorithms are used to predict air pollution level and comparisons between each results is done. The algorithms used are Multi-linear regression, SVM (support vector machines) and Random Forest. The major challenge is to find out the relationship or dependency between the features of the data set. Advanced Technology and tools like ensemble learning can be used to get high accuracy in predicting pollution in air.

The output in the form of the graphs is given as:

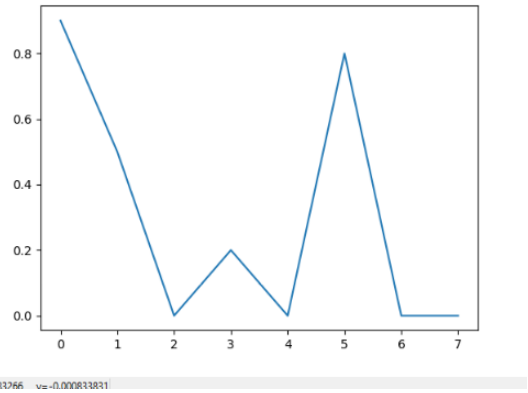
1) Representing Error in the form of graph between Predicted and Actual value after applying Multi-Linear Regression algorithm:



2) Representing Error in the form of graph between Predicted and Actual value after applying Support Vector Machine(SVM) algorithm:



3) Representing Error in the form of graph between Predicted and Actual value after applying Random Forest Algorithm:



IV- CONCLUSION

With the advancement of IoT infrastructures and Machine Learning Techniques Real-time air quality prediction and evaluation is desirable for future smart cities. Our study focuses on prediction of air pollution level of a particular or specific region. In air pollution prediction, model accuracy, efficiency and adaptability are key considerations.This paper reports our recent study in predicting air pollution level in Ghaziabad region, and compares result of different algorithms.

This system has the feature for the people to view the predicted amount of pollution on their mobile phones through websites. Letting civilians also involved in the process adds an extra value to it. As civilians are equally aware and curious about the environment, this concept of air pollution prediction system is beneficial for the welfare for the society. And it is implemented using the latest technology.The paper also highlights some observations on challenges and needs.

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the paper regarding Air pollution prediction system during B.tech Final Year.This project is supported and funded by CSTUP(Council of Science and Technology, UP) which is under the Department of Science and Technology, Govt. of U.P. (Reference link: https://www.cstup.org/). We would like to sincerely thank Ms.Sapna Yadav (Assistant Professor & Mentor) and Mr.Pankaj Agarwal (H.O.D) of Computer Science & Engineering Department,IMS Engineering College, Ghaziabad for her continuous support and guidance that led to the successful completion of our work and project for analysis and prediction of air quality.

REFERENCES

[1] *Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.*

[2] *Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340, 2004.*

[3] *Gu, Ke, Junfei Qiao, and Weisi Lin. "Recurrent Air Quality Predictor Based on Meteorology and Pollution Related Factors." IEEE Transactions on Industrial Informatics (2018).*

[4] *García Nieto, P.J., Sánchez Lasheras, F., García-Gonzalo, E. et al. "Estimation of PM10 concentration from air quality data in the vicinity of the major steel works site in the metropolitan area using machine learning techniques"Stoch Environ Res Risk Assess (2018), https://link.springer.com/article/10.1007/s00477-018-1565-6.*

[5] *Hu, Ke, Ashfaqur Rahman, Hari Bhrugubanda, and Vijay Sivaraman. "Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors." IEEE Sensors Journal 17, no. 11 (2017): 3517-3525.*

[6] *K. B. Shaban, A. Kadri, and E. Rezk, "Urban Air Pollution Monitoring System With Forecasting Models" IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606, Apr. 2016.*

[7] *Xiao Feng, Qi Li, Yajie Zhu, "Artificial Neural Network Forecasting of PM2.5 Pollution using Air Mass Trajectory based Geographic Model and Wavelet Transformation" Atmospheric Environment Journal, www.elsevier.com/locate/atmosenv, 2015.*

[8] *M. S. Baawain and A. S. Al-Serihi, "Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network," Aerosol and Air Quality Research, vol. 14, pp. 124–134, 2014.*

[9] *W.-Z. Lu and D. Wang, "Learning machines: Rationale and application in groundlevel ozone prediction," Applied Soft Computing`, vol. 24, pp. 135–141, Nov. 2014..*

[10]. *A. Sotomayor-Olmedo, M. A. Aceves-Fernández, E. Gorrostieta-Hurtado, C. Pedraza-Ortega, J. M. Ramos-Arreguín, and J. E. Vargas-Soto, "Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach," International Journal of Intelligence Science, vol. 3, no. 3, pp. 126–135, Jul. 2013.*