DOI: http://dx.doi.org/10.26483/ijarcs.v14i2.6972

ISSN No. 0976-5697

Volume 14, No. 2, March-April 2023

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# PREDICTION OF AIR POLLUTION AND AN AIR QUALITY INDEX USING MACHINE LEARNING TECHNIQUES

L. Ramesh,
Project Fellow,
Department of Computer Science,
University of Madras,
Chennai, 600025,India

S. Gopinathan
Professor & Principal Investigator,
Department of Computer Science,
University of Madras,
Chennai,600025,India.

***Abstract:*** Air pollution is the “world’s largest environmental health threat”, causing 7 million deaths worldwide every year. Its major constituents are PM2.5, PM10 and the harmful green house gases S02, N02, C0 and other effluents from vehicles and factories affecting not only humans but also other living organisms both on land and sea. The only effective solution to this global issue is to implement machine learning algorithms to predict the AQI (Air Quality Index) that can make the people aware of the condition of the air of a certain region such that certain actions could be issued by the government for the improvement of the air quality in the future. The prime objective behind this project is to predict the AQI based on the concentration of PM2.5, PM10, S02, N02, C0 as well as weather conditions like temperature, pressure and humidity .Hence the data set is combined from various web sources like cpcb and uci repository in order to bring accuracy in the prediction and to justify whether the Quality of air is suitable or not. This prediction will be brought about with the help of some supervised machine learning algorithms and the observation and the result will state which algorithm is giving better accuracy in prediction of AQI and which one is giving less error.

***Keywords:*** PM 2.5, machine learning algorithms, Air Quality Index, cpcb, prediction, accuracy.

## 1. INTRODUCTION

In this sophisticated era with the rapid growth of population and their demand, world has advanced in technology as well as industrialization. However, the dark side of this advancement is overlooked i.e. the unregulated emission of the harmful gases from vehicles, burning of fossil fuels as well as effluents from industries. This has lead to the global issue causing degradation of air quality called air pollution. Air Pollution refers to the release of pollutants into the air that are detrimental to human as well as other organisms. “According to the research by WHO[1] ( World Health Organization) approximately 7 million people die”[1] worldwide due to this global crisis of air pollution. Even though a lot of technologies are still working for it but still this crisis is affecting us globally.

The major constituent of such harmful gases are PM 2.5 (Particulate matter < 2.5 microns) and PM 10. They are the most hazardous ones for our health. The PM 2.5 [2] are the particles of size less than 2.5 microns which enter deeper in our lungs and cause various problems and issues like heart attacks, strokes, asthma etc and the PM 10 are the particles less than 10 microns (>2.5 microns) that affect the upper respiratory tract and cause nasal complications.

The Air Quality Prediction model intends to work on the concentration of various pollutants such as PM2.5, PM10,S02, N02, C0 and also on the weather conditions that also affects the AQI i.e. the Air Quality Index of a region scaling them to a range and defining whether it is healthy, satisfactory, moderate or unhealthy for the region. The various machine learning algorithms are applied after preprocessing the data and scaling the data properly.

## 2. LITERATURE SURVEY

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

Channappabhyri et al.,(2017)proposed of Predicting Trends in Air Pollution in Delhi using Data Mining. In this Research 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 Perception and Linear Regression [1].

Kularatna et al.,(2008) research work, 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].

Anupriya et al.,(2017) , Proposed research work 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].

Orru et al., (2018) research focused on, 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 data mining techniques. The time series data mining 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].

Buteau et al., (2018) explains, 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].

M.K.Woo et al.,(2018) Proposed work 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].

## 3. AIR POLLUTION PREDICTION SYSTEM

The air 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 the 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.

### 3.1. 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 [8]. As pollutants from vehicles, fireplaces, and industry are emitted into the air, the inversion traps these pollutants near the ground.

### 3.2. 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.

### 3.3. Relative Humidity

Humidity could affect the diffusion of contaminant.

### 3.4. 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.

### 3.5. 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 using work 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 [7]. 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.

## 4. PROPOSED ALGORITHM OF AIR QUALITY INDEX BASED COMPARATIVE ANALYSIS

1. Read the input dataset
2. Apply preprocessing techniques of Air pollution dataset.
3. To Apply model for training set and testing set level.
4. Check Air Quality Index for different Algorithms.
5. Find prediction score of machine learning algorithm
6. Predict mean squared error and mean absolute error of ML algorithm
7. End

## 5. DESCRIPTION OF THE WORK FLOW

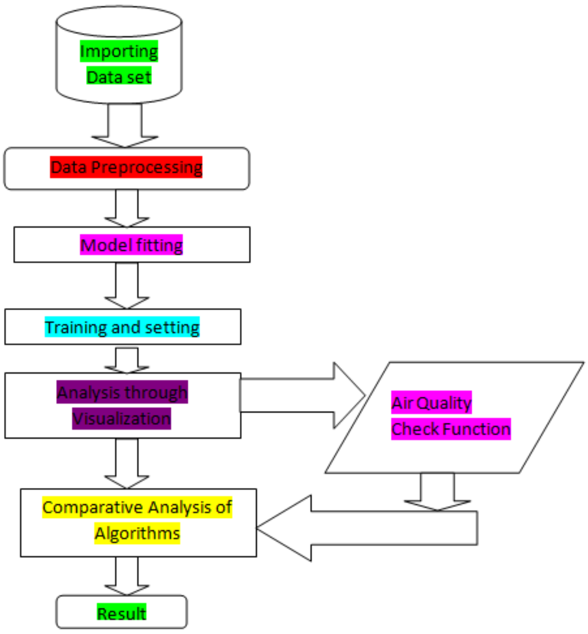
Importing Data set
Data Preprocessing
Model fitting
Training and setting
Analysis through Visualization
Air Quality Check Function
Comparative Analysis of Algorithms
Result

Fig 5.1: **Description of the Work Flow**

In the above block diagram describe overall work. The data was collected from various sources like Tamilnadu pollution control board and web links. The process is explained below. In this section, the present block diagram of Air Quality Index based comparative Analysis of Algorithms [13]. Fig 1 shows the general flow of the proposed scheme. The Air pollution of the input dataset is obtained. The preprocessing is a data to transform raw data into a standized data. To apply appropriate model by using Air quality index. Now testing and training set is applied. Finally to analysis through AQI for Machine learning techniques.

### 5.1 Dataset

The dataset has been referred from cpcb.nic.in as well as from the uci repository. The dataset plays a major role in accurate prediction also as the right features make the prediction more realistic and with least variance. The dataset has 1004 rows, 14 columns, the values of the features which may help in training of the model to successfully predict the AQI.

### 5.2 Preprocessing

Preprocessing is a data mining technique to transform our raw data into a standardized data, which would provide better result. We have often heard about "Garbage in Garbage Out ", the concept explains that the quality of the output and the prediction the will be expect depends largely on the quality of the input.

### 5.3 training and testing

The Dataset was splitted in such a way that 80% was training data and rest 20 % was testing data. It was done so that the model could be first trained and then could be tested on the testing data such that the accuracy score, precision and the error in prediction could be checked and proper results could be marked.

### 5.4. Analysis through Visualization

For proper understanding of the relation between the 'PM10 concentration on AQI ' and 'PM2.5 concentration vs AQI in the work drown the scatter plot which clearly shows that linear graph for the PM2.5 whereas, the PM10 has a little steep distribution i.e. the Gaussian distribution[9][12].

### 5.5 Comparative Analysis of Algorithms

After the Preprocessing and transforming the raw data into useful data, now split the dataset into two parts. One is the training part (comprising 80% of dataset) and the testing part (comprising of 20% of dataset). The splitting is done with the help of 'train_test_split' method from the sklearn. Model selection module.

## 6. RESULT AND DISCUSSION

To measure the performance of the enhanced system. In the research work we have tested the system with various methods and data sets. The Data sets are collected standard source from Tamil Nadu

### 6.1 Prediction Score

The prediction score is found using r2_score method for the test data and model. Score for training data. The various Algorithms and their Prediction score is as shown below:

| **Algorithms** | **Prediction Score** | |
|---|---|---|
| | **Training dataset** | **Testing dataset** |
| Multiple Linear Regression | 0.9506 | 0.9476 |
| Support Vector Machine | 0.9566 | 0.9578 |
| Decision Tree Regression | 0.9482 | 0.9349 |
| Random Forest Regression | **0.9998** | **0.9987** |

**Table 6.1 : Comparative Prediction score of both Training and Testing Dataset**

In Table 1, the experiments Air pollution of dataset was performed on total No. of Attributes 14 and total No. of Instances 1004. The input parameters used for the different machine learning algorithms. In the Multiple linear Regressions for prediction score are occurred 0.9506 training dataset and 0.9476 at the time of testing dataset. In Support Vector machine for prediction score are occurred 0.9566 of training dataset and 0.9578 of testing data set. In Decision tree machine for prediction score are occurred 0.9482 training dataset and 0.9349 at the time of testing dataset. In Random forest Regression for prediction score are occurred 0.9998 at the time of training dataset and 0.9987 at the time of testing dataset are performed.

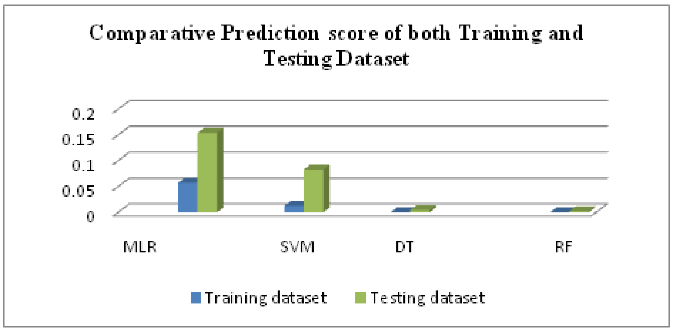

**Fig 6.1 Comparative Prediction score of both Training and Testing Dataset.**

The best accuracy in prediction is shown by Decision tree on the training dataset and Random Forest Regression for testing Dataset. The accuracy on Training dataset determines how well the model is trained. Whereas, the accuracy shown on testing dataset is the real prediction score to be taken into account. Hence best accuracy is shown by the Random forest Regression Algorithm [14].

### 6.2 Error in Prediction

| Algorithms | Error | |
|---|---|---|
| | Mean Squared Error | Mean Absolute Error |
| Multiple Linear Regression | 0.05776 | 0.15452 |
| Support Vector Machine | 0.01289 | 0.08328 |
| Decision Tree Regression | 0.00024 | 0.00469 |
| Random Forest Regression | 0.00012 | **0.00254** |

**Table 6.2 : Comparative analysis of Mean Squared Error and Mean Absolute Error**

In Table 2, the experiments Air pollution of dataset was performed on total No. of Attributes 14 and total No. of Instances 1004. The input parameters used for the different machine learning algorithms [11]. In the Multiple linear Regressions for prediction of Error Accuracy are occurred 0.05776 Mean squared errors and 0.15452 at the time of Mean Absolute error. In Support Vector machine for prediction of Error Accuracy are occurred 0.01289 of Mean squared errors and 0.08328 of Mean Absolute error. In Decision tree machine for prediction of Error Accuracy are occurred 0.00024 Mean squared errors and 0.00469 Mean Absolute errors. In Random forest Regression for prediction of Error Accuracy are occurred 0.00012 Mean squared errors and 0.00254 Mean Absolute errors are performed.

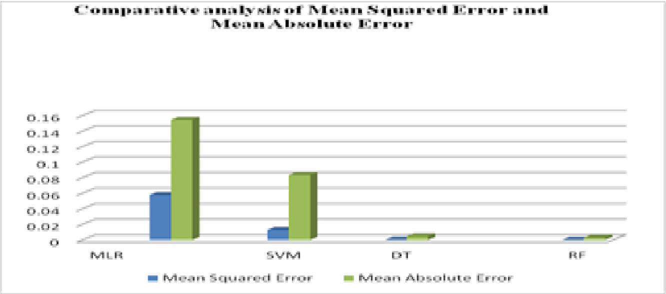

**Fig 6.2 Comparative analysis of Mean Squared Error and Mean Absolute Error**

Along with the Prediction score we need to also consider the error in prediction also. As the algorithm with minimum error stands to be better than others with greater error. We have Mean Squared Error that calculates the mean of the squared differences between actual value and predicted value [10]. Whereas Mean Absolute Error is the mean of the absolute difference between actual value and predicted value.

## 7. CONCLUSION

On the basis of all the observations, the Visualization graphs and Comparative Analysis of the Prediction and the error, the algorithm suited well for this prediction of air pollution is Random Forest Regression Algorithm as it provides an accuracy of **0.9998** on the testing data with the least 'Mean Squared Error '(MSE) of **0.00012** and ' Mean Absolute Error '(MAE) of **0.00254** . Mean Squared Error higher than Mean Absolute Error. Mean Absolute Error lower than Mean Squared Error. The above stated data is hereby collected from different source and experimented by the software programming R.

## 8. REFERENCES

[1] Dr. ChannappaBhyri, EliyazAhemad, "Design and Development of Industrial Pollution Monitoring System using LabVIEW and GSM," International Journal of Research and Scientific Innovation, Vol.4no. 7, 2017.

[2] N.kularatna and B.H.Sudhantha, "An environmental air pollution monitoring system based on IEEE 1451 standard for low cost requirements," IEEE Sensors journal,no. 8, pp.415-422, 2008.

[3] Anupriya V, "Smart Environmental Monitoring System using Labview," International Journal of Engineering and Computer Science, Vol. 6, no. 3, 2017.

[4] K. Orru, S. Nordin, H. Harzia, and H. Orru, "The role of perceived air pollution and health risk perception in health symptoms and disease: a population-based study combined with modelled levels of PM10," International Archives of Occupational and Environmental Health, March 2018.

[5] H. Orru, J. Idavain, M. Pindus, K. Orru, K. Kesanurm, A. Lang, and J. Tomasova,"Residents' self-Reported Health Effects and Annoyance in Relation to Air Pollution Exposure in an Industrial Area in Eastern-Estonia," Int. J. Environ. Res. Public Health, vol. 15, p. 252, February 2018.

[6] S. Buteau, M. Doucet, L. F. Tetreault, P. Gamache, M. Fournier, A. Brand, T. Kosatsky, and A. Smargiassi, "A population-based birth cohort study of the association between childhood-onset asthma and exposure to industrial air pollutant emissions," Environment International, vol. 121, Part. 1, pp. 23–30, December 2018.

[7] M. K. Woo, E. S. Young, M. G. Mostofa, S. Afroz, M. O. S. I. Hasan, Q. Quamruzzaman, D. C. Bellinger, D. C. Christiani, and M. Mazumdar, "Lead in Air in Bangladesh: Exposure in a Rural Community with Elevated Blood Lead Concentrations among Young Children," Int. J. Environ. Res. Public Health, vol. 15, Issue. 9, p. 1947, September 2018.

[8]Campbell-Lendrum, D., & Prüss-Ustün, A. (2018). Climate change, air pollution and noncommunicable diseases. Bulletin Of The World Health Organization, 97(2), 160-16.

[9]. Li, J., Li, X., & Wang, K. (2019). Atmospheric PM2.5Concentration Prediction Based on Time Series and Interactive Multiple Model Approach. Advances In Meteorology, 2019, 1-11

[10]. Soundari, M., Jeslin, M., & A.C, A. (2019). Indian Air Quality Prediction And Analysis Using Machine Learning. International Journal of Applied Engineering Research, 14(0973-4562), 1-6. Retrieved 22 July 2020.

[11]. C R, A., Deshmukh, C., D K, N., Gandhi, P., & astu, V. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends And Technology, 59(4), 204-207.

[12]. Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Journal Of Electrical And Computer Engineering, 2017, 1-14.

[13]. Sayyed, M., Sarode, A., Salunke, A., & Desai, S. (2020). A thorough Survey on prediction of Airpollution. Journal Of Emerging Technologies And Innovative Research, 7(3), 1-3. Retrieved 22 July 2020.

[14]. Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air Quality Prediction using Machine Learning Algorithms. International Journal Of Computer Applications Technology And Research, 8(9), 367-370.