International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Air Quality Index Prediction with Meteorological Data using Feature based Weighted Xgboost

NandigalaVenkatAnurag, YagnavalkBurra, S.Sharanya

Abstract—Over the recent years, air pollution or air contamination has become a concerning threat, being responsible for over 7 million deaths annually according to a survey conducted by "WHO"(World Health Organisation). The four air pollutants which are becoming a concerning threat to human health are namely respirable particulate matter, nitrogen oxides, particulate matter and sulphur dioxide. So efficient air quality prediction will enables mankind to foresee these undesirable changes made in the environment by keeping the pollutant emission under check and control.Machine learning algorithms are boon in these types of applications which demand high degree of human intervention ad computation. This work deploys feature based weighted XGBoost, that uses meteorological data and pollutant level to predict the Air Quality Index (AIQ) This model is tested to predict the AIQ of Velachery, a fast developing commercial station in South India and has shown remarkable decline in the error rate that its rivals.

Keywords— Machine Learning, Air Quality Index, Artificial Intelligence

I. INTRODUCTION

Air pollution is an alarming issue to mankind and earth. The global warming and depletion of ozone layer has caused adverse effects like rise in sea level, fast melting of glaciers in the Artic and Antarctic zones, rise in the temperature and unforeseen climatic changes. The pollution in air has seen a steep increase after the industrial revolution. The transition of population from rural to urban shifting is also seen as a prominent factor in increase in pollution level. Burning of fossil fuels like coal for domestic and industrial purposes makes incomplete combustion as a result of which the fine Particulate Matter(P.M 2.5) is suspended in the air [12]. This pollution emission worsens, when the winter sets in. Hence colder temperature aggregates air pollution during the winter. One of the earliest solutions was initiated by the government of Mongolia and the USA from 2013 to 2018 wherein they have been an issued a loan worth 15 million dollars from the World Bank. The solution was to replace these inefficient coal powered stove with the new clean combustion stoves. Due to the nature of these new stoves being issued to the citizens of Ulaanbaatar, they weren't widely accepted as the price of fuel shot up, it took more time to heat up the coils hence time-consuming process and the surface area of these stoves was relatively smaller versus the old traditional stoves. Further, attempts were made by the governments but only to result in vain and have lead to about, huge expenditure equivalent to 55 million USD. With the implementation of the machine learning techniques, this gives us the ability to forecast the air pollution levels in order to be able to alert crowd in cities if it was safe to step out especially during to winter, during which they exposed to a hazardous level of fine particulate matter (PM2.5) [13]. Table I shows the adverse effects of pollution in air.

Machine learning is a child of artificial intelligence that uses straightforward factual strategies, calculations, and current figuring power to predict the future outcome. It has found its application in almost all fields. Air quality prediction can be foreseen using machine leaning algorithms. The main hurdle in air quality prediction is the enormous number of factors that has to be considered. Each factor contributes to the pollution level in air its own unique way. Though many researches have been done in this domain, there is still a gap in identifying themost prominent factor that contributes to the air quality. These factors may vary from region to region. For instance in an industrial zone the PM 2.5 may be the most important factor but this may not be the case in rural areas.

This work focus on building an XGBoost model that considers meteorological data of Velachery, a commercial station in Tamil Nadu collected from the database of Central Control room for Air Quality for Air Quality Management. The model is built by considering the following elements in air: Carbon Monoxide, Benzene, Eth benzene, Nitrous oxide, MP Xylene, Nitrogen Oxides (Nox), Ozone, PM2.5, SO_2 and Toluene. Apart from the above mentioned pollutants, this model also considers the highly fluctuating meteorological parameters like pressure, relative humidity, wind speed, temperature and wind direction of the geographical region. Also, this work attempts to rank the most influential meteorological factor that highly contributes to the air quality. This information obtained from the comprehensive study of pollutant level and meteorological factors may be useful to mitigate the air pollution during the period.

Revised Manuscript Received on 30 May 2019.

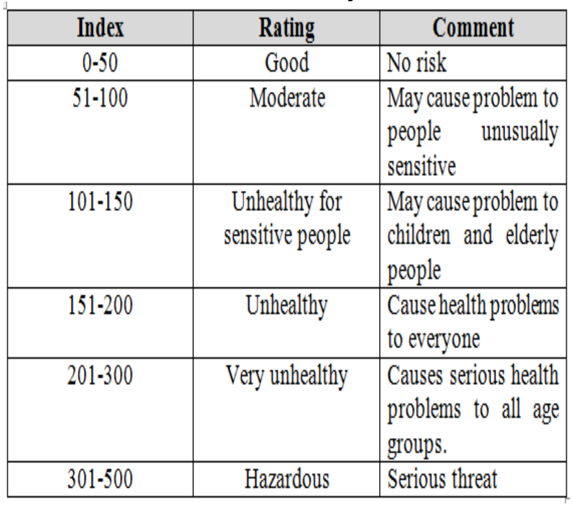
* Correspondence Author

NandigalaVenkatAnurag*, Department of Computer Science and Engineering, SRMIST, Chennai, India

YagnavalkBurra, Department of Computer Science and Engineering, SRMIST, Chennai, India

S.Sharanya, Department of Computer Science and Engineering, SRMIST, Chennai, India

Table I:Air Quality Index



Index	Rating	Comment
0-50	Good	No risk
51-100	Moderate	May cause problem to people unusually sensitive
101-150	Unhealthy for sensitive people	May cause problem to children and elderly people
151-200	Unhealthy	Cause health problems to everyone
201-300	Very unhealthy	Causes serious health problems to all age groups.
301-500	Hazardous	Serious threat

II. RELATED WORKS

Many works have been done in air quality prediction using machine learning and deep learning models. Some of the state of art techniques is discussed here.

Li et. al measured the air pollution using spatio-temporal interpolation methods [1]. Hsieh et al. used an framework based on affinity graph to deal eith real-time air quality of a geographical location and identifying the optimal locations to establish monitor stations [2]. Dong et built a hidden semi markov model predict PM2 .5 [3]. Thomas and Jacko deployed the basic regression and neural network to forsee the concentration of PM2 .5 in air [4]. A detailed feature analysis, was done by Zhou et. al through a probabilisticdynamic causal model that reveals the dynamic temporal dependencies of PM2 .5 [5]. Shabanet. al. devised a system for monitoring and forecasting air pollution to identify highly polluted area in a given city [6]. [7] fabricated a framework for estimating air pollution by considering traffic conditions and the available greenery. MdNazmulHoq e al. designed a mobile application to predict the asthma attack on highly populated cities [8]. Zhongang Qi designed a deep learning model based on interpolation to predict and analyse the air quality [9]. [10] discusses about deploying fusion based deep neural network that takes into account the spatial features to predict the air quality. Neural networks were also deployed to predict the air pollution [14][15].

The researches in air quality prediction are definitely a boon to mankind. But there are some research gaps that have to be focussed by the researchers. The above models consider only few features and the deep learning models may miss out or over weigh the features. Also prediction of air quality is of little use without giving the critical parameter that has to be monitored. Also the meteorological factors play a crucial role in prediction of air quality.

The proposed model is an attempt to alleviate the air quality prediction and monitoring. This model considers both pollutant level and meteorological parameters to predict the air quality. Also, the ranking of features clearly indicates the prominent factor that contributes to the pollution level in the air.

III. PROPOSED METHODOLOGY

The proposed work is based on XGBoost which is an ensembling model[11]. The novelty of the model is that the decision trees that were integrated to for a complete tree will be weighed based on the prominence of the feature being considered. The meteorological factors are ranked based on their contribution to the pollutant level. The factors are assigned weights between 0 and 1 based on the percentage of contribution.

The weighted or ranked features are span as decision trees in the ensemble XGBoost. Since XGBoost spans shallow interpretable trees, the prediction could be made very easily. The correlation among the factors also play a crucial role, since the weakly correlated features are given very lower significance.

For the data with m features, K additive functions and n examples, the output of ensemble tree is given as,

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

Where f_k ξ F.The pollutant levels are given into the tree, each pollutant will correspond to a function F. The final output prediction of the model with only XGBoost will be summation of the output of individual decision trees. The objective loss function for XGBoost is given by the following:

$$L(\phi) = \sum l(\hat{y}_i, y_i) + \sum \Omega(f_k)$$

The real pollutant and meteorological data is sparse. The feature ranking is made on the meteorological parameters using the Roulette wheel method and the weight for each feature is assigned based on its significance. The weight w_i : $R \rightarrow [0,1]$. The weighted meteorological parameters are applied on each additive function to yield the final prediction of the model. Now the loss function is modified as

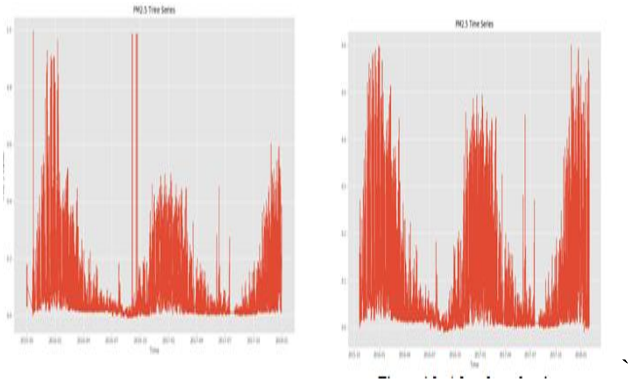
$$\hat{L}(t) = \sum_{i=1}^{n} w_i[g_i f_t(x_i) + 0.5 * h_i f_t^2(x_i)] + \Omega(f_t)$$

Where g_i is the differential value of the actual and predicted output of first order gradient and h_i is the second order gradient. The model could be built only using first order gradient function at the cost of convergence speed. The models output function is given below

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} w_i f_k(x_i)$$

IV. RESULTS AND DISCUSSION

The proposed model is deployed to predict the air quality index of the data obtained from Central Control room for Air Quality Management. Velachery which is a commercial station in Tamil Nadu served as testbed for the model. The data contains missing values, eradication of outliers and NaN data. The data cleansing for the PM2.5 is shown in figure 1.



a) Before cleansing **b) After cleansing**

Figure 1: Data cleaning

A. Exploratory Data Analysis (EDA)

The pollutant levels and the meteorological factors [12] are closely related to one another, it is important to perform EDA over the data to figure out the prominent feature that plays a pivotal role in the Air Quality Index Prediction (AQI). The important findings from the EDA (Figure 2) are listed below:

- As the wind speed increases the AQI decreases.
- Winter months foresee large values of AQI.
- The spikes in data can be observed during peak traffic hours from 9. A.M. till 5 P.M.

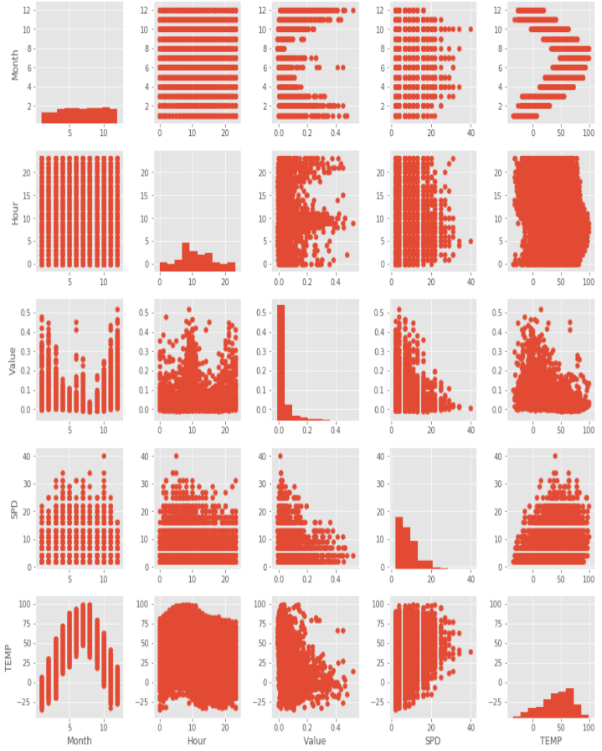


Figure 2: EDA for features

The findings from the EDA can be augmented information for AQI prediction and for ranking the features. The feature ranking is done through Roulette wheel selection method and the result of it is given in Figure 3.

Figure 3: Ranking Features

It is evident that wind speed (indicated as value 1 in figure 3) is a more dominant meteorological parameter that makes significant contribution to the AIQ prediction. This fact is also confirmed in EDA of the data. The next important factor influencing the AQI is the temperature. The weightage for each factor is estimated from the length of the bars.

The model is then trained with 80% of dataset and tested on 20% of the dataset. The resuls are shown in Figure 4.

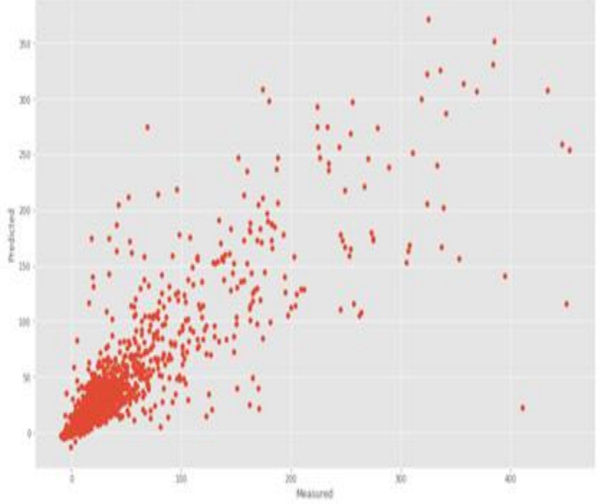


Figure 4: Feature based Weighted XGBoost

The proposed model is compared with other state of art machine learning algorithms like neural networks, decision tree and multiple linear regression in terms of Root Mean Square Error (RMSE) value. This model shows a low error rate of 15.97 against its rivals (given in Table II).

TABLE II Comparison of RMSE

Model	RMSE
Neural Networks [15]	24.14
Decision Tree	16.84
Multiple Linear Regression	18.72
XGBoost	15.97

V. CONCLUSION AND FUTURE ENHANCEMENTS

The alarming problem of air pollution caused serious changes to the earth. Hence to monitor the pollutant level feature based XGBoost model is deployed which considers the meteorological data along with pollutant level of Velachery in order to make the forecast of AQI. The model shows a reduced error rate when compared with other machine learning algorithms and also ranks the metrological factors based on their order of importance. The model resulted in lower RMSE values which makes it suitable for real me AIQ prediction.The model can be further extended to predict the AQI in a wider geographical area by augmenting additional factors.

REFERENCES

1. L. Li, X. Zhang, J. Holt, J. Tian, and R. Piltner, "Spatiotemporal interpolation methods for air pollution exposure," in Symposiumon Abstraction, Reformulation, and Approximation, 2011.
2. H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15, 2015, pp. 437– 446.
3. M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, and D. Kenski, "PM 2.5 concentration prediction using hidden semi-markovmodel-based times series data mining," Expert Syst. Appl., vol. 36, no. 5, pp. 9046–9055, Jul. 2009.
4. S. Thomas and R. B. Jacko, "Model for forecasting expressway pm2.5 concentration – application of regression and neural network models." Journal of the Air & Waste Management Association, vol. 57, no. 4, pp. 480–488, 2007.
5. X. Zhou, W. Huang, N. Zhang, W. Hu, S. Du, G. Song, and K. Xie, "Probabilistic dynamic causal model for temporal data," in Neural Networks (IJCNN), 2015 International Joint Conference on, July 2015, pp. 1–8.
6. Shaban, Khaled Bashir, Abdullah Kadri, and EmanRezk. "Urban air pollution monitoring system with forecasting models." IEEE Sensors Journal 16, no. 8 (2016): 2598-2606.
7. Jain, Varun, MansiGoel, MukulikaMaity, VinayakNaik, and RamachandranRamjee. "Scalable measurement of air pollution using COTS IoT devices." In Communication Systems & Networks (COMSNETS), 2018 10th International Conference on, pp. 553-556. IEEE, 2018.
8. MdNazmulHoq, RakibulAlam, Ashraful Amin, "Prediction of possible asthma attack from air pollutants: towards a high density air pollution map for smart cities to improve living", International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.
9. Zhongang Qi, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li ∗, Zhongfei (Mark) Zhang, "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality", IEEE Transactions On Knowledge And Data Engineering, 1041-4347, 2018.
10. Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, Yu Zheng, "Deep Distributed Fusion Network for Air Quality Prediction", In Proceedings of KDD'18, London, United Kingdom, August 19-23, 2019.
11. Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", ACM, KDD , August 13-17, 2016.
12. JaakkoKukkonen, LeivHårvardSlordal, RanjeetSokhi, " Analysis and evaluation of European air pollution episodes", Meteorology applied to urban air pollution problems, ISBN 954-9526-30-5, Demetra Ltd Publishers, Bulgaria, pp. 99-114, 2005.
13. Perez, P., Trier, A., and Reyes, " Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile", Atmospheric Environment, 34:1189–1196, 2005.
14. Gardner, Dorling and S.R. ,"Neural network modelling and prediction of hourly NOXand NO2concentrations in urban air in London", Atmospheric Environment 33(5), 709-719, 1999.
15. Cannon, A. J, " Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (CaDENCE)", Computers and Geosciences, 41:126–136, 2012.