NVEO
Natural Volatiles & Essential Oils

Prediction of Air Quality Index Using Machine Data Learning on Atmospheric

M. Mallegowda[1], Dr. Anita Kanavalli[2], Yash Verma[3], S Jaya Krishna Vamsi[4], Tata Mukesh[5] Vedant Saxena[6]

[1] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*
[2] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*
[3] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*
[4] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*
[5] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*
[6] *Department of Computer Science, Ramaiah Institute of Technology Bangalore Affiliated to VTU, India*

Abstract

Air Quality Index (AQI) is a standard measure of pollutant levels such as those of PM2.5, PM10, SO_2, NO_2, O3, CO, NH3 and Lead over a period. Nowadays, the atmosphere is getting polluted rapidly. So, there is an urge to know how the air is going to be around us in the near future. We are implementing an interactive and user-friendly web application, where users find out the AQI and also predict PM2.5 AQI for the next day. We have trained a XgBoost model. The application is also capable of plotting graphs of temperature, humidity and AQI. These graphs can be visualized based on monthly or weekly data. In this paper, we discuss how data was collected and cleaned, how feature engineering was applied as part of pre-processing and finally what all models we tried and came to the conclusion of XgBoost being the most suitable to our use-case.

Keywords: Air Quality Index (AQI); XgBoost; PM2.5; Machine Learning; Feature Engineering

Introduction

Air pollution and its prevention pose a major problem in our day-to-day life. Air pollution is one of the main causes of human's respiratory and cardiovascular system problems and causes an increased mortality rate as well. They increase the risk of diseases among the people. Many efforts from both local and state government are done in order to understand and solve the problem of air pollution aiming to improve public health. However, most of these measures are taken when the situation has worsened and there is no coming back. There have been loads of effort put by different organizations post the situation of severely bad air quality, however no one has tried to prevent this issue before-hand. As it is commonly known, "Better safe than sorry", this project will aim more at preventing the situation than at solving the problem of air quality after it occurs.

Related work

The paper [1] delved into a machine learning model based on Artificial Neural Network and they tested it using MATLAB. They have used the language R for writing their code. They are using historical AQI values to find out the AQI values for the future. This paper considers region location, date, time of the day, SPM values for training the machine learning model. Their analysis is only based on previous year values of AQI and not any other type of parameters or sensor data like temperature, humidity etc. Also,

they have no UI or user interface. The paper concluded with a machine learning model coded in R, running on MATLAB.

We have seen in paper [2] that the researchers have focused on calculating Air Quality Index (AQI) using various different machine learning models like Naïve Bayes, Random Forests, Decision Trees, etc. The parameters in the dataset include readings like previous values of NO2, SO2 and SPM readings. The dataset mostly focuses on the readings from a specific region, in this case it is Hyderabad, Andhra Pradesh. The paper also specifies the distribution of AQI values into different categories like “good”, “satisfactory”, “moderately polluted”, “poor” and “very poor”. The paper provided the accuracy of the different models when predicting the AQI values and found out which is the most accurate one. The paper has provided a comparison of various models which are working only for a specific region. Also, they only have a backend model to support their research and have not implemented an app to provide the real time values of AQI. We plan to make a general application which is not region exclusive, and also is easy for general users to use. From this paper we came to an understanding that non-linear machine learning models perform well on these kinds of dataset.

This paper [3] analyses the AQI levels across different regions of India and compares them using suitable statistical tools. They have further analyzed the causes of air pollution which contaminates air. In the end they have also tried to predict the AQI for the future years using the regression analysis method. Their research is more focused toward SO2, NO2, RSPM and PM values from 2015 data. They have categorized the AQI levels of different regions into defined divisions of quality of air, namely “Good”, “Moderate”, “Unhealthy for Sensitive Groups'', “Unhealthy”, “Very Unhealthy” and “Hazardous”. The paper also tried to predict future values of AQI for different regions using some statistical formulas. This paper was more focused on analyzing the present values of AQI and categorizing them into predefined categories of quality as mentioned before. However, they have not worked much in detail about the prediction model for finding out the future values of the AQI for different regions. They provided a statistical summary and plots regarding the AQIs but no interface for users.

After going through various research papers, we found that while some of the implementations had no well-defined user interface which is not user friendly, some other implementations had not been very accurate. Most of the implementations used only the previous values to form a time series or used artificial neural networks to train their model, however no consideration of the present atmospheric conditions were made. None of the research has applied hyper-parameter tuning as well. So, in this project we try to overcome these shortcomings and implement accordingly in order to achieve a more suitable and accurate model.

Methodology

There are 8 different kinds of AQI values depending on the type of pollutant chosen for calculation. The final AQI value is the value of that kind of AQI which is the worst in terms of air quality. More than 90% of the times that AQI is the PM2.5 AQI. Hence, in our approach we plan on predicting the PM2.5 AQI. PM2.5, also known as Particulate Matter 2.5 refers to the particles that have a diameter of less than 2.5 micrometers. For a better understanding, it is 100 times thinner than human hair. The most prominent source of PM2.5 is burning of fuel and chemical reactions that happen in the atmosphere.

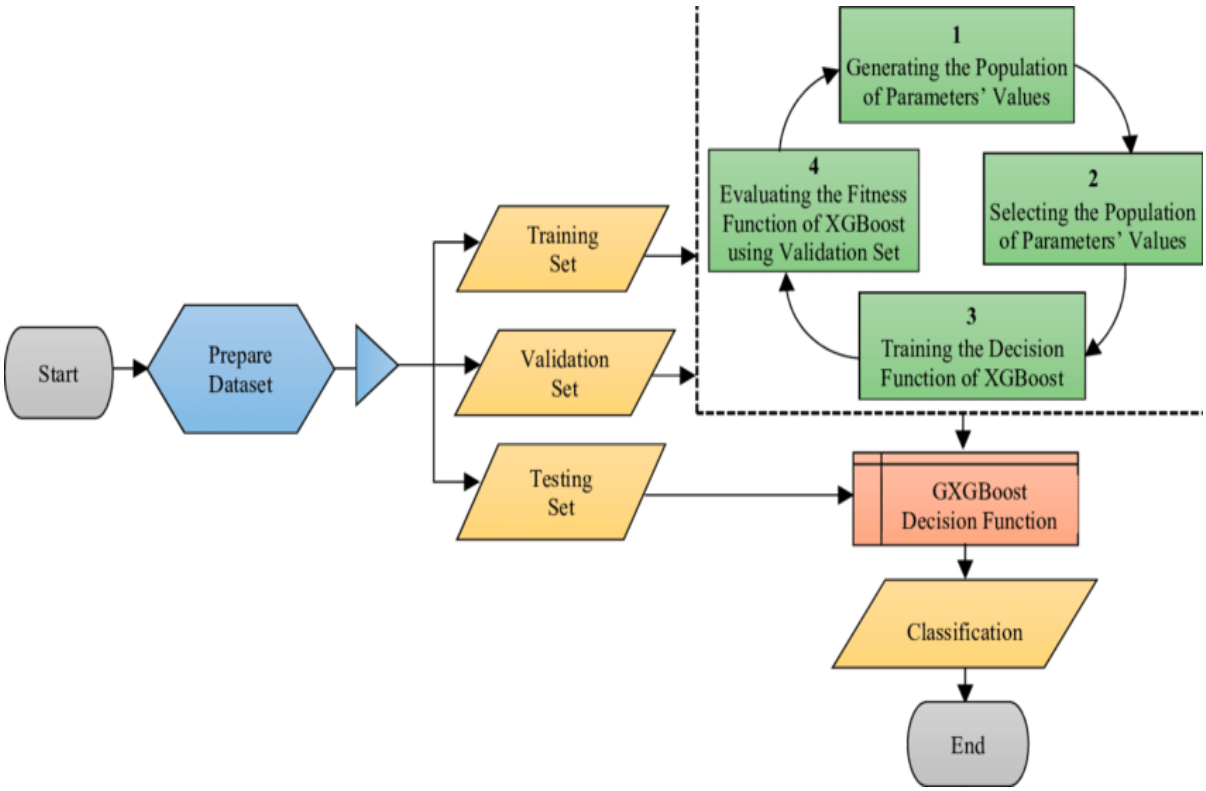
To begin with our model, the first and the most important step was the collection of the appropriate data. We collected meteorological data and previous years AQI values from January, 2015 to April, 2021 summing up to around 6 years of data. Web scraping was leveraged to perform data collection. The data was pre-processed into csv format and then cleaned properly by removing null and in-appropriate values. Since it is labelled data, supervised learning is used to train the model. Exploratory data analysis was performed on the data using various plots and other visualizations. This led us to the conclusion that a non-linear model is the way to go. Literature surveys on similar papers also suggested use of tree algorithms, hence we began training our model using Decision Trees and Random Forests Models.After achieving decent results, we moved onto a more advanced tree model called XgBoost and obtained an even better accuracy after hyperparameter tuning. The entire process from collecting data to training the model has been summarized in the following steps along with an algorithmic representation in Figure one.

1. Step 1 (Data Acquisition and pre-processing) : Past 6 years data has been collected through web scraping. The collected data has attributes of Average Temperature(T), Maximum Temperature (TM), Minimum Temperature (Tm), Atmospheric pressure at sea level in hPa (SLP), Average relative humidity in % (H), Total rainfall or snowmelt in mm (PP), Average visibility in Km (VV), Average wind speed in Km/h (V), Maximum sustained wind speed in Km/h (VM), Maximum speed of wind in Km/h (VG) and has predictor variable PM2.5(Particulate Matter) AQI. Rows with Null values for the predictor variable PM2.5 have been dropped. The VG column had to be dropped out of analysis because the majority of the data was missing from that column.

2. Step 2 (Exploratory Data Analysis): After cleaning the data, various plots were visualized for better understanding. Seaborn and matplotlib libraries were used. The plotting of the heatmap and importance level of all features along with the predictor variable PM2.5 AQI helped us understand how each feature was related to the predictor variable. We moved on with all columns due to correlation with the predictor variable. After plotting pair-plots between each feature and predictor variable, we could see that there was no linear relation between the features and predictor value. Hence, we proceeded with non-linear models.

3. Step 3 (Splitting Data) : All combined Data have been split to train data and test data in 7:3 ratio and with a random state of 43.

4. Step 4 (Training the model): Fit the training data by calling XGBRegressor over training data. XGBRegrssor is imported from the XgBoost module.



5.

6. Figure 1. XgBoost Algorithm Flowchart

7. Step 5 (Hyperparameter Tuning): Find the best values of parameters with help of the Sklearn model selection module. A pickle file is made after finding the best estimator. The pickle stores the model for faster prediction in the future.

8. Step 6 (Prediction) : The current weather data is collected through API's and passed to the trained model. The model returns the predicted PM2.5 AQI value for the next day.

9. The AQI values are categorized into different categories based on fixed ranges of value. The figure two is a representation of the AQI value ranges used in India. Figure three represents a mapping between PM2.5 concentration values to PM2.5 AQI values. We have categorized the AQI levels by following these standards.

Category	AQI (no units)	24-hour average PM 2.5 concentration (microgram/m^3)
Good	0-50	0-30
Satisfactory	51-100	30-60
Moderate	101-200	60-90
Poor	201-300	90-120
Very poor	301-400	120-250
Severe	401-500	250-380

Figure 2. AQI Category Ranges

Description	AQI	PM10 µg/m³ 24 hr avg	PM2.5 µg/m³ 24 hr avg
Good + Satisfactory	0-100	0-100	0-60
Moderate	101-200	101-250	61-90
Poor	201-300	251-350	91-120
Very Poor	301-400	351-430	121-250
Severe	401-500	431-550	251-350

Figure 3. PM2.5 concentration to AQI conversion

Results and Discussion

Out of all existing pollutants PM2.5 is the primary pollutant. Also, this concentration value has more significance in determining the air quality. This pollutant reduces the visibility of the atmosphere and sunlight. So, this is a great danger to humans and their health. Therefore, there is an urge to accurately predict PM2.5 concentration. So that the necessary measures can be taken by the government and precautions will be taken by the public.

The main aim to take up this project was to predict the PM2.5 AQI values of a city using environmental conditions. The parameters we considered to predict PM 2.5 are T, T_M, T_m, H, PP, VV, V, VM and trained a ML model with XgBoost technique to predict the PM2.5 AQI on the real-time environment values pulled from public API's.

In general regression, each data record is given the same weightage. All data records are given the same and equal importance. But this may not help in all scenarios. Sometimes, some records may be given higher weightage. So, boosting techniques can be used here so that some records get higher weightage.

When decision trees are used, many times the tree grows sequentially and at some points the weak learners are encountered. So, to avoid this we also use gradient boosting. In this approach, by giving weightages to the weak learners, they can be made into strong learners. So, at each level the tree is growing by learning from the previous tree. So, this is the motivation in choosing XGBOOST (Extreme gradient boosting) approach in training our ML model.

Performance of our model was based on errors between predicted and actual value on test data. We calculate the error by taking the difference between the predicted variable and actual variable. Since it is regression the error metrics that are tested on are:

Mean Squared Error (MSE): Popular error metric for regression models.

$$MSE = [\sum (y_i - \hat{y}_i)^2)]/ N. \quad (1)$$

In Eq. (1) y_i being actual vale and $\hat{y}_i$ being predicted value

Root Mean Squared Error (RMSE): RMSE is an extended version of MSE.

RMSE = sqrt ([$\sum (y_i - \hat{y}_i)^2$)] / N) (2)

In Eq. (2) y_i being actual vale and $\hat{y}_i$ being predicted value

Mean Absolute Error(MAE) = ($\sum$abs ($y_i - \hat{y}_i$)) / N (3)

In Eq. (3) y_i being actual vale and $\hat{y}_i$ being predicted value

The obtained values when the trained XgBoost model tested against these metrics:

Mean Absolute Error: 4.562

Mean Squared Error: 89.56

Root Mean Squared Error: 9.46

We have ended up with XgBoost technique only after trying with decision trees and random forests. Order of accuracy among all three models was XgBoost > Random Forests > Decision Trees. When decision trees were used, the feature split happened based on the GINI Index and information gain on the features.

The results obtained from when using Decision trees, random forests and XgBoost algorithms were compared on the basis of Mean Absolute Error, Mean Square Error and Root Mean Error. Graphs were plotted from the values of the errors obtained in each of the algorithm's implementations to better visualize the difference between those algorithms. The Graphs plotted are shown below

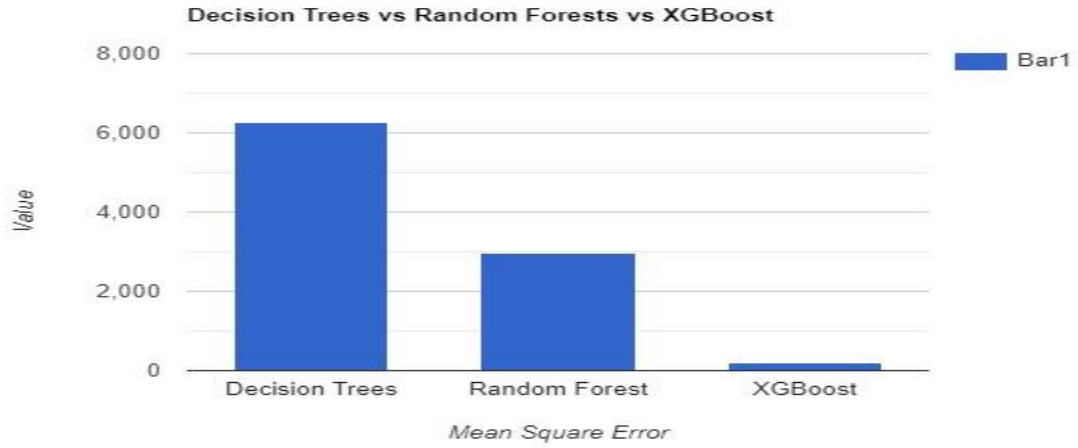


Figure 4. Comparison of MSE among three models

The obtained values when the trained Decision trees model tested against these metrics :

- Mean Absolute Error: 60.56
- Mean Squared Error: 6257.30
- Root Mean Squared Error: 79.10

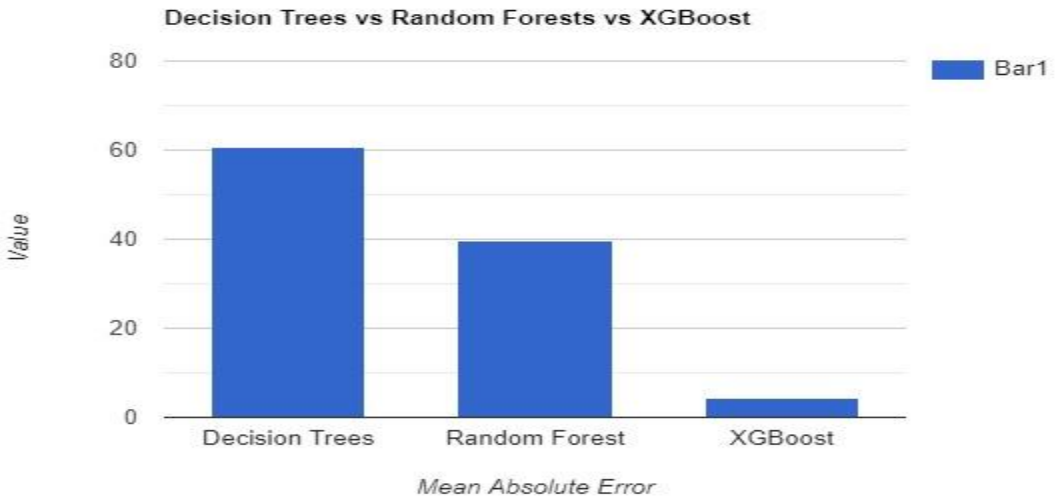


Figure 5. Comparison of MAE among three models

The obtained values when the trained Random Forests model tested against these metrics:

- Mean Absolute Error: 39.83
- Mean Squared Error: 2970.54
- Root Mean Squared Error: 54.50

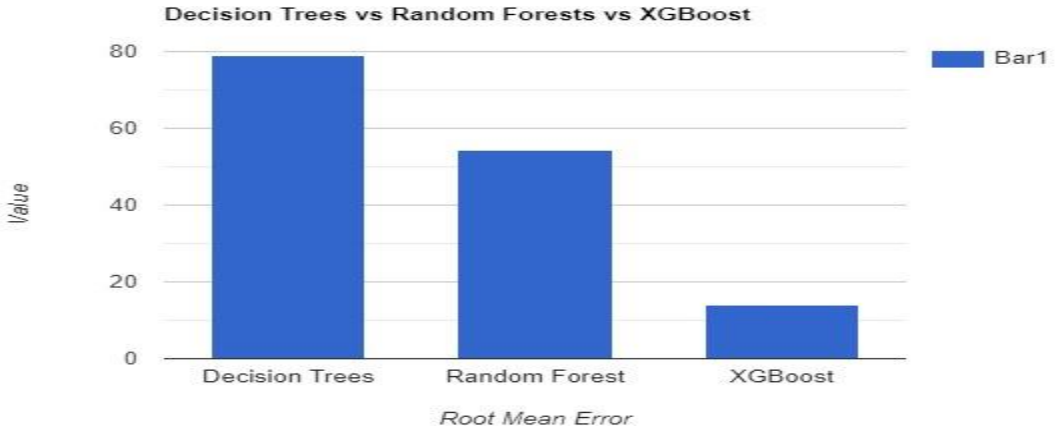


Figure 6. Comparison of RME among three models

Conclusion and Future Scope

We succeeded in developing an interactive web application where users and organizations can register successfully and get to know the predicted AQI Values for the next day. We were able to implement the visualization of AQI through graphs, and also for humidity and temperature both monthly and weekly. Users can also visualize the graphs both in line charts and bar graphs. They can also know the AQI value for the next and suitable measures are provided to them to take preventive measures and stay healthy.

In the future, we can extend this project to predict the AQI for some major centers within large cities and also for some major cities in the world. There is also a scope for registered users to get message or email alerts of the predicted AQI values for their next day so that they need not open the application every day. We can also extend this to provide an interactive map to the users where the user gets to tap a city to know the AQI value of that city.

There's an even interesting point of flaw in our model that the data used for training doesn't consider the covid and lockdown situation in India. So, the quality of the model could be further enhanced by using some more features like these.

References

1. Pooja Bhalgat, Sejal Pitale and Sachin Bhoite, "Air Quality Prediction using Machine Learning Algorithms" 2019 International Journal of Computer Applications Technology and Research Volume 8- Issue 09, 367-370, 0219, ISSN:-2319-8656
2. Krishna Chaitanya Atmakuri and Dr. K V Prasad, "A ComparativeStudy on Prediction of Indian Air Quality Index Using Machine Learning Algorithms" in Journal of Critical Reviews ISSN- 2394-5125, Vol 7, Issue 13, 2020. https://dx.doi.org/10.31838/jcr.07.13.058
3. Nikila Varshini.E, Sreeha.MR, Lhavanya Roobini. VN, Vijayarangam.J, Sujithra.M, "Analysis of Air Quality Index". Coimbatore Institute of Technology, Conference Paper July, 2018. https://doi.org/10.1155/2018/6421607
4. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072
5. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue) © Research India Publications. http://www.ripublication.com
6. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Web Site: www.ijettcs.org Email: editor@ijettcs.org Volume 7, Issue 1, January - February 2018 ISSN 2278-6856
7. Wei Jiang1 , Yandong Wang1 *, Ming-Hsiang Tsou2‡ , Xiaokang Fu1‡, "Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter)," in PLOS ONE | DOI:10.1371/journal.pone.0141185
8. Yves Rybarczyk 1,2 and Rasa Zalakeviciute 1,3,*, "Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review " in Appl. Sci. 2018, 8, 2570; doi:10.3390/app8122570
9. Colin Bellinger1*† , Mohomed Shazan Mohomed Jabbar1†, Osmar Zaïane1 and Alvaro Osornio-Vargas2, "A systematic review of data mining and machine learning for air pollution epidemiology" in Bellinger et al. BMC Public Health (2017) 17:907
10. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," International Journal of Environmental Science and Development vol. 9, no. 1, pp. 8-16, 2018
11. İ. Kök, M. U. Şimşek and S. Özdemir, "A deep learning model for air quality prediction in smart cities," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 1983-1990, doi: 10.1109/BigData.2017.8258144.

12. Liu, Huixiang & Li, Qing & Yu, Dongbing & Gu, Yu. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. Applied Sciences. 9. 4069. 10.3390/app9194069.