International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

AIR QUALITY PREDICTION USING MACHINE LEARNING

Pallavi Singh[1], Yashashwini R[1], Srinidhi Kulkarni[2] and Saravana M K[2]

[1]Under Graduate Student, [2]Associate Professor,
Dept. of Computer Science & Engineering
Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

DOI: http://dx.doi.org/10.52267/IJASER.2021.2405

ABSTRACT

In this paper, it is aimed to predict the Air Quality Index (AQI) by the use of Machine learning algorithms. To reach this, the key parameters have been selected which can affect the Air quality index are temperature, humidity, pressure, wind speed, PM10 and SO2 respectively. Air quality of certain states in India can be used as one of the major factors determining pollution index also how well the city's industries and population is controlled. Urbanized Air quality monitoring has been a constant challenge with the advent of industrialization. Air pollution causes conspicuous damage to the environment as well as to human health resulting in acid rain, heart diseases, global warming and skin cancer to all humankind. This paper addresses the challenge of predicting the Air Quality Index (AQI), with the goal to reduce the pollution before it gets unfavourable and also suggests mankind to move places in advance, using ensemble techniques for predicting the Air Quality Index (AQI). This paper investigates how effective some available prediction models are in predicting the Air Quality Index (AQI) values provided some input data, based on the pollution and meteorological information in India. We carry out regression analysis on the dataset, and our results shows which meteorological factors impact the AQI values most and how helpful the predictive models are to help in air quality prediction.

KEYWORDS: Machine learning, air quality index, random forest regressor, AQI

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

Air Quality Index - Particulate Matter	
301 – 500	Hazardous
201 – 300	Very Unhealthy
151 – 200	Unhealthy
101 – 150	Unhealthy for Sensitive Groups
51 – 100	Moderate
0 – 50	Good

1. INTRODUCTION

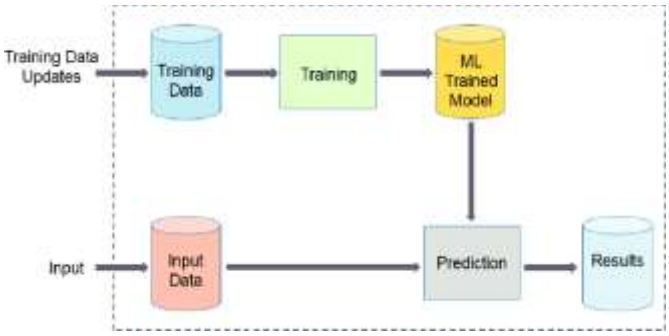
Air is one of the most essential natural resources for the existence and survival of the entire life on this planet. As this is the largest growing industrial nation, as known India is producing record amount of pollutants specifically Co2, pm2.5 etc and other harmful aerial contaminants. Air pollution in cities has become a cause for fear and has been a major topic of concern. All forms of life including plants and animals depend on air for their basic survival. The Indian air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants on the atmosphere. The main causes associated with air pollution are the burning of fossil fuels, agriculture, exhaust from factories and industries, residential heating, and natural disasters. We collect the data from the Indian government database and start calculating the individual index of the pollutant for every available data points and find their respective AQI for the region. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The prediction involves use of specialized algorithms. By predicting the air quality index, we can backtrack the major pollution causing pollutant and the location affected seriously by the pollutant across India. Some of the other environmental effects of air pollution are haze, eutrophication, and global climate change. By this we can extract various techniques to obtain heavily affected regions on a particular region. PM2.5 refers to tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. This can be determined based on a data set consisting of daily atmospheric conditions. In present scenario, Indian cities are failing in maintaining the WHO guidelines for monitoring outdoor air pollution for safe levels. The levels of PM10 and PM2.5 are also increased in air and reduced the air quality, results in adverse effect on living beings. This system exploits machine learning models to detect and predict the data set consisting of atmospheric conditions. Air pollution can cause long-term and short-term health effects. It's found that the elderly and young children are more affected by air pollution. The advancement in technology and research,

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

Alternatives to traditional methods have been proposed for evaluation and predications which use of machine learning approaches. Air quality evaluation is an important way to monitor and control air pollution. Air Quality Index (AQI), is used to measure the quality of air. One of the reason why we choose machine learning is to predict air quality index, this ability of adapting of machine learning algorithms. Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high. The air quality of a particular city selected by the user and groups it into different categories like good, satisfactory, moderate, poor, very poor, severe based on AQI (Air Quality Index).

2. METHODOLOGY



Challenges of collecting data and cleaning it.

Data collection is the most important part of data science from the integrity and correct of data is important to any kind of data work. The criteria that followed for data cleaning are:

• Validity of data: The data that is obtained should be relevant to the study.

• Uniformity of data: The data obtained from different resources should ideally be of same type.

• Consistency: The data collected should have same format.

The data that was used for the study is from government of India and some of the other 3rd party websites which provide the history of air quality data for different pollutants.

A. Data analysis process.

Data analysis is a process of collecting and organizing data to draw helpful conclusions from it. This process of data analysis uses analytical and logical reasoning to gain information from the data.

The objective of data analysis is to find meaning in data so that the derived knowledge can be used to make proper decisions.

International Journal of Applied Science and Engineering Review

IJASER

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

Methods of data analysis:

1. Collaborate your needs

Begin to analyse data or drill down into any analysis techniques, it's crucial to sit down collaboratively with all key, decide on your primary campaign or strategic goals, and gain a true understanding of the types of insights that will best benefit progress or provide with the heights of vision needed to evolve.

2. Establishing questions.

Once core objectives are outlined, consider which questions will need answering to help achieve the mission. Data analysis is the most important data analytics techniques that will shape the very foundations of success.

3. Setting up KPI's

As the data is ready, started to gather the raw data considering to offer potential value, and established clear-cut questions for insights to answer, a host of key performance indicators (KPIs) that will help track, measure, and shape progress in a number of key areas. KPIs are important to both analysis methods in qualitative and quantitative research. KPI is one of the primary methods of analysing data.

3. MODELLING AND ANALYSIS

Linear regression:

Linear Regression is an algorithm based on the machine learning are depends on supervised learning which performs a regression task. Depending on independent variables linear regression gives a target prediction value which is most likely used for finding the relationship among variables and forecasting. Depending on the connection among the established and the independent variables, different regression models differ, they are being considered and list of independent variables used.

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

$y = mx + c$

In the above expression y indicates labels to data and x indicates the input training data (input parameter). Value of x is used to predict the value of y which gives best fit line for finding the best m and c values during training the model.
c = intercept
m = slope of line

When the best m and c esteems, the best fit line. So, when long last utilizing model for expectation, it will foresee the estimation of y for the information estimation of x.

Random Forest Regressor:
Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The concept behind random forest is a simple but powerful one.

Need for random forest to perform well are:
There must be some actual signal in our features so that models built using those features do better than random guessing discretized levels. The process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently is inaccurate. Other researchers have worked on predicting concentrations of pollutants. It focuses on learning multiple tasks that have commonalities that can improve the efficiency and accuracy of the models. A variety of regularizations can be utilized to enhance the commonalities of the related tasks, including the nuclear norm, spectral norm, Fresenius norm, and so on. However, most of the former machine learning works on air pollutant prediction did not consider the similarities between the models and only focused on improving the model performance for a single task. Therefore, we decided to use meteorological and pollutant data to perform predictions of hourly concentrations on the basis of data models.

4. RESULTS AND DISCUSSION

The results were obtained from trying three different models that are related to time series model. Each model is evaluated with metrics root mean square error (RMSE). Root mean square error is the metrics used in every regression problem. RMSE tells the average distance between the value that was predicted from the model and the original value. Root mean square error is said to be good if the value of the metrics is closer to zero. Lower the normalized RMSE, better the model.

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

From the observation of forcast_x, multiple models were tried and tested and picked top three models which resulted in their corresponding RMSE. From the observation, Expo weighted was a better predicted model compared to the rest of the model

Table 1. Comparison of the model results

Sl No	Models Type	RMSE
1	Naive model	0.4
2	Half seas period mean	0.6
3	Expo weighted	0.2

5. CONCLUSION

The end result of this project will give a complete insight on how the atmosphere was before the pandemic started and the air quality during the pandemic with a detailed pictographic representation on all the pollutants as an individual. This project will also predict the future AQI level by state wise within India using advanced machine learning techniques.

6. FUTURE WORK

The current proposed system works on static data, where in the data obtained is from the year 2005 to 2015. The future work would be on streaming data that can actually predict the outcomes of Air Quality Index in real time which can in turn be used to alert people about the air quality in advance so as to prevent from causing health problems.

REFERENCES

[1] QING TAO1, FANG LIU 1, (Member, IEEE), YONG LI, AND DENIS SIDOROV…1 School of Automation, Central South University, Changsha 410083, China" Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU"IEEE 2009.

[2] Harsh Gupta, Dhananjay Bhardwaj, Himanshu Agrawal, Vinay Anand Tikkiwal, Arun Kumar" An IoT Based Air Pollution Monitoring System for Smart Cities" IEEE 2009.

[3] GALO D. ASTUDILLO 1 , LUIS E. GARZA-CASTAÑON 1 , AND LUIS I. MINCHALA AVILA 2 , (Senior Member, IEEE)" Design and Evaluation of a Reliable Low-Cost Atmospheric Pollution Station in Urban Environment" IEEE 2020.

[4] EdoardoArnaudo1,*, Alessandro Farasin 1,2 and Claudio Rossi 1" A Comparative Analysis for Air Quality Estimation from Traffic and Meteorological Data" MDPI 2020.

ISSN: 2582-6271

Vol.2 No.4; July-Aug 2021

[5] Xiankun Sun1, 2a ,Chengfan Li1,3b*,Lan Liu2c ,Jingyuan Yin1,4d,Yongmei Lei1e ,Junjuan Zhao1f ," Dynamic Monitoring of Haze Pollution Using Satellite Remote Sensing" IEEE 2019.

[6] Praveen V, Delhi Narendran T, Pavithran R, ChandrasegarThirumalai, IEEE Member " Data analysis using Box plot and Control Chart for Air Quality" IEEE 2017.

[7] BO LIU1 , (Senior Member, IEEE), SHUO YAN 1 , JIANQIANG LI 1 , (Senior Member, IEEE), GUANGZHI QU2 , (Senior Member, IEEE),YONG LI1 , JIANLEI LANG3 , AND RENTAO GU 4 "A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction"IEEE 2019.

[8] Pearl Pullan, ChitraGautam, VandanaNiranjan," Air Quality Management System"IEEE 2020.

[9] P. Vijayakumar, AbhinavKhokhar, Archit Pal and MohikaDhawan" Air Quality Index Monitoring and Mapping Using UAV" IEEE 2020.