International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Air Quality Index Prediction using Machine Learning

Harsh Jaiswal[1], Aman Singh[2], Akhilesh Chauhan[3]

[1]Harsh Jaiswal CSE department ITM Gida Gorakhpur
[2]Aman Singh CSE department ITM Gida Gorakhpur
[3]Akhilesh Chauhan CSE department ITM Gida Gorakhpur

Abstract - This paper presents an approach of employing Machine Learning for Air Quality prediction by utilizing input like temperature and wind speed. Supervised Machine Learning strategies are implemented through Random Forest Regressor algorithms to achieve more accurate output.
The AQI of a particular area is a measure of pollutants mixed in air of the respected region. As per the air quality standard pollutants are indexed in terms of their scale, these air quality indexes indicates the levels of major pollutants mixed in the atmosphere. Today, major cities experience levels of air pollution that are dangerously higher than the government-set air quality index standard. It significantly affects a person's health. The Machine Learning (ML) algorithms are capable of making predictions about air pollution. We are here to check the air quality index of a particular area to find air pollution level in that area.

Key Words: Machine Learning, Air Quality Index, Random Forest Regressor.

1.INTRODUCTION

It is a system that runs using machine learning algorithms to detect the air quality index [1]. Air pollution is one of the most important problems that the globe is now facing. Due to the economy's tremendous growth, industrial activity is expanding more frequently, which is causing air pollution to increase more quickly. As a one of the developing industrial nation, our country is producing record amount of pollutants on a daily basis. Need of knowing the AQI of any area is rising day by day. There are various gases available in the atmosphere which are the reason for pollution on our environment specifically like Carbon-dioxide, Nitrogen dioxide, Sulphur oxide, Chlorofluorocarbons, Particulate Matter, etc. and other harmful contaminant of air.

Machine Learning is a subset of artificial intelligence (AI) that mainly deals with algorithms which improve and optimize with the use of data and experience. Generally speaking, Machine Learning consists of two phases: training and testing [17]. It provides an efficient platform when it comes to solving health issues at an expedited rate. Machine Learning can be divided into two distinct categories: supervised learning and unsupervised learning. With supervised learning, a model is built using previously labeled data while with unsupervised learning model learns from unlabeled data.

We gathered the information from the database, which includes information on pollution concentrations that occur across the country. We begin by figuring out the individual pollutant index for each data point that is provided and then determining the corresponding AQI for the area. In order to estimate the air quality of India in any given place, we have developed a model that can predict the air quality index for each available data point in the dataset.

With this ml model, different information about the data is collected using different methodologies to determine the regions that are most severely affected in a specific location (cluster). By including the suggested parameter-reducing formulations into our model, we outperformed the traditional regression models in terms of performance. On the currently available dataset for predicting the air quality index for all of India, our model has a 96% accuracy rate.

2. RELATED WORK

Numerous works related to the AQI Prediction System utilizing different Machine Learning algorithms have been done and achieved various results in this field.
The paper [1] “AQI Prediction System” employed Decision Tree, Random Forest, and Naïve Bayes algorithms to forecast quality of air is suitable for health or not.

The paper [2] “A machine learning approach to predict air quality in california” used LR, NB, KNN, SVM, DT and RF algorithms for air quality prediction with competent data processing and implementation of Machine Learning algorithms with distinct parameters; KNN scored the highest accuracy of 87%.

The paper [3] “Development of machine learning-based predictive models for air quality monitoring and characterization” made a comparison of numerous Machine Learning models via performance metrics to air quality related problems with an accuracy of 89.34% achieved by SVM.

Moreover, the paper [4] “Air Quality Prediction using Machine Learning over Big Data” proposed a CNN-MDRP algorithm which combined structured and unstructured data and proved that CNN-MDRP is more precise than earlier prediction algorithms.

Additionally, the paper [5] “A Review of Air Quality Prediction Using Machine Learning and Data Analytics

Approach" utilized diverse Data Mining (DM) and Machine Learning methods to predict air quality index and applied the proposed system where needed.

Furthermore, the paper [6] "Application of Machine Learning Predictive Models in AQI" concentrated on SVM and LR algorithms to assess study models related to air quality. These models showed high applicability in classification.

Additionally, the paper [7] "Air Quality Outbreak Prediction with Machine Learning" utilized MLP as well as ANBEIS and presented a comparative analysis between ML models and soft ones for forecasting the quality of air outbreak while providing initial benchmarking to illustrate ML potential for future use.

Moreover, the paper [8] " A machine learning model for air quality prediction for smart cities" constructed a air quality index prediction system using NB algorithm that attained an accuracy of 88.163% among others.

The paper [9] "Air Quality Prediction Using Machine learning" suggested a robust model for predicting AQI which ranked Logistic Regression algorithm as having most efficiency with 82.89% accuracy followed by DT (80.40%) & NB (80.40%) & SVM (81.75%).

Additionally, the paper [10]"Implementation Of machine learning model to predict Air Quality Index" explored, recommended & applied a machine learning model in which Rapid Miner tool is used calculating high degree of correctness than MATLAB & Weka tool.

Moreover, the Paper [11] "Smart City Air Quality Prediction using Machine Learning" proposed multiple methods and factors that influences quality of air at some extent.

Besides, the Paper [12] "Modeling correlations among air pollution-related data through generalized association rules" used KNN, Naive Bayes and SVM Algorithms and collaborated with respect to accuracy using Air Quality dataset and achieved highest accuracy of 86.6 %Using Naive Bayes.

Additionally, The Paper [13] "Forecast urban air pollution in mexico city by using support vector machines" proposed method For Air Quality Prediction using machine learning and results showed great accuracy standards for better estimated results.

Also, The Paper [14]" Prediction of the level of air pollution using principal component analysis and artificial neural network techniques to predict quality of air implemented Using Grails Framework.

Lastly, The Paper [15]" Three hours ahead prevision of SO2 pollutant concentration using an neural based forecaster To Create Disease Prediction System With Better Accuracy which also provides motivational thoughts and images.

3. PROPOSED MODEL

Our proposed methodology includes the following steps:

i. First I will collect the records of pollutants in the atmosphere.
ii. Then I will perform preprocess operation to extract information from the data.
iii. Then I will perform the cleaning operation to solve the problem of missing value and to give it numerical value.
iv. Then select the attribute from the given dataset.
v. Then there is normalization of data is performed to make it in a range.
vi. The system then displays the results.

The flowchart of the methodology is shown below :-

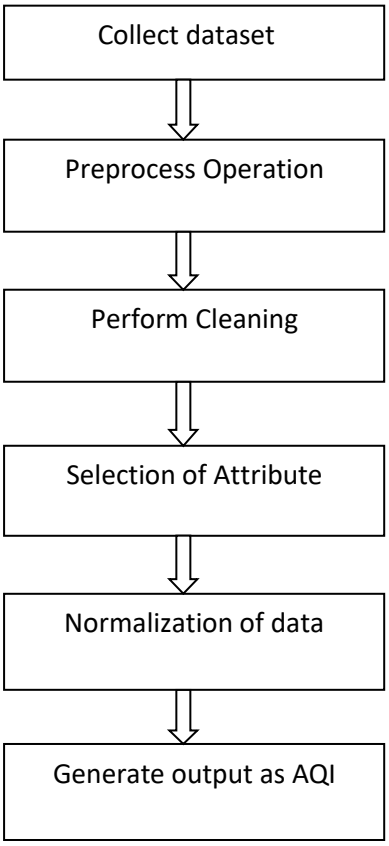

Fig 1: Flowchart of proposed model

ALGORITHM USED

As the name suggests, we use in our Random Forest Regressor for classification and for predicting the result.

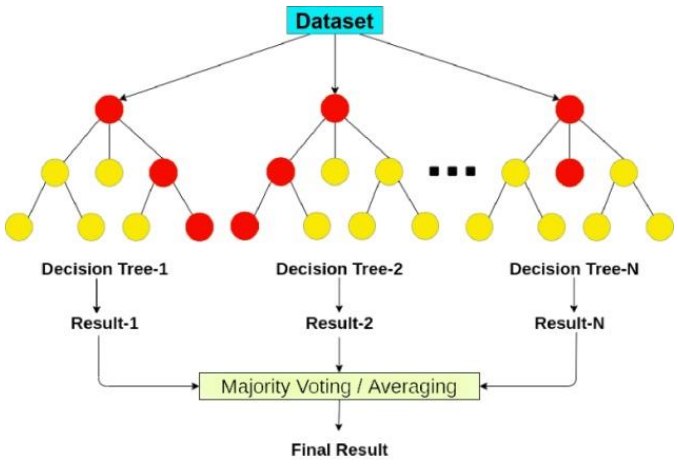

Fig -2: Random Forest Regressor

4. RESULTS ANALYSIS

Result analysis in our proposed system is an essential part of this research paper. By the analysis of results, we can compare that how much better this proposed system is performing. In result analysis we will see accuracy of air quality index that are predicted using our proposed system. We have taken datasets of 100 cases for result analysis. AQI prediction system leverages Random Forest Regressor (RFR) for classification for predicting the result. RFR is a form of regression algorithm which takes into account two or more predictor variables to determine response variable(Y).

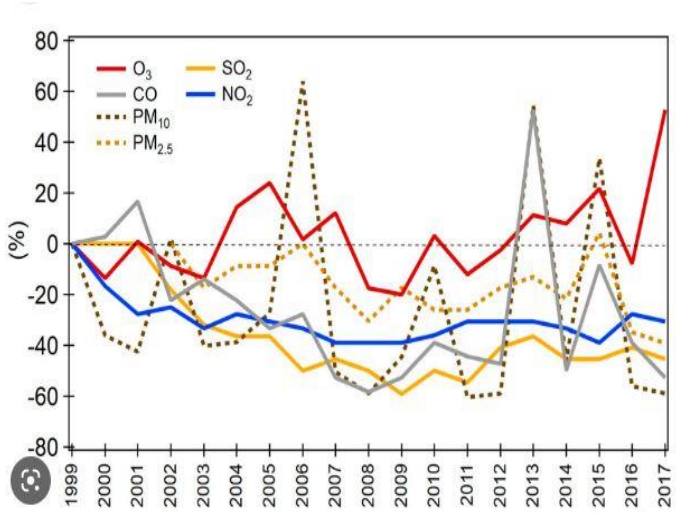

Fig -1: Output of AQI from year 1999 to 2017

Above diagram shows the accuracy of AQI from year 1999 to 2017 using Random Forest Regressor Algorithm.

	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
T	1.000000	0.967536	0.953719	-0.881409	-0.509299	0.640792	0.301994	0.287738	-0.631462
TM	0.967536	1.000000	0.892031	-0.822958	-0.586681	0.606945	0.292949	0.297011	-0.568409
Tm	0.953719	0.892031	1.000000	-0.917518	-0.287357	0.577240	0.296225	0.266782	-0.673824
SLP	-0.881409	-0.822958	-0.917518	1.000000	0.240256	-0.517915	-0.329838	-0.310704	0.623187
H	-0.509299	-0.586681	-0.287357	0.240256	1.000000	-0.455374	-0.380575	-0.362177	0.138005
VV	0.640792	0.606945	0.577240	-0.517915	-0.455374	1.000000	0.376873	0.342442	-0.573941
V	0.301994	0.292949	0.296225	-0.329838	-0.380575	0.376873	1.000000	0.775655	-0.268530
VM	0.287738	0.297011	0.266782	-0.310704	-0.362177	0.342442	0.775655	1.000000	-0.215854
PM 2.5	-0.631462	-0.568409	-0.673824	0.623187	0.138005	-0.573941	-0.268530	-0.215854	1.000000

Fig -2: Dataset for AQI Prediction

In the above chart we can see that nine factors are given and for these nine factors their accuracies are also given. These nine factors are processed using RFR algorithms for each data as shown in the row. The table shows the level of temperature, humidity, PM 2.5 level, etc.

The results of this study have important implications for the field of AQI prediction. The Air Quality Index Prediction system has the potential to improve the accuracy and speed of determining Air Quality. However, there are some limitations to the system, such as the reliance on accurate temperature and humidity level, which may not always be available.

Fig -3: The final/output page

5. CONCLUSIONS

In our research, we used a Random Forest Regressor Machine Learning algorithm to predict the AQI. And we also tested different algorithms like Linear Regression, Lasso Regression, Decision Tree Regression, KNN Regression etc. Despite testing these algorithms, I have found that the Random Forest Regressor gives higher accuracy than other algorithms. In this research, we found that the potential less accurate result is obtained if there is missing data, but if we can feed the system with one huge amount of data sets, this air quality index prediction can provide up to 96% accuracy.

Collecting a large number of datasets on pollutant is time consuming and cannot be done in a year or two. It takes

several years to collect these datasets and to train the system with these data searches. Doctoral students (PhD) can use this system for further research work. With the help of a AQI prediction system, it was possible to save people from polluted air. Components like gases and particulate matter affect the air's quality. When frequently breathed in, these contaminants lower the air quality, which can cause significant ailments. Only machine learning (ML) algorithms are capable of handling the meticulous analysis required to provide precise and efficient predictions from such a large body of environmental data. AQI prediction system only provides possible results, it does not guarantee that it will correctly predict the AQI each time, things may differ according to the weather condition. But it is far more accurate in predicting AQI from other systems. In our research we tested the accuracy of this system for 5 different locations value and our accuracy can be up to 96%.

REFERENCES

1. B. R. Gurjar, L. T. Molina, and C. S. P. Ojha, AQI Prediction System: health and environmental impacts. CRC press, 2010.
2. M. Castelli, F. M. Clemente, A. Popovi˘c, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in california,” Complexity, vol. 2020, 2020.
3. R. Sharda and R. Patil, “Neural networks as forecasting experts: an empirical test,” in Proceedings of the International Joint Conference on Neural Networks, vol. 2. IEEE, 1990, pp. 491–494.
4. C.-L. Hor, S. J. Watson, and S. Majithia, “Daily load forecasting and maximum demand estimation using arima and garch,” in 2006 International Conference on Probabilistic Methods Applied to Power Systems. IEEE, 2006, pp. 1–6.
5. R. G. Brereton and G. R. Lloyd, “Support vector machines for classification and regression,” Analyst, vol. 135, no. 2, pp. 230–267, 2010.
6. L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, G. Ricupero, and X. Xiao, “Modeling correlations among air pollution-related data through generalized association rules,” in 2016 IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2016, pp. 1–6.
7. U. A. Hvidtfeldt, M. Ketzel, M. Sørensen, O. Hertel, J. Khan, J. Brandt, and O. Raaschou-Nielsen, “Evaluation of the danish airgis air pollution modeling system against measured concentrations of pm2.5, pm10, and black carbon,” Environmental Epidemiology, vol. 2, no. 2, p. e014, 2018.
8. L. Pimpin, L. Retat, D. Fecht, L. de Preux, F. Sassi, J. Gulliver, A. Belloni, B. Ferguson, E. Corbould, A. Jaccard et al., “Estimating the costs of air pollution to the national health service and social care: An assessment and forecast up to 2035,” PLoS medicine, vol. 15, no.7, p. e1002602, 2018.
9. M. Castelli, F. M. Clemente, A. Popovi˘c, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in california,” Complexity, vol. 2020, 2020.
10. J. Zhu, R. Zhang, B. Fu, and R. Jin, “Comparison of arima model and exponential smoothing model on 2014 air quality index in yanqing county, beijing, china,” Appl. Comput. Math, vol. 4, pp. 456–461, 2015.
11. L. Y. Siew, L. Y. Chin, and P. M. J. Wee, “Arima and integrated arfima models for forecasting air pollution index in shah alam, selangor,” Malaysian Journal of Analytical Sciences, vol. 12, no. 1, pp. 257–263, 2008.
12. V. Kanawade, A. Srivastava, K. Ram, E. Asmi, V. Vakkari, V. Soni, V. Varaprasad, and C. Sarangi, “What caused severe air pollution episode of november 2016 in new delhi?” Atmospheric Environment, vol.222, p. 117125, 2020.
13. S. Arampongsanuwat and P. Meesad, “Prediction of pm10 using support vector regression,” in International Conference on Information and Electronics Engineering, IACSIT Press. Singapore, vol. 6, 2011.
14. R. Sharda and R. Patil, “Neural networks as forecasting experts: an empirical test,” in Proceedings of the International Joint Conference on Neural Networks, vol. 2. IEEE, 1990, pp. 491–494.
15. A. Azid, H. Juahir, M. E. Toriman, M. K. A. Kamarudin, A. S. M. Saudi, C. N. C. Hasnam, N. A. A. Aziz, F. Azaman, M. T. Latif, S. F. M. Zainuddin et al., “Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in malaysia,” Water, Air, & Soil Pollution, vol. 225, no. 8, pp. 1–14, 2014.

BIOGRAPHIES

Harsh Jaiswal is an undergraduate and his areas of interests are data structures, cloud computing, web development and machine learning, He has done training in Python from Coding Ninja and have the knowledge of Python, SQL, Web Development and Cloud Computing.

Aman Singh is an undergraduate and his areas of interests are machine learning, data structure and python. He has done training in Python from Tech Srijan.

Akhilesh Chauhan is an undergraduate and his areas of interests are machine learning and Web Development. He has a knowledge of Python and MySQL. He has done training in Python from Tech Srijan.