International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 2, March-April-2021, ISSN (Online): 2395-566X

Enhancing Air Pollution Predicton Using Artificial Neural Network with XGBOOST Algorithm

Asst. Prof. Christina Rini R, Aishwarya N, Meenakshi G, Saranya M
Department of Computer Science & Engineering,
Panimalar Institute of Technology, Chennai,
Tamil Nadu, India

***Abstract*-Air pollution has become a major public health concern in recent years. Despite substantial improvements in overall air quality in recent years, India remains the third most polluted country according to the latest edition of the World Air Quality Report [1]. The release of gases into the air that are harmful to human health and the environment is referred to as air pollution. Lung cancer, cardiovascular disease, respiratory disease, and metabolic disease have all been linked to high concentrations of fine particulate matter with a diameter less than 2.5 m (PM2.5). Experimental research with various machine learning algorithms has yielded promising results in the field of air pollution prediction. Predicting air quality by determining concentrations will assist different departments, including governments, in alerting people who are at high risk, reducing complications. An improved air pollution prediction approach based on the XGBOOST and ARTIFICIAL NEURAL NETWORK algorithms is used to forecast air quality. Our goal is to look into a machine learning-based approach for air quality forecasting using the AQI index and get the best results possible. In addition, we compare the performance of different ML algorithms from the dataset with the use of a GUI to predict air quality bycharacteristics.**

***Keywords*- Air quality forecasting, Artificial Neural Net- work, XGBoost.**

I.INTRODUCTION

Pollution, water resources, floods, and environmental in- formatics have all been studied extensively for the purposes of monitoring, modelling, and controlling environmental pro- cesses. Air pollution has become one of the most serious public health issues as a result of urbanisation and industrialization [1,2,3,4]. High PM2.5 concentrations have been linked to cancer, cardiovascular disease, respiratory disease, metabolic syndrome, and obesity. According to the World Health Organization, ambient air pollution causes approximately 4.2 million people to die prematurely each year.

PM2.5 or particulate matter with a diameter less than 2.5m, is the pollutant that poses the highest health risk. Because these particles are small and light, they can stay in the atmosphere for a long time. PM2.5 can reach the human body through the nose, pharynx, and throat, and then be absorbed through the trachea and bronchial alveoli, causing bronchitis, asthma, cardiovascular disease, heart disease, lung cancer, and other health problems (Song et al., 2017). Accordingly, exposure to air pollutants will increase the risk of disease and death (Hoek et al.,2013).

As a result, air pollution monitoring and control have become critical issues. Experts are continuing to develop a system that allows people to see the concentration ofair pollution at any time and from any location, allowing them to decide whether or not to go out or avoid polluted areas. However, we would rely on an air pollution forecast system to know the short-term or long-term air pollution status in advance.

As a result, predicting PM2.5concentrations is critical for social planning and management in order to reduce the negative effects of air pollution on public health. Aerosol Optical Depth (AOD) estimation using remote sensing technology has been successful in recent years, and this pa- rameter has become part of PM2.5 prediction research. Using regression and machine learning methods, as well as climatic variables and remote sensing data, several attempts have been madetopredict PM2.5 concentrations.For example,Lietal.[35] used the Moderate Resolution Imaging Spectroradio meter (MODIS) derived AOD product at 10 kmre solution (AOD10)in conjunction with meteoro logical data and PM2.5 historical observations to forecast PM2.5 in China.

Xiliang Ni etal.[36] Machine learning algorithms were used to estimatethe spatial distribution of PM2.5 concentration in the Beijing,Tianjin,and Hebei areas using satellite-derived MODIS AOD at 3km resolution (AOD03) and meteorological data.

Theimpactofair pollution has attracted the attention and concernofthepublic and government as a result of increasing environmental awareness and growing health consciousness.Thedistributionof air pollution in different locations demonstrates that the concentration of air

© 2021 IJSRET

pollution varies over time and location,as a result of climate and human activities.

1. Objectives:
Certain health problems are related to the air quality index.in order to prevent health issues due to bad air quality it is important to have an accurate estimate making use of high-performance models. Before dangerous levels are reached,certain preventive measure can be taken,like to stay inside in the house.The best solution would be to make the air cleaner for example by less pollution. However, that is not a solution that can be reached within in short time frame. But what we can do is to create awareness about the air quality bymaking certainpredictionsofforecomingairqualityusingwelltrained artificial intelligence algorithms so that they can be signaled when the level getsdangerous.

People can act upon this signal by taking preventive measures and so some health problems can be prevented. The training data that we will use to predict the air quality levels, consists of AQI index. The AQI is a linear feature of the pollutant concentration. The boundaries between AQI there is discontinuous jump between AQI categories unit to other. To calculate the AQI from the concentration the below equation isused.

2. Existing System:
The previous study used the LightGBM model to process high-dimensional data in order to predict PM2.5 concentra- tions in the next 24 hours based on datasets, which included forecasting data as one of the data sources for air quality pre- diction. For gradient enhancement, a residual learning-based iterative tree model called Gradient Boosting Decision Tree (GBDT) is used. LightGBM filters out data instances using a novel technique called Gradient-based One-Side Sampling (GOSS), which is slower than the XGBOOST algorithm. LightGBM also has a limited user base and usage due to a lack of documentation. Only 10,000+ predictive data sets are recommended.

II. PROPOSED SYSTEM

To predict the air quality index, the proposed system em- ploys supervised machine learning. To predict hourly PM2.5 concentrations, this technique employs the XGBoost (Extreme Gradient Boosting) rule. XGBoost is associate in nursing economicalrule and easy to use . It delivers high performanceand accuracy as compared to alternative algorithms. XGBoost makes use of data processing, that makes it a lot of quicker and permits it to run the model on completely different CPUs.

Understanding site specific pollution dispersion processes the factitious neural network(ANN) modelling ap- proach. The planned methodology can expeditiously use prog- no sticative knowledge to analyse high-dimensional options thorough. Hence, Improve the model's ability to understand data.

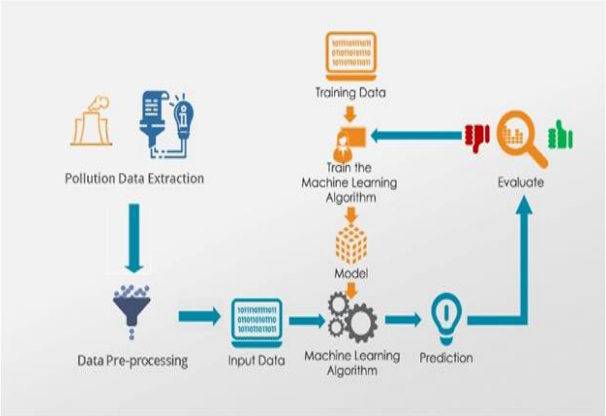

Fig 1.Architecture diagram

III. METHODOLOGIES

1. Data Visualization:
To begin, we remove some of the first rows from the dataset because their AQI values were not recorded. Then we try to visualise the data to see how different characteristics affect AQI levels over time. We try to discover the needed and unneeded features for our work as we plot the feature vs AQI graph. The plots of the characteristics that appear to influence AQI values.

2. Data Cleaning:
We can see from the plots that we need datetime, conds, dewptm, hum, pressurem, tempm, wdird, wspdm, month, day, hour, and AQI. The AQI is clearly lower at higher wind speeds. The speed of the wind may be a good predictor. Similarly, AQI values are slightly higher during the winter months, and wind direction plays a role as well.

All of the other characteristics have mostly missing values or have no discernible impact on the AQI value. As a result, some columns are removed. We then shift periods to generate past value features, allowing us to predict the AQI value for the current hour based on the previous 5 weather and AQI data. Then, if there are any missing values or NaNs, we either drop those rows or fill them with the mean column values, whichever seems most appropriate. Our data is now ready to be fitted into the models after we've completed these steps.Finally GUI simplifies the user needs by easily accessing the required data by selecting the required cities.

3. Data Scraping:
This module parses HTML code to discover relevant information and retrieves real-time data from the internet. The AQI which consists of each day data is stored in csv file format are s optimised and converted into monthly data before being compared to live data retrieved from the internet.

© 2021 IJSRET

4. XG BOOST:

Extreme gradient boosting, also known as XGBoost, is a well-known gradient boosting technique(ensemble) that improves the performance and speed of tree-based (sequential decision trees) machine learning algorithms. Tianqi Chen developed XGBoost, which was first maintained by the Distributed (Deep) Machine Learning Community (DMLC). It is the most widely used algorithm for applied machine learning in competitions, and it has grown in popularity as a result of winning solutions in structured and tabular data. It's a free and open-source software. Previously, only Python and R packages were available for XGBoost, but it has since been expanded to include Java, Scala, Julia, and other languages.

In Ensemble Learning, XGBoost is classified as a boosting technique. Ensemble learning combines multiple models into a collection of predictors to improve prediction accuracy. The errors made by previous models are attempted to be corrected by succeeding models by adding weights to the models in the boosting technique.

IV. ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a component of a computing system that mimics how the human brain analyses and processes data. Artificial intelligence (AI) is built on this foundation, and it solves problems that would be impossible or difficult to solve by human or statistical standards. ANNs have self-learning capabilities, allowing them to improve their performance as more data becomes available.

ANNs are made up of processing units, which have inputs and outputs. The ANN learns from the inputs to generate the desired output. Backpropagation is a set of learning rules that artificial neural networks use to guide them. ANNs have a wide range of practical applications, including finance, personal communication, business, education, and so on.

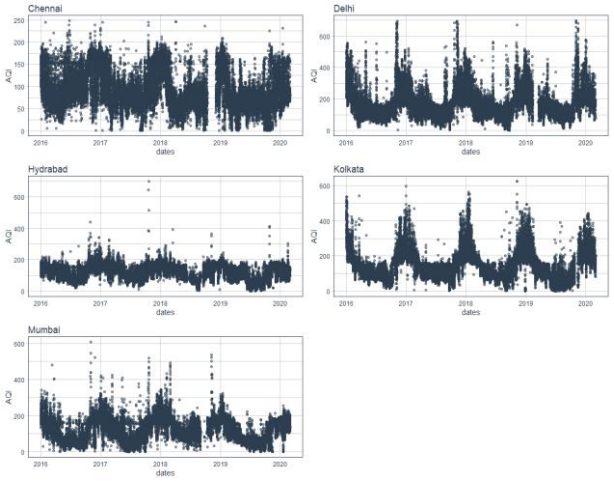

Fig 2.Graph based on AQI

V. CONCLUSION

PM2.5 is the most common pollutant in today's air, and its concentration is a key criterion for assessing air quality. PM2.5 not only decreases visibility and sunshine in the atmosphere, but it also causes more haze days and is harmful to human health.

As a result, accurate PM2.5 concentration predictions will assist the government in better understandingthe current state and trend of air quality, allowing it to formulate effective prevention and control measures.

The experimental results show that the PM2.5 hourly average concentration prediction model based on the XGBoost method outperforms other data mining techniques in terms of accuracy and low over-fitting probability.

REFERENCES

[1] K &Quot;Urban air quality analysis and prediction using machine learning",2019 IEEE,978-1-7281-0418-8/19/$31.00 2019 IEEE.
[2] Venkat Rao, Hari Kiran Reddy,"Air Quality Prediction Of Data Log By Machine Learning"2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS).
[3] T., Rajat Raj Niraj Tailor"Urban Air Quality Prediction Using Regression Analysis ",978-1-7281-1895-6/19/$31.00 c 2019 IEEE.
[4] Ms. Hable-KhandekarDr. "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring ",978-1-5386-4008-1/17/$31.00 ©2017 IEEE
[5] S. L.,"Air Pollution Prediction by Deep Learning Model" Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2.
[6] Aba Ameer, Munam Ali Shah, Abid Khan Song Maple Ul Islam Muhammad Nabeel Asghar, b "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities " 30, 2019, accepted June 3, 2019, date of publication June 26, 2019, date of current version September 20, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2925082
[7] X., T. , B., Li, S., H., S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257. . 20150257.
[8] Furui Liu, Hsun-Ping U-Air: When Urban Air Quality Inference Meets Big Data, KDD '13 Proceedings of

© 2021 IJSRET

the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1436-1444

[9] QI; WANG, Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. (2018). IEEE Transactions on Knowledge and Data Engineering. 30, (12), pp.2258-2297.

© 2021 IJSRET