RESEARCH ARTICLE OPEN ACCESS

AIR QUALITY PREDICTION USING MACHINE LEARNING TECHNIQUES

SREEJAHINEE G
UG Scholar-ECE
PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY
(An Autonomous Institution)-Dindigul
jahineegurusamy@gmail.com

P. G. AKILA
Assistant Professor/ECE
PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY
(An Autonomous Institution)-Dindigul
akilapg@psnacet.edu.in

RAMAVIDHYA B
UG Scholar-ECE
PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY
(An Autonomous Institution)-Dindigul
ramavidhya89@gmail.com

SALLY JANNET S
UG Scholar-ECE
PSNA COLLEGE OF ENGINEERING AND TECHNOLOGY
(An Autonomous Institution)-Dindigul
jannetsally001@gmail.com

ABSTRACT

By using machine learning, we anticipate the air quality index for a certain place in India. The Indian air quality index is a commonly used indicator of pollutant (so2, no2, etc.) Levels across time. Based on historical data from previous years and forecasting over a certain forthcoming year as a Gradient Decent Boosted Multivariable Regression Problem, we constructed a model to predict the air quality index. For our predictive problem, we use cost estimation to increase the model's efficacy. When given historical data on pollutant concentration, our model will be able to accurately estimate the air quality index for an entire county, any state, or any contiguous region. By including the suggested parameter-reducing formulations into our model, we outperformed the traditional regression models in terms of performance. Our model predicts the air quality index for the entirety of India with 96% accuracy, and we also use the XG Boost Algorithm technique to determine the order of preference based on how closely it approaches the optimal answer. By utilising machine learning to estimate the air quality index of a certain place, we forecast India's air quality. The Indian air quality index is a commonly used indicator of pollutant (so2, no2, etc.) Levels across time. Based on historical data from previous years and forecasting over a certain forthcoming year as a Gradient Decent Boosted Multivariable Regression Problem, we constructed a model to predict the air quality index. For our predictive problem, we use cost estimation to increase the model's efficacy. When given historical data on pollutant concentration, our model will be able to accurately estimate the air quality index for an entire county, any state, or any bounded region. By including the suggested parameter-reducing formulations into our model, we outperformed the traditional regression models in terms of performance. Using our model, we apply the XG Boost Algorithm technique to discover the order of preference by similarity to the ideal answer, and we predict the air quality index of the entire country of India with 96% accuracy using the currently available dataset.

KEYWORDS: Air quality, Pollutant (SO_2, NO_2), Gradient booster, XG Boost Algorithm Technique, Pollutant Concentration.

1. INTRODUCTION

Nowadays, accurate air pollution prediction and forecast become a challenging and significant task due to increased air pollution which acts as a fundamental problem in many parts of the world. Generally, the pollution is divided into two types: [1] natural pollution because of volcanic eruptions and forest fires resulting in emission of SO2, CO2, CO,

NO2, and Sulphate as air pollutants and [2] man-made pollution because of some human activities such as burning of oils, discharges from industrial production processes, and transportation emissions that have PM2.5 as its major air pollutant which has received much attention due to their destructive effects on human health, other kinds of creatures, and environment . Various studies testify that air pollution leads to respiratory and cardiovascular disease leading to death of animals and plants, acid rain, climate change, global warming, etc. thus making economic loses and the human life of a society difficult to survive in the world Regarding the effects of PM2.5 investigated over the last 25 years using the comparative analysis of ML techniques, Ameer et al. have estimated that approximately 4.2 million people have died due to long-term exposure of PM2.5 in the atmosphere, while an additional 250,000 deaths have occurred due to ozone exposure. In worldwide rankings of mortality risk factors, PM2.5 was ranked as 5th and accounted for 7.6% of total deaths all over the world. From 1990 to 2015, the number of deaths due to air pollution has increased, especially in China and India with more than 20% of 1.1 million deaths worldwide attributed to respiratory diseases. Hence, worldwide, huge number of research has been carried out on topics like air pollution levels and air quality forecasts to control air pollution more effectively. Extensive research 2 specifies that air pollution forecasting approaches can be imprecisely divided into three traditional classes: [1] statistical forecasting methods, [2] artificial intelligence method and [3] numerical forecasting methods.

The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. 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 major pollutants are Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Particulate matter (PM), Sulphur Dioxide (SO2), Carbon Dioxide (CO2) etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometre or less than that affects more to human health. Measures must be taken to minimize air pollution in the environment. Air Quality Index (AQI), is used to measure the quality of air. Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air. Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the 3 pollutants needs vigorous analysis. Convolution Neural networks, Recursive Neural networks, Deep Learning, Machine learning algorithms assures in accomplishing the prediction of future AI so that measures can be taken appropriately. Machine learning which comes under artificial intelligence has three kinds of learning algorithms, they are the Supervised Learning, Unsupervised learning, Reinforcement learning. In the proposed work we have used supervised learning approach. There are many algorithms under supervised learning algorithms such as Linear Regression, Nearest Neighbour, XG Boost Algorithm. Compared to all other algorithms Random Forest gives better results, so our approach selects Random Forest to predict the accurate air pollution.

Due to human activities, industrialization and urbanization air is getting polluted. The major air pollutants are SO2, CO2, CO, NO2, etc. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Earlier techniques such as probability, statistics etc. were used to predict the quality of air.

2. LITERATURE SURVEY

1. Outlook for clean air in the context of sustainable development goals

Air pollution is linked with many of the United Nations sustainable development goals. Strategies aiming at the improved air quality interact directly with climate mitigation targets, access to clean energy services, waste management, and other aspects of socio economic development. Continuation of current policies in the key emitting sectors implies that a number of sustainability goals

will likely not be met within the next two decades: emissions of air pollutants would cause 40% more premature deaths from outdoor air pollution than today, carbon emissions would rise globally by 0.4% per year, while nearly two billion people would not have access to clean cooking. This paper examines integrated policies to put the world on track towards three interlinked goals of achieving universal energy access, limiting climate change and reducing air pollution. Scenario analysis suggests that these goals can be attained simultaneously with substantial benefits. By 2040, emissions of main pollutants are projected to drop by 60-80% relative to today, and associated health impacts are quantified at two million avoided deaths from ambient and household air pollution combined. In comparison to costs needed for the decarbonization of global economy, additional investments in air pollution control and access to clean fuels are very modest against major societal gains.

2. Air pollution in China: Mapping of concentrations and sources,

China has recently made available hourly air pollution data from over 1500 sites, including airborne particulate matter (PM), SO2, NO2, CO2 and CO. We apply Kriging interpolation to four months of data to derive pollution maps for eastern China. Consistent with prior findings, the greatest pollution occurs in the east, but significant levels are widespread across northern and central China and are not limited to major cities or geologic basins. Sources of pollution are widespread, but are particularly intense in a northeast corridor that extends from near Shanghai to north of Beijing. During our analysis period, 92% of the population of China experienced >120 hours of unhealthy air (US EPA standard), and 38% experienced average concentrations that were unhealthy. China's population-weighted average exposure to PM2.5 was 52 μg/m3. The observed air pollution is calculated to contribute to 1.6 million deaths/year in China [0.7-2.2 million deaths/year at 95% confidence], roughly 17% of all deaths in China

3. An investigator digital forensics frequencies particle swarm optimization of detection and classification of APT attack in FOG computing environment

Though there are several approaches to detect the malware attacks in cloud, the detection techniques could not be applied in FOG based environment. This is because of its possession of distinct features. As FOG computing has been evolving, it is mandatory to develop detection and mitigation schemes of malware attacks. Thus, in this research, an approach for investigation of digital forensics has been developed, where it classifies and detects the APT attack named Shamoon attack from different attack types in FOG environment. Digital Forensics has been recently gaining focus in solving or investing the cybercrimes. Several researches have been developed in this field where they have analysed several security challenges. Previous technologies, to measure these attacks are completely based on methodology of pattern matching. If an attack is newly occurred, then the detection rate is very low and false negative will be very high. Thus, the challenges are highly increased as the data volume increases, and the technology used by attacker is continually developed. As there is a lack in detection technology and the deployment boards, and the low efficient models in FOG computing makes the challenge a difficult one. Thus a proposed scheme has been introduced where Frequency Particle Swarm Optimization (FPSO) has been utilized in investigating digital forensics Particle Swarm Optimization in order to detect and to classify the APT attack (Shamoon attack) in FOG environment. This approach uses four phases. In feature extraction, best set of features are extracted. Using FPSO 7 (Frequencies PSO), best weighed features are predicted. These weighed features are clustered using K-means clustering and classified using k-nearest neighbours (KNN) classifier. The performance of this approach is then evaluated using confusion matrix and results are provided. Finally, the proposed KNN-FPSO classifier is compared with other existing classifiers and the results are recorded

4. Effects of Air Pollution on Human health and Practical Measures for prevention in Iran

Air pollution is a major concern of new civilized world, which has a serious toxicological impact on human health and the environment. It has a number of different emission sources, but motor vehicles and industrial processes contribute the major part of air pollution. According to the World Health Organization, six major air pollutants include particle pollution, ground-level ozone, carbon monoxide, sulphur oxides, nitrogen oxides, and lead. Long and short term exposure to air suspended toxicants has a different toxicological impact on human including respiratory and cardiovascular diseases, neuropsychiatric complications, the eyes irritation, skin diseases, and long-term chronic diseases such as cancer. Several reports have revealed the direct association between exposure to the poor air quality and increasing rate of morbidity and mortality mostly due to cardiovascular and

respiratory diseases. Air pollution is considered as the major environmental risk factor in the incidence and progression of some diseases such as asthma, lung 8 cancer, ventricular hypertrophy, Alzheimer's and Parkinson's diseases, psychological complications, autism, retinopathy, fetal growth, and low birth weight. In this review article, we aimed to discuss toxicology of major air pollutants, sources of emission, and their impact on human health. We have also proposed practical measures to reduce air pollution in Iran

Air quality index and air pollutant concentration prediction based on machine learning algorithms

Air pollution has become an important environmental issue in recent decades. Forecasts of air quality play an important role in warning people about controlling air pollution. We used support vector regression (SVR) and random forest regression (RFR) to build regression models for predicting the Air Quality Index (AQI) in Beijing and the nitrogen oxides (NO) concentration in an Italian city, based on two publicly available datasets. The Root-Mean-Square Error (RMSE), correlation coefficient (r), and coefficient of determination (R2) were used to evaluate the performance of the regression models. Experimental results showed that the SVR-based model performed better in the prediction of the AQI (RMSE = 7.666, R2 = 0.9776, and r = 0.9887), and the RFR-based model performed better in the prediction of the NO concentration (RMSE = 83.6716, R2 = 0.8401, and r = 0.9180). This work also illustrates that combining machine learning with air quality prediction is an efficient and convenient way to solve some related environment problems

3. SYSTEM ANALYSIS

Existing system

The existing systems detect 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). The data is displayed on a monthly, weekly or daily basis. Also, once the values are forecasted, the values do not change with respect to the sudden change in the atmospheric conditions or unexpected increase in traffic. The values are detected for the whole city, and cannot be verified for the accuracy of the forecasted values afterward.

Disadvantages

Loading of all dataset is difficult and takes long time. Process is not accurate. Analysing the dataset takes long time.

Proposed system

Datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy. Advantages

These reports are to investigated through of machine learning techniques for air quality forecasting in operational conditions. Finally, it highlights some observations on future research issues, challenges, and needs.

4. SYSTEM DESIGN

Module design and organization

Data Pre-Processing Dataset splitting Model training Classification

Data pre-processing

Three common data pre-processing steps are: Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file. Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be removed from the data entirely.

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

Dataset splitting

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets. A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data. A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model over fitting, which is the incapacity for generalization we mentioned above.

Classification

Once all the crucial steps are performed including pre-processing, and feature extraction, we move towards classification. There are very many classification techniques proposed by various researchers. All these techniques have several pros and cons. There is a fluctuation in the performance of these techniques as well depending on the data and other prerequisite steps The most popular metrics for measuring classification performance include accuracy, precision, confusion matrix, log-loss, and AUC (area under the ROC curve).

• Accuracy measures how often the classifier makes the correct predictions, as it is the ratio between the number of correct predictions and the total number of predictions.

• Precision measures the proportion of predicted positives that are truly positive. Precision is a good choice of evaluation metrics when you want to be very sure of your prediction.

•The confusion matrix (or confusion table) shows a more detailed breakdown of correct and incorrect classifications for each class. Confusion matrix is useful when you want to understand the distinction between classes, particularly when the cost of misclassification might differ for the two classes, or you have a lot more test data on one class than the other. For example, the consequences of making a false positive or false negative in a cancer diagnosis are very different.

5. CONCLUSION

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings and takes a step towards fighting against it. The results show that machine learning models (logistic regression and auto regression) can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

6. FUTURE WORKS

Anticipated that their incorporation into deterministic air quality models will result in a much better spatio-temporal scenario that policymakers can use to regulate air pollution. According to this review, the majority of researches have focused on forecasting AQI and pollutants concentration levels, which will provide an accurate picture of AQI. Many researchers choose Artificial Neural Network (ANN), ARIMA Model, Linear Regression, and Logistic Regression for the prediction of AQI and air pollutants concentration. When projecting the AQI or the future concentration level of various pollutants, the future scope may take into account all elements, including meteorological parameters and air contaminants. As the data changes at specific intervals of time, we can also use real-time data analysis via the cloud to get better results for increased performance. To process enormous amounts of data and combine two or more machine learning algorithms, we can obtain more accurate results.

7. REFERENCE

Thus the regressing and prediction about air quality using scaling of parameters by values.

[1] P. Rafaj, G. Kiesewetter, T. Gül, W. Schöpp, J. Cofala, Z. Klimont, andP. Purohit, ``Outlook for clean air in the context of sustainable developmentgoals," Global Environ. Change, vol. 53, pp. 111, Nov. 2018.

[2] P. J. Landrigan, R. Fuller, N. J. R.Acosta, O. Adeyi, R. Arnold, A. B. Baldé,and R. Bertollini, ``The Lancet commission on pollution and health,"Lancet, vol. 391, no. 10119, pp. 462512, 2018.

[3] R. A. Rohde and R. A. Müller, ``Air pollution in China: Mapping ofconcentrations and sources," PLoS ONE, vol. 10, no. 8, Aug. 2015,Art. no. e0135749.

[4] State of Global Air 2019, Health Effects Institute, Boston, MA, USA, 2019.

[5]K. Al Hwaitat, S. Manaseer, R. M. Al-Sayed, M. A. Almaiah, and O. Almomani, ``An investigator digital forensics frequencies particle swarm optimization for detection and classi_cation of APT attack in FOG computing environment (IDF-FPSO)," J. Theor. Appl. Inf. Technol., vol. 98, no. 7, pp. 937_952, 2020.