November-December 2019
ISSN: 0193-4120 Page No. 5393 - 5398

Predicting the Quality of Air Using Supervised Techniques of Machine Learning

[1]G. Sai Kumar, [2]D. Mahalakshmi

[1]UG Student, Department of Computer Science and Engineering, Saveetha School of Engineering, saikumarg095@gmail.com
[2]Assistant professor, Department of Computer Science and Engineering, Saveetha School of Engineering, mahalakshmid.sse@gmail.com

Article Info
Volume 81
Page Number: 5393 - 5398
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *26 December 2019*

Abstract

Air contamination is the process of releasing the harmful gases into the atmosphere that are dangerous to human health and the whole planet. It is compared as one of most perilous threat that humans are never faced. It brings damage to all the animals and plants on the earth. To overcome this problem, the transport division has to analyze the air quality time to time using some machine learning techniques. Hence, predicting their quality using these techniques is became important these days. The main aim is to use classification techniques of machine learning (ML) in predicting air quality. The dataset of air quality is pre-processed with some of the techniques such as data preparing, data validation, and removal of missing values, bivariate and multivariate analysis. Now the quality of air is predicted using some supervised techniques such as Decision tree, support vector machines, Random forest, Logistic regression, K-Nearest neighbors. The various ML techniques are now compared with precision, Recall and F1 score. It is seen that decision tree performs very well than other techniques in air quality prediction. This implementation can help meteorological department in air quality prediction. In the next generation, some of the Artificial intelligence (A.I) techniques can be applied and optimized.

Keywords: *Air quality prediction, classification and machine learning techniques, Decision tree, predicting the accuracy.*

1. Introduction

Predicting the future from the past data is known as machine learning. It is a kind of A.I that gives the computers the capacity to learn without external programs. Python is being used in implementing some of the techniques of machine learning and the programs may also change when they exposed to the new data. Some of the special algorithms are used in training and prediction processes. The algorithm takes the training data, and this data gives the predictions for the new data that is being used for testing. Machine learning (ML) can be categorised into three techniques such as supervised, unsupervised and reinforcement learning. Supervised learning program is both

given the info information and the relating marking to learn information must be named by an individual in advance. Unsupervised learning is no marks. It gave to the learning calculation.

This calculation needs to make sense of the grouping of the information. At last, Reinforcement adapting progressively communicates with its condition and it gets positive or negative input to improve its performance. Scientists use various kinds of ML programs in identifying patterns in the python. Now at these levels, the program can be named as learn and predict using some the ML techniques such as supervised and unsupervised learning. In classification, we use it to predict the class of a given data. They are also known as labels. It is used to map the internal and external data. In ML, classification mainly means taking and learning the data and producing the new data based on the old data. Some of the examples are face identification, google or Alexa voice recognition.

2. Literature Review

[1]. Recurred neural networks are very useful in processing the data. The data that is coming up in the future is also very useful than the data that is present at this time. RNN can find the output by using the frames for some times. If it takes much time then the prediction accuracy may drop. While deferring the yield by certain casings has been utilized effectively to improve results for consecutive information, the ideal postponement is task ward and should be acquired by the experimentation strategy. Additionally, two separate systems, one for every bearing could be prepared on all info data and afterward the outcomes could be blended utilizing number juggling or geometric averaging for conclusive expectation. Be that as it may, it is hard to acquire ideal converging since various systems prepared on similar information can never again be viewed as free. To conquer these constraints, it proposed bidirectional repetitive neural system that can be prepared utilizing all accessible info data previously and eventual fate of a particular time period. Contamination information like some other sensor information isn't free from missing information and anomalous qualities. The inconsistencies may happen because of instrumental mistake or some other outer elements like power-shutdown or severance of availability and so forth. There were examples where contamination information was not detailed by a source checking station. These missing qualities were introduced utilizing moving normal of accessible information estimations of past three time occurrences. A worth lying outside the passable range for a parameter is treated as an anomalous worth. Anomalous esteems are likewise supplanted by moving normal of past three cases.

[2]. People who are working at the industries may face severe health problems due to the release of pollutants into the air surface. The air substances that releases into the atmosphere are very toxic and they cause severe skin diseases too. Many countries focussing in controlling the pollution and they are advertising through the media and some programmes are being conducted by government in controlling pollution. Now a days different techniques and technologies like Internet of things (IOT) and ML is used in predicting the air quality. The unavoidable nearness around us of different remote innovations, for example, radio Frequency Identification labels, sensors, actuators and cell phones establishes the foundation of the IOT idea. These items can send and get information self-sufficiently, accordingly opening new skylines for home, wellbeing, and mechanical applications. Actually, innovation progresses

alongside expanding request will cultivate an across the board sending of IOT administrations, which would profoundly change our organizations, networks and individual lives. Presently a day the air contamination in urban territories is a significant issue in created urban communities because of noteworthy effects of air contamination on general wellbeing, worldwide condition and the entire overall economy. The proposed work on an air contamination observing and expectation framework is empowers us to screen air quality with the assistance IoT gadgets. The framework uses air sensors to identify and transmit this information to microcontroller. At that point the microcontroller stores the information into the web server. For foreseeing the long short term memory (LSTM) is executed. It has a fast assembly and diminishes the preparation cycles with a decent precision.

[3]. External air quality plays a vital role in the human health. It causes death to many of persons and the cost of medicines also is very huge. External air also contains the pollutants such as PM 2.5 concentration. They contain many of the harmful gases such as nitrogen, ozone and carbon-monoxide. Huge levels of these toxins are delivered by anthropogenic exercises. While a great many people invest most of their energy inside, open air quality can influence indoor air quality to an enormous degree. What's more numerous patients, for example, asthmatics, patients with hypersensitivities and compound sensitivities, cardio therapy patients, heart and stroke patients, diabetics, pregnant ladies, the old and youngsters are particularly powerless to poor outside and indoor air quality. Much inquire about on the wellbeing impacts of outside air contamination have been distributed in the most recent decade. The objective of this audit is to briefly condense a wide scope of the ongoing exploration on wellbeing impacts of numerous kinds of open air contamination. A survey of the wellbeing impacts of major open air toxins including particulates, carbon monoxide, sulphur and nitrogen oxides, corrosive gases, metals, unpredictable organics, solvents, pesticides, radiation and bio mist concentrates is exhibited. Various examines have connected air poisons to numerous sorts of medical issues of many body frameworks including the respiratory, cardiovascular, immunological, haematological, neurological and conceptive/formative frameworks.

[4]. pollution in urban areas has a severe impact on the health problems in human beings. Pollutants in urban areas cause diseases such as asthma and some of the lung diseases. Ongoing considers have indicated generous confirmations that presentation to climatic contaminations has solid connects to antagonistic maladies including asthma and lung irritation. The modules are answerable for getting and putting away the information, pre-processing and changing over the information into valuable data, gauging the contaminations dependent on chronicled data, lastly showing the gained data through various channels, for example, versatile application, Web gateway, and short message administration. The focal point of this paper is on the observing framework and its determining module Progressing considers have demonstrated liberal affirmations that introduction to climatic pollutions has strong associates with opposing illnesses including asthma and lung disturbance. The modules are responsible for getting and taking care of the data, pre-processing and changing over the data into important information, gauging the defilements subject to chronicled information, in conclusion indicating the picked up information through different

November-December 2019
ISSN: 0193-4120 Page No. 5393 - 5398

channels, for instance, flexible application, Web entryway, and short message organization. The point of convergence of this paper is on the watching structure and its deciding module.

3. Methodology

Existing System

The pollution in urban areas causing severe effects. To avoid this some of the prediction methods should be used to identify harmful gases and to prevent them by implementing some techniques in decision making. Using the data of the environment and the techniques in the ML is used to detect them helps in avoiding the pollution. This can be done using some networks such as Convolution neural network and the other is long short term memory. LSTM is used to calculate the dependencies of atmospheric pollutants. This types of models can predict the PM 2.5 concentrate pollutants but the other are not identified. After the prediction is over using these techniques, they are compared with those of previous numeric models. Air pollution has severe impact on human health. So, we have to predict the environmental damaging pollutants. First we have to take some data from the environment and that is compared with that of the unsupervised techniques. LSTM is used to predict the quality of air from time to time. So we have to monitor the quality of air at all time. The pollution may be also from different forms such as the transportation, industries, electricity etc. Many of the scientists started using big data techniques to control the pollution by using the network sensing. By the end it is used to forecast also.

Drawbacks

- Multiple sites is used for training the data.
- The above air quality prediction is for only one city and if we have to predict other then we have to collect other data.

Proposed System

Many datasets are combined and they are made into a generalised dataset, then we have to apply some of the ML techniques and we have extract patterns, then we get the maximum results with accuracy. This is also known exploratory data analysis. This involves some steps such as:

- Wrangling up of data.
- Collection of data
- Data pre-processing
- Building the classification model
- Predictive model construction.

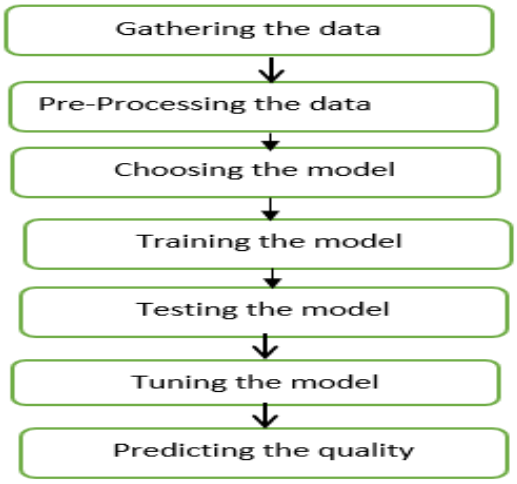

Figure 1: Process of Dataflow Diagram

Training the Dataset

- First we have to import the demo data set of various cities which is already exists in sklearn, it is basically a data in table with different varieties.
- To load the data using load_data() method, we have split it using train_test_split method. The value of X denotes the feature values and the Y denotes the Target values.
- This type divides the data into training and test data separately in different ratios.

November-December 2019
ISSN: 0193-4120 Page No. 5393 - 5398

- Now the training data is inserted into the algorithm. So the computer can be ready to trained using this data.

Testing the Dataset

- Here, we use numpy package. The numpy package consists of numeric values, it takes the numbers as input and produce the output as target values.
- Finally, we get the predicted value as 0. Now find the ratio between total no of predictions and the no of predictions identified. We get the maximum accuracy using this method because it compares the actual value vs the values predicted.

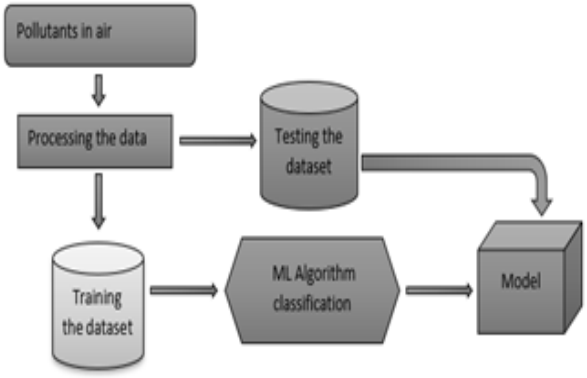

Figure 2: System Architecture

Pre-Processing the Dataset

Data which is collected may have some values missing that may cause inefficiency. To get the results better, the dataset should be processed to increase efficiency of the program.

Data cleaning: In this all the data which is noisy should be removed and all the values which are missing should also be removed. In the places where the missing values are identified, fill those areas with the mean values.

Data transformation: In this transformation, we can convert the data from one form to another form. Here, we can remove all the null values and all the duplicate values. We can determine structure of data and we can map them.

Data reduction: It is used for decreasing the size of data to increase its stability and efficiency. Only volume of the data should be reduced here. It involves various techniques like parametric and non parametric methods.

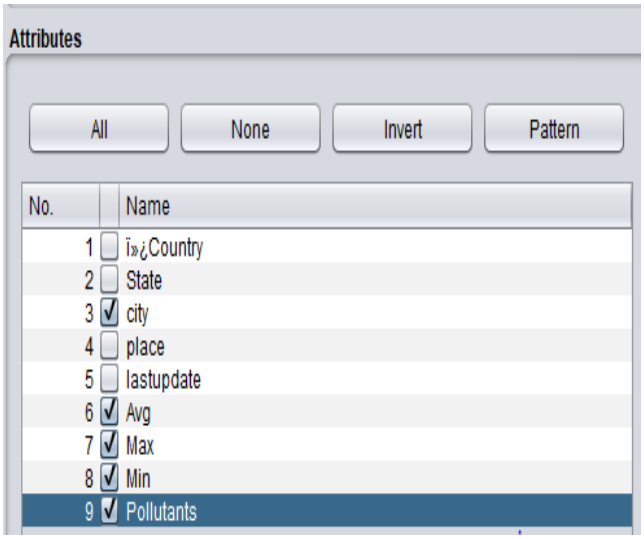

Figure 3: Selection of Attributes in Pre processing

Figure 4: Data visualisation of selected attributes in Weka tool.

4. Conclusion

The process is started from the data re-processing techniques such as data cleaning and processing by eliminating the missing values and finally we have to build a model. These types of applications helps the Indian

November-December 2019
ISSN: 0193-4120 Page No. 5393 - 5398

meteorological department in predicting the future pollution.

References

[1] C. Sun, M. E. Kahn, and S. Zheng, "Self-protection investment exacerbates air pollution.

[2] Q. Zhang et al., "Transboundary health impacts of transported global air pollution and international trade," Nature, vol. 543, no. 7647, p. 705, 2017.

[3] L. Gharibvand et al., "The association between ambient fine particulate air pollution and lung cancer incidence: results from the AHSMOG-2 study," Environmental health perspectives, vol. 125, no. 3, p. 378, 2017.

[4] A. Lee, A. Szpiro, S. Y. Kim, and L. Sheppard, "Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology," Environmetrics, vol. 26, no. 4, pp. 255–267, 2015.

[5] S. Park et al., "Predicting PM10 concentration in Seoul metropolitan sub-stations using artificial neural network (ANN)," Journal of hazardous materials, vol. 341, pp. 75–82,2018.

[6] I. Djalalova, L. DelleMonache, and J. Wilczak, "PM 2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model," Atmospheric Environment, vol. 108, pp. 76–87, 2015.

[7] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865–873, 2015.

[8] D.L. Yamins and J.J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," Nature neuroscience, vol. 19, no. 3, p. 356, 2016.