EasyChair Preprint

№ 9994

Air Quality Index Detection Using Random Forest Algorithm

A.Peter Soosai Anandaraj, Hari Krishnam Raju Keertipati and Adithya Gunda

April 27, 2023

AIR QUALITY INDEX DETECTION USING RANDOM FOREST ALGORITHM

Dr.Peter Soosai Anandaraj A[1] , **K.Hari Krishnam Raju2, G.Adithya3**

[1]*Associate professor,* [2,3] UG Student
Department Of Computer Science and Engineering,
Veltech Rangarajan Dr.Sagunthala R&D
Institute of Science and Technology,
Avadi, Chennai, Tamil Nadu, India.
[1]*anandsiriya@gmail.com*

ABSTRACT

Internet of Things (IoT) has various applications in our daily life from a fully automated Smart Building, to a simple Smartphone application that records the users health information. IoT is the driving force for rapid development of human life transforming it to be more intelligent, productive and organized. It sheds the light on creative and different methods of transferring, combining and conversion of different types of knowledge as users interact and practice them, thus allowing developers and hobbyists to create novel applications. The aim of this paper is to introduce a novel approach to a system, which controls simple electrical appliances such as a water kettle or a coffee machine, depending on the user's attention values measured using NeuroSky/MindWave Mobile EEG sensor. This novel approach of controlling simple home appliances is notonly a technological advancement in the area of IoT, it can be scaled to serve multiple purposes including the one proposed here to provide better assistant for disabled people, such that it breaks the barriers for the disabled people and allows them work their way around the house freely.

Keywords:

EEG, applications, Mind commands, Brain monitoring,

INTRODUCTION

The Air Quality Index (AQI) is a national system used to measure and report air quality. The AQI looks for five major air pollutants regulated by the Clean Air Act: particle pollution, ground-level ozone, carbon monoxide, nitrogen dioxide, and sulfur dioxide. The EPA(Environment Protection Act) takes daily readings of these pollutants and interprets it into a specific number ranging from zero to 500 and a specific color. Particulate matter is tiny particles in the air like dirt, dust, smoke, and soot which is reported as either PM 2.5 or PM 10. PM 2.5 particles are very tiny. Computation of the AQI requires an air pollutant concentration over a specified averaging period, obtained from an air monitor or model. Its air quality index values are typically grouped into ranges. Each range is assigned a descriptor, a color code. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. In this proposed model we will use Random Forest method to detect air quality index of a few cities in India.

TECHNOLOGIES USED:

IOT-Internet of things is a network of devices and sensor that are connected to the internet. where this are interrelated computing devices, mechanical and digital machines which provides with unique identifiers(UIDs) and it has the ability to transfer data over a network without requiring human to human or human to computer interaction.

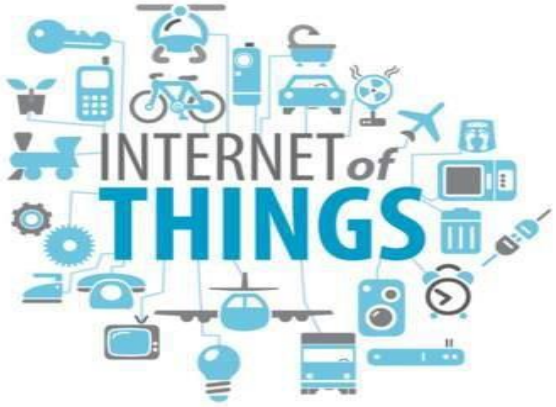


Figure 1 IOT

IOT is a computing concept that describes a future where everyday physical objects will be connected to the internet and be able to identify themselves to other devices. it significant because an object that can represent itself

digitally becomes something greater than the object by itself.

Where this aims to connect all devices to existing internet infrastructure. At present only mobile, computers, smart TV's are connected to internet. But by using IOT all devices can be connected like fan, lights..etc.

Figure 2 IOT applications

EXISTING SYSTEM:

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time.

PROPOSED SYSTEM:

As Machine Learning algorithms gain experience, they keep improving in accuracy and efficiency. Random Forest classifier uses recursive partitioning to generate many trees and then aggregate the results. Each tree is independently constructed using a bootstrap sample of the training data, which subdivides the parameter set first into several parts depending on one of the parameters, and subsequently repeats the process for each part. This lets them make better decision. In this project, a high amount of data of air in the surroundings is required which contains a millions of various gases or other impurities, Machine learning can analyze this data in a efficient way and gives a appropriate result and output.

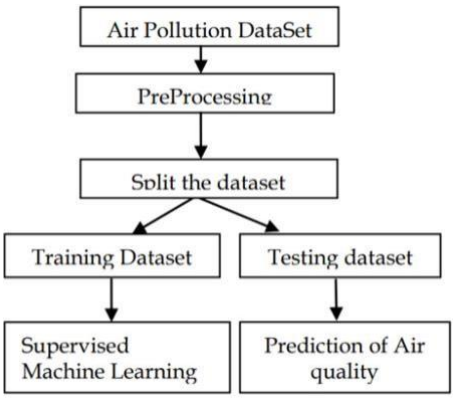
SYSTEM DESIGN

WORKING:

- **Collection**

Data Collection is the process of collecting and measuring information from a variety of sources. It must be collected and stored in a way that makes sense for the problem at hand.The dataset "data.xlsx" includes a
concentration of pollutants and
meteorological factors..

Figure 3 Block Diagram



- **Preprocessing of data**

Data cleaning is performed in preprocessing. It is very much customary to have missing values in the dataset. It may have happened during data collection. To solve this problem the rows with the missing data are eliminated. Object type is converted into numeric type because it is easy for a model to understand numerical inputs. Attribute selection will takes place in the preprocessing. The new attribute is selected from the given set of attributes. The attributes which majorly contribute to air pollution and the row-wise highest value is considered as Air Quality Index. Normalization takes place. It means scaling the data values in the specified range.

Algorithms

Random Forest Algorithm is used to predict the Air Quality Index. Random forest is another supervised learning algorithm that is used for both classifications as well as regression. Random Forest Algorithm constructs decision trees on the available data samples and then gets the prediction from each of them and finally designates the best solution by means of voting

MODULE DESCRIPTION

Our project has three modules mainly data collection, data preprocessing and data visualization. **Data Collection:**

```
raw_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218639 entries, 0 to 218638
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------  -----
 0   TEMP    208169 non-null  object
 1   CH4     95822 non-null   object
 2   CO      217310 non-null  object
 3   NMHC    95614 non-null   object
 4   NO      217227 non-null  object
 5   NO2     216681 non-null  object
 6   NOx     217228 non-null  object
 7   O3      196864 non-null  object
 8   PM10    215761 non-null  object
 9   PM2.5   215768 non-null  object
 10  RH      208143 non-null  object
 11  SO2     217046 non-null  object
dtypes: object(12)
memory usage: 20.0+ MB
```

Figure 4

Data Collection is the process of collecting and measuring information from a variety of sources. It must be collected and stored in a way that makes sense for the problem at hand.The dataset ”data.xlsx” includes a concentration of pollutants and meteorological factors. The total attributes in the dataset are twelve: Temperature, CH4 (Methane), CO (Carbon Monoxide), NMHC (Non Methane Hydro-Carbons),
NO (Nitrogen Monoxide), NO2 (Nitrogen Dioxide), NOx (Nitrogen Oxides), O3
(Ozone), PM10 (Particulate Matter), PM2.5, RH (Relative Humidity), and SO2 (Sulfur Dioxide)

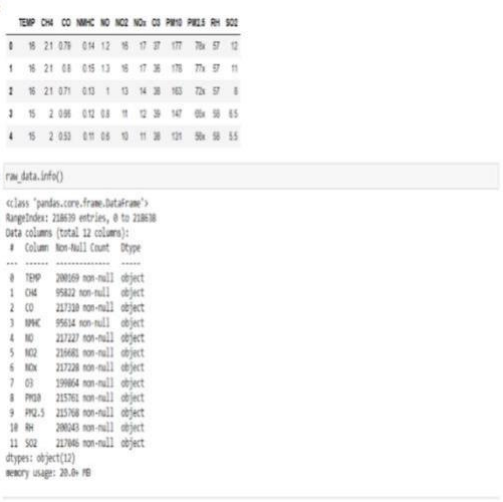
Data Preprocessing and Visualization:

Data visualization is the graphical representation of information and data and it plays an important role in the portrayal of both small-scale and large-scale data. Graphical elements like charts, graphs, and maps, data visualization tools provide an approachable way to see and fathom trends, outliers, and patterns in data..

A dataset can be viewed as a gathering of data objects, which are frequently also called a record, points, vectors, patterns, events, cases, samples, observations, or entities.

1.Cleaning
2.Attribute Selection
3.Normalization
4.Formatting Convert from one file format (xlxs) to another file format (CSV file).

Result and Discussion



```
raw_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218639 entries, 0 to 218638
Data columns (total 12 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------  -----
 0   TEMP    208169 non-null  object
 1   CH4     95822 non-null   object
 2   CO      217310 non-null  object
 3   NMHC    95614 non-null   object
 4   NO      217227 non-null  object
 5   NO2     216681 non-null  object
 6   NOx     217228 non-null  object
 7   O3      196864 non-null  object
 8   PM10    215761 non-null  object
 9   PM2.5   215768 non-null  object
 10  RH      208143 non-null  object
 11  SO2     217046 non-null  object
dtypes: object(12)
memory usage: 20.0+ MB
```

Figure 5 Working model

The proposed system is based on the Random forest Algorithm that creates many decision trees. Accuracy of proposed system is done by using random forest gives the ouput approximately 76 to 78 percent. Random forest implements many decision trees and also gives the most accurate output when compared to the decision tree. Random Forest algorithm is used in the two phases. Firstly, the RF algorithm extracts subsamples from the original samples by using the bootstrap resampling method and creates the decision trees for each testing sample and then the algorithm classifies the decision trees and implements a vote with the help of the largest vote of the classification as a final result of the classification.

CONCLUSION & FUTURE WORK

If there is increased awareness about Air Quality Index India and it’s health impacts depending on the various categories can help to reduce the incidence of air pollution to the most vulnerable people. Since acute exposure to acute exposure to air emissions may cause substantial harm to the health of the masses in general. Therefore, there are variables that can be taken to make people aware of the air-emission reports so that they can plan they’re outdoor activities accordingly to reduce the intake of highly polluted. If there is increased awareness about Air Quality Index India and it’s health impacts depending on the various categories can help to reduce the incidence of air pollution to the most vulnerable people. Since acute exposure to acute exposure to air emissions may cause substantial harm to the health of the masses in general. Therefore, there are variables that can be taken to make people aware of the air-emission reports so that they can

plan they're outdoor activities accordingly to reduce the intake of highly polluted.

REFERENCES

[1] K. Veljanovskal and A. Dimoski, "Air qualityindex prediction using simple machine learningalgorithms," International Journal of EmergingTrends Technology in Computer Science(IJETTCS), 2018

[2] J. Kotcher, E. Maibach, and W.T. Choi, "Fossilfuels are harming our brains: identifying keymessages about the health effects of air pollutionfrom fossil fuels," BMC public health, vol. 19, no.1, p. 1079, 2019.

[3] Khedo K.K., Perseedoss R., Mungur A. A Wireless Sensor Network Air Pollution Monitoring System. Int. J. Wirel. Mob. Netw. 2010;2:31–45. doi: 10.5121/ijwmn.2019.

[4] Ma Y., Richards M., Ghanem M., Guo Y., Hassard J. Air Pollution Monitoring and Mining Based on Sensor Grid in London. Sensors. 2008;8:3601–3623. doi: 10.3390/s80603601.

[5] P.-W. Soh, J.-W. Chang, and J.-W. Huang,"Adaptive deep learning-based air qualityprediction model using the most relevant spatial-temporal relations," IEEE Access, vol. 6, pp.38186–38199, 2018. [6] K. B. Shaban et al., "Urban air pollution monitoring system with forecasting models," IEEE Sensors Journal, vol. 16, no. 8, pp. 2598– 2606, April 2016. [7]Pallavi Pant, Raj M. Lal, Armistead G. Russell, Ajay S. Nagpure, AnuRamaswami, Richard E. Peltie, "Monitoring particulate matter in India: recent trends and future outlook", Air Quality, Atmosphere Health, 2018. [8]. YusefOmidiKhaniabadi, GholamrezaGoudarzi, Seyed Mohammad Daryanoosh, Alessandro Borgini, Andrea Tittarelli, Alessandra De Marco,

[9]U. A. Hvidtfeldt, M. Ketzel, M. Sørensen et al., "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, 2018.

[10]Ziyue Guan and Richard O. Sinnot, "Prediction of Air Pollution through Machine Learning on the cloud", IEEE/ACM5th International Conference on Big Data Computing Applications and Technologies (BDCAT),2019

[11]L. Pimpin, L. Retat, D. Fecht 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, Article ID e1002602, pp. 1–16, 2018.

[12]. BC. Liu, et al, "Urban air quality forecasting based on multi-dimensional collaborative Support Vector (SVR): A case study of Beijing-Tianjin-Method Shijiazhuang", PLOS, 20.