Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765851

AN EFFECTIVE PREDICTION OF AIR POLLUTION BY USING MACHINE LEARNING MODELS

Veeranjaneyulu.Y[1] , Mallikharjuna.T[1] , Sri Charan. T[1]
Dr.T. Nalini[2], Dr.P. Dinesh Kumar[2] , Dr.G.Victo Sudha George[2]
[2]Professor, Dept .of CSE, Dr.M.G.R. Educational and Research Institute
[1]Final Year Btech CSE, Dr.M.G.R. Educational and Research Institute
[1]veeraji0123@gmail.com , mallikharjunat007@gmail.com , tsricharan2@gmail.com
[2]nalini.cse@drmgrdu.ac.in ,dineshkumar.it@drmgrdu.ac.in
,victosudhageorge@drmgrdu.ac.in

Abstract- The regulation of air pollutant levels is rapidly increasing and its one of the most important tasks for the governments of developing countries, especially India. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it.. For Training purpose we are taking an AIR QUALITY DATA IN INDIA(2015-2020) city-day csv data set . And applied RANDOM FOREST, DECISION TREE, LOGISTIC REGRESSION, GAUSSIAN NAIVE BAYES, SUPPORT VECTOR MACHINE Algorithms for all of this RANDOM FOREST Algorithm got highest level of Accuracy 92.8234... For Testing purpose we are Building an IOT based device. This is a simple prototype for an Environmental IoT Air Pollution/Quality Monitoring System for monitoring the concentrations of major air pollutant gases. The system uses 3 sensors like PMS5003 PM2.5 Particulate Matter Sensor, MQ-135 Air Quality Sensor, BME280 Barometric Pressure Sensor. In this IoT project, you can monitor the pollution level from anywhere using your computer or mobile. PMS5003 PM2.5 Particulate matter sensor from Plant power measure particle concentration in PM1.0, PM2.5 & PM10. This MQ-135 Air Quality Sensor measures concentrations of gases such as CO, CO2, SO2, and NO2 and gives the result in PPM (Part per.Million).Similarly,BME280Measures environmental Temperature, Pressure & Humidity.

Keywords— Pollution detection, Pollution Prediction, Logistic Regression, Random forest, particular matter

I. INTRODUCTION

Particulate matter can be either human-made or naturally occur. Some examples include dust, ash and sea-spray. Particulate matter (including soot) is emitted during the combustion of solid and liquid fuels, such as for power generation, domestic heating and in vehicle engines. Particulate matter varies in size (i.e. the diameter or width of the particle). PM2.5 refers to the mass per cubic meter of air of particles with a size (diameter) generally less than 2.5 micrometers (µm). PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high. The regulation of air pollutant levels is rapidly becoming one of the most important tasks for the governments of developing countries, especially China. Submicron particles, such as ultrafine particles (UFP, aerodynamic diameter ≤ 100 nm) and particulate matter ≤ 1.0 micrometers

(PM1.0), are an unregulated emerging health threat to humans, but the relationships between the concentration of these particles and meteorological and traffic factors are poorly understood. To shed some light on these connections, we employed a range of machine learning techniques to predict UFP and PM1.0 levels based on a dataset consisting of observations of weather and traffic variables recorded at a busy roadside in Hangzhou, China. In this project we are going to make an IoT Based Air Pollution/Quality Monitoring System with ESP8266, PM2.5 Particulate Matter Sensor, MQ-135 Air Quality Sensor & BME280 Barometric Pressure Sensor. We will monitor the Air Quality on Thinspeak Server using the internet.

II . METHODOLOGY

There are two primary phases in these system:
Training phase1: The system is trained by using the data in the data set and fits the model based on the algorithm chosen accordingly and the accuracy is checked
Testing phase2: The various components are interfaced together and the project deliverables are built with the help of different circuit designs. The testing, debugging and troubleshooting of the design is performed to test the performance of the design under various conditions. If a circuit design fails to pass the tests, then a newer circuit design should be completed, implemented and tested.

And therefore the data that used in to tain is cvs data set and test with a physical device. The system is designed to detect and predict PM2.5 level and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

III. LITERATURE SURVEY

Gopalakrishnan (2021) [1] combined Google's Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable. The author developed a web application to predict air quality for any location in the city neighborhoods. Sanjeev (2021) studied a dataset that included the concentration of pollutants and meteorological factors. The author analyzed and predicted the air quality and claimed that the Random Forest (RF) classifier performed the best as it is less prone to over-fitting.

Monika Singh Et al. in August 2019 [2] proposed an Air Pollution Monitoring System. This system uses an Arduino microcontroller connected with MQ135 and MQ6 gas sensor which senses the different types of gases present in the environment. It was then connected to the Wi-Fi module which connects to the internet and LCD is used to display the output to the user and buzzer alerts when the ppm crosses certain limit. Their applications were industrial perimeter monitoring, indoor air quality monitoring, site selection for reference monitoring stations, making data available to user

Huixiang Liu (et al.2019) [3] have taken two different cities Beijing and Italian city for the study purpose. They have forecasted the Air Quality Index (AQI) for the city Beijing and predicting the concentration of NOxin an Italian City depending on two different publicly available datasets. The first Dataset for the period of December 2013 to August 2018 having 1738 instances is made available from the Beijing Municipal Environmental Centre which contains the fields like hourly averaged AQI and the concentrations of PM2.5, O3, SO2, PM10, and NO2 in Beijing. The second Dataset with 9358 instances is collected from Italian city for the period of March 2004 to February 2005. This dataset contains the attributes as Hourly averaged concentration of CO, Non methane Hydrocarbons, Benzene, NOx, NO2

Heidar Maleki (et al.2019) [4] predicted the hourly
concentration values for the ambient air pollutants NO2, SO2, PM10, PM2.5, CO and O3 for the stations Naderi, Havashenasi, MohiteZist and Behdasht in Ahvaz, Iran which is the most polluted city in the world. They have also calculated and predicted Air Quality Index (AQI) and Air Quality Health Index (AQHI) for the four air quality monitoring stations in Ahvaz mentioned above. They used Artificial Neural Network (ANN) machine learning algorithm for the prediction of air pollutants concentration (hourly) and two air quality indices AQI and AQHI over the August 2009 to August 2010. Input to ANN algorithms involves the factors such as meteorological parameters, Airpollutants concentration, time and date.

Aditya C R (et al.2018) [5] employed the machine algorithms to detect and forecast the PM2.5 concentration level on the basis of dataset containing atmospheric conditions in a specific city. They also predicted the PM2.5 concentration level for a particular date . First of all they classify the air as polluted or not polluted by using Logistic Regression algorithm and then Auto Regression algorithm was used to predict the future value of PM2.5 depending upon previous records.

Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai 2019 : An Empirical Study of PM2.5 Forecasting Using neural network.[6] In the recent years, a lot of efforts have been made to regulate air pollutant levels in most of the developed and developing countries. Fine particulate matter (PM2.5) is considered to be one of the major reasons behind deteriorating public health and a lot of efforts are being made to keep a check on PM2.5 levels. Accurately forecasting PM2.5 level is a challenging task and has been highly dependent on model based approaches. In this paper, we explore new possibilities to hourly forecast PM2.5. Choosing the right forecasting model becomes a very important aspect when it comes to improvement in prediction accuracy. We used Neural Network Autoregression (NNAR) method for the prediction task. The paper also provides a comparative analysis of prediction performance for additive version of Holt-Winters method, autoregressive integrated moving average (ARIMA) model and NNAR model. The experimentation and evaluation is done

IV.PROPOSED SYSTEM

The proposed system does two tasks

Detects the levels of air pollution values based on given atmospheric values.Predicts the level of air pollution values for a PM 2.5 Logistic regression is employed to detect whether a data sample is either polluted or not polluted. The primary goal is to predict air pollution level in City with the ground data. 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.

ADVANTAGES OF PROPOSED SYSTEM

To Easy to detect and predict pollution levels.
To Efficient approach for better output prediction.
Find the Accurate accuracy using various algorithm

V. PROBLEM STATEMENT

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. Provide limited accuracy as they are unable to predict the extreme points i.e. the pollution maximum and minimum. They use inefficient approach for better output prediction complex mathematical calculations

VI. BLOCK DIAGRAM

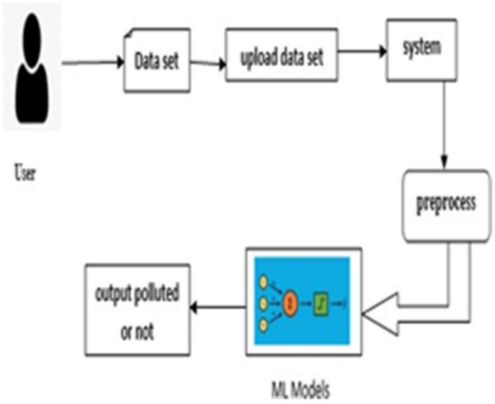


FIG 1: Proposed system Block Diagram

In fig 1 shows that how the data was uploaded and preprocessing was done by the system and ML models was trained the data set to show and shows the accuracy.

CIRCUIT DIAGRAM

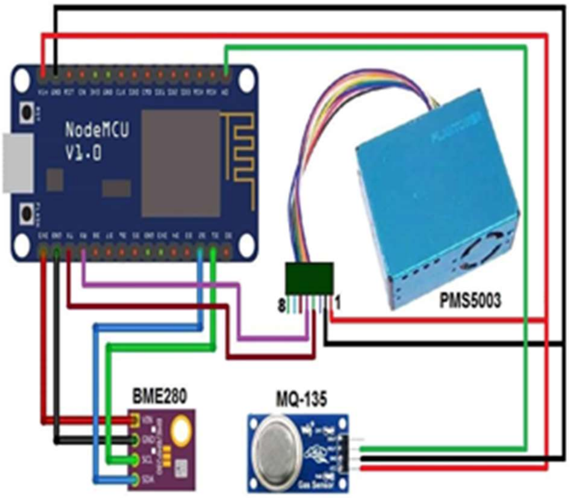


FIG 2: IOT based device circuit diagram

In fig 2 This is an circuit diagram that explains how the IOT Air Pollution/Quality Monitoring System was working and it shows how to connect all the sensors and how it all work together to get the Results

ARCHITECTURE DIAGRAM

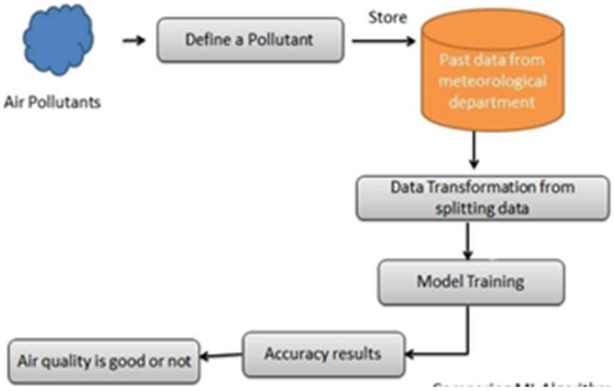

Air Pollutants
Define a Pollutant
Store
Past data from meteorological department
Data Transformation from splitting data
Model Training
Accuracy results
Air quality is good or not

FIG 3: System Architecture Diagram

1. Take dataset:
   System will take uploaded dataset.
2. Pre - processing :
If any null values are present in dataset that can removed in pre-processing step.
3. Training and testing:
System take data for training and testing based on user given test size.
4. Prediction:
System gives the predictions based on given characteristics.
5. predictions:
Using the machine leaning algorithms, we can predict the result

VII. EPERIMENTAL DATA SET

1. Data collection:
Users collect the data from the different websites and kaggle
2. Upload dataset:
User uploads the dataset in this step.
3. View dataset: Python
In this step uploaded dataset is viewed by the user.
4. Model performance:
In model performance step machine learning supervised algorithms are performed.
5. View predictions:
users view the predicted outputs

VIII. ALGORITHMS

LOGISTIC REGRESSION ALGORITHM

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

For example,

$$\textbf{LOGit (P)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{12} + \beta_2 \cdot x_{12} + \_ + \beta_p \cdot x_{in}$$

Logit function is used to generate log odds of anattribute that signifies the probability of the attribute. Log odds are an alternate way of expressing probabilities, which simplifies the process of updatingthem with new evidence.Based on logit function, the system classifies the training data to be either 0 (not polluted) or 1 (polluted) and verifies its accuracy using the test data. The result of the user input is also 0/1 and not the PM2.5 level.Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

DECISION TREE ALGORITHM

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal.A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stopIn general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop

RANDOM FOREST ALGORITHM

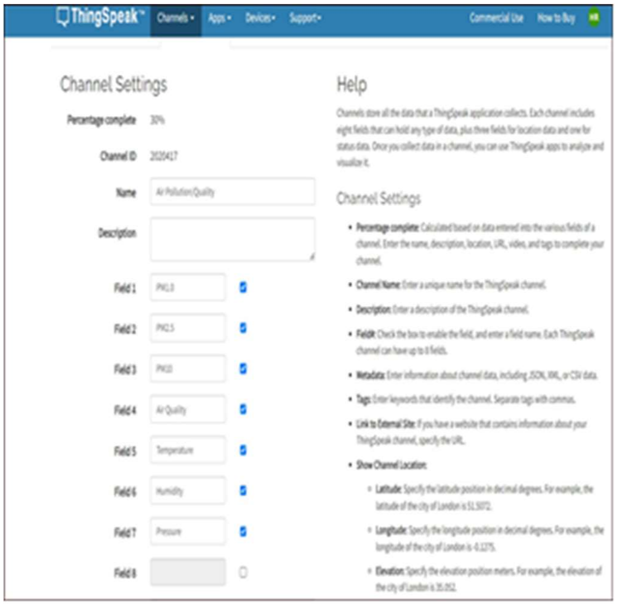
A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision. It generates predictions without requiring many configurations in packages (like Scikit-learn).

Features of a Random Forest Algorithm:

- It's more accurate than the decision tree algorithm.
- It provides an effective way of handling missing data.
- It can produce a reasonable prediction without hyper-parameter tuning.
- It solves the issue of over fitting in decision trees.
- In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. An overview of decision trees will help us understand how random forest algorithms work. A decision tree consists of three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm divides a training dataset into branches, which further segregate into other branches.

RESULT



Channel Settings

Percentage complete	30%
Channel ID	2020417
Name	Air Pollution/Quality
Description	
Field 1	PM1.0 ☑
Field 2	PM2.5 ☑
Field 3	PM10 ☑
Field 4	Air Quality ☑
Field 5	Temperature ☑
Field 6	Humidity ☑
Field 7	Pressure ☑
Field 8	☐

Help

Channels store all the data that a ThingSpeak application collects. Each channel includes eight fields that can hold any type of data, plus three fields for location data and one for status data. Once you collect data in a channel, you can use ThingSpeak apps to analyze and visualize it.

Channel Settings

- **Percentage complete:** Calculated based on data entered into the various fields of a channel. Enter the name, description, location, URL, video, and tags to complete your channel.
- **Channel Name:** Enter a unique name for the ThingSpeak channel.
- **Description:** Enter a description of the ThingSpeak channel.
- **Field#:** Check the box to enable the field, and enter a field name. Each ThingSpeak channel can have up to 8 fields.
- **Metadata:** Enter information about channel data, including JSON, XML, or CSV data.
- **Tags:** Enter keywords that identify the channel. Separate tags with commas.
- **Link to External Site:** If you have a website that contains information about your ThingSpeak channel, specify the URL.
- **Show Channel Location:**
  - **Latitude:** Specify the latitude position in decimal degrees. For example, the latitude of the city of London is 51.5072.
  - **Longitude:** Specify the longitude position in decimal degrees. For example, the longitude of the city of London is -0.1275.
  - **Elevation:** Specify the elevation position meters. For example, the elevation of the city of London is 35.052.

FEG 4: Creating a channels in Things speak

IN fig 4 we can see how the Things speak software was used and how we can create an separate channels in this system

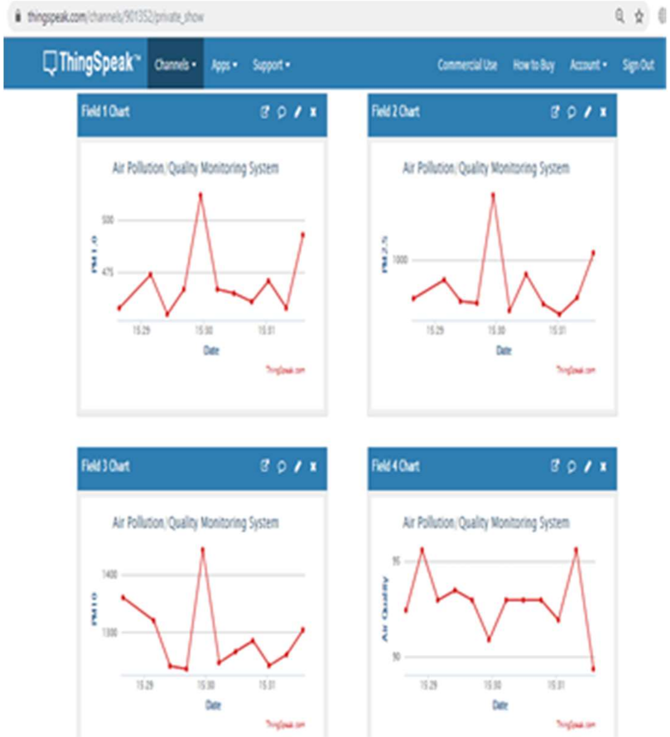


FIG 5: Checking PM2.5 and Air quality in Graph

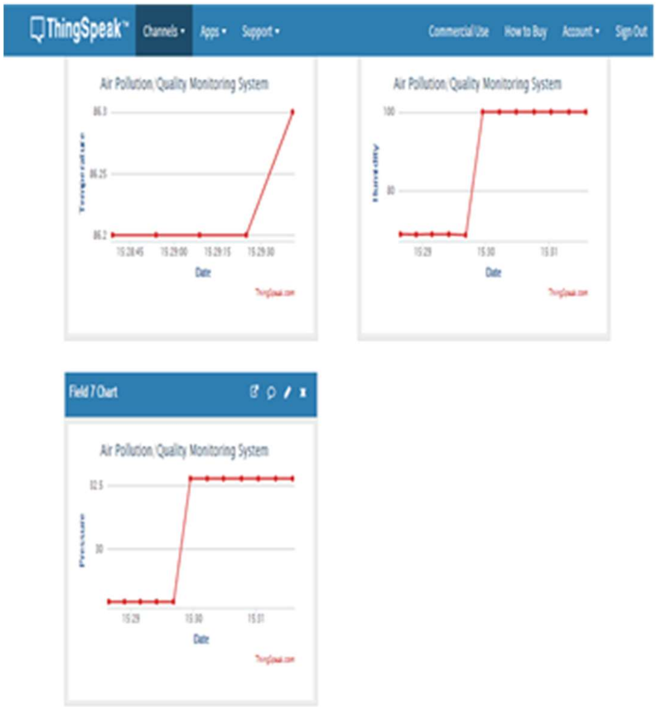
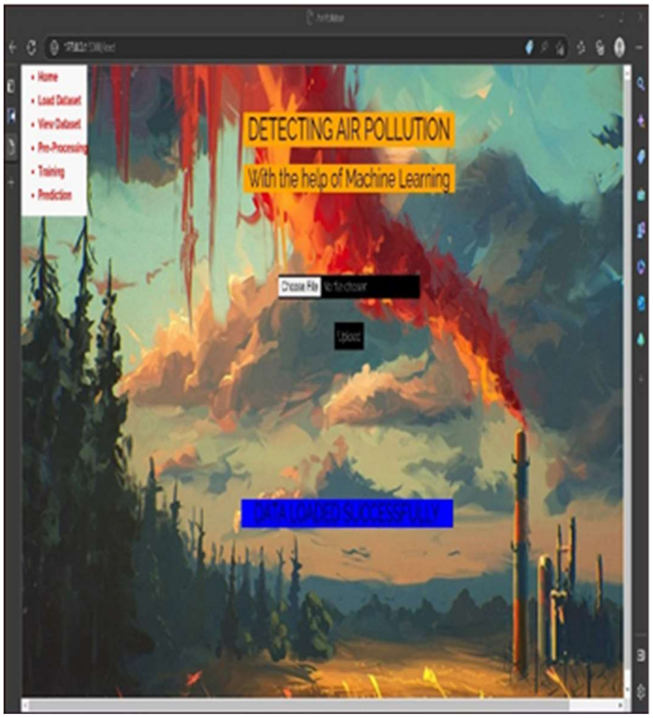
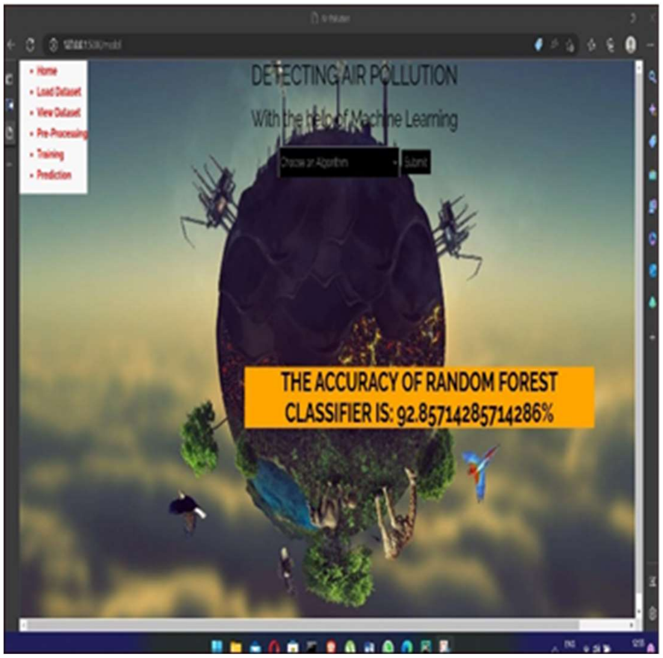


FIG 6: Checking Temparature. Humidity, Pressure



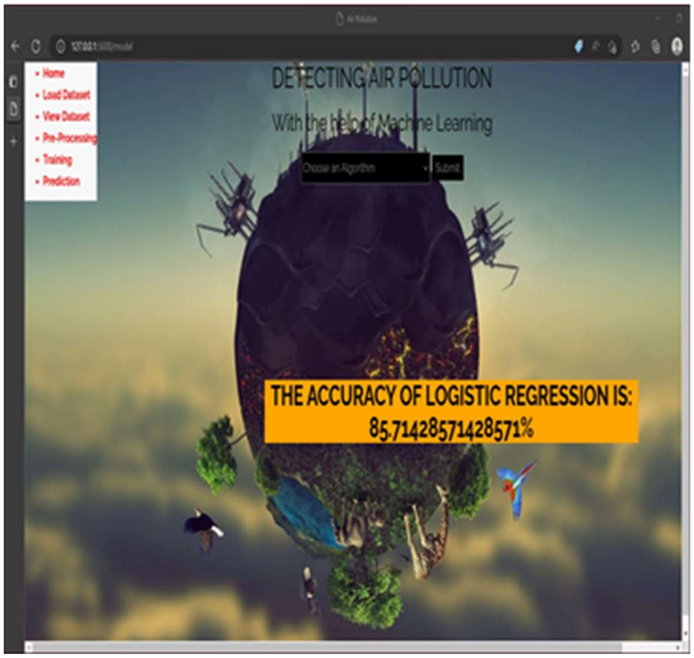
DATA LOADED SUCCESSFULLY

FIG 7: Data Set was uploaded



THE ACCURACY OF RANDOM FOREST CLASSIFIER IS: 92.857142857142 86%

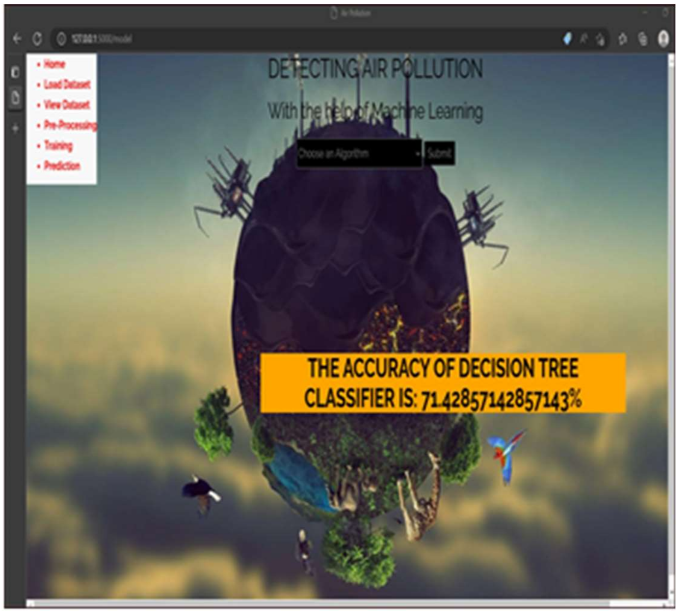
FIG 8: Checking the Accuracy of Algorithm



THE ACCURACY OF LOGISTIC REGRESSION IS: 85.71428571428571%

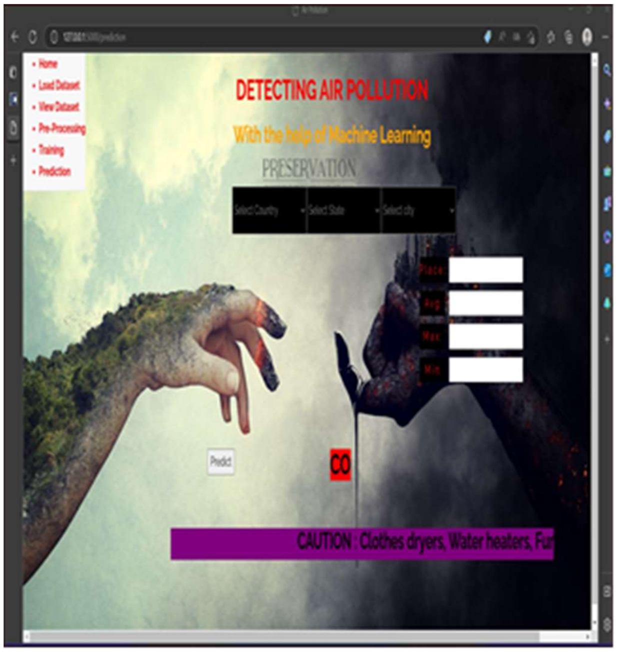
In the above fig 8 shows that Accuracy of the Random forest and its classifier is 92.85721..

FIG 9: Checking the Accuracy of Algorithm



THE ACCURACY OF DECISION TREE CLASSIFIER IS: 71.42857142857143%

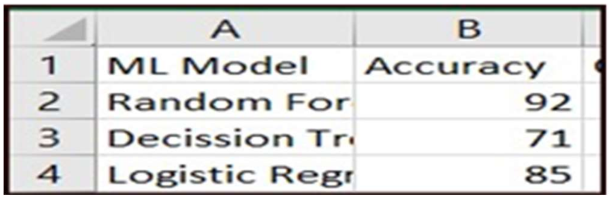
In the above fig 9 shows that Accuracy of Logistic regression and its classifier is 85.714285..



CAUTION : Clothes dryers, Water heaters, Fu

FIG 10: Checking the Accuracy of decision tree Algorithm

In the above fig 10 shows that Accuracy of Decision Tree and its classifier is 71.428571…



	A	B
1	ML Model	Accuracy
2	Random For	92
3	Decission Tr	71
4	Logistic Regr	85

FIG11: Predicting air quality at given max values

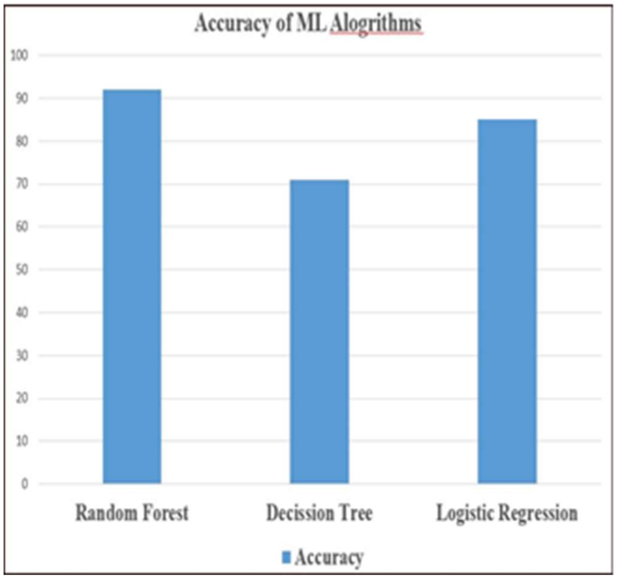


Fig 12 : Accuracy of three different ML Algorithms

We can see Fig 12 It shows the Accuracy of three different types of ML Algorithms and the classifiers

CONCLUSION

In this paper we designed to help a person to detect, monitor, and test air pollution in a given area. The kit has been integrated with a mobile application that helps the user. As a result, this paper is to check the quality of the exposed level in the air pollution by using an IOT based Device it will check that how much percentage of PM2.5 ,CO,CO2,NO2,SO2 PM1.0 ,PM10 was present in the air .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 is and takes a step towards fighting against it. Different types of algorithms was used to check the accuracy of the pollutants in the taken data set. In this application it monitors the pollutant level through that way of Graph and with the individual data and time .Design a tool which Sense how much carbon mono oxide (CO) ,PM 2.5,SO2,NO2 is present in air and display in the form of percentage ,Sense the temperature and display it in degree Detect the pollution in the air And it helps to people establish a data source for small localities Sense the Quality of Air and Display in the form of percentage .and Graph Effecientlly used to detect the quality of air and predict the level of PM2.5 in the future

REFERENCES

[1] Gopalakrishnan (2021) Hyperlocal air quality prediction using machine learning.Towards data science. https://towardsdat ascie nce.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71

[2]. Monika Singh, Misha Kumari, Pradeep Kumar Chauhan, (2019) 'IoT Based Air Pollution Monitoring System using Arduino', International Research Journal of Engineering and Technology, IRJET

[3]. Huixiang Liu (et al.2019) concentrations of PM2.5, O3, SO2, PM10, and NO2 in Beijing Dataset with 9358 instances is collected from Italian city for the period of March 2004 to

February 2005. This dataset contains the attributes as Hourly averaged concentration of CO, Non methane Hydrocarbons, Benzene, NOx, NO2

[3] HeidarMaleki (et al.2019) predicted the hourly concentration values for the ambient air pollutants NO2, SO2, PM10, PM2.5, CO and O3 for the stations Naderi, Havashenasi, MohiteZist and Behdasht in Ahvaz

[4]. Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai: An Empirical Study of PM2.5 Forecasting Using neural network. IEEE Smart World Congress, At San Francisco, USA [2017]

[5] Aditya C R (et al.2018) employed the machine algorithms to detect and forecast the PM2.5 concentration level on the basis of dataset containing atmospheric conditions in a specific city

[6] M. Caselli &amp; L. Trizio &amp; G. de Gennaro &amp; P. Ielpo. &quot;A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model.&quot; Water Air Soil Pollut (2009) 201:365–377.

[7] Nitin Sadashiv Desai, John Sahaya Rani Alex, (2017) ‘IoT based air pollution monitoring and predictor system on Beagle Bone Black’, International Conference on Nextgen Electronic Technologies, ICNET.

[8] K. S. E. Phala, A. Kumar, and Gerhard P. Hancke, ‘Air Quality Monitoring System Based on ISO/IEC/IEEE 21451 Standards’, IEEE.

[9] Poonam Pal, Ritik Gupta, Sanjana Tiwari, Ashutosh Sharma, (2017) ‘IoT based air pollution monitoring system using Arduino’, International Research Journal of Engineering and Technology IRJET.

[10] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.