International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

Air Quality Prediction by Classification of Supervised Machine Learning

T. R. Saravanan, V. Pavithra, G.Saranya

Abstract: ***Generally, air pollution refer to the release of various pollutants into the air which are threatening the human health and planet as well. The air pollution is the major dangerous vicious to the humanity ever faced. It causes major damage to animals, plants etc., if this keeps on continuing, the human being will face serious situations in the upcoming years. The major pollutants are from the transport and industries. So, to prevent this problem major sectors have to predict the air quality from transport and industries .In existing project there are many disadvantages. The project is about estimating the PM2.5 concentration by designing a photograph based method. But photographic method is not alone sufficient to calculate PM2.5 because it contains only one of the concentration of pollutants and it calculates only PM2.5 so there are some missing out of the major pollutants and the information needed for controlling the pollution .So thereby we proposed the machine learning techniques by user interface of GUI application. In this multiple dataset can be combined from the different source to form a generalized dataset and various machine learning algorithms are used to get the results with maximum accuracy. From comparing various machine learning algorithms we can obtain the best accuracy result. Our evaluation gives the comprehensive manual to sensitivity evaluation of model parameters with regard to overall performance in prediction of air high quality pollutants through accuracy calculation. Additionally to discuss and compare the performance of machine learning algorithms from the dataset with evaluation of GUI based user interface air quality prediction by attributes.***

Keyword: pollutants dataset, GUI results, machine learning, supervised algorithm.

I. INTRODUCTION

Machine learning method is used to predict the future from the past data. Machine learning is a component or sort of artificial intelligence that provide laptop with capability to learn without being programmed. Machine learning technique is the one that can change when exposed to new data. Machine learning is nothing but the ability to learn by itself and teaches the computer how to respond to an input by itself. It contains many algorithm into it.

Revised Manuscript Received on March 18, 2020.
* Correspondence Author
Dr.T.R.Saravanan*, department of computer science and engineering, JEPPIAAR SRR engineering college, Chennai, India. Email: saravanantrcse@gmail.com
V. Pavithra, department of computer science and engineering, JEPPIAAR SRR engineering college, Chennai, India. Email: pavicapricon3@gmail.com
G. Saranya, department of computer science and engineering, JEPPIAAR SRR engineering college, Chennai, India. Email: saranyanadar19599@gmail.com

The algorithm are classified into 3 main categories they are supervised learning, unsupervised learning and reinforcement learning. Supervised learning is the task of learning function that makes an input to an output based on example of input and output pairs. The main feature of supervised algorithm is getting to know set of rules is analysing the training records and produces an inferred characterises which can be used for mapping new examples. Unsupervised learning is used to draw inferences from dataset consisting of input data without labelled responses. Its main method is cluster analysis which can be used for exploratory data analysis to find hidden patterns or grouping in data .Finally, reinforcement learning concerned with how software agents ought to take action in an environment in order to maximize some notion of cumulative reward.

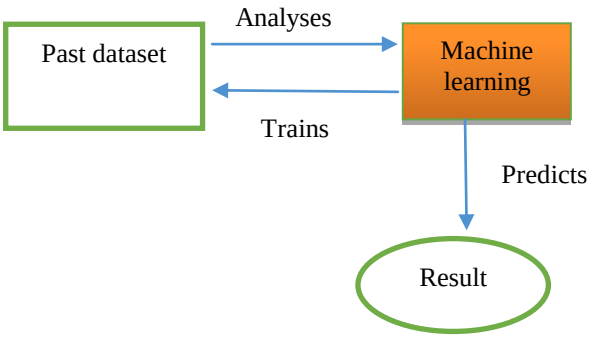

Fig: process of machine learning

To estimate the PM2.5 concentration by designing a photograph based method. By observation it is found that the quality of air in saturation map crediting entirely different approaches on high and low concentrations. They tend to loss their substances and mostly the pixel values tend to be zero under high concentration. When they try to make it similar the quality of structural information losses. The usage of weibull distribution is able to derive the value to estimate the color information. The photographic method is not sufficient to calculate PM2.5 concentration [1] [9]. It takes only one value of the pollutants in the air. It does not calculates all the pollutants. Air pollutants causes the major threats to the human society. They will be major for causing disease to human as well to the living organisms on earth. If this situation has to be overcome the pollution causing pollutants should be should be found out. [6][4]

To overcome this disadvantage from existing system we have implemented the technique of machine learning. [5] Various algorithm from the machine learning is used for getting the best result. The machine learning approach by user interface of GUI application is used as a proposed system. They are used to analyses the multiple dataset from the different sources and all are combined to form the generalized dataset. The process at first defines a problem of what the people are facing and preparing the dataset from the past report and then evaluating the dataset like removing null values, repeated values, maximum and minimum pollutants etc., then often obtaining the evaluation of algorithm will takes place. By comparing the same dataset with different machine learning algorithms it finds the best result from the comparison and predicts the result for analyzer by GUI interface.

II. LITERATURE REVIEW

GuanghuiYue ,KeGu, and JunfeiQiao, Member, have proposed that to estimate that PM2.5 concentration by designing a photograph-based method. It is found that the saturation map is sensitive to air quality, exhibiting entirely different appearances under high and low PM2.5 concentrations. To compute the gradient similarity between the saturation and grey-scale maps to quantify the structural information loss. Utilizing the Weibull distribution to fit the saturation map and able to derive a value to estimate the colour information. Finally, the PM2.5 concentration of an image can be estimated via the combination of the aforementioned two features followed by a nonlinear mapping procedure. Both numerical and visualized results on real captured data validate the effectiveness and superiority of the proposed method in comparison with the relevant state of-the-art methods. Air pollution has become a worldwide concerned issue and automatically estimation of air quality can provide a positive guidance to both individual and industrial behaviours. [1]

Ishan Verma, Rahul Ahuja and Hardik Meisheri, proposed the method which state about concept of Recurrent Neural Networks (RNN) has proved to be very efficient in processing temporal data It is difficult to obtain optimal merging since different networks trained on the same data can no longer be regarded as independent it proposed bidirectional recurrent neural network (BRNN) that can be trained using all available input information in the past and future of a specific time frame. [2]

Temesegan Walelign Ayele, Rutvik Mehta, proposed that nowadays it is better if every action is done using new technology in order to satisfy the demand of human being, Organization, Enterprise etc. Internet of Things (IoT) is one of the main communication developments in the last decade. Through this concept, it is possible to connect countless low-powered smart embedded objects to each other and to the Internet. [3]

Luke Curtis, William Rea, Patricia Smith-Willis, proposed that, the goal of this review is to concisely summarize a wide range of the recent research on health effects of many types of outdoor air pollution. An evaluation of the health consequences of main outdoor air pollution which includes particulates, carbon monoxide, sulphur and nitrogen oxides, acid gases, metals, volatile organics, solvents, pesticides, radiation and bio aerosols is presented. [4]

Khaled Bashir Shaban, Abdullah Kadri and Eman Rezk, proposed that, air quality data are collected wirelessly from monitoring motes that are equipped with an array of gaseous and meteorological sensors. These data are analysed and used in forecasting concentration values of pollutants using intelligent machine to machine platform. The platform uses ML-based algorithms to build the forecasting models by learning from the collected data. [5]

III. PROPOSED METHODOLOGY

Problem identification:

Monitoring and maintaining air satisfactory has turn out to be one of the most vital activities in many industrial and concrete areas today. The excellent of air is adversely affected because of various varieties of pollution due to transportation, electricity, fuel makes use of etc. The deposition of harmful gases is creating a serious threat for the quality of life in smart cities. With increasing air pollution, we need to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. The existing system with Concept of Recurrent Neural Networks (RNN) [2] has proved to be very efficient in processing temporal data it is difficult to obtain optimal merging since different networks trained on the same data can no longer be regarded as independent

For this they are using the sensor which are needed monitored regularly that they are working are not. [3][8] For this kind of problem usage of hardware's are avoided in proposed model. One of the existing system offers the prediction of air quality after 1 hour [7], they delay of such method are avoided by implementing anaconda navigator which gives the instant result.

Now a days the monitoring of air and preserving air quality has become the most essential activity on many areas like urban and industrial areas. The quality of air has become adversely affected due to the various forms of pollution from transportation, industries, coal products[10], the deposition of harmful gases on to the air causes the serious threat to human life with the increasing air pollution we need to find the solution. By collecting the information about the pollutants and provide the final report to the each area about the present condition of their area. Therecorda are submitted to Indian meteorological sector using machine learning techniques. The prediction are from air quality index value. The dataset about the air pollutants are taken as an input and then entering into the process of data processing.

Table: Air pollutants range estimated by government

AQI	Associated Health Impacts
Good (0–50)	Minimal impact
Satisfactory (51–100)	Sensitive people may suffer from minor breathing discomfort.
Moderately polluted (101–200)	May reason breathing pain to human beings with lung disease which include asthma, and soreness to people with heart disease , kids and older adults
Poor (201–300)	May reason respiration soreness to people on prolonged exposure, and discomfort to people with heart disease
Very poor (301–400)	May motive respiratory illness to the human beings on extended exposure. Effect may be extra suggested in humans with lung and heart diseases
Severe (401–500)	May cause respiratory impact even on healthful humans, and critical health influences on people with lung/heart disease. The health influences can be experienced even during light bodily activity

IV. BLOCK DIAGRAM

The dataset which are entering into the data processing phase are undergoing the process of finding the data shape, data type, elimination of null values etc., the output from the data processing is obtained as a proper dataset with correct values and no repeated values. Then these are entering into the training and testing of dataset.

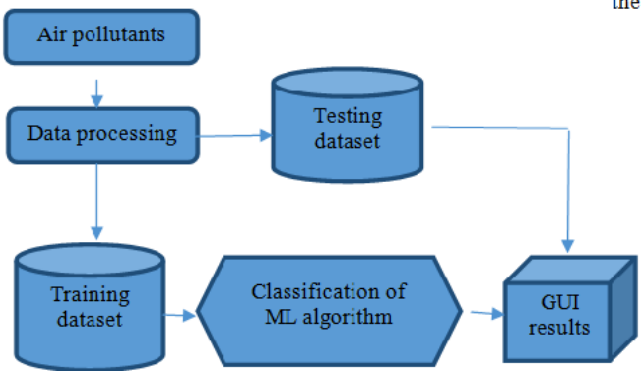

Fig: Architecture for proposed model

Here the dataset is trained to the machine and tested by machine. Then the trained dataset enter into the machine learning algorithms where many algorithms are compared for finding the best accuracy result. The supervised Machine learning algorithms are used and the final output is displayed in GUI interface.

V.PHASE

A. Data validation and pre-processing technique
B. Exploration data analysis of visualization and training a model by given attributes
C. Performance measurements of Logistic regression and Naive Bayes algorithms
D. Performance measurements of Random Forest and Support Vector Machines
E. Result in GUI

A.Data validation and pre-processing technique:

This process used to get the error rate of the machine learning model can be considered as close to the true error rate of the dataset. If the data volume is large enough to population, there is not necessary for validation technique. But in the real world, working with the sample of data that may not have a true representative of the given dataset. Finding the missing values, duplicated values and detail about data type i.e. whether float or integer etc... Data collection, data analysis and the process of addressing data content quality and structure can be add up to a time consuming process. The data cleaning process is done by using python's pandas library. They specifically focus on biggest data cleaning task, missing value and it is more quick clean data process.it takes less time for cleaning. Parting the library packages with given dataset. The analyses is done by their data shape, data type, evaluating missing values and duplicate values.

Variable identification with Uni-variant, Bi-variant and Multi-variant analysis:

It used to find the missing values of data frame, finding duplicate values, finding unique values, find count values of data frame. It explains about the data frame, given dataset. It eliminates the extra columns and rename and drop the given data frame.it specify the type of values.

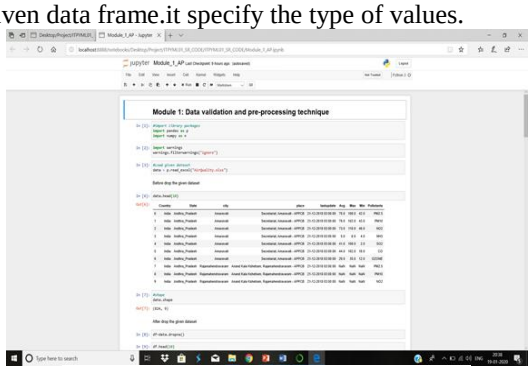

Fig: Importing the python packages

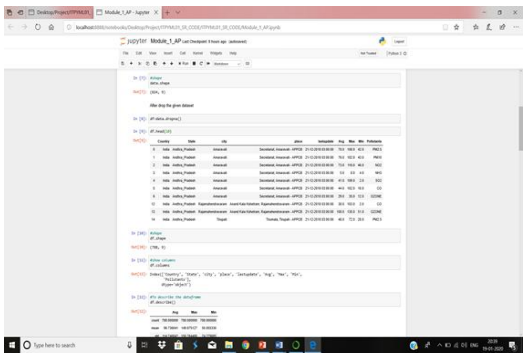

Fig: elimination of null values, repeated values

B. Exploration data analysis of visualization and training a model by given attributes:

Data visualization is a crucial skill in applied statistics and device learning. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

1. How to chart time series data with line plots and categorical quantities with bar charts.
2. How to summarize data distributions with histograms and box plots.
3. How to summarize the relationship between variables with scatter plots.

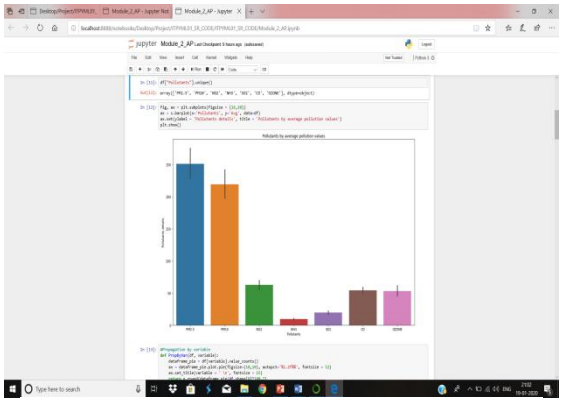

Fig: estimating the dataset values in bar chart

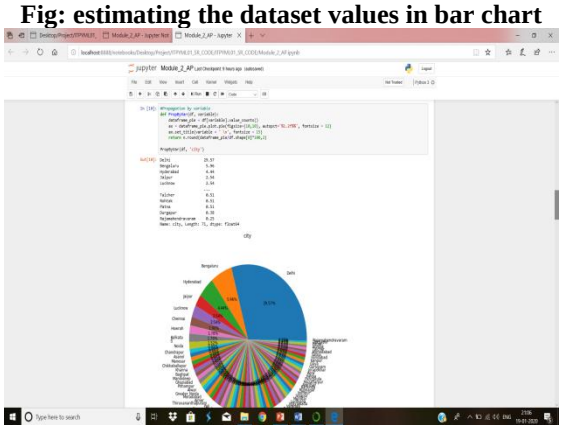

Fig: Estimating the country's dataset values in pie chart

C. Performance measurements of Logistic regression and Naive Bayes algorithms:

Calculation of accuracy from different algorithm

1. Algorithm

2.1 Logistic regression:

It is a statistical method for analyzing a statistics set in which there are one or more unbiased variables that determine a final results. The final results. The final results is measured with a dichotomous variable (in which there are most effective two feasible outcomes). The purpose of logistic regression is to find the satisfactory fitting version to explain the relationship among the dichotomous feature of interest (based variable =response or outcome variable) and a fixed of impartial (predictor or explanatory) variables. Logistic regression is a machine learning classification algorithm that is used to be expecting the opportunity of a categorical structured variable.

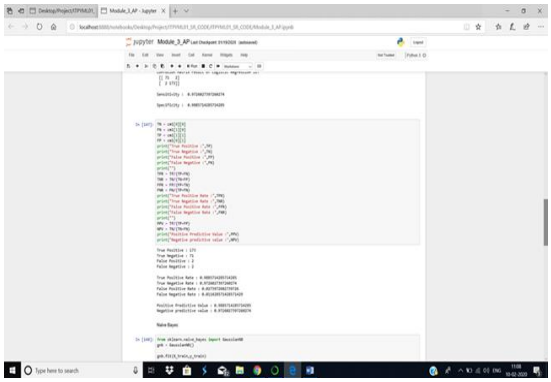

Fig: calculation of values by logistic regression algorithm and finding the accuracy value through algorithm

2.2 Naive Bayes:

Naïve Bayes model is simple to build and particularly useful for very huge statistics sets. Along with simplicity, its miles recognized to outperform even highly sophisticated classification methods. It is straight forward and fast to be expecting elegance of test information set. It also perform nicely in multi class prediction when assumption of independence holds, a Naïve Bayes classifier performs better evaluate to other models like logistic regression and need less training records. It perform property in case of categorical input variables compared to numerical variable, everyday distribution is assumed (bell curve, that's a robust assumption).

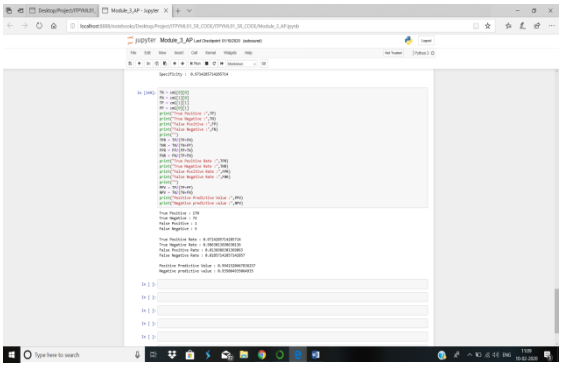

Fig: Calculation of accuracy values through naive Bayes algorithm

D. Performance measurements of Random Forest and Support Vector Machines:

2.3 Random forest:

Random forest is a sort of supervised system studying algorithm primarily based on ensemble gaining knowledge of ensemble learning is a type of gaining knowledge of in which you be part of different sorts of algorithm or equal algorithm multiple instance to shape a more effective prediction model. The random forest area set of rules may be used for each regression and class tasks.

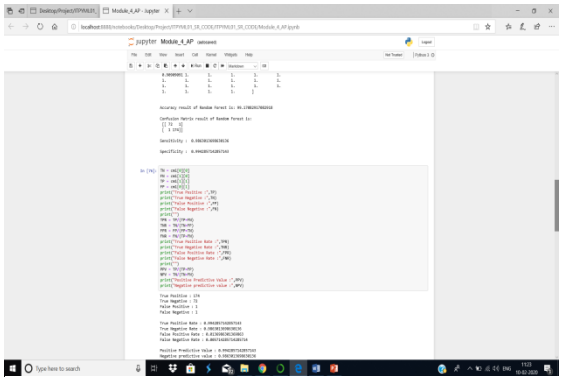

Fig: Calculation of accuracy values through Random forest algorithm

2.4 Support vector machines:

A classifier that categorizes the statistics set through setting a top-quality hyper plane bet ween records. This classifier has been pretty versatile inside the number of various kernelling functions that may be carried out and this model can yield an excessive predictability rate. Support Vector Machines are possibly one of the most famous and talked about machine learning algorithms. They have been extremely famous around the time they have been developed within the Nineteen Nineties and remain the go-to method for an excessive-appearing algorithm with little tuning.

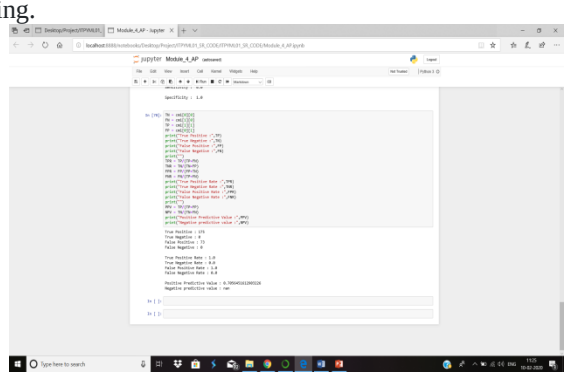

Fig: Calculation of accuracy values through supporting vendoring algorithm

The accuracy values are calculated for all algorithm though the general formula:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
True Positive Rate (TPR) = TP / (TP + FN)
False Positive (FPR) = FP / (FP + TN)
TP - True Positive FP - False Positive

VI. FLOWCHART

Machine learning needs data gathering have lot of past data's. Data gathering have sufficient historical data and raw data. Before data pre-processing, raw data can't be used directly. It's used to pre-process then, what kind of algorithm with model. Training and testing this model working and predicting correctly with minimum errors. Tuned model involved by tuned time to time with improving the accuracy.

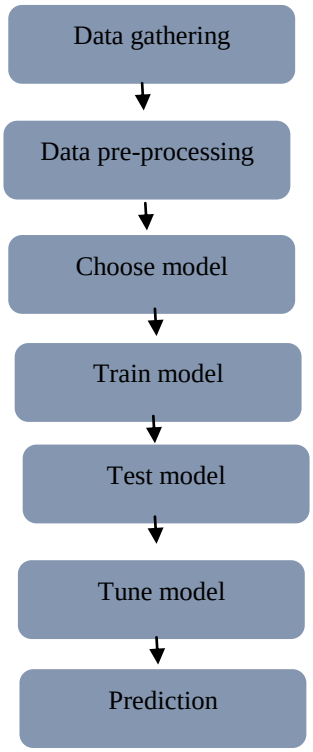

Fig Process of flowchart diagram

E. Result In Gui

The final output contain the module state, city, air quality index value by user and pollution prediction value, source for pollution and AQI stages are listed.

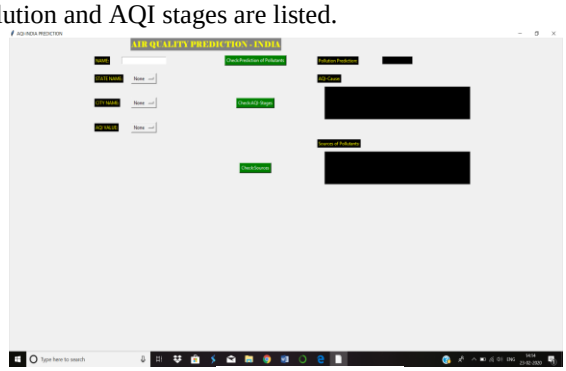

Fig: final output

VII. RESULT ANALYSIS

Final result is obtained through anaconda navigator which gives the line by line input immediately.

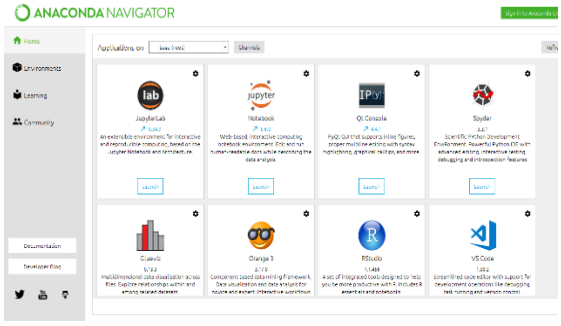

Fig: anaconda navigator

After launching the anaconda navigator then launch jupyter notebook.

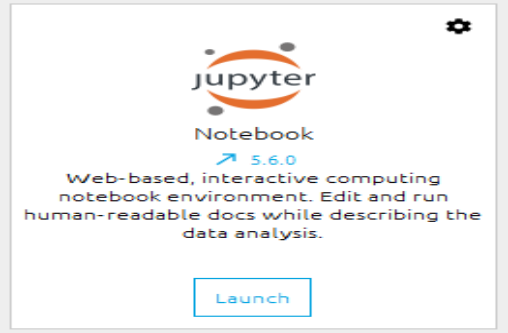

Fig: jupyter notebook

Then select the folder where it has been saved and run each module.

Fig: folders

After running the final module the user is provided with output page in GUI format.

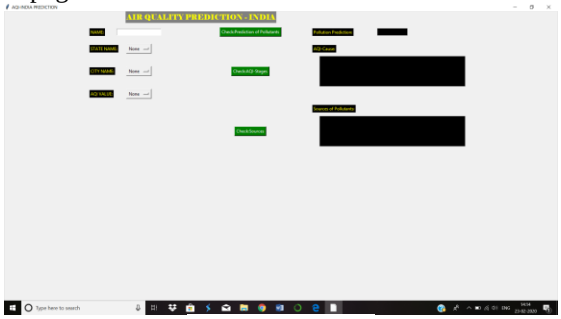

Fig: final output

VIII. CONCLUSION

The analytical process began from records cleaning and processing, missing value, exploratory analysis and subsequently model constructin g and evaluation. The pleasant accuracy on public take a look at set is better accuracy rating is could be discover out. This application can assist India meteorological branch in predicting the future of air best and its popularity and depends on that they can take action.

FUTURE ENCHANCEMENT

- India meteorological department wants to automate the detecting the air first-class is right or no longer from eligibility method (actual time).
- To automate this technique by display the prediction result in internet software or desktop application.
- To optimize the paintings to put into effect in Artificial Intelligence environment.

REFERENCES

1. Guanghui Yue , Ke Gu , and Junfei Qiao, Member," Effective and Efficient Photo -Based PM2.5 Concentration Estimation", website: http://www.ieee.org/publications_standards/publications/rights/index.html
2. Ishan Verma, Rahul Ahuja and Hardik Meisheri, "Air pollutant severity prediction using Bi-directional LSTM Network" website: https://ieeexplore.ieee.org/abstract/document/8609664
3. Temesegan Walelign Ayele, Rutvik Mehta, "Air pollution monitoring and prediction using IoT" website: https://ieeexplore.ieee.org/abstract/document/8473272
4. Luke Curtis, William Rea, Patricia Smith-Willis, "Adverse health effects of outdoor air pollutants", website: https://www.sciencedirect.com/science/article/pii/S0160412006000444
5. Khaled Bashir Shaban, Abdullah Kadri and Eman Rezk," Urban Air Pollution Monitoring System with Forecasting Models", website: https://ieeexplore.ieee.org/abstract/document/7370876
6. C. A. Pope, III, et al., "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution," J. Amer. Med. Assoc., vol. 287, no. 9, pp. 1132–1141, 2002.
7. K. Gu, J. Qiao, and W. Lin, "Recurrent air quality predictor based on meteorology- and pollution-related factors," IEEE Trans. Ind. Informat., vol. 14, no. 9, pp. 3946–3955, Sep. 2018.
8. G. Andria, G. Cavone, V. Di Lecce, and A. M. L. Lanzolla, "Model characterization in measurements of environmental pollutants via data correlation of sensor outputs," IEEE Trans. Instrum. Meas., vol. 54, no. 3, pp. 1061–1066, Jun. 2005.
9. K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
10. Y. Zhao, S. Wang, L. Duan, Y. Lei, P. Cao, and J. Hao, "Primary air pollutant emissions of coal-fired power plants in China: Current status and future prediction," Atmos. Environ., vol. 42, no. 36, pp. 8442–8452, 2008.

AUTHOR PROFILE

Dr. T. R. Saravanan received the Bachelor of Technology Degree in Information Technology from Adhiparasakthi Engineering College , University of Madras in 2004, Master of Information Technology Degree from Sathyabama University in 2007 and completed PhD in Computer science and Engineering from Sathyabama University in 2017.Currently he is working as Assistant professor in Department of Computer Science and Engineering at JEPPIAAR SRR Engineering College , Chennai, He has attended many Faculty training programs in advanced technologies in many leading Institutions, His Research interests include Computer networks, Location based networks, Data analysis, Machine learning, Big data and Data mining. He has received IBM Certification and NPTEL Certification. He is life time member of Indian institute of Technical Education and Institute of Engineers India. Mail id:saravanantrcse@gmail.com

V. Pavithra is currently pursuing her Bachelor of Engineering in Computer Science and Engineering at JEPPIAAR SRR engineering college, padur Chennai in 2020. She is passionate about the field of machine learning. She is strong developer in software design and also interested in android development. His research interest is machine learning and artificial intelligence. Mail id: pavicapricon3@gmail.com

Saranya is currently pursuing her Bachelor of Engineering in Computer Science and Engineering at JEPPIAAR SRR engineering college, padur Chennai in 2020. She is passionate about the field of machine learning. She is interested in application development. His research interest is network security and artificial intelligence. Mailid: saranyanadar19599@gmail.com