International Journal of Research (IJR)

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 5
May 2021

THE PREDICTION OF QUALITY OF THE AIR USING SUPERVISED LEARNING

Chakradhar Reddy K
Dept. of CSE
Hindustan Institute of Technology and Science
Chennai,India
chakradharreddy2016@gmail.com

Nagarjuna Reddy K
Dept. of CSE
Hindustan Institute of Technology and Science
Chennai, India
kallamnagarjuna99@gmail.com

Brahmaji Prasad K
Dept. of CSE Hindustan Institute of Technology and science, Chennai
kbprasad519@gmail.com

Dr.P. Selvi Rajendran
Professor, Dept. of CSE
Hindustan Institute of Technology and Science,Chennai
selvir@hindustanuniv.ac.in

***Abstract*:** In general, Pollution of the climate refers to the discharge of pollutants into the atmosphere that damage human health and the environment as a whole. It has the ability to be one of the most dangerous things humans have ever experienced. It damages livestock, crops, and forests, among other things. To avoid this issue in mostly urban areas, the popular approach such as machine learning techniques may be used to predict air quality from contaminants. As a consequence, the quality of the air assessment and forecasting has become an important field of study. The goal is to develop machine learning-based air quality forecasting techniques that are as accurate as possible. The supervised machine learning technique (SMLT) will be used to gather several pieces of information from the dataset, including variable recognition, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatment and analysis, data cleaning/preparation, and data representation. Our findings provide a valuable guide to sensitivity analysis of model parameters in terms of success in air quality pollution prediction through accuracy measurement. By comparing supervise classification machine learning algorithms and generating prediction results in the form of best accuracy, create a machine learning-based method for accurately predicting the Air Quality Index value. Furthermore, to compare and discuss various machine learning algorithms in order to determine the most accurate algorithm with the performance of a GUI-based user interface for air quality prediction.

Keywords: Supervised learning, Machine learning, Classification Model, Air Quality Prediction

I. INTRODUCTION

Supervised Machine learning is used to forecast the future using historical evidence [2,3]. Machine learning (ML) is a form of artificial intelligence (AI) that allows computers to learn without being specifically programmed. The fundamentals of Machine Learning, implementation of an easy machine learning algorithm using python, focuses on the event of Computer Programs that will alter when exposed to new data. It feeds the training data to an algorithm, which then uses the training data to make predictions on test datasets that has been replaced. Machine learning is usually divided into three groups. The supervised learning programme receives the input file, and the accompanying labelling to determine data must be done by a human being beforehand. There are no labels for unsupervised instruction. It was given to the algorithm for training. The input file's clustering must be determined by this algorithm. Finally, Reinforcement learning communicates with its environment in a complex manner and receives a positive or negative feedback to improve its efficiency. To find trends in Python that lead to actionable insights, data scientists employs the variety of machine learning algorithms. These algorithms are often divided into two categories based on how they "read" about data to shape predictions: supervised and unsupervised learning. The process of predicting the category of given data points is known as classification. Classification is a supervised learning technique in machine learning and statistics, in which the computer programme learns from the data input and then applies the learning to identify new observations.

II RELATED WORKS

1.Bi-directional LSTM Network.

As per the literature survey, the[6] recurrent neural networks (RNN) have proven to be very effective in the processing and analyzing of temporal data. Future input datasets that arrive after the current time case, on the other hand, is also useful for prediction. Although delaying the performance by a few frames has been shown to boost sequential data results, the optimal delay is task based and must be determined through trial and error. Alternatively, two different networks could be trained on all input data, one for each path, and the results combined for the final prediction will be done based on the arithmetic computation or geometric averaging. However, since multiple networks trained on the same data can no longer be considered separate, optimal merging is difficult to achieve. To address these limitations, it proposed a [7,8] bidirectional recurrent neural network (BRNN) that can be trained using all available input information in the past and future of a given time frame and this network can overcome the limitations of previous approach. Missing data and irregular values are typical in pollutant data, as they are in all sensor data. Instrumental error or other external factors such as power outages or loss of communication, for example, may cause irregularities. A source monitoring station may not have registered pollutant data in some cases. A rolling average of the available data values over the previous three time periods was used to measure these missing values. An abnormal value is one that falls outside of a parameter's permissible range. A rolling average of the previous three instances is often used to replace irregular values. It demonstrated an efficient method for predicting the magnitude of contaminants using Deep learning models and various sensor data six, twelve, and twenty-four hours ahead of time. By forecasting the magnitude of PM2.5 emissions, We used pollution data from New Delhi, India, to back up our claim. We present our findings in relation to a baseline scheme at various stations and for various time periods. Furthermore, the authors Ishan Verma et.al [1] presented an Ensemble method to solve this research problem that performs better in the majority of cases and is also more stable.

2. IoT based Approach

It goes without saying that those who develop a production line or plant would be indeed more at hazard of breathing in poisonous chemicals and gasses as a result of their delayed presentation to contamination. Contamination compounds the undesirable circumstance, which incorporates a negative effect on living creatures. It is one of the foremost squeezing issues for the complete planet. Contamination could be a worldwide issue that influences outside organizations, states, and, as a result, the media. Any utilize of natural resources that surpasses nature's capacity to reestablish itself can lead to contamination of plants, discuss, and water. Aside from human exercises, there are a couple of irregular characteristic interims that regularly result within the discharge of dangerous substances. Aside from human-caused occasions, characteristic calamities such as ejections can result in discuss contamination. Each day, it is best in case each activity is completed utilizing cutting edge innovation in arrange to meet the requests of people, organizations, and businesses. The Internet of Things (IoT) is one of the foremost important communication breakthroughs within the final decade. It is conceivable to connect an expansive number of low-powered keen implanted objects to each other and to the web utilizing this concept. The IoT plan is built on the omnipresent presence of different remote advances such as recurrence recognizable proof (RFID) labels, sensors, actuators, and cell phones. Undoubtedly, as innovation advances and request rise, IoT administrations will ended up more broadly accessible, possibly changing our businesses, social orders, and individual lives. Since of the significant impacts of emissions on public wellbeing, the characteristic environment, and thus the complete worldwide economy, contamination in urban zones is getting to be a major issue in created cities each day. With the assistance of IoT gadgets, the proposed work on a contamination observing and expectation framework permits us to track discuss quality. To identify and transfer information to the microcontroller, discuss sensors are utilized. The microcontroller then saves the details to an online server. The LSTM [1,4] is used to forecast the outcome. It's a fast convergence that cuts down on training cycles while maintaining accuracy.

3. Forecasting Models

It is widely assumed that urban pollution has a direct effect on human health, especially in developing and developed countries where air quality controls are not available or are only partially implemented or enforced. Getting and putting away information, preprocessing and interpreting information into usable substance, estimating contaminants based on verifiable information, and at long last showing the information through versatile apps, Web entrances, and brief message administrations are all duties of the modules. This paper centers on the observing framework and its determining module. Univariate and multivariate displaying are the two shapes of demonstrating that are followed. The research comes about appear that utilizing diverse highlights in multivariate displaying with the M5P calculation produces the most effortless determining comes about. This research work [13] comes about are regularly amazingly valuable for disturbing applications in sullied zones. Air quality is a serious concern that has a direct impact on human health. The framework learns from the collected data to construct forecasting models using machine learning-based algorithms.

III PROPOSED SYSTEM

In this section, various steps involved to develop the framework to assess the quality of the air is discussed.

1.Air Quality Prediction Exploratory Data Analysis.

Different datasets from different sources would be combined to make a generalized dataset, after which different machine learning calculations would be utilized to extract designs and get the foremost precise comes about.

2.Data Wrangling.

In this phase, the data will be loaded, checked for cleanliness, and then the dataset will be trimmed and cleaned for review. Verify that the document's steps are followed carefully and that cleaning decisions are justified.

3.Building the classification model

The decision tree algorithm prediction model is useful in predicting air quality problems for the following reasons: It provides better leads to classification problems. It excels at removing outliers, irrelevant variables, and a combination of continuous, categorical, and discrete variables during preprocessing.

It generates out-of-bag estimate error, which has been shown to be unbiased in numerous tests and is reasonably simple to tune.

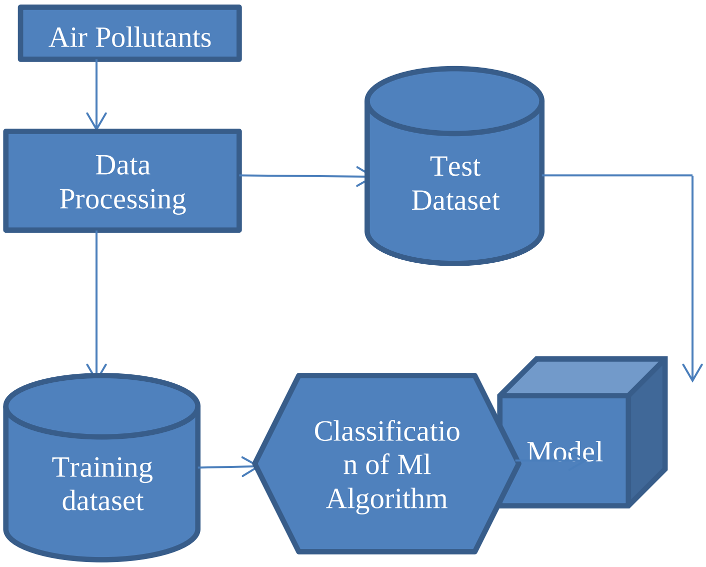


Fig 1. The proposed model's architecture

IV MODULE DESCRIPTION

1.Variable Identification Process

Machine learning validation approaches are used to calculate the error rate of the model built, which can be thought of as the dataset's on-the-edge-of-truth error rate. However, in real-world situations, it is essential to figure with information samples that are not truly representative of the population of a dataset. Duplicate the meaning and definition of the information form, whether it's a float variable or an integer, to find the missing value.

A variety of data cleaning tasks using Python's Pandas library, with a focus on the most common data cleaning job, missing values, and the ability to clean data more quickly. It tends to spend less time cleaning data and more time analyzing and modelling it.

The type of missing data will affect how missing values are filled in, how missing values are detected, and how simple imputation and precise statistical approaches are used to deal with missing data.

2.Exploration data analysis of visualization

In associated estimations and machine learning, data visualization can be a fundamental capacity. Data visualizations are habitually utilized to communicate and clarify key associations in plots and charts that are more visceral and locks in for accomplices than markers of connection or noteworthiness when combined with a little space data. Data visualization and exploratory data examination are regions in and of themselves, so some of the books indicated at the finest are worth examining in significance. Information may moreover be troublesome to get it sometime recently it is displayed in a visual format, such as charts and charts. It'll cover the different sorts of plots experience whereas visualizing information in Python, as well as how to utilize them to superior get it the possess information.

3.Using the best accuracy outcome to compare algorithms with predictions.

It'll be discovered how to make a test harness to fit multiple different machine learning algorithms in Python with scikit-learn. It's necessary to match the output of multiple different machine learning algorithms consistently. Using resampling methods like cross validation, we can get an estimate of how accurate each model could be on unseen results. Before settling on one or two to finalize, we can monitor the expected accuracy of your machine learning algorithms in a variety of ways. We'll learn how to do this in Python using scikit-learn in the next section.

Six algorithms are used in this research work: Logistic Regression is a technique for predicting the outcome of Bayesian naïve, Support Vector Machines based on Random Forest K-Nearest Neighbors Decision. The K-fold cross validation technique is used to validate and algorithm, which uses the same random seed to guarantee that the same splits are added to the training outcomes. Additionally, the train and test sets should be separated. By comparing precision, it is possible to predict the outcome.

4.Result Prediction

The algorithm's output must be categorized as variable data. The performance estimation parameter utilized for calculation is precision, review, f1-score, specificity, affectability, and accuracy.

False Positive Rate(FPR) = FP / (FP + TN)

Precision is defined as the ratio of true positives to the number of true and false positives.

Precision = TP / (TP + FP)

The recall is determined by dividing the total number of true positives and false negatives by the number of true positives.

TP / (TP + FN) = Recall

The F1 score is the average of precision and recall.

Accuracy: The percentage of total predictions that are accurate, or how much the model correctly predicts defaulters and non-defaulters overall.

Specificity: It is defined as the number of true negatives divided by the total number of true negatives and false positives.

Sensitivity: It is defined as the number of true positives divided by the total number of true positives and false negatives.

Confusion matrix: A confusion matrix is a table that is used to explain how well a classification model performs on a test dataset for which the correct values are identified.

Decision tree model: The model is built using a decision algorithm since it provides the best results. As a result, this algorithm is used in subsequent steps to predict air quality.

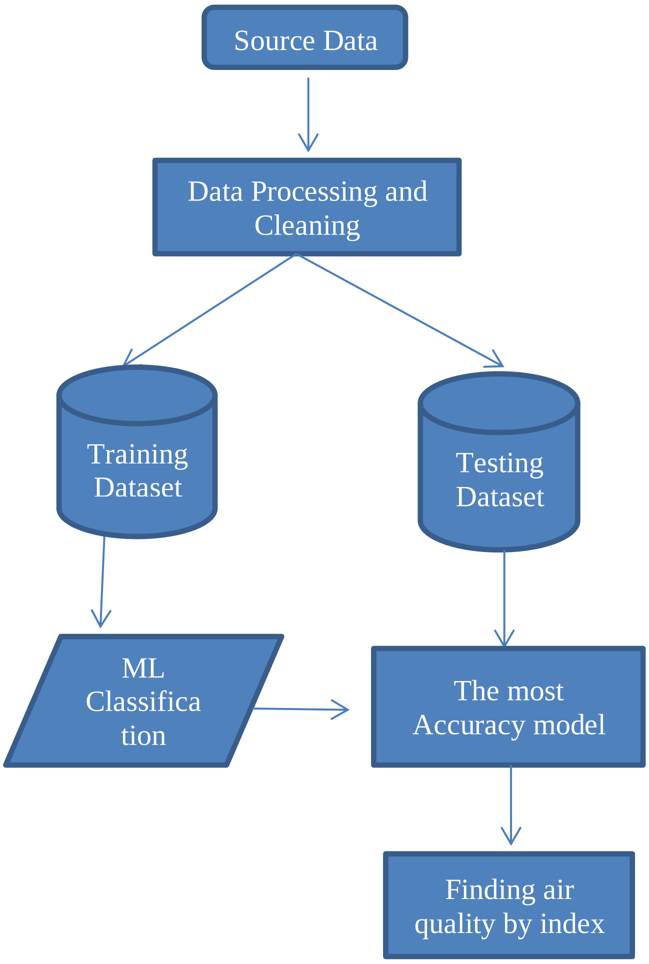


Fig 2. Air quality device work flow

V RESULTS

Using supervised machine learning algorithms, the presented system analyses the air pollution dataset in order to predict the pollutants' air quality with the greatest accuracy. The performance assessment analysis of ML algorithms is shown in Table I. The comparison of the confusion matrix parameter is shown in Table II. The following is a detailed description of the proposed system's procedure:

Step 1: The pre-processed air pollution dataset is obtained from various locations.

Step 2: Split the dataset into two sections after it has been pre-processed: training and testing.

Step 3: To train it, add the LR, SVM, DT, K-NN, NB, and RF ML algorithms to the training dataset.

Step 4: Compare the accuracy of algorithms to determine which is the most accurate, and then construct the model using the most accurate algorithm.

Step 5: Feed the test dataset into the prediction model to get the result. The accuracy of the ML algorithm is determined using TP, TN, FP, and FN. TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

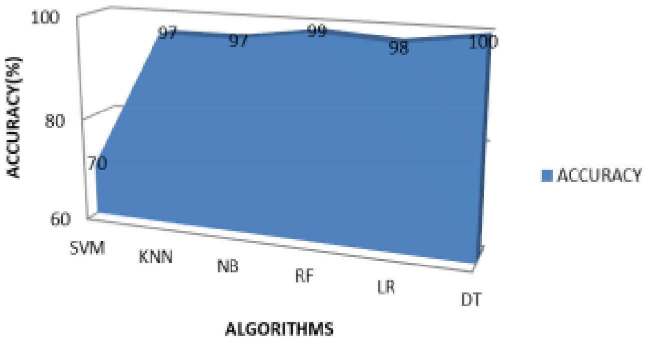
Table 1: The efficiency of machine learning algorithms is being measured.

PARAMETERS	**LR**	**NB**	**SVM**	**RF**	**KNN**	**DT**
PRECISION	0.97	0.94	0.0	0.99	0.96	1.0
RECALL	0.97	0.99	0.0	0.99	0.97	1.0
SENSITIVITY	0.97	0.98	0.0	0.98	0.97	1.0
SPECIFICITY	0.98	0.97	1.0	0.99	0.98	1.0
F1-SCORE	0.97	0.96	0.0	0.99	0.97	1.0
ACCURACY(%)	97.2%	97.4%	70.7%	99.1%	97.6%	99.8%

Table 2 Comparison of confusion matrix parameters

PARAMETER	**LR**	**NB**	**SVM**	**RF**	**KNN**	**DT**
TP	173	170	175	174	172	175
TN	71	72	0	72	71	73
FP	2	1	73	1	2	0
FN	2	5	0	1	3	0

ACCURACY OF THE ML ALGORITHM



. **Fig 3: Represents the accuracy of the ml algorithms**

VI CONCLUSION

Humans, trees, and livestock all suffer from the effects of air pollution. The proposed scheme is used in this paper to forecast the pollutants' air quality using supervised machine learning algorithms. A dataset of Indian air pollution is used to predict air pollution. The pre-processed dataset is divided into training and testing datasets in a 70:30 ratio, with 70% of the training dataset and 30% of the testing dataset. On the training dataset, use ML algorithms such as LR, SVM, NB, K-NN, RF, and DT to prepare the dataset for maximum accuracy. Precision, recall, f1-score, specificity, and sensitivity are all output measurement parameters that are measured for each algorithm. Confusion matrix parameters such as TP, TN, FP, and FN are calculated for each algorithm. LR is 98 percent accurate, NB is 95 percent accurate, RF is 99 percent accurate, SVM is 70% accurate, K-NN is 97 percent accurate, and DT is 100 percent accurate. Of the six ML algorithms, the Decision Tree algorithm is the most effective. The decision tree model is used to construct the prediction model, which predicts the pollutant present in the environment, as well as its causes and sources. This forecasting system aids asthmatic patients in avoiding polluted environments, as well as the metrological department in forecasting air quality. This air quality forecasting system can be modified for potential use in an Artificial Intelligence environment, and the mechanism can also be automated by displaying the prediction result in a site or desktop app.

FUTURE WORK

This work will be extended to apply for the Indian meteorological department by applying real time date air quality

By viewing the prediction result in a web or desktop application, this method can be automated.

To optimize the job in preparation for the introduction of Artificial Intelligence.

REFERENCES

.

[1] Ishan Verma Rahul Ahuja ,Hardik Meisheri ,Lipika Dey," Air Pollutant Severity Prediction Using Bi-Directional LSTM Network", 2018 IEEE/WIC/ACM International Conference on Web Intelligence, Year: 2018, Volume: 1, Pages: 651-654.

[2] Gaganjot Kaur Kang., Jerry Zeyu Gao., Sen Chiao., Shengqiang Lu., and Gang Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," in International Journal of Environment Science and Development, 2018.

[3] Guanghui Yue., Ke Gu., and Junfei Qiao, "Effective and Efficient Photo –Based PM2.5 Concentration Estimation," 2019.

[4] Ajitesh Kumar, Mona Kumari, Harsh Gupta," Design and Analysis of IoT based Air Quality Monitoring System", 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC).

[5] Ibrahim Yakut., Tugba Turkoglu., and Fijriye Yakut, "Understanding Customers Evaluations Through Mining Airline Reviews," in International Journal of Data Mining and Knowledge Management Process (IJDKP), 2015.

[6] Ke Gu., Junfei Qiao., and Weisi Lin, "Recurrent Air Quality Predictor Based on Meterology- and Pollution-Related Factors," 2017.

[7] Maple., Saif ul Islam., and Muhammad Nabeel Asghar,

"Comparative analysis of machine learning techniques for predicting air quality in smart cities," in Digital Object Identifier 8.1109/ACCESS.2017.Doi Number,2017.

[8] Xia Xi., Zhao Wei., Rui Xiaoguang., Wang Yijie., Bai Xinxin., Yin Wenjun., and don Jin, "A Comprehensive Evaluation of Air Pollution In 2015, he published "Improving Prediction with a Machine Learning Method."IEEE International Conference on Service Operations.

[9] L. Zhu and N. Laptev "Deep and confident estimation for time series at Uber,", 2017 IEEE International Conference on Data Mining Workshops (ICDMW). 103–110 in IEEE, 2017.

[10] Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, K. Katsouyanni et al.

[11] H. Hoos and K. Leyton-Brown "A contrast of linear regression, regularization,

[12] "An efficient approach for assessing hyperparameter importance," by, in International Conference on Machine Learning, 2014, pp. 754–762.

[13] Khaled Bashir Shaban; Abdullah Kadri; Eman Rezk," Urban Air Pollution Monitoring System With Forecasting Models", IEEE Sensors Journal (Volume: 16, Issue: 8, April15, 2016)

[14] A. Kendall and Y. Gal."What uncertainties do we need in Bayesian deep learning for computer vision?" by. in New Directions in N.

[15] P.Selvi Rajendran , Dr. T.P. Anithaashri, "CNN based Framework for identifying the Indian Currency Denomination for Physically Challenged People",IOP Conference Series: Materials Science and Engineering for the publication. Scopus

[16] P.Selvi Rajendran, Padmaveni Krishnan, D. John Aravindhar "Design and Implementation of Voice Assisted Smart Glasses for Visually Impaired People Using Google Vision API" 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA),2020

[17] V. Athira, P. Geetha, R. Vinayakumar, and K. Soman, "DeepAirNet: Applying recurrent networks for air quality prediction," Procedia Computer Science, vol. 132, pp. 1394–1403, 2018.