Prediction and Analysis of Pollutant using Supervised Machine Learning

Akiladevi R, Nandhini Devi B, Nivesh Karthick V, Nivetha P

Abstract**: *Air is the most essential natural resource for the survival of humans, animals, and plants on the planet. Air is polluted due to the burning of fuels, exhaust gases from factories and industries, and mining operations. Now, air pollution becomes the most dangerous pollution that humanity ever faced. This causes many health effects on humans like respiratory, lung, and skin diseases, which also causes effects on plants, and animals to survive. Hence, air quality prediction and evaluation as becoming an important research area. In this paper, a machine learning-based prediction model is constructed for air quality forecasting. This model will help us to find the major pollutant present in the location along with the causes and sources of that particular pollutant. Air Quality Index value for India is used to predict air quality. The data is collected from various places throughout India so that the collected data is preprocessed to recover from null values, missing values, and duplicate values. The dataset is trained and tested with various machine learning algorithms like Logistic Regression, Naïve Bayes Classification, Random Forest, Support Vector Machine, K Nearest Neighbor, and Decision Tree algorithm in order to find the performance measurement of the above-mentioned algorithms. From this, the prediction model is constructed using the Decision Tree algorithm to predict the air quality, because it provides the best and highest accuracy of 100%. The machine learning-based air quality prediction model helps India meteorological department in predicting the future of air quality, and its status and depends on that they can take action.***

Keywords: Prediction, Decision Tree algorithm, Air Quality Index, Air Pollution.

I. INTRODUCTION

In earlier days, the air is fresh and pure to breathe. But, the rapid increase of industries and the concentration of dangerous gases in the environment make the air more toxic to breathe. These gases cause many health effects on human beings like asthma, bronchitis, and other respiratory and lung diseases. Air pollution not only affecting humans but also it affects plants and animals [3]. So, an air quality forecasting system is essential to prevent this problem in these areas. In India, the vehicle contributes 35% of air pollution in major cities like Mumbai, Delhi, Bangalore, Chennai, and Kolkata. India is the fifth largest polluted country all around the world and it also estimated that every year 2 million people were died due to this pollution. In the year 2005 to 2006, about 8.9 million vehicles were sold and it reaches up to 15 million from 2010 to 2011. This shows that when the number of vehicles increased that leads to emission of hazardous gases which will finally affect the air quality.

In Delhi, November 2017 an event is known as Great Smog of Delhi causes air pollution beyond the acceptable level. The level of particulate matter PM 2.5 and PM 10 rises up to 999 micrograms per cubic meter, while the secure limits for those pollutants are 60 and 100 respectively.

There was much research for air quality forecasting, but many methods not given the most efficient model with the best accuracy. Most of the air quality forecasting models majorly use Machine Learning techniques [7]. There were many machine learning techniques to predict the model with the highest accuracy.

In this work, the proposed system is to develop the air quality forecasting system to predict the major pollutants present in those places. The prediction model is constructed using Machine Learning algorithms. This model will help people to prevent themselves from affecting air pollution.

II. LITERATURE SURVEY

The abundant air quality forecasting research works are carried out. Ishan Verma et al., [13] proposed a Bi-directional LSTM model to predict the severity level of air pollution. In this system, the prediction is improved using three Bi-directional LSTM that model short term, long term, and immediate effects of the severity level of PM 2.5.

An IoT based air pollution system is developed by Temesegan Walelign Ayele et al., [26] which is used to monitor the air, pollutant level in a particular location, and

the quality of the air is analyzed as well as predict the quality of air pollution. This system is developed using IoT along with ML algorithm more specifically Recurrent Neural Network-LSTM. In Delhi, PM 10 pollutant level is high in severity level. To predict and analyze the pollution level of pollutant PM 10 in Delhi, a system is developed by Aly Akhtar et al., [1] using Multi-Layer Perception, which is an Artificial Neural Network, Naïve Bayes, and Support Vector Machine. In this system, the accuracy of all the above-mentioned algorithms are compared to find the highest accuracy algorithm.

Revised Manuscript Received on June 22, 2020.

***** Correspondence Author

Akiladevi R*, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. E-mail: akiladevi.r@rajalakshmi.edu.in.

Nandhini Devi B, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. E-mail: nandhinidevi0305@gmail.com.

Nivesh Karthick V, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: niveshkarthick.v@gmail.com.

Nivetha P, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, Tamil Nadu, India. Email: nivethaparthasarathy99@gmail.com.

Among these algorithms, MLP achieved 98% accuracy so the model is constructed using the MLP algorithm for prediction of pollutant level in Delhi.

Shweta Taneja et al., [25] proposed a system for predicting trends in air pollution using data mining. This system uses two types of time series analysis they are linear regression and Multi-Layer Perception to understand different patterns in various types of pollutants. This system interprets the data and predicts the trends of air pollution. For urban PM 2.5 concentration, CNN and LSTM techniques were combined to construct the prediction model developed by Dongming Qin et al., [5]. This model uses two kinds of deep learning such as Convolution Neural Network as a base layer and Long Short Term Memory Neural Network as an output layer. The base layer is used to extract features of input data and the output layer is used to consider the time dependence of pollutants. Semantic ETL framework for air quality prediction and analysis is developed on the cloud platform by Yue Shan Chang et al., [30]. This system utilizes ontology to analyze the relationship of PM 2.5 from different data sources and merge the data together from the dataset. Then these dataset is analyzed and show the visualized result to make a prediction. Saba Ameer et al., [22] proposed a comparative analysis for the prediction of air quality forecasting system using Machine Learning techniques. An air quality forecasting system using an ensemble learning approach developed by Chao Zhang et al., [3]. Multi-channel Ensemble Learning via Supervised Assignment algorithm is used to predict and analyze the quality of air pollution.

III. MATERIALS AND METHODS

The dataset used for analysis is air pollution data in India. The data is collected between the periods of 1990 to 2018. The dataset is collected from the metrological department and it is in CSV file format. This dataset contains nine attributes like country, state, city, place, last update, average, maximum, minimum, and pollutants. The unit used to measure air pollution is micrograms per cubic meter. The ML method is useful to predict the present and future by analyzing the historical information. This historical data contains the air pollution data in earlier years. The ML algorithms like LR, NB, SVM, RF, KNN, and DT are used to measure the performance of the dataset.

A. Logistic Regression:

Logistic regression produces the result in a binary format that is used to predict the outcome of a categorical dependent variable. The outcome of the logistic regression should be discrete/categorical such as "one or zero", "yes or no", and "high or low". This algorithm uses sigmoid function which converts any value into a discrete value.

B. Naïve Bayes:

Naïve Bayes algorithm is a statistical classification technique based on Bayes theorem. Baye's theorem depends on the naïve assumption that the input variables are independent of each other. It is a simple and powerful algorithm for predictive analysis that uses probabilities of each attribute belonging to each class to make the prediction. Naïve Bayes classifier is easy to build and it is useful for a very large dataset.

C. K-Nearest Neighbor:

K-Nearest Neighbor is a simple algorithm that stores all the available data correspond to training data points in n-dimensional space and classifies the new data based on a similarity measure. Once an unknown discrete data is received, it analyzes the closest k number of instances saved and returns the most common class as the prediction and for real-valued data; it returns the mean of k nearest neighbors.

D. Support Vector Machine:

Support Vector Machine is a classification algorithm that is formally designed by a separate hyperplane. The aim of this algorithm is to segregate the given data points in the best possible way. SVM algorithm is suitable for high dimensional space. It uses s subset of training points in the decision function that makes it memory efficient.

E. Random Forest:

Random Forest is an ensemble machine learning algorithm that is used for both classification and regression. However, it is mostly used for classification tasks. The random forest algorithm is constructed by combining multiple decision trees, resulting in a forest of trees. Random Forest algorithm creates a decision tree on data items and then makes a prediction for each of them and finally selects the best solution. Hence, Random Forest is an ensemble method it reduces over-fitting by averaging the result.

F. Decision Tree:

The decision tree algorithm is a popular and simplest machine learning algorithm. It builds classification or regression models in the kind of a tree structure. The aim of the decision tree algorithm is to create a model that can be used to predict the class of the end result variable by learning the decision rules inferred from prior training data. This algorithm uses if-then rule which is mutually exclusive and exhaustive for classification technique.

IV. WORKING MODULES

The dataset is collected from different places that need to be converted into a generalized format, to recover from missing and null values. Then on this generalized dataset ML algorithms were applied in order to extract patterns and to find the highest accuracy. Figure 1, represents the complete workflow of the Air Quality System.

A. Air pollution dataset:

The dataset consists of 824 tuples and 9 attributes. The 9 attributes are country, stste, city, place, last update, minimum, maximum, average, and pollutants. These attributes are os string and numeric. The dataset is collected from https://www.kaggle.com/venky73/airquality.

B. Pre-processing dataset:

Data pre-processing is used to convert the raw format of data into an understandable format because the data in the real world is incomplete, noisy, and inconsistent data.

The generalized dataset undergoes pre-processing which helps to recover from missing values, null values, duplicate values, and convert the data into the numeric format. The AQI and class attribute is added based on the metrological data.

C. Splitting dataset:

Dataset is split into training and testing datasets. Generally, by default the dataset is split in the ratio of 80:20 but this system dataset is divided into the ratio of 70:30 that is 70% training dataset and 30% of the testing dataset.

D.Classification algorithm:

The dataset is trained by applying ML algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, K Nearest Neighbor, and Decision tree. The performance measurement parameter used for calculation is precision, recall, f1-score, specificity, sensitivity, and accuracy.

1) **Precision*:***
   Precision is defined as the ratio of true positive divided by the sum of a truly positive and false positive.
2) **Recall:**
   The recall is defined as the ratio of true positive divided by the sum of a truly positive and false negative.
3) **F1-score:**
   F1 score is defined as the mean between precision and recall.
4) **Specificity:**
   Specificity is defined as the ratio of true negative divided by the sum of a true negative and false positive.
5) **Sensitivity:**
   Sensitivity is defined as the ratio of true positive divided by the sum of a truly positive and false negative.
6) **Confusion matrix:**
   A confusion matrix is represented in the form of a table that is used to describe the performance of the classification model on a test dataset for which the correct values are known.

E. Decision tree model:

A decision algorithm is used to construct the model because it produces the highest accuracy. So this algorithm is used for further steps to predict the air quality.

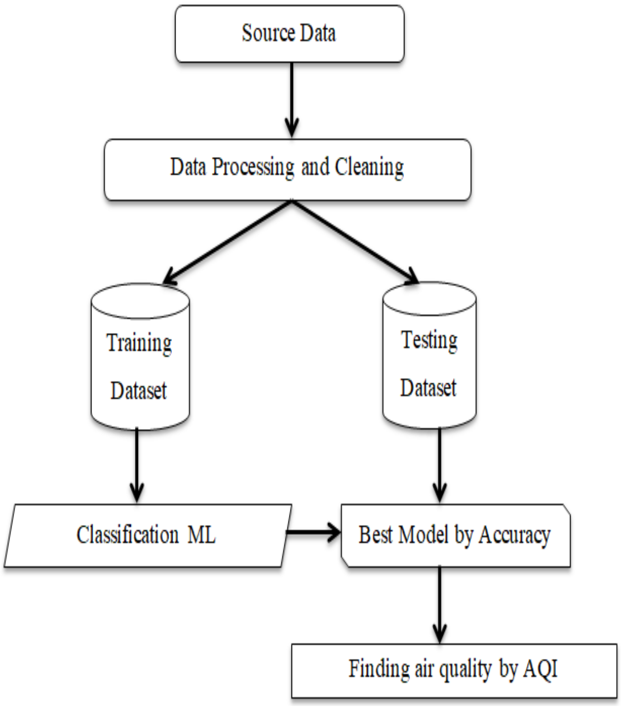

Fig 1: Workflow of Air Quality System

F. Prediction result:

CLASS 1: if the AQI value of the pollutant is <100 then the cause of the pollutant is minimal.

CLASS 2: if the AQI value of the pollutant is >100 then the cause of pollutant is severe.

V.EXPERIMENTS AND RESULTS

The proposed system is to analyze the air pollution dataset to predict the air quality of the pollutants with the highest accuracy using supervised machine learning algorithms. Table I shows the performance measurement analysis of ML algorithms. Table II shows the comparison of the confusion matrix parameter. The procedure of the proposed system is completely described below:

Step1: The air pollution dataset is collected from different places that are pre-processed.

Step 2: After the dataset is pre-processed divide the dataset into training and testing dataset.

Step 3: Train the dataset by applying LR, SVM, DT, K-NN, NB, and RF ML algorithms on the training dataset.

Step 4: Compare the accuracy of the algorithms to get the highest accuracy and then the model is constructed using the highest accuracy algorithm.

Step 5: Input test dataset to the prediction model in order to obtain the result.

Accuracy of the ML algorithm is calculated using TP, TN, FP, and FN. TP refers True Positive, TN refers True Negative, FP refers False Positive, and FN refers False Negative. Figure 2 represents the analysis of the classification algorithm to predict air quality.

Retrieval Number: A2837059120/2020©BEIESP
DOI:10.35940/ijrte.A2837.079220

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 1: Performance measurement of ML algorithms

PARAMETERS	LR	NB	SVM	RF	KNN	DT
PRECISION	0.97	0.93	0	0.98	0.95	1
RECALL	0.97	0.98	0	0.98	0.97	1
SENSITIVITY	0.97	0.98	0	0.98	0.97	1
SPECIFICITY	0.98	0.97	1	0.99	0.98	1
F1-SCORE	0.97	0.95	0	0.98	0.96	1
ACCURACY (%)	98%	97%	70%	99%	97%	100 %

Table 2 Comparison of confusion matrix parameters

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

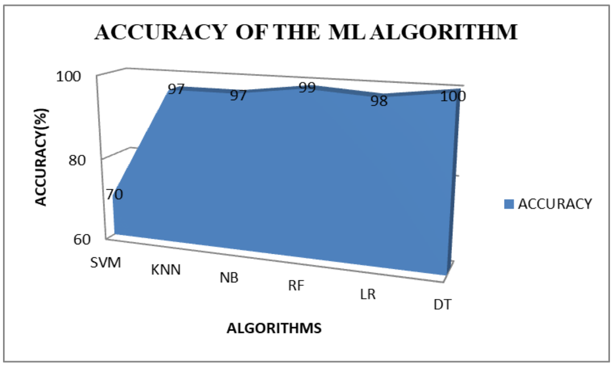

Fig 2: Represents the accuarcy of the ML algorithms

VI. CONCLUSION

Air pollution causes many health effects on human beings as well as affects plants and animals. In this paper, the proposed system is developed to predict the air quality of the pollutants using supervised machine learning algorithms. Indian air pollution dataset is used to predict air pollution. The collected dataset is pre-processed to recover from missing, null, and duplicate values. The pre-processed dataset is divided into training and testing datasets in the ratio of 70:30 that is 70% of training and 30% testing dataset. Apply ML algorithms such as LR, SVM, NB, K-NN, RF, and DT on the training dataset to train the dataset in order to obtain the highest accuracy. The performance measurement parameters like precision, recall, f1-score, specificity, and sensitivity are calculated for each algorithm. Confusion matrix parameters like TP, TN, FP, and FN are calculated for each algorithm. The accuracy achieved by LR is 98%, NB is 95%, RF is 99%, SVM is 70%, K-NN is 97%, and DT is 100%, from this six ML algorithm Decision Tree algorithm provide the highest accuracy. The prediction model is constructed using the Decision Tree algorithm to predict the pollutant present in the location, causes, and sources of the pollutant. This prediction system helps asthma affect a person to prevent themselves from the polluted area and also is developed to help the metrological department to predict air quality forecasting. In the future, this air quality forecasting system can be optimized to implement in the Artificial Intelligence environment and can also automate this system by showing the prediction result in either web or desktop application.

REFERENCES

1. Aly Akhtar., Sarfaraz Masood., Chaitanya Gupta., and Adil Masood, “Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron,” 2018.
2. Atakan Kurt and Ayse Betul Oktay, “Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks,” in *Expert Systems with Applications 37 (2010) 7986-7992*, 2010.
3. Chao Zhang., Junchi Yan., Yunting Li., Feng Sun., Jinghai Yan., Dawei Zhang., Xiaoguang Rui., and Rongfang Bie, “Early Air Pollution Forecasting as a Service: an Ensemble Learning Approach,” in *2017 IEEE 24th International Conference on Web Services*, 2017.
4. Dixian Zhu., Changjie Cai., Tianbao Yang., and Xun Zhou, “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization,” in *Big data and cognitive computing*, 2018.
5. Dongming qin., Jian Yu., Guojinian Zou., Ruihan Yong., Qin Zhao., and Bo Zhang, “A Novel Combine Prediction Scheme Based on CNN and LSTM for Urban PM2.5 Concentration,” in *Digital Object Identifier 10.1109/ACCESS.2019.DOI*, 2019.
6. Ebrahim Sahafizadeh and Esmail Ahmadi, “Prediction of Air Pollution of Boushehr City Using Data Mining,” in *2009 Second International Conference on Environmental and Computer Science*, 2009.
7. Elias Kalapanidas and Nikolaos Avouris, “Short-term air quality prediction using a case-based classifier,” in *Environmental Modelling and Software 16 (2001) 263-272*, 2000.
8. Emanuel Lacic., Dominik Kowald., and Elisabeth Lex, “High Enough? Explaining and Predicting Traveler Satisfaction Using Airline Reviews,” 2016.
9. Fang Mingjian., Zhu Guocheng., Zheng Xuxu., and Yin Zhongyi, “Study on air fine particles pollution prediction of main traffic route using artificial neural network,” in *2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, 2011.
10. 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.
11. Guanghui Yue., Ke Gu., and Junfei Qiao, “Effective and Efficient Photo –Based PM2.5 Concentration Estimation,” 2019.
12. 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.
13. Ishan Verma., Rahul Ahuja., Hardik Meisheri., and Lipika Dey, “Air Pollutant severity prediction using Bi-directional LSTM Network,” in *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018.
14. Kaixi Zhu., Mitchell P. Krawiee-Thayer., and Amir H. Assadi, “Stochastic Local-to-Global Methods for Air Quality Prediction,” 2017.
15. Ke Gu., Junfei Qiao., and Weisi Lin, “Recurrent Air Quality Predictor Based on Meterology- and Pollution-Related Factors,” 2017.
16. Khaled Bashir Shaban., Abdullah Kadri., and Eman Rezk, “Urban Air Pollution Monitoring System With Forecasting Models,” in *IEEE Sensors Journal*, 2016.
17. Ling Wang., Xi-yuvan Xiao., and Jian-yao Meng, “Prediction of Air Pollution Based on FCM-HMM Multi-model,” in *Proceedings of the 35th Chinese Control Conference*, 2016.
18. Luke Curtis., William Rea., Patricia Smith-Willis., Ervin Fenyves., and Yaqin Pan, “Adverse health effects of outdoor air pollutants,” in *Environment International 32 (2006) 815-830*, 2006.
19. MinHan Kim., YongSu Kim., SuWhan Sung., and ChangKyoo Yoo, “Data-Driven Prediction Model of Indoor Air Quality by the Preprocessed Recurrent Neural Networks,” in *ICROS-SICE International Joint Conference 2009*, 2009.
20. Nitin Sadashiv Desai and John Sahaya Rani Alex, “IoT based air pollution monitoring and predictor system on Beagle Bone Black,” 2017.
21. Praveen Kumar Sharma., Tanmay De., and Sujoy Saha, “IoT based Indoor Environment Data Modelling and Prediction,” 2018.

22. Saba Ameer., Munam Ali Shah., Abid Khan., Houbing Song., Carsten 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 10.1109/ACCESS.2017.Doi Number*, 2017.
23. Shajulin Benedict, “Revenue Oriented Air Quality Prediction MicroServices for Smart Cities,” 2017.
24. Shuting Li., Shujun Song., and Xin Fei, “Spatial Characteristics of Air Pollution in the Main City Area of Chengdu, China,” 2011.
25. Shweta Taneja., Dr. Nidhi Sharma., Kettun Oberoi., and Yash Navoria, “Prediction Trends in Air Pollution in Delhi using Data Mining,” 2016.
26. Temesegan Walelign Ayele and Rutvik Mehta, “Air pollution monitoring and prediction using IoT,” in *Proceedings of the 2nd Internaional Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant*, 2018.
27. Tsang-Chu Yu., Chung-Chin Lin., Ren-Gury Lee., Chao-Heng Tseng., and Shi-Ping Liu, “Wireless Sensing System for Prediction Indoor Air Quality,” 2012.
28. Xia Xi., Zhao Wei., Rui Xiaoguang., Wang Yijie., Bai Xinxin., Yin Wenjun., and don Jin, “A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method,” in *2015 IEEE International Conference on Service Operations and Logistics, And Informatics (SOLI)*, 2015.
29. Xinlong Tao., Jianqiang Yi., Zhiqiang Pu., and Tianyi Xiong, “State-Estimator-Integrated Robust Adaptive Tracking Control for Flexible Air-Breathing Hypersonic Vehicle With Noisy Measurements,” 2019.
30. Yue Shan Chang , Kuan-Ming Lin , Yi-Ting Tsai , Yu-Ren Zeng and Cheng-Xiang Hung, “Big data platform for air quality analysis and prediction” in the *27th Wireless and Optical Communications Conference (WOCC2018)* in 2018.
31. Zhiwen Hu., Zixuan Bai., and Kaigui Bian, “Real-Time Fine-Grained Air Quality Sensing Networks in Smart City: Design, Implementation and Optimization,” in *IEEE Internet of Things Journal*, 2019.
32. Ziyue Guan and Richard O. Sinnott, “Prediction of Air Pollution through Machine Learning Approaches on the Cloud,” in *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, 2018.

AUTHORS PROFILE

Nandhini Devi B, is currently pursuing her Bachelor of Engineering in Computer Science and Engineering at Rajalakshmi Engineering College, Anna University. She has completed 3 mini projects on HTML, CSS, and Javascript with Database connectivity. Her area of interest includes IoT, Networks, python, HTML, CSS, DBMS. She currently placed in TCS Ninja.

Nivesh Karthick V, is currently pursuing her Bachelor of Engineering in Computer Science and Engineering at Rajalakshmi Engineering College, Anna University. He has completed 2 mini projects on HTML, CSS, and Javascript with Database connectivity. His area of interest includes IoT, Networks, and DBMS.

Nivetha P, is currently pursuing her Bachelor of Engineering in Computer Science and Engineering at Rajalakshmi Engineering College, Anna University. She has completed 3 mini projects on HTML, CSS, and Javascript with Database connectivity. Her area of interest includes IoT, python, HTML, CSS, DBMS. She currently placed in TCS Ninja.