International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 7 ǁ July 2022 ǁ PP. 277-282

Analysis of the Dataset for Air Pollution by Air Quality Prediction Based On Supervised Machine Learning Approach

Arathi H L, Dr Sarojadevi H
Department of Computer Science and Engineering, Nitte Meenakshi Institute of technology

Abstract *--Typically, air pollution is defined as a type of contaminant that is emitted into the atmosphere and mostly impacts human health. Furthermore, it poses a serious threat to human health and occasionallyresults in fatalities. Not only are humans harmed by air pollution; animals, numerous forests, and several crops are all negatively impacted. Therefore, a project was created to analyze the air quality using variousmachine learning approaches in order to solve this issue. Consequently, one of the crucial issues to look into is air pollution. This investigation focused mostly on a data set based on several parameters, includingidentification and diversity of analysis by using the missing values to capture multiple photographs.*

Keywords: *Air pollution, Air Quality Index, Machine learning algorithms, PM10, PM2.5, SO2, CO, NO2,O3.*

Date of Submission: 01-07-2022 Date of acceptance: 11-07-2022

I. INTRODUCTION

Not just humans are harmed by air pollution; many animals, forests, and crops are as well. Therefore, a project was created to analyze the air quality using various machine learning approaches in order to solve this issue. Consequently, one of the crucial issues to look into is air pollution. Approximating a mapping function into input x and output y variables is the task of classification predictive modelling. Therefore, this can be used in machine learning for appropriate data analysis, air quality prediction, and air quality determination using various machine learning approaches. Machine learning is one of the types of artificial intelligence (AI), which primarily gives computers the capacity to Machine learning is a discipline that consistently emphasizes statistical data and computer judgement. One of the keyareas where machine learning, which always acts on the drive to learn, will be used is in data collecting.Predictive analysis of the data set is how this machine learning is always referred to. Machine learning is currently one of the most talked-about issues in the field of artificial intelligence and has been for some time. Starting a career in this field may be less challenging than it first appears, and it may be an appealing opportunity. It is not an issue even if you have no prior math or programming skills.

II. OBJECTIVES

The primary objective of this project was to detect air pollution using data analysis. To that end, a new machine learning algorithm was created in real time by tracking the air quality assessment, which has since replaced the appropriate result prediction provided by the analysis of supervised algorithms.

III. PROBLEM STATEMENT

The mainly air quality is detected by using the various forms of machine learning techniques. Also, the harmful air content gases lead to detection of the quality of the air gases in this world. The increase in the air pollution also increases the worst of the human health which leads to the death of the human beings. Efficient implementation of the air by the assurance of the quality models which mainly collects all the information and takes the assessment on that area.

IV. PROPOSED SYSTEM

To detect the air pollution if we apply the photo graphic method then the parameter evaluation is very difficult. To overcome that a machine learning algorithm is implemented to overcome this using GUI which combines all the dataset and different algorithms to apply the exact parameters.

Main advantages of the system are listed below;

1. Air quality investigation applicable effectively in machine learning
2. Observations, issues were analyzed by the different methods

3. Accurate analysis of the data-sets is done.
4. Air pollution detection is predicted and analyzed.

V. DATASET DETAILS

Variable	Description
Country	Home country (India)
State	Indian States name lists
city	City names for each state
place	Place names for each city
last update	Date and time (DD/MM/YYYY HH:MM)
Avg	Average range of pollutants
Max	Maximum range of pollutants
Min	Minimum range of pollutants
Pollutants	Pollutants name

Fig 1: Table shows details of dataset

A. Data Manipulation

In this phase, we will import the data, confirm its integrity, and then trim and clean the provided data-set in preparationfor analysis.

B. Data Gathering

In order to match the given data, the collected data set is divided into a Training and Test set. Usually, a 7:3 ratio is used to divide the training and test sets. The data model, which was created using algorithms, is applied to the training set, and the outcomes are forecasted for the test set based on their correctness.

C. Preprocessing

The collected data can have missing values, which would make it inconsistent. To improve performance and producebetter outcomes, data must be preprocessed.

D. Development of a Predictive Model

The gathering of data is essential for machine learning. For collecting data, there is a sufficient amount of raw data and information. Without pre-processing, raw information cannot be used directly. What model and algorithm are used to preprocess, then. This model has undergone testing and training, and it consistentlyproduces correct predictions.

VI. IMPLEMENTATION

Six modules make up the majority of this project: analysis, visualization, machine learning models for predicting pollution utilizing algorithms, and GUI-based air pollution check. Python is used to implement this project in Anaconda with Jupyter Lab.

A. Visualization of data

In applied statistics and machine learning, data visualization is a crucial ability. It is true that the focus of statistics is on the quantitative description and estimation of data. An essential set of tools for obtaining a qualitative understanding is provided by data visualization. This can be useful for discovering trends in a data set as well as for exploring and getting to know the data set.

B. Analysis of data

The material includes meteorological data on different pollutants for different Indian cities. The strategy is to employ many machine learning techniques to create the optimal machine learning model. To build our model, we used six distinct supervised machine learning methods.

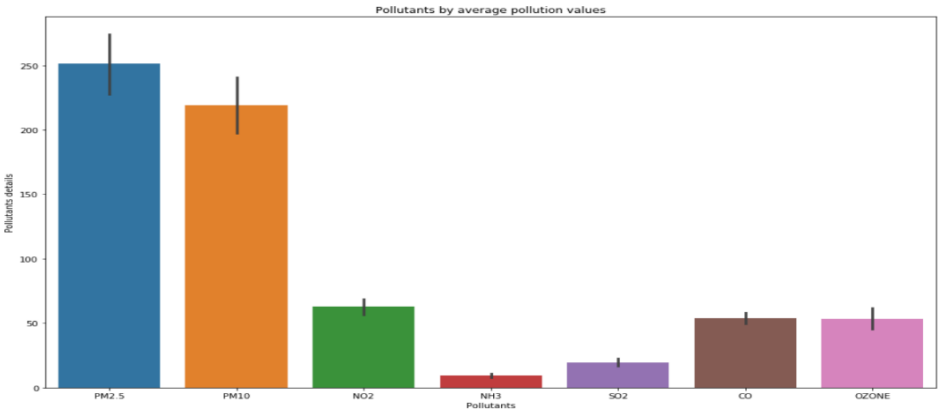

Fig 2. By average pollution values pollutants

The diagram demonstrates how pollution spread throughout our dataset

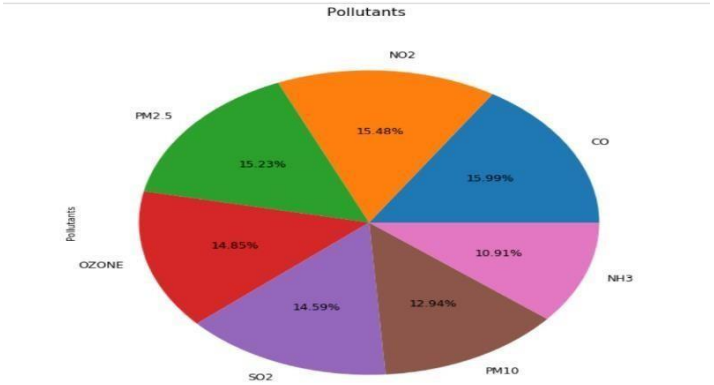

Fig 3. The spread of contaminants

The accompanying illustration depicts how contaminates are dispersed in relation to pollution ratings

C. Algorithms

• **Naive Bayes classifier**: This supervised approach uses the Naive Bayes classifier. It is a basic Bayesian classification technique. It presupposes those characteristics have strong (Naive) interdependence. Each quality on its own contributes to the maximizing the likelihood of it. It does not employ Bayesian methods and is suited for use with the Naive Bayes model. approaches. Naive Bayes classifiers are used to solve many challenging real-world problems. On our data set, this yielded a 97.58 percent accuracy rate.

• **K-nearest neighbour:** An approach to supervised classification algorithms is the K-nearest neighbour algorithm. Things are categorized based on their closest neighbour. It falls under the category of instance- based learning. The Euclidean distance is used to calculate a character's distance from its neighbours. It employs a set of points with titles and applies them to the labelling of another point. Various prediction techniques are applied to the data set after the missing values have been filled in. The accuracy of this algorithm on our data set is 97.58%.

• **Support vector Machine:** This classifier employs a fluttery plane to separate data points with the highest margin. Support vectors are the data points that are most closely related to the fluttery plane. A variety of kernels are regularly chosen for the hyper plane. In. This type of classifier uses less memory because they only use a subset of the showing points during the decision stage. SVM achieved an accuracy rate of 70.56 percent using our data set.

• **Random Forest:** The ensemble learning category includes this algorithm. Decision Trees are the basic building block of this method. Here, we divide the dataset into several smaller datasets, create a decision tree for each subset of the dataset, take into account the results of all the DTs, and make a forecast bymajority voting. Comparing the ensemble approach to a single DT always yields favorable results. Using the bagging method, manytrees are produced in this. On our data set, Random Forest attained an accuracyrate of 99.19%.

- **Decision Tree:** In which the training data is continually segmented based on a particular parameter, with you describing the input and the associated output. On our data set, Decision tree attained an accuracy rate of 100%.

- **Logistic Regression**: It is a method of data analysis where the outcome is influenced by one or more independent factors. The outcome is measured by a dichotomous variable (in which there are only two possible outcomes). Logistic regression achieved a 98.38 percent accuracy rate on our data set.

VII. RESULT AND DISCUSSION

On the data set, we have examined and assessed the classification reports from each of the six methods. The accuracy of each method has been compared to one another, as shown in the table below

Algorithm	Accuracy
Logistic Regression	98.38%
Decision Tree	100.0%
Random Forest	99.19%
Support vector machine	70.56%
Naïve Bayes	97.58%
K-Nearest Neighbor	97.98%

Fig 4. Accuracy result of each algorithm

A. GUI BASED AIR POLLUTION CHECK

Fig 5. Air pollution check window

The figure shows Login page for pollution check on basis of your choice

Fig 6. Login page

The figure shows Login page for pollution check on basis of your choice.

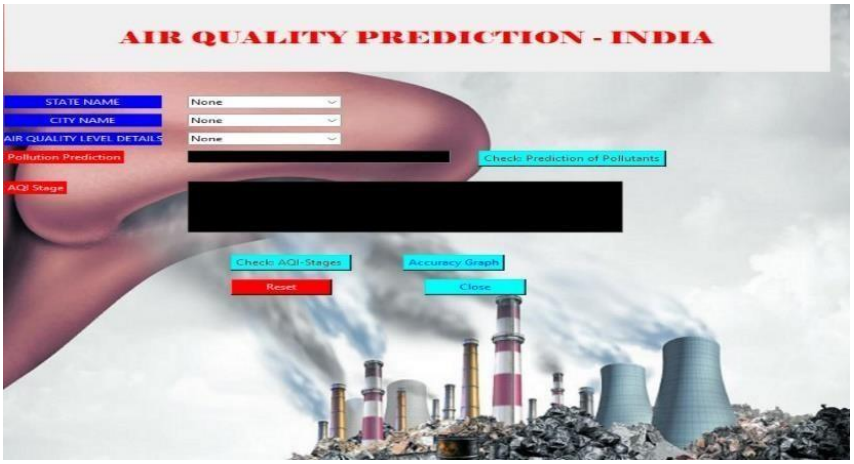

Fig 7. GUI Based pollution check

The figure shows GUI based pollution check. In this we have to enter the state name, city name, Air quality level details.

Fig 8. Pollution check

The figure shows once all the details entered then check for prediction of pollutants and check AQI stages then it shows cause and pollution prediction.

VII. CONCLUSION AND FUTURE WORK

For estimating air quality based on provided attributes, the decision tree algorithm has the best accuracy. With the aid of this application, the India Metrological Department may forecast the condition of the air and its future, and based on that, take appropriate action. India Metrological intends to automate the process of determining whether the air quality is good or terrible in real time as part of my project's future development. By displaying the prediction results in a desktop or web application, this procedure is automated to efficiently carry out the work in the environment of artificial intelligence.

REFERENCES

[1]. Ishan Verma, Rahul Ahuja, Hardik Meisheri, Lipika Dey, "Air pollutant severity prediction usingBi-directional LSTM Network " ,IEEE M International Conference on Web Intelligence,2018.
[2]. Sarita Jiyal, Rakesh Kumar Saini , "Prediction and Monitoring of Air Pollution Using Internet ofThings (IoT)", Sixth International conference on parallel, Distributed and Grid computing,2018.
[3]. Khaled Bashir Shaban, Abdullah Kadri and Eman Rezk, "Urban Air Pollution Monitoring SystemwithForecasting Models" ,IEEE,2019.
[4]. S. Jeya, Dr. L. Sankari, " Air pollution prediction by deep learning model",IEEE,2020.
[5]. Yu Jiao, Zhifeng Wang, Yang Zhang, Prediction of air quality index based on LSTM, IEEE,2019.
[6]. Shreyas Simu, Varsha Turkar, Rohit Martires, Vranda Asolkar, Swizel Monteiro, Vaylon Fernandes, andVassant Salgaoncar, Air pollution prediction using machine learning,IEEE,2021.
[7]. Wenjing Wang, Shengquan Yang, Research on air quality forecasting based on big data and neuralnetwork,IEEE,2020
[8]. Luke Curtis, William Rea, Patricia Smith-Willis, "Adverse health effects of outdoor air pollutants",IEEE,2019.
[9]. Sriram Krishna Yarragunta, Mohammed Abdul Nabi, "Prediction of Air Pollutants Using Supervised Machine Learning", IEEE,2021.

[10]. Vijay Siva Raman, Hari Bhrugubanda , SVR Based Dense Air pollution Estimation Model Using Static and Wireless Sensor Network, IEEE,2019.
[11]. Vijay Siva Raman, Air pollution detection by SVM algorithm,IEEE,2019.
[12]. Shaheduzzaman Chowdhury,MD. Shahedul Islam,MD.Kaiser Raihan and Mohammed Shahriar Arefin,"Design and Implementation of an IoT Based Air Pollution Detection and Monitoring System", IEEE,2019.
[13]. Jerry Gao, Chuanqi Tao, "Big Data validation and Quality Assurance – Issues, Challenges and Needs ", IEEE,2019.
[14]. Vijay Siva Raman, Hari Bhrugubanda , SVR Based Dense Air pollution Estimation Model Using Static and Wireless Sensor Network, IEEE,2019.
[15]. Xia Xi, Zhao Wei and Rui Xiaoguan , " A comprehensive evaluation of air pollution prediction improvementby a machine learning method", IEEE International Conference on Service Operations And Logistics, And Informatics,2019.
[16]. Hari Bhrugubanda, Vijay Sivaraman, "HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation from Fixed and Mobile Sensors" ,IEEE,2020.
[17]. Tapiwa M. Chiwewe and Jeofrey Ditsela, "Machine Learning Based Estimation of Ozone Using Spatio- Temporal Data from Air Quality Monitoring Stations", IEEE,2020.
[18]. Limei Ma, Yijun Gao, "Research on Machine Learning Prediction of Air Quality Index Based on SPSS",IEEE,2020.
[19]. Yurii Maslyiak, Andriy Pukas, Iryna Voytyuk, Mykola Shynkaryk, "Environmental Monitoring System for Control of Air Pollution by Motor Vehicles",IEEE,2018.
[20]. Paul D.Rosero-Montalvo,Jorge ,Air quality monitoring intelling system using machine learningtechniques,2018,IEEE.