iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** VI **Month of publication:** June 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.43993

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue VI June 2022- Available at www.ijraset.com*

# Prediction of Air Quality Index Using Supervised Machine Learning

Rajat R. Relkar[1], Vaibhav Deulkar[2], Ridam Gunjewar[3], Rahul Panghate[4], Pratik Gaurkar[5], Mukul Singanjude[6], Ritik Jarile[7], Prof. Preetee K. Karmore[8]
[1, 2, 3, 4, 5, 6, 7]*Department of Computer Science and Engineering*
[8]*Guide*
*Shri Vidyarthi Sudhar Sangh's*
*Dr. Babasaheb Ambedkar College of Engineering and Research, Nagpur*

***Abstract:*** ***The proposed system depicts various strategies utilized for forecast of Air Quality Index (AQI) utilizing supervised machine learning procedures. The system examines machine learning algorithm for air quality index by computing algorithm accuracy which will bring about the best precision. Moreover, the system exhibits different machine learning ac- curacy figures from the given dataset with assessment order report which recognizes the perplexity lattice. The outcome shows the adequacy of machine learning suggested calculation method that can be contrasted and best exactness with accuracy, Recall and F1 Score. The air pollution database contains data for each state of India. Four supervised machine learning algorithms, decision tree, random forest tree, Naïve Bayes theorem and K-nearest neighbor are compared and evaluated.***

## ORGANIZATION OF THE THESIS

In this thesis, Introduction and architecture of our system, the objective and the problem statement will be discussed in Chapter 1. The Review of Literature will be discussed in Chapter 2. Chapter 3 will consist of Work Done on various modules of the project. The System Design is explained in the Chapter 4. The Results obtained by usingvarious algorithms are mentioned in the Chapter 5. The Chapter 6 will consist of Conclusion. At last, we have mentioned the Appendix and the References.

## I. INTRODUCTION

Technological innovation in recent years has been a remarkable technological advancement. Instead of just writing instructions as a general rule, the philosophy of artificial intelligence, in which the system makes its own decisions, gradually affects all aspects of society. From the very first phase of the launch to the major platform vendors, machine learning in its segment has become a focus area for all companies.

Machine learning is a place where the artificial intelligence system collects data from the sensors and learns behavior in the environment. The ability to practice machine learning (ml) algorithms has been one of the reasons why machine learning is used to predict air quality indicators. The learning algorithms for the four machines used in the following system, decision tree, random forests, naïve bayes ,k nearest neighbors are compared.

Many researchers present other algorithms used in the proposed system but none of them compare their success as a single study under the same conditions and in the same data for all six. By collecting customer data and correlating it with behaviors over time, machine learning algorithms can learn associations and help teams tailor product development and marketing initiatives to customer demand.

In India, air pollution is considered a widespread problem. Daily the specified air quality indicators are significantly higher than the highest rates considered to be appropriate for public health care. The situation is worse in large urban areas such as Delhi where AQI has achieved the highest total time of 999 AQI. The central government and national authorities have implemented a number of measures to reduce air pollution.

The first phase requires air quality indicator prediction for the situation to improve. Six different project class dividers have been created based on different algorithms. Database analysis by supervised machine learning program (SMLT) to capture a few similar information, dynamic analysis, single dynamic analysis, dynamic dual analysis and various analyzes, short-term treatment and data analysis.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue VI June 2022- Available at www.ijraset.com*

In India, as in many different countries, the record revolves around six key elements polluted by particles less than 10 micrometers in diameter (PM10),particle less than 2.5 micrometers in diameter (PM2.5), carbon monoxide (CO), Ammonia (NH3), nitrogen dioxide (NO2), and ozone (O3).

At that point, the test station should have the option of supplying one with a specific toxicity and its norm in a certain period of CO and O3, which is generally considered to be in control for more than eight hours, and in the other three, a normal 24 hours. The unit of measurement per cubic meter is a microgram (or milligram, due to CO).

### *A. System Architecture*

The first step is to provide dataset to the prediction model by the user. The dataset supplied to machine learning model is used to train the model. Every new data detail filled at the time of application form acts as a test data set. The obtained data set and the previous data set are kept in Datawarehouse. Pre- processing and validation are the next stage. Pre-processing corresponds to the data transformations that are performed before the algorithm is processed.

Data Pre-processing is a method for transforming raw data into a clean data package. In other terms, once the data is obtained from different outlets, it is raw and can not be evaluated. The validation is the import of the dataset enabled library bundles. It analyses the variable identification by data shape, data type and evaluating the missing values and duplicate values.

A validation dataset is a sample of data held back from training the model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating the models.

After pre-processing the data, several machine learning algorithms such as decision tree, Random forest, Naïve Bayes, Logistic Regression, SVM , etc. are used to train the dataset and predict the AQI. The machine learning algorithm which gives the best accuracy is selected to builds the prediction model.

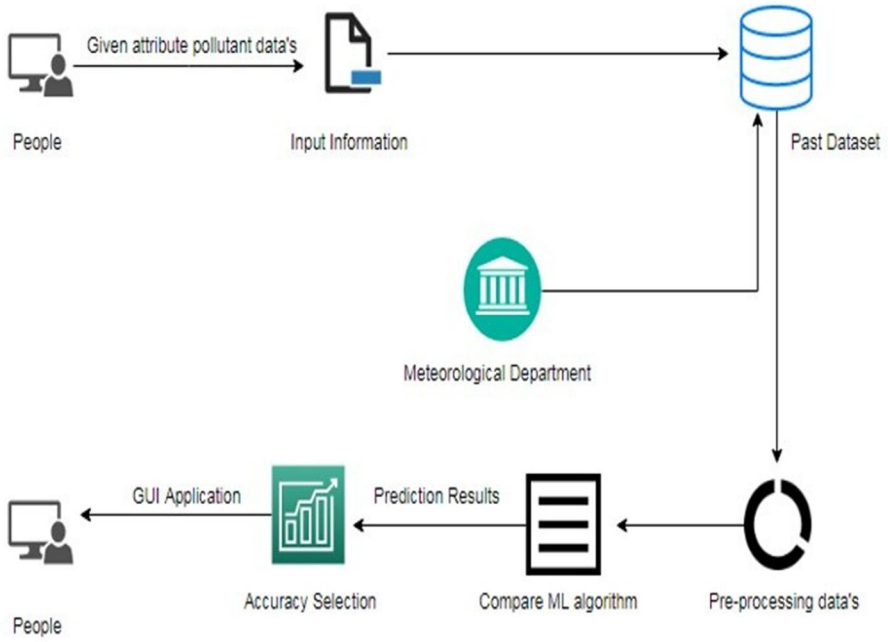



Figure 1.1: System Architecture Of The Proposed System

Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. Photo graphic method is not sufficient to calculate PM 2.5 and it takes only one pollutants of concentration . It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created.

It is necessary to regularly compare the output of several distinctly different learning algorithms and to build a test harness to easily compare multiple simultaneous learning algorithms in Python with scikit-learn. The work given is based on only one PM 2.5 and to collect more monitoring data from other cities to verify the generalization of the work and more factors such as geomorphic conditions. Finally, an interactive GUI (graphical user interface) is developed using Tkinter in python library.

### B. Objective

The goal is to develop a machine learning model for real-time air quality forecasting, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

### C. Problem Statement

The management and protection of air quality in many industrial and urban areas is actually a critical tasks, owing to specific styles of emissions induced by the usage of transport electric power, etc. the accumulation of toxic gasses presents a significant danger to the quality of life in smart cities. As air pollution rises, needed a effective air quality control models to gather information on air pollutant concentrations and to measure air pollution in all areas.

## II. REVIEW OF LITERATURE

### A. Photo–Based PM2.5 Concentration Estimation

A confluence of both features described above with a non-linear mapping procedure may project the PM2.5 concentration of the image. Photo graphic method is not sufficient to calculate PM2.5and it taken only one pollutants of concentration. Research is performed on the context of only one PM2.5 and to obtain further testing data from others, for example geomorphic conditions, to check that research is generalized. Thus the regularity of air pollutant data cannot be more accurately determined and the prediction results obtained.

### B. Effects of Air Pollution on Hospital Admissions

De Leon presented the effects of air pollution on daily hospital admissions for respiratory disease in Poisson regression analysis of day by day tallies of hospital confirmations, modifying for impacts of pattern, occasional and other recurrent variables, day of the week, occasions, flu pestilence, temperature, dampness, and auto connection. Variables of pollution were particulates black smoke (BS), nitrogen dioxide (NO2), Sulphur dioxide (SO2), and ozone (03).

### C. Data Cleaning for Data Quality

Challa presented Data cleaning techniques to achieve quality data. Now every second data is rapidly generated over the internet for a day and thus it has become a huge task to make the right decision. Surrounded by data but hungry for information, knowledge plays a vital part for decision-makers to generate more income in industry. Data collected from different sources may have dirty information, data cleaning should be performed before the data is loaded into the warehouse in order to obtain data quality.

### D. Survey on Big Data Analytics

Acharjya paper's primary objective is to research the potential impact of big data problems, issues in open research, and different apparatuses related with it. This paper provides important insights on big data and the principles behind its analysis, which is an important factor for this project because of large size of the sample data being used to provide conclusions.

### E. Spatial Characteristics of Air Pollution

It explains that air pollution is a major and pervasive influence in Chengdu, owing to its effect on human health and well-being. The main pollution induced by soil by the use of petroleum products was sulfur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2) and particulate matter (PM10). The outcomes showed that OK strategy was smarter to produce the forecast maps of toxin fixations than IDW technique. High spatial decent variety of NO2, SO2 and PM10 focuses existed, particularly in the east-west course. Between the six principle locales in Chengdu, Chengdu District was dirtied all the more truly in 2010.

## III. WORK DONE

### A. Exploration Data Analysis of Visualization

Data analysis is an essential capability in computational analytics and machine learning. In addition, statistics rely on detailed explanations and data estimates. Data visualization is an effective toolkit to achieve a contextual understanding. It can be helpful as you research and study a dataset which can help to detect trends, missing results, outliers and several other items.

Data visualizations are able to be utilized with a little technical awareness to communicate and explain core relations in more graphic and involved plots and maps than cumulative or concrete measures.

Only data becomes useless before it becomes readily visible, such as graphs and charts. The ability to easily imagine examples of data and others is essential for both applied analytics and machine learning. You can consider the various styles of pictures you need to recognize while you are visualizing Python data and how you can use them to fully visualize your personal data.

- How to model time series details using bar charts of lines and categorical numbers?
- How to sum up distributions of data using histograms and box plots.
- How to sum up variables 'connection to scatter plot.

Most machine learning algorithms are prone to attribute values in input data and delivery. Input data outliers will deceive and misrepresent the machine learning algorithms training cycle, which contributes to longer training times, less reliable models and eventually weaker performance.

Well before predictive models of training data are created, outliers will lead to misleading representations and misleading perceptions of gathered information. In descriptive statistics such as mean and standard deviation and in tracks such as histograms and scatterplots, Outliers will distort the overview distribution of attribute values by consolidating the body of evidence.

A histogram is a bar representation of data that varies over a range. It plots the height of the data belonging to a range along the y-axis and the range along the x-axis. Histograms are used to plot the data over a range of values. The Histograms use a bar representation to show the data belonging to each range. Use the iris data which contains the information about flowers to plot the histograms. Finally, outliers may be indicators of data cases, such as fraud detection abnormalities and computer protection, which are important to the issue. It can not adapt the model to the training data and it can not guarantee whether the model would function correctly on the real data. You need to be sure the model has the correct data characteristics so it will not have too much noise. Cross-validation is a method in which a model is educated using a data-set subset and then tested with an adjacent dataset sub-set.

### *B. Variable Identification Process*

Machine learning validation techniques are employed to obtain the machine learning (ML) error rate model, which can be considered as similar to the true data set error rate. You may not need testing strategies if the data volume is sufficiently high to represent the population. To locate the missing value, repeat the value and the data form definition, if the attribute is float or integer.

The data sample was used to provide an objective model fit evaluation on the testing dataset when adjusting hyper parameters of the model. The appraisal gets more skewed as expertise is integrated into the model setup on the validation dataset. To test a given model, the validation collection is used, but this is for regular evaluation. As machine learning developers, this data is used to fine-tune the hyperparameter of the model.

#### *1) Data Validation*

Data processing, data analysis, and the process of addressing data material, consistency, and configuration will add up to a time-consuming list of tasks. It helps to understand the data and its properties during the data recognition process; this information can help you select which algorithm to use to construct the model. By regression algorithm, for example, data from time series can be analyzed; classification algorithms can also be utilized to evaluate discrete data.

#### *2) Data Pre-Processing*

Pre-processing relates to the preparation of the data until it is passed to the algorithm. Preprocessing data is a method for transforming raw data into a clean collection of data. With other words, if the data were obtained with raw format from various outlets, this is not appropriate for study. The machine learning cycle will prove to be right in order to obtain improved outcomes from the implemented model. Random Forest algorithm does not accept null values. Certain Machine Learning models require details in a defined format. Therefore, null values from the initial raw data collection must be handled to operate random forest algorithms. And it is also critical that data sets are so structured that more than one machine learning and deep learning algorithm in a given dataset is implemented.

### *C. Decision vs Logistic Algorithm*

#### *1) Logistic Regression*

Analysis of logistic regression seems to be the most common approach of regression which could be used in binary dependent parameter modeling. Logistic regression is a computational methodology which often represents the associations between the independent variables x1 x2 ... xn and y that is the two alternative categories of discrete dependent variable coded in 0 or 1. Independent variables may be permanent, discrete, binary or combined.

That is, P(Y=1) as a function of X is predicted by the model of logistic regression.

Logistic regression Assumptions:

- The dependent variable must be binary for binary logistic regression.
- The desired result should be the level 1 factor for the binary regression of the dependent variable.
- Only the meaningful variables should be included.
- The separate variables should be mutually independent. In other words, the model ought to have little.
- The independent variables have a linear relationship with the log odds.
- Logistic regression requires sample sizes quite large.

*2) Decision tree*

The algorithm of the decision array falls into the supervised learning algorithm category. It works for both continuous and categorical variables of output.

Assumptions of Decision tree:

- At the beginning, considering the whole training set as the root.
- Attributes are considered to be categorical for finding information gain, given the attributes are continuous.
- Recursive distributions are made on the basis of attribute values.
- Using statistical methods for root or internal node ordering attributes.

Decision Trees is a data gathering categorization and analysis technique. Decision trees constitute the fundamental recursive basis for the sequence process of classification, in that one of the disjoint class decision-making structures contains nodes and leaves, is allocated a case identified with a collection of attributes. Each tree node includes testing a specific attribute and every tree leaf denotes a class. The test usually compares the value with a constant of an attribute. Leaf nodes are usually categorized in all cases whereas for classified set is entered or where probabilities are spread over any possible classification.

The process is continued until the termination is completed. It is designed in a recur- sive dividing-and-conquer top-down fashion. All characteristics should be categorical. They would otherwise be discretized in advance. Top of the tree characteristics have a greater effect on the classification and the knowledge gain principle is used. Too many branches can easily be over-fitted and can reflect disturbances caused by noise or bolts.

*D. Support Vector Machine vs Random Forest Algorithm*

*1) Support Vector Machines*

Support Vector Machine has been a collection of similar supervised learning methods for use in categorization and regression. This belongs to a family of generalized linear classifications. One unique aspect of SVM is that the empirical classification error is minimized and geometrical margin maximized at the same time.

- How to detangle the various names used for referring vector support machines.
- The representation that SVM will use when the model is actually stored on the disk.
- How to use a trained representation of the SVM model to make accurate predictions for new data.
- What to learn from the training data on an SVM platform.
- How would the data be better optimized for SVM?
- Where can you look for additional SVM information

*2) Random forest*

Random forests were a hybrid of trees forecasting flaws in forest general statement since each tree operates on the theory of the random variable within each tree. The error of generalizing forests depends on the strength of each of the forest's trees and the relationship between them on the individual since the percentage of forest trees is vital. By randomly selecting the characteristics for splitting every node, error levels are similar, but are richer in noise.

In carrying out the random forest algorithm the following are basicsteps:

- From the given dataset, choose N Random Data.
- Create a decision tree on the basis of those N records.
- Pick the number of trees you want and repeat steps 1 and 2 in the algorithm.
- In the event of a problem with regression, for a new record, the value for Y for each tree in the forest is expected. The final value can be calculated by taking the sum of all the forested trees 'values.

*E. K-Nearest Neighbor vs Naive Bayes Algorithm*

*1) K-Nearest Neighbor*

K-Nearest Neighbor is a supervised learning algorithms which hold all instances in n- dimensional space corresponding to training data points. When an unknown discrete data is obtained, it analyzes the nearest k number of stored instances (near neighbors) and finally returns the most previewed class and returns the average of k similar neighbors for real-value results.

In the nearest distance algorithm, each of the k neighbors assesses the contribution by distance by the subsequent question and gives more weight to the nearest neighbors. KNN is typically resilient to noisy data since the closest neighbors are combined. The algorithm of K-nearest is a classification algorithm, which is supervised: a group of defined points is needed and is used to learn how to label other points. This searches for a specific point which has such neighbors votes in view of the points that are labelled nearest to the new point (the "k" is the amount of neighborhoods that it checks). It aims at the neighbors' voting points, and the mark that is the most significant one of the neighborhood. Predict the testing collection for the whole training package. By filtering through the entire collection to find the k "closest" instances, KNN prediction for a new case. Close-up is evaluated across all apps of proximity (Euclidean).

*2) Naive Bayes algorithm*

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features.

Naive Bayes is a Bayes Theorem-based system of statistical classification. This is one of the basic guided algorithms for learning. A quick, accurate and reliable algorithm is the Naive Bayes classifier. High precision and efficient on broad datasets are the Naive Bayes classifier. The classification of Naive Bayes means that the influence of a certain function in a class is unconditional to others. Although these characteristics are interdependent, they are still independent. This presumption renders equations easier and is thus called naive.

## IV. SYSTEM DESIGN

The raw recorded data contains meteorological information for various cities in India. The idea is to use various machine learning techniques such as Logistic Regression, Support Vector Machines, Random Forests, K-Nearest Neighbors, Naïve Bayes Classifier and Decision Tree for model building and research purposes.

*A. Data Analysis and Visualization*

Data visualization is an essential specialist knowledge in applied statistics and the statis- tics of machine learning rely practically on quantitative explanations, and software analyses are valuable instruments for obtaining qualitative understanding. This helps to discover and know about a dataset and help to identify trends in fraudulent data transfers and more. Information visualizations are feasible with low domain knowledge to communicate and display core connections through graphs and maps which are more emotional and consumers than associative or substantive steps.

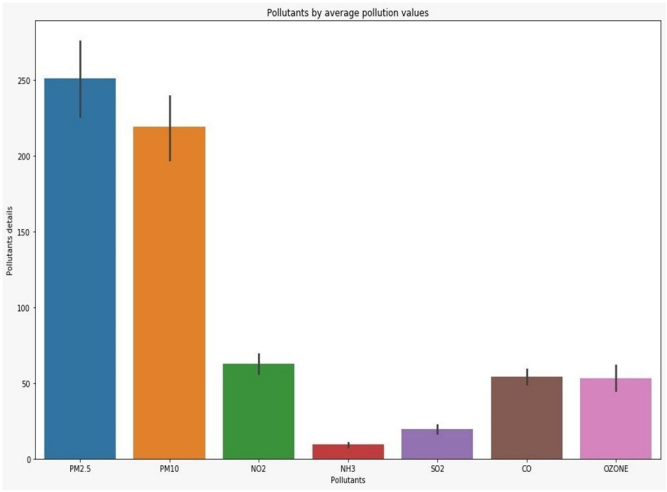



Figure 4.1: Demonstration of Pollutants vs Avg Pollution

For the estimation of results, the data collection is split into two groups, that is, a training collection, and a test set that typically utilizes 7:3 ratios to distinguish the training sample from the test sample. The data model built using Random Forest, Logistic Rectification, Decision Tree Algorithms, KNN, Support Vector Classifier (SVC) and Naive Bayes is applied to the training set and the Test Set Prediction is conducted on the basis of the precision and the test output.

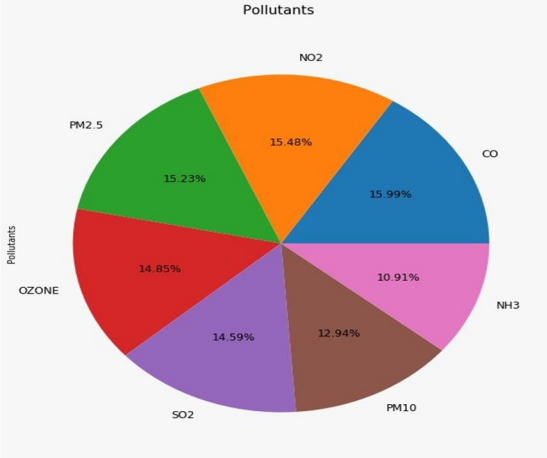



Figure 4.2: Demonstration of Concentration Pollutants

The raw recorded data contains meteorological information for various cities in India. The idea is to use various machine learning techniques such as Logistic Regression, Support Vector Machines, Random Forests, K-Nearest Neighbors, Naïve Bayes Classifier and Decision Tree for model building and research purposes.

## V. RESULTS AND DISCUSSION

Four Machine Learning techniques i.e. Naive Bayes method, Random Forests algorithm, K- Nearest Neighbor (KNN) and Decision Tree algorithm, were used for the purpose of building a prediction model.

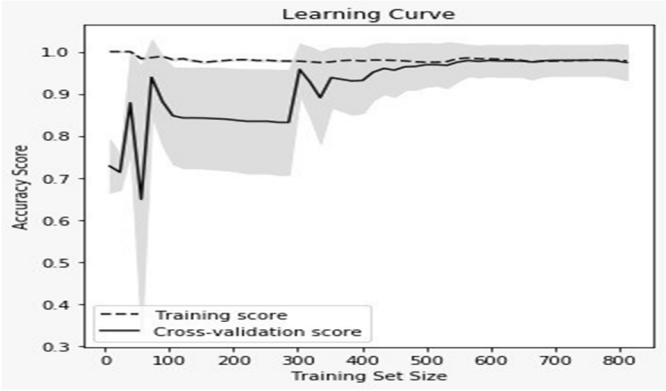
### *A. Naïve Bayes Algorithm*

Naïve Bayes is quite a basic learning algorithm that uses Bayes and strongly assumes that the attributes in the groups are conditionally separate. Combined with its analytical performance and many other attractive attributes, this leads to the use of naive Bayes in practice. After applying Naive Bayes Classifier, the prediction accuracy obtained was 97.38

Figure 5.1: Classification report of Naïve Bayes

```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       0.94      0.99      0.96        73
           1       0.99      0.97      0.98       175

    accuracy                           0.98       248
   macro avg       0.96      0.98      0.97       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         0.94117647 1.
 0.94117647 1.         1.         0.94117647 1.         0.94117647
 0.88235294 0.94117647 1.         1.         0.94117647 0.94117647
 1.         0.94117647 0.88235294 0.94117647 0.94117647 0.88235294
 1.         0.94117647 1.         1.         0.9375     1.
 0.875      1.         1.         1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Naive Bayes is: 97.38235294117646

Confusion Matrix result of Naive Bayes is:
 [[ 72   1]
 [  5 170]]

Sensitivity :  0.9863013698630136

Specificity :  0.9714285714285714
```

Figure 5.2: Naïve Bayes Accuracy graph



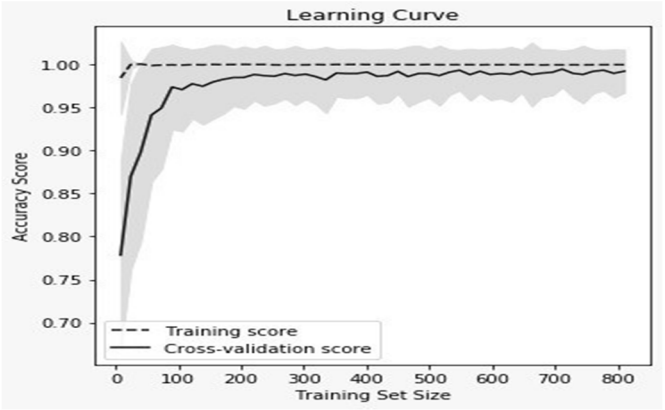

*B. Random Forest Algorithm*

Random forests were a hybrid of trees forecasting flaws in forest general statement since each tree operates on the theory of the random variable within each tree. The error of generalizing forests depends on the strength of each of the forest's trees and the relationship between them on the individual since the percentage of forest trees is vital. By randomly selecting the characteristics for splitting every node, error levels are similar to adaboost, but are richer in noise. After applying Random Forests algorithm, the prediction accuracy obtained was 99.16

Figure 5.3: Classification report of Random Forest

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       0.99      0.99      0.99        73
           1       0.99      0.99      0.99       175

    accuracy                           0.99       248
   macro avg       0.99      0.99      0.99       248
weighted avg       0.99      0.99      0.99       248

Cross validation test results of accuracy:
[0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 0.94117647 1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         0.94117647 1.         0.94117647
 1.         1.         1.         1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Random Forest is: 99.16911764705883

Confusion Matrix result of Random Forest is:
 [[ 72   1]
 [  1 174]]

Sensitivity :  0.9863013698630136

Specificity :  0.9942857142857143
```

Figure 5.4: Random Forest Accuracy graph



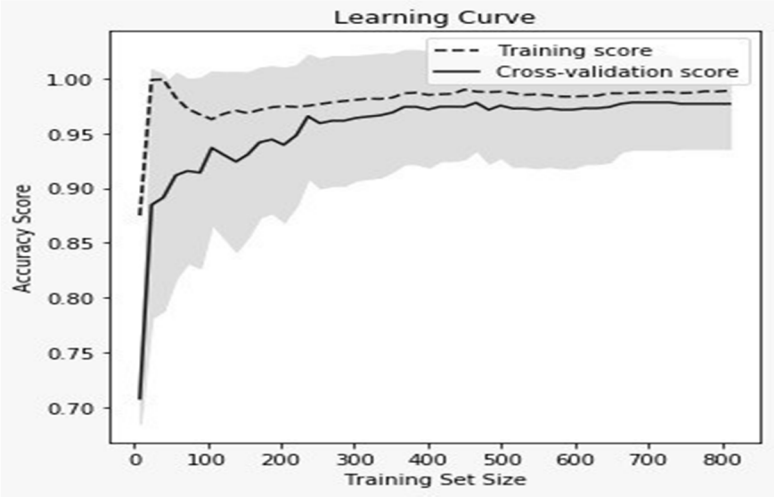
### *C. K-Nearest Neighbors Algorithm*

K-Nearest Neighbors is the simplest and most direct method for classification where the distribution of data has had no firsthand knowledge. This rule essentially preserves the whole training set in the process of learning and assigns a class representing the majority mark of its closest neighbors in the training set to each submission. The next most simple form of KNN is the nearest neighbor's law (NN) when N= 1. After applying KNN, the prediction accuracy obtained was 97.61

Figure 5.5: Classification report of K-Nearest Neighbors

```
Classification report of K-Nearest Neighbor Results:

              precision    recall  f1-score   support

           0       0.96      0.97      0.97        73
           1       0.99      0.98      0.99       175

    accuracy                           0.98       248
   macro avg       0.97      0.98      0.98       248
weighted avg       0.98      0.98      0.98       248

Cross validation test results of accuracy:
[1.         1.         1.         0.94117647 0.94117647 0.94117647
 1.         0.94117647 1.         0.94117647 1.         1.
 0.94117647 1.         1.         1.         1.         1.
 1.         1.         1.         1.         0.82352941 1.
 0.94117647 1.         0.94117647 0.94117647 0.94117647 0.94117647
 0.94117647 0.94117647 1.         0.9375     1.         1.
 0.875      1.         1.         1.         0.9375     1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of K-Nearest Neighbor is: 97.61764705882354

Confusion Matrix result of K-Nearest Neighbor is:
 [[ 71   2]
 [  3 172]]

Sensitivity :  0.9726027397260274

Specificity :  0.9828571428571429
```

Figure 5.6: KNN Accuracy graph



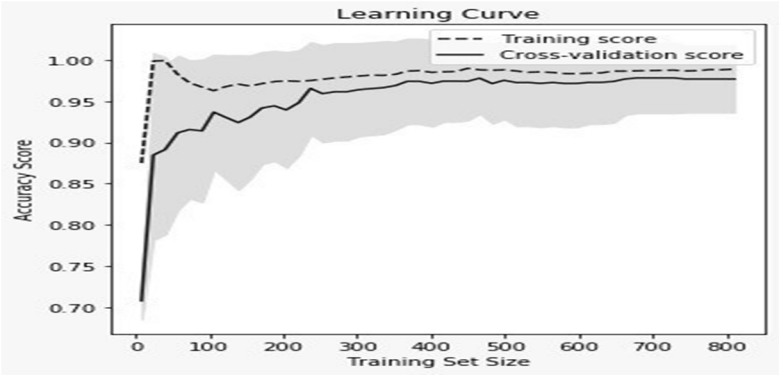
### *D. Decision Tree*

Decision Trees is a data gathering categorization and analysis technique. Decision trees constitute the fundamental recursive basis for the sequence process of classification, in that one of the disjoint class decision-making structures contains nodes and leaves, is allocated a case identified with a collection of attributes. Each tree node includes testing a specific attribute and every tree leaf denotes a class. The test usually compares the value with a constant of an attribute. Leaf nodes are usually categorized in all cases whereas for classified set is entered or where probabilities are spread over any possible classification. After applying Decision Tree algorithm, the prediction accuracy obtained was 99.88

Figure 5.7: Classification report of Decision Tree

```
Classification report of Decision Tree Classifier Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00        73
           1       1.00      1.00      1.00       175

    accuracy                           1.00       248
   macro avg       1.00      1.00      1.00       248
weighted avg       1.00      1.00      1.00       248

Cross validation test results of accuracy:
[1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         0.94117647
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.         1.         1.         1.         1.
 1.         1.        ]

Accuracy result of Decision Tree Classifier is: 99.88235294117646

Confusion Matrix result of Decision Tree Classifier is:
 [[ 73   0]
 [  0 175]]

Sensitivity :  1.0

Specificity :  1.0
```

Figure 5.8: Decision Tree Accuracy graph




### *E. Testing*

The classification report of all the four algorithms was analyzed and evaluated on the data set and compared each algorithm accuracy with each which other that can be seen in Table 1.

| Algorithm | Accuracy |
|---|---|
| Naive Bayes | 97.38 |
| Random Forest | 99.16 |
| K-Nearest Neighbors | 97.61 |
| Decision Tree | 99.88 |

Table 1: Comparison of Accuracy result of each algorithm

The Decision tree with 99.88 percent precision is perhaps the most effective methodology, whereas the least accurate algorithm is Naïve Bayes with an accuracy of 97.38. On the otherhand the Random Forest has almost near accuracy that of Decision Tree with an accuracy of 99.16, this is due to the fact that random forest is in itself a kind of decision tree and hence resembles almost same accuracy as decision tree. Also the logistic regression, Naïve Bayes and KNN have same accuracy of 97 approximately.

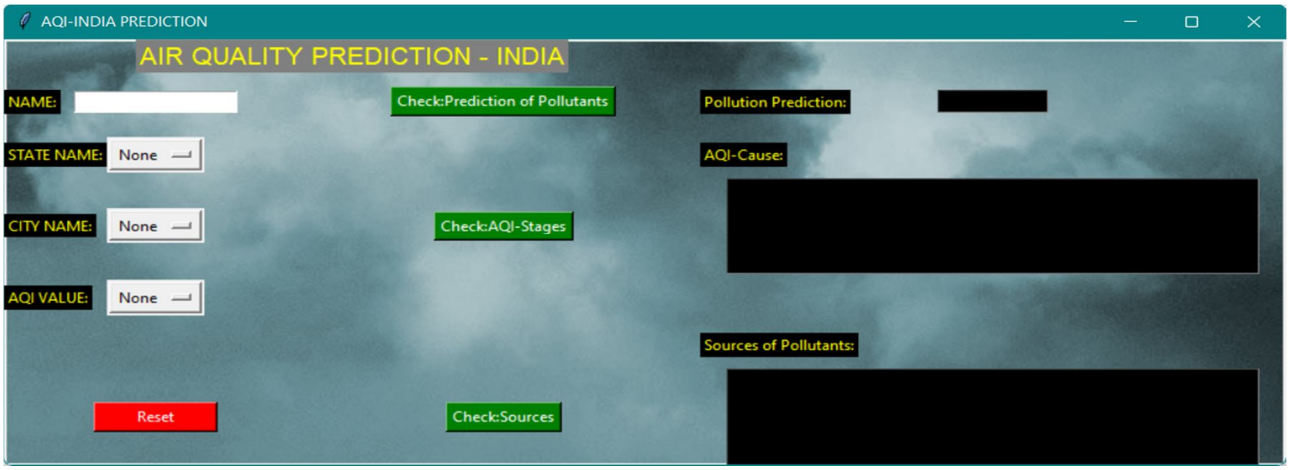



Figure 5.9: GUI Screenshot

## VI. CONCLUSION AND SUMMARY

Prevention of air pollution is the need of the hour, so a powerful machine learningsystem was established with the help of prediction model. Prediction of pollution events has become most important issue in major cities in India due to the increased urbanization of the population and the associated impact of traffic volumes. Data from a variety of heterogeneous resources were used and involved collection and cleansing for use in machine learning algorithms.

The number of model parameters and optimized outputs were reduced with help of structure regularization which in turn, alleviated model complexity. The Decision Tree Algorithm gave the best results among all the algorithms, with an overall accuracy of 99.8. The prediction model precision findings, helped in evaluating and contrasting current work on air quality assessment which is based upon Big Data Analytics and Machine Learning.

## APPENDIX A

### *A. Parameters*

#### *1) Precision and Recall*

Pattern detection is a fraction of the appropriate instances within the retrieved instances, with knowledge retrieval and classification (machine study), accuracy (also called optimistic predictive value), while recall (also known as sensitivity) is the fraction of the total number of the specific instances that were currently retrieved . Accuracy and alert are both dependent on awareness and significance calculation.

#### *2) Air Quality Index*

The state and local agencies use an Air Quality Index (AQI), in order to disclose to the general population how safe or how sterile the environment is actually predictable. Various nations have their own measures of air quality, which are compatible with various national criteria of air quality.

#### *3) Confusion Matrix*

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

#### *4) F1 Score*

The indicator of precision of a check is the F1 (also F-point or F-mount). It takes into consideration both the precision p and the warning r of the check in order to measure the score: p is the proportion of correct positive results determined by the number of possible results recorded by the classifier and r is the number of correct positive results segregated by all samples involved.

*5) A.5 PM2.5*

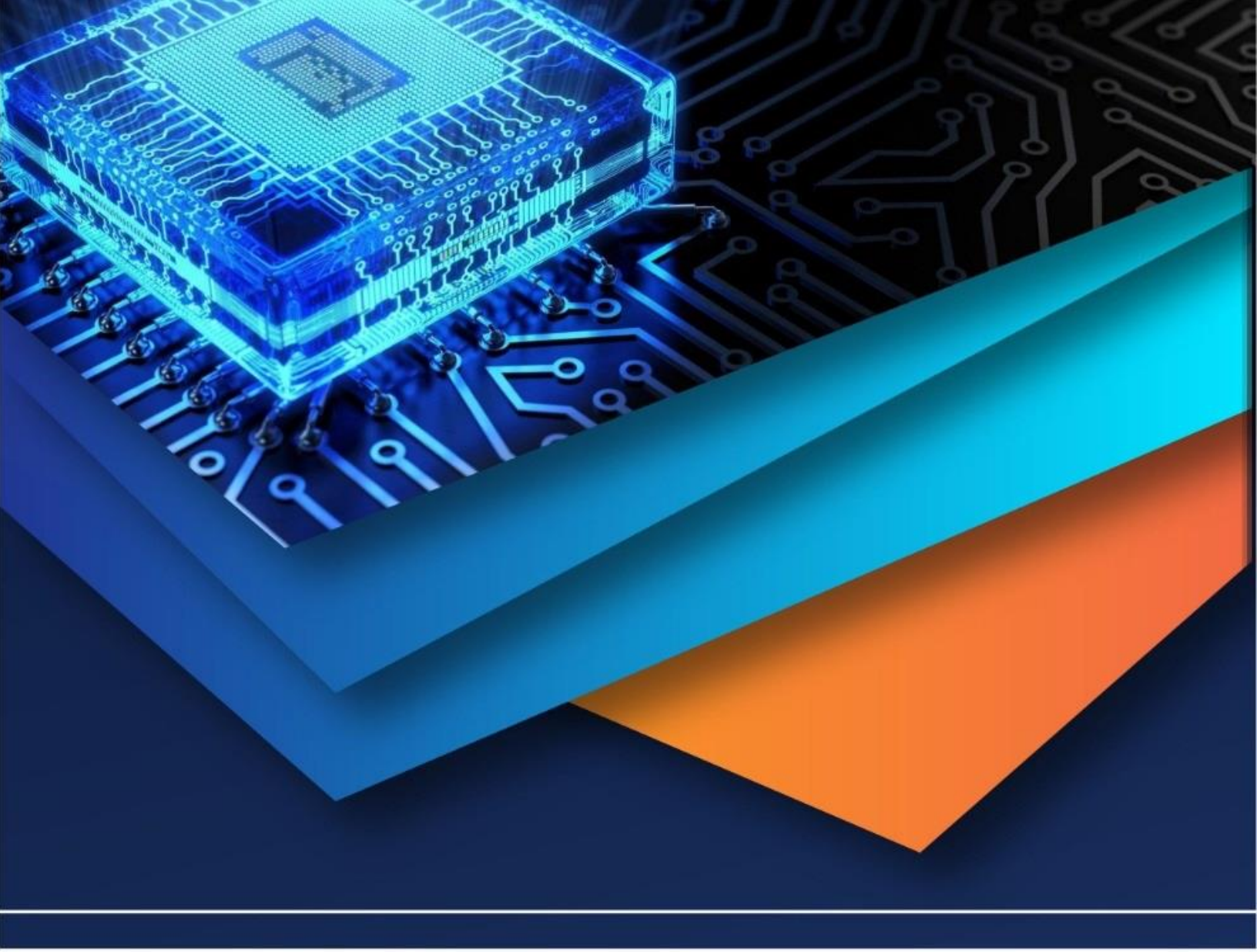
PM2.5 applies to particle atmosphere that is smaller than 2.5 micrometers in diameter and is about 3 percent of the thickness of the hair of humans. The particles usually known as PM2.5 are so tiny that even an electron microscope can be observed. These are also smaller than PM10, which are 10 micrometers or fewer and are known as tiny particles.

*6) A.6 Learning Curve*

In machine learning, the validity and testing performance for an estimator for a variety of simulation samples is seen in the learning curve (or testing curve). This is a guideto understand how much more training data will improve a machine understand algorithm, and if the estimator has a variance flaw or bias mistake. It is also a aid. When both the validity score and the training scores are converged to an extremely low value with the growing scale of the training collection, further training results does not gainmuch.

## REFERENCES

[1] Acharjya, D. P. "A survey on big data analytics: challenges, open research issues and tools." , (2019)
[2] Challa, J. S., Goyal, P., Nikhil, S., Mangla, A., Balasubramaniam, S. S., and Goyal N. "Dd-rtree: A dynamic distributed data structure for efficient data distribution among cluster nodes for spatial data mining algorithms." , 2016 IEEE International Conference on Big Data (Big Data). 27–36.
[3] De Leon, A., Anderson, H., Bland, J., Strachan, D., and Bower, J. "Effects of air pollution on daily hospital admissions for respiratory disease in london between 1987-88 and 1991-92." (1996) Journal of epidemiology and community health, 50 Suppl 1, s63–70.
[4] Li, S., Song, S., and Fei, X. "Spatial characteristics of air pollution in the main city area of chengdu, china." (2011) 19th International Conference on Geoinformatics, 1–4.
[5] Qin, D., Yu, J., Zou, G., Yong, R., Zhao, Q., and Zhang, B. "A novel combined prediction scheme based on cnn and lstm for urban pm2.5 concentration." (2019) IEEE Access, 7, 20050–20059.
[6] Yue, G., Gu, K., and Qiao, J. "Effective and efficient photo -based pm2.5 concentrationestimation." (2019) IEEE Transactions on Instrumentation and Measurement,PP,1–10



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY