International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Prediction of Air Pollution by using Machine Learning Algorithm

[1]Dr.CHIRAPARAPU SRINIVAS RAO, [2] PENMATSA LAKSHMI SOWJANYA

[1]Associate Professor, S.V.K.P & Dr K.S. Raju Arts & Science College(A), Penugonda, W.G. District, Andhra Pradesh, chiraparapu@gmail.com

[2]PG, Scholar, S.V.K.P & Dr K.S. Raju Arts & Science College(A), Penugonda, W.G. District, Andhra Pradesh, sowjanyapenmatsa1@gmail.com

ABSTRACT

controlling and defensive the higher air greatness has gotten one in everything about first imperative occasions in different creating and metropolitan districts at the present. The greatness of air is adversely contacting collectible to the different styles of tainting influenced through the transportation, power, powers consumptions, and so forth. In our country population is a big problem as day by day population is increasing, so the rapid increasing in population and economic upswing is leading environment problems in city like air pollution, water pollution etc. In some of air pollution and air pollution is direct impact on human body. As we know that major pollutants are arising from Nitrogen Oxide, Carbon Monoxide & Particulate matter m=(PM), SO2 etc. Carbon Monoxide is arising due to the deficient Oxidization of propellant like as petroleum, gas, etc. nitrogen oxide (NO) is arising due to the ignition of thermal fuel; Sulphur Dioxide(So2) is major spread in air, So2 is a gas which is present more pollutants in air, it's affect more in human body. the predominance of air is overstated by multidimensional impacts containing spot, time and vague boundaries. The goal of this improvement is to take a gander at the AI basically based ways for air quality expectation. In this paper we will predict of air pollution by using machine learning algorithm

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

1.INTRODUCTION

The Environment describe about the thing which is everything happening in encircles the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries there are 75 % NO or other gas like CO, SO2 and other particle is exist in environment. The expanding scene, vehicles and creations square measure harming all the air at a feared rate.

Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of Delhi

2.LITERATURE SURVEY

2.1 INTRODUCTION

In literature, a lot of work is done in the study and analysis of air pollution as well as predicting the future trends. In [2] Linear regression-based air pollution prediction is done. It suggests cloud data for data analytics which can be used for taking the decision to minimize pollution. But they have used BI service and Microsoft Azure for analysis which is very expensive services. The model is not very accurate because of linear regression-based model. In [1] machine learning based air pollution prediction is done. It suggests multilayer perceptron which results in very accurate result. But it takes large datasets and long duration for training. In [3] Recurrent Neural Network based model for air pollution prediction is done. It suggests using machine learning algorithm and recurrent neural network for prediction which generates most accurate result but, its very expensive to implement.

2.2 ALGORITHMS:

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not “learn” until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training **dataset**

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (**Weka 3.6.0**, **R 2.9.2**, **Knime 2.1.1**, **Orange 2.0b** and **RapidMiner 4.6.0)**. We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point *x* and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

The Air Pollution Forecasting System: Air Quality Index (AQI) is a record that gives the public the degree of contamination related with its wellbeing impacts. The AQI centers around the different wellbeing impacts that individuals may encounter dependent fair and square and long stretches of introduction to the poison concentration. The AQI values are not quite the same as nation to nation dependent on the air quality norm of the country.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

3.EXIXTING SYSTEM

The higher the AQI level more noteworthy is the danger of wellbeing related problems. The by and large point of this venture is to make a student calculation that will have the option to foresee the hourly contamination focus. Additionally, an Android application will be built up that will provide the clients about the constant contamination convergence of PM2.5 alongside the hourly forecasted value of the toxin fixation from the student calculation. The Android application will also recommend data of the less dirtied[1].

Disadvantages

- The system is not implemented Stepwise Multiple Linear Regression Method.
- The system is not implemented Instance-Linear Regression Model

Proposed System

1) Data assortment: There is a different method from which we collected data from various dependable sources like Delhi Gov. site.

2) Exploratory examination: We research and explore examination with various parameter like ID of outliners, consistency check, missing qualities, and so on, it's totally occurred in this period of the venture.

3) Data Manipulation control: In period of data control stage the required missing data need to insert in utilizing the mean estimations of that characteristic of information. [2]

4) Prediction of boundaries utilizing by gauge model: For appropriate data indirect relapse we have to keep future qualities for different boundaries just

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

5) Implementation of straight relapse: Whenever all the boundaries become in active mode or they are accessible mode, the direct relapse calculation would be used in anticipate the air quality index (AQI).

6) Data accuracy investigation: We have to analyze that used model is being fit for overall data or not so we have to cross check root mean error, absolute percentage error then after we have to assume this factor is good for accuracy or not

Advantages

- The proposed system implemented Linear Regression is basically use for predicting the real values data y using continuous parameter.
- Stepwise Multiple Linear Regression Method is used for continuous data testing and training in effective way.

ARCHITECTURE

System Architecture mainly consist s of 2 modules and database to store all the data .Those are:

- Remote User
- Service provider

The Remote User module can perform the following operation: Register and login, view your profile, Predict Air pollution Type

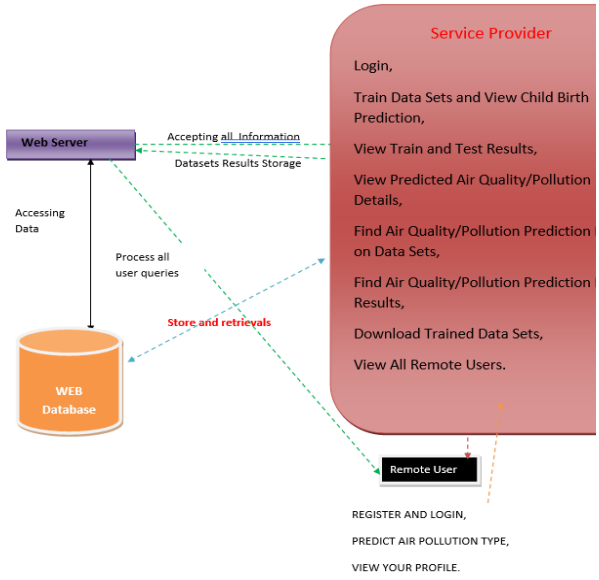
The Service provider module can perform the following operations: Login, Browse and Train &Test data sets, View Trained And tested Accuracy in Bar Charts, view Trained and Tested Accuracy Results, find air quality pollution predict ratios on data sets , Air quality pollution predict ratios Results, View All Remote Users.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Architecture Diagram



Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train Data Sets and View Child Birth Prediction, View Train and Test Results, View Predicted Air Quality/Pollution Details, Find Air Quality/Pollution Prediction Ratio on Data Sets, Find Air Quality/Pollution Prediction Ratio Results, Download Trained Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT AIR POLLUTION TYPE, VIEW YOUR PROFILE.

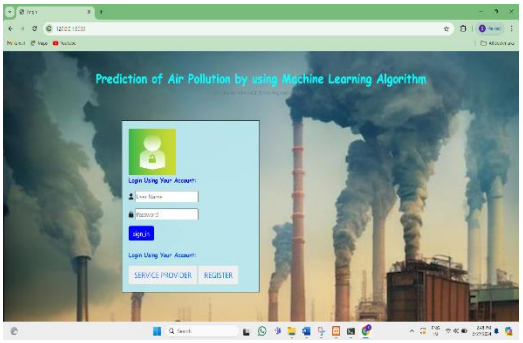
International Journal of Engineering Research and Science & Technology

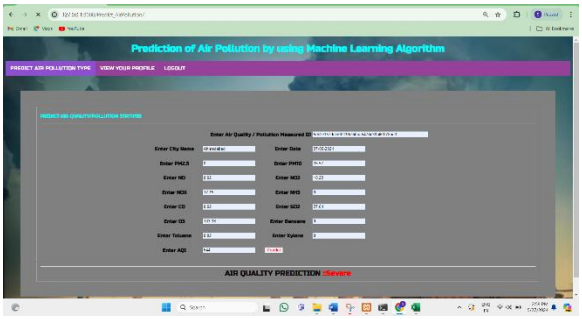
ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

SCREENS

User Login:





Output Screen:

CONCLUSION

Precision of our model is very acceptable. The anticipated AQI has a precision of 96%. Future upgrades incorporate expanding the extent of district and to incorporate whatever number locales as could be allowed as of now this venture targets foreseeing the AQI estimations of various areas of close by New Delhi. Further, by utilizing information of various urban areas the extent of this venture can be exhausted to anticipate AQI for different urban communities also.

REFERENCES

[1] Ni, X.Y.; Huang, H.; Du, W.P. “Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data.” Atmos. Environ. 2017, 150, 146-161.

[2] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.

[3] Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd),

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

Akshaya A.C. "Indian Air Quality Prediction And Analysis Using Machine Learning". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

[4] Suhasini V. Kottur , Dr. S. S. Mantha. "An Integrated Model Using Artificial Neural Network

[5] RuchiRaturi, Dr. J.R. Prasad ."Recognition Of Future Air Quality Index Using Artificial Neural Network".International Research Journal ofEngineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations," IEEE ACCESSJuly 30,

[6] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu ." Detection and Prediction of Air Pollution using Machine Learning Models". International Journal o f Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018

[7] Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie." Air Quality Prediction: Big Data and Machine Learning Approaches". International Journal o f Environmental Science and Development, Vol. 9, No. 1, January 2018

[8] PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG," 2018.Digital Object Identifier10.1109/ACCESS.2018.2849820.

[9] GaganjotKaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Gang Xie,"Air Quality Prediction: Big Data and Machine Learning Approaches," International Journal of Environmental Science and Development, Vol. 9, No. 1, January2018.

[10] Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner and Daniel Price, "Air Quality Simulations using Big Data Programming Models," IEEE Second International Conference on Big Data Computing Serviceand Applications,2016.