AIR QUALITY PREDICTION USING MACHINE LEARNING ALGORITHMS

Dr. V. Anantha Krishna[1], Harika Koganti[2], M. Madhumathi[3], V. Dharani[4]

Article History: **Received:** 08.02.2023 **Revised:** 23.03.2023 **Accepted:** 08.05.2023

Abstract

Air quality prediction is an important process that must be taken by the state authorities as the health of human is sincere concern. By measuring index of air, we can prevent humans from getting affected by various diseases. The diseases that effect the humans health are lungs cancer, brain disease and even a person may die because of this air quality index can be determined using machine learning algorithm. Although, different researches are happening on this issue but there are no accurate results and are not that successful. Kaggle dataset is being used in this project and this dataset is divided into training and testing. The algorithms used in this project are Linear Regression, Random Forest and C 4.5 Decision tree. Through this paper represents our hard work to help in managing this problem. By calculating air quality index we can enhance the situation. By this project, the main objective is to forecast the index value of air and let the people in the society know the amount of pollution that is present in the air. In this paper, it is shown that many machine learning algorithms are used based on comparative analysis.

Keywords: Machine Learning, Air Quality Index, Random Forest, Decision Tree, Artificial Neural Network, Meteorologists, Data Processing, Feature Selection, Data Splitting, Sulphur Oxide, Carbon Monoxide, Nitrogen Oxide, Chlorofluorocarbons.

[1]Professor, Computer Science and Engineering, Sridevi Women's Engineering College,Hyderabad, India
[2]Computer Science and Engineering, Sridevi Women's Engineering College, B.Tech IVYear Hyderabad, India
[3]Computer Science and Engineering, Sridevi Women's Engineering College, , B.Tech IVYear Hyderabad, India
[4]Computer Science and Engineering, Sridevi Women's Engineering College, B.Tech IVYear, Hyderabad, India

Email: [1]krishnaananthav@gmail.com, [2]Kogantiharika5@gmail.com, [3]Madhumaliptel55@gmail.com, [4]Vellankidharani22@gmail.com

DOI: 10.31838/ecb/2023.12.s3.275

1. Introduction

Although there are various types of pollutions such as water, soil, air pollution. The most effective pollution among there is air pollution which is health hazardous to human health. So, it is really very important to predict air quality index value. So, air pollution must be decreased as soon as possible and through this project, we educate humans regarding the pollution in air. So, the higher authorities of the state must take their as an challenging issue and do the needful to help the people inhale fresh air.

Some of the Primary Pollutants are Chlorofluorocarbons (CFC), nitrogen dioxide and carbon family gases.

Some of the Secondary Pollutants are Ground Level Ozone, Acid Rain.

Air pollution prevention is the big issue since many years. But it still remain huge problem in the society. Air pollution gets new diseases to human in respiratory and heart diseases. With these diseases these occurs increased risk and increased rate of death. The government must take some important and tips to understand and forecast air quality index for better human health. Since, air pollution is the most concerned issue now a days, air quality must be checked continuously and predicted to ensure better life for humans. The Environmental Protection Agency (EPA) users index value to predict the condition of air. Air quality index is measured accurately with accurate senior readings and calculations.

Here, it is shown that many machine learning algorithms are used based on comparative analysis. All the algorithms are performed in the dataset and the best algorithm is selected by the accurate results that is provide.

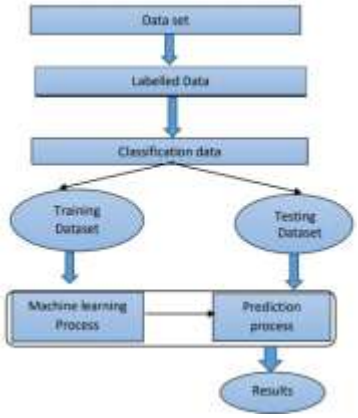
By predicting the index value of air and pollution in the air, it will be easier for the people to take precautious accordingly. By predicting the index value of air quality. This project prevent the air quality is good, moderate (or) unhealthy.

Existing System

The Existing system does not provide or predict the accurate values as they are using the unnecessary data in the process. In some other existing systems they use temperature to predict the air quality value which does not provide the accurate results as the temperature may change every minute. The state and municipal governments make many efforts to comprehend and forecast the air quality index with the goal of enhancing public health but the results are still not accurate.

Proposed System

In proposed system we use only the important data without considering the null values and outliers in the data. Additionally, by creating a data source for tiny communities which are typically overlooked in favor of big cities this will assist people in gathering information.



Algorithms Used

Linear Regression: Linear Regression is a supervised learning method, which is one of the algorithms in machine learning. It is used in performing a regression task. It performs predictive analysis task as it is a statistical method. In this project, linear regression algorithms show relation between the air quality index (dependent variable) and one or more variables that are independent that is nitrogen oxide, sulphur oxide values etc.

This air quality index result is based on the independent variables.

Figure no: 1

Decision Tree: A decision tree is a probability tree which makes you take a decision about some problems.
It is a supervised learning algorithm which is used to solve classification and regression problems .

In this project, the prediction is made after computing all features and predict the result whether the air quality is good, moderate or harmful for living beings.

Figure no: 2

Random Forest: Random Forest is a supervised learning technique under machine learning algorithms which can be used for regression as well as classification.
In this project, Random Forest combines multiple decision trees and their output to reach a single result. Additionally random forest often a high level of accuracy as it reduces the risk of overfitting and the required training time. It classifiers the pollution in air and predicts whether the air is safe for living beings or not.

Figure no: 3

Artificial Neural Network: ANN uses the brain as a basis to develop techniques that are used to create and process complex patterns and predicting problems.
Artificial Neural Network is used to make the computer where the computer can be able to understand and judge things and take accurate decisions that is similar to humans. Here, the ANN trains the system with the 80% of the air quality data and the decisions are made on the remaining 20% of testing data and such decisions are made like humans.

Figure no: 4

Xgboost: Xgboost is used to combine weak models and their prediction to stronger model and stronger prediction. Xgboost means "Extreme Gradient Boosting" as it has a greater capability to manage larger dataset it is one of the most used and widely used algorithms. It can be used as classification and regression algorithms. The main feature of xgboost is its ability to handle large datasets. Since the air quality data is large, the extension for this project is using xgboost.

It also efficiently handles the missing values and null values and also the datasets are directly used without performing any pre-processing techniques.
Though, if the pre processing is not done perfectly. The xgboost can handle and recover the data if there are any null valves, outliers or any missing values. It also has support to parallel processing which makes the system to train systems with huge datasets in short time.

Figure no: 5

Adaboost: Adaboost is also known as Adaptive boosting which is an ensemble modelling techniques used in machine learning for finding the best model.

Adaptive Boosting is widely known for its process in combining various "weak classifiers" into a "strong classifier"
In this project, if there are any weak prediction, there are togetherly combined to make a strong prediction on the dataset.

Figure no: 6

Modules

Data Exploration: Data exploration is the foremost thing in analyzing the data. Where the analysis use data visualization and some technique to understand data and its characterization such as quantity, type and accuracy. Here, In this project ,the data exploration checks for the data quality and its size to better understand the dataset.

Data Preprocessing: Data preprocessing is the process cleaning of learning data by removing null values, missing values and outliers. By doing this, it makes the project in increasing the accuracy of the machine learning model and as such efficiency.
A real world data may contain noisy data , outliers, missing values that cannot be directly used for algorithms in machine learning.
In this project, the data set is first preprocessed and the data is leaved by removing theses unnecessary data which makes the predictions accurate.
This preprocessed data will then be processed by applying machine learning algorithms and helps to predict accurate results.

Data Splitting: Data splitting is mostly used by people in order to train the system and test the system by dividing the data into training data and testing data.
The training data is used in training the system and the testing data is used in testing the system process and comparing.
Here, the air quality dataset is spitted into train data and test data. The training data trains the system, so that if there is any similar data provided by the user, the machine will compare the data and give the results. This obtained result is compared with training data to check whether the results are predicted accurately or not.

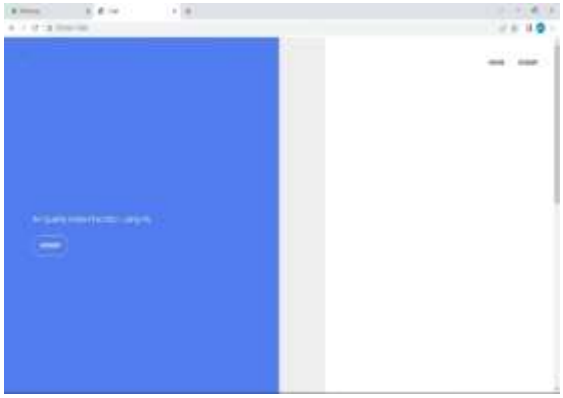
Model Generation: Model Generation is done to predict the accurate output.

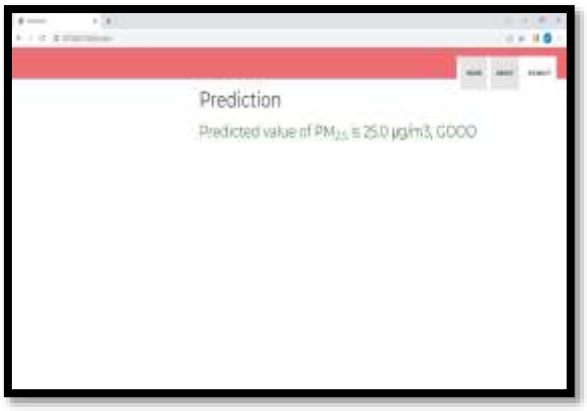
User Module: User Module is where, the user is able to give inputs to the system to predict the accurate results. Here the user is able to register and login to the system website and then give the data as input.

Data Prediction: Final prediction of the results will be displayed.

Implementation

In the implementation First python version 3.7 downloaded along with some functions which are needed for the project. Then by running the project on anaconda prompt. We will be displayed the link of the air quality website. By opening the website, the home page and then the login page will be displayed. Here are some of the screens regarding the implementation.





2. Conclusion

The proposed system will make it easier for regular people & meterologists to identify and forecast pollution levels and take appropriate action in response to those predictions. The algorithms used here are Linear Regression, Random Forest ,Decision Tree, ANN, Xgboost, Adaboost. Xgboost, ANN, Random Forest, Decision Tree gives more accurate air quality index values when compared to other algorithms**.**

Future Scope

Many other algorithms in machine learning and the meteorologists combined togetherly can make a collaboration to predict. The accurate air quality index values. Also, other gases can be identified in the future and can be used to predict more accurate air index value.

3. References

Database Systems design, Implementation, and Management, Peter Rob & Carlos Coronel 7th Edition.

Oracle for Professionals, The X Team, S.Shah and V. Shah, SPD.

Software Engineering, an Engineering approach- James F. Peters, Witold Pedrycz, John Wiley.

Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms,2018, International Journal of Emerging Trends &

Technology in Computer Science (IJETTCS).

Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura and Juan Felipe Franco, Machine Learning Techniques for PM10 Levels Forecast in Bogotá,2018, IEEE

Lidia Contreras Ochando, Cristina I. Font Julian, Francisco Contreras Ochando, Cesar Ferri, Airvlc: An application for real-time forecasting urban air pollution,2015, Proceedings of the 2nd International Workshop on Mining Urban Data, Lille, France.