IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

Detection and Prediction of Air Pollution Using Machine Learning Models

[1]Mr. Sandip Pawar, [1]Mr. Kunal Kamle, [1]Mr. Shubham Pawar, [1]Mr. Ganesh Kale, [2]Prof. Kavita Patil

UG Student, Department of Information Technology Engineering, ZCOER's, Pune, India[1]

Assistant Professor, Department of Information Technology Engineering, ZCOER's, Pune, India [2]

ABSTRACT:- in the populated and developing\countries, governments consider the regulation of air\as a major task. The burning of fossil fuels, traffic patterns, and industrial factors like power plant emissions all have a significant impact on air pollution. Particulate matter (PM 2.5) requires more attention among all the particulates that affect air quality. When it's level is high\sin the air, it produces major effects for people's health. Thus, it's crucial to control it by regularly monitoring its level in the atmosphere. In this study, logistic regression is used to determine whether or not a data sample is polluted. Based on the previous PM2.5 readings, autoregression is used to forecast future PM2.5 levels. Knowing the level of PM2.5 in upcoming months, weeks, or years allows us to lower it to below the dangerous range. On the basis of a data set that includes daily atmospheric conditions in a particular city, this system makes an effort to anticipate PM2.5 level and identify air quality.

KEYWORDS: Pollution detection, Pollution Prediction, RF Regression, RF Classifier, Autoregression

I. INTRODUCTION

There can be both naturally occurring and artificial particles. Examples include dust, ash, and sea spray. Particulate matter (including soot) is emitted into the air when solid or liquid fuels are burned for electricity generation, house heating, or in automobile engines. The diameter or width of the particles, or the size of the particles, varies. The mass of airborne particles having a diameter of less than 2.5 micrometres (m) per cubic metre of air is referred to as PM2.5[13]. Another name for it is PM2.5, or fine particulate matter (2.5 micrometres is one 400th of a millimetre). Fine particulate matter (PM2.5) is a significant contributor to the pollutant index because it poses a substantial risk to public health when airborne levels are high.

Human, animal and even plant health are affected by the chemicals and particles that make up air pollution. Humans are susceptible to various dangerous diseases caused by air pollution such as pneumonia, lung cancer, heart disease and bronchitis. Affected are smog, aerosol generation, hazy eyes, global warming, acid rain, premature death and other current environmental problems. All due to poor air quality. Scientists are aware that air pollution can have a negative impact on historic sites. The increase in greenhouse gases is caused by atmospheric emissions from factories and power plants, and exhaust gases from agriculture, etc. Greenhouse gases adversely affect the climate, which in turn affects the rate of plant development. The interaction between plants and soil is also affected by greenhouse gas and inorganic carbon emissions. Climate change will not only affect people and animals, but it will also have a major impact on agricultural factors and productivity.

The related repercussions also include financial losses. A measurement metric known as the Air Quality Index (AQI) has a direct connection to public health. An International Journal of Environmental Science and Technology 1 3 higher AQI level denotes a population at risk of more exposure. Therefore, the desire to accurately anticipate the AQI drove scientists to track and model air quality. With an increase in industrial and motorized activity, monitoring and forecasting AQI, particularly in metropolitan areas, has become an essential and difficult undertaking. Though the concentration of the deadliest pollutant, PM2.5, is shown to be multiplied in developing countries, the majority of air quality studies and research efforts focus on these nations. A few researchers attempted to do the study of Indian city air quality prediction. Following a review of the research, it was decided that study and prediction of the AQI for India would be a good way to close this gap.

1.1 **Work Related to This:-** Many past studies have looked into the possibility of utilising machine learning algorithms to forecast air quality. Objectives have occasionally been divided into distinct tiers in an effort to predict objectives. The case-based reasoning (CBR) method, a lazy learning approach, was used to elaborate on the effects on

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

air pollution only from meteorological parameters such as temperature, wind, precipitation, sun radiation, and humidity. It was also used to categorise air pollution into four levels (low, medium, high, and alarm). An example of inductive learning is this system. The fuzzy lattice neuro-computing classifier was used to predict and categorise O 3 concentrations into three levels: low, mid, and high. This classification was based on climatic parameters and other pollutants including SO2, NO, NO2, and others.

1.2 Problem Statement:- Air pollution is one such form that refers to the contamination of the air, irrespective of indoors or outdoors. It occurs when pollutants enter the atmosphere and make it difficult for plants, animals, and humans to survive as the air becomes dirty. The sustainment of all living things is due to a combination of gases that collectively form the atmosphere. The imbalance caused by the changes in these gases can be harmful to survival.

II. LITERATURE REVIEW

A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere Saurabh Kumar, Shweta Mishra, and Science Direct 2020

The air quality in Delhi, the capital of India, has been dangerous for a number of years. Asthma and other breathing-related issues have been identified in many persons. The main cause of this has been the atmosphere's high concentration of potentially fatal PM2.5 particles. In order to protect the locals from numerous health-related disorders, a decent model that can predict the concentration level of these dissolved particles may help to better equip the people with prevention and safety measures. By using time series analysis and regression using a variety of atmospheric and surface parameters, including wind speed, air temperature, pressure, and others, this work intends to forecast the PM2. 5 concentration levels in various parts of Delhi on an hourly basis. The Indian Meteorological Department has put up several weather monitoring stations throughout the city to provide the data for the analysis (IMD). It is suggested to utilize Extra-Trees regression and AdaBoost for further boosting in a regression model. Experimental work is conducted to compare with recent efforts, and the findings show that the suggested model is effective.

Regional air quality forecasting using spatiotemporal deep learning S Abirami, P Chitra, Journal of Cleaner Production, Elsevier 2021

Accelerated urbanization and industrialization have led to poor air quality, threatening lung health. Monitoring, modeling, and forecasting air quality can raise awareness and protect people from air pollution. Various air quality monitoring stations monitor a region's air quality. These stations' air quality data are highly dynamic, nonlinear, and stochastic. Deep learning methods that can abstract data well capture its spatiotemporal properties. DL-Air is a hierarchical deep learning model that forecasts air quality. The encoder encodes geographical data relationships. The second component, STAA-LSTM, identifies temporal relations and the amount of relationship between detected spatiotemporal relation and forecast. The STAA-LSTM predicts latent space spatiotemporal linkages. Decoder decodes these relations to get the actual forecast. The proposed methodology was used to anticipate Delhi's air quality. DL-Air has 30% lower RMSE and MAE, 37% lower AAD, 11% higher R2 and 8% greater AQI category prediction accuracy than baseline techniques. DL-predicted Air's performance is stable across Delhi's seasons.

Potential of machine learning for prediction of traffic related air pollution An Wang, Junshi Xu, Transportation Research Part D, Science Direct 2020

The technique of land use regression, sometimes known as LUR, has seen widespread application in the study of the spatial distribution of air pollution. However, it may be difficult to represent regional background and non-linear interactions using linear techniques. Recently, methods based on machine learning have been applied in the field of air quality prediction. This study investigates the boundaries of LUR approaches and the potential of two different machine learning models—Artificial Neural Networks (ANN) and gradient boost—by using data from a mobile campaign of fine particulate matter and black carbon in Toronto, Canada.

The data was collected during the course of the study. In addition to that, a moving camera was utilized to record the traffic in real time. The performance of the models that were built for fine particulate matter was superior to those that were developed for black carbon. Machine learning demonstrated higher performance compared to LUR for the same pollutants, suggesting that LUR performance may benefit from an understanding of how explanatory factors were portrayed in machine learning models. This study investigates the performance of various models in the context of how they capture the relationship between air quality and various predictors in order to shed light on the black-box nature of machine learning algorithms. Specifically, this research looks at how well these models capture the relationship between air quality and various predictors.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain) A. Suárez Sánchez, Elsevier 2021

This study's goal is to locally scale a support vector machine (SVM) technique-based regression model of air quality for the Spanish city of Avilés. Any substance that may result in or contribute to an increase in mortality or serious illness, or that may provide a current or potential risk to human health, is referred to as a hazardous air pollutant or toxic air contaminant. A highly nonlinear model of the air quality in the Avilés urban nucleus (Spain) based on SVM techniques is developed using experimental data of nitrogen oxides (NOx), carbon monoxide (CO), sulphur dioxide (SO2), ozone (O3), and dust (PM10) for the years 2006– 2008. The local scale estimation of the dependency between primary and secondary pollutants in the Avilés metropolitan area is one goal of this model. Finding the elements that have the biggest impact on air quality is a second goal with the intention of suggesting changes to lifestyle and health. In order to protect the health of healthy individuals, the National Ambient Air Quality Standards (NAAQS) of the United States determine the limit values of the major air pollutants. The term "criteria pollutants" applies to them. To accurately anticipate the relationships between the principal pollutants in the urban region of Avilés, this support vector regression model incorporates the key ideas of statistical learning theory. The conclusions of this work are then drawn utilizing the support vector regression (SVR) technique on the basis of these numerical computations.

A model for particulate matter (PM2.5) prediction for Delhi based on machine learning a approaches, Adil Masooda , Kafeel Ahmad, Science Direct 2020

The exposures have caused significant negative effects on people's health. As a result, it is essential to be able to make precise predictions regarding the magnitude of PM2.5 concentrations in order to formulate emission reduction plans for the control of air quality. In light of this, just a handful of machine learning strategies have been utilized to attempt to forecast the daily concentrations of PM2.5 in Delhi. On the basis of the inputs of a wide variety of meteorological and pollutant parameters corresponding to the period of two years spanning 2016-2018, two distinct models, namely SVM and ANN, were constructed. The performance evaluation of the models for predicting PM2.5 has been carried out, and the results have been talked about.

III. OBJECTIVES

- The primary objective is to develop a machine learning model that predict and detect air pollution or quality.
- This system attempts to predict PM2.5 level and detect air quality based on a data set consisting of daily atmospheric conditions in a specific city.
- The results show that machine learning models (logistic regression and linear regression) can be efficiently used to detect the quality of air and predict the level of PM2.5 in the future.
- The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that.

Algorithm steps:-

- Step 1: Read the dataset.
- Step 2: Data pre-processing and remove Nan values From dataset.
- Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.
- Step 4: Feature selection are applied for the proposed models.
- Step 5: Accuracy and performance metrics has been calculated to know the efficiency for algorithms.
- Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

IV. RESEARCH METHODOLOGY

Machine learning is defined as an automated process that mines patterns from a dataset. The art of developing and using models that make predictions based on patterns extracted from data form the past is called predictive data analytics.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

- **Random Forest Algorithm:-** Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification & Regression problems in ML.

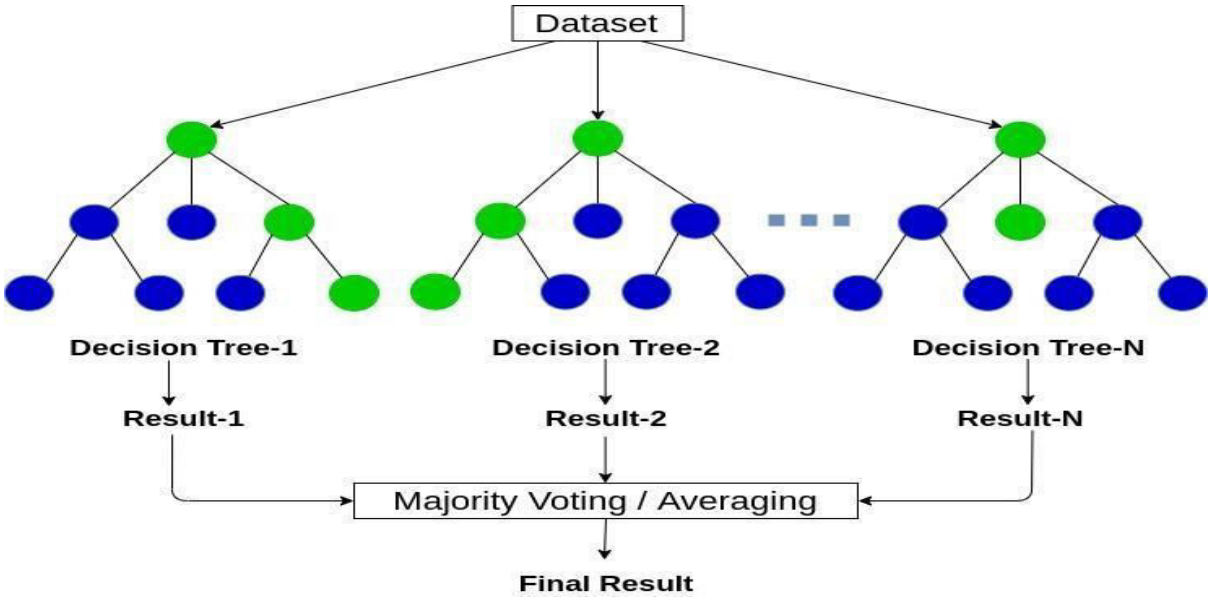


Figure1.1: Random Forest Algorithm

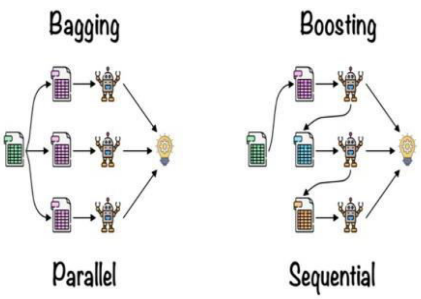
- **Assumptions for Random Forest:**

Because the random forest combines many trees to anticipate the class of the dataset, some decision trees may predict the proper output while others may not. But when all the trees are taken into account, they accurately predict the outcome. The following two suppositions for a better Random forest classifier are as a result: There must be some actual values in the dataset for the feature variable to forecast actual outcomes rather than a hypothetical consequence. There must be very minimal correlation between the forecasts of each tree. There must be some actual values in the dataset for the feature variable to forecast actual outcomes rather than a hypothetical consequence. The correlations between the predictions of each tree must be very small.ers.

- Working of Random Forest Algorithm:

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble imply means combining multiple models; thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:



1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest
2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOSTS, XG BOOST.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

- Training algorithm:- For Air pollution prediction, there are several different optimization algorithms used models development.

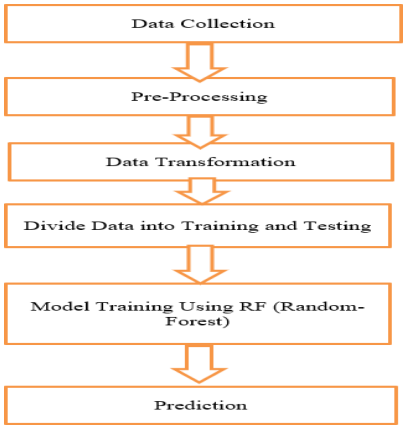


Fig 1.2: The Training Architecture

Data Preprocessing:-

1. Noise Removal:-

It is very important for making the data useful because noisy data can lead to poor results. In telecom dataset, there are a lot of missing values, incorrect values like ‘‘Null’’ and imbalance attributes in the dataset.

2. Feature selection:-

Feature Selection is a crucial step for selecting the relevant features from a dataset based on domain knowledge. A number of techniques exist in the literature for feature selection in the context of churn predictions.

II) Training the Network:-

The primary goal of training is to minimize an error using error techniques. In this Project We will training machine learning model using random forest algorithm. this is classifier and regression algorithm .

III) Testing:-

Through this process, the RF fined the predicted and compares it with the input values using data that was not used in training or validation process. At this stage no adjustment occurs to weights.

- **SYSTEM DESIGN:-**

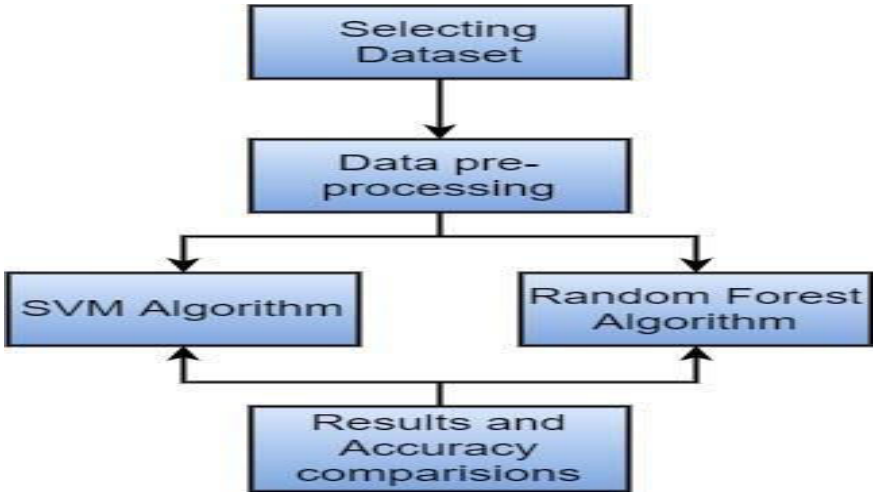


Fig. 1.3 The System Design

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

- Performance Evaluation Matrix:

In this study, the proposed churn prediction model is evaluated using accuracy, precision, and recall, f-measure, and ROC area. Equation 1 calculates the accuracy metric. It identifies a number of instances that were correctly classified.

Accuracy = (TP + TN) (TP + TN + FP + FN) …….(1)

Here ‘‘TN’’ stands for True Negative, ‘‘TP’’ stands for True Positive,
“FN’’ stands for False Negative and

‘‘FP’’ stands for False Positive.

TP Rate is also known as sensitivity. It tells us what portion of the data is correctly classified as positive.

For any classifier, the TP rate must be high. TP rate is calculated by using Equation 2.

TP Rate = True Positives Actual Positives (2)

FP Rate tells us which part of the data is incorrectly classified as positive. The result of the FP rate must be low for any classifier. It is calculated by using Equation3.

FP Rate = False Positives Actual Negatives (3)

Accuracy, also known as Positive Predictive Value (PPV), indicates which part of the prediction data is positive. It is calculated by using Equation 4.

Precision = True Positive (True Positive + False Positive)……(4)

The recall is another measure for completeness i.e. the true hit of the algorithm. It is the probability that all the relevant instances are selected by the system. The low value of recall means many false negatives. It is calculated by using Equation 5.

Recall = (True Positive) (True Positive + False Negative)……..(5)

The F-measure value is a trade-off between correctly classifying all the data points and ensuring that each class contains points of only one class. It is calculated by using Equation 6.

F − measure = 2 ∗ Precision ∗ Recall Precision + Recall ……(6)

ROC area denotes the average performance against all possible cost ratios between FP and FN. If the ROC area value is equal to 1.0, this is a perfect prediction. Similarly, the values 0.5, 0.6, 0.7,
0.8 and 0.9 represent random prediction, bad, moderate, good and superior respectively. Values of ROC areas other than these indicate something is wrong.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

- Result:-

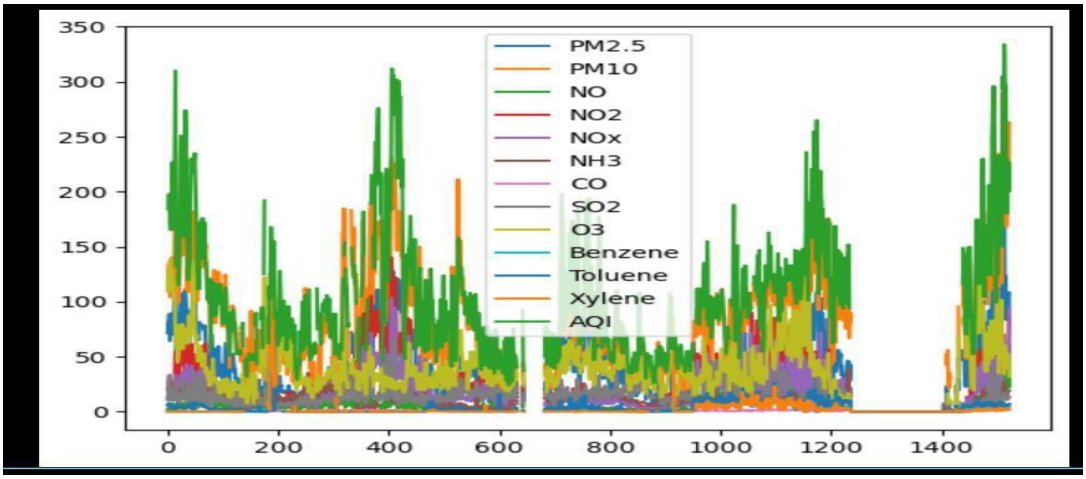


Figure1.4: Over All Data

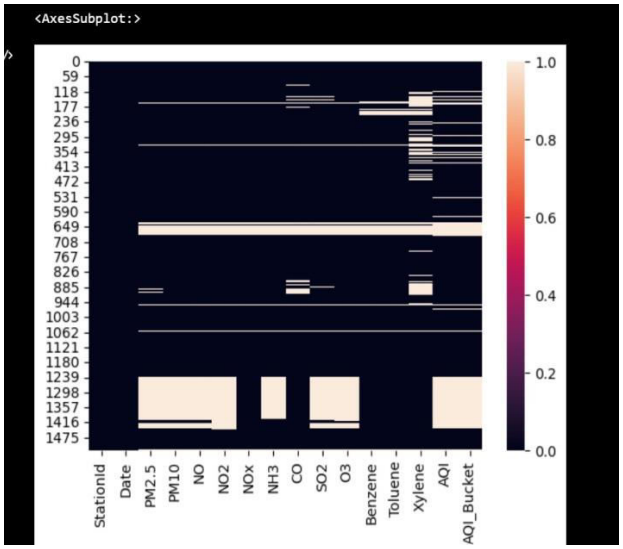
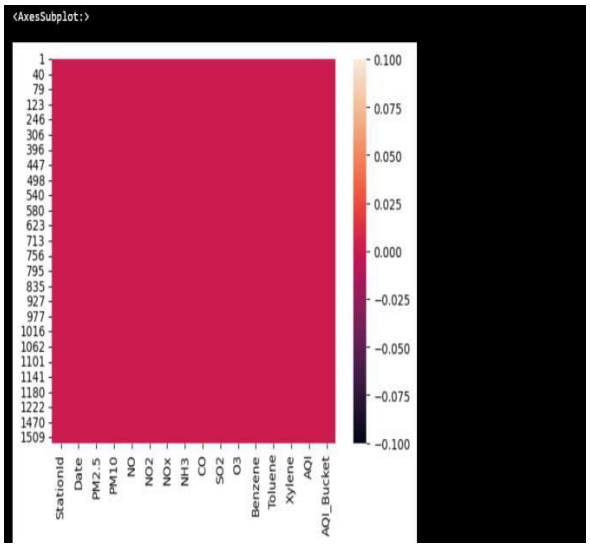


Figure1.5: Noisy



Data Figure1.6: Clean Data

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

GUI:-

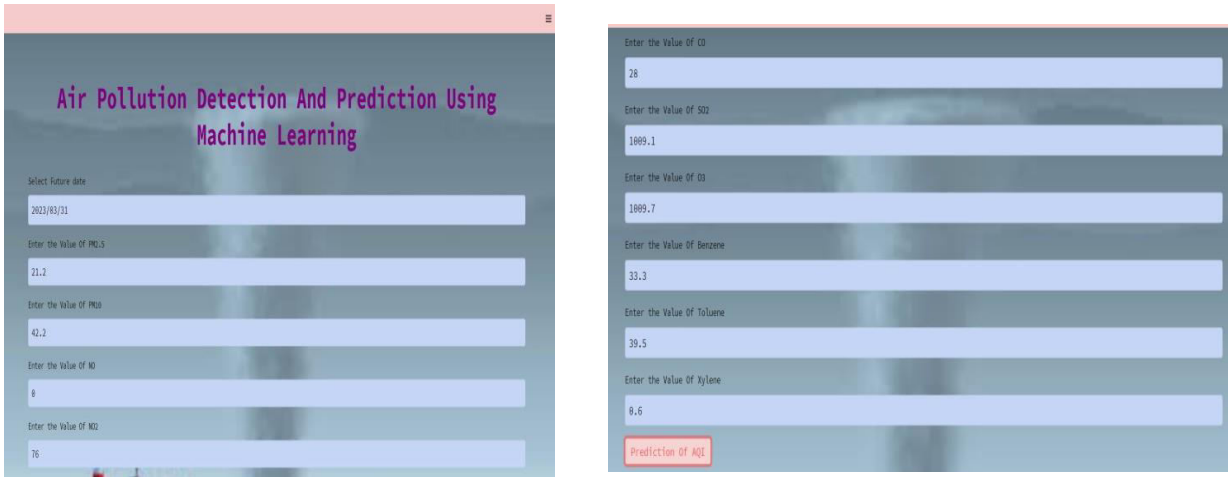
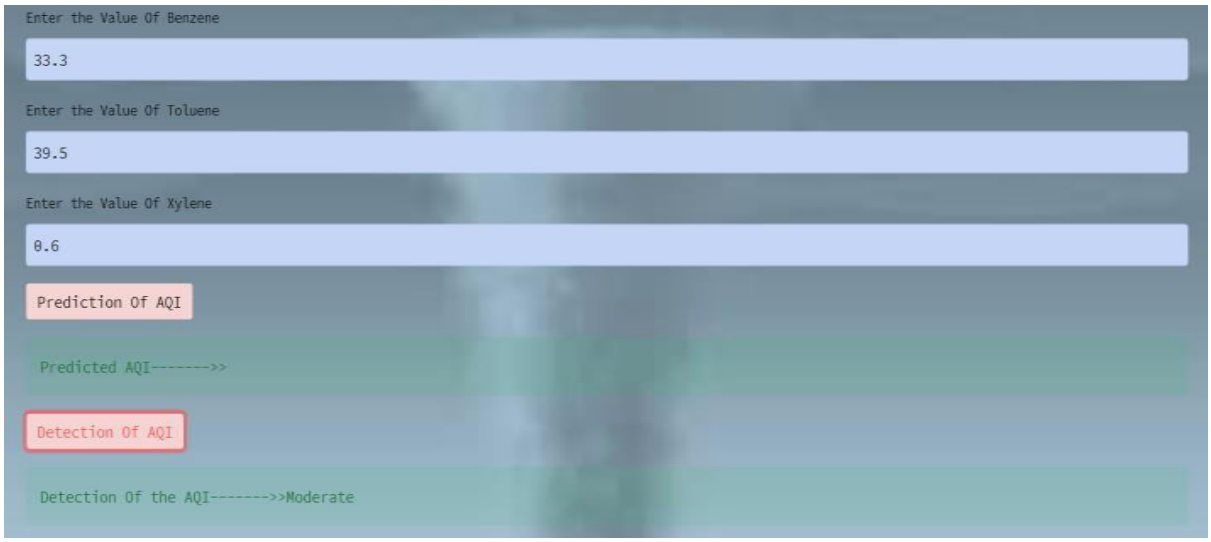


Figure1.7: User interface



REFERENCES

1. Alade IO, Rahman MAA, Saleh TA (2019b) Modeling and prediction of the specific heat capacity of Al2 O3/water nano fluids using hybrid genetic algorithm/support vector regression model. Nano-Struct Nano-Objects 17:103–111. https://doi.org/10. 1016/j.nanoso.2018.12.001
2. Al-Jamimi HA, Al-Azani S, Saleh TA (2018) Supervised machine learning techniques in the desulfurization of oil products for environmental protection: a review. Process Saf Environ Prot 120:57–71. https://doi.org/10.1016/j.psep.2018.08.021
3. Al-Jamimi HA, Bagudu A, Saleh TA (2019) an intelligent approach for the modeling and experimental optimization of molecular hydro-desulfurization over AlMoCoBi catalyst. J Mol Liq 278:376–384. (https://doi.org/10.1016/j.molliq.2018.12.144)
4. Ayturan YA, Ayturan ZC, Altun HO, Kongoli C, Tuncez FD, Dursun S, Ozturk A (2020) Short-term prediction of PM2.5 pollution with deep learning methods. Global NEST J 22(1):126–131

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205431 |

5. Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. https://doi.org/10. 1186/s12889-017-4914-3 Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learningbased prediction of air quality. Appl Sci 10(9151):1– 17. https:// doi.org/10.3390/app10249151

6. Fahad S, Sönmez O, Saud S, Wang D, Wu C, Adnan M, Turan, V (2021a) Plant growth regulators for climate-smart agriculture (1st ed.). CRC Press. https://doi.org/10.1201/9781003109013
7. Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221
8. Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ ICACCCN51052.2020.9362912
9. Air pollution and its impact on business: the silent pandemic. https://www.cleanairfund.org/wp-content/uploads/2021/04/01042021_Business-Cost-of-Air- Pollution_LongForm-Report.pdf Deshpande T (2021) India Has 9 Of World's 10 most-polluted cities, but few air quality monitors. India spend. https://www.india spend.com/pollution/india-has-9-of-worlds-10-most- pollutedcities-but-few-air-quality-monitors-792521
10. Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. 2003.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details