PREDICTION OF AIR QUALITY USING ML ALGORITHMS

[1]JONNALAGADDA VIJAYA SAI SUDHEER, [2]JONNALAGADDA GOPI CHAND,
[3]DIVI SUDHEER, [4]TANIKONDA DANIEL RAJU, [5]MS. FEMIMOL.R

[1,2,3,4]students, [5]assit Professor,
COMPUTER SCIENCE DEPARTMENT
Bharath Institute of Science &Technology affiliated to
Bharath Institute of Higher Education and Research,
Chennai, Tamil Nadu, India

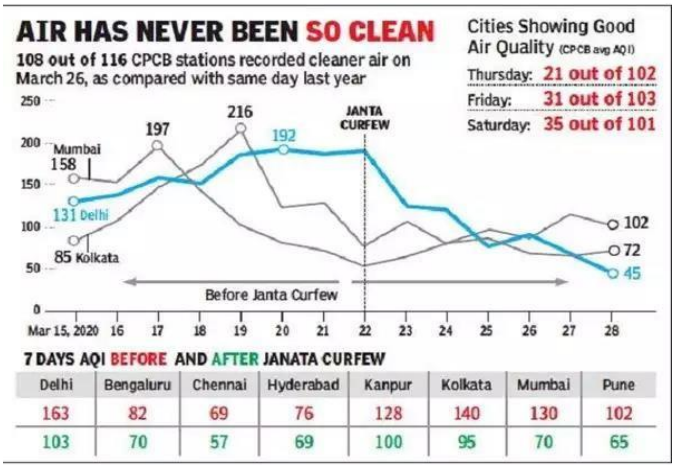
***Abstract* - Air pollution is a growing issue these days. It is necessary to monitor air quality and keep it under control for a better future and healthy living for all. Here we propose a prediction of air quality using ML. The system predicts air quality. Examining and protecting air quality has become one of the most essential activities for the government in many industrial and urban areas today. The meteorological and traffic factors, burning of fossil fuels, and industrial parameters play significant roles in air pollution. With this increasing air pollution. The deposition of this harmful gases in the air is affecting the quality of people's lives, especially in urban areas. Sulphur dioxide irritates the skin and mucous membranes of the eyes, nose, throat, and lungs. By using the machine learning regression models on AQI and classification models on AQI range, the system predicts the air quality in the respective locality.**

***Key Words*: Air pollution, AQI range, machine learning**

1. INTRODUCTION:

Nowadays the air condition is very polluted. In recent years, car emissions,chemicals from factories, smoke, and dust are everywhere Air pollution cannot be detected by human feelings. Air pollution may contain a lot of dangerous substances suchas ozone, particulate matter Sulphur dioxide, nitrogen dioxide, carbon monoxide, and lead. Humans need an atmosphere of air that is free from contaminants. This is very crucial for human life and health. Any change in the natural composition of air may causegrave harm to life forms on Earth. The effect of Air Pollution ranges from difficulty in breathing, coughing etc., polluted air can also impair visibility. Air Pollution is one of the serious and major environmental problem worldwide. People are suffering from health problems as a result of prolonged exposure to polluted environments. Health problems have been growing at faster rate especially in urban areas of developing countries where industrialization and growing number of vehicles leads to release of lotof gaseous pollutants.

Fig:1.1 Air Quality Index Analysis report



Delhi	Bengaluru	Chennai	Hyderabad	Kanpur	Kolkata	Mumbai	Pune
163	82	69	76	128	140	130	102
103	70	57	69	100	95	70	65

Harmful effects of pollution include mild allergic reactions such as irritation of the throat,eyes and nose as well as some serious problems like bronchitis, heart diseases, pneumonia, lung and aggravated asthma. According to a survey, due to air pollution 50,000 to 100,000 premature deaths per year occur in the U.S. alone. Whereas in EU number reaches to 300,000 and over 3,000,000 worldwide. NASA on 19th November 2021 released a report upon the issue of pollution in Delhi.

According to the news, in the year 2019,16.7 lakh people died due to polluted air in India. This is a big Disaster, which the administration is trying to fix, butthis work is not looking easy. The number of vehicles increased rapidly as soon as the lockdown lifted. Continued exposure to environments with poor air quality is a major public health concern in developed and developing countries. It is estimated that the pollutants responsible for poor air quality cause nearly 2.5 million premature deaths peryear world-wide. Significantly, around 1.5 million of these deaths are due to polluted indoor air, and it is suggested that poor indoor air quality may pose a significant health risk to more than half of the world's population

According to the World Health Organization (WHO), each year air pollution isresponsible for nearly seven million deaths around the globe. Nine out of ten human beings currently breathe air that exceeds the WHO's guideline limits for pollutants, withthose living in low- and middle-income countries suffering the most. Smog can ir ritate the eyes and throat and also damage the lungs, especially those of children, senior citizens, and people who work or exercise outdoors. It's even worse for people who haveasthma or allergies: these extra pollutants can intensify their symptoms and trigger asthma attacks. The tiniest airborne particles in soot, whether gaseous or solid, are especially dangerous because they can penetrate the lungs and bloodstream and worsen bronchitis, lead to heart attacks, and even hasten death. A number of air pollutants pose severe health risks and can sometimes be fatal even in amounts.

AQI	Remark	Color Code	Possible Health Impacts
0-50	Good		Minimal impact
51-100	Satisfactory		Minor breathing discomfort to sensitive people
101-200	Moderate		Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor		Breathing discomfort to most people on prolonged exposure
301-400	Very Poor		Respiratory illness on prolonged exposure
401-500	Severe		Affects healthy people and seriously impacts those with existing diseases

Fig:1.2 AQI Range Chart

Almost 200 of them are regulated by law; some of the most common are mercury, lead, dioxins, and benzene. Another category of toxic compounds, Polycyclic Aromatic Hydrocarbons (PAHs), are by-products of traffic exhaust and wildfire smoke.In large amounts they have been linked to eye and lung irritation, blood and liver issues,and even cancer. In one study, the children of mothers exposed to PAHs during pregnancy showed slower brain-processing speeds and more pronounced symptoms of ADHD.People experience a wide range of health effects from being exposed to air pollution. Effects can be broken down into short-term effects and long-term effects. Short-term effects, which are temporary, include illnesses such as pneumonia or bronchitis. They also include discomfort such as irritation to the nose, throat, eyes, or skin.

Air pollution can also cause headaches, dizziness, and nausea. Bad smells made by factories, garbage, or sewer systems are considered air pollution, too.These odors are less serious but still unpleasant.Long-term effects of air pollution can last for years or for an entire lifetime. They can even lead to a person's death. Long-term healtheffects from air pollution include heart disease, lung cancer, and respiratory diseases suchas emphysema.Air pollution can also cause long-term damage to people's nerves, brain, kidneys, liver, and other organs. Some scientists suspect air pollutants cause birth defects.Nearly 2.5 million people die worldwide each year from the effects of outdoor orindoor air pollution

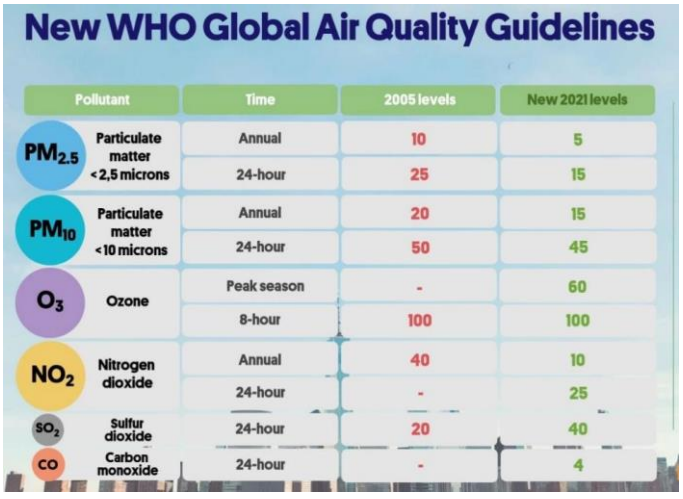


Fig:1.3 Annual AQI report

Accordingly, epidemiology studies have documented adverse respiratory andcardiovascular effects for populations living in close proximity to major roadways. Studies have shown specific adverse health effects of being exposed to Particulate Matter(PM), including heart diseases, cancer risk and adverse birth outcomes.Children, pregnant women, elderly and people with existing health issues are some of the vulnerable population groups to adverse traffic-related air pollution.Studies have shownan association between traffic-related air pollution and reduced fetal growth, preterm birth and post-term low birth weight and susceptibility to

asthma.We cannot completelyeradicate this problem but we may take several measures so that we can reduce air pollution that is caused by the vehicles.We Proposed this System so that it will help to monitor the air pollution that was caused by the Vehicles and to inform to higher authorities regarding to this.so that they will take appropriate measures.
We cannot completely eradicate this problem but we may take several measures sothat we can reduce air pollution that is caused by the vehicles. We Proposed this Systemso that it will predict the air quality. Air pollution is a growing issue these days. It is necessary to monitor air quality and keep it under control for a better future and healthyliving for all. Here we propose an air quality predicting system using machine learning.

2. LITERATURE SURVEY:

Tapiwa M, et.al. [1] machine-learning techniques such as Linear Regression (LR) models and Artificial Neural Networks (ANNs) to forecast or estimate air quality has been done before even though the focus has largely been on different pollutants and with different methodologies. Historical air quality data from the city of Johannesburg and the Vaal Triangle in South Africa was collected. The data was analysed and machine-learningtechniques were applied to the data to generate prediction models for ground level pollution.The results obtained are encouraging with the prediction accuracy ranging from fair to good

A. Masih [2] The major techniques applied for pollutant concentration estimation or forecasting; and whether these techniques were based on Linear Regression, Neural Network, Support Vector Machine or Ensemble learning algorithms. The results obtained suggest that, machine learning techniques are mainly conducted in continent Europe and America. Further more a factorial analysis named multi-component analysis performed showthat pollution estimation is generally performed by using ensemble learning and linear Regression based approaches, whereas, forecasting tasks tend to implement neural networksand support vector Machines based algorithm

C.Amuthadevi1, et.al. [3] Development of air quality monitoring (AQM) models using different machine learning approaches (i) detection and classification of gases by implementing classifiers such as ANNs, k-Nearest Neighbors (KNN) algorithm, hybrid support vector machine (HSVM) etc., and (ii) estimation of gas molecule concentrations byusing multivariable regression models such as linear regression, ANN, Deep learning Long- Short-Term Memory (DL-LSTM) .Many air monitoring places were established for monitoring and collecting the air-pollution related data for research. By periodic monitoring,the attentions of air pollutants in subsequent hour otherwise the subsequent-day could be predicted. Based on these predictions, government can be aware of it and control vehicles atthat location

Dr.D.J.Samatha,et.al. [4] ML models are able to achieve higher accuracies with large datasets, than classic statistical methods. Such models have long been used for AQI forecasting tasks. ML models are nonlinear, nonparametric in nature and hence are better able to handle the complexity of nonlinear elements like pollutant levels in the air. The results show that machine learning models (logistic regression and auto regression) can be efficiently used to detect the quality of air and predict the level of AQI in the future. The proposed system will help common people as well as those in the meteorological departmentto detect and predict pollution levels and take the necessary action in accordance with that.

SriramKrishna Yarragunta, et.al.[5] The supervised machine learning technique (SMLT) was used to analyse a dataset and capture multiple pieces of information, including variablerecognition, uni-variate analysis, bi-variate and multivariate analysis, missing value treatments, and data analysis. The agenda of our work is not only to bring awareness but alsoto minimize pollution through proper measures and ensure that the vehicles are emitting thepollutants within the range of regular pollution check. This can lead to a pollution free regionin the area.

Elia Georgiana Dragomir.[6] Air quality index prediction using K-nearest neighbour technique Data mining techniques, as artificial neural networks, genetic algorithms, decisiontrees, k -nearest neighbour, logistic regression have been successfully used in air quality prediction problems.In this paper, we have presented an experiment done in order todetermine the particularities of applying this technique for air quality analysis. Aiming at generating a prediction for the air quality index, training data that were collected in June 2009 were used as input data for the algorithm. The experimental results show that among the parameters that have been selected for this experiment, there is a strong correlation, and,therefore, these can be used in the forecasting procesess
Mahalingam U, et.al.[7] A machine learning model for air quality prediction for smart cities.This paper addresses the challenge of predicting the Air Quality Index (AQI), with the aim to minimize the pollution before it gets adverse, using two Machine Learning Algorithms: Neural Networks and Support Vector Machines. The air pollution databases were extractedfrom the Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate change, Government of India. The proposed Machine Learning (ML) model is promising in prediction context for the Delhi AQI. The results show improvement of the prediction accuracy and suggest that the model can be used in other smart cities as well.

Kingsy Grace R, et.al. [8] A comprehensive review of wireless sensor networks based air pollution monitoring systems. This paper presents a comparative study about the literature for air pollution monitoring systems based on the classification such as stationary air pollution monitoring systems, dynamic air pollution monitoring systems and pollution data analysis techniques. These pollution monitoring systems are compared based on the methodologies followed, microcontroller used, communication device used, pollutants analyzed using sensors, evaluation attributes, tested location and performance of the system.

Gallego E, et.al. [9] Outdoor air monitoring. Use of a semiconductor gas sensor for activatinga sampler during pollution episodes. Simultaneous 24 h and episodic samples were taken during 15 days in El Morell. Higher levels of VOCs, not specific compounds, trigger the activation of the sampler. Sensor interferences derived from relative humidity were not observed. The study proposes an expansion of the field of use of semiconductor gas sensors.Hence, these aspects validate the use of the evaluated sensor for its application for the activation of samplers in air quality evaluations when episodic event occur, an interesting and innovative technique.

3. SYSTEM ARCHITECTURE

The following flow chart demonstrates the architecture of proposed system.

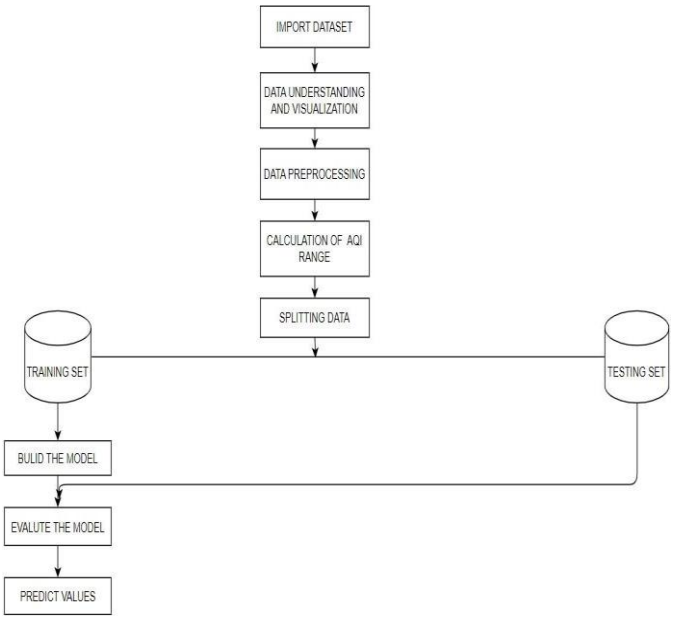


Fig :3.1 System Architecturee

3.1 DATA SOURCE:

To predict the air quality index of a particular region ,we need the date, AQI, AQI Range

As we taken the data from a particular location that is New-Delhi embassy road. The AQI formulae will be applied in order to calculate the AQI by using the KNN classification algorithm for a particular years. Several datasets will be imported inside the directory and null values will be set to the infinite data. The predicted and actual values will be represented using the Box-Plot analysis in order to remove the outliers.

The dateset consists of around 2843 records of the New-Delhi embassy road in India. This dateset consist of 3 attributes listed below.

1) Date
2) AQI
3) AQI RANGE

3.2 PREPROCESSING THE DATA:

A preliminary processing of data in order to prepare it for the primary processing or for further analysis.

Three steps in preprocessing of data for data cleaning 1.Missing value

2 Noisy Data

3 Removing unwanted data

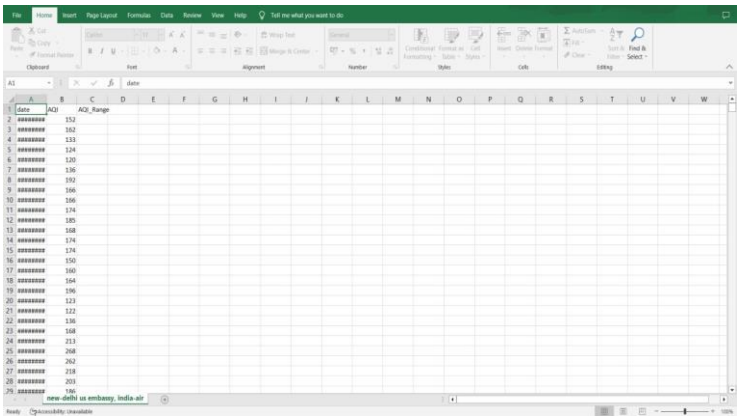
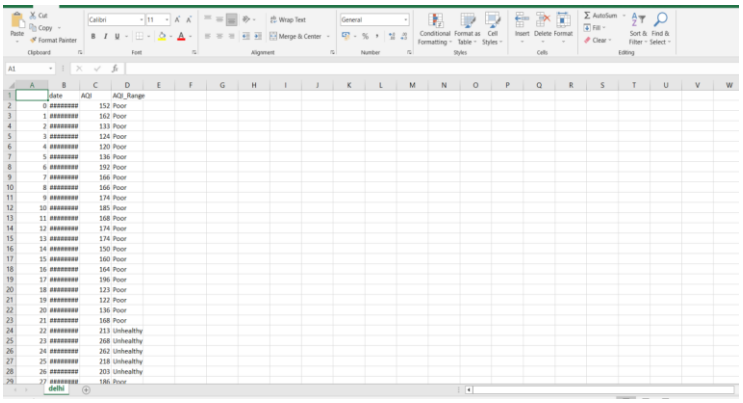


Fig :3.2 Before Pre-processing the Dataset



	A	B	C	D
1		date	AQI	AQI_Range
2	0	########	152	Poor
3	1	########	162	Poor
4	2	########	133	Poor
5	3	########	124	Poor
6	4	########	120	Poor
7	5	########	136	Poor
8	6	########	192	Poor
9	7	########	166	Poor
10	8	########	166	Poor
11	9	########	174	Poor
12	10	########	185	Poor
13	11	########	168	Poor
14	12	########	174	Poor
15	13	########	174	Poor
16	14	########	150	Poor
17	15	########	160	Poor
18	16	########	164	Poor
19	17	########	196	Poor
20	18	########	123	Poor
21	19	########	122	Poor
22	20	########	136	Poor
23	21	########	168	Poor
24	22	########	213	Unhealthy
25	23	########	268	Unhealthy
26	24	########	262	Unhealthy
27	25	########	218	Unhealthy
28	26	########	203	Unhealthy
29	27	########	186	Poor

Fig:3.3 After Pre-processing the Dataset

3.3 CALCULATION OF AQI Range

```
In [62]: def AQI_Range(x):
             if x<=50:
                 return "Good"
             elif x>50 and x<=100:
                 return "Moderate"
             elif x>100 and x<=200:
                 return "Poor"
             elif x>200 and x<=300:
                 return "Unhealthy"
             elif x>300 and x<=400:
                 return "Very unhealthy"
             elif x>400:
                 return "Hazardous"

         df['AQI_Range'] = df['AQI'] .apply(AQI_Range)
         df.head()
```

Fig:3.4 Calculation of AQI range

3.4 PREDICTION OF AIR QUALITY:

we spitted the data set into two parts of first 75% and rest 25% data into test and train datasets to identify the huge seasonal variations and trend. By using KNN classification algorithm, we can predict the air quality.

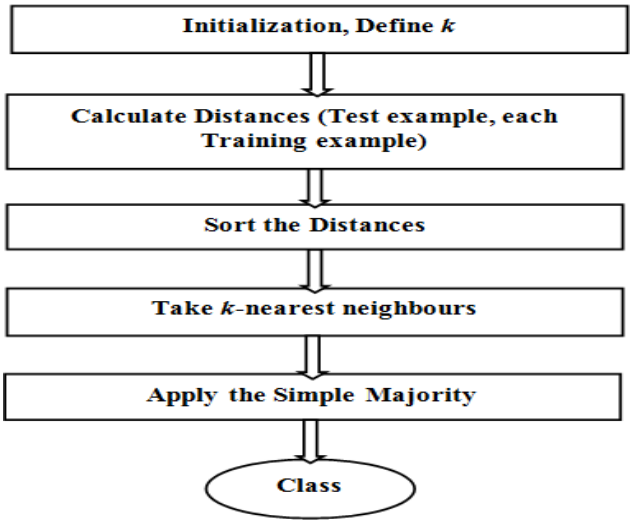


Fig:3.5 KNN Algorithm steps

3.5 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 and Regression problems in ML. Itis based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the mod

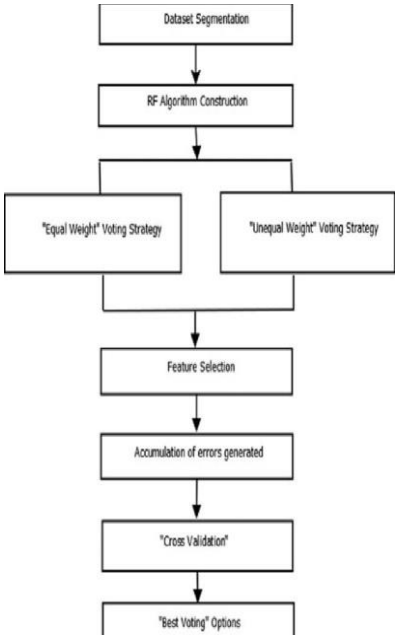


Fig:3.6 Random Forest Regression Algorithm

4.SYSTEM DESIGN

4.1 USE-CASE DIAGRAM:

A use case describes a function that a system performs to achieve the user's goal.
A use case must yield an observable result that is of value to the user of the system

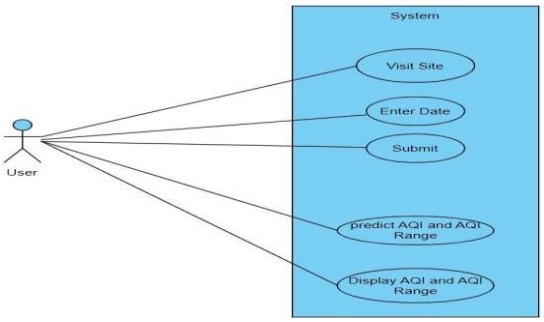


Fig:4.1 Use-Case Diagramm

4.2 CLASS DIAGRAM:

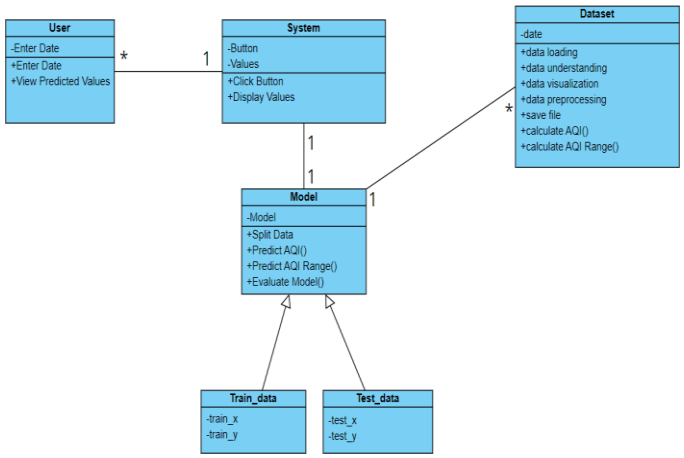


Fig:4.2 Class Diagram

4.3 ACTIVITY DIAGRAM:

Activity diagrams provide a way to model the workflow of a business process. An activity diagram is typically used for modeling the sequence of workflows

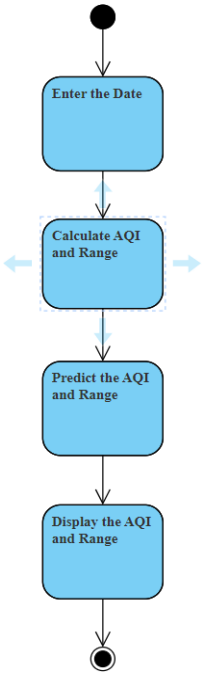


Fig:4.3 Activity Diagram

4.4 SEQUENCE DIAGRAM:

A Sequence diagram is a graphical view of a scenario that shows object interaction in a time- based sequence what happens first, what happens next. Sequence diagrams establish the roles of objects and help provide essential information to determine class responsibilities and interfac**es.**

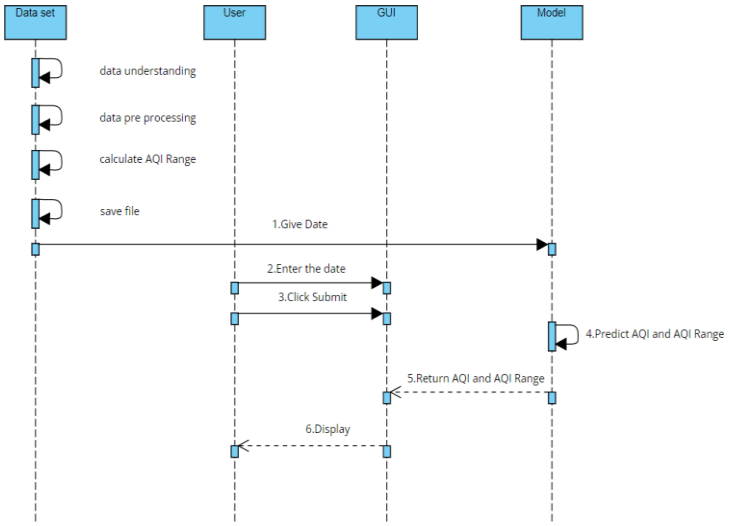


Fig:4.4 Sequence Diagram

5. METHODOLOGY

5.1 KNN ALGORITHM

Object model describes the structure of the system in terms of objects, attributes, associations, and operations. During requirements and analysis the object model starts as theanalysis object model and describes the application concepts relevant to the syste K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parmeteric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

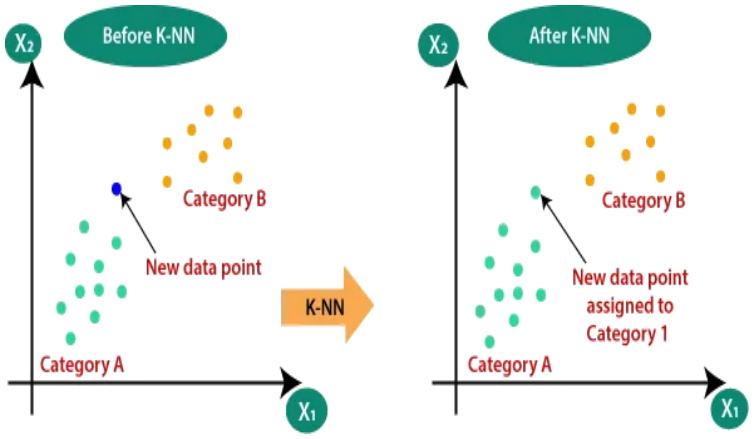


Fig:5.1 KNN Algorithm

5.2 RANDOM FOREST REGRESSION ALGORITHM:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the model.
Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.
The Working process can be explained in the below steps and diagram:
Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.
Random Forest is capable of performing both Classification and Regression tasks. It is capable of handling large datasets with high dimensionality.
It enhances the accuracy of the model and prevents the overfitting issu

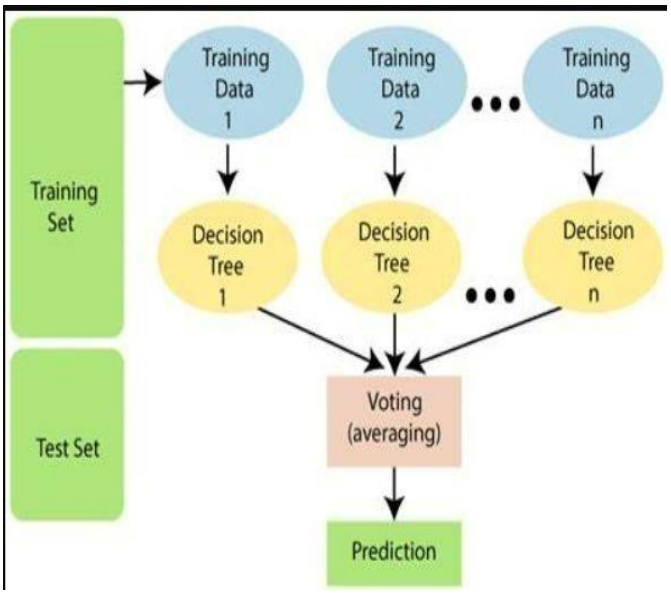


Fig:5.2 Random Forest Algorithm

5.3 CLASSIFICATION ALGORITHM

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.

In Classification, a program learns from the given data set or observations and then classifies new observation into a number of classes or groups.

Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc.

Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

The main goal of the Classification algorithm is to identify the category of a given data set, and these algorithms are mainly used to predict the output for the categorical data

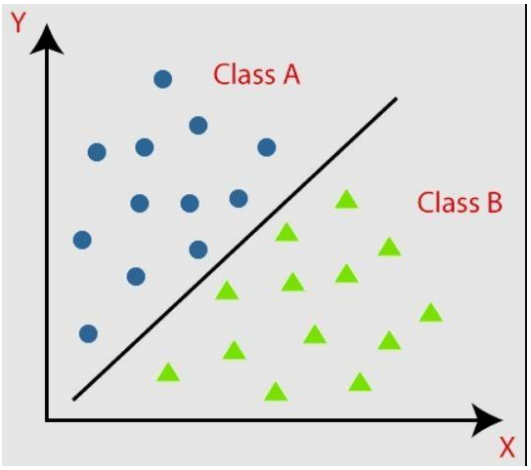


Fig:5.3 Classification Algorithm

6.OUTPUT SCREEN

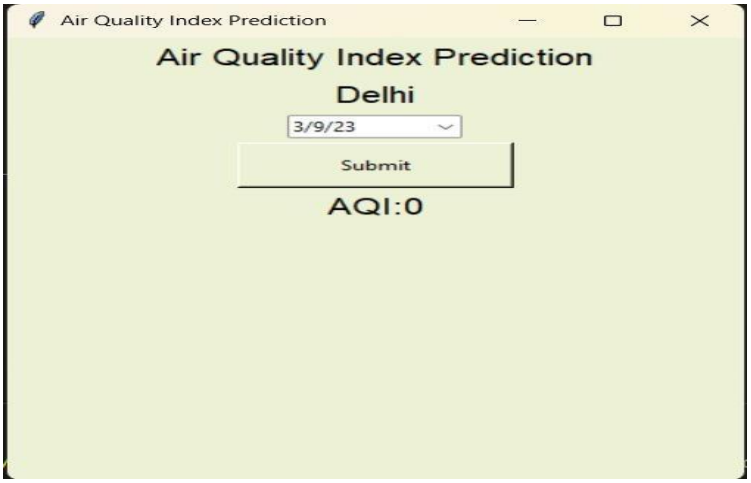


Fig:6.1 Before Giving the Date

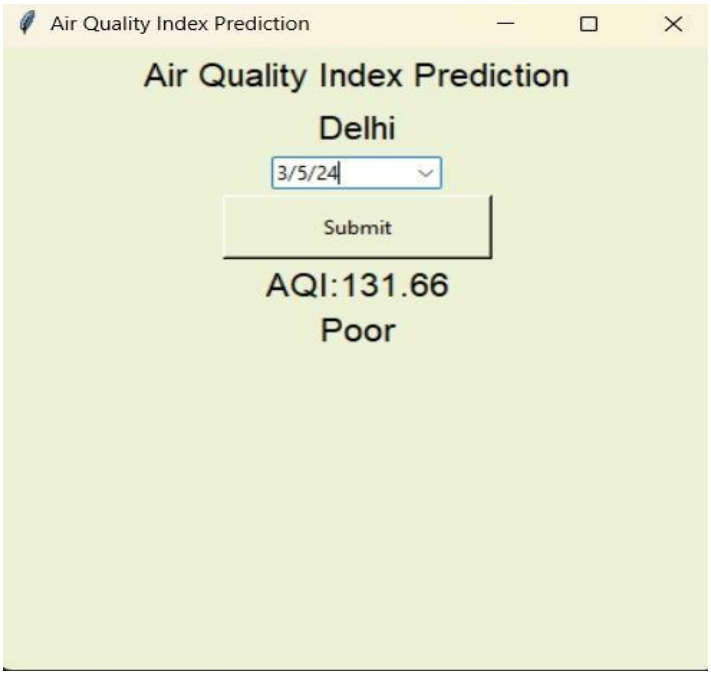


Fig:6.2 After Giving the Date

REFERENCES:

1. T M. Chiwewe and J. Ditsela, "Machine learning based estimation of Ozone using spatio- temporal data from air quality monitoring stations," 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), 2016, pp. 58-63, doi: 10.1109/INDIN.2016.7819134.[1]
2. Masih A. Machine learning algorithms in air quality modeling. Global Journal of Environmental Science and Management. 2019 Oct 1;5(4):515-34.[2]
3. Amuthadevi C, Vijayan DS, Ramachandran V. Development of air quality monitoring (AQM) models using different machine learning approaches. Journal of Ambient Intelligence and Humanized Computing. 2021 Jan 3:1-3.[3]
4. Naidu DS, Aruna R. Study of Air Quality Detection using Machine Learning Techniques. International Journal of Scientific and Academic Research (IJSAR), eISSN: 2583-0279. 2022 Sep 16;2(8):1-8.[4]
5. S. Yarragunta, M. A. Nabi, J. P and R. S, "Prediction of Air Pollutants Using Supervised Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1633-1640, doi: 10.1109/ICICCS51141.2021.9432078.[5]
6. Dragomir EG. Air quality index prediction using K-nearest neighbor technique. Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII. 2010 Jun 1;1(2010):103- 8.[6]
7. Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G. A machine learning model for air quality prediction for smart cities. In2019 International conference on wireless communications signal processing and networking (WiSPNET) 2019 Mar 21 (pp. 452-457). IEEE.[7]

8. Kingsy Grace R, Manju S. A comprehensive review of wireless sensor networks based air pollution monitoring systems. Wireless Personal Communications. 2019 Oct;108(4):2499- 515.[8]
9. Gallego E, Folch J, Teixidor P, Roca FJ, Perales JF. Outdoor air monitoring: Performance evaluation of a gas sensor to assess episodic nuisance/odorous events using active multi-sorbent bed tube sampling coupled to TD-GC/MS analysis. Science of the total environment. 2019 Dec 1;694:133752.[9]
10. Liu Y, Uthra DR. Bluetooth based smart home control and air monitoring system. International Journal of Advanced Research in Engineering and Technology (IJARET). 2020;11(5).[10]
11. Goh CC, Kamarudin LM, Zakaria A, Nishizaki H, Ramli N, Mao X, Syed Zakaria SM, Kanagaraj E, Abdull Sukor AS, Elham MF. Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. Sensors. 2021 Jul 21;21(15):4956.[11].