# AIR POLLUTION PREDICTION USING RANDOM FOREST AND DECISION TREE ALGORITHMS

For the partial fulfillment of the requirements for the award of the degree of

### BACHELOR OF TECHNOLOGY

In

### COMPUTER SCIENCE AND ENGINEERING

Submitted by

|  |  |
| --- | --- |
| **ALLAM VINEELA** | **319129510001** |
| **BONDA PRIYANKA** | **319129510009** |
| **BHADRI SAI GANESH** | **319129510008** |
| **JAMPANA SAI VARAMA** | **319129510021** |
| **PEDAGADA SIVATEJA** | **319129510032** |
| **GANDI SASI** | **319129510052** |



Under the Esteemed Guidance of

Dr.PRASADA RAO.M**,** Ph.D.

Assistant Professor

### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING WELLFARE INSTITUTE OF SCIENCE TECHNOLOGY AND MANAGEMENT

(Approved by AICTE, New Delhi& Affiliated to Andhra University) Pinagadi, Pendurthi, Visakhapatnam – 531173.

### WELLFARE INSTITUTE OF SCIENCE TECHNOLOGY AND MANAGEMENT

(Approved by AICTE and Affiliated to Andhra University) Pinagadi, Pendurthi, Visakhapatnam – 531173, Andhra Pradesh, India.



**CERTIFICATE**

This is to certify that the project report entitled “AIR POLLUTION PREDICTION USINGRANDOM FOREST AND DECISION TREE ALGORITHMS ” being submitted by

|  |  |
| --- | --- |
| **ALLAM VINEELA** | **319129510001** |
| **BONDA PRIYANKA** | **319129510009** |
| **BHADRI SAI GANESH** | **319129510008** |
| **JAMPANA SAI VARAMA** | **319129510021** |
| **PEDAGADA SIVATEJA** | **319129510032** |
| **GANDI SASI** | **319129510052** |

In partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering to the Andhra University is a record of bonafide project work carried out under the guidance of **Dr.PRASADA RAO.M,** . The results embodied in this project report have not been submitted to any university or Institute for the award of any degree.

### INTERNAL GUIDE HEAD OF THE DEPARTMENT

**Dr.PRASADA RAO.M,Ph.D Mrs.N.SUNEETHA,M.Tech(Ph.D)**

### Assistant Professor, Assistant professor,

**Department of CSE, Department of CSE,**

### WISTM Engg. College WISTM Engg. College

**EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

This satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all the effort with the success. I would like to thank **Dr.JOSHUA ARUMBAKAN, Ph.D., VICE PRINCIPAL** of Welfare institute of Science Technology and Management, for his kind cooperation. It gives me boundless pleasure to avail this opportunity to express my deep sense of gratitude and whole hearted thanks to Head of the Department **Mrs. N. SUNEETHA, Assistant Professor,** Department of Computer Science and Engineering and my project guide **,Dr.PRASADA RAO.M(Ph.D), M.Tech, Assistant professor,** Department of Computer Science and Engineering, for their valuable suggestions and co-operation during the development of this project. I am grateful to all my friends for their constructive suggestions to do this project in effective manner.

|  |  |
| --- | --- |
| **ALLAM VINEELA** | **319129510001** |
| **BONDA PRIYANKA** | **319129510009** |
| **BHADRI SAI GANESH** | **319129510008** |
| **JAMPANA SAI VARAMA** | **319129510021** |
| **PEDAGADA SIVATEJA** | **319129510032** |
| **GANDI SASI** | **319129510052** |

# DECLARATION

We hereby declare that the project report entitled “**AIR POLLUTION PREDICTION USING RANDOM FOREST AND DECISION TREE ALGORITHMS**” has been done

under the esteemed guidance of **Dr.PRASADA RAO.M(Ph.D), M.Tech, Assistant professor,** Department of Computer Science and Engineering is dissertation of my own work except where specifically ask to the contrary and his submitted to the Department of Computer Science and Engineering, Welfare Institute of Science Technology and Management for the partial fulfillment of the requirement for the award of B.Tech degree.

|  |  |
| --- | --- |
| **ALLAM VINEELA** | **319129510001** |
| **BONDA PRIYANKA** | **319129510009** |
| **BHADRI SAI GANESH** | **319129510008** |
| **JAMPANA SAI VARAMA** | **319129510021** |
| **PEDAGADA SIVATEJA** | **319129510032** |
| **GANDI SASI** | **319129510052** |

# ABSTRACT

The Project entitled “**AIR POLLUTION PREDICTION USING RANDOM FOREST AND DECISION TREE ALGORITHMS.**” Due to human

activities, industrialization and urbanization air is getting polluted. The major air pollutants involved are Carbon monoxide(co), Nitric oxide(NO),Nitrogen dioxide(N02)

,Ozone(O3).The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. In Earlier techniques such as Probability, Statistics etc. were used to predict the pollutants involved in air, but those methods are very complex to predict .So we are and going to use Data mining techniques for the better approach to predict the air pollution and quality of air which are need to predict air, relative humidity by considering various parameters. such as CO,NO2, NO, Temperature etc. So,to predict the pollution of air whether its increasing year by year or decreasing.We are going to use Decision tree and Random forest algorthims in Data mining techniques .

## TABLE OF CONTENTS

1. [INTRODUCTION 01-02](#_TOC_250026)
2. LITERATURE SURVEY 03
3. [SYSTEM ANALYSIS 04](#_TOC_250025)
   1. [Existing System 04](#_TOC_250024)
   2. [Proposed System 05](#_TOC_250023)
   3. [Feasibility Study 05-06](#_TOC_250022)
      1. [Economical Feasibility 06](#_TOC_250021)
      2. [Technical Feasibility 06](#_TOC_250020)
      3. [Social Feasibility 06](#_TOC_250019)
   4. System Requirement Specification 07
      1. [Functional Requirements 07](#_TOC_250018)
      2. [Non-Functional Requirements 08](#_TOC_250017)
   5. [Hardware Requirements 09](#_TOC_250016)
   6. [Software Requirements 10](#_TOC_250015)
4. [SYSTEM DESIGN](#_TOC_250014)
   1. [System Architecture 11-12](#_TOC_250013)
   2. [Modules 13-14](#_TOC_250012)
5. [UML Diagrams](#_TOC_250011)
   1. [USECASE Diagram 16](#_TOC_250010)
   2. [Class Diagram 17](#_TOC_250009)
   3. [Sequence Diagram 18](#_TOC_250008)
   4. [Activity Diagram 19](#_TOC_250007)
   5. [Component Diagram 20](#_TOC_250006)
   6. [Deployment Diagram 21](#_TOC_250005)
   7. [Database Design 22](#_TOC_250004)
6. [SYSTEM IMPLEMENTATION 23-34](#_TOC_250003)
7. OUTPUT SCREENSHOTS 35-39
8. [SYSTEM TESTING 40-42](#_TOC_250002)
9. CONCLUSION 43
10. [FUTURE SCOPE 44](#_TOC_250001)
11. [REFERENCES 45](#_TOC_250000)

## INTRODUCTION

Air pollution is one of the major hazards among the environmental pollution. As each of the living organism needs fresh and good quality air for every second. None of the living things can survive without such air. But because of automobiles**,** agricultural activities, factories and industries, mining activities, burning of fossil fuels our air is getting polluted. These activities spread sulphur dioxide, nitrogen dioxide, Ozone, carbon monoxide,particulate matter pollutants in our air which is harmful for all living organism. The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution.

The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea; Particulate matter with a diameter 2.5 micrometer or less than that affects more to human health.Measures must be taken to minimize air pollution in the environment. Air Quality Index(AQI), is used to measure the quality of air. Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air.

Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the pollutants needs vigorous analysis. Convolution Neural networks, Recursive Neural networks, Deep Learning, Machine learning algorithms assures in accomplishing the prediction of future AQI so that measures can be taken appropriately.Machine learning which comes under artificial intelligence has three kinds of learning algorithms, they are the Supervised Learning, Unsupervised learning, Reinforcement learning. In the proposed work we have used supervised learning approach. There are many algorithms under supervised learning algorithms such as Linear Regression, Nearest Neighbor, SVM, kernel SVM, Naive Bayes and Random Forest. Compared to all other

algorithms Random forest gives better results, so our approach selects Random Forest to predict the accurate air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. If the humidity is more, we feel much hotter because sweat will not evaporate into the atmosphere. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. Carbon Monoxide is produced due to the deficient Oxidization of propellant such as petroleum, gas, etc. Nitrogen Oxide is produced due to the ignition of thermal fuel; Carbon monoxide causes headaches, vomiting; Benzene is produced due to smoking, it causes respiratory problems; Nitrogen oxides causes dizziness, nausea;

## LITERATURE REVIEW

In this the author Pasupuleti et al. (2020) the decision tree, linear regression, random forest. Major air pollutants are taken and meteorological condition are taken using Arduino Platform. Random forest gives better accurate result due to over fitting that reduces errors But drawback is Random forest uses more memory and high cost.

Haitian Jung & Yingchun Wang(2020) had predicted the air quality index using XG Boost.It uses the weak classifier and shortcoming of the previous weak classifier to form a strong classifier thus reducing the error between predicted and actual values .It uses the K- cross validation .The mean absolute error and coefficient of determination is determined to predict the difference between actual and predicted value.The drawback faced is that it takes the previous value and is affected by outlier unwanted pollutant in the air.

Desislava Ivanova and Angel Elenkov (2019) had used Raspberry Pi platform with Multi layer perceptron algorithm of machine learning to predict the air pollutant accurately.Multi layer perceptron overcome problem of classification which is used for discrete values and regression which is used for continuous value. In this author uses discrete values and had used multi layer perceptron using back propagation and therefore input did not pass to the activation function resulting in 0 or 1indicating how big the difference between the predicted and actual value. The coefficient of determination(R2) obtained is better but more can be improved by increment feeding.

In Elseviere 2018,The tittle is Forecasting air pollution load in Delhi using data analysis tools.The methods used are Time series regression.With limitations are Here we are using time series with regression.so again regression is limited to the linear relationship.It is easily affected by outliers.

Soubhik et al. (2018) had compared various algorithms like Linear regression, Neural network regression, Lasso regression, Elastic Net regression, Decision Forest, Extra trees, Boosted decision tree, XG Boost, KNN, and Ridge regression to predict air pollutant level.Better accuracy is obtained by extra trees because features are arranged in decreasing order of importance to predict the upcoming value.

## SYSTEM ANALYSIS

### EXISTING SYSTEM:

In Earlier techniques such as Probability, Statistics etc. were used to predict the pollutants involved in air, but those methods are very complex to predict .sometimes the data is manipulated easily and data is not shown accurately .So we are going to use Data mining techniques and Decision tree , Random forest algorithms for the better approach to predict the air pollution .

### Disadvantages:

* + - It is time consuming.
    - It consumes lot of manpower to better results
    - Percentage of accuracy is less.
    - Reports take time to produce.

### PROPOSED SYSTEM

Conventional techniques are too complex and broad to handle and analyze the large amounts of information gathered for heart disease prediction. Text mining provides the tools and procedures for turning these mountains of data into data that can be accessed for decision- making. Using data mining techniques accelerates and improves sickness prediction. In this post, we will look at a few studies that used one or more data mining algorithms to predict heart illness. The outcome of implementing algorithms in one research [10] and another [6] is nearly 97%. to assure that recommendations created that used algorithms for data mining deliver accurate findings. Because once data mining software disease management records, the findings may be just as precise like when cardiovascular disease is identified.

### ADVANTAGES OF THE PROPOSED SYSTEM

Using data mining techniques accelerates and increases sickness prediction accuracy.

1. In this study, we look at a few papers that use one or more data mining algorithms to predict cardiac illness.
2. In one article [10] and in another [6,] the result of applying neural networks is about 97%. to verify that predictions generated using a data mining algorithm deliver reliable results.
3. Applying data mining techniques to heart disease treatment data can yield findings that are as reliable as those obtained when diagnosing heart disease

## FEASIBILITY STUDY

The viability of the project is reviewed at this stage, and a business proposal with a very basic project plan and some cost forecasts are submitted. During system analysis, the feasibility of the proposed system must be assessed. This ensures that the offered solution will not impose a burden on your organization. Understanding the key technical parameters is critical for the feasibility analysis. The three factors listed below are critical to the site inquiry:

Three key considerations involved in the feasibility analysis are

* ECONOMICAL FEASIBILITY
* TECHNICAL FEASIBILITY
* SOCIAL FEASIBILITY

### ECONOMICAL FEASIBILITY

Additional investigation is underway to figure out the economic impact of the system on the firm. The corporation has a limited budget for the research and development of the system. Evidence must be provided to back up the expenses. Because the majority of the technologies used are freely available, the system design stayed below budget. Only specialist items were required to be purchased.

### TECHNICAL FEASIBILITY

Additional research is being carried out to investigate the service's specifications or practicality. Any technology created should not impose an undue strain on the available technical resources. As a result, the available technical capabilities will be severely limited. In the end, the consumer will be confronted with unrealistic expectations. The built system should be in high demand since its implementation needs either few or no revisions.

### SOCIAL FEASIBILITY

The author’s purpose is to study how much the user accepts the system. The above page includes the steps required for the user to successfully run the system. Instead of being afraid of the system, the user must recognize it as necessary. The only things that influence the amount of user acceptance are the strategies employed to enlighten and familiarise the user with the technology. He has to be more comfortable because he will be the program's ultimate user and he will be able to provide some useful feedback.

## SYSTEM RQUIREMENT SPECIFICATIONS

### 3.4.1. FUNCTIONAL REQUIREMENTS

In computer science, acceptance criteria specify the purpose of a software system or one of its components. A function's description provides a list of its inputs, behaviour, and consequences. Functional specifications describe what a system is intended to perform and may contain computations, architectural details, data management, and management, as well as specifics. Use stories to describe how the system will use the functionality as well as provide all of the behavioral needs. Requirements are typically phrased as "organization shall perform" (program shall do demand). The system design includes explicit information about the approach for implementing design criteria.

In requirement analysis, functional specifications specify specified operating outputs. The architecture of an application is determined by functional requirements. A requirements analyst creates use cases after gathering and confirming the list of functional requirements. The hierarchy of significance for acceptance criteria is user or partner request -> innovation -

> application case -> management rule.

The functional requirements determine an application's architecture. A requirements analyst receives and validates a set of functional requirements before creating use cases. Functional demands for the project might include technical details, data processing, and other specialized features. The user receiving the information is the project's primary goal. The following are our system's characteristics and capabilities:

* We provide all available information about heart disease.
* Each operator might input his values into the same system to predict whether heart disease will be present.
* For classification, we employ a range of machine-learning methods.
* Using the optimal classification strategy that would have been determined through comparison, the recommended application may correctly forecast heart disease.

### 3.4.2 NON-FUNCTIONAL REQUIREMENTS

The system's general features or qualities are defined by non-functional rules. Non- functional requirements place restrictions on the product being developed and the development process, and they specify demands that are outside the product's scope of control. NFR includes, among others, standards for performance, dependability, usefulness, and safety. Project concerns about costs, time, and schedule are occasionally viewed as non-functional needs.

### Performance requirements

The necessary resource requirements, as well as any throughput, transaction rates, or other performance-related details. Before attempting to train the system with that dataset, the user will attempt to get all the data from the KAGGLE website for this project.

### Modifiability

limitations regarding how many-hour shifts any technological improvements should take. Frequent personnel effort is one sign (person-months).

### Portability

We may share the application with a new demography platform thanks to the same work. Person months or the percentage of modules that need to be changed are two of the most often measured statistics.

### Reliability

Limitations on how frequently the program breaks down A frequent unit of measurement is the probability of failure (mean time between failures). Failure must have a specific definition. Furthermore, availability and reliability are two very different types of criteria and should never be confounded. Include information about the consequences of software failure, methods for preventing failure, a method for detecting errors, and a method for detecting and fixing errors.

### Security

A limitation or requirement relating to the safety of your system and the data on it. Information measurement might be stated in a variety of ways to circumvent the system (power, quality of play, and length). Never discuss remedies in a document detailing needs (such as passwords).

### Usability

stipulations of how difficult it will be to utilize and master the technology. Often, the parameters are given in terms of learning time or other comparable units.

### Legal

There may be legal concerns around data privacy, intellectual property rights, the export of forbidden technology, and other things.

## HARDWARE REQUIREMENTS

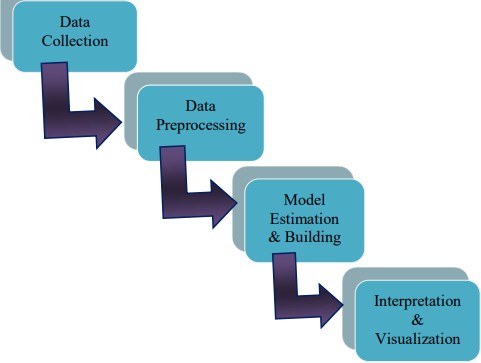
* Processor : Core I3
* RAM : 4 GB(Min)
* Hard Disk : 100 GB

## SOFTWARE REQUIREMENTS

* Operating System : Windows7 (Min)
* Coding Language : Python( panda, NumPy)
* Dataset : Random forest, Decision tree
* [https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set.](https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set) Dataset.

# SYSTEM DESIGN

## SYSTEM ARCHITECTURE



Trained data set

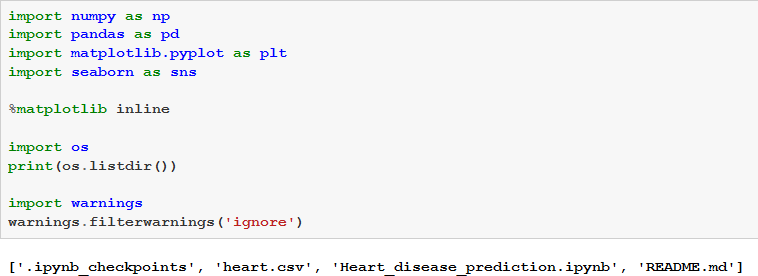
Fig.4.1 System Architecture

## MODULES

The theoretical design is converted into a programmed method during the implementation phase. The application will be divided into several modules during this stage, which will then be programmed for deployment. The front end of the program makes use of Google Collaboratory, while the dataset for UCI Heart Patients Records was utilized as the back-end database. The current application is being implemented in this case using Python. The application's main five modules are listed below:

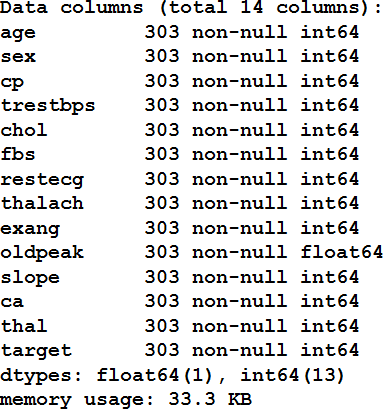
* + 1. Import Necessary Libraries
    2. Load Dataset Module
    3. Pre-processing Data
    4. Train the Model
    5. **IMPORT NECESSARY LIBRARIES**

To build the model, we must first import all the necessary libraries into this module. Here, we attempt to use every package that is accessible for effectively transforming data. We try to import the NumPy module and use the Matplot to plot the data in graphs and charts since the system can easily recognize the numerical values in the data.



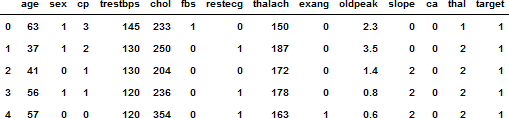
### LOAD DATASET MODULE

In this module, we try to load the dataset that was downloaded or collected from the UCI repository. Each Characterstic comprises data that was tested and gathered based on the unique patient id.



### DATA PRE-PROCESSING MODULE

Here in this section, we try to pre-process the input dataset and find out if there are anymissing values or incomplete data present in the dataset. If there is any such data present in the dataset, the application will ignore those values and load only valid rows which have all the valid inputs.



### TRAIN THE MODEL USING ML SEVERAL ML ALGORITHMS

To find the best algorithms for accurately and efficiently identifying and categorizing the input dataset, we also try to train the current model using several ML classification techniques on a particular dataset. Here, we attempt to use the provided dataset with the methods listed below, including

* + - 1. logistic regression
      2. Simple Bayes
      3. Support Vector Machine
      4. Nearest Neighbor
      5. Decision Tree
      6. Random Forest
      7. XG Boost
      8. Neural Networks

### PERFORMANCE ANALYSIS MODULE

We compare each classification technique in this module on a particular input dataset to see which one is best at generating accurate results. Finally, we will select the approach that offers the quickest and most accurate results. As we can see, Random Forest performs better in terms of accuracy than other machine learning algorithms.

## UML DIAGRAMS

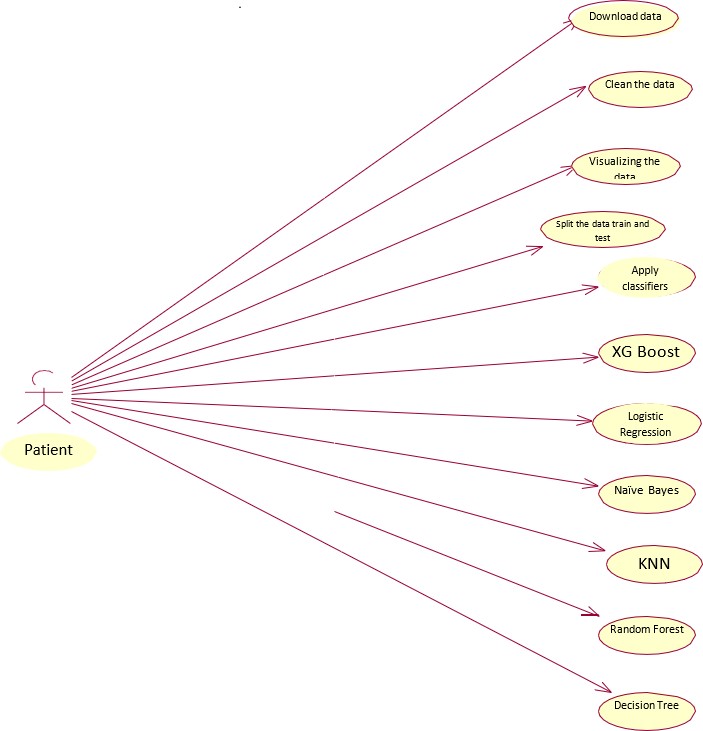
No one diagram can fully describe all of a System's components, according to a core premise of UML. UML is made up of nine diagrams that may be used to represent a system at different points in its software life cycle.

There are two types of components in software packages: structural or "static" components and behavioral or "dynamic" components. In addition to these two, a software system contains a third attribute that pertains to implementation. Before assigning each of these characteristics to one of the three categories for UML diagrams, let's quickly review each one.

* Use case diagram
* Class diagram
* Object diagram
* State diagram
* Activity diagram
* Sequence diagram
* Collaboration diagram
* Component diagram
* Deployment diagram

### USE CASE DIAGRAM

The use case diagram illustrates the key elements and processes that make up the System. The methods are known as "use cases," and the essential elements are known as "actors." The use case diagram displays the actors involved in each use case.



Data set

Fig.5.1 Use Case Diagram Admin/User

use instances A use case is shown as a horizontal ellipse and shows a sequence of actions that provide an actor with a measurable benefit. Actors are any individual, team, or external system that engages in one or more interactions with the system.

### CLASS DIAGRAM

The class diagram, which also specifies the complex architecture of the system, aids in the improvement of the use case diagram. The class diagram is used to divide the players in the use case diagram into several related classes. There may be an "a" or "has-a" connection between the classes.

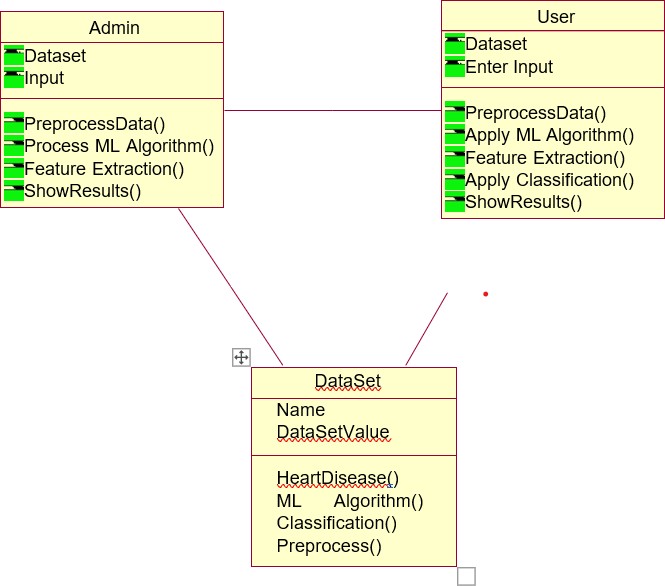


Fig..5.2 Class Diagram

### 5..3 SEQUENCE DIAGRAM

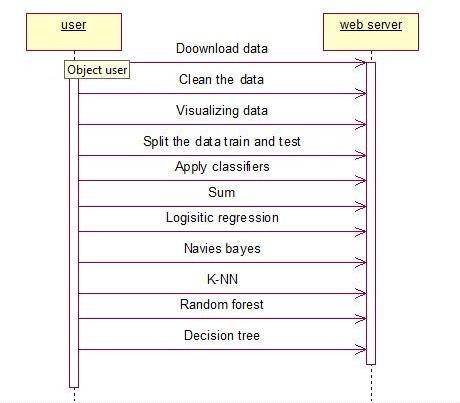
The use case diagram may be improved with the help of the class diagram, which also outlines the intricate system architecture. The actors in the use case diagram are categorized into distinct classes using the class diagram. The classes may be connected by an "is-a" or "has-a" relationship.

fig. 5.3 Sequence Diagram

### ACTIVITY DIAGRAM

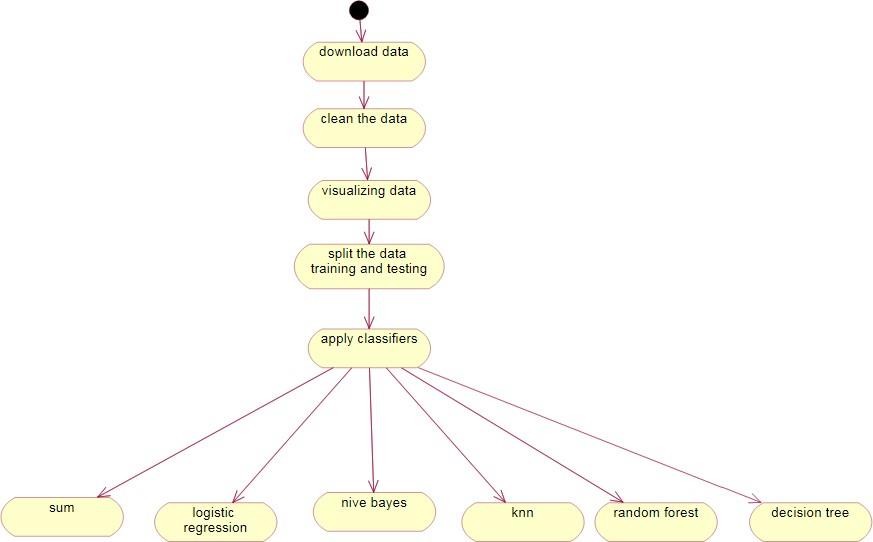


Fig. 5.4 Activity Diagram

The activity diagram displays the system's process flows. Similar elements seen in a state diagram, such as activities, actions, transitions, starting and ending states, and guard conditions, also appear in activity diagrams.

### COMPONENT DIAGRAM

This diagram's flow depicts how a group of components are organized and interdependent. It displays a system's static implementation view.



|  |  |  |
| --- | --- | --- |
|  |  | |
|  |  |  |
|  |  | |
|  |  |  |
|  |  | |

Fig.5.5 Component Diagram

### DEPLOYMENT DIAGRAM

The deployment diagram shows how the runtime components of the application are configured. Unquestionably, the best time to use this visual is when a system is finished and ready for deployment. The word "deployment" itself contains the graphic's goal. Deployment diagrams are used to show the physical components that receive software components. Component diagrams and deployment diagrams have many similarities. The following objectives are achieved via deployment diagrams:

* Display the hardware topology of a system.
* Describe the physical components used in the deployment of software components.
* Describe the processing runtime nodes.

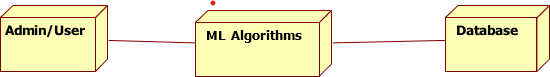


Fig.5.6 Deployment Diagram

## DATABASE DESIGN

The database needs to be properly developed before the data can be saved there since there is so much data associated with the suggested system. Users have freedom in data storage, retrieval, and information production thanks to database management systems. A bridge connects the DBMS and application programs and determines which data are writing data to magnetic storage media is controlled by the computer's operating system, which also controls how they are processed

### 4.4.1 Normalization

The normalization theory is based on the notion of normal forms. A relation is said to be in a certain normal form if it satisfies a specific set of conditions.

### First Normal form

A relation R would be in first normal form only if and only if all of the underlying domains contained only atomic values.

### Second Normal form

A connection is said to be in the second normal form when every non-key attribute is entirely dependent on the primary key and the relationship is the in first normal form.

### Third Normal form

A relation is said to be in normalized form if and only if it is in second normal form and none of the non-key attributes transitively rely on the primary key.

### Boyce and Codd Normal Form (BCNF)

Boyce and Codd A more complex version of the Normal Form is the Fourth Normal form. This form addresses a certain class of anomaly that 3NF does not address. If there aren't any candidate keys that overlap, a 3NF table is said to be in BCNF. The following conditions must be satisfied for a table to be included in Complementary distribution:

* For any functional dependency, R must be in third normal form and X must be a super Key (X Y).

### Fourth Normal Form (4NF)

When Multi-valued Dependency exists in any connection, the fourth Normal Form is involved. The following two requirements must be met for a table to meet the Fourth Normal Form:

# 6 . SYSTEM IMPLEMENTATION

## 6.1 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI). Understanding data structure and incorporating it into models that humans can understand and utilize is the core goal of machine learning. Machine learning is unique from traditional computational techniques even though it is a subfield of computer science. Algorithms are sets of purposefully crafted instructions that computers use to do computations or address issues in traditional computing. Instead, computers may train on data inputs and utilize statistical analysis to create numbers that fall within a specific range thanks to machine learning algorithms. Machine learning is a branch of science that examines how computers learn from experience. For many scientists, the phrases "machine learning" and "artificial intelligence" are synonymous since the capacity to learn is the basic attribute of an entity considered to as intelligent in the broadest sense of the word. The objective of machine learning is to create computer systems that are flexible and can learn from their experiences. Two types are available.

* + **Supervised Learning** In reinforcement methods, the system must inductively "learn" a function known as the target function, which is an expression of a model describing the data. The value of a variable, known as the dependent variable or output variable, is predicted using a collection of variables, referred to as independent variables, input variables, characteristics, or features. The collection of possible input values for the function, or its domain, is referred to as instances. Each circumstance is described by a collection of characteristics. A subset of all circumstances for which the output variable value is known constitutes training data or examples. The learning system analyses potential functions, referred to as hypothesis functions, denoted by h, to infer the best target function from a training set. The two kinds of learning issues employed in supervised learning are regression and classification.
  + **Unsupervised Learning** The system searches for relationships between variables or the underlying structure of the data during unsupervised learning. In such cases, the training data consists of samples without any corresponding labels. Association rule mining is primarily impacted by database research, while machine learning emerged much later.

### CNN MODEL

The experiment made advantage of a freely accessible database on heart disease. The dataset has 303 records in total, which were divided into the training set and the test set.

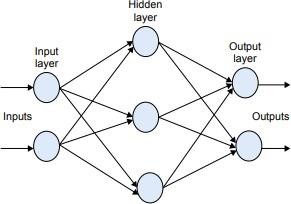


Fig. Multiplayer perceptron Neural Network

(45%) compared to the test group's (65%) A data mining application called Weka 3.6.11 was used for the experiment. Additionally, a multilayer perceptron neural network (MLPNN) with backpropagation was used in the training process (BP).

## MLPNN

The MLPNN is one of the most significant models in artificial neural networks. The MLPNN consists of one input layer, one or more hidden layers, and one output layer. 3 In MLPNN, input nodes first transmit values to the first hidden layer, then second-layer nodes.The first hidden layer passes values to the second, and so on until the outputs as seen in the figure are produced.

### BP Network

In many different situations, it has been demonstrated that the BP algorithm is an effective multilayer perceptron training method. 4 The disparity between actual and predicted values is calculated by the BP network and communicated backward from output nodes to previous layer nodes. The two components that make up the BP learning approach are propagation and weight updating. 4 This learning strategy initially feeds the network with training data before comparing the network's actual and predicted outputs. The error made by each neuron is then discovered. And use this knowledge, the algorithm determines the output that each neuron

should produce, the amount that output must be raised or lowered to achieve the desired output, and finally updates the weights. The entire process is done to improve transaction efficiency.

### The risk factor for heart disease

1. **Family history of heart disease:** - Most individuals are aware that heart disease sometimes runs in families. that if someone has a family history of heart illness, they may be more prone to heart attacks, strokes, and other heart issues. Smoking: Peripheral artery disease, heart attacks, and strokes are all significantly influenced by smoking. Over 40% of all smoking- related fatalities are caused by conditions of the heart and blood vessels. A smoker's risk of suffering a heart attack considerably drops after just one year of quitting.
2. **Cholesterol:** - Atypical blood lipid (fat) levels put people at risk for developing heart diseases. Lipids, particularly the soft, waxy molecule known as cholesterol, are found in the circulation and every cell of the body. Triglycerides, the most common kind of body fat, together with high levels of LDL (low-density lipoprotein) cholesterol hasten atherosclerosis, which raises the risk of heart disease.
3. **High blood pressure: -** High blood pressure (HBP), sometimes known as hypertension, is a medical condition that is commonly misdiagnosed. When we have high blood pressure, our blood vessel walls are more susceptible to being harmed and stretched out of shape. Additionally present are an elevated risk of heart attack, stroke, heart failure, renal failure, and peripheral vascular disease.
4. **Obesity:-** Obesity, which pertains to one's current state of health, is defined as being significantly above one's ideal healthy weight. Anyone who is overweight has a higher risk of having health problems including high blood pressure, diabetes, heart disease, and more.
5. **Lack of physical exercise: -** Exercise inactivity raises the risk of coronary artery disease (CAD). Lack of exercise increases the risk of diabetes and high blood pressure, which increases the likelihood of coronary artery disease (CAD).

## Logistic Regression (Scikit-Learn)

The 4-step modeling technique of Python's sci-kit-learn module, which makes building a machine learning classifier straightforward, is one of its most outstanding aspects. The coding process may be used with any of Sklearn's classifiers, even though this course uses the Logistic

Regression classifier (Decision Tree, K-Nearest Neighbor,s, etc). In this session, we'll show you how to use logistic regression to predict digit labels from images. The picture above depicts a collection of training digits (observations) from the MNIST dataset, organized by category membership (labels 0–9). After a model with logistic regression has been trained, a picture may be used to predict an image label (labels 0-9).A toy dataset (the digits dataset) was used to rapidly illustrate Scikit-four-step Learn's modeling structure and the operation of the stochastic regression method. The second half of the class covers the MNIST dataset to briefly illustrate how changing a model's default parameters could impact performance (both in timing and accuracy of the model). Logistic regression is a statistical model used for predicting the probability of a binary outcome (such as a 0 or 1, yes or no, true or false). It is a type of regression analysis that uses a logistic function to model the relationship between a set of independent variables and a binary dependent variable.

## Naive Bayes Classification using Scikit-learn

Assume you're a product manager who wants to categorize customer reviews into positive and bad comments. Or As a loan manager, do you want to know whether loan

applications are risky or safe? As a healthcare analyst, you want to predict who will get diabetes. Every circumstance involves the problem of categorizing reviews, loan applications, and patients.

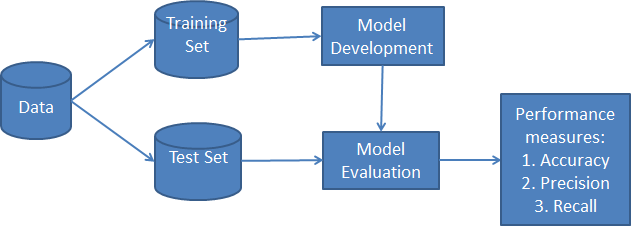


**Classification Workflow**

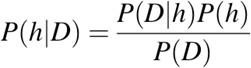
Large volumes of data may be handled with Naive Bayes, the simplest and fastest categorization method. The naive Bayes classifier is effectively used in several applications, such as spam filtering, text categorization, sentiment analysis, and recommender systems. The probability theory's Bayes theorem is used to predict an unknown class. Naive Bayes is a probabilistic algorithm used for classification tasks in machine learning. It is based on Bayes' theorem and assumes that the features are independent of each other, which is why it is called "naive."The basic idea of Naive Bayes is to calculate the probability of each class given a set of features, and then choose the class with the highest probability as the predicted class for the input data.

Understanding the problem, locating possible characteristics, and labeling them are the first steps in categorization. Features are those characteristics or features that affect the results of the label. In the case of a loan distribution, for instance, bank management can determine a customer's occupation, income, age, geography, previous loan history, transaction history, and credit score. These characteristics are known as features that help the model classify customers.

The categorizing process consists of a learning phase and an evaluation phase. During the learning phase, a given dataset is utilized to train the classifier's model, and performance is evaluated during the evaluation phase. Performance is evaluated depending on a variety of criteria, including memory, accuracy, and error.



### What is a Naive Bayes Classifier?

Naive Bayes is the name of a statistical classification technique built on the Bayes Theorem. This is one of the simplest supervised learning techniques. The naive Bayes classifier is a practical, reliable, and fast method. Naive Bayes classifiers function rapidly and precisely on large datasets.Presumptuous The Classification algorithm operates upon the assumption that the effects of one attribute on a class are independent of those of other qualities. For instance, the eligibility of a loan applicant is influenced by their geography, age, history of loans and transactions, and income. Even though these attributes are related, they are nonetheless considered independently. Considering that it makes calculations simpler, this assumption is viewed as naive. It is known as class conditional independence.

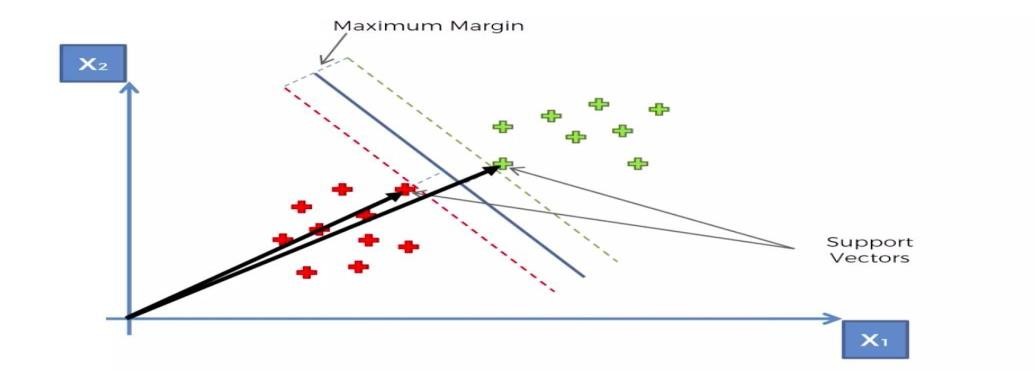
* P(h) is the probability that hypotheses h and I are true (regardless of the data). This is referred to by the h prior probability.
* "Potential of the Data" is what P(D) stands for (regardless of the hypothesis). This is referred to as the prior probability.
* P(h|D): the chance that, given the facts D, hypothesis h will come true. This is

referred to as posterior probability.

* P(D|h) measures how likely it is that data d would be present if hypothesis h were correct. The likelihood in the future this is.

## Support Vector Machine

The Support Vector Machine (SVM), a supervised machine learning method, may do classification, regression, and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. A single class label will be applied to all of the data points that fall on one side of the line, and a second class label will be applied to all of the points that fall on the other side. Although it appears simple, there are countless lines to choose from. How can we identify the line that will categorize the data the most precisely? This circumstance makes use of the LSVM approach. The LSVM method will select a line that is as distant from the nearest samples as is feasible in addition to dividing the two classes. In actuality, the term "support vector" in the phrase "support vector machine" refers to two position vectors drawn from the origin to the points that determine the decision boundary.



## K Nearest Neighbor Algorithm

K-Nearest Neighbors, sometimes known as KNN for short, is one of the most widely used machine learning algorithms. KNN is a non-parametric lazy learning technique. If a technique does not make any assumptions about the underlying data, it is said to be non-parametric. In other words, it makes a decision based on how closely the selection resembles other data points, independent of the attribute that the numerical values represent. This definition states that a lazy learning algorithm has little to no training phase. Consequently, we can swiftly categorize new data points as soon as they occur.

### Some pros and cons of KNN Pros:

* + No data assumptions are made.
  + Simple algorithm that is simple to comprehend
  + Regression and classification techniques are applicable.

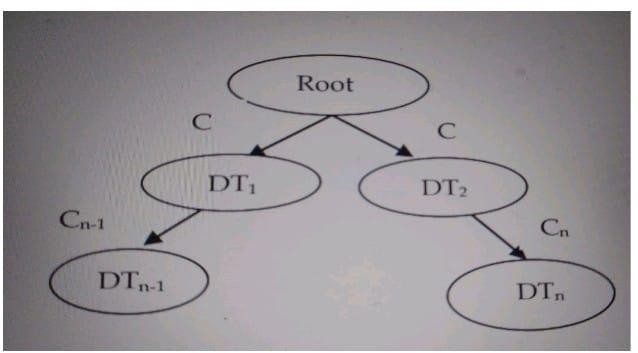
### Cons:

* Huge storage stipulation: To compute the nearest K neighbors, all of the training data must be stored in memory.
* sensitive to little details
* considering that we're calculating the distance to the nearest K points, sensitive to the data's size.

## Decision Tree Classification

Well, how to build and fine-tune Decision Tree Classifiers using Python's Scikit-Learn module, along with metrics for attribute selection. As a marketing manager, you want a set of customers who are most likely to purchase your products. You may cut your marketing costs by determining your audience. As a loan manager, you must identify risky loan applications to lower the rate of defaults. When clients are split into categories of potential and non-potential clients or safe or risky loan applications, a classification challenge arises. The two stages of the categorization process are learning and prediction. The learning phase's given training data is used to build the model. In the prediction step, the modes utilized the response for the given data. The decision tree is one of the most popular and simple classification methods. It might be applied to classification and regression issues. Decision Tree is one of the supervised learning algorithms which it is used to represent the decision that is made based on the condition. It is used for both classification and regression. The Decision tree is always constructed from top to bottom. The first node from the top is called as root node. The last nodes is called as a leaf node. Internal nodes are present in between the root node and leaf nodes. Based on some condition the internal nodes are split and finally, the decisions are made. In the real time as the number of variables increases tree grows larger and algorithm becomes complex. In Decision tree we have two types, they are classification and regression trees. Classification tree is used to classify the dataset. so that it is easy to analyze the data. But using this algorithm we cannot make a prediction. The Regression tree is a tree mainly used to predict continuous values.Growth of tree depends on factors like the attribute which is chosen to make a prediction,condition used for the split the tree; deciding when to stop or terminate the

growth of the tree.

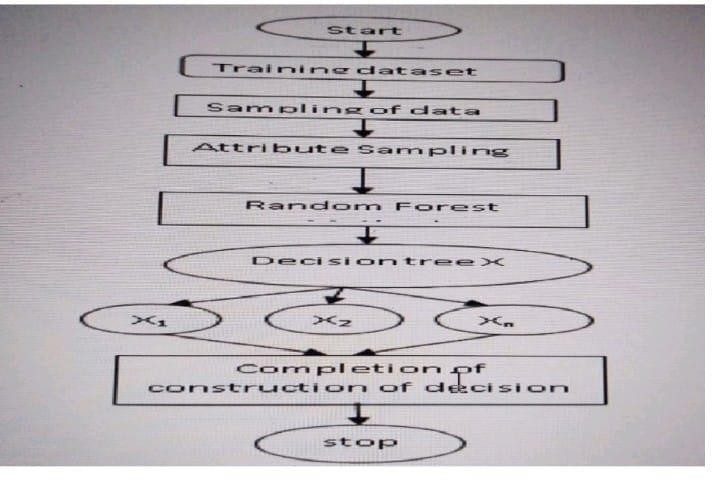


## Fig: 6.1 DECISION TREE

**Random Forest Algorithm**

Random forest is a supervised machine learning technique built on ensemble learning. You can combine different sorts of algorithms or employ the same strategy more than once in ensemble learning to get a more accurate prediction model. The random forest approach combines many algorithms of the same type, or various decision trees, into a forest of trees, hence the name "Random Forest." Using the random forest method, regression and classification tasks may both be completed. t is defined as a set of decision trees to do regression and classification. Classification is used to find out the majority voting. Regression is used to calculate the mean value. This algorithm is more accurate, robust, and can handle a variety of data such as binary data, categorical data, and continuous data. Random Forest is nothing but multiple decision trees. 75% of the dataset is considered for the training. The training data is subjected to sampling and based on attribute sampling different decision trees are constructed by applying

the Random Forest.



## XG Boost Algorithm

### Fig:6.2 RANDOM FOREST

A distributed gradient boosting library called XG Boost was created to be exceptionally efficient, flexible, and portable. It builds machine learning methods using the Gradient Boosting framework. XG Boost provides a parallel tree-boosting method (also known as GBDT or GBM) that effectively addresses a range of data science problems.

## WEKA Software:

Weka is a popular open-source machine learning software tool that provides a collection of algorithms for data mining and data analysis. Weka stands for Waikato Environment for Knowledge Analysis and was developed at the University of Waikato in New Zealand. It was first released in 1997 and has since been widely used by researchers and practitioners in various fields, including bioinformatics, marketing, finance, and more.Weka includes a graphical user interface that allows users to explore their data, preprocess it, and apply various machine learning algorithms to build models. The software also supports several

data formats, including CSV, ARFF, and SQL databases. Weka offers a wide range of classification, regression, clustering, and association rule mining algorithms that can be used to build predictive models from data.

One of the main benefits of using Weka is that it is open-source and free to use. Additionally, it has a large user community and comprehensive documentation, making it easy for users to get started with the software. Weka also has a plugin system that allows users to extend its capabilities with new algorithms and tools. Overall, Weka is a powerful and flexible tool for data mining and machine learning, and is widely used by researchers and practitioners in various fields.

1. Algorithms: Weka includes a wide range of machine learning algorithms for classification, regression, clustering, and association rule mining. These include popular algorithms such as decision trees, random forests, support vector machines, k-nearest neighbors, and more.
2. Preprocessing: Weka offers several preprocessing tools to help users clean and prepare their data for analysis. These include tools for attribute selection, normalization, discretization, and more.
3. Visualization: Weka includes several visualization tools that allow users to explore their data and model results. These include scatterplots, boxplots, ROC curves, and decision trees.
4. Integration: Weka can be easily integrated with other tools and programming languages such as R and Python. This allows users to leverage the strengths of Weka alongside other tools in their workflow.
5. Community: Weka has a large and active user community that provides support and resources to users. The Weka website includes a forum, documentation, and tutorials to help users get started with the software.
6. Extensions: Weka has a plugin system that allows users to extend its capabilities with new algorithms and tools. There are several plugins available for Weka, including plugins for deep learning, time series analysis, and more.
7. Education: Weka is often used in machine learning courses and workshops as a tool for teaching data mining and machine learning concepts. The software is user-friendly and provides a great platform for hands-on learning and experimentation.

Overall, Weka is a powerful and versatile tool for machine learning and data mining that is widely used by researchers and practitioners in various fields. Its range of algorithms, preprocessing tools, visualization tools, and community support make it a popular choice for both beginners and experienced users.

# 7. OUTPUT SCREENS

* 1. Home Page of Weka:

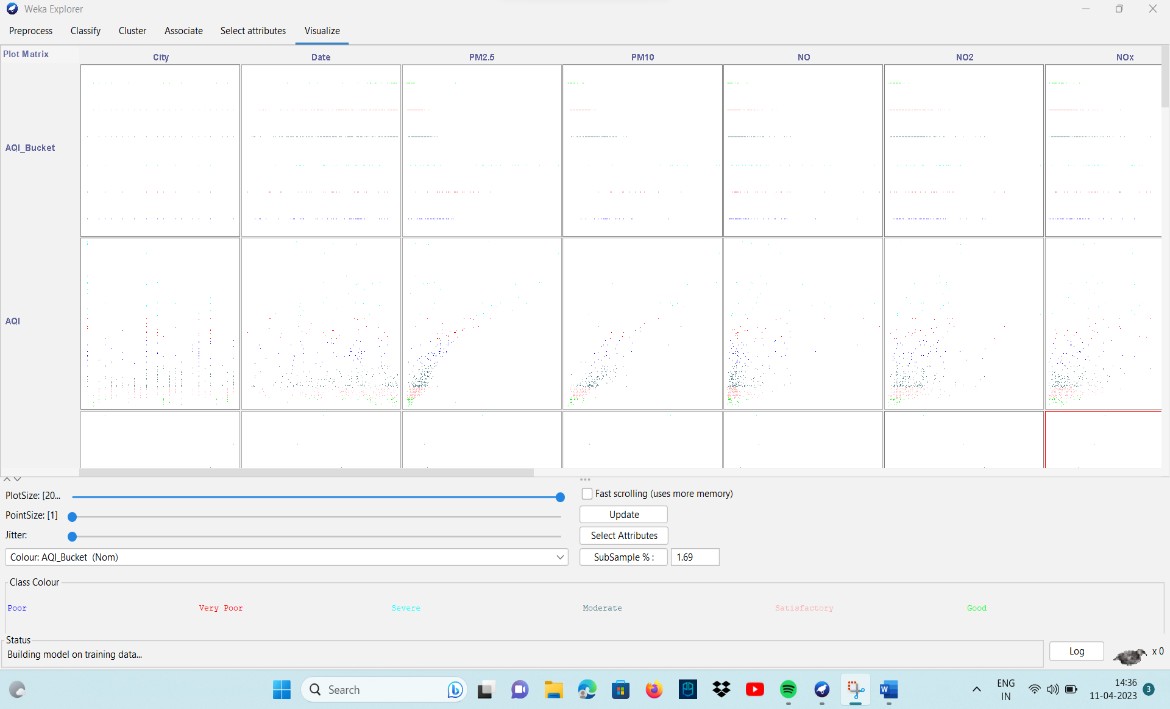


Fig: 7.1 Home Page of Weka

* 1. Representation of particles present in the air

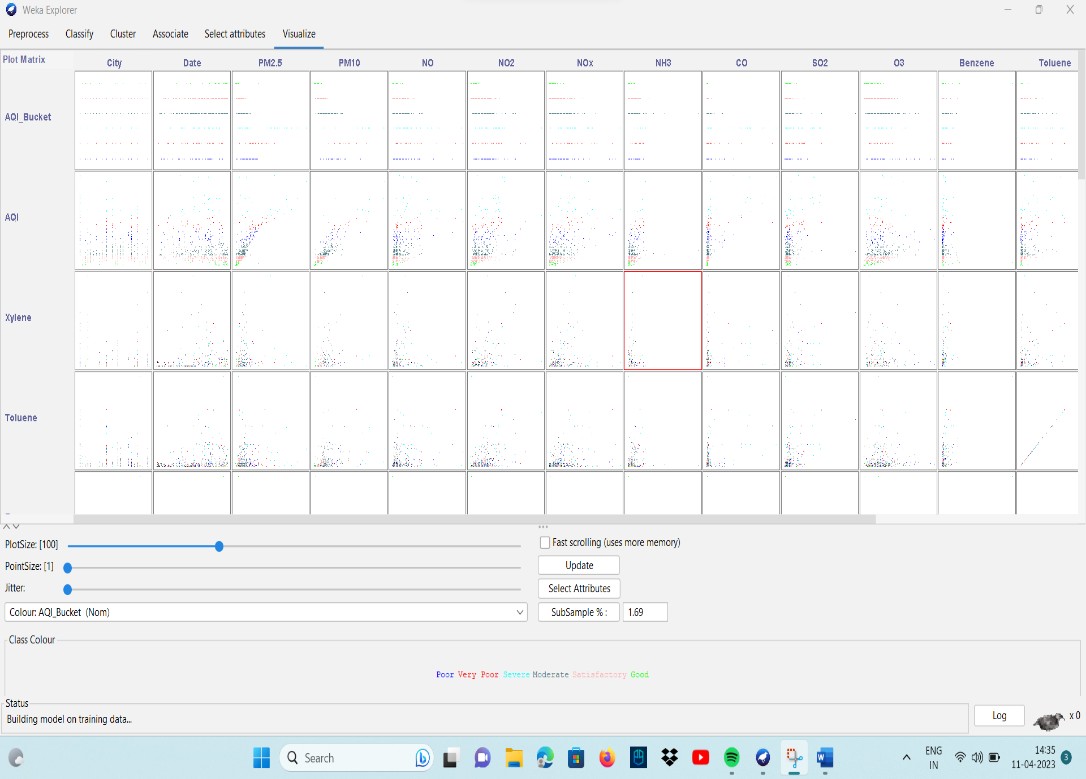


Fig: 7.2 Representation of particles present in the air

### SELECTING DATA BASE

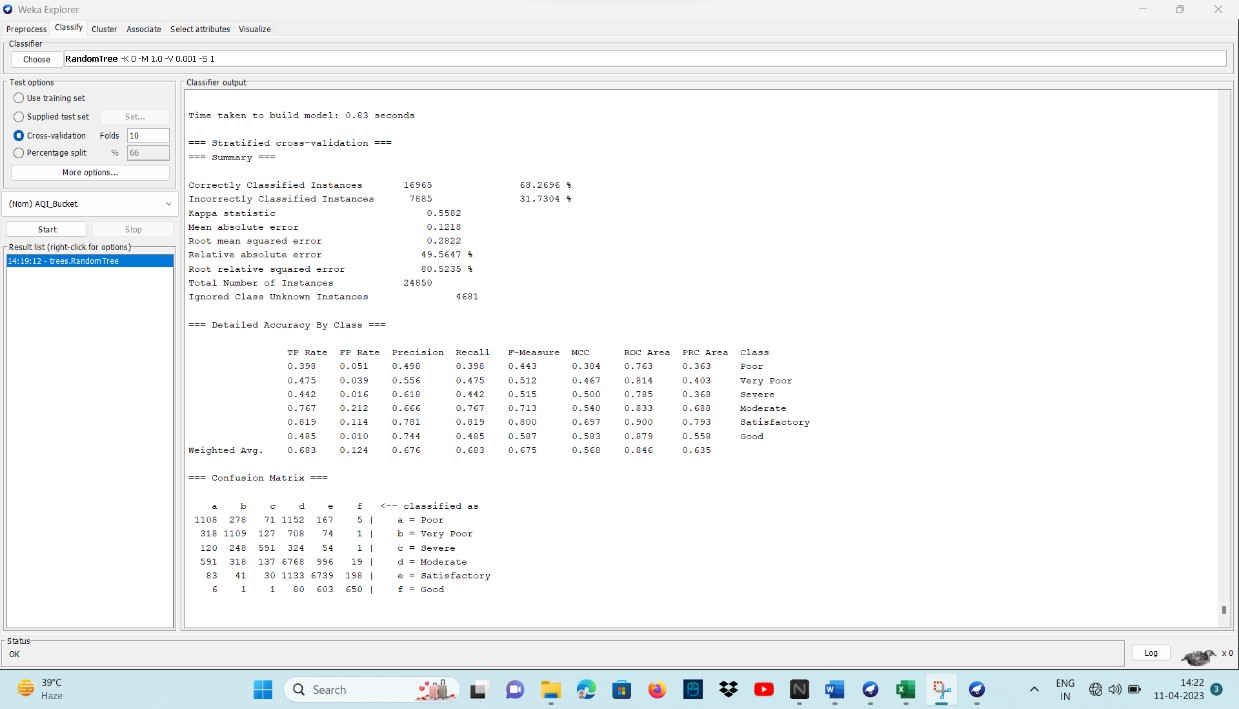


Fig:7.3 Selecting a database

* 1. AQI BUCKET Of Dataset

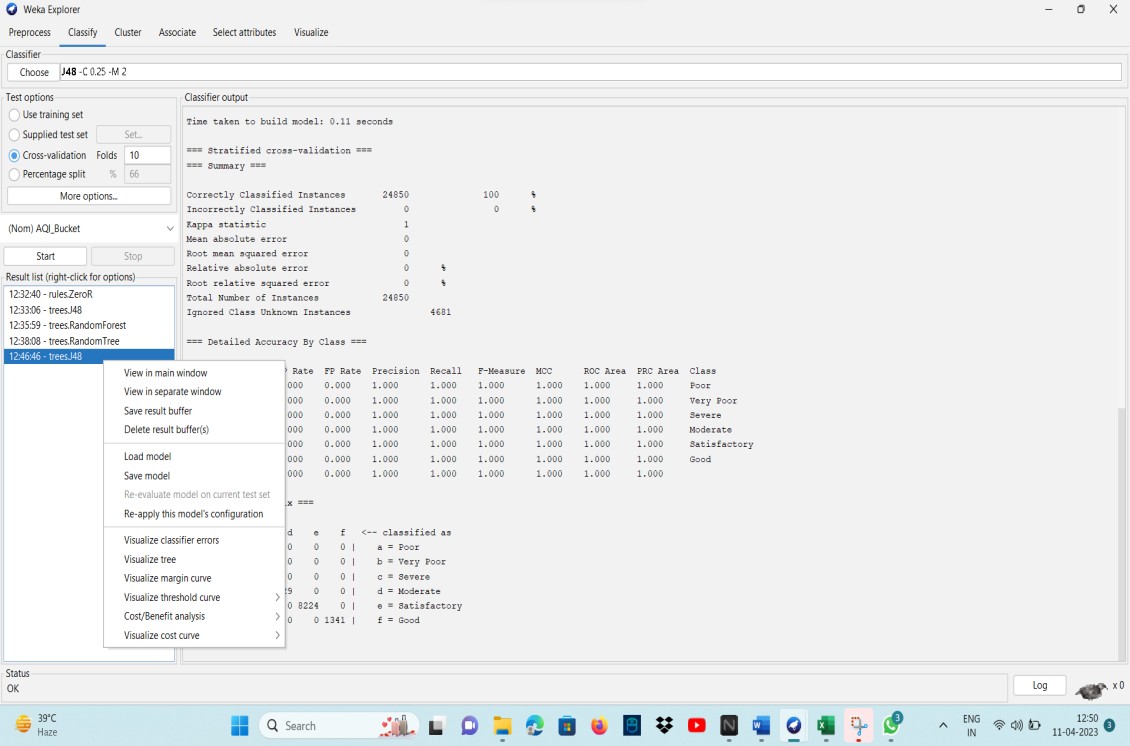


Fig: 7.4 AQI BUCKET Of Dataset

* 1. AQI BUCKET of date with pm 2.5

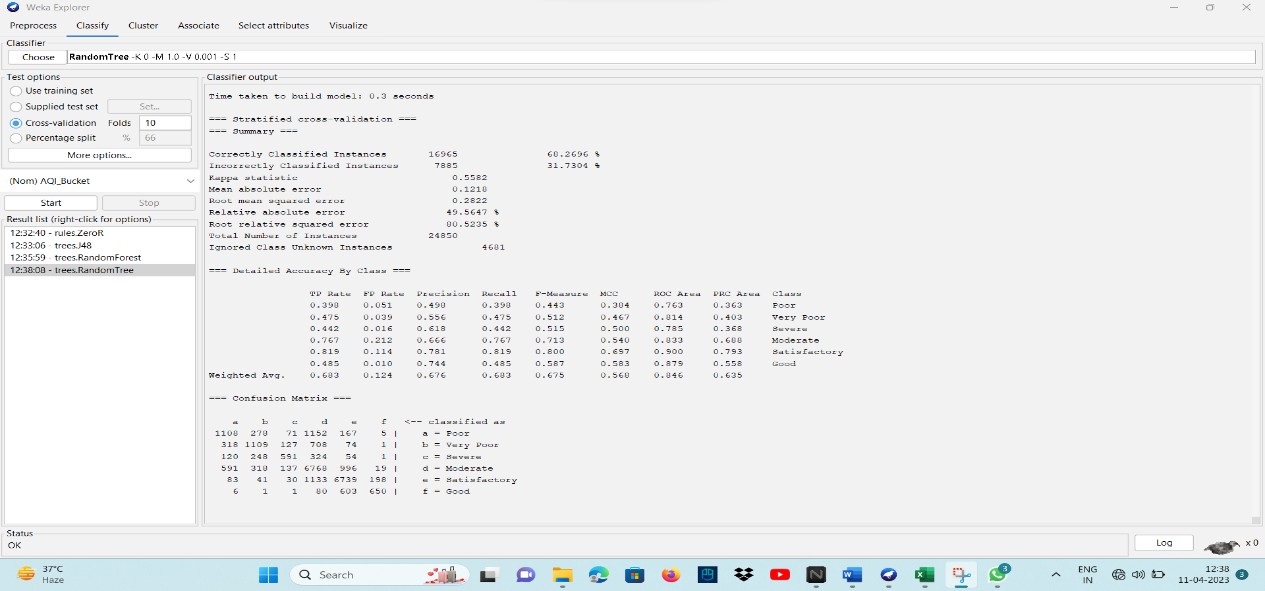


Fig: 7.5 AQI BUCKET of date with pm 2.5

* 1. OUTPUT OF RANDOM FOREST

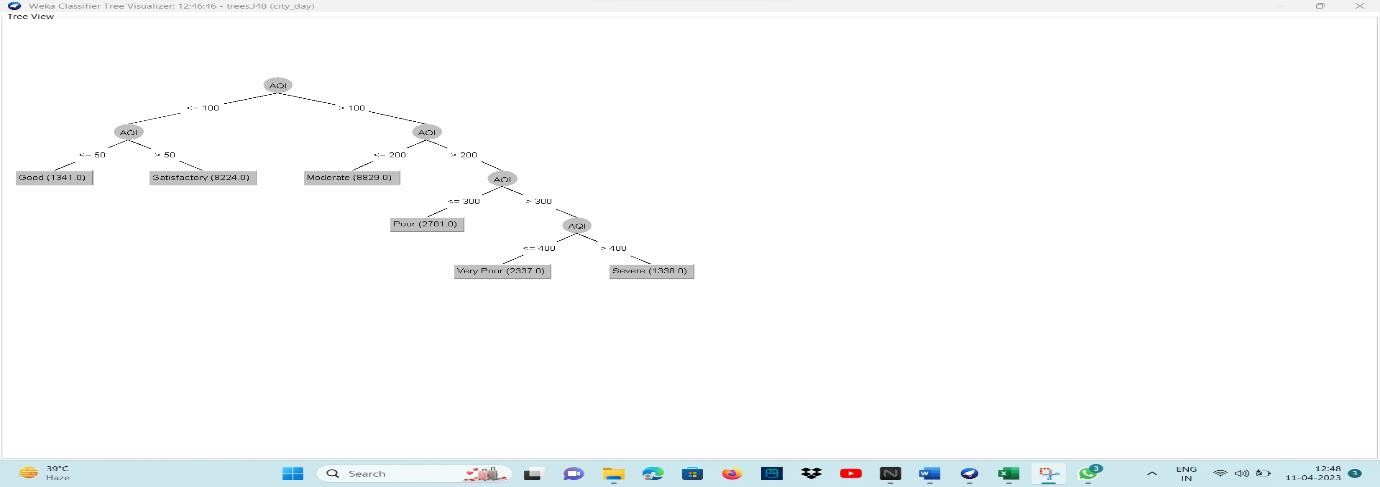


FIG: 7.6 OUTPUTOF RANDOM FOREST

* 1. OUTPUT OF RANDOM TREE

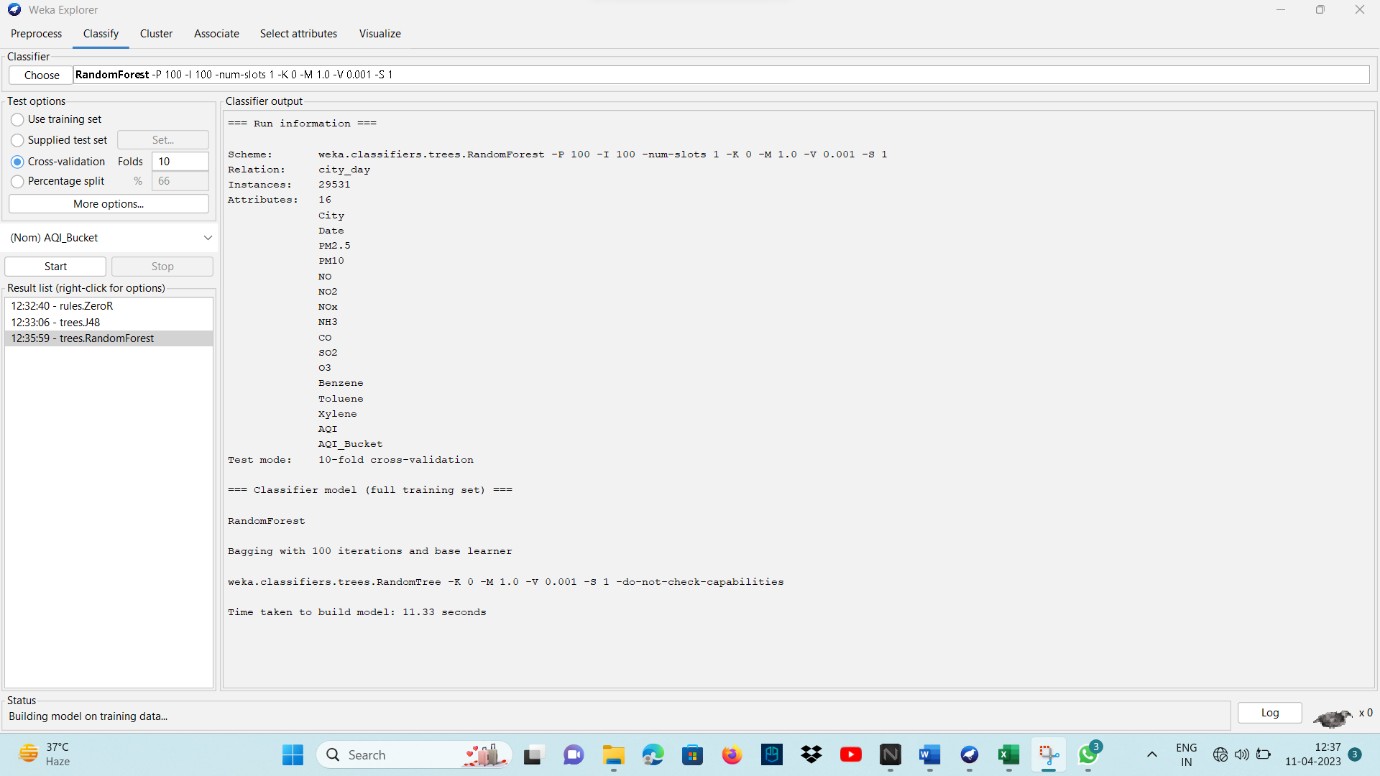


FIG:7.7 Output of random Tree

* 1. OUTPUT OF DECISION TREE

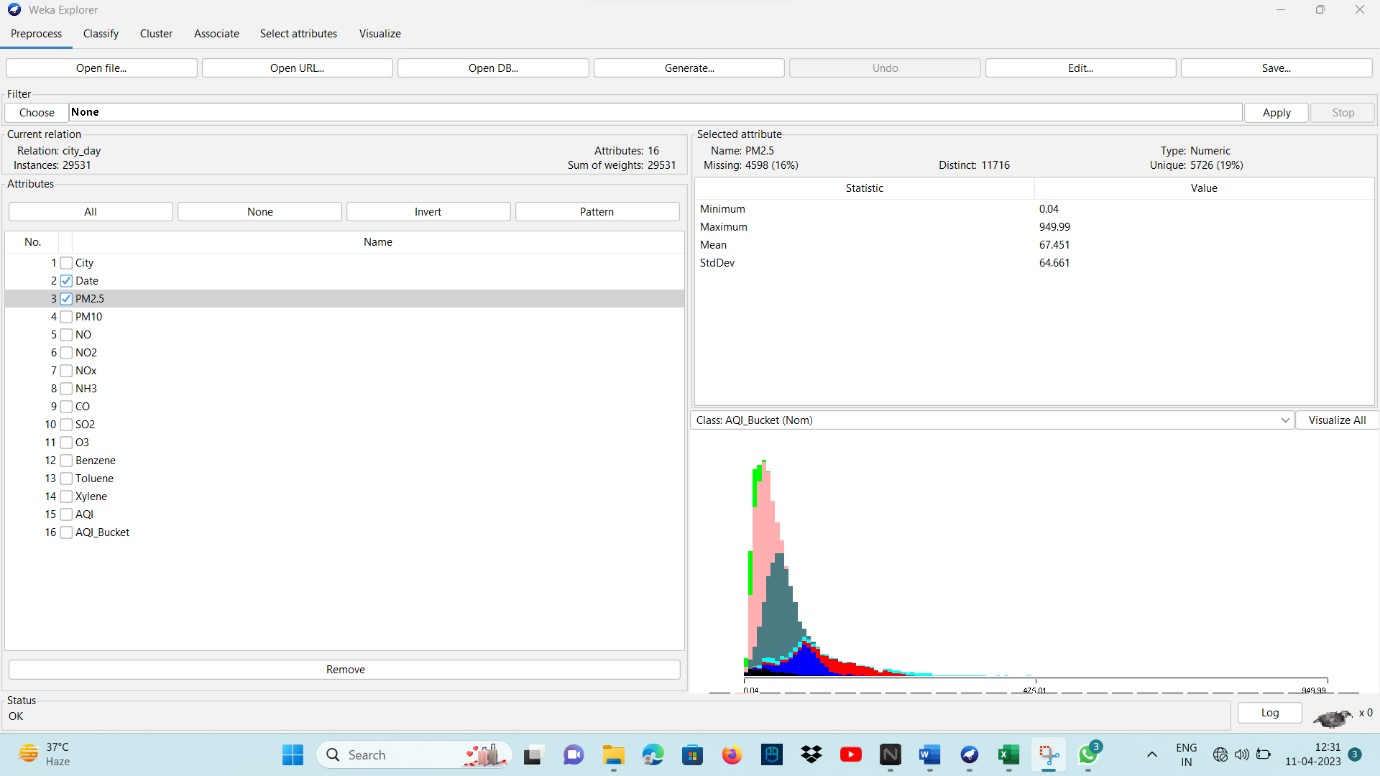


Fig: 7.8 Output Of Decision Tree

* 1. Detailed output of random tree with cross validation

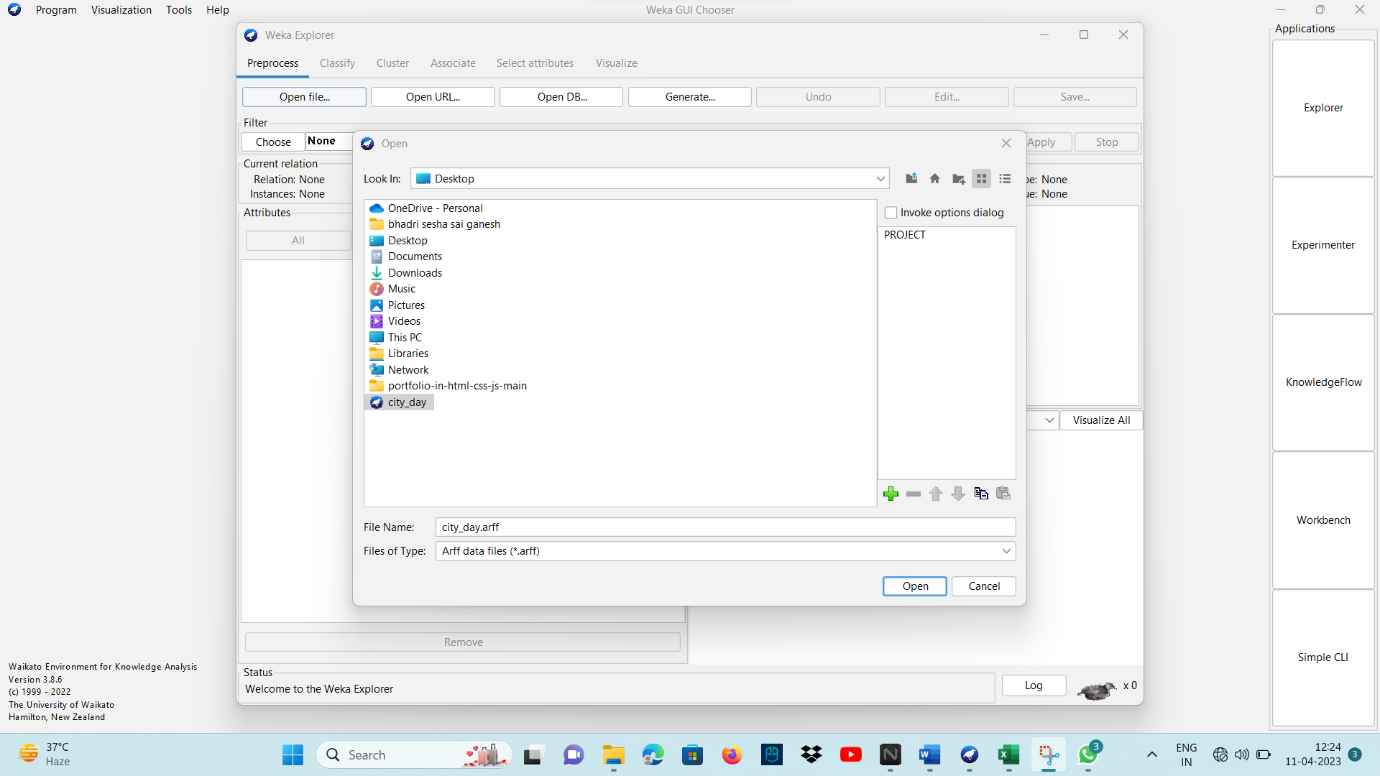


Fig:7.9 cross validation

* 1. Visualising the pollutants from dataset to plots of graphs.

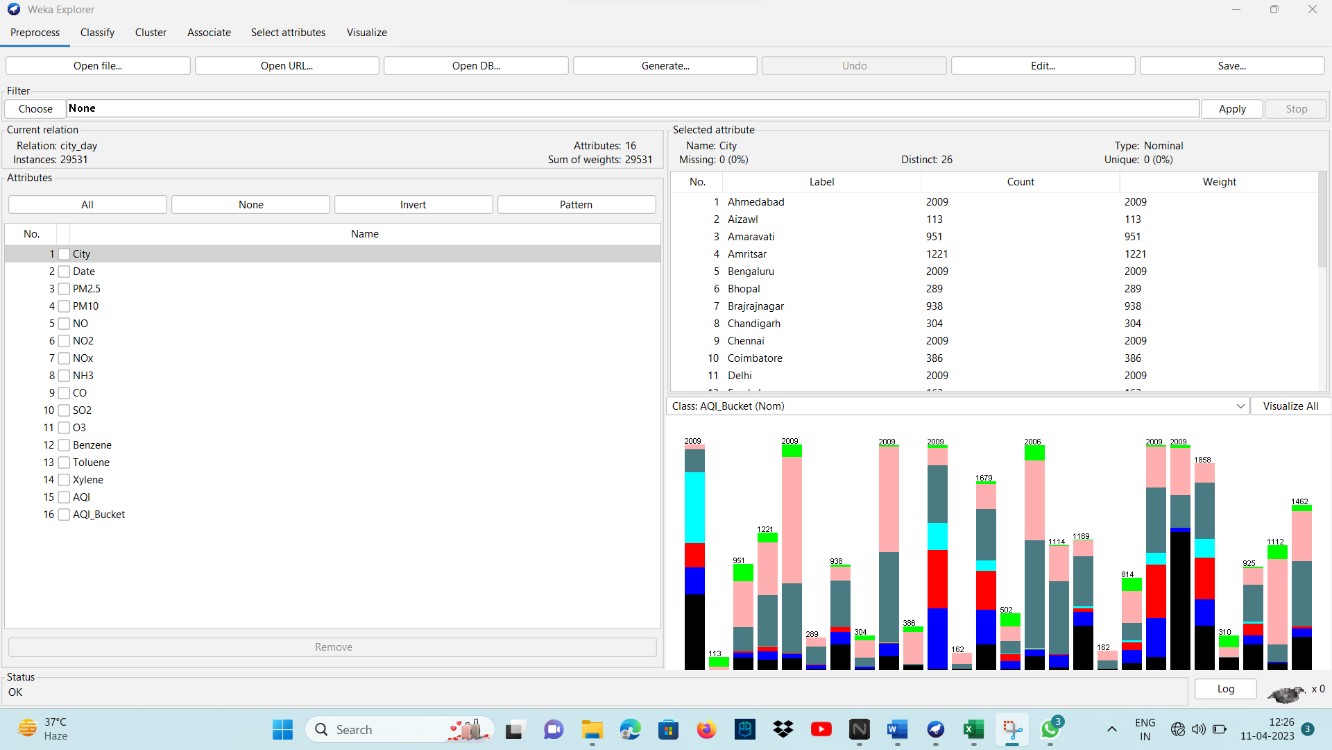


Fig: 7.10 Graphs

1. **SYSTEM TESTING**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

## TYPES OF TESTS

**Unit testing**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

## Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

## Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is cantered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected. Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised. Systems/Procedures: interfacing systems or procedures must be invoked.Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

## System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

## White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

## Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

## Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

## Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

## Test objectives

* All field entries must work properly.
* Pages must be activated from the identified link.
* The entry screen, messages and responses must not be delayed.

## Features to be tested

* Verify that the entries are of the correct format
* No duplicate entries should be allowed
* All links should take the user to the correct page

## Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications,

e.g. components in a software system or – one step up – software applications at the company level – interact without error.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

## Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

## CONLCUSION

The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Air pollution prediction, is used to measure the predict the quality of the air. So, it predict the polluted air whether its increasing year by year or decreasing. We are going to use Decision tree and Random forest algorithms in Data mining techniques. To give the accuracy in data and amount of pollution increasing year by year or decreasing year by year with the analysis made from past few years.

## FUTURE SCOPE

Integration with IoT and mobile devices for real-time monitoring and prediction. Expansion of models to incorporate geographic and temporal variability.

Incorporation of data on multiple pollutants for a more comprehensive air quality assessment.Collaboration with public health agencies and policymakers to inform interventions and policy decisions.

# REFERENCES

[1] Kumar, N. Komal, et al. "Predicting Non-Small Cell Lung Cancer: A Machine Learning Paradigm." Journal of Computational and Theoretical Nanoscience 15.6-7 (2018): 2055-2058.

[2] Vigneswari, D., et al. "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus." 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, 2019.

[3] Kumar, N. Komal, et al. "An Optimized Random Forest Classifier for Diabetes Mellitus." Emerging Technologies in Data Mining and Information Security.Springer, Singapore, 2019. 765-773.

[4] Squire, Kenneth J., et al. "Photonic crystal-enhanced fluorescence imaging immunoassay for cardiovascular disease biomarker screening with machine learning analysis." Sensors and Actuators B: Chemical 290 (2019): 118-124.

[5] Alizadehsani, Roohallah, et al. "Machine learning-based coronary artery disease diagnosis: A comprehensive review." Computers in biology and medicine (2019): 103346.

[6] Kumar, N. Komal, and D. Vigneswari. "Hepatitis-Infectious Disease Prediction using Classification Algorithms." Research Journal of Pharmacy and Technology 12.8 (2019): 3720- 3725.

[7] Tang, Xiaochen, et al. "A Real-time Arrhythmia Heartbeats Classification Algorithm using Parallel Delta Modulations and Rotated Linear-Kernel Support Vector Machines." IEEE Transactions on Biomedical Engineering (2019).

[8] Abdar, Moloud, et al. "A new machine learning technique for an accurate diagnosis of coronary artery disease." Computer methods and programs in biomedicine 179 (2019): 104992.

[9] Chen, Rui, et al. "Using Machine Learning to Predict One-year Cardiovascular Eventsin Patients with Severe Dilated Cardiomyopathy." European Journal of Radiology (2019).

[10] Dhar, Sanchayita, et al. "A Hybrid Machine Learning Approach for Prediction of Heart Diseases." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.