# RAINFALL PREDICTION USING MACHINE LEARNING

*A Project Report submitted in the partial fulfillment of the requirements for the award of the Degree of*

**BACHELOR OF TECHNOLOGY**
**In**
**COMPUTER SCIENCE & ENGINEERING**

**Submitted By**


| | |
|---|---|
| **K.RENUKA DEVI** | **(Regd No : 20NE1A0572)** |
| **A.PADMA SREE** | **(Regd No : 20NE1A05B7)** |
| **N.BALA MARY SOWMYA** | **(Regd No : 20NE1A05B4)** |
| **K.SANTOSH DUTT** | **(Regd No : 20NE1A0580)** |


**Under The Esteemed Guidance Of**

**Mrs.SHAMMI SHAIK** M.Tech
**Asst.Prof**



**Department of Computer Science & Engineering**

**TIRUMALA ENGINEERING COLLEGE**

**(Approved by AICTE &Affiliated to JNTU, KAKINADA, Accredited by NAAC&NBA)**

**Jonnalagadda, Narasaraopet, GUNTUR(Dt.),A.P.**


**2020-2024**

# TIRUMALA ENGINEERING COLLEGE

**(Approved by AICTE &Affiliated to JNTUKAKINADA, Accredited by NAAC&NBA)**

**Jonnalagadda, Narasaraopet-522601, Guntur (Dist) A.P**

## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

## CERTIFICATE

This is to certify that the project report entitled **"RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES"** is the bonafide work carried out by **K.RENUKA DEVI (20NE1A0572), A.PADMA SREE (20NE1A05B7), N.BALA MARY SOWMYA (20NE1A05B4), K.SANTOSH DUTT (20NE1A0580)** in partial fulfillment of the requirements for the award of **"Bachelor of Technology"** degree in the **Department of CSE** from J.N.T.U.KAKINADA during the year 2023-2024 under our guidance and supervision and worth of acceptance of requirements of the university.

**Project Guide**

**Mrs. Shammi Shaik**

**M.Tech**

**Head of the department**

**Dr. N. Gopala Krishna**

**MTech, Ph.D, MISTE**

**Project Coordinator**

**Mr.S. Anil Kumar**

**M.Tech, (Ph.D)**

**External Examiner**

# ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairman **Sri**. **Bolla Brahma Naidu,** our secretary **Sri. R. Satyanarayana**, who took keen interest in our every effort through out this course. We owe out gratitude to our principal sir **Dr. Y. V. Narayana** for his kind attention and valuable guidance through out the course.

We express our deep felt gratitude to our **H.O.D Dr. N. Gopala Krishna,** and **Mr. S. Anil Kumar** coordinator of the project for extending their encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us through out the project work.

We wish to express our sincere deep sense of gratitude to our**, Mrs. Shammi Shaik** for significant suggestions and help in every respect to accomplish the project work. Her persisting encouragement, ever lasting patience and keen interest in discussion shave benefited us to be extent that cannot be spanned by words to our college management for providing excellent lab facilities for completion of project within our campus.

We extend our sincere thanks to all other teaching and non-teaching staff of department of CSE for their cooperation and encouragement during our B.Tech course.

We have no words to acknowledge the warm affection,constant inspiration and encouragement that we received from my parent.

We affectionately acknowledge the encouragement received from my friends and those who involved in giving valuable suggestions had clarifying out doubts which hadreally helped us in successfully completing our project.

*By*

**K.RENUKA DEVI** **(20NE1A0572)**

**A.PADMA SREE** **(20NE1A05B7)**

**N.BALA MARY SOWMYA** **(20NE1A05B4)**

**K.SANTOSH DUTT** **(20NE1A0580)**

# ABSTRACT

# ABSTRACT

India is an agricultural country and its economy is largely based upon crop productivity and rainfall. For analyzing the crop productivity, rainfall prediction is require and necessary to all farmers. Rainfall Prediction is the application of science and technology to predict the state of the atmosphere. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre planning of water structures. Using different data mining techniques it can predict rainfall. Data mining techniques are used to estimate the rainfall numerically. This paper focuses some of the popular data mining algorithms for rainfall prediction. Random Forest, K-Nearest Neighbor algorithm, Logistic regression, Decision Tree are some of the algorithms have been used. From that comparison, it can analyze which method gives better accuracy for rainfall prediction.

# INDEX


| CONTENT | PAGENO |
|---|---|
| **1. INTRODUCTION** | **1-2** |
| **2. LITERATURE SURVEY** | **3-5** |
| **3. SYSTEM ANALYSIS** | **6-11** |
| **3.1** Existing System | 6 |
| **3.2** Proposed System | 7 |
| **3.3** System Architecture | 8 |
| **3.4** Feasibility Study | 9 |
| 3.4.1 Economical Feasibility | |
| 3.4.2 Technical Feasibility | |
| 3.4.3 Social Feasibility | |
| **3.5** Requirements Specifications | 11 |
| 3.5.1 Hardware Requirements | |
| 3.5.2 Software Requirements | |
| **4. DESIGN** | **12-19** |
| **4.1** Flow Chart | 12 |
| **4.2** Data Flow Diagrams | 13-14 |
| **4.3** UML Diagram | 14-19 |
| 4.3.1 Use case Diagram | |
| 4.3.2 Class Diagram | |
| 4.3.3 Activity Diagram | |
| 4.3.4 Sequence Diagram | |

**5. IMPLEMENTATION** **20-36**

**5.1** Python 20-28

**5.2** Sample Code 28-36

**6. OUTPUT SCREENS** **37-54**

**7. SYSTEM TESTING** **55-58**

**7.1** Testing plan 55

**7.2** Basics of software testing 55-56

**7.3** Types of testing 56-58

**7.4** Test cases 58

**8. CONCLUSION** **59**

**9. FURTHER ENHANCEMENTS** **60**

**10. BIBILIOGRAPHY** **61-62**

# INTRODUCTION

# 1. INTRODUCTION

## 1.1 OBJECTIVE OF THE PROJECT

The goal is to develop a machine learning model for Rainfall Prediction to potentially replace the up datable supervised machine learning classification models by predicting results in the form of best accuracy by comparingsupervised algorithm.

### Necessity

This prediction helps in predicting the rainfall and it helps in Over coming the crop productivity and to predict the state of atmosphere in agricultural countries. These models are very easy to use. It can work accurately and very smoothly in a different scenario. It reduces the effort workload and increases efficiency in work. Inaspects of time value, it is worthy.

### Software development method

In many software applications program different methods and cases are followed such as, Waterfall model, Iterative model, Spiral model, V- model and Big Bang model. we used waterfall model in this application. we tried to use test case and cases of ware approaches.

### Layout of the document

This documentation starts with form a introduction. After introduction analysis and design of the project are described. In analysis and design of the project have many parts such as project proposal, mission, goal, target audience, environment. Use cases and test cases are explained below respectively. Finally, thisdocumentation finished with result and Conclusion part.

## 1.2 OVERVIEW OF THE DESIGNED PROJECT

At first, we take the data set from out resource then we have to perform data-prepossessing, visualization methods for cleaning and visualizing the data set respectively and we applied the Machine Learning algorithms on the data set and then we plot confusion matrix of each technology at last we compare those models and draw the ROC curve for the best per forming model and also we plot classificationreport forth at model.

# LITERATURE SURVEY

# 2. LITERATURE SURVEY

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes are search proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information.It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent event of them.

## Review of Literature Survey

.

[1] Measurable investigation shows the idea of ISMR, which can't be precisely anticipated by insights or factual information.Hence, this review exhibits the utilization of three techniques: object creation, entropy, and artificial neural network (ANN). In view of this innovation, another technique for anticipating ISMR times has been created to address the idea of ISMR.This model has been endorsed and supported by the studio and exploration data. Factual examination of different information and near investigations showing the presentation of the normal technique.

[2] The primary impact of this movement is to exhibit the advantages of AI calculations, just as the more prominent degree of clever frame work than the advance drain fall determining methods. We analyze and think about the momentum execution (Markov chain stretched out by rainfall research) with the forecasts of the six most

AI machines:Genetic programming, Vector relapse support, radio organizations, M5organizations, M5 models, models-Happy. To work with a more itemized appraisal, we led rainfall over view utilizing information from 42 metropolitan urban communities.

[3] RF was utilized to anticipate assuming that it would rain in one day, while SVM was utilized to foresee downpour on a blustery day. The limit of the Hybrid model was fortified by the decrease of day-by-day rainfall in three spots at the rainfall level in the eastern piece of Malaysia. Crossover models have likewise been found to emulate the full change, the quantity of days straight, 95% of the month -to-month rainfall, and the dispersion of the noticed rainfall.

[4] In India, farming is the back bone. Down pour is a significant plant.These days, climate is a major issue. Climate gauging gives data on rainfall estimating and crop security. Numerous strategies have been created to recognize rainfall. Machine Learning calculations are significant in foresee in grain fall.

[5] Climate sooner or later. Climatic still up in the air utilizing various sorts of factors all over the place of these, main the main highlights are utilized in climate conjectures. picking something like this relies a great deal upon the time you pick. Under lying displaying is utilized to incorporate the fate of demonstrating, AI applications, data trade, and character examination.

[6] Contrasted with different spots where rainfall information isn't accessible, it consumes a large chunk of the day to build up a solid water overview for a long time. Improving complex neural organizations is intended to be a brilliant instrument for anticipating the stormy season. This downpour succession was affirmed utilizing a complex and the arrangement of informational collections for transient arranging are clear in the examination of different organizations, like Ada naive. Ada SVM.

[7] In this paper, Artificial Neural Network (ANN) innovation is utilized to foster a climate anticipating strategy to distinguish rainfall utilizing Indian rainfall information. A long these lines, Feed Forward Neural Network (FFNN) was utilized

Back propagation Algorithm. Execution of the two models is assessed dependent on emphasis examination, Mean Square Error(MSE) and Magnitude of Relative Error(MRE). This report like wise gives a future manual for rainfall determining.

[8] This page features rainfall investigation speculations utilizing Machine Learning. The principle motivation behind utilizing this program is to secure against the impacts of floods. This program can be utilized by conventional residents or the public authority to anticipate what will occur before the flood. The flood card, then, at that point, furnish them with the vital help by moving versatile or other important measures.

# SYSTEM ANALYSIS

# 3. SYSTEM ANALYSIS

## 3.1 EXISTING SYSTEM

Agriculture is the strength of our Indian economy. Farmer only depends up on monsoon to be their cultivation. The good crop productivity needs good soil, fertilizer and also good climate. Weather fore casting is the very important requirement of the each farmer.Due to the sudden changes in climate/weather, The people are suffer economically and physically. Weather prediction is one of the challenging problems in current state. The main motivation of this paper to predict the weather using various data mining techniques. Such as classification, clustering, decision tree and also neural networks. Weather related in formation is also called the meteorological data. In this paper the most commonly used weather parameters are rainfall, wind speed, temperature and cold.

### 3.1.1 Disadvantages

The biggest disadvantage of this approach is that it fails when it comes for long term estimation.

### Scope of the project

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are limited to precision,sensitivity,F1 - score**.**

### Overview of the project

The overview of the project is to provide a best machine learning algorithm to the user.Therefore, the user can directly know whether the rainfall is occur or not through his best model**.**

## 3.2 PROPOSED SYSTEM

### 3.2.1 Exploratory Data Analysis of Rainfall Prediction

Multiple datasets from different sources would be combined to form a generalized data set, and then different machine learning algorithms would be appliedto extract patterns and to obtain results with maximum accuracy.

### 3.2.2 Data Cleaning

In this section of the report will load in the data,check for cleanliness, and then trim and clean given data set for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

### 3.2.3 Data Collection

The data set collected for predicting given data is split into Training set and Test set. Generally, we split the data set into Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

### 3.2.4 Building the classification model

For predicting the rainfall, ML algorithm prediction model is effective becauseof the following reasons: It provides better results in classification problem.It is strong in pre processing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables. It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

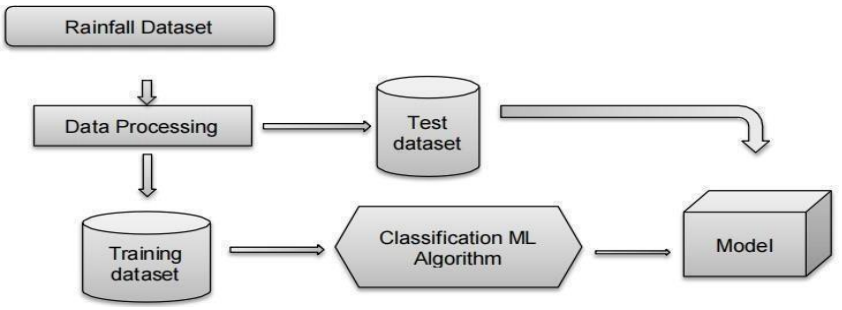



**Fig 3.2 : Architecture of Proposed model**

### 3.2.5 Advantages

1. Performance and accuracy of the algorithms can be calculated and compared.
2. Numerical Weather Prediction
3. Statistical Weather Prediction
4. Synoptic Weather Prediction

## 3.3 SYSTEM ARCHITECTURE

A modular design reduces complexity, facilities change (a critical aspect of software maintainability), and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity that is software is divided in to separately named and addressable components called modules that are integrated to satisfy problem requirements.Modularity is the single attribute of software that allows a program to be intellectually manageable. The five important criteria that

us to evaluate a design method with respect to its ability to define an effective modular design are: Modular decomposability, Modular Composability, Modular Understand ability, Modular continuity, Modular Protection.

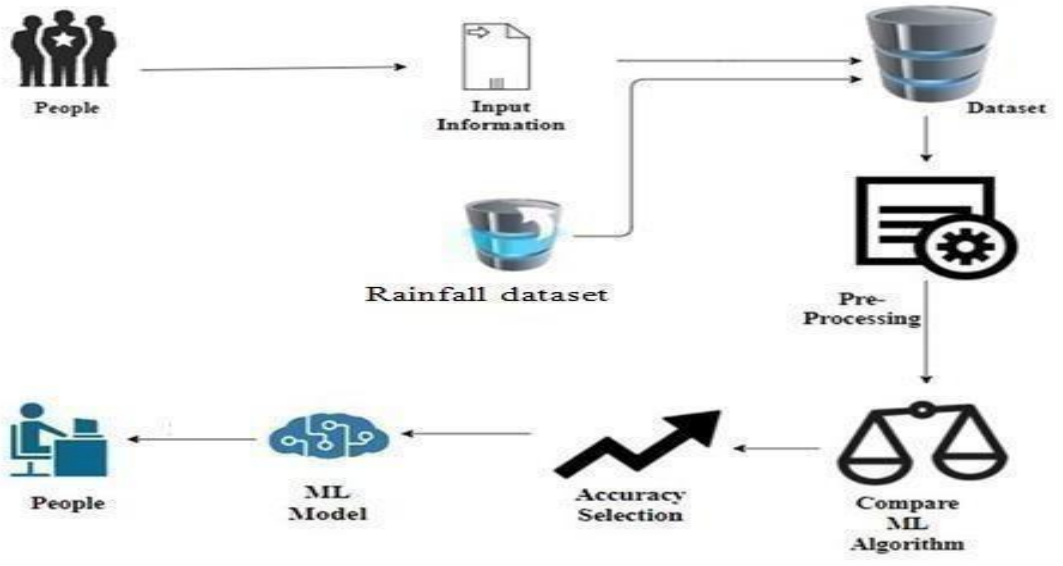



**Fig 3.3 : System architecture**

## 3.4 FEASABILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is Put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be Carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. The feasibility

study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components.

### 3.4.1 Economical feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour in to there search and development of the system is limited. The expenditures must justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

### 3.4.2 Technical feasibility

This study is carried out to check the technical feasibility, that is, the technical Requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands On the available technical resources. This will lead to high demands being Placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

### 3.4.3 Social feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity.

.

## 3.5 REQURIMENTS SPECIFICATIONS

### 3.5.1 Hardware requirements:

| | | |
|---|---|---|
| Processor | : | Intel |
| RAM | : | 2GB |
| Hard Disk | : | 80GB |

### 3.5.2 Software requirements:

| | | |
|---|---|---|
| OS | : | Windows |
| Technology | : | Machine Learning using Python |
| Web Browser | : | Chrome, Microsoft Edge |
| Code editor | : | Visual Studio Code, Google Colab, Anaconda or Jupyter notebook. |

# DESIGN

# 4. DESIGN

## 4.1 FLOW CHART

A flow chart is a diagram that represents a process, work flow, or algorithm. It uses standardized symbols to represent different steps in the process and how they relate to each other. Flow charts are commonly used in software development, business process modeling, and other fields to visually represent the steps of a process for better understanding and analysis. Flow charts can vary in complexity, from simple diagrams that outline a basic process to more detailed charts that include multiple decision points, loops and branching paths.

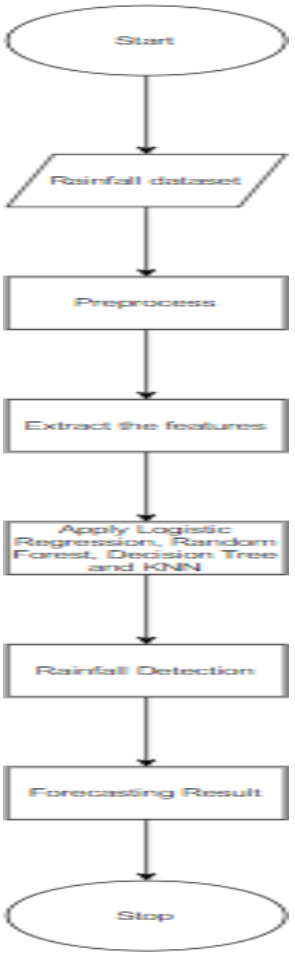



**Fig 4.1 :Flow Chart**

## 4.2 DATA FLOW DIAGRAMS

In Software engineering DFD(data flow diagram) can be drawn to represent the system of different levels of abstraction. Higher-level DFDs are partitioned in to low levels- hacking more information and functional elements. Levels in DFD are numbered 0, 1, 2 or beyond. Here, we will see mainly 3 levels in the data flow diagram, which are:0- level DFD, 1-level DFD,and 2-level DFD. Data Flow Diagrams (DFD) are graphical representation sofa system that illustratethe flow of data with in the system.

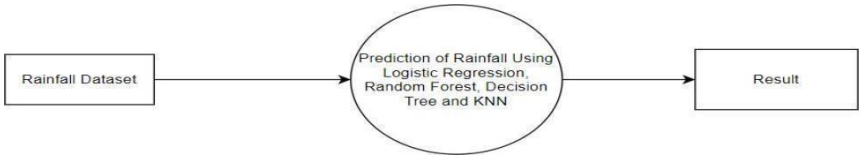



**Fig 4.2 : Level-0 data flow diagram**

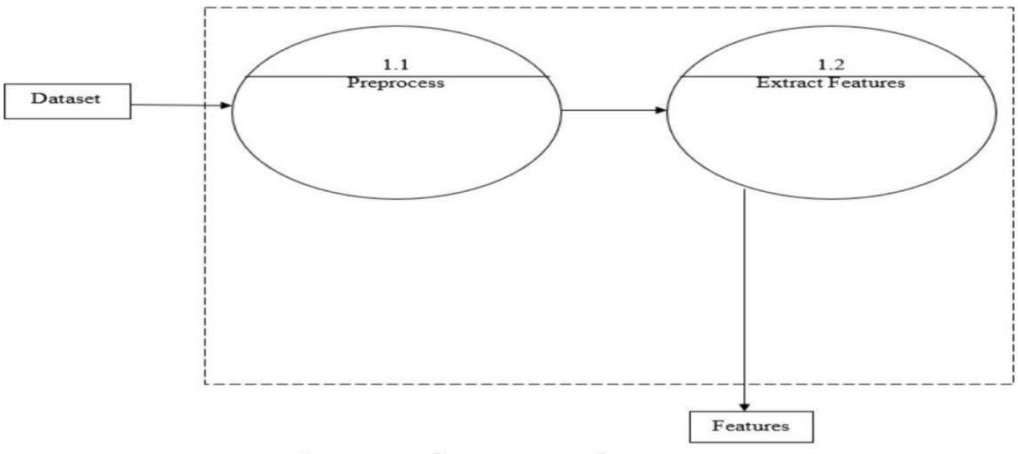



**Fig 4.2 : Level-1 data flow diagram**

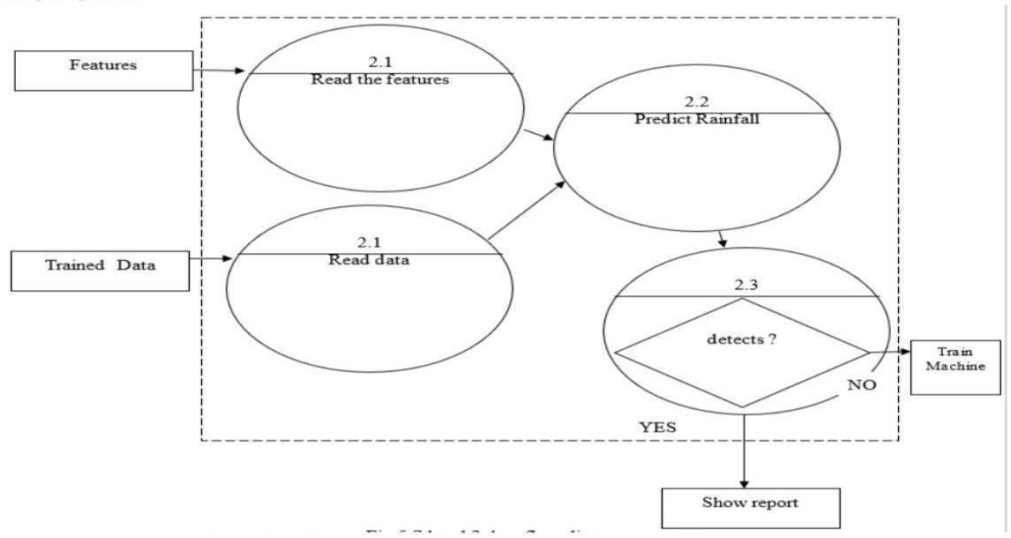



**Fig 4.2 : Level-2 data flow diagram**

## 4.3 UML DIAGRAMS:

. UML stands for Unified Modeling Language. UML is a standardized general- purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.The goal is for UML to become a common language forcreating models of object Oriented computer software. In its current formUML is comprised of two major components: a Meta-model and anotation.In the future, some form of method or process may also be added to; or associated with,UML.The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well As for business modeling and other non- software systems.The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects orientedsoftware and the software development process. The UML uses mostlygraphical notations to express the design of software projects.There are many diagrams like Use case diagram, class diagram, sequence diagram, Activitydiagram, interaction diagram, state machine

diagram, component diagram, object Diagram,deployment diagram, data flow diagram and package diagram. Mainly these diagrams are Divided in to behaviour diagrams, structural diagrams and Also interaction overview diagrams of UML.

## GOALS:

The Primary goals in the design of the UML areas follows:

1. Provide users a ready-to-use, expressive visual modeling Languagesthat they can develop and exchange meaningful models.
2. Provide extensibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frame works, Patterns and components.
7. Integrate best practices.
8. UML diagrams should be understandable to user and stake holders so that they can easily understand about the structure of the whole project.9. They should be very simple not much too complex.

### 4.3.1 Use case diagram

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner.

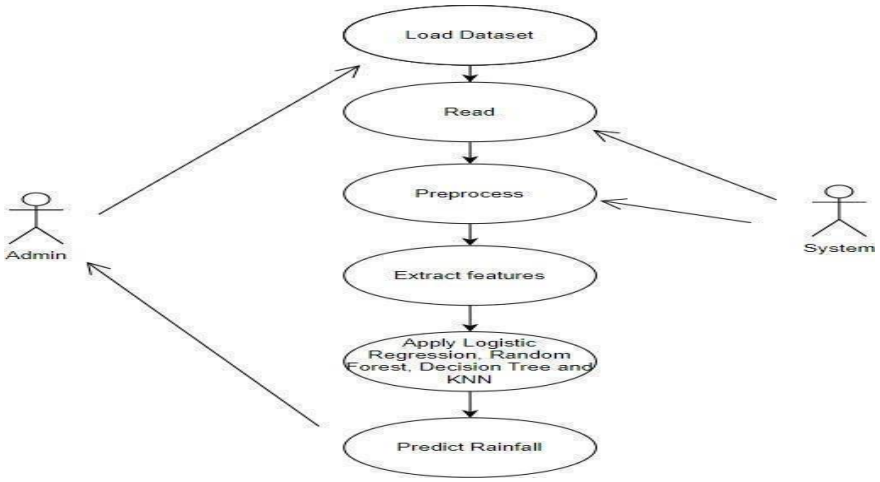



**Fig 4.3.1 : Use case diagram**

### 4.3.2 Class diagram

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.

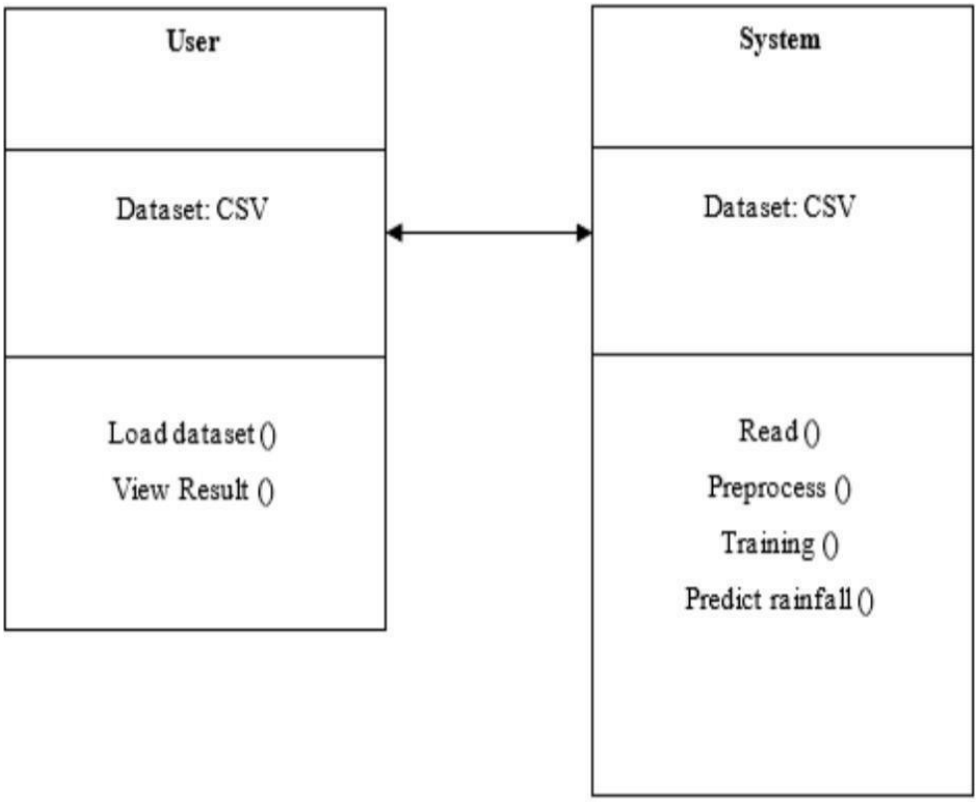



**Fig 4.3.2 : Class diagram**

### 4.3.3 Activity diagram

Activity is a particular operation of the system. Activity diagrams are notonly used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart.

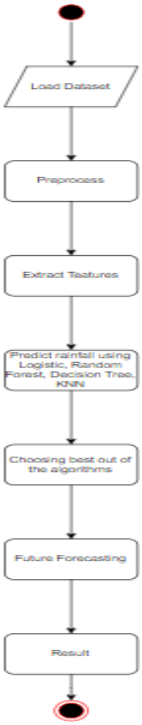



**Fig 4.3.3 : Activity diagram**

### 4.3.4 Sequence diagram

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior with in your system. Other dynamic modeling techniques Include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

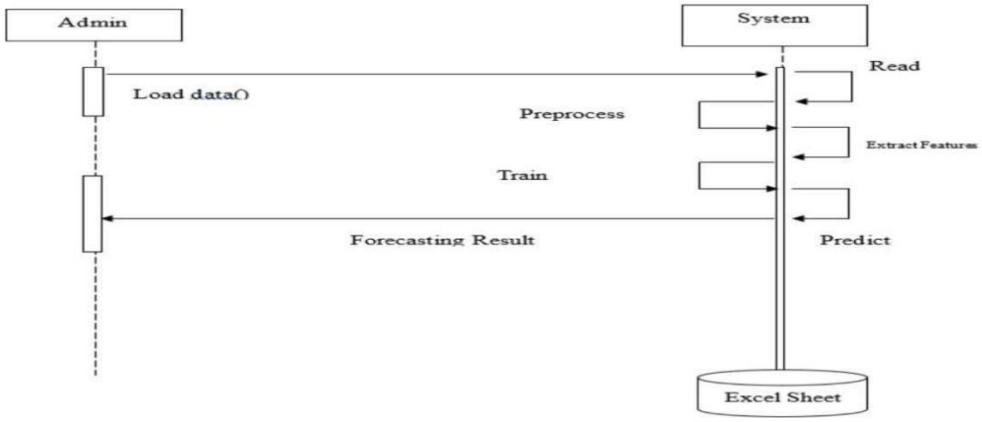



**Fig 4.3.4 : Sequence diagram**

# IMPLEMENTATION

# 5. IMPLEMENTATION

## 5.1 PYTHON

Python is a high-level, interpreted, interactive and object- oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

**Python is Interpreted** − Python is processed at run time by the interpreter. Youdo not need to compile your program before executing it. This is similar to PERL and PHP.

**Python is Interactive** –You can actually it at a Python prompt and interact with the interpreter directly to write your programs.

**Python is Object-Oriented** –Python supports Object-Oriented style or technique of programming that encapsulates code with in objects.

**Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

**History of Python**

Python was developed by Guido van Rossum in the late eighties and earlynineties at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python is derived from any other languages, including ABC, Modula-3,C,C++, Algol-68,Small Talk, and Unix shell.

**Python features**

Python's features include:

**Easy-to-learn**−Python has few key words, simple structure.

**Easy-to-read** −Python code is more clearly defined and visible to the eyes.

**Easy-to-maintain**−Python's source code is fairly easy-to-maintain.

**Abroad standard library**−Python's bulk of the library is very portable and cross plat form compatible on UNIX, Windows, and Macintosh.

**Interactive Mode**−Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

**Portable –** Python can run on a wide variety of hardware plat forms and has the same interface on all platforms.

**Extendable**− You can add low-level modules to the Python interpreter. These Modules enable programmers to add to or customize their tools to be more efficient.

**Databases**− Python provides inter faces to all major commercial data bases.

**GUI Programming**− Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the Window system of Unix.

**Steps to download and install python**

Download the Latest version of the Python executable installer (https://www.python.org/downloads/).Watch the PIP list where pipis the package installer for python. Now upgrade the pip and set up tools using the command.

**Pip install—upgrade pip and Pip install—upgrade setup tools**

**IDE installation for python**

IDE stands for Integrated Development Environment. It is a GUI (Graphical User Interface) where programmers write their code and produce the final products. Best IDE is Pycharm. So download the pycharm new version and install the software.

(https://www.jetbrains.com/pycharm/download/)

**Python libraries needed**

There are many libraries in python. In those we only use few main libraries needed like:

NUMPY LIBRARY
PANDAS LIBRARY
MATPLOT LIBRARY
SEABORN LIBRARY

**Numpy library:**

Numpy is an open-source numerical Python library. NumPy contains a multi- dimensional array and matrix data structures. It can be utilized to perform anumber of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines like mean, mode, standard deviation etc…,

Installation-(https://numpy.org/install/)

**Pip install numpy**

Here we mainly use array, to find mean and standard deviation.

**Pandas library:**

Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called theData Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables. There are several ways to create a Data Frame.

Installation-(https://pandas.pydata.org/getting_started.html)

**Pip install pandas**

Here we use pandas for reading the csv files, for grouping the data, for cleaning the data using some operations.

**Matplotlib library:**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Use interactive figures that can zoom, pan, update,visualize etc..,

Installation-(https://matplotlib.org/users/installing.html

**Pip install matplotlib**

Here we use pyplot mainly for plotting graphs.matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot Function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels.

**Seaborn library:**

Sea born package was developed based on the Matplotlib library . It is used to create more attractive and in formative statistical graphics.While sea born is a different package, it can also be used to develop the attractiveness of matplotlib graphics. Installation-**(https://seaborn.pydata.org/installing.html)**

**Pip install seaborn**

**LOGISTIC REGRESSION :**

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest(dependent variable=response or out come variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).In other words, the logistic regression model predicts P(Y=1) as a function of X.Logistic regression Assumptions:

1. Binary logistic regression requires the dependent variable to be binary.
2. For a binary regression, the fact or level1of the dependent variable should represent the desired outcome.
3. Only the meaningful variables should be included.
4. The independent variables should be independent of each other.That is, the model should have little.
5. The independent variables are linearly related to the log odds.
6. Logistic regression requires quite large sample sizes.

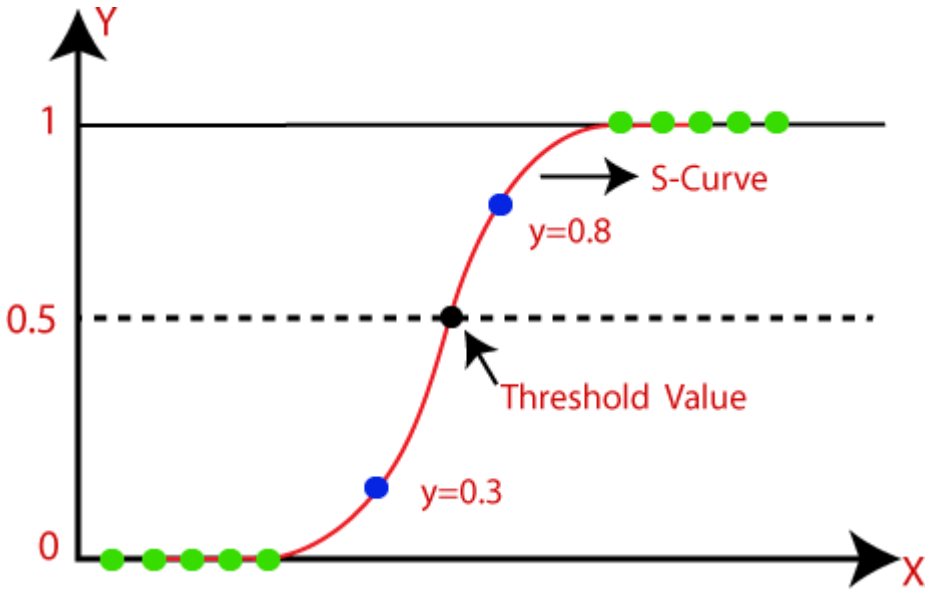



**Fig 5.7.5 : Logistic regression**

GIVEN INPUT EXPECTED OUTPUT

input: data

output: getting accuracy

**RANDOM FOREST :**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction(regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine earning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees,resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.The following are the basic steps involved in performing the randomforest algorithm:

1. Pick N random records from the data set.

2. Build a decision tree based on these N records.

3. Choose the number of trees you want in your algorithm

In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or,in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

GIVEN INPUT EXPECT OUTPUT
input: data
output: getting accuracy

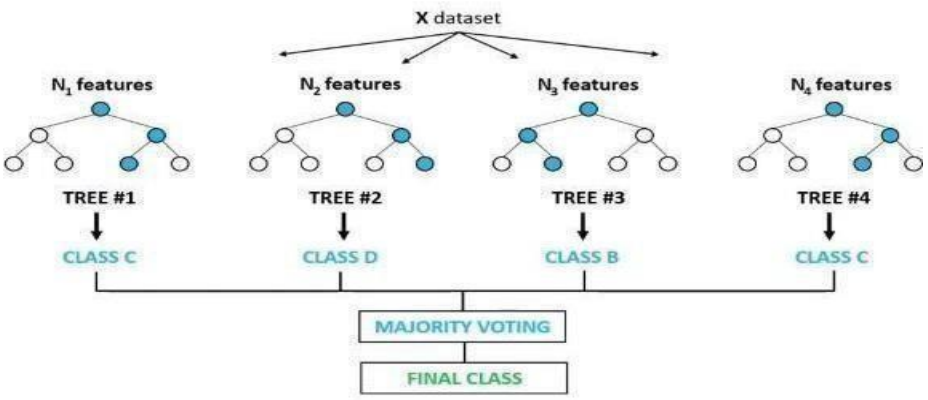



**Fig 5.7.5 : Random forest**

## K-NEAREST NEIGHBOR :

Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case in to the category that is most similar the available categories. Itstores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easilyclassified 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.

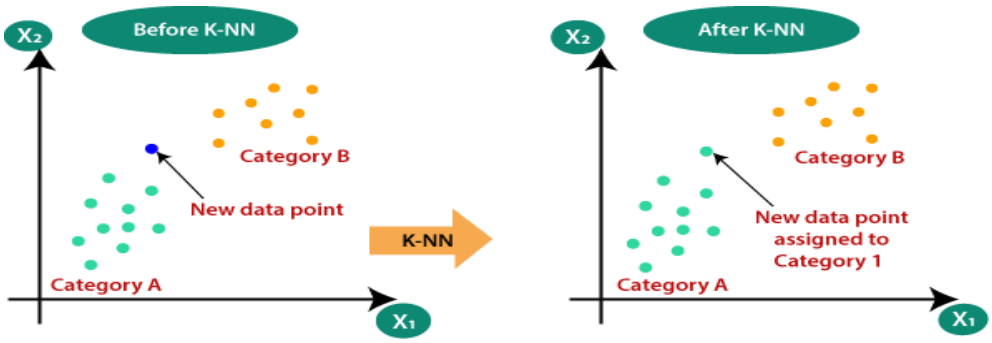



**Fig 5.7.5 :K- Nearest neighbor**

GIVEN INPUT EXPECT OUTPUT

input: data

output: getting accuracy

**DECISION TREE :**

Decision tree builds classification or regression models in the form ofa tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node.Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time.

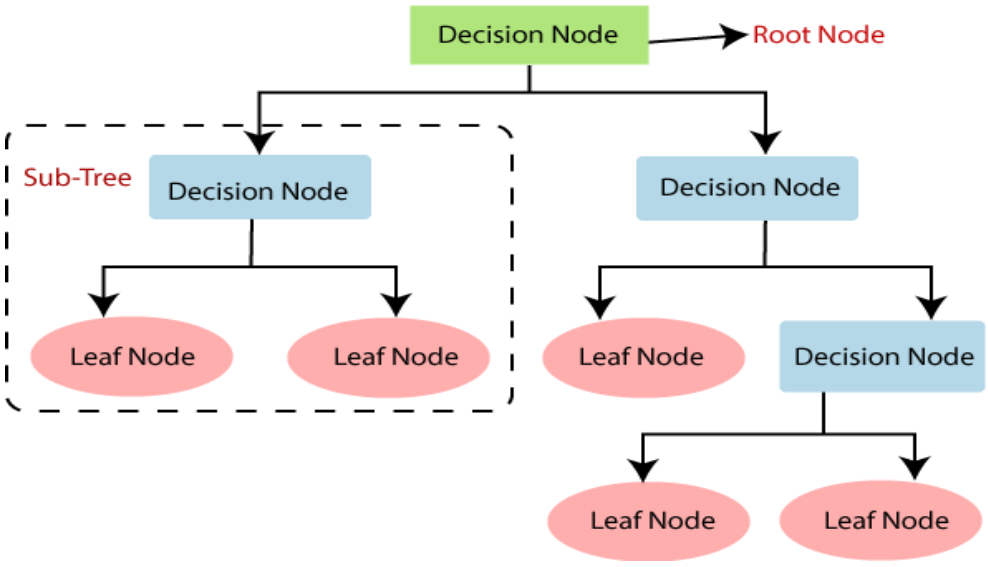



**Fig 5.7.5 : Decision tree**

GIVEN INPUT EXPECT OUTPUT

input: data

output: getting accuracy

## 5.2 Sample Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb

# LOAD THE DATASET
data = pd.read_csv(r"C:\Users\hp\OneDrive\Desktop\weatherAUS.csv")
data.head()
data.shape
data.info()
data.describe().T

#DATA CLEANING
data.isnull().sum()
data.columns
#REMOVING UNNECESSARY SPACES IN COLUMN NAMES
data.rename(str.strip,axis='columns', inplace=True)
data.columns
#EXPLORATORY DATA ANALYSIS
plt.pie(data['RainTomorrow'].value_counts().values,
        labels = data['RainTomorrow'].value_counts().index,
        autopct='%1.1f%%')

plt.show()
```

```
features = list(data.select_dtypes(include = np.number).columns)
print(features)


# PLOT HISTOGRAMS FOR COLUMNS

for i in data.columns:
    if data[i].dtypes == "float64":
        plt.figure(figsize=(4, 3))
        sb.histplot(data[i])
        plt.title(i)
        plt.show()

#PLOT BARPLOTS FOR COLUMNS
for i in data.columns:
    if data[i].dtypes == "float64":
        plt.figure(figsize=(4, 3))
        sb.boxplot(data[i])
        plt.title(i)
        plt.show()

# EXCLUDE NON-NUMERIC COLUMNS FROM CORELATION
CALCULATION
numeric_data = data.select_dtypes(include=[np.number])
plt.figure(figsize=(10,10))
sb.heatmap(numeric_data.corr(),
           annot=True,
           cbar=False)
plt.show()

# CROSSTAB OF RAINTOMORROW AND RAINTODAY
print(pd.crosstab(data["RainTomorrow"],  data["RainToday"]))
```

```
#MODEL TRAINING
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler

# Drop rows with missing values in the target variable
data = data.dropna(subset=['RainTomorrow'])

# Convert 'RainTomorrow' to binary
data['RainTomorrow'] = data['RainTomorrow'].map({'Yes': 1, 'No': 0})

# Convert 'Date' to datetime and extract year, month, and day
data['Date'] = pd.to_datetime(data['Date'])
data['Year'] = data['Date'].dt.year
data['Month'] = data['Date'].dt.month
data['Day'] = data['Date'].dt.day
data = data.drop('Date', axis=1)

# Convert categorical variables to dummy variables
data = pd.get_dummies(data)

# Handle missing values
X = data.drop('RainTomorrow', axis=1)
y = data['RainTomorrow']

imputer = SimpleImputer(strategy='mean')
X = imputer.fit_transform(X)

scaler = StandardScaler()
X = scaler.fit_transform(X)
```

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

#LOGISTIC REGRESSION

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix

# Create a logistic regression model
model = LogisticRegression(max_iter=1000)

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(conf_matrix)

# Plot confusion matrix
plt.figure(figsize=(4, 3))
sb.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
```

```
plt.title('Confusion Matrix')
plt.show()
#RANDOM FOREST

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Create a Random Forest Classifier model
model = RandomForestClassifier(n_estimators=100, random_state=0)

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(conf_matrix)
# Plot confusion matrix
plt.figure(figsize=(4, 3))
sb.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('Confusion Matrix')
plt.show()
#KNN
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Create a KNN Classifier model
model = KNeighborsClassifier()

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(conf_matrix)
# Plot confusion matrix
plt.figure(figsize=(4, 3))
sb.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('Confusion Matrix')
plt.show()
#DECISION TREE

from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Create a Decision Tree Classifier model
```

```
model = DecisionTreeClassifier(random_state=0)

# Train the model on the training set
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print('Confusion Matrix:')
print(conf_matrix)
# Plot confusion matrix
plt.figure(figsize=(4, 3))
sb.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g', cbar=False)
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('Confusion Matrix')
plt.show()
#MODEL EVALUATION

#plot ROC curve

from sklearn.metrics import roc_curve, roc_auc_score

# Calculate probabilities for positive class
y_prob = model.predict_proba(X_test)[:, 1]

# Calculate ROC curve
```

```
fpr, tpr, _ = roc_curve(y_test, y_prob)

# Calculate AUC score
roc_auc = roc_auc_score(y_test, y_prob)

# Plot ROC curve
plt.figure(figsize=(4, 3))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve
(area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()

from sklearn.metrics import classification_report
# Generate the classification report
report = classification_report(y_test, y_pred, output_dict=True)
report_df = pd.DataFrame(report).transpose()

# Plot the classification report using a heatmap
plt.figure(figsize=(6, 4))
sb.heatmap(report_df.iloc[:-1, :].T, annot=True, fmt=".2f", cmap="Blues")
plt.title("Classification Report")
plt.show()
```

# OUTPUT SCREENS

# 6. OUTPUT SCREENS

| | Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustDir | WindGustSpeed | WindDir9am | ... | Humidity3pm | Pressure9am | Pressure3pm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | NaN | NaN | W | 44.0 | W | ... | 22.0 | 1007.7 | 1007.1 |
| 1 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | NaN | NaN | WNW | 44.0 | NNW | ... | 25.0 | 1010.6 | 1007.8 |
| 2 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | NaN | NaN | WSW | 46.0 | W | ... | 30.0 | 1007.6 | 1008.7 |
| 3 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NaN | NaN | NE | 24.0 | SE | ... | 16.0 | 1017.6 | 1012.8 |
| 4 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | NaN | NaN | W | 41.0 | ENE | ... | 33.0 | 1010.8 | 1006.0 |

5 rows × 24 columns

**Fig 6.1 : Data set**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 24 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           142193 non-null  object
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RISK_MM        142193 non-null  float64
 23  RainTomorrow   142193 non-null  object
dtypes: float64(17), object(7)
memory usage: 26.0+ MB
```

**Fig 6.2 : Info of Data set**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **MinTemp** | 141556.0 | 12.186400 | 6.403283 | -8.5 | 7.6 | 12.0 | 16.8 | 33.9 |
| **MaxTemp** | 141871.0 | 23.226784 | 7.117618 | -4.8 | 17.9 | 22.6 | 28.2 | 48.1 |
| **Rainfall** | 140787.0 | 2.349974 | 8.465173 | 0.0 | 0.0 | 0.0 | 0.8 | 371.0 |
| **Evaporation** | 81350.0 | 5.469824 | 4.188537 | 0.0 | 2.6 | 4.8 | 7.4 | 145.0 |
| **Sunshine** | 74377.0 | 7.624853 | 3.781525 | 0.0 | 4.9 | 8.5 | 10.6 | 14.5 |
| **WindGustSpeed** | 132923.0 | 39.984292 | 13.588801 | 6.0 | 31.0 | 39.0 | 48.0 | 135.0 |
| **WindSpeed9am** | 140845.0 | 14.001988 | 8.893337 | 0.0 | 7.0 | 13.0 | 19.0 | 130.0 |
| **WindSpeed3pm** | 139563.0 | 18.637576 | 8.803345 | 0.0 | 13.0 | 19.0 | 24.0 | 87.0 |
| **Humidity9am** | 140419.0 | 68.843810 | 19.051293 | 0.0 | 57.0 | 70.0 | 83.0 | 100.0 |
| **Humidity3pm** | 138583.0 | 51.482606 | 20.797772 | 0.0 | 37.0 | 52.0 | 66.0 | 100.0 |
| **Pressure9am** | 128179.0 | 1017.653758 | 7.105476 | 980.5 | 1012.9 | 1017.6 | 1022.4 | 1041.0 |
| **Pressure3pm** | 128212.0 | 1015.258204 | 7.036677 | 977.1 | 1010.4 | 1015.2 | 1020.0 | 1039.6 |
| **Cloud9am** | 88536.0 | 4.437189 | 2.887016 | 0.0 | 1.0 | 5.0 | 7.0 | 9.0 |
| **Cloud3pm** | 85099.0 | 4.503167 | 2.720633 | 0.0 | 2.0 | 5.0 | 7.0 | 9.0 |
| **Temp9am** | 141289.0 | 16.987509 | 6.492838 | -7.2 | 12.3 | 16.7 | 21.6 | 40.2 |
| **Temp3pm** | 139467.0 | 21.687235 | 6.937594 | -5.4 | 16.6 | 21.1 | 26.4 | 46.7 |
| **RISK_MM** | 142193.0 | 2.360682 | 8.477969 | 0.0 | 0.0 | 0.0 | 0.8 | 371.0 |

**Fig 6.3 : Describing the Data set**

```
Date                0
Location            0
MinTemp           637
MaxTemp           322
Rainfall         1406
Evaporation     60843
Sunshine        67816
WindGustDir      9330
WindGustSpeed    9270
WindDir9am      10013
WindDir3pm       3778
WindSpeed9am     1348
WindSpeed3pm     2630
Humidity9am      1774
Humidity3pm      3610
Pressure9am     14014
Pressure3pm     13981
Cloud9am        53657
Cloud3pm        57094
Temp9am           904
Temp3pm          2726
RainToday        1406
RISK_MM             0
RainTomorrow        0
dtype: int64
```

**Fig 6.4 : Data Cleaning**

```
Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',
       'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',
       'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
       'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',
       'Temp3pm', 'RainToday', 'RISK_MM', 'RainTomorrow'],
      dtype='object')
```

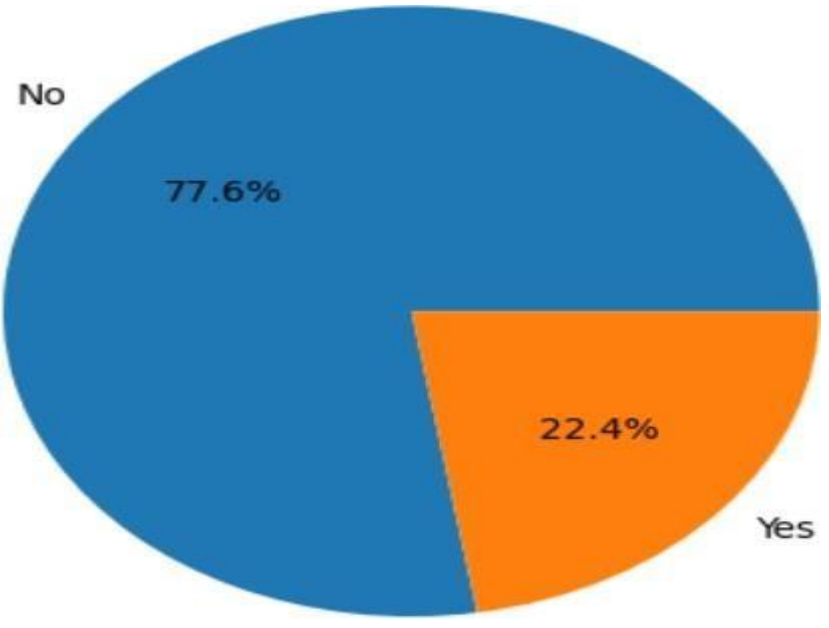
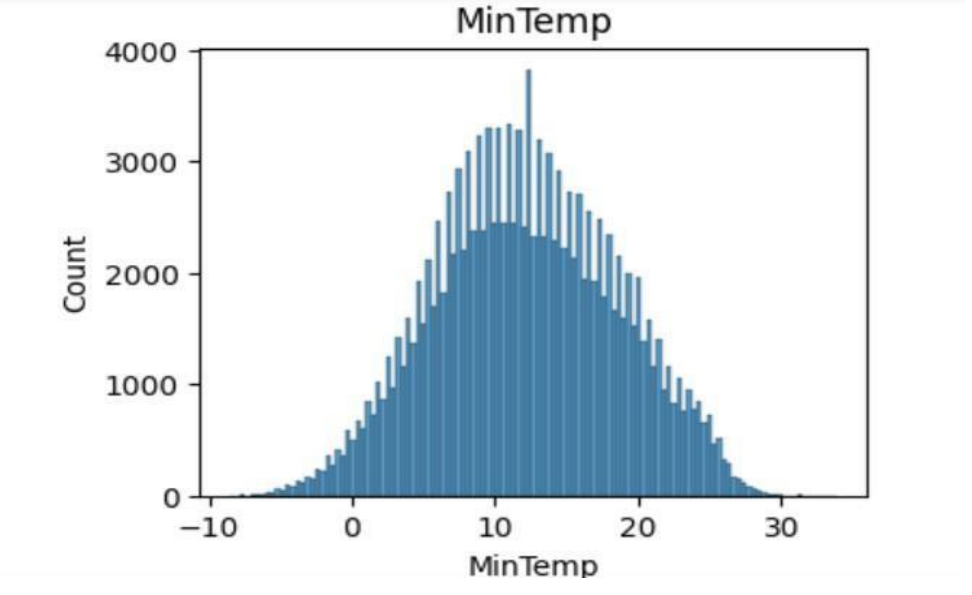
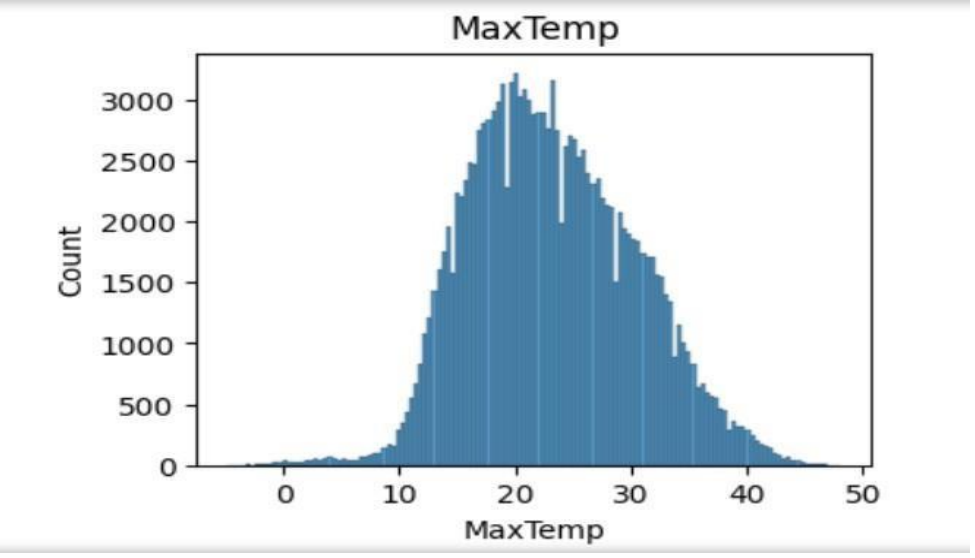
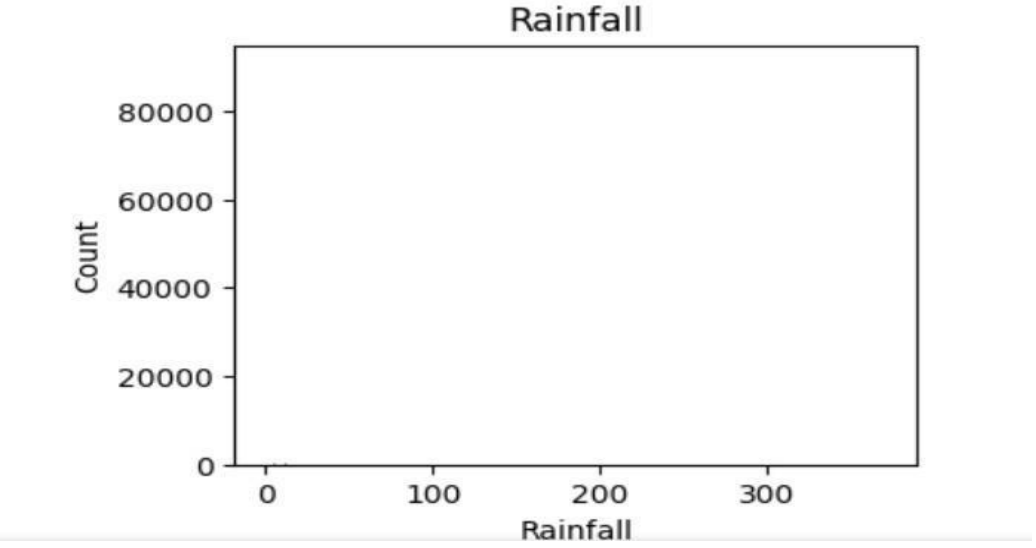
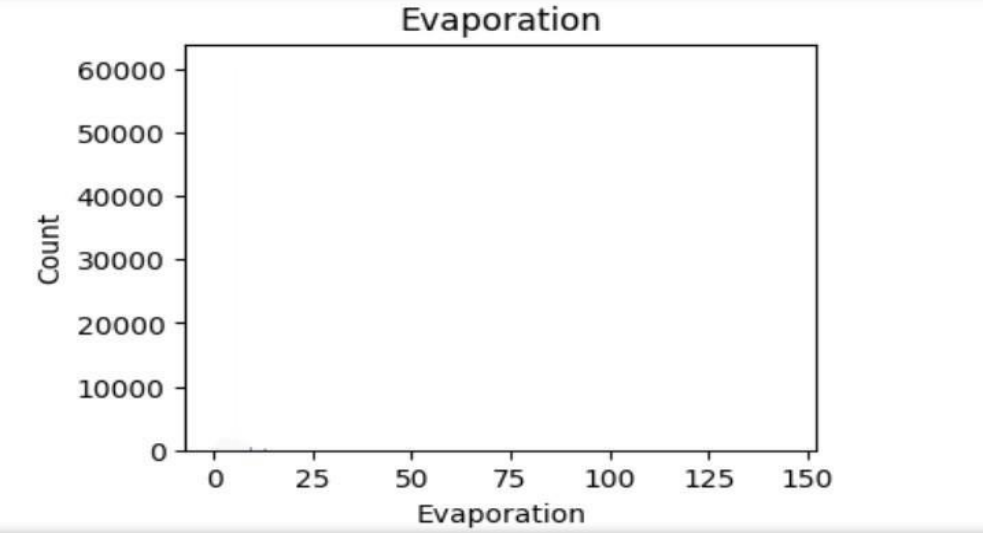


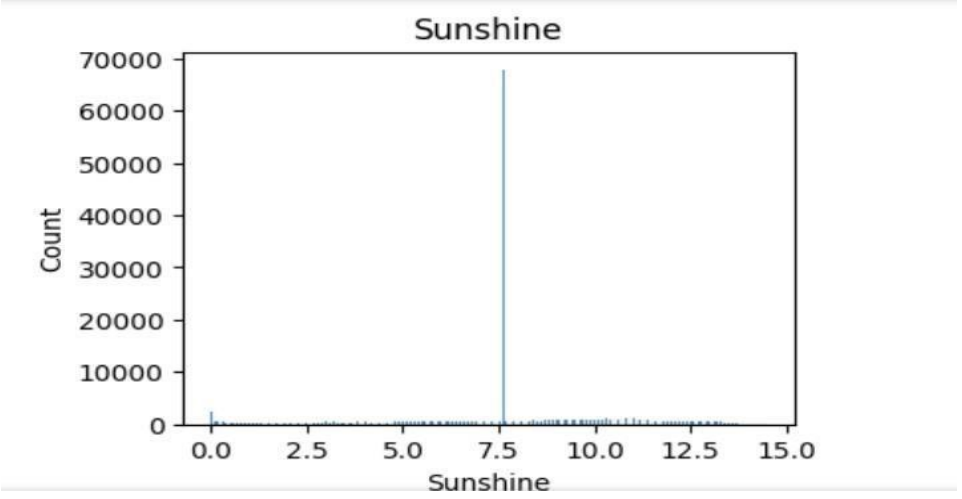

**Fig 6.5 : Exploratory Data Analysis**

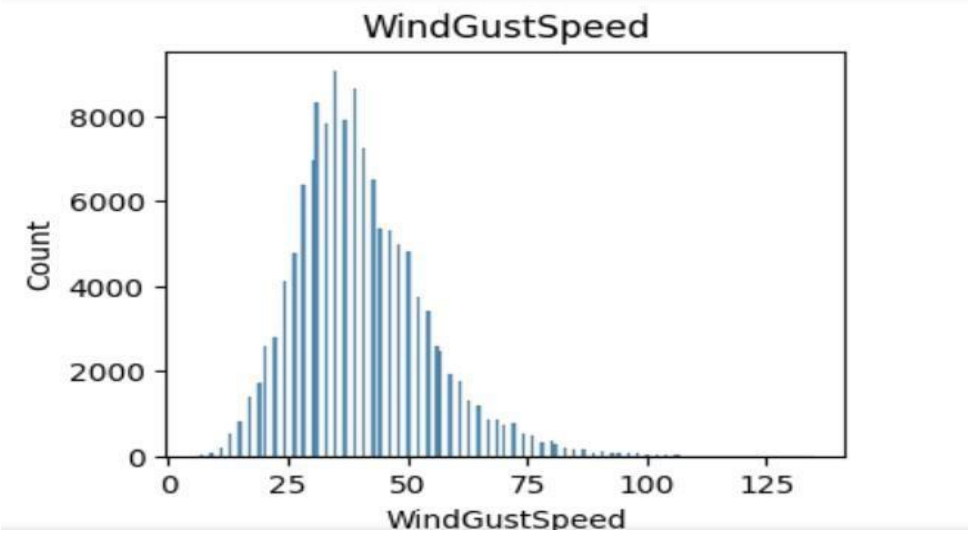


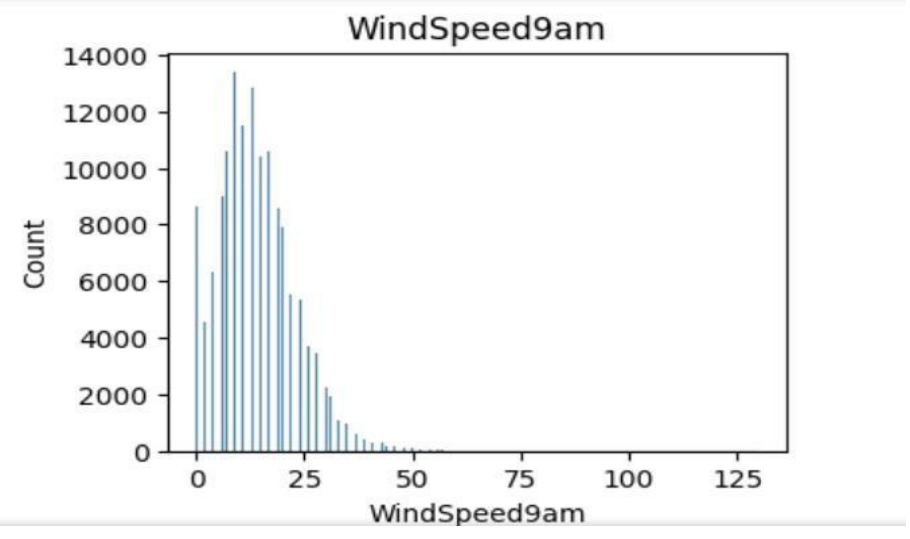

MaxTemp
Count
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
MaxTemp

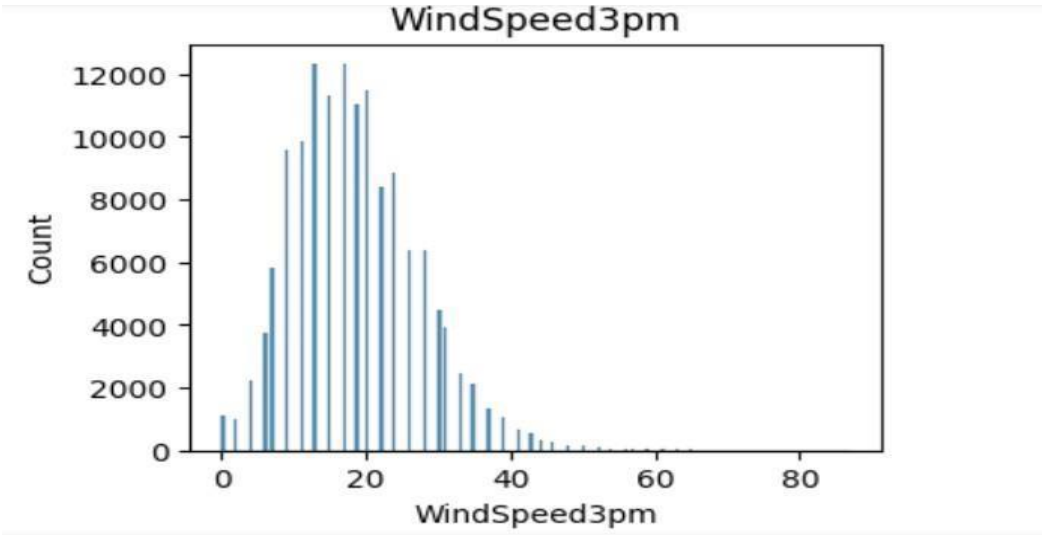

Rainfall
Count
80000
60000
40000
20000
0
0
100
200
300
Rainfall

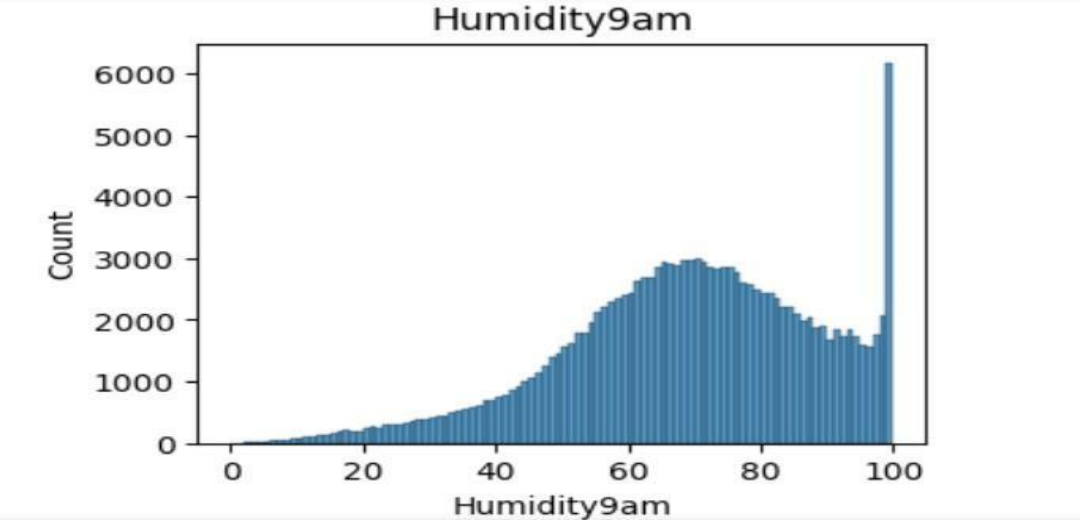

Evaporation
Count
60000
50000
40000
30000
20000
10000
0
0
25
50
75
100
125
150
Evaporation

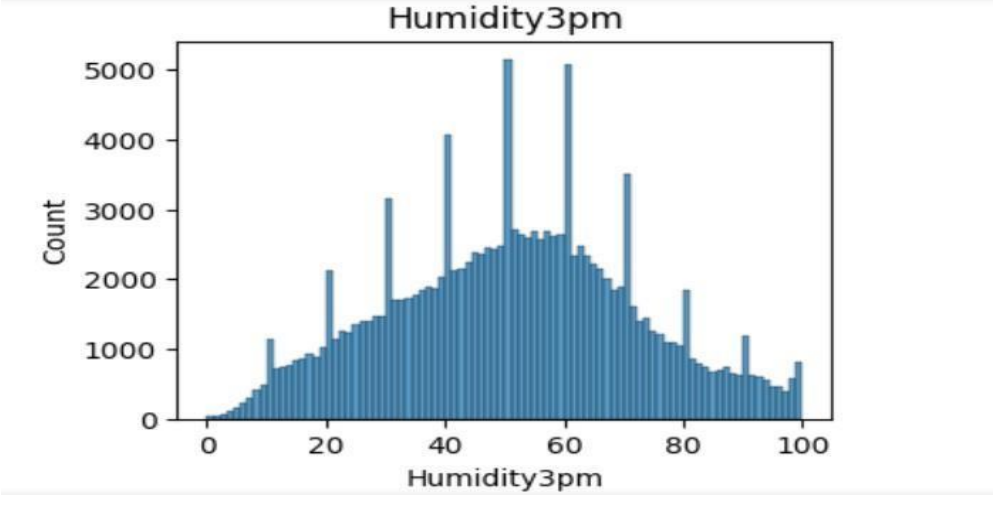

Sunshine
70000
60000
50000
40000
30000
20000
10000
0
Count
0.0
2.5
5.0
7.5
10.0
12.5
15.0
Sunshine

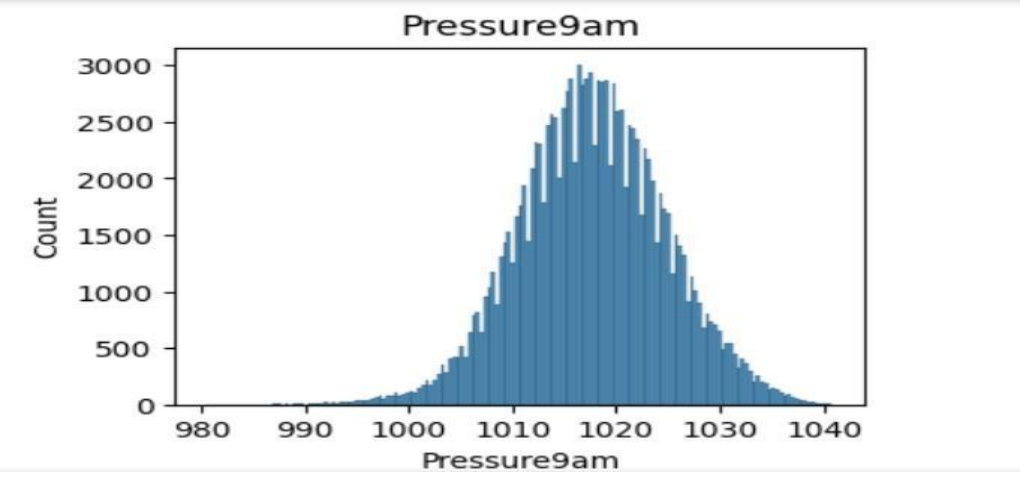

WindGustSpeed
8000
6000
4000
2000
0
Count
0
25
50
75
100
125
WindGustSpeed

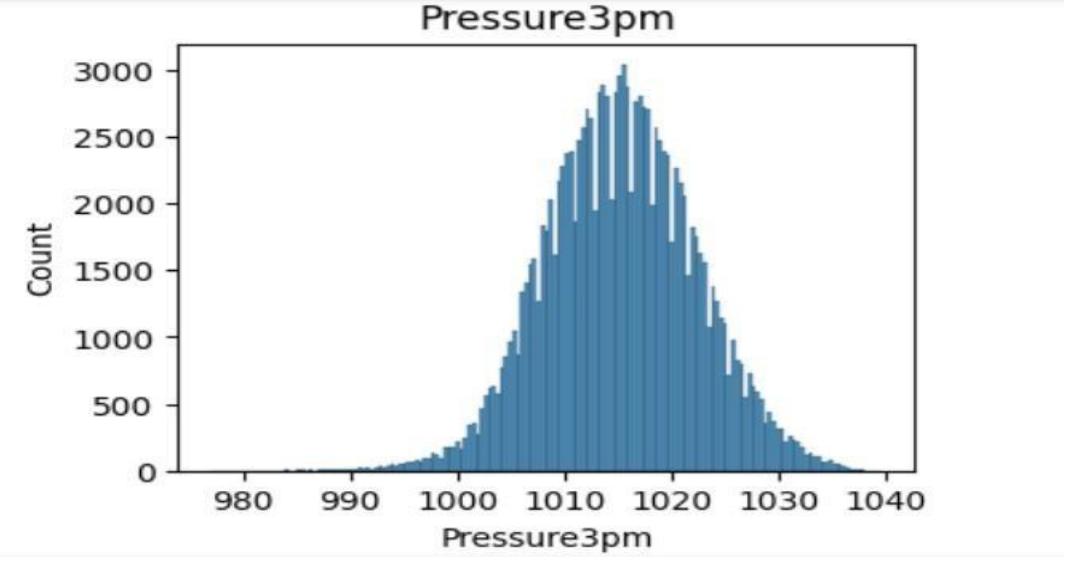

WindSpeed9am
14000
12000
10000
8000
6000
4000
2000
0
Count
0
25
50
75
100
125
WindSpeed9am

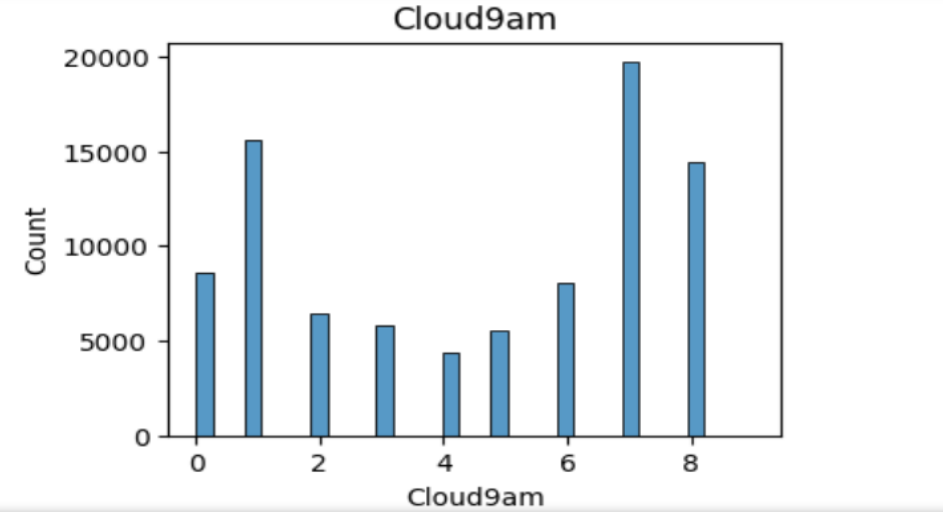

WindSpeed3pm
Count
12000
10000
8000
6000
4000
2000
0
0
20
40
60
80
WindSpeed3pm

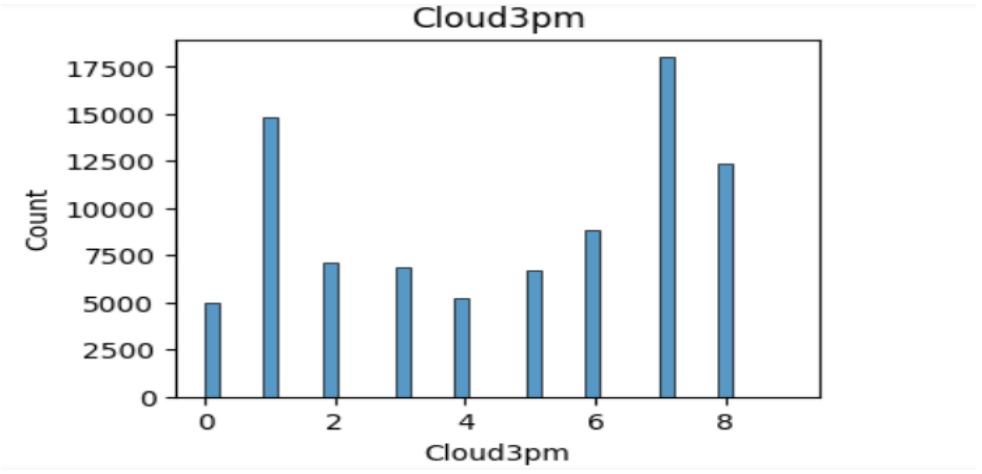

Humidity9am
Count
6000
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
Humidity9am

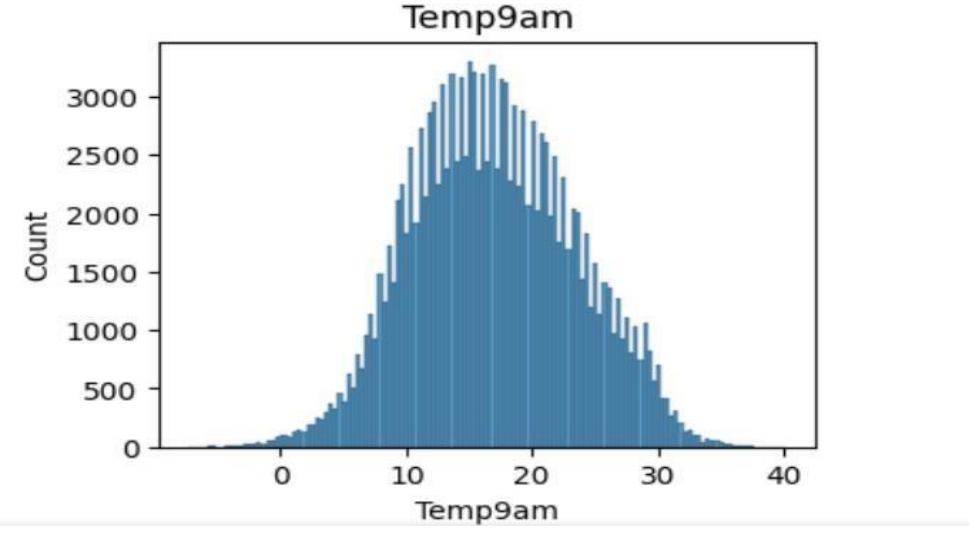

Humidity3pm
Count
5000
4000
3000
2000
1000
0
0
20
40
60
80
100
Humidity3pm

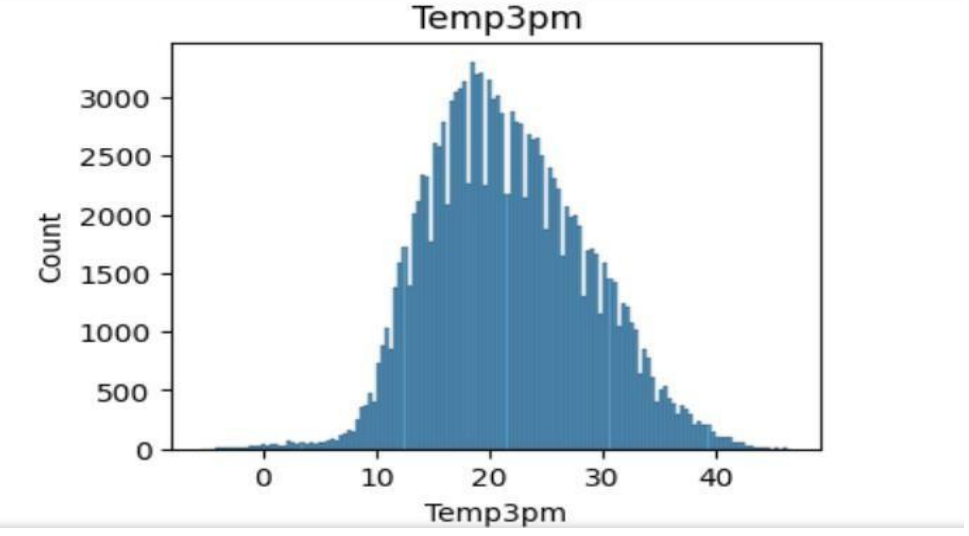

Pressure9am
Count
3000
2500
2000
1500
1000
500
0
980
990
1000
1010
1020
1030
1040
Pressure9am

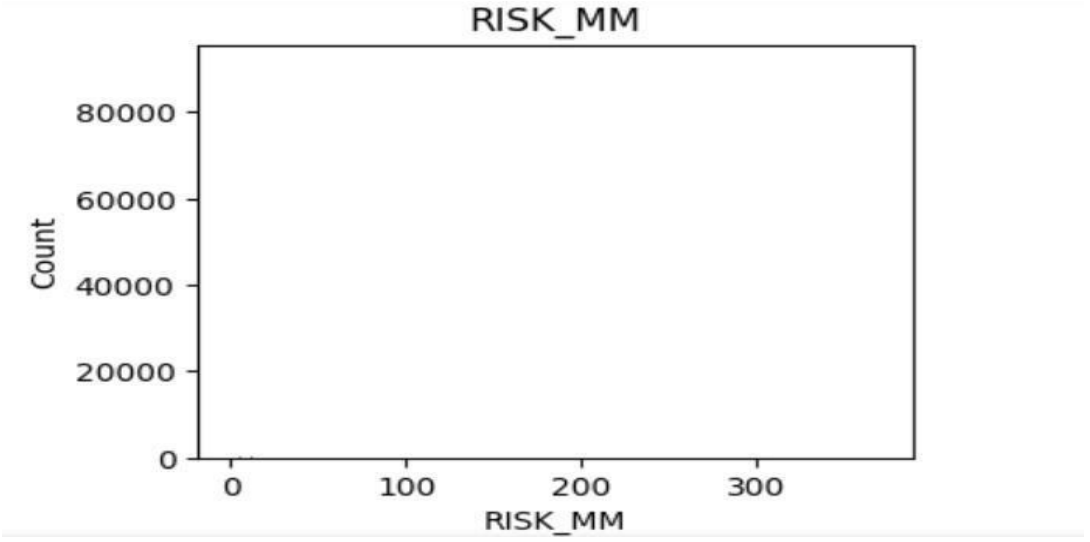

Pressure3pm
Count
3000
2500
2000
1500
1000
500
0
980
990
1000
1010
1020
1030
1040
Pressure3pm

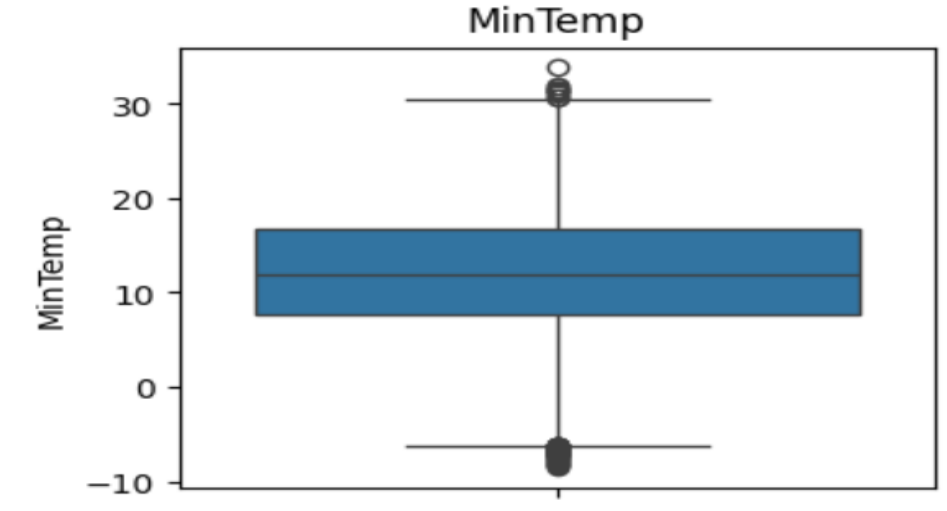

Cloud9am
Count
20000
15000
10000
5000
0
0
2
4
6
8
Cloud9am

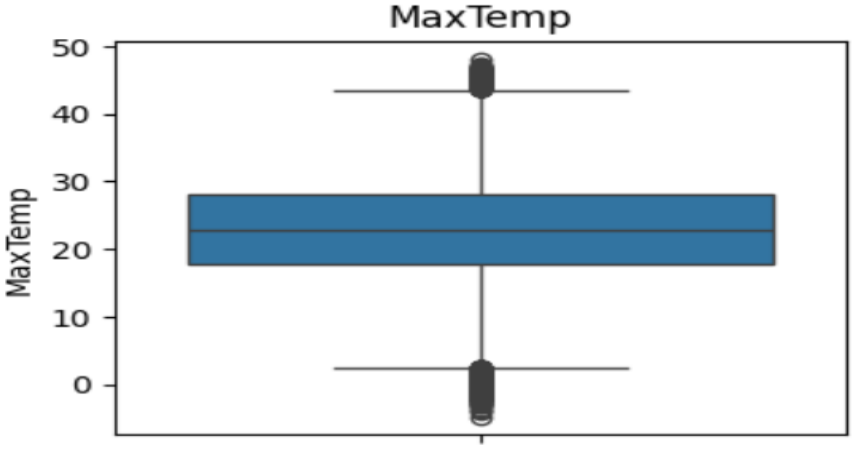

Cloud3pm
17500
15000
12500
10000
7500
5000
2500
0
Count
0
2
4
6
8
Cloud3pm

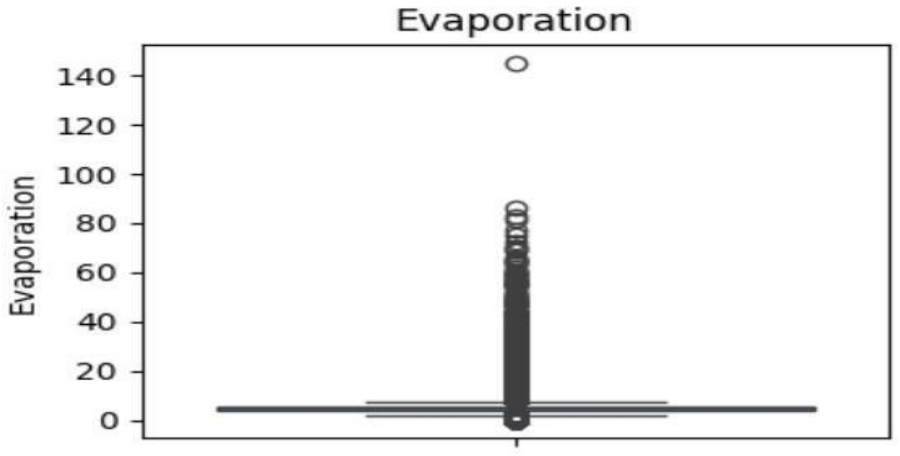

Temp9am
3000
2500
2000
1500
1000
500
0
Count
0
10
20
30
40
Temp9am

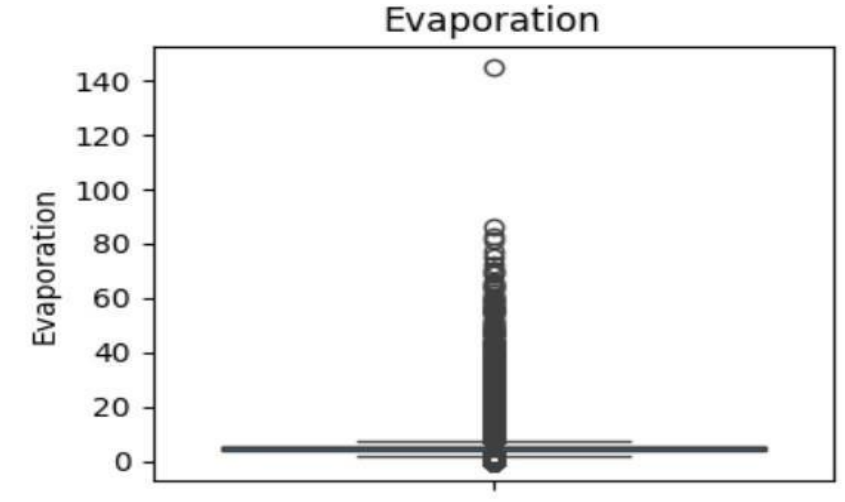

Temp3pm
3000
2500
2000
1500
1000
500
0
Count
0
10
20
30
40
Temp3pm

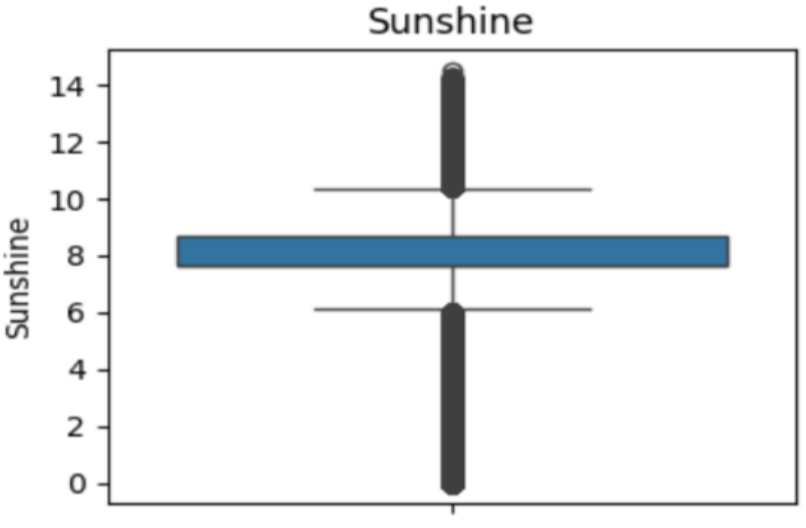

RISK_MM
80000
60000
40000
20000
0
Count
0
100
200
300
RISK_MM

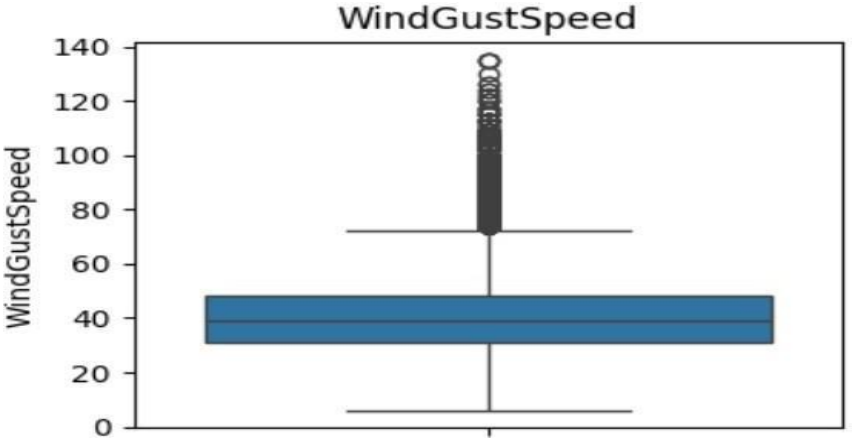

MinTemp
30
20
10
0
−10
MinTemp

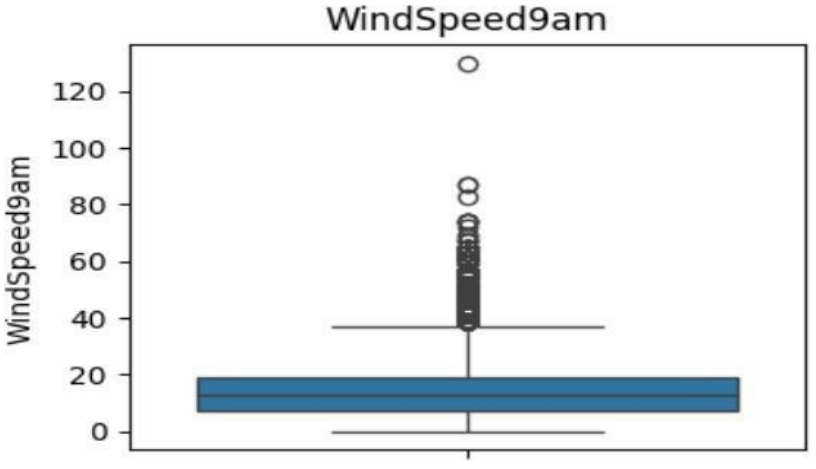

MaxTemp
50
40
30
20
10
0
MaxTemp

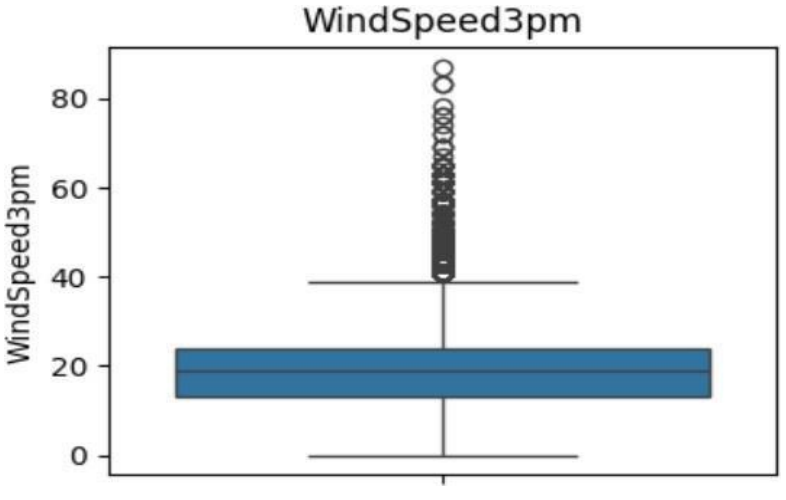

Evaporation
Evaporation
140
120
100
80
60
40
20
0

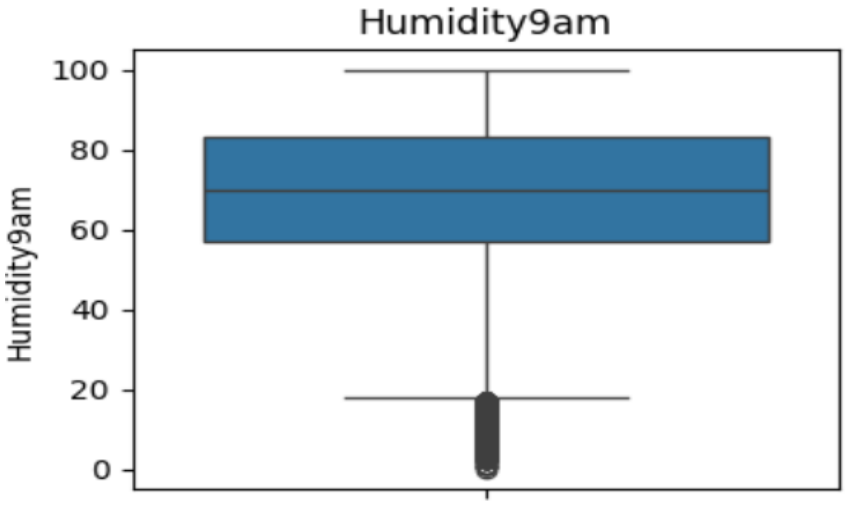

Evaporation
Evaporation
140
120
100
80
60
40
20
0

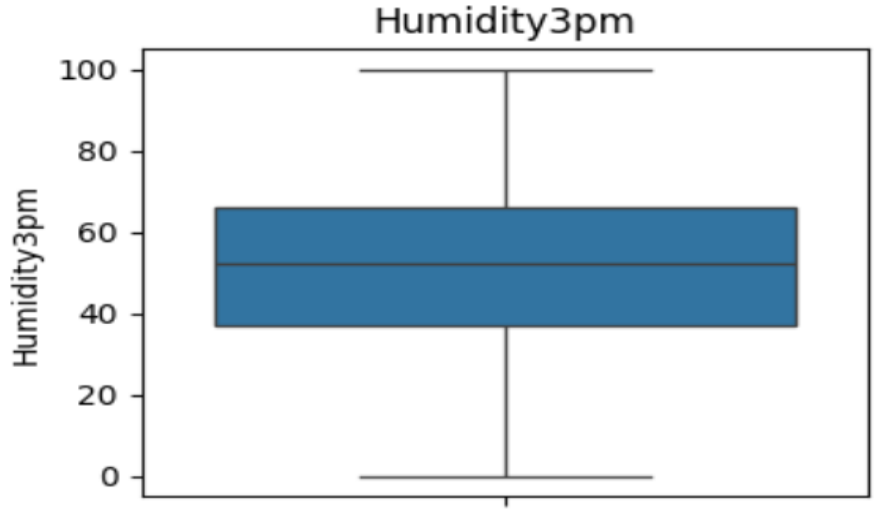

Sunshine
Sunshine
14
12
10
8
6
4
2
0

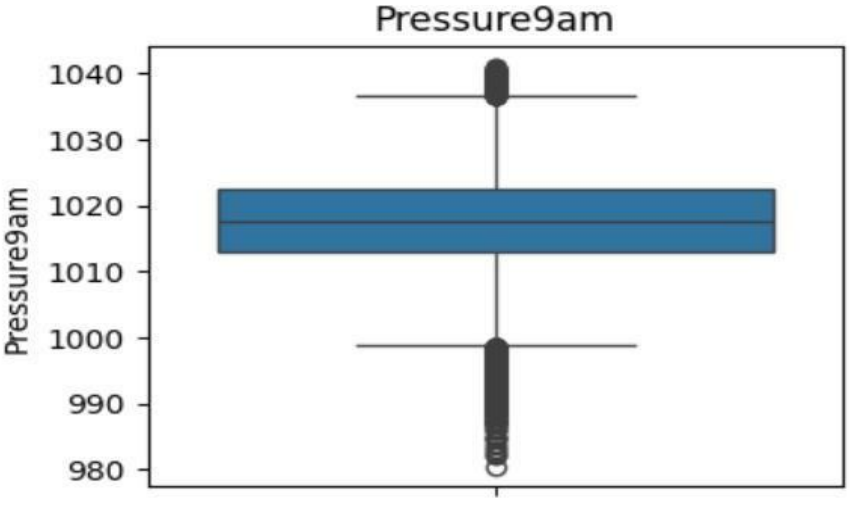

WindGustSpeed
WindGustSpeed
140
120
100
80
60
40
20
0

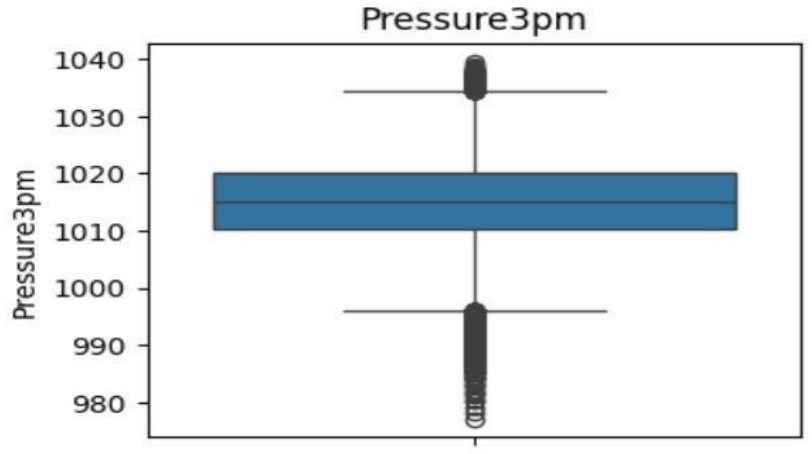

WindSpeed9am
WindSpeed9am
120
100
80
60
40
20
0

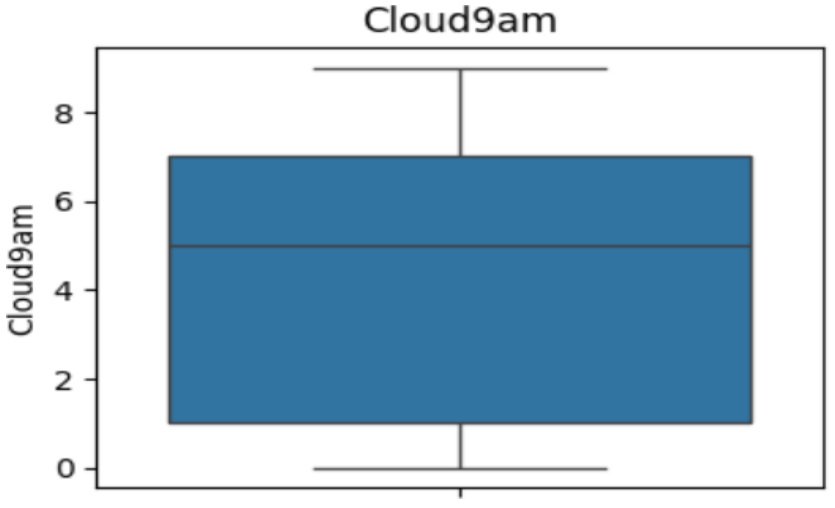

WindSpeed3pm
WindSpeed3pm
80
60
40
20
0

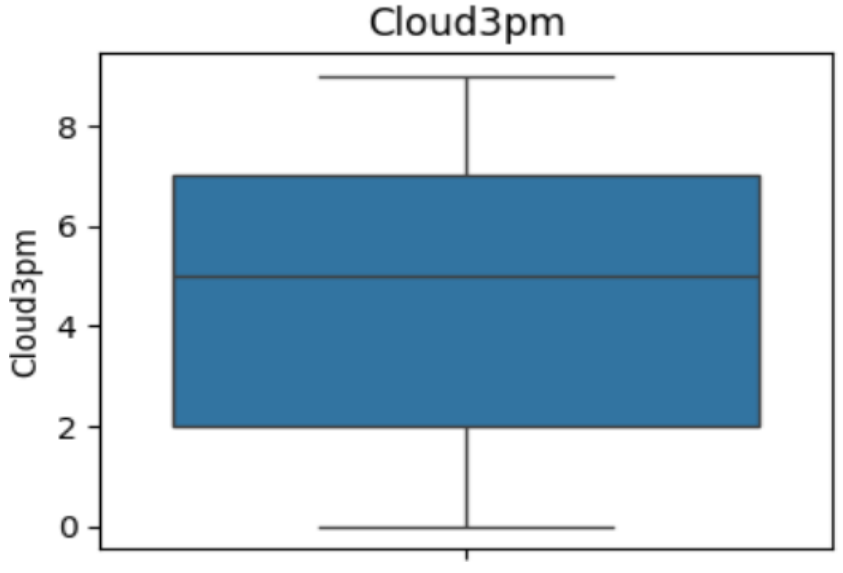

Humidity9am
Humidity9am
100
80
60
40
20
0

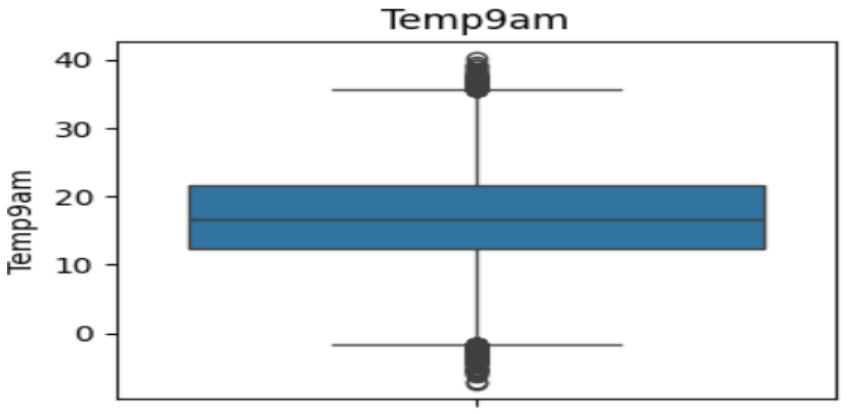

Humidity3pm
Humidity3pm
100
80
60
40
20
0

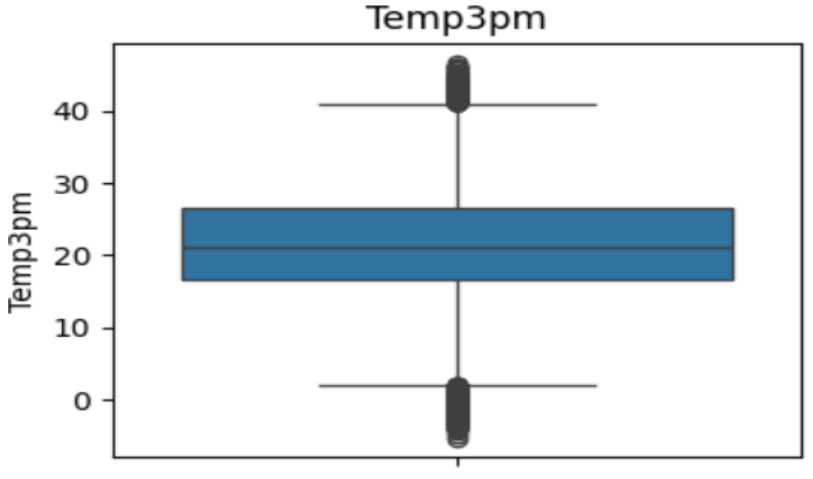

Pressure9am
Pressure9am
1040
1030
1020
1010
1000
990
980

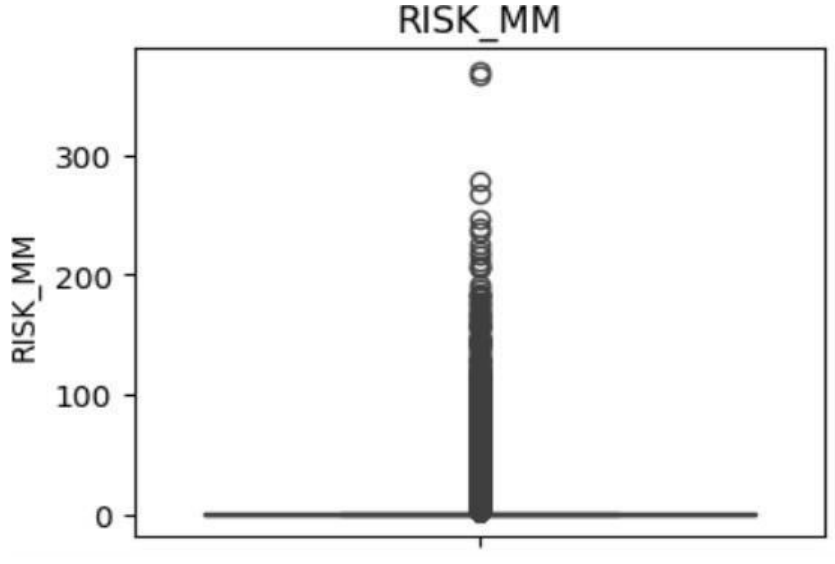

Pressure3pm
Pressure3pm
1040
1030
1020
1010
1000
990
980


Cloud9am
Cloud9am
8
6
4
2
0


Cloud3pm
Cloud3pm
8
6
4
2
0


Temp9am
Temp9am
40
30
20
10
0


Temp3pm
Temp3pm
40
30
20
10
0


RISK_MM
RISK_MM
300
200
100
0

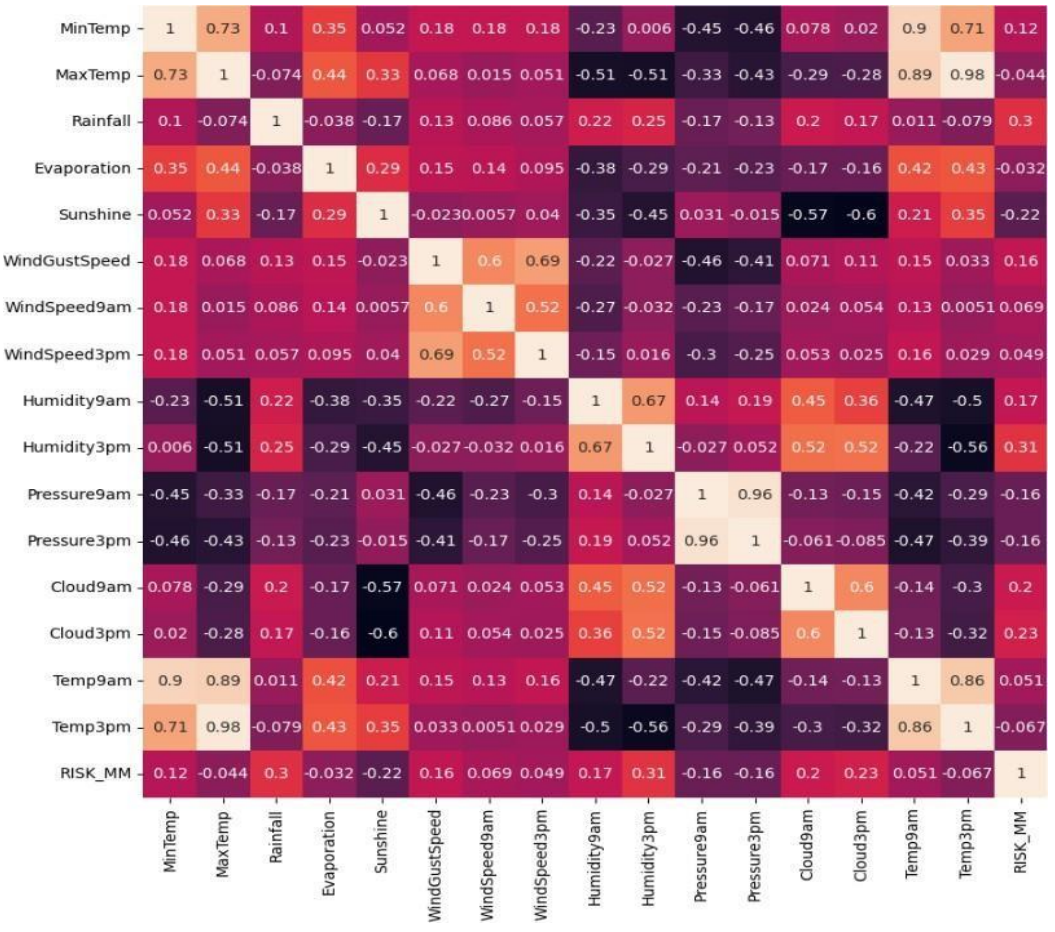



**Fig 6.6 : Exclude non numeric columns from correlation calculation**

| RainToday | No | Yes |
|---|---|---|
| RainTomorrow | | |
| No | 92728 | 16858 |
| Yes | 16604 | 14597 |

**Fig 6.7 : Cross tab of RainTomorrow and RainToday**

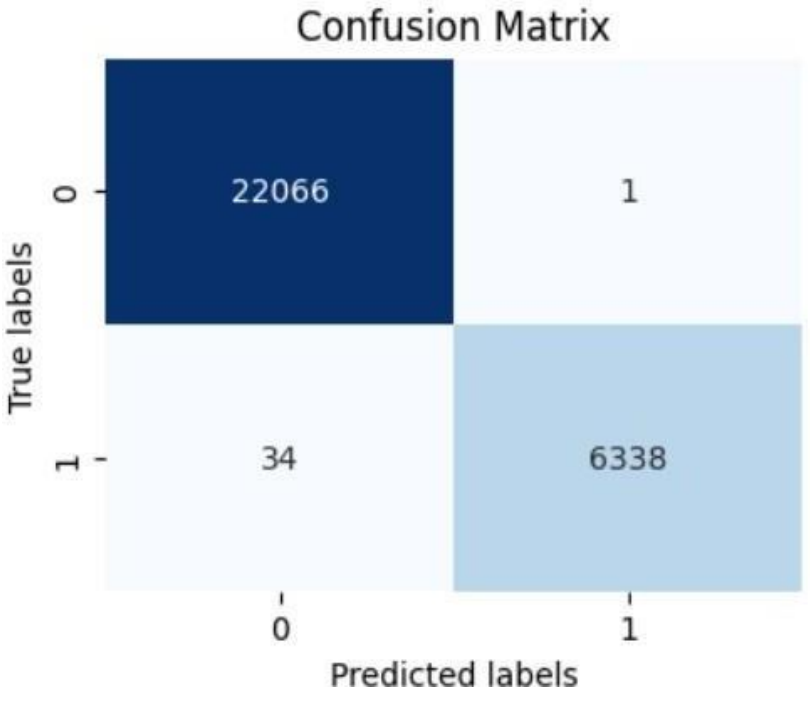



**Fig 6.8 : Confusion Matrix of Logistic Regression**

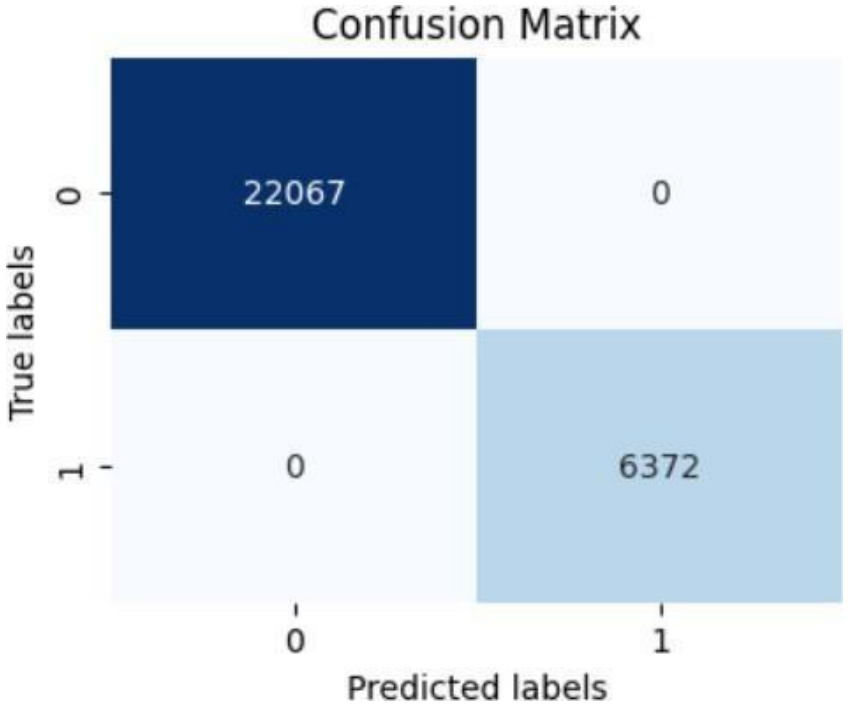



**Fig 6.9 : Confusion Matrix of Random Forest**

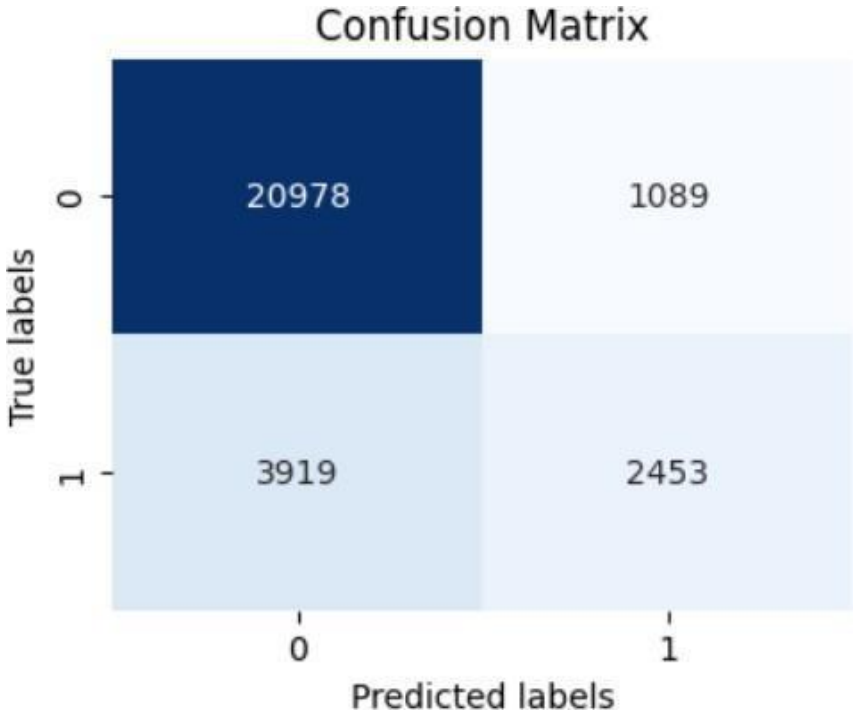



**Fig 6.10 : Confusion Matrix of K-Nearest Neighbor**

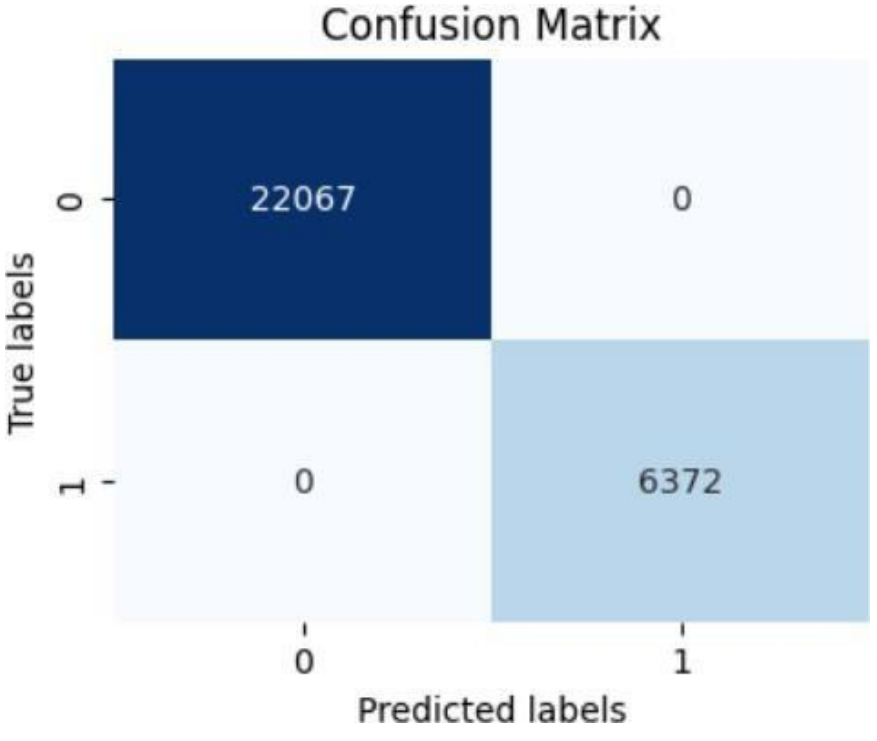



**Fig 6.11 : Confusion Matrix of Decision Tree**

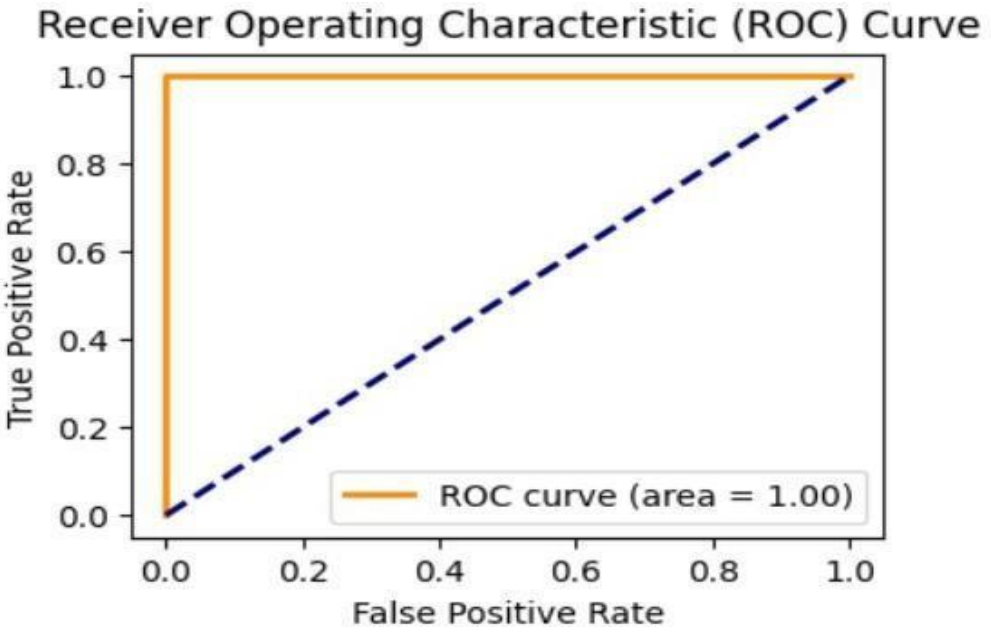



**Fig 6.12 : ROC Curve of Decision Tree**


**Fig 6.13 : Classification Report of Decision Tree**

# SYSTEM TESTING

# 7. SYSTEM TESTING

## 7.1 TESTING PLAN

Software testing is the process of evaluation a software item to detect differences between given input and expected output. Also to assess the feature of a software item. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

### Verification :

Verification is the process to make sure the product satisfies the conditions imposed at the start of the development phase. In other words, to make sure the product behaves the way we want it to.

### Validation :

Validation is the process to make sure the product satisfies the specified requirements at the end of the development phase. In other words, to make sure the product is built as per customer requirements.

## 7.2 BASICS OF SOFTWARE TESTING

There are two basics of software testing: black box testing and white box testing.

**Black box testing :**

Black box testing is a testing technique that ignores the internal mechanism of the system and focuses on the output generated against any input and execution of the system. It is also called functional testing.

**White box testing :**

White box testing is a testing technique that takes into account the internal mechanism of a system. It is also called structural testing and glass box testing. Black box testing is often used for validation and white box testing is often usedfor verification.

## 7.3 TYPES OF TESTING :

There are many types of testing like :
Unit Testing
Integration Testing
Functional Testing
System Testing
Stress Testing
Performance Testing
Usability Testing
Acceptance Testing
Regression Testing
Beta Testing

**Unit Testing :**

Unit testing is the testing of an individual unit or group of related units. It falls under the class of white box testing. It is often done by the programmer to test that the unit he/she has implemented is producing expected output against given input.

**Integration Testing :**

Integration testing is testing in which a group of components are combined to produce output. Also, the interaction between software and hardware is tested in integration testing if software and hardware components have any relation. It under both white box testing and black box testing.

**Functional Testing :**

Functional testing is the testing to ensure that the specified functionality required in the system requirements works. It falls under the class of black box testing.

**System Testing :**

System testing is the testing to ensure that by putting the software in different environments (e.g., Operating Systems) it still works. System testing is done with full system implementation and environment. It fall sunder the class of black box testing.

**Stress Testing :**

Stress testing is the testing to evaluate how system behaves under unfavorable conditions. Testing is conducted at beyond limits of the specifications. It fall sunder the class of black box testing.

**Performance Testing :**

Performance testing is the testing to assess the speed and effectiveness of the system and to make sure it is generating results within a specified time as in performance requirements. It falls under the class of black box testing.

**Usability Testing :**

Usability testing is performed to the perspective of the client, to evaluate how the GUI is user-friendly? How easily can the client learn? After learning how to use, how proficiently can the client perform? How pleasing is it to use its design? This fall sunder the class of black box testing.

**Acceptance Testing :**

Acceptance testing is often done by the customer to ensure that the delivered product its the requirements and works as the customer expected. It fall sunder the class of black box testing.

**Regression Testing :**

Regression testing is the testing after modification of a system, component, or a group of related units to ensure that the modification is working correctly and is notdamaging or imposing other modules to produce unexpected results. It falls under the class of black box testing.

## 6.1 TEST CASES :

| S.No | Test case name | Input | Expected result | Actual result | Pass/Fail/ Not Exectued |
|---|---|---|---|---|---|
| 1. | TC1 | Data set | No missing values | No missing values | Pass |
| 2. | TC2 | Data set after removing missing values | Train and test | Train and test | Pass |
| 3. | TC3 | Trained and tested dataset | Model Evaluation | Model Evaluation | Pass |
| 4. | TC4 | Model accuracy | Comparing accuracy | Comparing accuracy | Pass |
| 5. | TC5 | Best accuracy | Classification report for best model | Classification report for best model | Pass |

# CONCLUSION

# 8. CONCLUSION

This project represented the Machine Learning Approach for predicting the rainfall by using 4 ML algorithms like Logistic Regression, Random Forest Classifier, Decision Tree and KNN. Comparing the 4 algorithms and choosing the best approach for rainfall prediction. This project provides a study of different types of methodologies used to forecast and predict rainfall and Issues that could be found when applying different approaches to forecast in rainfall.

Because of nonlinear relationships in rainfall data sets and the ability to learn from the past, makes a superior solution to all approaches available. The future work of the project would be the improvement of architecture for light and other weather scenarios. Also, can develop a model for small changes in climate in future. An algorithm for testing daily basis data set instead of accumulated data set could be of paramount Importance for further research.

# FURTHER ENHANCEMENTS

# 9. FURTHER ENHANCEMENTS

Rainfall prediction to connect with cloud.Creating pickle file and deployment using flask.Developing a website using pickle file.To optimize the work to implement in Artificial Intelligence environment.

The future work of the project would be the improvement of architecture for light and other weather scenarios. Also, can develop a model for small changes in climate in future. An algorithm for testing daily basis data set instead of accumulated data set could be of paramount Importance for further research.

# BIBLIOGRAPHY

# 10. BIBLIOGRAPHY

[1] singh, p., 2018. indian summer monsoon rainfall (ismr) forecasting using time series data: a fuzzy-entropy-neuron based expert system. Geo science frontiers, 9(4),pp.1243-1257.

[2] cramer, s., kampouridis, m., freitas, a. and alexandridis, a., 2017. an extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives.expert systems with applications,85,pp.169-181.

[3] pour,s.,shahid,s.and chung,e.,2016.a hybrid model for statistical down scaling of daily rainfall.procedure engineering,154,pp.1424-1430.

[4] manjunath n, muralidhar b r, sachin kumar s, vamshi k and savitha p, 2021.rainfall prediction using machine learning and deep learning techniques. [online] irjet.net.available at:<https://www.irjet.net/archives/v8/i8/irjet-v8i850.pdf>[accessed20january2022].

[5] tanvi patil and dr. kamal shah, 2021. weather forecasting analysis using linear and logistic regression algorithm. [online] irjet.net. Available at:<https://www.irjet.net/archives/v8/i6/irjetv8i6454.pdf>[accessed 20 ja nuary2022].

[6] n.divya prabha and p.radha, 2019.prediction of weather and rainfall forecasting using classification techniques. [online] irjet.net.available at:<https://www.irjet.net/archives/v6/i2/irjet-v6i2154.pdf>[accessed 20 january 2022].waghm are,d.,2021.machine learning technique for rainfall prediction.

[7] international journal for research in appliedscience and engineering technology,9(vi),pp.594-600.

[8] yashasathreya, vaishalibv, sagark and srinidhihr, 2021 flood prediction and rainfall analysis using machine learning [online]irjet.net.availableat:<https://www.irjet.net/archives/v8/i7/irjet-v8i7432.pdf>[accessed20january2022].

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# RAINFALL PREDICTION USING MACHINE LEARNING

**Kanasani Renuka Devi[1], Alapati Padma Sree[2],**

**Narra Bala Mary Sowmya[3] , Katuri Santosh Dutt[4] , Mrs. Sk. Shammi[5]**

[1,2,3,4] *Student, Department of Computer Science and Engineering, Tirumala Engineering College*
[5] *Professor, Department of Computer Science and Engineering, Tirumala Engineering College*

---

**Abstract -** India is an agricultural country and its economy is largely based upon crop productivity and rainfall. For analyzing the crop productivity, rainfall prediction is require and necessary to all farmers. Rainfall Prediction is the application of science and technology to predict the state of the atmosphere. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre planning of water structures. Using different data mining techniques it can predict rainfall. Data mining techniques are used to estimate the rainfall numerically. This paper focuses some of the popular data mining algorithms for rainfall prediction. Random Forest, K-Nearest Neighbor algorithm, Logistic regression, Decision Tree are some of the algorithms have been used. From that comparison, it can analyze which method gives better accuracy for rainfall prediction.

**Key Words:** Rain fall, Agricultural, Economy , Farmers, Water resources.

## 1. INTRODUCTION

### 1.1. OBJECTIVE OF THE PROJECT:

The goal is to develop a machine learning model for Rainfall Prediction to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

#### 1.1.1. Necessity:

This prediction helps in predicting the rainfall andit helps in Over coming the crop productivity and to predict the state of atmosphere in agricultural countries. These models are very easy to use. It can work accurately and very smoothly in a different scenario. It reduces the effort workload and increases efficiency in work. In aspects of time value, it is worthy.

#### 1.1.2. Software development method:

In many software applications program different methods and cases are followed such as, Waterfall model, Iterative model, Spiral model, V- model and Big Bang model. we used waterfall model in this application. we tried to use test case and cases of ware approaches.

#### 1.1.3 Layout of the document:

This documentation starts with form a introduction. After introduction analysis and design of the project are described. In analysis and design of the project have many parts such as project proposal, mission, goal, target audience, environment. Use cases and test cases are explained below respectively. Finally, this documentation finished with result and Conclusion part.

## 2. LITERATURE REVIEW

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes are search proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information.It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent orel event of them.

## 3. SYSTEM ANALYSIS

The project proposal is the term of documents. A project can describe the project proposal. It is the set of all plans of a project. Like, how the software works, what are the steps to complete the entire projects, and what are the software requirements and analysis for this project. In our project, we are doing all the steps and also risk and reward and other project depend encies in the project proposal.

To compare several machine learning models like logistic

regression, random forest, knn and decision tree. Plotting confusion matrix for each model after cleaning the data set so that we can easily find the best model among them. After finding best model we will draw ROC curve and classification report for that best fit model to predict train fall which is very essential for farmers.

The goal is to develop a machine earning model for predicting the rainfall.

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are limited to precision,sensitivity,F1-score.

The overview of the project is to provide a best machine learning algorithm to the user.Therefore, the user can directly know whether the rainfall is occur or not throught his best model.

## 4. EXISTING SYSTEM

Agriculture is the strength of our Indian economy. Farmer only depend suponmon soon to be their cultivation. The good crop productivity needs good soil, fertilizer and also good climate. Weather fore casting is the very important requirement of the each farmer.Due to the sudden changes in climate/weather, The people are suffer economically and physically. Weather prediction is one of the challenging problems in current state. The main motivation of this paper to predict the weather using various data mining techniques. Such as classification, clustering, decision tree and also neural networks. Weather related in formation is also called the meteorological data. In this paper the most commonly used weather parameters are rainfall, wind speed, temperature and cold.

### 4.1 PREPARING THE DATASET

This data set contains 145460 records of features extracted from kaggle, which is having Rain To morrow as a target column containing 2 values.

## 5. PROPOSED SYSTEM

### 5.1 Exploratory Data Analysis of Rainfall Prediction:

Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

### 5.1.2. Data Cleaning

In this section of the report will load in the data, check for cleanliness, and then trim and clean given data set for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

### 5.1.3 Data Collection

The data set collected for predicting given data is split into Training set and Test set. Generally, we split the dataset into Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result.

### 5.1.4 Building the classification model

For predicting the rainfall, MLalgorithm prediction model is effective because of the following reasons: It provides better results in classification problem.It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables. It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

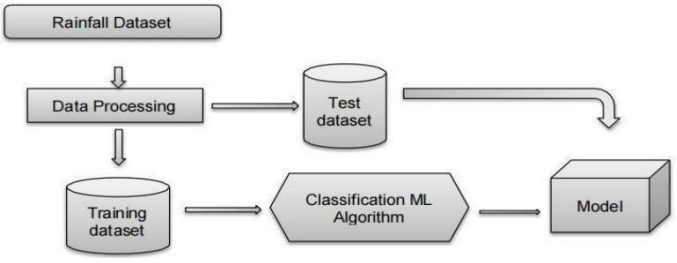
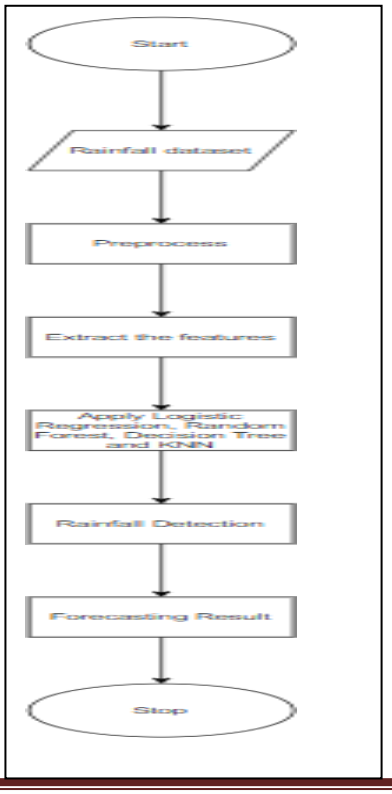



**Fig 5 : Architecture of Proposed model**

### 5.1.5 Advantages

- Performance and accuracy of the algorithms can be calculated and compared.
- Numerical Weather Prediction
- Statistical Weather Prediction

### 5.1.6 FLOW CHART



## 6. METHODS AND ALGORITHMS USED

**SYSTEM STUDY:**

To develop his model we use new modern technologies which are Machine Learning using Python for predicting rainfall.

**System requirement specifications:**

**Hardware requirements:**

- Processor :Intel
- RAM :2GB
- Hard Disk :80GB

**Software requirements:**

- OS :Windows
- Framework: Flask
- Technology: Machine Learning using Python
- Web Browser: Chrome, Microsoft Edge
- Codeeditor : Visual Studio Code, Google Colab,

**SOFTWARE ENVIRONMENT:**

Python is a high-level, interpreted, interactive and object- oriented scripting language. Python is designed to be highly readable. It uses English key words frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

**Python is Interpreted** − Python is processed at run time by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

**Python is Interactive** –You can actually it at a Python prompt and interact with the interpreter directly to write your programs.

**Python is Object-Oriented** –Python supports Object-Oriented style or technique of programming that encapsulates code with in objects.

**Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

## 7. SYSTEM TESTING

**TESTING PLAN**

Software testing is the process of evaluation a software item to detect differences between given input and expected output. Also to assess the feature of a software item. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

**Verification :**

Verification is the process to make sure the product satisfies the conditions imposed at the start of the development phase. In other words, to make sure the product behaves the way we want it to.

**Validation :**

Validation is the process to make sure the product satisfies the specified requirements at the end of the development phase. In other words, to make sure the product is built as per customer requirements.

**BASICS OF SOFTWARE TESTING :**

There are two basics of software testing: black box testing and white box testing.

**Black box testing :**

Black box testing is a testing technique that ignores the internal mechanism of the system and focuses on the output generated against any input and execution of the system. It is also called functional testing.

**White box testing :**

White box testing is a testing technique that takes into account the internal mechanism of a system. It is also called structural testing and glass box testing. Black box testing is often used for validation and white box testing is often used for verification.

### 1. TYPES OF TESTING :

There are many types of testing like :

- Unit Testing
- Integration Testing
- Functional Testing
- System Testing
- Stress Testing
- Performance Testing
- Usability Testing
- Acceptance Testing
- Regression Testing
- Beta Testing

**Unit Testing :**

Unit testing is the testing of an individual un it or group of related units. It falls under the class of white box testing. It is often done by the programmer to test that the unit he/she has implemented is producing expected output against given input.

**Integration Testing :**

Integration testing is testing in which a group of components are combined to produce output. Also, the interaction between software and hardware is tested in integration testing if software and hardware components have any relation. It may fall under both white box testing and black box testing.

**Functional Testing :**

Functional testing is the testing to ensure that the specified functionality required in the system requirements works. It falls under the class of black box testing.

**System Testing :**

System testing is the testing to ensure that by putting the software in different environments (e.g., Operating Systems) it still works. System testing is done with full system implementation and environment. It fall sunder the class of black box testing.

**Stress Testing :**

Stress testing is the testing to evaluate how system behaves under unfavorable conditions. Testing is conducted at beyond limits of the specifications. It fall sunder the class of black box testing.

**Performance Testing :**

Performance testing is the testing to assess the speed and effectiveness of the system and to make sure it is generating results within a specified time as in performance requirements. It falls under the class of black box testing.

**Usability Testing :**

Usability testing is performed to the perspective of the client, to evaluate how the GUI is user-friendly? How easily can the client learn? After learning how to use, how proficiently can the client perform? How pleasing is it to use its design? This fall sunder the class of black box testing.

**Acceptance Testing :**

Acceptance testing is often done by the customer to ensure that the delivered product meets the requirements and works as the customer expected. It fall sunder the class of black box testing.

**Regression Testing :**

Regression testing is the testing after modification of a system, component, or a group of related units to ensure that the modification is working correctly and is not damaging or imposing other modules to produce unexpected results. It falls under the class of black box testing.

## 8. RESULTS, DISCUSSIONS AND PERFORMANCE ANALYSIS

**PERFORMANCE ANALYSIS :**

Website performance optimization, the focal point of technologically superior website designs is the primary factor dictating the Rainfall occurred or not. After all, unimpressive website performance kills admission process when the torture of waiting for slow Web pages to load frustrates visitors in to seeking alternatives – impatience is a digital virtue! And also the ml algorithms used in our project will give the best accurate result to the user for Rainfall prediction.

We created the following six chapter in-depth speed optimization guide to show you how important it is to have a fast loading, snappy website! Countless research papers and benchmarks prove that optimizing your sites' speed is one of the most affordable and highest ROI providing investments!

Lightning-fast page load speed amplifies visitor engagement, retention, and boosts sales. Instantaneous website response leads to higher conversion rates, and every1second delay in page load decreases customer satisfaction by 16 percent, page views by 11 percent and conversion rates by 7 percent according toarecent Aberdeen Group research.

| Algorithm | Accuracy |
|---|---|
| Logistic Regression | 0.8 |
| Random Forest | 1.0 |
| K Nearest Neighbors | 0.9 |
| Decision Tree Classifier | 1.0 |

**DISCUSSIONS :**

While discussions provide avenues for exploration and discovery, leading a discussion can be anxiety-producing: discussions are, by their nature, unpredictable, and require us as instructors to surrender a certain degree of control over the flow of information. Fortunately, careful planning can help us ensure that discussions are lively without being chaotic and exploratory without losing focus. When planning a discussion, it is helpful to consider not only cognitive, but also social/emotional, and physical factors that can either foster or inhibit the productive exchange of ideas.

## 9. CONCLUSION

This project represented the Machine Learning Approach for predicting the rainfall by using 4 ML algorithms like Logistic Regression, Random Forest Classifier, Decision Tree and KNN. Comparing the 4 algorithms and choosing the best approach for rainfall prediction. This project provides a study of different types of methodologies used to forecast and predict rainfall and Issues that could befound when applying different approaches to forecast in rainfall.

Because of nonlinear relationships in rainfall data sets and the ability to learn from the past, makes a superior solution to all approaches available. The future work of the project would be the improvement of architecture for light and other weather scenarios. Also, can develop a model for small changes in climate in future. Analgorithm for testing daily basis data set instead of accumulated data set could be of paramount Importance for further research.

## 10. REFERENCES


- Xiong, Lihua, and Kieran M. OConnor. "An empirical method to improve the prediction limits of the GLUE methodology in rainfallrunoff modeling." Journal of Hydrology 349.1-2 (2008): 115-124.
- Schmitz, G. H., and J. Cullmann. "PAI-OFF: A new proposal for online flood forecasting in flash flood prone catchments." Journal of hydrology 360.1-4 (2008): 1-14.
- Riordan, Denis, and Bjarne K. Hansen. "A fuzzy casebased system for weather prediction." Engineering Intelligent Systems for Electrical Engineering and Communications 10.3 (2002): 139-146.
- Guhathakurta, P. "Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network." Current Science 90.6 (2006): 773-779.
- Pilgrim, D. H., T. G. Chapman, and D. G. Doran. "Problems of rainfall-runoff modelling in arid and semiarid regions." Hydrological Sciences Journal 33.4 (1988): 379-400.
- Lee, Sunyoung, Sungzoon Cho, and Patrick M. Wong. "Rainfall prediction using artificial neural networks." journal of geographic information and Decision Analysis 2.2 (1998): 233- 242..
- French, Mark N., Witold F. Krajewski, and Robert R. Cuykendall. "Rainfall forecasting in space and time using a neural network." Journal of hydrology 137.1-4 (1992): 1-31.
- Charaniya, Nizar Ali, and Sanjay V. Dudul. "Committee of artificial neural networks for monthly rainfall prediction using wavelet transform." Business, Engineering and Industrial Applications (ICBEIA), 2011 International Conference on. IEEE, 2011.
- Noone, David, and Harvey Stern. "Verification of rainfall forecasts from the Australian Bureau of Meteorology‟s Global Assimilation and Prognosis(GASP) system." Australian Meteorological Magazine 44.4 (1995): 275-286.
- Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." Neural networks 2.5 (1989): 359-366.
- Haykin, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.
- Rajeevan, M., Pulak Guhathakurta, and V. Thapliyal. "New models for long range forecasts of summer monsoon rainfall over North West and Peninsular India." Meteorology and Atmospheric Physics 73.3-4 (2000) 211-

DOI: 10.55041/IJSREM31965

ISSN: 2582-3930

Impact Factor: 8.448

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

# CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

**Kanasani Renuka Devi**

in recognization to the publication of paper titled

## RAINFALL PREDICTION USING MACHINE LEARNING

published in IJSREM Journal on Volume 08 Issue 04 April, 2024

Editor-in-Chief
IJSREM Journal

www.ijsrem.com

e-mail: editor@ijsrem.com

DOI: 10.55041/IJSREM31965

ISSN: 2582-3930

Impact Factor: 8.448

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

# CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

## Alapati Padma Sree

in recognization to the publication of paper titled

## RAINFALL PREDICTION USING MACHINE LEARNING

published in IJSREM Journal on Volume 08 Issue 04 April, 2024

Editor-in-Chief
IJSREM Journal

www.ijsrem.com

e-mail: editor@ijsrem.com

DOI: 10.55041/IJSREM31965

ISSN: 2582-3930

Impact Factor: 8.448

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

# CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

## Narra Bala Mary Sowmya

in recognization to the publication of paper titled

## RAINFALL PREDICTION USING MACHINE LEARNING

published in IJSREM Journal on Volume 08 Issue 04 April, 2024

Editor-in-Chief
IJSREM Journal

www.ijsrem.com

e-mail: editor@ijsrem.com

DOI: 10.55041/IJSREM31965

ISSN: 2582-3930

Impact Factor: 8.448

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

# CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

**Katuri Santosh Dutt**

in recognization to the publication of paper titled

## RAINFALL PREDICTION USING MACHINE LEARNING

published in IJSREM Journal on Volume 08 Issue 04 April, 2024

Editor-in-Chief
IJSREM Journal

www.ijsrem.com

e-mail: editor@ijsrem.com