RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

by

TANNERU NAVANISH (38110591)
KANCHARLA NARAYAN PAVAN (38110229)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MAY - 2022

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **TANNERU NAVANISH (38110591), KANCHARLA NARAYAN PAVAN (Reg. No: 38110229)** who carried out the project entitled "**RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES**" under my supervision from November 2021 To April 2022

Internal Guide
Dr. Mercy Paul Selvan M.E., Ph.D.,

External Guide (if Applicable)

Head of the Department
Dr. L. LAKSHMANAN M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **TANNERU NAVANISH, KANCHARLA NARAYAN PAVAN** hereby declare that the Project Report entitled **RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES** done by me under the guidance of **Dr. Mercy Paul Selvan, M.E., Ph.D., (Internal)** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

DATE: 15-03-2022 T. Navanish, K.N.Pavan

PLACE: CHENNAI **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. T.Sasikala M.E., Ph.D, Dean, School of Computing Dr. L. Lakshmanan M.E., Ph.D. , and Dr. S. Vigneshwari M.E., Ph.D. Head of the Department, Dept. of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr. Mercy Paul Selvan, M.E., Ph.D., for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways for the completion of the project.

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, SVM, Decision Tree are some of the algorithms have been used. From that comparison, it can analyze which method gives better accuracy for rainfall prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-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 and it helps in overcoming the crop productivity and to predict the state of atmosphere in agricultural countries. This Application is 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. I used waterfall model in this application. I tried to use test case and case software approaches.

1.1.3 Layout of the document:

This documentation starts with formal 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 in chapter 2 and chapter 3 respectively. Finally, this documentation finished with result and Conclusion part.

1.2 OVERVIEW OF THE DESIGNED PROJECT:

At first, we take the dataset from out resource then we have to perform data-preprocessing, visualization methods for cleaning and visualizing the dataset respectively and we applied the Machine Learning algorithms on the dataset then we generate the pickle file for best algorithm and flask is used as user interface for displaying the result.

CHAPTER-2
LITERATURE SURVEY

General

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 a research 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 or relevant 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 framework than the advanced rainfall determining methods. We analyze and think about the momentum execution (Markov chain stretched out by rainfall research) with the forecasts of the six most not able AI machines: Genetic programming, Vector relapse support, radio organizations, M5

organizations, M5models, models - Happy. To work with a more itemized appraisal, we led a rainfall overview 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 backbone. Downpour 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. Machine7Learning calculations are significant in foreseeing rainfall.

[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. Underlying 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 perceptron neural organization. Estimations like MSE (Early Modeling), NMSE (Usually Early Error), and the arrangement of informational collections for transient arranging are clear in the examination of different organizations, like Adanaive. AdaSVM.

[7] In this paper, Artificial Neural Network (ANN) innovation is utilized to foster a climate anticipating strategy to distinguish rainfall utilizing Indian rainfall information. Along these

lines, Feed Forward Neural Network (FFNN) was utilized utilizing the Backpropagation Algorithm. Execution of the two models is assessed dependent on emphasis examination, Mean Square Error (MSE) and Magnitude of Relative Error (MRE). This report likewise 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.

CHAPTER-3

AIM AND SCOPE OF THE PRESENT INVESTIGATION

3.1 PROJECT PROPOSAL:

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 my project, I am doing all the steps and also risk and reward and other project dependencies in the project proposal.

3.1.1 Mission:

An online Web based machine learning application is very popular and well known to everyone. Now a day's everybody wants to get it and work with it. Rainfall prediction is mostly useful for farmers in preplanning of water structure, crop productivity and it determines the rainfall for effective use of water. This simple method gives fast and accurate results in enhancing the rainfall.

3.1.2 Goal:

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

3.2 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.

3.3 OVERVIEW OF THE PROJECT:

The overview of the project is to provide a web-based machine learning application to the user. Therefore, the user can directly know whether the rainfall is occur or not through our website over the internet.

3.4 EXISTING SYSTEM:

Agriculture is the strength of our Indian economy. Farmer only depends upon monsoon to be their cultivation. The good crop productivity needs good soil, fertilizer and also good climate. Weather forecasting is the very important requirement of the each farmer. Due to the sudden changes in climate/weather, The people are suffered 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 information is also called the meteorological data. In this paper the most commonly used weather parameters are rainfall, wind speed, temperature and cold.

3.4.1 Disadvantages:

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

3.5 PREPARING THE DATASET:

This dataset contains 145460 records of features extracted from kaggle, which were then classified into 2 classes:

- Rainy day
- Sunny day

3.6 PROPOSED SYSTEM:

3.6.1 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.

3.6.2 Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the

document steps carefully and justify for cleaning decisions.

3.6.3 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 accuracy, Test set prediction is done.

3.6.4 Building the classification model

For predicting the rainfall, ML algorithm 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.

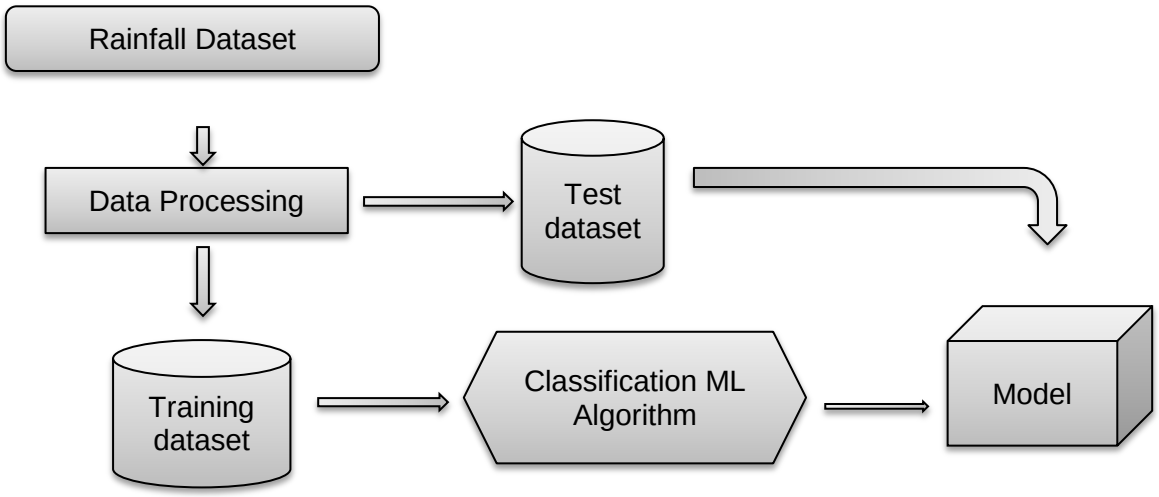


Fig: 3.1: Architecture of Proposed model

3.6.5 Advantages:

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

3.8 FLOW CHART:

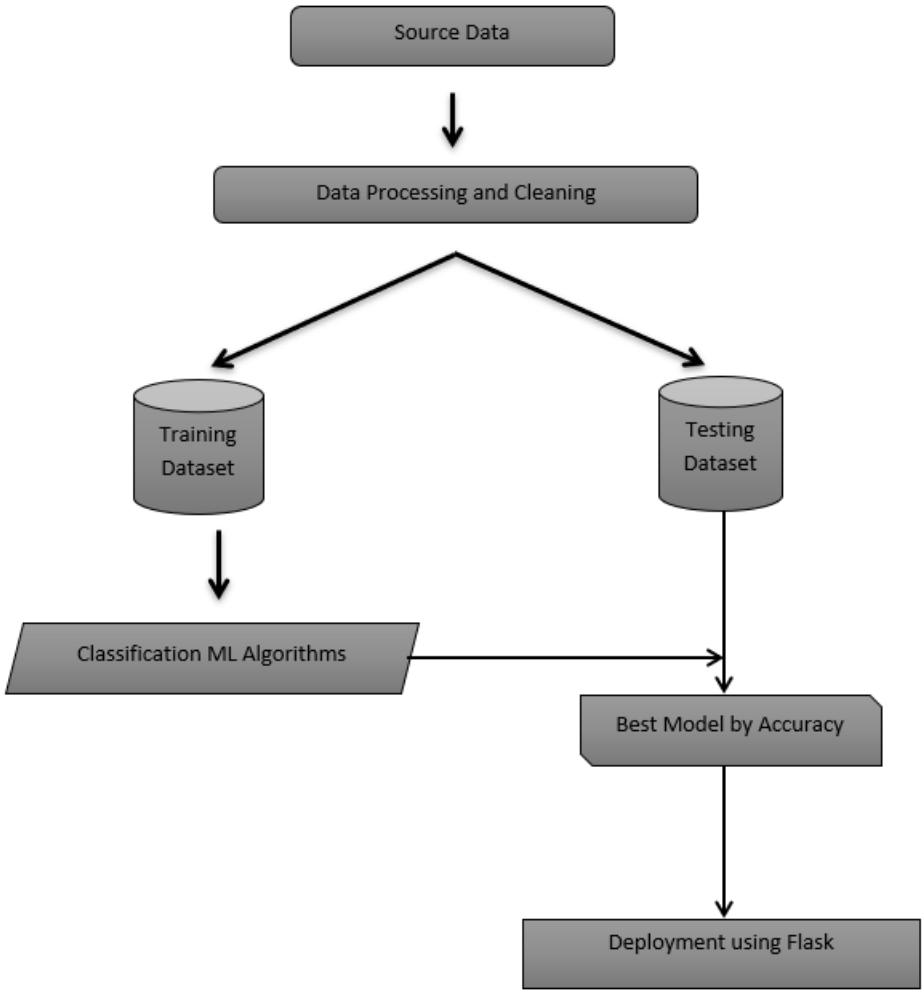


Fig: 3.2: FLOW CHART

CHAPTER-4
EXPERIMENTAL OR MATERIALS AND METHODS
ALGORITHMS USED

4.1 SYSTEM STUDY:

To develop this model we use new modern technologies which are Machine Learning using Python for predicting and Flask is for user interface.

4.1.1 System requirement specifications:

a) Hardware requirements:

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

b) Software requirements:

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

4.2 SYSTEM SPECIFICATIONS:

4.2.1 Machine Learning Overview:

Machine learning is a field of study that looks at using computational algorithms to turn empirical data into usable models. The machine learning field grew out of traditional statistics and artificial intelligences communities. Through their business processes immense amounts of data have been and will be collected. This has provided an opportunity to re-invigorate the statistical and computational approaches to autogenerate useful models from data. Machine learning algorithms can be used to (a) gather understanding of the cyber phenomenon that produced the data under study, (b) abstract the understanding of

underlying phenomena in the form of a model, (c) predict future values of a phenomena using the above-generated model, and (d) detect anomalous behavior exhibited by a phenomenon under observation.

4.2.2 Flask Overview:

Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application.

4.3 STEPS TO DOWNLOAD & INSTALL PYTHON:

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

Pip install --upgrade pip and Pip install --upgrade setuptools

4.3.1 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/)

4.3.2 PYTHON FILE CREATION

GO *To FILE MENU > CREATE > NEW > PYTHON FILE >(Name Your Python File as "HOUSE PRICE PPREDICTION" > SAVE*

4.4 PYTHON LIBRARIES NEEDED

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

4.4.1 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 a number 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.

4.4.2 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 the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables. There are several ways to create a DataFrame.

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.

4.4.3 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, etc.

4.4.4 SEABRON LIBRARY

Seaborn package was developed based on the Matplotlib library. It is used to create more attractive and informative statistical graphics. While seaborn 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
```

4.4.5 SCIKIT-LEARN LIBRARY

Scikit-learn is a free machine learning library for the Python. It features various algorithms like support vector machine, random forests, regression and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

```
Pip install Scikit-Learn
```

Installation-(https://scikit-learn.org/stable/install.html)

Here use scikit-learn's regression methods for prediction purpose.

4.4.6 FLASK

Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it

```
pip install flask
```

has less base code to implement a simple web-Application.

Here we use flask for the user-interface.

4.5 MODULES:

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

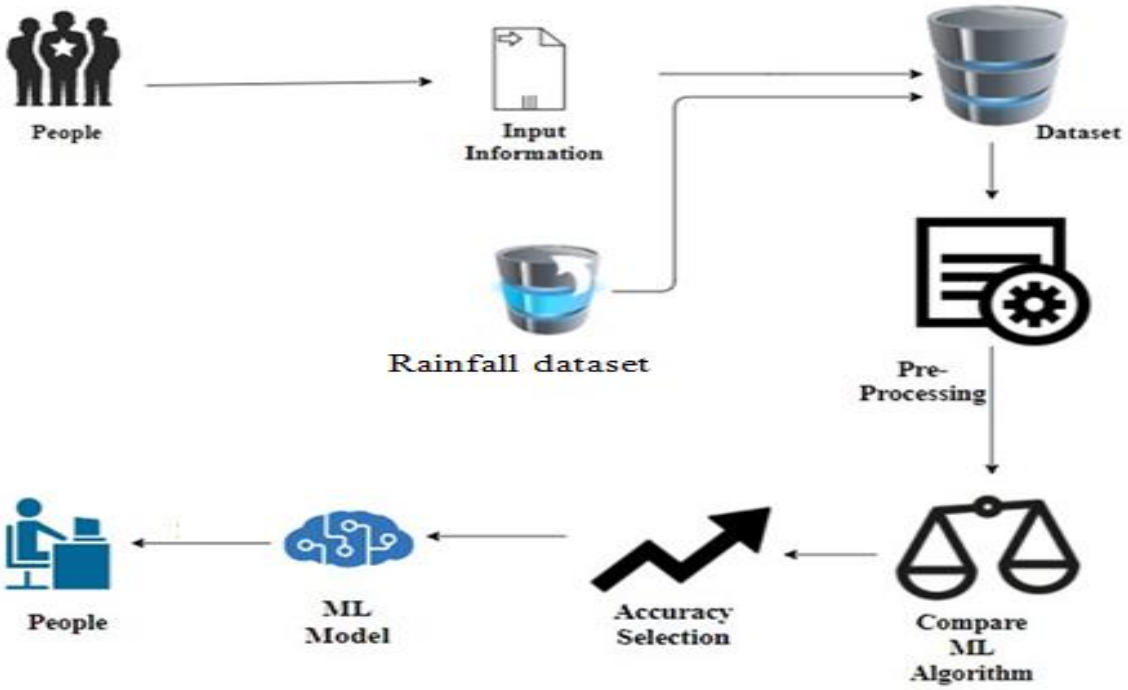


Fig: 4.1: SYSTEM ARCHITECTURE

4.6. UML DIAGRAMS

4.6.1 Use Case Diagram

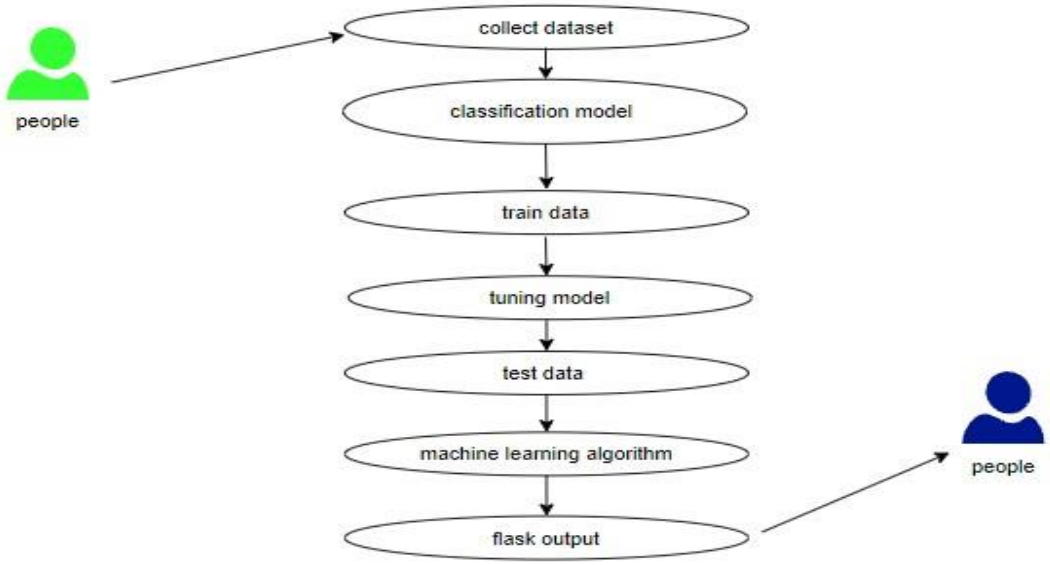


Fig: 4.2: USE CASE DIAGRAM

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.

3.6.2 Class Diagram

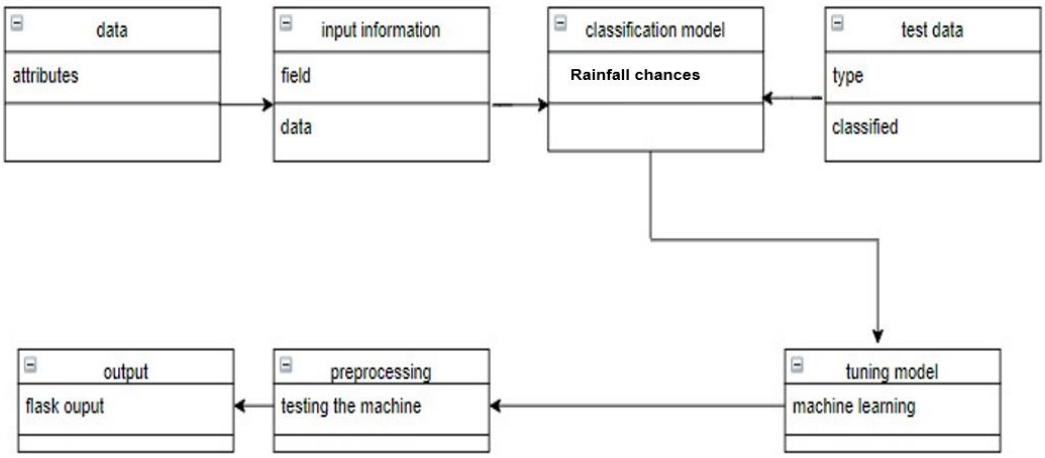


Fig: 4.3: CLASS DIAGRAM

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.

4.6.3 Activity Diagram

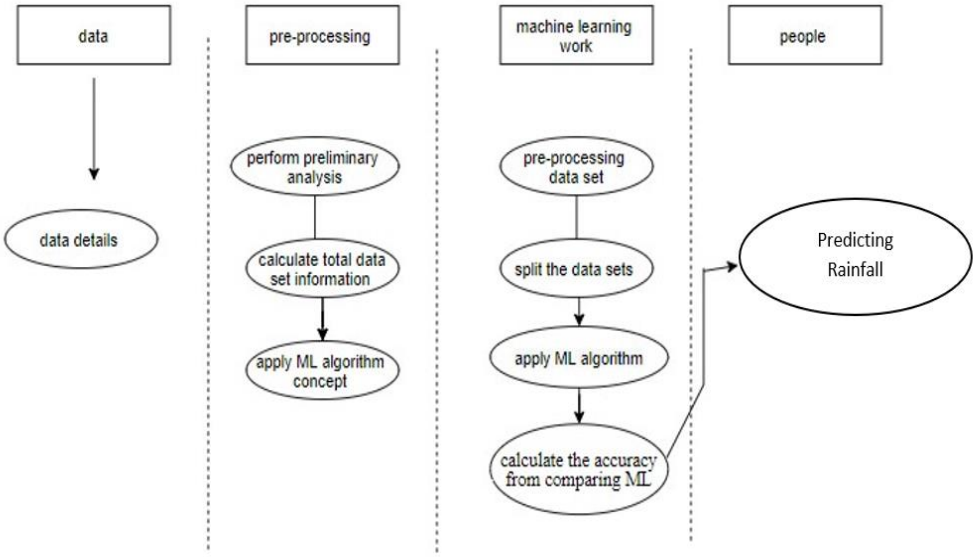


Fig: 4.4: ACTIVITY DIAGRAM

Activity is a particular operation of the system. Activity diagrams are not only 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 but it is not. It shows different flow like parallel, branched, concurrent and single.

4.6.4 Sequence Diagram

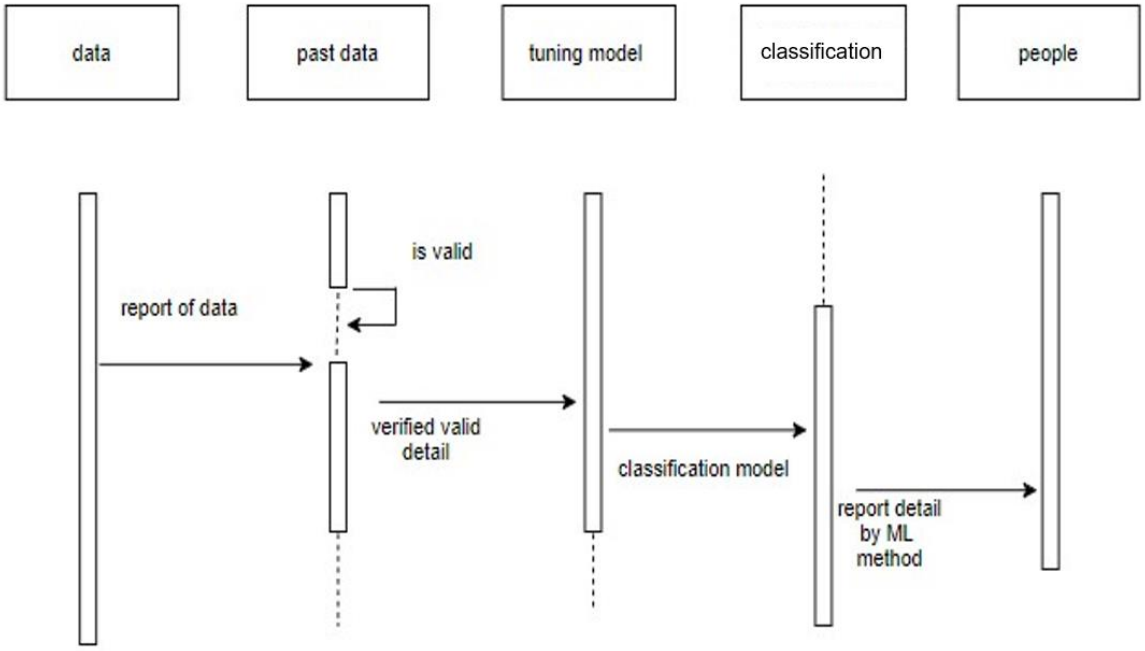


Fig: 4.5: SEQUENCE DIAGRAM

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 within 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.

4.6.5 Entity Relationship Diagram (ERD)

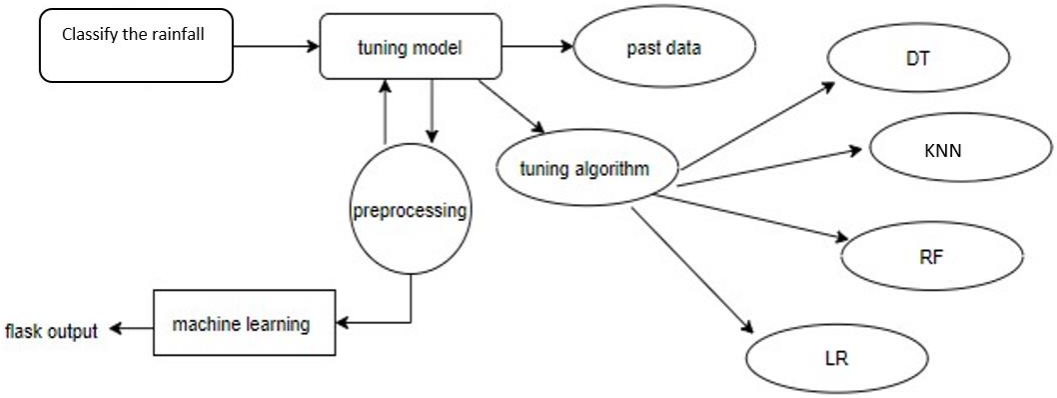


Fig: 4.6: ENTITY RELATIONSHIP DIAGRAM

An entity relationship diagram (ERD), also known as an entity relationship model, is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An ERD is a data modelling technique that can help define business processes and be used as the foundation for a relational database. Entity relationship diagrams provide a visual starting point for database design that can also be used to help determine information system requirements throughout an organization. After a relational database is rolled out, an ERD can still serve as a referral point, should any debugging or business process re-engineering be needed later.

The following are the modules of the project, which is planned in aid to complete the project with respect to the proposed system, while overcoming existing system and also providing the support for the future enhancement.

4.7 MODULE DETAILS:

4.7.1 Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It’s important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it’s important to understand the sources of missing data. Here are some typical reasons why

data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose & read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

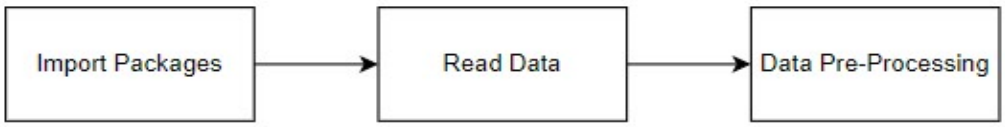
4.7.2 Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to

analyze the uni-variate, bi-variate and multi-variate process.

The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

MODULE DIAGRAM



GIVEN INPUT EXPECT OUTPUT

input: data

output: removing noisy data

4.7.3 Exploration data analysis of visualization

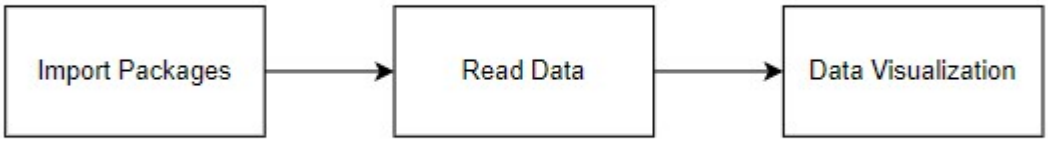
Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and

others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

MODULE DIAGRAM



GIVEN INPUT EXPECT OUTPUT

input: data

output: visualized data

4.7.4 Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different

techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

In the example below these 5 different algorithms are compared:

- Logistic Regression
- Random Forest
- Decision Tree
- Support Vector Machine
- K Nearest Neighbor

The K-fold cross validation procedure is used to evaluate each algorithm,

importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It needs the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression

model by comparing the best accuracy.

True Positive Rate (TPR) = TP / (TP + FN)
False Positive Rate (FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct.
Precision = TP / (TP + FP)
Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)
Recall = TP / (TP + FN)
Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4.7.5 ALGORITHM AND TECHNIQUES

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.

- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

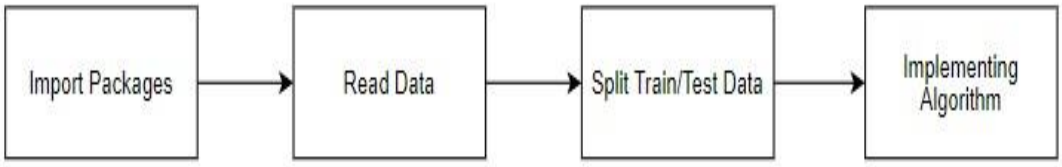
Logistic Regression:

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 outcome 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:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

MODULE DIAGRAM



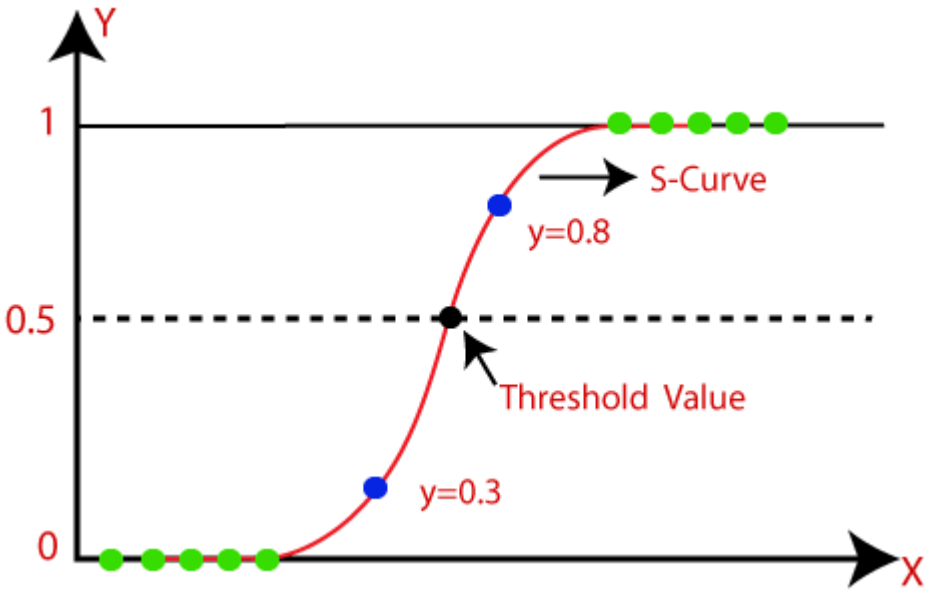


Fig: 4.7: LOGISTIC REGRESSION

GIVEN INPUT EXPECT OUTPUT

input: data

output: getting accuracy

Random Forest Classifier:

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 learning 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 random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

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.

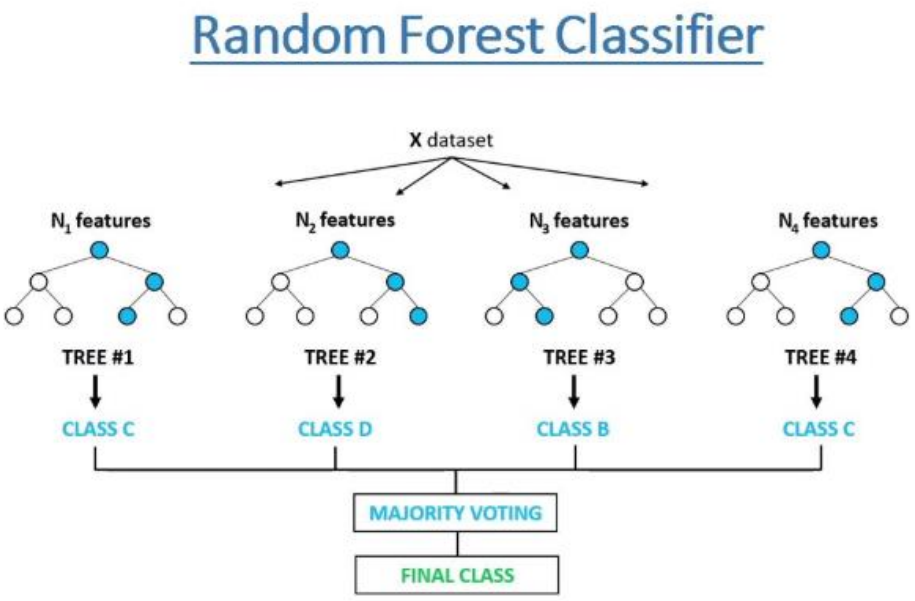


Fig: 4.8: RANDOM FOREST CLASSIFIER

GIVEN INPUT EXPECT OUTPUT

input: data

output: getting accuracy

Decision Tree :

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a 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. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

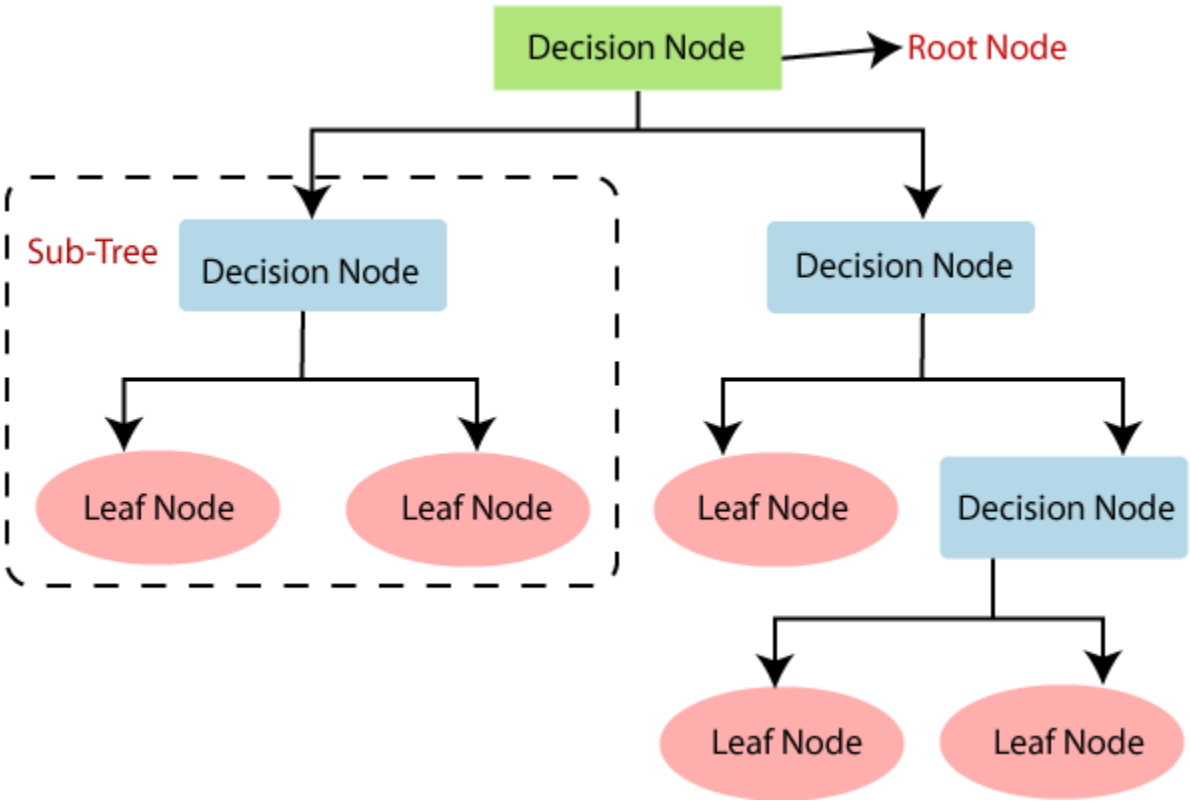


Fig: 4.9: DECISION TREE CLASSIFIER

K-Nearest Neighbor

K-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 into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

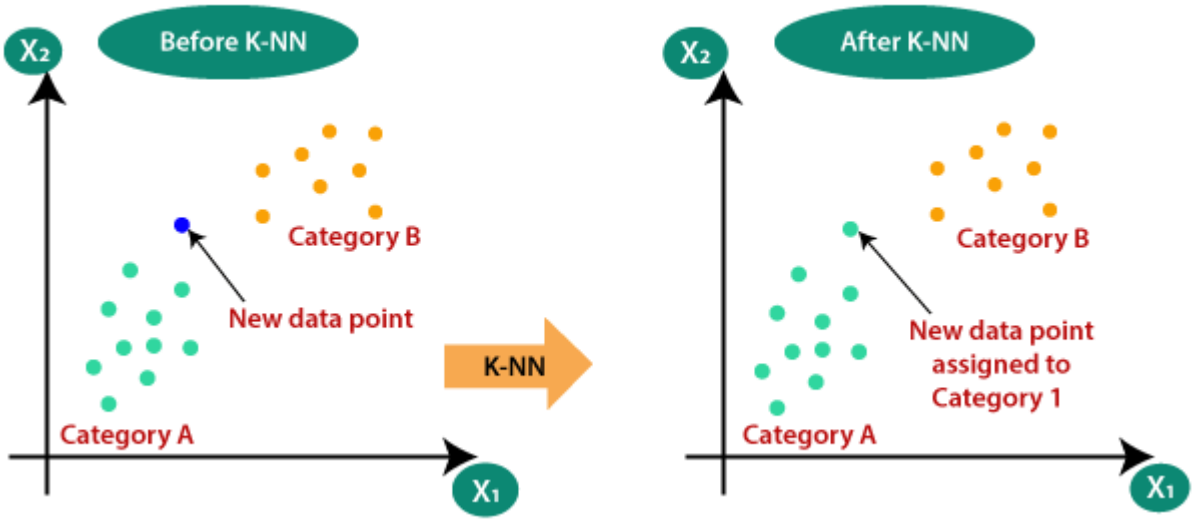


Fig: 4.10: K-NEAREST NEIGHBOR

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. Primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes. So that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

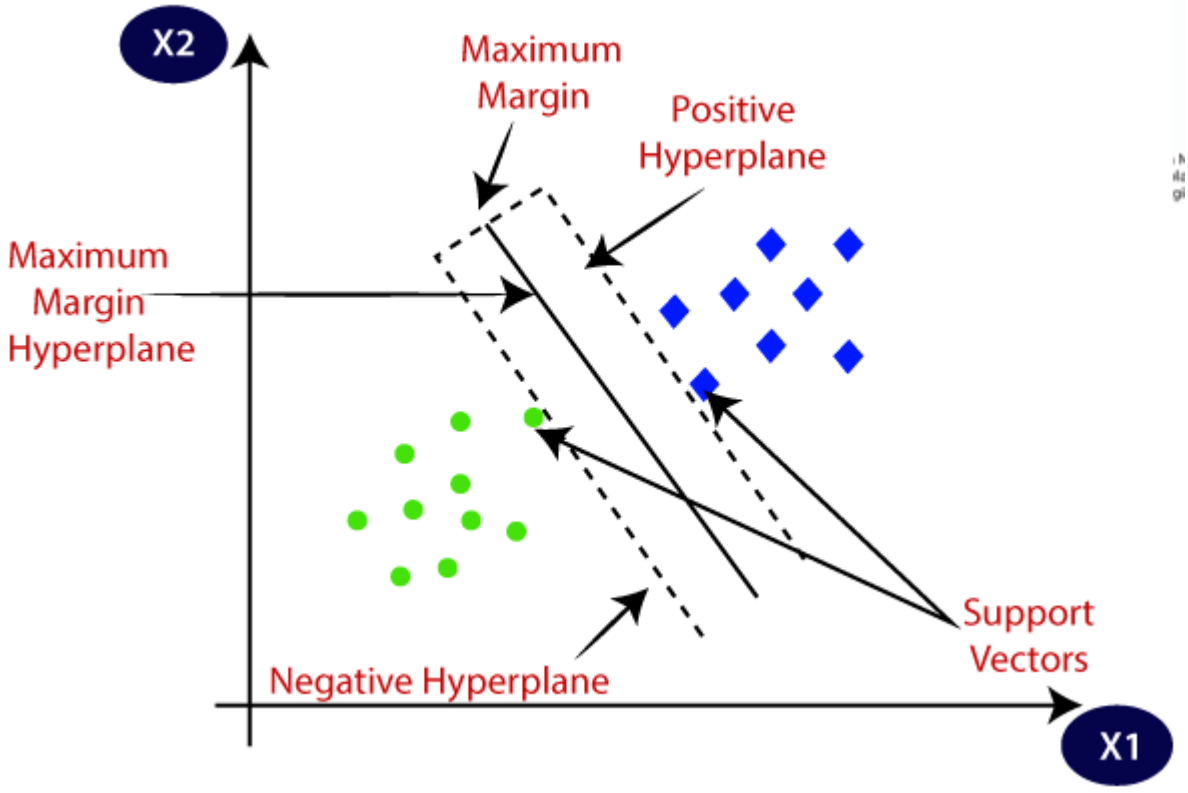


Fig: 4.11: SUPPORT VECTOR MACHINE

4.7.6 Deployment Using Flask (Web Framework):

Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Flask was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts formed in 2004. According to Ronacher, the idea was originally an April Fool's joke that was popular enough to make into a serious

application. The name is a play on the earlier Bottle framework.

When Ronacher and Georg Brand created a bulletin board system written in Python, the Pocoo projects Werkzeug and Jinja were developed.

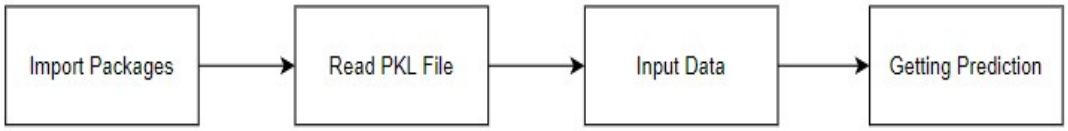
In April 2016, the Pocoo team was disbanded and development of Flask and related libraries passed to the newly formed Pallets project.

Flask has become popular among Python enthusiasts. As of October 2020, it has second most stars on GitHub among Python web-development frameworks, only slightly behind Django, and was voted the most popular web framework in the Python Developers Survey 2018.

The micro-framework Flask is part of the Pallets Projects, and based on several others of them.

Flask is based on Werkzeug, Jinja2 and inspired by Sinatra Ruby framework, available under BSD licence. It was developed at pocoo by Armin Ronacher. Although Flask is rather young compared to most Python frameworks, it holds a great promise and has already gained popularity among Python web developers. Let’s take a closer look into Flask, so-called “micro” framework for Python.

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data values

output : predicting output

CHAPTER-5

RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

5.1 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 into 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 every 1 second delay in page load decreases customer satisfaction by 16 percent, page views by 11 percent and conversion rates by 7 percent according to a recent Aberdeen Group research.

Algorithm	Accuracy
Logistic Regression	**0.83**
RandomForest	**0.86**
KNeighborsClassifier	**0.84**
DecisionTreeClassifier	**0.78**

Table 5.1: Algorithms Accuracy

5.2 DISCUSSION:

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.

CHAPTER-6

SUMMARY AND CONCLUSION

6.1 SUMMARY:

This project objective is to predict the Rainfall. So this online Rainfall prediction system will helps the farmers to analyzing the crop productivity, preplanning the water structure and estimate the rainy or not.

6.2 CONCLUSION:

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can helps in predicting the Rainfall.

6.3 FUTURE WORK:

- Rainfall prediction to connect with cloud.
- To optimize the work to implement in Artificial Intelligence environment.

REFERENCES:

[1] singh, p., 2018. indian summer monsoon rainfall (ismr) forecasting using time series data: a fuzzy-entropy-neuro based expert system. geoscience 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 downscaling of daily rainfall. procedia 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> [accessed 20 january 2022].

[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/irjet-v8i6454.pdf> [accessed 20 january 2022].

[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].

[7] waghmare, d., 2021. machine learning technique for rainfall prediction. international journal for research in applied science 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.available at: <https://www.irjet.net/archives/v8/i7/irjet-v8i7432.pdf> [accessed 20 january 2022].

APPENDIX:

A. SOURCE CODE:

Jupyter Note Book

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
dataset=pd.read_csv("weatherAUS.csv")
pd.set_option('display.max_columns',70)
```

```
dataset.shape

dataset.isnull().sum()

#missing value percentage
round(dataset.isnull().sum()*100/len(dataset),2)
```

```
dataset.dtypes

for i in dataset.columns:
    print(i,"=" ,len(dataset[i].unique()))

dataset["RainToday"].value_counts()
```

```
for i in dataset.columns:
    if dataset[i].dtypes=="float64":
        sns.histplot(dataset[i])
        plt.show()

plt.figure(figsize=(14,9))
sns.heatmap(dataset.corr(),annot=True)

sns.barplot(x=dataset["year"],y=dataset["Rainfall"])
sns.barplot(x=dataset["month"],y=dataset["Rainfall"])

plt.figure(figsize=(12,8))
sns.countplot(x=dataset["month"],hue=dataset["RainToday"])
```

```
pd.crosstab(dataset["RainTomorrow"],dataset["RainToday"])

#replace Na in numerical columns with mean for columns with Na ratio higher
than 3%:
dataset['WindGustSpeed'].fillna(np.mean(dataset['WindGustSpeed'].dropna().valu
es), inplace = True)
dataset['Pressure9am'].fillna(np.mean(dataset['Pressure9am'].dropna().values),
inplace = True)
dataset['Pressure3pm'].fillna(np.mean(dataset['Pressure3pm'].dropna().values),
inplace = True)
dataset['WindGustDir']= dataset['WindGustDir'].fillna('Unknown')
dataset['WindDir9am']= dataset['WindDir9am'].fillna('Unknown')

from sklearn import preprocessing
a=preprocessing.normalize(dataset[names])
len(a)

d = preprocessing.normalize(scale_feat)
scaled_df = pd.DataFrame(d, columns=names)
scaled_df
```

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier()

from sklearn.model_selection import train_test_split as tts
x_train,x_test,y_train,y_test=tts(x,y,test_size=0.2,random_state=100)

from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.metrics import confusion_matrix

def result(algo):
    var=algo
    var.fit(x_train,y_train)
    y_pred=var.predict(x_test)
    return accuracy_score(y_pred,y_test) , f1_score(y_pred,y_test)
,confusion_matrix(y_pred,y_test)

result(RandomForestClassifier())
result(DecisionTreeClassifier()
```

```
from flask import Flask,render_template,url_for,request,jsonify
from flask_cors import cross_origin
import pandas as pd
import numpy as np
import datetime
import pickle
import sklearn

app = Flask(__name__, template_folder="templates")
app.static_folder="static"

model = pickle.load(open("rain_model2.sav", 'rb'))
print("Model Loaded")

@app.route("/",methods=['GET'])
@cross_origin()
def home():
    return render_template("index.html")

@app.route("/predict",methods=['GET','POST'])
@cross_origin()
def predict():
    if request.method == "POST":

        minTemp = float(request.form['mintemp'])
        maxTemp = float(request.form['maxtemp'])
        windGustSpeed = float(request.form['windgustspeed'])

        input_cat = [rainToday,location , month ]
        input_num=[humidity3pm,pressure3pm,windGustSpeed,humidity9am,pressure9
am,temp3pm,minTemp
                   ,maxTemp,temp9am,windSpeed3pm,windSpeed9am,rainfall]
        norm_arr=sklearn.preprocessing.normalize([input_num])

        arr1=np.insert(norm_arr,2,rainToday)
        arr2=np.insert(arr1,9,location)
        arr3=np.insert(arr2,13,month)

        pred = model.predict([arr3])
        output = pred
        if output == 0:
            return render_template("sunny.html")
        else:
            return render_template("rainy.html")
    return render_template("predict.html")
```

INDEX.HTML

```
<!DOCTYPE html>
<html lang="es" dir="ltr">
<head>
    <title>Rainfall Prediction</title>
</head>
<body>
  <div class="photo"></div>
    <section class="text-gray-600 body-font">
  <div class="container mx-auto flex px-5 py-15 items-center justify-center
flex-col">
    <img class="lg:w-2/6 md:w-3/6 w-5/6 mb-10 object-cover object-center
rounded-xl" alt="img" src="{{ url_for('static', filename = 'umb.jpg') }}" >
    <div class="text-center lg:w-2/3 w-full">
      <h1 class="title-font sm:text-4xl text-3xl mb-4 font-medium text-gray-
900">Will it rain tomorrow?</h1>
      <p class="mb-8 leading-relaxed sm:text-2xl">Stay Ahead of time and find
out now with ML powered predictions </p>
      <div class="flex justify-center">
      <a href="/predict"> <button class="inline-flex text-white bg-indigo-500
border-0 py-2 px-6 focus:outline-none hover:bg-indigo-600 rounded text-
lg">Predict</button></a> </div> </div> </div> </section> </body> </html>
```

Predication.HTML

```
<html lang="en">
<head>
    <meta charset="UTF-8">
    <link rel="stylesheet" href=
{{url_for('static',filename='predictor.css')}}>
    <title>Rain Prediction</title>
</head>
<body>
    <section id="prediction-form">
    </br>
    <div class="col-md-6 my-2 d-flex align-items-end justify-content-around">
       <a href="./"> <button type="submit" class="btn btn-info button"
style="margin-right: 100%;">Back</button></a>
    </div>
        <form class="form" action="/predict", method="POST">
            <h1 class="my-3 text-center">Predictor</h1>
            <div class="row">
                <div class="col-md-6 my-2">
                    <div class="md-form">
        <label for="mintemp" class="mintemp"> Minimum temprature (°C)</label>
```

```
        <input type="text" class="form-control" id="mintemp" name="mintemp">
</div> </div>
                <div class="col-md-6 my-2">
                    <div class="md-form">
                        <label for="temp9am" class="temp9am">Temperature 9am
(°C)</label>
                        <input type="text" class="form-control" id="temp9am"
name="temp9am">
                    </div>
    <div class="col-md-6 my-2 d-flex align-items-end justify-content-around">
<button type="submit" class="btn btn-info button" style="margin-left:
90%;">Predict</button>
        </div> </div> </form>   </section> <div>
 <h1><center> {{ prediction }} </center></h1> </div> </body> </html>
```

Rainy DAY

```
<!DOCTYPE html>
<html lang="en">
<head>
    <title>rainy</title>
</head>
<body>
    <div class="bg-image"></div>
    <div class="bg-text">
      <div class="col-md-6 my-2 d-flex align-items-end justify-content-
around"> <a href="/predict"> <button type="submit" class="btn btn-info button"
style="margin-right: 100%;">Back</button></a>
      </div>
      <h1>TOMORROW WILL BE A RAINY DAY</h1>
      <p>MAKE SURE TO CARRY AN UMBRELLA WHEN YOU GO OUT </p>
    </div>   </body> </html>
```

Sunny DAY

```
<html lang="en">
<head> <title> Sunny Day </title> </head>
<body>
    <div class="bg-image"></div>
    <div class="bg-text">
      <div class="col-md-6 my-2 d-flex align-items-end justify-content-
around"> <a href="/predict"> <button type="submit" class="btn btn-info button"
style="margin-right: 100%;">Back</button></a>
      </div>
      <h1>TOMORROW WILL BE A Sunny DAY</h1>
      <p> GO OUT AND MAKE THE BEST OF IT </p>
    </div>   </body> </html>
```

B. SCREENSHOTS

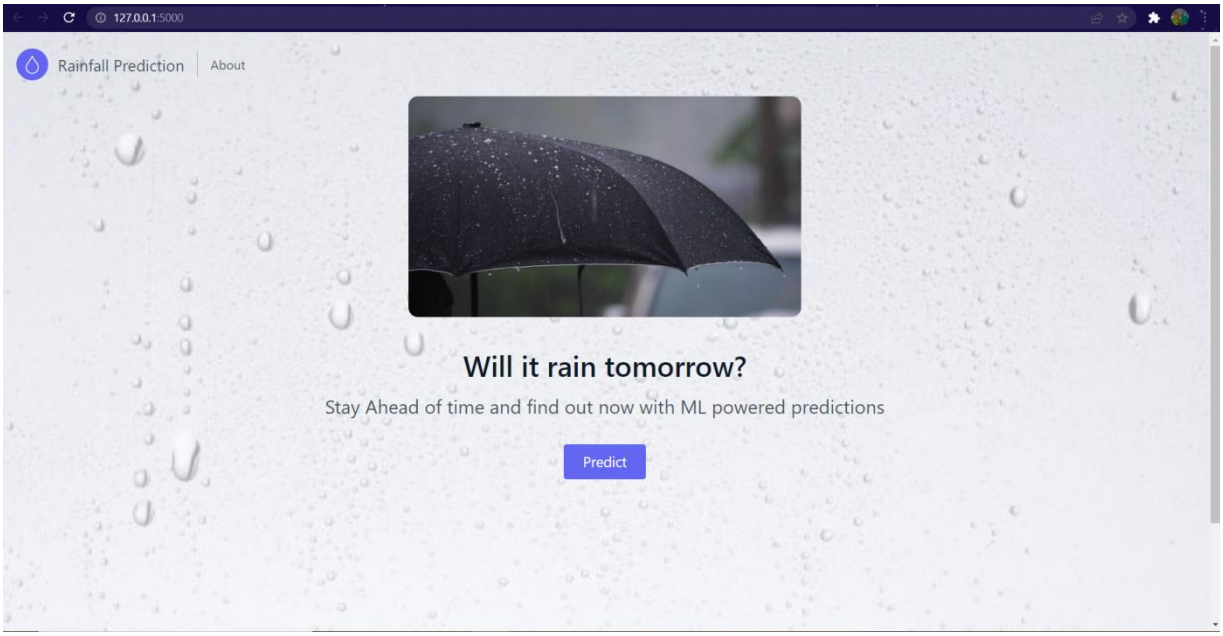


Fig: 6.1: HOME PAGE

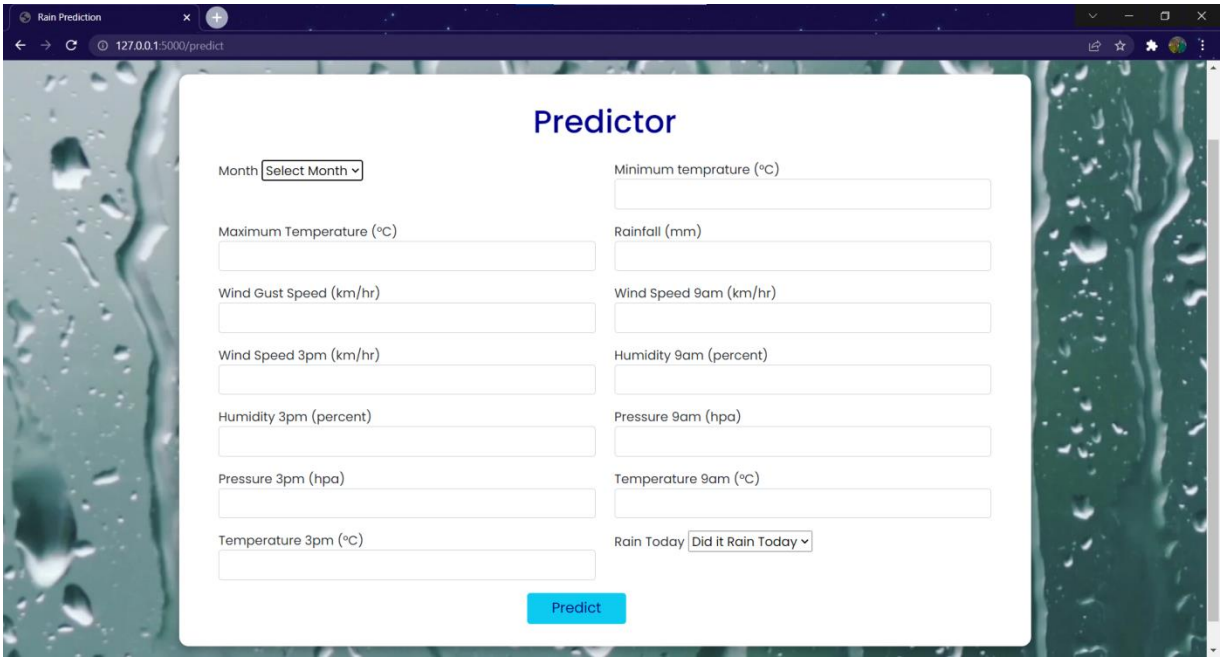


Fig: 6.2: INPUTS PAGE

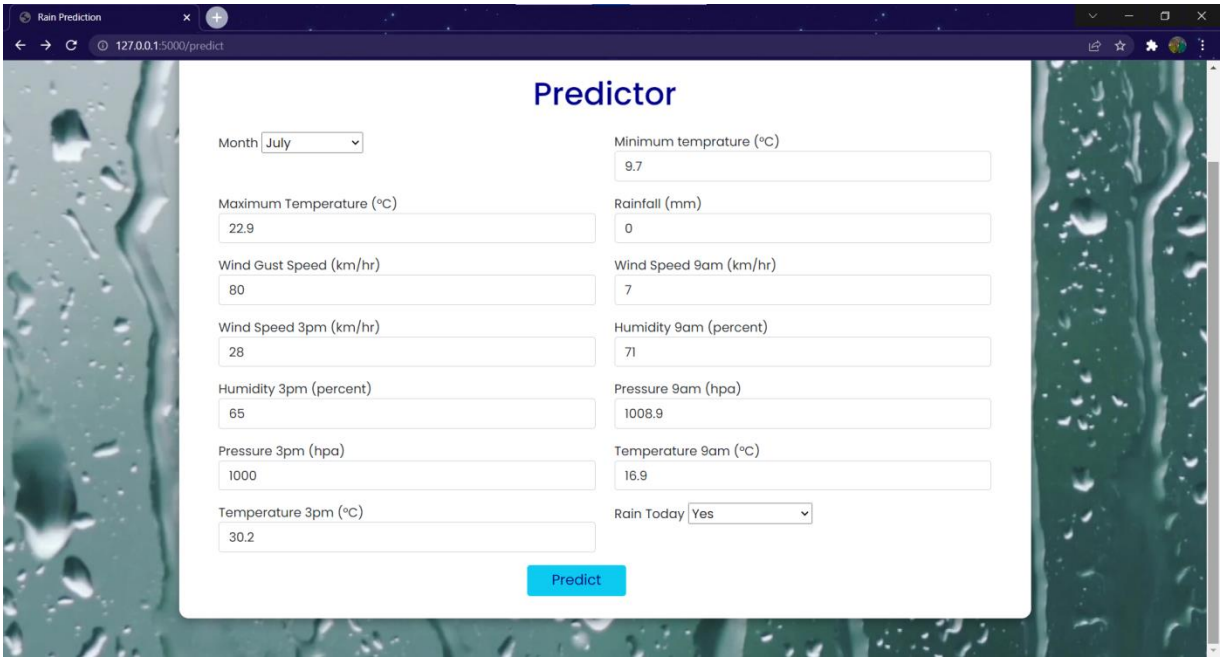


Fig: 6.3: INPUTS GIVEN BY THE USER

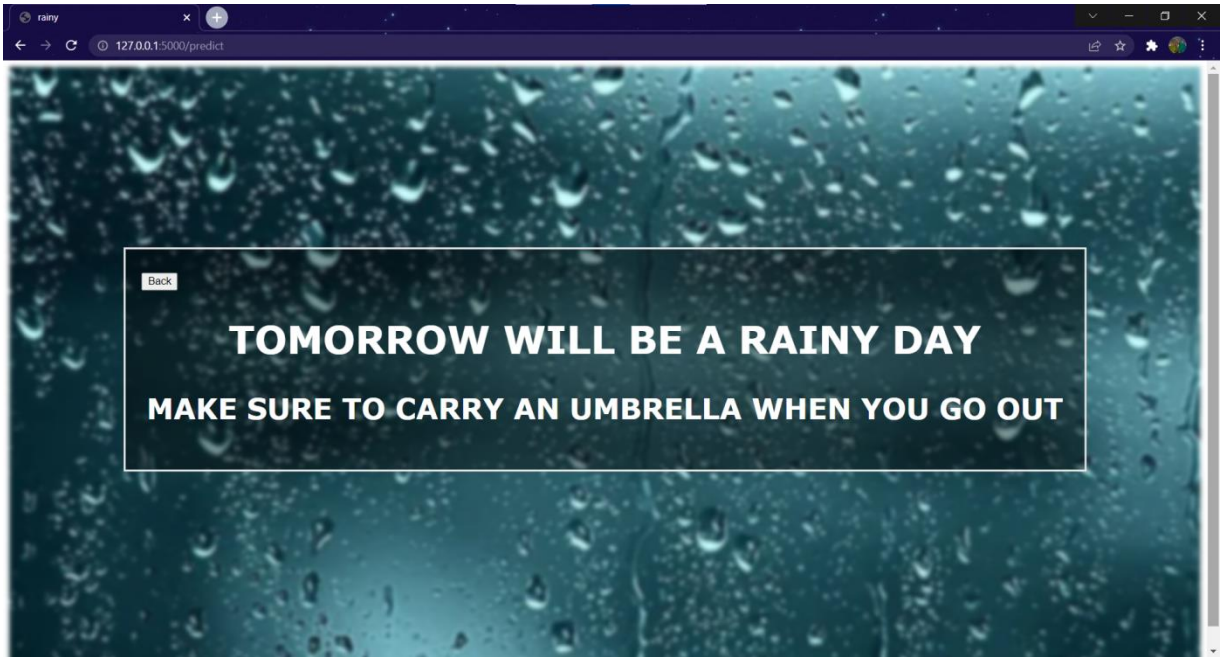


Fig: 6.4: Rainy Day

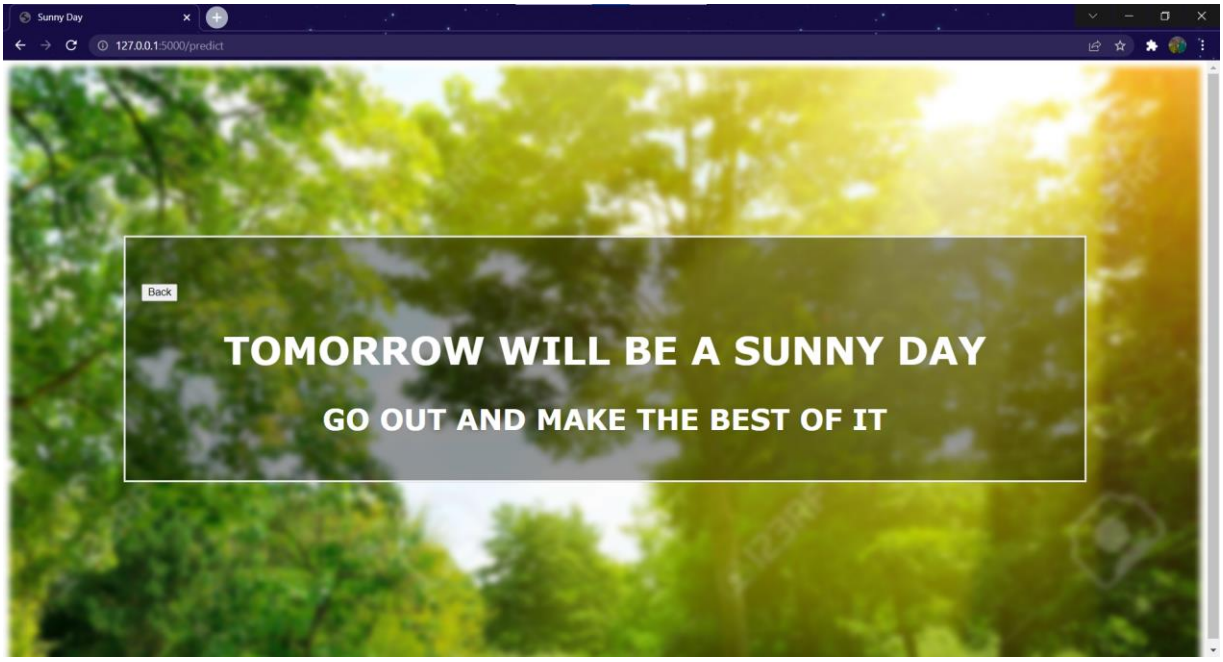
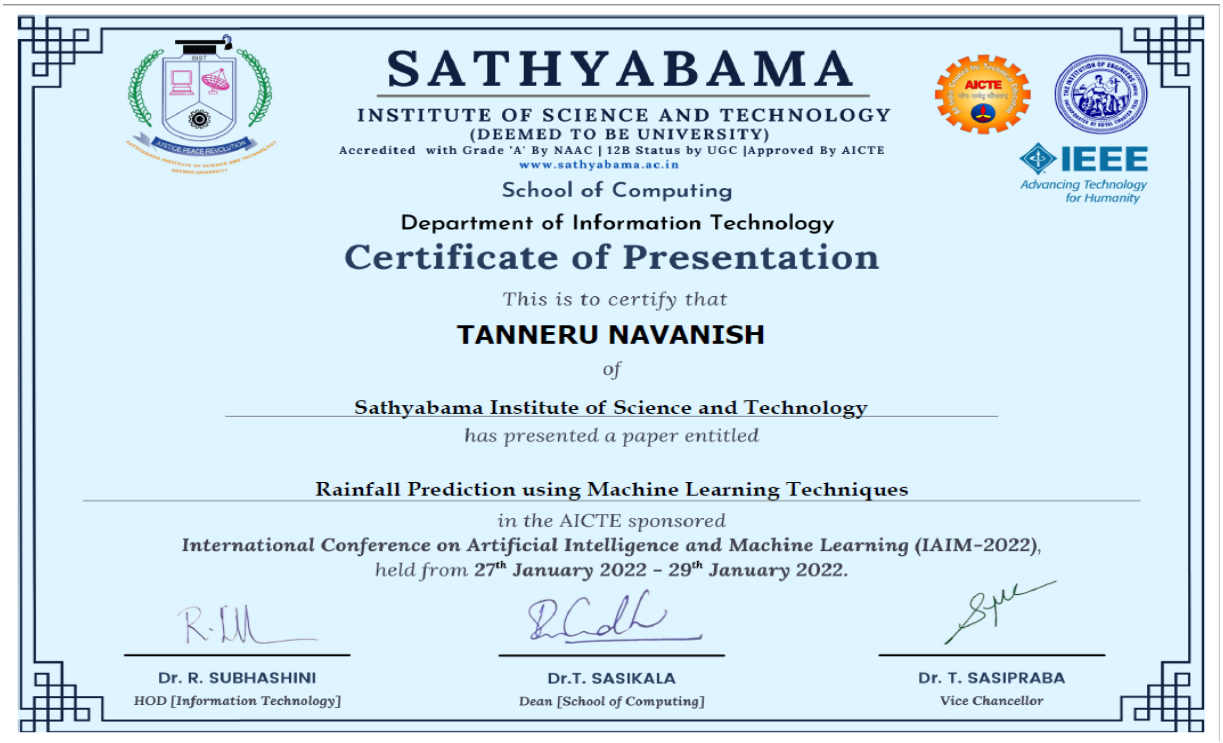


Fig: 6.5: Sunny Day

PAPER PUBLICATION PARTICIPATION CERTIFICATION



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade 'A' By NAAC | 12B Status by UGC |Approved By AICTE
www.sathyabama.ac.in
School of Computing
Department of Information Technology

Certificate of Presentation

This is to certify that

TANNERU NAVANISH

of

Sathyabama Institute of Science and Technology

has presented a paper entitled

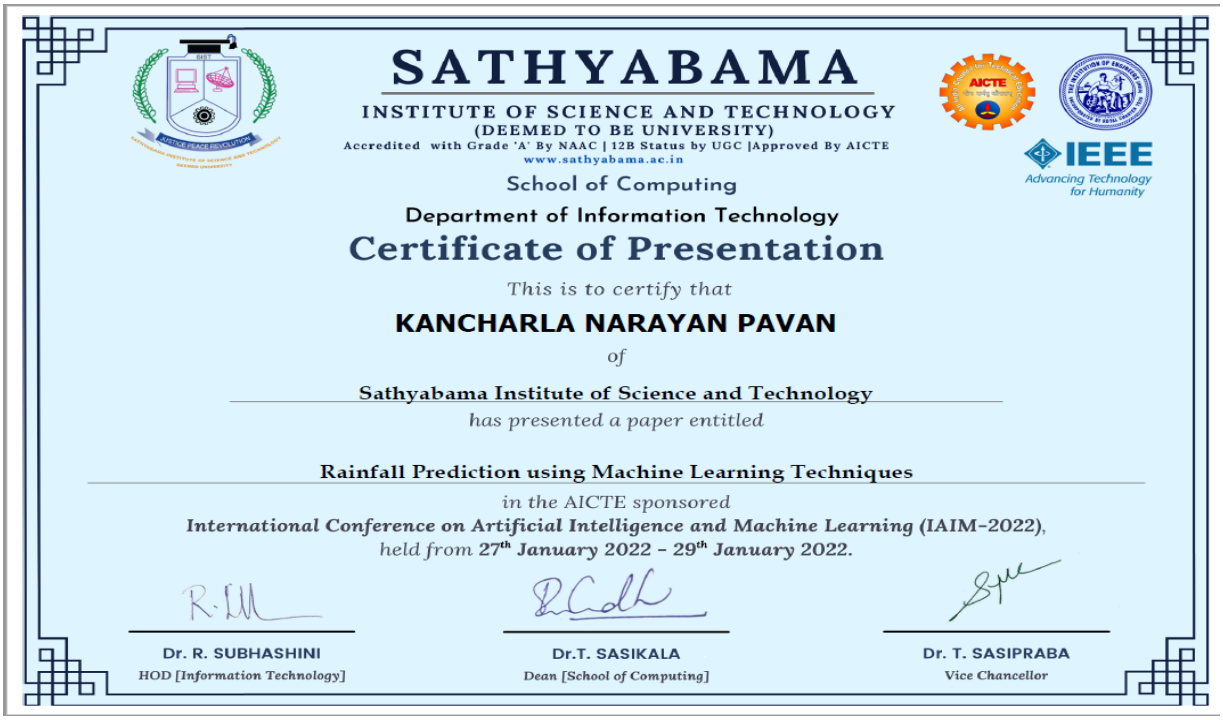
Rainfall Prediction using Machine Learning Techniques

in the AICTE sponsored
International Conference on Artificial Intelligence and Machine Learning (IAIM-2022),
held from 27th January 2022 - 29th January 2022.

Dr. R. SUBHASHINI
HOD [Information Technology]

Dr.T. SASIKALA
Dean [School of Computing]

Dr. T. SASIPRABA
Vice Chancellor



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade 'A' By NAAC | 12B Status by UGC |Approved By AICTE
www.sathyabama.ac.in
School of Computing
Department of Information Technology

Certificate of Presentation

This is to certify that

KANCHARLA NARAYAN PAVAN

of

Sathyabama Institute of Science and Technology

has presented a paper entitled

Rainfall Prediction using Machine Learning Techniques

in the AICTE sponsored
International Conference on Artificial Intelligence and Machine Learning (IAIM-2022),
held from 27th January 2022 - 29th January 2022.

Dr. R. SUBHASHINI
HOD [Information Technology]

Dr.T. SASIKALA
Dean [School of Computing]

Dr. T. SASIPRABA
Vice Chancellor

PUBLICATION WITH PLAGIARISM REPORT

RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES.pdf

by Rainfall Prediction Using Machine Learning Techniq Rainfall Prediction Using Machine Learning Techniq

---

Submission date: 20-Jan-2022 12:54AM (UTC+1000)
Submission ID: 1744148458
File name: RAINFALL_PREDICTION_USING_MACHINE_LEARNING_TECHNIQUES.pdf (1.16M)
Word count: 2330
Character count: 13555

RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES

Tanneru Navanish[1], Kancharla Narayan Pavan[2], Dr. Mercy Paul Selvan [3]

[12]Student,[3]Professor Sathyabama Institute of Science and Technology,Chennai, India

Abstract:

India is a farming nation and its economy is to a great extent dependent on rainforest creation. Downpour estimates are vital and fundamental for all ranchers to examine crop yields. Unsurprising rainfall is the capacity to foresee the climate with the assistance of science and innovation. It is essential to know how much rainfall to utilize water assets, horticultural creation and water arranging proficiently. Various strategies for information mining can foresee rainfall. Information extraction is utilized to appraise rainfall. This article features probably the most well-known rainfall forecast calculations. Logistic Regression, KNeighborsClassifier, Random Forest Classifier and Certificate Tree are a portion of the calculations contrasted with this record. According to a relative perspective, it is feasible to break down how rainfall is accurately anticipated.

Keywords: Rainfall Prediction, model evaluation, oversampling, under sampling , Logistic Regression, KNeighborsClassifier, Random Forest Classifier.

INTRODUCTION

Rainfall arranging is perhaps the most troublesome task. Albeit numerous calculations have been set up, it is truly challenging to precisely anticipate rainfall. In a nation like India, the attention is on crop achievement, disappointment, and water deficiencies at whatever year. A slight change in the stormy season will contrarily affect horticulture. Appropriate rainfall arranging is fundamental to forestall debacle. Legitimate climate anticipating sometimes, like floods and dry seasons, can assist with overseeing farming and forestall fiascos. This article investigates various calculations. Information mining procedures are all around used to recognize rainfall.

LITERATURE REVIEW

Indian Summer Monsoon Rainfall (ISMR) Forecasting using Time Series Data: A Fuzzy-Entropy-Neuro based Expert System

Pritpal Singh / 30 July 2018

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

An Extensive Evaluation of Seven Machine Learning Methods for Rainfall Prediction in Weather Derivatives

Sam Cramer a , Michael Kampouridis a , Alex A. Freitas a , Antonis K. Alexandridis b/ 2017

The primary impact of this movement is to exhibit the advantages of AI calculations, just as the more prominent degree of clever framework than the advanced rainfall determining methods. We analyze and think about the momentum execution (Markov chain stretched out by rainfall research) with the forecasts of the six most notable AI machines: Genetic programming, Vector relapse support, radio organizations, M5 organizations, M5 models, models - Happy. To work with a more itemized appraisal, we led a rainfall overview

utilizing information from 42 metropolitan urban communities.

A Hybrid Model for Statistical Downscaling of Daily Rainfall Sahar Hadi Poura , Shamsuddin Shahida , Eun-Sung Chungb / 2019

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.

RAINFALL PREDICTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES: A REVIEW, 2021

In India, farming is the backbone. Downpour 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. Machine7Learning calculations are significant in foreseeing rainfall.

Weather Forecasting Analysis using Linear and Logistic Regression Algorithm Tanvi Patil 1, Dr. Kamal Shah/ 2021

The reason for the framework is to anticipate the 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. Underlying displaying is utilized to incorporate the fate of demonstrating, AI applications, data trade, and character examination.

PREDICTION OF WEATHER AND RAINFALL FORECASTING USING CLASSIFICATION TECHNIQUES N. DIVYA PRABHA, P. RADHA/ 2019

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 perceptron neural organization. Estimations like MSE (Early Modeling), NMSE (Usually Early Error), and the arrangement of informational collections for transient arranging are clear in the examination of different organizations, like Adanaive. AdaSVM.

RAINFALL PREDICTION BY USING TIME-SERIES DATA IN ANALYSIS OF ARTIFICIAL NEURAL NETWORK MODELS Senthamil Selvi S & Seetha/ 2019

In this paper, Artificial Neural Network (ANN) innovation is utilized to foster a climate anticipating strategy to distinguish rainfall utilizing Indian rainfall information. Along these lines, Feed Forward Neural Network (FFNN) was utilized utilizing the Backpropagation Algorithm. Execution of the two models is assessed dependent on emphasis examination, Mean Square Error (MSE) and Magnitude of Relative Error (MRE). This report likewise gives a future manual for rainfall determining.

Flood Prediction and Rainfall Analysis using Machine Learning YashasAthreya, VaishaliBV, SagarK, SrinidhiHR/ 2021

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.

EXISTING SYSTEM

Agribusiness and the Economic Power of India. Ranchers possibly plant when its downpours. To get a decent collect, you really want great soil,

composts and a decent environment. The climate is vital for each rancher. Unexpected changes in the climate are harming the populace monetarily and genuinely. The climate is one of the most troublesome issues today. The principle motivation behind this climate gauge page is to utilize various techniques for information mining. Isolating, combining, getting trees and nets. Meteorological data is additionally called meteorological data. The most generally utilized boundaries in this article are rainfall, wind speed, temperature, and temperature.

PROPOSED SYSTEM

Rainfall is significant for food conveyance, water assets the board, and every single ecological action. Long haul dry spells or substantial rainfall during extreme development and advancement can altogether diminish yields. India is a rural nation and its economy is to a great extent dependent on farming creation. In this way, rainfall determining is turning out to be an ever-increasing number of significant in farming nations like India. Rainfall gauging has become one of the world's most squeezing science and innovation issues in ongoing hundreds of years.

ADVANTAGES OF PROPOSED SYSTEM

Numerical Weather Prediction
Statistical Weather Prediction
Synoptic Weather Prediction

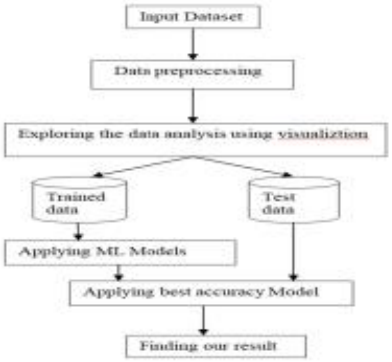
SYSTEM ARCHITECTURE

IMPLEMENTATION

This proposed method is been developed using various AI strategies like Random Forest, Logistic Regression, and, KNeighborsClassifier to predict and recommend for predication of rainfall.

In Dataset the key attributes we are using as follow: Maximum Temperature, Minimum Temperature,Wind Speed, Pressure etc.,

WorkFlow Diagram



The above are the module we are doing for preprocessing our dataset:

MODULES

Data Collection

Data Cleaning

Data Selection

Data Transformation

Data Mining Stage

DataSet

Data Collection

The data utilized in this action was gathered by the meteorological association. The conversation covers the period from 2012 to 2015. This segment of the review talks about the accompanying systems: Cleaning the dad, picking the dad, changing the dad, cleaning the dad.

Data Cleaning

The primary components of coordinated media in this class are the quest for missing data, the quest for bogus data, and the obliteration of weeds. At last, the information cleaning framework was changed into an effective information mining framework.

Data Selection

At this stage, the data identified with the examination tree is the choice tree and is extricated from the informational index. The meteorological informational index had ten attributes that pre-owned two elements later on. Because of the idea of cloud information, all qualities are the equivalent, and the vast majority of the qualities that are absent in sunlight-based information are not utilized in the investigation.

Data Transformation

"It basically came to our notification then, at that point. This is the phase of changing over the chose data into an effective information mining framework. The information document is put away as a Commas Separated Value (CVS) record, and the information is standardized to limit the effect of the information estimation.

Data Mining Stage

The information mining stage is separated into three sections. All calculations were utilized to investigate the information at each stage. The exploratory strategy utilized in this review is separated into the level of trains in the informational collection, their cross-approval, and the excess rate.

Summarize the Data Visualization :

The data set Consist of various attributes like Temperature,Wind Speed, Pressure etc., Depending Up on these attributes the following are differ:

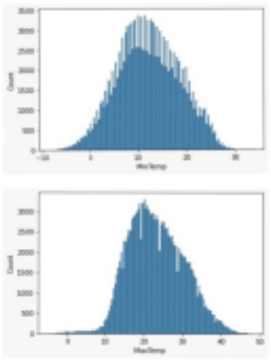


Fig1 : Relation on Count of theTemperature For Rainfall

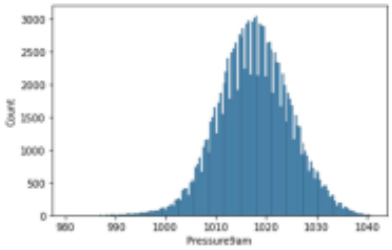


Fig2: Relation on Count of the Pressure for Rainfall

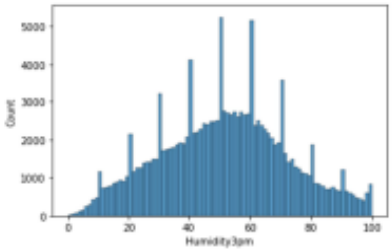


Fig3: Relation on Count of the Humidity for Rainfall

Figure 1,2,3 Tells the relationship Among the rainfall and their Count For the Given DataSet and below Figure 4 is the Heat map.

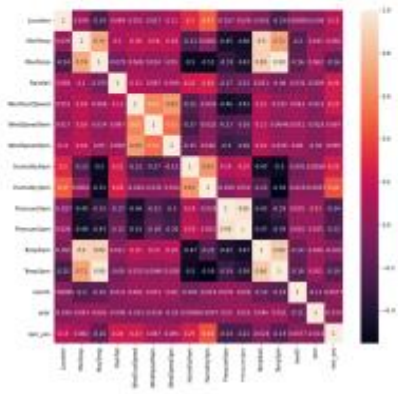


Fig4: Correlation between all the features

Machine learning techniques used:

Because more than one class may be given to a single instance, multi-label classification (MLC) is the best and optimum option. The MLC is embedded into the algorithms used for categorization and model building, which are as follows:

1. Logistic Regression
2. Random Forest
3. KNeighborsClassifier

Logistic regression:

This is a AI strategy used to predict and describe the connection between independent and dependent data values. It predicts the dependent variable by analyzing the relation between the other independent variables. Here is the Accuracy and F1_Score the particular machine learning algorithm.

Accuracy	F1_Score
0.83672635411	0.49511774842

Random Forest:

This is a AI strategy used to solve problems like regression and classification. The algorithm gives an outcome that is based on the predictions of the decision trees by taking the average or mean of the output from different trees. It can be used to predict what will happen in the future. Here is the Accuracy and F1_Score for the particular machine learning algorithm.

Accuracy	F1_Score
0.85578423359	0.60700546669

KNeighborsClassifier:

This is a AI strategy used to solve problems like regression and classification. The algorithm takes nearest values around the unknown variable. Here is the Accuracy and F1_Score for the particular machine learning algorithm.

Accuracy	F1_Score
0.8375808009	0.5655802861

RESULT

With the created Random Forest model we acquired an accuracy of 85.57% with a voting classifier where we created an array called Random forest. which consists of our models and we made the voting parameter set to hard, which tells that our model to make the predictions by the highest votes now, we can implement Random Forest model and we predicted the accuracy score with our test data and here is the individual accuracy of the algorithms.

Algorithm	**Accuracy**
LR	**83.67%**
RF	**85.57%**
KNN	**83.75%**

CONCLUSION

This record tells the best way to concentrate on how much rainfall checked utilizing a machine preparing machine to decrease rainfall data. We utilized different calculations to decide the normal rainfall. We looked at the SVM, Good Forest, Navie Bayes, and MLP (Multilayer perceptron) classifications. From Figure 3 above, we can

presume that normal timberland is a reasonable AI calculation for anticipating rainfall in India.

At present, AI is utilized for. industry. As how much data builds, how much data increments, so we use machines to more readily comprehend the data. The climate conjecture assists a ton with getting a decent score, and the rainfall gives a decent figure. Later on, we intend to build the work on harvest and yield estimating and downpour gauging.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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..

[7] 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.

[8] 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.

[9] 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.

[10] Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." Neural networks 2.5 (1989): 359-366.

[11] Haykin, Simon. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

[12] 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): 211-225.

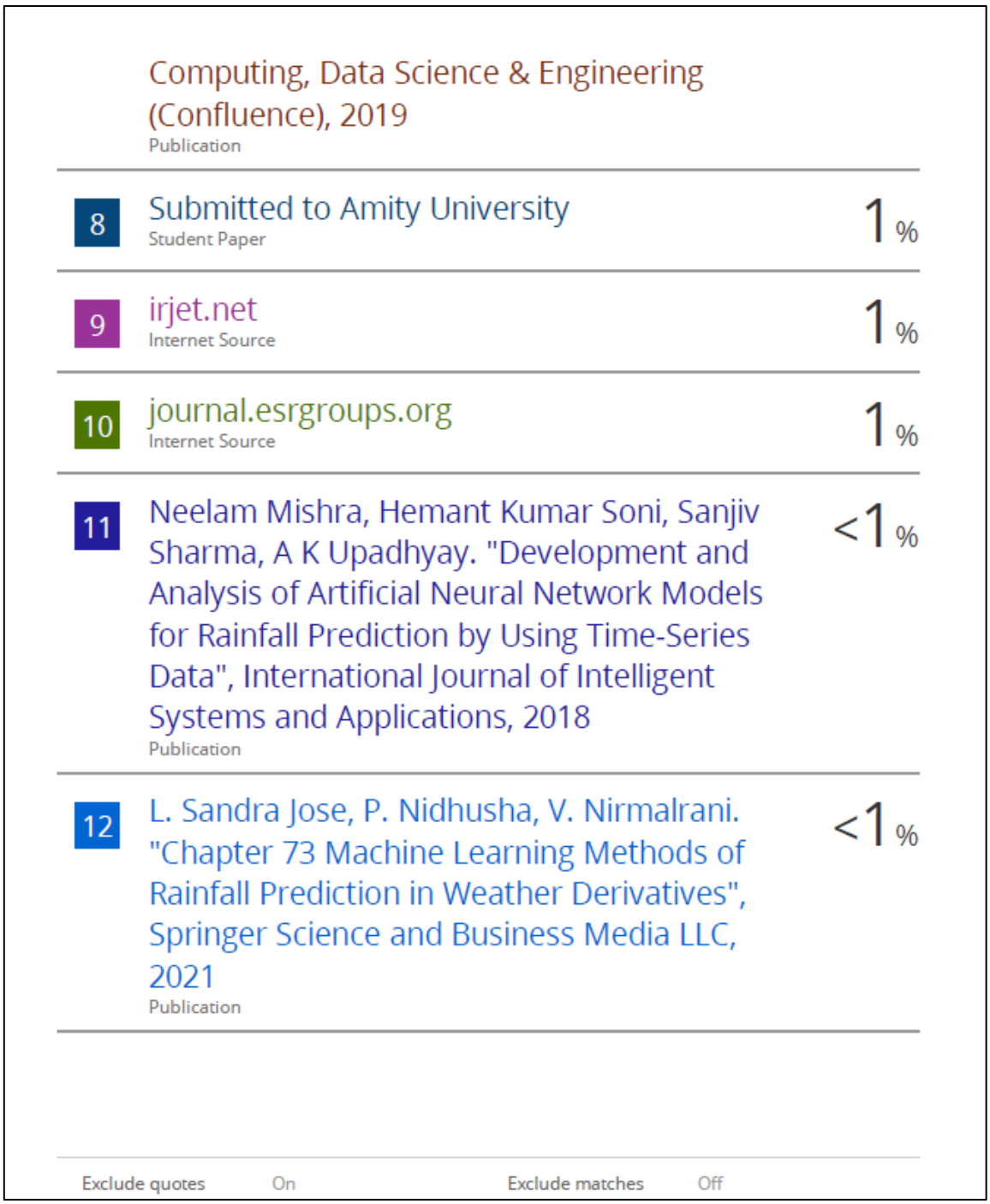
RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES.pdf

ORIGINALITY REPORT

10%	7%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.irjet.net Internet Source	2%
2	academic.oup.com Internet Source	1%
3	www.jetir.org Internet Source	1%
4	Submitted to Aston University Student Paper	1%
5	doaj.org Internet Source	1%
6	Sahar Hadi Pour, Shamsuddin Shahid, Eun-Sung Chung. "A Hybrid Model for Statistical Downscaling of Daily Rainfall", Procedia Engineering, 2016 Publication	1%
7	Gurpreet Singh, Deepak Kumar. "Hybrid Prediction Models for Rainfall Forecasting", 2019 9th International Conference on Cloud	1%


Computing, Data Science & Engineering (Confluence), 2019
Publication
8
Submitted to Amity University
Student Paper
1 %
9
irjet.net
Internet Source
1 %
10
journal.esrgroups.org
Internet Source
1 %
11
Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, A K Upadhyay. "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data", International Journal of Intelligent Systems and Applications, 2018
Publication
<1 %
12
L. Sandra Jose, P. Nidhusha, V. Nirmalrani. "Chapter 73 Machine Learning Methods of Rainfall Prediction in Weather Derivatives", Springer Science and Business Media LLC, 2021
Publication
<1 %
Exclude quotes
On
Exclude matches
Off