|  |  |  |
| --- | --- | --- |
| **D:\PROJECT\rabiproject\DOCUMENT\images.jpg** |  | **docu0005** |

**RAINFALL PREDICTION USING DATA SCIENCE TECHNIQUE**

**A PROJECT REPORT**

***Submitted by***

**V AVANTIKA 211419205023**

**L HARRIAT LINDA 211419205067**

**R NIVESHAA 211419205121**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

***in***

**INFORMATION TECHNOLOGY**

**PANIMALAR ENGINEERING COLLEGE, POONAMALLEE**

**ANNA UNIVERSITY : CHENNAI 600 025**

**JUNE 2022**

**BONAFIDE CERTIFICATE**

Certified that this project report **“RAINFALL PREDICTION USING DATA SCIENCE TECHNIQUE”**is the bonafide work of **“V AVANTIKA (211419205023),**

**L HARRIAT LINDA (211419205067**), **R NIVESHAA (211419205121)”** who carried out the project under my supervision.

**SIGNATURE SIGNATURE**

**Dr. M. HELDA MERCY M.E., Ph.D., Ms. S. KUMARI M.E., (Ph.D.,)**

**HEAD OF THE DEPARTMENT SUPERVISOR**

Associate Professor

Department of Information Technology Department of Information Technology

Panimalar Engineering College Panimalar Engineering College

Poonamallee, Chennai - 600 123 Poonamallee, Chennai - 600 123

Submitted for the project and viva-voce examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_

**SIGNATURE SIGNATURE**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**DECLARATION**

I hereby declare that the project report entitled “**RAINFALL PREDICTION USING DATA SCIENCE TECHNIQUE**” which is being submitted in partial fulfilment of the requirement of the course leading to the award of the ‘Bachelor Of Technology in Information Technology ’ in **Panimalar Engineering College, Autonomous Institution Affiliated to Anna university- Chennai** is the result of the project carried out by me under the guidance and supervision of **Ms. S. KUMARI M.E., (Ph.D.,) in the Department of Information Technology**. I further declared that I or any other person has not previously submitted this project report to any other institution/university for any other degree/ diploma or any other person.

Date**:**

Place**:** Chennai AVANTIKA V

HARRIAT LINDA L

NIVESHAA R

It is certified that this project has been prepared and submitted under my guidance.

Date: **Ms. S. KUMARI M.E., (Ph.D.,)**

Place: Chennai (Associate Professor / IT )

**ACKNOWLEDGEMENT**

A project of this magnitude and nature requires kind co-operation and support from many, for successful completion . We wish to express our sincere thanks to all those who were involved in the completion of this project.

Our sincere thanks to **Our Honorable Secretary and Correspondent, Dr. P. CHINNADURAI, M.A., Ph.D.,** for his sincere endeavor in educating us in his premier institution.

We would like to express our deep gratitude to **Our Dynamic Directors , Mrs. C. VIJAYA RAJESHWARI and Dr. C. SAKTHI KUMAR, M.E.,M.B.A.,PhD.,**and **Dr. SARANYA SREE SAKTHIKUMAR .,B.E.,M.B.A.,** **Ph.D** for providing us with the necessary facilities for completion of this project.

We also express our appreciation and gratefulness to **Our Principal Dr. K. MANI, M.E., Ph.D.,**  who helped us in the completion of the project. We wish to convey our thanks and gratitude to our head of the department, **Dr. M. HELDA MERCY, M.E., Ph.D.,** Department of Information Technology, for her support and by providing us ample time to complete our project.

**Ms. S. KUMARI M.E., (Ph.D.,)** Associate Professor, Department of Information Technology for her guidance throughout the course of our project. Last, we thank our parents and friends for providing their extensive moral support and encouragement during the course of the project.

**ABSTRACT**

Rain forecasting is a quite difficult task. **rain forecasting** is the application of science and technology to predict the **conditions of the atmosphere** for a given **location**and **time.** **rain forecasts**are made by collecting **quantitative data**about the **current state of the atmosphere** at a given place and using meteorology to project how the atmosphere will change. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of rain fall by accuracy calculation. To propose a machine learning-based method to accurately predict the rain fall Index value by prediction results in the form of best accuracy from comparing supervised classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given dataset with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

#### **TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** | V |
|  | **LIST OF FIGURES** | IX |
| **1** | **INTRODUCTION** | **2** |
|  | 1.1 ARTIFICIAL INTELLIGENCE | 3 |
|  | 1.2 MACHINE LEARNING | 3 |
|  | 1.3 PREPARING DATA SET | 4 |
|  | 1.4 NEED OF THE PROJECT | 4 |
|  | 1.5 OVERVEIW OF THE PROJECT | 4 |
| **2** | **LITERATURE SURVEY** | **6** |
|  | 2.1 RAINFALL PREDICTION USING DATA MINIG TECHNIQUES | 6 |
|  | 2.2 RAINFALL PREDICTION USING REGRESSION MODEL | 7 |
|  | 2.3 FEASIBILITY STUDY | 8 |
|  | Technical Feasibility | 8 |
|  | Operational Feasibility | 8 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **3** | | **SYSTEM DESIGN** | | **10** | |
|  | | 3.1 PROPOSED SYSTEM ARCHITECTURE DESIGN | | 10 | |
|  | | 3.2 WORK FLOW DIAGRAM FOR PROPOSED SYSTEM | | 11 | |
|  | | 3.3 USECASE DIAGRAM | | 12 | |
|  | | 3.4 CLASS DIAGRAM | | 13 | |
|  | | 3.5 LIST OF MODULES | | 14 | |
|  | | Module description | | 15 | |
|  | | Data pre processing | | 15 | |
|  | | Exploration of data analysis of  visualization | | 19 | |
|  | | Implementation of algorithm | | 22 | |
|  | | logistic regression | | 25 | |

|  |  |  |
| --- | --- | --- |
| **4** | **REQUIREMENT SPECIFICATION** | **29** |
|  | 4.1 SYSTEM CONFIGURATION | 29 |
|  | 4.2 PROJECT REQUIREMENTS | 29 |
|  | Hardware requirements  Software requirements  4.3 SOFTWARE REQUIREMENTS  Anaconda navigator  Jupyter notebook  Python | 30  30  30  30  31  34  34 |
| **5** | **IMPLEMENTATION** | **38** |
|  | 5.1 SAMPLE CODE | 38 |
|  |
| **6** | **TESTING AND MAINTAINANCE** | **47** |
|  | 6.1. UNIT TESTING | 47 |
|  |  |  |
|  | 6.2. SYSTEM TESTING | 47 |
|  | 6.3 INTEGRATION TESTING | 48 |
|  |  |  |
| **7** | **CONCLUSION AND FUTURE ENHANCEMENT** | **49** |
|  | 7.1 CONCLUSION | 49 |
|  | 7.2 FUTURE ENHANCEMENTS | 49 |
|  | **REFERENCES** | 51 |
|  |  |  |

### **LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE No** | **NAME OF THE TABLE** | **PAGE**  **NO** |
| 3.1 | Architectural design of rainfall detection | 11 |
| 3.2 | Workflow of rainfall detection | 12 |
| 3.3 | Use case diagram of rainfall detection | 12 |
| 3.4 | Class diagram of rainfall detection | 13 |
| 3.5 | Data validation(a) | 17 |
|  | Data validation(b) | 18 |
| 3.6 | Exploration of data(a)  Exploration of data(b) | 20  20 |
| 3.7 | Logistic regression(a)  Logistic Regression(b) | 20  26 |
| 4.1 | Anaconda Navigator | 27 |

# **CHAPTER 1**

**INTRODUCTION**

Rain forecasting is a quite difficult task. **rain forecasting** is the application of science and technology to predict the **conditions of the atmosphere** for a given **location**and **time**. **rain forecasts**are made by collecting **quantitative data**about the **current state of the atmosphere**at a given place and using meteorology to project how the atmosphere will change. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset.

**1.1 ARTIFICIAL INTELLIGENCE**

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Artificial intelligence (AI) is [intelligence](https://en.wikipedia.org/wiki/Intelligence) demonstrated by [machines](https://en.wikipedia.org/wiki/Machine), as opposed to the natural intelligence [displayed by humans](https://en.wikipedia.org/wiki/Human_intelligence) or [animals](https://en.wikipedia.org/wiki/Animal_cognition). Leading AI textbooks define the field as the study of "[intelligent agents](https://en.wikipedia.org/wiki/Intelligent_agent)" any system that perceives its environment and takes actions that maximize its chance of achieving its goals. Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the [human mind](https://en.wikipedia.org/wiki/Human_mind), such as "learning" and "problem solving", however this definition is rejected by major AI researchers.

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

**1.2 MACHINE LEARNING**

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

**1.3 PREPARING THE DATASET:**

This dataset contains 142193 records of features extracted from, which were classified into 2 classes:

* Chance of Rain Tomorrow
* No Chance of Rain Tomorrow

**1.4 NEED OF THE PROJECT**

The predicting Rainfall problem, random forest 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.

**1.5 OVERVIEW OF THE PROJECT**

* These reports are to the investigation of applicability of machine learning techniques for rainfall forecasting in operational conditions.
* Finally, it highlights some observations on future research issues, challenges, and needs.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 TITLE** : **Rainfall Prediction using Data Mining Techniques**

**Author**: Jyothis Joseph, Ratheesh T K

**Year**: 2013

Rainfall becomes a significant factor in agricultural countries like India. Rainfall prediction has become one of the most scientifically and technologically challenging problems in the world. A wide variety of rainfall forecast methods are available. There are mainly two approaches to predict rainfall. They are Empirical method and dynamical method. The empirical approach is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. The most widely used empirical approaches used for climate prediction are regression, artificial neural network, fuzzy logic and group method of data handling. This paper uses data mining techniques such as clustering and classification techniques for rainfall prediction. Rainfall predictions are made by collecting quantitative data about the current state of the atmosphere. Numerous machine learning algorithms can learn complex mappings from inputs to outputs, based solely on samples and require limited. Accurate prediction of rainfall is a difficult task due to the dynamic nature of the atmosphere. To predict the future’s rainfall condition, the variation in the conditions in past years must be utilized. We have proposed the use of linear regressions by making use of various parameters such as temperature, humidity and wind

**2.2 TITLE: Rainfall Prediction using Regression Model**

**Author**: J.Refonaa, M. Lakshmi, Raza Abbas, Mohammad Raziullha

**Yea**r: 2019

Artificial intelligence is been widely used in all the applications and weather forecasting is not an exception. When it comes to weather forecasting, rainfall prediction is one of the most widely used research areas as numerous lives and property damages occur due to this. Intense rainfall has abundant impacts on society and on our daily life from cultivation to disaster measures. Previous rainfall prediction models that are widely used, makes use of many the complicated blend of mathematical instruments which was insufficient to get a higher classification rate. In this project, we propose new novel methods for predicting monthly rainfall using linear regression analysis. Rainfall predictions are made by collecting quantitative data about the current state of the atmosphere. Numerous machine learning algorithms can learn complex mappings from inputs to outputs, based solely on samples and require limited. Accurate prediction of rainfall is a difficult task due to the dynamic nature of the atmosphere. To predict the future’s rainfall condition, the variation in the conditions in past years must be utilized. We have proposed the use of linear regressions by making use of various parameters such as temperature, humidity and wind. The proposed model tends to forecast rainfall based on the previous records of a particular geographic area, therefore, this prediction will prove to be much reliable. The performance of the model is more accurate when compared with traditional rainfall prediction systems.

###### **2.3 FEASIBIITY STUDY**

A feasibility study is carried out to select the best algorithm that meets performance requirements. The main aim of the feasibility study activity is to determine that it would be technically feasible to detect the fake news. The document provides the feasibility of the project that is being designed and lists various areas that were considered very carefully during the feasibility study of this project such as Technical feasibility and Operational feasibility.

**Technical Feasibility**

The system must be evaluated from the technical point of view first. The assessment of this feasibility must be based on an outline design of the system requirement in the terms of input, output, programs and procedures. Having identified an outline system, the investigation must go on to suggest the type of equipment, required method developing the system, of running the system once it has been designed. The project should be developed such that the necessary functions and performance are achieved within the constraints.

The project is developed within latest technology. Through the technology may become obsolete after some period of time, due to the fact that never version of same software supports older versions, the system may still be used. So, there are minimal constraints involved with this project. As the system has been developed using Java the project is technically feasible. 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 being placed on the client. The developed system must have modest requirements, as only minimal or null changes are required for implementing this system.

**Operational 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. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system

**CHAPTER 3**

## SYSTEM DESIGN

### **3.1 PROPOSED SYSTEM ARCHITECTURE DESIGN**

This research purposes a rainfall prediction framework (Figure 1) using a machine learning fusion technique for smart cities. The proposed framework mainly consists of two layers: training and testing. Both of these layers further include multiple stages. The first stage of the training layer deals with the extraction of weather attributes from technologically advanced sensors in the smart city. However, in this research, we have extracted a real-time pre-labeled dataset of rainfall prediction from a weather forecasting website [15] of the city of Lahore. The dataset consists of 25,919 instances and 11 features, out of which 10 features are independent and 1 is dependent (output class). The data preprocessing stage consists of three activities: (1) cleaning, (2) normalization, and (3) splitting. The data cleaning process aims to remove the missing values in the dataset by using the technique of mean imputation. The normalization technique brings the attribute values within a particular range. These cleaning and normalization activities aid the classifiers in obtaining maximum accuracy. In the third activity of the pre-processing stage, cleaned and normalized data is divided into two subsets: training data and test data, with a 70:30 ratio of class split rule. After performing the tasks of pre-processing activities, the dataset is ready for the stage of classification, where training and test datasets are both given as input to four classification techniques (DT, NB, KNN, and SVM). All of these algorithms are optimized iteratively during training and testing in order to achieve higher accuracy. After the classification process, the trained models are given as input to the fuzzy layer, which deals with the development and implementation of fuzzy logic for final prediction.

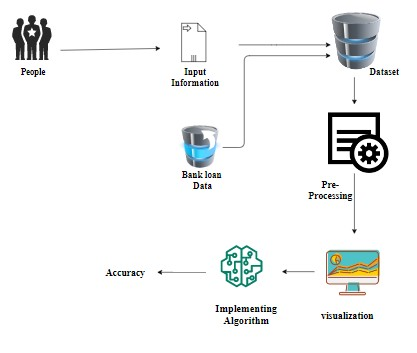


Fig 3.1 architecture design of rainfall detection

**3.2 WORK FLOW DIAGRAM FOR PROPOSED SYSTEM**

Raw, unprocessed reflectivity products can already provide useful visual information about the spatial distribution of rainfall fields. However, in order to use weather radar observations for quantitative studies (e.g. in hydrological modelling or for assimilation into Numerical Weather Prediction models), the data has to be carefully processed in order to account for typical errors sources such as ground echoes (clutter), attenuation of the radar signal, or uncertainties in the Z/R relationship.

Moreover, it might be necessary to transfer the data from polar coordinates to Cartesian grids, or to combine observations from different radar locations in overlapping areas on a common grid (composition). And in the end, you would typically like to visualise the spatial rainfall distribution on a map. Many users also need to quantify the potential error (uncertainty) of their data-based rainfall estimation

Source Data

Accuracy

Testing Dataset

Training Dataset

Classification ML Algorithms

Data Processing and Cleaning

Fig 3.2 workflow of rainfall detection

**3.3 USECASE DIAGRAM**

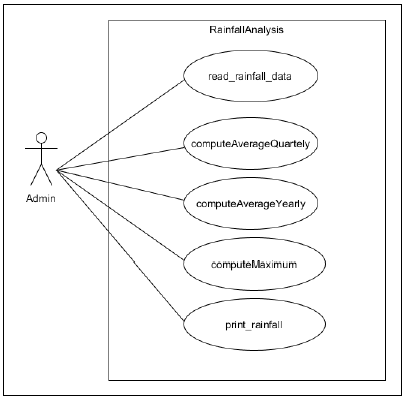


Fig 3.3 Use Case diagram of rainfall detection

**3.4 CLASS DIAGRAM**

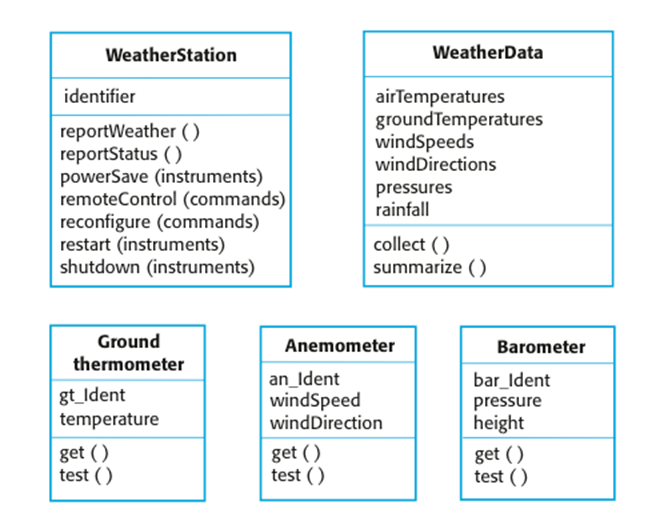


Fig 3.3 Class diagram of rainfall detection

**3.5 List Of Modules:**

* Data Pre-processing
* Data Analysis of Visualization
* Implementing Algorithm

**Module description:**

**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’s [Pandas library](https://pandas.pydata.org/) and specifically, it focus on probably the biggest data cleaning task, **missing values** and it able to **more**[**quickly clean data**](https://www.dataoptimal.com/data-cleaning-with-python-2018/). It wants to **spend less time cleaning data,** and more time exploring and modeling.

Some of the 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:

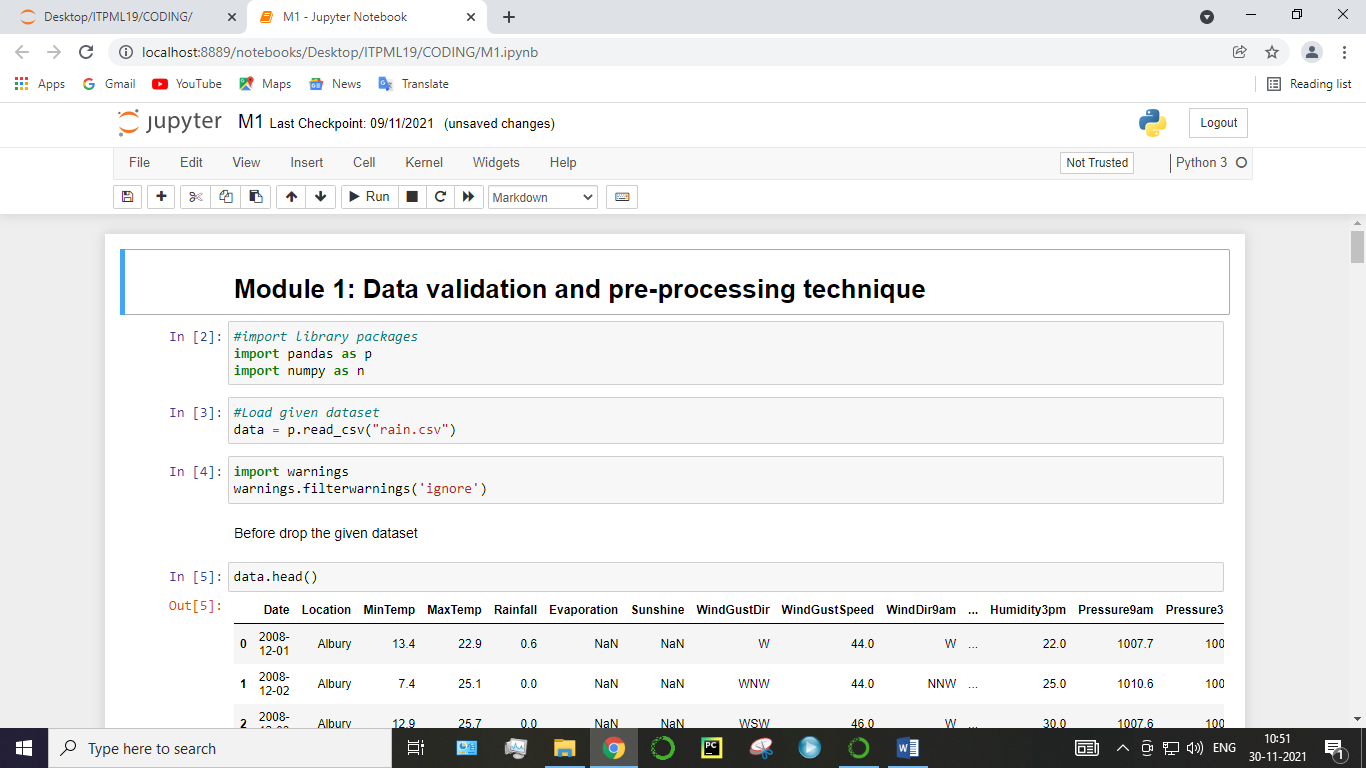
* User forgot to fill in a field.
* Data was lost while transferring manually from a legacy database.
* There was a programming error.
* Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

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

* import libraries for access and functional purpose and 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

**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.



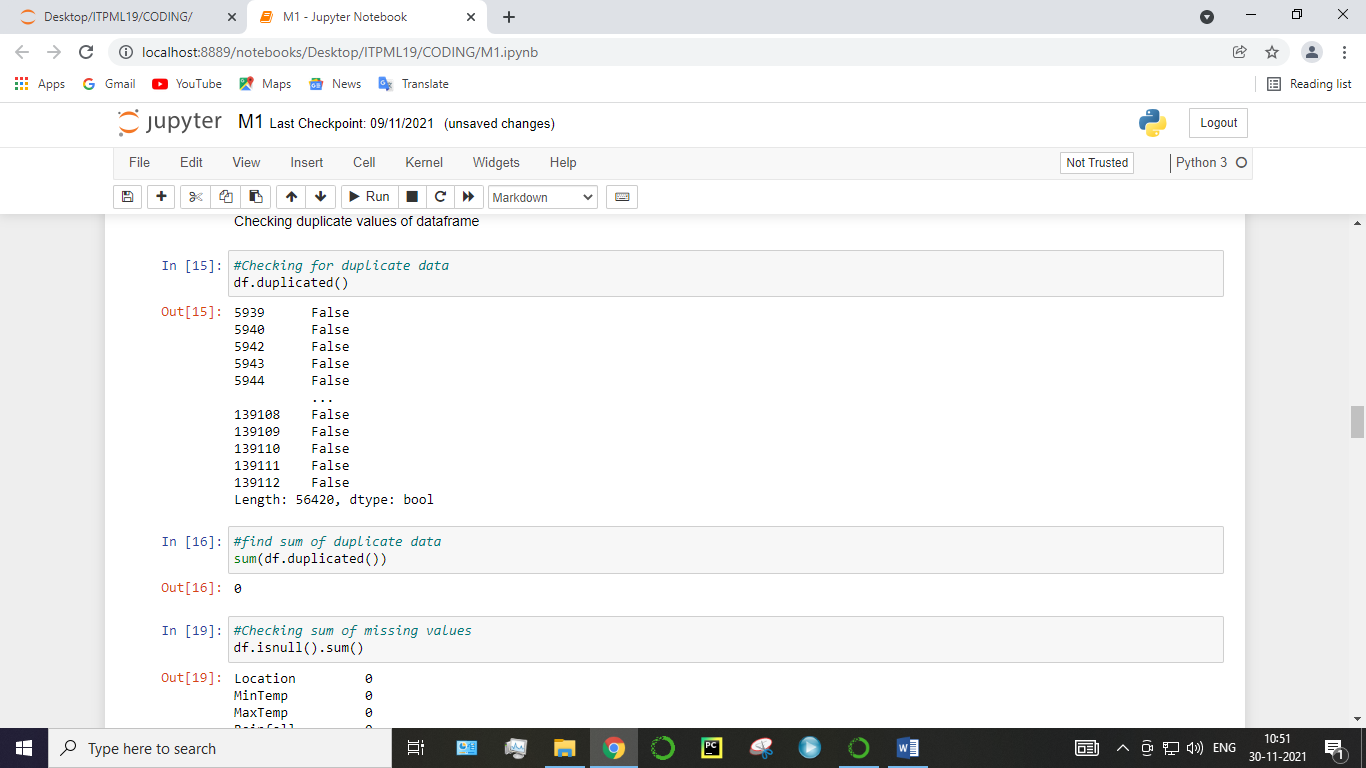
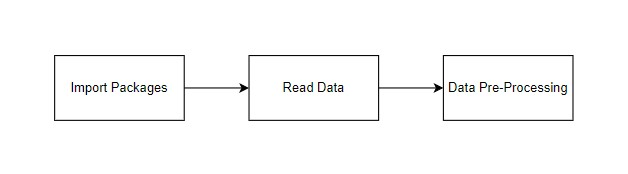


Fig 3.5.1 data validation(b)

**Module Diagram:**

****

**GIVEN INPUT EXPECTED OUTPUT**

**input :** data

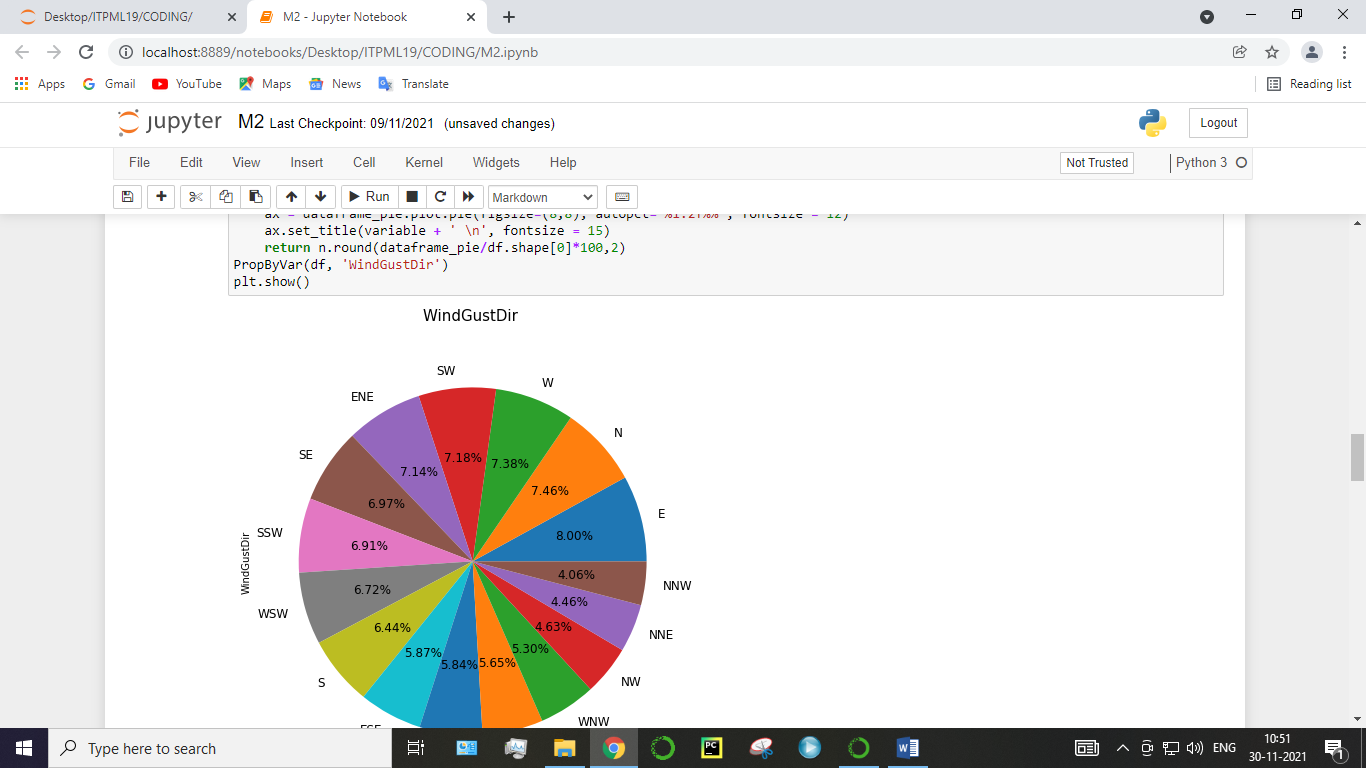
**output :** removing noisy data

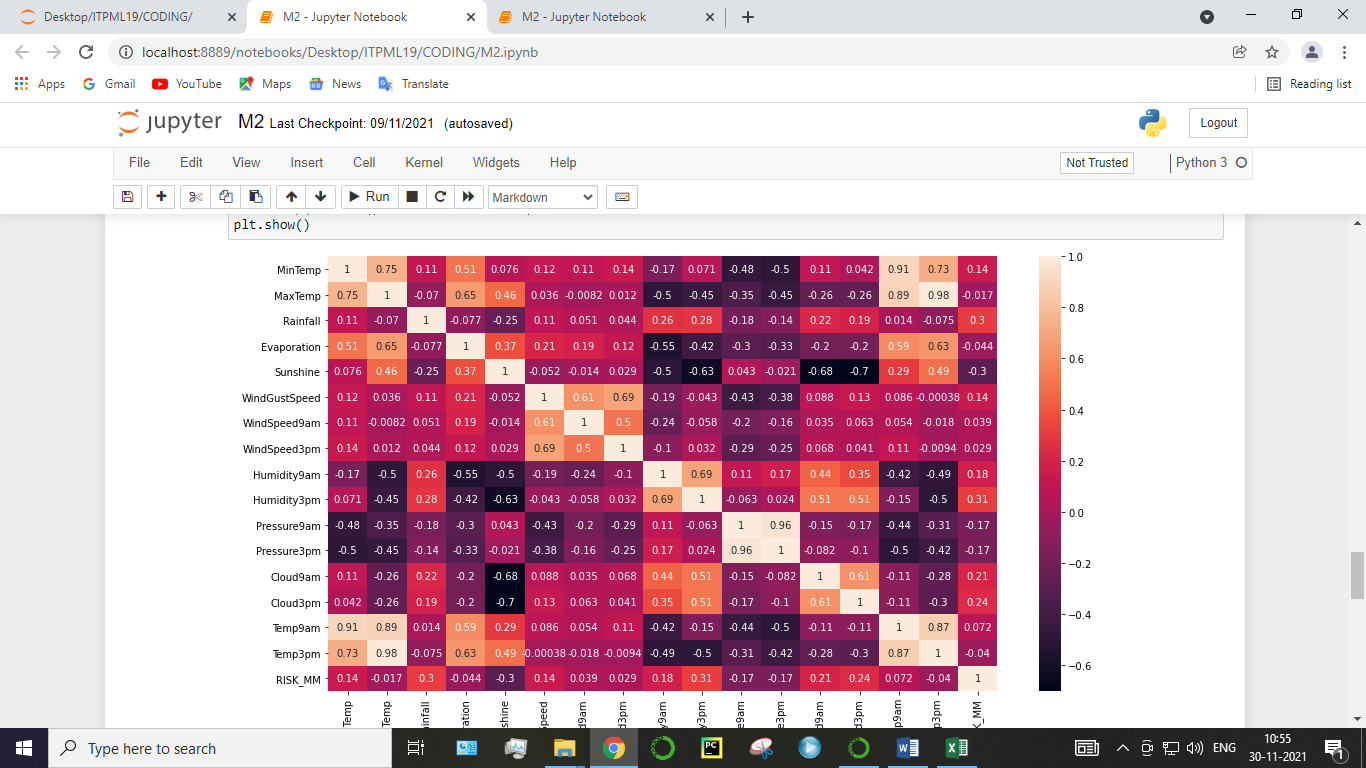
**Exploration data analysis of visualization**

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.





**GIVEN INPUT EXPECTED OUTPUT**

**input :** data

**output :** visualized data

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.

**False Positive (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.

**Implementing Algorithm**

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.

In the example below **1** different algorithms are compared:

* Logistic Regression

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.

**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 need 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)

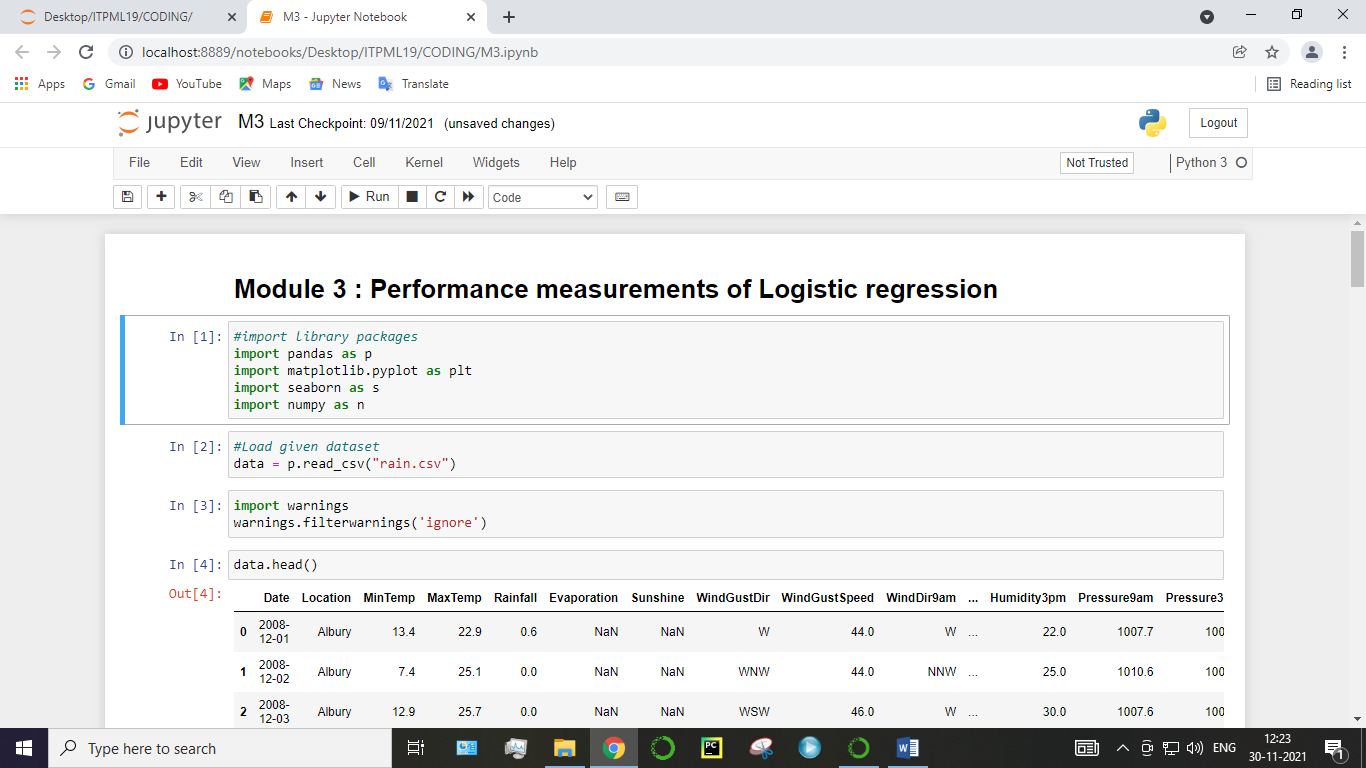
**MODULE – 3**

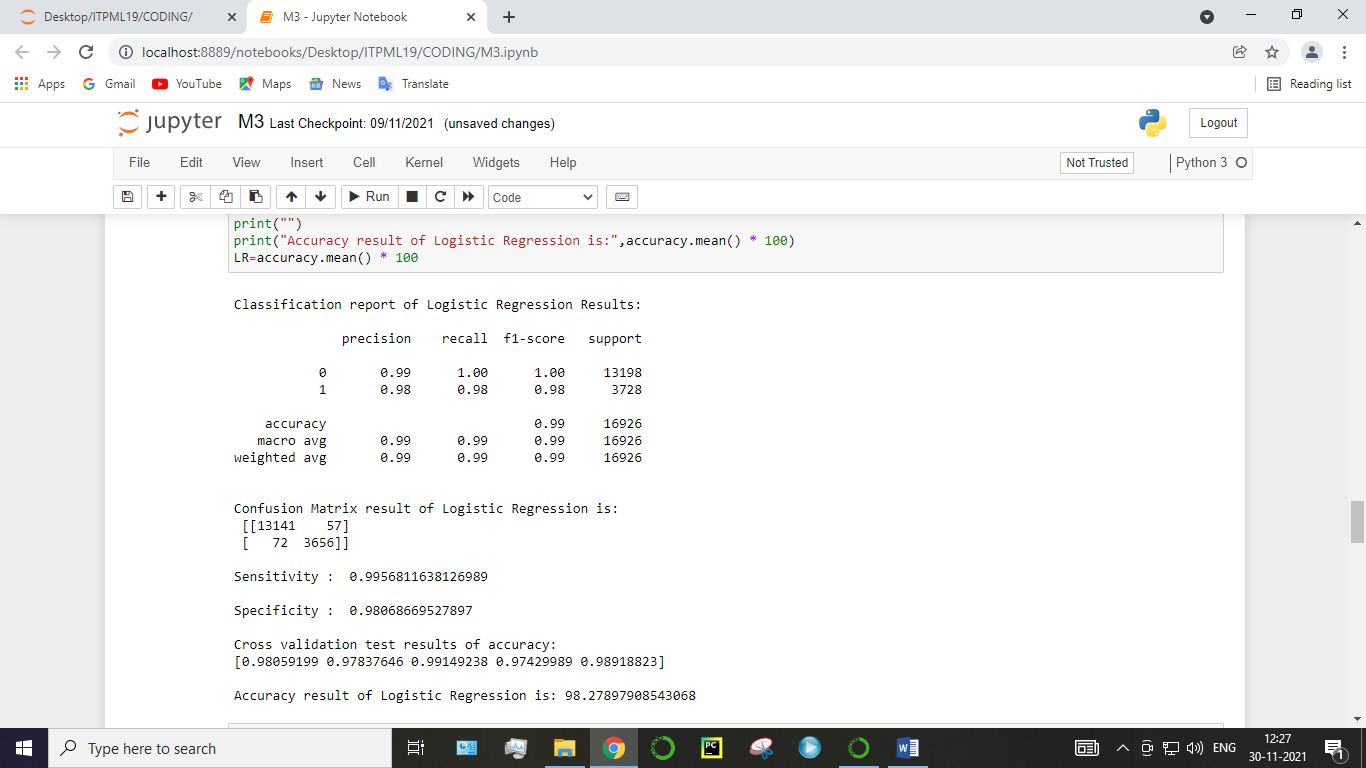
**Logistic Regression:**

It is a statistical method for analysing 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.





**CHAPTER 4**

## REQUIREMENTS SPECIFICATION

### **4.1 SYSTEM CONFIGURATION**

This project can run on commodity hardware. We ran entire project on an Intel I5 processor with 8 GB Ram, 2 GB Nvidia Graphic Processor, It also has 2 cores which runs at 1.7 GHz, 2.1 GHz respectively. First part of the is training phase which takes 10-15 mins of time and the second part is testing part which only takes few seconds to make predictions and calculate accuracy.

**4.2 PROJECT REQUIREMENTS**

**General:**

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Hardware requirements

2. Software requirements

**1. Software Requirements:**

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

**2. Hardware requirements:**

Processor : Pentium IV/III

Hard disk : minimum 80 GB

RAM : minimum 2 GB

**4.3 SOFTWARE DESCRIPTION**

**ANACONDA NAVIGATOR**

Anaconda is a [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source) distribution of the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)) programming languages for [scientific computing](https://en.wikipedia.org/wiki/Scientific_computing) ([data science](https://en.wikipedia.org/wiki/Data_science), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) applications, large-scale data processing, [predictive analytics](https://en.wikipedia.org/wiki/Predictive_analytics), etc.), that aims to simplify [package management](https://en.wikipedia.org/wiki/Package_management) and deployment. Package versions are managed by the [package management system](https://en.wikipedia.org/wiki/Package_manager) “Conda”. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the [Conda](https://en.wikipedia.org/wiki/Conda_(package_manager)) package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the conda install command or using the pip install command that is installed with Anaconda. [Pip packages](https://en.wikipedia.org/wiki/Pip_(package_manager)) provide many of the features of conda packages and in most cases they can work together. Custom packages can be made using the conda build command, and can be shared with others by uploading them to Anaconda Cloud, [PyPI](https://en.wikipedia.org/wiki/Python_Package_Index) or other repositories. The default installation of Anaconda2 includes Python 2.7 and Anaconda3 includes Python 3.7. However, you can create new environments that include any version of Python packaged with conda.

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository.

Anaconda. Now, if you are primarily doing data science work, Anaconda is also a great option. Anaconda is created by Continuum Analytics, and it is a Python distribution that comes preinstalled with lots of useful python libraries for data science.

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

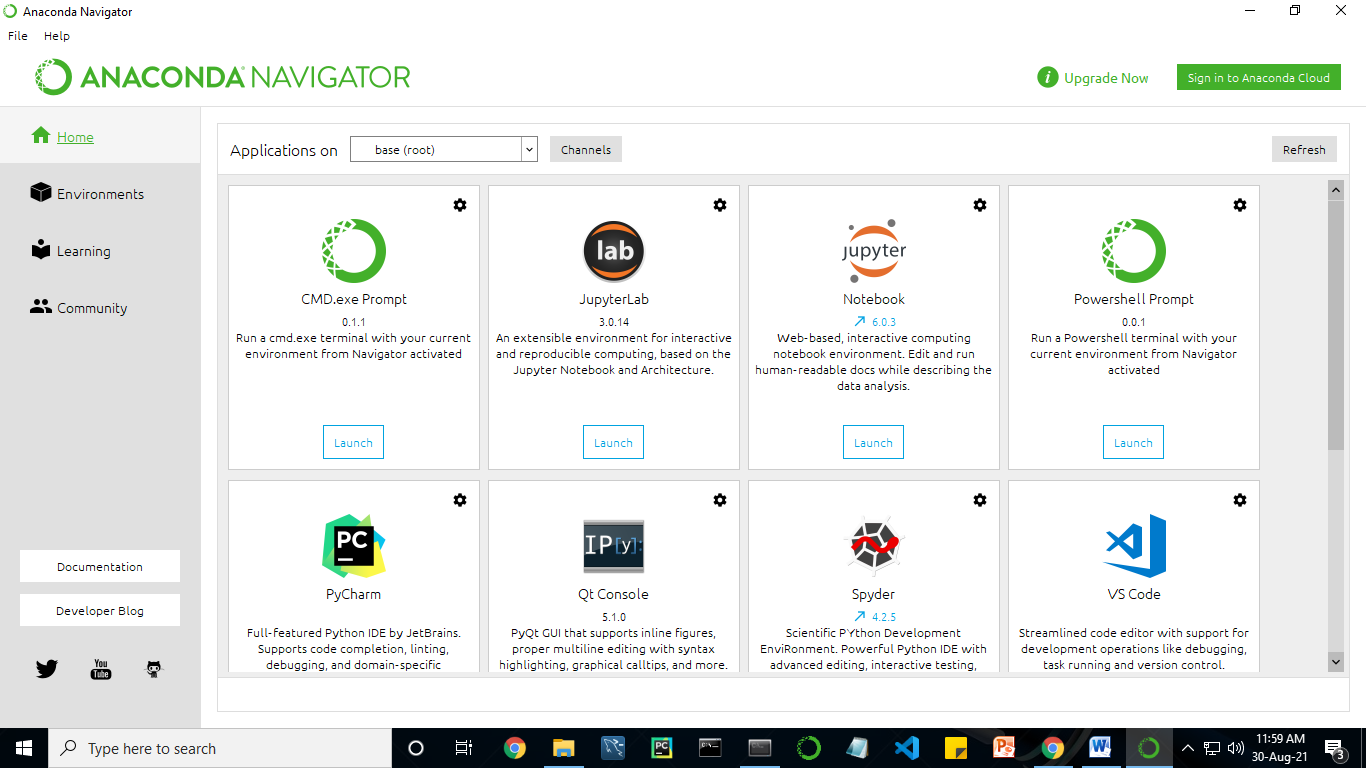
In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

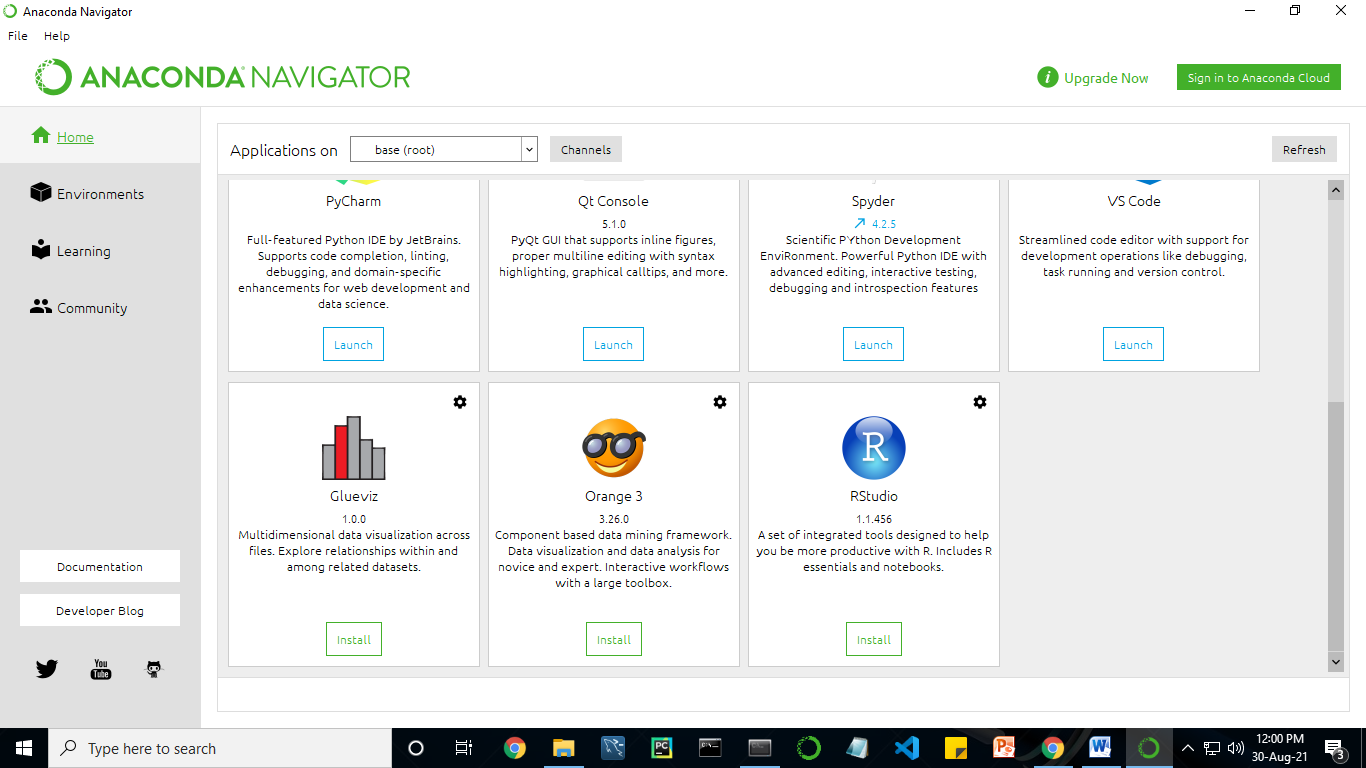
The command-line program conda is both a package manager and an environment manager. This helps data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

The following applications are available by default in Navigator:

* [JupyterLab](https://jupyterlab.readthedocs.io/en/stable/)
* [Jupyter Notebook](https://jupyter.readthedocs.io/en/latest/)
* [Spyder](https://www.spyder-ide.org/)
* [PyCharm](https://www.jetbrains.com/pycharm/documentation/)
* [VSCode](https://code.visualstudio.com/docs)
* [Glueviz](http://glueviz.org/en/stable/)
* [Orange 3 App](http://orange.biolab.si/docs/)
* [RStudio](http://docs.rstudio.com/)
* Anaconda Prompt (Windows only)
* Anaconda PowerShell (Windows only)





Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution.

Navigator allows you to launch common Python programs and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository.

Anaconda comes with many built-in packages that you can easily find with conda list on your anaconda prompt. As it has lots of packages (many of which are rarely used), it requires lots of space and time as well. If you have enough space, time and do not want to burden yourself to install small utilities like JSON, YAML, you better go for Anaconda.

**4.3.2 JUPYTER NOTEBOOK**

* The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

**4.3.3 PYTHON**

Python is an [interpreted](https://en.wikipedia.org/wiki/Interpreted_language) [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) [general-purpose programming language](https://en.wikipedia.org/wiki/General-purpose_programming_language). Its design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with its use of [significant indentation](https://en.wikipedia.org/wiki/Off-side_rule). Its [language constructs](https://en.wikipedia.org/wiki/Language_construct) as well as its [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aim to help [programmers](https://en.wikipedia.org/wiki/Programmers) write clear, logical code for small and large-scale projects.

Python is [dynamically-typed](https://en.wikipedia.org/wiki/Type_system#DYNAMIC) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigms), including [structured](https://en.wikipedia.org/wiki/Structured_programming) (particularly, [procedural](https://en.wikipedia.org/wiki/Procedural_programming)), object-oriented and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). It is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

[Guido van Rossum](https://en.wikipedia.org/wiki/Guido_van_Rossum) began working on Python in the late 1980s, as a successor to the [ABC programming language](https://en.wikipedia.org/wiki/ABC_(programming_language)), and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as [list comprehensions](https://en.wikipedia.org/wiki/List_comprehension) and a garbage collection system using [reference counting](https://en.wikipedia.org/wiki/Reference_counting). Python 3.0 was released in 2008 and was a major revision of the language that is not completely [backward-compatible](https://en.wikipedia.org/wiki/Backward_compatibility). Python 2 was discontinued with version 2.7.18 in 2020.Python consistently ranks as one of the most popular programming languages

**Python Packages**:

➢ FLASK

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries.

➢ PANDAS

The name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.” But I think it'sjust a cute name to a super- useful Python library!

➢ NUMPY

➢ TENSORFLOW

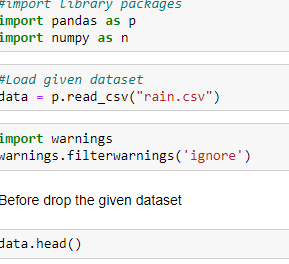
It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow.

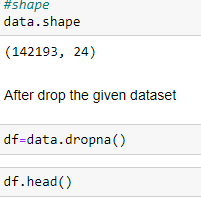
**CHAPTER 5**

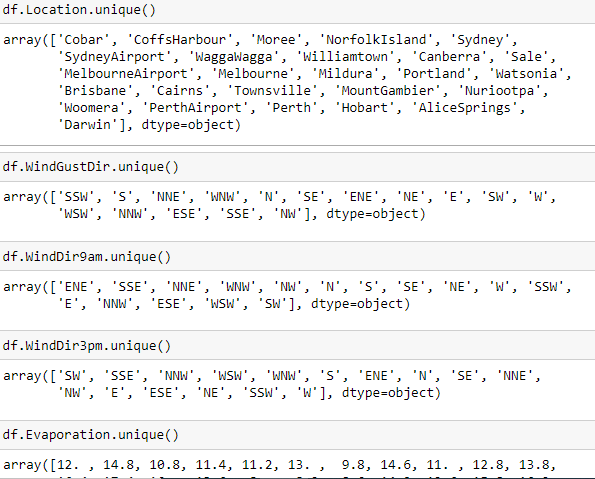
## IMPLEMENTATION

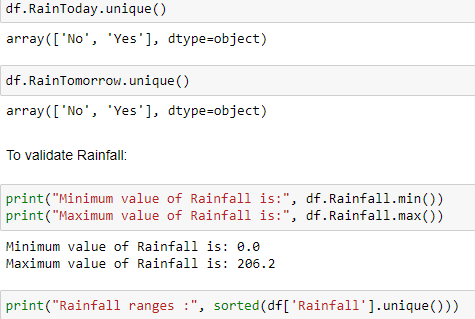
**5.1 SAMPLE CODE:**

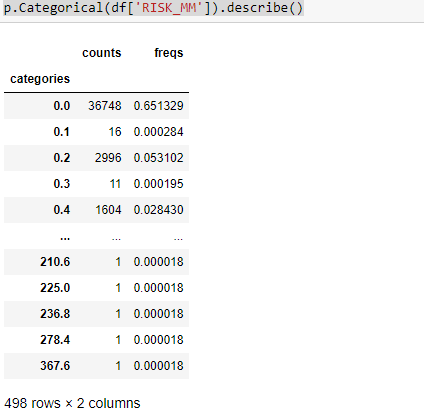
# Module 1: Data validation and pre-processing technique





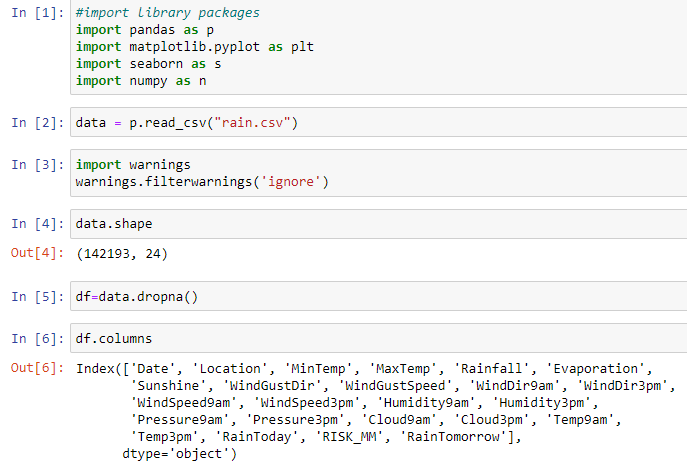


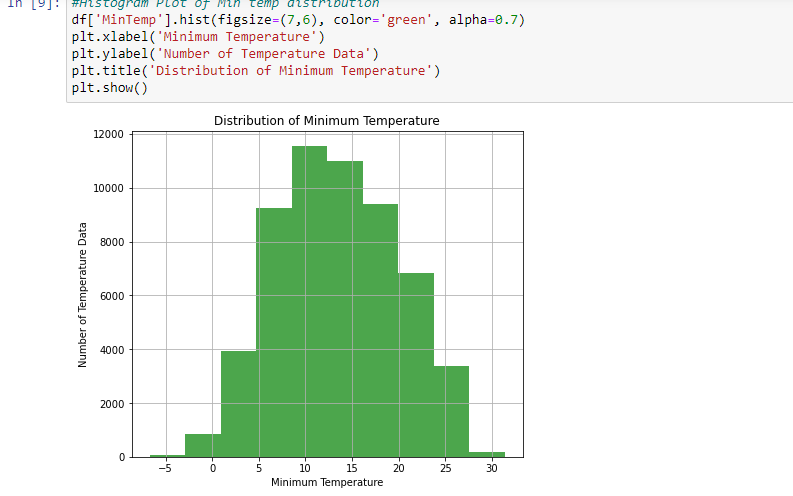


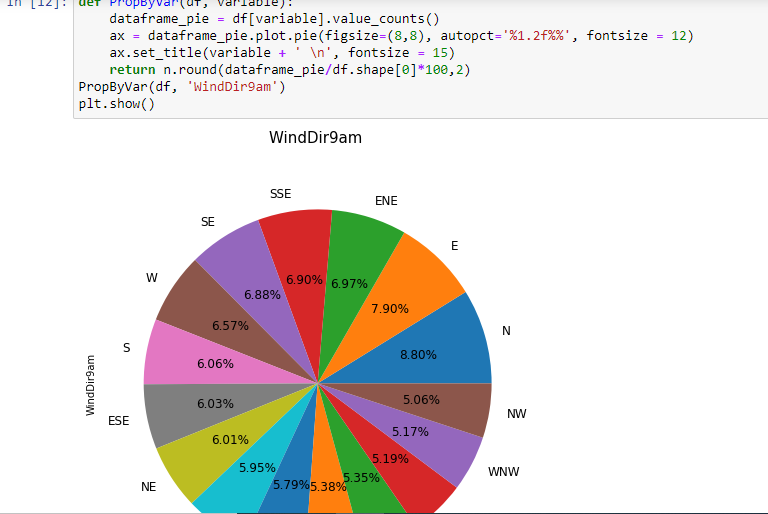


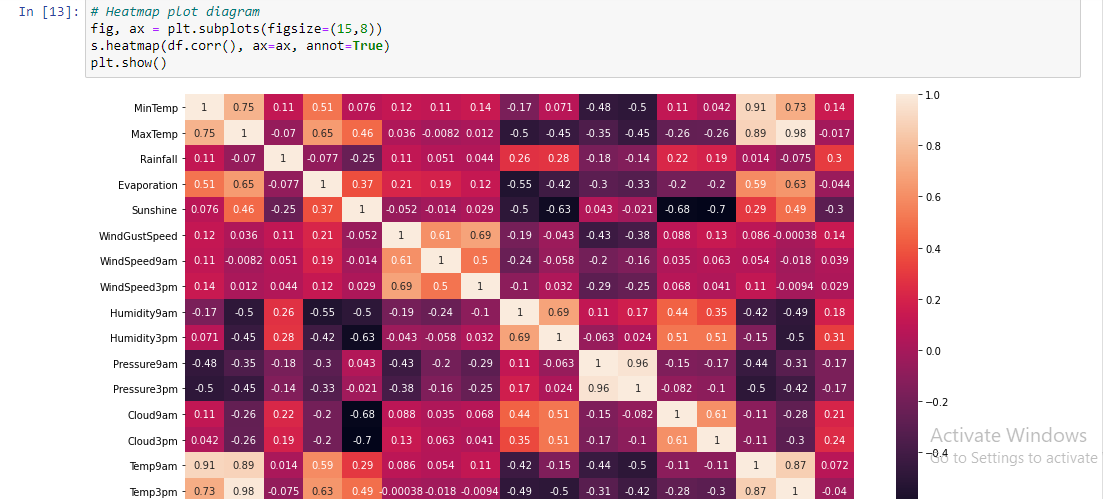


# Module 2: Exploration data analysis of visualization and training a model by given attributes

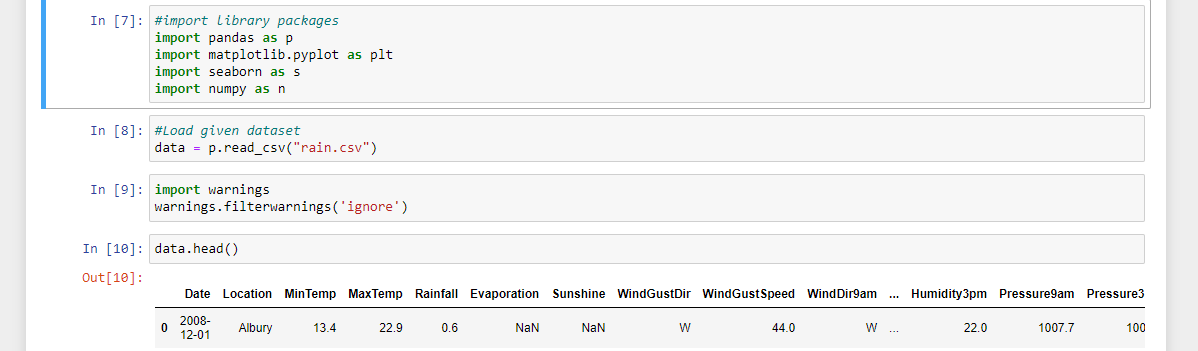


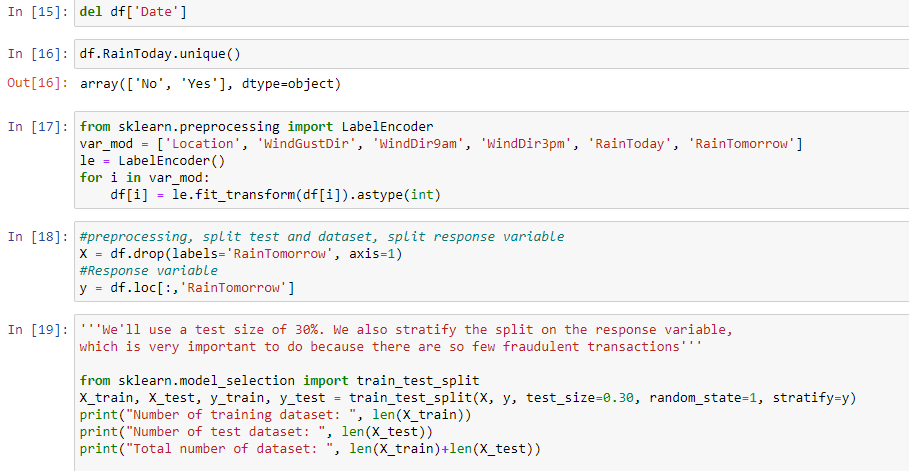




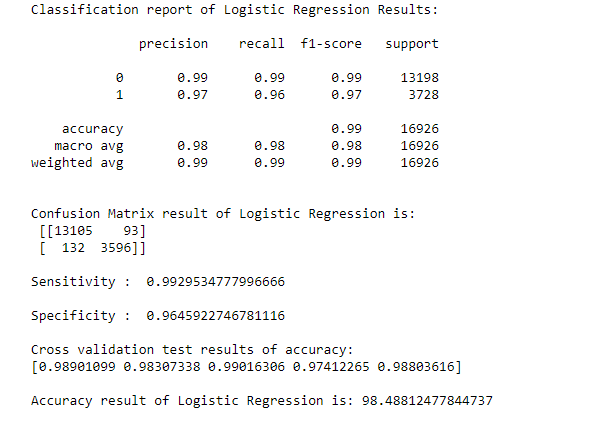


# Module 3 : Performance measurements of Logistic regression







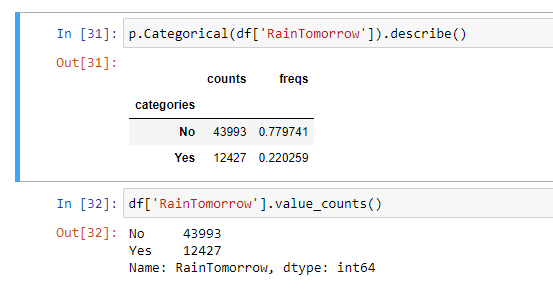


**CHAPTER 6**

## TESTING AND MAINTAINANCE

### **6.1. UNIT TESTING**

Unit testing is conducted to verify the functional performance of each modular component of the software. Unit testing focuses on the smallest unit of the software design (i.e.), the module. The white-box testing techniques were heavily employed for unit testing.



### **6.2. SYSTEM TESTING**

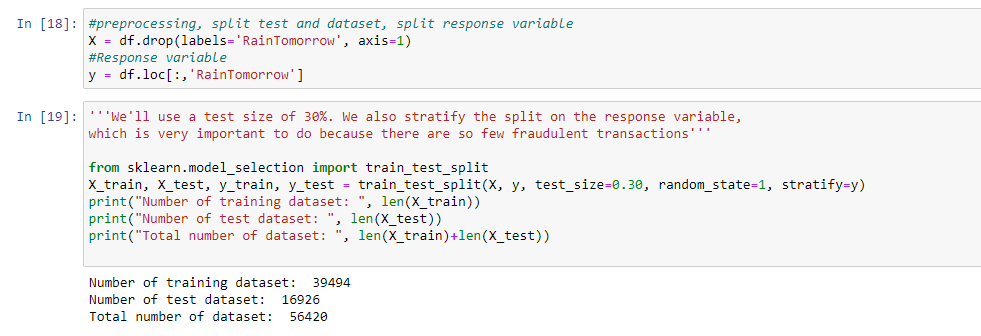
Testing is performed to identify errors. It is used for quality assurance. Testing is an integral part of the entire development and maintenance process. The Goal of the testing during phase is to verify that the specification has been accurately and completely incorporated into the design, as well as to ensure the correctness of the design itself. For example, the design must not have any logic faults in the design is detected before coding commences, otherwise the cost of fixing the faults will be considerably higher as reflected. Detection of design faults can be achieved by means of inspection as well as walk through. Testing is one of the important steps in the software development phase.

Testing checks for the errors, as a whole of the project testing involves the following test cases:

* Static analysis is used to investigate the structural properties of the Source code.
* Dynamic testing is used to investigate the behavior of the source code by executing the program on the test.

**6.3. INTEGRATION TESTING**

It also known as integration and testing (I&T) -- is a type of software testing in which the different units, modules or components of a software application are tested as a combined entity. After completing the unit testing for every module integration take place to combine the modules and check the module performance.

****

**CHAPTER 7**

## CONCLUSION AND FUTURE ENHANCEMENTS

**7.1 CONCLUSION:**

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the Rainfall using machine learning algorithm with different results. This brings some of the following insights about rainfall prediction.

* These reports are to the investigation of applicability of machine learning techniques for rainfall forecasting in operational conditions.
* Finally, it highlights some observations on future research issues, challenges, and needs.

### **7.2 FUTURE ENHANCEMENTS**

* Disaster management wants to automate the detecting the rain happened or not from eligibility process (real time).
* To automate this process by show the prediction result in web application or desktop application.
* To optimize the work to implement in Artificial Intelligence environment

# 

# REFERENCES

## REFERENCES

1. S. Kogan, T. J. Moskowitz, and M. Niessner, “Rainfall: Evidence from Financial Markets,” 2019, [https://ssrn.com/abstract=3237763.](https://ssrn.com/abstract=3237763)
2. H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic rainfalldetection model based on bidirectional encoder representations from transformers (bert),” *Applied Sciences*, vol. 9, no. 19, 2019.
3. S. Akhtar, F. Hussain, F. R. Raja et al., “Improving mispronunciation detection of arabic words for non-native learners using deep convolutional neural network features,” *Electronics*, vol. 9, no. 6, 2020.
4. Survey on RainfallDetection using Machine learning Algorithms, ICACT – 2021.
5. RainfallClassification Using Random Forest and Decision Tree,December, 2020.

1. Priyanka S, "Detection of Fake Profiles on Twitter using Random Forest & Deep Convolutional Neural Network", 2019.

1. Sam F., "Decision Tree Classification with Differential Privacy: A Survey", 2019.

M. Sumithra and Dr. S. Malathi, “A Novel Distributed Matching Global and Local

Fuzzy Clustering (DMGLFC) FOR 3D Brain Image Segmentation for Tumor Detection”, IETE Journal of Research, doi.org/10.1080/03772063.2022.2027284, 2021

1. B.Buvanswari and T.Kalpalatha Reddy, “A Review of EEG Based Human Facial Expression Recognition Systems in Cognitive Sciences” International Conference on Enenrgy, Communication,Data analytics and SoftComputing(ICECDS),CFP17M55PRJ:978-1-5386-1886-8”,August 2017.

1. Chethana, C., Subbiah Swaminathan, S. Sharanyaa, E. Sathish, R. Prathipa, and Anuradha Thakare. "Application Of Reverse Engineering in the Process of Utilization of Human Brain in Artificial Intelligence." Journal of Optoelectronics Laser 41, no. 3 (2022): 89-93.

1. M. Sumithra and Dr. S. Malathi, ” Modified Global Flower Pollination Algorithmbased image fusion for medical diagnosis using computed tomography and magnetic resonance imaging”, International Journal of Imaging Systems and Technology, Vol. 31, Issue No.1, pp. 223-235, 2021

1. K. Sridharan , and Dr. M. Chitra "SBPE: A paradigm Approach for proficient

Information Retrieval , Jokull Journal" , Vol 63, No. 7;Jul 2013