

**PROFESSIONAL TRAINING REPORT**  
**at**  
**Sathyabama Institute of Science and Technology**  
**(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award  
of Bachelor of Technology Degree in Information Technology

By

**ASWIN KUMAR.K**

**38120015**



**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**SCHOOL OF COMPUTING**  
**SATHYABAMA INSTITUTE OF SCIENCE AND**  
**TECHNOLOGY**  
**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**  
**CHENNAI – 600119. TAMILNADU**



---

## SCHOOL OF COMPUTING

### BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Bonafide work of who carried out the project entitled "PREDICTION OF FLOOD BY RAINFALL USING MLP CLASSIFIER" under our supervision.

**Internal Guide**

**Dr.P.AJITHA M.E., Ph.D.,**

**Head of the Department**

**Dr. R. Subhashini M.E., Ph.D.,**

---

Submitted for Viva voce Examination held on \_\_\_\_\_

**Internal Examiner**

**External Examiner**

## **DECLARATION**

I hereby declare that the Project Report entitled “**PREDICTION OF FLOOD BY RAINFALL USING MLP CLASSIFIER**“ done by me under the guidance of Dr.P.Ajitha M.E.,Ph.D., at Sathyabama Institute of Science and Technology (Deemed to be University), Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119 is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

**DATE:**

**Place:** Chennai

**SIGNATURE OF THE CANDIDATE**

## **ACKNOWLEDGEMENT**

I am pleased to acknowledge our 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 our thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean, School of Computing** and **Dr. R. SUBHASHINI M.E., Ph.D., Head of the Department, Department of Information Technology** for providing us the necessary support and details at the right time during the progressive reviews.

I would like to express our sincere and deep sense of gratitude to our Project **Dr.P.AJITHA M.E.,Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express our thanks to all Teaching and Non-teaching staff members of the Department of **INFORMATION TECHNOLOGY** who were helpful in many ways for the completion of the project.

# TRAINING CERTIFICATE

## COURSE COMPLETION CERTIFICATE

This is to Certify that

Mr. / Ms. **ASWIN KUMAR**

has successfully completed the training on ZCF

**ZOHO CREATOR FOUNDATION PROGRAM**

conducted by **ITKESPL**, Authorized Training Partner for Zoho

Certification Number: ZCF-02017

ISSUED BY:  
ITK EDUCATION SERVICES PVT LTD

JOSEPH M  
DIRECTOR OPERATIONS



## INTERNSHIP CERTIFICATE:



**RETECH**  
SOLUTIONS PVT. LTD.  
An ISO 9001:2015 certified company

INT: RE20S035

**TO WHOMSOEVER IT MAY CONCERN**

This is to certify that **ASWIN KUMAR K** student of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY, BTECH/IT** has undergone the Internship Training through online in our concern entitled **PYTHON** from **19<sup>th</sup> Aug 2020** to **21<sup>st</sup> Aug 2020** in relevant departments related to their academic studies.

During the above period, the performance was good & we wish great success in all your future endeavours.

*J. meru*

*Technology drives the world. We are the people who are behind the technology.*

**Authorised Signature**



No-31, 1<sup>st</sup> Floor, Alagesan Street, Tambaram West, Chennai-45.

Office : 044-42077204 | +91-8122266653

[www.retechsolutions.in](http://www.retechsolutions.in)

## **ABSTRACT**

### **PREDICTION OF FLOOD**

Floods are among the most destructive natural disasters, which are highly complex to model. The research on the advancement of flood prediction models contributed to risk reduction, policy suggestion, minimization of the loss of human life, and reduction the property damage associated with floods. To mimic the complex mathematical expressions of physical processes of floods, during the past two decades, neural network methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions. To prevent this problem to predict flood happen or not by rainfall dataset with investigate the neural network based techniques. The analysis of dataset by Multi-Layer Perceptron Classifier (MLP) to capture several information's like, variable identification, missing value treatments, data validation and data cleaning/preparing will be done on the entire given dataset. To performance in prediction of flood happen or not by accuracy calculation with evaluation classification report, identify the confusion matrix and the result shows that the effectiveness of the python framework flask based application by given attributes.

| <b>CHAPTER NO</b> | <b>TITLE NAME</b>                        | <b>PAGE NO</b>         |
|-------------------|------------------------------------------|------------------------|
|                   | <b>ABSTRACT</b>                          | <b>5</b>               |
| <b>1</b>          | <b>INTRODUCTION</b>                      |                        |
|                   | 1.1 DOMAIN OVERVIEW                      | <b>9</b>               |
|                   | 1.2 PROJECT DESCRIPTION                  | <b>10</b>              |
| <b>2</b>          | <b>SYSTEM CONFIGURATION</b>              |                        |
|                   | 2.1 HARDWARE REQUIREMENT                 | <b>11</b>              |
|                   | 2.2 SOFTWARE REQUIREMENT                 | <b>11</b>              |
| <b>3</b>          | <b>SYSTEM ANALYSIS</b>                   |                        |
|                   | 3.1 EXISITING SYSTEM                     | <b>12</b>              |
|                   | 3.1.1 DRAWBACKS OF EXISTING<br>SYSTEM    | <b>12</b><br><b>13</b> |
|                   | 3.2 PROPOSED SYSTEM                      | <b>13</b>              |
|                   | 3.2.1 ADVANTAGES OF PROPOSED<br>SYSTEM   |                        |
| <b>4</b>          | <b>SYSTEM DESIGN</b>                     |                        |
|                   | 4.1 SYSTEM FLOW DIAGRAM                  | <b>16</b>              |
|                   | 4.2 DATABASE DESIGN                      | <b>18</b>              |
|                   | 4.3 INPUT DESIGN                         | <b>20</b>              |
|                   | 4.4 OUTPUT DESIGN                        | <b>22</b>              |
|                   | 4.5 CODING                               | <b>24</b>              |
| <b>5</b>          | <b>SYSTEM TESTING AND IMPLEMENTATION</b> |                        |

| <b>FIGURE NO</b> | <b>FIGURE NAME</b>                   | <b>PAGE NO</b> |
|------------------|--------------------------------------|----------------|
| 3.1              | DATA FLOW                            | 16             |
| 4.1              | OUTPUT                               | 34             |
|                  | 5.1 SYSTEM TESTING                   | <b>27</b>      |
|                  | 5.2 SYSTEM IMPLEMENTATION            | <b>29</b>      |
| <b>6</b>         | <b>SCOPE FOR FUTURE ENCHANCEMENT</b> | <b>31</b>      |
| <b>7</b>         | <b>CONCLUSION</b>                    | <b>32</b>      |

## **CHAPTER 1**

### **1.INTRODUCTION**

#### **1.1 DOMAIN OVERVIEW:**

##### **Data Science:**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

##### **Data Scientist:**

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data.

##### **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.

#### **1.2 PROJECT DESCRIPTION**

- Flood is one of the biggest natural disaster causing many lives as well as damages. Different types of floods like

river flood, urban flood, coastal flood, and flash flood have been observed over the years.

- Our approach for short-term flash flood prediction in urban areas is to establish a theoretical model incorporating the factors influencing flood and use the power of machine learning techniques to estimate flood ahead of time. The organization of this paper is as follows.

## **CHAPTER 2**

### **SYSTEM CONFIGURATION**

#### **2.1 HARDWARE REQUIREMENT**

This section gives the details and specification of the hardware on which the system is expected to work.

|                    |                                       |
|--------------------|---------------------------------------|
| Processor          | : Pentium IV/III                      |
| Speed              | : 2.00 GHZ                            |
| RAM                | : 2GB                                 |
| Hard Disk Capacity | : 80 GB                               |
| Monitor            | : 14” Colour Monitor                  |
| Keyboard           | : Microsoft Natural PS/2 IBM Keyboard |
| Mouse              | : Wireless Dell Mouse                 |

#### **2.2 SOFTWARE REQUIREMENT**

This section gives the details of the software that are used for the development.

|                  |                                 |
|------------------|---------------------------------|
| Language         | : PYTHON                        |
| Tool             | :Anaconda with Jupyter notebook |
| Operating system | : Windows                       |
| Documentation    | : Microsoft Word                |

## **CHAPTER 3**

### **SYSTEM ANALYSIS**

#### **3.1 EXISTING SYSTEM**

Rapid urbanization, climate change, and extreme rainfall have resulted in a growing number of cases of urban flash floods. It is important to predict the occurrence of a flood so that the aftermath of it can be minimized. As the name suggests, an urban flash flood occurs in an urban area in a very short span of time. To reduce the impact of these events, short-term forecasting or nowcasting is used for prediction of the very near future incident. In orthodox methods of flood forecasting, current weather conditions are examined using conventional methods such as the use of radar, satellite imaging and calculations involving complicated mathematical equations. However, recent developments in Information and Communication

##### **3.1.1 DRAWBACKS OF EXISTING SYSTEM**

The existing system has following disadvantage:

- They did not mentions any accuracy.

- It can't thereby better determine the regularity of rainfall data and achieve more accurate prediction results of flash flood.

### **3.2 PROPOSED SYSTEM**

Flood is one of the biggest natural disaster causing many lives as well as damages. Different types of floods like river flood, urban flood, coastal flood, and flash flood have been observed over the years. A flash flood is a direct response to a rainfall having very high intensity in small time. This kind of flood is seen typically in urban areas where the underlying ground cannot cope, or drain excess water away fast enough via the sewage system and draining canals in a short amount of time. In recent years, we have seen the impact of floods in cities such as mumbai, chennai, ahmedabad and kerala. Poor urban planning, inaccurate and delayed forecasting and inadequate

#### **3.2.1 ADVANTAGES OF PROPOSED SYSTEM**

.

The propose system has following advantage

- These reports are to the investigation of applicability of machine learning techniques for Flood prediction
- . •Finally, it highlights some observations on future research issues, challenges, and needs.
- ML is a field of artificial intelligence (AI) used to induce regularities and patterns, providing easier implementation with low computation cost, as well as fast training, validation, testing, and evaluation, with high performance compared to physical models, and relatively less complexity.

List Of Modules:

- Data Pre-processing
- Data Analysis of Visualization
- Implementing Algorithm with best accuracy

Project Requirements General:

Requirements are the basic constraints 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. Functional requirements
2. Non-Functional requirements
3. Environment requirements
  - A. Hardware requirements
  - B. software requirements

Functional requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

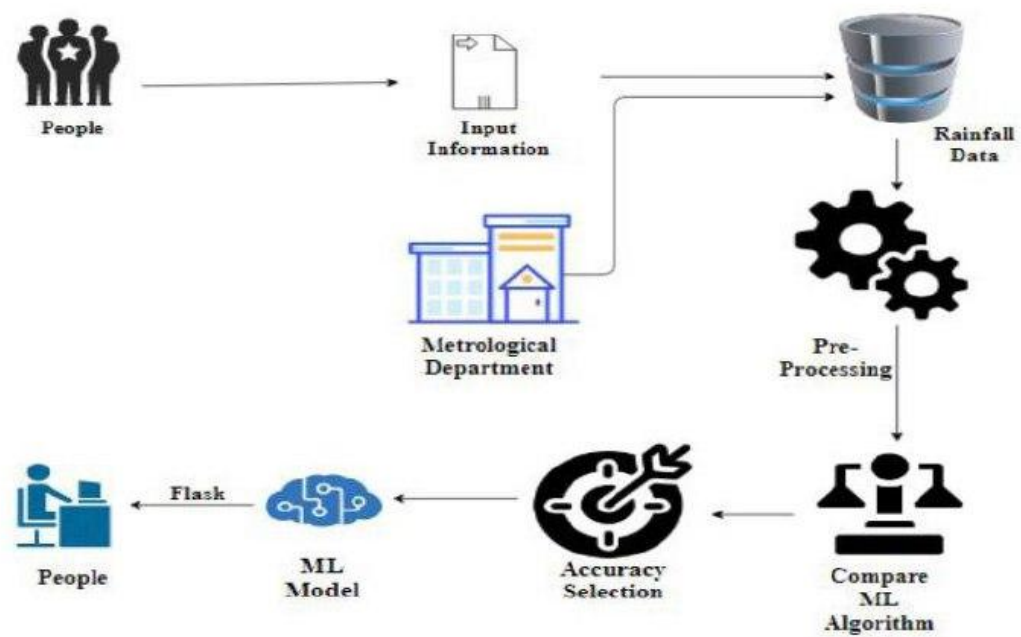
Non-Functional Requirements:

Process of functional steps:

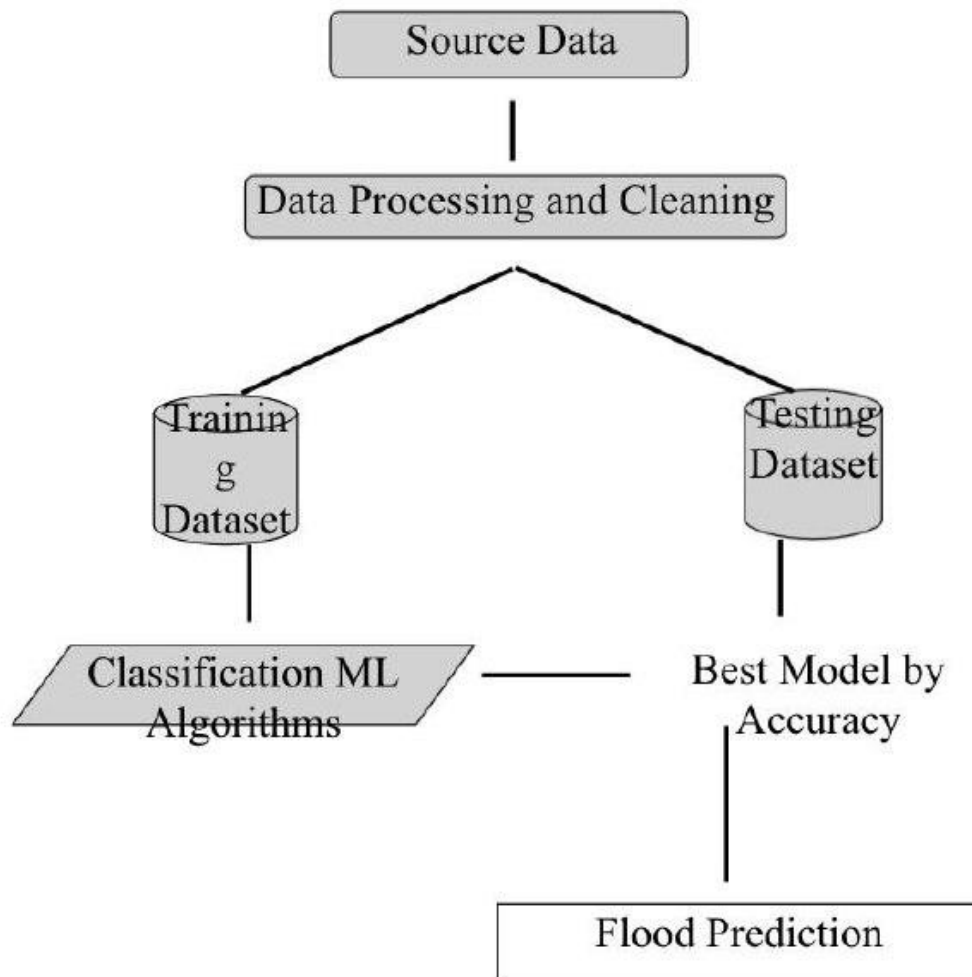
1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

### 4.3 INPUT DESIGN

#### System Architecture

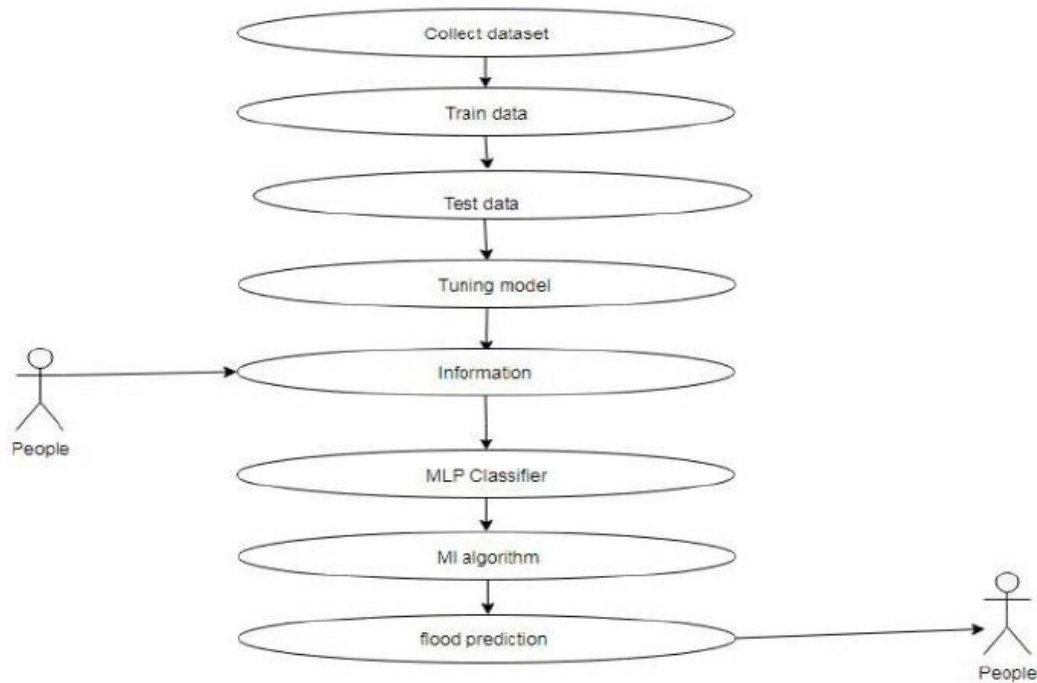


## Work flow diagram



Workflow Diagram

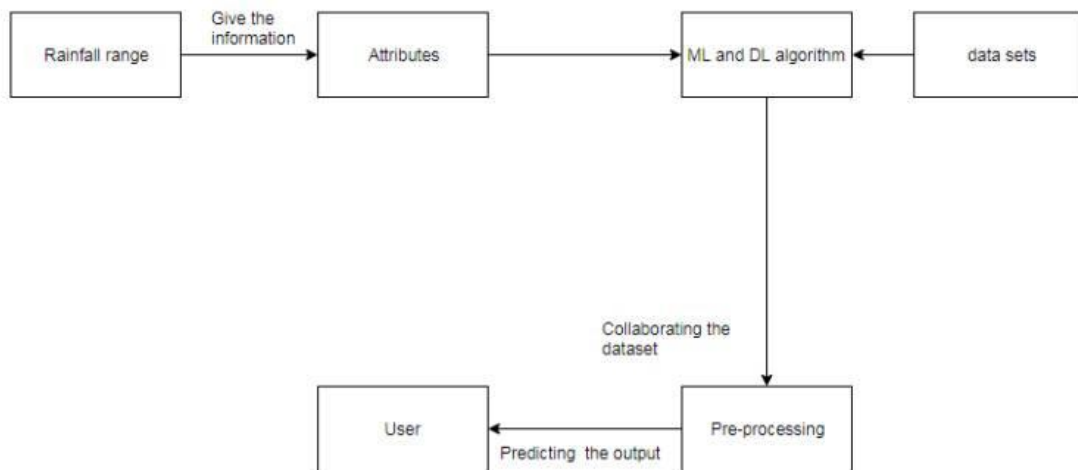
## Use Case Diagram



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 use cases are nothing but the system functionalities written in an organized manner. Now the second things which are relevant to the use cases are the actors (people).

### 4.4 OUTPUT DESIGN

## Collaboration Diagram:



### 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](#) 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.

## 4.5 CODING

Pre-processing

```
import pandas as pd
import numpy as np
```

```
data = pd.read_csv('demo.csv')
```

```
data.head()
```

```
data.shape
```

```
df = data.dropna()
df.head()
df.shape
df.columns
df.info()
```

```
df.describe()
df.isnull()
df.isnull().sum()
df.STATE_UT_NAME.unique()
df.columns
df.flood.unique()
df.JAN.unique().mean()
df.FEB.unique().mean()
df.ANNUAL.unique().min()
df.ANNUAL.unique().max()
df.ANNUAL.unique().mean()
df.Avg_june10days.unique()
df.corr()
```

## Visualization

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv('demo.csv')
import warnings
warnings.filterwarnings('ignore')
df.head()
df.shape
df.columns
plt.hist(df['STATE_UT_NAME'],color="red")
plt.xlabel("STATE_UT_NAME")
plt.show()
plt.hist(df['ANNUAL'],color="blue")
plt.xlabel("ANNUAL")
plt.show()
plt.figure(figsize=(10,6))
```

```

plt.plot(df['ANNUAL'])
plt.show()
fig, ax = plt.subplots(figsize=(18,13))
sns.heatmap(df.corr(), ax=ax, annot=True)
boxplot = df.boxplot(column=['ANNUAL'], by="flood",figsize=(8,6),
    fontsize=15)
plt.show()
def PropByVar(df, variable):
    dataframe_pie = df[variable].value_counts()
    ax = dataframe_pie.plot.pie(figsize=(8,12), autopct='%1.2f%%',
    fontsize = 12)
    ax.set_title(variable + ' \n', fontsize = 15)
    plt.show()
    return np.round(dataframe_pie/df.shape[0]*100,2)
PropByVar(df,"flood")
plt.show()
plt.figure(figsize=(12,8))
plt.scatter(df["ANNUAL"],df["flood"],color="black")
data1=df.iloc[:,9]
data2=df.iloc[:,9:-1]
a=pd.plotting.scatter_matrix(data1, alpha=0.05, figsize=(18,10),
    diagonal='hist')
axis = 'off'

plt.show()

```

## Algorithm Implementation

### Performance measurements of Logistic regression

#import library packages

import pandas as p

import matplotlib.pyplot as plt

import numpy as n

import seaborn as sns

```

data = p.read_csv("demo.csv")
import warnings
warnings.filterwarnings('ignore')
data.head(5)
data.tail()
data["flood"].value_counts()
data.isnull().sum()
data.duplicated().sum()
data.columns
data.shape
del data['JAN']
del data['Unnamed: 0']
del data['FEB']
del data['MAR']
del data['APR']
del data['MAY']
del data['JUN']
del data['JUL']
del data['AUG']
del data['SEP']
del data['OCT']
del data['NOV']
del data['DEC']
del data['Avg_june10days']
del data['maytojune']
data.head()
from sklearn.preprocessing import LabelEncoder

var_mod=["STATE_UT_NAME"]
le = LabelEncoder()
for i in var_mod:
    data[i] = le.fit_transform(data[i]).astype(int)
data.head()

```

```

#According to the cross-validated MCC scores, the random forest is
the best-performing model, so now let's evaluate its performance on
the test set.
from sklearn.metrics import confusion_matrix, classification_report,
matthews_corrcoef, cohen_kappa_score, accuracy_score,
average_precision_score, roc_auc_score
X = data.drop(labels='flood', axis=1)
#Response variable
y = data.loc[:, 'flood']
#We'll use a test size of 30%. We also stratify the split on the response
variable, which is very important to do because there are so few
fraudulent transactions.
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=1, stratify=y)
Logistic regression
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score

logR= LogisticRegression()

logR.fit(X_train,y_train)

predictLR = logR.predict(X_test)

print("")
print('Classification report of Logistic Regression Results:')
print("")
print(classification_report(y_test,predictLR))
x = (accuracy_score(y_test,predictLR)*100)

print('Accuracy result of Logisticregression is:', x)

print("")

```

```

cm2=confusion_matrix(y_test,predictLR)
print('Confusion Matrix result of Logistic Regression is:\n',cm2)
print("")
sensitivity2 = cm2[0,0]/(cm2[0,0]+cm2[0,1])
print('Sensitivity : ', sensitivity2 )
print("")
specificity2 = cm2[1,1]/(cm2[1,0]+cm2[1,1])
print('Specificity : ', specificity2)
print("")

accuracy = cross_val_score(logR, X, y, scoring='accuracy')
print('Cross validation test results of accuracy:')
print(accuracy)
#get the mean of each fold
print("")
print("Accuracy result of Logistic Regression is:",accuracy.mean() *
100)
LR=accuracy.mean() * 100

def graph():
    import matplotlib.pyplot as plt
    data=[LR]
    alg="Logistic Regression"
    plt.figure(figsize=(5,5))
    b=plt.bar(alg,data,color=("b"))
    plt.title("Accuracy comparison FLOOD",fontsize=15)
    plt.legend(b,data,fontsize=9)

```

```

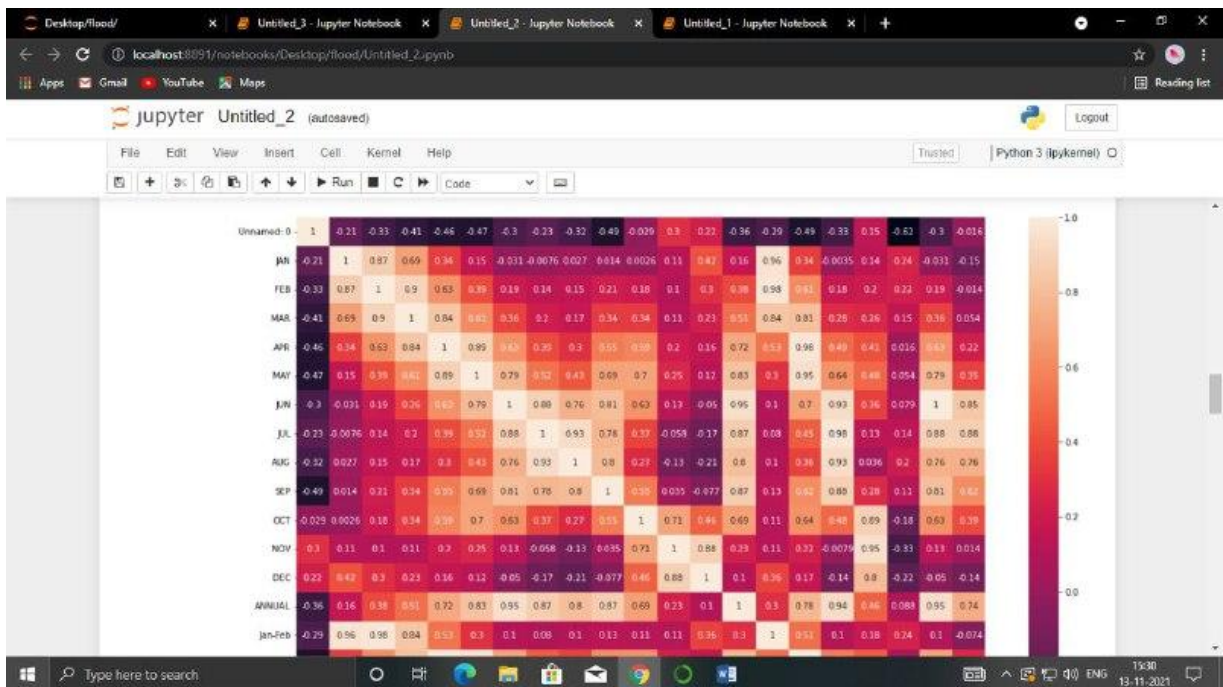
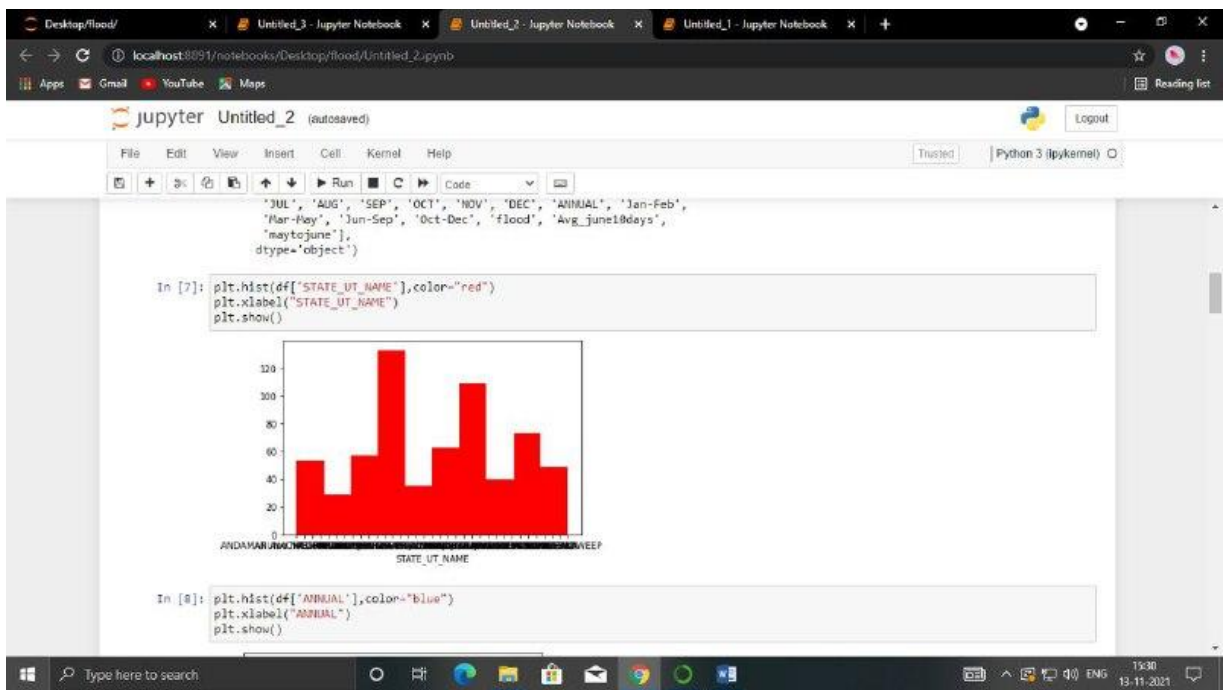
graph()
TN = cm2[1][0]
FN = cm2[0][0]
TP = cm2[1][1]
FP = cm2[0][1]
print("True Positive :",TP)
print("True Negative :",TN)
print("False Positive :",FP)
print("False Negative :",FN)
print("")
TPR = TP/(TP+FN)
TNR = TN/(TN+FP)
FPR = FP/(FP+TN)
FNR = FN/(TP+FN)
print("True Positive Rate :",TPR)
print("True Negative Rate :",TNR)
print("False Positive Rate :",FPR)
print("False Negative Rate :",FNR)
print("")
PPV = TP/(TP+FP)
NPV = TN/(TN+FN)
print("Positive Predictive Value :",PPV)
print("Negative predictive value :",NPV)

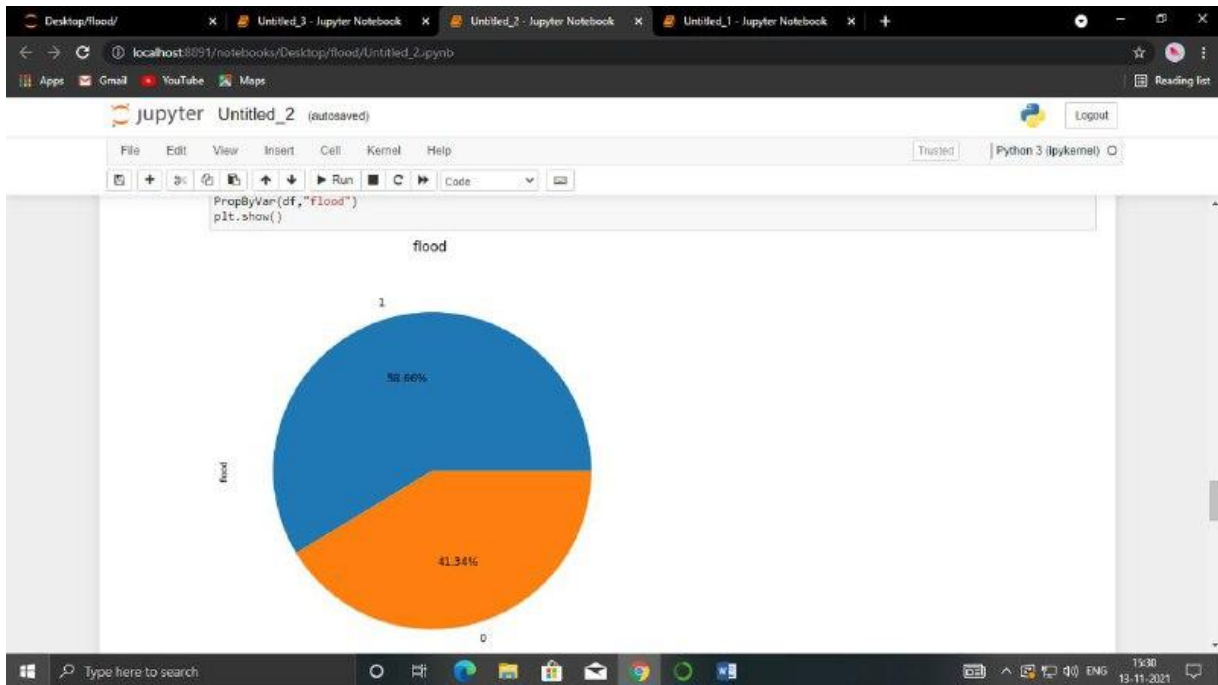
cm2=confusion_matrix(y_test, predictLR)
print(cm2)
sns.heatmap(cm2/n.sum(cm2), annot=True, cmap = 'Blues',
annot_kws={"size": 16}, fmt='.2%')
plt.show()

```

## CHAPTER 5

### 5.1 SYSTEM TESTING





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 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 is 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.

## **5.2 SYSTEM IMPLEMENTATION**

To practically implement this method, the cells are created using the features said earlier. Elevation of any location can be made available via altitude or elevation maps that are obtained by satellites. This elevation can be either in form of height above sea level or relative elevation against neighbouring cells. Soil index is measured for each cell and drainage outlet is checked along with their capacity. To measure the water sent to drainage and water coming from the adjacent cells, open flow methods are used such as Weir and Flume having the hydraulic structure as well as methods like Dye Testing and Manning's theoretical equation[10]. To check the validity

of the proposed model, it has to be tested on real data and the performance of the model must be examined. Since no actual data is present, a synthesis is done where we let the pattern of water level at time  $t+1$  to generate using the values of net residue water in the cell and rainfall forecast. The 6-dimensional input to our system is listed as rainfall, water from adjacent cells, water absorbed, a surplus of water when sent to drainage against drainage capacity, current water level in the cell, and rain forecast whereas the output is the stage of a flood. The input is considered over the sliding window of  $w = 2$  hour and the forecast is done ahead at time  $t + 1$  hours which we call as future lookahead time. The stages of a flood are decided by the water level in the cell. The water level is divided into 4 bands and these 4 bands depict the stages of a flood from low intensity to high. One-hot encoding is used to convert the numeric value to a class.

## **CODING**

Coding is the process of whereby the physical design specifications created by the analysis team turned into working computer code by the programming team.

## **TESTING**

Once the coding is being and proceed in parallel , as program module can be tested.

## **INSTALLATION**

Installation is the process doing which the current system is replaced by the new system. This includes conversion of existing data, software, and

documentation and work procedures to those consistent with the new system.

## **DOCUMENTATION**

It is result from the installation process, user guides provides the information of how the use the system and it's flow.

## **TRAINING AND SUPPORT**

Training plan is a strategy for training user so they quickly learn to the new system. The development of the training plan probably began earlier in the project.

# **CHAPTER 6**

## **SCOPE OF FUTURE ENHANCEMENT**

- Disaster management wants to automate the detecting the flood 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

## 7.CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the flash flood using machine learning algorithm with different results. So, the best result is MLP algorithm.



### 1.1 FLOOD REDICTION



**SAY NO TO FLOOD**

**BE PREPARED**



**BE PREPARED**