

# Flood Prediction Using Machine Learning Models

by

Miah Mohammad Asif Syeed

18101393

Maisha Farzana

18101665

Ishadie Namir

18101043

Ipshita Ishrar

18101573

Meherin Hossain Nushra

18101493

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
January 2022

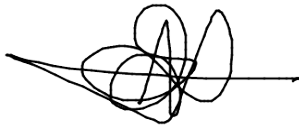
© 2022. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

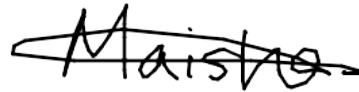
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

## Student's Full Name & Signature:



---

Miah Mohammad Asif Syeed  
18101393



---

Maisha Farzana  
18101665



---

Ishadie Namir  
18101043



---

Ipshita Ishrar  
18101573



---

Meherin Hossain Nushra  
18101493

# Approval

The thesis/project titled “Flood Prediction Using Machine Learning Models” submitted by

1. Miah Mohammad Asif Syeed (18101393)
2. Maisha Farzana (18101665)
3. Ishadie Namir (18101043)
4. Ipshita Ishrar (18101573)
5. Meherin Hossain Nushra (18101493)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 18, 2022.

## Examining Committee:

Supervisor:  
(Member)



---

Tanvir Rahman  
Lecturer  
Department of Computer Science and Engineering  
Brac University

Program Coordinator:  
(Member)

---

Md. Golam Rabiul Alam, PhD  
Associate Professor  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Sadia Hamid Kazi, PhD  
Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

# Ethics Statement

Hereby, we, the members, consciously assure that the following is fulfilled for the manuscript, “Flood Prediction Using Machine Learning Models”.

1. The contents of this paper are unique work of the writers’ and it has not been published before.
2. Only the authors’ own research and analysis is presented in the work with utmost accuracy.
3. Contributions by the co-authors and co-researchers are rightly acknowledged in the study.
4. All sources are appropriately mentioned (correct citation). Text that is literally copied must be identified as such by using quote marks and providing suitable reference.
5. The authors actively participated and put in effort leading to the article and any public responsibility related to it’s content will be accepted.

Violations of the Ethical Statement standards may have serious repercussions. We agree with the preceding declarations and certify that this submission adheres to BRAC University’s rules.

# Abstract

Floods are one of nature's most catastrophic calamities which cause irreversible and immense damage to human life, agriculture, infrastructure and socio-economic system. Floods are one of Bangladesh's most common natural catastrophes, causing modest to large-scale devastation every year. As a poor-economy developing country, taking structural steps to manage floods in the world's great rivers is a major problem. Several studies on flood catastrophe management and flood forecasting systems have been conducted. The accurate prediction of the onset and progression of floods in real time is challenging. To estimate water levels and velocities across a large area, it is necessary to combine data with computationally demanding flood propagation models. This paper aims to reduce the extreme risks of this natural disaster and also contributes to policy suggestions by providing a prediction for floods using different machine learning models. This prediction will be done by analyzing different parameters like temperature, area, water level, soil moisture, rainfall, etc which are some of the hydrological and climatic factors that influence flooding. This research will use Binary Logistic Regression, K-Nearest Neighbour (KNN), Support Vector Classifier (SVC), Decision tree Classifier and Stacked Generalization (Stacking) to provide an accurate prediction. With the outcome, a comparative analysis will be conducted to understand which model delivers a better accuracy.

**Keywords:** Binary Logistic Regression, Support Vector Classifier(SVC), K-Nearest Neighbor(KNN), Decision Tree Classifier(DTC), Flood Prediction, Rainfall.

## Acknowledgement

To begin, glory be to Allah, who ensured that our thesis was completed on schedule and without serious setbacks.

We would like to offer our heartfelt appreciation to Tanvir Rahman, our distinguished mentor and advisor, for his advice and assistance during our work.

Finally, we'd want to express our gratitude to our most deserving parents. We might not have been able to accomplish this work without their relentless support. We are approaching graduating as a result of their goodwill and efforts.

# Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
Nomenclature	x
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Cause and Effects of Floods . . . . .	1
1.3 Aims and Objectives . . . . .	2
<b>2 Related Work</b>	<b>5</b>
<b>3 Research Methodology</b>	<b>7</b>
3.1 Dataset Preprocessing . . . . .	8
3.1.1 Dataset Cleaning . . . . .	8
3.1.2 Feature Engineering . . . . .	8
3.1.3 Feature Encoding . . . . .	8
3.1.4 Feature Scaling . . . . .	8
3.2 Machine learning Models . . . . .	8
3.2.1 Binary Logistic Regression . . . . .	8
3.2.2 Support Vector Classifier . . . . .	10
3.2.3 K-Nearest Neighbor . . . . .	11
3.2.4 Decision Tree Classifier . . . . .	11
<b>4 Implementation</b>	<b>13</b>
4.1 Dataset . . . . .	13
4.1.1 Source . . . . .	13
4.1.2 Dataset Description . . . . .	13
4.2 Pre-processing Steps . . . . .	14



4.2.1	Data Cleaning . . . . .	14
4.2.2	Feature Engineering . . . . .	14
4.2.3	Feature Encoding . . . . .	15
4.2.4	Feature Scaling . . . . .	15
4.3	Programming Language and Computational Speed . . . . .	16
4.4	Application of the Machine Learning Models . . . . .	16
<b>5</b>	<b>Results and Discussion</b>	<b>17</b>
5.1	Classification Result . . . . .	17
5.1.1	Timeline: 1980 to 2020 . . . . .	17
5.1.2	Timeline: 2011 to 2020 . . . . .	20
<b>6</b>	<b>Conclusion</b>	<b>24</b>
	<b>Bibliography</b>	<b>26</b>

# List of Figures

1.1	Economic Damages due to floods worldwide from 1900 to 2016 . . . . .	3
3.1	Workflow . . . . .	7
3.2	Contrasting linear to logistic regression . . . . .	9
3.3	Separating the Data in a Feature Space . . . . .	10
4.1	Rainfall Dataset . . . . .	13
4.2	Flood Dataset . . . . .	14
4.3	Dataset after categorizing monthly rainfall data with flood data . . . . .	15
4.4	Dataset after feature encoding . . . . .	15
5.1	Monthly Mean Precipitation (Timeline: 1980-2020) . . . . .	17
5.2	Classification Report of Binary Logistic Regression (Timeline: 1980-2020) . . . . .	18
5.3	Classification Report of Support Vector Classifier (SVC) (Timeline: 1980-2020) . . . . .	18
5.4	Classification Report of K-Nearest Neighbors (KNN) (Timeline: 1980-2020) . . . . .	19
5.5	Classification Report of Decision Tree Classifier (DTC) (Timeline: 1980-2020) . . . . .	19
5.6	Confusion matrix of the used models (Timeline: 1980-2020) . . . . .	19
5.7	ROC Curves of used models (Timeline: 1980-2020) . . . . .	20
5.8	Monthly Mean Precipitation (Timeline: 2011-2020) . . . . .	20
5.9	Classification Report of Binary Logistic Regression (Timeline: 2011-2020) . . . . .	21
5.10	Classification Report of Support Vector Classifier (SVC) (Timeline: 2011-2020) . . . . .	21
5.11	Classification Report of K-Nearest Neighbors (KNN) (Timeline: 2011-2020) . . . . .	22
5.12	Classification Report of Decision Tree Classifier (DTC) (Timeline: 2011-2020) . . . . .	22
5.13	Confusion matrix of the used models (Timeline: 2011-2020) . . . . .	22
5.14	ROC Curves of used models (Timeline: 2011-2020) . . . . .	23

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$\alpha$  Alpha

$\beta$  Beta

$\epsilon$  Epsilon

$\mu$  Mu

$\sigma$  Sigma

*ANN* Artificial neural network

*AUC* Area under the ROC Curve

*BPN* Back Propagation Neural Network

*DNN* Deep Neural Network

*EKF* Extended Kalman Filter

*GDP* Gross domestic product

*KNN* K-Nearest Neighbor

*ML* Machine Learning

*NARX* Nonlinear Autoregressive Network with Exogenous Inputs

*RMSE* Root Mean Square Error

*ROC* Receiver Operating Characteristic

*SVC* Support Vector Classifier

*SVM* Support Vector Machine

*USD* United States dollar

# Chapter 1

## Introduction

### 1.1 Motivation

Floods are among the world's most prevalent and devastating natural disasters. When water is inundated over normal dry land, it causes a flood. It is caused by an overflow of water from a lake, river, or ocean that submerges neighboring land. Because of seasonal fluctuations in input water volume and evaporation rate, water levels in lakes and rivers fluctuate dramatically. Floods can occur in any place where rain falls, although rain is not the only cause. Floods usually build over a period of hours or days, providing locals plenty of time for preparing or evacuating the affected area. Others appear out of nowhere and without warning. Fast-moving rivers, sometimes referred to as “Flash Flood”, can wipe out everything in their path, including people and property.

The inability of watercourses to drain water away under extremely high rainfall is the most prevalent cause of floods. Floods, on the other hand, are not necessarily caused by excessive rainfall. They can result from other natural or man-made phenomena. A storm surge linked with a tropical cyclone, a tsunami, or a high tide, for example, can create inundation in coastal areas, especially when rivers are flowing at higher level than typical river levels. Flooding, resulting from extreme hydro and meteorological events taking place in unexpected magnitudes and frequencies causes loss of lives, livelihood and infrastructure. Most part of Bangladesh is low lying and 80% of the landmass is floodplain, leaving the country vulnerable to the threat of flood [1]. Floods have huge devastation potential. Many constructions are unable to withstand the power of water when a river overflows its banks or the sea moves inland. Bridges, buildings, trees, and automobiles may all be lifted and hauled away. In July 2007, severe flooding in Bangladesh destroyed more than a million homes [2]. Infact, when flood water recedes, they can inflict much more harm than just structural damages. Hazardous materials such as sharp debris, pesticide, gasoline and untreated sewage can contaminate the water which in turn would contribute to the outbreak of many deadly water-borne diseases like cholera, typhoid, malaria etc.

### 1.2 Cause and Effects of Floods

Climate change is expected to affect floods in various ways. Low-lying areas and areas with poor drainage systems are most prone to floods. A survey shows that

around 4.84 million people in India, 3.84 million in Bangladesh, 3.28 million people in China are exposed to flood every year [3]. Also, there are other countries where urban areas are prone to flood as well. Areas that have elevations of less than 10m above sea level are also at risk of flooding e.g. Netherlands, Monaco, Bahrain, etc. In the years 1997–2008, floods in Australia caused the lives of at least 73 people [4]. Floods in the United States kill roughly 100 people each year and cost \$7.5 billion in damage. Furthermore, the Yellow River flood of China in 1931 was one of the worst natural disasters ever recorded, killing about a million people and displacing millions more [2]. The World Resource Institute finds that by the end of 2030, floods will affect over 147 million people around the world, with damages to urban properties ranging from \$174 billion to \$712 billion per year [5].

Floods have been causing enormous damage to the economy and social life all around the world. Last year, its effect worsened due to the trigger of the coronavirus crisis which has swept across the world, wreaking havoc on the economy and health. In Vietnam, there have been 90 deaths and at least 34 persons have gone missing. In Laos, 100 villages were inundated, and at least 10,000 hectares of farmed land were swamped. In Cambodia, there were 25 people killed and 40 thousand people displaced [6]. Since late June last year, flooding in Bangladesh has affected nearly 1 million people in 13 districts, according to disaster management authorities. Severe rains in the country and in river catchments in neighboring India caused flooding in numerous northern districts of the country. More than 3.3 million people have been evacuated from flood-affected areas, while another 7,31,958 individuals have been left without access to safe drinking water. Since June 2020, 93 individuals have died, the majority of them through drowning, and 41 of those deaths were children [7].

Because the effects of natural disasters such as floods are largely consistent year after year, one would expect the recovery procedures to be well-known as well. However, the same challenges resurface with each severe flood. Floods have a multi-faceted impact that includes both economic and social dimensions. Forecasting river water levels after heavy rain is crucial for social safety, environmental challenges and water management. Mathematical models based on physical considerations or statistical analysis have been developed for these aims. In both cases, the forecasts they provide are not entirely accurate and they are time-consuming. People have attempted to prevent and manage floods for thousands of years. But floods being a natural phenomena, it is hard to predict it only using statistical methods. And thus, using machine learning tools will be a better approach to forecast floods and manage its risks as much as possible.

### **1.3 Aims and Objectives**

Flood is one of the highly devastating calamities which causes serious social and economic losses and also causes destruction in different parts of the world every year. Flood damage accounts for over a third of the economic damages caused by natural disasters around the world, and floods, together with strong wind, are the most common natural catastrophes. The Organization for Economic Cooperation and Development estimates that floods globally inflict \$40 billion in damage each

year. Death tolls have increased to 100 people a year in recent times [8].

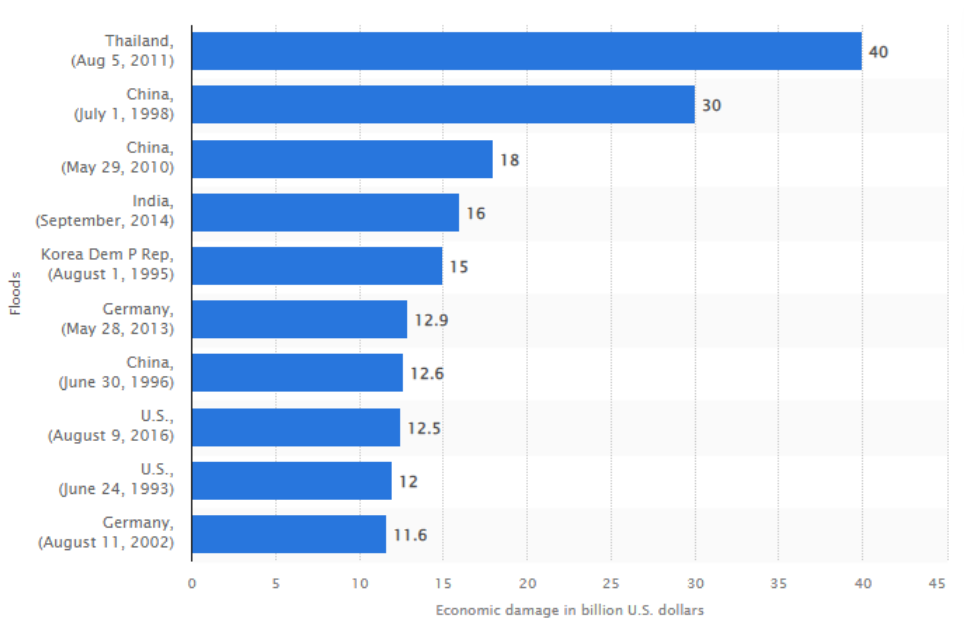


Figure 1.1: Economic Damages due to floods worldwide from 1900 to 2016

The southeast asian region is particularly more prone to flooding due to its geographical location threatening half of India’s population on a regular basis. The Yangtze River basin in China was flooded, resulting in an estimated 3,500 deaths, 5 million dwellings destroyed, and 200 million people displaced. According to statistics released in July 2014 by the Center for Research on the Epidemiology of Disasters, severe floods all around the world have caused significant economic damage. Flooding in Thailand on 5 August 2011 was estimated to have cost almost \$40 billion, making it the most economically devastating event to occur between 1900 and 2013. [9]. Bangladesh faced its one of the most dreadful floods in 1998 which submerged two third of the country for more than two months wrecking many of the country’s recent social and economic advances [10].

In terms of climate change, Bangladesh is one of the most susceptible countries in South Asia, particularly when it comes to flood. A study [11] claims that Most of Bangladesh’s land is low-lying and over 80% of the nation is located in flood plains. According to [11], unstable and migratory rivers, tidal effects, sea-level rise, deforestation, damming of rivers are mainly responsible for flood occurrence in Bangladesh. Also, natural flow of water is harmed because of unplanned development with poor engineering workmanship in the flood prone areas. The research work [11] mentions that during the time period (1971-2014) flood has occurred in Bangladesh 78 times in total and due to those floods 41,783 people have lost their lives. Also, in Bangladesh, almost 50% of the population is employed in the agricultural department and so contributes to 80% of the total export. Thus, it plays an important role to increase the economic growth and due to flooding many agriculture farms face huge economic losses which affects the GDP negatively. Moreover,

Bangladesh faces some secondary effects after floods, e.g. pollution of drinking water, destruction of structures such as bridges and roads, waterborne diseases etc.

Bangladesh is a very small nation with an increasing population, which suggests a higher population density. This increased population density along with river bank erosion, contributes to the decrease in land area. Flood dangers are on the rise due to a variety of factors, including increased urbanization in flood-prone areas, a poor integration between land management and flood risk management and aging stormwater infrastructure that cannot handle increased runoff loads. So it is important that a devastating natural disaster like flood is predicted beforehand and necessary evacuation steps are taken based on that to reduce the number of casualties and destruction. To forecast floods, this paper proposes multiple logistic regression as one of the machine learning models with higher accuracy. It will try to find out if Binary Logistic Regression can give a better accuracy to predict floods. It will also provide a comparative analysis between KNN, SVC and Decision Tree and find out which one gives a better performance.

The primary goal of this study is to improve the precision with which floods can be forecast. Binary classification may be used to classify the flood dataset, which consists of labeled training data. It may be identified which model parameters should be used to predict the unknown labels on other data by fitting the training dataset to the model. So far, greater accuracy has been obtained by employing classic ML methods such as SVC, KNN, and Logistic Regression. The primary goal of this research will be predicting the floods with a higher accuracy from the datasets using Binary Logistic Regression, the model which is going to be proposed will be using rainfall and flood data of a particular area which have been collected with the help of monitoring systems.

The next objective is to categorize the datasets depending on various parameters. With a filtered data set explored, an appropriate column should be decided to use as a target column for data modeling. Before putting the data into an algorithm to train, missing values and categorical features have to be handled because most of the models assume that the data is numerical and contains no missing values.

The sole objective of this research is to get maximum accuracy using Binary Logistic Regression. After training using the training data, the points in the validation set are used to compute the accuracy of the classifier.

# Chapter 2

## Related Work

The research paper [12] proposes a flood prediction model using the NARX structure of the Artificial Neural Network model. This paper used several upstream river water levels as the input variable as it plays a crucial role for floodings. With this sample, the model was first trained and the rest of the samples were then validated through this trained model. It has a 73.54% success rate in predicting the flood 5 hours in advance. The research paper [13] also uses the NARX neural network to model and predict flood water level. Four variables were employed in this model: river water levels upstream, water levels at the flood site, and differences in water levels. The data were then classified into the training and testing data ready to be fed to the NARX model. This model, according to the authors, has an accuracy rate of 87% when used to forecast flood water levels.

According to a research work [14] that was carried out in India, in recent years, a number of machine learning algorithms have been used to anticipate floods by taking into account variables such as rainfall and humidity as well as temperature and water velocity. Specifically, Deep Neural Network (DNN) and other machine learning algorithms (Support Vector Machine, K-Nearest Neighbor, and Naive Bayes) are compared for accuracy and error in this paper's flood prediction. This paper claims that it took temperature and rainfall as parameters and a confusion matrix was developed to figure out the best way to detect floods. Among the ML methods the accuracy rate of SVM is 85.57%, Naïve Bayes has 87.01% accuracy, KNN has 85.73% of accuracy and on the other hand DNN has 91.18% of accuracy.

Another study [15] attempted to forecast flood levels using an Artificial Neural Network technology known as Back Propagation Neural Network (BPN). Water level data from different stations of Johor, Malaysia are fed into the BPN model. The result from the model was not quite satisfactory and thus to improve the result Extended Kalman Filter (EKF) was added. The paper claims that by adding EKF with the result of the BPN model gave better accuracy. The paper [16] also uses the EKF algorithm to predict the flood water level. It describes that the EKF method is divided into two stages: prediction and update. Previous data estimations are used in the prediction stage, and the forecasted values are rectified using the feedback correction step in the update stage. The data used here is the water level which is to be used for forecasting the flood level. According to this paper, this algorithm has a calculated Root Mean Square Error (RMSE) of 0.9236 m.



The paper [17] gives a comparative analysis of the forecasting accuracy of SVM and ANN using the datasets of flood occurrences in Bangladesh. According to this research, SVM has a prediction accuracy as good as ANN and sometimes for higher lead, it gives better performance than ANN. The water level of five stations in Bangladesh were used as the data to train and verify the SVM. ANN was found to be more time consuming which requires more trial and error than the SVM architecture. Limited availability of data is also a factor that hinders the accuracy of prediction done using the ANN model. SVM provides better results in a shorter time window with limited dataset, which is the first and foremost concern for predicting floods. The paper also discusses the computational advantages offered by SVM, as it can represent input variables in dual form. This could prove to be advantageous in real time forecasting problems which requires huge data handling.

# Chapter 3

## Research Methodology

The purpose of this proposed paper is to answer if a higher accuracy rate using Binary Logistic Regression and lessening the error can be achieved. To determine and evaluate the accuracy, several data pre-processing techniques and machine learning models are applied.

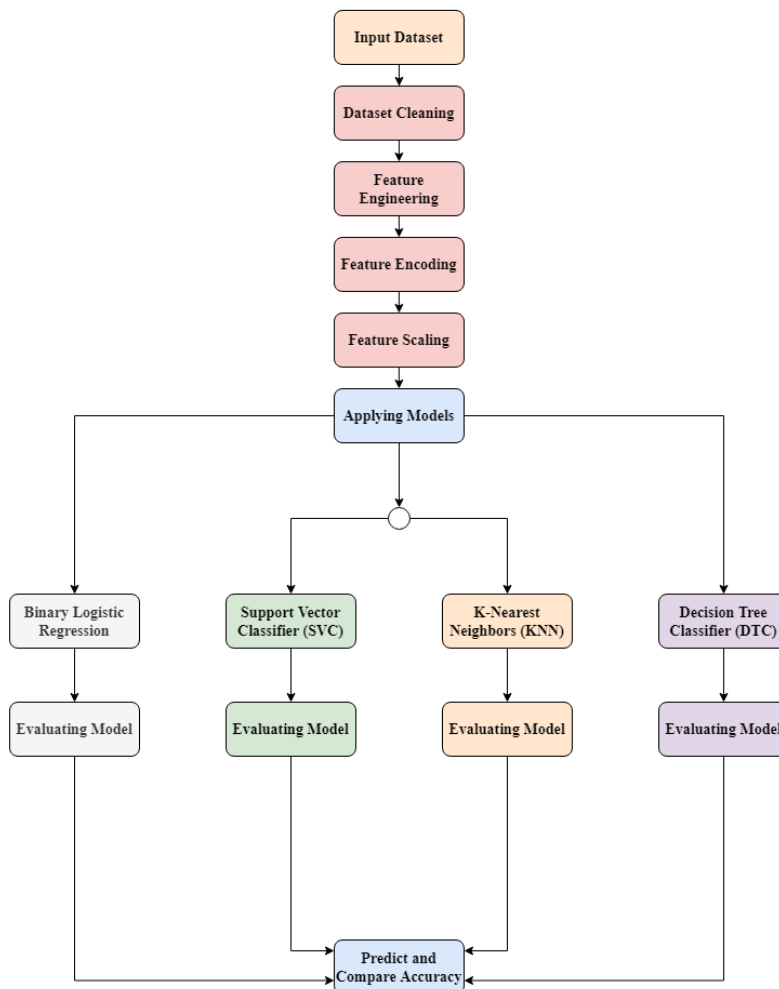


Figure 3.1: Workflow

## **3.1 Dataset Preprocessing**

### **3.1.1 Dataset Cleaning**

The discipline of cleaning up data is an essential one. It's a technique for deleting incomplete or irrelevant data from a dataset or table by finding and correcting any inaccuracies.

### **3.1.2 Feature Engineering**

Feature engineering is a machine learning preprocessing procedure that transforms raw data into features that may be used to build a predictive model using a machine learning model. It primarily seeks to enhance the models' performance. It aids in better representing an underlying problem to predictive models, which improves the model's accuracy for unobserved data. While the feature engineering process identifies the most effective predictor variables for the model, the predictive model contains predictor variables and an outcome variable [18].

### **3.1.3 Feature Encoding**

In machine learning algorithms, only numerical values can be used. Therefore, the categorical values of the relevant attributes must be converted to numerical values and this process is referred to as feature encoding.

### **3.1.4 Feature Scaling**

Analysis of the sample data and the selection of features that will make a machine learning model correct is facilitated by a process called feature scaling. Machine learning models like Linear and Logistic Regression, KNN, Support Vector Classifier etc require the data to be scaled. Scaling has no effect on algorithms based on trees, such as the Decision Tree Classifier, because the algorithm's split on a feature is unaffected by its other features.

## **3.2 Machine learning Models**

Binary Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Decision Tree Classifier (DTC) have all been used to predict values from the training set.

### **3.2.1 Binary Logistic Regression**

The Binary Logistic Regression, like all other regressions analyses, is a predictive analysis that is used to describe data and explain the relationship between one dependent binary variable and one or more independent variables. The dependent variable has two categories, generally which are 1 for the occurrence of an event

and 0 for its absence. A Logistic Regression can be interpreted as a specific case of generalized linear models with a dichotomous dependent variable [19]. A classical linear model can be denoted in the following manner:

$$Y = \alpha + \beta X + \varepsilon$$

where  $Y$  is the dependent variable,  $\alpha$  is the  $Y$  intercept when  $X$  is equals to zero,  $X$  is the independent variable,  $\beta$  is the regression coefficient representing the variation in  $Y$  due to the change in values of  $X$  and  $\varepsilon$  is the error of the model. To categorize or limit the range of values for the dependent variable, the logistic regression model fits best. A graphical comparison between linear regression and logistic regression is shown below:

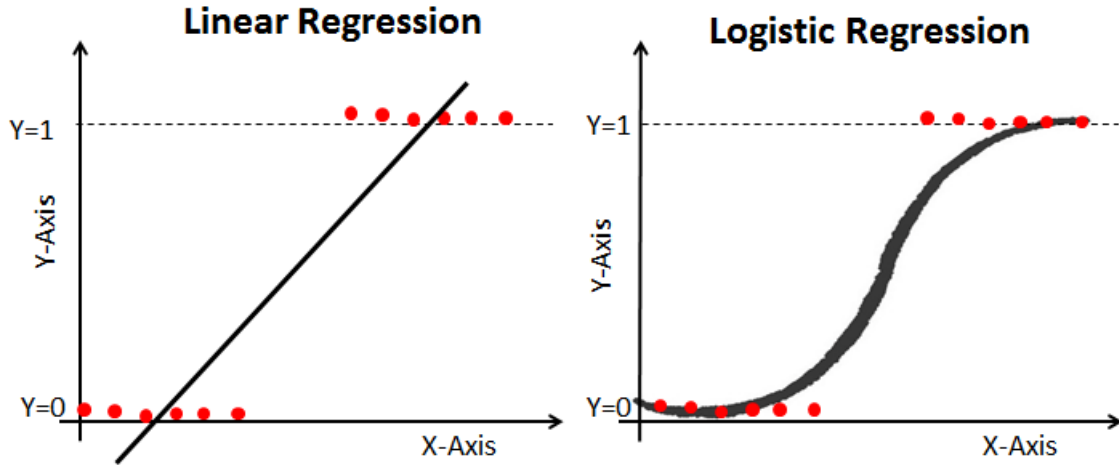


Figure 3.2: Contrasting linear to logistic regression

The derivation of the base equation for the logistic regression model is as follows: Replacing  $Y$  with a probability  $P$  that takes the range of probability to be within 0 and 1, the odds of  $P$  are taken,

$$\frac{P}{1-P} = \alpha + \beta X$$

In this equation, the range gets restricted which decreases the number of datapoints which eventually decreases the correlation. To avoid this, the log of the odds need to be taken and exponent has to be added to both sides and the solution for  $P$  is:

$$P = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

This is the sigmoid function for the logistic regression model used to predict any dichotomous dependent variable [20]. According to the dataset, the independent variable is the amount of annual rainfall and the dependent variable remains whether there will be a flood based on the rainfall or not.

### 3.2.2 Support Vector Classifier

The Support Vector Classifier (SVC) is a machine learning algorithm that uses both regression and classification. Structural Risk Minimization Principle is roughly implemented in this way. The SVC, in contrast to other models, aims to fit the best line within a threshold value rather than to reduce error between real and predicted values. According to this research article [14], a training SVC algorithm creates a model which designates the data to one class or another developing a binary and non probabilistic linear classifier. These classes are clearly separated from each other with the help of a gap or a spatial line.

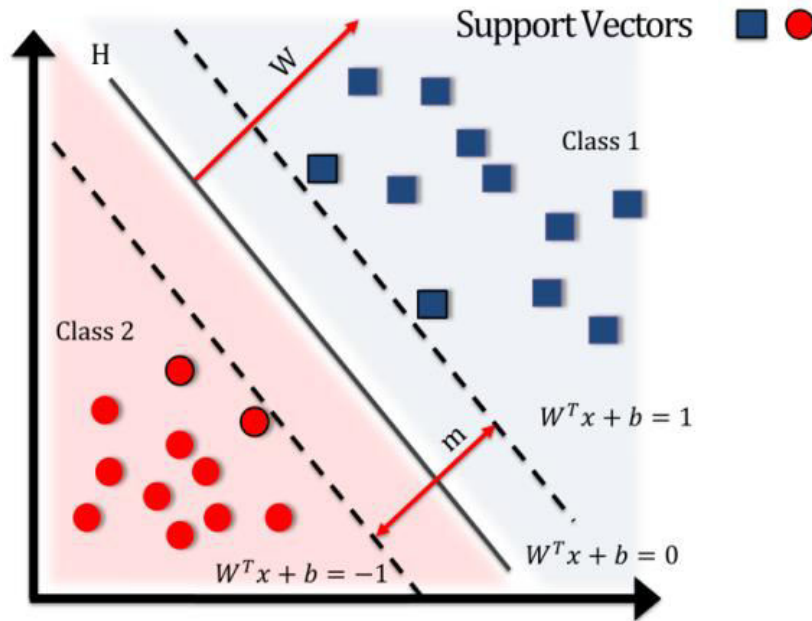


Figure 3.3: Separating the Data in a Feature Space

SVC predicts the newer data and the class they belong to considering the distance of these classes from the line. The basic idea is to perform linear regression to find a decision function for a given sample  $x$ ,

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b$$

The term  $\alpha_i$  are called the dual coefficients which are upper-bound by C, and b is an independent term that has to be estimated.  $K(x_i, x)$  is the kernel where, x is the input vector [21].

### 3.2.3 K-Nearest Neighbor

The KNN algorithm is a sort of supervised machine learning method that is being used to solve both classification and regression predicting problems. The KNN approach uses feature similarity to forecast the values of new data points, which means that the new data point will be assigned a value based on how closely it resembles the points in the training set. According to this research paper [14], the K nearest neighbor (KNN) is a non-parametric algorithm that can be used for regression predictive problems. The KNN method assumes the resemblance between the new data and the existing data and places the new data in the category that is most comparable to the existing categories. The KNN algorithm calculates the distance between a new data point and all previous data points in the training set. There are a variety of distance functions for calculating the distance but Euclidean is the most widely utilized method.

Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$$

### 3.2.4 Decision Tree Classifier

Decision Trees Classifier (DTC) are mainly used for regression and classification which is a non-parametric supervised learning method. By using decision trees, models can be created which will be able to predict target variables through learning simple decision rules deduced from the features of the dataset [22]. Decision trees help to explain the decisions of predictive models graphically and in decision trees the internal nodes denote the test on the features, branches denote the outcome and the leaf nodes denote the final decision that is derived from computing the features [23]. One of the most important factors of decision trees is to create a sequence of splits and for splitting it separates the data into two groups that are the purest. For calculating the purity of groups, decision trees calculate entropies of those groups. Entropy of a decision tree with C classes:

$$Entropy = \sum_i^c -p_i \log_2 p_i$$

Here,  $p_i$  means the probability of randomly picking an element of class  $i$ . The entropy values are ranged between 0 to 1, where 1 means maximum impure groups and 0 means full pure groups. Another statistical property of decision trees is information gain which represents decrease in entropy. It calculates the difference between the dataset's entropy before and after splitting depending on specified feature values.

$$InformationGain = Entropy(before) - \sum_{j=1}^k Entropy(j, after)$$

Here, *before* refers to the dataset prior to the split,  $k$  refers to the number of subsets formed by the split, and  $(j, after)$  refers to subset  $j$  following the split [24].

# Chapter 4

## Implementation

### 4.1 Dataset

#### 4.1.1 Source

The data was collected from Bangladesh Meteorological Department, Dhaka, Bangladesh, who are responsible for monitoring and issuing forecasts of all natural disasters to keep casualties to a minimum. They use rainfall data, satellite images and various other parameters to issue accurate weather forecasts.

#### 4.1.2 Dataset Description

The dataset consists of the daily rainfall index and flood index of 34 stations in Bangladesh from 1980 to 2020.

Station	Year	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31		
0	Dhaka	1980	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Dhaka	1980	2	0	0	0	0	14	1	0	0	0	0	0	0	0	0	0	2	15	0	0	0	0	0	0	0	0	0	0	0	0	0	NaN	NaN
2	Dhaka	1980	3	30	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	3	0	0	1	1		
3	Dhaka	1980	4	0	0	0	1	0	1	0	0	0	0	0	6	0	1	0	0	0	77	0	43	0	1	0	0	0	4	0	11	2	NaN		
4	Dhaka	1980	5	1	0	11	0	21	60	2	4	74	0	25	25	5	1	15	4	0	21	2	28	1	0	0	0	19	1	32	39	2	21		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
15823	Sitakunda	2020	8	0	2	3	8	14	0	9	22	0	3	1	53	11	2	2	60	113	142	17	9	2	54	65	6	29	7	9	14	0	0		
15824	Sitakunda	2020	9	0	1	0	0	0	0	3	0	0	100	12	14	0	3	9	39	0	0	0	14	3	14	43	28	164	68	35	0	0	0	NaN	
15825	Sitakunda	2020	10	0	0	0	22	21	4	0	0	0	0	0	9	0	2	0	0	0	29	0	0	0	0	108	101	0	0	0	0	0	0	44	
15826	Sitakunda	2020	11	30	4	1	18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NaN	
15827	Sitakunda	2020	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

15828 rows x 34 columns

Figure 4.1: Rainfall Dataset

The index indicates how much precipitation is received corresponding to the long term average for a specified area and timeframe. In the rainfall dataset, the data has been recorded on a daily basis.



	Station	Year	Flood
0	Dhaka	1980	YES
1	Dhaka	1981	NO
2	Dhaka	1982	NO
3	Dhaka	1983	YES
4	Dhaka	1984	YES
...	...	...	...
1314	Sitakunda	2016	YES
1315	Sitakunda	2017	YES
1316	Sitakunda	2018	NO
1317	Sitakunda	2019	YES
1318	Sitakunda	2020	YES

1319 rows × 3 columns

Figure 4.2: Flood Dataset

The flood dataset has records of whether floods happened around the recorded stations from 1980-2020.

## 4.2 Pre-processing Steps

As mentioned in the methodology subsection, data preprocessing, the dataset will go through : data cleaning, feature engineering, feature encoding and feature scaling.

### 4.2.1 Data Cleaning

Firstly, in the dataset, unequal days of each month were handled. For example, some months have 31 days and some have 28 or 30 days. So, there was no data of rainfall for those specific days. Additionally, some days of particular stations did not have any data. Therefore, data imputation was applied to handle this issue. For instance, in the month of February,1980, the month ended at the 29th date, so for the 30th and 31st dates, the value zero for rainfall was added. Furthermore, in the dataset, the rainfall value for the 14th and 15th of 1983 November, Dhaka station, was missing, so, the value zero was also added here.

### 4.2.2 Feature Engineering

From the rainfall dataset, monthly rainfall data has been calculated and added into a new column. Then, the monthly rainfall data was set according to the particular stations and years. At this stage, the dataset contains the column of stations, years and all the 12 months. After that, the flood data that was collected was merged as a new feature into the rainfall data according to the stations and years.

	Station	Year	January	February	March	April	May	June	July	August	September	October	November	December	Annual	Flood
0	Dhaka	1980	3	32	54	147	414	323	380	269	296	300	0	0	2218	YES
1	Dhaka	1981	10	42	109	274	272	168	356	188	320	82	9	35	1865	NO
2	Dhaka	1982	0	15	81	104	154	514	136	346	258	146	51	0	1805	NO
3	Dhaka	1983	14	61	138	318	348	300	179	437	322	253	0	18	2388	YES
4	Dhaka	1984	13	1	5	124	707	637	694	311	478	58	0	0	3028	YES
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1314	Sitakunda	2016	3	285	55	65	328	949	1304	359	378	217	109	0	4052	YES
1315	Sitakunda	2017	0	0	119	419	239	1025	631	772	835	179	1	42	4262	YES
1316	Sitakunda	2018	18	0	0	246	279	250	1313	454	170	218	0	3	2951	NO
1317	Sitakunda	2019	0	35	20	222	286	576	1114	424	333	170	114	14	3308	YES
1318	Sitakunda	2020	46	0	0	171	244	711	1472	657	550	340	54	0	4245	YES

1319 rows x 16 columns

Figure 4.3: Dataset after categorizing monthly rainfall data with flood data

### 4.2.3 Feature Encoding

The dataset that is used in this research paper has two attributes that have string type data which are - ‘Station’ and ‘Flood’. As machine learning models give better results for numerical values, the string type datas are encoded. The attribute ‘Station’ contains all the names of the stations from where the daily rainfall data has been collected. Then by performing label encoding, the categorical values of the ‘Station’ column have been transformed into numerical values without adding any additional column. In this dataset, the attribute ‘Flood’ has two unique values, ‘YES’ and ‘NO’ and to encode these values, binary encoding is used. After encoding the values of the feature ‘Flood’, the value ‘YES’ is replaced by 1 and the value ‘NO’ is replaced by 0.

	Station	Year	January	February	March	April	May	June	July	August	September	October	November	December	Annual	Flood
0	1	1980	3	32	54	147	414	323	380	269	296	300	0	0	2218	1
1	1	1981	10	42	109	274	272	168	356	188	320	82	9	35	1865	0
2	1	1982	0	15	81	104	154	514	136	346	258	146	51	0	1805	0
3	1	1983	14	61	138	318	348	300	179	437	322	253	0	18	2388	1
4	1	1984	13	1	5	124	707	637	694	311	478	58	0	0	3028	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1314	34	2016	3	285	55	65	328	949	1304	359	378	217	109	0	4052	1
1315	34	2017	0	0	119	419	239	1025	631	772	835	179	1	42	4262	1
1316	34	2018	18	0	0	246	279	250	1313	454	170	218	0	3	2951	0
1317	34	2019	0	35	20	222	286	576	1114	424	333	170	114	14	3308	1
1318	34	2020	46	0	0	171	244	711	1472	657	550	340	54	0	4245	1

1319 rows x 16 columns

Figure 4.4: Dataset after feature encoding

### 4.2.4 Feature Scaling

Standard Scaler has been used on the dataset to make it unbiased and relevant to the models used. The data is scaled by centering them around the mean with a unit standard deviation. The formula for standardization can be defined as:

$$X = \frac{X - \mu}{\sigma}$$

where,  $\mu$  is the mean of the feature values and  $\sigma$  is the standard deviation of the feature values [25]. There are no restrictions on the range of the values. The dataset has been split into a train and test set with a ratio of 80:20. Then, the features have been scaled using the standard scaler.

### **4.3 Programming Language and Computational Speed**

Python language was used to test and implement a proposed model for flood detection. The models were tested with various computational calculations. For the large computational activities, Python along with Google Compute Engine Backend along with around 16 gigabytes of RAM has been used as the main programming language for the applied models since it can handle large computational activities that require optimizing.

### **4.4 Application of the Machine Learning Models**

For this research, we have used Binary Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbor (KNN) and Decision Tree Classifier (DTC) for predicting our data. The data has been split into 80:20 ratio for training and testing the models. The classification has been done on 16 columns: Station, Year, 12 months and Flood index. Firstly the models have been implemented on the whole dataset which consists of data from 1980-2020. Later, the same models have been implemented with a shorter timeline of 10 years, 2011-2020 to check the accuracy and compare with the previous implementation.

# Chapter 5

## Results and Discussion

### 5.1 Classification Result

For the prediction, four different types of classifiers have been used. The classifiers are - Binary Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) and Decision Tree Classifier (DTC).

#### 5.1.1 Timeline: 1980 to 2020

At first, the models were applied to the data that had been collected within the timeline of 1980 to 2020. The result of the applied models is given in Table 5.1.

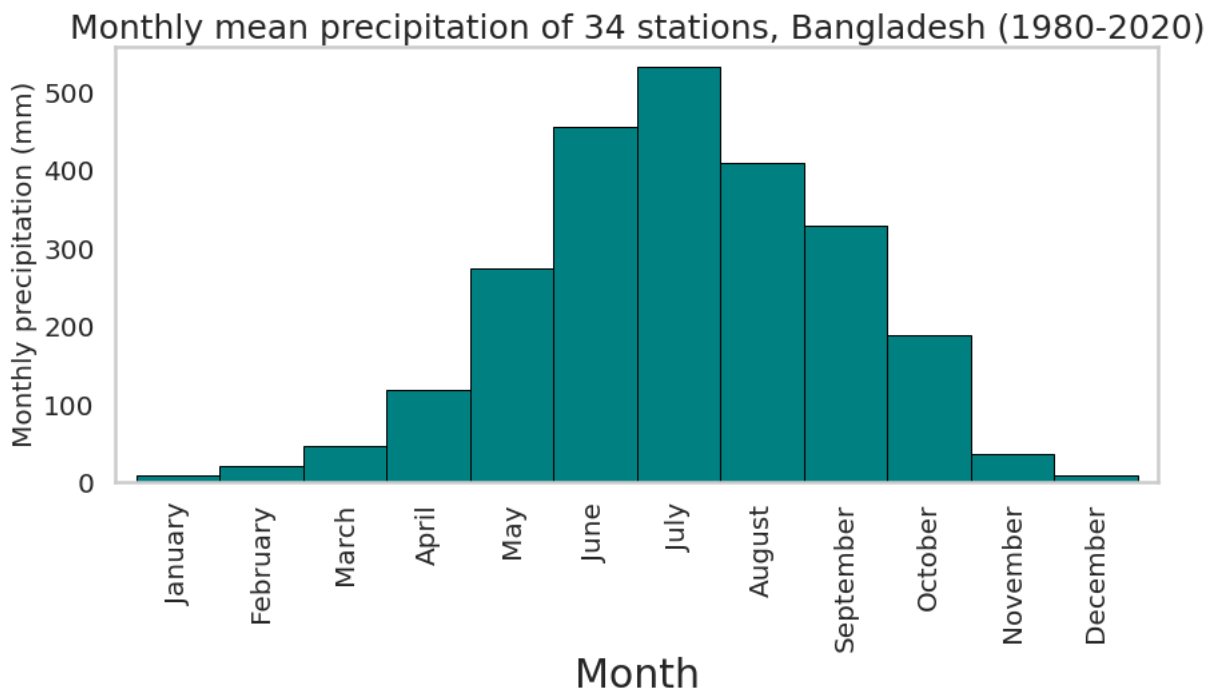


Figure 5.1: Monthly Mean Precipitation (Timeline: 1980-2020)

<b>Machine Learning Models</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
Binary Logistic Regression	0.8561	0.75	0.55
Support Vector Classifier (SVC)	0.8409	0.7647	0.4333
K-Nearest Neighbors (KNN)	0.8371	0.7576	0.4167
Decision Tree Classifier (DTC)	0.7879	0.5303	0.5833

Table 5.1: Results of Implemented Models (Timeline: 1980-2020)

From the table, Binary Logistic Regression has the highest accuracy rate of 0.8561 with a precision and recall score of 0.75 and 0.55 respectively.

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
0	0.88	0.95	0.91	204
1	0.75	0.55	0.63	60
<b>accuracy</b>			0.86	264
<b>macro avg</b>	0.81	0.75	0.77	264
<b>weighted avg</b>	0.85	0.86	0.85	264

Figure 5.2: Classification Report of Binary Logistic Regression (Timeline: 1980-2020)

After that, the Support Vector Classifier (SVC) has the highest accuracy of 0.8409 with a precision of 0.7647 which is higher than the Binary Logistic Regression.

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
0	0.85	0.96	0.90	204
1	0.76	0.43	0.55	60
<b>accuracy</b>			0.84	264
<b>macro avg</b>	0.81	0.70	0.73	264
<b>weighted avg</b>	0.83	0.84	0.82	264

Figure 5.3: Classification Report of Support Vector Classifier (SVC) (Timeline: 1980-2020)

Then, K-Nearest Neighbors (KNN) has an accuracy of 0.8371, precision and recall score respectively 0.7576 and 0.4167.

	precision	recall	f1-score	support
0	0.85	0.96	0.90	204
1	0.76	0.42	0.54	60
<b>accuracy</b>			<b>0.84</b>	<b>264</b>
<b>macro avg</b>	<b>0.80</b>	<b>0.69</b>	<b>0.72</b>	<b>264</b>
<b>weighted avg</b>	<b>0.83</b>	<b>0.84</b>	<b>0.82</b>	<b>264</b>

Figure 5.4: Classification Report of K-Nearest Neighbors (KNN) (Timeline: 1980-2020)

Lastly, Decision Tree Classifier (DTC) has an accuracy of 0.7879 which is the lowest among the models that have been used but it has the highest recall score of 0.5833.

	precision	recall	f1-score	support
0	0.87	0.85	0.86	204
1	0.53	0.58	0.56	60
<b>accuracy</b>			<b>0.79</b>	<b>264</b>
<b>macro avg</b>	<b>0.70</b>	<b>0.72</b>	<b>0.71</b>	<b>264</b>
<b>weighted avg</b>	<b>0.80</b>	<b>0.79</b>	<b>0.79</b>	<b>264</b>

Figure 5.5: Classification Report of Decision Tree Classifier (DTC) (Timeline: 1980-2020)

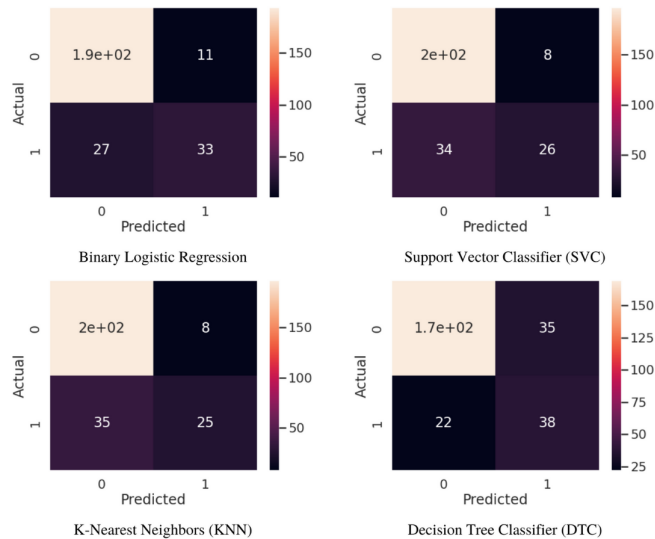


Figure 5.6: Confusion matrix of the used models (Timeline: 1980-2020)

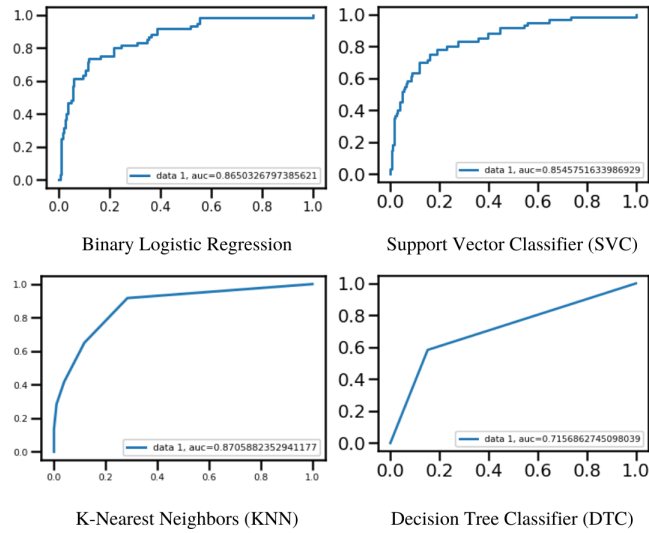


Figure 5.7: ROC Curves of used models (Timeline: 1980-2020)

From the ROC Curves, it is shown that the K-Nearest Neighbors (KNN) model has the highest AUC score of 0.87 but from table 1 and figure 13, K-Nearest Neighbors (KNN) model has less accuracy and f1 score than Binary Logistic Regression Model (Accuracy: 0.8561, f1 score: 0.86).

### 5.1.2 Timeline: 2011 to 2020

To check if the accuracy gets better, this time, the data of only the last 10 years (2011-2020) has been used with the same models. The result of the applied functions is given in Table 5.2.

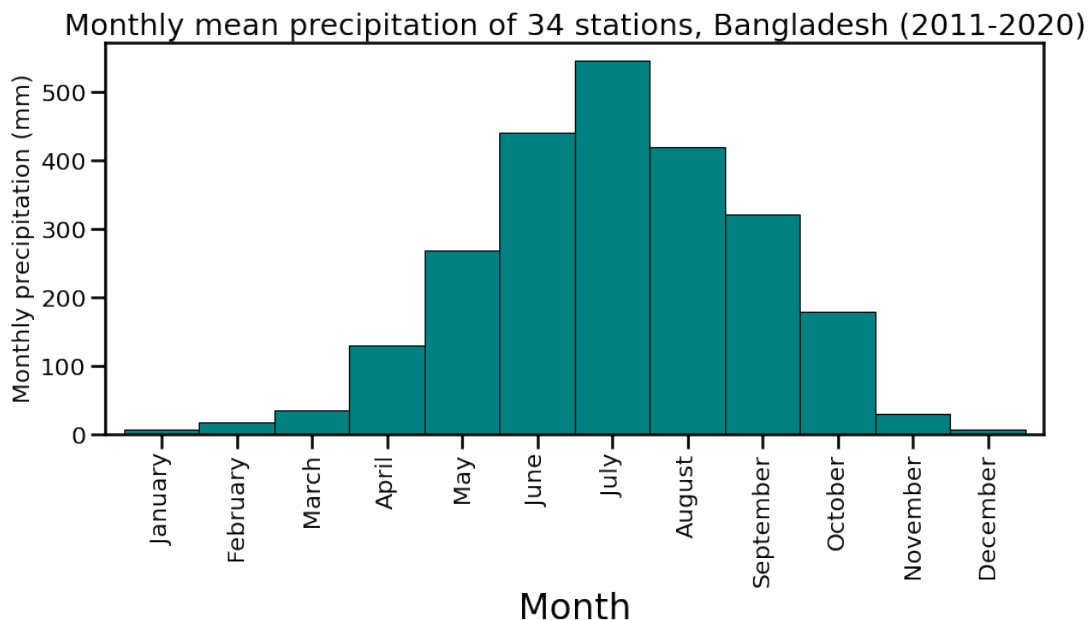


Figure 5.8: Monthly Mean Precipitation (Timeline: 2011-2020)

<b>Machine Learning Models</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
Binary Logistic Regression	0.8676	0.6154	0.667
Support Vector Classifier (SVC)	0.8088	0.4667	0.5833
K-Nearest Neighbors (KNN)	0.8235	0.50	0.50
Decision Tree Classifier (DTC)	0.8088	0.4545	0.4167

Table 5.2: Results of Implemented Models (Timeline: 2011-2020)

From the table, Binary Logistic Regression has the highest accuracy rate of 0.8676 with a highest precision and also highest recall score of 0.6154 and 0.6667 respectively.

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	<b>0.93</b>	<b>0.91</b>	<b>0.92</b>	<b>56</b>
<b>1</b>	<b>0.62</b>	<b>0.67</b>	<b>0.64</b>	<b>12</b>
<b>accuracy</b>			<b>0.87</b>	<b>68</b>
<b>macro avg</b>	<b>0.77</b>	<b>0.79</b>	<b>0.78</b>	<b>68</b>
<b>weighted avg</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>	<b>68</b>

Figure 5.9: Classification Report of Binary Logistic Regression (Timeline: 2011-2020)

After that, the Support Vector Classifier (SVC) has accuracy of 0.8088 with a precision of 0.4667 and recall score of 0.5833.

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	<b>0.91</b>	<b>0.86</b>	<b>0.88</b>	<b>56</b>
<b>1</b>	<b>0.47</b>	<b>0.58</b>	<b>0.52</b>	<b>12</b>
<b>accuracy</b>			<b>0.81</b>	<b>68</b>
<b>macro avg</b>	<b>0.69</b>	<b>0.72</b>	<b>0.70</b>	<b>68</b>
<b>weighted avg</b>	<b>0.83</b>	<b>0.81</b>	<b>0.82</b>	<b>68</b>

Figure 5.10: Classification Report of Support Vector Classifier (SVC) (Timeline: 2011-2020)

Then, K-Nearest Neighbors (KNN) has an accuracy of 0.8235 which is the highest after Binary Logistic Regression, with equal precision and recall score of 0.50.



	precision	recall	f1-score	support
0	0.89	0.89	0.89	56
1	0.50	0.50	0.50	12
<b>accuracy</b>			0.82	68
<b>macro avg</b>	0.70	0.70	0.70	68
<b>weighted avg</b>	0.82	0.82	0.82	68

Figure 5.11: Classification Report of K-Nearest Neighbors (KNN) (Timeline: 2011-2020)

Lastly, Decision Tree Classifier (DTC) has an accuracy of 0.8088 which is equal to Support Vector Classifier (SVC), respectively precision and recall score of 0.4615 and 0.4167.

	precision	recall	f1-score	support
0	0.89	0.88	0.88	56
1	0.46	0.50	0.48	12
<b>accuracy</b>			0.81	68
<b>macro avg</b>	0.68	0.69	0.68	68
<b>weighted avg</b>	0.82	0.81	0.81	68

Figure 5.12: Classification Report of Decision Tree Classifier (DTC) (Timeline: 2011-2020)

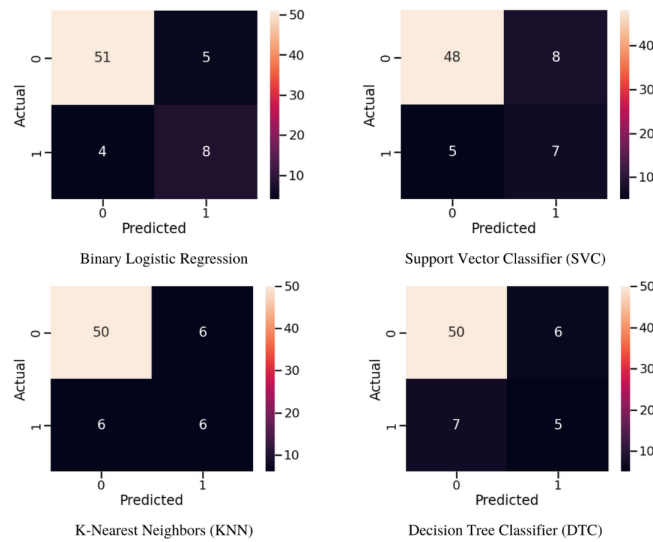


Figure 5.13: Confusion matrix of the used models (Timeline: 2011-2020)

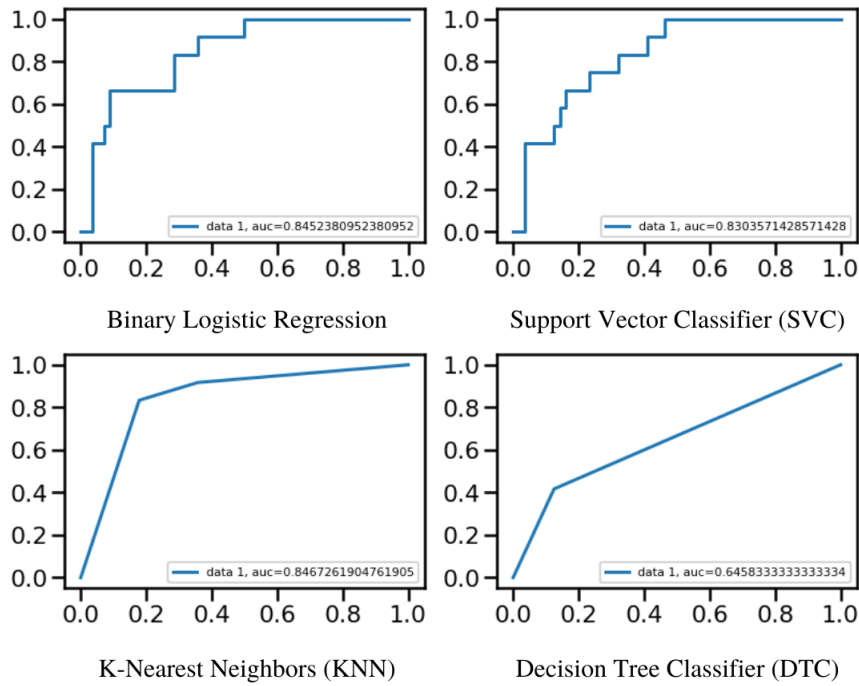


Figure 5.14: ROC Curves of used models (Timeline: 2011-2020)

From the ROC Curves, it is shown that the Binary Logistic Regression model and the K-Nearest Neighbors (KNN) model have almost AUC scores (0.845 and 0.846 respectively). As the Binary Logistic Regression model has a better accuracy (0.8676) and f1 score (0.87), this model is considered to be the better one among the four models.

From both of the timeline, all the models gave better accuracy on the 10 years of rainfall data (2011-2020) than the whole timeline (1980-2020). As climate changes over the years and depending on other parameters, the thresholds for floods are changing. That's why the shorter timeline of data gives slightly better accuracy.

# Chapter 6

## Conclusion

Floods can be devastating in certain years due to high rainfall and upstream water movement. Floods in 1988, 1998, 2004, 2007, and 2017 are examples of the intensity of floods in Bangladesh in terms of duration and damage. The 2020 floods have also emerged as one of these natural catastrophes, as the shocks are being felt by a vast number of people over an extended period of time. The most important thing to learn is how to better manage floods while causing the least amount of harm to people's lives and property, as well as how to lessen flood-induced vulnerability [24]. And for that, it is important to have an idea of a flood which is about to hit beforehand, so that necessary precautions can be taken. There are various factors that need to be kept in mind in order to predict the occurrence of a flood. For example, rainfall, water level of the river basins, geographical structure of the said location etc. Current researches have achieved very high accuracy rates to predict these floods, by correlating some of the aforementioned factors or more and by using different methods i.e. Artificial Neural Network (ANN), Deep Neural Network (DNN) etc.

Climate is an ever changing factor and with it as time goes on, other factors dependent on it also change. For instance, over the last 40 years, Bangladesh has changed a lot in terms of geography and industry, both of which have an effect on climate change. So, over time, the amount of rainfall and water level also have changed. Additionally, the population of Bangladesh also has increased over the years and with the number of housing structures. Also, the population density of Bangladesh is very high, which means that building structures are a lot in number. Due to this changing geographical structure of the country with time, the drainage system of the country is changing as well. Water level is dependent on the drainage system, as the rain water collects and discharges on a nearby river. So, all the factors that a flood occurrence is related to, are all related to one another. The previously mentioned research was done by taking time to collect these data for the respected location and finding the correlation between the factors.

Since, these factors change over a long time period, in this research, the models gave higher accuracy with a shorter time range. Also, due to time constraint only the rainfall data along with flood occurrence was found. But in the future, this research paper would attempt to develop the models further by adding the other factors and correlating them.

# Bibliography

- [1] M. Kabir and M. N. Hossen, “Impacts of flood and its possible solution in bangladesh,” *Disaster Advances*, vol. 12, pp. 48–57, Oct. 2019.
- [2] N. G. Society, *Flood*, Nov. 2011. [Online]. Available: <http://www.nationalgeographic.org/encyclopedia/flood/>.
- [3] T. Luo, A. Maddocks, C. Iceland, P. Ward, and H. Winsemius, “World’s 15 countries with the most people exposed to river floods,” Mar. 2015. [Online]. Available: <https://www.wri.org/insights/worlds-15-countries-most-people-exposed-river-floods>.
- [4] G. FitzGerald, W. Du, A. Jamal, M. Clark, and X.-Y. Hou, “Flood fatalities in contemporary australia (1997–2008),” *Emergency Medicine Australasia*, vol. 22, no. 2, pp. 180–186, 2010, ISSN: 1742-6723. DOI: 10.1111/j.1742-6723.2010.01284.x. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1742-6723.2010.01284.x>.
- [5] E. Holden, “Flooding will affect double the number of people worldwide by 2030,” *The Guardian*, Apr. 2020, ISSN: 0261-3077. [Online]. Available: <https://www.theguardian.com/environment/2020/apr/23/flooding-double-number-people-worldwide-2030>.
- [6] AsiaNews.it, *Floods in southeast asia: Dozens of deaths, thousands of homeless people*. [Online]. Available: <https://www.asianews.it/news-en/Floods-in-Southeast-Asia:-dozens-of-deaths,-thousands-of-homeless-people-51349.html>.
- [7] [Online]. Available: <https://reliefweb.int/report/bangladesh/bangladesh-monsoon-floods-2020-coordinated-preliminary-impact-and-needs-assessment>.
- [8] C. Nunez, *Learn about how floods happen and the damage they cause*, May 2021. [Online]. Available: <https://www.nationalgeographic.com/environment/article/floods>.
- [9] [Online]. Available: <https://www.ncdc.noaa.gov/billions/events>.
- [10] [Online]. Available: <https://reliefweb.int/report/bangladesh/dec-bangladesh-1998-flood-appeal-final-report-independent-evaluation>.
- [11] T. H. Dewan, “Societal impacts and vulnerability to floods in bangladesh and nepal,” *Weather and Climate Extremes*, SI: IGBP APN, vol. 7, pp. 36–42, Mar. 2015, ISSN: 2212-0947. DOI: 10.1016/j.wace.2014.11.001. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2212094714000930>.
- [12] F. A. Ruslan, A. M. Samad, Z. M. Zain, and R. Adnan, “5 hours flood prediction modeling using nnarx structure: Case study kuala lumpur,” in *2014 IEEE 4th International Conference on System Engineering and Technology (ICSET)*, vol. 4, Nov. 2014, pp. 1–5. DOI: 10.1109/ICSEngT.2014.7111798.

- [13] —, “Flood water level modeling and prediction using narx neural network: Case study at kelang river,” in *2014 IEEE 10th International Colloquium on Signal Processing and its Applications*, Mar. 2014, pp. 204–207. DOI: 10.1109/CSPA.2014.6805748.
- [14] S. Sankaranarayanan, M. Prabhakar, S. Satish, P. Jain, A. Ramprasad, and A. Krishnan, “Flood prediction based on weather parameters using deep learning,” *Journal of Water and Climate Change*, vol. 11, no. 4, pp. 1766–1783, Nov. 2019, ISSN: 2040-2244. DOI: 10.2166/wcc.2019.321. [Online]. Available: <https://doi.org/10.2166/wcc.2019.321>.
- [15] R. Adnan, F. A. Ruslan, A. M. Samad, and Z. Md Zain, “Flood water level modelling and prediction using artificial neural network: Case study of sungai batu pahat in johor,” in *2012 IEEE Control and System Graduate Research Colloquium*, Jul. 2012, pp. 22–25. DOI: 10.1109/ICSGRC.2012.6287127.
- [16] —, “Extended kalman filter (ekf) prediction of flood water level,” in *2012 IEEE Control and System Graduate Research Colloquium*, Jul. 2012, pp. 171–174. DOI: 10.1109/ICSGRC.2012.6287156.
- [17] S.-Y. Liong and C. Sivapragasam, “Flood stage forecasting with support vector machines1,” *JAWRA Journal of the American Water Resources Association*, vol. 38, no. 1, pp. 173–186, 2002, ISSN: 1752-1688. DOI: 10.1111/j.1752-1688.2002.tb01544.x. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.2002.tb01544.x>.
- [18] [Online]. Available: <https://www.javatpoint.com/feature-engineering-for-machine-learning>.
- [19] A. A. T. Fernandes, D. B. Figueiredo Filho, E. C. d. Rocha, and W. d. S. Nascimento, “Read this paper if you want to learn logistic regression,” *Revista de Sociologia e Política*, vol. 28, Jan. 2021, ISSN: 0104-4478, 1678-9873. DOI: 10.1590/1678-987320287406en. [Online]. Available: <http://www.scielo.br/j/rsocp/a/RWjPthhKDYbFQYydbDr3MgH/?lang=en>.
- [20] Aug. 2021. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/>.
- [21] [Online]. Available: <https://scikit-learn.org/stable/modules/svm.html>.
- [22] [Online]. Available: <https://scikit-learn.org/stable/modules/tree.html>.
- [23] [Online]. Available: <https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/>.
- [24] F. Khatun, *Living with floods and reducing vulnerability in bangladesh*, Aug. 2020. [Online]. Available: <https://www.thedailystar.net/opinion/macro-mirror/news/living-floods-and-reducing-vulnerability-bangladesh-1950277>.
- [25] Apr. 2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/>.