Estimating flood susceptibility of Bangladesh in the future year using machine learning

by

Sakib Bin Alim
21141068
Rakebun Islam Lucky
21141071
Aunindya Arif Ahmed
17101225
Prethu Nahian
17301191

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

School of Data and Sciences
Brac University
June 2021

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:

Sakib Bin Alim
21141068

Rakebun Islam Lucky
21141071

Aunindya Arif Ahmed
17101225

Prethu Nahian
17301191

Approval

The thesis/project titled "Estimating flood susceptibility of Bangladesh in the future year using machine learning" submitted by

1. Sakib Bin Alim (21141068)
2. Rakebun Islam Lucky (21141071)
3. Aunindya Arif Ahmed (17101229)
4. Prethu Nahian (17301191)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 6, 2021.

Examining Committee:

Supervisor:
(Member)

MD Saiful Islam
Lecturer
School of Data and Sciences
BRAC University

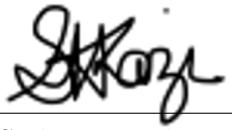
Co-Supervisor:
(Member)

Shehran Syed
Vice chancellor's Fellowship Lecturer
School of Data and Sciences
Brac University

Co-Supervisor:
(Member)

Marum Monem Anik
Lecturer
School of Data and Sciences
Brac University

Head of Department:
(Chair)



Dr Sadia Hamid Kazi
Chairperson and Associate Professor
School of Data and Science
Brac University

Abstract

Being a riverine country with more than 400 rivers, flood is a common phenomenon for Bangladesh. As, the land is less than five meters above sea level, and also due to heavy rainfall during monsoon season, it makes the country an easy target of flooding and about 30% of the total area is in danger level during this period. Additional to the yearly flooding, every 4 to 5 years there is a major flood occurs which covers more than 60% of the country. As of 22 July, 2020 alone, 102 upazila and 654 unions have been inundated in flood, affecting 3.3 million people, leaving 731,958 people water logged and a total of 93 deaths [2]. The aim of this research is to predict Bangladesh's susceptibility to flooding so that the government as well as the people of this country can take necessary steps to lessen the effect. To predict the probability of flood we will be using some machine learning algorithm namely Linear Regression model, Random forest Regressor, Naive Bayes Theorem and Artificial Neural Network. This study is based on the data set from 1991-2013 water level and weather variables from Khulna districts Rupsa-Pasur station.

Keywords: Flood Susceptibility, Machine Learning, Flood in Bangladesh, Linear Regression Model and Random forest, Naive Bayes Theorem, Artificial Neural Network.

Dedication

We would like to dedicate this thesis paper to our respected parents, and all the friends and family members who supported us in the undergraduate run. Moreover, we would like to dedicate this paper to our respected faculty members who helped us throughout the years inside and outside classroom with their knowledge and how to implement it in real life, also to the respected co supervisors and specially our supervisor Md. Saiful Islam sir. Last but not the least, we also want to mention the names of (Late) Md. Abdul Mottalib sir (May he rest in peace) and Mahbubul Alam Majumder sir for showing us and the whole department, the right path to gain knowledge.

Acknowledgement

We would like to acknowledge that this thesis report has been done under the supervision and guidance of our honorable teacher, MD. Saiful Islam sir. Firstly, we would like to thank the Almighty for providing us with the opportunity to do the research successfully, for giving us the confidence and guidance by Md. Saiful Islam sir. Then, we would like to thank our co supervisor Mr. Maruf Monem Anik and Mr. Shehran Syed for helping with our ideas and executing them. Moreover, our parents for the endless support and prayers. Finally, we would like to thank the General Manager of Department of Disaster Management and the in charge officer of Hydrology department, Bangladesh Water Development Board (BWDB), Dhaka for giving us their precious time and our required data set.

Table of Contents

List of Figures

List of Tables

Nomenclature

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

ANN Artificial Neural Network

$BWDB$ Bangladesh Water Development Board

$corr$ Correlation

CV Cross Validation

DDM Department of disaster management

KNN K-Nearest Neighbor

LR Linear Regression

MAE Mean Average Error

ML Machine Learning

MSE Mean Squared Error

NB Naive Bayes

RF Random Forest

$RMSE$ Root Mean Square Error

SD Standard Deviation

SVM Support Vector Machine

Chapter 1

Introduction

1.1 Overview

A flood is an overflow of water that occurs and submerge lands that are usually dry and mostly occurs when the water level rises over the top of river banks. Floods are one of the most devastating natural disasters that occur on an extensive level and their post-disaster effects remain for a long time. River floods can happen for several reasons, among them one of the most important and frequent reason is excessive rainfall, a result of tropical storm. Floods can happen slowly or suddenly, Slow flooding events occur mostly on large riverside area due to overflow causing damage in a vast area.

Bangladesh, a country owing its position in the South-Asian region on the global map, is formed by a delta plain at the confluence of the Padma (Ganges), Jamuna (Brahmaputra) and Meghna rivers. It is neighboured by India mostly and a small amount of border with Myanmar. It is one of the most affected and vulnerable countries to the climate change, which arises its vulnerability to drought and flood. Specially, floods are the most occurring natural disaster in Bangladesh due to climate change. In last few decades, floods caused more damage than any other natural disaster in Bangladesh. In 1998, over 75% of the total area of Bangladesh was flooded, including its capital. It was similar as the previous devastating flood of 1988. During the devastating flood 30 million became homeless and thousands were dead. Due to contamination in the drinkable water, it creates scarcity of drinking water. In, 1998, there was a massive flood that flooded half of the country under water. More than 1000 people died during the flood. It resulted in more than 400 factories closure and almost 700,000 hectares of crops were destroyed. As a result, it turned into a 20% reduction in the economic production of Bangladesh.[17] There were few more devastating floods in 2004,2007, 2015, 2017 and 2020. A combination of heavy rainfall within and outside the country and synchronization of peak flows of the major rivers contributed to the flood. These all portrays a picture how Bangladesh is continuously facing cumulative effects of flood.

For this paper our field of interest is assessment of water level and other key factors of the flood in Bangladesh. Later on in this study we show the relation between these factors with floods and these are being exploited to predict the upcoming floods. Although, several studies proved the efficiency of predicting the flood using machine learning, which would be discussed in Chapter 2, we believe this type of work is new to Bangladesh region.

1.2 Current Scenario

Being situated on the Brahmaputra River Delta (also known as Ganges Delta), Bangladesh becomes an easy target of flooding. However, the many river-streams flowing into the Bay of Bengal does not help in this scenario. Bangladesh is only less than 5 metres above sea level and the heavy rainfall, water flashing from the nearby hills and the inflow of water from the upstream catchments causes the frequent flooding nearly every year. In the monsoon season rainfall caused by the relief of Himalayas makes June to September flood prone season.[17]
In the 20th century, there were 18 major floods occurred, among them 1987, 1988 and 1951 had catastrophic consequences. In 1987, throughout July and August, the flood covered around 57,300 square kilometres (22,100 square miles) of land which is about 40% of the total area of Bangladesh.As for this catastrophic effect it was estimated as once in a 30-70 year event.[17]
The catastrophic flood of 1988, which lasted throughout August and September, inundated about 82,000 square kilometres (32,000 square mile) of land that is about 60% of the total area of the country, making its a 50-100 year events. The flood was caused by rainfall along with the very high flows of the tree major rivers of the country in only 3 days of time. [17]
In 1998, over 75% of the total area of the country including half of Dhaka was flooded,which was much too similar to the catastrophic flood of 1988, in terms of the extent of the flooding.This flood ws caused by a combination of heavy rainfall within and outside the country and with the synchronization overflows of the major rivers, causing 30 million people homeless. During this event the death toll reached over a thousand and there was cholera and triphoid outbreaks as well.However, only few hospitals were functional because of damage from the flooding during that time, and those that were open could only do so much as there were too many patients to handle. This flood caused damage in every section which includes about the damage of 700,000 hectares of crops, 400 factories were forced to close, and there was a 20% decrease in economic production.[1]
Although the flood of 1999 (July-September) was not as serious as the flood of 1998, it still caused pretty serious damage, causing many deaths and leaving many people homeless. In 2004 there was a repetition of 1998 and 1999 causing two thirds of the country under water.[2]
In case of 2017, the major flood occurred in two different timelines. The first time was less severe than the 2nd time. In 2017 first flow, there were total number of 8 districts affected, which is very small amount in terms of the 2nd flow, which was 32. The number of upozillas was 71 first time where it was 208 the second time. Even the amount of affected crops was more than double in the 2nd time. Moreover, the number of affected houses was 26288 the first time and it went up to 633792 by the 2nd time. A total number of 10 people died by the first time and it went up to 147 by the 2nd time. In terms of animals, the numbers were also very differentiable. 1743 and 84770. In short, this year was a hectic one because the damage was very high if we match it with 2020.
But when we look at 2019 stats, there are numbers similar to both 2017 and 2020 in different cases. The number of affected districts was 28, which was close to the stat of 2020. The number of affected families and affected people were also close to 2020. However, the number of affected crops was very low in terms of both 2017

and 2020. On the other hand, the death of people was over a 100 which was close to the amount of people died in 2017, and the death of animal was 22,688 which was much less than both 2017 and 2020. The number of institution affected was 5507 and the number of roads affected was 292, which is also similar to 2017, and much less than 2020.

In the most recent flood of 2020, along with the Covid pandemic, people had to face the flood challenge as well. Due to the heavy rainfall, there was a quite severe situation in that year. The flood damaged total of 33 districts, inside of that there was total of 151 upozillas. A total number of 1343121 affected families and 4314793 affected people (on record) were affected by the flood. Moreover, a total number of 130365 acres of crops were affected partially and 81259 acres of crops were affected fully. Furthermore, a total 183342 number of houses were fully affected and a number of 554480 houses were affected partially. Total 42 people and over 60,000 animals died due to the flood on record. Total 73 institutions were fully affected due to the flood and a number of them were destroyed by the flow of river waves. Almost 2000 bridges were affected by the flood and went under the water. In some case, the railing of the bridges were not even visible. Throughout the country there was around 506 km of dams that got destroyed by flood in 2020.

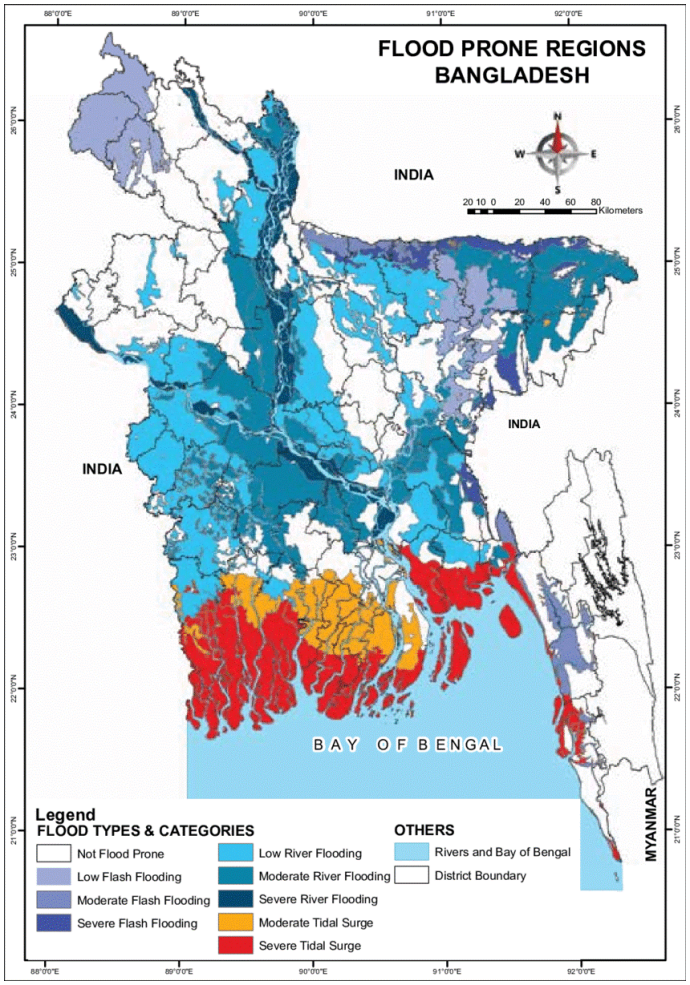


Figure 1.1: Flood Prone Regions of Bangladesh

Apart from the major floods, if we talk about the areas that are most likely to be affected due to flash floods are mostly located in the northern belt of Bangladesh,

namely Netrokona, Sylhet, Habiganj, Moulvibazar and Sunamganj Districts. As for the southeast part like Bandarban, Chittagong and Cox's Bazar are also in flood danger zone. Additionally, Bangladesh is covered with the net like river/water distributaries system including three major river.
Furthermore, if we were to estimate the district that are more likely prone to frequent flooding, are those that consists rivers and/or closer to the sea area. These are Gaibandha, Barishal, Narayanganj, Kurigram, Rangpur, Lakshmipur, Shariatpur, Tangail, Pabna, Madaripur, Bogra, Comilla, Manikganj, Kushtia, Sirajganj, Chapai Nawabganj, Dhaka, Natore, Sherpur, Chandpur, Faridpur, Munshiganj, Narshingdi, Kishoreganj, Rajshahi, Jamalpur, Gajipur and Kurigram. Due to the pre-existing underlying vulnerabilities, floods in Northern part of Bangladesh often has catastrophic effects. As for the southwest part namely Jhenaidah, Jessore, Khulna and Satkhira Districts are mainly faces flood as a result of bad drainage system and unabated canals. [8]
Although, peoples are adjusting their lifestyle accordingly, as the yearly frequent flooding specially during monsoon season has become an regular phenomenon. However, it has been observed that constructing elevated houses and roads to avoid the initial effects of floods are quite effecting, but the amount of the elevated houses are nearly not enough to fit all the affected people. Therefore, with that in our mind we aim to give people the idea about the upcoming flood and its effect on people, so they can build necessity amount of elevated houses and roads to keep all the affected people safe. Additionally, organizations that are willing to help by raising funds can also have an idea about how much they need to raise to offer help to those affected people.

1.3 Research Objectives

Flooding has become an yearly event for the people of this country, disrupting the normalcy of life. Even though people are taking many prevention step, it is nearly not enough, causing a havoc and damaging on a larger scale in almost every sector. Given that scenario, we tried to build a model that could give an idea about the flood yet to come and be able to prepare enough to face the flood and have comparatively lesser damage. The main purpose of this research is to provide people a warning of the upcoming flood. The objectives of this study are given below:

- Our main objective of this research is to build and ML based model that predicts the probability of flood that might occur. This will assist people to take necessary steps to prevent the damages
- The model will predict the chances of the flood to happen based on the previous record of water level, temperature, humidity, cloud coverage etc.
- The model will provide actual numbers to give people a glimpse of the upcoming flood and what to expect.
- Introduce a model that will predict the probability along with the maximum water level for the upcoming months. However, the model will not predict other weather variables.

- The model will operate on real world data and will provide the prediction on maximum accuracy.

1.4 Challenges Faced

Initially, after being decided upon our topic with great many difficulties, we had to decide what type of attributes we need for our data-set. As per our research topic, initially we wanted to mainly focus on the damages that were caused by the flood in Bangladesh. So, we wanted to collect the data-sets of damages due to flood previously, but we could not find anything online. We even contacted many authors who previously worked on related topics but we were without any luck. Then we visited DDM (Department of disaster management) and after many visit we were able to collect a data-set that contains overall damages of Bangladesh. Unfortunately, we were provided with the hard-copy of the list of damages on yearly basis and we had to go through with the lengthy process of digitization ourselves. But the data-set was not very fitting and we were not able to train the data-set other than only two machine learning algorithms. As of the flood prediction itself, we were facing the same problem with data-set collection. We could not find anything online, so we had to visit BWDB (Bangladesh Water Development Board) and collect the data from there. However, we found out that the charged amount of the data-set was too high to work for the whole country. So, we decided to reduce our range and timeline and focus only one station in particular, as we could not afford to pay that much amount for the data-set.

1.5 Benefits of Estimating Flood Susceptibility

Both rural and urban areas are affected by floods causing damage to infrastructure, personal assets and even human life. Children and senior citizens are the most exposed and vulnerable during floods. Most of the death during floods caused by drowning, electrocution and waterborne diseases like diarrhoea, cholera, dysentery, typhoid etc. Deaths due floods are declining because necessary steps are taken to prevent. So, it is visible to decline the direct deaths from flooding, estimation of flood susceptibility is needed. There are large-scale of effect on housing and households due to flooding. In the villages most of the houses are made by metal sheet ("Tin Sheets") and mud walls. As a result, a lot of households are washed away or damaged by the rising water. So, a huge number of people become homeless. Sometime, the hospitals, clinics, educational institutes and many more government offices are affected by floods that leaves further consequences. So, estimating flood is way to reduce these damages on rural citizen's household and also prepare different solution if it damages the hospitals, clinics, educational institutions. The effects of damages on educational institutes disrupts the academic session. Sometime, floods make the roads unusable and disrupts the transportation system. In the rural areas, agriculture is one of the most important largest employment sectors of Bangladesh by employing at least 43% of the total population and providing almost 15% to the GDP. Most of the flood results in damaging the summer crops during June-August, cause during this period there are high possibilities of flooding. Sometime, the loss is too much for the country and it results in food shortage. Fish agriculture is also

a leading profession here. Bangladesh is also renowned for a leading fish producing nation. It is one of the most valuable agricultural crops. Due to flooding, sometime Bangladesh needs foreign aid to face the loss of crops and fisheries. It results in economic loss for both farmers and the country.[7] It results in lowering employment in agricultural sector. Flood also results in increasing in the price of household costs. So, the rebuilding become more costly for Government and NGOs.[4] To reduce the effects and damages caused by floods, the estimation of flood susceptibility is necessary. Machine learning made it easier to predict the accuracy on large number of data. By estimating the rainfall, we can predict the chances of flooding. Floods cause a huge damage towards personal and national assets. By implementing flood susceptibility, many lives were saved for last few years for the early flood warning. It makes it easier for the timely operation of flood control structures such as temporary flood defences. People and authority can remove the properties to some safe places to reduce the loss during flood. The authority can easily install flood resilience actions.

1.6 Thesis outline

The aim of this research is to build a model that can predict the vulnerability of flood based on the previous record. The goal was to formulate the best model and predict with as much accuracy as possible that can help with the flood damage mitigation and create awareness among people.
To begin with, in the first chapter (Chapter 1) gives the gentle introduction and the current scenario of Bangladesh in terms of flood. The motivations behind this study and problem statement was develop to give people the upcoming scenario of flood.
In the second chapter (Chapter 2), is provided with the details literature review. All the related works in this field is discussed briefly.
Thirdly, in Chapter 3, the background analysis of the models that are going to be build is discussed. The supervised algorithms that are going to be pipe-lined later on is introduced in this chapter.
Next in Chapter 4, the research methodology and workflow is proposed. From data collection, pre-processing and to feature selection and engineering that were conducted was elaborated here.
Furthermore, in Chapter 5, the model implementation and optimization process was elaborated. The techniques to optimize the model with cross validation was briefed in this chapter.
Finally, in Chapter 6, the results analysis from the models were conducted. Visual comparison between the models and results were shown in this chapter. And to end the study conclusion and future work were drawn in the final chapter (Chapter 7).

Chapter 2

Literature Review

Many recent studies focused on the flood. Because, it is creating a lot of chaos recently. After 2000, many researchers are emphasizing this topic. As a result, there are a lot of paper regarding this issue. So, there are a lot of chance to take precautions cause, it is easy to predict time using previous researches. While there has been much research on the time and places of flood, few researchers have focused on the damage or losses caused by flood.

Flood susceptibility prediction
The study integrated machine learning methods with re-sampling algorithms for flood susceptibility prediction, proposed a Generalized Additive Model (GAM) for flood susceptibility predictions. In their paper they used multivariate adaptive regression splines (MARS) for modeling complex nonlinear process and reproducing the simple, piece-wise linear models, Boosted regression trees (BRT) to construct multiple regression models to improve the performance of the predictions. The authors also used Bootstrapping algorithm for the flood susceptibility predictions over the B (parameter) and finally Random sub-sampling to splits the data into train and test portions and repeats the process in B iterations. This study indicates that employing re-sampling approaches improved the machine learning models performance. The BT (bootstrapping) algorithm outperformed the RS (random sub-sampling) algorithm in terms of performance evaluation. This study also shows that the BT-GAM and BT-MARS models had similarly close results. And finally the BT-MARS model reproduced lower area percentage of a very high susceptibility class.[12] This research is on the flood susceptibility modeling in Teesta River basin.[15] Teesta River is one of the largest rivers in Bangladesh. Every year Bangladesh faces a huge loss due to flood in this river. Another research paper is on the Quishui River. Quishui River also known as Zhuoshui River which is one of the largest river of China. Flood on these rivers cause a huge impact on the economic, casualties of these countries.[15] [16] The researchers used data collected from the Bangladesh Meteorological Department (BMC) [15] and Land Cover maps which was from the United States Geological Survey (USGS) website.[15] They also set 207 flood point to collected the data from the field.[15] From these 207 points they collected the data of water level, soil types, distance to rivers, rainfall.[15] They used previous data from 23 flood report for their research.[16]
Mainly this research focused on flood susceptibility based on the Bagging Algorithm.[16] In this paper, they imposed the most recent phase of novel ensemble

machine learning algorithm to predict and identify the flooding sites. To find out more accurate results they used REPtree, random tree, M5Ptree, random forest algorithms and to find the factors of flooding they used the Gain Ration Statistical technique. They generated receiver operating characteristic curve (ROC) from the collected data for validation of the flood susceptible models.[15] On the other hand, this research paper is forecast flood based on hydro-graph generalization and random forest.[15] Hydro-logical model were used as the main technical approach for real time flood forecasting. For technical and data conditions, they could not use hydro-logical model for arid and semi-arid areas. Real-time flood forecasting were used in the selected area. They created a hybrid flood forecasting model combining the hydro-graph and random forest and used the data they collected for validation and development of the model.[16] From their research, they found out that all models achieved more than 0.85 area under the curve indicating highly accurate flood models.[15] For the first time to compare the models with each other they applied The Freidman test, Wilcoxon signed-rank test, Kruskal–Wallis test and Kolmogorov–Smirnov test. The finding from all the algorithms stated that more 800 km2 area as the most vulnerable zone.[15] However, they found out hydro-graph model provided more accurate predictions.[16] With the help of 30 years long data they found it. Finally, these researches focused on the probability of flood in their selected rivers and areas. Though there are a lot of areas they did not do anything. For instance they should also collect more data from different reliable sources rather than imposing everything in only one single data-sheet.[16] They should also use present data. On the other paper the researches should have collect more data from field rather than only collect data from 207 selected areas. Overall from these researches we can find out some new information. From their findings we can know that among all the findings from the flood susceptibility models M5P was the most accurate.[15] So, we can say that M5Ptree is the best method among those four.[12] Also, for future researches it is really helpful. Cause they don't have to use these 4 models to find accurate result, because, comparing it with the empirical model and the hydro-graph showed more accurate predictions for flood forecasting.[16]
Snehil and Ruchi Goel in their study 'Flood damage analysis using machine learning techniques" used Gaussian Naive Bayesian for both classification and regression, Decision tree to predict value of target variable, Support vector machine (SVM) which performs really complex data transformations and separates the data based on the labels defined, K nearest neighbor (KNN), and finally Random forest for classification and regression problem. In their study they showed that Random forest and KNN performs best and equally well. Furthermore, Random forest performs better than normal regression trees. Random forest, KNN and normal regression also outperform the Gaussian Naive Bayesian. And finally discretization of the data and small data-set in combination with its sensitivity to low-quality data are the reasons for the poor performance of the Gaussian Naive Bayesian.[13]

Mathematical model for flood loss estimation
In the study, a physical based distributed hydro-logic model is used to simulate in land as well as river overflow flood. Tangible and intangible damages were classified and expressed in mathematical model. The concept of unit loss model has been used to formulate the mathematical model of flood loss categories. They also performed different object functions which are stage-damage, derived from the normalized and

average data. And finally, integration of the overflowing water inundation simulation model and the model for estimation of loss has been done for the dynamic exchange of simulated parameter. The physical based distributed hydro-logic has considered five major hydro-logic cycle processes. i) Interception and evaporation, ii) river-flow, iii) overland-flow, iv) unsaturated-zone flow, v) saturated-zone flow. Moreover, there are two model that can estimate the linkage effect and inter-sectorial relationship of the flood within economy. [3]

Prediction and analysis of flood affected household
In this study, they used linear regression to set the extent to where there is a connection and relation between a dependent and one or more independent variables, Random forest to get the prediction from each data sample and select the best solution by creating decision trees, artificial neural network and k-fold cross validation. Linear regression was used for comparing the model's performance with the other two algorithms as the base classifier. Most of the studies in the study consider the building-level damage and limited set of areas for flood. [9]

Flood susceptibility assessment using ml and multi-criteria decision analysis
In this they used artificial neural network to explore relationships between inputs and outputs, analytical hierarchy process to determine the weighting indicates the importance of each factor to the occurrence of floods. They also used logistic regression to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. They used frequency ratio and integrated model as well. This study shows that the prediction rate of the Logistic Regression model is better than other models. This work have considerable management implications for disaster management of a highly populous country.[10]

Machine learning to evaluate impacts of flood protection
Part of the dataset used in the study is available covering a certain areas that almost covers one third of the areas geographically. In this research they used machine learning approach to infer, whether the dam has caused welfare both inside and outside the embankment over time for socio economic status and household. We can see a machine learning approach to determine the differences in mortality and migration patterns of the peoples' household both inside and outside the collective dams over time. They also did a classification approach to a lot of differences in people migrating, and mortality patter of people that live inside and outside the collective dam or embankment over time. They also did a classification approach to classify two classes which includes i) Logistic regression, ii) SVM (Support vector system), iii) Random Forest. The ROC comparison across 4 years in suggests that there are difference in people living inside vs outside the embankment and there are discrimination over time-based on socio-economic variables. The study reinforces findings around the timely impacts of the collective dams or embankment but does not find that they are continued after 30 years or so, with certain well-being indicators that evens out mortality and migration. The household inside the embankment holds a huge amount of agricultural land within 10 years of embankment's construction. [14]

Chapter 3

Background Analysis

3.1 Linear Regression

A linear approach that is based on the relationship between the input variables (x) and an output variable (y), which is presumably a linear relationship. Based on the independent variables, the model targets a prediction value assuming the relation between them is linear, which can be used for forecasting and finding the relationship between variables. Thus, the model can be represent as:

$$\hat{y} = \alpha x + C \tag{3.1}$$

Where,

- $\hat{y}$ = The predicted value of the response variable y
- x = Predictor variable
- C = Intercept
- α = Slope

The models residual is the difference between y and $\hat{y}$. However, given a training data set, the Residual Sum of Squared (RSS) could be minimized using the equation 3.2.

$$RSS = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{3.2}$$

Hence, the cost function is the Mean Squared Error (MSE) between the predicted and true value.

$$MSE = \frac{1}{n} \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{3.3}$$

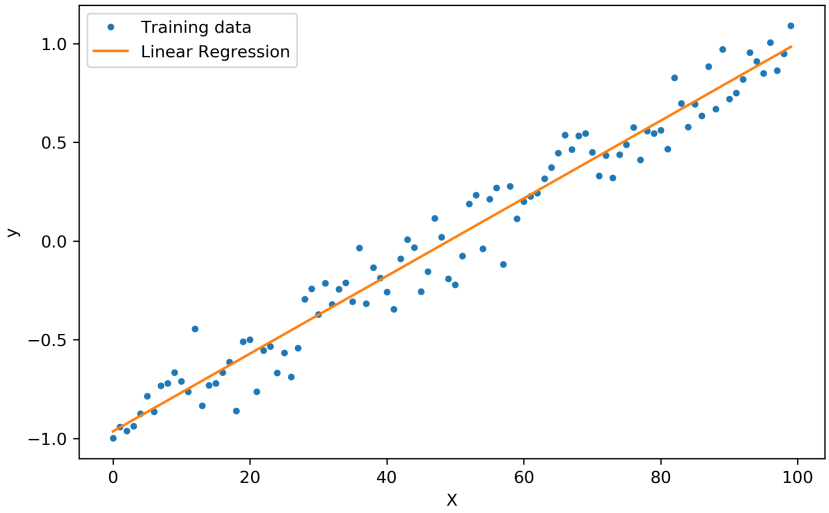


Figure 3.1: Linear Regression Model

3.2 Random Forest

Random Forest is an ensemble method that could be use for both classification and regression problem. The algorithm works by building multiple decision tress on the training data and predicts the output by the mean/average value of each tree. This model usually uses bootstrap aggregation, or begging to train the algorithm. Although, it works similar to decision trees, given that each node has been split by the measure of node impurity.[11] Here, the impurity of the node could be measured using the Errors Sum of Squares (ESS) which is given below.

$$ESS = \sum_{i=1}^{n(S_1)} y_i - \bar{y_1} + \sum_{i=1}^{n(S_2)} y_i - \bar{y_2} \tag{3.4}$$

S_1 and S_2 are the two children node after dividing. In order to find the node with lowest ESS, each variables were search using brute force. While RF has a major over-fitting problem, it can be solved by data bootstrapped. However, the reason RF performs better is because a large number of trees without any correlation are operating as a committee, outperforming any other individual sentient trees. The key for this to work is the low correlation between the models. This has an advantage because models with no correlation tends to provide better accuracy in terms of ensemble predictions rather than individual predictions. The reason behind this model to work better is because the trees work together against individual error. Therefore, the prerequisite for this model to perform well are as follows:

- Features should contain real signal to build the models rather than random guesses so that the models performs better.
- Individual trees needs be uncorrelated and so as their predictions and errors.

As Random Forest builds multiple decision tree in the process, it needs to row sampling and feature sampling randomly in order to form data set for the model.

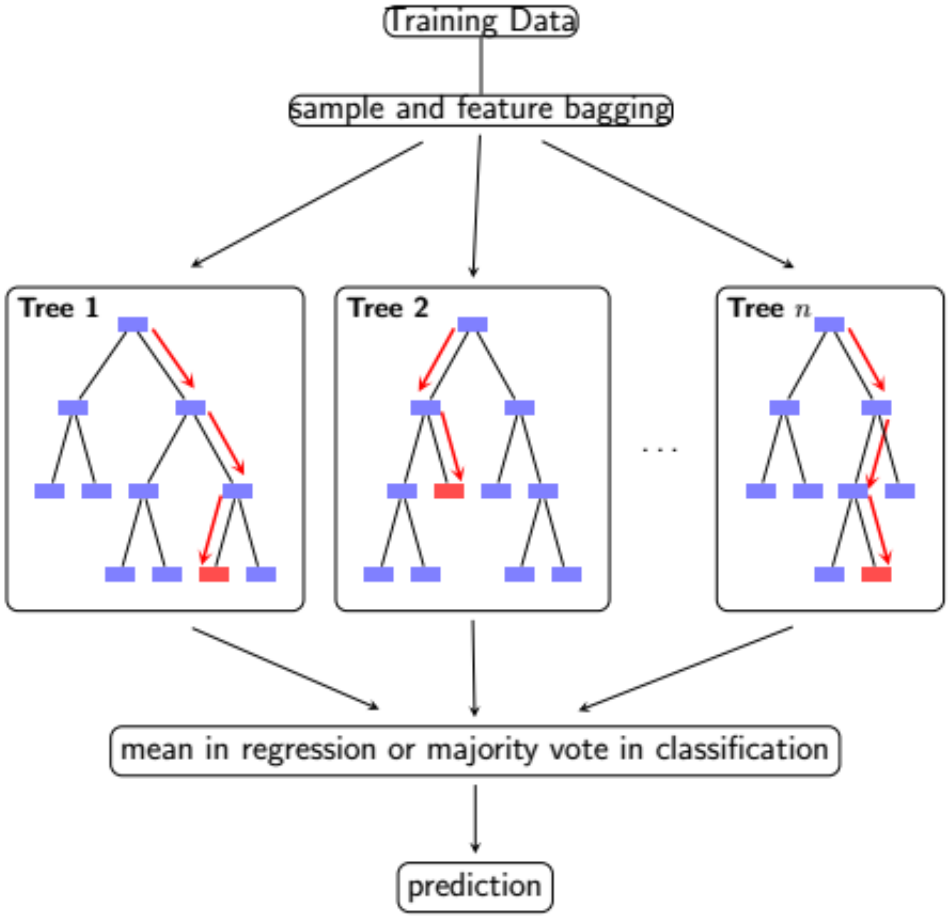


Figure 3.2: Random Forest Model

3.3 Naive Bayes Classifier

Similar to the previous algorithm, Naive Bayes can also work for both classification and regression analysis. This ML algorithm is based on Bayesian Theorem which assumes that predictors are independent. To simply put it, the classifier assumes that the features are unrelated with one another and could work independently. Given a large data set, this model could be extremely useful as it can perform better than some highly sophisticated models and additionally it even easy to build.[5] Bayes theorem provides a way of calculating probability as the form of equation given below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(c)} \tag{3.5}$$

Where,

- $P(c|x)$: Probability of class(c = target) given predictor (x = attributes)
- $P(c)$: Initial probability of class.
- $P(x|c)$: Probability of predictor given class.
- $P(x)$: Initial probability of predictor.

3.4 Artificial Neural Network

Artificial Neural Network (ANN) mirrors the structure of biological neural networks of the human brain on a computational basis. Much likely to human brain structure that consisted interconnected neurons, ANN too have neurons known as nodes that are internally linked with each other on various layers of the network. Ann construction includes an input layer, one or several hidden layer and as output layer. However, the learning of ANN model could be supervised, unsupervised and reinforcement strategy.[6]

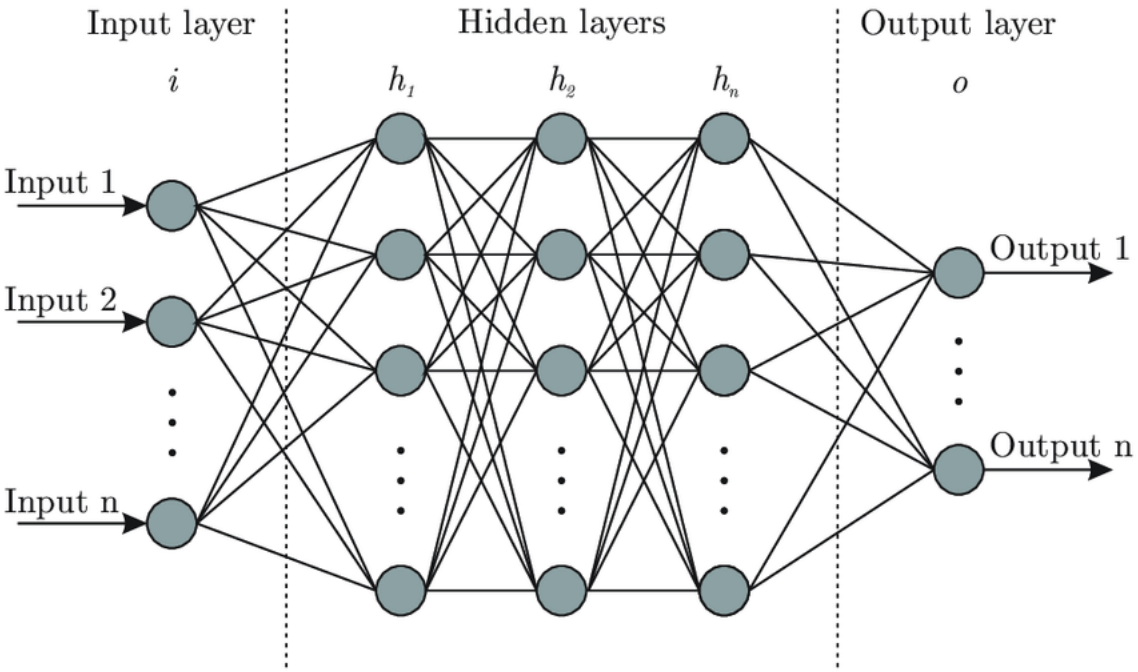


Figure 3.3: Artificial Neural Network Architecture

In the model, predictor variables are being fed and input layer, while the output layer gives the prediction value. In the middle section hidden layer/s work as an nonlinear transformation between the input layer and the output layer. For the learning, the popular method is backpropagation. At every neuron, output are calculated using Equation (3.6).

$$y_k(i) = \theta * [\sum_{j=1}^{n} x_j(i) * w_{jk}(i) - th_k] \tag{3.6}$$

Here, $y_k(i)$ is the output of neuron k in the i^{th} iteration. x_1,x_2,...,x_n are the inputs from the previous layer. $w_{jk}(i)$ is the weight between input x_j and output neuron y_k, th_k is the threshold and θ is the activation function.[6]

Chapter 4

Research Methodologies

4.1 Data

For the purpose of our research we used data-set from 1991 to 2013, which was collected from the Department of Disaster Management (DDM). This data-set contains water levels and weather variables which has the potential to cause flood. Table 4.1 shows all attributes of the data-set and descriptive statistics (Minimum, Maximum, Mean and Standard Deviation) of their value.

Attributes	**Min**	**Ma**	**Mean**	**Standard Deviation**
Maximum Temperature	27.5	40.5	34.4576	2.7450
Minimum Temperature	10.8	27.7	21.7157	4.9602
Average Temperature	19.3	33.25	28.0867	3.6971
Relative Humidity	66.0	92.0	80.6108	5.7077
Wind Speed	0.1	3.4	1.2424	0.7165
Cloud Coverage	0.2	6.8	3.3084	1.7004
Maximum Water level	1.6	4.75	3.2187	0.5760
Minimum Water level	-2.26	2.36	-0.4997	0.5668
Average Water level	0.04	3.57	1.3352	0.5687
Rainfall	0.0	846.0	145.6995	157.9184

Table 4.1: Descriptive Statistics (Minimum, Maximum, Mean & Standard Deviation) of the Data

4.2 Project Work-flow

The work-flow of this project consists of using previous months weather and water level reports to develop a model that can predict the probability of flood and maximum water level for the next month. This is a supervised (both the features and the target are given), regression problem (the target is a continuous variable); given a set of data(x) with target(y) included. Even though the general structure of a machine learning project stays relatively constant, the exact implementation details can vary:

- Data Collection

- Data cleaning & formatting
- Exploratory Data Analysis
- Feature engineering & selection
- Establishing a baseline and compare several machine learning models
- Evaluate the best model on the testing set
- Interpret the model results to the extent possible

4.3 Data Pre-Processing

Initially, we collected two different data set from the Department of Disaster Management (DDM), one of them is water level data set and another data set is about weather. So firstly we had to merge both the data set. However, in the water level data set, there were few missing rows (few years data were absent), containing 203 rows, whereas the weather data set contains 276 rows. Following is the list of missing year and months from the water level data set.

Year	Month
1992	4-12
1993	1-3
1996	12
1997	3
2000	6,7
2001	2
2002	9
2003	6,11
2004	5-12
2005	1-12
2006	1-12
2007	1-12
2008	5-8
2009	6

Table 4.2: List of missing years and months

Keeping in mind that both the data set has uneven number of rows, therefore, while combining the data set we only merged those rows they both have in common. The merge was completed using Python that performed an inner merge to only combine the values in the column names that match both the data frames. Other than the few missing rows, there were no other missing values in the data set. As this is the real world data, we did not want to mess up the original value by further pre-processing step, and the machine learning model we used did not require any feature scaling as well. Following is the list of initial columns after performing the merge:
Finally, as per our research topic is to predict the upcoming flood, we needed to adjust the data set accordingly so that the model could read the previous months values and based on that be able to predict the next/current months possibility

Maximum Temperature	float64
Minimum Temperature	float64
Rainfall	int64
Relative Humidity	int64
Wind Speed	float64
Cloud Coverage	float64
Bright Sunlight	float64
Maximum Water level	float64
Minimum Water level	float64
Average Water level	float64

Table 4.3: List of initial columns in the combined data set

of flooding. Therefore, we up-shifted the value of the maximum water level and average water level by 1 and put them in a different column naming the next months maximum water level and next months average water level.

4.4 Feature Selection & Engineering

To figure out the relation between the attributes, scatter plot has been plotted for each features. Figure 4.1 shows the relations between the features and next months average water level and Figure 4.2 shows the relations with next months maximum water level.

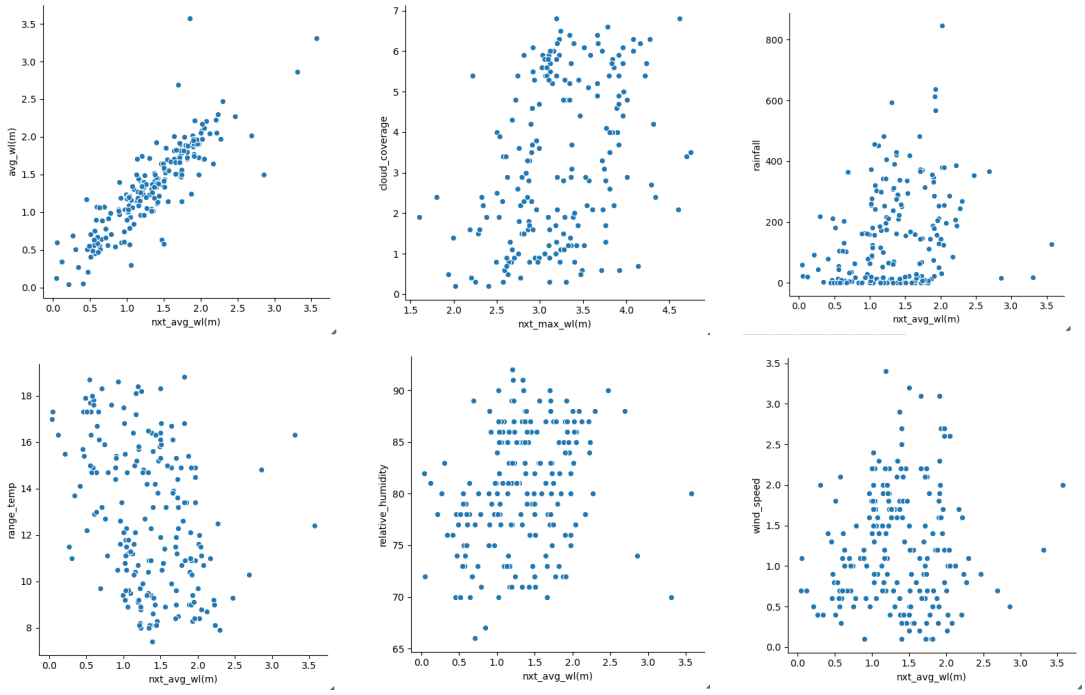


Figure 4.1: Scatter Plots between next months average water level (meter) and current months average water level (meter), cloud coverage, rainfall, range of temperature, relative humidity and wind speed (from left up to right bottom respectively)

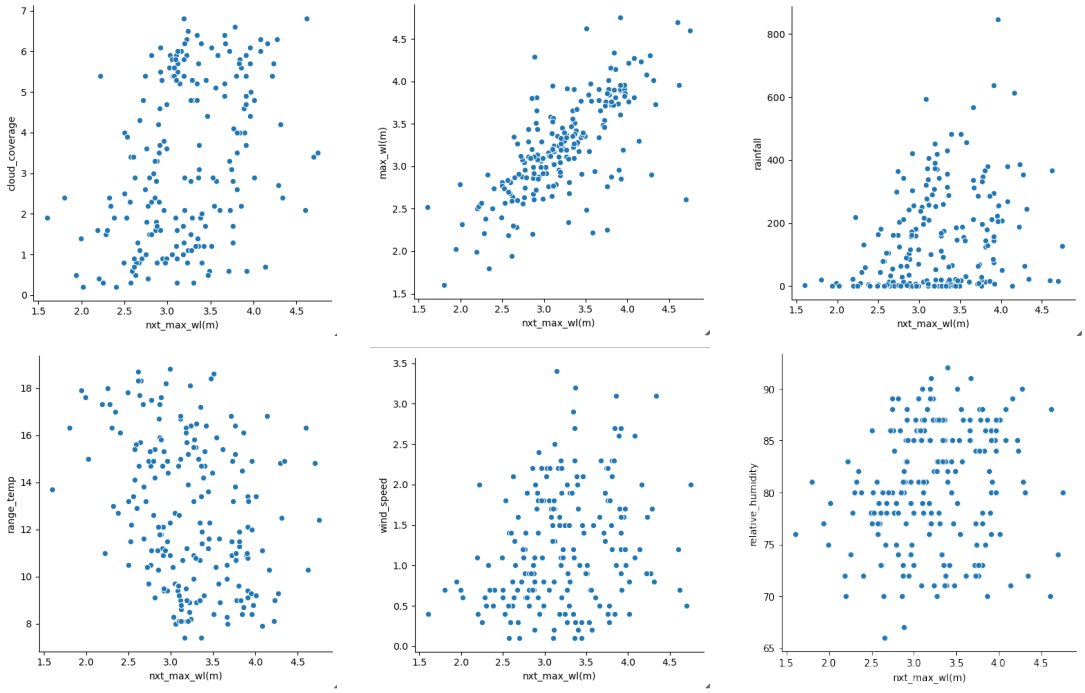


Figure 4.2: Scatter Plots between next months maximum water level and current months cloud coverage, maximum water level (meter), rainfall, range of temperature, wind speed and relative humidity (from left up to right bottom respectively)

- **Feature selection:** The process of choosing relevant features for the given data set based on the higher variance or other various factors. Features with most relevancy are chosen and the least relevant features are removed in order to make the model more accurate.
- **Feature engineering:** The process of deriving more features from the raw data to improve the machine learning algorithm work better. Forming new features allows the model to absorb a mapping of features with the target. However, feature engineering is purely based on intuition, knowing what the model want to predict and what kind of features could help the model.

From the figure 4.1 and 4.2, it could be seen that rainfall does not have a good correlation with the water level. So, not including rainfall in the model will not create any issues. Same goes for the bright sunlight feature, which does not have any relation with flood nonetheless. Therefore, rainfall and bright sunlight has been removed from the features. Additionally, a new binary feature has been added as the probability of flood, where if the maximum water level is higher than 3m it takes 1 or else 0. Additionally, we included few more columns as the average of monthly temperature and water level, as well as the range between the minimum and maximum temperature.
After carrying out Feature Engineering we needed to remove multi-linearity, such as finding features which are highly col-linear due to some underlying similarity, hence keeping them is redundant. Features with col-linearity between them above a certain threshold was removed from the feature set. Heat-map illustrating the correlations between the features were generated to find out multi-linearity or col-linearity.

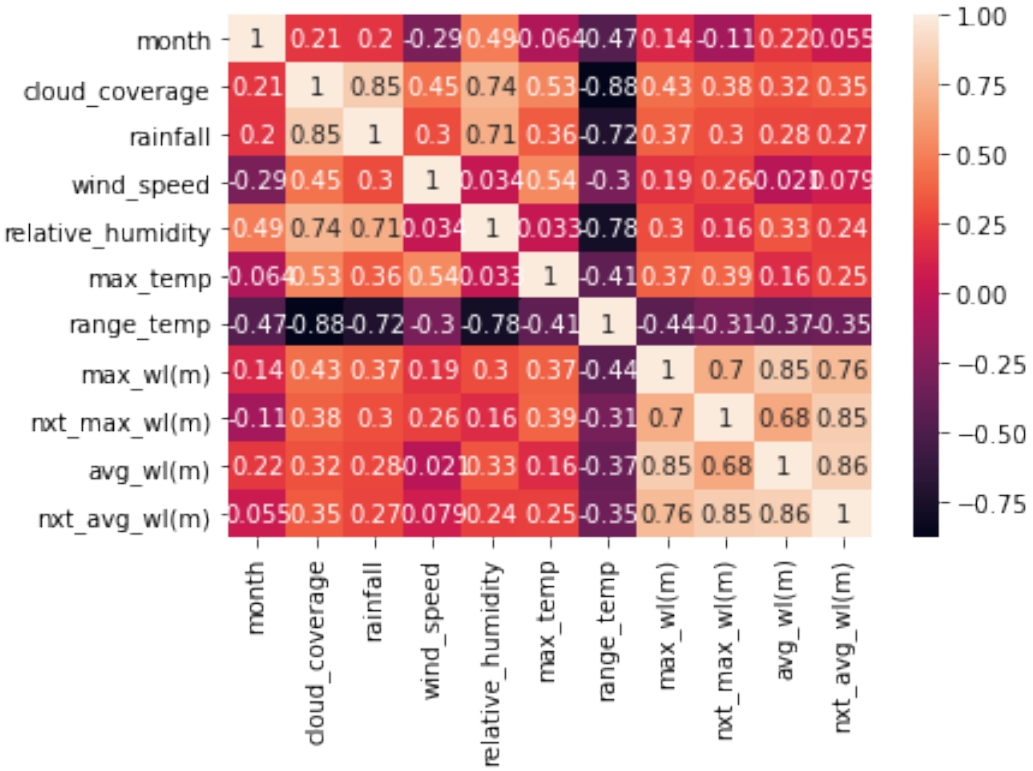


Figure 4.3: Heat-map of Numeric Features

However, after conducting the iterative process of feature selection and feature engineering we ended up with the following sets of features:

Maximum Temperature	float64
Minimum Temperature	float64
Range of Temperature	float64
Average Temperature	float64
Relative Humidity	int64
Wind Speed	float64
Cloud Coverage	float64
Maximum Water level	float64
Minimum Water level	float64
Average Water level	float64
Probability	int64

Table 4.4: List of initial columns in the combined data set

4.5 Train-Test Split

Heat-map of Numeric FeaturesFirstly, the features(X) and the target(Y = "Next months maximum water level" for LR and RF; Y = "Probability" for ANN and NB) were separated and then the train-test split were carried out. The ratio between Test set and Training set is 30% and 70%. Train set would be used to train the

model. And accuracy evaluation would be done using the test set, which is unseen data for the trained model.

Chapter 5

Model Implementation and Optimization

5.1 Work Flow Overview

Our primary goal is to find the perfect model for the data set and fit the training set as best as possible. Minimizing the true error on the test set and avoid over fitting the data. However, to work with the real world data with less variances, our main concern was to deliver as much accuracy as possible. In Figure 5.1 is an overview of our workflow. The details are given below:

- Collecting the real world data and processing the raw data to fit the model.
- Detailed Exploratory Data Analysis of the data set to figure out the correlation and variances between the features.
- Performing feature selection & feature engineering multiple times. Feature selection was based on the correlation and variances.Features with least correlation were being removed. Feature engineering was conducted by adding multiple variables extracting from the raw data to help out with the model.
- Train test split was carried out by 70:30 ratio. Train set were used to train the model and test set were held out to final evaluation of the optimized model.
- Machine learning algorithms were being fed the training set to build the model.
- Cross validation were performed to optimize the model along with the tuning of the features recursively.
- Final model evaluated on hold out test set. Results obtained and analyzed. Comparative analysis conducted between the models and the final model.

However, some of the steps were done repetitively. Going back and forth between steps to find out the best features to build the maximum optimized model.

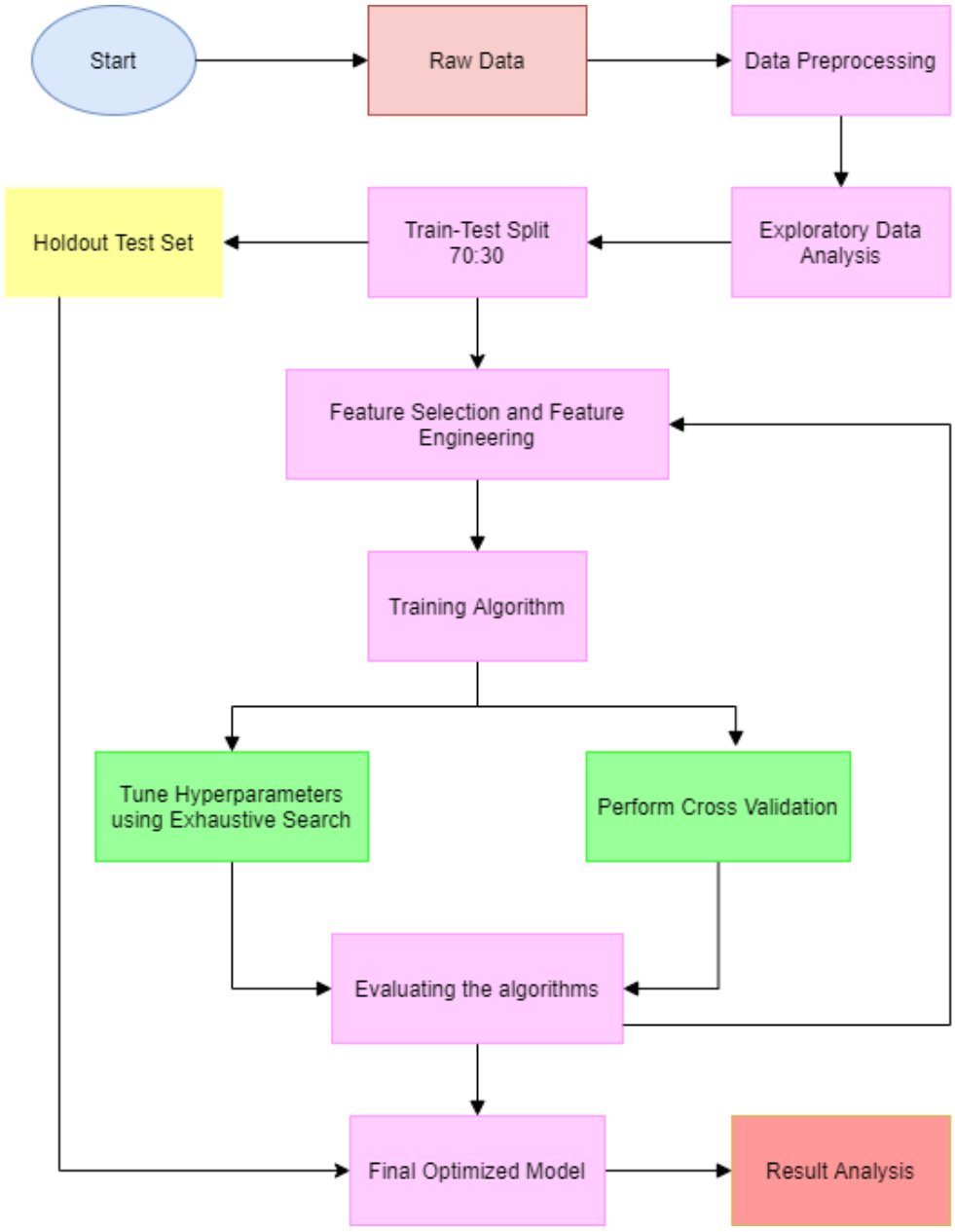


Figure 5.1: Work flow diagram

5.2 Model Implementation

Model implementation for this research was quite a bit of challenge as we worked on real world data. The data set were not modified for the use of machine learning algorithm and we had to create the workable data set along the path of implementing our models per requirement. However, as we used four different algorithms, the implementation of each differed from the others. The process has been described below:

- **Linear Regression Model:** Initially we fed the data set prior to feature scaling and feature engineering, and we were getting very poor accuracy. However, after details data visualization and analyzing the relation between the features, we added few more variables to fed the model namely average temperature, average water level, range of temperature etc. The results improved but it consisted overfitting problem. After analysis we came to know that we are predicting the average water level but also fedding the maximum water level both of which are the same catergory. However, as we wanted to predict the flood prior to the time being we decided to upshift the column we wanted to predict by 1, and by doing so the model could be fed previous months data and be able to predict the upcoming months water level and thus the probability of the flood.

- **Random Forest:** Similar to the linear regression model, random forest had the overfitting issue initially. However, by shifting the predicted column upward, the problem fixed and we got the desired result.

- **Naive Bayes:** With our initial data set naive bayes could not be implemented. After some research, we added a column in the data set as the probability of flood. In the probability column the values were added as 1 and 0, where it depended on the maximum water level, if its greater than 3.05m than 1 or else 0. According to DDM, 3.05 meter was marked as the danger point of khulna for water level, therefore if the water level crosses that points it could interpret as flooding. Additionally, for the model to work we up-shifted the probability column as well so that it could predict the next months probability.

- **Artificial Neural Network:** Finally, for ANN the model were fed the final data set with next months probability as predicted variable. We added two hidden layers in the model and 6 units for the input layer. However, with more than 2 hidden layers the accuracy were getting lower (around 34%), same with the different units for each of the layers than we initially assigned. Therefore, we finalized the model with two hidden layers in it and 6 units for input layer and 16 units for hidden layers.

5.3 Evaluating and Comparing Machine Learning Models

For our research we build several machine learning models which were trained and evaluated for our supervised regression task and compared them with the baseline

as well as with each other. Four different machine learning models were trained and evaluated from the Sci-kit library. The models are as follows:

1. Linear Regression
2. Random Forest
3. Naive Bayes Theorem
4. Artificial Neural Networks

The models are being compared using their Mean Absolute Error (MAE) on the test set. A baseline model using the mean value of the predictor class was scored as 45.19.

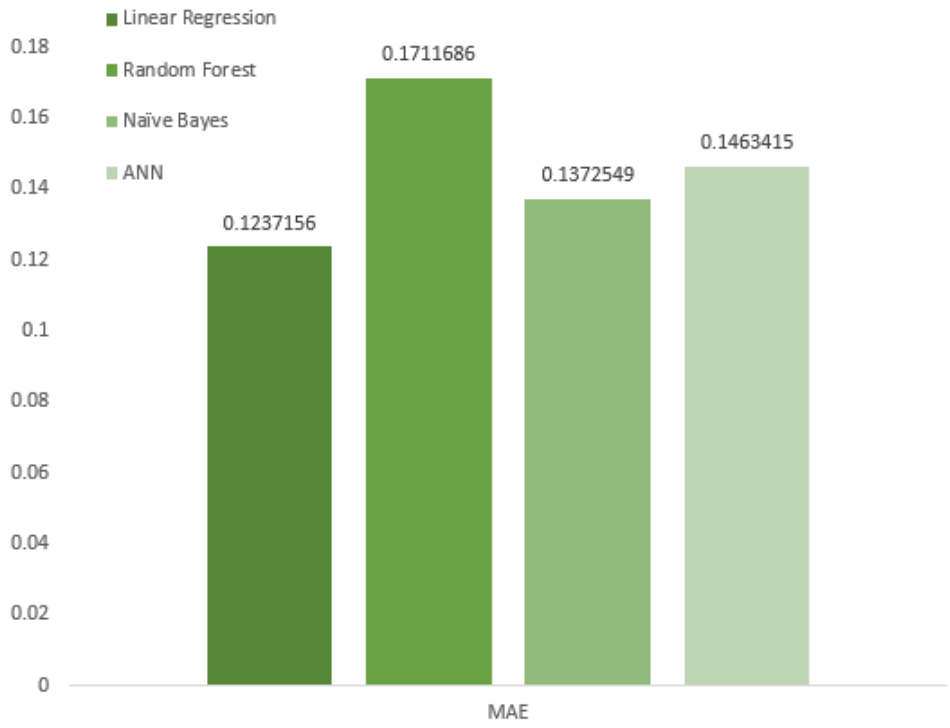


Figure 5.2: Mean Absolute Error on test set

From the Figure 5.2, it can be shown that all of the models performed significantly better than the baseline MAE. Although, it is not a very good comparison given that the model is trained on real-life data, but from the errors it can be inferred as the model is performing well. Additionally, the baseline model RMSE is scored as 58.35, which can be compared with the models RMSE given in the Figure 5.3. It clearly shows that the models are outperforming the baseline RMSE as well.
From Figure 5.2 and Figure 5.3, it can be observed that all models performing exceptionally better than the baseline model. Again, if we compare the models with each other we can see that Linear Regression provides with the lesser error both in MAE and in RMSE. However, Naive Bayes performing better for MAE than Random Forest, but in RMSE Random Forest is outperforming NB.

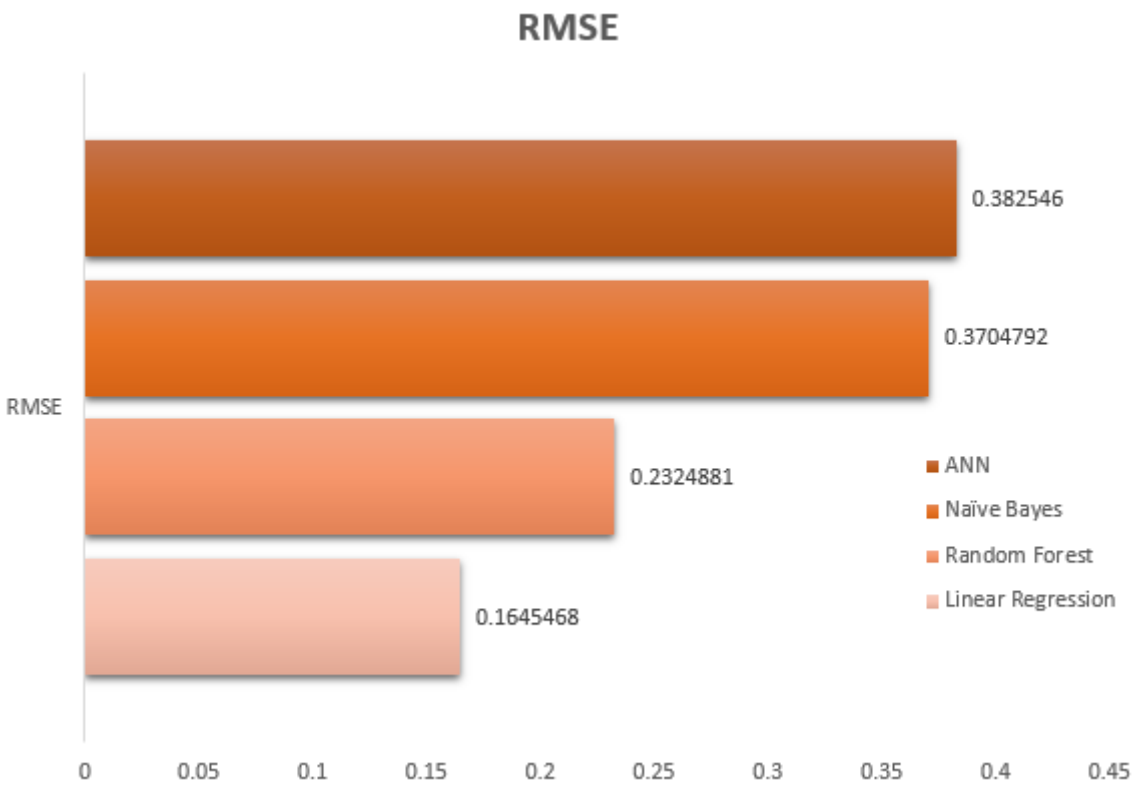


Figure 5.3: Root Mean Squared Error on test set

5.4 Model Optimization

Model optimization refers to finding the set of input parameters that could results in a maximum output for the model. It is important to optimize the model as it minimizes the cost function. However, it can be done by adjusting the hyperparameter. Hyperparameters are the settings for the machine learning algorithms and needs to be set before training the model, i.e. the number of trees in the random forest. Although, there are many approach to tune the hyperparameters, we used the exhaustive search and cross validation for our model.

- Exhaustive Search: This is similar to brute-force search, as the process is to look for the best optimal match by checking all the combinations. It is a simple method and can be performed manually. Generally, exhaustive search performed better when we have limited knowledge of the best model hyperparameters and we can use random search to narrow down the options.

- Cross Validation: This is a statistical technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance. This can assess the performances of the hyperparameters. In our research we used 5-fold cross validation, where the data was randomly divided into 5 fold and the CV is performed for each fold. At the end of the CV, the errors were measured by averaging the values.

After performing cross validation and exhaustive search we had to perform feature engineering few more times. However, in cross validation Linear Regression per-

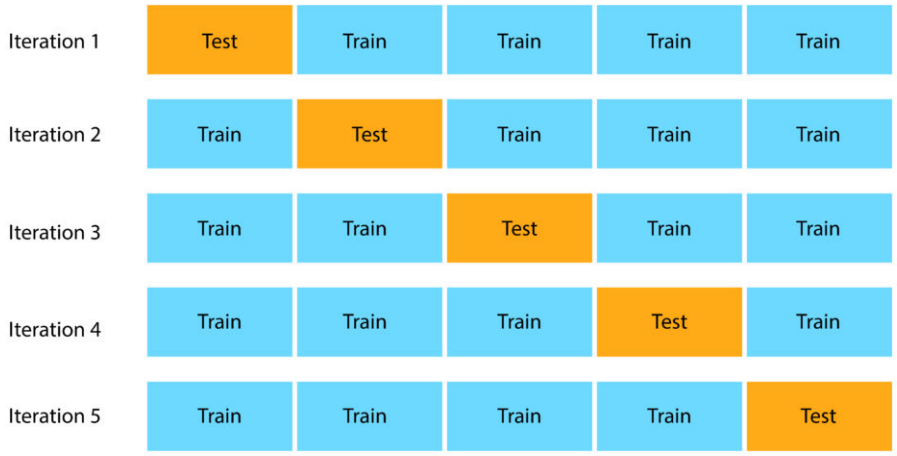


Figure 5.4: Cross-Validation Overview

formed better than other models. It was also found that Linear Regression gives lesser error on both MAE and RMSE.

Chapter 6

Experimental Results and Analysis

6.1 Comparative Analysis of Supervised Models

Initially, we fit our training data to few existing models used in literature are:

- Linear Regression
- Random Forest Regressor
- Naive Bayesian Theorem
- Artificial Neural Networks

Initially, before tuning and feature engineering the scores were pretty low for all the models. Given is the accuracy on the training set before feature engineering and cross validation.

Models	**MSE**	**RMSE**	**MAE**	**Accuracy**
Linear Regression	23.73004	48.71348	41.76449	16.78%
Random Forest Regressor	23.86267	48.84943	36.39411	88.59%
Naive Bayes Theorem	37.25490	61.03679	37.25490	67.55%

Table 6.1: Comparison of MSE, RMSE, MAE, R Squared between models

After conducting the Cross Validation and tuning our parameters for all the models the results improved by a greater margin.

Models	**MSE**	**RMSE**	**MAE**	**R Squared**
Linear Regression	2.70756	16.45468	12.37156	86.12021
Random Forest Regressor	5.40507	23.24881	17.11686	80.69225
Naive Bayes Theorem	13.72549	37.04792	13.72549	42.41935
Artificial Neural Network	14.63414	38.25460	14.63414	34.92063

Table 6.2: Comparison of MSE, RMSE, MAE, R Squared between models

However, if we compare our models with the baseline model we can clearly see from the Figure 6.1 that every model outperforms the baseline model with a greater margin. Although all the model performs way better than the baseline model, LR gives the least error from all the other models, leaving NB a close second in terms of errors.

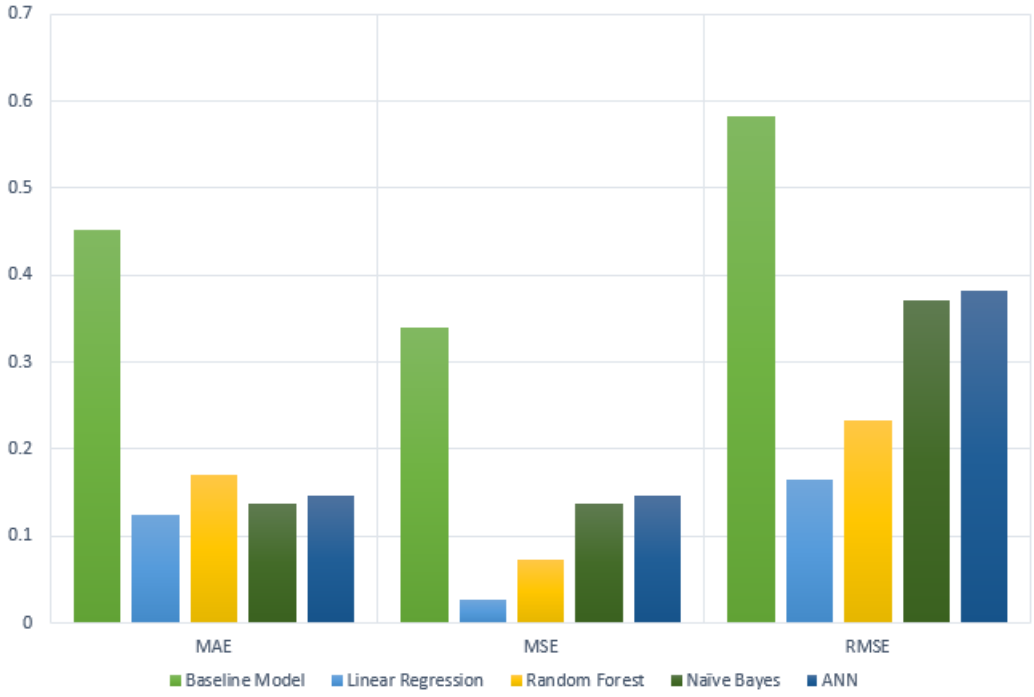


Figure 6.1: Comparison with baseline model

6.2 Final Model Evaluation

The final optimized model was fit into training set and was evaluated on the test. The model outperformed all the previous models in terms of MAE, RMSE, MSE and R Squared.

Set	Models	Accuracy
Training Set	Linear Regression	76.09%
	Random Forest	95.89%
	Naive Bayes	75.49%
	Artificial Neural Network	85.36%
Test Set	Linear Regression	86.12%
	Random Forest	80.69%
	Naive Bayes	86.27%

Given that the models were trained and tested on real-world data, the accuracy were expected low but the models worked pretty good. As, the relation between features were not very col-linear and the variances were not that high. However, from the table above, we can see that Random Forest outperforms all the other models in the training set, while Naive Bayes and Linear Regression works better on test set.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This paper has derived a machine learning based model to predict flood susceptibility using district level data. We were able to provide an illustration of the importance of flood damage prediction and give an analysis. Most of the studies in the literature worked with the previous years flood susceptibility and the damages caused by flood. Our paper differs as it tries a prediction for the upcoming flood and in our better knowledge this is the first work. However, in our study we observed that while Random Forest worked better on training data set, but Linear Regression and Naive Bayes outperformed others on the test set. Although some focused on the damage caused by flood every year and to predict the damage so the authority and people can take precautions to face it. This field of inquiry is very important as its main concern with helping the people at the end and reduce the destruction by taking precautions. For future improvement these researches can be used for flood preparation. However, we aim to provide people with actual number of loss the country is going to face in the upcoming disasters for they could take the necessary preparation to lessen the loss.

7.2 Future Work

However, this study has some limitations. Estimation of flood susceptibility in a division was determined. The inability to obtain more division's data, made us unable to display the flood susceptibility for the other areas of Bangladesh. Although, the data of Khulna division provided enough information that were necessary for flood estimation in that area. Simultaneously, we would try to collect more data of other divisions of Bangladesh, so that we can test our model on a large scale. As a consequence, we would be able to find out more about the chances of heavy rainfall in all the vulnerable and most affected divisions. This would provide us a much more accurate data. That can be a great for the development of any strategies by the authorities. Hence, we would like to adjourn with a statement that this paper provides a clear demonstration of a gripping approach in predicting an estimation of flood susceptibility. It can a play a valuable role as every year the floods are becoming more stronger as a consequence of global warming. It will be able to generate a reliable estimation of floods based on the results, the authority can come up with a effective strategy.

Bibliography

[1] Reuters, *Bangladesh floods recede, but death toll rises - bangladesh*, Sep. 1998. [Online]. Available: https://reliefweb.int/report/bangladesh/bangladesh-floods-recede-death-toll-rises.

[2] "World: South asia bangladesh floods rise again," *BBC News*, Aug. 1998. [Online]. Available: http://news.bbc.co.uk/2/hi/south_asia/157254.stm.

[3] D. Dutta, S. Herath, and K. Musiake, "A mathematical model for flood loss estimation," *Journal of hydrology*, vol. 277, no. 1-2, pp. 24–49, 2003.

[4] A. Haque and S. Jahan, "Impact of flood disasters in bangladesh: A multi-sector regional analysis," *International Journal of Disaster Risk Reduction*, vol. 13, pp. 266–275, Sep. 2015. DOI: 10.1016/j.ijdrr.2015.07.001.

[5] S. Ray, *Learn naive bayes algorithm: Naive bayes classifier examples*, Sep. 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.

[6] F. Bre, J. M. Gimenez, and V. D. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings*, vol. 158, pp. 1429–1441, Nov. 2017. DOI: 10.1016/j.enbuild.2017.11.045.

[7] M. C. Montoya, *Emergency flood damage rehabilitation in bangladesh*, Jul. 2017. [Online]. Available: https://www.adb.org/publications/emergency-flood-damage-rehabilitation-bangladesh.

[8] A. C. P. Start Network, *Bangladesh flooding disaster summary sheet - 20 may 2018 - bangladesh*, May 2018. [Online]. Available: https://reliefweb.int/report/bangladesh/bangladesh-flooding-disaster-summary-sheet-20-may-2018.

[9] K. K. Ganguly, N. Nahar, and B. M. Hossain, "A machine learning-based prediction and analysis of flood affected households: A case study of floods in bangladesh," *International journal of disaster risk reduction*, vol. 34, pp. 283–294, 2019.

[10] M. Rahman, C. Ningsheng, M. M. Islam, A. Dewan, J. Iqbal, R. M. A. Washakh, and T. Shufeng, "Flood susceptibility assessment in bangladesh using machine learning and multi-criteria decision analysis," *Earth Systems and Environment*, vol. 3, no. 3, pp. 585–601, 2019.

[11] T. Yiu, *Understanding random forest*, Aug. 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[12] E. Dodangeh, B. Choubin, A. N. Eigdir, N. Nabipour, M. Panahi, S. Shamshirband, and A. Mosavi, "Integrated machine learning methods with resampling algorithms for flood susceptibility prediction," *Science of the Total Environment*, vol. 705, p. 135 983, 2020.

[13] R. Goel *et al.*, "Flood damage analysis using machine learning techniques," *Procedia Computer Science*, vol. 173, pp. 78–85, 2020.

[14] A. Manandhar, A. Fischer, D. J. Bradley, M. Salehin, M. S. Islam, R. Hope, and D. A. Clifton, "Machine learning to evaluate impacts of flood protection in bangladesh, 1983–2014," *Water*, vol. 12, no. 2, p. 483, 2020.

[15] S. Talukdar, B. Ghose, R. Salam, S. Mahato, Q. B. Pham, N. T. T. Linh, R. Costache, M. Avand, *et al.*, "Flood susceptibility modeling in teesta river basin, bangladesh using novel ensembles of bagging algorithms," *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 12, pp. 2277–2300, 2020.

[16] T. Tang, Z. Liang, Y. Hu, B. Li, and J. Wang, "Research on flood forecasting based on flood hydrograph generalization and random forest in qiushui river basin, china," *Journal of Hydroinformatics*, vol. 22, no. 6, pp. 1588–1602, 2020.

[17] *Floods in bangladesh*, May 2021. [Online]. Available: https://en.wikipedia.org/wiki/Floods_in_Bangladesh.