

A Machine Learning Approach to Analyze and Predict
Rainfall
in Different Regions of Bangladesh

by

Arnob Rahee

18101225

Md. Montasir Nafiz

17109028

Sania Azhmee Bhuiyan

18101486

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
August 2021

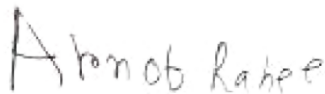
© 2021. Brac University
All rights reserved.

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:



Arnob Rahee
18101225



Sania Azhmee Bhuiyan
18101486



Md. Montasir Nafiz
17109028

Approval

The thesis/project titled “A Machine Learning Approach to Analyze and Predict Rainfall in Different Regions of Bangladesh” submitted by

1. Arnob Rahee (18101225)
2. Sania Azhmee Bhuiyan (18101486)
3. Md. Montasir Nafiz (17109028)

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 06, 2021.

Examining Committee:

Supervisor:
(Member)



Md. Ashraful Alam, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Co- Supervisor:
(Member)

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

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
Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Ethics Statement

Abstract

Rainfall has always been important in context of Bangladesh as almost 43% of the population depends on agriculture for their livelihood. Global warming has been taking a toll on environment and rainfall patterns have been changing around the world. Almost half the population depends on rainfall for irrigating their lands and grow crops. If rainfall can be predicted precisely then people involved with agricultural sector will be benefited. In this research, we analyzed the rainfall statistics on the basis of Bangladesh Meteorological Department's data of rainfall of last 66 years. With Mann-Kendall Trend Test with 5% level of significance we tested the trend of 6 divisional stations of Bangladesh. Later we utilized three regression models to predict rainfall on basis of data from 1948 to 2014. We have also implemented those 3 regression models on 6 regional station data to understand if there is any change in accuracy. Trend tests showed no significant change in rainfall patterns in last 30 years. We also broke down the data to understand the hydrological regions of Bangladesh and the rainfall by stations.

Keywords: Rainfall Analysis; Machine Learning; Rainfall in Bangladesh; Regression; K-Nearest Neighbour; Random Forest; Decision Tree

Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

At first, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

We would like to thank to our advisor Md. Ashrafal Alam sir for his kind support and advice in our work. He helped us whenever we needed help. He has been nothing but pragmatic and helping whenever we had any problems

We would also like to thank our co-advisor Md. Golam Rabiul Alam sir for his help whenever we needed it.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Dedication	vi
Acknowledgment	vii
Table of Contents	viii
List of Figures	xi
List of Tables	xiii
Nomenclature	xiv
1 Introduction	1
1.1 Rainfall in Bangladesh	2
1.2 Problem Statement	6
1.3 Aims and Objectives	6
1.4 Overview	6
2 Literature Review	7
2.1 Random Forest	7
2.2 K-Nearest Neighbour	8
2.3 Decision Tree	9
2.4 Mann-Kendall Trend Test	11
2.5 Related Work	12
3 Dataset Analysis	13
3.1 Dataset	16
3.2 Data Preprocessing	17
3.3 Data Visualization	18
3.3.1 Highest Rainfall by Months	18
3.3.2 Highest Rainfall By Years	19
3.3.3 Highest rainfall by Stations	21

3.4	Trend Analysis	22
4	Methods and Implimentation	26
4.1	Dataset Description	26
4.2	Steps in creating a machine learning model:	27
4.2.1	Matplotlib:	29
4.2.2	Developing a Rainfall Prediction Model	29
4.3	K-Nearest Neighbor	29
4.3.1	How it works	29
4.3.2	Pros and Cons of Priori:	29
4.3.3	Limitations of KNN	30
4.3.4	Optimization of KNN	30
4.4	Implementation of KNN	31
4.5	Implementing KNN on Station Data	33
4.5.1	KNN on Rainfall data of Dhaka	33
4.5.2	KNN on Rainfall data of Chittagong	33
4.5.3	KNN on Rainfall data of Sylhet	34
4.5.4	KNN on Rainfall data of Barisal	34
4.5.5	KNN on Rainfall data of Rajshahi	35
4.5.6	KNN on Rainfall data of Khulna	35
4.6	Random Forest	35
4.6.1	How Random Forest Works	36
4.6.2	Limitations of Random Forest	36
4.6.3	Optimization of Random Forest	36
4.7	Implementation of Random Forest	37
4.8	Implementing RF on Station Data	38
4.8.1	Random Forest on Rainfall data of Dhaka	39
4.8.2	Random Forest on Rainfall data of Chittagong	39
4.8.3	Random Forest on Rainfall data of Sylhet	40
4.8.4	Random Forest on Rainfall data of Barisal	40
4.8.5	Random Forest on Rainfall data of Rajshahi	41
4.8.6	Random Forest on Rainfall data of Khulna	41
4.9	Decision Tree	42
4.9.1	How it works	42
4.9.2	Limitation of Decision Tree	42
4.9.3	Optimization	43
4.10	Implementation of Decision Tree	43
4.11	Implementing Decision Tree on Station Data	44
4.11.1	Decision Tree on Rainfall data of Dhaka	44
4.11.2	Decision Tree on Rainfall data of Chittagong	45
4.11.3	Decision Tree on Rainfall data of Sylhet	45
4.11.4	Decision Tree on Rainfall data of Barisal	46
4.11.5	Decision Tree on Rainfall data of Rajshahi	46
4.11.6	Decision Tree on Rainfall data of Khulna	47
4.12	Comparison between Models	47

5	Conclusion	53
5.1	Conclusion	53
5.2	Limitations	54
5.3	Future Work	54
	Bibliography	56

List of Figures

1.1	Rainfall Map of Bangladesh (Mean Annual)[23]	4
1.2	Hydro logical zones of Bangladesh[16]	5
2.1	Structure of Random Forest	8
2.2	KNN Model[22]	9
2.3	Weather Prediction Using Decision Tree[20]	10
3.1	Location of BMD Stations	14
3.2	Location of the stations with mean annual rainfall[4]	15
3.3	Unmodified Dataset	16
3.4	Modified Dataset	17
3.5	Separated Dataset	18
3.6	Highest Rainfall by Months	19
3.7	Treemap of Highest Rainfall by Years	20
3.8	Rainfall in 1988-2008 period	20
3.9	Highest raifall in stations by year	21
3.10	Highest raifall in stations by year	21
3.11	Lowest rainfall by station	22
3.12	Rainfall Trend Dhaka	23
3.13	Rainfall Trend Chittagong	23
3.14	Rainfall Trend Sylhet	24
3.15	Rainfall Trend Barisal	24
3.16	Rainfall Trend Rajshahi	25
3.17	Rainfall Trend Khulna	25
4.1	Steps of Creating a ML Model	28
4.2	KNN Prediction Results	32
4.3	KNN on Rainfall data of Dhaka	33
4.4	KNN on Rainfall data of Chittagong	33
4.5	KNN on Rainfall data of Sylhet	34
4.6	KNN on Rainfall data of Barisal	34
4.7	KNN on Rainfall data of Rajshahi	35
4.8	KNN on Rainfall data of Khulna	35
4.9	Random Forest Regression Results	38
4.10	Random Forest on Rainfall data of Dhaka	39
4.11	Random Forest on Rainfall data of Chittagong	39
4.12	Random Forest on Rainfall data of Sylhet	40
4.13	Random Forest on Rainfall data of Barisal	40
4.14	Random Forest on Rainfall data of Rajshahi	41

4.15	Random Forest on Rainfall data of Khulna	41
4.16	Decision Tree Results	44
4.17	Decision Tree on Rainfall data of Dhaka	44
4.18	Decision Tree on Rainfall data of Chittagong	45
4.19	Decision Tree on Rainfall data of Sylhet	45
4.20	Decision Tree on Rainfall data of Barisal	46
4.21	Decision Tree on Rainfall data of Rajshahi	46
4.22	Decision Tree on Rainfall data of Khulna	47
4.23	R-square value comparison Dhaka	49
4.24	R-square value comparison Chittagong	49
4.25	R-square value comparison Sylhet	50
4.26	R-square value comparison Barisal	50
4.27	R-square value comparison Rajshahi	51
4.28	R-square value comparison Khulna	51

List of Tables

3.1	Mann-Kendall Test Results	22
4.1	R-square value comparison of 3 models	48

Nomenclature

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

BMD Bangladesh Meteorological Department

KNN K-Nearest Neighbor

MKTest Mann-Kendall Test

RF Random Forest

WKNN Weighted K-Nearest Neighbors

Chapter 1

Introduction

Bangladesh is situated in sub-tropical climate where rainfall is a regular natural phenomena. Rainfall is very important in the context of Bangladesh as it sits on a low Bengal Delta where 700 rivers of a total length of 24,140 km. run through the country[18]. Because of geographical position a significant amount of population depends on agriculture as their primary source of employment. As of 2017, 42.7 percent of the total population in Bangladesh depends on agriculture for their livelihoods[19]. Agriculture contributes 14.2 percent of the total GDP of Bangladesh[19]. For Agriculture, rainfall is one of the most important aspects. It is important for irrigation and growth of the agricultural products. Every year at the monsoon season some parts of the Bangladesh experience flood. For a country where a significant amount of people depend on rain for their livelihood, rainfall prediction is necessity. Bangladesh's GDP is largely dependent on rain. The variability in rainfall effects the country as a whole. The monsoon season stretches from June to September. During this time 75 percent to 80 percent of the rain occurs. The rainfall in the monsoon season affects our economy in various sectors. It affects the crop harvest, investment in agriculture and people's migration to cities. Different regions grow different crop according to the rainfall in those regions. Excess rainfall causes flood in North-eastern regions whereas sometimes lack of sufficient rainfall causes crops to die of heat. The prediction of rainfall depends on how the trend of rainfall is, the predictor-prediction relationship, inter correlation between the predictor and variables that causes rainfall. It is known that the rainfall patterns are getting disturbed because of global warming. For a partly agriculture based country it is an alarming news. Before predicting the rainfall, we need to understand if the pattern of rainfall is changing. The implications of changing environment because of global warming can be significant for Bangladesh where hydrological disasters are very common phenomena. Global warming has significantly changed the seasonal cycles in different regions, change in precipitation has been induced in several regions. With time it is going to be more severe. Bangladesh is also getting a few signals with little changes in seasonal cycles, rising water level causing flood etc. We need to know if it is also changing the rainfall patterns. Machine learning been Along with these the threat of natural disasters are ever increasing and are becoming more and more deadly as rainfall patterns that had existed started to change causing huge floods in many regions and also creating huge droughts in some places as rainfall is drying up in some areas whereas increasing at a deadly rate in others. Thus accurate prediction of rainfall is vital for the very survival of Bangladesh because knowing how much

rain would fall in a given month or day would give its people adequate amount of time and give them the valuable means to collect resources that can be used to prevent natural disasters. Thus using the dataset that dates all the way back to 1948 till the year 2016 and the latest machine learning models we tried to derive accurate amount of rainfall a Station would experience in a given month. In this thesis paper we tried to accurately measure and predict the amount of rainfall a Station would experience in a given month using three models KNN, Random Forest Regression and Decision Tree. The difference between the predicted result and the actual amount of rainfall the fell was plotted on a graph. The trend of rainfall in major parts of the country was calculated and plotted on a graph. At first we had to pre-processed the data we had before feeding it to the machine learning models as the dataset we had obtained from Bangladesh weather forecast had lots of errors. Then we applied this dataset to each of our models and accuracy of each model was calculated Later we compared the results that we found of all the models and it was revealed that amongst all the model Random Forest Regression was the best and decision tree by far was not upto the mark. The map of the annual rainfall of Bangladesh was also produced that gave us a clearer picture as to how much rainfall each Station experiences in an year.

1.1 Rainfall in Bangladesh

Bangladesh is situated in sub-tropical monsoon climate. where there is typically a hot rainy season and a dry winter season. Bangladesh's rainfall pattern can be divided into 4 part, where there is cold dry season in December to February, pre-monsoon in March to May, monsoon season in June to September and post-monsoon season in October to November[18]. In general, Bangladesh has 6 seasons and is a tropical country. April to May is hot summer season, June to July is monsoon season, August to September is autumn season, October and November is post-autumn season, December to January is winter and February to March is spring season. With time the difference between seasons is fading and one season overlaps another sometimes. The temperature is increasing gradually and winters are getting colder and longer [10]. Dominantly there are three seasons in Bangladesh, hot and stormy summer season from March to May, rainy and wet season from June to October and cold dry season is November to February[18]. The pre-monsoon season comprises of high temperature and thunderstorms. The hottest time of the year is usually during April when the temperature reaches almost 35 degrees in some regions[21]. Northern part of Bangladesh is usually the hottest at that period. Bangladesh's summer is usually hot and humid. After April the temperature reduces a bit because of the cloud forming from Bay of Bengal region. Thunderstorms occur in this time period in different parts of country. Wind direction is not from any particular direction in this period specially during the starting of this period. Rainfall is not very consistent in this period. The hot weather causes formation of cloud during this period. Rainfall in this season is usually the thunderstorms that are very windy, 10 to 25 percent of the total rainfall in Bangladesh occurs during this period that is mostly caused by the storms. The monsoon season starts from June. Southern or South-Western wind blows towards Bangladesh during this period. This period in our country is

very humid and wet. Due to the tropical depression that enters from Bay of Bengal region, in this period the country experiences heavy rainfall and consistent rainfall through days. The rainfall in this period accounts for 70 to 85 percent of the rainfall of the total rainfall that occurs in Bangladesh [18]. This period is very important for the agricultural sector as rainfall helps farmers to irrigate their land and grow their products. Bangladesh being an agriculture dominated country monsoon season is the most important season of the seasons for people here. At some time, it rains without respite for days causing different parts of country flooding. North-Eastern and South-Eastern parts of the country is prone to flood in this period. In some regions monsoon consisting of haors, baors and beels the monsoon season is usually longer than the other parts, those parts remain under water for a significant amount of time due to heavy rainfall and floods from rivers in those regions. Post-monsoon and winter season is usually dry and cold and temperature drops below 20 degrees in most of the regions. North-eastern and North-western regions experience the lowest temperature while south-western regions are a little warmer because of being closer to Bay of Bengal. North-Eastern regions experience cold wind originated from Himalayas but temperature never goes down to negative. During January, the temperature in North-Eastern region goes down to 3 to 4 degrees which is the lowest in Bangladesh [20]. During the winter Bangladesh does not experience much rainfall. Usually it is the driest period of the time. Rainfall remains one of the most important aspects for Bangladesh as a significant amount of the population depends on the rainfall to irrigate their agricultural land. Bangladesh is situated in Bengal Delta because of which lots of rivers run through the land of Bangladesh. According to rainfall the regions of Bangladesh can be divided into different climate zones. Northern zone consists of Rajshahi and Rangpur which are the driest of the regions and experience higher temperature. Sylhet experiences the highest amount of rainfall in Bangladesh. Sylhet is situated just near the southern foothills of Himalayan range where the winds blow northeast and east, it receives highest rainfall of all regions. In Figure 1, the map shows, the region with average rainfall every year[23]. Bangladesh is sits on a low land where it is prone to floods during monsoon seasons. The south-western side of Bangladesh is largely jungle coastline and south-eastern side is comprised of hilly jungles. South-Western regions experience tropical cyclones originated from Bay of Bengal. 5 percent of the tropical cyclones in the whole world originates from Bay of Bengal region. Wind blow towards Bangladesh is usually divided into two branches, wind from Arab sea and wind from Bay of Bengal. The wind blow towards Himalaya region during summer in this region, the wind reaches Himalayas and make it rain over the region. Wind flow from south west flowing over country and Bay of Bengal influences south western monsoon in Bangladesh during monsoon season In figure 2, we can see the hydrological map provided by the NWRD,2014[16]

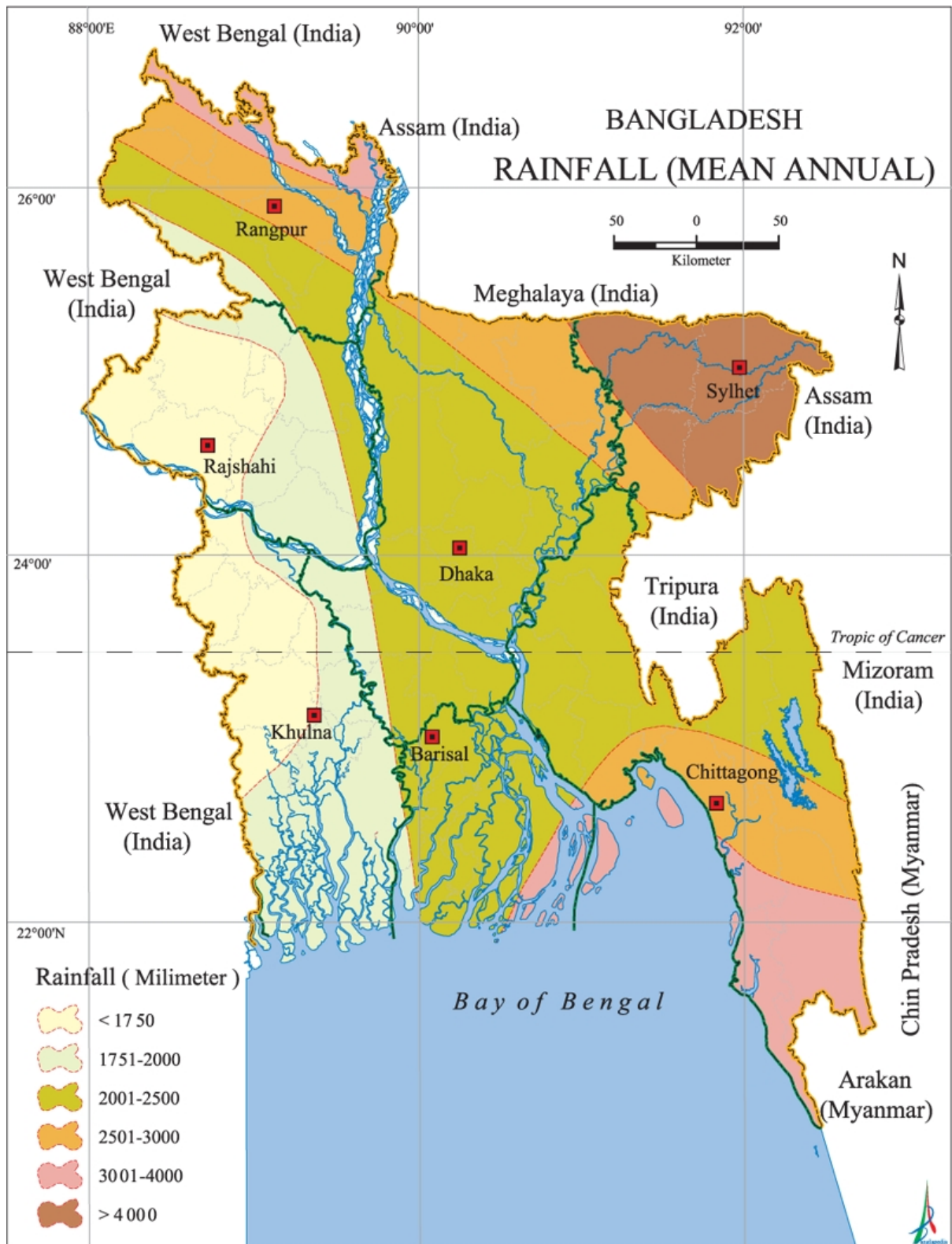


Figure 1.1: Rainfall Map of Bangladesh (Mean Annual)[23]

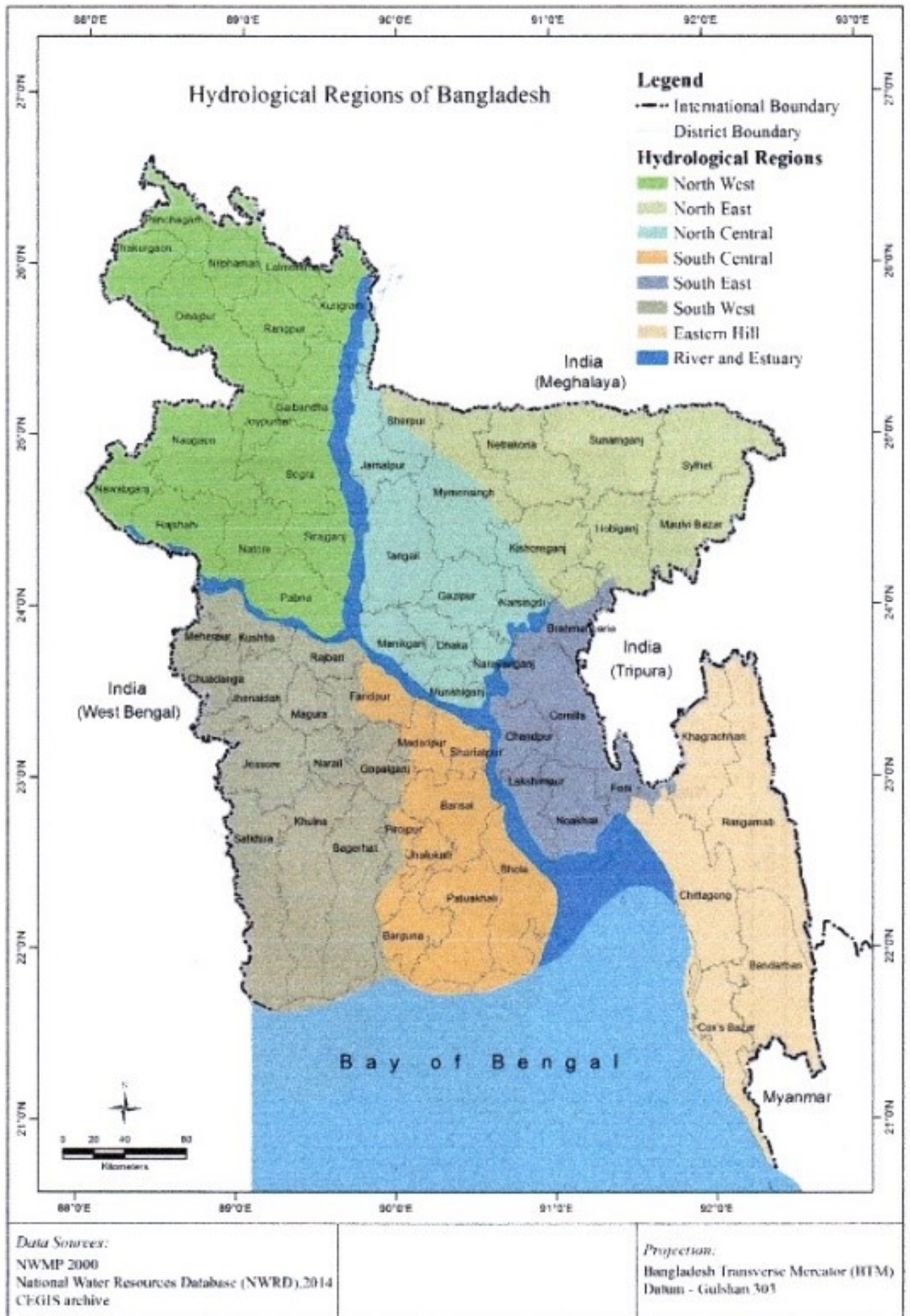


Figure 1.2: Hydro logical zones of Bangladesh[16]

1.2 Problem Statement

Rainfall in Bangladesh has been an important aspect for the economy and life of general people. Global warming has been changing precipitation patterns in different regions. It is important to understand if the rainfall patterns are changing in our region too. If the patterns are similar to then the prediction models can be developed with the historical data. Rainfall prediction has been necessity for years as Bangladesh faces a lots of different natural calamities which are related to the amount of rainfall that occurs in this region. If there is lack of rainfall cyclones may be developed in Bay of Bengal, if there is excess rainfall then north eastern regions might get flooded. Agricultural aspect is also a very relevant when we talk about rainfall prediction. If there is lack of rain irrigation of crops can be hampered, if there is excess rain then flood can ruin acres of land with crops. Rainfall prediction can help agricultural department to take steps according to the increase and decrease in rainfall. It may help saving the crops and the people in rural areas prone to flood.

1.3 Aims and Objectives

- To understand if there is any significant change of trends in Rainfall patterns by analyzing district station data
- To predict the rainfall in Bangladesh using KNN, Decision tree and Random forest regression techniques on (1948-2014) rainfall data to understand the accuracy of the models
- Comparing results obtained by using different techniques to understand which technique is more suitable

1.4 Overview

Chapter 1: In chapter 1, we discussed about the problem, gave an overview of rainfall in Bangladesh and Showed maps of rainfall zones in Bangladesh

Chapter 2: In chapter 2, we reviewed relevant papers on rainfall prediction and analysis, discussed the works done previously with the methods used

Chapter 3: In chapter 3, we discussed about the pre-processing of dataset, analyzed data for better understanding and finally analyzed the trend of rainfall in divisional stations

Chapter 4: In chapter 4, we discussed the use of methods, implemented methods on our dataset and analyzed the results

Chapter 5: In chapter 5, we conclude our research, discussed about the limitations and our plans for the future with this research

Chapter 2

Literature Review

This part mainly discusses the different methods that have been implied on the backdrop of research. It delves down into a finer detail about the previous studies scrutinizing the pros and cons of different methods, their limitation and extension. For both classification and regression analysis, many of these regression approaches are statistically proven to give better results. The most difficult part is feeding real-world time series data into these algorithms and training them. They should have all of the data set's relevant features. While there are numerous regressions approaches available. Despite their ability to perform admirably, many people fall short of their expectations when it comes to achieving the desired outcome.

2.1 Random Forest

The history of decision trees and how they're trained, as well as how a group of decision trees becomes a Random Forest. Many forecasters utilize regression trees to accurately anticipate load demand these days. They now employ regression trees to accurately anticipate load demand. As long as we cannot reach an acceptable result, the training data is separated into sub units of data in a decision tree. We can see that decision trees are widely used. Many decision-making areas are automated. Previously, the rules that constituted the human commands that altered the decision tree. However, due to the time commitment, the challenges became increasingly difficult as the procedure progressed as a method and the complexity of the problem. To address this issue, an automated rule extraction mechanism was developed. Hunt and his colleagues examine one of the earliest studies in the subject of automated decision tree generation[18]. Many studies have been undertaken in order to determine the criteria that should be considered while creating a decision tree that works with the least amount of mistake. Breiman et al. proposed the CART (Classification and Regression Trees) algorithm[17]. When it comes to creating decision trees, there are a few things to keep in mind. Load forecasting, as we all know, is a challenging problem to handle. Regressive decision trees are used because this is a quantitative regression problem. This is done by looking for the highest point in a sequence of statistics to predict future loads. If the prediction error does not decrease below the minimal minimum, the training dataset is split. The decision tree's root node holds all of the training data. The process of generating a decision tree starts at this node. We can see that the predicted value is the average of all the target values identified in the training data at this point.

The prediction error is calculated using the Mean Squared Error. The tree error is derived using the weighted average of errors at each node, as stated in the following equation, given N is the total number of samples in the tree.

$$MSE_{\text{tree}} = 1/N \sum_{\text{tree}} \sum_{D_i} (y_i - \hat{y}_i)^2 [15] \quad (2.1)$$

Below the figure shows the structure of Random Forest Model:

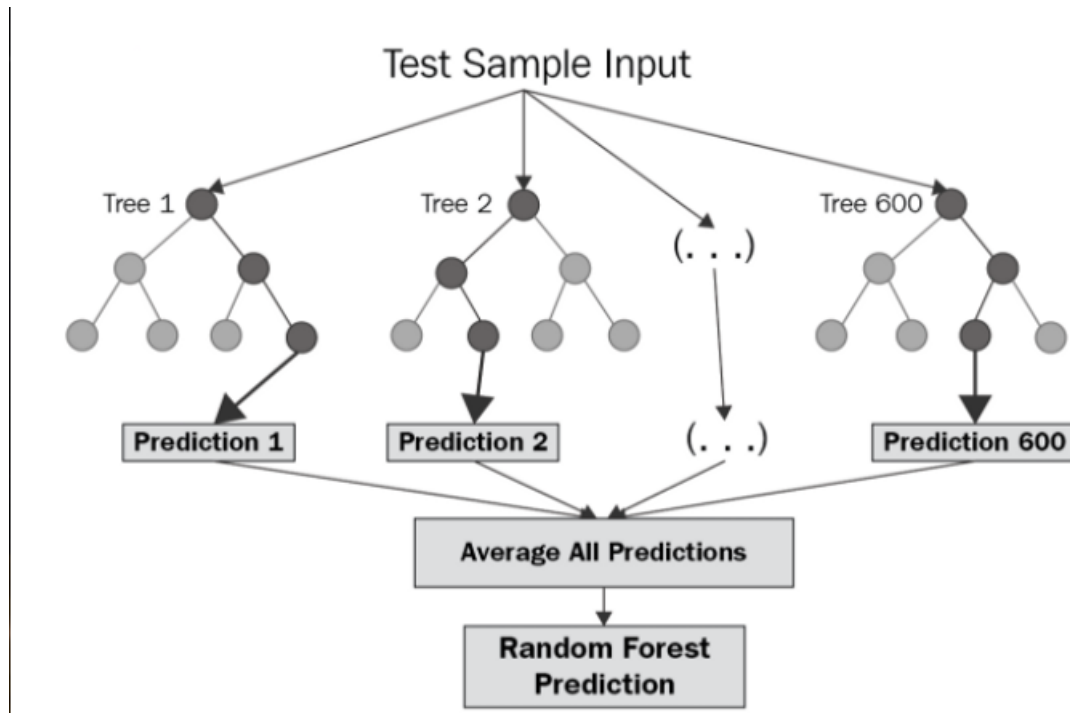


Figure 2.1: Structure of Random Forest

2.2 K-Nearest Neighbour

The K-Nearest Neighbors algorithm is a machine learning technique that can be used for classification and regression. It's a basic method that saves all of the available examples and uses a similarity metric to model the new input or case. The K in KNN refers for the number of nearest neighbors that are used to predict new data. True, it's a hyperparameter that should be used to get the best possible fit for the data set. Hyper parameters are those that cannot be determined directly from data; K controls the shape of the decision border in a standard training cycle[14]. The new data is labeled based on the average value of the K nearest neighbors' regression analysis. Unlike statistical methods that seek to distinguish a model from the given data, the KNN method uses the training set as the model. In any KNN methodology, distance traveled, and other parameters are defined by the number of neighbors, kind of neighbors, and other factors. There are some disadvantages to using machine learning in this way. There are a number of variables that can be used.

These conditions must be applied correctly. Certainly, we can improve precision. In a real-world scenario, using data comparable to ours. We can't test everything since we can't use the trial and error method to determine which of the possible settings performs best. As a result, we employ the most effective methods. By choosing the right parameters may go through some of the key factors that have a big impact on the accuracy of a KNN model. K is recognized as one of the most essential components in the model that can have a significant impact on the outcome. The accuracy of the predictions A lesser value of k can show for any given issue domain, a huge difference between projected and actual results A big value of k, on the other hand, will carry out a large-scale model bias. As a result, we must select a suitable value for k, which is neither excessively enormous nor excessively little. We can observe how the k parameter selects the training set and predicts the unknown value by referring to the image below. Our goal is to use cross validation to find a suitable value for k so that we can develop an ideal model. Below is a figure of KNN Model[22]

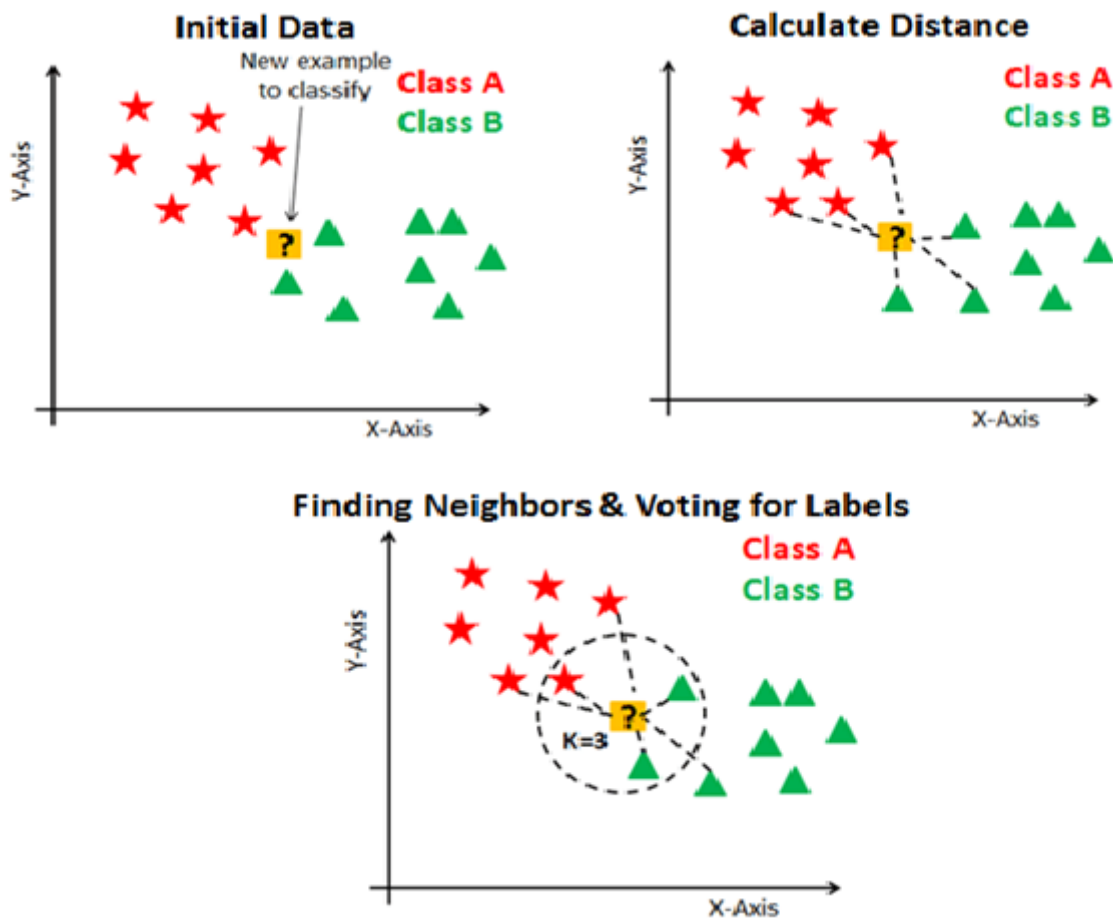


Figure 2.2: KNN Model[22]

2.3 Decision Tree

A decision tree is a powerful tool for solving classification and regression problems that has its roots in machine learning theory. Unlike other classification strategies

that use many features (or bands) to address a problem, categorization can be completed in a single decision step. The tree containing all of the data is supported by a multistage or hierarchical decision process. A root node (which includes all data), a set of internal nodes (splits), and a set of exterior nodes (the number of terminal nodes) make up the tree (leaves)[2]. Each node in the decision tree structure makes a binary choice that separates one or more classes from the rest of the classes. Moving down the tree until you reach the leaf node is how most processing is done. This is referred to be a top-down strategy. Decision tree regression is a type of decision tree classifier that can be used to approximate real-valued functions such as class proportions. The construction is completed. Binary recursive regression trees are also based on binary recursive regression trees. Partitioning is an iterative procedure for dividing data into compartments. To begin, all of the training samples are used. Determine the tree's structure. After that, the algorithm breaks the data into binary splits and chooses the best one. The division of data into two sections in such a way that the sum of the squared departures from the mean is minimized. In the various parts After that, the dividing process is applied to each of the newly created branches When the whole structure is attained, the tree may suffer from over-fitting because it was built using training samples. This may cause the categorization accuracy to degrade. When applied to unknown data, the tree may result in little ability to generalize. As a result, a pruning technique is frequently employed in conjunction with validation data collecting and a cost complexity factor chosen by the user. Pruning is done to reduce the sum of the output variable variance and increase the variance of the output variable nodes representing the cost of complexity per node, and the product of the number of terminals and the validation data cost complexity factor, which is a parameter during the process of pruning. The most recently grown node is eliminated first, then the next node, and so on. The figure below shows a data mining process used by this method as similar to our research case showing how the weather predictors and target formulated to give an idea of future forecasting.



Figure 2.3: Weather Prediction Using Decision Tree[20]

2.4 Mann-Kendall Trend Test

Trend analysis is often used to understand the changes in hydrological and climatic time series pattern. Different statistical tests are being used in the current world to determine the changes in pattern in time series data. Mann-Kendall trend test is mostly used for non-parametric tests to understand trend of time series data. In the MK test, preemptive assumption or null hypothesis is that the data is randomly distributed and independently arranged. Positive autocorrelation within the existing data, on the contrary, improves the likelihood of finding trend on data even if there is no trend and vice versa. Despite the fact that it is a renowned fact among researchers, some researches still looked through the possibilities of it and the autocorrelation in data is often overlooked. When determining if there is any monotonic trend upwards or downwards exists in a time series, the MK Test is used. The testing data does not have to be regularly distributed or linear. It does necessitate the absence of autocorrelation in it.

Preemptive assumption or the null hypothesis is that no trend exists there, while on the other hand alternative hypothesis is that trend exists there on the two-sided test or in the one-sided test a upward or a downward trend exists. The MK Test employs the following statistic for a time series x_1, \dots, x_n :

$$S = \sum_{i=1}^{n-1} \sum_{j=k+1}^n \text{sign}(x_j - x_i) \quad [1] \quad (2.2)$$

If the value of S is greater than 0 then the observations from later are greater than the earlier observations on the time series and if the value of S is lesser than the 0 then the observations of the earlier observations are greater than the later observations.

Variance of S is determined by:

$$\text{var} = \frac{1}{18} \left[n(n-1)(2n+5) - \sum_t f_t(f_t-1)(2f_t+5) \right]_{[1]} \quad (2.3)$$

Here, t differentiates of the set of tied ranks f_t is the frequency of rank t appearing. The Mann-Kendall test uses following test statistic to determine z,

$$= \begin{cases} (S-1)/se, S > 0 \\ 0, S = 0 \\ (S+1)/se, S < 0 \end{cases} [1] \quad (2.4)$$

Here, se is equal to sqrt of var, if there is no monotonic trend according to the null hypothesis then for a time series which has more than 10 elements in it $z \sim N(0,1)$ then it has a normal distribution for z.

2.5 Related Work

One of the studies used five data mining approaches to estimate rainfall in the city of Lahore: Support Vector Machine, Naive Bayes, k Nearest Neighbor, Decision Tree, and Multilayer Perceptron are examples of machine learning algorithms they have used [28]. The estimated data of rainfall compiled of 12 years and used the following methods to compare and contrast between the methods most appropriate for that particular work. A classification frame has been used pertaining of table, graphs and frameworks[10]. According to their findings, the applied classification techniques performed well for the no-rain class, but not so well for the rain class. Secondly, in another paper they used similar but slightly different data mining techniques like decision trees, lazy learning, artificial neural networks, clustering and regression algorithms to predict suitable weather condition. This paper[7] concentrated on many characteristics such as temperature, thunderstorms, rainfall etc. Throughout the paper, the researchers have compared in between the algorithms which gave them the most accurate prediction. They discovered that when the size of the training set grows larger, the accuracy increases at initially, but subsequently decreases after a certain point[7]. Thirdly, we have another paper[6] that unlike the previous two studies where they have taken many algorithms to observe the trend, this concentrated mainly on Mann Kendal Test to predict climate change on certain states of northern United states. We get another paper[13] similar to this that coalesced Mann Kendal and added another data mining technique Sen's T slope in in Yangtze river basin, china. However, one of the drawbacks of these paper was the deficiency of their data set and the credibility to check the forecasts with more variants and variables. The majority of the papers mentioned the Mann Kendall methodology, however we did find one work that talked about utilizing KNN to predict up to seventeen climatic variables at once, such as Mean Temperature, Max Temperature, Min Temperature, SST, SLP, Gust, and so on[3]. None of the prior created systems have been able to accurately forecast such a large number of qualities at the same time. The Climate Prediction Tool (CPT) [from NOAA] has just been updated to operate with many attributes, although this new version is only available in English. The outcomes were simple to comprehend and proved to be a huge success[5].

Chapter 3

Dataset Analysis

Our primary dataset was taken from Bangladesh Meteorological Department (BMD). BMD keeps rainfall data from 1948. They have 35 stations spread over the whole country where they collect daily rainfall data. This rainfall data is available on their website and anyone can request for the dataset from BMD. Bangladesh Meteorological Department (BMD) is a government organization that is responsible for recording the meteorological observation and providing forecast for natural disasters to the disaster management department. It falls under the administrative control of defense ministry. It started its work with first station in Satkhira in 1887 under British rule, in 1947 it was renamed as Pakistan Meteorological Services (PMS) later in 1971 it was named as Bangladesh Meteorological department.

Bangladesh Meteorological department has 35 stations across Bangladesh. Those stations collect data of every day rainfall around those stations. Those data are put together into a dataset and kept on their website. Anyone can avail those datasets upon request to BMD. We requested for rainfall dataset of Bangladesh. We got rainfall dataset from 1948 to 2014. In this dataset we have data from 35 stations. Data is organized in arrays of Stations, year, month and daily rainfall in mm.

We collected few other data from the website such as data of 2015, 2016 manually for analysis purposes. We did not use merge that data with the existing dataset.

We also collected detailed dataset of Chittagong and Barisal but we couldn't use that data due to unavailability of data from other stations. In fig.1 we see the location of the stations across Bangladesh. [12] In fig.2 we can see the position of those stations and elevation of those stations. We also got mean annual rainfall of those stations.[4]



Figure 3.1: Location of BMD Stations

Station	Latitude (N)	Longitude (E)	Elevation (m)	Mean annual precipitation (mm)
Barisal	23°43'	90°22'	2.10	2111.6
Bhola	22°41'	90°39'	4.30	2265.0
Bogra	24°51'	89°22'	17.90	1719.7
Chandpur	23°14'	90°42'	4.88	2018.0
Chittagong (City)	22°21'	91°49'	33.20	2951.6
Chittagong (AP)	22°13'	91°48'	5.50	2856.5
Chuadanga	23°39'	88°49'	11.58	1524.1
Comilla	23°26'	91°11'	7.50	2257.7
Cox's Bazar	21°27'	91°58'	2.10	3612.8
Dhaka	23°47'	90°23'	8.45	2062.8
Dinajpur	25°39'	88°41'	37.58	1851.1
Faridpur	23°36'	89°51'	8.10	1883.2
Feni	23°02'	91°25'	6.40	2967.9
Hatiya	22°27'	91°60'	2.44	3128.3
Ishurdi	24°09'	89°20'	12.90	1457.3
Jessore	23°12'	89°20'	6.10	1610.4
Khepupara	21°59'	90°14'	1.83	2775.4
Khulna	22°47'	89°32'	2.10	1725.2
Kutubdia	21°49'	91°51'	2.74	2769.9
Madaripur	23°10'	90°11'	7.00	2030.5
Maijdee Court	22°52'	91°60'	4.87	2988.8
Mongla	22°28'	89°36'	1.80	1948.6
Mymensingh	24°44'	90°25'	18.00	1741.3
Patuakhali	22°20'	90°20'	1.50	2684.7
Rajshahi	24°22'	88°42'	19.50	1555.2
Rangpur	25°44'	89°16'	32.61	2075.1
Rangamati	22°38'	92°09'	68.89	2402.8
Sandwip	22°29'	91°26'	2.10	3507.5
Satkhira	22°43'	89°05'	3.96	1683.7
Sitakunda	22°38'	91°42'	7.30	2984.4
Srimangal	24°18'	91°44'	21.95	2206.3
Syedpur	25°45'	88°55'	39.60	2217.9
Sylhet	24°54'	91°53'	33.53	3963.0
Tangail	24°15'	89°56'	10.20	1898.7
Teknaf	20°52'	92°18'	5.00	3884.5

Figure 3.2: Location of the stations with mean annual rainfall[4]

Then, we separated data for different stations into separate datasets. As we mentioned in chapter 1, Bangladesh is divided into different hydrological zones. It is important to find data for different stations in different zones to understand the rainfall in those regions. In figure 6, we have shown how the data from different stations look after separating it from the modified dataset,

The image shows a large table with 29 columns and 48 rows. The first column is 'Station' (all 'Chittago'), the second is 'Year' (ranging from 1943 to 1953), and the third is 'Month' (ranging from 1 to 12). The remaining 26 columns contain numerical data representing rainfall values. The data is organized such that each row corresponds to a specific station, year, and month combination, with the numerical values representing the rainfall amount for that specific entry.

Figure 3.5: Separated Dataset

3.3 Data Visualization

We have a dataset We have a dataset of 66 years of rainfall in Bangladesh. Before we dive into the prediction with machine learning models we tried to get some overview with statistical models to get a better understanding of what we can expect.of 66 years of rainfall in Bangladesh. Before we dive into the prediction with machine learning models we tried to get some overview with statistical models to get a better understanding of what we can expect.

3.3.1 Highest Rainfall by Months

As we discussed in Rainfall in Bangladesh, monsoon in Bangladesh usually stretches from June to September. Figure 8 shows the amount rainfall by months in last 66

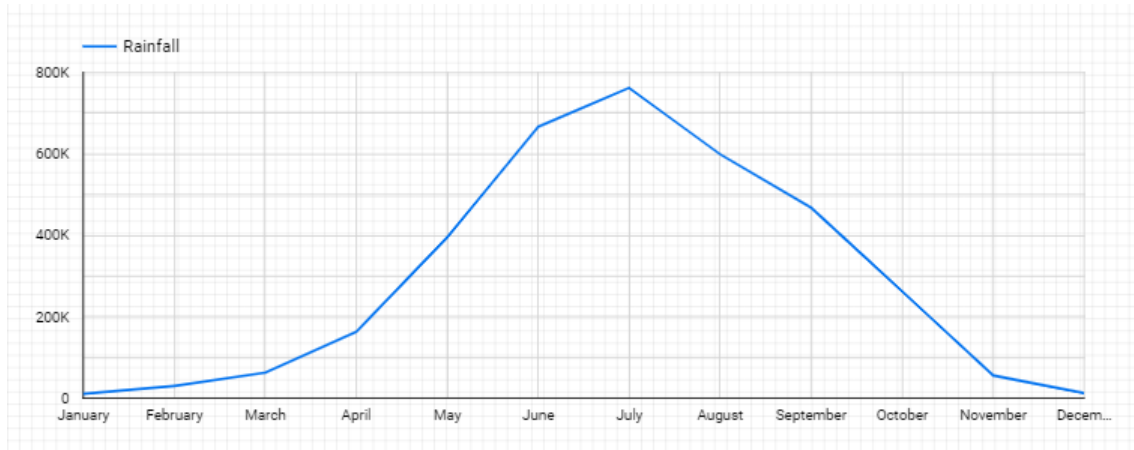


Figure 3.6: Highest Rainfall by Months

years.

As we can see from December to February there is very little amount of rainfall in Bangladesh. This period is Dry cold winter period in Bangladesh. From March it starts to pick up till May. This period is pre-monsoon period when rainfall trend starts to pick up. As we can see, the rainfall amount spikes up during June to July and starts going down from August. July is the month when Bangladesh experiences the highest amount of rainfall. From June to September it is usually monsoon period. We can see till September the rainfall amount is high. After September, next two months October and November are post-monsoon period. During this period rainfall is not very high. Rainfall amount starts going down from October and by the end of November the cold-dry period starts, hence it completes a circle.

3.3.2 Highest Rainfall By Years

If we analyze dataset for 66 years we can how the rainfall amount every year in this periods. Figure 9 shows the tree map of years with highest rainfall.

As we can see, year 1991 experienced the highest amount of rainfall, followed by 1999, 2002, 1998. All the Higher rainfall years sit between 1988 and 2008 (figure 10), this 20 years period. We can assume there was a change during this period in our hydrological cycle but some of the lowest rainfall by years also is happened within this period. So, understanding the hydrological pattern would take more research.

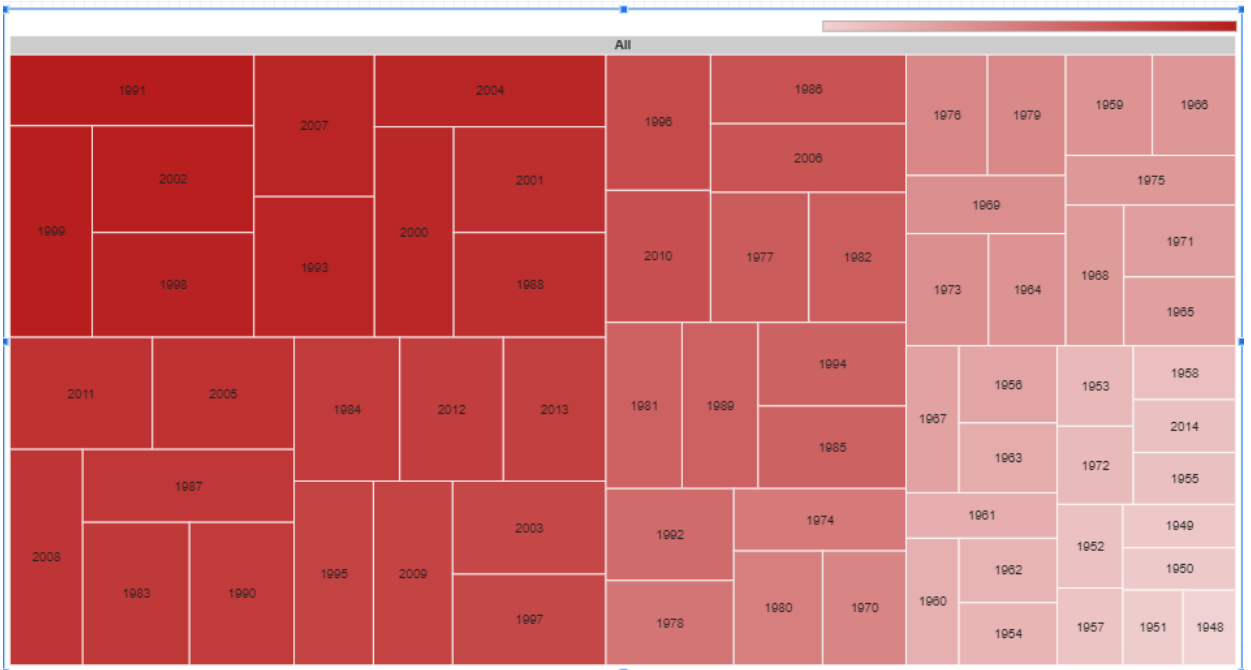


Figure 3.7: Treemap of Highest Rainfall by Years

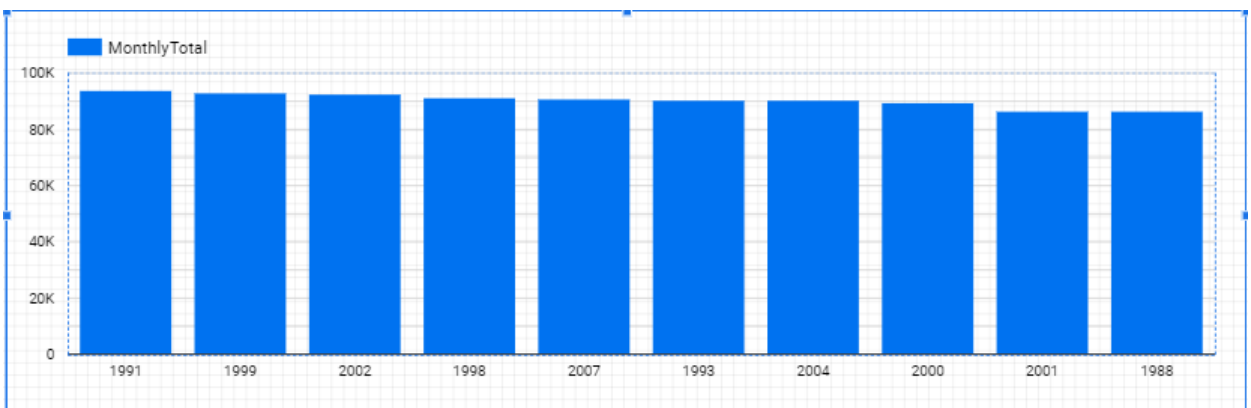


Figure 3.8: Rainfall in 1988-2008 period

3.3.3 Highest rainfall by Stations

According to the hydrological mapping of Bangladesh, North-eastern and South-western stations experience the highest amount of rainfall. Figure 11 shows the amount of rainfall by years in these stations in descending order.

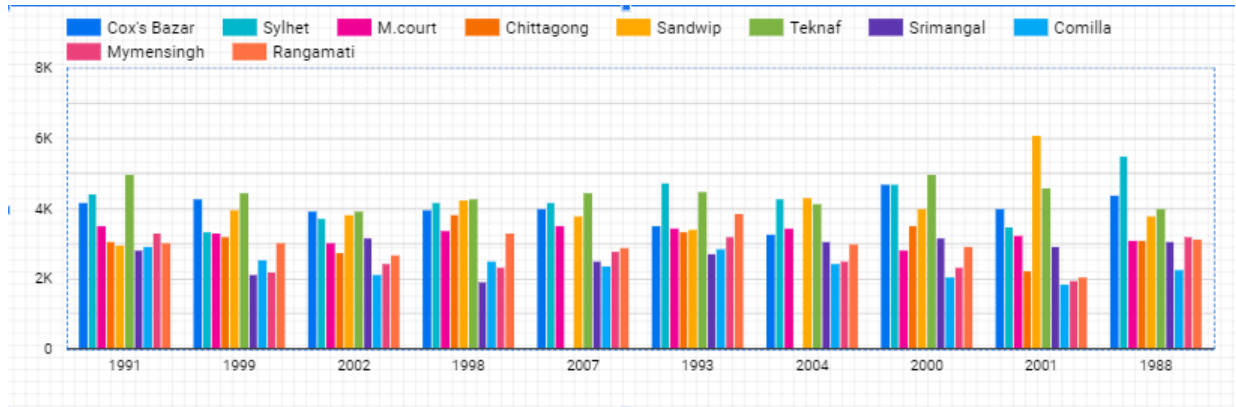


Figure 3.9: Highest rainfall in stations by year

We can see Teknaf experienced the highest amount of rainfall during 1991, the year severe cyclone hit Bangladesh. That year southern side of Bangladesh suffered because of the cyclone. Other than the years of cyclones, we can see Sylhet experiences the highest amount of rainfall in Bangladesh. Sylhet is situated just near the southern foothill of Himalayas where the winds blow northeast and east. Figure 12 shows the amount of rainfall the stations experience every year

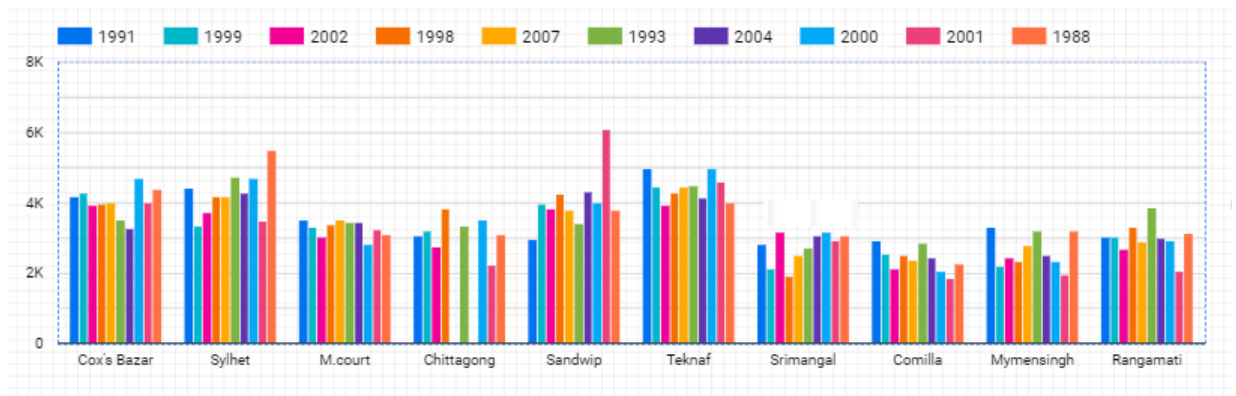


Figure 3.10: Highest rainfall in stations by year

As we can see, Sylhet and Cox's Bazar experience the highest rainfall among all the stations. Sylhet averages almost 4,000 mm. of rainfall every year. Almost all the highest rainfall regions are situated in or near Sylhet and Chittagong region. Let's look at the lowest rainfall regions in Bangladesh in contrary. Figure 13 shows the lowest rainfall by station

As we can see, northern regions experience lowest amount of rainfall and are usually the driest. Tangail experiences the lowest rainfall followed by Sydpur and Rajshahi.

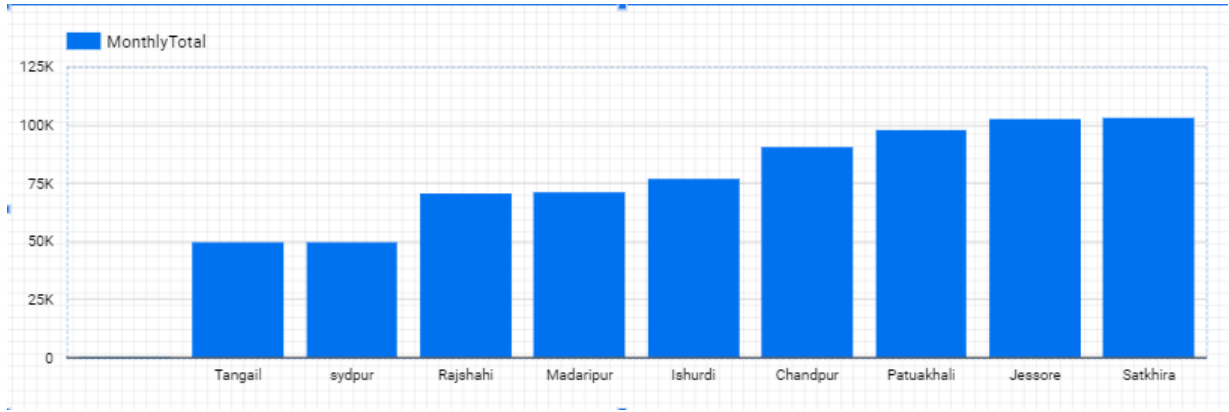


Figure 3.11: Lowest rainfall by station

3.4 Trend Analysis

To understand the trend of Bangladeshi rainfall, we decided to apply Mann Kendall test on the data of 6 divisional stations which fall into different hydrological zones. We have separated data for 30 years of Dhaka, Chittagong, Sylhet, Barisal, Rajshahi and Khulna stations and applied Mann Kendall test on the data. We had to add the monthly total rainfall and convert it into yearly data to make the data suitable for Mann Kendall Test. Table 1 shows the results of results obtained for Mann Kendall Trend test.

Stations	MK Test Results		
	MK Test (Z)	Trend	Results
Dhaka	-1.68	Down	NS
Chittagong	0.75	Up	NS
Sylhet	-1.86	Down	NS
Barisal	-1.32	Down	NS
Rajshahi	-1.78	Down	NS
Khulna	1.11	Up	NS

Table 3.1: Mann-Kendall Test Results

From the results we can see the trends of Dhaka, Sylhet, Barisal and Rajshahi is going downwards whereas the trend in Khulna and Chittagong is going up. We have tested the data on at 5% level of significance. None of the stations show significant change in terms of rainfall. Sylhet shows the highest downward trend at Z score of -1.86 and Khulna shows the highest upward trend at Z score of 1.11. Southern stations show upward trends whereas Central and Northern stations show downwards trend. As none of the trends have significant increase and decrease it is safe to say there was not any drastic change in the patterns of rainfall. Below we show the trend lines of chosen stations for better understanding:

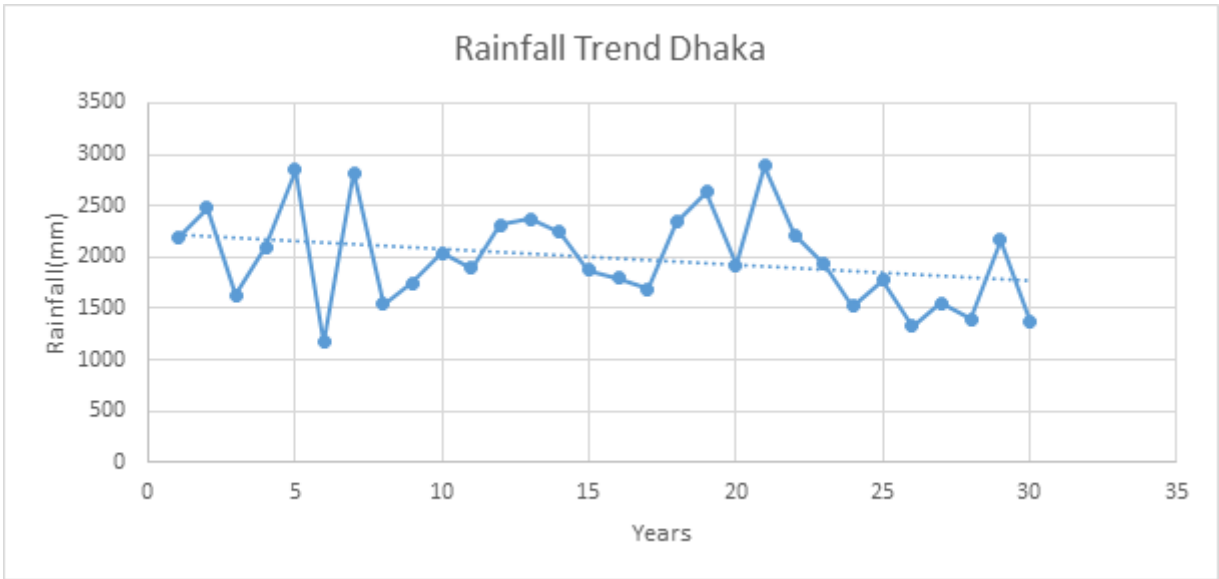


Figure 3.12: Rainfall Trend Dhaka

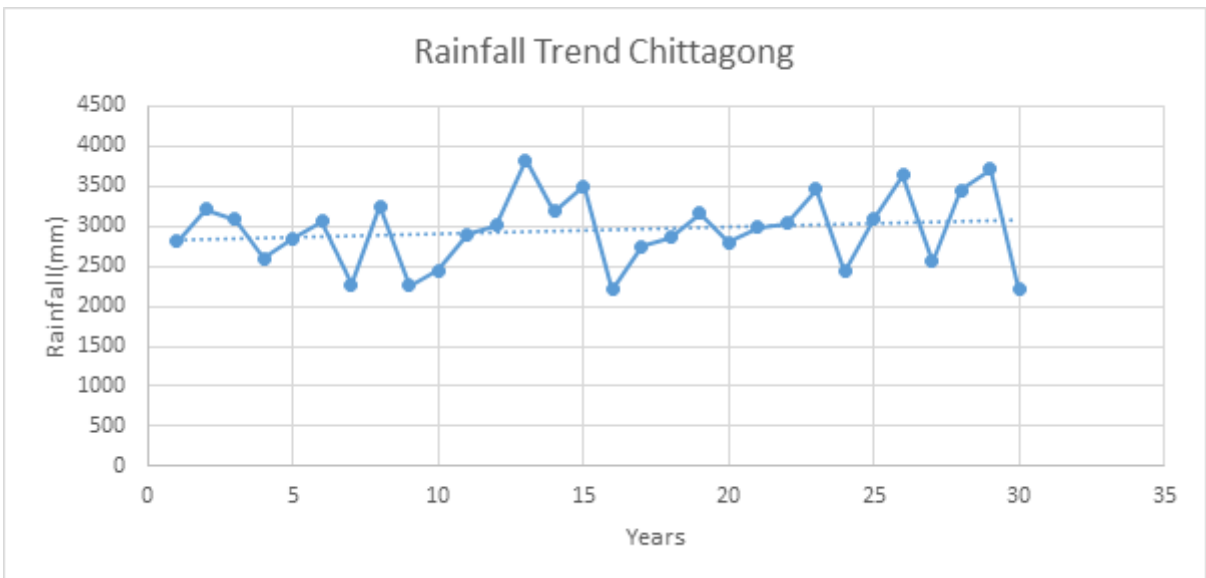


Figure 3.13: Rainfall Trend Chittagong

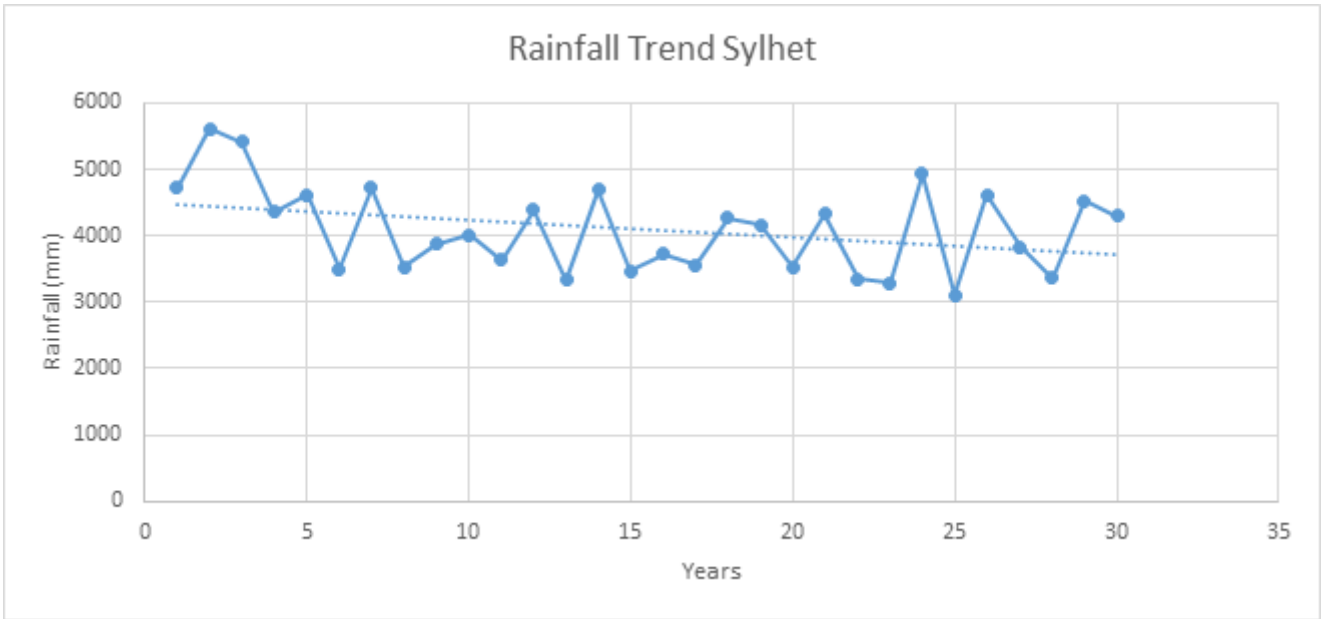


Figure 3.14: Rainfall Trend Sylhet

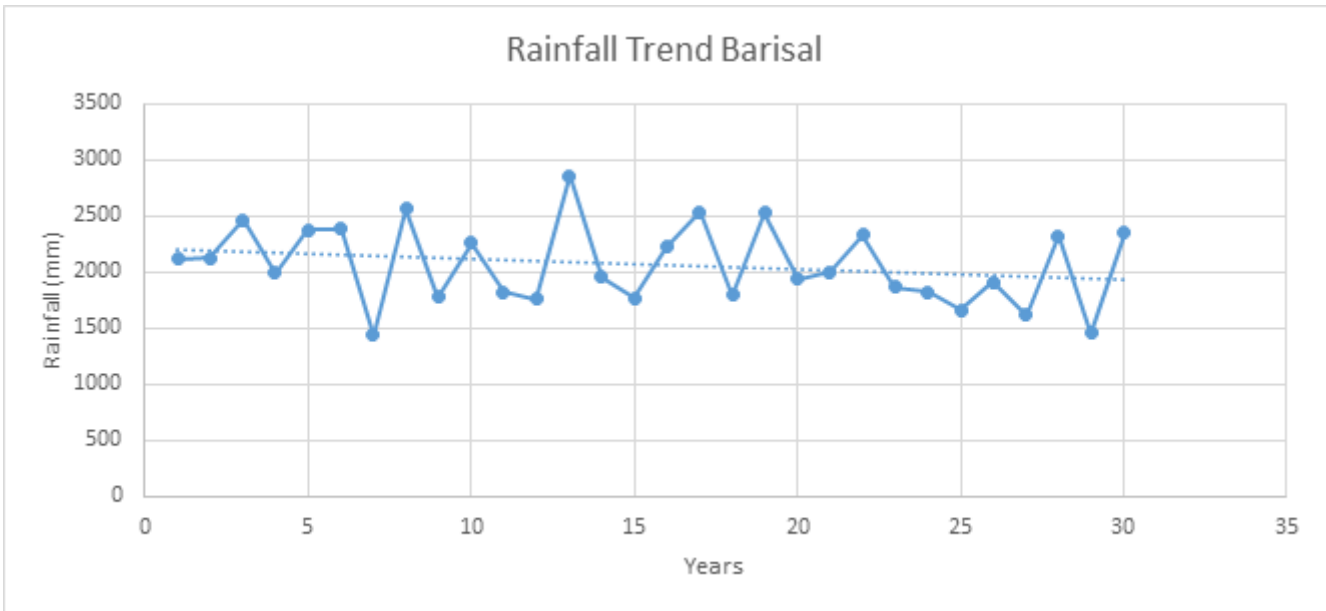


Figure 3.15: Rainfall Trend Barisal

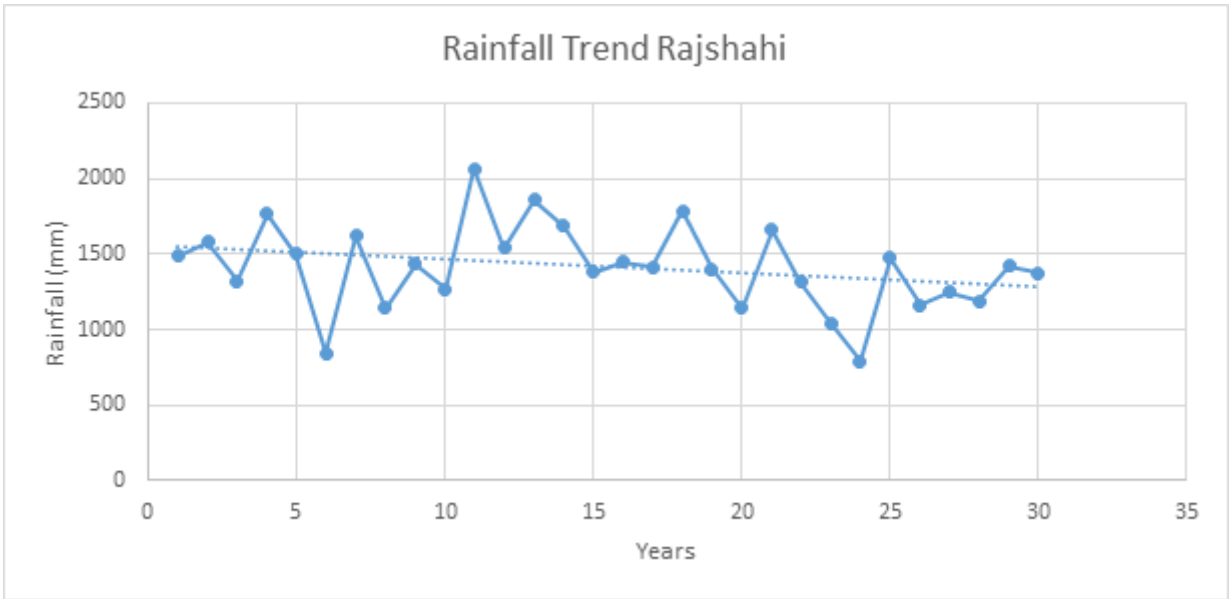


Figure 3.16: Rainfall Trend Rajshahi

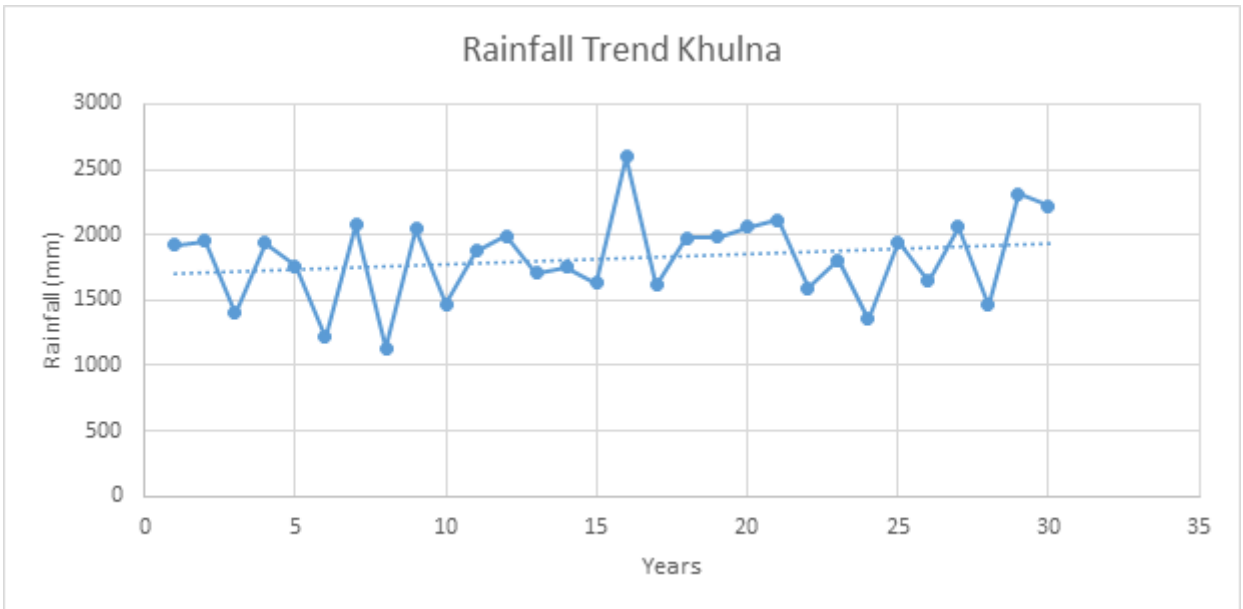


Figure 3.17: Rainfall Trend Khulna

Chapter 4

Methods and Implimentation

Our method of approach was developing a predictive model which would takes in values from our dataset and uses various machine learning models to predict the amount of rainfall which falls in thirty two districts of Bangladesh in every month of every year. The rainfall was predicted by dividing the dataset into two parts training and testing later we used different machine learning models such as KNN (K-nearest neighbor) algorithm, Random forest Regression and Decision making tree to predict our y-value which is “The annual rainfall of a station in every month”.

4.1 Dataset Description

The dataset we used had records of daily amount of rainfall of every Station starting from 1948 to 2014. However, we only needed the Station index, Station name, year, month and monthly total at the end of our csv file. For this reason the amount of daily rainfall data that we had in our dataset was dropped. Here, is an elaborate description of how our dataset looked like and how we modified it to serve our purpose.

- The dataset consisted of the record of rainfall that fell in on a daily basis in 35 stations of Bangladesh from 1948-2014
- It had columns such as Station name, year, month and days from 1-31.
- The rainfall here was measured in mm.

Due to the huge amount of data in our dataset it gave us an extra leverage of achieving a very high accuracy. However, for achieving high amount of accuracy we had to do some feature reduction which also helped us in storage and reduced our computational time by a great margin. The day’s column of this dataset was dropped as we want to measure monthly annual rainfall not the total amount of rainfall that falls on a daily basis. Later during our project we had to encode the Station name as it is a categorical variable. A categorical variable is a type of variable whose value takes on a form of label. Most machine learning algorithms such as keras and others takes numerical inputs thus the challenge here for us was to transpose this categorical value into a numeric one which could be used for our machine to train dataset. To do this we used label encoder from sklearn library.

Label encoding transforms the labels into numeric form so that the data gets machine

readable and our machines can be trained properly. Encoding categorical variables makes machine learning algorithms easier to make predictions as converting it into numeric form which is machine readable increases the decision power of our machines on how to operate on these type of variables and apart from this encoding categorical variables is an very effective part of pre processing. Data preprocessing is the preparation of raw data making it suitable for machine learning. This step is the very first and the most crucial step in the creation of a machine learning model as real life data contains various types of noises, missing values thus most datasets are not suitable for the process of machine learning. In order to make this suitable preprocessing is a very required step. Preparation of a machine learning algorithm follows:

- Import the dataset
- Importing libraries
- Handling of missing values
- Encoding categorical values
- Choosing dependent and independent variables
- Splitting dataset into training and testing
- Then feeding it to a machine learning algorithm

Let's go back to Label encoding in this paragraph we will describe how it works. Label encoding transforms categorical variables into numerical variables the choice of values are between 0 and number of classes minus 1. This Label encoder is imported from sklearn library. Sklearn library provides extremely efficient tools for encoding categorical values. If a label repeats itself the value assigned is the one which the tool assigned the label earlier. This method of assigning numerical values is very critical as machines understand numeric language only and machines can now come up with a suitable pattern of prediction by looking at our dataset.

4.2 Steps in creating a machine learning model:

The steps describe in the flowchart below is extremely critical at how machine learning algorithms are to be designed

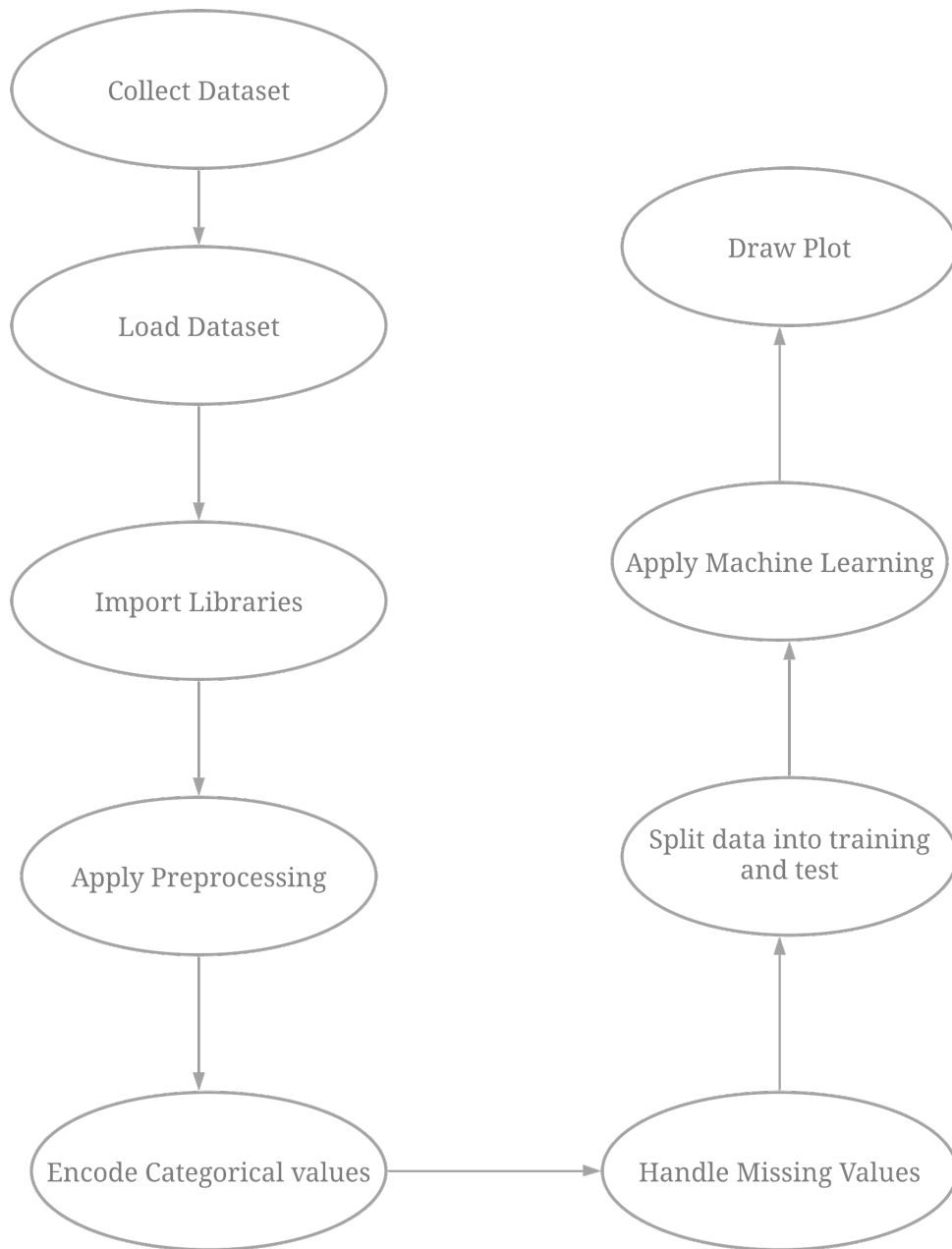


Figure 4.1: Steps of Creating a ML Model

4.2.1 Matplotlib:

Matplotlib is a package in python which is used for data visualization. Matplotlib is a library that works cross platform and used when we need to make arrays and dataset into 2d plots. The API is extremely object oriented and is used in IPython shells, web applications and in Jupyter notebook.

4.2.2 Developing a Rainfall Prediction Model

Three algorithms were used in developing this model and in all these cases the independent variable was taken as Station name, month number and year number. The dependent variables in all these cases were the total monthly rainfall in every month of every year. The models used in this case were KNN, Decision Tree and Random forest Regression. The outputs of KNN, Decision Tree were pretty similar with very negligible differences which prove the authenticity of our dataset.

4.3 K-Nearest Neighbor

The k-nearest algorithm is a very simple artificial based machine learning algorithm. This algorithm was developed by Dudani in 1976 and an improved version was introduced by Adam Jozwik. Several decades into the future it remains as one of the top choices for data mining and application into medical science as it can also detect cancer cells at a very early stage[8]

4.3.1 How it works

When instances are classified KNN's primary aim then is to make its nearest neighbor a concrete object in the sense of distance. The class of the new instance is determined according to the frequency of its appearance among its k-neighbours. The "choice of the value of k is chosen as a priori" [8]. The main idea in priori machine learning is to apply a machine learning method to a machine learning problem itself. This method is called priori because the data does not originate from any measurement or from any other observation. Prior algorithms come with lots of pros and cons.

4.3.2 Pros and Cons of Priori:

The pros of priori are:

- It is very simple to understand.
- The resulting output is intuitive and easier for end users to understand.
- Priori models are unsupervised thus it does not require a labeled data.
- The algorithm is very exhaustive so it finds all types of rules both specified and unspecified

The cons of priori are:

- If the size of the dataset is small the algorithm might find false connections
- Needs lots of computation time as it scans the entire database repeatedly

KNN chooses nearest neighbor on priori basis although various other techniques such as heuristics and cross validation method has also been suggested. The value chosen must not be a multiple of the number of classes in order to avoid tie votes. Thus in cases of binary classification a non even value of k are selected so that a majority emerges.[11] “The performance of KNN is highly dependent on measure used to calculate distances” [8]. A non parametric approach gives it the flexibility to classify without making any assumptions about the Y functions[11]. However, this algorithm comes with lots of limitations.

4.3.3 Limitations of KNN

There are some limitations to K-NN, such as:

- KNN is extremely slow as it reviews instances every time.
- KNN is very vulnerable to change in dimensions.
- It takes irrelevant parameters into calculation.
- The wrong choice of distance can degrade its performance.

4.3.4 Optimization of KNN

“The KNN algorithm has several limitations and thus needs improvement. Introduction of Bailey method which introduces weights to classical KNN to present nearest k-weighted neighbors. WKNN assigned weights to each calculated value then closest neighbors value is calculated and class is presented to the model.” [8]. Other than this there is a condensed nearest neighbor (CNN) algorithm which removes duplicate data, removes unnecessary variables.

Reduced Nearest Neighbors (RNN) contains all steps of CNN and it also eliminates patterns which contradicts/ does not incline with the results.

The advantage of WKNN over classical KNN is that the samples close to the instance possess a higher weight than the ones that are further apart. Thus distances between k-neighbors are standardized by (K+1)th neighbor and all the standardized values are taken at an interval of [0,1][11] . The weights are calculated agreeing to a kernel function K, The reason of such capacities is to weight the observations relative to a reference point so that the closer an occasion is to the reference, the more prominent its weight will be[11].

“Yong, et al. [24] connected an unused demonstrate of KNN, moved forward by clustering, in content classification. The training instances of each course are clustered by k-means calculation, and coming about cluster centers are taken as the new

training dataset. Each preparing occurrence is weighted by a esteem speaking to its importance. Su [25] proposed a strategy to recognize flooding assaults in real-time, based on irregularity location by genetic weighted KNN. A hereditary calculation is utilized to prepare an ideal weight vector for highlights; in the mean time, an unsupervised clustering calculation is connected to diminish the number of occurrences within the testing dataset, in arrange to shorten preparing and execution time, as well as to advance the system's by and large precision.”[8]

4.4 Implementation of KNN

Before applying KNN model the dataset had to be pre-processed and categorical values had to be encoded using label encoder that was imported from sklearn library. To handle missing values in our dataset we used the fillna() method of python. This fillna() method of pandas replaces missing values with the value that is passed as a parameter. The independent variable X was chosen to be from row 0 to row 3 which is from Station name to the month number. Variables in x='Station name', 'Year', 'Month'. Station name being a categorical value was thus encoded using label encoders. And variable y was chosen to be 'Monthly Total'. In this case the amount of rainfall each Station experienced in mm. Both x and y was to numerical arrays using array() method imported from numpy. Then from sklearn model selection train test split data the data was split into training part and testing part and later from Kn neighbors regressor X train and y train was passed as a parameter and later predicted results were calculated from that along with R-squared score which came out as 0.6333. The difference between our predicted values and actual values was later plotted on the graph below:

The blue lines are actual and orange is predicted. Notice that there is a certain difference between the two proving that the KNN model despite feeding it with large datasets proves that is not 100 percent accurate.

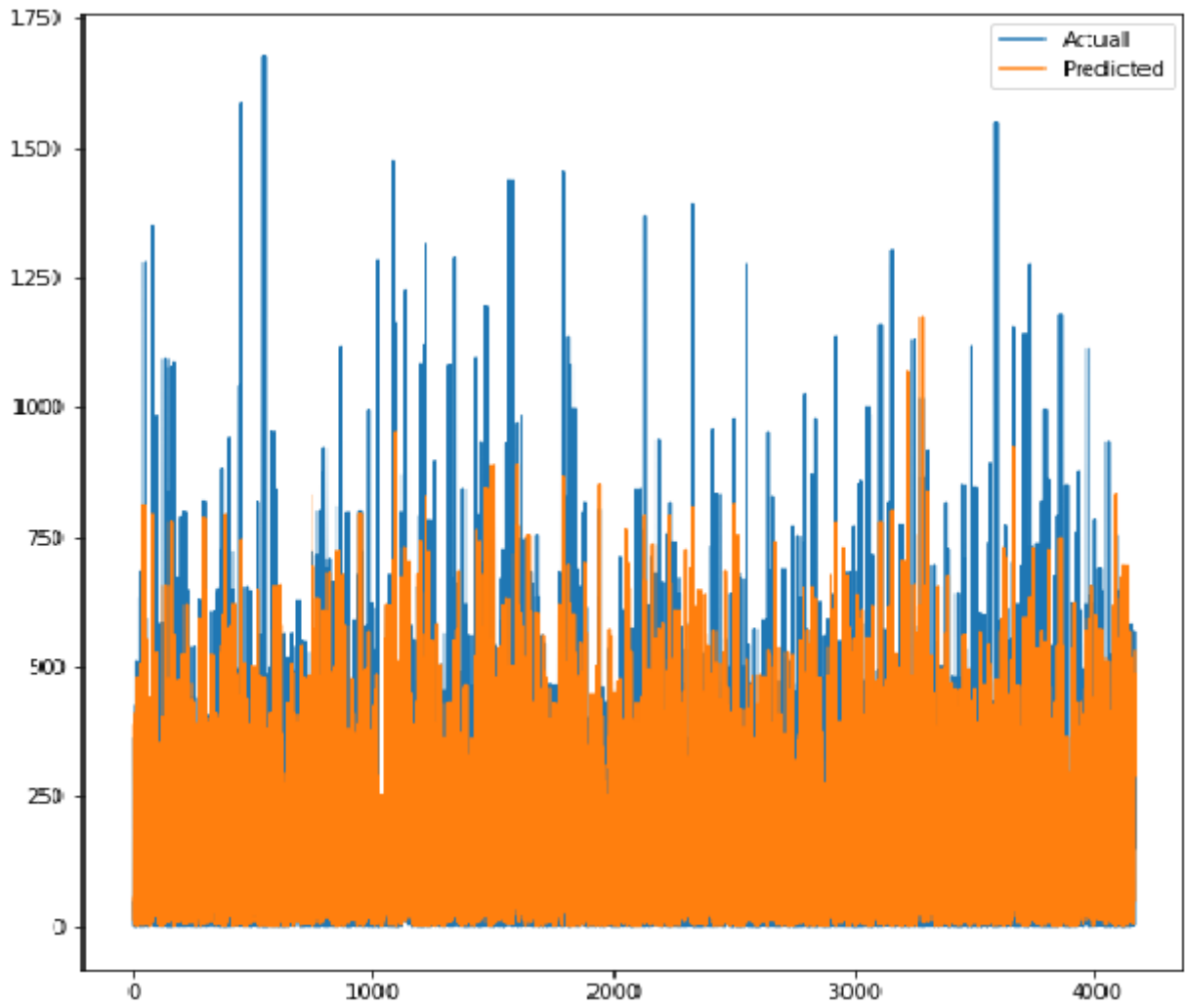


Figure 4.2: KNN Prediction Results

4.5 Implementing KNN on Station Data

We will now implement K-Nearest Neighbour on datasets we have separated for divisional districts.

4.5.1 KNN on Rainfall data of Dhaka

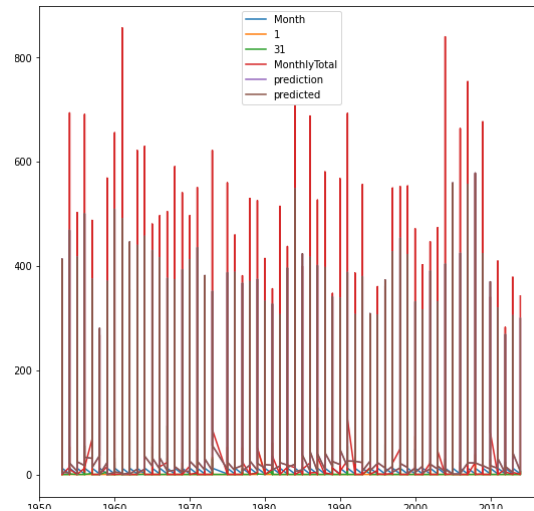


Figure 4.3: KNN on Rainfall data of Dhaka

4.5.2 KNN on Rainfall data of Chittagong

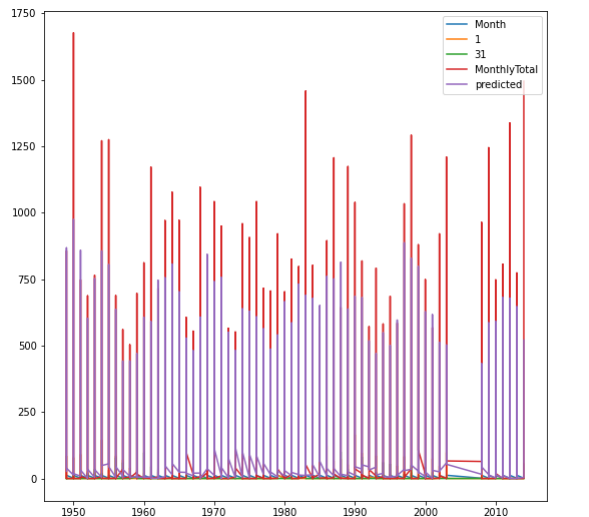


Figure 4.4: KNN on Rainfall data of Chittagong

4.5.3 KNN on Rainfall data of Sylhet

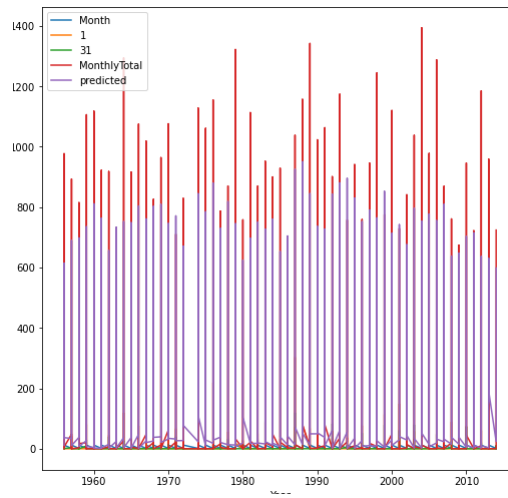


Figure 4.5: KNN on Rainfall data of Sylhet

4.5.4 KNN on Rainfall data of Barisal

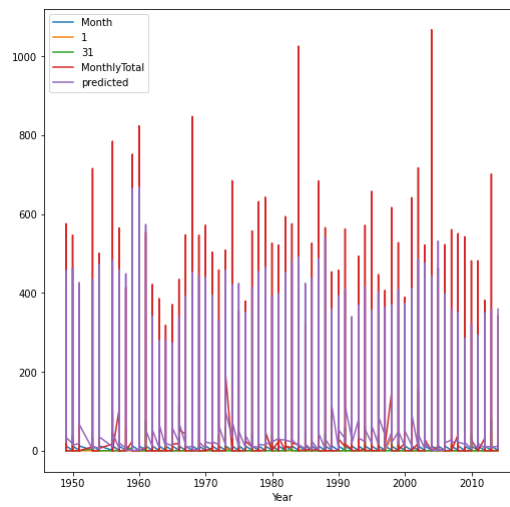


Figure 4.6: KNN on Rainfall data of Barisal

4.5.5 KNN on Rainfall data of Rajshahi

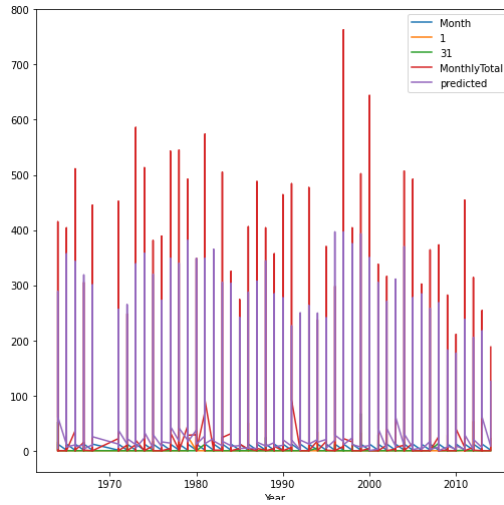


Figure 4.7: KNN on Rainfall data of Rajshahi

4.5.6 KNN on Rainfall data of Khulna

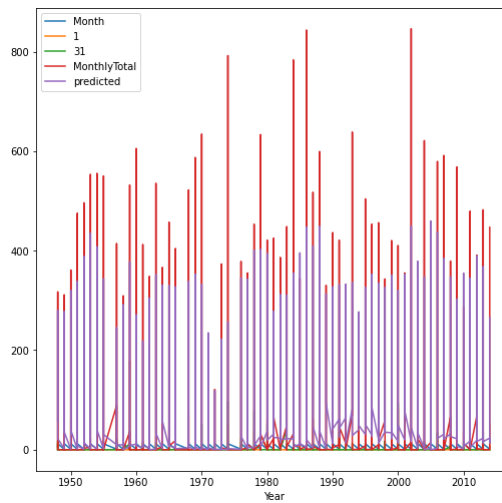


Figure 4.8: KNN on Rainfall data of Khulna

4.6 Random Forest

Since the introduction of the random forest algorithm. It has experienced a wide range of applications. The definition of the Random Forest Algorithm was first introduced by Breiman. The algorithm is defined as follows. The Random Forest classifier is a type of classifier that consists of a collection of tree-structured classifiers $t(x, b)$, $b = 1, \dots, B$ where b are independently distributed random vectors and individual trees casts a unit of vote to the for the most popular class at input X .

4.6.1 How Random Forest Works

The algorithm is supervised learning algorithm which means that it learns from the training dataset it gets. In supervised learning algorithms make predictions based upon the training dataset it gets and the learning stops when the algorithm reaches a satisfying level of performance. Most learning algorithms are supervised learning algorithms. The “forest” that this algorithm builds is based upon the ensemble of decision trees and is trained with “bagging method”. The bagging method is the ensemble algorithm that divides a training dataset into different subsets and then applies multiple models on different datasets and later combines the results of all of these. This bagging method makes random forest algorithm particularly effective against predictive datasets. In summary, random forest algorithm builds more than one decision trees and later merges them together to get a more accurate and stable picture of its assigned predictions. One of the main advantages of this model is that if there are enough trees on the forest than the classifier does not overfit. Overfitting is the statistical error that occurs when a function is too closely aligned to a very limited number of data points. It also has the ability to look out for both noticeable patterns and variables on a dataset for the development of individual decision trees[9]

The random forest regressor is widely in financial markets such as banking, stock market and e-commerce.

It is also widely used in the field of medicine.

The advantage of this model is its accuracy and so in banking sectors it can easily distinguish whether a customer is fraud or loyal.

In e-commerce it can help companies to fond potential customers.

In medicine it can help companies identifying the best chemical combinations and also by analyzing a patient’s medical records it can help doctors pin point the disease a patient is suffering from

4.6.2 Limitations of Random Forest

The limitations of Random Forest regressor are:

- Large number of trees make it really slow and very ineffective.
- Requires huge computational power and resources.
- Requires a lot of computational time.
- “Due to ensemble of decision trees it suffers from interpretability and fails to determine the significance of each variable.”[9]

4.6.3 Optimization of Random Forest

The accuracy and the timing of random forest algorithm can be increase by several ways which are:

1. Specifying the depth of each tree- Trees in this algorithm by default are expanded until all leaves are pure or contains less than the amount of sample that is present.

Using hyperparameters the optimal maximum depth of the algorithm can be found.

2. Increasing or decreasing the size of the estimator- As the number of trees increases so does the accuracy but this causes a huge problem which is the computational time becomes slower. Thus for making the computational time faster the number of estimators need to be lowered however, is higher accuracy is needed the number of trees and the estimator would need to be increased.

3. Specify the maximum number of features to be included at each node split- This method is purely dependent on the dataset that if fed to it. If independent variables are highly correlated then it is best to decrease the maximum number of features however if the accuracy of the training algorithm is very low then it is best to increase the number of features to be included.

4.7 Implementation of Random Forest

Before applying Random Forest model the dataset had to be pre-processed and categorical values had to be encoded using label encoder that was imported from sklearn library. To handle missing values in our dataset we used the fillna() method of python. This fillna() method of pandas replaces missing values with the value that is passed as a parameter. The independent X was chosen to be from row 0 to row 3 which is from Station name to the month number. Variables in x='Station name', 'Year', 'Month'. Station name being a categorical value was thus encoded using label encoders. And variable y was chosen to be 'Monthly Total'. In this case the amount of rainfall each Station experienced in mm. Both x and y was to numerical arrays using array() method imported from numpy. Then from sklearn.model selection train test split data the data was split into training part and testing part and later from sklearn.ensemble we RandomForestRegressor. The dataset was split into training and testing part and later was trained from RandomForestRegressor method in python in a loop from 1 to 100 and the r-squared value was calculated every time. The value of R-squared ranged from 0.62 to 0.68. The difference between our predicted values and actual values was later plotted on the graph below:

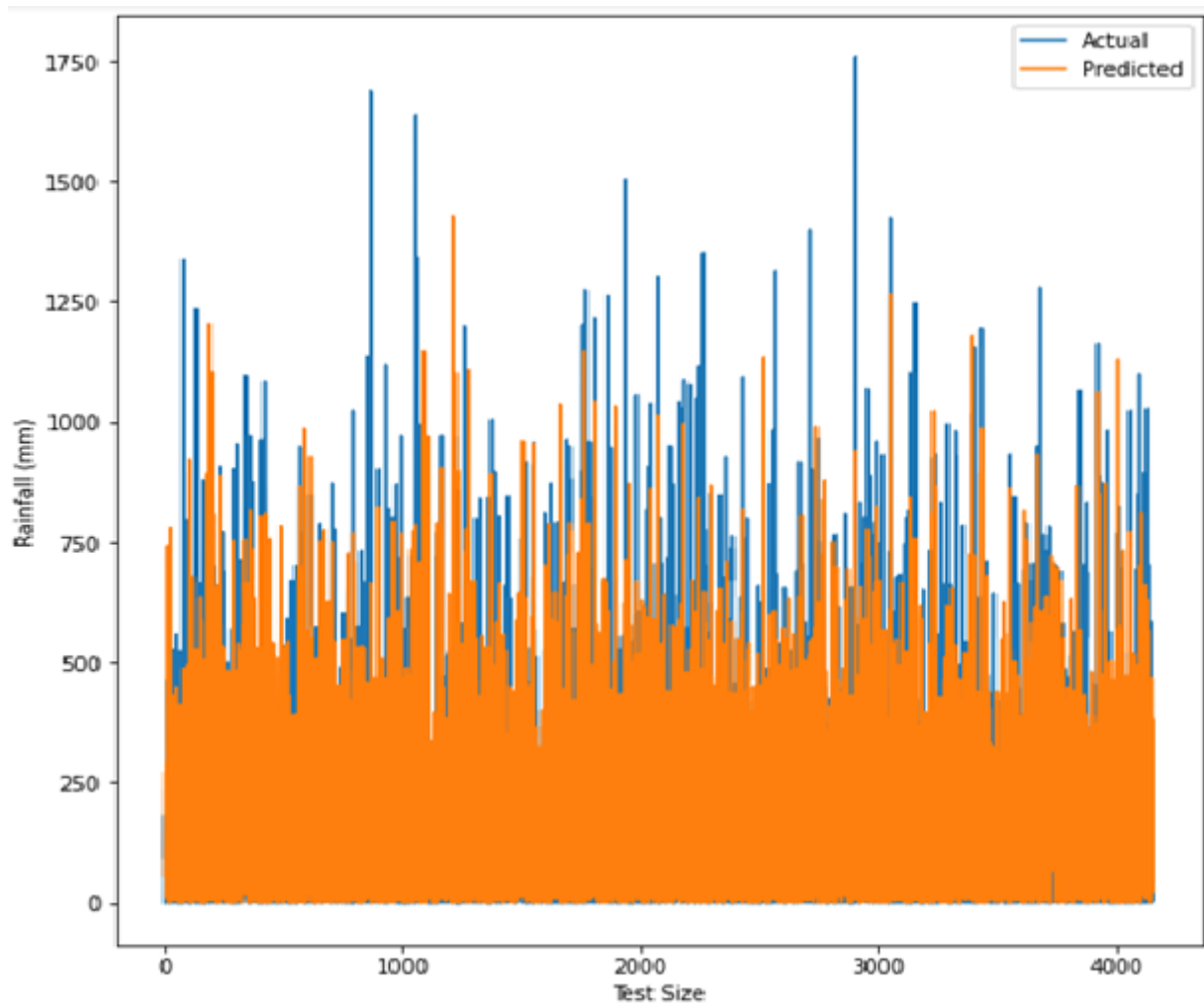


Figure 4.9: Random Forest Regression Results

The blue lines are actual and orange is predicted. Notice that there is a certain difference between the two proving that the Random Forest Regressor model despite feeding it with large datasets proves that is not 100 percent accurate.

4.8 Implementing RF on Station Data

We will now implement Random Forest on datasets we have separated for divisional districts.

4.8.1 Random Forest on Rainfall data of Dhaka

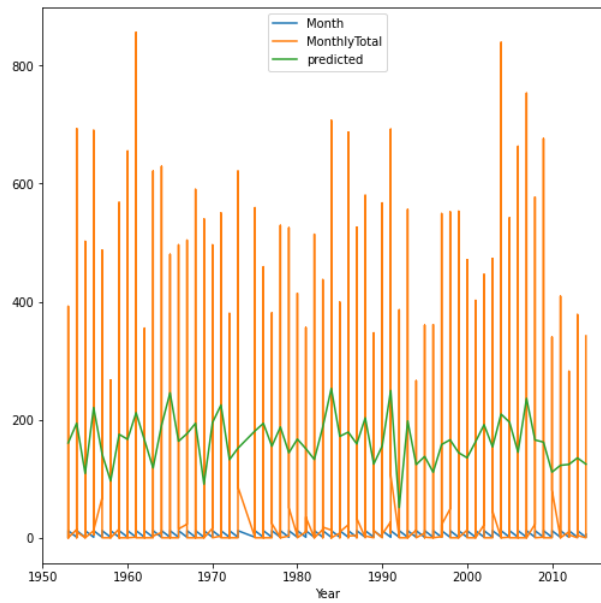


Figure 4.10: Random Forest on Rainfall data of Dhaka

4.8.2 Random Forest on Rainfall data of Chittagong

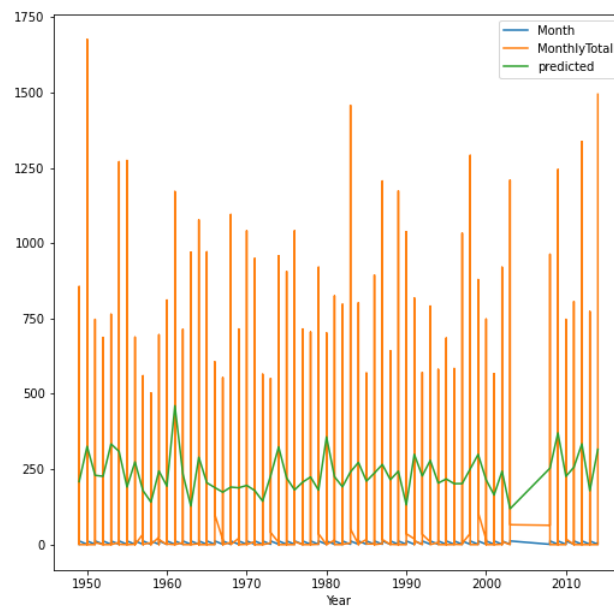


Figure 4.11: Random Forest on Rainfall data of Chittagong

4.8.3 Random Forest on Rainfall data of Sylhet



Figure 4.12: Random Forest on Rainfall data of Sylhet

4.8.4 Random Forest on Rainfall data of Barisal

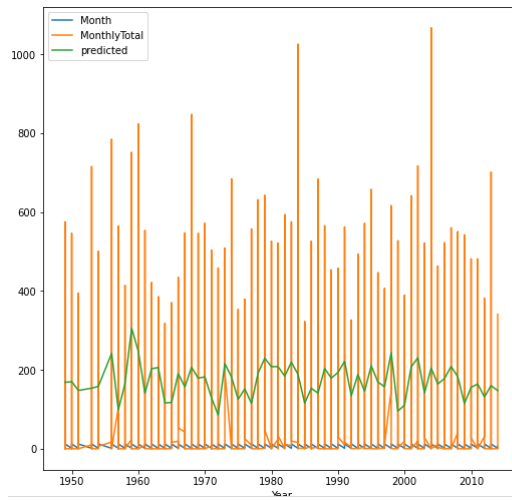


Figure 4.13: Random Forest on Rainfall data of Barisal

4.8.5 Random Forest on Rainfall data of Rajshahi

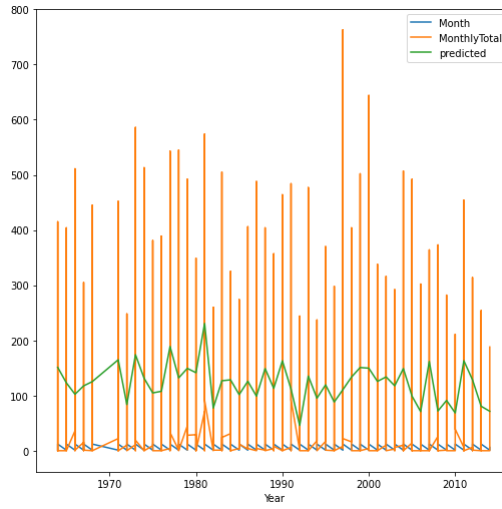


Figure 4.14: Random Forest on Rainfall data of Rajshahi

4.8.6 Random Forest on Rainfall data of Khulna

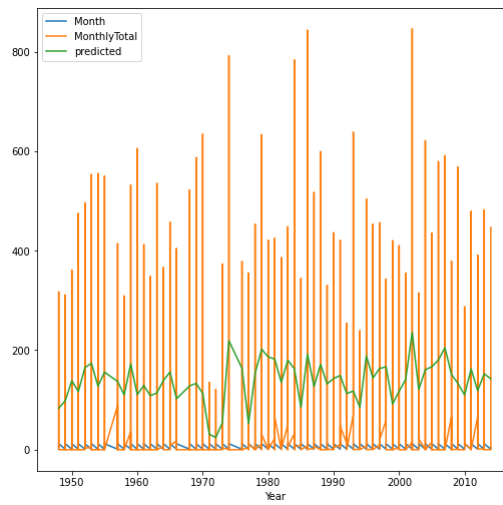


Figure 4.15: Random Forest on Rainfall data of Khulna

4.9 Decision Tree

Decision tree is a classifier which works in a recursive partition of an instance space. It is made up of nodes forming a rooted tree and the starting node is called a root. Nodes with out-going edges are called internal or test nodes all other types of nodes are called leaves or terminals. In decision tree, each instance is split into two or more sub spaces[12] In case of numeric traits, choice trees can be geometrically interpreted as a collection of hyperplanes, each orthogonal to one of the tomahawks. Naturally, decision-makers lean toward less complex choice trees, since they may be considered more comprehensible. Moreover, agreeing to Breiman et al. (1984) the tree complexity features a vital impact on its exactness[24]. The tree complexity is expressly controlled by the ceasing criteria utilized and the pruning method employed. As a rule the tree complexity is measured by one of the following metrics: the full number of hubs, add up to number of clears out, tree profundity and number of properties utilized. Choice tree acceptance is closely related to rule induction. Each way from the root of a choice tree to one of its clears out can be transformed into a run the show basically by conjoining the tests along the way to form the predecessor portion, and taking the leaf's lesson forecast as the course esteem.[24] In machine learning decision tree are supervised. Here the given data is continuously splitted out the leaves are the decisions or the final result and decision nodes are where the data is split.

4.9.1 How it works

In machine learning the decision tree algorithm selects the best attribute using Attribute Selection Measures(ASM) to split the records. This selection measure is a heuristic process for selection of the splitting criteria. The splitting criteria in this process would be best if after splitting each partition is pure and the partition is pure if all the tuples that fall into the partition belongs to the same class. The process begins with the assignment of a rank to each attribute that describes the training tuple. And the attribute for having the best score is chosen as splitting attribute for all the other tuples.[24]

4.9.2 Limitation of Decision Tree

There are some limitations to decision tree. The limitations of Decision Tree is discussed below:

- They are very unstable in nature small changes in data can cause huge structural changes
- They are less efficient in determining the prediction of a continuous variable this is because decision trees lose information during the process of categorization.
- Easily overfits a model.

4.9.3 Optimization

Pre-pruning and post-pruning are done to avoid overfitting. Pre-pruning stops the tree growth before it completely classifies the training set and post-pruning allows the tree to perfectly classify the dataset and after which it post-prunes it. Optimal decision tree- The recursive heuristic method of the decision tree creates split in isolation which may not capture the underlying characteristics of a dataset. The optimal decision tree resolves this by creating the entire dataset at once to achieve global optimality. (Dimitris B, Drunn, Williamson J Optimal classification trees ,2017) . These optimal decision trees provides higher predictive power along with smaller trees using less computer resources and increasing the accuracy of the model.

4.10 Implementation of Decision Tree

Before applying Decision Tree model the dataset had to be pre-processed and categorical values had to be encoded using label encoder that was imported from sklearn library. To handle missing values in our dataset we used the fillna() method of python. This fillna() method of pandas replaces missing values with the value that is passed as a parameter. The independent X was chosen to be from row 0 to row 3 which is from Station name to the month number. Variables in x='Station name', 'Year', 'Month'. Station name being a categorical value was thus encoded using label encoders. And variable y was chosen to be 'Monthly Total'. In this case the amount of rainfall each Station experienced in mm. Both x and y was to numerical arrays using array() method imported from numpy. Then from sklearn.model selection train test split data the data was split into training part and testing part and later from .Then from sklearn.tree we import decision tree. The dataset was split into training and testing part and later was trained from decision tree method in python in a loop from 1 to 100 and the r-squared value was calculated every time . The value of R-squared ranged from 0.56 to 0.610. Two curves were plotted to show the difference between our predictive results and actual rainfall. The graph is plotted below:

The red line is the predictive one and blue line is the actual one. Despite having a large dataset decision tree is very poor in prediction

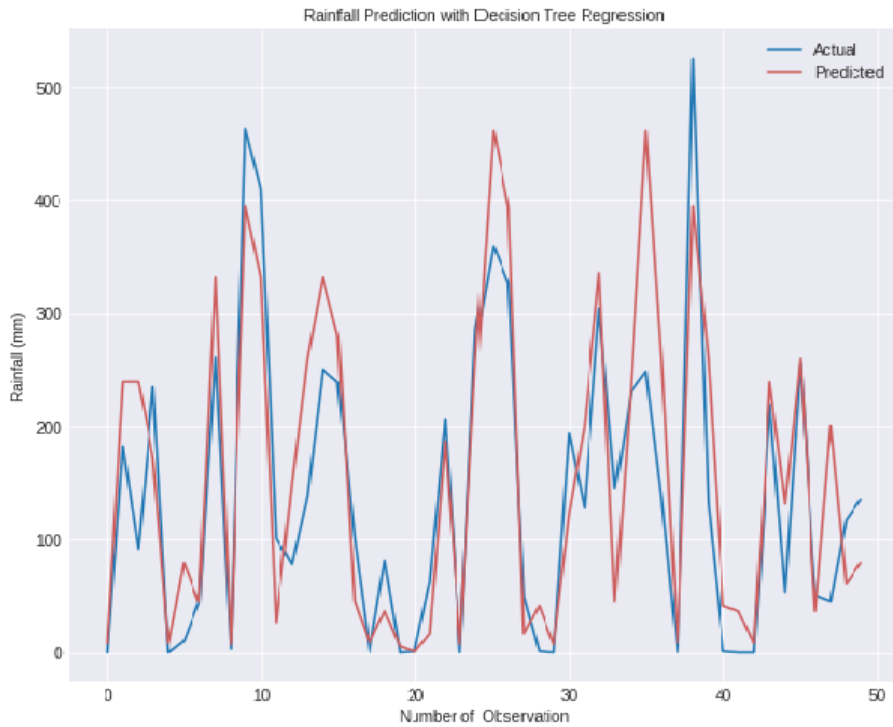


Figure 4.16: Decision Tree Results

4.11 Implementing Decision Tree on Station Data

We will now implement Decision Tree on datasets we have separated for divisional districts.

4.11.1 Decision Tree on Rainfall data of Dhaka

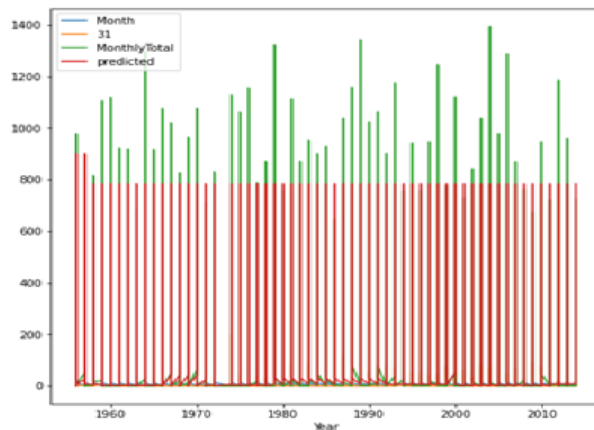


Figure 4.17: Decision Tree on Rainfall data of Dhaka

4.11.2 Decision Tree on Rainfall data of Chittagong

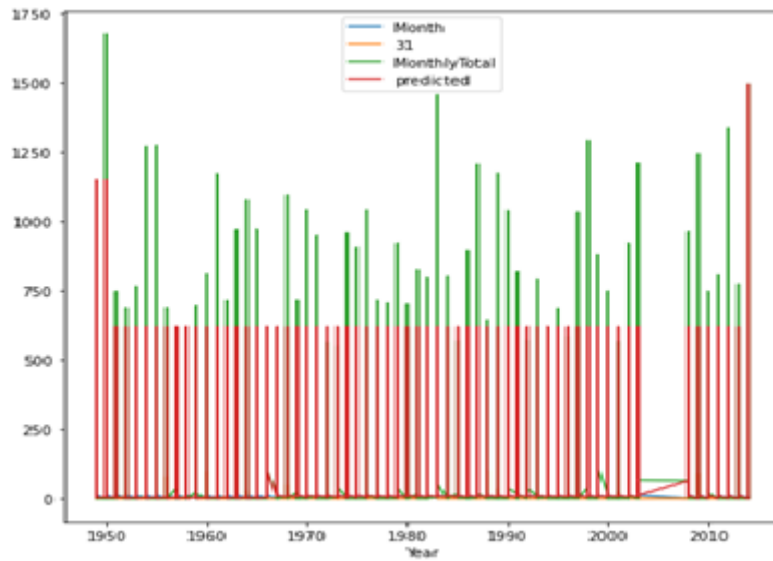


Figure 4.18: Decision Tree on Rainfall data of Chittagong

4.11.3 Decision Tree on Rainfall data of Sylhet

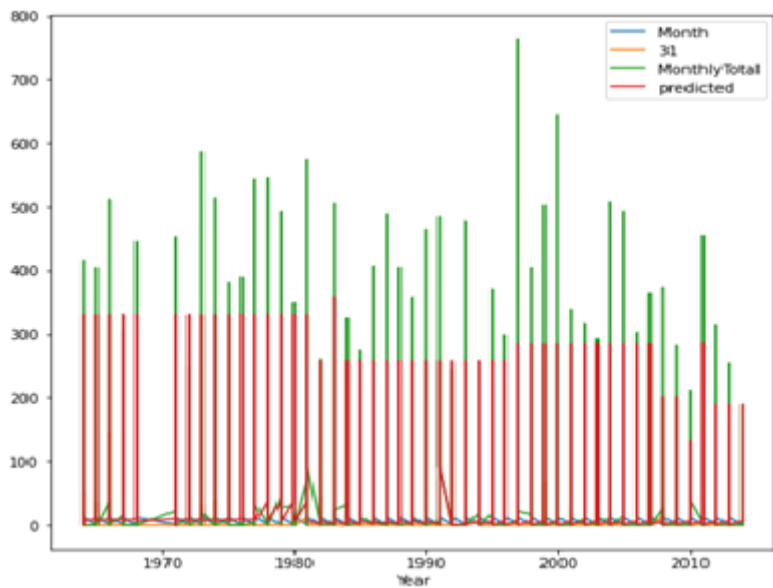


Figure 4.19: Decision Tree on Rainfall data of Sylhet

4.11.4 Decision Tree on Rainfall data of Barisal

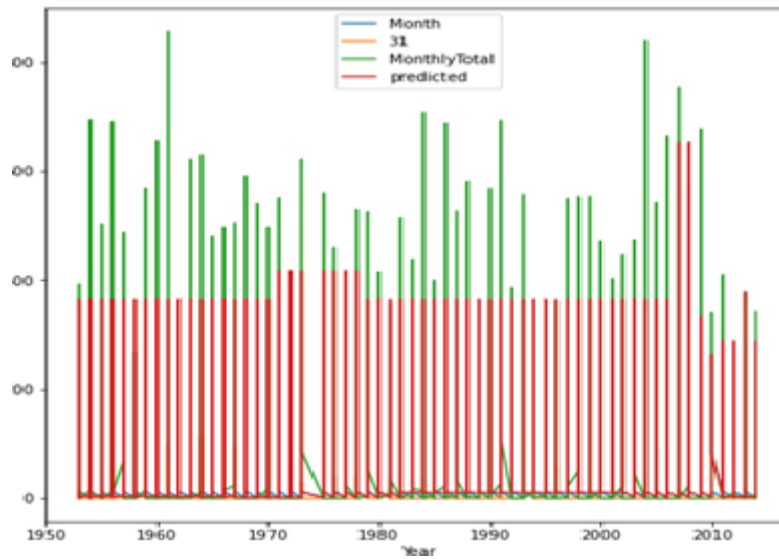


Figure 4.20: Decision Tree on Rainfall data of Barisal

4.11.5 Decision Tree on Rainfall data of Rajshahi

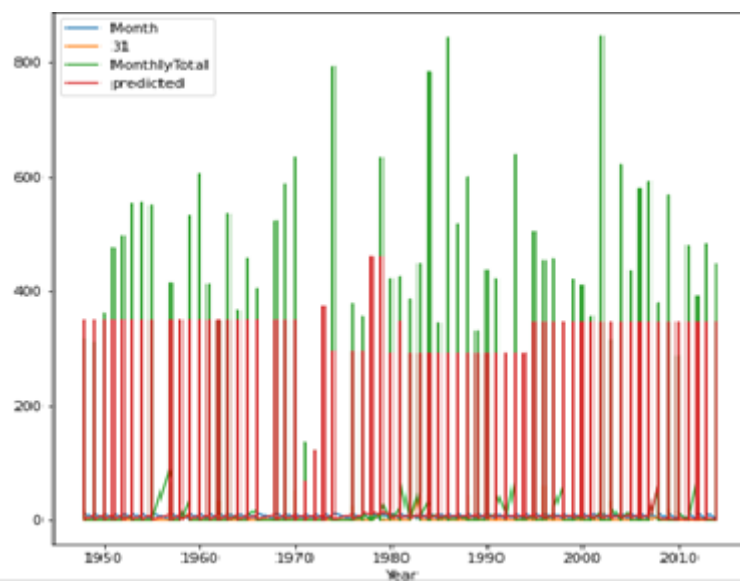


Figure 4.21: Decision Tree on Rainfall data of Rajshahi

4.11.6 Decision Tree on Rainfall data of Khulna

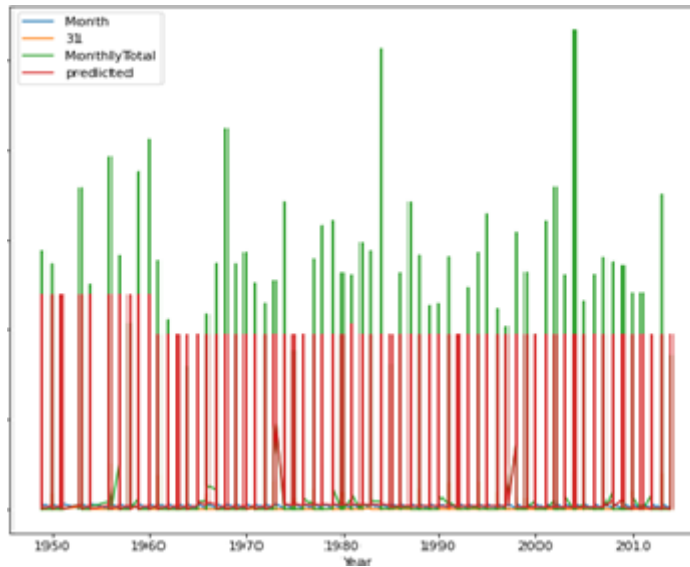


Figure 4.22: Decision Tree on Rainfall data of Khulna

4.12 Comparison between Models

Amongst the three models the performance of the KNN and the Random Forest Regressor was more or less equal as both the values had similar R-squared results. The decision tree in this case performed poorly in comparison with the other two. The reason for that is the decision tree approach to the problem unlike the two algorithms is very heuristical. Thus it can easily overfit the result. But in this case the fitting was underfit as if rainfall values even had slight changes it can cause huge structural change in the model.

On the contrary, when we applied KNN and Random Forest regression on district data we can see KNN performed better than Random Forest. Decision Tree regression did good in the comparison.

Now let's take a look at the r-squared value of all the models for the first 10 and the last 10 iterations that are compiled below in this table

Iterations	R-Squared Value		
	K-Nearest Neighbor	Random Forest	Decision Tree
1	0.641	0.634	0.601
2	0.632	0.660	0.575
3	0.635	0.646	0.587
4	0.615	0.633	0.565
5	0.633	0.656	0.590
6	0.627	0.635	0.592
7	0.634	0.646	0.589
8	0.616	0.633	0.590
9	0.619	0.636	0.563
10	0.624	0.637	0.586
90	0.626	0.629	0.584
91	0.623	0.632	0.593
92	0.598	0.617	0.559
93	0.602	0.629	0.569
94	0.625	0.626	0.566
95	0.612	0.609	0.561
96	0.641	0.664	0.591
97	0.605	0.613	0.580
98	0.633	0.654	0.605
99	0.633	0.650	0.589

Table 4.1: R-square value comparison of 3 models

From the table it is evident that KNN and Random Forest is a better suited model to predict rainfall in comparison to decision tree and KNN. The accuracy score of KNN and Random Forest is also pretty close but accuracy of decision tree is very low compare to the other two.

Now we look at the graphs of r-square scores of 6 regional stations we implemented regression models on:

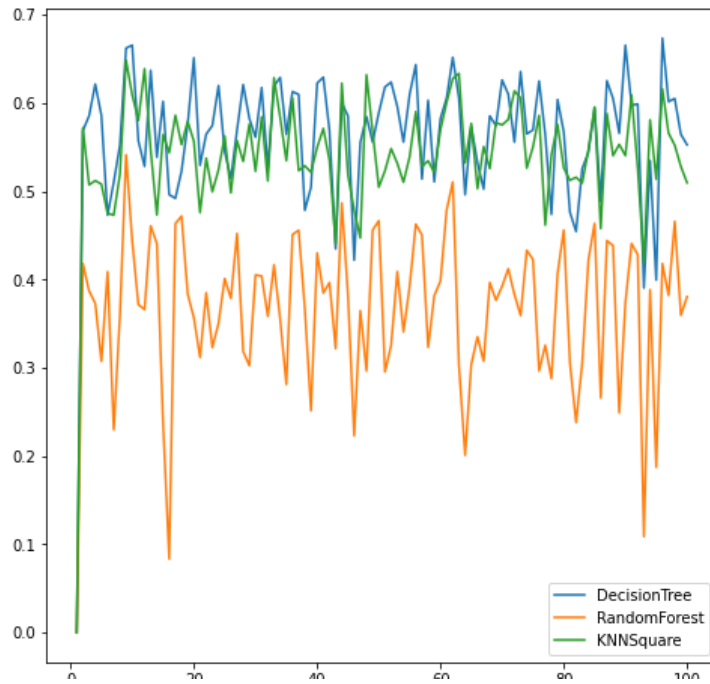


Figure 4.23: R-square value comparison Dhaka

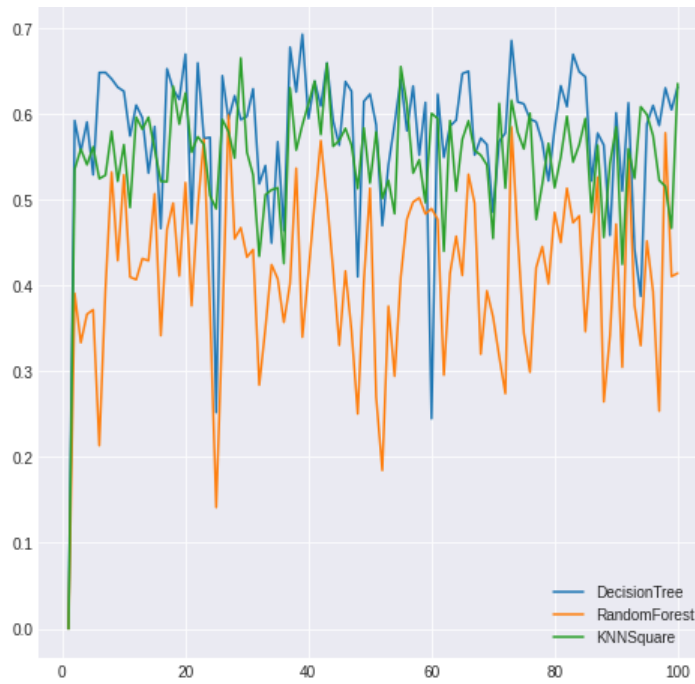


Figure 4.24: R-square value comparison Chittagong

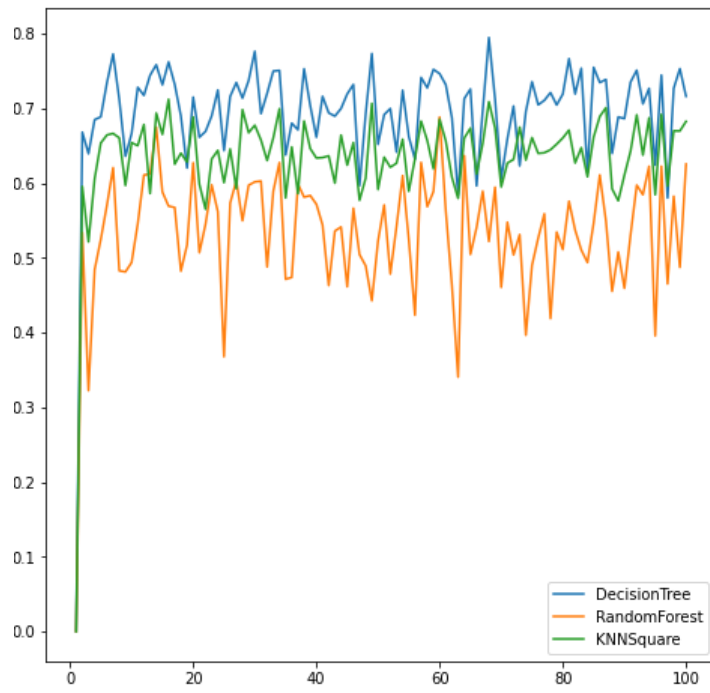


Figure 4.25: R-square value comparison Sylhet

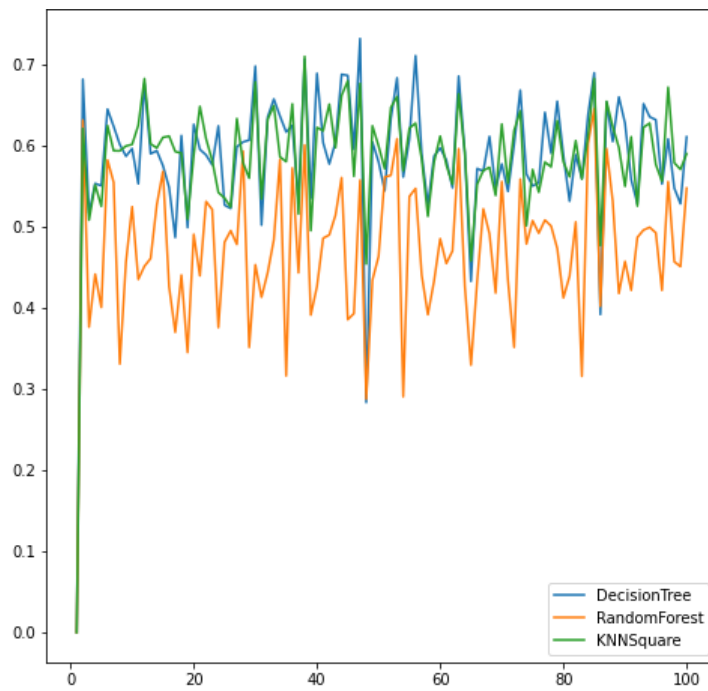


Figure 4.26: R-square value comparison Barisal

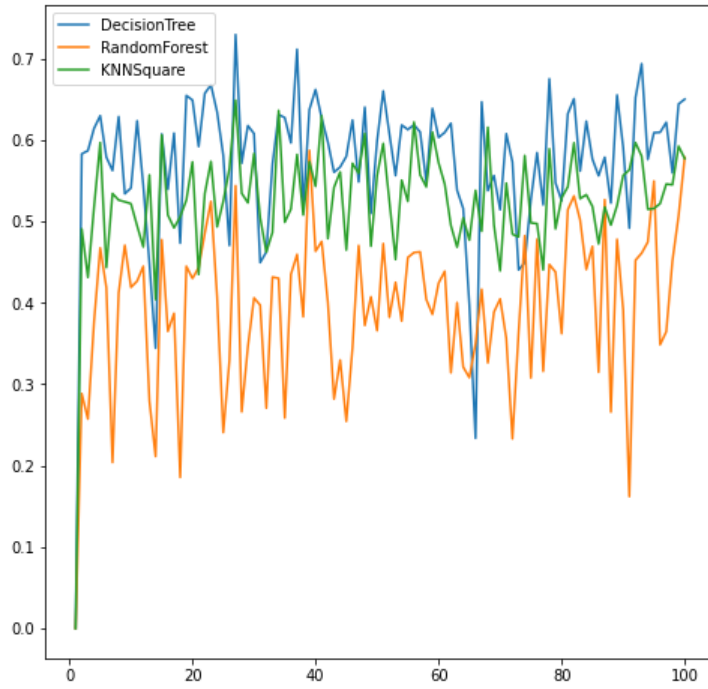


Figure 4.27: R-square value comparison Rajshahi

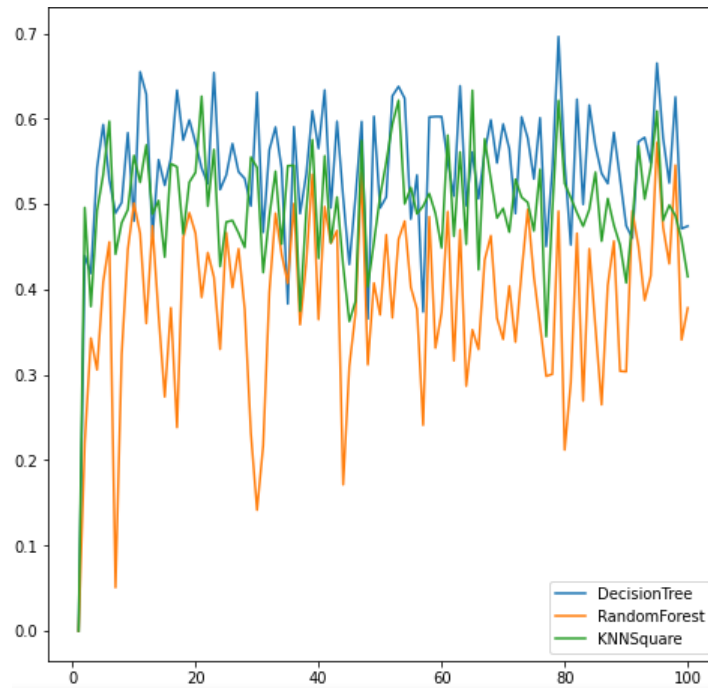


Figure 4.28: R-square value comparison Khulna

So, It is evident from the R-square score evaluation that Random Forest works better on when it is implemented on 35 stations altogether but loses accuracy when implemented on a single station.

KNN works moderately on both the tests, on 35 stations altogether and on the regional station

Decision tree on the other hand has poor results when implemented on 35 station data but gives better results when implemented on a single station data.

Chapter 5

Conclusion

5.1 Conclusion

In this research we tried to understand the trend of the rainfall in Bangladesh and predict the rainfall using machine learning techniques. Due to rise of global warming many countries are facing climate change. Lack of rain is hampering the growth of crop. In a country where agriculture is one of the primary source of earning livelihood and food it is important that the climate change is understood and actions has to be taken before patterns start changing. We analyzed data in this research and data shows the rainfall trend is going down in four regions and going up in 2 regions. Although there is a change in rainfall patterns the change is not significant for now so it is understood from the primary analysis that the rainfall trend in Bangladesh is still similar to the past. We have visualized the dataset to understand the rain statistics in different regions. It shows the eastern region and south eastern region experience the most amount of rain whereas Northwestern regions experience the least. The map of the annual rainfall of Bangladesh was also produced that gave us a clearer picture as to how much rainfall each Station experiences in an year. The results of all our models has been compromised due to poor book keeping by the authorities as variables such as humidity and temperature was absent and it plays a significant part as to how much rainfall a Station would experience. So, our models could not generate the accuracy we hoped for nevertheless the accuracy they generated was all more better than the expected.

Our main goal was to predict the rainfall of Bangladesh using the data of 66 years from 1948 to 2014. We have applied 3 regression model such as KNN, Random Forest and Decision Tree. The best result was obtained by applying Random Forest prediction, KNN showed almost similar results but prediction with decision tree was not up to the mark. Rainfall prediction has been an important research for a while now. We have focused on analyzing the rainfall data and predicting the rainfall which is a very complex field. In this paper we have done:

- Analyzing data of 66 years to understand the rainfall patterns.
- Applied Mann-Kendall test on 6 divisional station's data to understand trend.
- Used KNN, Random Forest and Decision tree on the rainfall data for prediction.
- Applied Random Forest and KNN on rainfall data of 6 district station for better understanding

The models used were validated by dataset from Bangladesh Meteorological depart-

ment. Random Forest regression performed better than KNN and Decision Tree regressions

5.2 Limitations

Rainfall prediction is very complex and tricky thing to execute. Rainfall depends on lots other variables such as humidity, precipitation data, temperature etc. No data can be found with temperature, humidity and precipitation data of every day for a significant amount of time. Without these data it is very tough to predict rainfall in different regions very precisely. Meteorological department has started keeping those data from 2008. With the daily temperature, humidity and wind data we can predict the rainfall more precisely in future.

5.3 Future Work

One of the reason that makes the rainfall prediction important is the yearly flood occurrence in Bangladesh. Bangladesh suffers from flood almost every year due to heavy rainfall and overflow of the rivers due to rainfall. We plan to collect the data of water levels of the rivers flowing through the Bangladesh and the elevation of the lands from water body level from relevant government organization. We plan to incorporate rainfall prediction with the prediction of increasing level of water flowing in river and predict when a region can experience flood. Flood prediction can be very tricky due to lots of variable being involved with it but with ever improving machine learning and neural networking sector we hope to achieve what can make lives of people better.

Bibliography

- [1] L. Breiman, J. H. Friedman, and R. A. Olshen, “classification and regression trees. belmont, ca: Wadsworth,” *International Group*, p. 432, 1984.
- [2] M. XU, P. WATANACHATURAPORN, P. VARSHNEY, and M. ARORA, “Decision tree regression for soft classification of remote sensing data,” *Remote Sensing of Environment*, vol. 97, no. 3, pp. 322–336, 2005. DOI: 10.1016/j.rse.2005.05.008.
- [3] Z. Jan, M. Abrar, S. Bashir, and A. M. Mirza, “Seasonal to inter-annual climate prediction using data mining knn technique,” *Wireless Networks, Information Processing and Systems*, pp. 40–51, 2008. DOI: 10.1007/978-3-540-89853-5_7.
- [4] R. Hoque, D. Nakayama, H. Matsuyama, and J. Matsumoto, “Flood monitoring, mapping and assessing capabilities using radarsat remote sensing, gis and ground data for bangladesh,” *Natural Hazards*, vol. 57, no. 2, pp. 525–548, 2010. DOI: 10.1007/s11069-010-9638-y.
- [5] L. Mckinna and Y. Everingham, “Seasonal climate prediction for the australian sugar industry using data mining techniques,” *Knowledge-Oriented Applications in Data Mining*, 2011. DOI: 10.5772/13904.
- [6] N. Karmeshu, “Trend detection in annual temperature precipitation using the mann kendall test – a case study to assess climate change on select states in the northeastern united states,” *University of Pennsylvania Scholarly Commons*, 2012.
- [7] D. Chauhan and J. Thakur, “International journal on recent and innovation trends in computing and communication,” *Data Mining Techniques for Weather Prediction: A Review*, vol. 2, no. 8, Aug. 2014.
- [8] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, “Breast cancer detection using k-nearest neighbor machine learning algorithm,” *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*, 2016. DOI: 10.1109/dese.2016.8.
- [9] A. Pretorius, S. Bierman, and S. J. Steel, “A meta-analysis of research in random forests for classification,” *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2016. DOI: 10.1109/robomech.2016.7813171.

- [10] S. Aftab, M. Ahmad, N. Hameed, M. Salman, I. Ali, and Z. Nawaz, “Rainfall prediction in lahore city using data mining techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 4, 2018. DOI: 10.14569/ijacsa.2018.090439.
- [11] W. Cherif, *Optimization of k-nn algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis*, Mar. 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050918301376?via%3Dihub>.
- [12] J. Lee, C.-G. Kim, J. Lee, N. Kim, and H. Kim, “Application of artificial neural networks to rainfall forecasting in the geum river basin, korea,” *Water*, vol. 10, no. 10, p. 1448, 2018. DOI: 10.3390/w10101448.
- [13] Ali, Kuriqi, Abubaker, and Kisi, “Long-term trends and seasonality detection of the observed flow in yangtze river using mann-kendall and sen’s innovative trend method,” *Water*, vol. 11, no. 9, p. 1855, 2019. DOI: 10.3390/w11091855.
- [14] B. N. Mahmud, Z. Ferdoush, and L. T. Mim, “Modelling and forecasting energy demand of bangladesh using ai based algorithms,” 2019.
- [15] S. Ronaghan, *The mathematics of decision trees, random forest and feature importance in scikit-learn and spark*, Nov. 2019. [Online]. Available: <https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>.
- [16] Cdkn, *Bangladesh hydrological zone*, Feb. 2020. [Online]. Available: https://cdkn.org/2020/02/feature-bangladeshs-delta-plan-offers-major-opportunity-for-climate-compatible-development/bangladesh-hydrological-zone/?loclang=en_gb.
- [17] X. Qiao and F. Chang, “Research on underground location algorithm based on random forest and environmental factor compensation,” 2020. DOI: 10.21203/rs.3.rs-26537/v2.
- [18] *Geography of bangladesh*, May 2021. [Online]. Available: https://en.wikipedia.org/wiki/Geography_of_Bangladesh.
- [19] [Online]. Available: <https://www.cia.gov/the-world-factbook/countries/bangladesh/>.
- [20] [Online]. Available: http://www.saedsayad.com/decision_tree_reg.htm.
- [21] *Climate - bangladesh*. [Online]. Available: <https://www.climatestotravel.com/climate/bangladesh>.
- [22] *Knn classification using scikit-learn*. [Online]. Available: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>.
- [23] *Rainfall*. [Online]. Available: <http://en.banglapedia.org/index.php/Rainfall>.
- [24] L. Rokach and O. Maimon, “Decision trees,” *Data Mining and Knowledge Discovery Handbook*, pp. 165–192, DOI: 10.1007/0-387-25465-x_9.