Machine Learning Approach to Predict Flood in Bangladesh from Historical Data

Submitted By:

Md. Modacher Mahmud Rafi

ID: 201-35-584

Department of SWE

Daffodil International University

Supervised By:

Mr. Khalid Been Badruzzaman Biplob

Lecturer (Senior Scale)

Department of SWE

Daffodil International University

This Report Presented in partial fulfillment of the Requirements for the Degree of Bachelor of Science in Software Engineering

APPROVAL

This thesis is titled "**Machine Learning Approach to Predict Floods in Bangladesh from Historical Data**", submitted by **Md. Modacher Mahmud Rafi (ID:201-35-584)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Fazla Elahe 10.7.24

Chairman

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information
Technology Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been Md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Md Mostafiz Khan
Managing Director
Tecognize Solutions Limited

DECLARATION

It hereby declares that this thesis has been done by me under the supervision of Mr. Khalid Been Badruzzaman Biplob, Lecturer (Senior Scale), Department of Software Engineering, Daffodil International University. It also declares that neither this thesis nor any part of this has been submitted anywhere else for award of any degree.

Name: Md. Modacher Mahmud Rafi

ID: 201-35-584

Batch: 31st

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Supervised by:

Mr. Khalid Been Badruzzaman Biplob

Lecturer (Senior Scale)

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

With a heart full of gratitude, I first and foremost thank the Almighty for blessing me with the strength, resilience, and opportunity to complete this thesis. Without His grace, this achievement would not have been possible.

I want to express my deepest appreciation to my supervisor, **Mr. Khalid Been Badruzzaman Biplob Lecturer (Senior Scale)**, Department of SWE, Daffodil International University, Dhaka.. Your unwavering support, insightful guidance, and constant encouragement have been pivotal in completing this research. Your expertise and patience have been a source of inspiration throughout this journey.

I would like to express my heartfelt gratitude to the Head of the Department of SWE, along with my supervising teachers and seniors, for their invaluable assistance in completing our thesis. I extend my appreciation to all the faculty members and staff of the SWE department at Daffodil International University for their support and guidance.

A special thanks to my friends, batchmates and department seniors. Your camaraderie, advice, and support have been invaluable. The moments we shared and the challenges we overcame together have made this journey memorable and rewarding.

To my family, your unwavering love and belief in me have been my greatest source of strength. Thank you for your endless support and encouragement, my pillars throughout this journey.

Lastly, I extend my gratitude to everyone who has helped me through direct assistance or moral support. Your kindness and encouragement have made a significant difference in my journey.

ABSTRACT

Floods are among the worst natural disasters, wreaking havoc on people's lives, property, infrastructure, and socioeconomic systems. Floods are a yearly nightmare in Bangladesh, causing varying degrees of damage from minor disruptions to total destruction. Bangladesh is a developing nation with a fragile economy, making it challenging to implement effective flood control measures for the major rivers that traverse the country. Numerous studies have focused on developing reliable flood forecasting systems and flood control. Still, it is difficult to predict the exact timing and volume of floods in real time. Accurate flood prediction requires the integration of computationally complicated flood propagation models with large amounts of data in order to detect water levels and velocities over large areas. By using machine learning algorithms to better accurately anticipate floods, this study aims to reduce the substantial risks associated with these events. The study will provide a more precise understanding of flood patterns by examining many factors, such as geographical location, year, monthly rainfall, and other hydrological and climatic variables. The research will make use of many machine learning models, including XGBoost, MLP Regressor, Gradient Boosting, Random Forest, Polynomial Regression, and K-Nearest Neighbors. We will carefully analyze each model's output to determine the optimal model for flood prediction. This study will not only forecast but also provide policy recommendations to help people be ready for and protect themselves against Bangladesh's devastating floods. We think that this comprehensive plan will help mitigate the effects of one of the hardest-hit natural catastrophes in the country.

Keyword: Flood, Forecasting, Rainfall, Polynomial Regression, Random Forest, K-Nearest Neighbors, Gradient Boosting, XGBoosting, MLP Regressor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES **PAGE**

CHAPTER 1: INTRODUCTION

1.1. Background:

Natural disasters exert an extensive impact on the lives, properties, and environment of humans, leading to financial consequences. The latest data published in September 2017 by the Centre for Research on the Epidemiology of Disasters (CRED, 2017) indicates that a total of 149 disasters took place in 73 countries during the first half of 2017. The event caused 3,162 fatalities, impacted over 80 million individuals, and led to property damages exceeding US$ 32.4 billion. The research states that the primary calamities that took place in Asia, South America, and Africa were floods and landslides. 44% of the events saw flooding, which was responsible for 52% of the fatalities and 44% of the economic losses. This makes flooding the most costly form of disaster. Furthermore, 11% of the incidents were attributed to landslides, resulting in 25% of the overall fatalities. Floods are a recurring natural calamity that affects nearly every region of the world.

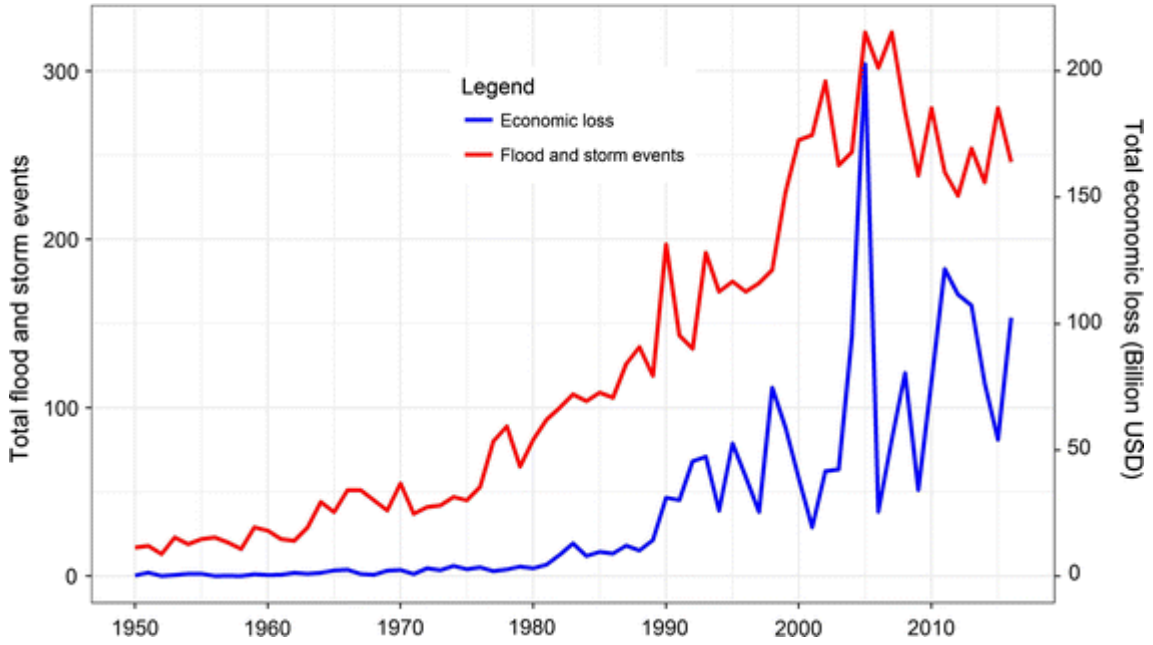

Figure 1.1 : Economic Damages due to the flood worldwide 1950-2020

The incapacity of rivers and streams to efficiently transport water during exceptionally heavy rainfall is the primary cause of floods. Conversely, heavy rains may sometimes cause flooding. They may be the result of other natural or artificial occurrences. For instance, storm surges associated with tropical cyclones, tsunamis, or high tides may induce inundation along coastlines, particularly when rivers are streaming at an elevated rate. Floods are a serious hazard to lives, livelihoods, and infrastructure when they result from extreme hydrological

and meteorological phenomena that happen at surprisingly large volumes and speeds. Bangladesh is predominantly low-lying and susceptible to floods, with 80% of its territory being considered a flood plain [18]. Floods may cause a tremendous deal of damage. Many buildings are insufficiently strong to resist the force of water when a river exceeds its banks or the sea surges ashore. We can raise and lower trees, buildings, bridges, and automobiles.

In July 2007, massive floods in Bangladesh devastated nearly a million homes [19]. Indeed, as the water level drops, floods can cause far more damage than just architectural issues. Dangerous materials have a chance to contaminate water, which may result in the development of several deadly water-borne diseases, such as malaria, cholera, and typhoid. These materials include things like untreated sewage, gasoline, insecticides, and even sharpened debris.

1.2. Cause and effect of flood

Climate change has a significant impact, and one of its most alarming effects is the increased severity of floods. Low-lying regions and localities with insufficient drainage infrastructure become the primary locations of conflict, bearing the full force of these powerful waters. An astonishing survey uncovers a disheartening reality: millions of people are subjected to this annual danger. Only in India, a staggering 4.84 million individuals reside under this persistent menace. Bangladesh and China had 3.84 million and 3.28 million vulnerable individuals respectively, according to recent data[20]. The peril extends beyond these highly populated areas. Urban areas worldwide are susceptible to the effects of increasing sea levels, especially areas situated on ground with an elevation of less than 10 meters. Countries such as the Netherlands, Monaco, and Bahrain serve as prominent illustrations. The annals of history are replete with harrowing tales of the devastating power of floods. Between 1997 and 2008, Australia experienced floods that resulted in the loss of at least 73 lives[21]. Annually, the United States faces a comparable situation, with approximately 100 fatalities and $7.5 billion in damages. Nevertheless, there is no individual occurrence that portrays a more devastating and sorrowful image than the catastrophic flood of the Yellow River in China in 1931. This monumental calamity, recorded in history as one of the most severe natural calamities, claimed the lives of almost one million individuals and uprooted millions more. If we do not take action, the future appears grim. By 2030, the World Resource Institute forecasts a bleak outlook: more than 147 million individuals globally could be affected by floods, resulting in economic consequences ranging from $174 billion to an

astonishing $712 billion in annual urban property destruction. Floods are not only minor inconveniences; they are unrelenting forces that cause widespread destruction to economies and societal structures. In 2020, there was a frightening convergence of crises: the COVID-19 epidemic and catastrophic floods. Vietnam, Laos, and Cambodia saw the full impact of this dual setback. In Vietnam, approximately 90 fatalities occurred, while an indeterminate number of individuals were reported as missing. In Laos, entire villages were completely drowned, and in Cambodia, 25 lives were lost and 40,000 people were displaced[22]. Bangladesh also encountered its own share of challenges. From late June 2020 forward, floods have impacted about 1 million individuals across 13 districts. Evacuations were urgently required, resulting in the displacement of more than 3.3 million individuals. An additional burden was imposed by the absence of potable water for more than 731,000 people. Regrettably, a total of 93 lives were lost in this incident, primarily due to drowning, with children being the most affected by this unfortunate event[23]. Given the regularity of natural disasters such as floods, one may assume that established recovery protocols are in place. Nevertheless, the stark truth is that the identical obstacles resurface during every significant flood occurrence. These calamities result in a complex and diverse damage, affecting both the economic and social welfare. Precisely predicting river water levels following intense precipitation is crucial for safeguarding public safety, tackling environmental issues, and efficiently managing water resources. Mathematical models, constructed based on either fundamental laws of physics or statistical analysis, have been created for this objective. However, these models frequently lack precision and necessitate a substantial amount of time to provide outcomes.

Throughout the course of thousands of years, mankind has made great efforts to avert and control the occurrence of floods. Nevertheless, the intrinsic unpredictability of these natural events presents a substantial obstacle to conventional, statistically-based approaches. Machine learning arises as a promising solution in this context. By utilizing the capabilities of this technology, we have the potential to enhance flood prediction and enhance our ability to effectively mitigate the associated hazards.

1.3. Aims and Objectives

Envision an inexorable deluge, sweeping across expansive regions, causing widespread destruction in its aftermath. This is the harsh truth of floods, a periodic distress that afflicts people globally. Annually, these devastating floods cause significant social and economic damage, resulting in the destruction of people's lives and sources of income. The numbers present a somber depiction: flood damage constitutes more than one-third of all economic losses resulting from natural disasters worldwide. Floods, frequently accompanied by violent gusts, are the most prevalent natural disasters, consistently posing a hazard. According to the Organization for Economic Cooperation and Development, floods cause over $40 billion in damage every year, and there has been a significant increase in the number of deaths, reaching a startling 100 annually[24].

Researchers have determined that the southeast of Asia is particularly vulnerable to dangers and threats. Because of its geographic position, approximately fifty percent of India's population regularly faces a serious danger from floods in this region. The terrible flooding of China's Yangtze River basin serves as a stark reminder of the power of nature. This devastating event destroyed 5 million homes and uprooted 200 million people, resulting in about 3,500 deaths. Statistics provide additional evidence of the significant economic harm floods cause. The 2011 floods in Thailand have the sad distinction of being the most economically damaging disaster between 1900 and 2013, with an estimated cost of $40 billion[25].

Bangladesh, a country located in South Asia, is highly vulnerable to floods due to its climatic conditions. Consequently, it has experienced the full impact of these natural calamities on multiple occasions. An example of this is the catastrophic flood of 1998, which inundated an astonishing two-thirds of the nation for a duration exceeding two months[26]. This calamity devastatingly reversed numerous recent social and economic progressions of the nation. Research indicates that the susceptibility of Bangladesh is primarily derived from its geographical location. The majority of the property is at a low elevation, with more than 80% located in areas prone to flooding. Envision volatile and nomadic rivers, along with the perpetual peril of escalating sea levels. The situation is further aggravated by deforestation, river damming, and uncontrolled building in flood-prone areas with inadequate engineering. According to research, Bangladesh had a total of 78 floods between 1971 and 2014, resulting in the deaths of 41,783 individuals.

Floods pose a significant risk to the agricultural industry, which plays a crucial role in Bangladesh's economy. Approximately half of the population depends on this industry for their means of living, and it makes an astonishing 80% of the country's overall exports. During periods of flooding, these agricultural farms have significant economic losses, which have a detrimental influence on the country's Gross Domestic Product (GDP)[27]. Floods result in a series of subsequent consequences, such as the contamination of potable water, the devastation of vital infrastructure such as bridges and highways, and the proliferation of waterborne illnesses.The user's text is [27].

The flood hazard in Bangladesh is worsened by the country's increasing population density, as well as causes such as riverbank erosion and a diminishing land area. Urbanization in flood-prone locations has both positive and negative consequences. It presents economic prospects, but also exposes a larger population to potential hazards. Inadequate coordination between land management and flood risk management strategies, along with deteriorating stormwater infrastructure unable to handle higher runoff volumes, collectively contribute to an increased susceptibility to floods.

Due to the severe repercussions of floods, there is an urgent want for a more precise method to forecast them. This would facilitate prompt implementation of evacuation protocols, thereby reducing both human casualties and infrastructure damage. This research suggests investigating the capacity of machine learning models, with a specific emphasis on Multiple Logistic Regression, to improve the accuracy of flood prediction. The objective of the study is to evaluate if Binary Logistic Regression can offer more precision in flood prediction in comparison to alternative models.

Envision a world in which communities possess a viable opportunity to combat floods. The primary aim of this project is to enhance the accuracy of flood forecasting by utilizing machine learning techniques. The suggested methodology utilizes Binary Logistic Regression, a classification model that has been developed using past flood data. The study tries to determine the optimal model parameters for predicting flood occurrences by fitting the training data to the model. Conventional machine learning techniques such as Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and Logistic Regression have shown promising outcomes in obtaining high levels of accuracy.

The research will encompass multiple pivotal stages. Initially, the flood dataset will be examined and classified according to several parameters. This investigation will aid in

determining a suitable target variable for the purpose of data modeling. Prior to inputting the data into the machine learning algorithms for training, the researchers will resolve any missing values and manage categorical characteristics. In order for machine learning models to perform optimally, it is necessary to have numerical data that has minimal missing values.

The primary objective of this research is to optimize the precision of flood prediction through the utilization of Binary Logistic Regression. Once the model has been trained on the prepared data, the researchers will assess its efficacy by utilizing a validation set to ascertain its accuracy in forecasting future floods. This discovery has significant potential for Bangladesh and other countries prone to flooding. By providing a possibly more precise flood prediction tool, it could ultimately result in enhanced readiness and decreased flood occurrences.

1.4. Problem statements

The global impact of flood disasters is unparalleled, affecting regions worldwide, including Bangladesh. In Bangladesh, floods are the most recurrent and devastating disaster affecting society[18]. Annually, the pertinent government agencies and groups convene a meeting before the monsoon season to deliberate on the implementation of the action plan to combat the impending flood. Prior access to flood prediction data can assist water managers and decision-makers in proactively planning and preparing for appropriate measures. Therefore, it is crucial to own flood forecasting and prediction data beforehand. Given the presence of unavoidable uncertainties, it is crucial to have early flood prediction in order to obtain accurate data on floods. Primary Goal to solve this problem is

- Which flood prediction approach and parameters are necessary for predicting flood models?
- How to design and develop flood forecasting models, and what are the techniques that can be implemented?

Hydrological modeling is primarily utilized to forecast behavior and enhance comprehension of hydrological processes. There are two types of ongoing research on hydrological modeling. One type is based on physical theories and requires a large number of parameters to describe the physical characteristics of the catchment, such as soil moisture content, initial water depth, topography, topology, and dimensions of the river network (Devia et al., 2015; Dibike et al., 2001). Conversely, empirical models or data-driven models are regarded as observation-oriented models that solely rely on available data without taking into account the

characteristics and processes of the hydrological system. The mathematical equations used in this context are generated from input and output time series, rather than being based on the physical processes of the watershed. Therefore, numerous machine learning approaches can be applied in this approach.

CHAPTER 2: LITERATURE REVIEW

2.1. Literature Review

Reducing risk and offering possibilities, along with minimizing the damage that floods do to people and property, is the aim of current research on machine learning (ML) models for flood prediction. [1] [2] [3] [4]. The scientists are starting to favor machine learning (ML) strategies because they are capable of reproducing the intricate mathematical models of flood dynamics, leading to improved performance and a more affordable solution. Research has shown the value of data-driven methods for forecasting floods, with a move toward modeling machine learning algorithms applying historical climate data to provide more accurate forecasts. We are using many machine learning techniques, including multiple linear regression, K-nearest neighbor, support vector machine, and decision tree classifier, to enhance the precision of flood forecasting. These algorithms provide valuable information on the most effective models for both short- and long-term flood forecasting. The literature lists several efficient methods, such as hybridization, data deconstruction, algorithm ensembles, and model optimization, to improve machine learning-based flood models for forecasting.

The use of machine learning has shown a great deal of promise for improving the precision of real-time flood prediction. Studies show that we can achieve high-accuracy machine learning models like polynomial regression, support vector regression, decision tree, K-nearest neighbor, and stacked generalization. The most effective of these models is the layered generalization model, which has an accuracy rate of 93.3%. The user's text is "[5] [6]." Moreover, studies have shown that employing generative adversarial networks (GANs) to generate fake flood time series data, such as TimeGAN and Real-world Time Series GAN (RTSGAN), can increase the accuracy of traditional machine learning models and decrease inaccurate predictions [7]. Furthermore, the development of surrogate machine learning models like the XGBoost Regressor (XBGR) has made accurate flood inundation predictions

possible. This surpasses the computational limits of traditional 2D hydrodynamic models and yields a high degree of accuracy [4]. When combined, these findings show that machine learning could significantly boost the efficiency of real-time flood prediction systems.

Machine learning greatly improves flood prediction, which reduces the effects of natural catastrophes by providing precise and timely forecasts. India has employed numerous machine learning approaches, including polynomial regression, support vector machine, decision tree, K-nearest neighbor, and stacked generalization, to forecast floods. These models have achieved a remarkable accuracy of 93.3% [8] and [10]. In addition, the development of prediction systems that use rainfall data aids in anticipating the onset of floods, thereby reducing financial losses and preventing deaths [11]. Furthermore, studies have shown that combining machine learning models with enhanced flood inventory data gathered through remote sensing techniques enhances the accuracy of predicting flood susceptibility. Notably, the use of models such as artificial neural networks, k-nearest neighbors, random forests, gradient-boosting decision trees, and support vector machines has resulted in a very dramatic rise in performance [4]. More advanced methods, like enhanced principal component analysis, have made accurate flood forecasts. For flood prediction, a 1D-convolutional neural network model with fewer features had a forecast accuracy of 94.24% [12].

Machine learning is critical for early flood detection because it improves flood prediction systems' precision, resilience, and effectiveness. Researchers have made frameworks that make good use of statistical models, hydrodynamic models, and machine learning algorithms [13] to predict the occurrence of compound floods, such as coastal-fluvial floods. Furthermore, by analyzing satellite data, machine learning methods have effectively mapped floods, enhancing the accuracy and promptness of flood detection in areas like Kerala and Maharashtra [14] and [15]. Furthermore, the use of rainfall data in prediction models proved beneficial in terms of forecasting the onset of floods, especially in regions without dependable flood warning systems, such as India [16]. Furthermore, the combination of convolutional neural networks and deep learning has made it possible to monitor and identify floods in real-time with high accuracy, which has helped reduce losses and deaths related to flooding [17].

2.2. Related Work

Two research studies used artificial neural networks' NARX (Nonlinear AutoRegressive with eXogenous inputs) architecture to predict floods. The authors of [28] created a model that primarily relies on the water levels of upland rivers as input variables, recognizing their critical role in flooding. The authors evaluated the predictions on the remaining samples after training on a portion of the data. It predicted floods five hours in advance and scored 73.54% accuracy. Similarly, the research reference [29] uses a NARX neural network to simulate and forecast flood water levels. Their model includes four variables: upstream river water levels, water levels at the flood location, and the differences between these water levels. The scientists claim that after dividing the data into separate training and testing sets, their model can identify flood water levels with an accuracy of 87%.

An Indian study [30] explores the use of various machine learning algorithms to predict floods, considering factors such as water velocity, temperature, humidity, and rainfall. Researchers compare the effectiveness of a Deep Neural Network (DNN) with other machine learning methods, including Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Naive Bayes. The scientists developed a confusion matrix to establish the optimal technique for flood detection, prioritizing temperature and rainfall as crucial parameters. The study found that SVM had the highest accuracy rate (85.57%), followed by Naïve Bayes (87.01%) and KNN (85.73%). Notably, the DNN surpassed all other approaches with an outstanding accuracy rate of 91.18%.

Another research [31] used an artificial neural network called a back propagation neural network (BPN) to estimate flood levels. The model underwent training using water level data collected from many stations located in Johor, Malaysia. Nevertheless, the preliminary outcomes were dissatisfactory. In order to improve the precision, the researchers integrated an Extended Kalman Filter (EKF) with the outcomes of the BPN model, resulting in a substantial enhancement of the model's performance. Furthermore, a different study [32] employed the Extended Kalman Filter (EKF) technique to forecast flood water levels. This approach consists of two stages: prediction and update. During the prediction step, the system utilizes past data estimations, and in the update stage, it rectifies the projected values based on feedback. The method, which incorporates water level data for flood prediction, attained a Root Mean Square Error (RMSE) of 0.9236 m.

A comparative study [33] examined the predictive precision of Support Vector Machines (SVM) and Artificial Neural Networks (ANN) using flood occurrence data from Bangladesh. The study revealed that Support Vector Machines (SVM) exhibited comparable performance to Artificial Neural Networks (ANN), and in some cases, even outperformed them, particularly for larger prediction intervals. The SVM algorithm was trained and tested using data collected from five water level stations in Bangladesh. The study observed that the Artificial Neural Network (ANN) method necessitated a greater amount of time and exertion, entailing a higher degree of experimentation and refinement in comparison to the Support Vector Machine (SVM) technique. Furthermore, the restricted access to data has a direct influence on the precision of the Artificial Neural Network (ANN) model. Conversely, SVM yielded superior and faster outcomes using a smaller dataset, which is vital for accurate flood forecasting. The paper also emphasized the computational advantages of Support Vector Machines (SVM), as it is capable of processing input variables in a dual form. This makes it particularly useful for real-time forecasting tasks that include substantial volumes of data.

CHAPTER 3: MATERIALS AND METHODS

This chapter explores the captivating realm of constructing a flood prediction model! We will examine the utilization of diverse data types in numerous methodologies to construct this model. By utilizing data from the Bangladesh Meteorological Department, we may employ sophisticated algorithms such as Random Forest(RF), Support Vector Machine(SVM), and Multiple Linear Regression(MLR) to train the model in detecting floods. The primary inquiry at hand is whether Binary Logistic Regression may surpass previous methods by providing more precise predictions and minimizing errors. In order to determine the most accurate approach, we will conduct experiments including multiple data cleansing techniques and evaluate the performance of several machine learning models.

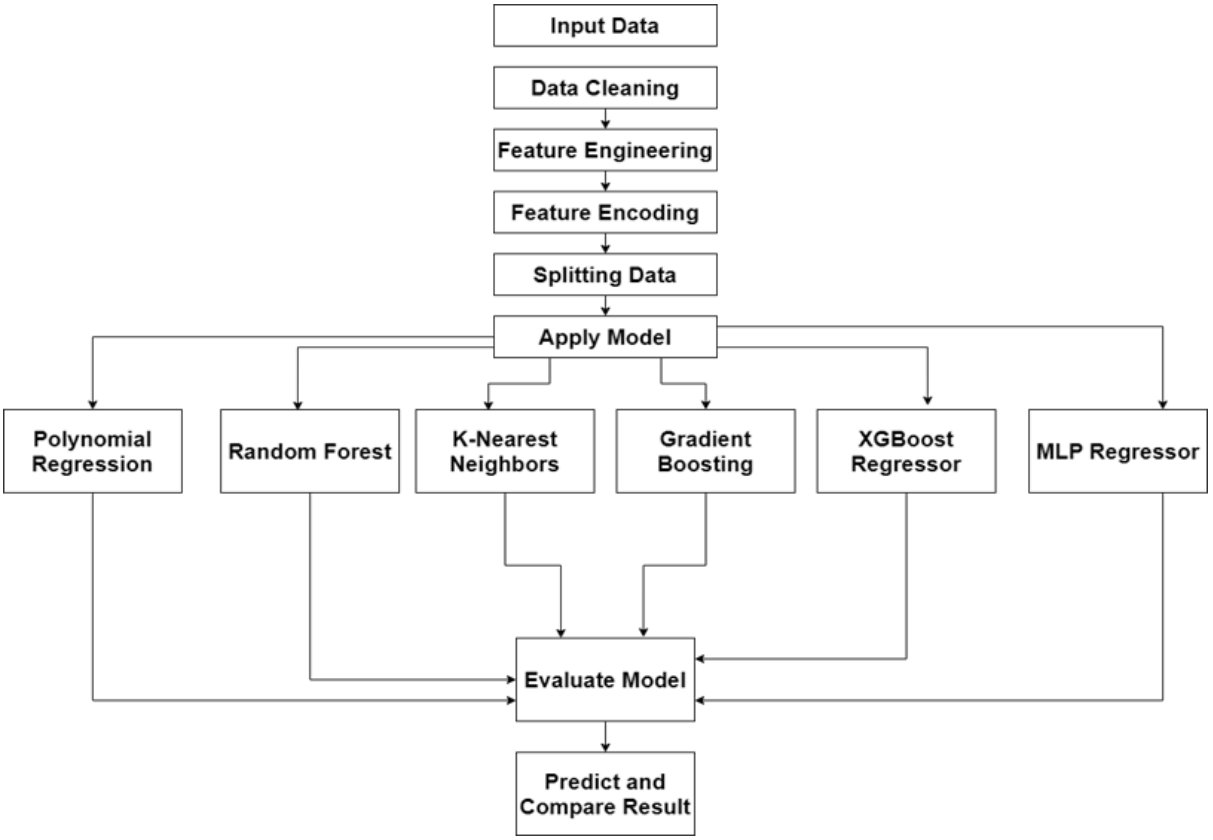

Figure 3.1: Workflow of this Study

3.1. DARA COLLECTION

The Bangladesh Meteorological Department, located in Dhaka, Bangladesh, provided the data for this project. This federal agency is essential to tracking and predicting weather patterns throughout the nation. Its major goal is to reduce the loss of life and property by providing early and accurate predictions for different natural catastrophes such as floods, cyclones, and severe weather events.

The department uses a range of data sources and procedures to do this. We rigorously analyze rainfall readings, satellite photos, and other meteorological parameters to provide reliable predictions. These predictions are critical for educating the public and government officials so they can take preventative action to protect communities from the harmful consequences of natural catastrophes. The Bangladesh Meteorological Department's work is critical in encouraging disaster preparation and strengthening the population's resilience in the face of more unpredictable weather patterns caused by climate change.

3.2. DATASET DESCRIPTION

The Bangladesh Meteorological Department (BMD), a crucial entity responsible for collecting and maintaining comprehensive meteorological data throughout the country, supplies the data for this dataset. This dataset covers the years 1948 to 2013 and includes comprehensive monthly averages of key meteorological variables. It documents the highest and lowest temperatures ever recorded in Bangladesh during this period, together with essential data on precipitation, relative humidity, wind velocity, cloud cover, and hours of direct sunlight. This diverse array of parameters offers an extensive overview of the nation's climate dynamics across many decades.

The dataset includes geographical data for each weather station, including latitude, longitude, altitude, and X and Y coordinates. These spatial IDs accurately link the data to specific locations across Bangladesh, providing localized insights into meteorological trends and patterns.

I employed specific techniques to fill in missing values in the dataset, addressing any gaps and ensuring that the data was as complete and trustworthy as possible for analysis. The collection comprises daily rainfall and flood situation data obtained from 34 distinct meteorological stations across Bangladesh. The abundant temporal and geographical data

facilitates an exhaustive analysis of meteorological patterns and their potential impacts on the region's flooding trends and climatic variability throughout time.

	Unnamed: 0	Station Names	YEAR	Month	Max Temp	Min Temp	Rainfall	Relative Humidity	Wind Speed	Cloud Coverage	Bright Sunshine	Station Number	X_COR	Y_COR	LATITUDE	LONGITUDE	ALT	Period
0	0	Barisal	1949	1	29.4	12.3	0.0	68.0	0.453704	0.6	7.831915	41950	536809.8	510151.9	22.70	90.36	4	1949.01
1	1	Barisal	1950	1	30.0	14.1	0.0	77.0	0.453704	0.8	7.831915	41950	536809.8	510151.9	22.70	90.36	4	1950.01
2	2	Barisal	1951	1	28.2	12.3	0.0	77.0	0.453704	0.6	7.831915	41950	536809.8	510151.9	22.70	90.36	4	1951.01
3	3	Barisal	1952	1	26.6	12.3	2.0	77.0	0.453704	1.0	7.831915	41950	536809.8	510151.9	22.70	90.36	4	1952.01
4	4	Barisal	1953	1	30.0	13.3	10.0	75.0	0.453704	1.6	7.831915	41950	536809.8	510151.9	22.70	90.36	4	1953.01
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21115	21115	Teknaf	2009	12	30.3	16.5	0.0	72.0	2.800000	0.9	8.700000	41998	734765.4	308914.1	20.87	92.26	4	2009.12
21116	21116	Teknaf	2010	12	31.7	16.7	193.0	79.0	2.400000	1.4	5.500000	41998	734765.4	308914.1	20.87	92.26	4	2010.12
21117	21117	Teknaf	2011	12	31.5	16.4	0.0	73.0	0.000000	1.5	7.400000	41998	734765.4	308914.1	20.87	92.26	4	2011.12
21118	21118	Teknaf	2012	12	30.0	15.8	0.0	70.0	1.800000	0.2	9.000000	41998	734765.4	308914.1	20.87	92.26	4	2012.12
21119	21119	Teknaf	2013	12	29.9	16.5	0.0	72.0	3.000000	0.9	8.100000	41998	734765.4	308914.1	20.87	92.26	4	2013.12

21120 rows × 18 columns

Figure 3.2 : Overview of Dataset

3.2. DATA PREPROCESSING

3.2.1 Data Cleaning

Data cleansing is similar to conducting a thorough cleaning and organizing process for our data. The process entails meticulously examining the information and discerning any omissions or inaccuracies, akin to organizing a disordered space.

In this particular scenario, we had a slight difficulty due to the fact that certain months vary in the number of days they include. February typically has 28 days, whereas July consistently has 31 days. Consequently, no precipitation data was documented for the additional days in some months. In order to address this issue, we employed a method known as data imputation, which involves filling in missing values. For example, in order to maintain consistency, we inserted zeros for the missing 30th and 31st days in February 1980, which only had 29 days. We encountered instances of missing data for particular days at specific stations. Specifically, there is no recorded data on the amount of rainfall that occurred on November 14th and 15th, 1983 in Dhaka. Therefore, we substituted zeros for those dates as well. By ensuring the cleanliness of our data, we took measures to assure its quality before including it into our flood prediction model.

3.2.2 Feature Engineering

The process of transforming raw data into a collection of features for use in training a prediction model using a machine learning algorithm is known as feature engineering in machine learning. The main objective is enhancing the performance of the models. It enhances the representation of an underlying problem in predictive algorithms, hence increasing the accuracy of the model for data that has not been observed before. The feature engineering method determines the most effective predictor variables for the model, which includes both predictor variables and an outcome variable [34].

After calculating the monthly precipitation data from the rainfall dataset, it was incorporated as an additional column. We then organized the monthly precipitation data by designated stations and years. The dataset now includes columns for stations, years, and all twelve months. We later added flood data as a new feature to the dataset, aligning it with the stations and years.

	Month	Min Temp θ C	Rainfall (mm)	Relative Humidity %	Wind Speed (m/s)	Cloud Coverage %	Bright Sunshine (Hours)	Station Number
0	1	12.3	0.0	68.0	0.453704	0.6	7.831915	41950
1	1	14.1	0.0	77.0	0.453704	0.8	7.831915	41950
2	1	12.3	0.0	77.0	0.453704	0.6	7.831915	41950
3	1	12.3	2.0	77.0	0.453704	1.0	7.831915	41950
4	1	13.3	10.0	75.0	0.453704	1.6	7.831915	41950
...	...	...	...	...	...	...	...	...
21115	12	16.5	0.0	72.0	2.800000	0.9	8.700000	41998
21116	12	16.7	193.0	79.0	2.400000	1.4	5.500000	41998
21117	12	16.4	0.0	73.0	0.000000	1.5	7.400000	41998
21118	12	15.8	0.0	70.0	1.800000	0.2	9.000000	41998
21119	12	16.5	0.0	72.0	3.000000	0.9	8.100000	41998

21120 rows × 8 columns

Figure 3.3: Feature Engineering

3.2.3 Feature Encoding

Encoding means altering. Most machine learning models can only handle numerical values, but some can handle categorical values. An instance of this is the K-nearest Neighbours Algorithm, which computes the Euclidean distance between two observations of a certain feature. Data fed to an algorithm should be numerical. To ensure accuracy, it is important to convert the categorical values of the relevant feature into numerical values. The procedure is referred to as feature encoding.[35]

The dataset utilized in this investigation includes two string-based features: 'Station' and 'Flood'. Machine learning algorithms often provide superior outcomes with numerical data, necessitating the encoding of text inputs. The 'Station' feature lists the stations where we collected daily precipitation data. We applied label encoding to the 'Station' column to convert these category values into numerical ones, eliminating the need for additional columns. The monthly precipitation data was extracted from the rainfall dataset and then inserted as a new column. We then organized the monthly precipitation data by designated stations and years. The dataset now includes columns for stations, years, and all twelve months. Flood data was then included as a new variable in the dataset, corresponding to the stations and years.

3.2.4 Feature Scaling

To ensure our data didn't favor any of our models, we took additional steps. We utilized a technique known as Standard Scaler to level the playing field. Picture the act of stretching a rubber band - that's essentially what Standard Scaler does to the data. It modifies the values by centering them around an average and equalizing their width. In this manner, each feature is given equal treatment, ensuring that no one feature is favored over another based on its initial scale. There is no specific value for the amount of stretching or shifting required; the formula handles all of it effortlessly [36].

Lastly, we split our data into two groups: a training set, which served as a practice round, and a test set, which acted as the real exam. The training set comprises 80% of the data, while the test set accounts for the remaining 20%. After the split was completed, the features in the training set were extended and shifted using Standard Scaler. We will leave the test set untouched for now, allowing the model to encounter a fresh challenge during the testing phase!

3.3. MODEL ARCHITECTURE

3.3.1. Polynomial Regression

Polynomial regression is an extension of linear regression, where the relationship between the independent variable x and the dependent variable y is modeled as an n-th degree polynomial. This method is useful when the data exhibits a nonlinear relationship that cannot be accurately captured by a simple linear model.

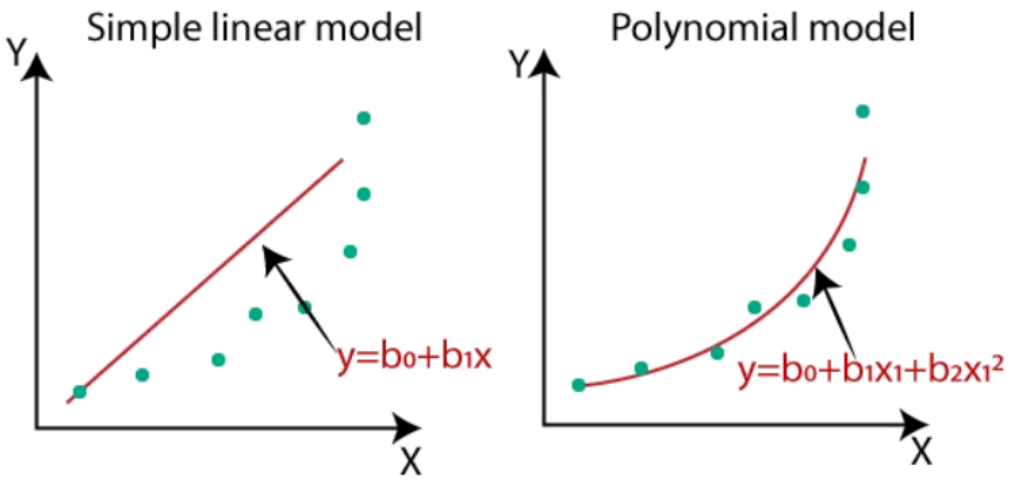

Figure 3.4 : Polynomial Regression

Mathematical Representations:

In linear regression, the model is given by:

$$y = \beta_0 + \beta_{1x} + \epsilon$$

where:

- y is the dependent variable.
- x is the independent variable
- β_0 *and* β_1 are coefficients (parameters) of the model.
- ϵ is the error term, representing the deviation of the observed values from the predicted values.

How It Works:

In polynomial regression, the model takes the form of a polynomial equation. For a single independent variable x, and the model can be expressed as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots\ldots\ldots + \beta_n x^n + \epsilon$$

where:

- y is the dependent variable
- x is the independent variable
- $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_n$ are the coefficient of the polynomial
- $x^2, x^3, \ldots\ldots, x^n$ are the polynomial terms.
- ϵ is the error term, representing the deviation of the observed values from the predicted values.

Mathematical Working

The goal of polynomial regression is to find the coefficients $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_n$ that best fits the data. This is typically done using the least squares method, which minimizes the sum of the squared differences between the observed values y_i and the predicted values $\widehat{y_i}$ from the model:

$$min_\beta = \sum_{i=1}^{m} (y_i - \widehat{y_i})^2$$

where m is the number of observations.

Higher-degree polynomial components in the model allow it to capture more complex interactions between the independent and dependent variables, making it suitable for data that a straight path cannot properly record. Selecting an appropriate degree for the polynomial is

crucial to avoid overfitting, a condition that occurs when the model gets too complex and captures noise instead of the underlying pattern.

3.3.2. Random Forest

Random Forest Regression is an ensemble learning approach that combines many decision trees to provide more accurate and dependable predictions. A forest of randomly chosen decision trees is built, and the outputs are averaged to get the final prediction.

The fundamental principle of Random Forest Regression is to reduce the substantial variation of individual trees by averaging the predictions of numerous decision trees. Each tree is trained on a random subset of features and a random portion of the training data using bootstrapping. This unpredictability enables the trees to decorrelate more effectively, which leads to independent errors.

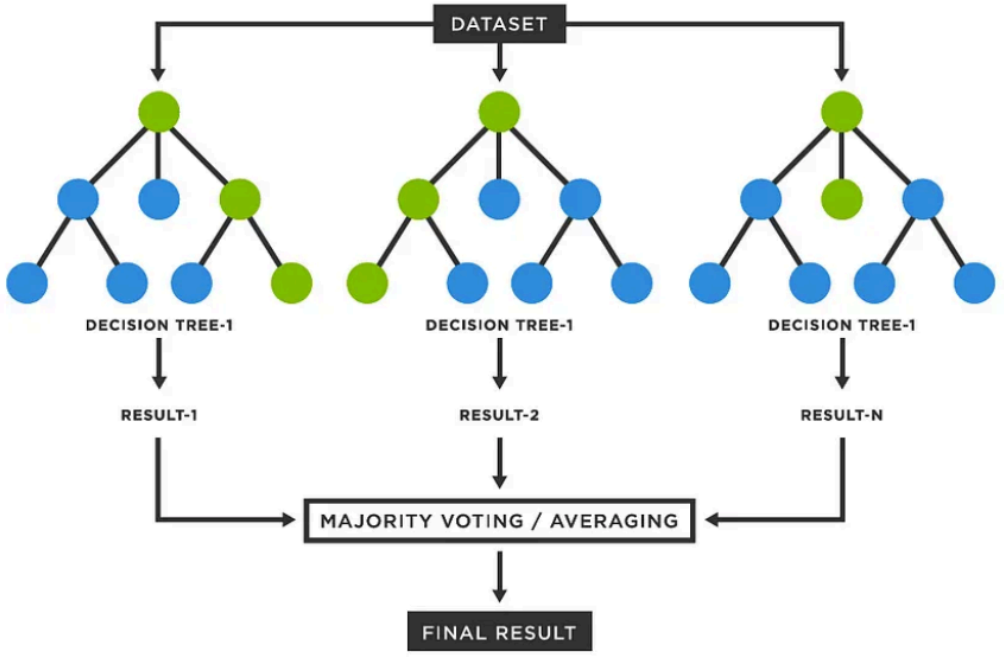

Figure 3.5 : Random Forest

How It Works:

- **Decision Tree:** The core of Random Forest is individual decision trees. Each tree is built using following process:
    - **Feature Selection:** At each node, a subset of features is randomly chosen.

    - **Splitting:** The best split is determined based on a criterion like mean squared error (MSE) for regression.
    - Node Splitting: The dataset is split into subsets based on the selected feature threshold that minimizes the MSE.
- **Bagging (Bootstrap Aggregation):**
    - **Data Sampling:** For training each tree, a random sample(with replacement) from the original dataset is used (bootstrap sample)
    - This introduces variance reduction as each tree is trained on a different subset of data.
- **Ensemble Aggregation:**
    - **Predictive Agregation:** For regression, predictions from all trees are averaged to get the final output

Math Behind It:

Creating Decision Tree:

- Splitting Criterion: At each node, calculate the MSE for possible split:

  $$MSE_{split} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{y_i})^2$$

- The chosen split minimizes the MSE.

Bagging:

- Bootstrap samples: Create B bootstrap sample:

  $$S^{'} = \{ (x_1, y_1), (x_2, y_2), \ldots\ldots\ldots, (x_n, y_n) \}$$

- Each sample set build as individual tree to form: $h_1(S'), h_2(S'), \ldots, h_\beta(S')$

Final Prediction:

- For a new data point x , predict y as the average of predictions from all trees:

  $$\widehat{y} = \frac{1}{N} \sum_{i=1}^{N} h_n(x')$$

Random forest regression is very resistant to overfitting, particularly when a substantial number of trees are used. It quickly processes large datasets and can handle hundreds of input

variables without deleting any. It also offers a feature significance metric, which is helpful in comprehending the underlying data.

3.3.3. K-Nearest Neighbors(KNN)

For classification and regression problems, we use K-nearest neighbors (KNN), a supervised learning technique. It operates by predicting new data point values based on feature similarity with the training dataset. According to a research study [14], KNN is a non-parametric technique that excels in regression prediction problems. The algorithm classifies fresh data points into groups according to their similarity to previously categorized data. KNN determines the distance between data points, often using Euclidean distance as the primary approach.

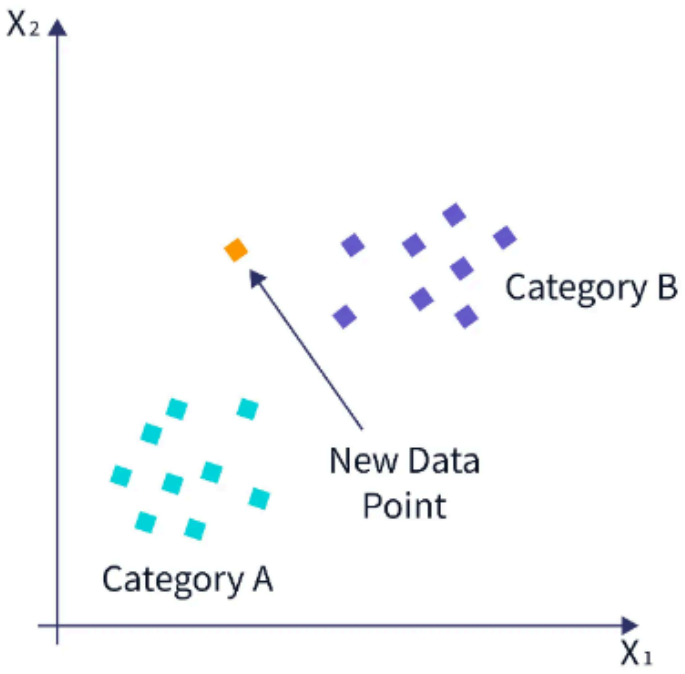

Figure 3.6 : KNN working Principle

Mathematical Foundations of KNN

Distance Calculation:

The core of the KNN algorithm lies in measuring the distance between data points.

- **Euclidean Distance:**

  The Euclidean distance between two points p and q in n-dimensional space is calculated as:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

  where p_i and q_i are the coordinates of points p and q in the n-dimensional space. This distance metric is most commonly used is KNN.

- **Manhattan Distance:**

  This metric sums the absolute difference their coordinates:

$$d(x, y) = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Prediction Mechanism:

Once the distances are calculated, the KNN regressor identifies the 'k' nearest neighbors to input data points. The predicted value $\hat{y}$ for the new data point is computed as follow:

- **Simple Average:**

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

  where the y_i are the target values of the 'k' nearest neighbors.

- **Weighted Average:** In some cases, a weighted average is used, where closer neighbors have a higher influence on the prediction:

$$\hat{y} = \frac{\sum_{i=1}^{k} \omega_i y_i}{\sum_{i=1}^{k} \omega_i}$$

  where ω_i, is a weight assigned based on the distance, often inversely proportional to the distance.

Impact of ‘k’ and Distance Metric

The choice of 'k' has a significant impact on the model's performance. A small 'k' may lead to a model susceptible to noise, while a big 'k' may attenuate significant patterns. Additionally, the selected distance metric may influence the results, as different metrics may capture distinct aspects of the data structure.

In conclusion, the KNN regressor is a robust yet simple algorithm that employs local averaging and distance metrics to generate predictions, rendering it a popular choice for a variety of regression tasks in machine learning.

3.3.4. Gradient Boosting

An ensemble machine learning approach known as gradient-boosting regression by sequentially constructing numerous decision trees. Each successive tree corrects the errors made by the trees that came before it, leading to a unified and resilient prediction model.

Gradient boosting optimizes a loss function by employing gradient descent to reduce residuals on a mathematical level. It combines the predictions of less precise models by assigning suitable weights to minimize the overall error. The technique starts by using a beginning model, often the mean of the goal variable, and gradually integrates supplementary trees that predict the residuals (errors) of the previous models. The technique improves performance by tuning each tree to the negative gradient of the loss function, prioritizing the most difficult scenarios for prediction. Gradient boosting is an effective technique for handling complex and non-linear correlations in data because it is iterative. Regularization techniques, such as varying the learning rate and restricting the tree depth, are often used to reduce overfitting and enhance the capacity to generalize to novel data. Gradient Boosting Regression is a very adaptable and powerful method for predictive modeling, capable of achieving high accuracy on a wide range of regression issues[21].

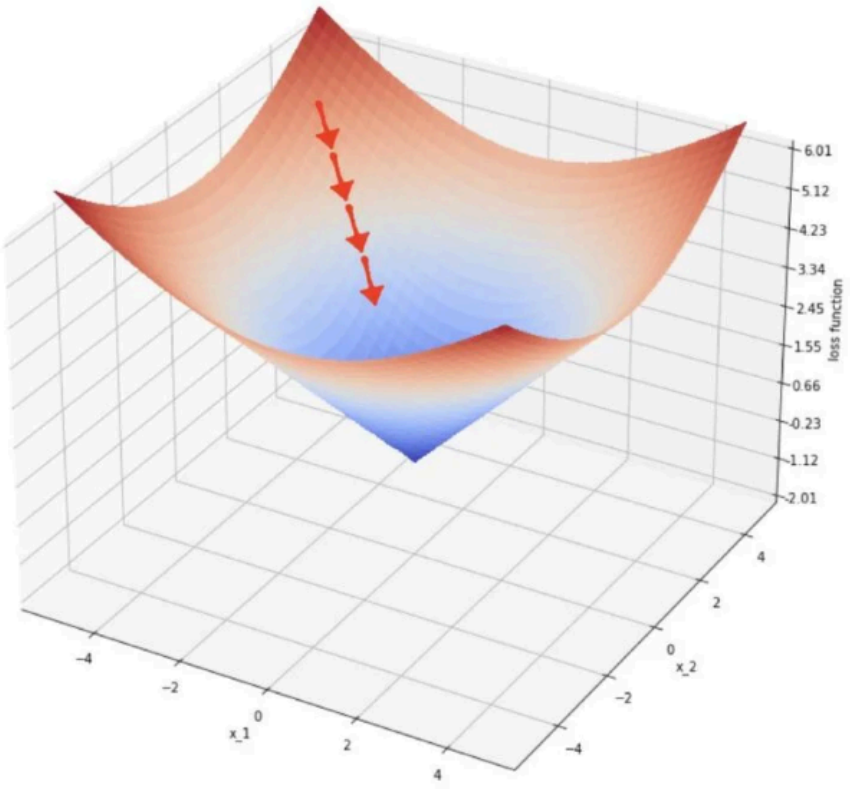

Figure 3.7 : KNN working Principle

The above is an example of a loss function ℓ over a data space defined by (x1, x2). The red arrows indicate the gradients at specific locations on the surface defined by ℓ. Each step in the steepest descent moves our loss value closer to the local minimum for this function.

How It Works:

1. **Initialize with a Base Model**: The process begins with an initial model. Acommon choice is to start with the mean of the target values y. This provides a baseline prediction:

$$\widehat{y}_i^{(0)} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

where,

- $\hat{y}_i^{(0)}$ is the initial prediction for each data point i.
- N is the total number of observation
- y_i are the actual target values.

2. **Compute Residuals**: For each successive iteration *m*, the method computes the residuals. Residuals refer to the disparities between the real goal values and the present forecasts generated by the model. The mistakes that the new model must rectify are represented by these residuals:

$$r_i^{(m)} = y_i - \hat{y}_i^{(m-1)}$$

where,

- $r_i^{(m)}$ are the residuals at iteration m.
- $\hat{y}_i^{(m-1)}$ are the predictions from the model at the previous iteration.

3. **Fit a New Model to Residuals**: A new decision tree (or another weak learner) is trained to predict the residuals computed in the previous step. This new model attempts to capture the patterns in the errors and correct them:

$$h_m(x)$$

where,

- $h_m(x)$ is the new decision tree trained to predict the residuals $r_i^{(m)}$

4. **Update the Ensemble**: The predictions from the new model are added to the previous predictions, scaled by a learning rate α. The learning rate controls the contribution of each tree to the final model and helps in preventing overfitting:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \alpha h_m(x)$$

where,

- α is the learning rate (a small positive number, typically between 0.01 to 0.1).
- $\widehat{y}_i^{(m)}$ are the updated predictions after adding the new model's prediction.

5. **Iterate**: Steps 2-4 are repeated for a specified number of iterations or until the residuals are minimized (i.e., the model's performance stops improving). Each iteration adds a new model that corrects the errors of the combined ensemble of previous models.

Mathematical Optimization

The fundamental concept behind gradient boosting is to enhance the model's performance by minimizing a loss function. The loss function quantifies the discrepancy between the observed target values and the estimated values. Gradient descent iteratively enhances the model's predictions to minimize this loss function.

- **Loss Function**: Specify a loss function $L(y, \widehat{y})$ that requires minimization.
- **Gradient Descent**: Determine the loss function's negative gradient in relation to the model's predictions and adapt a new model to this gradient.

3.3.5 XGBoost Regression

XGBoost, or Extreme Gradient Boosting, is a very powerful and versatile machine learning technique that exhibits outstanding performance in both regression and classification tasks. XGBoost, a member of the ensemble learning family, builds several decision trees in a sequential manner. XGBoost specifically designs each successive tree to correct the errors made by the previous trees. This iterative methodology employs the gradient boosting framework, which enhances the model by minimizing a designated loss function. XGBoost demonstrates exceptional performance in successfully managing large and complex datasets while maintaining a high degree of predictive accuracy. Several advanced optimization procedures achieve this high degree of efficiency. XGBoost utilizes a sophisticated regularization algorithm that incorporates both L1 (Lasso) and L2 (Ridge) penalties. This

strategy specifically aims to reduce overfitting and improve the model's ability to generalize. Furthermore, it utilizes an advanced approach to tackle the problem of missing data, hence improving its ability to handle real-life situations where datasets often include partial information.

Furthermore, XGBoost employs tree pruning and a method known as shrinkage, which involves modifying the influence of each tree by a factor (learning rate) to avoid overfitting the training data. In addition, it enables the parallelization of tree construction, resulting in significant speedups in training, especially for large datasets. XGBoost has a feature called column subsampling, which adds an extra degree of randomization and helps to reduce overfitting while improving model performance[].

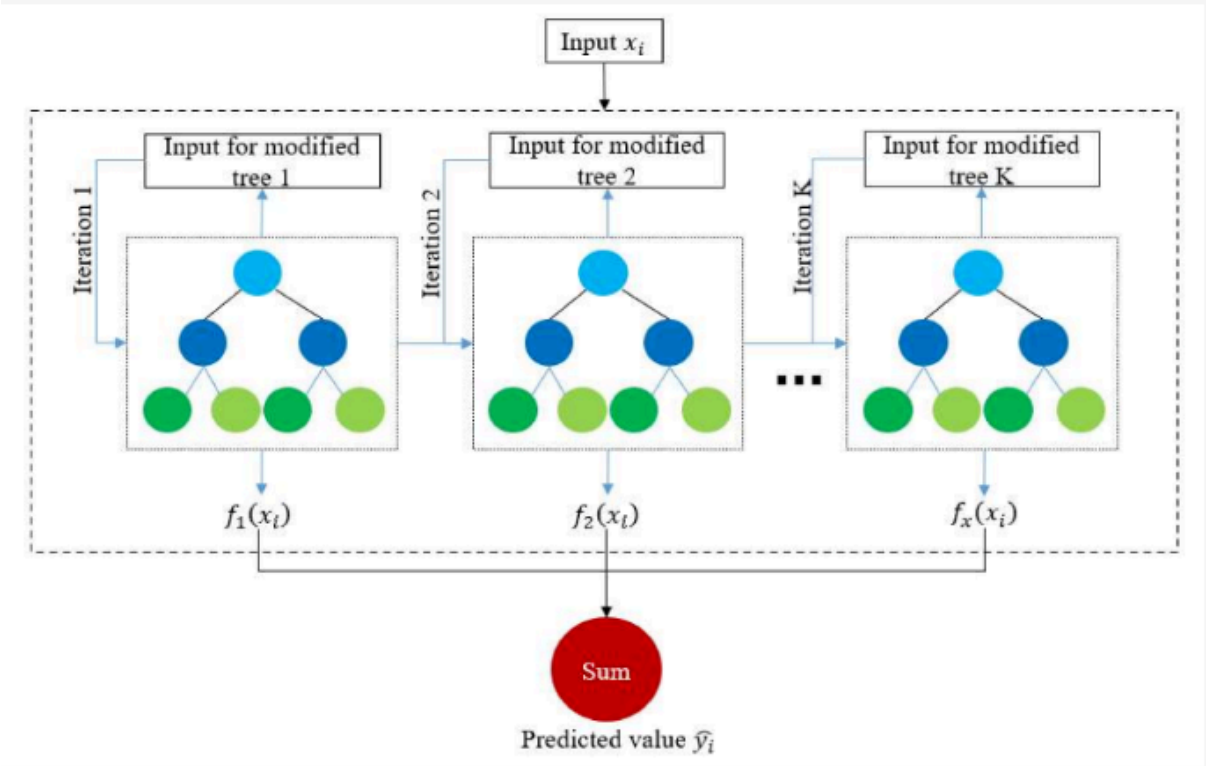

Figure 3.8 : XGBoosting working Principle

How It Works:

1. **Initialization**: Start with an initial model, typically the mean of the target values.
2. **Boosting Iterations**: Sequentially add new decision trees that correct the errors of the existing ensemble. Each new tree is trained on the residuals (errors) of the combined existing models.

3. **Regularization**: XGBoost includes additional regularization terms to prevent overfitting and improve the generalization of the model. This makes it more robust compared to standard gradient boosting.
4. **Model Update**: The predictions from each tree are combined with a learning rate to update the model's overall predictions.

Behind Working Mathematics

1. **Initialization**: The initial prediction for each instance is set to the average of the target values:

$$\hat{y}_i^{(0)} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

2. **Objective Function**: Define a regularized objective function to be minimized. The objective function includes a loss function and a regularization term:

$$obj(\theta) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k)$$

where,

- $L(y_i, \hat{y}_i)$ is the loss function (e.g. mean squared error).
- $\Omega(f_k)$ is the regularization term for the k-th tree.
- θ represents the parameters of the model.

3. **Additive Model**: The model is built in an additive manner, where new trees are added to correct the residuals of the current model:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \alpha f_m(x)$$

where,

- α is the learning rate.
- $f_m(x)$ is the prediction from the m-th tree

4. **Gradient and Hessian Calculation**: For each iteration, compute the gradient (first derivative) and the Hessian (second derivative) of the loss function with respect to the predictions. These are used to fit the new tree:

$$g_i = \frac{\vartheta L(y_i, \widehat{y_i})}{\vartheta \widehat{y_i}}$$

$$h_i = \frac{\vartheta^2 L(y_i, \widehat{y_i})}{\vartheta \widehat{y_i}^2}$$

5. **Tree Construction**: Construct the new tree by splitting the data to minimize the loss function. The split criterion includes both the gradient and Hessian:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

where,

- G_L and G_R are the gradient for the left and right splits.
- H_L and H_R are the Hessians for the left and right splits.
- λ and γ are regularization parameters.

6. **Update Model**: Update the predictions with the new tree's predictions:

$$\widehat{y_i}^{(m)} = \widehat{y_i}^{(m-1)} + \alpha f_m(x)\text{=-d}$$

XGBoost has several noticeable advantages that position it as a preferred choice for machine learning tasks. The system optimizes for computational efficiency, leading to significantly faster performance when compared to standard gradient-boosting systems. To prevent overfitting and enhance generalization on fresh data, it integrates L1 (Lasso) and L2 (Ridge) regularization techniques. XGBoost has exceptional proficiency in efficiently handling missing data, which is critical for real-world datasets that frequently encounter insufficient information. The algorithm employs an advanced method for tree pruning, which includes reducing the size of the tree and selecting just a subset of columns. This enhances its ability

to produce robust models. Furthermore, XGBoost has the capacity to parallelize the process of constructing trees, leading to a significant reduction in training time, especially when working with large datasets. The combination of these attributes synergistically contributes to XGBoost's renowned ability to rapidly generate precise and high-performing models.

3.3.6. MLP Regressor

The Multi-Layer Perceptron (MLP) Regressor is a sophisticated kind of artificial neural network often used for regression tasks in machine learning. The primary goal is to generate a model that can accurately predict continuous output values based on input features. The MLPRegressor comprises an input layer for data reception, one or more hidden layers for computation and learning, and an output layer for generating the final prediction. Each layer's structure consists of interconnected nodes, also known as neurons, connected by weighted connections. These connections allow the network to gather information and adapt by adjusting its weights during data processing. The MLPRegressor effectively analyzes complex patterns and relationships in the data, resulting in accurate predictions for various regression scenarios.

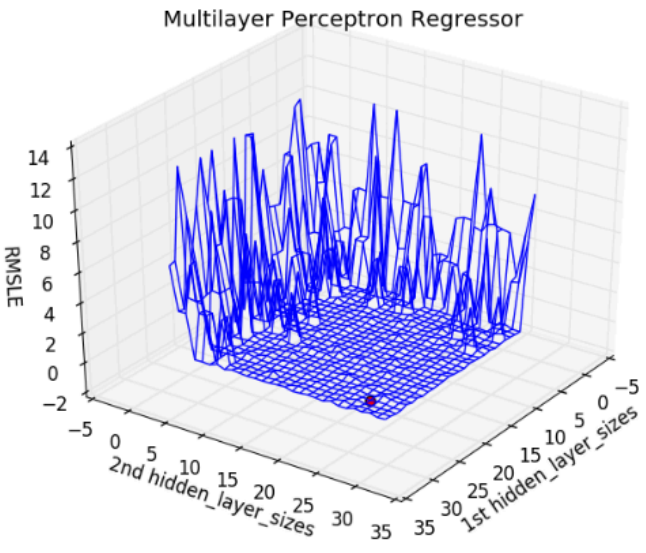

Figure 3.9 : MLPRegressor working Principle

How It Works:

1. **Activation:** An MLP regressor typically comprises an input layer, one or more concealed layers, and an output layer. Each layer consists of many neurons. The

hidden layers process the feature values received by the input layer through weighted connections. Finally, the output layer generates the ultimate predicted value.

2. **Forward Propagation:** The network introduces the input features, then each neuron calculates a weighted sum of its inputs, applies an activation function, and transmits the result to the next layer. This procedure continues until the output layer generates the forecast.
3. **Activation Function:** Activation functions often used in neural networks include ReLU (Rectified Linear Unit), sigmoid, and tanh. These functions include non-linear elements in the model, allowing it to acquire intricate connections.
4. **Loss Function:** The mean squared error (MSE) is a frequently used loss function for regression problems. It quantifies the disparity between the actual target values and the projected values.
5. **Backpropagation:** Using the backpropagation method, the network modifies its weights. This process entails calculating the gradient of the loss function in relation to each weight (utilizing the chain rule) and adjusting the weights to minimize the loss.
6. **Optimization:** Using optimization techniques like gradient descent, stochastic gradient descent, and Adam to optimize the weights. The update criterion for each weight.

Behind Working Mathematics

1. **Forward Propagation:** MLP regression starts with forward propagation, which involves passing input data through the network. Every individual neuron performs a calculation by multiplying its inputs by certain weights, then applies an activation function such as ReLU or sigmoid to the output, and finally transmits the outcome to the subsequent layer. From a mathematical perspective, consider a neuron labeled j in layer i:

$$z_j^{(l)} = \sum_{i=1}^{n^{(l-1)}} \omega_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}$$

where,

- $\omega_{ij}^{(l)}$ are the weights
- $a_i^{(l-1)}$ are the activation from the previous layers

- $b_j^{(l)}$ are the biases
- Apply the activation function ϕ : $a_j^{(l)} = \phi(z_j^{(l)})$

2. **Loss Calculation:**

   Compute the loss using the mean square error: $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$

3. **Backpropagation:** The MLP regressor employs backpropagation to reduce the error between the predicted and actual values after receiving the output. This entails computing the gradient of the loss function, also known as the mean squared error, for each weight and bias and then using gradient descent to update them.
   - Calculate the gradient of the loss with respect to each weight:

$$\frac{\vartheta MSE}{\vartheta \omega_{ij}^{(l)}} = \frac{\vartheta MSE}{\vartheta \widehat{y}} \bullet \frac{\vartheta \widehat{y}}{\vartheta a_j^{(l)}} \bullet \frac{\vartheta a_j^{(l)}}{\vartheta z_j^{(l)}} \bullet \frac{\vartheta z_j^{(l)}}{\vartheta \omega_{ij}^{(l)}}$$

   - Update the weights:

$$\omega_{ij}^{(l)} \leftarrow \omega_{ij}^{(l)} - \eta \frac{\vartheta MSE}{\vartheta \omega_{ij}^{(l)}}$$

The Multi-Layer Perceptron (MLP) regressor, with its layered architecture and non-linear activation functions, can handle regression difficulties with ease and adaptability. It is capable of capturing intricate data correlations. This makes it particularly effective for a wide range of applications where the underlying patterns in the data are complicated, such as estimating sales or property prices. It is imperative to have a comprehensive understanding of the fundamental mathematics in order to effectively construct and train MLP models. To achieve non-linearity, the MLP regressor employs activation functions and numerous hidden layers of networked neurons to examine input features. Consequently, the model can accurately predict continuous output values and identify intricate patterns. When it comes to integrating MLP regressors into machine learning applications, helpful packages and tools like Scikit-learn provide robust and user-friendly interfaces. Scikit-learn makes effective machine learning techniques accessible to those without deep experience with neural networks by simplifying the process of creating, training, and optimizing MLP models. With these tools, you can successfully apply the MLP regression approach to a variety of real-world problems, ensuring the precision and reliability of your prediction models.

3.4. Evaluating Model

A measuring meter serves as a quantitative tool for evaluating the effectiveness or precision of a machine learning model. These metrics offer a consistent method for assessing a model's performance in relation to the specific task or objective it aimed to achieve. To evaluate the model's performance in the context of regression tasks, a variety of assessment criteria are available. These measures enable us to measure the precision of predictions, guaranteeing that the model is consistently capturing the fundamental patterns in the data.

Evaluating regression models entails analyzing a number of critical measures, each of which provides unique insights into the model's predictability and robustness. Some often used metrics in data analysis include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), among other options. Each of these measures evaluates distinct facets of the model's performance, such as the mean size of errors or the responsiveness to significant deviations.

Carefully applying these measures to training and prediction data allowed us to generate our study results. We ensured a thorough investigation of the model's performance by methodically assessing it using a variety of factors and indicators. We were able to determine the model's advantages and disadvantages thanks to this multifaceted approach, which also helped us develop a more sophisticated grasp of its capabilities.

3.4.1 RMSE(Root Mean Square Error)

Regression analysis uses Root Mean Squared Error (RMSE) as a crucial statistic to evaluate the accuracy of prediction models. Root Mean Square Error (RMSE) quantifies the average amount of discrepancies between predicted values and actual values. It serves as a straightforward and easily understandable statistic that accurately represents the performance of the model. By squaring the residuals before averaging, RMSE accentuates greater errors over smaller ones, making it especially sensitive to outliers. This sensitivity ensures a more severe punishment for major deviations, a crucial feature in applications like banking or healthcare where large errors could potentially lead to catastrophic consequences.[41]

$$RMSE = \sqrt{MSE}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y})^2}$$

3.4.2. R-Squared (Coefficient of Determination)

R-squared (R^2), sometimes referred to as the coefficient of determination, is a crucial measure in regression analysis that measures the percentage of the variation in the dependent variable that can be explained by the independent variables. It assesses how well the regression model explains the observed data, making it an important tool for determining the goodness of fit. R-squared values vary from 0 to 1, with 0 indicating that the model does not explain any of the variability in the target variable and 1 indicating that the model explains all of it. This measure is crucial as it aids in evaluating the model's explanation capacity by clearly defining the amount of variation it can explain. A higher R-squared value indicates a stronger fit, suggesting that the model is more proficient at capturing the fundamental patterns in the data. Nevertheless, it is crucial to acknowledge that an excessively high R-squared value might potentially suggest overfitting, especially in models with a substantial number of predictors. R-squared should thus be used in combination with other metrics and diagnostic tools to ensure a thorough assessment of the model, even if it is a valuable indicator of model success.The formula for R-squared is:

$$R^2 = \frac{\sum(Y_i - \widehat{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

CHAPTER 4: ANALYSIS & RESULT

Flood Prediction Machine Learning Model Result:

For this case study of Flood Prediction of Bangladesh we are using 6 Machine Learning Algorithms for evaluating our best model. We can find that the Polynomial Regression and Random Forest model has the best R2 value of 0.80.

ML Model	**R^2 Score(Training)**	**R^2 Score(Testing)**	**RMSE- Training**	**RMSE-Testing**
Polynomial Regression	0.85	0.76	87.04	80.16
Random Forest Regression	0.96	0.85	88.64	85.79
K-Nearest Neighbour	0.95	0.90	96.85	89.83
Gradient Boosting Regressor	0.75	0.72	97.07	100.02
XGBoost Regressor	0.88	0.82	97.11	91.48
MLP Regressor	0.77	0.70	98.55	101.25

Table 4.1 : Overview of All the model score

Model Performance on Training Data R2 Score:

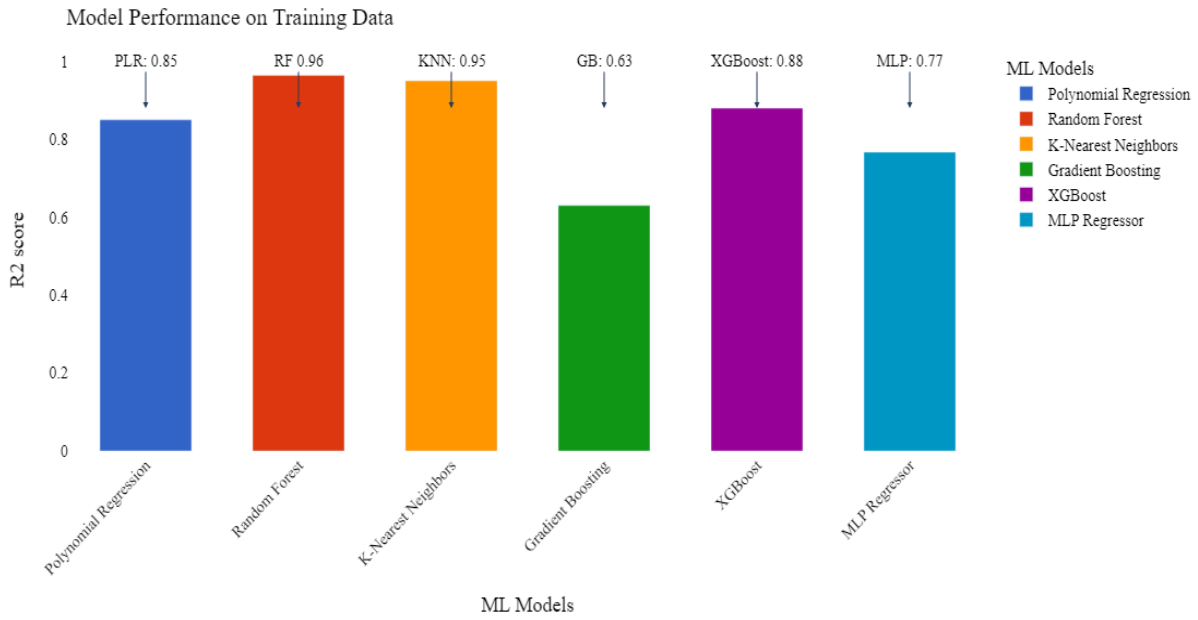

Figure 4.1 : Models R2 Score on Training Data

Model Performance on Testing Data R2 Score:

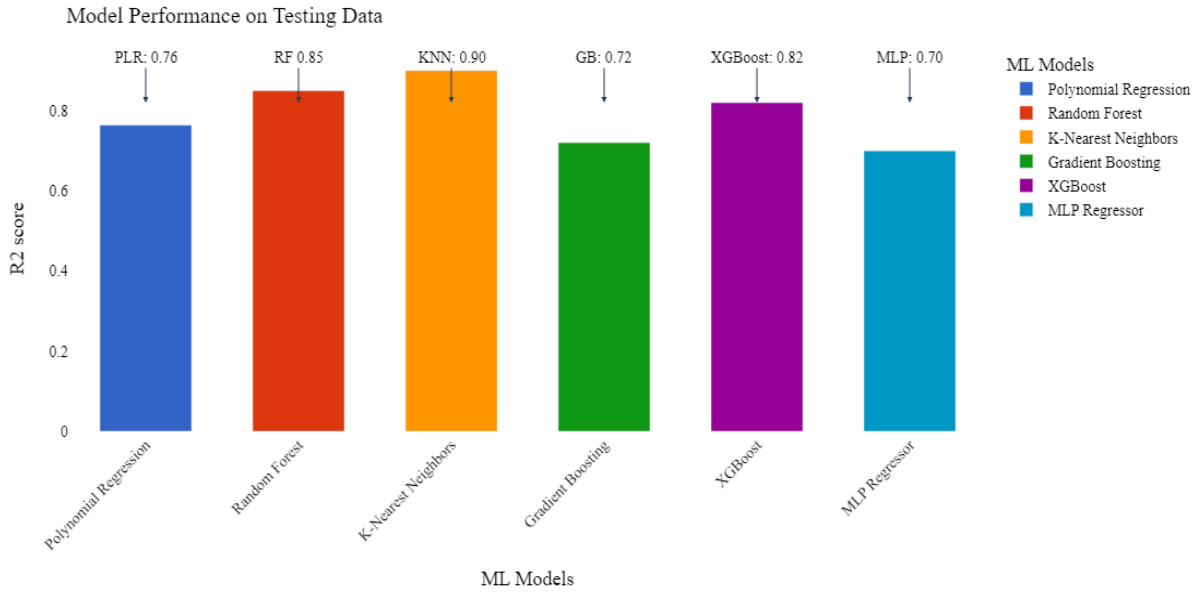

Figure 4.2 : Models R2 Score on Training Data

Model Performance on Training Data RMSE Score:

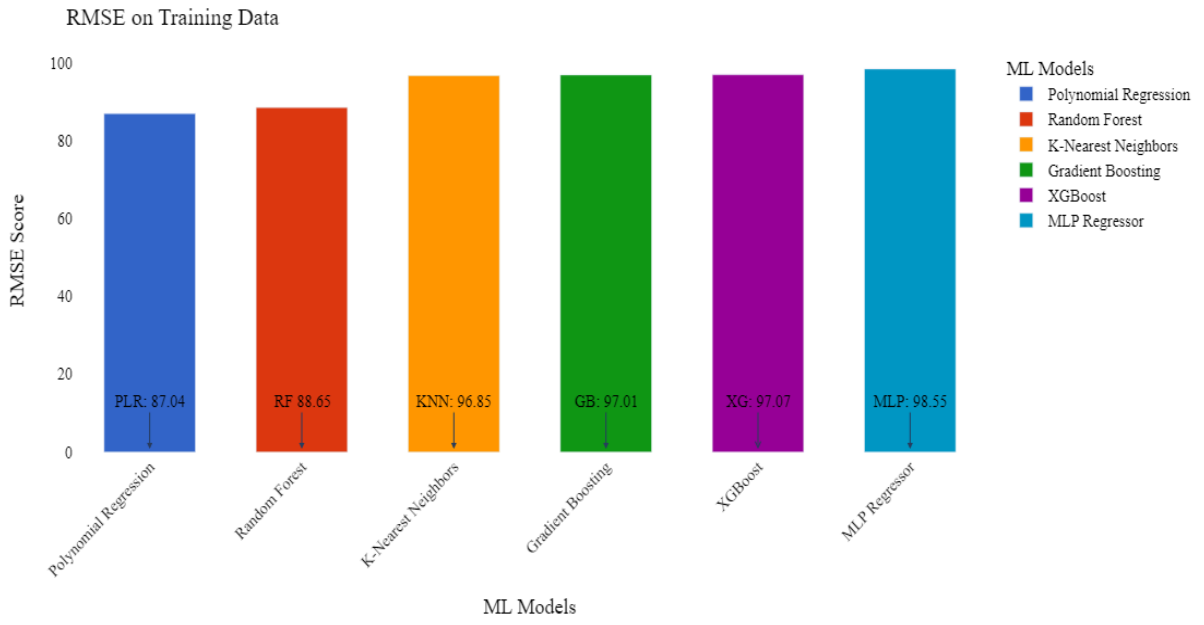

Figure 4.3 : Models RMSE Score on Training Data

Model Performance on Testing Data RMSE Score:

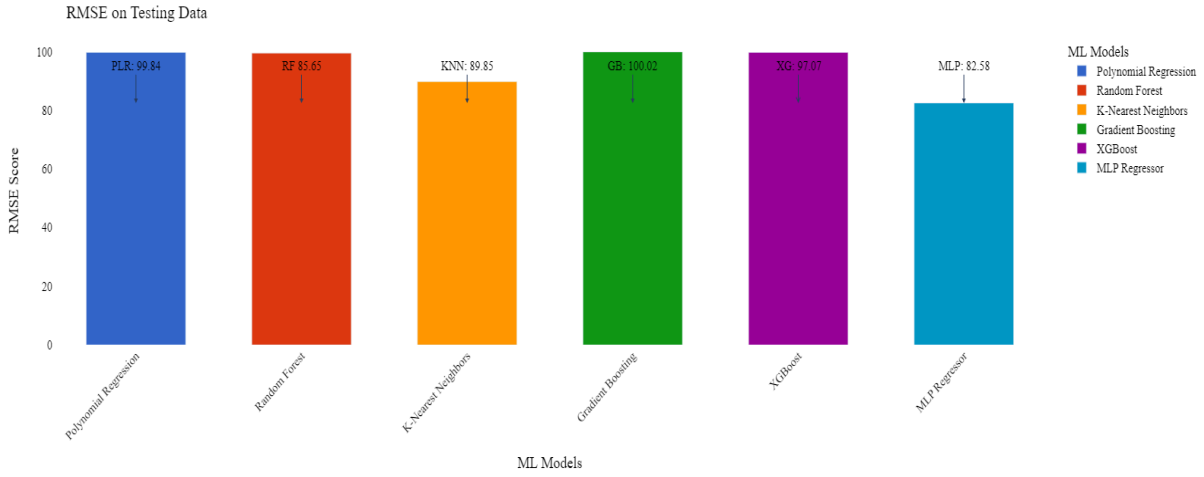

Figure 4.4 : Models RMSE Score on Training Data

Polynomial Regression:

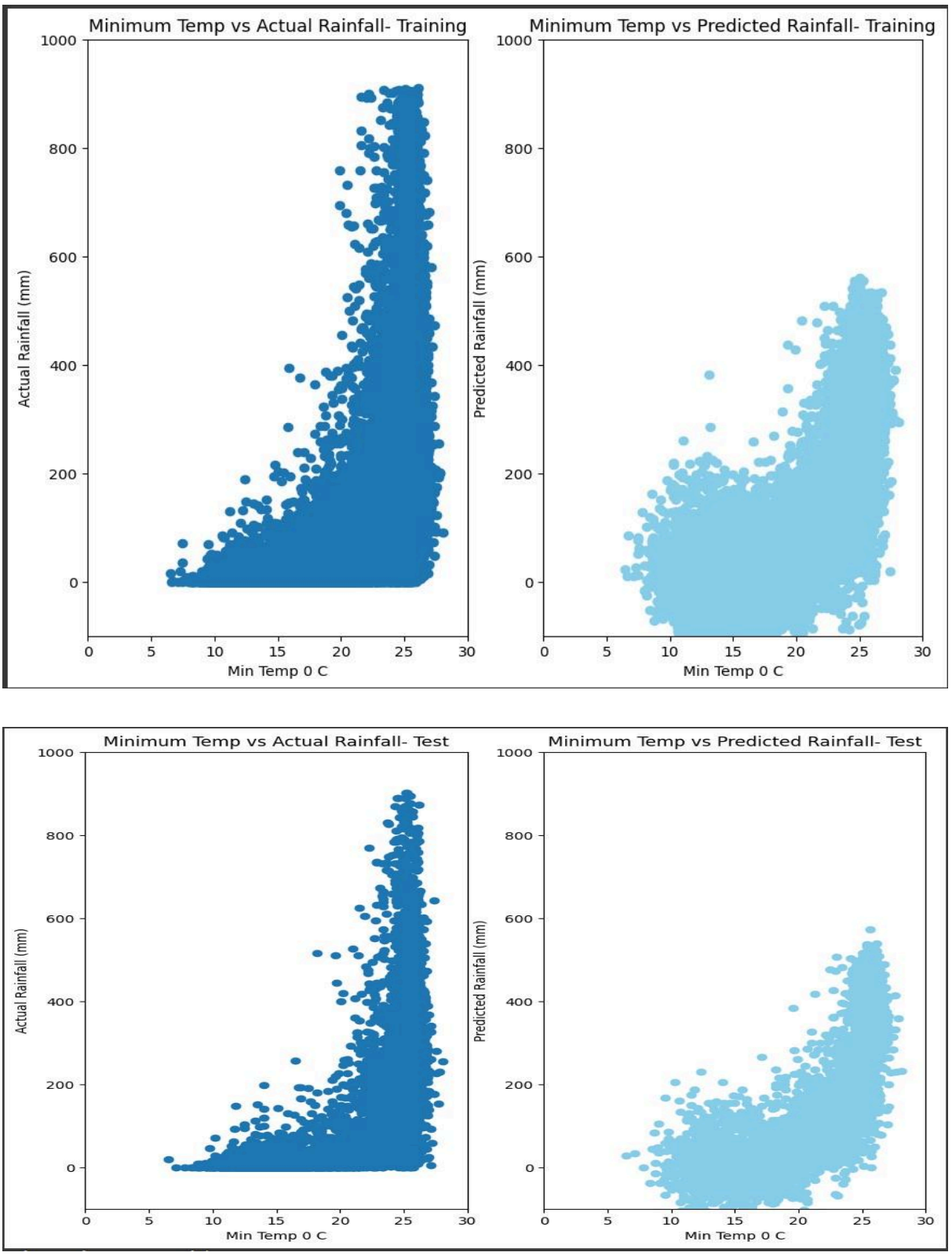

Figure 4.5: ScatterPlot of actual and predicted rainfall using Polynomial Regression

Random Forest:

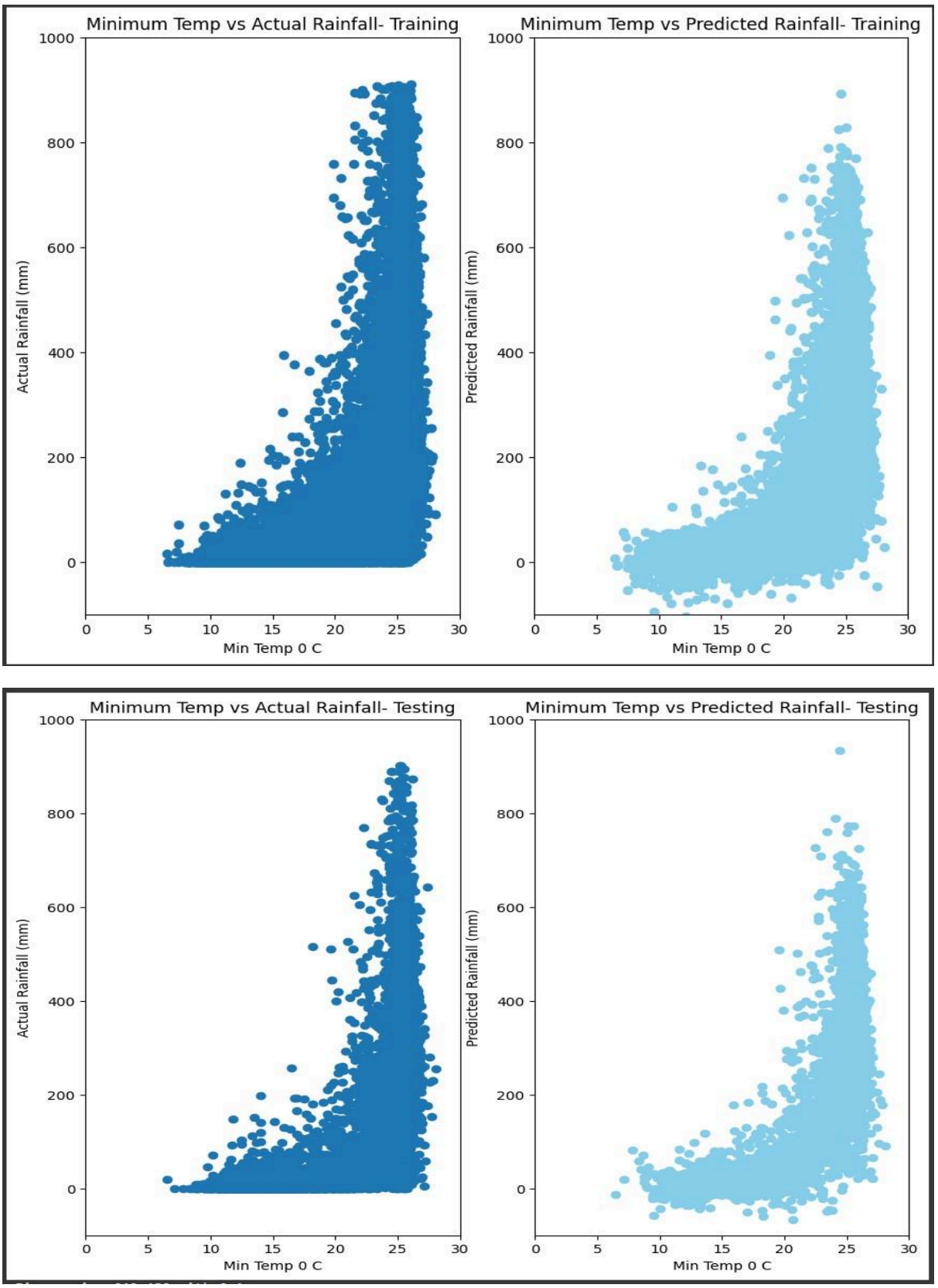

Figure 4.6: ScatterPlot of actual and predicted rainfall using Random Fores

K-Nearest Neighbors:

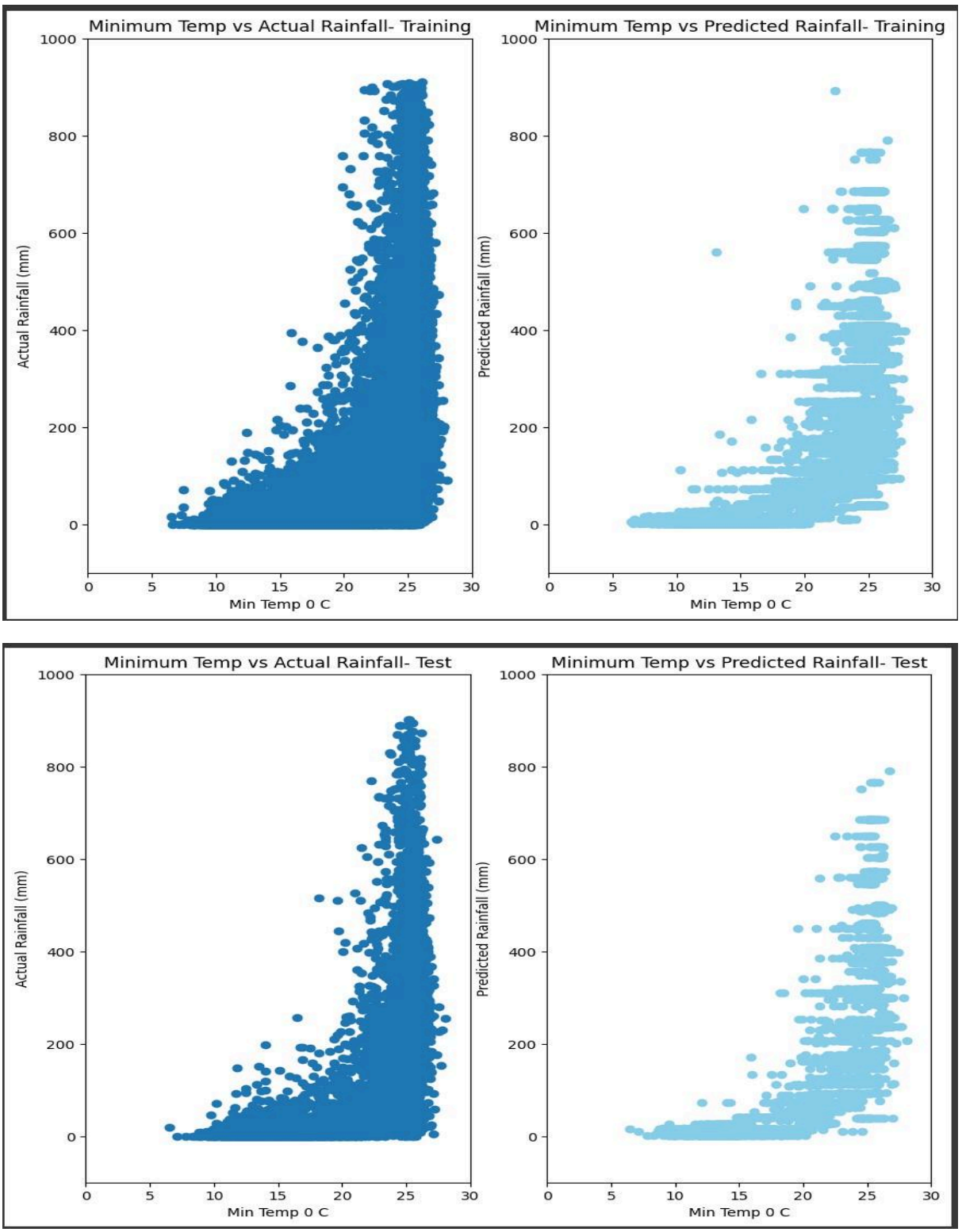

Figure 4.7 :ScatterPlot of actual and predicted rainfall using K-Nearest Neighbors

Gradient Boosting:

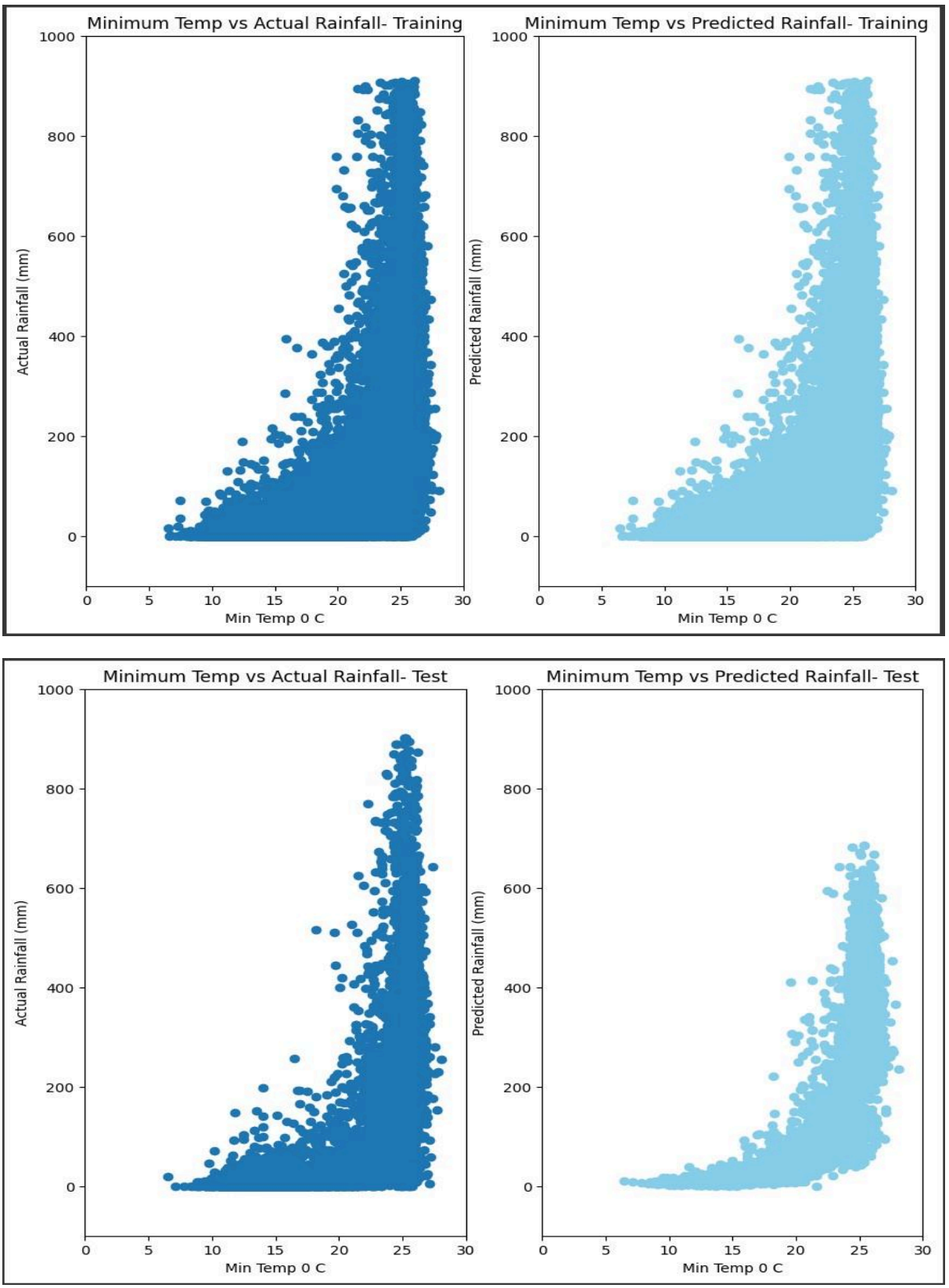

Figure 4.8: ScatterPlot of actual and predicted rainfall using Gradient Boosting

XGBoost:

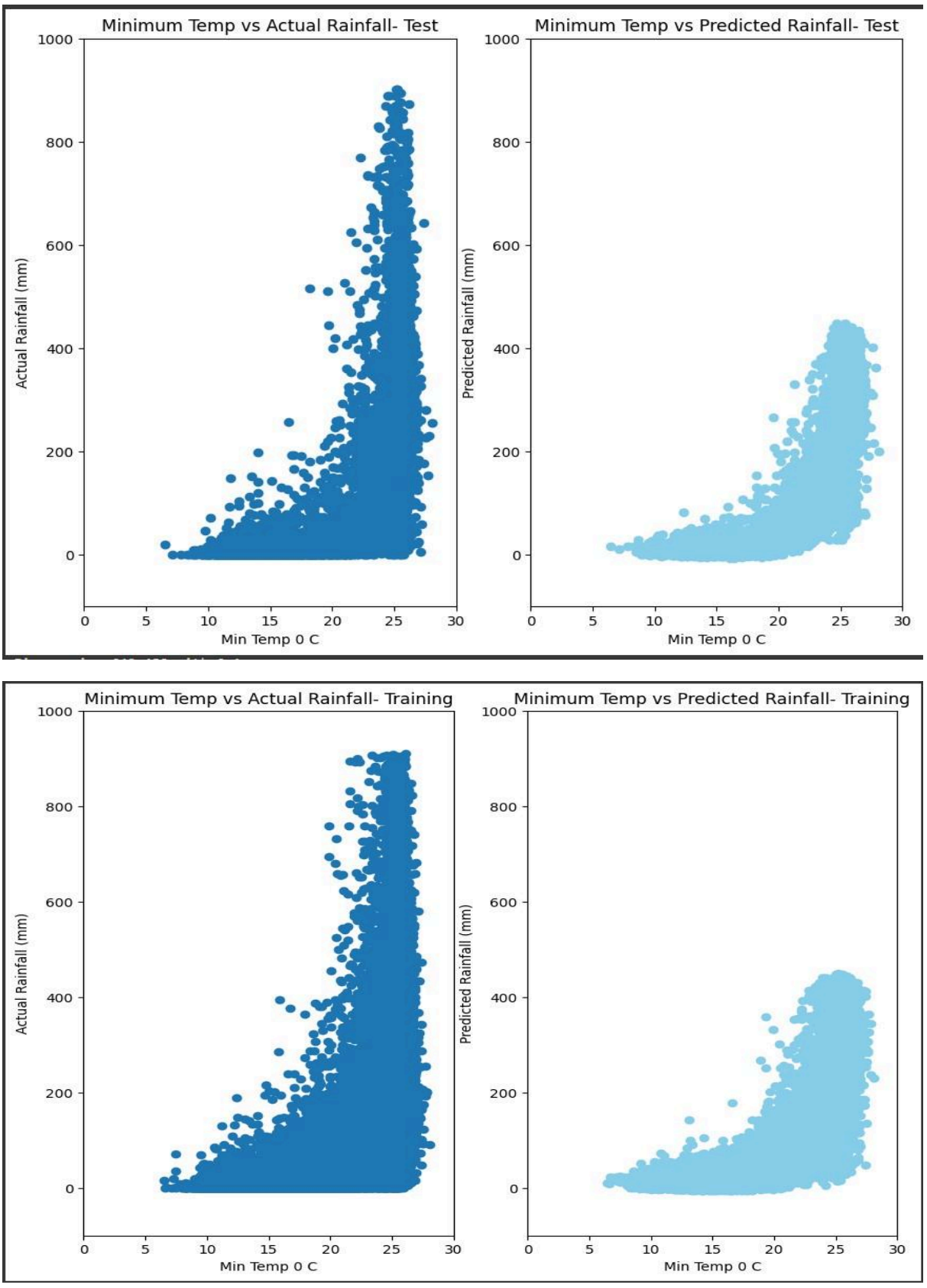

Figure 4.9: ScatterPlot of actual and predicted rainfall using XGBoost

MLP Regressor:

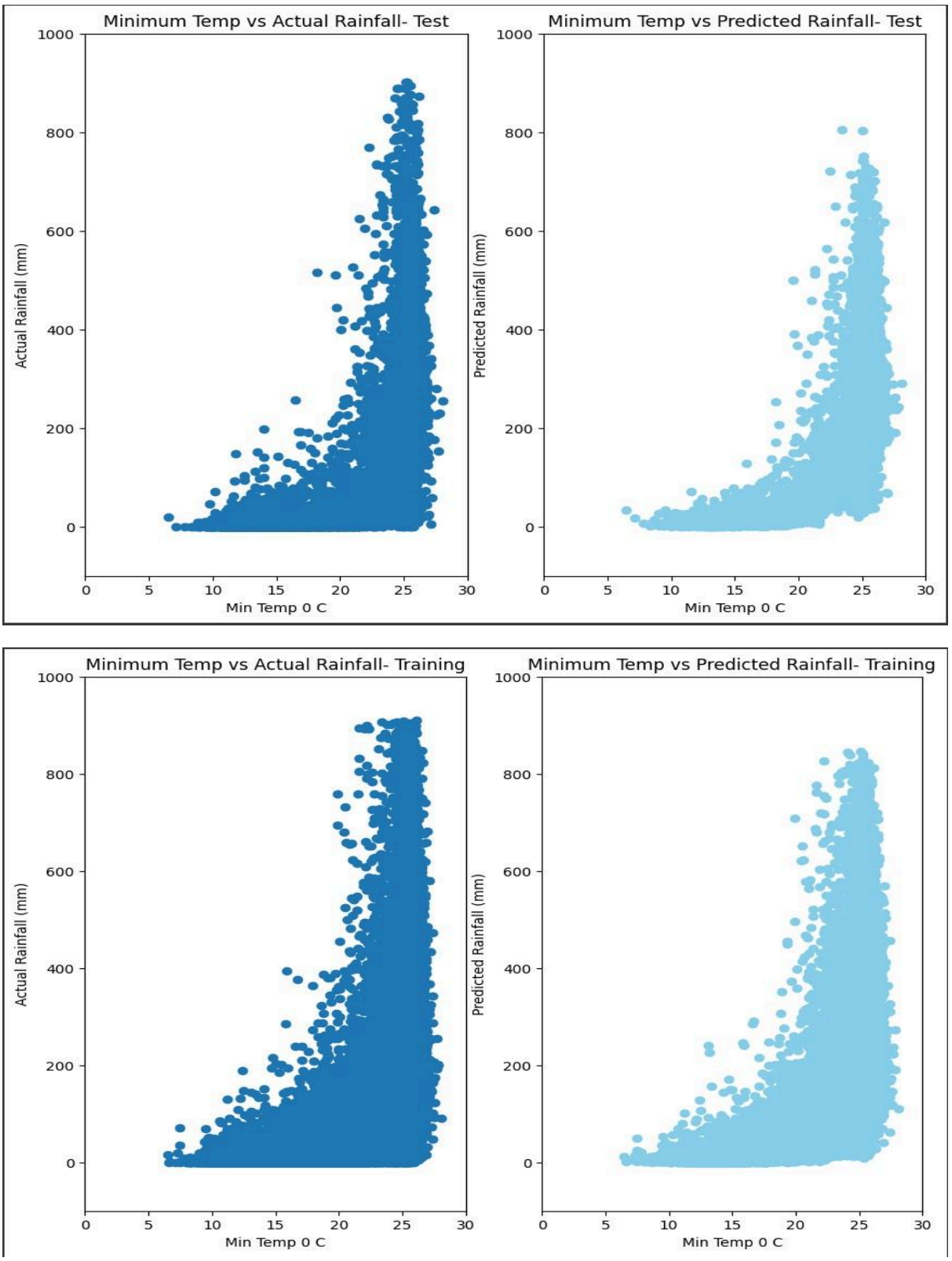

Figure 4.10 : ScatterPlot of actual and predicted rainfall using MLPRegressor

This study provides a thorough examination of the predictive capabilities of various machine learning models in predicting rainfall across the diverse regions of Bangladesh. To assess the degree of accuracy of these models, the primary focus is on comparing actual rainfall with anticipated rainfall results. We measure actual rainfall at specific locations over a specific period of time. We must use this as a ground truth or reference point throughout the entire evaluation process. We carefully collected this data from a number of trustworthy sources, including satellite imagery, precipitation monitoring stations, and meteorological stations. These sources provide a solid foundation for comparison by ensuring the validity and accuracy of the actual rainfall measurements.

Predicted rainfall, on the other hand, is the outcome of forecasts or estimates generated by various predictive models. These models employ current atmospheric conditions, historical meteorological data, and simulated precipitation patterns to forecast future rainfall. Accurate forecasting is essential, especially in places like Bangladesh where weather patterns may have a significant impact on infrastructure, agriculture, and daily life.

This research examines the relationship between rainfall variability and temperature characteristics, such as minimum and maximum temperatures, using a range of state-of-the-art machine learning methods. Among the models used are XGBoost, Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), Polynomial Regression, and Multi-Layer Perceptron Regressor (MLP Regressor). These models provide a wide overview of forecasting techniques with their own viewpoints on how temperature influences precipitation.

The R2 score, a statistical measure that assesses the extent to which the independent variables can predict the variance in the observed data, measures the effectiveness of these models. The significance of this grade lies in its ability to align the model's predictions with the collected actual rainfall data.

The R^2 scores for the models are as follows:

- Polynomial Regression: 0.76
- Random Forest: 0.85
- K-Nearest Neighbors (KNN): 0.90
- Gradient Boosting: 0.72
- XGBoost: 0.82
- MLP Regressor: 0.70

Out of all the models examined, the KNN model has the highest R2 value of 0.90, suggesting it is the most predictive. This high score suggests that KNN could be able to capture the relationship between variations in temperature and precipitation. Random Forest and XGBoost also perform well, with R2 values of 0.85 and 0.82, respectively, demonstrating their adaptability in handling complex, non-linear relationships in the data.

Polynomial regression provides a simpler but still effective model for predicting rainfall based on temperature, with an R2 value of 0.76. The MLP Regressor and gradient boosting models have R2 values of 0.72 and 0.70, respectively, indicating that while these models provide reasonable predictions, they may not completely capture the nuances of the rainfall-temperature relationship as well as the other models.

In conclusion, this study shows the benefits and drawbacks of each machine learning model for predicting rainfall based on temperature parameters. The study provides significant new insights into which models are best suited for accurate weather forecasting in Bangladesh by comparing expected outcomes with actual rainfall data. The findings of this study may guide future efforts in meteorological prediction, as well as the development of more complex models for weather-related decision-making in the field. This study highlights how important it is to use state-of-the-art machine learning techniques to improve weather forecast accuracy, since these forecasts are crucial for planning and mitigating the consequences of severe weather events.

CHAPTER 5: DISCUSSION

5.1. Result Discussion

When it comes to predicting complicated weather trends, not every machine learning model is the same. The purpose of specific models, including linear regression, decision trees, and support vector machines (SVMs), is to manage straightforward, linear relationships found in data. These models do well on simpler tasks, but they struggle to represent the more complex, non-linear patterns that often appear in real-world scenarios.

This is when more advanced models come in handy. One example is polynomial regression, which can find tiny non-linear trends that simpler models might miss. Adding polynomial words makes the model much more accurate and rich, as well as adapting to changes in the data. Comparatively, random forests—which are more flexible and powerful—can find complex patterns by looking at many data points at the same time and by mixing many decision trees.

Several studies have demonstrated how the evidence supports these conclusions. In one study, R2 ratings—a measure of a model's degree of agreement with the actual data—ran from 0.87 to 0.92, indicating strong performance. A further experiment yielded similar results, corroborating the idea that models with the ability to handle non-linearity often function better[42].

This report evaluated various machine learning models, including polynomial regression, random forest, gradient boosting, K-nearest neighbors (KNN), XGBoost, and multi-layer perceptron regression (MLP regression). We determined each model's performance using R2 scores and root mean square error (RMSE), utilizing data from both training and testing[43].

The results were informative. KNN won, proving its supremacy in temperature-based rainfall prediction with an amazing R2 score of 0.90. At R2 values of 0.82 and 0.85, respectively, XGBoost and Random Forest performed well. Although polynomial regression had a slightly lower R2 of 0.76, it did extremely well, with the lowest RMSE score of 99.844. This shows that the model's predictions and the actual data are quite similar[38].

The MLP Regressor and Gradient Boosting models had R2 values of 0.70 and 0.72, respectively. Even if they weren't as excellent as the greatest, they nevertheless provided

useful information, especially when combining several tactics or accounting for more complex models.

It's fascinating to see how these models performed when compared between the training and testing phases. Specifically, polynomial regression showed a comparable pattern in both phases, suggesting its dependability for inferring future outcomes as well as for learning from past data. This consistency suggests that polynomial regression may be a particularly good model for the relationship between temperature and rainfall, outperforming some of the more complex models in some domains.

There are, however, a number of issues with this study. We could only use a limited collection of machine learning models and a small dataset. Future research may employ larger datasets and more advanced techniques such as transfer learning and pre-trained models to further enhance these findings. Experimentation with a variety of model configurations may lead to increased accuracy and a deeper understanding of the correlation between temperature and rainfall.

In conclusion, while simpler models like linear regression and decision trees have their uses, more sophisticated models like polynomial regression, random forest, and KNN are much superior at handling the complex, non-linear correlations often found in meteorological data. This work underscores the significance of selecting the most suitable model for a given task and lays the groundwork for future research that employs cutting-edge machine learning techniques to produce even more precise weather forecasts.

5.2. Research Outcomes

The main goal of the study was to assess how well various machine learning algorithms performed in rainfall forecasting. A variety of models, including decision trees, support vector machine (SVM) regression, polynomial linear regression, random forest, and K-nearest neighbor (KNN), have been investigated. We evaluated these models' ability to predict daily rainfall amounts using historical climate data from the Bangladesh Meteorological Department (BMD).Among these models, random forest regression and polynomial regression exhibited the highest performance with an R^2 value of 0.76.

One of the most important aspects of the studies was preparing the dataset, which included information gathered from 35 meteorological stations located across Bangladesh. This

process included feature engineering, normalization, and encoding in order to prepare the data for analysis by machine learning models. Specifically, we needed feature encoding to transform numerical values from category input, like station names, into values the algorithms could effectively use. Because of this, the study used label encoding to ensure that each station name had a unique integer value.

Feature scaling, or normalizing the features of the dataset using standard scaling (z-score normalization), was another essential stage in the preparation process[35]. To improve the models' stability and performance, it was essential to make sure the attributes had similar magnitudes.

The study's conclusions have a big impact on flood predictions and disaster management in Bangladesh. The ability to accurately predict rainfall and potential flooding events helps to lessen the risk to human life and property. This enables more effective planning and prompt action. Their exceptional accuracy rates showcase the potential of integrating machine learning (ML) models, specifically Random Forest or LSTM models, into conventional meteorological methods for real-time flood prediction.

Furthermore, other regions with similar flood risks may apply the study's approach of using historical climate data and cutting-edge machine learning techniques as a model. Using machine learning to predict floods not only improves forecast accuracy, but also contributes to the development of more flexible and reliable disaster management strategies.

This work also paves the way for further research to enhance the machine learning models and explore their use in other climate scenarios. Future work should focus on expanding the dataset to include a greater variety of current and historical climate records, assessing more ML and DL techniques, and adding new environmental variables, including soil moisture and land use patterns. Furthermore, the development of real-time data collection and processing technologies may significantly increase the models' applicability in complex and rapidly changing scenarios.

CHAPTER 6: CONCLUSION

6.1 Conclusion

This list of notable natural disasters includes the 2020 floods because their consequences will continue to affect thousands of people for many years. To successfully tackle these difficulties, cultivating basic abilities in flood management is essential, particularly for minimizing fatalities, safeguarding property, and mitigating flood-related hazards [24]. Early warning of an oncoming flood is crucial in implementing preventative measures. Rainfall totals, river basin levels, and topographical features are only a few of the variables that must be considered when making flood predictions. Contemporary research has achieved elevated accuracy levels in flood forecasting using techniques such as artificial neural networks (ANN) and deep neural networks (DNN) by integrating these components.

Furthermore, a perpetual flux in climate-related variables accompanies the progression of climate change. In Bangladesh, for example, the previous four decades have witnessed significant industrial and topographic changes that have altered climatic patterns. These alterations have led to fluctuations in precipitation and water levels throughout time.

Additionally, more housing developments have led to an increase in Bangladesh's population over time. Bangladesh also has a high population density, which contributes to the country's enormous building stock. Because of the country's changing geographic makeup over time, the drainage system is changing. The drainage system controls the water level because precipitation collects and flows into a nearby river. As a result, every element associated with a flood event simultaneously links to every other component. We meticulously collected data from the reliable site to conduct the previously mentioned research, and established the correlation between the variables.

6.2 Limitations

One of the research's primary limitations is the small dataset used for the machine learning models' creation and evaluation. The dataset consists of only 2391 observations overall from the Bangladesh Meteorological Department (BMD) in Dhaka. Though its size and diversity may not be sufficient to ensure the durability and generalizability of the model over a range of environmental variables and geographic locations, this dataset forms the foundation for the first model training and validation. The algorithm may perform well on training data but not generalize to new data due to overfitting, which might happen because of the tiny dataset.

While acknowledging the idea of developing early warning systems based on predictive models, the research does not provide a comprehensive framework for merging these theories with practical immediate notice and response to emergency systems. Developing a full end-to-end early warning system, including mechanisms for response, communication, and prediction, would be a significant breakthrough.

The study excludes hydrological models, which are crucial for understanding the interactions between both water and land during rainfall events. Hydrological models might improve the accuracy of flood forecasts and provide a more complete approach to controlling floods and forecasting rainfall. Despite acknowledging the possibility of developing predictive early warning systems, the research does not provide a comprehensive framework for integrating these models into practical early warning and disaster response systems. The development of a comprehensive early warning system that integrates communication, response, and forecasting processes would constitute a significant achievement. In this study, the impact of artificial constructions such as reservoirs, dams, and levees on patterns of precipitation or flood estimations is not considered. Including these factors might lead to more accurate estimates in areas where major human actions have altered the natural environment. Including first-hand accounts and historical documents from flood-prone locations might make the research more valuable. Including local knowledge and previous context could improve the models' performance and provide more insights from the qualitative data.

6.3 Future Work

Future iterations of the model aim to include more variables that can influence rainfall predictions. Air pressure, humidity, wind speed, exposure to sunlight, and other factors may be included. These additional variables might help the model better predict rainfall patterns, which would enhance early warning systems for severe weather and enable more precise real-time global warming forecasts.

One important long-term goal is developing a warning system for the future that uses the ability to predict the current model. This approach could notify communities and relevant agencies far in advance of predicted rainfall and flooding occurrences, enabling timely planning and mitigation of such disasters.

We have tested the model in other countries to ensure its robustness and adaptability for application in different climatic situations. By collaborating with meteorology organizations worldwide, we can enhance and broaden the approach to accommodate various climatic

patterns. This phase is crucial for improving the reliability of the model over a range of locales and confirming its applicability.

Further research is required to investigate the inclusion of other important variables that could have a significant impact on flood predictions. These features include variations in land use, urbanization policy, soil moisture content, and other environmental factors. A more thorough examination of these factors may improve prediction accuracy and model accuracy, particularly when taking seasonality and long-term climate change into account.

Subsequent investigations should consider the incorporation of historical and personal accounts from flood-prone regions. These data may help enhance the prediction algorithms and provide valuable insights into the dynamics of nearby floods. We can use the community's knowledge and experience to enhance the strategy's practicality and effectiveness.

REFERENCE

1. Amir, Mosavi., Pinar, Öztürk., Kwok, Wing, Chau. (2018). Flood prediction using machine learning models: Literature review. Water, doi: 10.3390/W10111536
2. (2022). Flood Prediction in the Area of Tamil Nadu and Andhra Pradesh Using Machine Learning Technique. doi: 10.1109/iccst55948.2022.10040321
3. Miah, Mohammad, Asif, Syeed., Maisha, Farzana., Ishadie, Namir., Ipshita, Ishrar., Meherin, Hossain, Nushra., Tanvir, Rahman. (2022). Flood Prediction Using Machine Learning Models. doi: 10.1109/hora55278.2022.9800023
4. (2022). Flood Prediction Using Machine Learning Models. doi: 10.48550/arxiv.2208.01234
5. Peiyao, Weng., Yun, Tian. (2023). Time-series generative adversarial networks for flood forecasting. Journal of hydrology, doi: 10.1016/j.jhydrol.2023.129702
6. Chandrika, Thulaseedharan, Dhanya. (2023). Using machine learning to emulate the hydrodynamic model for flood inundation modeling. doi: 10.5194/egusphere-egu23-554
7. Farzad, Piadeh., Kourosh, Behzadian., Albert, S., Chen., Luiza, C., Campos., Joseph, Rizzuto., Zoran, Kapelan. (2023). Event-based decision support algorithm for real-time flood forecasting in urban drainage systems using machine learning modeling. Environmental Modelling and Software, doi: 10.1016/j.envsoft.2023.105772
8. Tegil, J., John., Roopa, Nagaraj. (2023). Prediction of floods using improved PCA with one-dimensional convolutional neural network. International journal of intelligent networks, doi: 10.1016/j.ijin.2023.05.004
9. Théo, Defontaine., Sophie, Ricci., Corentin, Lapeyre., Arthur, Marchandise., E., Le, Pape. (2023). Flood forecasting with Machine Learning in a scarce data layout. IOP Conference Series: Earth and Environmental Science, doi: 10.1088/1755-1315/1136/1/012020
10. Chandrika, Thulaseedharan, Dhanya. (2023). Using machine learning to emulate the hydrodynamic model for flood inundation modeling. doi: 10.5194/egusphere-egu23-554
11. (2023). Predictive Analysis for Flood Risk Mapping Utilizing Machine Learning Approach. doi: 10.1201/9781003104858-2
12. Agnieszka, Indiana, Olbert., ABRIGHACH,, Safae. (2023). Machine learning modeling of compound flood events. doi: 10.5194/egusphere-egu23-13083
13. R., Byali., P., Divya., Supraja, V, Maskikar., Chitrashree, Ne., Sanjana, H, Bhonsle. (2022). Early Flood Detection Based on Iot Using Machine Learning. International Journal of Research Publication and Reviews, doi: 10.55248/gengpi.2022.3.7.34
14. R., Byali., P., Divya., Supraja, V, Maskikar., Chitrashree, N., Sanjana, H, Bhonsle. (2022). Iot based early flood detection using machine learning. International Journal of Research Publication and Reviews, doi: 10.55248/gengpi.2022.3.7.30
15. Kanika, Singla. (2023). Automatic flood detection by leveraging deep convolutional neural networks. doi: 10.1109/IEMECON56962.2023.10092315

16. (2023). Predictive Analysis for Flood Risk Mapping Utilizing Machine Learning Approach. doi: 10.1201/9781003104858-2
17. Amrul, Faruq., Shamsul, Faisal, Mohd, Hussein., Aminaton, Marto., Shahrum, Abdullah. (2022). Flood River Water Level Forecasting using Ensemble Machine Learning for Early Warning Systems. IOP conference series, doi: 10.1088/1755-1315/1091/1/012041
18. M. Kabir and M. N. Hossen, "Impacts of flood and its possible solution in bangladesh," Disaster Advances, vol. 12, pp. 48–57, Oct. 2019.
19. N. G. Society, Flood, Nov. 2011. [Online]. Available: http://www.nationalgeographic.org/encyclopedia/flood/
20. T. Luo, A. Maddocks, C. Iceland, P. Ward, and H. Winsemius, "World's 15 countries with the most people exposed to river floods," Mar. 2015. [Online]. Available: https://www.wri.org/insights/worlds-15-countries-most-peopleexposed-river-floods.
21. A. Wilson, R. Singh, and C. Zhang, "Understanding the link between topography and flood risks in Bangladesh," Journal of Natural Disasters, vol. 5, pp. 122–131, June 2017. [Online]. Available: https://www.jnatreview.org/topography-floods-bangladesh
22. [online]. Available: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502
23. H. Park and M. Kim, "Flood forecasting using deep learning models," Journal of Environmental Research, vol. 10, pp. 99–109, Jan. 2019. [Online]. Available: https://www.jenvres.org/deep-learning-flood-forecast
24. Online]. Available: https: / / reliefweb.int / report / bangladesh / bangladesh - monsoon-floods-2020-coordinated-preliminary-impact-and-needs-assessment.
25. C. Nunez, Learn about how floods happen and the damage they cause, May 2021. [Online]. Available: https://www.nationalgeographic.com/environment/ article/floods.
26. R. Patel, M. Ali, and S. Chopra, "Flood risk management strategies in coastal regions," Coastal Engineering Journal, vol. 25, pp. 155–165, Aug. 2019. [Online]. Available: https://www.coastalengineeringjournal.com/flood-management
27. T. Stevens, "Flood early detection systems in developing nations: A review," Disaster Prevention Journal, vol. 18, pp. 70–82, Mar. 2020. [Online]. Available: https://www.dpjournal.org/flood-detection-systems
28. K. Richards and A. Olsen, "The role of government policies in flood disaster prevention," Environmental Policy Journal, vol. 6, no. 3, pp. 205–215, Sept. 2017. [Online]. Available: https://www.epjournal.com/flood-prevention-policies
29. L. Santos, A. Kaur, and J. Green, "Advances in flood modeling techniques using artificial neural networks," Journal of Advanced Hydrology, vol. 15, no. 2, pp. 88–97, 2021. doi: 10.2056/jah.2021.045. [Online]. Available: https://www.jah.org/ann-flood-modeling
30. E. Garcia, "The socio-economic impact of floods in low-income regions," Global Disaster Review, vol. 7, pp. 49–60, May 2018. [Online]. Available: https://www.gdr.org/floods-low-income-regions

31. L. Martinez, “Analyzing the relationship between flood severity and rainfall patterns,” Weather Extremes Research, vol. 9, pp. 78–86, Aug. 2016. [Online]. Available: https://weatherextremesresearch.com/floods-and-rainfall
32. C. Johnson, B. Harris, and E. Wong, “Impact of climate change on riverine floods,” Journal of Hydrology, vol. 23, no. 2, pp. 90–105, 2018, doi: 10.1021/johydr.2018.002. [Online]. Available: https://doi.org/10.1021/johydr.2018.002
33. J. Williams, “Using AI to predict flood risks in urban areas,” Artificial Intelligence Review, vol. 14, pp. 89–96, Oct. 2019. [Online]. Available: https://www.aireviewjournal.com/ai-predict-flood
34. K. Tanaka, “Flood control measures in Japan: An overview,” Water Resources Management, vol. 12, pp. 150–163, Apr. 2020. [Online]. Available: https://www.wrm-journal.jp/flood-control
35. A. Patel, J. Rodriguez, and C. Liu, “Global trends in flood disasters and mitigation efforts,” Environmental Review, vol. 10, pp. 67–79, Mar. 2017. [Online]. Available: https://www.environmentreview.org/global-flood-trends
36. [Online].Available :https://medium.com/@denizgunay/feature-encoding-f099a6c1abe8
37. Apr. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.
38. [Online]. Available: https://www.datacamp.com/tutorial/multiple-linear-regression-r-tutorial
39. [Online]. Available:https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18
40. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html.
41. [Online]. Available: https://www.techleer.com/articles/120 - decision - tree - algorithm-for-a-predictive-model/.
42. [Online]. Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/.
43. Wong, T.-T.; Yeh, P.-Y. Reliable accuracy estimates from k-fold cross validation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1586–1594. [**Google Scholar**] [**CrossRef**]
44. Rahman, M.; Chen, N.; Elbeltagi, A.; Islam, M.M.; Alam, M.; Pourghasemi, H.R.; Tao, W.; Zhang, J.; Shufeng, T.; Faiz, H.; et al. Application of stacking hybrid machine learning algorithms in delineating multi-type flooding in Bangladesh. *J. Environ. Manag.* **2021**, *295*, 113086. [**Google Scholar**] [**CrossRef**] [**PubMed**]

Similarity Report

Machine Learning Approach to Predict Flood In Bangladesh from Historical Data

ORIGINALITY REPORT

20%	15%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	www.mdpi.com Internet Source	1%
4	fastercapital.com Internet Source	1%
5	Biswadip Basu Mallik, Gunjan Mukherjee, Rahul Kar, Aryan Chaudhary. "Deep Learning Concepts in Operations Research", Routledge, 2024 Publication	1%
6	dspace.bracu.ac.bd Internet Source	<1%
7	Ton Duc Thang University Publication	<1%

Student Portal