American International University-Bangladesh (AIUB)

RAINFALL PREDICTION IN BANGLADESH USING MACHINE LEARNING APPROACH

Project Thesis

Name	ID
Bishal Paul	17-35836-3
Avijit Dey	18-36613-1
Somya Dhar	18-36396-1
Mehedi Hasan Pranta	18-36233-1

A Thesis submitted for the degree of Bachelor of Science (BSc)in Computer Science and Engineering (CSE) at American International University Bangladesh in 2 July, 2022
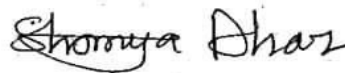
Declaration

This thesis is composed of our original work and contains no material previously published or written by another person except where due reference has been made in the text. We have clearly stated the contribution of others to our thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financialsupport, and any other original research work used or reported in our thesis. The content of our thesis is the result of work we have carried out since the commencement of the Thesis.

We acknowledge that the copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate we have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Bishal paul
17-35836-3
Department of Computer Science

Somya Dhar
18-36396-1
Department of Computer Science

Mehedi Hasan Pranta
18-36233-1
Department of Computer Science

Avijit Dey
18-36613-1
Department of Computer Science

Approval

The thesis titled **"Rainfall Prediction in Bangladesh using Machine learning approach"** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science on **July 2, 2022,** and has been accepted as satisfactory.

....

Sajib Hasan

Assistant Professor

Department of Computer Science

American International University-Bangladesh

....

Dr. Mohammad Mahmudul Hasan

Assistant Professor

Department of Computer Science

American International University-Bangladesh

....

Dr. Md. Mahbub Chowdhury Mishu

Assistant Professor & Head (Undergraduate)

Department of Computer Science

American International University-Bangladesh

....

Professor Dr. Tafazzal Hossain

Dean

Faculty of Science & Information Technology

American International University-Bangladesh

....

Dr. Carmen Z. Lamagna

Vice Chancellor

American International University-Bangladesh

Acknowledgments

Abstract

Bangladesh is an agricultural country and the agriculture sector is the main driving force of this country. Many people's livelihoods are depending on such a sector. Rainfall is one of the most important drivers of plants and the environment which are great complementary of agriculture phenomena. Besides this domains like defense, airline services, and the fishing industry also rely on rainfall. Irregular and unexpected rainfall can lead to catastrophic consequences. That's why it is important to predict the rainfall and take precautionary steps in accordance with this prediction. The rainfall pattern is fluctuating in recent years. Wind speed, temperature, humidity, and air pressure these parameters have a remarkable impact on rainfall. Numerous times experts have used a number of scientific techniques in this particular field. Another noteworthy prediction stuff that is mainly adopted by data scientists nowadays is the approach of Machine Learning. The rainfall data is collected over a period of time by using different machine learning techniques, a model is fit into it which is then used for making predictions. The current paper discusses such methods and draws a comparative line to find out which algorithm fit the situation best. Some of the algorithms discussed in the paper are Naïve Bayes' algorithm, Artificial Neural Networks, Decision Trees, and Random Forests Algorithm.

Keywords: Rainfall bd, Machine learning Rainfall, RFBD, Weather Forecasting, Numerical Weather Prediction.

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

AI	Artificial intelligence
ANN	Artificial neural network
DT	Decision trees
RFC	Random Forest classifier
GNB	Gaussian Naive Bayes
RR	Ridge Regression
NWP	Numerical Weather Prediction

TABLE OF CONTENTS

Chapter 1

Introduction

Rainfall is an important role in forming fauna and vegetation of natural life. It is now not just vast for human beings but additionally for animals, plants, and all residing things. Through this research, such risks can be resolved. Thus, it is honestly vital to predict the variation in rainfall. Rainfall prediction will now not solely facilitate important troubles as such, floods, drought, and problems that are related to agriculture, but also it will assist in remedying for humans by way of priorly informing them about the prediction. It will assist humans to deal with warm and moist weather. Now flood and drought are very frequent as in Sylhet, Barisal, Dinajpur has confronted worst herbal disaster in June, July month. Due to such kind of heavy rainfall, roads and bridges had been totally destroyed and travelers were trapped this catastrophe should not be estimated by the government, massive industries, critical-path administration entitles, as well as the scientific community before its occurrence. These additionally would possibly additionally lead to the landslide which is moreover a most serious geo-hazard inflicting the loss of existence and property all over the world.

For the remaining many years, scientists and engineers are efficaciously produced quite a few fashions for making the right predictions in several fields. Machine gaining knowledge is additionally a discipline that is widely used for prediction purposes or classifying things. There is a range of methods, recorded from KNN, an extra intricate method such as SVM and ANN (Artificial Neural Network) [1]. For metrology predictions, ANN is pictured as an alternative approach as hostile to the ordinary approach and is based totally on self-adaptive mechanisms that study from examples and capture useful interconnections between data, even if the interconnection between the data is unknown or tough to describe.

In the latest times, Deep getting to know will become one of the profitable mechanisms in ANN to resolve complicated issues and deal with exquisite amounts of data. Deep studying is basically a series of a multilayer structures that are trained. The necessary changes which have an effect on the model are weight and getting to know the charge of the layers. The deep getting to know the strategy has been extensively utilized in fields like pc vision, image recognition, herbal language processing, and bioinformatics [2].

In our experimental study, we use the rainfall data gathered from the decent internet site of the Bangladesh government. The data accrued involves greater than a decade of the size of rainfall all over Bangladesh. As the world, is shifting toward the issue of water, and in Bangladesh specific, rainfall prediction is the most necessary thing. So, in this paper, we strive to optimize the end result and come across the model which is properly gorgeous for the rainfall prediction in Bangladesh-specific areas only.

In our experimental study, we have a tendency to use the rain knowledge accrued from the real net website of the Bangladeshi government. the data accrued consists of exceedingly a decade of mensuration of rain all advised over the Asian state. due to the fact, that the world is transferring towards the challenge of water and in Asian countries, precise rain prediction is a most critical factor. So, at some stage in this paper, we have a propensity to try and optimize the quit end result and to are in search of out the mannequin that is right remarkable for the rain prediction in Asian nation-specific areas entirely.

Climate alternate is a big hassle that influences mankind. Human beings around the world are going thru serious significance due to this neighborhood climate change. Climate exchange potential that the alternate in the world or regional climate and specifically the local weather change is coming into consideration from the mid to late twentieth century onward [3]. This is all going on due to the excessive air pollution degree in the environment and is attributed mostly to the elevated level of atmospheric carbon dioxide (CO2) produced by means of the use of fossil fuels. Climate alternate also referred to as world warming. Global warming refers to the upward jostle in common floor temperature on earth Due to which the total earth is changing [4].

1.1 Aim of the Study

The purpose of the study we have a tendency to use the rain knowledge gathered from the legitimate internet site of the Bangladeshi government. the records collected consist of pretty a decade of mensuration of rain all knowledgeable over the Asian nation. due to the truth, the world is transferring toward the scenario of water and in the Asian nation-specific rain prediction is a most essential factor. So, via this study, we have the disposition to attempt and optimize the end result and to discover out the model that is nicely excellent for the rain prediction in Asian nation-specific areas entirely.

1.2   Significance of Study

The discovery is significant in terms of its critical contribution to the fields of agriculture, water reserve management, flood prediction, and administration intended to ease human beings by keeping them informed about the weather and rainfall forecasts. In addition, it is crucial that agricultural sectors use it to protect their crops and ensure the production of seasonal fruits and vegetables through accurate rainfall forecasting. Therefore, it would be truly amazing if the workforce could be predicted beforehand and necessary actions could be done [5]. The most fascinating subject is that water is an important resource for ecology. If we collect the water from the rain in the right way, we can use it. Another scenario that presents challenges for the management of water resources is incorrect rainfall forecasting.

1.3   Limitations of the Study

i. The data sample only includes monthly statistics; it does not include forecasts for daily output.

ii. Weather and climate variations may have an impact on how well the expected output is predicted.
iii. Geographically dispersed and various sites used for the study's data processing could potentially have an effect on the correlation efficiency used to gauge how well the ANN and NPL performed.

iv. A Jupyter notebook will be used to run the system covered in this study.

1.4   Problem Statement

Today, because of climate change, it is increasingly difficult to make accurate predictions of the local weather. The world's most pressing issue at the moment is whether to trade. People are trying to figure out the local weather exchange patterns since it affects everything from infrastructure to manufacturing. Creating a prediction of rainfall is a challenging task with a good accuracy rate, just like making a prediction of rainfall. It is impossible to predict rainfall using conventional methods, so scientists use computers to learn and conduct in-depth research to identify patterns that can be used to predict rainfall.

In most cases, farmers will be impacted by a poor rainfall forecast because their entire crop depends on rain, and agriculture is a crucial component of every economy. Therefore, it is quite good to anticipate the rain with accuracy. There are many different methods for learning on a computer, but the success of rainfall predictions depends on their accuracy more often than not. Rainfall has a wide range of effects on the earth, including drought, flooding, extreme summer heat, etc. It will also affect the availability of water around the world [3].

1.5   The Study region and data

All cities' climate information, including humidity, air pressure, wind speed, wind direction, and temperature, will be examined. Tall mountains and substantial annual rainfall may be found at Sylhet, Barisal, and Chittagong, among other places. Despite the low humidity, some of the expected results from the data gathered for the Bangladeshi government's climate websites have been liked. We will discuss the different types of weather in Bangladesh and see how the amount of rain changes as a result of the weather that changes with the seasons. We will determine future rainfall, as well as what kind of relationship exists between weather, rain, humidity, wind, and temperature and pressure.

Chapter 2

Literature review

2.1 Climate Vs Weather

Climate means the average weather prediction at a certain area for a long amount of time. On the other hand, weather refers to the atmospheric situation for a certain area for a specific time. Generally, weather means the day-to-day atmospheric condition for a specific region. Five major components are responsible for the climate which are the atmosphere, cryosphere, hydrosphere, geosphere, cryosphere, and biosphere. Six major components are responsible for the weather which are heat, air pressure, nebulosity, breeze, steam, and precipitation. Moreover, weather includes rain, sunlight, water, rock, wind, flood, storms, etc. Climate determines by calculating the statistics of the weather for more than 30 years. Contrariwise, weather determines by calculating the regular meteorological data. Basically, the climate is a statistical analysis process that is measured over a long time. On the other side, the weather is day to day measurement of the atmosphere for a short term like minutes to a week. Climate is very much effective for the long time planning of a country's development [6][7]. Besides this, weather is very much effective for human's daily activities. Moreover, it is responsible to determine climate. For climate change, human beings or nature may be responsible. But only nature is responsible for the weather change. Climate change affects many humans and animals' lifestyles. But the weather does not affect humans and animals' lifestyles.

2.2 Forecast Vs Prediction

The forecast is an estimated process to determine the result by using the data from past datasets, events, and experiences. On the other side, prediction means an idea that indicates an event can happen or can't happen in the coming future. As forecast gives a calculated result, we get an actual read for the future. On the contrary, as prediction gives us a guessing result, we get an idea about our future [8]. The forecast is a calculative statement. But prediction is a potential statement. As a forecast comes after an estimating result, it can create a business demand. Though prediction may or may not make a business demand, it has a great experimental demand. To determine the forecast, we use various Machine Learning (ML) algorithms. On the other hand, we use statistical analysis to

determine the prediction level. The forecast is the result of scientific analysis. But prediction relies on the subjective concern. We can analyze the error in the forecast. But error analysis is not possible in prediction [9].

2.3 Numerical Weather Prediction

Numerical Prediction means predicting the value numerically or in numbers. This numerical possibility is determined for a specific object. It reveals the possibility of a number of an event occurring. The prediction occurs or the decision is taken based on previous records which is known. Accuracy, robustness, interpretability, etc. are major concerns in numerical prediction [10]. In the perspective of machine learning, numerical prediction can be done through a regression algorithm. Which basically, tries to sketch a relationship between known and target variables to predict the value [11]. The most common regression algorithms are decision tree, linear regression, logistic regression, etc. NWP (Numerical Weather Prediction) is performed based on previous data related to the ocean and atmosphere. Parameterization plays a momentous role in this process. NWP deploys a set of equations that impart the drift of multiple weather parameters [12]. These essential equations are altered in the conformation of computer language in order to predict the weather. Various predictor models are used in predicting rainfall. Numerical models observe the current weather situation to predict the weather.

2.4 Rainfall

Rainfall is collapsing of rain or precipitation of it as drops of water. It is a very natural process. The amount of rainfall is measured in millimeters/inch in a specific area for a specific period. It is one of the leading ingredients of the water hoop. The water of ponds, rivers, and other sources is vaporized due to the hot summer season. Afterward this water showers as rain [12].

Precipitation (water from clouds) falls as rain (water on land or water) upon the surface of the Earth. Air masses that cross over warm bodies of water or over a wetland surface create a thunderstorm. Water vapor, or moisture, is transported upward into air masses where it condenses into clouds through atmospheric turbulence and convection. These clouds later release the water vapor, which becomes rain [12]. Raindrops with a diameter greater than 0.02 inches (0.5 mm) usually compose falling rain. It is called a drizzle when there are several smaller raindrops together.

Rainfall plays a significant role in the hydrologic cycle as a continual flow of water below, on, and above the Earth's surface - this includes subsurface water storage, ice caps, glaciers, rivers, lakes,

oceans, and rivers. When water vapor escapes from these containers, it condenses in clouds and travels over a variety of distances before falling as rain on Earth. Such precipitation is not considered rain if the clouds evaporate before reaching the surface.

2.4.1 Types of Rainfall

Rain plays a vital role for the growth our agriculture. For example, Ireland, Columbia as well as Bangladesh are familiar as a green country as there happens heavy rain. Water is very essential element for agriculture as well as our economy. Besides the river side crops growth is very high comparatively other land and this water depends on falling of rain. Basically, rainfall is cyclic process.

There are three main categories of precipitation -

• Relief Rainfall

• Convectional Rainfall

• Frontal Rainfall

2.4.2 Relief Rainfall

Generally, relief rainfall happens near the mountains and sea. When moist air rises up from the sea to the mountain and after reaching the higher point of the mountain the moist air convert to the cool water and precipitation occurs, then relief rainfall happens. When the moist air turns around to cool water then the cloud is formed and the weight of the cloud increased continuously. After a certain period, the rainfall is occurred which we called relief rainfall. When it occurs facing the frontside of the mountain, then it is called by the windward side and if it occurs facing the opposite side, then it is called by the leeward side. Another term of relief rainfall is orographic rainfall. Basically, when the moist vapor is raised up and the vapor is blocked by the mountain's obstacle, then the orographic rainfall occurs.

2.4.3 Convectional Rainfall

When the water of the earth or planet's surface is heating up because of the temperature of the sun and water is evaporating to the vapor and occurs the precipitation then it is called Conventional Rainfall. Basically, this type of rainfall happens in the summer season. When the air is heated up, the weight of its' respective water has been lighter than before. Then it goes the upper and formed into cloud. Thus, conventional rainfall is occurred. Basically, it (conventional rainfall) occurs in a hot summer day. The rainfall of Bangladesh is a perfect example of conventional rainfall as summer is our main season.

2.4.4 Frontal Rainfall

When the hot water vapor meets with cold cloud and hit each other than frontal rainfall occurs. In this rainfall, the cold breeze (cloud) comes to downward and the hot breeze (water vapor) goes to upward. Then the hot breeze and clod cloud became engaged in the clash and it starts the precipitate heavily. Typically, frontal rainfall occurs in our country in the months of April, May, and June. So, it can be a good example of frontal rainfall.

2.5 Rainfall prediction

Rainfall prediction is predicting the probability of rain in the upcoming time. This analysis is done based on multiple weather parameters like humidity, air pressure, solar radiation, temperature, etc. Previous data are also taken into consideration in forecasting rainfall. One of the main goals of this forecast is to determine the relationship between these variables and the likelihood of rain.

2.6 Weather Parameter

Weather parameters are associated with rain. Some of them are listed below.

2.6.1 Air temperature

Air temperature is a type of measurement by which we can measure how hot or cool our surrounding is. The unit of the measurement value is degrees Celsius (°C) or degrees Fahrenheit (°F) and it is measured with a thermometer [13].

2.6.2 Atmospheric (barometric) pressure

We live inside the air. Each and every moment we are pressurized by air as it has a weight. The forces exerted by the weight of the air on the surface are called atmospheric pressure. We can call it barometric pressure and it is measured by a barometer [14].

2.6.3 Humidity

Humidity means a type of measurement by which we can measure how much water vapor is existing in the atmosphere. It is measured with a hygrometer. During a given temperature, what contributes to the formation of clouds is the amount of water vapor present in the air [13]. As the air becomes moister, high saturation levels will be accommodated. Precipitation is more prevalent in humid weather compared to dry weather.

2.6.4 Precipitation

When any liquid water of frozen water or ice falls into the earth it is called precipitation. Basically, it is a systematic process of how water vapor goes into the atmosphere and again falls down to the earth because of the influence of gravitational force. It is measured with a rain gauge which is a cylindrical instrument [15].

2.6.5 Wind

When air or other gases move naturally with respect to the planet's ground is called wind. It can happen in various limits. At the time of a thunderstorm, it can happen for ten to fifteen minutes. Again, the rising of the temperature of the surface is responsible for natural wind which is lasting for hours. There are various kinds of winds that are separated by wind speed, wind strength, wind direction, and range [13]. The wind has various kinds of names depending on the speed and strength such as thunderstorms, hailstorms, tornadoes, hurricanes, cyclones, etc. The wind is very much effective for seeds, birds, insects, spores, pollen, etc. to travel from one place to another place. But sometimes it hampers the human being, animals, birds, insects, and so on.

It is the origin of the wind from where it generates. The direction of the wind determines the precipitation. Therefore, it is also important to analyze the direction of the wind as a parameter for rainfall prediction.

2.7 Supervised Learning

Supervised Learning means learning or training which is supervised. Supervised Learning is a type of machine learning in which each data has a label and a target variable to predict. Based on the dataset, the models are trained [16]. Here the label refers to, any kind of information that is pledged with the data. In supervised learning, there is a training dataset to teach the model. In such kind of learning, the

machine predicts the output based on previously provided data. Here machines play the role of an observer to train models. The model learns from a specific feature from the dataset as well as tries to act so in accordance with that.

2.8 Logistic Regression

The logistic Regression algorithm is an important part of supervised Machine Learning. Though is a regression algorithm it works like a classification algorithm. The logistic algorithm can work with both continuous and discrete values but we will get only discrete values as output. There are three types of logistic regression Binary Logistic Regression, Multinomial Logistic Regression, and Ordinal Logistic Regression. We will discuss Binary Logistic Regression in this report. The formula of the logistic model is:

$$f(x) = \frac{1}{1 + e^{-(b0+b1x)}}$$

Y

S-Curve

Predicted Y lies within 0 and 1 range

Y=0
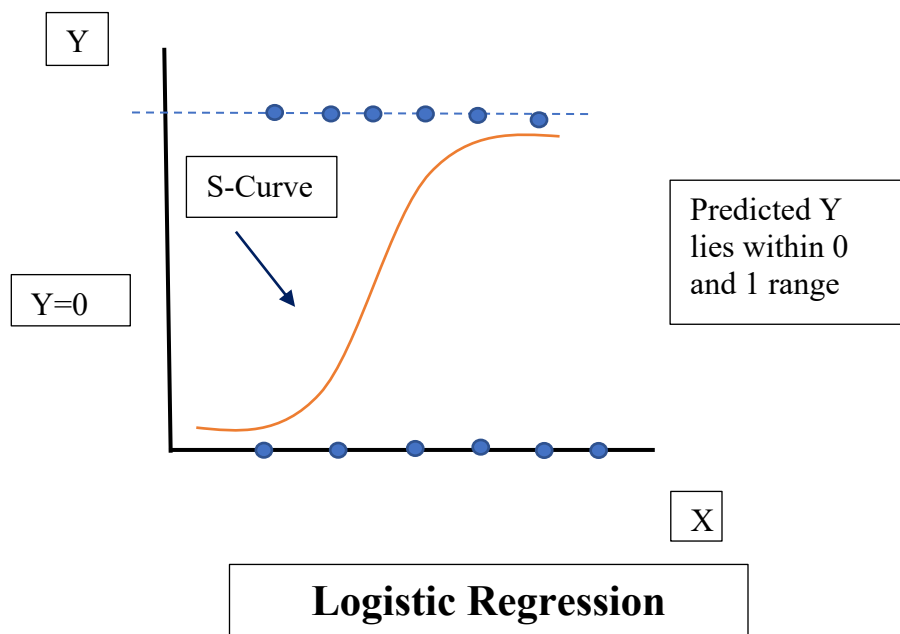
X

Logistic Regression

Figure 1: Logistic Regression Curve

From the formula and graph, we can see the output of the Binary Logistic Regression Algorithm will be a fractional value which will be within 0 and 1. Basically, we can predict the result by calculating the corresponding value. There have only two probabilities, one is "YES" and another is "NO". If the final output is greater, than 0.5 and less than or equal to 1, the machine will predict the result as "YES" / 1 and output is less than 0.5 and greater than or equal to 0, the machine will predict the result as a "NO" / 0. It is so much popular algorithm in machine learning to predict the result based on binary (0 and 1) output [17].

2.9 KNN Algorithm

K-NN is a classification algorithm. In the K-NN Algorithm, it stores all the cases according to their availability and then classifies them into new cases based on the similarity measures. By K-NN algorithm can compute the nearest distance data among all the reference data. Then based on the nearest data machine can predict the best possible result from the dataset. Thus K-NN works. There have a few steps to how K-NN works:

At first, assume a value of K of the neighbors

Then calculate the value of K of the neighbors by following the Euclidean distance formula which is (distance=$\sqrt{((x\_2-x\_1)^2+(y\_2-y\_1)^2)}$.

Determine the K nearest neighbors from the Euclidean formula.

Count the values of the data point for every category from the K neighbors.

Then the new data point in the category in which the neighbors are maximum.

2.10 Naïve Bayes Algorithm

Naïve Bayes is a type of probabilistic classifier which classify by applying the Bayes Theorem. In statistics, Naïve Bayes is also known as simple bias and independence Bayes. In Naïve Bayes each of the features utilizes on one's own. That means, these individual features don't depend to any other feature. That's why it is also called independence Bayes.

2.11 Linear Regression Algorithm

A Linear Regression algorithm is a simple technique of supervised Machine Learning. Linear Regression is the process to create the relationship between the dependent variable and the independent variable using a best-fitting straight line. It minimizes the sum of the squared difference between each data point and the line. The relationship between the is denoted by this formula.

Y = mx + C
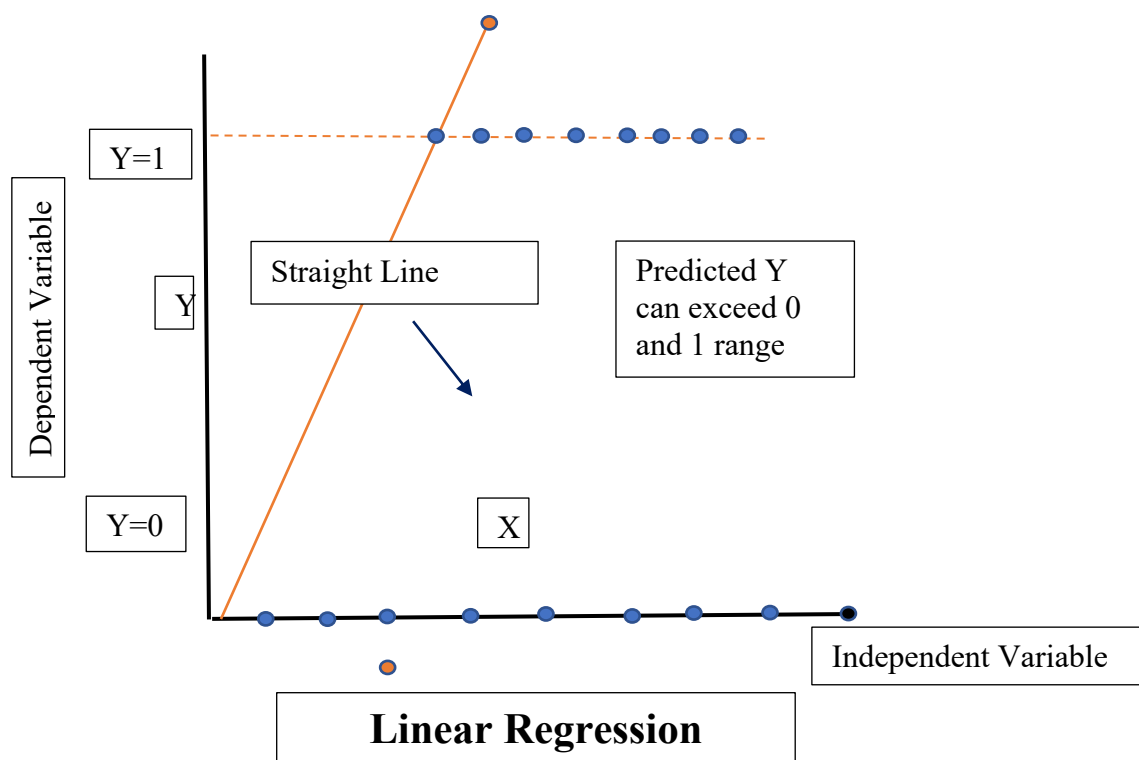
Figure 2: Linear Regression Curve

2.12 Neurons

Neurons are part of brain which assist in making sense. It is one of the most vital organs of a human body, which is responsible for receiving external input from the outer world to the body. It is responsible for signal processing in our body. The feelings we feel indeed this organ makes it happened to us. Through the collaboration of multiple cells or parts of brain this process is occurred. It contains various parts mentioned below.
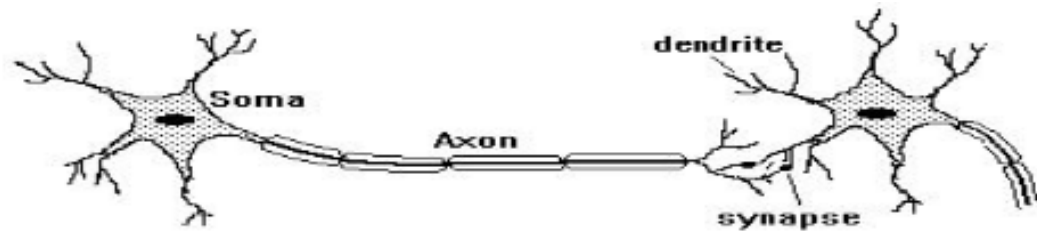
Soma is the part which contains nucleus. Axon is like a tail in looking. A lot of axons are connected through a fatty type of element known as myeline. It is responsible for all of the expectation of a specific action. Dendrite receives the necessary inputs. Neuron's may contain other set of dendrites which is called dendritic trees. Synapse is the organs which makes it possible to communicate within all of the internal organs of neuron.

2.13 ANN

Artificial Neural Network (ANN) means a network consisted with neurons which is artificial or intangible, which simulates the way, how human brain executes. The concept of ANN was invigorated by neurons of human brain. It is basically a blueprint for prediction or perception based on previously defined activation function. ANN utilizes algorithm which can learn its own and can make adaptation. This network learns from existing information that are provided in the dataset. Signals traverse from first to last through all of the layers of a neural network. These layers work to get the job done. One ANN can have any numbers of layers and any amounts of inputs may belong to a layer. ANN arrive in a conclusion after collectively estimating value. ANN is structured with two major phases called Forward propagation and Back propagation. Forward propagation phase incorporates adding bias in inputs, multiplying the weights, applying activation function and thereafter proceeding the inputs to the subsequent layer. The main objective of activation function is to let the data be informed about non linearity. Which assist in identifying the underlying types in data. The backpropagation phase's assignment is to find the optimal solution. To do so it's required optimization functions. Optimization function's major motive is to find the best suited value for the parameters. ANN can be used for regression and classification both purposes. This one is considered as one of the greatest advantages of ANN. The amount of error of an ANN is detected through comparing the targeted output and the actual output. Usage phenomenon of ANN includes spam detection in emails, business intelligence, chatbots etc. ANN works better than pristine machine learning models when volume of data get extended [32].

2.13.1 Structure of ANN

Ann comprised with huge numbers of neurons. These are input layer, hidden layer and output layer. The input layer is responsible for receiving the inputs of the algorithm. The hidden layer commits the action of processing the inputs. Afterwards, the finalized outputs are sent to the output layer. All of the nodes of the network are interconnected. This interconnection can also be classified in two types. One is feedforward neural network another one is recurrent neural network. In feedforward ANN the strolling of the inputs is unidirectional. Most of the time this type is particularly used in pattern recognition. In recurrent ANN sometimes the inputs proceed to the inverse of its direction.
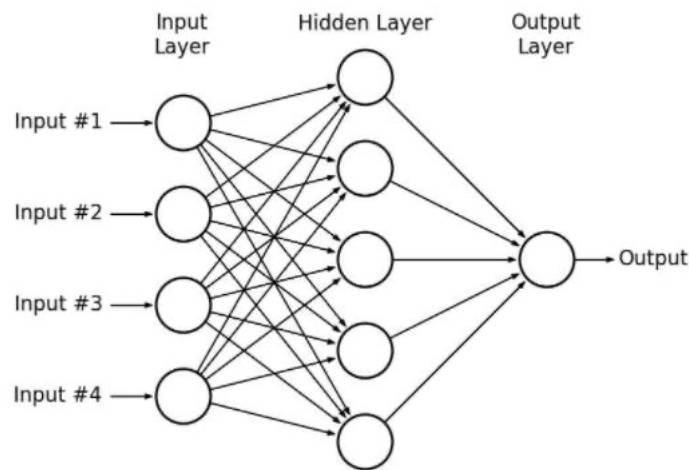
Figure 4: Structure of ANN (Ahn, 2017)

2.13.2 Feedforward neural network

Feedforward neural network (FNN) has multilayer perceptron and single layer perceptron. It was the first one as invention in ANN and simplest too. In feedforward ANN the strolling of the inputs is unidirectional. In these types of ANN there is not any kind of hoop among the connections of nodes. Most of the time this type of ANN is particularly used in pattern recognition, classification tasks etc. [33].

2.13.3 Recurrent neural network

In Recurrent Neural Network (RNN) the strolling of the inputs is multidirectional. It takes input, then move forward and then also provide feedback, means move in inverse direction to provide that feedback. Most of the times this type of neural network is used to solve temporary solution. As such image captioning, translation of languages etc. [34].

2.14 Decision Tree Algorithm

The decision tree algorithm refers to a graphical representation of all possible data sets and reaches the best possible output by following some specific conditions. It starts the journey from the root and goes to the next stage by maintaining some conditions. Then again, the algorithm checks whether it has a leaf or not. It then it chooses the next leaf by following the corresponding condition. If there has no leaf, then the algorithm closes its' operation and declares the outcome. Thus, it operates the algorithm. To determine the outcome, it has to be the representation named Sum of Product or SOP. Another name for SOP is Disjunctive Normal Form (DNF). It is a type of Random Forest Algorithm that refers to a Supervised Learning technique and includes a group of decision trees with different datasets and gives a predictive result with the highest accuracy. The complexity time is less than the others and can give high accuracy when a large dataset is missing. It works the result which is determined from the number of decision trees. Predict the Precipitation, may be one of the best algorithms as we have to maintain a large number of datasets and among the data set a big portion can miss. So, it will be the best option for rainfall prediction [18].

2.15 Ridge Regression

Ridge Regression is a type of Regularization Machine Learning method by which we can predict the best cost-effective result from a large number of datasets. The general regularization formula is:

$Y = \alpha 1 X 1 + \alpha 2 X 2 + \alpha 3 X 3 + ...$
Here, Y= dependent variable
X = independent variable
The magnitude of α is creasing continuously and by using regularization we can reduce the value of α and get the lowest complexity after applying of this formula. So, the best outcome depends on the lowest magnitude of the last coefficient of X.
For Ridge Regression the formula is:

Ridge, $R = loss + \alpha ||W||^2$
$\alpha ||W||^2 = Penalty$
$||W|| = W1 + W2 + W3 + ...$

Here, loss = Difference between the actual value and predicted value. We can get the best outcome if there has the lowest difference between the actual value and the predicted value. α is a constant and W is the coefficient of the vectors which is given in the dataset. If the penalty increase, the loss reduces. Thus, we got the predicted outcome with the highest accuracy. This model is very effective to reduce complexity as well as cost.

2.16 Necessity of Train and Test

Train and test sets are used to enumerate the effectiveness or accuracy of a machine learning model. It can be considered a determiner of the execution of a machine learning model. One of the key benefits of train and testing is, it allows us to compare the perfection of our model with others. The main focus of machine learning algorithms is to fulfill the goal to predict something. The train and test set to facilitate establishing the goal as well as checking the successfulness of that particular goal [19]. In data, there is a tangible effect that is random in nature. We are not informed about this pattern of randomness. That's why we need to train our model based on given data. Here the train portion is used to make the model compatible and the test portion is used to weigh the model. The model notice patterns inside the data of the train set and try to learn from it. The model is then judged in an unbiased manner through the test set. This prediction is suppositional and we don't know how proper it is. That's why the test is set to come in the deal [20]. After auditing the preciseness of our model, we can tune hyperparameters of our model accordingly in order to optimize or betterment. Hereafter, we can enhance other aspects also of our model by fostering the validation result. Train and test procedure is encouraged when the dataset is large.

2.17 Deep Learning for Numerical Weather Detection

It has already been shown in the research described in the preceding section that DL ideas may be successfully used to solve difficulties pertaining to weather forecasting. However, the few attempts that have been made to completely substitute the NWP process with a DL system have been restricted to short-term forecasting of 24 hours or less or have only employed a small portion of the meteorological data that is now accessible. In this part, we go through some of the difficulties that must be addressed before a whole complete DL weather forecasting system can produce outputs that are as good as the existing NWP. Weather forecast results can include a category index, a series of variables at a single location or aggregated over a region, a map of a specific weather variable such as temperature, or aggregated statistics of a particular variable over a specified period of time. With the existing NWP, we are accustomed to using a single forecasting system to provide the whole range of forecast products needed for end-user requests or for system assessment and future enhancement. Due to its ability to

reproduce target values, such DL weather forecasting systems could provide a significant output in computational resources and an absence of model bias [35].

2.18 Related Paper's Review

Machine Learning techniques for Big Data processing were described by Steve Oberlin and colleagues (2012). [21] He employed Machine Learning and several Artificial Intelligence approaches to vast and robust data sets. One use of Machine Learning is the recommendation engines used by Netflix to determine audience ratings and preferences. IBM's "Watson" employs several Machine Learning techniques to analyze and represent human language and answer inquiries in informatics and data mining. Linear regression, data massaging, perception, and k means are a few of the tactics he uses to reveal links and patterns in data. The nature of the prediction dictates the choice of a Machine Learning algorithm. Type or classification can be used to estimate the forecast. He also explained how adding features increases algorithm complexity and computing costs. A machine learning approach developed by Jainender Singh and colleagues (2014) may offer potential solutions to security challenges in applications, technologies, and theories. Using machine learning algorithms such as Support Vector Machine, Naive Bays classifiers, and clustering techniques, he emphasized mining sparse, incomplete, and uncertain data, which will yield optimal results when hidden patterns are discovered from data sets. It would provide insight into information in domains such as health, education, trade, and many others [22]. Junfei Qiu et al. (2017) offered some of the most recent developments in Machine Learning for Big Data processing. Representation Learning is a novel advanced learning approach that extracts meaningful information from data representations while building classifiers and predictions [23]. It seeks to capture a large amount of data in order to improve computing and statistical efficiency. Representation learning has three subtopics: feature selection, feature extraction, and metric learning. Another sophisticated Machine Learning approach used for massive data processing is active learning, which is used for biological deoxyribonucleic acid identification and picture categorisation. It is an example of semi-supervised Machine Learning in which it asks users to obtain the required output from a subset of crucial labelled samples accessible, hence saving costs while providing improved accuracy and optimal outcomes. Aside from explaining machine learning for Big Data processing, he also discussed the challenges and concerns associated with it. Some of the fundamental issues regarding Big Data include its diverse nature, data created at breakneck speed, ambiguity and incomplete data, and its immensity. He also provided solutions to the problem. A potential approach for parallel and distributed large-scale data processing is alternating direction methods of multipliers. It efficiently separates the many variables, assisting in the solution of enormous amounts of data. The Extreme Learning Method (ELM) has been developed to handle high-speed data and offers quick learning, excellent performance, and minimal

human influence. Yasir Safeer et al. (2010) introduced k-means clustering, a machine learning algorithm, for identifying a document from a large number of unstructured text documents. He suggested a method for representing documents that would enhance clustering outcomes. He spoke about the stream of document clustering, employed k-means, created an algorithm for improved document representation, and suggested using a systematic domain dictionary to improve document similarity findings [24]. According to Rohit Bhatnagar, et al., in 2018, Big Data Processing, Analytics & Machine Learning were discussed. Big Data analytics and machine learning are both developing fields that work well together. He talked about numerous machine learning trends for extensive data in the future. Data Meaning suggests ways to improve the intelligence of machine learning so that it can recognise text or other types of data. Integration is another trend utilised to handle and combine data [25]. Machine learning techniques like classification, regression, and cluster analysis are used to do analytics, forecast the future based on current trends, and identify correlations between the provided data sets. Machine learning algorithms were used by Alexandra L'Heureux et al. (2017) to provide novel approaches to processing large amounts of data. Traditional tools are now unable to handle Big Data's storage, transmission, or efficiency because of its unique properties [26]. Computational efficiency would be impossible due to the Support Vector Machine's enormous rise in size, space, and time complexity. Map Reduce addresses the Curse of Modularity, which occurs when an algorithm's specified boundary collapses as data size increases. Large data sets are processed using the following parallelism using this customisable and scalable approach. It uses an iterative process. K-means can also be utilised to overcome the Curse of Modularity's drawbacks. Online learning is one of the Machine Learning paradigms that will close the efficiency gaps brought on by Big Data. It aids in the processing of massive amounts of data. It can manage messy and loud data because of its adaptive nature.

Chapter 3

Methodology

It is possible to predict hydrologic responses of a watershed to changes by simulating precipitation. However, in order to model precipitation, it is necessary to first define the random variables that describe the precipitation process, group these variables into groups of statistically non-different months, estimate parameter values using the method of moments, and then fit a theoretical probability distribution to each random variable grouping. In order to replicate a series of precipitation occurrences based on the probability features of the previous historical record, the fitted theoretical probability distributions may be computed and then applied to a precipitation model. The following discussion will cover the data analysis algorithms that were built utilizing 38 years of climate data from various stations from 1981 to 2019.

According to a review of 38 years of monsoon rainfall data, there is no long-term variation or trend in monsoon rainfall averaged across the country. Even while overall rainfall in Bangladesh has not changed, yearly rainfall in particular meteorological sub-divisions has changed significantly. Rainfall in Barisal, Chittagong, Sylhet, Khulna, and Mymensingh is on the decline.

3.1 Data Collection

For this study, the uncooked information has been accrued from the regional meteorological station at Dhaka, Sylhet, Barisal, Chittagong, etc. All of these data were obtained from the Bangladesh Meteorological Office's website, where various sorts of raw data from various meteorological stations at various periods were made available. There are Eight data features Station id, days, maximum temperature, minimum temperature, humidity, observed wind, Observed Pressure, and Rainfall Categorical have been included. All the values of the weather variables have been recorded by the metrological station every day for all weather stations. As a result, the information was recorded in tabular form in the CSV file. There are 12 months and days of the month arranged according to climate variables in the row of tablelands. A 38-year period of uncooked data (1981-2019) from the station has been used for the study.

3.2   Data Processing

The data pre-processing step covered the records conversion, managing lacking values, specific encoding, and splitting dataset for education and testing dataset. A total of 38 years (1981–2019) of information were amassed from the meteorology office. Since the statistics have been raw, they contained lacking values, and wrongly encoded values so the lacking values of the target variable were removed and the other aspects had been filled with the usage of the imply of the data.

In the climate station, the raw records were additionally arranged on a 12 monthly based, and the attributes in rows that want to combine and rearrange points in columns. Thus, statistics were transformed from excel facts to CSV statistics files.

Encrypting, the dataset used is organized for the model. The main facets for rainfall forecasting have been selected and the dataset splitting as an 80:20 ratio for testing was viewed as an input for the model. The information incorporates the great parameters of the rainfall. Station id, Days, Maximum temperature, Minimum temperature, Humidity, observed wind, Observed Pressure, and Rainfall Categorical are considered. The analytical facts studied and examined in this study will be used as input for processing the output as rainfall prediction.
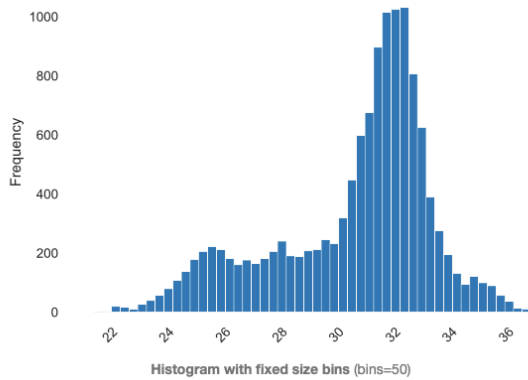
Figure 5: observedTempMax
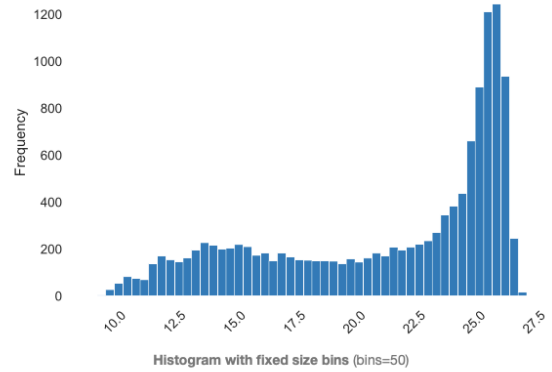
Figure 6: observedTempMin
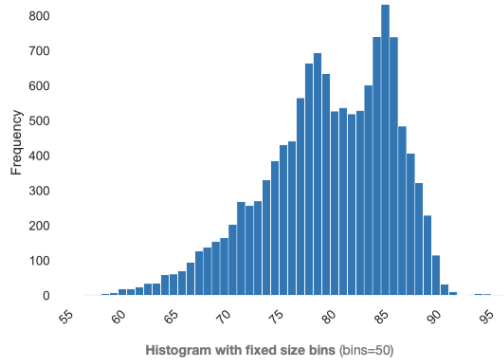
Figure 7: observedHumidity
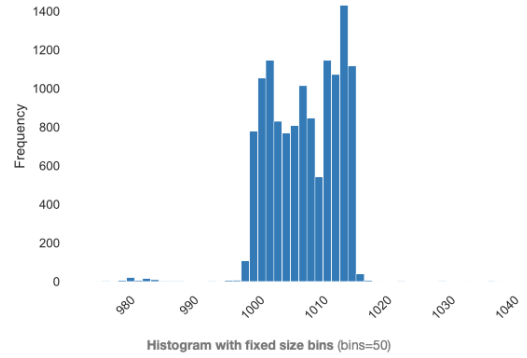
Figure 8: observedPressure
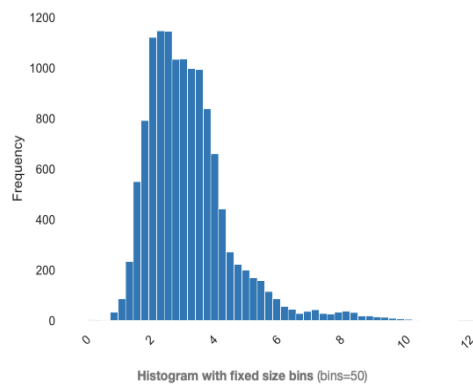
Figure 9: observedWind
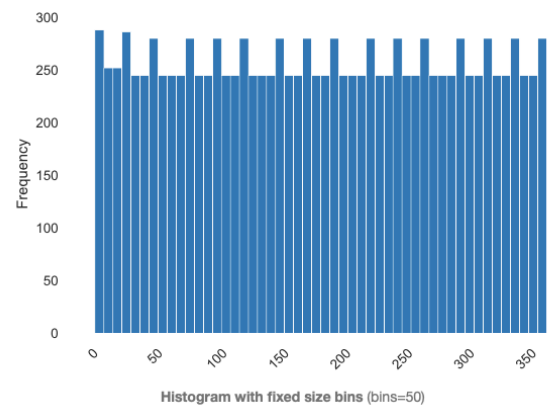
Figure 10: days
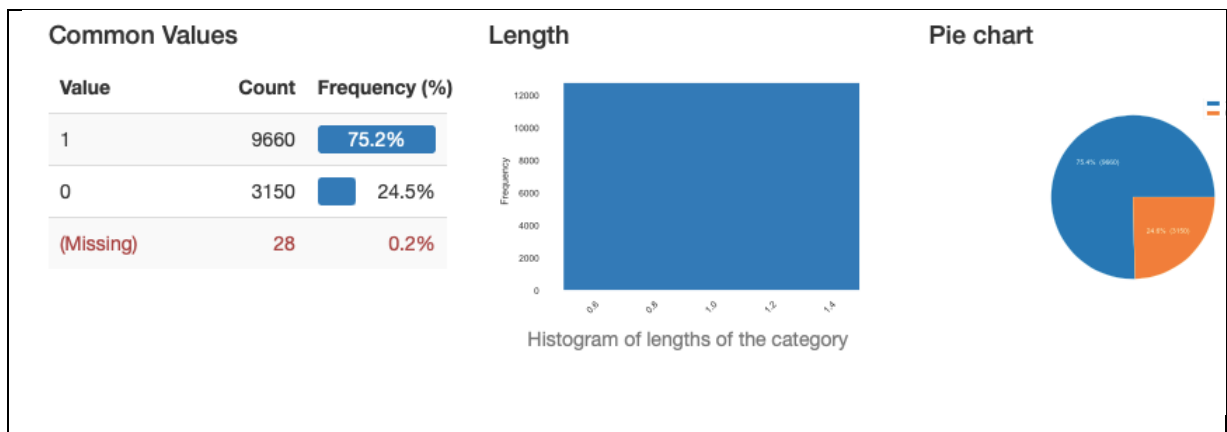
Figure 11: observedRainfallcatagorical
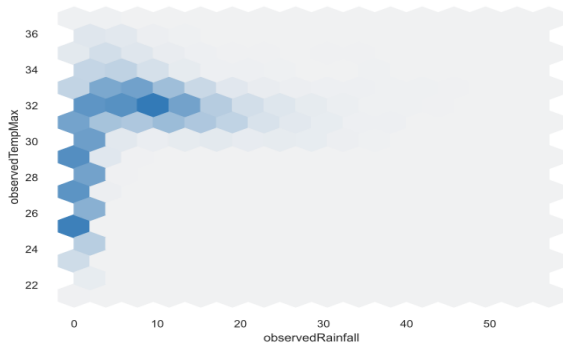
Figure 12: Interactions by observedTempMax
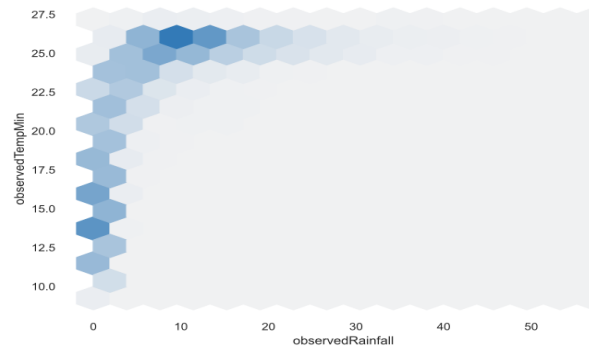
Figure 13: Interactions by observedTempMin
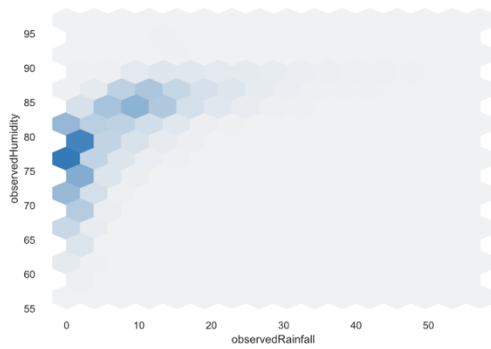
Figure 14: Interactions by observedHumidity
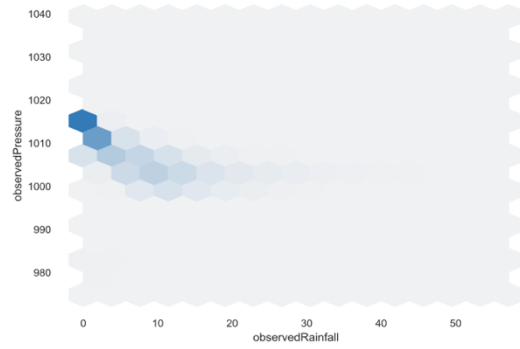
Figure 15: Interactions by observedPressure

Figure 16: Correlations

The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.
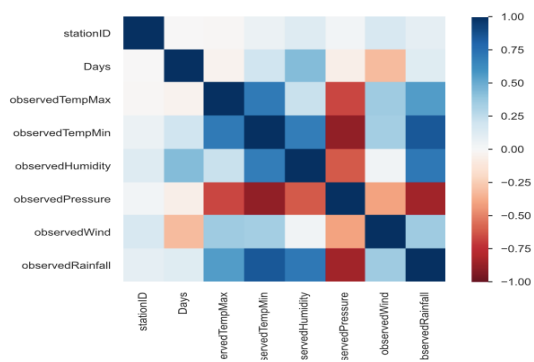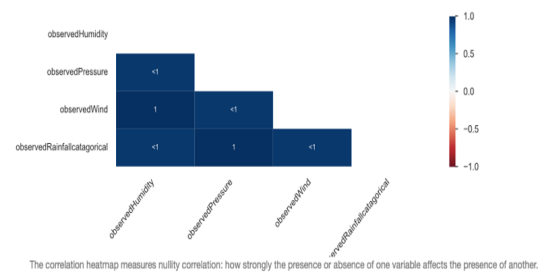
Figure 17: Missing values

3.3   Methods

We analyze the rainfall forecasting is done by using the methods of artificial intelligence, neural network, and Machine learning in some journals. Artificial intelligence and neural network are more difficult compared to Machine learning because artificial intelligence involves some algorithms [28].

The Data used in the present study are collected from the weather station in different states. We take 38 years of data during three months June, July, August is explored because these three months are the Rainy Season for our state.

3.4   Tools and Libraries Used

3.4.1 Keras

For simple or complex neural network implementation and calculation, Python has produced the high-level learning API known as Keras. An open-source software framework called Keras gives artificial neural networks a TensorFlow interface. Keras works on top of the TensorFlow library since it serves as an interface for the library. TensorFlow version 2.4 is the only tool that supports deep neural networks. It is user-friendly, modular, and extensible, and provides quick access to deep neural networks. Keras supported multiple backends up until version 2.3, including TensorFlow, Microsoft Cognitive Toolkit CNTK, Theano, mxnet, and plaidML. With an emphasis on contemporary deep learning, Keras is a high-level API of TensorFlow and an unrestricted, positively effective interface for addressing machine learning challenges. It offers necessary. It provides fundamental concepts and structural elements for rapidly creating and deploying machine learning. Low-level operations like convolution and tensor products are not carried out by Keras [30].

3.4.2 TensorFlow

Machine learning models and complex numerical problems can be solved with TensorFlow, a low-level software library created by Google. There's nothing more open-source or free than TensorFlow when it comes to machine learning software. It may be applied to a variety of applications with a focus on deep neural network training and inference. The core parts of TensorFlow, a symbolic math framework used by Google for both research and production, are data flow and differential programming.

3.4.3 Scikit- Learn

The most effective and reliable library for computers learning Python is Scikit-learn, which is a crucial component. By using a Python consistency interface, it lays out the need for systematic tools for computer analysis and statistical modelling, including classification, regression, clustering, and dimensionality reduction. This library was created using NumPy, SciPy, and Matplotlib [30] and is primarily written in Python. The library also makes it possible to do data processing operations including imputation, standardization, and normalizing. The performance of the model may frequently be greatly enhanced by completing these activities. In addition, there are also a number of packages available in Scikit-learn for creating linear models, tree-based models, clustering models, and many other models.

3. 5   Workflows

At first, we collected the dataset from the department of meteorology Bangladesh. Then we processed the data through data cleaning and doing group by in panda data frame. Then in the data partition phase, we partitioned the dataset in train, test, and split. Then we committed feature extraction and relevant feature selection. Afterward, we implemented Decision trees (Random - forest classifier), Gaussian Naive Bayes, Artificial neural networks, Linear Regression, Logistic Regression, and Ridge Regression. Through these models we predicted rainfall. Our model was intellectual enough to classify and identify the pattern of rainfall. The actual result and prediction results were very close in the case of almost all of the models. The accuracy was percipient enough. Then we fetched for errors. We calculated means square error, standard deviation error, and cross-

validation error. The error rate was high for linear regression and ridged regression models and the rest of the model's error was tiny [29].
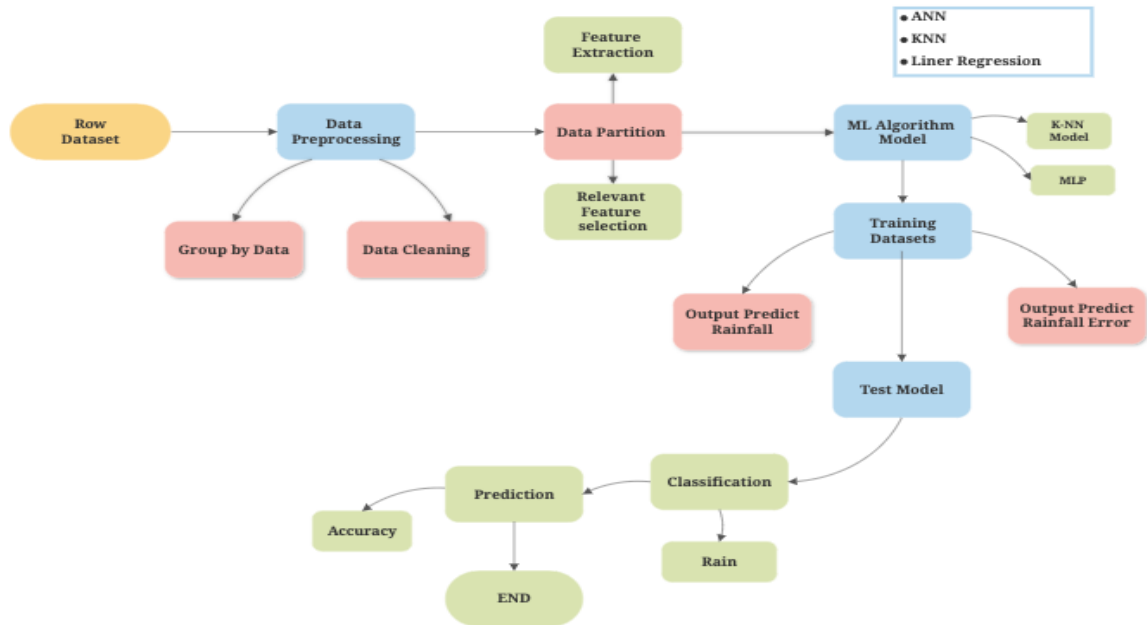

Figure 18: Machine Learning Model

Chapter 4

Result And Analysis

In order to predict rainfall, we use raw weather data from our local weather office. Following the cleaning and preprocessing of the data set, we used several different machine algorithms to generate predictions amongst the test data set below.

4.1   Data Correlations

In order to understand any type of data relationship, it is imperative to understand correlations. A correlation between two variables is when there is some movement between them that indicates they are related. We can see in the figure below that our targeted column is 'observed rainfall Categorical' and that 'observed Pressure' is negatively correlated (-0.576958) with rainfall. Therefore, we move forward without the observed pressure column in order to achieve a better result.

	stationID	Days	observedTempMax	observedTempMin	observedHumidity	observedPressure	observedWind	observedRainfallcatagorical
stationID	1.000000	-0.008462	0.042371	-0.000458	0.009811	-0.006394	0.387652	0.000351
Days	-0.008462	1.000000	0.045088	0.205650	0.437585	-0.072143	-0.266540	0.035526
observedTempMax	0.042371	0.045088	1.000000	0.866812	0.221166	-0.634708	0.298645	0.651053
observedTempMin	-0.000458	0.205650	0.866812	1.000000	0.571570	-0.785690	0.286132	0.714239
observedHumidity	0.009811	0.437585	0.221166	0.571570	1.000000	-0.532134	0.005046	0.366324
observedPressure	-0.006394	-0.072143	-0.634708	-0.785690	-0.532134	1.000000	-0.299707	-0.576958
observedWind	0.387652	-0.266540	0.298645	0.286132	0.005046	-0.299707	1.000000	0.250445
edRainfallcatagorical	0.000351	0.035526	0.651053	0.714239	0.366324	-0.576958	0.250445	1.000000

Figure 19: Correlations of Weather parameters in full data set.

4.2   Analysis and comparison of results

Our aim was to find the best model that maximizes specificity while maintaining acceptable sensitivity using some machine learning prediction algorithms and strategies. We used several techniques and different amounts of data for training data modeling. In the beginning, we collect raw data sets from our local weather department. Our next step was to preprocess the data by using different techniques. The data set of every station is taken for 366 days of the year in order to get information about minimum temperature, maximum temperature, humidity pressure, and wind. There for our aim is to observe the targeted rainfall for our chosen weather parameters. As

part of the preparation of the data set, we used different types of normalization or scaling. The following are some tables that provide details about the results.

A full set of data is taken into consideration for training and testing machine learning models

	By applying MinMax scaler	**By applying standard scaler**
Decision trees (Random Forest classifier)	Accuracy_score: 0.8739266198282591	Accuracy_score: 0.8739266198282591
Gaussian Naive Bayes	Accuracy_score: 0.8727556596409055	Accuracy_score: 0.8727556596409055
Artificial neural network	Test Accuracy: 0.8708040714263916 Test Loss: 0.273854523897171	Test Accuracy: 0.8778298497200012 Test Loss: 0.26832425594329834
Linear Regression	Variance score: 0.5236907253191425 Linear regression r2 score: 0.5237 Error: 0.2974 Linear cross_val_score mean: 0.522 std: (0.032)	Variance score: 0.5236907253191425 Linear regression r2 score: 0.5237 Error: 0.2974 Linear cross_val_score mean: 0.522 std: (0.032)
Logistic Regression	Variance score: 0.8696330991412958 Logistic regression r2 score: 0.2977 Error: 0.3611 Logistic cross_val_score: 0.283 std: (0.089) Logistic cross_val_score(accuracy): 0.867 std: (0.022)	Variance score: 0.8711943793911007 Logistic regression r2 score: 0.3061 Error : 0.3589 Logistic cross_val_score: 0.298 std: (0.086) Logistic cross_val_score(accuracy): 0.870 std: (0.022)
Ridge Regression	Variance score: 0.4651062253132259 Ridge regression r2 score: 0.4651 Error: 0.3151 Ridge cross_val_score mean: 0.463 std: (0.024)	Variance score: 0.46510622531322576 Ridge regression r2 score: 0.4651 Error: 0.3151 Ridge cross_val_score mean: 0.463 std: (0.024)

Table 1: Comparison results of different data scaler

According to the table above, there are two columns. We have taken the whole dataset for both columns, but trained models using MinMax scalars normalization for the first column. The MinMax Scalar is sensitive to outliers because it scales the data to a fixed range (0 to 1). As opposed to standardization, this bounded range leads to smaller standard deviations, which lowers the impact of outliers. The accuracy_score obtained by applying MinMax scaler Decision trees (Random Forest Classifier) is 0.8739266198282591, which is higher than others. An artificial neural network's accuracy (accuracy_score:0.87080) and Gaussian naive bias' accuracy (accuracy_score:0.872755) are close to that. On the other hand, the accuracy of Logistic regression (accuracy_score: 0.86963) also is fine. We can observe from the difference between the min-max scalar and the standard scale that the accuracy of the artificial neural network (accuracy_score:0.877829) and logistic regression (accuracy_score: 0.87119437) has grown somewhat, while the error and loss have dropped slightly. Overall, these two algorithms' accuracy has increased by one percent, and loss decreased by one percent. And we can see that the accuracy and losses of the other algorithms are the same for both scalers.

o   Assuming one station has data, and another has data for the rainy season for all stations.

	When selected rainy season weather data for all station	When considering only single Station for data set
Decision trees (Random Forest classifier)	Accuracy_score: 0.9907407407407407	Accuracy_score: 0.8783783783783784
Gaussian Naive Bayes	Accuracy_score: 0.9708994708994709	Accuracy_score: 0.9054054054054054
Artificial neural network	Test Accuracy: 0.9907407164573669 Test Loss: 0.03988247737288475	Test Accuracy: 0.9189189076423645 Test Loss: 0.3092549443244934
Linear Regression	Variance score: 0.04850479872212354 Linear regression r2 score: 0.0485 Error: 0.0934 Linear cross_val_score mean: 0.014 std: (0.072)	Variance score: 0.7114916088038657 Linear regression r2 score: 0.7115 Error: 0.2564 Linear cross_val_score mean: 0.656 std: (0.205)
Logistic Regression	Variance score: 0.9907407407407407 Logistic regression r2 score: -0.0093	Variance score: 0.9459459459459459 Logistic regression r2 score: 0.7628 Error: 0.2325

	When selected rainy season weather data for all station	**When considering only single Station for data set**
	Error: 0.0962 Logistic cross_val_score: 0.091 std: (0.303) Logistic cross_val_score(accuracy): 0.991 std: (0.007)	Logistic cross_val_score: 0.669 std: (0.435) Logistic cross_val_score(accuracy): 0.934 std: (0.087)
Ridge Regression	Variance score: 0.03691408562448917 Ridge regression r2 score: 0.0369 Error: 0.0940 Ridge cross_val_score mean: 0.020 std: (0.033)	Variance score: 0.5916049995146753 Ridge regression r2 score: 0.5916 Error: 0.3051 Ridge cross_val_score mean: 0.593 std: (0.163)

Table 2: Comparison of results data table

There are two columns, according to the above table. The largest distinction between the two columns is the data collection size. We generated the models using only the specified rainy season data from the data set on the left side, and we trained the models using only one station from the right-side column. We achieved the best results from all data set formats when we constructed a data set containing Rainy Season's data and used it to train machine learning models. Using a decision tree and Logistic regression, the accuracy is the same (accuracy_score:0.9907407407407407), but with higher accuracy. This is also the highest accuracy score for artificial neural networks, 0.9907407164573669. Although Naive Bayes (accuracy_score:0.9708994708994709) is slightly less accurate than artificial neural networks, it is still an ideal score. In contrast, ridge regression (r2 score: 0.0369) and linear regression (r2 score: 0.0485) are smaller in accuracy and more error-prone.

When we did model training with all-day data of only one station instead of model training for the whole data set, the models gave better results than before. Using the full data set, we get an accuracy of 87 % for ANN, Logistic regression, and Gaussian Naive Bayes models. Still, when we use only station data (Dinajpur station), we get an accuracy of more than 90%, and Logistic regression is 94%. Additionally, the accuracy of an artificial neural network increased by 4%. Despite this, decision tree algorithms remain almost as same with fewer data (366). We get lower accuracy from a machine learning algorithm when using a station dataset (Dinajpur station) than when using a rainy season dataset. However, linear regression and Ridge Regression are also less accurate and more error-prone. We were surprised to find that when we calculated cross-validation accuracy for Logistic regression, we got Logistic cross_val_score (accuracy): 0.934 for Dinajpur station and 0.991 for rainy season data with a standard deviation of the error (0.007).

If we add up the entire result Analysis, we can obtain the following overall rating for our data set and research.

- **ANN > Logistic Regression > DT(RFC) > Naive Bayes > Linear Regression > Ridge Regression**

4.3 Best Machine learning algorithm model for our data set

We used various machine learning algorithms with different techniques as part of our paper studies and implementation to find a better prediction model and obtaining better results. Among all algorithms and techniques, we find that decision trees (random forest classifier), artificial neural networks, and logistic regression provide a higher level of test accuracy for our dataset. In all our tests, all machine learning algorithms gave good results and accuracy. Still, we found that the machine learning algorithms below provide the best information for our datasets, which is what we discuss below.

4.3.1 Decision Trees (Random Forest Classifier)

A random forest classifier is an appropriate grouping of decision trees, the results of which are aggregated into a single final output. From everything we have examined thus far, it is clear that the Decision Tree has produced outstanding results throughout this study. Here we can see the Feature importance's weather parameters,

When Full dataset training array ([0.27645116, 0.45781952, 0.14499606, 0.12073326])
When rainy season dataset training array ([0.24772622, 0.34109379, 0.21758822, 0.19359177])

We can see from this array that observedTempMin (0.24772622), observedTempMax (0.3410937), and observed humidity (0.21758822) are more important for the Random Forest classifier. We can see that temperature is a factor in rainfall.

Furthermore, we can observe that the random forest classifier prediction and the actual value are nearly identical. They are compelling models because they limit overfitting without substantially increasing error due to bias.

4.3.2 Artificial neural network

Among all of the algorithms we studied in this study, ANN is a powerful machine learning method. Despite the fact that our model isn't completely /perfectly tuned or hyper-tuned, it works brilliantly with our simple neural network. The most intriguing aspect of our model is that it does not Overfitted. The neural network training accuracy is greater than 85% in all formats, with a training loss of less than 0.3. However, we have the best accuracy in the rainy season data set, with a loss of 0.0533 and an accuracy of 0.9905.
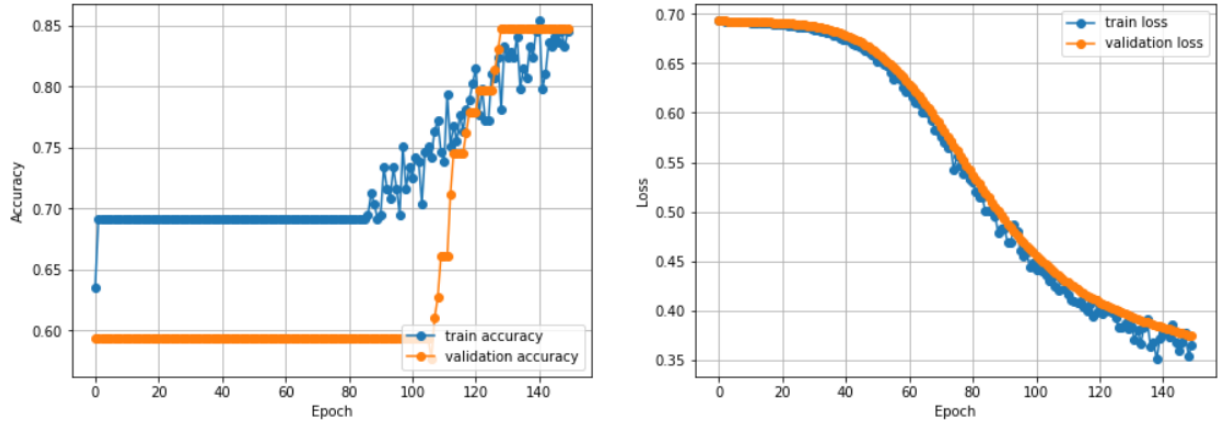
Figure 20: ANN model training with a single station data
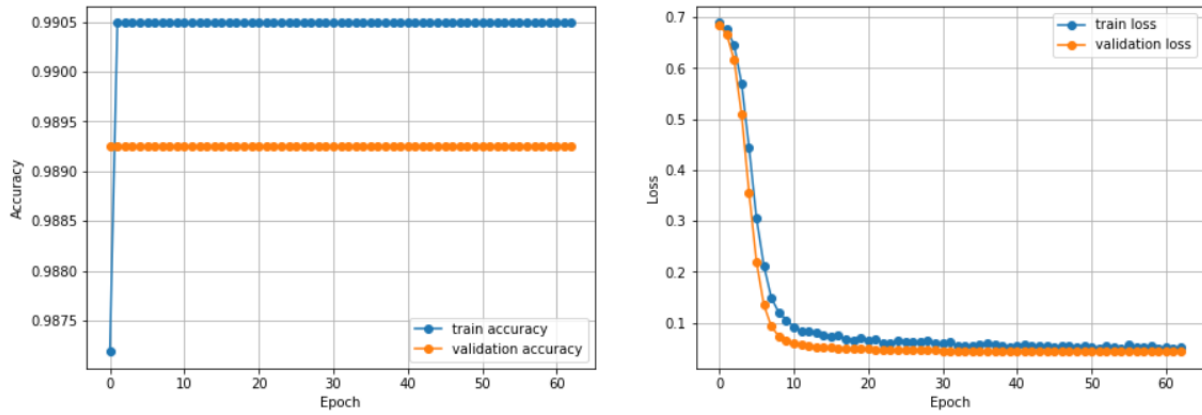
Figure 21: ANN model training with the selected rainy seasonal dataset
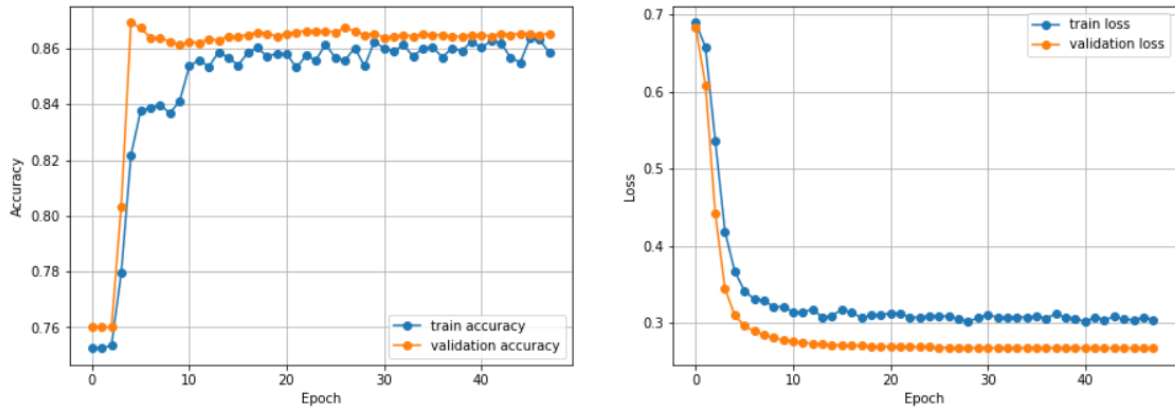
Figure 22: ANN model training with the full dataset

Using artificial neural networks, we extracted patterns and detected trends from our weather data set. In addition, the predicted value is close to the actual value with less error. While the Decision Tree accuracy is quite a bit higher than the Neural Network accuracy, the Neural Network has been

able to capture the patterns of our data sets very nicely, and the Prediction on the test dataset was very accurate.

4.3.3 Logistic Regression

According to theory, our data set would be best suited to a Logistic regression. In logistic regression, categorical dependent variables are predicted as output. It is, therefore, necessary that the outcomes be categorical or discrete. There are two possible answers: Yes or No, 0 or 1, true or false, etc. logistic regression gives the probabilistic values, which lie between 0 and 1, instead of the exact values. As we can see from the confusion matrix for chosen rainy season days on the dataset, confusion matrix: array ([[0, 0],

[14, 1498]], dtype=int64),

our model prediction is also good for Logistic regression approach. We can observe that the confession algorithm has true positive (TP) = 0 and true negative (TN) = 1498 in this confusion matrix. As a result, the Logistic regression r2 score is negative, resulting in a Logistic regression score = -0.0093. On the other hand, this Logistic cross_val_score(accuracy): 0.991 std: (0.007) indicates that our model score is good.

And this one station's logistic regression predicted confusion matrix: array ([[23, 1],

[3, 47]], dtype=int64)

data set (Dinajpur station). In addition, we can see that false positive and false negative values are respectively 3 and 1, which indicates that more data is needed for a better result in logistic regression. Compared to other models, the logistic regression model predicts more accurately with less data training.

Chapter 5

Discussion & Conclusion

5.1 Significance of Result:

The environmental characteristics used in this analysis, which were obtained from a meteorological station and measured by measuring devices, were examined for their applicability to the effects of rainfall. The applicability of the environmental characteristics was determined based on the experiment's Pearson correlation values, which are displayed in Tables 1 and 2 for the daily rainfall prediction. This study examined the rainfall prediction using environmental characteristics with a correlation coefficient larger than 0.2. Similarly, uses a degree of correlation between each measure to identify the five key environmental features: temperature, relative humidity, dew point, solar radiation, and precipitable water vapour. According to the study's experiment correlation, there is a powerful negative association between observedPressure' and relative humidity of about -0.5.

In order to execute the experiment, the researcher employed the year, temperature, cloud cover, and year attributes. Instead of examining the relationships between these environmental factors, the researcher used the monthly and annual rainfall data to forecast the average yearly rainfall.

In order to train and evaluate the three machine learning models (RFC, GNB, ANN, LR (Linear Regression), LR (Logistic Regression), and RR for the prediction of daily rainfall quantity, this study employed the pertinent environment variables. This machine learning models' effectiveness was evaluated using the MinMax scaler and the conventional scaler, respectively.

The researcher used the year, temperature, cloud cover, and year features to carry out the experiment. The researcher projected the average annual rainfall using the monthly and annual rainfall data rather than looking at the correlations between these environmental elements.

This work used the relevant environmental factors to train and assess the three machine learning models (RFC, GNB, ANN, LR (Linear Regression), LR (Logistic Regression), and RR for the

prediction of daily rainfall quantity. The MinMax scaler and the traditional scaler were both used to assess the efficacy of these machine learning models. We determined the cross-validation accuracy for logistic regression, and the results were 0.934 for the Dinajpur station and 0.991 for the rainy season data with an error standard deviation of 0.007. The accuracy score of 0.99 for artificial neural networks is likewise the highest. Naive Bayes accuracy:0.97 is still a perfect score even if it is a little less accurate than artificial neural networks.

5.2   Limitations:

In terms of data collection and processing, and training machine learning algorithms, we have a number of limitations that have resulted in more or less accurate results.

The data set we used had too many missing values, which is one of the problems we encountered. We have filled in all the missing values by using the fill function with the above value. Thus, it affects our accuracy and error rate. During the collection of the dataset, we encountered some limitations. This had an impact on the training of the machine learning algorithms. One column even has a negative correlation with targeted columns, so we must drop it. In the case of only one station, we had very fewer data to train the machine learning algorithms. We would be able to train our models better if there were more data. The dataset was preprocessed and some duplicate rows were detected. As a result, those rows had to be dropped. Preprocessing the data involved storing it in the database, retrieving it, and creating a new data sheet based on the group data. Feature extraction data has been minimally analyzed, so it was challenging to compare.

During the development of the artificial neural network algorithm, we did not hyper-tune or perfectly tune it. Therefore, there would be some limitations. As a result, Traning accuracy levels appear to be less than ideal. We did not choose all of the settings that are connected to improved training when we trained it with our regression algorithms for others and Algorithms. For improved training, artificial neural networks and logistic regression require perfect normalization, however, we did some basic normalizations.

5.3 Future work:

Hopefully, we can use our machine learning algorithms in the future to precisely predict whether or not it will rain, and how much it will rain, based on this Research. We have thus far focused on yes or no (0 or 1) predictions of Rainfall, i.e., whether it will rain. It is also possible to predict how much rain will fall in the future. We can estimate how much it will rain if it rains. We can calculate how much rain will fall in advance to predict floods. Our algorithms can be set up so that they can be trained better if we set every option. Our training accuracy will increase if we apply all of the best normalization techniques to our data. Our artificial neural network models can be hyper-tuned and perfectly tuned. To create hidden layers and final layers we can use the right loss function and other things. We can test different test data so we can observe that our machine learning algorithms can accurately predict how much data will vary. Additionally, we can add other weather parameters to our data set that are not included in our existing data set. We can use a near-perfect technique to preprocess our data. We can also do numerous data analyses and feature extractions. We will conduct more research and develop various sorts of machine learning algorithms to get much better outcomes. We may do several forms of climate or weather analysis to forecast rain and winter, as well as various calamities. So that our people might be protected from floods and various natural disasters.

5.4 Conclusion

In Bangladesh rainfall is the most common and important incident not only for human being but also for animals, birds, insects, and trees. The full ecosystem depends on natural water. To run the ecosystem smoothly needs the proper use of water and the preparation to avoid uncertain situations like flood, drought, tornadoes, hurricanes, etc. Is this paper we discussed on some ML algorithms and analysis them to find out the best result by which we can predict whether there is any probability to happen rain or not. We analysis the Logistic Regression Algorithm, Gaussian Naïve Bayes Algorithm, Linear Regression Algorithm, Decision Tree Algorithm, ANN Algorithm and Ridge Regression Algorithm to find out best prediction of precipitation. We found the highest accuracy of the prediction by analysis the ANN Algorithm. The prediction of precipitation is very

important for our cattle and agriculture because it helps us to protect our crops and cattle from sudden floods, landslides, etc.

For this study, we built a system to predict precipitation using Artificial Neural Networks, decision trees (random forest classifier), and logistic regression for stations such as Dinajpur. After completing the training, and testing we compare the results to check the efficiency of the models. Finally, after successful training, testing, and comparing the actual and predicted data among the results obtained from the different ML algorithms, we got the best efficient result by using ANN Algorithm with the minimum error.

Ultimately, our research has been successful in every aspect we set out to accomplish. Nevertheless, it's not the end of the story. The current flood situation in Bangladesh reminds us how important and helpful it would be for us and flood-affected people to know how much rain will fall in the future. Moreover, if they had been warned in advance that more rain or flooding would occur, they might have been prepared in advance, and the damage to the country and the nation might have been reduced.

References

[1] Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S., "A Rainfall Prediction Model using Artificial Neural Network," *IEEE Control and System Graduate Research Colloquium*, 1- 4, 2012.

[2] Ahn, J. "Analysis of a neural network model for building energy hybrid controls for the in-between season," *Architecture of Complexity*, (pp. 1-4), 2017.

[3] Selase, A. E., Agyimpomaa, D. E., Selasi, D. D., & Hakii, D. M. "Precipitation and Rainfall Types with Their Characteristic Features," *Journal of Natural Sciences Research, 5*(20), 1-3, 2015.

[4] "Advancing the science of climate change. National Research Council; Division on Earth and Life Studies; Board on Atmospheric Sciences and Climate; America's Climate Choices," *Panel on Advancing the Science of Climate Change* ,2010

[5] Sharma, A., & Nijhawan, G. "Rainfall Prediction Using Neural Network," *International Journal of Computer Science Trends and Technology, 3*(3), 1-4, 2015.

[6] Jayanta Kumar Basak, Rashed Al Mahmud Titumir and Nepal Chandra Dey, "Climate Change in Bangladesh: A Historical Analysis of Temperature and Rainfall Data," Vol. 02, Issue 02, pp. 41-46, January,2013.

[7] Jayanta Kumar Basak, Rashed Al Mahmud Titumir and Nepal Chandra Dey, "Climate Change in Bangladesh: A Historical Analysis of Temperature and Rainfall Data," Vol. 02, Issue 02, pp. 41-46, January,2013.

[8] J. Scott Armstrong, "Research Needs in Forecasting," Vol. 04, pp. 449-465, January,1988

[9] Brijesh Kumar Bhardwaj, Saurabh Pal, "Data Mining: A prediction for performance improvement using classification," Vol. 9, No. 4, April, 2011

[10] CMAK Zeelan Basha, Nagulla Bhavan, Ponduru Bhavya, Sowmya, "Rainfall Prediction Using Machine Learning & Deep Learning Techniques," *International Conference on Electronics and Sustainable Communication Systems (ICESC 2020).*

[11] K.Sarvani, Y.Sai Priya, Ch.Teja, T.Lokesh, E.Bala Bhaskara Rao , " RAINFALL ANALYSIS AND PREDICTION USING MACHINE LEARNING TECHNIQUES ," Issue 06, *Journal of Science and Engineering,* Vol 12.

[12] ZANYAR RZGAR AHMED, "RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES," *NEAR EAST UNIVERSITY.*

[13] Awkash Kumar, Rashmi S Patil, A. K. Dikshit, Rakesh Kumar, "Analysis of Weather Parameters for Three Hourly Intervals," Volume 7, Issue 3, March 2016.

[14] Lyudmyla Didyk, Yuriy Gorgo, Alina Prigancova, Igor Tunyi, Magdalena Vaczyova, Sergey Mamilov and Joris Dirckx, "The Effects of Atmospheric Pressure Fluctuations on Human Behavior Related to Injury Occurrences: Study on the Background of Low and Moderate Levels of Geomagnetic Activity," Volume 2012.

[15] Aida Tayebiyan, Thamer Ahmad Mohammad, Abdul Halim Ghazali, Marlinda Abdul Malek, Syamsiah Mashohor, "Potential Impacts of Climate Change on Precipitation and Temperature at Jor Dam Lake," Vol 24 (1), pp. 575 – 586, 2016.

[16] Juliana Aparecida Anochi, Vinícius Albuquerque de Almeida, Haroldo Fraga de Campos Velho, "Machine Learning for Climate Precipitation Prediction Modeling over South America," *National Institute for Space Research*, São José dos Campos 12227-010, Brazil.

[17] A.H.M. Rahmatullah Imon, Manos C Roy, S. K. Bhattacharjee , "Prediction of Rainfall Using Logistic Regression," 3, 655-667, 2012.

[18] Rajesh Kumar, "Decision Tree for the Weather Forecasting," *International Journal of Computer Applications (0975 – 8887)*, Volume 76– No.2, August, 2013.

[19] Qingyi Feng, Ruggero Vasile , Marc Segond et al.:ClimateLearn, "A machine-learning approach for climate prediction using network measures" gmd-2015-273.

[20] Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, Siva Sai Maturi, "Prediction Of Rainfall Using Machine Learning Techniques"2020.

[21] Steve Oberlin, "Machine Learning, Cognition, and Big Data" CA Technologies, 2012.

[22] Jainendra Singh et.al., "BIG DATA ANALYTICS: CHALLENGES, TECHNOLOGIES AND KEY APPLICATIONS," 2014
[23] Junfei Qiu et al., "A survey of machine learning for big data processing" 2016.

[24] Yasir Safeer, et.al. "Clustering Unstructured Data (Flat Files) - An Implementation in Text Mining Tool," 2010.

[25] Rohit Bhatnagar, et.al., "Machine Learning and Big Data Processing: A Technological Perspective and Review," 2018.

[26] Alexandra L'Heureux, et.al., "Machine Learning With Big Data: Challenges and Approaches," 2017.

[27] Vijayan R, Mareeswari V, Mohankumar P, Gunasekaran G, Srikar K, "Estimating rainfall prediction using machine learning techniques on a dataset," *International Journal of Science and Technology*, Res. 2020;9(06):440–5 JUNE,2020.

[28] Manandhar S, Dev S, Lee YH, Meng YS, Winkler S., "A data-driven approach for accurate rainfall prediction," *IEEE Trans Geoscience Remote Sens*. *2019*;5(11):9323–31.

[29] Zeelan BCMAK, Bhavana N, Bhavya P, Sowmya V., "Rainfall prediction using machine learning & deep learning techniques," *Proceedings of the International Conference on Electronics and Sustainable*

[30] *Tutorials Point*. [Last accessed on 2022 1 june]. Available from: https://www.tutorialspoint.com/keras/keras_discussion.html

[31] Sampliner. [Last accessed on 2022 1 june]. Available from: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow

[32] *Towards Data Science*. [Last accessed on 2022 1 july]. Available from: https://towardsdatascience.com/an-introduction-to-artificial-neural-networks-5d2e108ff2c3

[33] *Towards Data Science*. [Last accessed on 2022 1 july]. Available from: https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7

[34] *IBM*. [Last accessed on 2022 1 july]. Available from: https://www.ibm.com/cloud/learn/recurrent-neural-networks

[35]   Schultz M. G., Betancourt C., Gong B., Kleinert F., Langguth M., Leufen L. H., Mozaffari A. and Stadtler S. 2021Can deep learning beat numerical weather prediction?Phil. Trans. R. Soc. A.379:2020009720200097

APPENDICES

APPENDIX A

Dataset of Observed Weather

station ID	Days	observedTemp Max	observedTemp Min	observedHumidity	observedPressure	observedWind	observedRainfallcatagorical
10120	1	23.18205128	10.73589744	77.5862069	1016.255172	3.968965517	0
10120	2	22.9974359	10.73846154	78.65517241	1016.424138	4.017241379	0
10120	3	22.27692308	10.56666667	80.55172414	1016.227586	4.306896552	0
10120	4	22.11538462	10.71282051	79.86206897	1016.027586	4.234482759	1
10120	5	22.32564103	10.27435897	79.75862069	1015.627586	4.434482759	0
10120	6	22.64871795	10.02820513	79	1015.765517	3.410344828	0
10120	7	22.51282051	10.31025641	79.65517241	1016.141379	3.972413793	0
10120	8	22.53333333	10.18205128	78.93103448	1015.827586	3.806896552	1
10120	9	22.09230769	10.49487179	80.03448276	1015.786207	3.679310345	1
10120	10	21.91794872	9.892307692	79.4137931	1015.903448	3.55862069	0
10120	11	22.10512821	10.01025641	80.24137931	1015.724138	4.117241379	0
10120	12	22.03076923	10.04358974	79.5862069	1015.431034	3.765517241	0
10120	13	22.32307692	9.679487179	78.82758621	1015.065517	3.993103448	1
10120	14	22.24102564	9.556410256	79.86206897	1015.210345	4.175862069	0
10120	15	22.43846154	9.679487179	79.65517241	1015.665517	3.655172414	0
10120	16	22.93589744	9.994871795	78.17241379	1015.306897	3.8	1
10120	17	22.83076923	10.26153846	77.96551724	1015.486207	3.803448276	1

10120	18	23.11282051	9.976923077	77.27586207	1015.496552	3.693103448	0
10120	19	23.00769231	10.11794872	77.93103448	1015.051724	4.744827586	0
10120	20	22.3538461	10.54615384	80.0344827	1015.2620689	4.18620689	1