PROJECT REPORT

ON

RAINFALL PREDICTION USING MACHINE LEARNING

NIELIT Guwahati
July, 2021

Submitted By:

Biki Deka

Souvik Das

Nikhil Kumar Yadav

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

ASSAM ENGINEERING COLLEGE

JALUKBARI- 781013,

GUWAHATI

CERTIFICATE

Acknowledgement

I would like to express my sincere thanks to Mr. David Ray and Mr. Apurba Dey for his valuable guidance and support in completing this project.

I would also like to express my gratitude towards our principal Dr. Atul Bora and Head of the Department Prof. Dinesh Shankar Pegu for giving me this great opportunity to do a project on "Rainfall Prediction using Machine Learning in Python" . Without their support and suggestions, this project would not have been completed . I would like to extend my deep appreciation to all my group members, without their support and coordination we would not have been able to complete this project. I hope we will achieve more in our future endeavours.

Place : AEC, Jalukbari

Date : 22/07/2021

Declaration

We declare that this written submission represents our ideas in our own words and where other's ideas or words have not been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honestyand integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above willbe cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Name of the student)
Biki Deka (AEC)
Souvik Das (AEC)
Nikhil Kumar Yadav (AEC)

Abbreviations

AI: Artificial intelligence

ML: Machine Learning

RFC: Random Forest Classifier

ROC: receiver operating characteristic

Chapter 1

Introduction & Objective

What is machine learning?

According to Arthur Samuel, Machine Learning algorithms enable the computers to learn from data and even improve themselves, without being explicitly programmed.

Importance of rainfall predictions are as follows: -

Rainfall plays an important role in forming of fauna and flora of natural life. It plays a significant role in agriculture and farming and therefore, water is one of the most natural resources on earth. The changing climatic conditions and the increasing greenhouse emissions have made it difficult for the human beings and the planet earth to experience the necessary amount of rainfall that is required to satisfy the human needs and its uninterrupted use in everyday life. Therefore, it has become significant to analyse the changing patterns of the rainfall and try to predict the rain not just for the human needs but also to predict for natural disasters that could cause by the unexpected heavy rainfall. To be more specific and aware of the devastating climatic changing and stay updated by predicting rainfall has been the focus of computer scientist and engineers.

- Rainfall prediction is important as heavy rainfall can lead to many disasters.
- The prediction helps people to take preventive measures and moreover the prediction should be accurate.
- Prediction mostly short-term prediction can give us the accurate result.
- Heavy rainfall is a cause for natural disasters like flood and drought that square measure encountered by individuals across the world each year thereby, accuracy of rainfall statement has a great importance.
- The dynamic nature of atmosphere, applied mathematics techniques fail to provide sensible accuracy for precipitation statement.

- The prediction of precipitation using machine learning techniques may use regression. Intention of this project is to offer non-expert easy access to the techniques; approaches utilized in the sector of precipitation prediction and provide a comparative study among the various machine learning techniques.

Objective of the project

The main objective of the study is the prediction of the rainfall using historical monthly data based on the Machine learning methodologies or algorithms to predict the higher outcomes. The similar data will be grouped together for the accurate and precise information that will predict the rainfall more correctly and with perfect figures. The accurate and exact predictions will help in developing the more appropriate strategies for agriculture and water reserves and will also be informed about the flood to implement precautionary measures. The data for the rainfall prediction is collected from the Kaggle dataset of Australia i.e., weatherAUS.csv file. Here, the monthly data with all parameters of rainfall including wind speed, direction, air pressure, humidity and temperature. The aim of the proposed study is too effective and efficient in predicting the rainfall with good accuracy and precision. Rainfall prediction is significant not only on the micro but also on the macro level. The study is of significance with respect to its vital contribution in the field of agriculture, water reserve management, flood prediction and management with an intention to ease the people by keeping them updated with the weather and rainfall prediction. It is also important to be utilized by the agricultural industries for keeping their crops safe and ensure the production of seasonal fruits and vegetables by updated rainfall prediction. The study will also be significant for the flood management authorities as more precise and accurate prediction for heavy monsoon rains will keep the authorities alert and focused for an upcoming event that of which the destruction could be minimized by taking precautionary measures. The rainfall prediction will impressively help in dealing with the increasing issue of water resource management; as water is a scarce resource and it needs to get saved for the benefit of human beings themselves. Also, it will help the people to manage and plan their social activities accordingly.

Chapter 2

Theoretical Background

The aim of the study is the prediction of the rainfall using historical data based on artificial intelligence methodologies such as Logistic Regression and Decision Tree. The extraction procedures/algorithms will produce the output by classification of the data according to the categories. The similar data will be grouped for the accurate and precise information that will predict rainfall more correctly and with perfect figures. The accurate and exact predictions will help in developing the more appropriate strategies for agriculture and water reserves and will also be informed about the flood to implement precautionary measures. This is the data includes all parameters of rainfall including wind speed, direction, air pressure, humidity, temperature etc. The aim of the proposed study is too effective and efficient in predicting the rainfall with accuracy and precision.

The predictive model is used for prediction of the rainfall. The first step is converting data in to the correct format to conduct experiments then make a good analysis of data and observe variation in the patterns of rainfall. We predict the rainfall by separating the dataset into training set and testing set then we apply machine learning approaches and statistical techniques and compare and draw analysis over various approaches used. With the help of numerous approaches, we attempt to minimize the error.

Machine Learning Algorithms Used:

Logistic Regression: Logistic Regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), x. Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function. Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

Decision Tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be *"learned"* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called *recursive partitioning*. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

Random Forest Regression: Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is Aggregation.

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap andAggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model.

Chapter 3

Project Execution

Data Exploration: The dataset consists of different data like temperature, windspeed, humidity etc. of different Australian cities. Its size is (142193, 24). Our target feature is “Rain Tomorrow”. But it is an unbalanced dataset because of class imbalance of target feature. The ratio between “0” and “1” in target feature is 78.2. So, we need to balance the dataset.

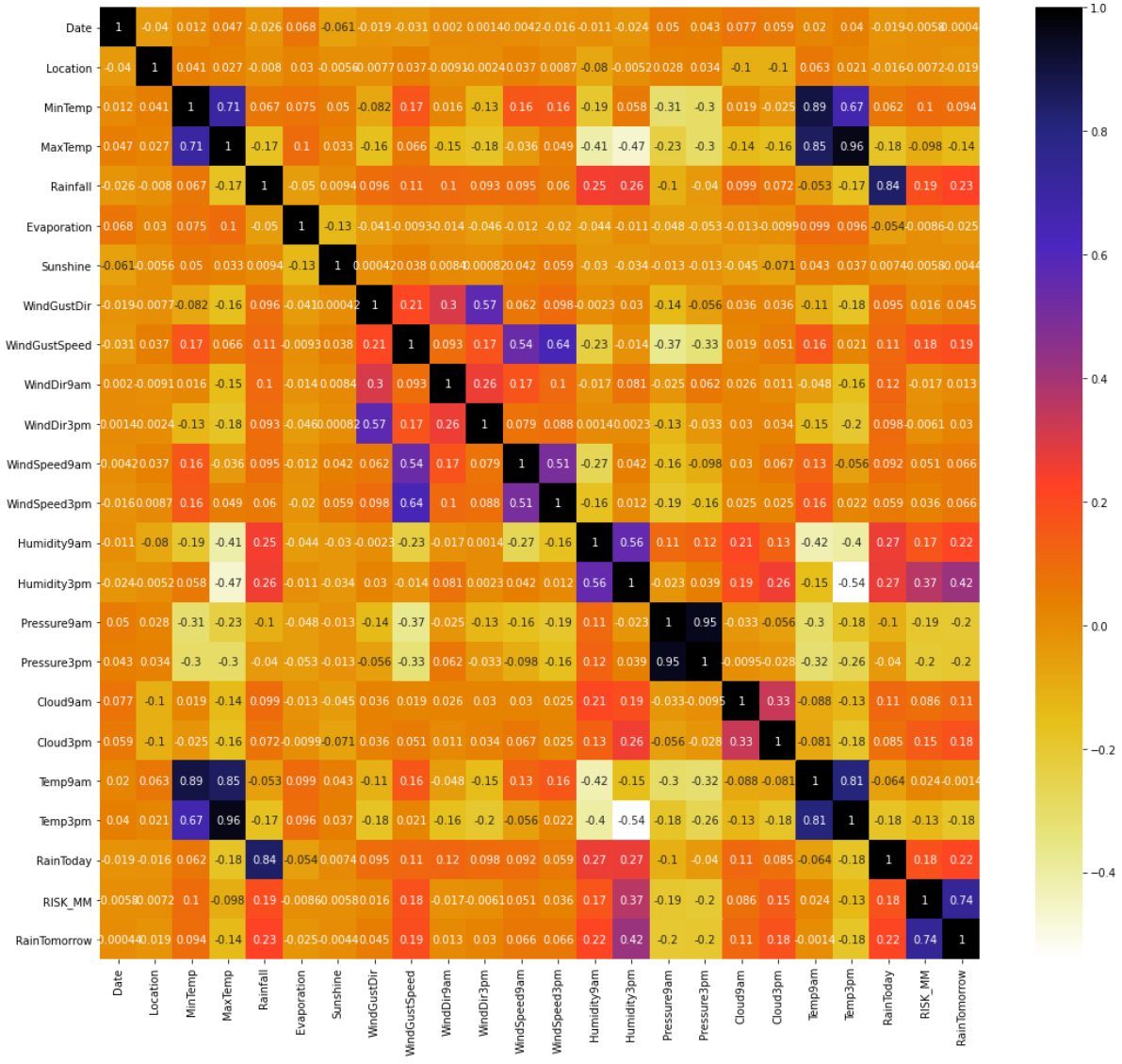
Handling Class Imbalance for Rainfall Prediction: Using sklearn’s (a python library for implementing machine learning models and it can do different data handling tasks) resample method we have up sampled the minority class in target feature. We up sampled because dataset was small.

Filling missing values: Almost all the features have missing values. But none of the features have more than 50% missing data. So, we can consider to fill the missing values instead of ignoring the features with most missing values.

We will impute the categorical columns with mode and numerical columns with mean. Also, categorical features are converted to numeric using Label encoder.

Detecting and removing outliers: An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data. We will detect outliers using the interquartile range (IQR) and remove them to get the final working dataset. Whichever data do not fall into the IOR we will remove them. For this sklearn have a function quantile () which will give the interquartile range.

Correlation between features: Two features sometime co relate to each other. In that case one feature changes in relation to the other. Pearson correlation coefficient is used to check correlation if two feature has correlation 1 then one of them is removed from the dataset.



This above is a heatmap of correlation. The following features have strong correlation.

- MaxTemp and MinTemp
- Pressure9h and pressure3h
- Temp9am and Temp3pm
- Evaporation and MaxTemp
- MaxTemp and Temp3pm

But nowhere in the correlation value is equal to a perfect “1”. So, we are not removing any features.

Training Rainfall Prediction Model with Different Models: We will divide the dataset into training (75%) and test (25%) sets respectively to train the rainfall prediction model. For best results, we will standardize our X_train and X_test data:

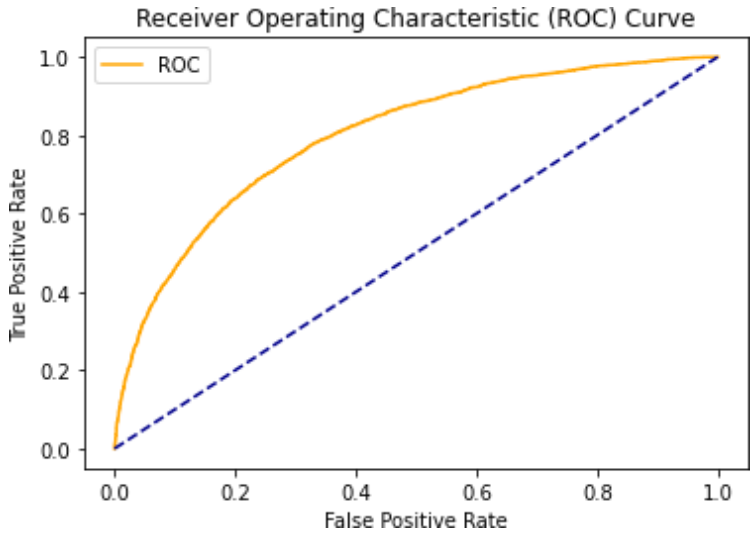
Three different models have been used are:

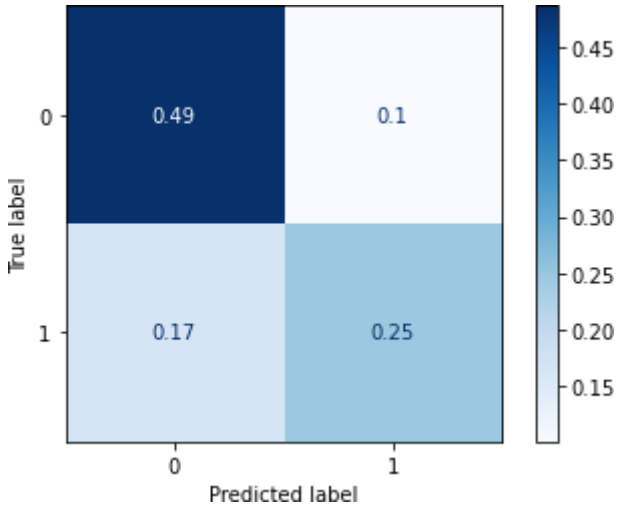
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Model	Accuracy	ROC_AUC	Time taken
Logistic Regression	0.733489	0.713486	0.822428
Decision Tree	0.811272	0.807737	0.161130
Random Forest	0.897179	0.892464	10.557426

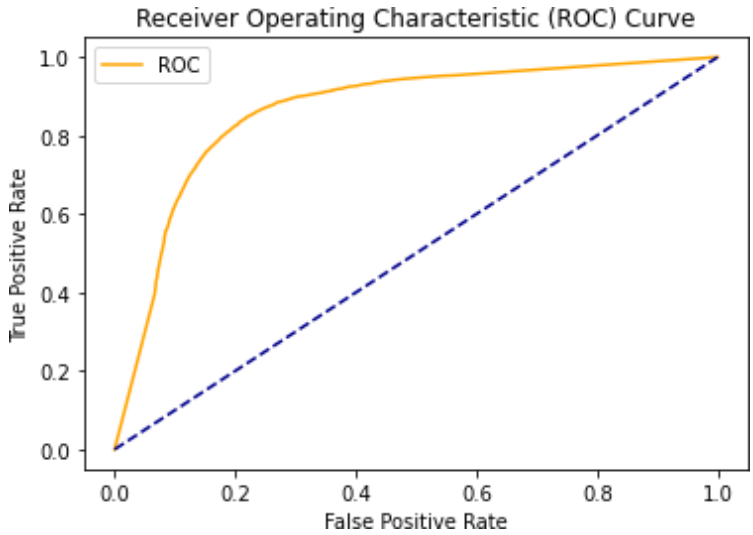
ROC-AUC curve and confusion metrics:

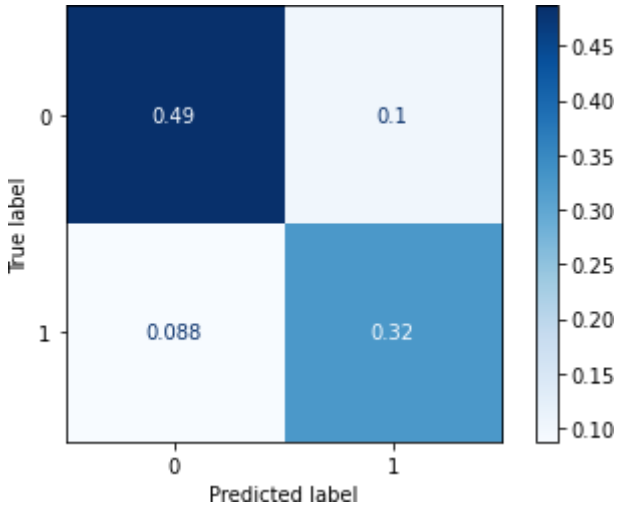
Logistic Regression



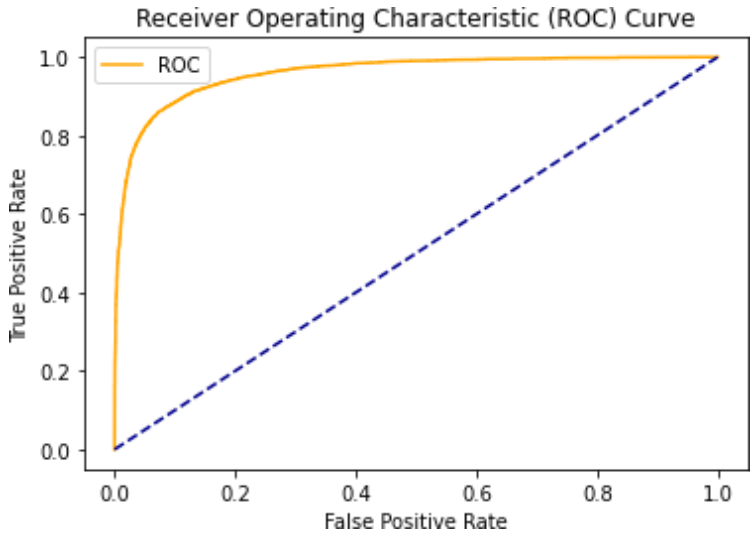


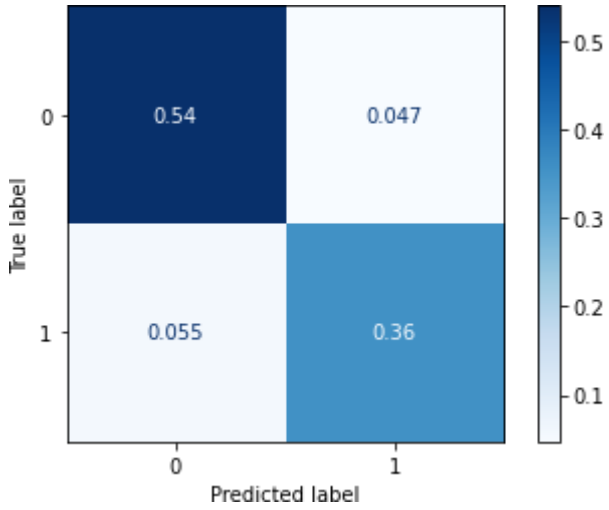
Decision Tree Classifier



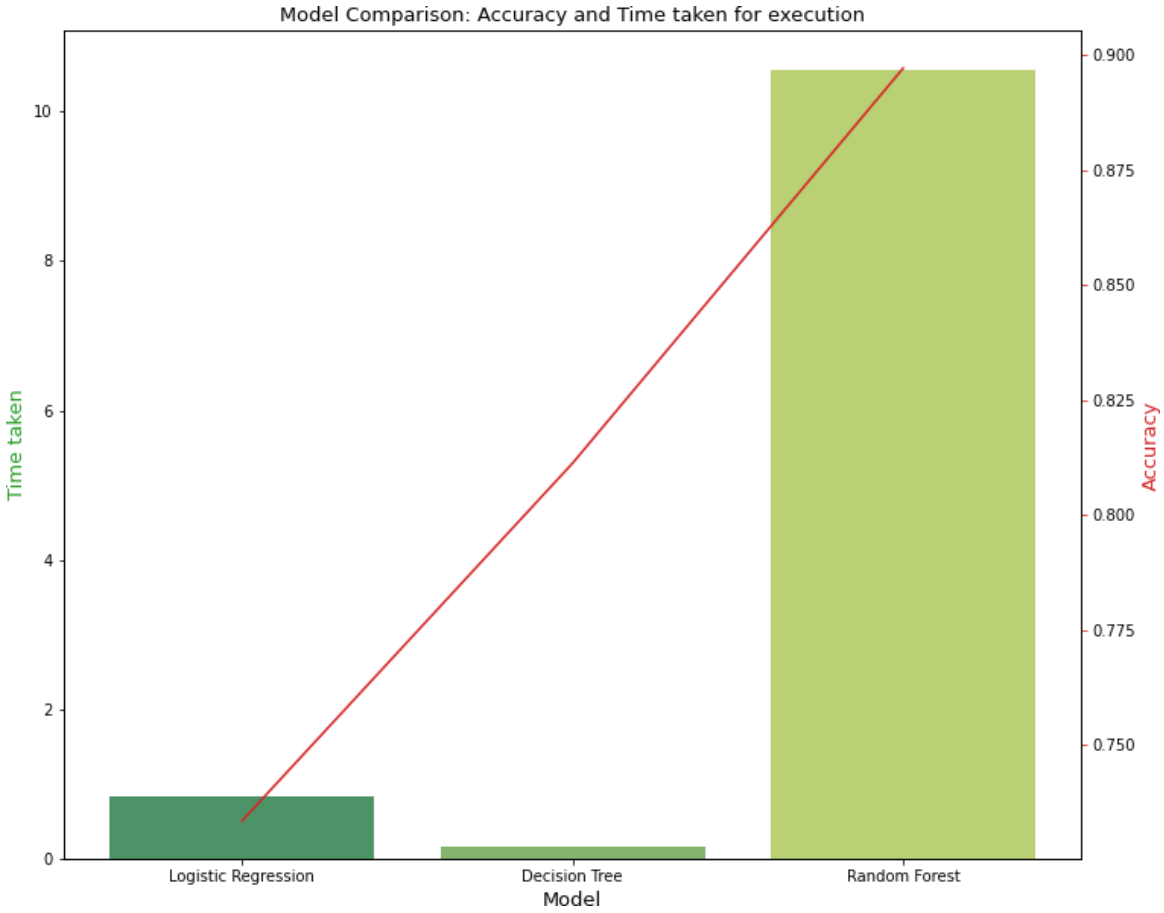


Random Forest Classifier





Rainfall Prediction Model Comparison: Now we need to decide which model performed best. For balanced dataset accuracy metric is enough for this task. So, although our dataset was unbalanced initially but as we have balanced the dataset by oversampling so we can use accuracy.



Finally, we can see random forest performed best than all the other models.

Chapter 4

Conclusion and Future Prospects

The estimation of rainfall is of great importance in terms of water resources management, human life and their environment. It can be met with the incorrect or incomplete estimation problems because rainfall estimation is affected from the geographical and regional changes and properties. This paper presented review of different methods used for rainfall prediction and problems one might encounter while applying different approaches for rainfall forecasting. Rainfall is one the most significant natural phenomenon that is not only important for the human beings only but the living beings. Due to the changing climatic conditions, rainfall cycles are also changing and the temperature of the earth is rising. The changing temperature is also affecting the agriculture, industry and sometimes may cause flooding and land slide. Therefore, it is essential for the human beings to keep a check upon this natural phenomenon in order to survive. The water is a scarce natural resource without which human life is impossible and also there is no substitute to this natural resource. Thus, predicting the rainfall for agriculture and water reserves, also it also good for keeping human beings alert of natural disasters like flood and landslide. However, to overcome these issues and meet the demands, a system to forecast rainfall is essential using machine learning models that is popular within the modern technology.

The estimation of rainfall is of great importance in terms of water resources management, human life and their environment. It can be met with the incorrect or incomplete estimation problems because rainfall estimation is affected from the geographical and regional changes and properties. This paper presented review of different methods used for rainfall prediction and problems one might encounter while applying different approaches for rainfall forecasting.

Future prospect : The future enhancement of this project can be an approach towards about how to reduce the percentage of errors present. Along with that one of the major enhancements will be to decrease the ratio for train data to test data, so that it will assist in improving the level of prediction within the available time and complexity. The accuracy of the algorithm can be additionally tested on increase in the complexity. Many other types of errors can be calculated in order to test the accuracy of any of the above algorithms.

Henceforth, algorithm for testing daily basis dataset instead of accumulated dataset could be of paramount Importance for further research. More the accuracy of the system used for rainfall prediction, smarter will be the agriculture. Along with that, this will be an efficient tool for people in coastal areas of the country thereby making them well aware of the situation in advance.

The development of advanced weather-casting systems for severe weather events is ongoing in several countries. Researchers are exploring several approaches to the problem of very short-range forecasts that are highly specific in time and space. These approaches vary widely, ranging from extrapolation to expert systems to explicit numerical modelling of storm cells. They all share three common needs: data, data and even more data!

The observational data must be sufficient to characterize the heavy rainfall and its environment in great detail. Herein lies the dilemma: how to ensure that the measurement systems will be available where and when they will be needed. Part of the answer lies in determining in advance what locations will be served by a weather-casting capabilities. But an equally important question is, “Who will be responsible for supporting these weather-casting systems?” . Will they be a public or a private enterprise? Or will there be public private partnerships that emerge to meet these needs? One thing seems certain: weather-casting will be an ever more important and valuable component of the weather forecasting paradigm

References

[1] Thirumalai, Chandrasegar, et al. "Heuristic prediction of rainfall using machine learning techniques." 2017 International Conference on Trends in Electronics and Informatics (ICEI). IEEE, 2017.
[2] Geetha, A., and G. M. Nasira. "Data mining for meteorological applications: Decision trees for modeling rainfall prediction." 2014 IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2014
[3] Parmar, Aakash, Kinjal Mistree, and Mithila Sompura. "Machine learning techniques for rainfall prediction: A review." 2017 International Conference on Innovations in information Embedded and Communication Systems. 2017.
[4] Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi. "Rainfall prediction for the Kerala state of India using artificial intelligence approaches." Computers & Electrical Engineering 70 (2018): 66-73.
[5] Pinky Saikia Dutta, Hitesh Tahbilder, "Prediction of Rainfall Using Data Mining Technique Over Assam", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 5 No.2 Apr May 2014.
[6] M.Kannan, S.Prabhakaran, P.Ramachandran, "Rainfall Forecasting Using Data Mining Technique",