Application of Machine learning models in the prediction of floods (Using Bangladesh as a case Study)

Applied AI (Machine Learning Methods) in Flood Prediction

Ayomiposi Adebayo [1]
✉ adebayoayomiposi25@gmail.com
[1] *Department of Data Science and Business Analytics, University of Plymouth*

Introduction

Flood has been identified as one of the most occurring natural disasters that results in the loss of human lives, farmlands, agriculture, households, socio-economic systems, and the environment which in turn affects the economy of nations at large(Parag et al. 2021). Machine learning methods has been efficient on providing flood prediction models using findings from data to develop a model that could be replicated on new datasets based on the accuracy and precision of the prediction model(Felix and Sasipraba 2019). The most common cities that has been recorded to be prone to flooding are India, Bangladesh, and China. Every year, roughly 4.84 million people in India, 3.84 million people in Bangladesh and 3.28 million people in China are affected by Flooding. Over 80% of lands in Bangladesh are prone to floods and flooding was recorded 78times between 1971 and 2014 killing about 41,783 people(Dewan 2015).

Machine Learning Models Employed for the Prediction of Floods

Machine Learning Models used for Flood Prediction in this study are:

- K-Nearest Neighbor(KNN): This is a non-parametric supervised machine learning method that is most commonly used for both regression and classification issues. This model checks for the similarities between a new data from an existing data and place the new data points in classes that are similar to the existing data(Sankaranarayanan et al. 2019)

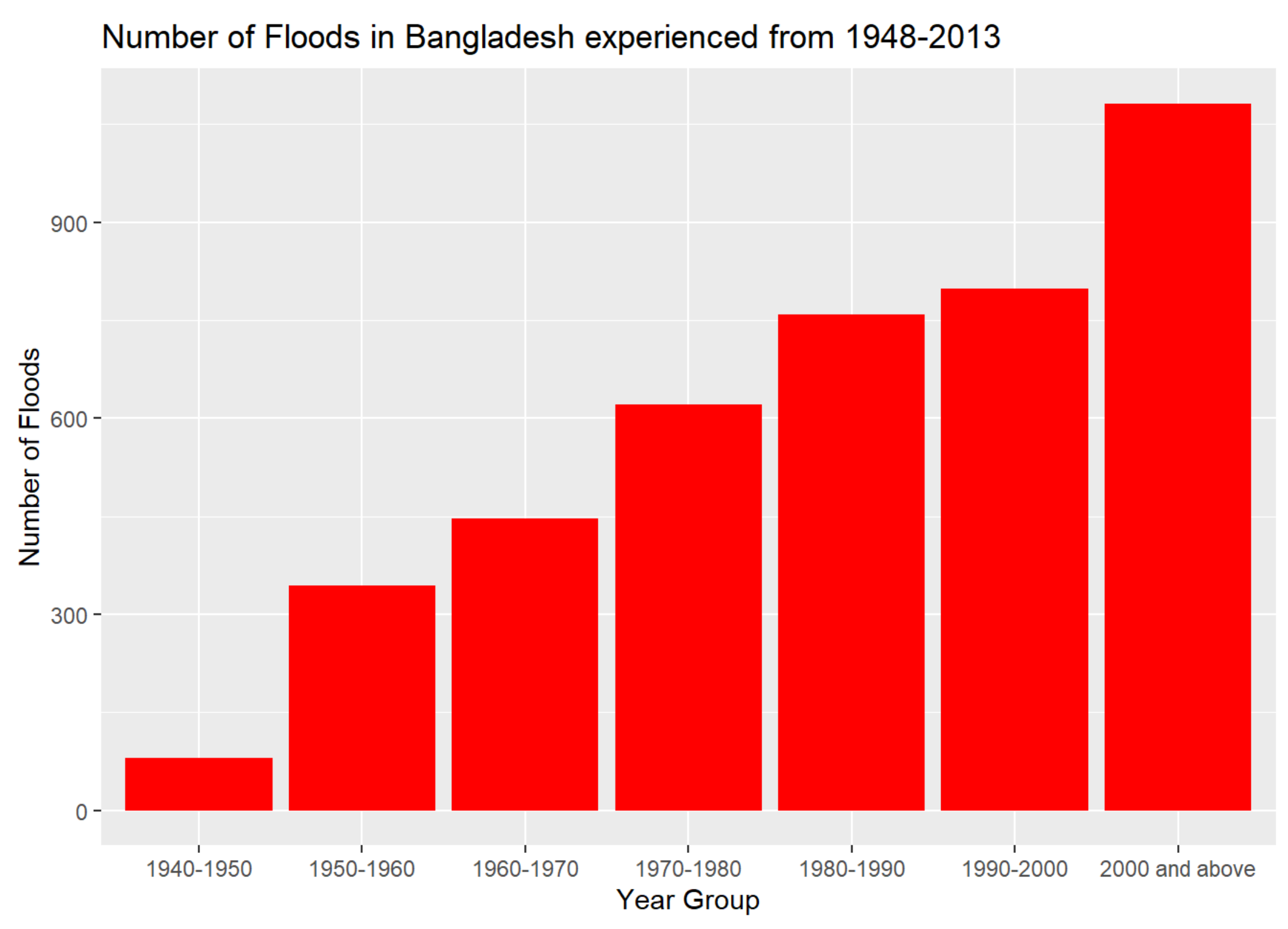
- Logistic Regression: Logistic regression is a statistical approach for analyzing the interaction of dependent and one or more independent variables in linear and nonlinear regression (Grover 2020). Logistic regression is employed to forecast the likelihood of a flood happening depending on several elements involving rainfall, the humidity of the soil, and river velocity(Lee and Kim 2021).

- Decision Tree: A decision tree is a supervised classification machine learning technique that uses a tree to illustrate the basic trends in a dataset and how they interact, with each branch representing potential decision-making or likelihood of an event's fate. Potential flood risks can be discovered by simulating various scenarios and actions(Tehrany, Pradhan, and Jebur 2013)

Methods and Results

Dataset Extraction and Preparation: The historical weather data was from Bangladesh Meteorological Department (BMD) and uploaded by some researchers on Kaggle. The data was collected from 32 stations and contains 20,544 observations within year 1948-2013. Important features in the dataset includes: Rainfall, Cloud Coverage, Rainfall, Flood status, Minimum Temperature, Maximum Temperature, Relative Humidity, Wind speed, Month when the flood occurred etc(Reza 2020).

Below shows the distribution of flood by Year, Rainfall and Month



Dataset Processing: This involves feature engineering, encoding, and scaling. The features of interest in the data were changed into numeric variable types, missing values were replaced with zero, flood status variable was encoded into binary values "0 and 1" against "Yes and No". The dataset was then splitted using the ratio 70:30 to train and test set.

Feature Selection: The features employed in this prediction model were selected using correlation analysis to establish a relationship between the dependent and independent variables using Pearson and Spearman Rank Correlation. Rainfall and Cloud Coverage variables were selected as they had correlation coefficient values close to 1 according to the table below:

The results gotten from the Pearson Correlation Analysis:

```
##      Rainfall Cloud_Coverage         Month
##     0.7688156      0.5736647     0.1054965
```

The results gotten from the Spearman Rank Correlation Analysis:

```
##      Rainfall Cloud_Coverage         Month
##     0.6626058      0.5635919     0.1054965
```

Exploratory Analysis: The two selected features were then subjected to some exploratory analysis to shows the existing relationship between the selected independent variables(Rainfall and Cloud_Coverage) and the dependent variable (Flood_status) using boxplot.

Results derived from the plots below showed that cities with more rainfall and Cloud_Coverage had encountered more flooding compared to the cities with less rainfall and Cloud_Coverage.

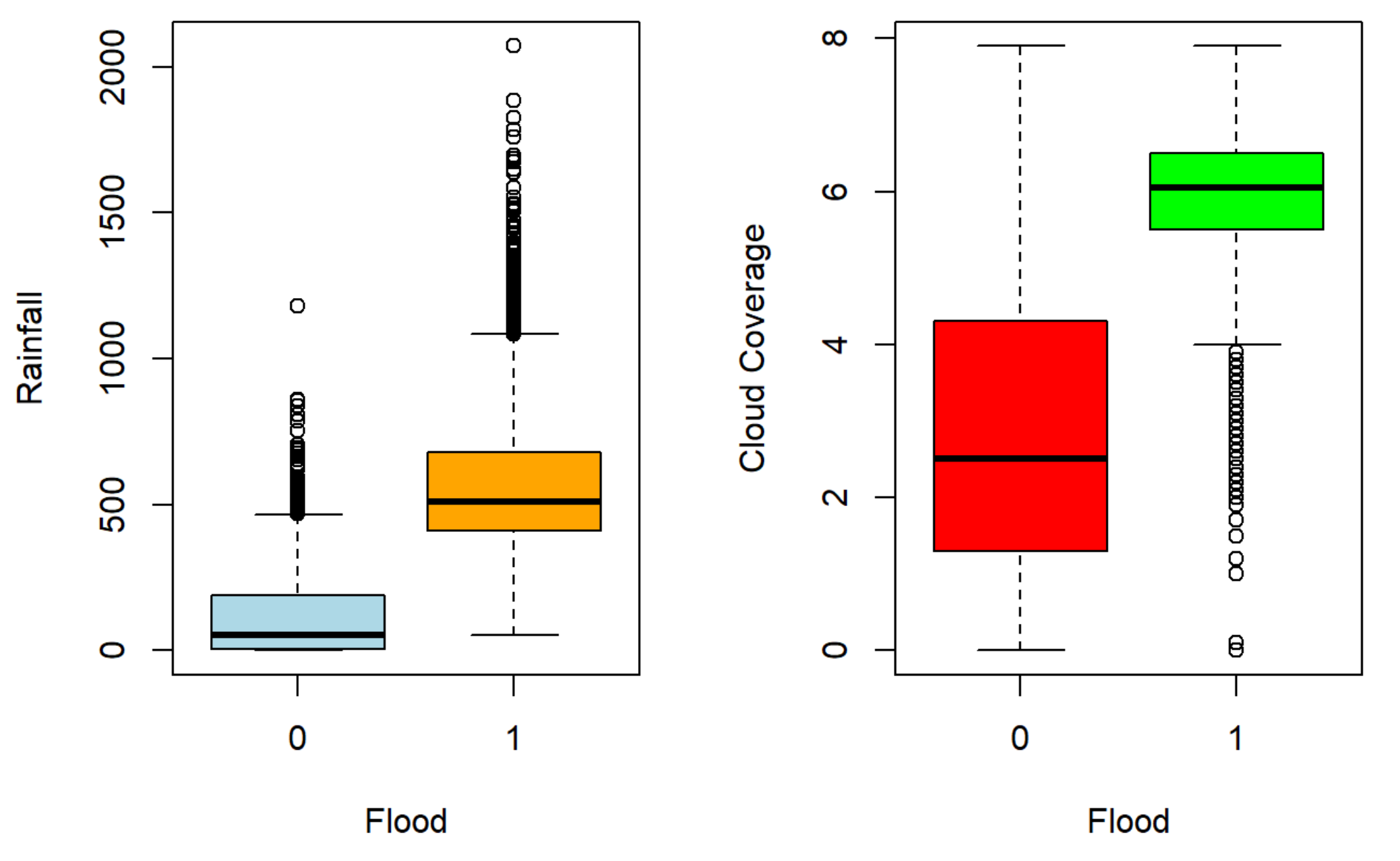


Figure 1: Boxplot Showing the distribution of Rainfall and Cloud Coverage against Flood

The flood column was presented using a scatter plot

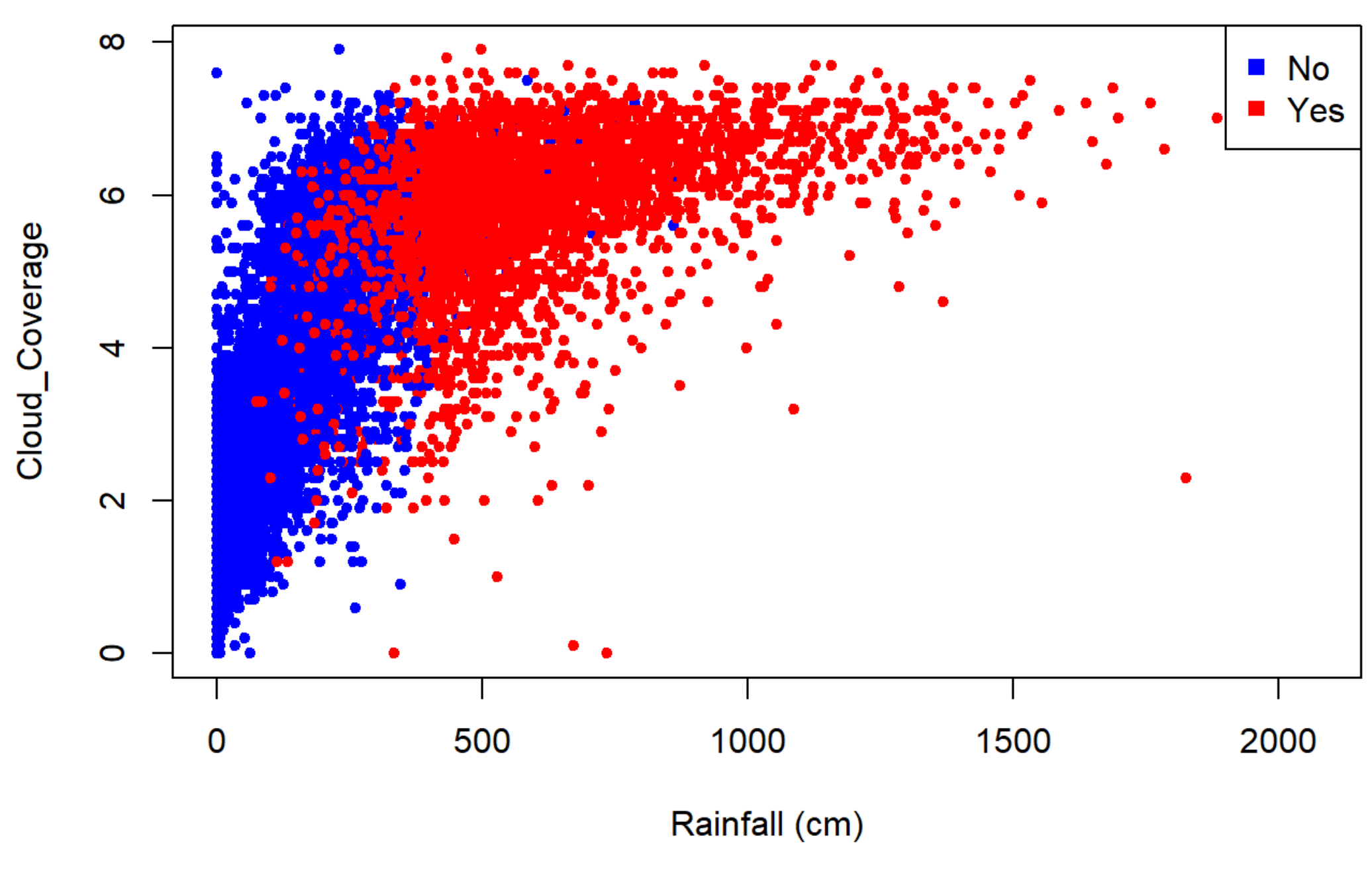
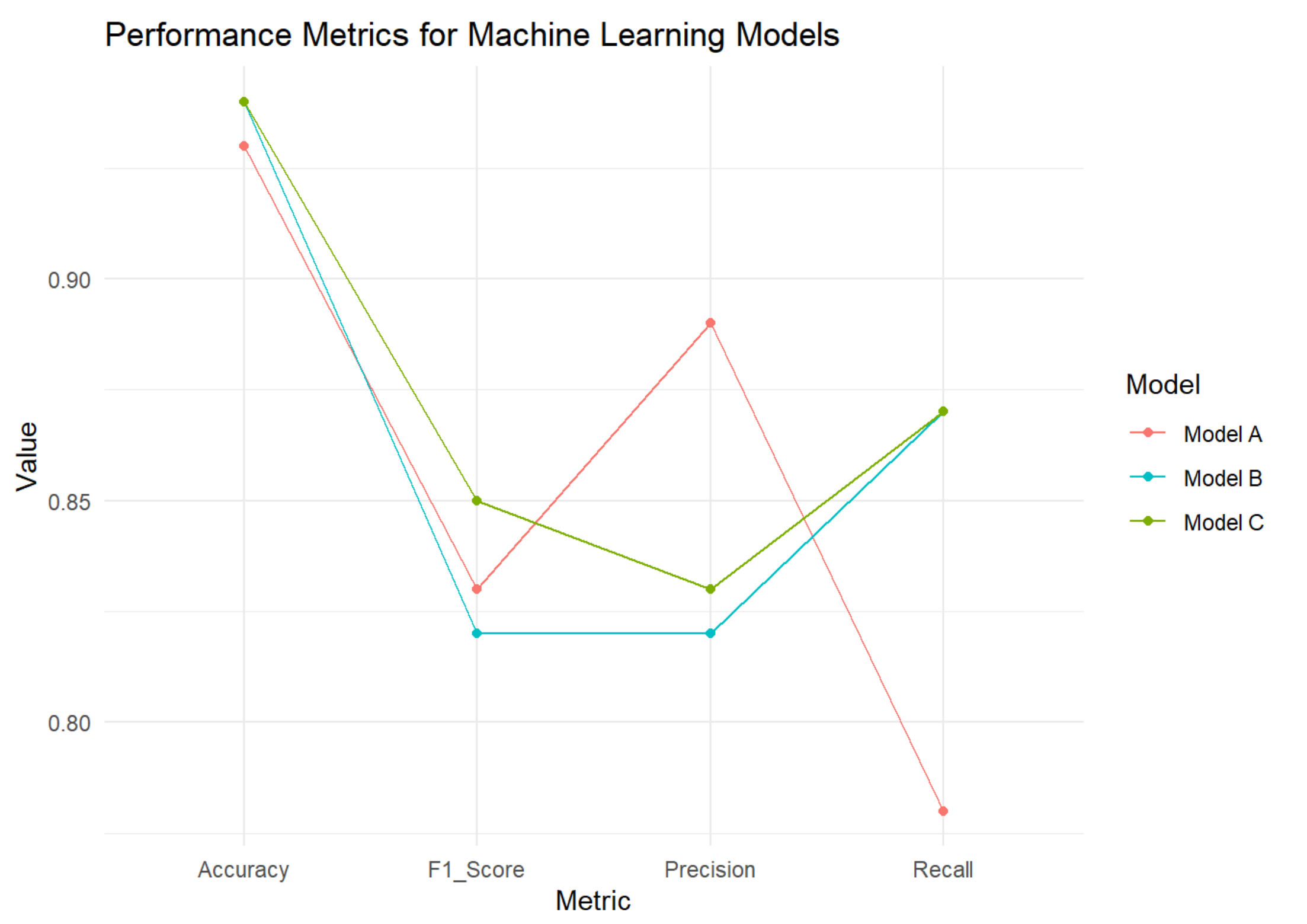


Figure 2: Scatterplot Showing the distribution of Flood by Rainfall and Cloud Coverage

Machine Learning Models: Binary Logistic Regression, Decision Tree Classifier and K-Nearest Neighbors were the models employed to predict flood in the training set. This model was then deployed in the test data and evaluated using the accuracy, precision, recall and f1-score metrics.



KNN Model was chosen as the most appropriate model because it had the highest accuracy score, Recall and F-measure score.The F-measure score provides a balance between the recall(minimize false negatives) and precision (minimize false positives) while accuracy checks how accurate the classification of the prediction model.

The most appropriate K was 15 with an accuracy of 94% and the classification output is displayed below:

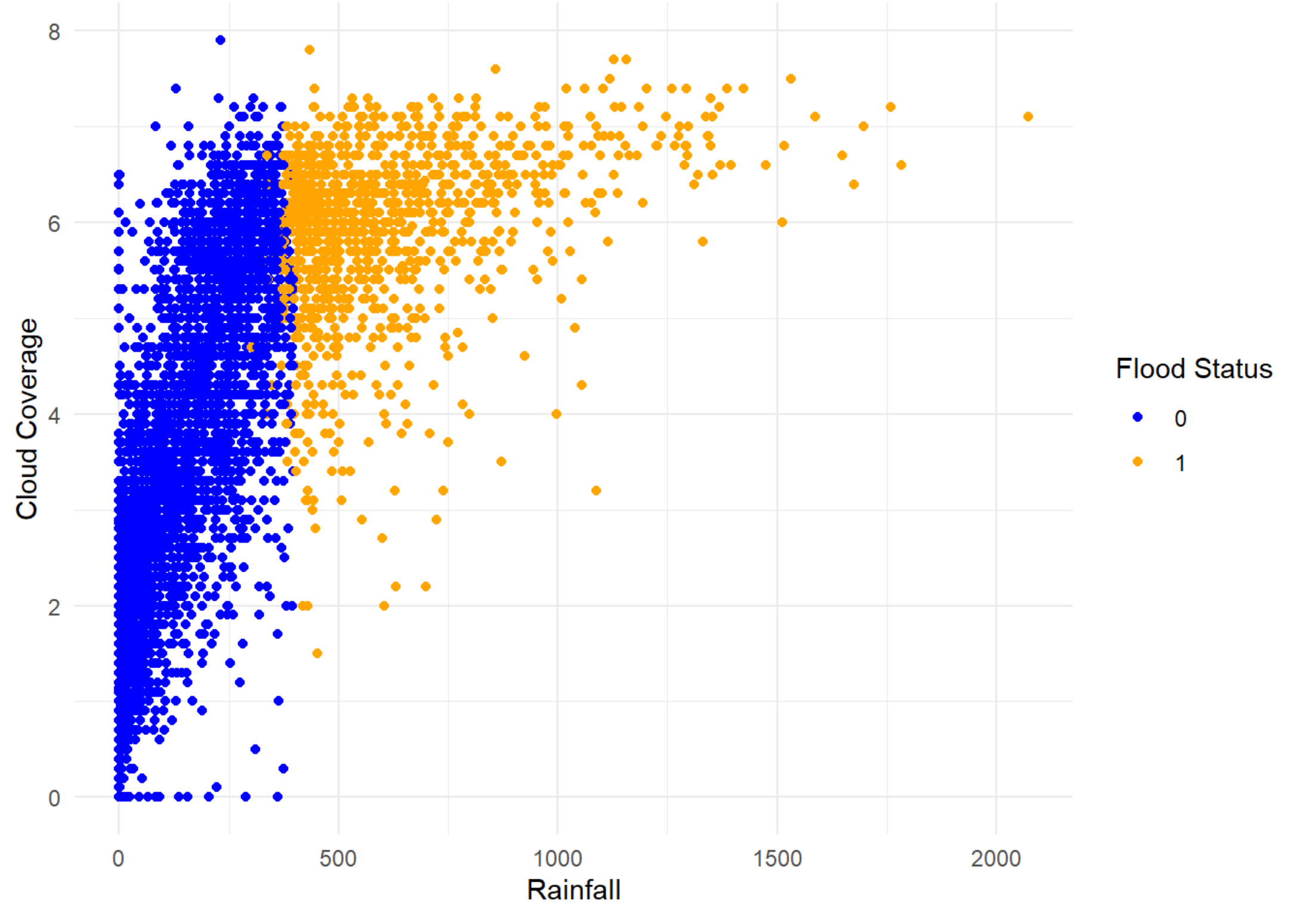


Figure 3: KNN Classification using K=15

Conclusion

- This study sought to predict floods using Decision Tree, Bayesian Logistic Regression, and K-Nearest Neighbors Models that has been used in previous studies for flood prediction. The dataset was processed based on the intended purpose of the analysis while the relevant features were determined using correlation analysis. Overall, all the models performed excellently, K-Nearest Neighbors (KNN) demonstrated the highest accuracy, recall, and F1 score, making it the most suitable model for flood prediction in this study. However, it is important to note that the dataset used in this study covered a period from 1948 to 2013, which may limit the model's ability to capture recent patterns and trends of floods in Bangladesh.

- Therefore, its important that more up-to-date data are collected on floods to improve the accuracy and reliability of the predictive models. Also, further analysis can be done in determining other relevant features with a greater effect flood prediction. Developing real time monitoring system which employs advanced procedures that provides data on remote sensing, river gauges and weather sensors would help to enhance the model's predictive power and reliability.

- Overall, this research provided information on the efficacy of machine learning models in the prediction of floods, ongoing research and data collection efforts must be done to improve the accuracy and applicability of these models in real-world flood scenarios.

References

Dewan, T H. 2015. "Societal Impacts and Vulnerability to Floods in Bangladesh and Nepal." *Weather and Climate Extremes* 7: 36–42.
Felix, A Y, and T Sasipraba. 2019. "Flood Detection Using Gradient Boost Machine Learning Approach." *IEEE Xplore*, 42–47.
Grover, K. 2020. "Advantages and Disadvantages of Logistic Regression." *Open Genus IQ: Computing Expertise & Legacy.* Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression.
Lee, J, and B Kim. 2021. "Scenario-Based Real-Time Flood Prediction with Logistic Regression." *Multidisciplinary Digital Publishing Institute*. https://www.mdpi.com/2073-4441/13/9/1191.
Parag, G, G Aditya, C Hitesh, G Gita, M Asima, H Basavaraj, and S I Yashwant. 2021. "Flood Forecasting Using Machine Learning: A Review." Conference on Smart Computing; Communications (ICSCC).
Reza, R B. 2020. "65 Years of Weather Data Bangladesh Pre-Processed." *Kaggle*. https://www.kaggle.com/emonreza/65-years-of-weather data-bangladesh-preprocessed.
Sankaranarayanan, S, M Prabhakar, S Satish, P Jain, A Ramprasad, and A Krishnan. 2019. "Flood Prediction Based on Weather Parameters Using Deep Learning". Journal of Water and Climate Change." *Journal of Water and Climate Change* 11 (4): 1766–83. https://doi.org/https://doi.org/10.2166/wcc.2019.321.
Tehrany, Mahyat Shafapour, Biswajeet Pradhan, and Mustafa Neamah Jebur. 2013. "Spatial Prediction of Flood Susceptible Areas Using Rule Based Decision Tree (DT) and a Novel Ensemble Bivariate and Multivariate Statistical Models in GIS." *Journal of Hydrology* 504: 69–79. https://doi.org/https://doi.org/10.1016/j.jhydrol.2013.09.034.