Enhanced Flood Prediction Using Logistic Regression: A Comprehensive Analysis of Hydrological and Environmental Factors

CSE 303
Statistics for Data Science
Section:02

Submitted To:

Dr. Mohammad Manzurul Islam
Assistant Professor
Department of Computer Science & Engineering
East West University

Submitted By:

Tasfia Tahsin Annita
2021-3-60-031
Samura Rahman
2021-3-60-064

Date of Submission:
21/09/2024

1 CONTENTS

2 INTRODUCTION

Flood has been an ongoing crisis in Bangladesh especially in last month, august 2024. Large number of people have lost their lives during the last flood crisis. If there was a system to predict floods that might occur in future, many people could have been saved which shows us how essential flood prediction can be for mitigating the adverse effects of floods on people, infrastructure and the environment. Our main motive to work on this flood prediction dataset was to create a model that could predict the possibility of flood occurring using surrounding features of a particular area. The dataset we chose is comprehensive and contains features pertinent to flood prediction. Our primary goal is to achieve a high accuracy to predict floods so that valuable insights for early warning systems can be provided, alerting the area to be prepared for flood disaster.

3 METHODOLOGY

For our prediction of flood, we started by surfing on Kaggle for an efficient dataset on flood prediction and we succeeded on finding a proper dataset when we came across a playground taking place on Kaggle. In order to predict flood probabilities, our first motive was to prepare the dataset which represented various environmental and infrastructural factors. Our dataset is free of null values and is entirely numeric. The dataset is already split into two different files for training and testing. We dropped the irrelevant column and performed exploratory data analysis alongside a correlation matrix heatmap to understand the relationships among the features. We trained our model using logistic regression and we used a threshold to classify the outcomes of flood prediction, and the threshold was chosen based on the median of the flood probability column. We used evaluation matrices such as accuracy, precision, recall, F1-score, confusion matrix, and ROC_AUC score which gave us AUC score. High AUC indicates that our model is performing well to predict flood.

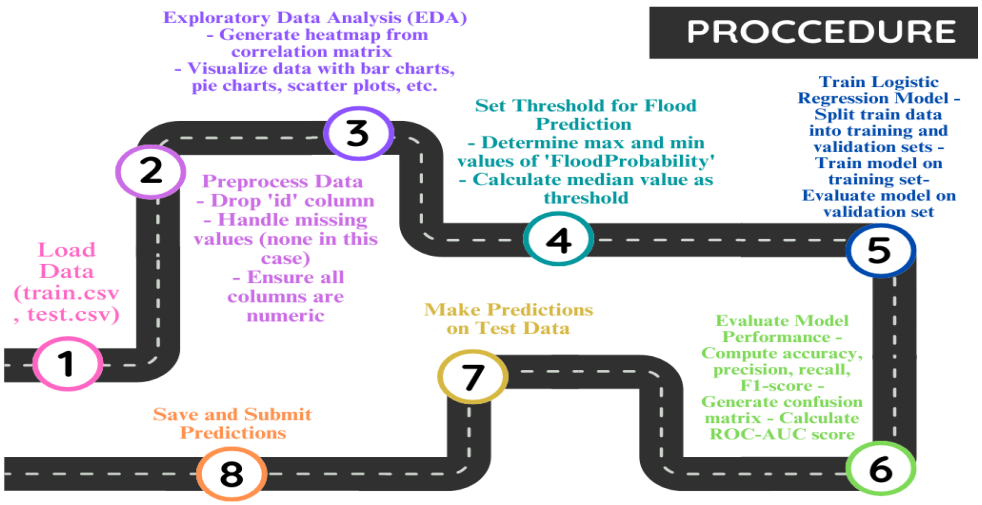

Figure0: Overview

3.1 DATA DESCRIPTION

Our dataset contains 1,117,957 rows and 22 columns. Both the train.csv and test.csv consist of 22 columns and in total they have 1,117,957 rows. The attributes present in the dataset are:

1. id: distinctive identifier for the columns.
2. MonsoonIntensity: A score representing the intensity of the monsoon (integer).
3. TopographyDrainage: A score representing the drainage quality of the topography (integer).
4. RiverManagement: A score reflecting the effectiveness of river management (integer).
5. Deforestation: A score indicating the extent of deforestation (integer).
6. Urbanization: A score measuring the level of urbanization (integer).
7. ClimateChange: A score reflecting the impact of climate change (integer).
8. DamsQuality: A score representing the quality of dams (integer).
9. Siltation: A score indicating the extent of siltation (integer).
10. AgriculturalPractices: A score reflecting the quality of agricultural practices (integer).
11. Encroachments: A score for encroachments on natural spaces (integer).
12. IneffectiveDisasterPreparedness: A score indicating the ineffectiveness of disaster preparedness measures (integer).
13. DrainageSystems: A score representing the quality of drainage systems (integer).
14. CoastalVulnerability: A score measuring the vulnerability of coastal areas (integer).
15. Landslides: A score indicating the risk or occurrence of landslides (integer).
16. Watersheds: A score reflecting the condition of watersheds (integer).
17. DeterioratingInfrastructure: A score representing the deterioration of infrastructure (integer).
18. PopulationScore: A score measuring population density or impact (integer).
19. WetlandLoss: A score reflecting the loss of wetlands (integer).
20. InadequatePlanning: A score indicating the inadequacy of planning measures (integer).
21. PoliticalFactors: A score reflecting political factors affecting the situation (integer).
22. FloodProbability: A continuous variable representing the probability of flooding (float).

We have a total of 3 csv data files.

1. To train the flood prediction machine learning model Train.csv.

2. To Test our model's performance, we used test.csv.

3. There is an unseen data of which we are supposed to predict the label named submission.csv

As we know, our dataset was solely numeric, which means no mapping was required as there were no categorical values, no mapping is needed. Also, our dataset set does not contain any null values and thus there was no need for replacing values. Our dataset (both train and test) was generated from a deep learning model trained on the Flood Prediction Factors dataset. Feature distributions are close to, but not the same, as the original. We want to predict the probability of flood occurring using logistic regression.

3.2 DATA PREPROCESSING

At the very beginning, we checked is there was any null value present in our data. We found that there was no null value present so we checked all the columns' data type. From the data type we found all the columns were either integer type or float type. Then we formed the correlation matrix and saw that the id column had zero correlation will all other features and label. And thus, we dropped the 'id' column. There was no need for splitting our data as our data was divided in two parts from the very beginning, train.csv and test.csv.

3.3 EXPLORATORY DATA ANALYSIS (EDA)

EDA is conducted to gain information regarding the relationships between the attributes. Our key visualization and finding from the project include:

3.3.1 Correlation Matrix:

A correlation matrix is a table of correlation coefficients for a set of variables used for a set of variables used to determine if a relationship exists between the variables. The coefficient indicates both the strength of the relationship as well as the direction.

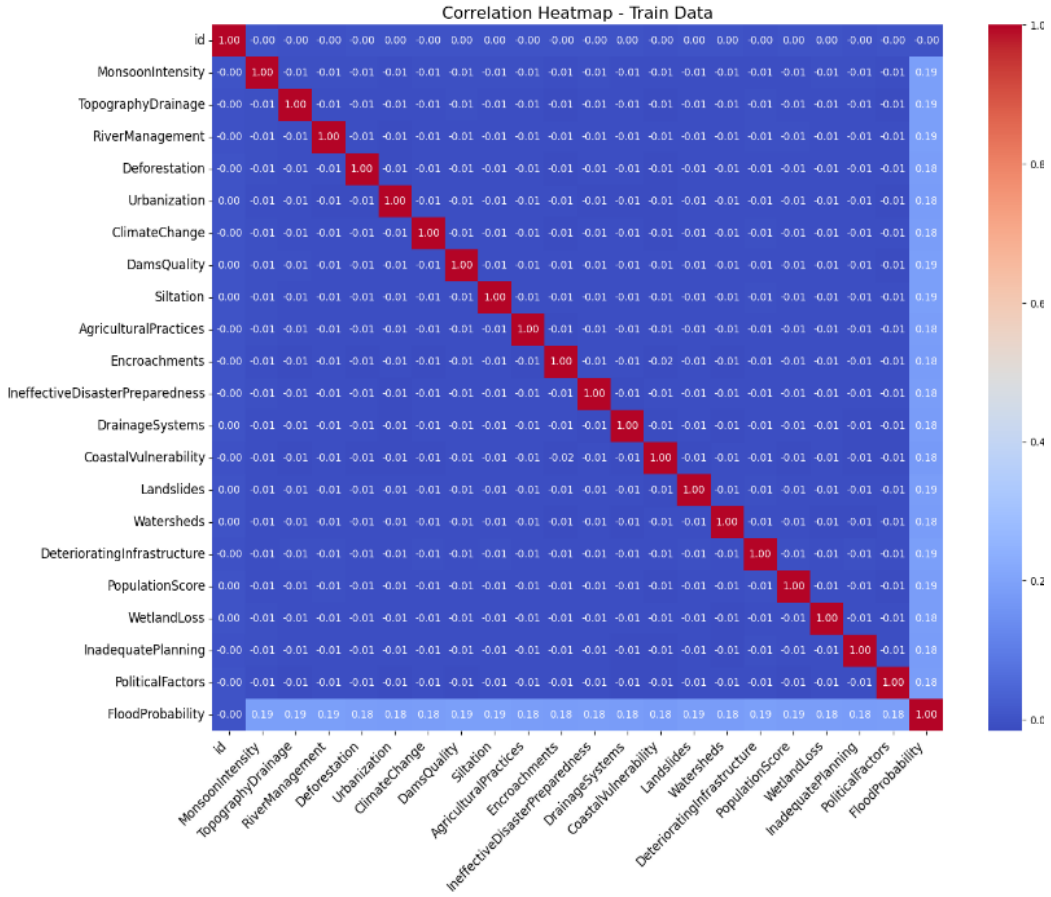

Figure1: Correlation Matrix

Figure1 is a heatmap of correlation that shows the relationship between the variables in the dataset. This heatmap shows that the diagonal elements give a perfect correlation of 1.00 as each variable is correlating with itself. The other elements show a very slight negative correlation of -0.01 which shows no significant liner relationship between all the variables.

In the heatmap, red indicates a high positive correlation and blue indicated low or negative correlation. All the features have a slight positive correlation between 0.18 to 0.19 when correlating with flood prediction, showing that all the features have a positive relation with flood prediction. The lack of strong correlation implies that the variables give unique information that are useful got feature selection and avoids multicollinearity in models.

3.3.2 Histogram:

The histograms present in figure2 shows the histogram of all the columns present in our dataset against their frequencies. Each histogram shows the distribution of data points for specific variable. Id has a uniform distribution with values between 0 and 1,000,000 and is dropped later on. All the other features except flood probability have skewed distributions and mostly are right-skewed with higher frequency at lower values. Only the flood probability which is our targeted column has a relatively normal distribution with values centered around 0.5 to 0.54.

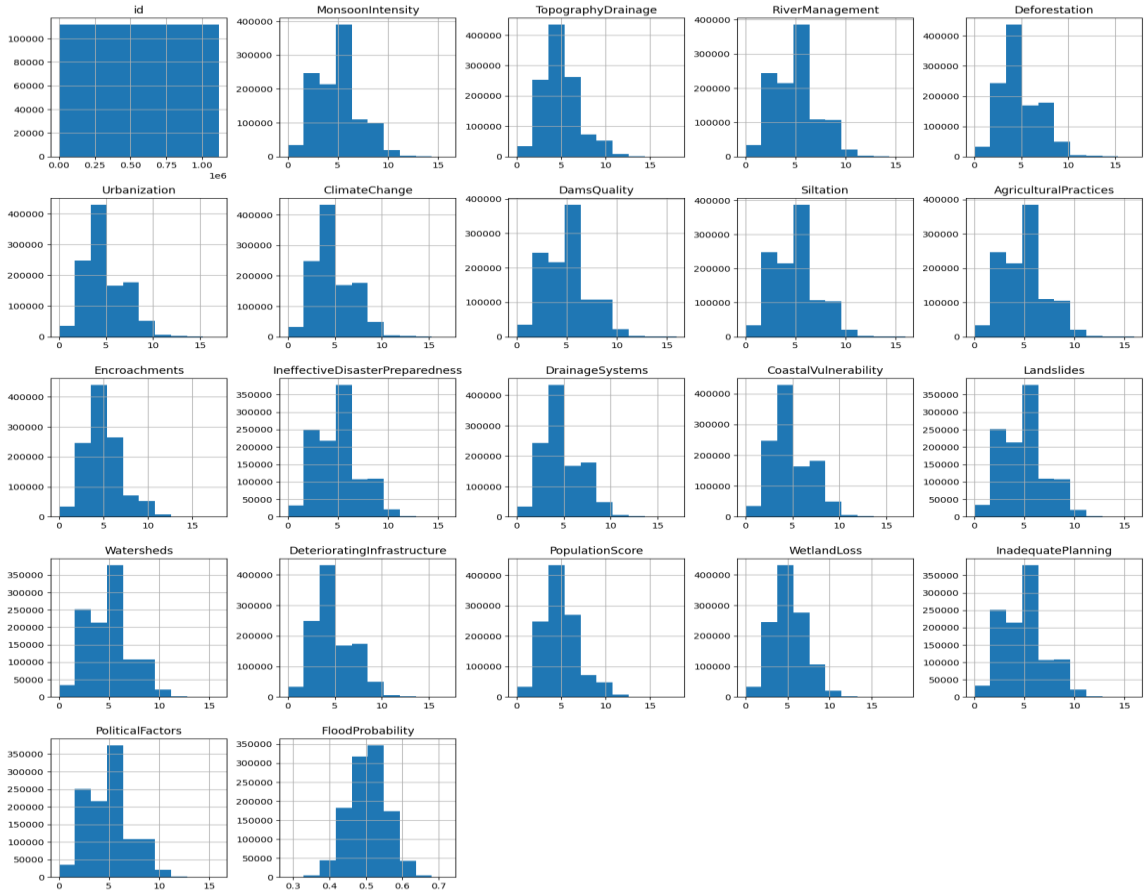

Figure2: Histogram of all the features and their frequency

3.3.3 Bar Plots:

Figure3 has two bar plots, one for flood probability and other one for political factor. Here we can see that the bar plot for flood probability takes the shape of a bell curve while the bar plot for political factors takes a right-skewed shape.

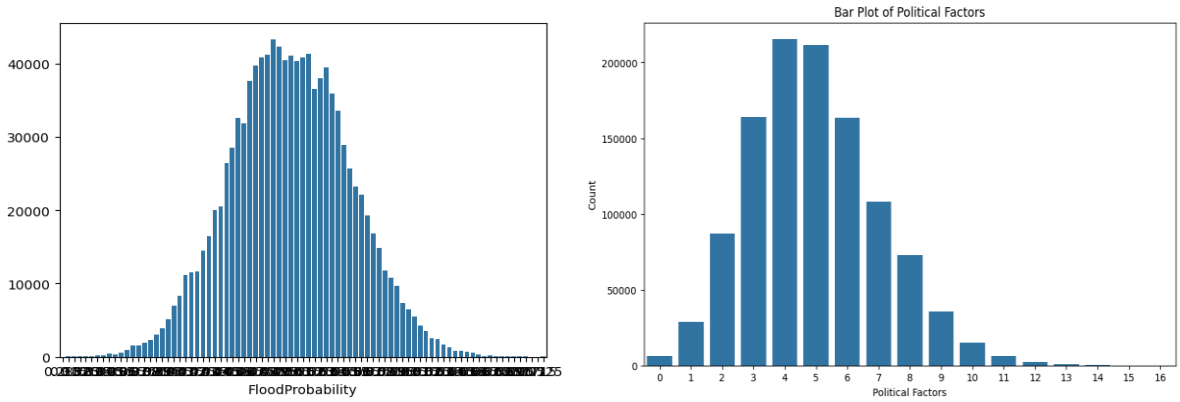

Figure3: Bar plot of flood probability and political factors

3.3.4 Box Plots:

- Figure4 shows a box plot that displays the distribution of Flood Probability in the dataset. The plot shows key statistics like the median, quartiles, and potential outliers for flood probability values. Using the box plot we can quickly understand the central tendency, variability, and any extreme values (outliers) in the flood probability data, which can help us to assess its overall distribution and spread.

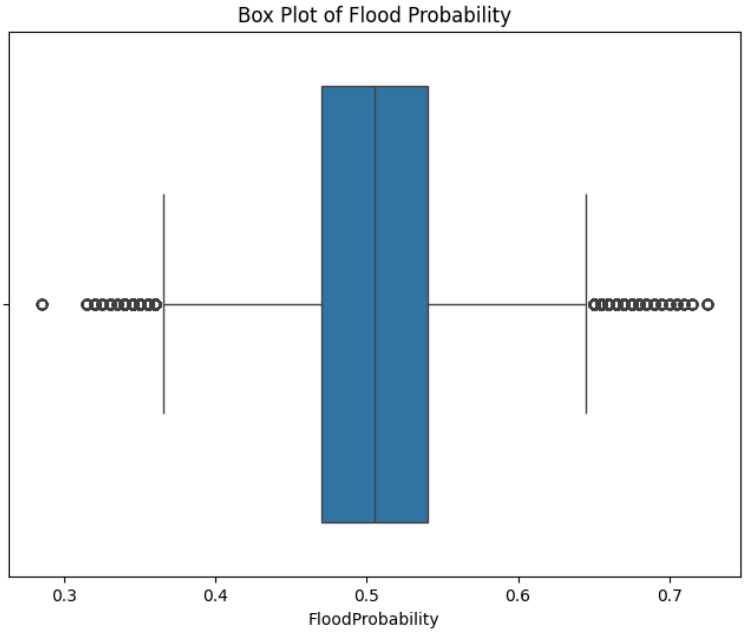

Figure4: Boxplot of Flood Probability

From the boxplot of flood probability, we can see the targeted column has a large number of outliers and so median will be suitable for threshold.

- Figure5 shows a box plot diagram of the relationship between Deforestation (x-axis) and Climate Change (y-axis). The box plot provides a visual summary of how Climate Change values are distributed for different levels or categories of Deforestation. It can also help identify trends, variability, and potential outliers in the data.

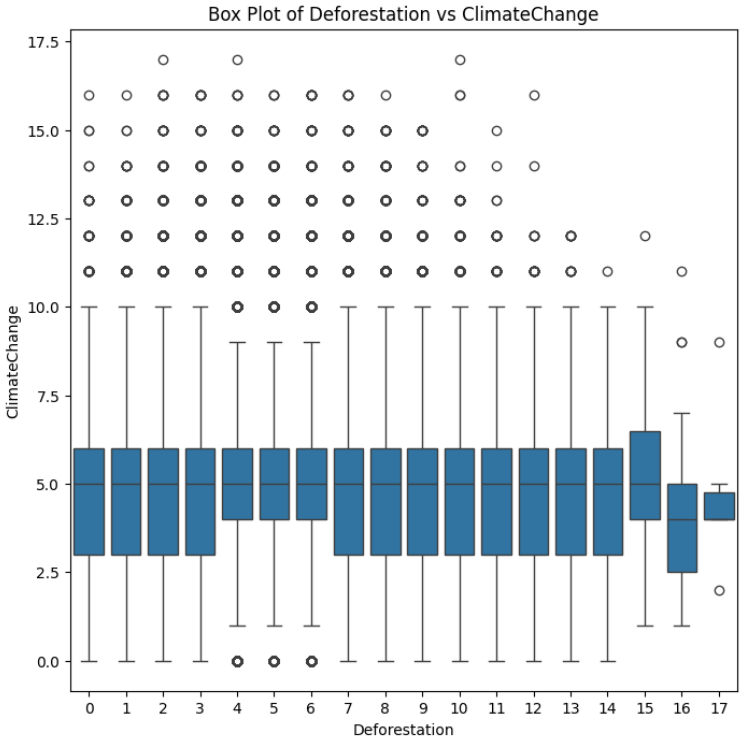

Figure5: Boxplot of deforestation vs climate change

It offers a clear way to see if higher deforestation is associated with changes in climate, making patterns and distributions easy to interpret. Since our attributes have same value of correlation with one another, their comparison with one another will result in a similar pattern. Also, from the comparison of climate change and deforestation we found a lot of outliers.

3.3.5 Scatterplots:

Figure6 has two scatter plots which explore the relationship between Flood Probability and two attributes: Ineffective Disaster Preparedness and Inadequate Planning. The first plot shows how Ineffective Disaster Preparedness relates to Flood Probability. The second plot displays the relationship between Inadequate Planning and Flood Probability, helping us see if inadequate planning is linked to increased flood risk. These visualizations highlight how weaknesses in preparedness and planning might affect flood probabilities. But we can see, despite the attribute being different in both scatterplots, the shape is the same for both. This is due to the similar value of correlation. And we can assume all the scatterplot might look the same against flood probability.

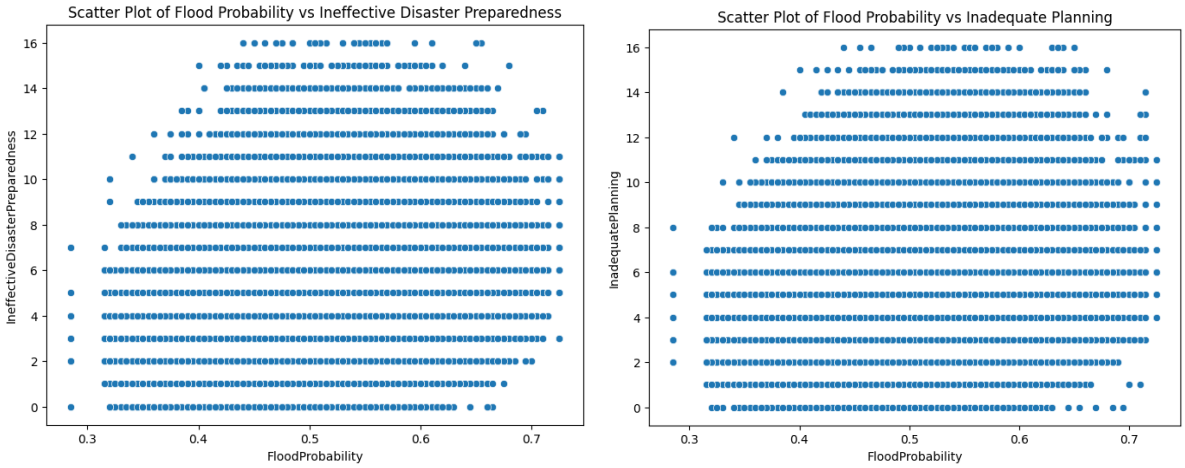

Figure6: Scatterplots of Ineffective disaster preparedness against flood probability and Inadequate Planning against flood probability

3.3.6 Pie Chart:

Figure7 shows a pie chart of the distribution of different Flood Probability values in our training dataset. Each slice represents the proportion of instances for each flood probability and are labeled with percentages. Using the chart, we can quickly understand how flood risk is distributed across the dataset, such as the flood probability of 0.49 is the most common but the presence of 0.485, 0.505. 0.52, 0.515 and 0.48 are present in equal amounts.

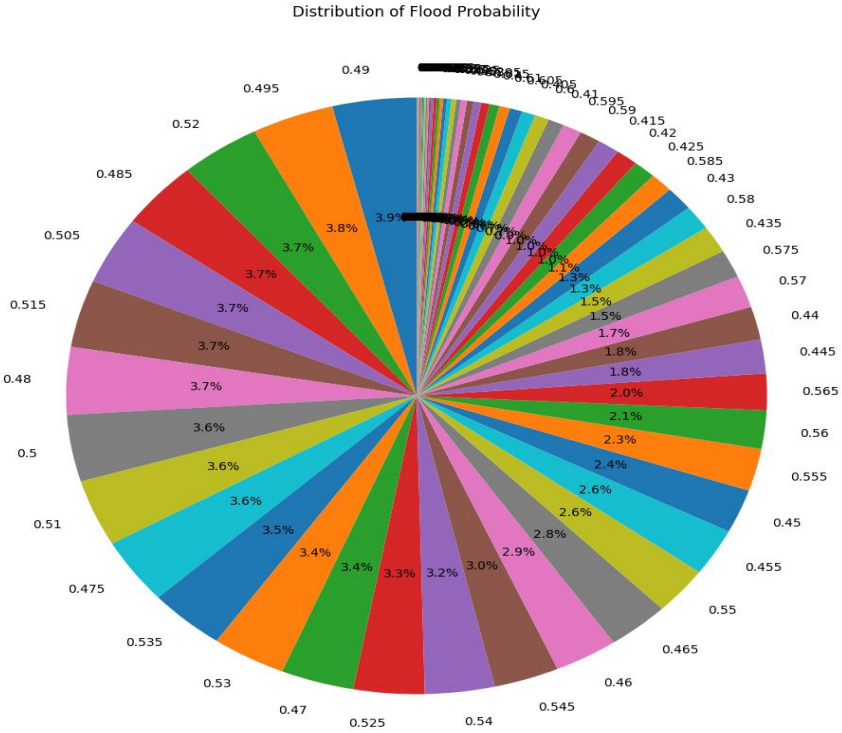

Figure7:Pie chart for flood probability

3.3.7 Count plot:

Figure8 displays a bar chart (count plot) which tells us how Flood Probability is distributed across different Watersheds in our training dataset. Here x-axis represents the Watersheds, and the bars are split by Flood Probability values (shown by different colors). The height of each bar shows how many instances of each flood probability occur in each watershed. Using this chart, we can easily compare the flood risk across different watersheds and see which ones are more prone to flooding. Here watersheds are greater in number around 0.48 and 0.56 flood probabilities and so we can say during flood, more watersheds are readily available.

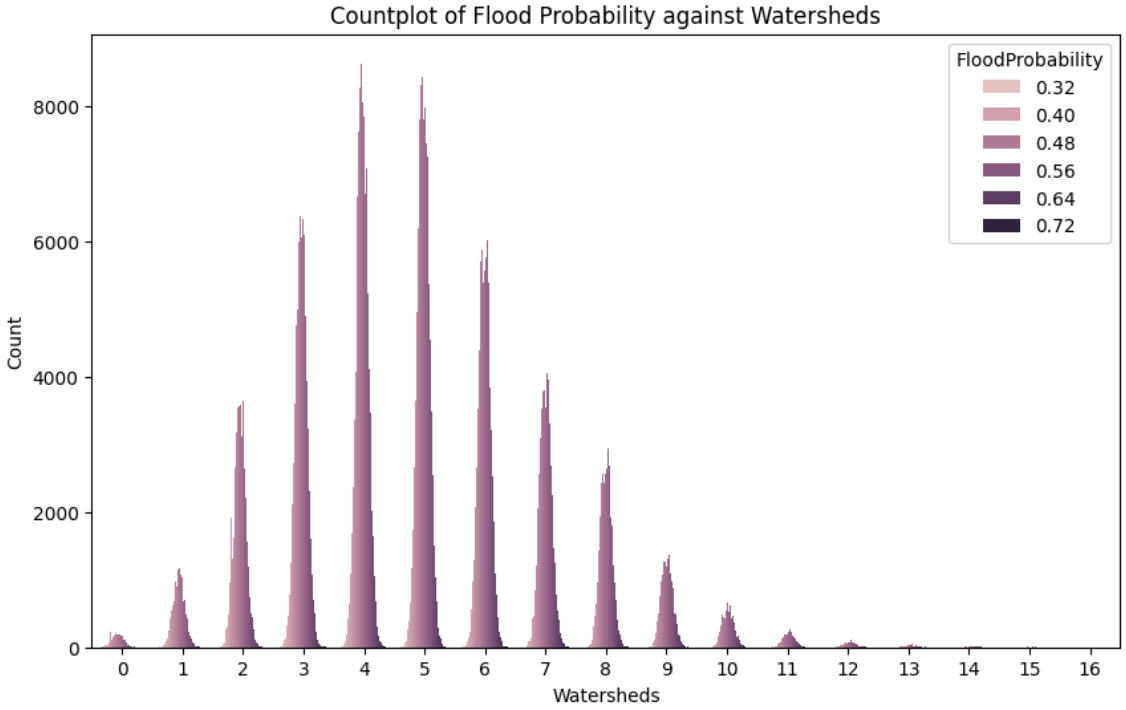

Figure8:Countplot of flood probability against watersheds

3.3.8 3D Scatter plot:

Figure9 displays a 3D representation of the first 3 columns of our training dataset which shows how Monsoon Intensity, Topography Drainage, and River Management influence Flood Probability. The x, y, z axes show the first 3 features while the color of each point indicates the flood risk. For the color of the points, we used "inferno" color map. From this plot we can identify how different combinations of monsoon intensity, drainage and river management affect the flood probabilities faster which makes it easier to spot patterns and make decisions about flood predictions.

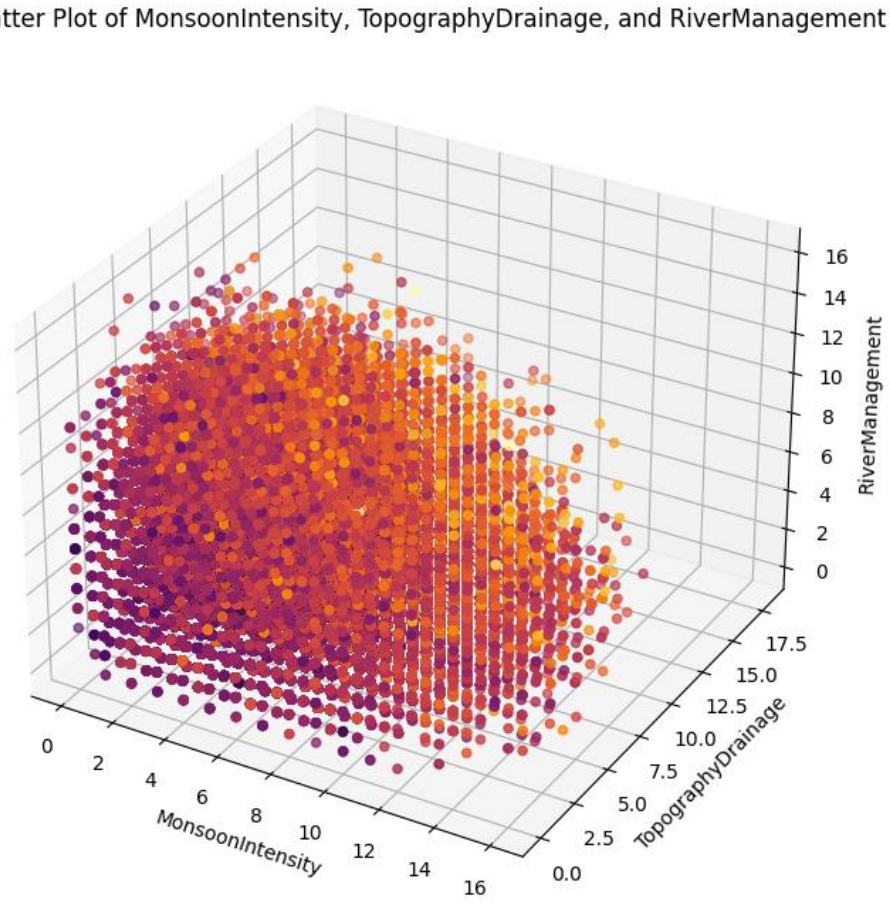

Figure9: 3D Scatter Plot of first 3 Columns

3.4 MACHINE LEARNING MODELS

We used logistic regression to predict the chances of flood occurring in a particular area. Logistic regression, a supervised machine learning algorithm, is used for binary classification. Logistic Regression finds relation between data factors by using mathematics and then it uses these relationships to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes like yes or no. We chose this model because flood prediction is a binary classification problem where it will predict if flood occurs or not. Table1 shows the values of accuracy, precision, recall, f1-score and ROC-AUC score we found using our model.

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
88.74%	92.42%	84.84%	88.47%	92.47%

Table1: Model Evaluation

3.4.1 Confusion Matrix

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

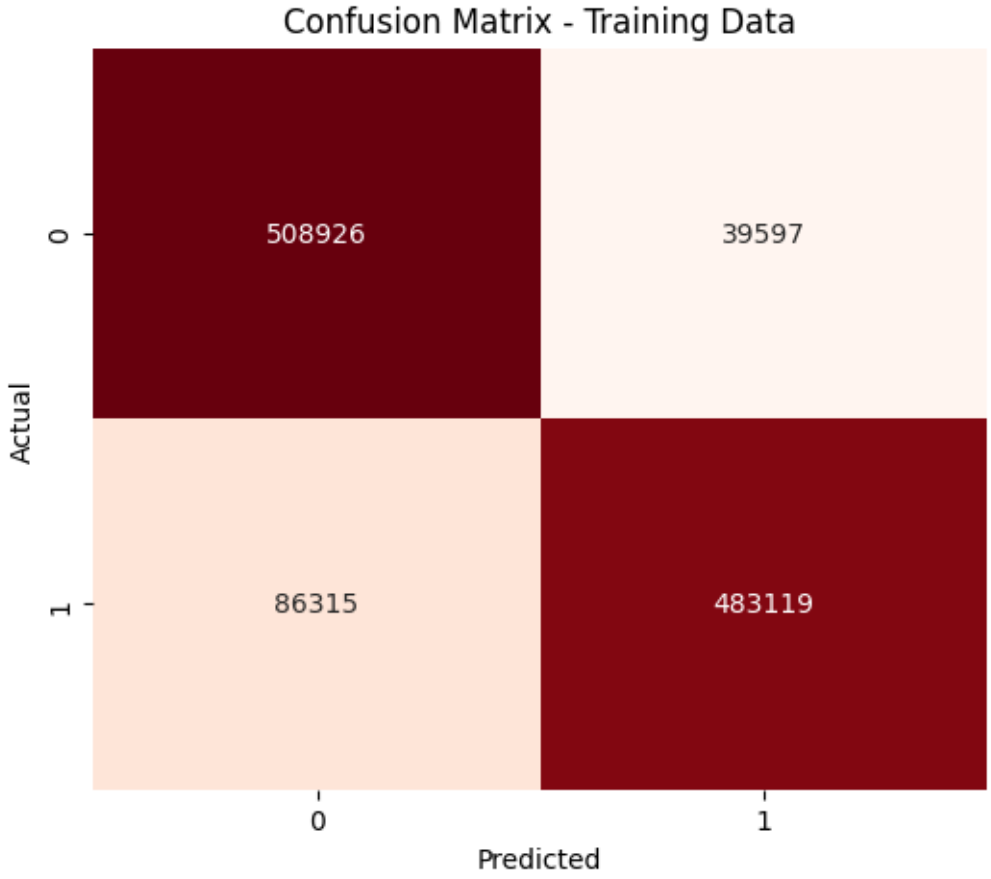

Figure10: Confusion Matrix

Here is how we can interpret the matrix.

The rows here show the true labels (0 or 1) and the columns shows the predicted labels (0 or 1). Here in the figure10 top left tile shows the true negatives, correctly predicted 0s which is 508926. The top right shows the false positives, predicted 1 but actually 0, which is 39,597. The bottom left shows the false negatives, predicted 0 but actually 1, which is 86,315. The bottom right shows the true positives, correctly predicted 1, which is 483,119.

3.4.2 ROC-AUC Curve:

The ROC (Receiver Operating Characteristic) curve visualizes the trade-off between sensitivity and specificity, giving insight into the model's ability to classify flood vs non-flood cases. The AUC (Area Under the Curve) score of our model is 0.92 which shows that our model has a high ability to detect flood and non-flood instances with a 92% chance of correctly classifying them. From the figure11 the dotted straight line labeled as “Random Guess” on the curve is a diagonal line from (0,0) to (1,1) of the plot which shows a model with no biasness in guessing the class labels. Since our ROC curve is further from the diagonal line and closer to the top left corner, we can say that our model is strong at predictive performance.

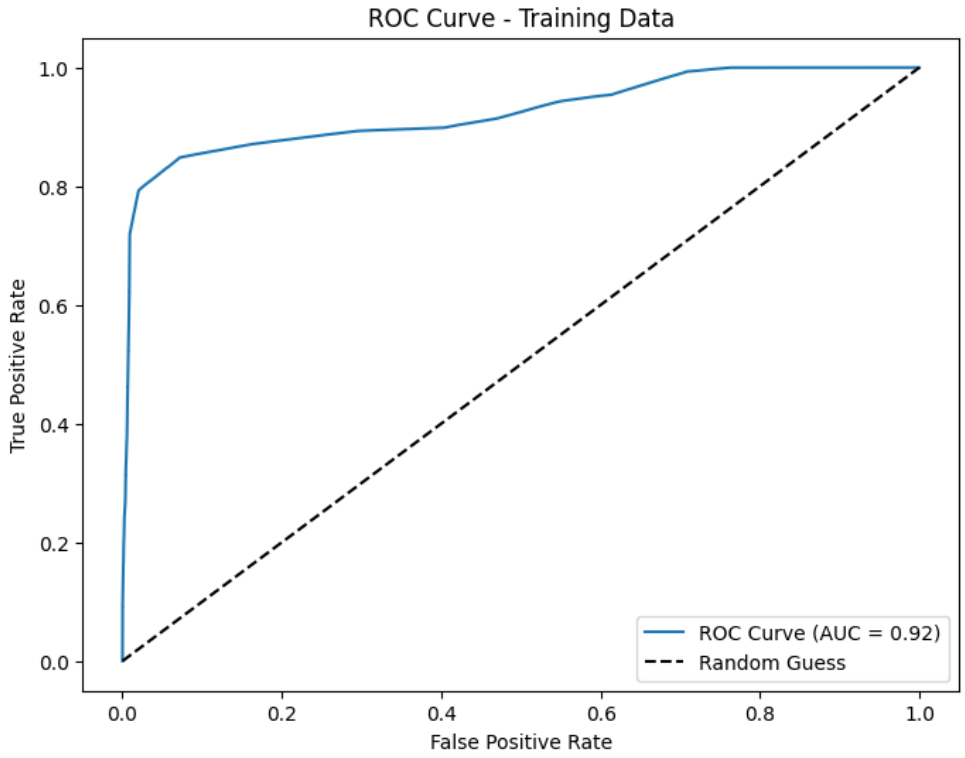

Figure11: ROC-AUC Curve

3.4.3 Additional Models

We tried to use linear regression and SVM models using this dataset but couldn't succeed in doing so. This is because linear Regression predicts a continuous numerical value while we need a binary outcome of flood or non-flood. SVM is an effective model but creates complex decision boundaries and are not easily understandable like logistic regression. Thus, we decided to use logistic regression to keep things simple and easy to comprehend.

4 CONCLUSION

In this project, we have developed a flood prediction model with the help of logistic regression in order to determine if flood can occur or not based on several environmental and socio-economic factors. We had no null values of duplicated values, thus making our data preprocessing easier and received a high training accuracy of 88.74% alongside an AUC score of 0.92 which shows that our model is strong at predictive performance. Our model's effectiveness in classifying flood and non-flood was confirmed by the ROC curve. Our project brings out the utility of logistic regression in flood prediction and provides valuable insights to the policymakers and authorities of disaster management to mitigate flood risks. Further future work can improve the predictive accuracy by implementing advanced modeling techniques.