IS 517 Final Project Report

Group 5 – Flood Prediction Using Environmental and Infrastructural Factors

1. Introduction

Floods are among the most devastating natural disasters, causing environmental degradation, property loss, and human displacement globally. The increasing unpredictability of climate change and urbanization necessitates robust data-driven flood prediction systems.

Our project aims to predict flood probability using environmental and infrastructural variables and to classify regions into flood risk categories. This work can help local governments and businesses allocate resources effectively for disaster mitigation and planning.

2. Research Question

Our primary research objectives include:

Identifying key environmental and infrastructural contributors to flood probability.

3. Dataset and Preprocessing

We used the publicly available Flood Prediction Dataset from Kaggle. The dataset contains 50,000 rows and 21 columns, including 20 predictor variables encompassing both environmental and infrastructural factors. Each row represents a specific region and timeframe, with variables such as Monsoon Intensity, Urbanization, Drainage Systems, and more. The target variable, FloodProbability, is a continuous value ranging from 0 to 1, representing the likelihood of a flood occurring. We carried out comprehensive data preprocessing steps as outlined below:

3.1 Handling Missing Values

To ensure data completeness without introducing bias, we addressed missing values by imputing them with the median of the respective variables. By using median imputation, we preserved the central tendency of each feature while maintaining the overall data structure, allowing for accurate and consistent downstream analysis.

3.2 Outlier Treatment

We used Winsorization to cap extreme values based on the IQR method, reducing the influence of outliers without removing any data points. This helped stabilize predictions, especially for models sensitive to extreme values. As shown in *Figure 1*, the Monsoon Intensity variable's distribution became more uniform after capping. The same approach was applied to other numeric features such as FloodProbability, Urbanization, and Siltation.

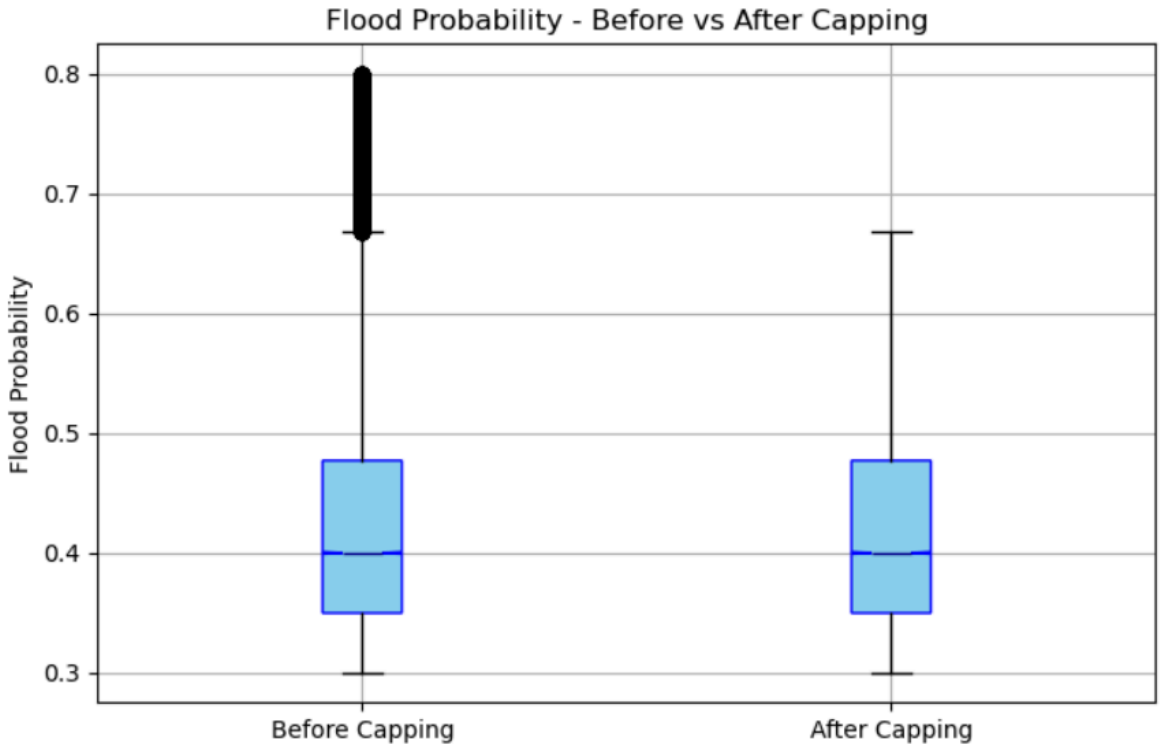

Figure 1. Boxplot of Monsoon Intensity Before and After Capping

3.3 Skewness Correction

Siltation and FloodProbability showed right-skewed distributions, which can affect linear model performance. To address this, we applied log transformations, reducing skewness and stabilizing variance. This improved model interpretability and ensured more balanced feature contribution. Figures 2 and 3 show the distributions before and after transformation.

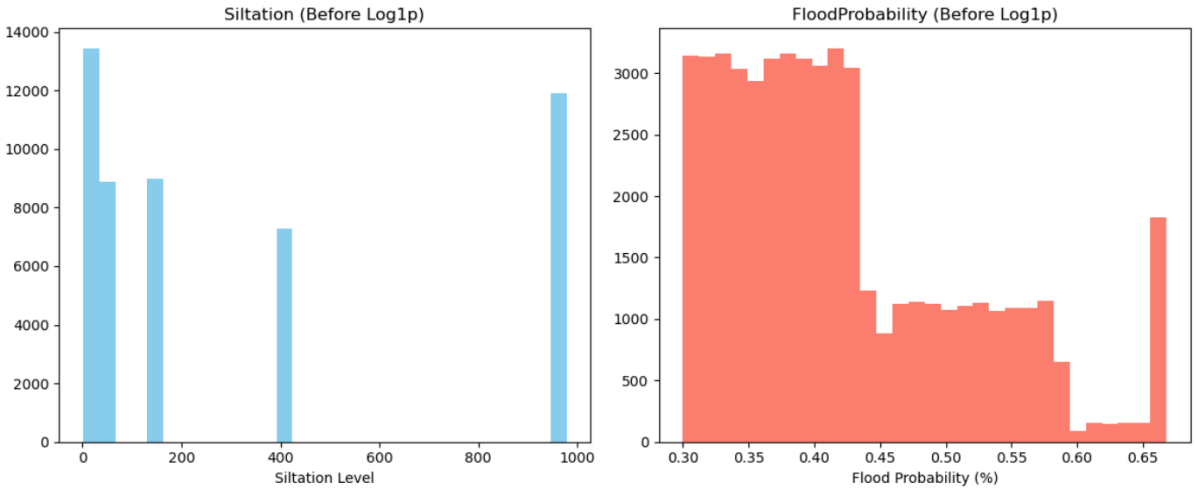

Figure 2. Distribution of Siltation and FloodProbability Before Log Transformation

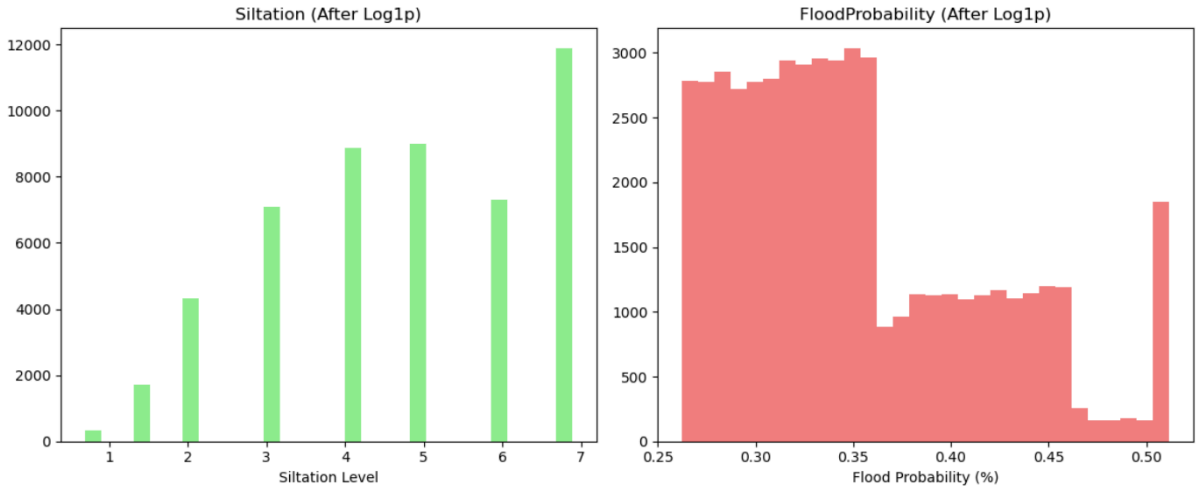

Figure 3. Distribution of Siltation and FloodProbability AfterLog Transformation

3.4 Multicollinearity Check

We conducted a multicollinearity analysis using the Variance Inflation Factor (VIF) to detect redundant or highly correlated predictors. All features returned VIF values below the common threshold of 5, indicating acceptable levels of multicollinearity. For example, Urbanization had a moderate VIF of 3.31, while PopulationScore and Deforestation had lower values of 2.6 and 1.7, respectively.

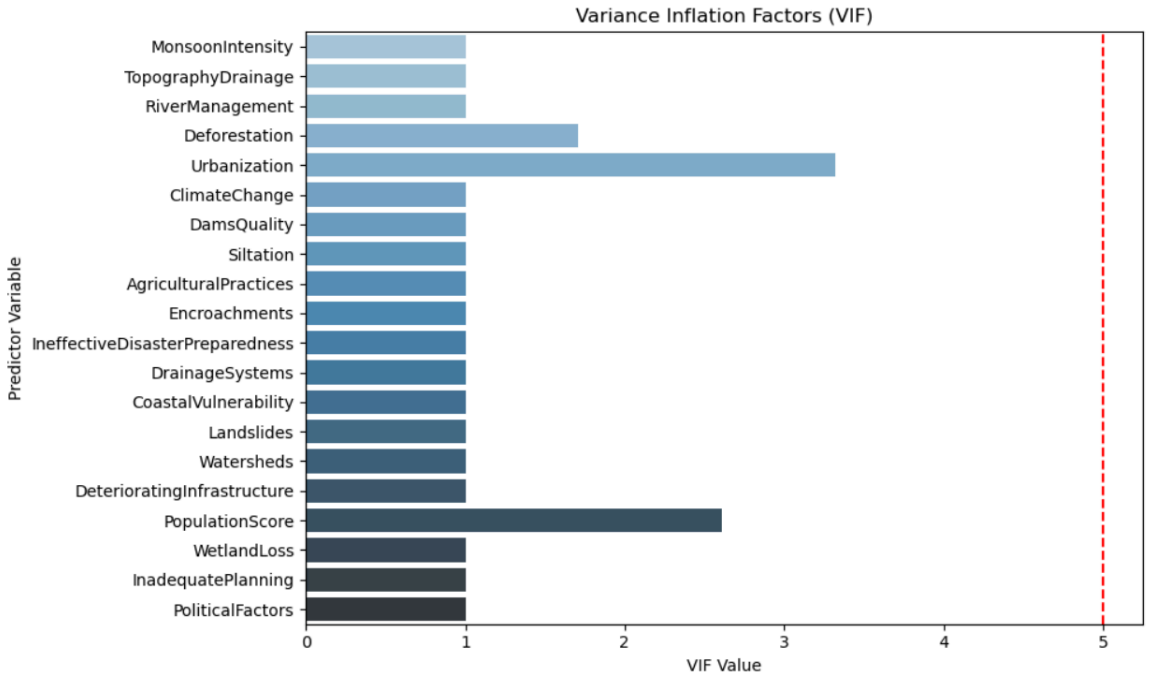

Figure 4. Variance Inflation Factor (VIF) Scores for Predictor Variables

3.5 Feature Engineering

We performed feature engineering by creating composite indicators from existing variables. Specifically, we derived new features such as Urban_Drainage, Dams_Encroachments, and

Climate_Political. These engineered variables captured complex relationships and interactions among original predictors, providing more nuanced representations of infrastructural integrity and environmental stressors.

4. Modelling

Research Question 1: What are the most significant factors contributing to flood probability?

To address this question, we applied a multi-step modeling approach using linear regression, ridge regression, and random forest. Each method offered different perspectives, allowing us to identify the most influential environmental and infrastructural predictors of flood probability.

Step 1: Linear Regression Analysis

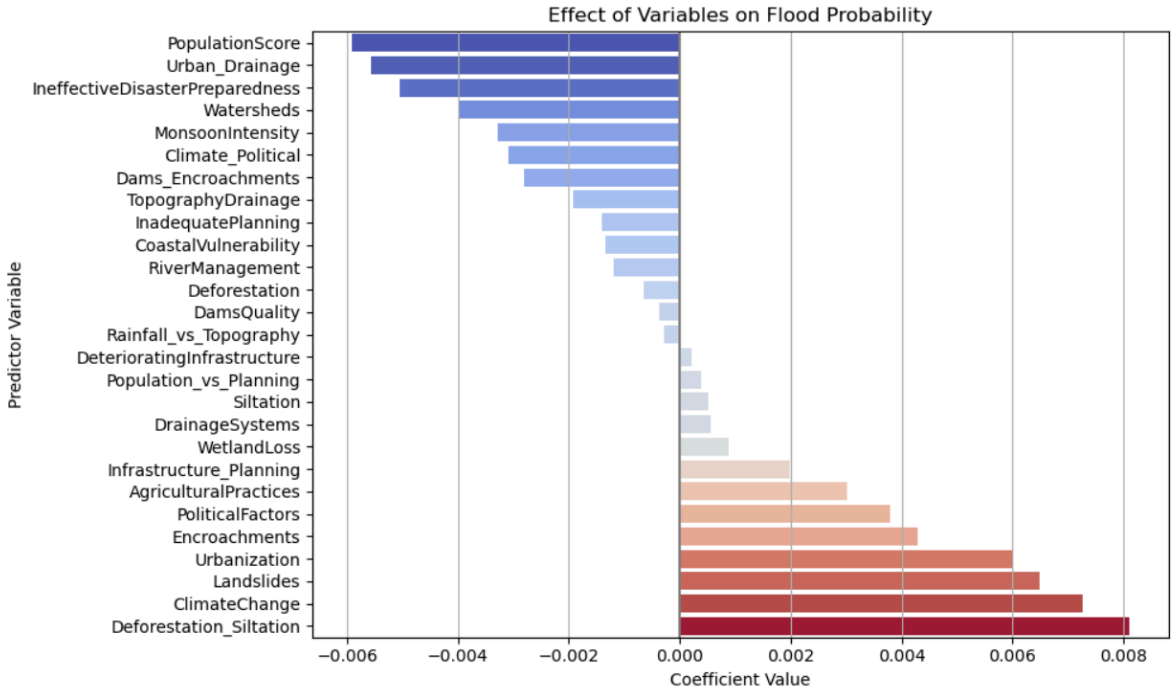

Figure 5. Effects of variables on Flood Probability using MLR

We first used Multiple Linear Regression to examine linear relationships between predictors and flood probability. ClimateChange and Landslides showed strong positive effects, while some variables had negative coefficients due to multicollinearity and shared variance. These should not be seen as less important but rather as adjustments made by the model to balance overlapping influences. The accompanying bar chart Figure 5 ranks variable influence within the linear framework.

Step 2: Model Assumption Diagnostics

Before accepting the linear model's findings, we conducted diagnostic checks to ensure the reliability of its statistical outputs. These included:

1. Residual vs Fitted Plot and QQ Plot: These suggested that linearity and normality assumptions were reasonably met. The absence of funnel shapes or curves indicated no strong signs of nonlinearity or heteroscedasticity. However, the dense, centered residuals may reflect the effect of log transformation compressing the target variable into a narrower range.

2. ANOVA Test: Identified ClimateChange as a marginally significant predictor with a p-value of 0.09.
3. Breusch-Pagan Test: This returned a p-value of 0.4081, indicating no significant heteroscedasticity. This supports the assumption of constant residual variance, validating the reliability of standard errors and p-values in the regression results.

The linear regression model yielded very low explanatory power, with an R-squared of 0.00022 and adjusted R-squared of -0.00017. Despite satisfying diagnostic assumptions, the model failed to explain variability in flood probability, suggesting that linear methods are inadequate for capturing its complexity. To address potential multicollinearity, we applied Ridge Regression with cross-validation. While the technique successfully shrank coefficients, most were near zero (±1e-39), offering no meaningful improvement in model fit. The R-squared remained unchanged, confirming that linear approaches are insufficient for this task.

This finding led us to pursue non-linear models, such as Random Forests, which are better suited for capturing complex interactions.

Step 3: Random Forest for Nonlinear Relationships and Feature Importance

Recognizing the limitations of linear models in capturing complex interactions, we advanced to Random Forest Regression. This ensemble-based model does not assume linearity and can naturally handle multicollinearity and non-additive effects. The resulting bar chart Figure 7, illustrated the most important predictors based on their contribution to reducing impurity in the decision trees.

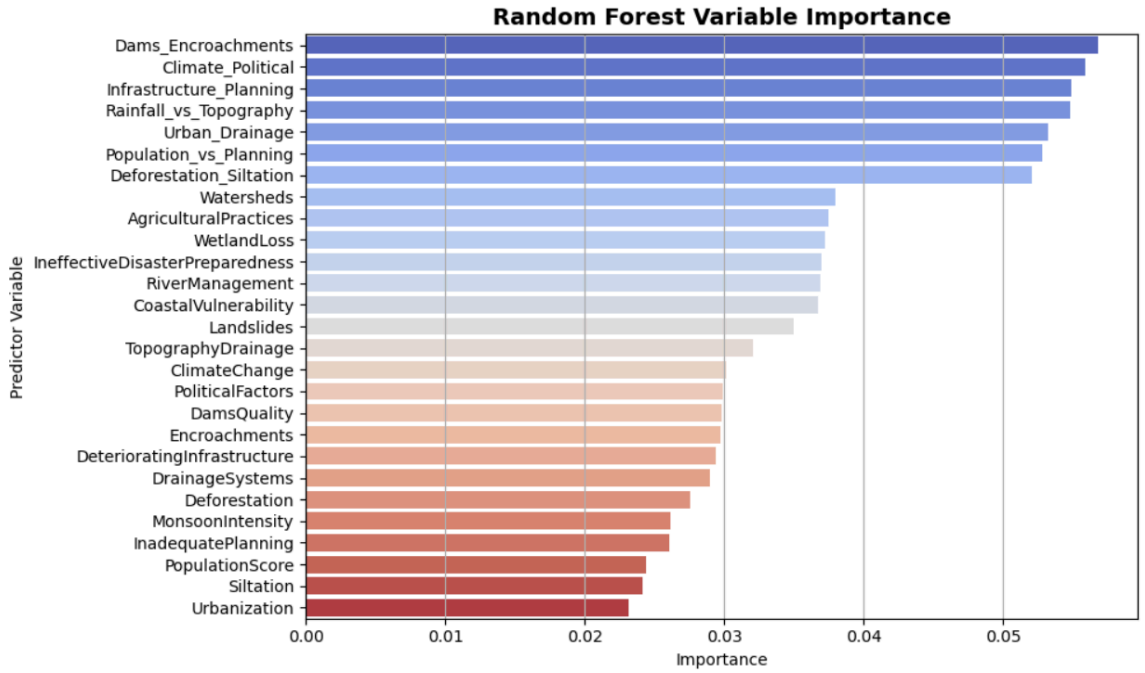

Figure 6. *Effects of variables on Flood Probability using Random Forest*

Although Random Forest is well-suited for modeling non-linear relationships and managing multicollinearity, our model returned a mean squared error of 1.039 and explained only -3.93% of the variance in flood probability. While overall predictive performance was limited, the feature importance scores still offered valuable exploratory insights. Population Score, Urbanization, and Deforestation consistently ranked as the most influential variables, reinforcing their relevance to flood

risk. The model's poor fit may stem from the log transformation compressing target variance, weak predictor signals, or inherent noise in the data. These results highlight the need for more granular spatial or temporal features and suggest that classification approaches, as used in RQ2, may be more effective than direct regression.

5. Results and Interpretation

This study shows how predictive modeling supports flood risk assessment and mitigation. We identified key contributors, urbanization, population score, and deforestation, using linear and ensemble models.

6. Literature Survey

Compared to existing Kaggle notebooks on flood prediction, our project adopts a broader and more practical approach. While many models rely solely on meteorological data like rainfall and temperature, we integrate both environmental and infrastructural factors such as urbanization, drainage systems, and deforestation. In contrast to models that overlook multicollinearity or class imbalance, we applied advanced techniques Multiple and Ridge Regression, Random Forest for regression, and Logistic Regression, LDA, and XGBoost for classification alongside SMOTE, skewness correction, and VIF analysis. Most importantly, our model offers actionable insights for policy and planning, moving beyond academic modeling toward real-world impact.

7. Limitations and Future Work

- Data Limitations: The dataset lacked spatial and temporal granularity. The inclusion of GIS layers or seasonal data could improve model performance.
- Model Improvements: Future work could include hyperparameter tuning, ensemble stacking, and deep learning methods.
- Broader Scope: Integrating real-time weather data and elevation maps would provide dynamic and localized predictions.

Citation: Phrasing in this report was assisted by OpenAI's ChatGPT to enhance clarity and readability. All analysis, interpretations, and content reflect the authors' original work.