Flood Prediction using Machine Learning

Ahmed Tahlil Kadir[1], Masiat Mohammad Momshad[2], Riyadus Salehin Fahmid[3], and Sabik Swanan[4]
Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
{[1]ahmed.tahlil.kadir, [2]masiat.mohammad.momshad, [3]riyadus.salehin.fahmid, [4]sabik.swanan}@g.bracu.ac.bd

***Abstract*—Floods are increasingly destructive due to climate change, necessitating advanced prediction methods beyond traditional statistical models. This study employs machine learning to predict flood probability using the flood prediction dataset from kaggle which had 50,000 instances and 21 features. Four models—SVR, Decision Tree, Random Forest, and KNN—were evaluated using MAE, RMSE, and R^2. KNN performed best (MAE: 0.0186, RMSE: 0.0238, R^2: 0.7776), with Random Forest close behind, while Decision Tree lagged. Results demonstrate machine learning's effectiveness in modeling flood risks, offering a scalable solution for climate-vulnerable regions.**

***Index Terms*—Flood, Probability, Machine Learning, Prediction, Regression**

I. Introduction

Floods are among the most destructive natural disasters, with their frequency and severity rising due to climate change. Traditional prediction models, based on historical data and statistical methods, often fall short in capturing the complex, non-linear interactions of multiple flood drivers. These models struggle to adapt to evolving environmental conditions and human impacts. In contrast, machine learning offers a powerful alternative, utilizing large datasets from sources like weather stations and satellite imagery to uncover intricate patterns and enhance prediction accuracy. Its adaptive nature makes it especially suited for dynamic climate scenarios. We made use of a few mentionable machine learning models like SVR, KNN Regressor, Decision Tree Regressor and Random Forest Regressor to analyse the contributing factors and forecast the flood probabilities.

II. Dataset Description

The dataset used in this study comprises 50,000 instances with 21 features, designed to model and predict flood probability based on a variety of environmental, infrastructural, and socio-political factors. Each feature is a critical contributor to flood risk and is represented as an integer score reflecting its intensity or impact. The independent variables include factors such as Monsoon Intensity, River Management, Deforestation, Urbanization, Drainage Systems, Climate Change, and others, totaling 20 predictors. The target variable, FloodProbability, is a continuous value ranging from 0 to 1, indicating the likelihood of a flood occurring given the input conditions. All data is structured numerically, with 20 integer features and 1 floating-point output, making the dataset well-suited for regression modeling. This synthetic yet comprehensive dataset enables robust analysis of complex interactions among multiple flood-contributing factors.

III. Exploratory Data Analysis

A. Data Anomalies

- Null Values: No null values were found in the dataset.
- Duplicate Rows: No duplicate rows were found, all rows were unique.

B. Feature Correlation Analysis

A correlation analysis was performed to determine which features had the strongest relationship with the target feature "FloodProbability".

- Heatmap to show all pairwise correlations.

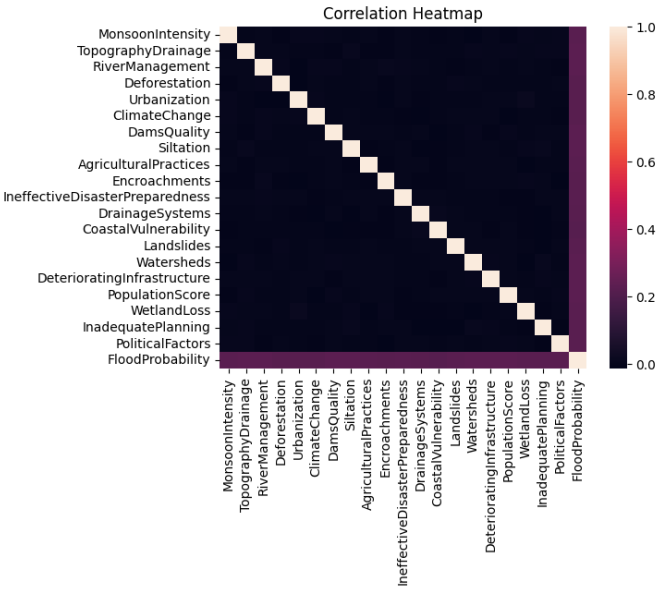

Fig. 1. Correlation Heatmap

- Bar plot showing correlation values of each feature with the target feature.

IV. Dataset Preprocessing

A. Feature Scaling

Standard Scaling has been applied to normalize the features as models like SVM and KNN require feature scaling for effective prediction as they do work on distance based algorithms. First, the scaler was fitted on the training data to compute the mean and standard deviation. Then, the same transformation is applied to both the training and testing sets using, ensuring consistency and preventing data leakage from the test set.

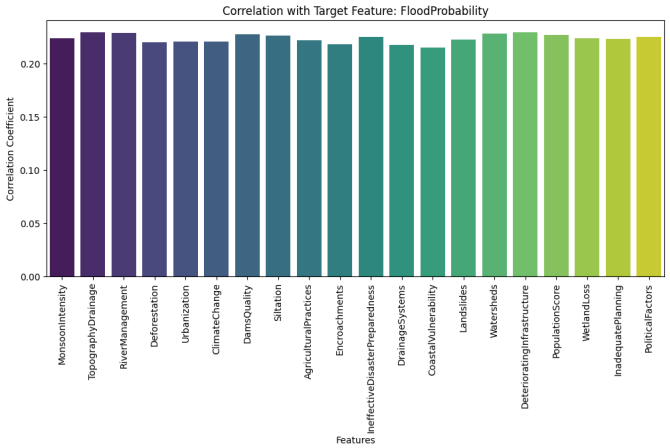

Fig. 2. Correlation Barplot

B. Data Splitting

The dataset was divided into training and testing sets with a test size of 0.2.

- Training Data: 80% of the full dataset
- Testing Data: 20% of the full dataset

V. Model Analysis

A. Model Training and Testing

In this section, we outline the models selected for our project, along with the rationale behind their selection and performance evaluation.
The following machine learning regression models were trained:

- Support Vector Regression (SVR):
  - Used for high-dimensional feature spaces
  - Sensitive to data scale and kernel choice
- Decision Tree Regressor:
  - Captures non-linear relationships
  - Prone to overfitting unless pruned
- Random Forest Regressor:
  - Ensemble method combining multiple trees
  - Reduces variance and improves generalization
- K-Nearest Neighbors Regressor (KNN):
  - Predicts based on closest feature similarity
  - Sensitive to feature scaling and choice of k

Each model was trained on the training set and tested on the test set. Evaluation metrics used include:

- Mean Absolute Error (MAE): Measures average magnitude of errors
- Root Mean Squared Error (RMSE): Penalizes large errors
- R^2 Score: Measures proportion of variance explained by the model

B. Model Selection Process

The models were chosen to balance interpretability, computational efficiency, and predictive power. Each algorithm was evaluated based on its theoretical suitability for the dataset and its ability to handle various feature characteristics, including numerical and continuous variables.

C. Model Performance

After comparing model performances, it was found that:

- KNN consistently outperformed other models, showing the lowest error values and highest R^2, indicating robust predictions and strong generalization.
- Random Forest gave results close to KNN but could not outperform it in terms of MAE, RMSE, and R^2 Score.
- SVR showed decent results with an R^2 Score of 0.7090 and errors similar to as in KNN and Random Forest.
- Decision Tree didn't seem to do much as it could not provide any optimal results.

The predicted flood probabilities from each model were plotted against the actual values. These plots revealed how closely each model's predictions matched the ground truth and highlighted any overfitting or underfitting.

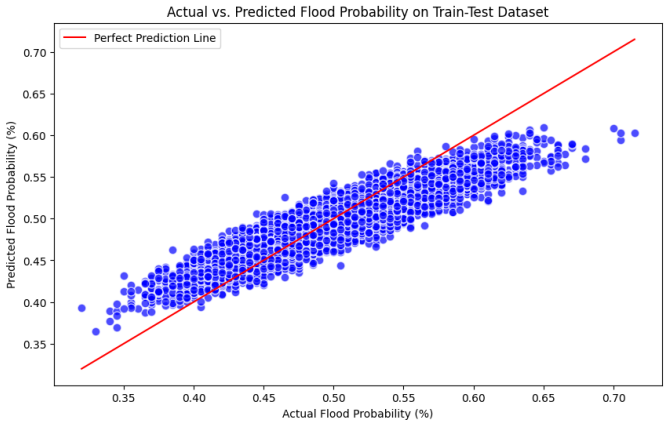

Fig. 3. Actual Probability vs Predicted Probability Correlation

D. Comparison Analysis

A comparison analysis table for the results of the selected models:

TABLE I
Evaluation Metrics Table

Model	MAE	RMSE	R^2 Score
SVR	0.0207	0.0272	0.7090
Decision Tree	0.0372	0.0469	0.1346
Random Forest	0.0205	0.0261	0.7315
KNN	0.0186	0.0238	0.7776

VI. Conclusion

This project successfully demonstrates the use of supervised machine learning models to predict flood probability based on environmental and geographic features. Random Forest emerged as the most effective model with high accuracy and low prediction error. EDA and Preprocessing, such as

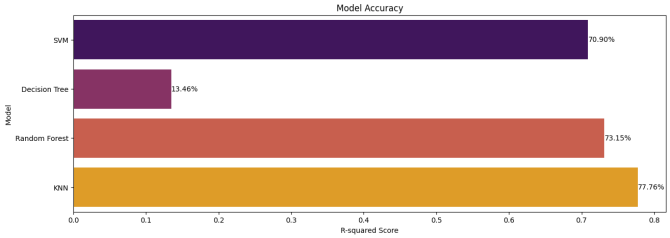

Fig. 4. Model Accuracy (R^2 Score)

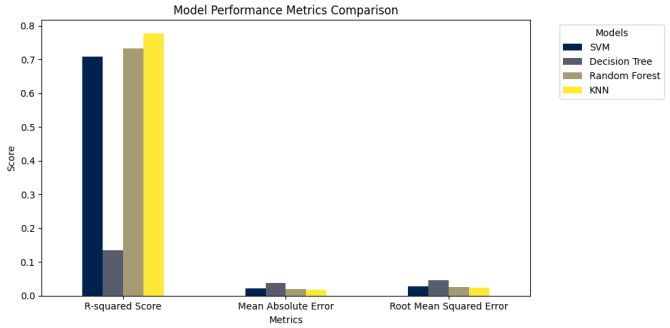

Fig. 5. Evaluation Metrics Comparison

correlation analysis and feature scaling respectively, helped ensure high data quality and effective prediction. Future improvements could include hyperparameter tuning (e.g., GridSearchCV), cross-validation, and feature engineering.