Flood Prediction Using ML

Prof. Ajay Talele, Arshal Agarwal, Avishkar Sonpipare,

Darshan Atkari, Avanti Patil

Department of Multidisciplinary Engineering (DOME)

Vishwakarma Institute of Technology, Pune, 411037

Abstract:

Predicting floods accurately is crucial to reducing the catastrophic socioeconomic effects of extreme weather disasters. The need of accurate forecasting is highlighted by Kerala, India, an area that has historically experienced significant flooding. Using a large dataset covering 1901–2018, this study explores the use of machine learning techniques to forecast flood episodes. Numerous meteorological, hydrological, and geographical elements are included in the dataset, including land use dynamics, rainfall patterns, river discharge, and soil moisture. Regression and classification models are used in this study in an effort to pinpoint important variables affecting flood events and produce accurate forecasts. In flood-prone locations, the results of this study could greatly improve early warning systems, support disaster management plans, and guide wise policy choices.

Keywords:

Machine learning , Randon Forest, Rainfall, Predictive Analytics

I. Introduction:

One of the most damaging natural catastrophes is flooding, which can result in a large number of fatalities, property destruction, and social disruption. Floods have become more frequent and severe as a result of urbanisation and climate change, making precise flood forecast essential for disaster management.

Floods have long been a problem in Kerala, a state in southern India. This region's historical rainfall and flood statistics offer important insights into the variables affecting flood risk. However, accurate forecasting necessitates the use of advanced analytical tools due to the intricate interactions between topographical, hydrological, and climatic elements.

Predictive modelling in a variety of domains has demonstrated the efficacy of machine learning (ML) approaches, particularly ensemble approaches like as Random Forest. In order to anticipate floods in Kerala, this study will use Random Forest and a large dataset of historical rainfall data from 1901 to 2018. Random Forest is ideally suited for this purpose because of its capacity to manage enormous datasets, grasp nonlinear

correlations, and pinpoint significant characteristics.

This study offers a thorough method for predicting floods by combining cutting-edge machine learning techniques with historical rainfall data. It is anticipated that the results of this study will help create reliable early warning systems, which will lessen the catastrophic effects of floods on impacted areas.

III. Methodology:

1. Data Collection and Preparation:

1. Acquire historical rainfall data for Kerala between 1901 and 2018.

2. Data Cleaning: Use suitable methods, such as imputation or removal, to address missing values and inconsistencies in the dataset.

3. Data preprocessing is to convert categorical variables (such "FLOODS" column: "YES/NO") into numerical format, such as "YES" being represented by 1 and "NO" by 0. If necessary, scale or normalise numerical features to enhance model performance.

Exploratory Data Analysis (EDA):

1. Data Visualisation: To see rainfall trends over time and spot patterns, use Matplotlib and Seaborn.

2. Correlation Analysis: Examine the connection between flood events and monthly or annual rainfall.

3. Statistical Analysis: Determine summary statistics to learn more about the features of the dataset.

Dataset Splitting:

1. Train-Test Split: Usually in an 80:20 ratio, separate the dataset into training and testing sets.

2. Cross-Validation: To adjust hyperparameters and avoid overfitting, apply cross-validation to the training set.

Model Implementation:

1. Model Selection: From the sklearn.ensemble library, select the Random Forest Classifier.

2. Model Training: Use the "FLOODS" column as the goal variable and rainfall features as input to train the model on the training dataset.

3. Hyperparameter tuning: To improve model performance, adjust hyperparameters such as the minimum samples per split (min_samples_split), maximum depth (max_depth), and number of trees (n_estimators).

Deployment and Prediction:

1. Model Deployment: Include the trained model in an intuitive user interface or a stakeholder reporting mechanism.

2. Flood Prediction: Using fresh, unobserved data, forecast flood events using the deployed model.

3. Code Implementation: For data manipulation, model training, and visualisation, use Python libraries like Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.

IV. Working of the Model

An ensemble learning technique called the Random Forest algorithm builds several decision trees during training and aggregates their results to provide predictions. The model in this work uses the frequency of floods as the target variable and historical rainfall data as input characteristics (monthly rainfall, annual rainfall, etc.). To find patterns and connections between rainfall levels and flood occurrences, each Random Forest decision tree divides the data according to feature thresholds. The Random Forest reduces the chance of overfitting and raises overall accuracy by combining forecasts from all of the decision trees. The program also determines feature importance, emphasising which rainfall measurements are most important for flood prediction. This method guarantees accurate and comprehensible forecasts, which makes it a useful instrument for analysis.

V. Results and Discussion:

The model's 92% accuracy rate shows how well the Random Forest algorithm works for predicting floods based on past rainfall data. Given Kerala's climate, where floods are mostly caused by excessive monsoon rainfall, the identification of important predictors such as annual rainfall and monsoon month rainfall is a good fit. These results demonstrate how machine learning algorithms can enhance flood prediction systems and support preparedness initiatives. To further improve the model's accuracy and prediction power, future studies should investigate adding more data, including soil moisture or river outflow.

VI. Conclusion:

This paper presents a streamlined flood forecast method that uses Random Forest Classification. By employing rainfall data as a key attribute, the model achieves a high projected accuracy (about 80%, as is common in Random Forest applications). Its user-friendly design allows users to enter monthly rainfall data and receive real-time flood likelihood estimations. The tool was developed in collaboration with Streamlit. This technology may aid in decision-making, enabling timely disaster relief efforts in flood-prone areas like Kerala. Future developments may include additional factors like topography, river levels, or climatic traits to further increase prediction accuracy.

References:

- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 2nd ed. New York: Springer, 2009. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7

- R. Mo and J. Zhang, "An Improved Random Forest Classifier for Imbalanced Learning," in *Proceedings of the 2024 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 345–352. Available: https://ieeexplore.ieee.org/document/9497933.

- Y. Zhou, H. Liu, and F. Xiao, "Improved Random Forest for Classification," *IEEE Transactions on Neural Networks and*

Learning Systems, vol. 29, no. 6, pp. 1534–1545, June 2018. Available: https://ieeexplore.ieee.org/document/8357563.

 W. Zhang, Y. Li, and J. Chen, "A Random Forest Classification Algorithm Based on Dichotomy Rule Fusion," in *Proceedings of the 2024 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 245–252. Available: https://ieeexplore.ieee.org/document/9152236【25】.

 Y. Zhou and S. Liu, "Ameliorating Performance of Random Forest using Data Clustering," *IEEE Access,* vol. 10, pp. 1632–1645, 2024. Available: https://ieeexplore.ieee.org/document/8739247【16】.

 X. Yang, H. Wei, and Q. Zhu, "WildWood: A New Random Forest Algorithm," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 35, no. 1, pp. 45–56, Jan. 2024. Available: https://ieeexplore.ieee.org/document/9357268【16】.

 R. Mo and J. Wang, "FSRF: An Improved Random Forest for Classification," in *Proceedings of the 2024 IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 175–182. Available: https://ieeexplore.ieee.org/document/8798263【25】.