DTP Project

jeanpierre Nkundmahaoro

July 2025

TOPIC: Forecasting Local Climate Risks Using Weather Data.

1 Introduction

Climate change is causing more floods and droughts in many places around the world. These events can destroy homes, farms, roads, and even put people's lives in danger.
This project is about using old weather data to predict future floods. We want to build a model (a smart system) that looks at:

- MonsoonIntensity
- TopographyDrainage
- Deforestation
- Urbanization
- ClimateChange
- DamsQuality
- Siltation
- AgriculturalPractices
- Encroachments
- IneffectiveDisasterPreparedness
- DrainageSystems
- CoastalVulnerability
- Landslides
- Watersheds
- DeterioratingInfrastructure

- PopulationScore
- WetlandLoss
- InadequatePlanning
- PoliticalFactors

Using these, the model will tell us the chance of a flood happening in a certain place. The goal is to help people and local leaders prepare early so they can protect their communities and reduce damage from floods.

1.1 Problem

Floods are becoming more common because of climate change. They can damage homes, farms, roads, and even cause deaths. Many communities do not have early warning systems, so they are not prepared when floods happen. This puts people's lives and property at risk.

1.2 Solution

This project builds a predictive model that uses weather data (like MonsoonIntensity, TopographyDrainage, Deforestation, Urbanization, ClimateChange, DamsQuality, Siltation, AgriculturalPractices, Encroachments, IneffectiveDisasterPreparedness, DrainageSystems, CoastalVulnerability, Landslides, Watersheds, DeterioratingInfrastructure, PopulationScore, WetlandLoss, InadequatePlanning, PoliticalFactors) to calculate the chance of a flood happening. The goal is to help communities get early warnings so they can prepare in advance, reduce damage, and stay safe.

1.3 Project objectives

- To present the results in a way that helps local communities and decision-makers take early action.
- To build a predictive model that uses weather data to estimate the probability of floods.
- To create a model that can be reused or improved in other regions with similar challenges.

2 Methodology

This project uses several environmental, human, and infrastructure-related factors to predict the likelihood of floods. Key weather variables like Monsoon Intensity and Climate Change influence rainfall and storm patterns. Physical

factors such as Topography, Drainage, Watersheds, and Coastal Vulnerability affect how water flows or accumulates. Human activities like Deforestation, Urbanization, Agricultural Practices, and Wetland Loss increase surface runoff and reduce natural water absorption. Infrastructure-related issues such as Dams Quality, Siltation, Drainage Systems, and Deteriorating Infrastructure can worsen flood impacts. Finally, social and political factors—including Population Score, Ineffective Disaster Preparedness, and Poor Planning—affect how communities respond to flood risks. Together, these variables provide a complete view for accurate flood prediction.

2.1 Study approach

This study follows a structured approach to build a flood prediction model. First, historical data on weather, environmental conditions, and human activities are collected and cleaned to remove errors and missing values. Then, exploratory data analysis (EDA) is used to understand patterns and relationships between variables such as monsoon intensity, deforestation, urbanization, and flood occurrence.

2.2 Data Source and Collection

The data used in this study was collected from Kaggle, an open online platform that provides high-quality datasets for data science and machine learning projects. The dataset includes historical records on rainfall, temperature, humidity, and several environmental and human-related factors such as monsoon intensity, urbanization, deforestation, drainage systems, and population exposure. These variables were downloaded in spreadsheet format and then cleaned, organized, and prepared for analysis. Kaggle offers reliable and diverse datasets contributed by researchers and institutions, making it a suitable source for building and testing predictive models related to flood risks.

2.3 Data Cleaning

Data cleaning is an important step to make the dataset accurate and ready for analysis. In this study, data cleaning involved checking for and handling missing values, such as empty cells in rainfall or temperature columns. Missing values were either filled using the average (mean) or removed if the data point was not useful.

Next, we looked for incorrect or outlier values, such as negative rainfall or extremely high humidity, and corrected or removed them. We also made sure all data types were correct—for example, changing text values like "Yes"/"No" into numbers like 1 and 0 where it is necessary, or converting date columns into proper date formats.

Finally, we organized the columns and rows clearly, removed duplicate entries, and ensured that each record matched the right flood probability. This clean dataset is now ready for accurate analysis and model building.

2.4 Exploratory data analysis

In this part, we understand the structure of the data and relationships between variables that may influence flooding. This is what we check:

- Look at summary statistics for each variable.
- Visualize the data: histograms, boxplots, and correlations.
- Identify trends, outliers, or patterns.

2.5 Feature engineering

In this part, we prepare and transform data features to improve model performance.

- We create the binary target variable Flood (already done in Step 2).
- We check if any variables need transformation (e.g., scaling, converting categorical variables).
- We might create new features or combine existing ones if useful (for now, we keep original variables).
- Remove variables that won't be used in modeling (e.g., if they are identifiers or irrelevant).

2.6 Model building

This study uses Logistic Regression as the main predictive model to estimate the probability of flooding based on environmental and weather-related variables. Logistic regression is a statistical method used for binary classification, meaning it helps predict whether a flood is likely to happen (Yes = 1) or not (No = 0), based on input data such as rainfall, temperature, humidity, and other factors.

This model was chosen because it is simple, easy to interpret, and effective for understanding how each variable affects the chance of flooding. The output of the model gives a probability score between 0 and 1, which represents the likelihood of a flood occurring under certain conditions. If the probability is above a certain threshold (e.g., 0.5), the model predicts a flood; otherwise, it predicts no flood.

Logistic regression formular

2.6.1 Probability form (Display mode):

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

2.6.2 Logit Form (Linear in predictors):

$$\log\left(\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

2.7 Logistic Regression Model Formula

The logistic regression model estimates the probability (P) of an event (such as flooding) occurring, using the following formula:

$$\begin{aligned} P(\text{Flood}) = \frac{1}{1+\exp(-(} & \beta_0 + \beta_1 \cdot \text{MI} + \beta_2 \cdot \text{TD} + \beta_3 \cdot \text{DF} + \beta_4 \cdot \text{UR} \\ & + \beta_5 \cdot \text{CC} + \beta_6 \cdot \text{DQ} + \beta_7 \cdot \text{SI} + \beta_8 \cdot \text{AP} + \beta_9 \cdot \text{EN} \\ & + \beta_{10} \cdot \text{IDP} + \beta_{11} \cdot \text{DS} + \beta_{12} \cdot \text{CV} + \beta_{13} \cdot \text{LS} + \beta_{14} \cdot \text{WS} \\ & + \beta_{15} \cdot \text{DI} + \beta_{16} \cdot \text{PS} + \beta_{17} \cdot \text{WL} + \beta_{18} \cdot \text{IP} + \beta_{19} \cdot \text{PF})) \end{aligned}$$

where:

- MI: Monsoon Intensity
- TD: Topography Drainage
- DF: Deforestation
- UR: Urbanization
- CC: Climate Change
- DQ: Dams Quality
- SI: Siltation
- AP: Agricultural Practices
- EN: Encroachments
- IDP: Ineffective Disaster Preparedness
- DS: Drainage Systems
- CV: Coastal Vulnerability
- LS: Landslides
- WS: Watersheds
- DI: Deteriorating Infrastructure
- PS: Population Score
- WL: Wetland Loss

IP: Inadequate Planning

PF: Political Factors

and where: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \ldots, \beta_{19}$ are coefficients.

2.8 Model Evalaution

We now assess how well the logistic regression model performs. Since this is a classification problem (Flood = 0 or 1), we'll evaluate it using:Accuracy, Confusion Matrix,Precision, Recall, F1-Score

where

- Accuracy: Measures how often the model's predictions are correct overall.
- Confusion Matrix: Gives a detailed summary of prediction performance by showing:
- F1-Score:Balance between precision and recall

2.9 Interprentation

From the output of summary(model), each coefficient () shows the effect of the corresponding variable on the log-odds of flood occurrence.

where

- A positive coefficient means an increase in that variable increases flood risk.
- A negative coefficient means an increase in that variable reduces flood risk.

3 Impact

This model developed in this project has a significant real-world impact which are follow:

- The forecasts help raise public awareness about the factors contributing to climate risks and empower citizens to take preventive actions.
- Local governments and emergency response teams can use the predictions to allocate resources, plan evacuation routes, and build resilient infrastructure.
- By accurately predicting flood probabilities based on climate and environmental data, the model can serve as an early warning tool for at-risk communities.
- Data-driven insights support government and NGO decision-making for sustainable urban planning, wetland protection, and climate adaptation programs.

4 Scalability

The predictive flood risk model developed in this study offers valuable tools for communities to prepare for climate-related disasters through proactive planning and early warning systems. By identifying key environmental drivers such as Monsoon Intensity, Urbanization, Wetland Loss, and Infrastructure Deterioration, the model enables targeted interventions and better disaster response.
Importantly, the model is highly scalable. It can be adapted and retrained using localized weather and environmental data to suit different geographic regions around the world. This flexibility aligns with global initiatives such as Microsoft's AI for Earth and the Data Science Global Impact Challenge, aiming to harness data science to address environmental sustainability challenges.