Data Collection and Preprocessing Phase

Date	20 June 2025
Project Title	Rising Waters: A Machine Learning Approach to Flood Prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan:

Section	Description
Project Overview	The machine learning project aims to predict the likelihood of a flood event based on key environmental and climatic parameters. Using a dataset with features such as rainfall patterns, temperature, humidity, cloud cover, and runoff levels, the objective is to build a robust classification model that determines whether a flood is expected (1) or not expected (0). This predictive capability enables early warnings, disaster preparedness, and more informed decision-making in flood-prone regions.
Data Collection Plan	● □ Searched for datasets related to meteorological and hydrological data influencing flood occurrences. ● □ Prioritized datasets with a wide range of environmental parameters across different seasons and regions. ● □ Focused on obtaining data that represents real-world weather variation to reduce overfitting and improve model generalization.

Raw Data Sources Identified	The raw data sources for this project include publicly available datasets from meteorological repositories and environmental research studies, supplemented with curated synthetic samples to balance the classes and enhance model learning. The final dataset includes attributes such as annual rainfall, seasonal rainfall distribution (e.g., Jan–Feb, Jun–Sep), average temperature and humidity, cloud cover, and surface runoff. These variables were used in the machine learning pipeline to build, evaluate, and deploy models such as XGBoost, Random Forest, Decision Tree, and KNN for flood prediction.,

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises key flood prediction variables such as climatic and hydrological indicators, including temperature, humidity, cloud cover, annual and seasonal rainfall distributions (e.g., Jan–Feb, Mar–May, Jun–Sep, Oct–Dec), average June rainfall, and subsurface runoff.	https://www.kaggle.com/datasets/arbethi/rainfall-dataset	Excel Sheet	16 kB	Public

Kaggle Dataset	This data concerns rainfall in India from 1901-2015	https://www.kaggle.com/datasets/arbethi/rainfall-dataset	Excel Sheet	503 kB	Public