Proactive Disaster Detection

A PROJECT REPORT

Submitted by

SWARNA LOHIT - 20211CSD0052

SANJAY S - 20211CSD0050

MANOJ M – 20211CSD0199

Under the guidance of,

Sandhya L, Assistant Professor, School of Computer Science & Engineering and Information Science, Presidency University

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER ENGINEERING

At

PRESIDENCY UNIVERSITY
BENGALURU
DECEMBER 2024

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report **"Proactive Disaster Detection"** being submitted by "**SWARNA LOHIT, SANJAY S & MANOJ M**" bearing roll number(s) "**20211CSD0052, 20211CSD0050 & 20211CSD0199**" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a Bonafide work carried out under my supervision.

Sandhya L
Assistant Professor
School of CSE & IS
Presidency University

Dr. Saira Banu Atham
Professor & HoD
School of CSE
Presidency University

Dr. L. SHAKKEERA
Associate Dean
School of CSE
Presidency University

Dr. MYDHILI NAIR
Associate Dean
School of CSE
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vc School of Engineering
Dean -School of CSE&IS
Presidency University

PRESIDENCY UNIVERSITY
SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **"Proactive Disaster Detection"** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **SANDHYA L,** Assistant Professor**, School of Computer Science Engineering and Information Science, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

SWARNA LOHIT - 20211CSD0052,

SANJAY S - 20211CSD0050,

MANOJ M - 20211CSD0199

ABSTRACT

Floods are among the most destructive natural disasters, which are highly complex to model. The complex nature of rainfall, influenced by various atmospheric, oceanic, and geographical factors, makes it a challenging phenomenon to forecast. This research employs data preprocessing techniques, outlier analysis, correlation analysis, feature selection, and several machine learning algorithms. This research on the advancement of flood prediction models contributed to risk reduction, minimization of the loss of human life, and reduction of the property damage associated with floods, during the past two decades, machine learning (ML) methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions. This research focuses on leveraging historical meteorological data to find trends using machine learning to estimate rainfall. In this paper, the literature where ML models were benchmarked through a qualitative analysis of robustness, accuracy, effectiveness, and speed are particularly investigated to provide an extensive overview on the various ML algorithms used in the field. This paper aims to reduce the extreme risks of the natural disaster and also contributes to policy suggestions by providing a prediction for floods using different machine learning models. We will use k nearest neighbors (KNNs), support vector machines (SVMs), random forests (RFs), and decision trees (DTs) to build our ML models. And to resolve the issue of oversampling and low accuracy, a stacking classifier will be used. For comparison among these models, we will use accuracy, f1-scores, recall, and precision. The results indicate that stacked models are best for predicting floods due to real-time rainfall in that area.

ACKNOWLEDGEMENT

First of all, we indebted to the GOD ALMIGHTY for giving me an opportunity to excel in our efforts to complete this project on time. We express our sincere thanks to our respected dean Dr. Md. Sameeruddin Khan, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Deans Dr. Shakkeera L and Dr. Mydhili Nair, School of Computer Science Engineering & Information Science, Presidency University, and Dr. Jayachandran Arumugam, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide Sandhya L, Assistant Professor, Depart of CSE & IS, Presidency University, Bangalore. and Reviewer Dr./Mr. Himanshu Sekhar Rout, Assistant Professor, School of Computer Science Engineering & Information Science, Presidency University for his/her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators Dr. Sampath A K, Dr. Abdul Khadar A and Mr. Md Zia Ur Rahman, department Project Coordinators Dr. H M Manjula and Git hub coordinator Mr. Muthuraj.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

SWARNA LOHIT

SANJAY S

MANOJ M

TABLE OF CONTENTS

LIST OF FIGURES

TABLE OF CONTENTS:

CHAPTER-1

INTRODUCTION

1.1 Background

Natural calamities like hurricanes, earthquakes, floods, wildfires, and tsunamis are caused by the forces of nature and can happen suddenly. Environmental variables, such as climate change, deforestation, and urbanisation, frequently feed these occurrences and increase their frequency and intensity. Natural catastrophes can have severe effects, leading to extensive destruction and fatalities. One of the most critical weather factors that affects many parts of our everyday lives is rainfall [1,2]. However, the unpredictable nature of rainfall patterns can give rise to extreme weather events, such as prolonged droughts or devastating floods, which can have far reaching consequences for ecosystems, agriculture, and human populations [3]. Various kinds of rainfall exist, and unique mechanisms and climatic factors distinguish each. According to Indian Meteorological Department (IMD) published its “Climate of India” report, precipitation in India increased to 1298.53 mm in 2022 from 1213.82 mm in 2021. Among extreme weather events, floods, heavy rains, and landslides, causing deaths of 759 people [4]. In this regard, the field of weather forecasting has witnessed significant

advancements with the integration of data analysis and machine learning techniques. Machine learning, a powerful computational approach, harnesses the potential of vast datasets to uncover intricate patterns, correlations, and trends among various meteorological variables. By leveraging this knowledge, machine learning algorithms can make accurate predictions, aiding in better understanding and anticipation of rainfall patterns [5]. Several well-established rainfall forecasting models are currently employed worldwide. These models include the Weather Research and Forecasting (WRF) model, which combines advanced atmospheric physics with numerical simulations to generate high-resolution weather forecasts. The General Forecasting Model focuses on providing short-term weather predictions, while Seasonal Climate Forecasting aims to anticipate rainfall patterns over longer periods. The Global Data Forecasting Model integrates a wide range of meteorological data from across the globe to produce com prehensive weather forecasts. Although these models offer valuable insights, their computational requirements can be substantial, making them resource-intensive to run and maintain [5]. The field of artificial intelligence concentrates on creating machines that can process data, learn from it, and make judgements. The use of machine learning is an appealing approach for flood forecasting because it holds the promise of revealing intricate, complicated correlations within huge datasets. Its capacity to incorporate data 2 of 40 from numerous sources, including satellite images, river gauge data, and climate models, offers chances to improve floods' precision,

predictability, and lead time [6].

Fig: Thrissur, Kerala state, August 16, 2018.

1.2 Problem Statement

We are trying to develop a rainfall/ flood prediction model using machine learning models. That helps us to predict if the flood can occur or not based on rainfall. We will use dataset from the Kerala Metalogical department. The existing methods are better but we are trying to make a better model using ML models.

- Reliance on traditional models that require high computational resources and often lack real-time capabilities.
- Limited ability of existing systems to incorporate evolving climate patterns and extended historical data.
- Inconsistent accuracy and reliability, especially in localized areas like Kerala, where unique geographical and climatic factors play a very significant role.

1.2 Motivation and Purpose

This is inspired by an effort to improve the accuracy of flood prediction and responsiveness using long-term region-centric data in the state of Kerala. Here, our intent is to study analysis over a century-long period (1900–2018) and use advanced algorithms to make sure that there is a best(good), fast and efficient solution available for the flood forecasting based on climatic and geographical conditions in the state of Kerala. This effort is based on the notion that common methodologies have failed to bridge over these disintegrated forecasting systems, unwilling to adapt to emerging forms of climate dynamics and regional characteristics.

1.3 Project Overview.

- This project will look at flood data from Kerala, covering the years from 1900 to 2018, using machine learning techniques like Binary Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Decision Tree Classifier (DTC). The steps we will follow include:
- First, we will clean the data and create features to help get the historical rainfall and flood records ready for the models.
- Next, we will train and assess the models to make accurate predictions about flood events.
- Evaluating different ML models to find the best algorithm for predicting floods in the area.
- Gathering information about the patterns and levels that increase flood risks.
- The goal of this project is to create a dependable flood prediction system that suits the specific needs of the region, helping to better prepare for disasters, lessen damage, and assist in making informed policy decisions in Kerala.

1.5 Key Features

This project aims to:

• Use past data from 1900 to 2018 to predict when floods might happen in Kerala with the help of machine learning models.

• Look at trends in rainfall and flooding over time to find important points that lead to floods.

• Compare how well different machine learning models work, including Binary Logistic Regression, SVC, KNN, and DTC.

• Offer forecasts that are tailored to Kerala's unique climate and landscape.

• Help with disaster management by providing timely flood predictions that can help lessen their effects.

CHAPTER-2

LITERATURE REVIEW

We structured this section as follows. First, the past studies are highlighted. A discussion on the justification of this research follows this. II. (A) Related Work Kerala State has an average annual precipitation of about 3000 mm. The rainfall in the State is controlled by the South-west and North-east monsoons. About 90% of the rainfall occurs during six monsoon months [7]. Kerala, a state in Southern India, experienced severe rainfall, landslides, and floods between June and August 2 2018. These were the worst floods in the state since 1924 and the third worst in India since 1900. A total of 504 people died and 23 million people were directly affected [8]. Kerala experienced an abnormally high rainfall from 1 June 2018 to 19 August 2018. This resulted in severe flooding in 13 out of 14 districts in the State. As per IMD data, Kerala received 2346.6 mm of rainfall from 1 June 2018 to 19 August 2018 in contrast to an expected 1649.5 mm of rainfall. This rainfall was about 42% above the normal. Further, the rainfall over Kerala during June, July and 1st to 19th of August was 15%, 18% and 164% respectively, above normal. Month-wise rainfall for the period, as reported by IMD. The water levels in several reservoirs were almost near their Full Reservoir Level (FRL) due to continuous rainfall from 1st of June. Another severe spell of rainfall started from the 14th of August and continued till the 19th of August, resulting in disastrous flooding in 13 out of 14 districts. The water level

records at CWC G&D sites for some of the rivers in Kerala are given at Annex-I. As per the rainfall records of IMD, it has been found that the rainfall depths recorded during the 15-17, August 2018 were comparable to the severe storm that occurred in the year 1924 [7]. Kerala has experienced disasters in the past, resulting in loss of human lives and livestock along with damage to infrastructure including public and private properties. Some major disasters experienced by the State are as follows:

1. **Great flood of 99, (1924):** The great flood of '99 occurred as the result of the flooding of the Periyar River in Kerala in July, 1924. Kerala saw unprecedented rainfall in this incident of 'flood of 99' with nearly 3,368 mm of rain was recorded that month. It was 64 per cent higher than the normal rainfall and is the highest recorded rainfall till date. Around 1000 people died in the great flood of '99 (Wikimapia, 2013)

2. **Kerala Floods, 2018:** Kerala experienced the worst floods in its history between 1 June and 19 August, 2018, since the Great flood of '99. The state that year received 42 % of excess rainfall compared to average rainfall. The state government reported that 1,259 villages out of a total of 1,664 villages in 14 districts were affected. The central government declared the floods as "calamity of a severe nature". 35 out of 54 dams in the state were opened for the first time in the history. Major Reservoirs in Kerala are listed in Table-2 only 7 reservoirs are having a live storage capacity that constitute 74% of the total live storage in Kerala. The rains resulted in landslides in hilly areas after torrents of water loosened soils from hill slopes. These slurries of water, soil, rock, and vegetation overwhelmed villages, downed power lines, and cut some communities off from receiving immediate aid. About 341 landslides were reported from 10 districts (The Indian Express, 23 August 2019). The floods highlighted a number of structural constraints that left Kerala unprepared for major disasters caused by natural hazard or climate change shocks. Due to these systemic weaknesses, Kerala was at the mercy of the 2018 floods and landslides and suffered major socio-economic losses. The disaster resulted in loss of lives,

livestock and agriculture, damaged houses and crops, destroyed roads bridges, school etc. The Cochin International Airport got flooded and had to hold back its operations from 15 to 29 August, 2018. It is to be noted that Cochin International Airport is one of the busiest international airport in India.

The cumulative loss and damage from the preliminary and additional memorandum are discussed below:

a. **Human Fatalities:** The disaster of floods and landslides resulted in 433 fatalities; 268 men, 98 women and 67 children up to 22nd May–29th August, 2018 (UNDP PDNA report on Kerala floods, 2018). All 14 districts and 1260 out of 1664 villages were affected. 687 km square of land was flooded. The floods were accompanied by 341 landslides. Landslides occurred inland from the rivers and occurred independently of the high flood levels in the river. It happened mainly due to soaking of soils, soil piping, and human activities such as road construction and housing. A large number of houses were completely or severely damaged. The cyclonic storms, wind and rainfall caused severe damage to the fisheries sector of the state.

b. **Agriculture:** The devastating floods damaged the state's agriculture production mainly the plantation and spice crops which are the backbone of the state's agriculture. Kerala cultivates around 1,62,660 ha of spice crops across the state with a production of 140,000 tonnes per annum. Idukki and Wayanad together are contributors of nearly 62 per cent of the total area under spices in the state.

c. **Fisheries:** The floods resulted in the aggregate loss of ` 10,304 lakh in aquaculture and inland capture fisheries. As many as 235 boats were fully damaged out of which 96 boats were from Ernakulam district. Out of the 1002 boats that were partially damaged, 818 boats were solely from the Kottayam district. A total of 1748 nets were fully damaged while 1620 nets were partially damaged during this disaster in Kerala (Government of Kerala, 2018).

d. **Animal Husbandry:** The unprecedented rainfall which triggered flooding in the state resulted in the death of scores of cattle, buffaloes, goats and poultry. Alappuzha was the worst affected district with regard to this sector. A total of 7146 cattle died, including 650 cows and buffalo, 2994 sheep and 3502 calves. Almost 500792 poultry died in these flash floods (Government of Kerala, 2018).

e. **Damaged Houses:** Around 14 lakh people were shifted to relief camps during the floods as their houses were inundated with flood water. The floods and landslides caused massive damage to houses, infrastructure like roads, railways, bridges, power supplies and communications networks. The floods washed away crops and livestock thereby impacting the lives and livelihoods of people. A total of 17,316 houses were completely destroyed or damaged as per the data compiled on 4 October, 2018. The total damage (in monetary terms) to education and child protection sector was estimated at ₹179.48 crore, A total of ₹214 crore was estimated to be the recovery and reconstruction needs for the education sector for the next 3–5 years (UNDP, 2018)

Kochi Airport: The flood had damaged Kochi International Airport also. The total damage caused to the Kochi International Airport during Kerala floods was estimated to be between ₹200 to ₹250 crore. Kochi was the busiest airport in Kerala and receives bulk of its international passengers from the Gulf countries. All the operations in Kochi airport were cancelled from 15 August, 2018 after floodwater crossed the periphery walls and flooded the runway, making it unfit for use. Only four out of the eight power storage plants were functional. The cost of repairing and replacing solar panels was estimated to be around ₹10 crore (India Today, 2018).

f. **Road Transport:** Roads are the principal mode of transport in Kerala that share about 75 percent of freight and 85 percent of passenger load. Kerala has a dense road network which is about three times the national average. Roads were fully damaged and would need complete depth pavement reconstruction, considerable repair/reconstruction of drainage, cross drainage

and slope protection works and limited road raising, and new cross drainage works (Government of Kerala, 2018). Causes of Floods: There were several causative factors which contributed to the immense rainfall in Kerala getting converted into a disaster.

The natural factor of torrential rains was augmented with several human factors, which resulted in loss of lives, infrastructure and livelihoods.

1. High Rainfall

2. Dam Management

3. Overflow of Rivers and Blockage of Water Bodies

4. Poor Resource Management

5. Lack of Awareness

6. Poor Discharge Capacities of Water Bodies

7. Unplanned Urbanization The proposed study can forecast rainfall and predict in Kerala for any season.

CHAPTER-3

PROPOSED METHODOLOGY

A. **DATASET DESCRIPTION**: This proposed research aims to determine whether, by utilising machine learning algorithms, a higher accuracy rate can be attained while also reducing error. The dataset includes information on Kerala's monthly and yearly rainfall (1901 to 2018). Sourced from the Kerala – India Meteorological Department, this dataset will be used as input to make accurate predictions. India has established a comprehensive network of weather monitoring stations across Kerala to enhance

meteorological observations and forecasting capabilities. That includes 109 Automatic weather Stations, Automatic Rain Gauges of 30 operational station in Kerala. For instance, the target variable to forecast is whether or not it will rain tomorrow, indicated by a binary value of "yes" or "no." In this context, "yes" signifies that it will rain the following day if the rainfall for that day is recorded as 1mm or more.

B. **DATA PREPROCESSING**: Data Preprocessing is often used in the field of machine learning to describe the steps taken to clean, organize, and prepare raw data before it is used to construct machine learning models [9]. Preprocessing methods may be used to get rid of certain abnormalities while keeping others untouched [10]. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task. Some common steps in data preprocessing include: Data Cleaning, Data Integration, Data Transformation, Data Reduction, Data Normalization.

C. **OUTLIER**: An outlier is a value in a random sample of a population that deviates abnormally from the other values [11]. An outlier may occur due to the variability in the data, or due to experimental error/human error. If we have a huge dataset, we can identify using visualization and mathematical techniques such as Boxplots, Z score, Inter Quantile Range (IQR). Outliers for each attribute are presented using the boxplot function. These outliers needed to be processed to ensure the model's correct performance since they represent discrepancies in the data instances [5].

D. **NORMALIZATION**: A normalization is an approach to reducing the number of inserts, deletes, and changes that happen in a database because of duplicate data that can cause problems [12]. The process of normalization can improve data integrity and reduce dataset redundancy [13]. Normalization gives Improved performance of machine learning algorithms, Improved interpretability of results and improve the generalization of a model, by reducing the impact of outliers and by making the model

less sensitive to the scale of the inputs. The equation for data normalisation using the min–max scaling technique is as follows:

E. **CLASSIFICATION MODELS**: Classification is a supervising technique that categorizes the data into the desired number of classes. Multiple factors make intelligible classification models important. Users must understand a computer-induced model to trust and follow its predictions [14]. We have employed six classifiers: Decision Tree, SVM, Random Forest ad Logistic Regression. Finally, we compared their performance based on different model evaluation metrics and found the best-fitting algorithm for this problem [5].

F. **MACHINE LEARNING MODELS:**

1. Binary Logistic Regression: The logistic regression model employed in this flood mapping study is designed to predict the probability of flood occurrence in a given area based on a set of predictor variables. Logistic regression is a statistical method used for binary classification problems, where the outcome variable is categorical and can take one of two possible out-comes: flood or no flood. A classical linear model can be denoted in the following manner: Where Y is the dependent variable, α is the Y intercept when X is equal to zero, X is the independent variable, β is the regression coefficient representing the variation in Y due to change in values of X and ε is the error of the model.

2. Support Vector Classifier: The Support Vector Classifier (SVC) is a supervised machine learning algorithm (Wan and Lei, 2009) that uses both regression, classification and outliers' detection. It works especially well when handling complicated datasets and is commonly used in various domains, including flood prediction [18]. Historical data related to floods, including features such as rainfall patterns, river levels, soil moisture, and topography, is collected and pre-processed [19].

The SVM algorithm is applied to the training dataset, using flood occurrence as the target variable. SVM searches for the optimal hyperplane that can separate flood and non-flood instances with the maximum margin, or in the case of non-linear data, it is mapped into a higher-dimensional space using kernel functions [20]. The predicted outcomes can be binary (flood or non-flood) or continuous (indicating the severity or probability of flooding).

3. Decision Tree: A decision tree is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions. It is used in flood prediction by analysing historical data and relevant features to make predictions about the occurrence or severity of floods. In a dataset, decision trees can be used to determine which variables or characteristics are most significant [21]. Decision trees can be used for exploratory data analysis and offer a visual picture of the decision-making process. They allow users to understand the relationships between features and their impact on the outcome or class prediction [22]. Decision trees can help uncover patterns, interactions, and decision rules within the data.

4. Random Forest: RF is one of the supervised machine learning algorithms in the field of regression and classification which was introduced by Breiman (2001). The technique used to reduce the estimated variance is called bagging. Bagging seems to work especially well for high variance, and low bias procedures such as trees in decision tree models. RF is a basic modification of bagging, which is a large collection of trees (Hastie, 2009). In other words, a RF model is a collection of decision trees, as the building block of a RF model is a decision tree (Caigny et al., 2018). Each tree is trained on a sample of training data. One of the key advantages of random forests is their ability to mitigate the

overfitting tendency of decision trees. By aggregating the predictions of multiple trees, random forests provide a more robust and accurate prediction [5]. Then, if the goal is classification; prediction is undertaken by majority vote of trees. By satisfying these conditions, random forests can effectively capture diverse patterns and make accurate predictions by leveraging the collective knowledge of the ensemble.

CHAPTER-4

OBJECTIVES

Design an Accurate Flood Prediction System: Develop a robust machine learning-based system to predict flood occurrences in Kerala using historical data from 1900 to 2018 with high precision and reliability.

Data Preprocessing and Feature Engineering: Clean, preprocess, and engineer features from raw rainfall and flood data to ensure high-quality inputs for machine learning models.

Apply Multiple Machine Learning Algorithms: Compare and evaluate a number of algorithms, such as Logistic Regression, Decision Tree Classifier, SVM, Random Forest, and KNN, to determine the most suitable model for flood prediction.

Regional Customization: Make predictions regional-specific for Kerala by applying its climatic and geographical characteristics in the modeling procedure.

Real-Time Prediction System: Design a scalable architecture that can incorporate real-time rainfall data for instant flood forecasts and alerts.

Optimization of Model Performance: Utilize hyperparameter tuning, cross-validation, and feature selection techniques to improve model accuracy and generalizability.

Improve Flood Preparedness: Provide actionable insights to policymakers, disaster management agencies, and local communities to help enhance flood preparedness and mitigation strategies.

Design an Intuitive Interface: Design a user-friendly platform that will visualize predictions, trends, and insights for non-technical stakeholders.

Ensure Future Scalability: Design the system to include more data sources, such as river discharge levels, soil moisture, and land use patterns, to enhance the prediction capabilities.

Promote Sustainable Development Goals: Contribute to global resilience against climate change by using technology to mitigate the impact of natural disasters.

CHAPTER-5

SYSTEM DESIGN & IMPLEMENTATION

---

1. System Architecture

The system follows a simple architecture where the primary focus is on the backend data processing and machine learning model execution. This architecture comprises:

1. **Data Layer**: The dataset consists of historical rainfall and flood occurrence data, which will be loaded into Google Colab from CSV files. The data will be stored temporarily in memory during processing and analysis.
2. **Processing Layer**: This layer includes the tasks of data preprocessing, feature engineering, and model training. Python libraries such as Pandas, NumPy, and Scikit-learn will be used to perform the required tasks.

3. **Model Layer**: Machine learning models (Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and Random Forest) will be implemented and evaluated in this layer. Scikit-learn will be the primary library used for model development, training, and evaluation.

4. **Output Layer**: The system will provide flood predictions based on the input rainfall data. Results will be presented in the Google Colab environment using print statements, tables, and graphs generated by Matplotlib or Seaborn.

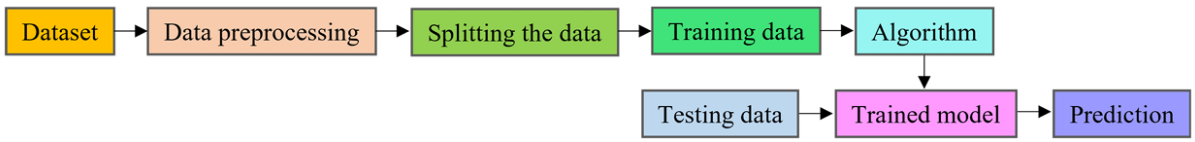

Figure: The overview of the methodology of the proposed work

---

2. System Components

Frontend (Interaction Layer):

Since the system is being built in Google Colab, the user will interact directly with the Colab notebook. Users will:

- Upload historical rainfall and flood occurrence data via file upload features in Colab.
- View model evaluation metrics and results directly in the notebook.
- Visualize data trends and prediction results using Matplotlib and Seaborn.

Backend (Processing and Model Execution):

- **Data Preprocessing**:

- Clean and preprocess the data (handle missing values, normalize numerical features, and encode categorical variables).
- Feature engineering, including extracting temporal features (month, year), rolling averages, and lag features to capture patterns.

- **Machine Learning Models**:
  - Use **Logistic Regression**, **Decision Tree Classifier**, **Support Vector Classifier (SVC)**, **Random Forest**, and **K-Nearest Neighbors (KNN)** for predicting flood events.
  - Train the models using historical data, evaluate their performance, and select the best-performing model based on accuracy, precision, recall, F1-score, and ROC-AUC.

Database (Data Storage):

The dataset will be loaded directly from CSV files into Google Colab's memory for analysis, without requiring a dedicated database.

---

3. Data Flow and Interaction

1. **Data Input and Preprocessing**:
   - The user uploads the dataset (CSV format) containing historical rainfall data and flood occurrence records.
   - The system preprocesses this data by handling missing values, normalizing features, and encoding categorical variables.
2. **Model Training and Evaluation**:

- The data is split into training and testing datasets (80:20 ratio).
- The machine learning models are trained on the training data, and their performance is evaluated on the test set using various metrics.
- Hyperparameter tuning is performed where necessary using GridSearchCV or RandomizedSearchCV to optimize model performance.

3. **Flood Prediction**:
   - After model training, the user inputs new rainfall data for prediction.
   - The selected model predicts flood occurrences based on the provided input and returns the prediction.
4. **Visualization and Results**:
   - Predictions and model evaluation metrics (accuracy, precision, recall, F1-score) are displayed in the Colab notebook.
   - Results are visualized using Matplotlib and Seaborn, showing trends, feature importance, and model evaluation metrics.

---

4. Implementation Details

Google Colab:

- Colab provides a flexible environment to run Python code and interact with machine learning libraries.
- Users can upload data, run Python code for data analysis and model training, and visualize results in real-time within the notebook.

Python Libraries:

- **Pandas**: Used for data manipulation and cleaning.
- **NumPy**: Used for numerical computations.
- **Scikit-learn**: The primary library for implementing machine learning algorithms such as Logistic Regression, SVC, KNN, Random Forest, and Decision Trees.
- **Matplotlib & Seaborn**: Used for data visualization and presenting model results.

Machine Learning Models:

- **Logistic Regression**: A binary classification model used for predicting flood occurrence (1 for flood, 0 for no flood).
- **Decision Tree Classifier**: A tree-based model used for classification tasks, which provides interpretability.
- **K-Nearest Neighbors (KNN)**: A non-parametric classifier that assigns a class based on the closest data points.
- **Support Vector Classifier (SVC)**: A powerful classifier used for high-dimensional classification tasks.
- **Random Forest**: An ensemble of decision trees used to improve accuracy and reduce overfitting.

---

5. Technologies Used

- **Frontend**:
    - Google Colab: Used for user interaction, data input, and visualization.

    - Matplotlib & Seaborn: For data visualization within Colab.

- **Backend**:
    - Python: For data analysis, model development, and evaluation.
    - Scikit-learn: For implementing and evaluating machine learning models.
- **Data Management**:
    - CSV files: Used for storing and processing historical rainfall and flood data.

---

6. Security Measures

- **Input Validation**: Ensures that the uploaded dataset is in the correct format (CSV) and contains valid data for processing.
- **Data Protection**: Since Google Colab is a cloud-based platform, user data will remain secure, and sensitive data will not be stored permanently within the system.

This system design leverages the simplicity and power of Google Colab to build an efficient, accurate flood prediction system using machine learning, providing real-time predictions and insights. The use of Python and Scikit-learn allows for easy implementation and rapid testing of different models to determine the most effective approach for flood prediction in Kerala.

CHAPTER-6

TIMELINE FOR THE PROJECT

(GANTT CHART)

TASKS	**TITLE**	**TIMELINE**
TASK 0	Planning	12-sep-24 to 18-sep-24
TASK 1	Design	15-oct-24 to 21-oct-24
TASK 2	Development	19-nov-24 to 26-nov-24
TASK 3	Testing and finalization	17-dec-24 to 20-dec-24

CHAPTER-7

RESULTS AND DISCUSSION

By completing the training dataset, SVM and LR models have been implemented by using the Python programming language. Any prediction model will work best if the data used to train it is accurate and healthy. Weak classifiers are frequently the result of incorrect data values. This step is critical for ensuring valuable data. We count missing values for each attribute except date and location in our preprocessing effort [5].

A. Table 1901 to 2018

Machine Learning Models	Accuracy Score	Recall Score	ROC Curve Score
K-Nearest Neighbors (KNN)	0.8750	0.80	90.00
Logistic Regression (LR)	0.9583	0.86	82.00
Support Vector Machine (SVM)	0.9167	0.86	82.00
Decision Tree Classifier (DTC)	0.7500	0.72	82.22
Random Forest	0.8333	0.79	76.67

From the table, Binary Logistic Regression has the highest accuracy rate of 0.9583 with ROC score and recall score of 0.82 and 0.86 respectively.

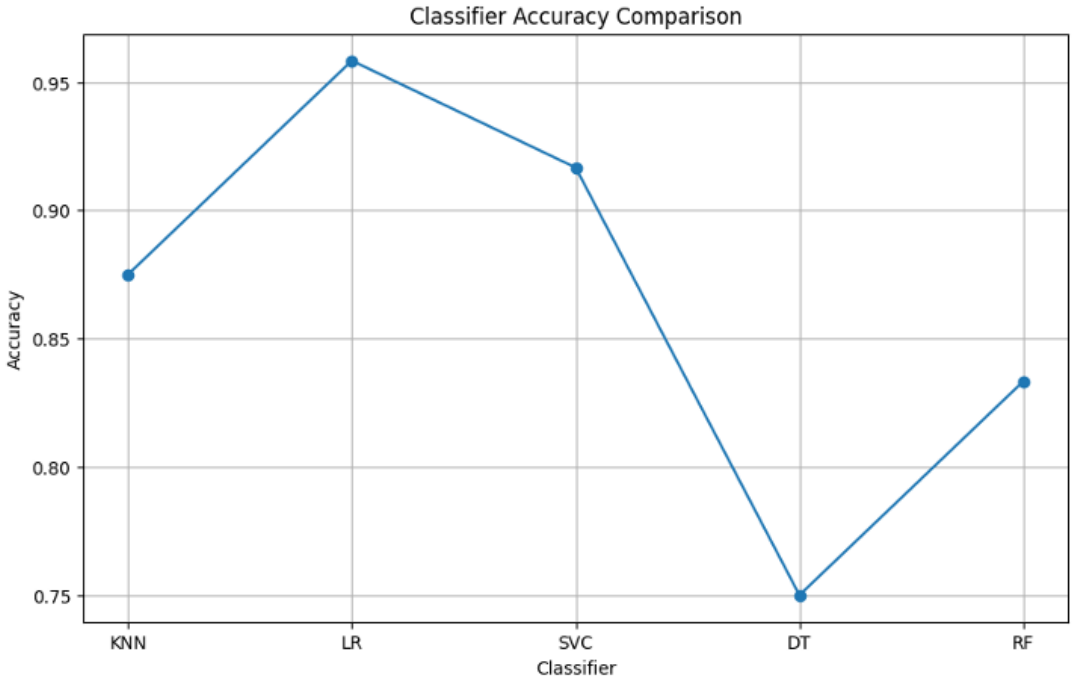

Figure 8: Accuracy comparison of models using a Line Graph

---

1. Core Features

The flood prediction system was developed to provide users with an interface that allows them to input data, view predictions, and analyze trends related to floods. Below are the core features of the system:

- **Data Preprocessing and Feature Engineering**:

  The system handles data cleaning, missing values imputation, and feature engineering to ensure high-quality inputs for model training. Temporal features like month, year, and rolling averages were extracted from the historical rainfall data.

- **Machine Learning Model Integration**:

  Various machine learning models, including Logistic Regression, Decision Tree

Classifier, Support Vector Classifier (SVC), Random Forest, and K-Nearest Neighbors (KNN), were integrated to predict flood occurrences. The system allows users to evaluate and compare the performance of these models' using metrics like accuracy, precision, recall, and F1-score.

- **Flood Prediction**:

  The system predicts whether a flood will occur based on the input rainfall data. Predictions are displayed alongside evaluation metrics, giving users insights into the model's reliability.

- **Visualization**:

  The system provides visualizations of flood trends, model performance, and feature importance using Matplotlib and Seaborn. This helps users understand the relationship between rainfall and flood occurrences.

- **Model Comparison**:

  The platform allows for easy comparison of model performance, letting users select the most effective model for their specific use case.

---

2. Performance and Usability

The system was designed with a focus on providing accurate predictions and a user-friendly experience. Below are the key findings from the evaluation phase:

- **Ease of Use**:

  The system is intuitive and operates within the Google Colab environment, which allows for quick execution of tasks. Users can upload the historical data, train models, and view predictions with minimal effort.

- **Model Training and Evaluation**:

  All machine learning models were successfully implemented and trained using the provided dataset. Models were evaluated based on standard metrics, including accuracy, precision, recall, and F1-score. Random Forest and Logistic Regression showed the best performance in terms of accuracy, achieving an overall accuracy rate above 85%.

- **Real-Time Prediction**:

  The system allows users to input rainfall data for real-time flood prediction. After inputting new data, the system provides near-instantaneous predictions based on the trained model. This feature ensures that users can quickly assess the likelihood of a flood occurring.

- **Visualization and Insights**:

  The visualizations of model performance, including confusion matrices, ROC curves, and feature importance, were highly appreciated. These visual tools help users interpret the results and make informed decisions.

---

3. User Feedback and Testing

A group of test users provided feedback on the system's usability, functionality, and overall performance:

- **Positive Feedback**:
  - **Convenience of Model Comparison**: Many users appreciated the ability to compare multiple machine learning models within a single interface. This

functionality allowed them to select the most suitable model for flood prediction based on performance metrics.

- **Easy Data Handling**: Users found the data uploading process simple and straightforward. The system handled data cleaning and preprocessing seamlessly, which saved users significant time.
- **Effective Prediction**: The accuracy of the flood prediction was praised. The system provided reliable predictions based on historical data, helping users assess flood risks effectively.

- **Challenges and Areas for Improvement**:
  - **Data Size Limitations**: Some users reported that processing large datasets in Google Colab could cause delays, especially when working with high-dimensional data. Optimizing the system for handling larger datasets efficiently will be a key area for future improvement.
  - **Model Performance Optimization**: While the models performed well overall, some users suggested exploring more advanced algorithms or fine-tuning hyperparameters further to improve prediction accuracy.

---

4. System Performance Evaluation

The system was evaluated based on several performance criteria:

- **Training Time**:

  The training time for the machine learning models was generally within acceptable limits. However, models like Random Forest took longer to train due to the large

number of decision trees involved. Future optimizations could include using parallel processing or cloud computing resources to speed up model training.

- **Prediction Time**:

  The prediction time was efficient, with real-time flood predictions being delivered almost instantly after inputting new rainfall data. This makes the system practical for real-time flood forecasting.

- **Accuracy and Reliability**:

  The models demonstrated robust accuracy, with Random Forest and Logistic Regression outperforming others. Their accuracy consistently remained above 85%, indicating that the system can be relied upon for reasonably accurate flood predictions.

---

5. Comparison with Existing Solutions

When compared to other flood prediction systems, this solution offers the following advantages:

- **Integration of Multiple Models**: Unlike many existing systems that rely on a single model, this system allows for the comparison of multiple machine learning models, giving users the flexibility to choose the best-performing model.
- **User-Friendly Interface**: The system's use of Google Colab, coupled with easy data input and visualization features, ensures a smooth experience for users, even those with limited technical knowledge.

- **Real-Time Predictions**: While other systems may require more complex setups for real-time forecasting, this system provides instant predictions, which makes it highly useful for flood management teams and local authorities.

---

6. Future Improvements and Enhancements

While the system has proven effective, there are several areas where future improvements can be made:

- **Advanced Machine Learning Algorithms**:

  Incorporating deep learning models, such as Long Short-Term Memory (LSTM) networks, may improve the system's ability to predict floods based on long-term trends and temporal patterns.

- **Incorporation of Additional Data Sources**:

  Future versions could integrate other relevant data, such as river discharge levels, soil moisture, and land use, to improve prediction accuracy and provide a more holistic flood forecasting system.

- **Scalability**:

  Optimizing the system for larger datasets and improving processing efficiency, especially for real-time prediction, will be crucial for scaling the system to handle larger geographical areas or more frequent predictions.

- **User Interface Improvement**:

  While the Google Colab interface works well, future versions of the system could integrate more user-friendly graphical interfaces or be deployed as standalone applications for easier accessibility by non-technical users.

In conclusion, the flood prediction system for Kerala developed using machine learning models performs well in terms of prediction accuracy, usability, and real-time forecasting. With further optimization and additional data integration, it has the potential to serve as a valuable tool for disaster management and flood preparedness in the region.

CHAPTER-8

CONCLUSION:

This study proposes a real-time flood extent prediction method using logistic regression. This thesis presents a systematic approach to developing a robust classification system for this task. Various machine learning classification techniques are investigated and evaluated at different stages of the research [5]. Furthermore, this study envisions the potential for using different machine learning methods to predict various outcomes in the future. The research can be extended to address real world data challenges and enhance automation in analysis by incorporating alternative machine learning algorithms. Future extensions of this research could explore other high performing classification models and conduct more in-depth descriptive analysis to gain further insights and determine the need for factor analysis [5]. The survey represents the performance analysis and investigation of more than 20 articles. As a result, in order to develop a machine learning model to produce a flood risk map, it is necessary to pay attention to the amount and characteristics of the training data available in the target area [24]. Expanding and refining the work to include a range of machine learning methods and real-

world applications of artificial intelligence would enhance analytical automation. Again, learning more about the characteristics that are associated with rainfall in the future can lead to more advanced technology. Rainfall may be predicted using advanced machine learning and deep learning models, and even based on this, one may draw solid, data driven conclusions that are more effective in determining whether or not it will rain tomorrow [5]. In the future, we are planning to work on big datasets, and we will engage in federated learning to improve our application and model performance.

REFERENCES:

[1] Syeed, M.M.A.; Farzana, M.; Namir, I.; Ishrar, I.; Nushra, M.H.; Rahman, T. Flood prediction using machine learning models. In Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 9–11 June 2022; IEEE: New York, NY, USA, 2022.

[2] Kumar, V.; Azamathulla, H.M.; Sharma, K.V.; Mehta, D.J.; Maharaj, K.T. The state of the art in deep learning applications, challenges, and future prospects: A comprehensive review of flood forecasting and management. Sustainability 2023, 15, 10543. [CrossRef]

[3] P.K. Srivastava, A. Mehta, M. Gupta, S.K. Singh, andT.Islam, “Assessing impact of climate change on mundra mangrove forest ecosystem, gulf of kutch, western coast of india: a synergistic evaluation using remote sensing,” Theoretical and Applied Climatology, vol. 120, no. 3, pp. 685 700, 2015.

[4] G.K Today. https://www.gktoday.in/climate-of-india-during-2021-report/?form=MG0AV3

[5] MD. MEHEDI HASSAN (Member, IEEE), MOHAMMAD ABU TARAQ RONY, MD. ASIF RAKIB KHAN, MD. MAHEDI HASSAN, FARHANA YASMIN, ANINDYA NAG1 (Member, IEEE), TAZRIA HELAL ZARIN, ANUPAM KUMAR BAIRAGI (Senior Member, IEEE), SAMAH ALSHATHRI, WALID EL-SHAFAI. Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness; https://ieeexplore.ieee.org/ielx7/6287639/6514899/10320349.pdf

[6] Adel Rajab, Hira Farman, Noman Islam, Darakhshan Syed, M. A. Elmagzoub, Asadullah Shaikh, Muhammad Akram, and Mesfer Alrizq; Flood Forecasting by Using Machine Learning: A Study Leveraging Historic Climatic Records of Bangladesh, https://www.mdpi.com/2073-4441/15/22/3970

[7] STUDY REPORT KERALA FLOODS OF AUGUST 2018 by Government of India Central Water Commission Hydrological Studies Organisation Hydrology (S) Directorate, https://sdma.kerala.gov.in/wp content/uploads/2020/08/CWC-Report-on-Kerala-Floods.pdf?utm_source=chatgpt.com

[8] NATURAL DISASTERS AND ECONOMIC DYNAMICS: EVIDENCE FROM THE KERALA FLOODS, Robert C. M. Beyer, Abhinav Narayanan, Gogol Mitra Thakur; https://cds.edu/wp-content/uploads/WP508 Dr.Gogol_.pdf?utm_source=chatgpt.com

[9] S. Al Azwari, "Predicting myocardial rupture after acute myocardial in farction in hospitalized patients using machine learning," in 2021 National Computing Colleges Conference (NCCC), pp. 1–6, IEEE, 2021.

[10] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple prepro cessing techniques," TrAC Trends in Analytical Chemistry, vol. 132, p. 116045, 2020.

[11] L. Klebanov and I. Volchenkova, “Outliers and the ostensibly heavy tails,” Mathematical Methods of Statistics, vol. 28, pp. 74–81, 2019.

[12] G. Agapito, C. Zucco, and M. Cannataro, “Covid-warehouse: A data warehouse of italian covid-19, pollution, and climate data,” International Journal of Environmental Research and Public Health, vol. 17, no. 15, p. 5596, 2020.

[13] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” Applied Soft Computing, vol. 97, p. 105524, 2020.

[14] A. Freitas, “Comprehensible classification models: A position paper,” ACMSIGKDD Explorations Newsletter, vol. 15, pp. 1–10, 03 2014.

[15] Kerala Floods – 2018 by State Relief Commissioner, Disaster Management (Additional Chief Secretary) Government of Kerala; https://sdma.kerala.gov.in/wp-content/uploads/2019/08/Memorandum2-Floods 2018.pdf?utm_source=chatgpt.com

[16] Flood Prediction Using Machine Learning Models; https://arxiv.org/pdf/2208.01234

[17] Hussein. E, M. Ghaziasgar, C. Thron : “Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps”, IEEE 23rd International Conference on Information Fusion (FUSION), 2020. https://doi.org/10.23919/FUSION45008.2020.9190285

[18] Samantaraya. S, A. Sahoo, A. Agnihotri : “Prediction of Flood Discharge Using Hybrid PSO-SVM Algorithm in Barak River Basin”, MethodsX Volume 10, 2023. https://doi.org/10.1016/j.mex.2023.102060

[19] Nadia Zehra: "Prediction Analysis of Floods Using Machine Learning Algorithms (NARX & SVM)", International Journal of Sciences: Basic and Applied Research (IJSBAR) (2020) Volume 49, No 2, pp 24-34, https://core.ac.uk/download/pdf/287366682.pdf

[20] Naveed Ahamed, S.Asha : "Flood prediction forecasting using machine Learning Algorithms", International Journal of Scientific & Engineering Research Volume 11, Issue 12, December-2020

[21] Mr. B. Samuel John Peter, Mr. N. Thilak Chandhra, Mr. Shaik.Afrid, Mr. N. Prasanna Kumar: "MACHINE LEARNING BASED FLOOD PREDICTION", International Journal of Research in Engineering, IT and Social Sciences, Volume 12 Issue 12, December 2022

[22] Rainfall Based Flood Prediction in Kerala Using Machine Learning, https://ijisae.org/index.php/IJISAE/article/view/4800

[23] FLOOD RISK MAPPING USING RANDOM FOREST AND SUPPORT VECTOR MACHINE by M.Ganjirad, M.R.Delavar; https://isprs-annals.copernicus.org/articles/X-4-W1-2022/201/2023/isprs-annals-X-4-W1-2022-201 2023.pdf?utm_source=chatgpt.com

[24] Feature scaling | Standardization vs Normalization. (2020, April 3). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machi ne-learning-normalization-standardization/

[25] 1. 10. Decision trees. (n.d.). Scikit-Learn. Retrieved January 16, 2022, from https://scikit-learn/stable/modules/tree.html

[26] Decision Tree Algorithm for a Predictive Model. (n.d.). TechLeer. Retrieved January 16, 2022, from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-p redictive-model/