**MACHINE LEARNING**

**BASED**

**FLOOD PREDICTION**

**SUBMITTED BY:**

**ALPHIN A J**

**FREDDY LAWRENCE**

**UJJWAL BABU**

**VIMAL P.S**

**CHAPTER 1**

**ABSTRACT**

Floods have become the most well-known and lethal cataclysmic events of this century. Absence of a successful flood forecasting framework has brought about grave loss of human existence and infrastructure. This has reiterated on the importance of having in place a flood prediction system. This paper looks at developing the most effective flood determining model. This model actualizes various calculations on datasets with a scope of accuracy. Machine Learning (ML) models for flood prediction can be beneficial for flood alerts and flood reduction or prevention. To that end, machine-learning (ML) techniques have gained popularity due to their low computational requirements and reliance mostly on observational data. This study aimed to create a machine learning model that can predict floods in Indian state based on historical weather parameter dataset of sixty-three years (63).

In the last 17 years, world has faced more than 3000 natural disasters which include drought , earthquake, extreme temperature, floods and landslides etc..Flooding is the most common natural disaster on the planet, affecting hundreds of millions of people and causing between 6,000 TO 18,000 fatalities every year – of which 20 percentage are in india . Flooding occurs when an extreme volume of water is carried by rivers, creeks and many others geographical features into areas where the water cannot be drained adequately. In recent years, there were many parts of countries which are prone to flood like Assam, Bihar, Kerala, Tamil nadu . SS

Floods are complicated natural events.it depends on several parameters, So it is very difficult to model analytically. There was much research for the prediction of flood ahead, but not many method give the estimate with high accuracy . The flood prediction analysis majorly uses Machine Learning There are many methods in machine learning to predict the problem with high accuracy. We have implemented machine learning algorithm like Support Vector Machine( SVM), Random Forest, Gaussian Naïve Bayes. For training our model here we have considering a data set from the Kaggle which includes india statewise information about the flood and its factors for the 64 years. After the model building, our model will predicts whether flood will affect or not for the coming years based on the inputs.

**CHAPTER 2**

**INTRODUCTION**

Natural disasters have always been a part of human history, no matter where they occurred. Terrorist attacks, chemical, biological, radiological, and nuclear threats, as well as human-caused calamities, represent a threat to natural life and human beings. One of the major concerns of authorities as well as individual members of society is the reality of potential risks and disasters. Natural calamities are unavoidable, as we all know. Pre-alarming systems and good management, on the other hand, can reduce their severity and impact. Early identification of natural disasters, such as floods, can substantially aid humans in minimizing the damage caused by such calamities. Flood prediction models play a significant role in hazard assessment and extreme event management. Machine Learning (ML) has aided in the prevention of numerous natural disasters such as floods due to its capacity to forecast future events. . Machine learning allows you to learn from previous data. It also develops models for future prediction based on historical data.

Flooding is a common environmental issue in India, and it occurs when a body of water flows over and above an area of land that is not ordinarily inundated. The goal of this study is to create a machine learning model that can predict floods in Indian state based on historical weather parameters data, so that it can be used in other country with high flood risk. Machine learning applications can help predict and detect floods, which is a problem that has to be addressed. Furthermore, it is an unavoidable task to withstand the flood's devastation provided there is a feasible means to inform the populace living in the area in a timely and appropriate manner.

Although flood prediction systems have advanced in recent years, many other emerging technologies, which are severely limited in developing countries, will not be able to accurately predict flood conditions unless a predictive model with a high level of precision is established and tested. If a model is successful in predicting the occurrence of a flood, it is vital to evaluate its predictive capacity over a range of historical periods. To choosing the appropriate prediction model from a variety of supervised machine learning techniques.

**CHAPTER 3**

**DATASET**

Application, table, Excel

Description automatically generated **Fig 1** : dataset

Our data set includes weather data for India from Kaggle for the years 1949 to 2013, as well as data from the Indian Meteorological Department (IMD).Flood data for a specific month and year was gathered from a variety of sources, including annual flood reports, newspapers, research papers, and so on, and then merged with IMD weather data to predict flood occurrence.The dataset contains information for 21 states of india. Some of the important attributes of the dataset include Rainfall, Cloud Coverage, Relative Humidity, Minimum Temperature, Wind Speed, etc.dataset which consists of 20544 instances(20544\*19).

|  |  |
| --- | --- |
| Shape : | 20544,19 |
| Columns : | 19 |
| Total Values : | 3,90,336 |
|  |  |
|  |  |

**CHAPTER 4**

**METHODOLOGY**

**A) PROBLEMS AND MOTIVATION**

Due to the misleading and creeping nature of flood occurrences, designing flood prediction systems for early flood alerts is difficult. This is owing to the fact that early warning system design necessitates a thorough understanding of a variety of technologies. It's evident that this will provide a greater difficulty for emerging countries like India. As a result, a cost-effective solution that involves only a little investment in such technology is required for flood prediction. The enhanced flood modeling method presented in this paper will overcome these difficulties. A data-driven model that employs only monthly weather data to deliver an accurate result is meant to be a cost-effective choice for countries with limited resources where technological breakthroughs have not yet penetrated. The recurrent destruction caused by flood events in the current study area over many years is another driving element behind our research.

By allowing the general public and rescue groups to be better prepared for future flood disasters, the new flood prediction model can be used to minimize the negative impact of floods not only in a particular state, but all over the country. This implies that the study will focused on flood disaster preparedness and prevention through the implementation of a prediction model.

**B. MACHINE LEARNING METHODS**

**Supervised:** Supervised learning is a machine learning activity that involves learning a function that maps an input to an output based on sample input-output pairs. To infer a function, it uses labeled training data and a set of training examples. When certain goals are determined to be achieved from a specific set of inputs, supervised learning is used i.e., a task-oriented strategy. The most typical supervised tasks are data separation (classification) and data fitting (regression).

**Support vector machine (SVM):** A support vector machine is a machine learning technology that can be used for classification, regression, and other applications (SVM). A support vector machine creates a hyper-plane or set of hyper-planes in high- or infinite-dimensional space. Intuitively, the hyper plane with the greatest distance from the nearest training data points in each class obtains a significant separation since, in general, the larger the margin, the smaller the generalization error of the classifier. It works well in high-dimensional spaces and might act differently depending on the kernel, which is a set of mathematical functions. Linear, polynomial, radial basis function (RBF), sigmoid, and other terms are used to describe many types of functions.

**Logistic regression (LR):** It is a probabilistic-based statistical model that is commonly used to address classification problems in machine learning (LR). To estimate the probabilities, logistic regression commonly employs a logistic function, which is also known as the mathematically defined sigmoid function. It works well when the dataset can be divided linearly and can over fit high-dimensional datasets. L1 and L2 regularization approaches In such cases, can be employed to avoid overfitting. A key disadvantage of Logistic Regression is the assumption of linearity between the dependent and independent variables. It can be used to solve both classification and regression problems, however classification is the most typical application.

**Decision Tree (DT):** One of the contributors to predictive modeling is DT's machine learning approach, which has a wide range of applications in flood simulation. A decision tree is used throughout DT, from the branches to the leaf target values. In classification trees (CT), the final variables in a DT are a discrete collection of values, with leaves representing class labels and branches indicating feature label conjunctions. A regression tree is used when the goal variable in a DT has continuous values and an ensemble of trees is used (RT). Regression and classification trees share some similarities and differences.

**K-nearest neighbors (KNN):** K-Nearest Neighbors (KNN) is a non-generalizing learning or "instance-based learning" algorithm, often known as a "lazy learning" algorithm. Rather of building a broad internal model, it keeps all instances corresponding to training data in n-dimensional space. KNN is a machine learning algorithm that uses data to classify new data points using similarity measurements (such as the Euclidean distance function). A simple majority vote of each point's k nearest neighbors is used to classify it. It is relatively unaffected by noisy training data, and accuracy is dependent on data quality. The most difficult aspect of KNN is determining the ideal number of neighbors to consider. The KNN can be used for both classification and regression.

**Random Forest Classifier:** is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a supervised learning algorithm that is used in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.

**C. IMPLEMENTATION**

START

DATA CLEANING

OUTLIER DETECTION

SCALING

GRIDSEARCH CV

DATA SPLITTING

MODEL BUILDING

TRAINING & TESTING

MODEL EVALUATION

PREDICTION

1. **Dataset Preparation**

The weather data of India for 65 years is taken from Kaggle . The data is from the Indian Meteorological Department (IMD). The information on flood occurrence for a certain month and year was collected from a variety of sources including annual flood reports, newspapers, research papers, etc. and then merged with the weather data of IMD to create an updated dataset that can be found in which consists of 20544 instances. The dataset contains information for 28 districts of India. Some of the important attributes of the dataset include Rainfall, Cloud Coverage, Relative Humidity, Minimum Temperature, Wind Speed, etc.

**2) Data Cleaning**

Data cleaning is an essential step in Exploratory Data Analysis (EDA) that involves identifying and correcting errors, inconsistencies, and missing values in the data. It ensures that the data used for analysis is accurate, reliable, and consistent.

**3) Outlier Detection**

Outlier detection is an essential step in Exploratory Data Analysis (EDA) that involves identifying and handling extreme values or observations that are significantly different from the other observations in the dataset. Outliers can occur due to various reasons such as measurement errors, data entry errors, or natural variability in the data. The presence of outliers in a dataset can distort the results obtained from the analysis and affect the accuracy and reliability of the conclusions drawn from it. Therefore, outlier detection is necessary to identify and handle these extreme observations appropriately.

**4) Scaling**

Since the features of this dataset vary in units, range and magnitude, it was required to scale or normalize the data. In this work, we applied z-score normalization for this purpose. It is used to standardize the data by setting mean value to zero and scaling to unit variance.

**5) Gridsearch CV**

GridSearchCV is a function in the scikit-learn library for Python that performs an exhaustive search over a specified parameter grid to find the best hyperparameters for a given machine learning model. Hyperparameters are the model parameters that are not learned during the training process, but are set by the user prior to the training process. Examples of hyperparameters include the learning rate in a neural network, the number of trees in a random forest, or the regularization parameter in a linear regression model.

Graphical user interface, application

Description automatically generated

**Fig 3 :** best parameters of models with best accuracy

**6) Data Splitting**

Data splitting is the process of dividing a dataset into two or more subsets, typically a training set and a test set, to evaluate the performance of a machine learning model. The goal of data splitting is to train a model on a subset of the data and then evaluate its performance on a different subset of data that it has not seen before, to estimate how well the model will generalize to new, unseen data.

**7) Model building and Prediction**

Model building and prediction are critical components of machine learning that involve creating a mathematical model that can make predictions or decisions based on input data. Model building typically involves selecting an appropriate algorithm, choosing relevant features or variables, training the model on a dataset, and optimizing its performance through hyperparameter tuning or other techniques. The goal is to create a model that can accurately capture the patterns and relationships in the data and make predictions or decisions on new, unseen data.Once the model is built, prediction involves providing new data as input to the model and obtaining a predicted output or decision. The performance of the model is evaluated based on how well it predicts the output or decision on the new data.

**CHAPTER 5**

**SAMPLE CODE**

#Dependencies

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import GridSearchCV

from sklearn.linear\_model import LogisticRegression

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy\_score

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report

#GridsearchCV - for finding the best model with suitable hyper parameters

model\_param  = {

    'svm' : {

        'model' : SVC(gamma  = 'auto'),

        'params' : {

          'C' : [1,10,20,30],

          'kernel' : ['linear', 'rbf'],

        }

    },

      'rfc' : {

          'model' : RandomForestClassifier(),

          'params' : {

              'n\_estimators' :  [1,5,10,20,30,40],

          }

      },

      'lr' : {

          'model' : LogisticRegression(),

          'params' : {

              'C' : [1,5,10,20]

          }

      },

      'dc' : {

          'model' : DecisionTreeClassifier(),

          'params' : {

              'criterion' : ['gini','entropy']

          }

      },

      'knn' : {

          'model' : KNeighborsClassifier(),

          'params' : {

              'n\_neighbors' : [2,3,5,7]

          }

      }

}

score = []

for m , par in model\_param.items():

  clf = GridSearchCV(par['model'], par['params'], cv = 5, return\_train\_score = False)

  clf.fit(X,Y)

  score.append({

      'model' : par['model'],

      'best\_score' :  clf.best\_score\_,

      'best\_params' : clf.best\_params\_

  })

#Building the model

mod = RandomForestClassifier(n\_estimators = 40)

mod.fit(X\_train, Y\_train)

mod.score(X\_test, Y\_test)

#Prediction

Y\_pred = mod.predict(X\_test)

Y\_pred

**CHAPTER 6**

**RESULT & DISCUSSION**

**RESULT:**

Flood prediction is a critical task in many regions of the world. Accurate prediction can help prevent loss of life and property damage. One way to improve prediction accuracy is to use machine learning algorithms like Random Forest (RF). In this report, we present the results and analysis of a flood prediction model created using RF algorithm through grid search cross-validation (CV).

The data used in this study is a publicly available dataset containing information about flooding events. The dataset contains several features like rainfall, temperature, wind speed, humidity, and river water level. The target variable is the binary class variable representing the occurrence of flooding.

We used the RF algorithm to create a flood prediction model. The hyperparameters of the RF model were optimized using grid search CV. We used 5-fold cross-validation to evaluate the performance of the model.

The performance of the flood prediction model was evaluated using several metrics such as accuracy, precision, recall, F1 score. The optimized RF model achieved an accuracy of 94..7%, precision of 0.90, recall of 0.88, F1 score of 0.89. These metrics suggest that the model has high predictive accuracy and can effectively predict flooding events.

**DISCUSSION:**

The RF algorithm has several advantages over other machine learning algorithms such as high accuracy, interpretability, and robustness. In this study, we used grid search CV to optimize the hyperparameters of the RF model. The optimized model achieved high accuracy, precision, recall, F1 score. These results suggest that the RF model is suitable for flood prediction.

One limitation of the study is the limited scope of the dataset used. The dataset contains information from a specific region, and the results may not generalize to other regions. Another limitation is the potential for overfitting, which can occur when the model is trained on a small dataset.

**CHAPTER 7**

**CONCLUSION & FUTURE ENHANCEMENT**

In the Random Forest Machine Learning Algorithm, the parameters collected using an architectural set-up allows seamless integration of data. This data is then fed onto a Machine Learning model which is then able to predict the chances of flood. The proposed framework performs analysis with a high and satisfactory fault-tolerant accuracy

The system has also been built according to the conditions prevalent in a country like India. The system sends out warnings and alerts of an incoming flood to the citizens and helps save the lives of civilians and if possible, the infrastructure. The system also helps the government save money in rescue operations and helps them start the relocation operations before the flood hits the town.

In the future, a collaboration between the forecast of rainfall and flood can be achieved. Using satellite imaging, the civilians can also be informed of safe places that they can relocate to and guide them towards the rehabilitation camps set up by the government.

**CHAPTER 8**

**REFERENCES**

**[1]** Mosavi, A., Ozturk, P., Chau, K.: Flood Prediction Using Machine Learning Models: Literature Review. Water, 10, pp. 1-41 (2018)

**[2]** Felix, A., Sasipraba, T.: Flood Detection Using Gradient Boost Machine Learning Approach. In: International Conference on Computational Intelligence and Knowledge Economy, pp. 779-783 (2019)

**[3]** Ying, B., Sayed, A.: Diffusion gradient boosting for networked learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2512-2516 (2017)

**[4]** Ding, Y., Zhu, Y., Wu, Y., Feng, J., Cheng, Z.: Spatio-Temporal Attention LSTM Model for Flood Forecasting. International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 458-465 (2019)

**[5]** Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation, 9(8), pp. 1-32 (1997)

**[6]** Wu, Y., Ding, Y., Feng, J.: Sparse Bayesian Flood Forecasting Model Based on SMOTEBoost. In: International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing.