Flood Prediction using Machine Learning

Kishan Kashyap M[1], Ajay Karthik K[1], ShivaKumar G.S[2]

[1]Student **-** Department of Computer Science and Engineering, Srinivas Institute of Technology, Karnataka, India.

[2]Professor - Department of Computer Science and Engineering, Srinivas Institute of Technology, Karnataka, India.

---***---

Abstract - *The unusual rainfall and global climate change has led to floods in different parts of the world. Floods are one of the worst affecting natural phenomena which causes heavy damage to property, infrastructure and most importantly human life. To prevent such disasters Machine learning model is created to predict the floods that can occur in the future. It's hard to create a predictive model because of its complexity. In this system the rainfall data is fed into four different Machine Learning models prior to this process, the data is cleaned and preprocessed, the dataset for training is split into Training set and Test set in the ratio of 7:3. Then the accuracy of each model is compared and the confusion matrix parameters are taken to evaluate and analyze. At the end the best model is chosen by comparing the accuracy.*

Key Words*:* Floods, Rainfall, natural disasters.

1. INTRODUCTION

Floods are among the most destructive natural disasters and it causes lots of damage to property and human life. The yearly data shows that the amount of rainfall is increasing and it's due to climate change. Flood is predicted in several locations using some advanced technologies which just helps the people to be prepared for upcoming disasters. It is very difficult to create a predictive model using machine learning. Machine learning gives computers the capability to learn without being explicitly programmed. Machine learning has a role in preventing many natural disasters like earthquakes, floods and many more. Machine learning make decisions using past data and these data are fed into the algorithms and the output is predicted. Machine learning(ML) can be classified into three categories Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is the types of machine learning in which machines are trained using well “labelled” training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. The Supervised learning can be again further divided into two types Regression and Classification. Regression algorithms are used if there is a relationship between the input variable and output variable. It is used for the prediction of continuous variables. Classification algorithms are used when the output variable is categorical which means there are two classes such as Yes-No, Male-Female, True-False. Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead models itself find the hidden patterns and insights from the given data. The Unsupervised learning algorithm can be further categorized into two types: Clustering and Association. Clustering is a method of grouping objects into clusters such that objects with most similarities remain into a group and have less or no similarities with the objects of another group. Association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. The aim of this project is to develop a flood prediction model which is real time. This could be helpful in the areas where the flash flood occurs. This system takes the input as the rainfall data all over the india process it using different machine learning models and the best model is determined with the help of accuracy of different algorithms, which would help people prior, save lives and also save lots of meteorological efforts.

2. EXISTING SYSTEM

Flood forecasting is highly complicated and expensive. The weather and rainfall is a factor of predicting the flood. The advanced technology uses simulations supported by physics and differential equations. The satellite images are used to get the rainfall data. In recent times rapid urbanization, global climate change and extreme rainfall have resulted in flash floods. In orthodox methods of flood forecasting, using satellite images and radar also involving mathematical equations, current weather conditions are detected. At present machine learning technologies are implemented to detect such kinds of natural disasters. The floods are predicted by considering the parameters causing the flash flood. There are some drawbacks in machine learning that lead to wrong predictions of floods. The results cannot be accurate in predicting flash floods.

3. PROPOSED SYSTEM

The aim of this project is to get all the rainfall data of India and from a dataset containing yearly rainfall data. By providing real time input to different models of machine learning, those are Logistic Regression, Support Vector Machine, K-Nearest Neighbors and Decision Tree Classifier. The input provided to models are pre-processed and patterns are extracted by getting maximum accuracy. The data provided is split into a Training set and Test set. It is split in the ratio of 7:3. The all four models are used to predict and by comparing all the results of model and considering the confusion matrix of all the models the accuracy is determined. The best model is chosen by comparing the accuracy of each model.

4. OBJECTIVE

The objective of Flood Prediction using Machine Learning is to design a model to predict the flood using the rainfall data. The prediction of different models is taken and compared within each other to find the best model that has high accuracy. The flood can be predicted in different states of India in different months. The confusion matrix of different models in Machine learning is considered to evaluate the accuracy and precision of the system.

5. SYSTEM IMPLEMENTATION

The prediction accuracy of the different models is evaluated using data validation, and the results are compared to get accuracy. The accuracy of the training dataset, accuracy of the testing dataset, false-positive rate, specification, precision, and recall are calculated by comparing algorithms using python code.

The steps involved are:

- Define a problem
- Preparing data
- Evaluating algorithms
- Predicting results
- Predicting results

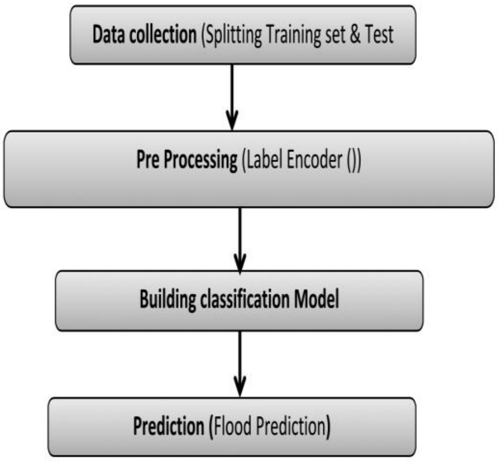
Data collection (Splitting Training set & Test
Pre Processing (Label Encoder ())
Building classification Model
Prediction (Flood Prediction)

Fig 1: System architecture

The goal of cleaning data is to detect and remove errors. The process of transforming data prior to feeding it to the algorithm is called data pre-processing. For better results such a process takes place where raw data is converted to clean data and fed to an algorithm. Some of the machine learning models need data in some format, random forest algorithm does not support null values. Therefore to execute random forest algorithm the raw data has to be transformed into null free data set.

6. ALGORITHMS AND TECHNIQUES

6.1 Logistic Regression

Logistic Regression may be a machine learning algorithm that predicts the probability of a categorical variable. It is a statistical way of analyzing a group of knowledge that comprises quite one experimental variable that determines the result. The outcome is then measured with a dichotomous variable. The goal of this algorithm is to seek out the simplest model to explain the connection between a dichotomous characteristic of interest and a group of independent variables. In this algorithm, the dependent variable is a binary variable that contains data coded as 1 or 0. In other words, the logistic regression model predicts P(Y=1) as a function of X.

6.2 Support Vector Machines

SVM uses a classifier that categorizes the info set by setting an optimal hyperplane between data. This classifier is chosen as it is incredibly versatile in the number of different kernel functions that can be applied, and this model can yield a high predictability rate. Support Vector Machine is one among the foremost popular and widely used clustering algorithms. It belongs to a gaggle of generalized linear classifiers and is taken into account as an extension of the perceptron. It was developed in the 1990s and continues to be the desired method for a high-performance algorithm with a little tuning.

6.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one among the supervised machine learning algorithms that stores all instances like training data points in an n-dimensional space. For real-valued data, the algorithm returns the mean of k nearest neighbors, and in case of receiving unknown discrete data, it analyses the closest k number of instances that is saved and returns the most common class as the result of the prediction. In the distance-weighted nearest neighbor algorithm, the contribution of each of the k neighbors is weighed according to their distance, giving higher weight to the closest neighbors. The K-Nearest Neighbor algorithm is a classification algorithm and is robust to noisy data as it averages the k-nearest neighbors. The algorithm first takes a bunch of labeled points and analyses them to find out the way to label the opposite points. Hence, to label a new point, it looks at the closest labeled points to that new point and has those neighbors vote, so whichever label most of the neighbors have is the label for the new point. This algorithm makes predictions about the validation set using the whole training set. Only by rummaging through the whole training set to seek out the closest instances, the new instance is predicted. Closeness is a value that is determined using a proximity measurement across all the features involved.

6.4 Performance Analysis Metrics

- **True Positive:** It is an outcome where the model correctly predicts the positive class. The outcome is considered as true positive when the system can correctly predict that an incident has indeed occurred.
- **True Negative:** It is an outcome where the model correctly predicts the negative class. The outcome is considered as true negative when the system can correctly predict that the particular incident has not occurred.
- **False Positive:** False Positive is an accuracy measure where the model mis-predicts the positive class. The outcome is considered as False Positive when the system cannot correctly predict that the particular incident has occurred.
- **False Negative:** False Negative is an accuracy value where the model mis-predicts the negative class. The outcome is considered as False Negative when the system cannot correctly predict that the particular incident has not occurred.
- **Sensitivity:** Sensitivity is a measure of the proportion of true positive values, that is, the actual number of positive cases that are correctly predicted as positive. It is also known as Recall value. There exists another proportion of actual positive cases that are mis-predicted, which can be represented in the form of a false negative rate. Therefore, the sum of sensitivity and false-negative rate value is 1.

Mathematically sensitivity can be calculated as:

Sensitivity/Recall = (True Positive) / (True Positive + False Negative)

The higher value of sensitivity would mean a higher value of the true positive and lower value of false negative. The lower value of sensitivity would mean a lower value of the true positive and higher value of false negative.

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is calculated by dividing the number of correctly predicted positive observations by the total number of predicted positive observations. High precision relates to the low false-positive rate.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model correctly predicts)

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is calculated by dividing the number of correctly predicted observations to the total number of observations in an actual class.

F1 Score: F1 Score is defined as the weighted average of Precision and Recall values. Hence both false negative and false positives values are taken into account. When there's an uneven class distribution, the F1 Score value is generally more useful when compared to Accuracy value. On the other hand, Accuracy value worksheets best when the values of false positive and false negatives have a similar cost. If both values are different, then Precision and Recall values are taken into account.

GENERAL FORMULA:

F- Measure = 2TP / (2TP + FP + FN)

F1-SCORE FORMULA :

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

7. RESULTS AND DISCUSSION

ML Models	Precision	Recall	F1-Score	Sensitivity	Specificity	Accuracy (%)
Logistic Regression	0.95	0.98	0.96	0.96	0.99	99.39
Support Vector Machine	0.93	0.88	0.90	0.88	0.99	98.37
K-Nearest Neighbours	0.9	0.81	0.85	0.81	0.99	97.47
Decision Tree Classifier	0.72	0.75	0.73	0.75	0.97	95.07

Table 1: Comparison of Accuracy Results

Table 1 compares the accuracy and precision results of the Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree Classifier and Random Forest Classifier algorithms. From the above calculated values we can observe that Support Vector Machine and Logistic Regression has comparatively better results.

Algorithm	TP	TN	FP	FN	TPR	TNR	FPR	FNR	PPV	NPV
LR	76	752	4	1	0.98	0.99	0.01	0.02	0.95	0.99
SVC	68	751	5	9	0.88	0.99	0.01	0.12	0.93	0.98
KN - N	63	749	7	14	0.81	0.99	0.01	0.19	0.9	0.98
DTC	58	734	22	19	0.75	0.97	0.03	0.25	0.72	0.97

Table 2: Comparison of Confusion Matrix Parameters

The above table shows the comparison of confusion matrix parameters of Logistic Regression, Support Vector Classifier, K-Nearest Neighbors, Decision Tree Classifier, Random Forest Classifier algorithms. From the Calculated result we can conclude that Logistic Regression has produced best results.

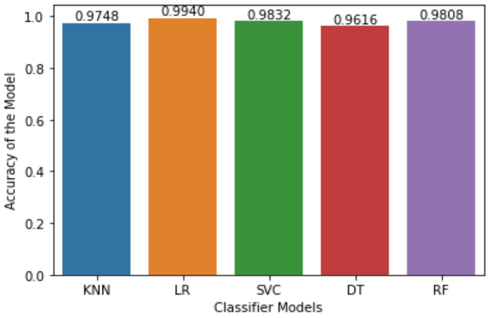

Fig 1: Accuracy vs Algorithms

8. CONCLUSIONS

The machine learning system ignited by data cleaning and processing, replacing or removing the null values, model building and evaluation. At the end the flood prediction model has given different accuracy results from four different models. From the above results and analysis, the best algorithm for flood prediction is Logistic Regression with (99%).

REFERENCES

[1] An Innovative Flood Prediction System Using Improved Machine Learning Approach (csfjournal.com)

[2] https://encyclopedia.pub/73

[3] PyQt5 Reference Guide — PyQt v5.15 Reference Guide (riverbankcomputing.com)

[4] Requests: HTTP for Humans™ — Requests 2.25.1 documentation (python-requests.org)

[5] Tutorials — Matplotlib 3.4.2 documentation

[6] https://www.irjet.net/archives/V7/i5/IRJET-V7I51189.pdf

[7] Phyo Pa Pa Tun, Myint Myint Sein, “Flood Prediction System for Middle Region of Myanmar”, 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE 2018).