Comparative Study Of Machine Learning Algorithms for Flood Detection

KIRAN DOMINIC
Master of Computer Applications.
Amal Jyothi College of Engineering.
Kanjirapally,India
kirandominic2022@mca.ajce.in

RINI KURIAN
Assistant professor in Computer Applications.
Amal Jyothi College of Engineering.
Kanjirapally,India
rinikurian@amaljyothi.ac.in

Abstract**—Flood is the hike of water level of water streams and bodies. The rise of water is temporary and may last for hours or in extreme cases last for weeks. This is caused by heavy rain coastal rains and situations like opening of dams etc. Flood can happen quickly slowly or quickly or can happen without any warnings. Flood kills and make more damage than other natural disasters. Floods are so powerful that a height of one foot can pound a person. The escape and recovery become difficult by the risen water. Water makes great damage and difficulties in transportation facilities which arduous the backing of personals and commodities. Damages in transport facilities may cause isolation of areas. The response time is very crucial in managing and recovery. The prediction of floods can massively decrease the damage. This paper discusses about the prediction of floods using different machine learning algorithms and developing efficacious and operative Computer Application that alerts the people which affects the flood.**

Keywords—Machine Learning, Flood detection, Decision tree algorithm, Logistic Regression, Random Forest, Ensemble Learning, KNN

Introduction

Flood is a common natural disaster in many regions of the globe. The sudden rise in the water level causes the water to flow to the land. The land usually dry and rising water makes damage very quickly. The increased flow can damage bridges and roads which increases the severity of the natural disaster. Sudden rain caused by low pressure areas in ocean which causes heavy rain which may lead to floods in some geographical regions. Kerala, the southern part of India faced this disaster in sequence for the last couple of years. In 2018, more than 400 people died when heavy rains flooded the state.[1] Floods can cause collateral damage very quickly. The flood water washes away the homes and other belongings in no time. The time for a taking an action to manage the flood is very crucial in escape and recovery. The early warning systems are implemented. However, warnings based only on natural hazard monitoring do not offer sufficient protection[2] Predicting the chance of occurrence of floods can save lives and prevent loss of commodities. Predicting the occurrences early can bring valuable time in managing the disaster. Advancement in the field of artificial intelligence and use of appropriate algorithms in machine learning, which is the branch of artificial intelligence makes the prediction of floods more accurate and easier. These techniques can improve over time. These systems learn from the past data and apply those on the future predictions and make them more accurate. The data for analyzing and predicting can be collected from different sources. The measure of rain, height of the water level can be collected using rain gauges and other sensing devices. Other parameters like the slope of the terrain, elevation, stream power index topographic wetness index is also influential in the results. The delay in alerting the public and its effectiveness is the setback of early waning systems. This problem can be overcome using an application platform to alert the public. This can also be used very effectively using the global positioning systems which are very common in the handheld devices like mobile phones and tablets. Using global positioning system alerts can be send to the target people, police force, emergency response teams etc.

I. Machine Learning

Machine learning is a collection of methods that enable computers to automate data-driven model building and programming through a systematic discovery of statistically significant patterns in the available data[3]. It is the branch of artificial intelligence. These models improvise in future using the past data. These systems can intellectually predict using data models. It can also take decision on situations

DOI: 10.5281/zenodo.6369640
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

which are not foreseen in the learning process. The prediction and decision are almost automated and these systems needs very less human interaction. Advanced analyzing power and better sources of collecting huge data which is substantial for the decision making and the computing resources to analyze and collected data makes the process easier. The important aspect of machine learning is that the iterative aspect. When different data models are given to the models they evolve independently. Previous computations performed can be used to learn more reliable reputable results. In recent days machine learning gained more pace. There are various fields where machine learning is applied effectively. Weather predictions use the help of machine learning to make the predictions. With the advancements in the field of machine learning methodology prediction of flood can be made possible by less effort and very less human interaction. Advancements in the monitoring of rain and water level detection could be used to reduce the impact of sudden floods. There are two types of machine learning techniques can be used to make the prediction.

1. Supervised learning.
2. Unsupervised learning.

In Supervised Learning, an algorithm is made to learn to map an input to a certain output. That is done on labelled datasets that have been collected over the course of time. The algorithm is learned successfully if the mapping is done correct. If the mapping is not done correct then required alterations can be done to the algorithm in order to correctly learn. Trained data models made with Supervised Learning algorithms can be used for predictions on new data that can be collected in the future. Supervised Learning gives experience to the algorithm which can be used to predict the outputs for new unseen data Experience helps in optimizing the performance of the algorithm. Supervised Learning is classified into 2 types — Regression and Classification. Regression is when the algorithm learns from the labelled datasets and then a continuous-valued output is predicted for the newly given dataset. It is used whenever a number is required as on output. Linear regression and Logistic regression are examples of regression algorithm.

Classification is the type of learning where the algorithm maps new data to any one of the two classes in our dataset. The classes can be 1 or 0 and ‘Yes’ or ‘NO’. Decision Tree, Naive Bayes Classifier, Support Vector Machines etc are examples of Classification algorithm.

In unsupervised learning the data consist of values without any labe1s and the output is not pre-determined. The model predicts on the basis of self-learning. The main purpose of these mode k is to predict, Classify, detect, segmentation and The most common use of machine learning is analysis, recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

II. Motivation

The climate changes cause unpredictable natural events some of which are disastrous. The unexpected rainfall causes accumulation of water that normal nature of the rivers and water bodies cannot handle. This leads to the flash floods and flood causes severe damage to the natural ecosystem and also taking tolls on the human lives as well as animal life. Predicting and alerting the people in the flood raising areas can help to minimize the impact of the disaster. The response team and the government bodies can communicate to the public via the platform.

I. PROPOSED METHODOLOGY

The purpose of this paper is f a Machine Learning mode1 which can predict whether flood can happen and suggesting an application to alert the people in the regions that might affect the flood. The model should be able to anticipate when the flood will occur. To make the forecast, data must be gathered. The model is made using monitoring rain gauge. The datasets will be bifurcated into training dataset and testing data sets. A dataset with limited data cannot provide accuracy with learning so we need to train the mode1wdh more data in the dataset. The results obtained from this algorithm mode 1 can be analysed to create a flood prediction. If the prediction model can detect a flash flood the value will be set to 1 else the value will be 0. The algorithm used for the prediction is decision tree algorithm (DT). DTs are classified as fast algorithms; they became very popular in ensemble forms to model and predict floods.

. The following will be the structure of this research:

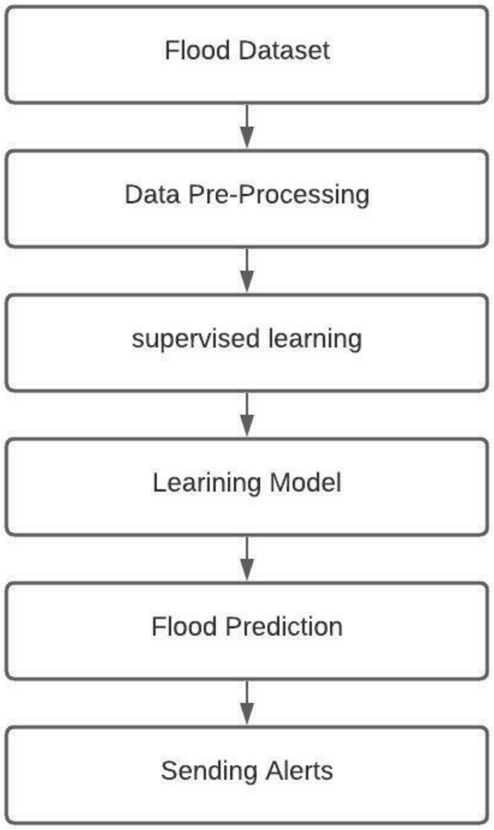

Figure1. Flood prediction methodology.

The steps to follow are

1.Data Collection

2. Defining data

3.Pre-Processing.

4.Building Model

5. Analysis

6.Results.

With the algorithm we follow the be low steps

1. Import the libraries
2. Import the dataset
3. Define the dataset
4. Dataset training and testing
5. Algorithm execution
6. Results comparison and evaluation

A. Table 1: Dataset Description

S No.	Attributes
1	Subdivision
2	year
3	January
4	February
5	March
6	April
7	May
8	June
9	July
10	August
11	September
12	October
13	November
14	December
15	Annual Rainfall
16	Floods

The next attribute is the class variable for each of the attribute in the dataset. This attribute represents the values 0 and 1 which Represents weather to send the alert message or not.

Data Processing: This is the most vital process in the whole machine learning methodology. The missing data and the impurities in the dataset can reduce the quality of the output that can be generated from the dataset. Data preprocessing can be performed in order to increase the effectiveness of the data that is received after the data processing technique. On our dataset we can perform data preprocessing by following the be low methods

1).Removing the missing values — All the entries that have 0 as an entry needs to be removed. Having) is not a valid instance. So all the entries with 0 needs to be removed. We create a feature subset by removing all the irrelevant entries.

2).Data Splitting —After cleaning the data, it must be normalised for both the training and test modes1. The training data set is used to train the algorithm after the split data is ready. Following the training process, a training mode1 is constructed based on the aspects of the training.

Machine Learning: Once the data is ready we exercise Machine Learning process. We apply various classification and algorithms to prognosticate diabetes prone patients. The performance of these methods are probed to find the inaccuracy and identify the major features which can help us in our augury of diabetes. The following techniques can be used —

1). Random Forest — It is a machine learning algorithm that is used to classify and predict data. In comparison to other models, it gets the highest accuracy score. This approach can easily handle any data set of any size. By minimizing variance, the Random Forest tree technique can substantially increase the performance of the Dec is ion Tree algorithm. It works by training and outputting the mode of the classes or the regression of the discrete trees while constructing a slew of decision trees.
Algorithm:

1. Select 'R' features from 'M', total features where R<<M.
2. Find the best node from R features.
3. Split the node into sub nodes using the best split method.
4. Repeat the above steps until '1' number of nodes is reached.
5. Build a forest by repeating the procedures above a number of times until you have "n" trees.

2). Decision Tree — It is a supervised learning method which is also a basic classification method. It is used to categorize the response variable. It has a tree like structure. It describes classification process based on the input features.

B. Algorithm:

1. With the nodes as input feature construct a tree
2. Select the input feature with the highest information gain to predict the output.
3. For each feature in each node of the tree the highest information gain is calculated.
4. Repeat step 2 to form a subtree using the feature which is not used in the above node.

3).K-Nearest Neighbor It is also known as the KNN algorithm. It is a supervised machine learning algorithm. It works on the principle that items of same attributes stay near to each other. The idea of similarity measure is used to group a new work. It notes the records and categorize them on the basis of the similarity measure. The algorithm finds the nearest data points in the training dataset to prognosticate a new data point. K is the positive number of nearest neighbors. The distance between the neighbors is defined in Euclidean distance. Between two points P and Q, the Euclidean distance is defined as:

1. A sample dataset from the Flood data set is taken.
2. A test data set of attributes and rows is taken.
3. The Euclidean distance is calculated using the formula below.

$$d(p,q) = d(q,p)$$

$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$

4. A random value of K is selected.
5. The nth column of each neighbor is found using the Euclidean distance and the minimum distance.
6. The output values is found out.

If the values are the same ,then flood is detected if not no chance for flood.

4). Logistic Regression

Logistic regression works very similar to linear regression, but with a binomial response variable. The greatest advantage when compared to Mantel-Haenszel OR is the fact that you can use continuous explanatory variables and it is easier to handle more than two explanatory variables simultaneously.[4]

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.

5). Ensemble Learning

An ensemble-based system is obtained by combining diverse models (henceforth classifiers). Therefore, such systems are also known as multiple classifier systems, or just ensemble systems [5]

BUILD THE MODEL

The model building is an important stage in the prediction of flood In this spell we implement the algorithms we have discussed above.

Procedure —

1. Necessary libraries are imported
2. The flood dataset is imported
3. Missing data is removed with the pre-process of the data
4. 4. In a 4:1 ratio, divide the dataset into training and test data sets.
5. Select any of the algorithms, Random Forest, Decision Tree, Logical Regression, Ensemble Learning.
6. Based upon the selected algorithm using the training set build a classifier mode1
7. Evaluate the classifier mode1 on the test set
8. Using the performance values received for each classifier conduct a comparison evaluation

9. Find the algorithm with the best performance based on the obtained performance values.

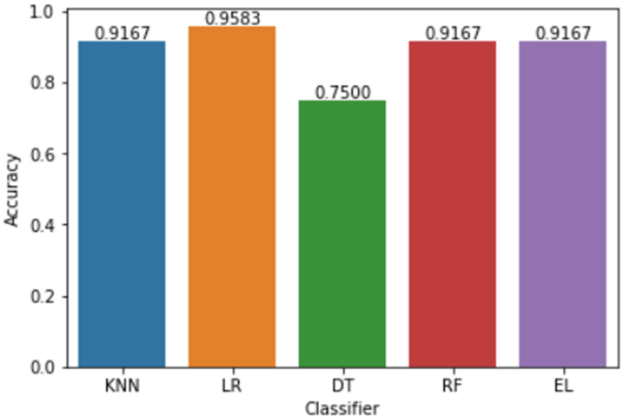

Figure plotting the accuracy obtained by different model on the flood dataset.

VI. RESULTS

From the algorithms compared Logical regression has the most accuracy. Hence helps the system to predict more accurate results. After comparing the output of the algorithms selected Logical regression is apt for the flood dataset. Random Forest, KNN, Ensemble Learning also acquired higher accuracy.

VII. CONCLUSION

With the advancements in the field of machine learning methodology prediction of flood can be made possible by less effort and very less human interaction. And this data along with right algorithms can save lives. An easily reachable alert system can spread the occurrence of flood and can inform public which thereby increases the chance of reacting faster and survival. The algorithms with higher accuracy can be used in machine learning to predict the occurrence of flood and can be used to alert people. These techniques can reduce the impact of natural calamities and can save lives.

.

[1] " Kerala floods: At least 26 killed as rescuers step up efforts"BBC News,October 18, 2021, https://www.bbc.com/news/world-asia-india-58940880.

[2] Shoyama, K., Cui, Q., Hanashima, M., Sano, H., & Usuda, Y. (2021). Emergency flood detection using multiple information sources: Science of The Total Environment, 767, 144371. doi:10.1016/j.scitotenv.2020.144

[3] Bhavsar, Parth (2017). Data Analytics for Intelligent Transportation Systems || Machine Learning in Transportation Data Analytics. , (), 283–307. doi:10.1016/B978-0-12-809715-1.00012-2

[4] S. Sperandei
Understanding logistic regression analysis
Biochemia Medica, 24 (2014), pp. 12-18, 10.11613/bm.2014.003
[5] Robi Polikar (2009) Ensemble learning. Scholarpedia, 4(1):2776
.' doi:10.4249/scholarpedia.2776'