Flood prediction forecasting using machine Learning Algorithms

Naveed Ahamed, S.Asha
Vellore Insitute of Technology,Chennai, India
asha.s@vit.ac.in

Abstract

Floods are very harmful for nature, which are very complex to model. The flood prediction model will give risk reduction & it minimizes the future loss of human life. On 18 May 2016 a south Indian state Kerala was affected by flood. Machine learning is a method which provides intelligence to predict the result in future. The performance comparison of ML models is based on the speed, time and accuracy of the result. There exist a lot of machine algorithms which generate models with more accuracy. For flood prediction classification algorithms like decision tree and linear regression are used in this research. This paper will present the dataset of Kerala flood 2016 which is provided by government.

Keywords: *Machine learning Algorithms, Predictive Analytics, Flood Forecasting.*

1. Introduction

Machine learning provides capabilities to learn from past data. Also based on past data it generates models for future prediction. This technique will be very useful for flood prediction. Earlier we need to give instruction to system for generating output and result. But now with the help of machine learning technique it generates models and gives result itself. Most of the work related to machine learning for flood either predict and helps to make precaution measure and suggest future flood. Kerala, one of the Southern state of India, experienced once in a century flood. There was high damage to the life and property. This motivated us to do study on this pattern of rainfall in the state of Kerala.

2. Machine Learning Algorithms

In machine learning algorithm, two variables like x and y to are used. Variable x in function (f) is mapped to an output variable (Y): Y = f(X). The various algorithms used for prediction are discussed below.

2.1 Logistic Regression

When the nature of the dependent variable is binary logistic regression is used. It is used to solve binary classification problems. This algorithm is best compared to linear regression algorithm because predicting the value of binary variable linear regression was not suitable. It will predict values only in particular range like outside 0 and 1. In logistic regression linear relationship doesn't required. There is no multi collinearity in this algorithm.

2.2 Linear Regression

Linear regression algorithm is the most frequently used model for predictive analysis. Therefore, in regression simple regression to analysis the relationship between the dependent(y) variables and the independent variables to predict the accuracy outcomes and represent it in a statistical graphical representation.

2.3 Decision Tree

The final predicting model is decision tree. Generally, the decision tree is a predicting tool which split the data continuously according to the given certain data parameters. It is a type of supervised learning where a non-parametric method is approached for regression and classification problems. The model first targets the variables to predict value of the variables from the given data by analyzing the decision rules. By this way the accuracy and the output are determined by this decision model.

The following python concepts are used to predict the rainfall.

- Sklearn
  The Sklearn is a library for python which feature algorithm like SVM, Random forest etc for machine learning analysis. It is used to build models.

- NumPy
  In python we used NumPy library for scientific computing. It is a core library which provides tools and high performance for a given array objects.

- Panda
  Panda library is a open source library that is used to make analysis of data and to use easily. It provides high performance and easy to use data structure.

- Matplotlib
  Matplotlib is a python library used to create plots and graphs. It provides variety of bar charts, histogram and error charts

- Seaborn
  Seaborn library is based on data visualization like matplotlib. It's used to represent statistical plotting.

3. Proposed system architechture

Figure 1 shows the proposed system architecture. Labeled data set are used for training. The features are extracted from the training data set and the set of features are given as input to the machine learning algorithm. Figure 2 shows the steps involved in the prediction model.

3.1 Report

A report is a specific form of describing and identifying and examining issues arise in case of an event. Such an event may be described as high rainfall and change in weather pattern which may increase the chance of flood occurrence.

3.2 Analysis

A report in ML defines the accuracy of the data. The data are examined and evaluated by breaking its variables to uncover their dependencies. These help to understand better about the data.

3.3 Monitor

The data are constantly measured, and the performance is monitored to provide the accuracy of the output from the given dataset.

3.4 Prediction

The obtained historical dataset is trained, and an algorithm is applied to obtain an output. The forecast method is used for the prediction analysis.

3.5 Simulate

The simulation in machine learning helps us to forecast the changes that have never happened before and to obtain scenarios outside the historical bounds.

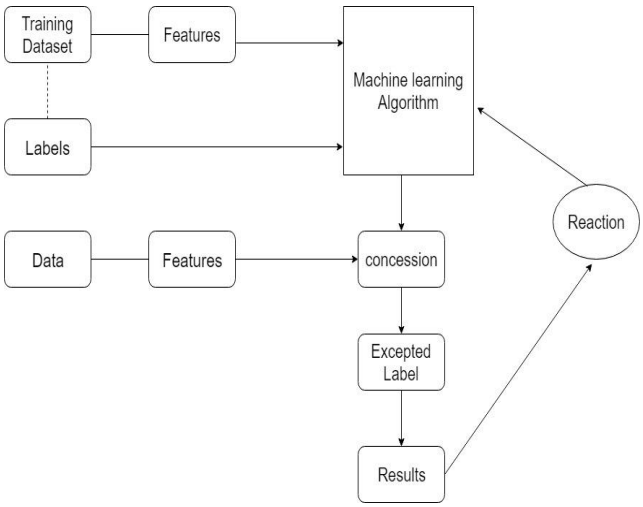

Fig.1: Proposed rainfall prediction model architecture

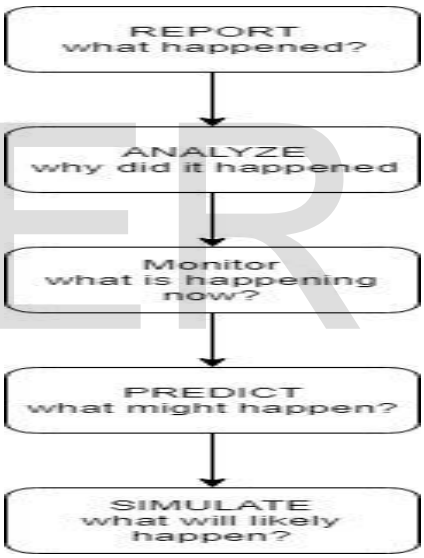

Fig.2: Prediction Methodology

4. Conclusion and Results

Figure 3 shows the rainfall data for 3 months from June to September for the year 2016. Figure 4 shows the sample rainfall dataset. Figure 5 shows the histogram for the dataset. Figure 6 shows the distance plot for the data. Figure 7 shows the logistic regression binary pattern for the 3 months data, and figure 8 shows the coefficients of the linear regression model. Figure 9 gives the accuracy of the decision tree prediction model.

In this paper, the simple machine learning algorithm to predict the accuracy of the flood occurrence is implemented. The desired algorithm shows the results of occurrence of flood in the upcoming year. When compared with the other algorithms, the decision tree algorithm gives more accurate results and provide high performance accuracy and easy to understand. The decision tree also generate model for nonlinear dataset. This nonlinear can be applied to find the accuracy of linear or logistic dataset. As the compared results shows that the decision tree gives more accuracy compared to other simple machine learning algorithm.

As the gathered dataset can provide huge volume of variables it can't be implemented in a simple machine learning algorithm. For a huge amount of data set it can be implemented in neural network which will provide more accuracy and output of the provided dataset. As the neural network uses fuzzy state machine act it can produce multiple results with different probabilities. It can provide historical dataset with more mutable and adaptable form.

```
x=pd.read_csv("kerala.csv")
y=pd.read_csv("kerala.csv")

y1=list(x["YEAR"])
x1=list(x["Jun-Sep"])
z1=list(x["JUN"])
w1=list(x["MAY"])
x2=list(x["Mar-May"])
x3=list(x["Jan-Feb"])
plt.plot(y1, x1,'*',color="Red")
plt.plot(y1, z1,'*',color="Green")
plt.plot(y1, w1,'*',color="Blue")
plt.show()
```

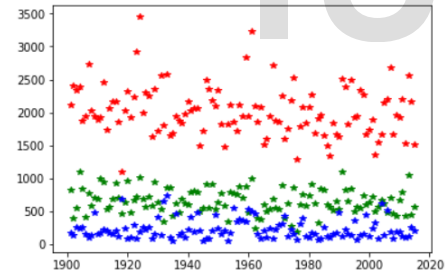

Fig. 3: Data for 3 months

```
In [1]: import pandas as pd
        from sklearn.metrics import accuracy_score
        from sklearn.model_selection import train_test_split
        from sklearn.tree import DecisionTreeClassifier

In [2]: data = pd.read_csv('./kerala.csv')

In [3]: data.columns

Out[3]: Index(['SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL',
               'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'ANNUAL', 'Jan-Feb', 'Mar-May',
               'Jun-Sep', 'Oct-Dec'],
              dtype='object')

In [4]: data.head()

Out[4]:
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	73.4	386.2	2122.8	666.1
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	9.3	275.7	2403.4	638.2
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	21.7	336.3	2343.0	570.1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	26.7	339.4	2398.2	365.3
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	23.4	378.5	1881.5	458.1

Fig: 4 Sample Datasets

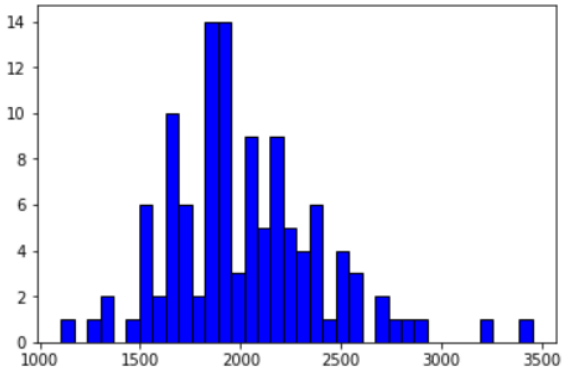

Fig: 5 Histogram June2016 – September 2016

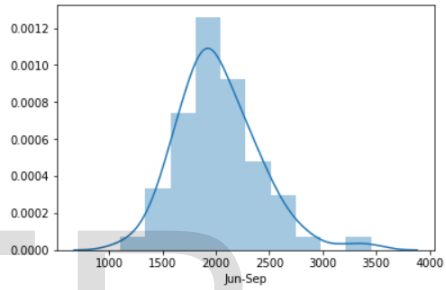

Fig: 6 Distplot June 2016 – Sep2016

```
In [249]: y1

Out[249]: array(['0', '1', '0', '0', '0', '0', '1', '0', '0', '0', '0', '1', '0',
                 '0', '0', '0', '0', '0', '0', '0', '0', '0', '1', '1', '0', '0',
                 '0', '0', '0', '0', '1', '0', '1', '0', '0', '0', '0', '0', '0',
                 '0', '0', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0',
                 '0', '0', '0', '0', '0', '0', '1', '0', '1', '0', '0', '0', '0',
                 '0', '0', '1', '0', '0', '0', '0', '0', '0', '1', '0', '0', '0',
                 '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '1',
                 '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0',
                 '0', '0', '1', '0', '0', '0', '0', '0', '1', '0', '0'],
                dtype=object)
```

Fig.7: Logistic Regression Binary Prediction

```
: cofficient

: array([-0.07014037,  0.08356004,  0.10655933, -0.09041044, -0.13188365,
         -0.30738758,  0.62613572,  0.3150817 ,  0.84194645,  0.22037796,
         -0.39360273, -0.24616949, -0.13986746,  0.46095697])
```

Fig: 8: Coefficients

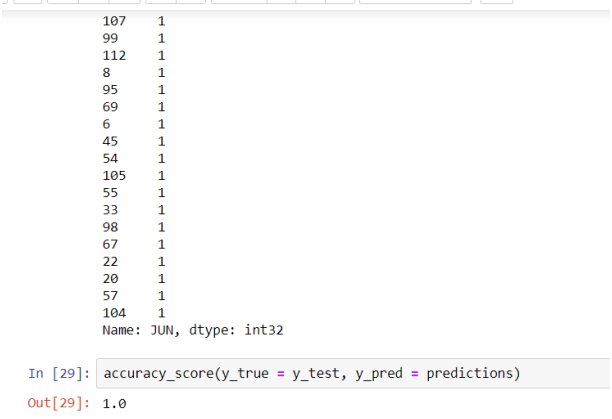

Fig: 9: Decision tree prediction and accuracy result

REFERENCES

1. *E. Toth*, A. Brath, A. Montanari , "Comparison of short-term rainfall predection models for real time flood forecasting".*
2. *BAXTER E. VIEUX, " Estimation of Rainfall for Flood Prediction from WSR-88D Reflectivity:* A Case Study".
3. Hsu, K., Gupta, H.V., Sorooshian, S., 1995, "Artificial neural network modeling of the rainfall–runoff process. Water Resource" Res. 31 (10), 2517–2530.
4. Zealand, C.M., Burn, D.H., Simonovic, S.P., J. Hydrol 1999, " Short term streamflow forecasting using artificial neural networks.. 214, 32–48.
5. .Rodriguez-Iturbe, I., Gupta, V.K., Waymire, E.C., 1984, "Scale considerations in the modeling of temporal rainfall. Water Resource",Res. 20 (11), 1611–1619.