Flood Forecasting by Using Machine Learning

M.Datta charan
MCA ,Chaitanya Bharathi Institute of Technology
Hyderabad,Telangana State ,India

***Abstract*—**This research paper conducts a comparative analysis of machine learning algorithms, such as logistic regression, support vector machines (SVM), and k-nearest neighbors (KNN), for flood forecasting using historical weather and rainfall data. By evaluating their performance and accuracy in predicting flood occurrences, the study aims to offer insights that can enhance proactive flood management and disaster mitigation efforts. These findings are crucial for informing timely actions to minimize the impact of floods.

***Keywords*—**KNN, SVM, LOGISTIC REGRESSION

INTRODUCTION:

Flood Forecasting can be defined as a process of estimation of time and duration based on topographical characteristics of any river basin which reduces the hazards to human life and environment also. Flood Forecasting technique challenge to predict occurrence and magnitude with time of flash flood. Floods happened due to continued precipitation with respective time. Ordinary rainfall also contributes to transforming with time into deadly floods. Flood forecasting techniques play a vital and important role to mitigate the hazards for non-structural structures with cost- effective management. Flood forecasting stations cover the network of flood prone areas to give flood warning to administration. Forecasting inflow is also used for the operation of hydraulic structures such as dams on which opening and closing of gates on spillways. Flood forecasting techniques and Flood warning systems require different types of flood architectures of flood. Flood may be reduced by constructing structures such as dams, weirs, dykes but cannot eliminate the risk. Flood forecasting techniques are able to mitigate the hazards for population and environment in real time with an early warning [1]. Flood forecasting has been approached through rainfall – runoff and flood routing models. Flood forecasts predict inflow at selected locations with HFL value at selected locations of river with time depending on watershed or catchment area. Later, the downstream side predicts the flood will be limited with travel time with an assessment of uncertainties to properly support the decision-makers activities [9]. Flood forecasting using machine learning algorithm (MLA) method to learn and improve system scale to mitigate flood hazards according to the climate change with the help of AI. Flood forecasting used for creating the machine learning algorithm with past and real records of flood with real time data using rain gauges for different coming back time. The dataset sources are rainfall-runoff,water levels using automatic rain gauges with satellites technology, infiltration rate etc.

II. PROPOSED METHOD

The aim of this proposed paper is to investigate whether Binary Logistic Regression can achieve higher accuracy rates and reduce errors. The dataset comprises monthly rainfall index records for India, alongside yearly flood occurrence data, serving as input for precise predictions.

A. Dataset

The data was collected from India Meteorological Department, India, who are responsible for monitoring and issuing forecasts of all natural disasters to keep casualties to a minimum. They use rainfall data,satellite images and various other parameters to issue accurate weather forecasts.

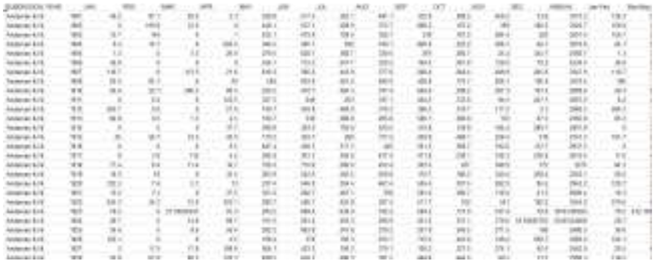


Fig. 1. Rainfall Dataset

Data Cleaning: Firstly, in the dataset, unequal days of each month were handled. For example, some months have 31 days and some have 28 or 30 days. So, there was no data of rainfall for those specific days. Additionally, some days of particular stations did not have any data. Therefore, data imputation was applied to handle this issue. For instance, in the month of February,1980, the month ended at the 29th date, so for the 30th and 31st dates, the value zero for rainfall was added. Furthermore, in the dataset, the rainfall value for the 14th and 15th of 1983 November, delhi station, was missing, so, the value zero was also added here. Feature Engineering: From the rainfall dataset, monthly rainfall data has been calculated and added into a new column. Then, the monthly rainfall data was set according to the particular stations and years. At this stage, the dataset contains the column of stations, years and all the 12 months. After that, the flood data which was collected, was merged as a new feature into the rainfall data according to the stations and years.

Feature Encoding: The dataset that is used in this research paper has two attributes that have string type data which are - ‘Station’ and ‘Flood’. As machine learning models give better results for numerical values, the string type data are encoded. The attribute ‘Station’ contains all the names of the stations from where the daily rainfall data has been collected. Then by performing label encoding, the categorical values of the ‘Station’ column have been transformed into numerical values without adding any additional column. In this dataset, the attribute ‘Flood’ has two unique values, ‘YES’ and ‘NO’ and to encode these values, binary encoding is used. After encoding the values of the feature ‘Flood’, the value ‘YES’ is replaced by 1 and the value ‘NO’ is replaced by 0.

Feature Scaling: Standard Scaler has been used on the dataset to make it unbiased and relevant to the models used. The data is scaled by centering them around the mean with a unit standard deviation. The formula for standardization can be defined as:

$$X = \frac{X - \mu}{\sigma}$$

where, μ is the mean of the feature values and σ is the standard deviation of the feature values [10]. There are no restrictions on the range of the values. The dataset has been split into a train and test set with a ratio of 80:20. Then, the features have been scaled using the standard scaler.

B. Machine Learning Models

Binary Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) have all been used to predict values from the training set.

Binary Logistic Regression: The Binary Logistic Regression, like all other regressions analyses, is a predictive analysis that is used to describe data and explain the relationship between one dependent binary variable and one or more independent variables. The dependent variable has two categories, generally which are 1 for the occurrence of an event and 0 for its absence. A Logistic Regression can be interpreted as a specific case of generalized linear models with a dichotomous dependent variable [11]. A classical linear model can be denoted in the following manner:

$$Y = \alpha + \beta\ X + \varepsilon$$

where Y is the dependent variable, α Y is the intercept when X is equal to zero, X is the independent variable, β is the regression coefficient representing the variation in Y due to the change in values of X ε and is the error of the model. To categorize or limit the range of values for the dependent variable, the logistic regression model fits best. A graphical comparison between linear regression and logistic regression is shown below: Replacing Y with a probability P that takes the range of probability to be within 0 and 1, the odds of P are taken,

$$P\ /\ 1 - P = \alpha + \beta X$$

In this equation, the range gets restricted which decreases the number of datapoints, eventually decreasing the correlation. To avoid this, the log of the odds need to be taken and exponent has to be added to both sides and the solution for P is: $P = 1\ 1 + e\ -(\alpha + \beta X)$ This is the sigmoid function for the logistic regression model used to predict any dichotomous dependent variable [12]. According to the dataset, the independent variable is the amount of annual rainfall and the dependent variable remains whether there will be a flood based on the rainfall or not. Support Vector Classifier: The Support Vector Classifier (SVC) is a machine learning algorithm that uses both regression and classification. Structural Risk Minimization Principle is roughly implemented in this way. The SVC, in contrast to other models, aims to fit the best line within a threshold value rather than to reduce error between real and predicted values. According to this research article [13], a training SVC algorithm creates a model which designates the data to one class or another developing a binary and non probabilistic linear classifier. These classes are clearly separated from each other with the help of a gap or a spatial line. SVC predicts the newer data and the class they belong to considering the distance of these classes from the line. The Basic idea is to perform linear regression to find a decision function for a given sample x,

$$i \epsilon SV \sum y\ i\ \alpha\ i\ K(x\ i\ , x) + b$$

The term α are called the dual coefficients which are i upper-bound by C and b is an independent term that has to be estimated. $K(x$ is the kernel where, is the input i , $x)$ x vector [9].

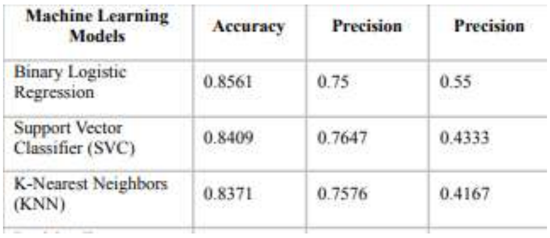
K-Nearest Neighbor: The KNN algorithm is a sort of supervised machine learning method that is being used to solve both classification and regression predicting problems. The KNN approach uses feature similarity to forecast the values of new data points, which means that the new data point will be assigned a value based on how closely it resembles the points in the training set. According to this research paper [13], the K nearest neighbor (KNN) is a non-parametric algorithm that can be used for regression predictive problems. The KNN method assumes the resemblance between the new data and the existing data and places the new data in the category that is most comparable to the existing categories. The KNN algorithm calculates the distance between a new data point and all previous data points in the training set. There are a variety of distance functions for calculating the distance but Euclidean is the most widely utilized method.

III. RESULTS

For this research, we have used Binary Logistic Regression, Support Vector Classifier (SVC), K-Nearest Neighbor (KNN) for predicting our data. The data has been split into 80:20 ratio for training and testing the models. The classification has been done on 16 columns: Station, Year, 12 months and Flood index. Firstly the models have been implemented on the whole dataset which consists of data from 190-2018. Later, the same models have been implemented with a shorter timeline of 10 years, 2011-2020 to check the accuracy and compare with the previous implementation.

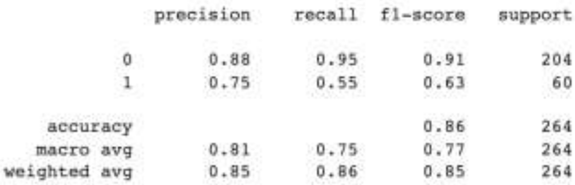
A. Timeline:

TABLE I.



Machine Learning Models	Accuracy	Precision	Precision
Binary Logistic Regression	0.8561	0.75	0.55
Support Vector Classifier (SVC)	0.8409	0.7647	0.4333
K-Nearest Neighbors (KNN)	0.8371	0.7576	0.4167

From the table, Binary Logistic Regression has the highest accuracy rate of 0.8561 with a precision and recall score of 0.75 and 0.55 respectively.



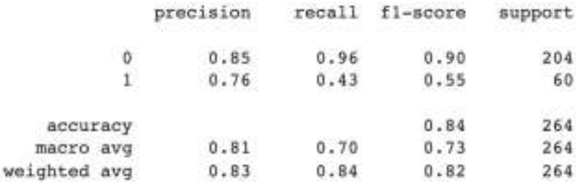
```
              precision    recall  f1-score   support

           0       0.88      0.95      0.91       204
           1       0.75      0.55      0.63        60

    accuracy                           0.86       264
   macro avg       0.81      0.75      0.77       264
weighted avg       0.85      0.86      0.85       264
```

Fig. 4. Classification Report of Binary Logistic Regression.

After that, the Support Vector Classifier (SVC) has the highest accuracy of 0.8409 with a precision of 0.7647 which is higher than the Binary Logistic Regression.



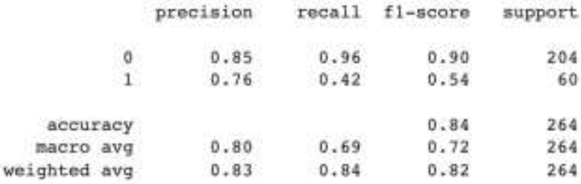
```
              precision    recall  f1-score   support

           0       0.85      0.96      0.90       204
           1       0.76      0.43      0.55        60

    accuracy                           0.84       264
   macro avg       0.81      0.70      0.73       264
weighted avg       0.83      0.84      0.82       264
```

Fig. 4. Classification Report of Support Vector Classifier.

Then, K-Nearest Neighbors (KNN) has an accuracy of 0.8371, precision and recall score respectively 0.7576 and 0.4167.



```
              precision    recall  f1-score   support

           0       0.85      0.96      0.90       204
           1       0.76      0.42      0.54        60

    accuracy                           0.84       264
   macro avg       0.80      0.69      0.72       264
weighted avg       0.83      0.84      0.82       264
```

Fig. 5. Classification Report of K-Nearest Neighbors (KNN)

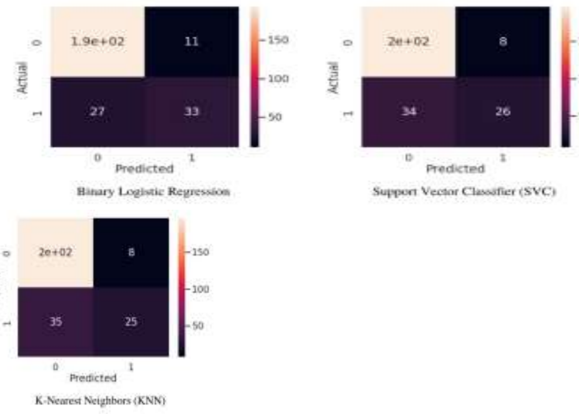


Fig. 7. Confusion matrix of the used models .

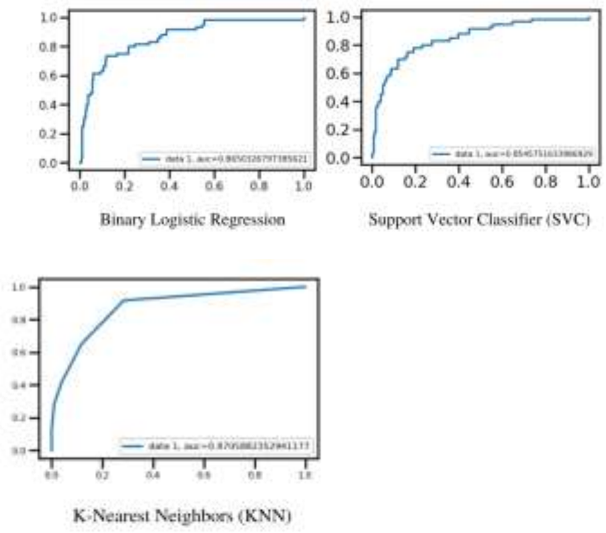


Fig. 8. ROC Curves of used models

IV. DISCUSSION

For the timeline 1901-2010, from the ROC Curves of Fig. 10, it is shown that the K-Nearest Neighbors (KNN) model has the highest AUC score of 0.87 but from Table 1 and figure 13, K-Nearest Neighbors (KNN) model has less accuracy and f1 score than Binary Logistic Regression Model (Accuracy: 0.8561, f1 score: 0.86). As higher the f1 score the better, it can be stated that Binary Logistic Regression is better than any other model with a better accuracy and other scores. For the timeline 2011-2020, from the ROC Curves it is shown that the Binary Logistic Regression model and the K-Nearest Neighbors (KNN) model have almost AUC scores (0.845 and 0.846 respectively). As the Binary Logistic Regression model has a better accuracy (0.8676) and f1 score (0.87), this model is considered to be the better one among the four models. From both of the timelines, most of the models gave better accuracy on the 10 years of rainfall data (2011-2020) than the whole timeline (1901-2020). Binary Logistic Regression has given the highest accuracy of 86.76% (Timeline: 2011-2020) among all with a better accuracy, recall and f1-score.

V. CONCLUSION

As climate changes over the years and depending on other parameters, the thresholds for floods are changing. That's why the shorter timeline of data gives slightly better accuracy. Since these change over a long time period, in this research, the models gave higher accuracy with a shorter time range. Also, due to time constraint only the rainfall data along with flood occurrence was manageable. There are more factors related to flood like, river water level, temperature, humidity, other natural disasters etc. In the future, this research paper would attempt to develop the models further by adding the other factors and correlating them.

REFERENCES

[1] Pappenberger, F.; Cloke, H.L.; Parker, D.J.; Wetterhall, F.; Richardson, D.S.;Thielen, J. The monetary benefit of early flood warnings in Europe. Environ. Sci. Policy 2015, 51, 278–291.

[2] Krzysztofowicz, R. Bayesian system for probabilistic river stage forecasting. J. Hydrol. 2002, 268, 16–40.

[3] Todini, E. Role and treatment of uncertainty in real-time flood forecasting. Hydrol. Process. 2004, 18, 2743–2746.

[4] Clark, M.P.; Slater, A.G. Probabilistic quantitative precipitation estimation in complex terrain. J. Hydrometeorol. 2006, 7, 3–22.

[5] Vrugt, J.A.; Robinson, B.A. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resources. 2007, 43.

[6] Ebtehaj, M.; Moradkhani, H.; Gupta, H.V. Improving robustness of hydrologic parameter

estimation by the use of moving block bootstrap re-sampling. Water Resour. Res. 2010, 46.

[7] He, X.; Refsgaard, J.C.; Sonnenborg, T.O.; Vejen, F.; Jensen, K.H. Statistical analysis of the impact of radar rainfall uncertainties on water resources modeling. Water Resources. 2011, 47.

[8] Legleiter, C.J.; Kyriakidis, P.C.; McDonald, R.R.; Nelson, J.M. Effects of uncertain topographic input data on two dimensional flow modelling in a gravel-bed river. Water Resour. Res. 2011, 47.

[9] Sikorska, A.E.; Scheidegger, A.; Banasik, K.; Rieckermann, J. Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models. Hydrol. Earth Syst. Sci. 2012, 16, 1221–1236.

[10] Montanari, A.; Koutsoyiannis, D. A blueprint for process-based modeling of uncertain hydrological systems. Water Resour. Res. 2012, 48.

[11]Prof.Parthasarathi Choudhury and A. Sankarasubramanian(2009), “River Flood Forecasting Using Complementary Muskingum Rating Equations”,Journal of Hydrologic Engineering, Vol. 14, No. 7, July 1, 2009.

[10] Feature scaling | Standardization vs Normalization. (2020, April3). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machi ne-learning-normalization-standardization/ [11] Fernandes, A. A. T., Figueiredo Filho, D. B