International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

DYNAMIC AND BOOSTING FLOOD PREDICTION THROUGH MACHINE LEARNING

Mrs.R.Prema[1], Sara Mohan Satya Ram[2], G Anil Govind[3]

[1]*MrsR.Prema Assistant professor, SCSVMV University*
[2]*Sara Mohan Satya Ram Student, SCSVMV University*
[3]*G Anil Govind Student, SCSVMV University*
[1-3]*Dept. of Computer Science Engineering, SCSVMV University, TN, India*

Abstract - *Floods are one of the most catastrophic natural disasters, and, due to their complex nature, it is not easy to create a predictive model. The advanced research works on flood prediction models have contributed to risk reduction, policy suggestion, minimization of the loss of human life, and reduced property damage associated with floods. In general, neural networks are used in the development of prediction systems, to mimic the complex mathematical expressions of the physical processes of floods providing better performance and cost-effective solutions. The best algorithm for prediction of flood occurrence is analyzed by comparing MLP with Logistic Regression, Support Vector Machine, K Nearest Neighbor algorithm and accuracy values are calculated with an evaluation of classification report and by precisely examining the confusion matrix parameters. The proposed system analyses the dataset using Multilayer Perceptron Classifier (MLP) algorithm to train the predictive model, and with a developed graphical user interface, real time flash flood predictions are made.*

Key Words: Deep Learning, Neural Networks, Multilayer Perceptron, Classification, Floods, Rainfall

1. INTRODUCTION

It is seen that now-a-days Machine Learning is playing a huge role in every field, it also includes research about various events depending on the previous event related data. It is the ability to learn based on experiences which is only possible when we have some original, precise, complete data. Thus, In-order to apply ML we need to assemble all the required data. The data then needs to be pre-processed so that it could be carried forward for further operations or functioning. It is necessary to maintain knowledge about the residing or available data to use it in the best way possible. We can use or try different Algorithms to get the accuracy based on the existing data. Flood prediction is an important consideration, due to changing climatic conditions. We have used Logistic Regression to come-up with the best outcome. We have considered TELANGANA STATE for the maximum use of the built- system. Floods are the most damaging natural disaster in this world. On the occasion of heavy flood, it can destroy a whole community. It is crucial to develop a flood prediction system as a mechanism to predict and reduce the flood risk. It proves necessary for alerting resident to take early action such as evacuate quickly to a safer and higher place. Aim is to specify the contribution of ML in different models. The dataset for the amount of rainfall in various states in India is provided on data.gov.in. We have provided dataset consisting of rainfall details of TELANGANA of previous 115 years, it clearly defines the annual as well as the monthly rainfall data which proves this system more accurate, and it confirms its reliability, efficiency and confident dependence as well. This model gives us a well defined idea of how the Logistic Regression Algorithm works well with a precise data. This algorithm solves half of the case because of its binary classified nature. The goal of this particular system is to contribute to development of ML as well and to improve the conditions of Living Life in case of the calamity (Flood).

1.1 Iiterature Review

The flash flood explored a small but ample flood forecasting research work carried out. Akshya et al. [1] proposed the classification of the flood-hit region using the aerial images from the satellite. A hybrid ML approach, SVM classifier with k-means provides better accuracy but after flood region classified. The prediction system designed using IoT and Neural Networks for smart prediction of the flood by Bande et al., [27] suggested that the data collected from the sensor using IoT and Wi-Fi with an Artificial Neural Network (ANN) approach was used for communication of the data analysis in flood prediction. The Wi-Fi is highly impossible due to network during rainfall, the high cost of IoT, and communication. Cruz et al. [8] studied the flood prediction system by Multi-Layer ANN. Rain gauge, soil moisture, and water level are the parameters used in this model. Moreover, the actual setup tested and the model showed 2.2645 of Root Mean Square Deviation rate across the whole data set, which

implies an overall small difference between the flood predicted, and actual flood level. The Multi-layered artificial neural network gives very well validation, but this study requires more parameters for calculation of prediction.

The coastal cities flood loss prediction by Cui et al. [20] designed the AHP-GM-ANN model would reduce the predictive error to get reliable results significantly. The model solved the problem in nonlinear relationship variables and also improved the quantitative system accuracy in the predictive method. Kartika et al. [18] implemented the system to predict the flood using Radial Basis Function (RBF). It determines the next month's water level and daily rainfall. But it may be impossible to predict the rainfall due to climate change. Kaur et al. [9] implemented the hybrid algorithm in standalone and cloud environments for efficiency. The automated hybrid algorithms are a Genetic Algorithm (GA) and SVM. The evaluated model shows the results in a cloud environment, which yields the highest accuracy of 86.36%, but in a standalone, it is less inaccurate. Case study of flood prediction system by Ruslan et al. [7] studied the design of prediction models using Multiple Input Single Output (MISO) ARX and ARMAX model structure in Pahang for 7 hours and shows the comparison of the prediction performances. When the prediction time is longer, the model output can be unreliable.

[11] implemented the LSTM model for solving. The model gives context-awareness, which in turn provides information about data processing and real-time predictions of the events to the user. Sardjono et al. [30] applied an ANN for flood systems mitigation in place of Jakarta City. The data patterns were determined as a result of this flood model. The result of Jakarta are expected to use in prediction and flood behavior in the future. Ramli et al. [6] presented the neural network autoregressive model for floods prediction in Kuala Lumpur. The model showed an accuracy of 73.54% as a maximum. The authors suggested that by using different prediction models, the prediction time can reduce further.

As per the literature survey, authors have not combined the classification algorithm and unfocused on the accuracy, which reduces the execution time for optimization. Hence, this research work mainly focuses on predicting the floods accurately with minimum execution time, more accuracy, and best features.

1.2 Implemented Methodology

The DATA concerned with factors that affect flood will be provided. The provided data will be analyzed and used for the training. The data will be analyzed using basic approaches of Machine Learning-Linear/Logistic Regression. After the training the model will be tested. Based on the given input the model will predict if flood will occur or not. Real time data of rainfall from the month of March to May (for present year in TELANGANA state)Real time data of average rainfall in the first 10 days of June(for present year in TELANGANA state)Real time data of average increase in rainfall from the months of May of June(for present year in TELANGANA state). Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two potential qualities, for example, pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log- odds to probability is the logistic function, hence the name. The unit of measurement for the log- odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model. The defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the chances of the given result at a consistent rate, with every free factor having its own parameter; for a binary dependent variable this generalizes the odds ratio.

The binary logistic regression model has augmentations to multiple degrees of the needy variable: clear cut yields with multiple qualities are demonstrated by multinomial strategic relapse, and if the different classes are requested, by ordinal strategic relapse, for instance the corresponding chances ordinal calculated model. The model itself essentially models likelihood of yield as far as information, and doesn't perform measurable order (it's anything but a classifier), however it

very well may be utilized to makea classifier, for example by picking a cutoff esteem and grouping contributions with likelihood more noteworthy than the cutoff as one class, underneath the cutoff as the other; this is a typical method to make a binary classifier.

Flood column shows the classification of occurrence of the flood if the amount is greater than threshold value.

2. FLOOD PREDICTION MODEL

Flood prediction model can play a main role in providing appropriate information of possible impending floods in populated locations. The development of such models can reduce the damage in such areas. More importantly, a prediction system developed, it can effectively lower the risk of harm and loss of life. If neural network models can provide sufficiently accurate forecasts, even one day ahead, the lead time for flood warning can be extendedand the possible flood emergency measures can be better planned and executed. This system is applied to one region, as a proof of concept. It is primarily focused on the use of neural networks trained on historical rainfall values to predict the future values. The advantage of the proposed method is that it requires very few variables and very little knowledge

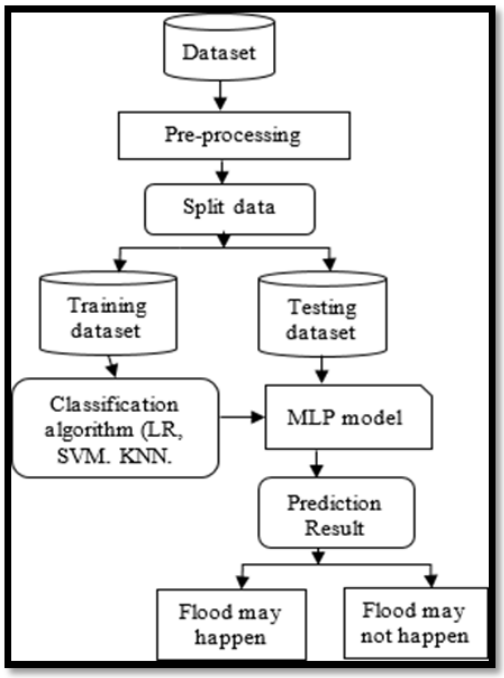

Fig1

Process of Prediction of Flood

3. Algorithms & Techniques

3.1 Algorithms

3.1.1 Logistic Regression: Logistic Regression is a machine learning algorithm that predicts the probability of a categorical dependent variable. It is a statistical way of analyzing a set of data that comprises more than one independent variable that determines the outcome. The outcome is then measured with a dichotomous variable. The goal of this algorithm is to find the best model to describe the relationship between a dichotomous characteristic of interest and a set 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.

3.1.2 Support Vector Machines: SVM uses a classifier that categorizes the data 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 of the most popular and widely used clustering algorithms. It belongs to a group of generalized linear classifiers and is considered 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.

3.1.3 K-Nearest Neighbor (KNN): K-Nearest Neighbor is one of the supervised machine learning algorithms that stores all instances corresponding to 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 which 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 learn how to label the other 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 the most of the neighbors have is the label for the new point. This algorithm makes predictions about the validation set using the entire training set. Only by searching through the entire training set to find the closest

instances, the new instance is predicted. Closeness is a value that is determined using a proximity measurement across all the features involved.

3.2 Data Validation/ Cleaning/Preparing Process:

This process involves variable identification by data shape, data type, and evaluating the missing values, duplicate values. The process of data cleaning varies from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making. A sample of data, known as the validation dataset, is held back from training the model, that can be used to make the best use of validation and test datasets when evaluating your models, giving an estimate of model skill while tuning models and procedures.

3.3 Data Pre-processing: Pre-processing is the process where the data is transformed prior to feeding it to the algorithm. This is done in order to convert the raw data into a clean dataset for better results after feeding it to the algorithm. As the data gathered are from different sources, and not uniform, it is not feasible for analysis. To achieve better results from the applied model in the Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning models need information in a specified format; for example, the Random Forest algorithm does not support null values. Therefore, to execute random forest algorithms, null values have to be managed from the original raw data set.

3.4 Exploration Data Analysis of Visualization: Data Visualization is considered as a very important skill in applied statistics which focus on quantitative descriptions and estimations of data, and machine learning. Data visualisation provides an important suite of tools for gaining a qualitative understanding which can be useful when exploring and getting to know a dataset which can help with identifying patterns, corrupt data, outliers and much more. Data visualization can also be used to demonstrate key relationships in plots and charts that are more visceral.

3.5 Outlier Detection Process: Many machine learning algorithms are known to be sensitive to range and distribution of attribute values in the input data. Outliers in those data can mislead and skew the training process of the algorithms which will result in longer training times, less accurate models and ultimately poorer results. Even before the process of training the model, the outliers can result in misleading representations hence resulting in misleading interpretations of collected data. Outliers are capable of skewing the summary distribution of attribute values in statistics like mean, standard deviation and in plots such as scatter plots, histograms, compressing the body of the data. Outliers are also capable of representing examples of data instances which are relevant to the problem such as anomalies in the case of fraud detection and computer security. It couldn't fit the model on the training data, and this can affect the model to work inaccurately in case of real data. Hence, we must ensure that our model gets the correct patterns from the data, and not receiving too much noise. Cross Validation is a process where we train our model using the subset of the dataset and then evaluating using the complementary subset of the dataset. Cross validation is advantageous as it provides a more accurate estimate of out of sample accuracy and also for its more efficient use of data as every observation is used for both training and testing.

3. 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 mispredicts 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 mispredicts 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 mispredicted, which can be represented in the form of a false negative rate. Therefore, the sum of sensitivity and false-negative rate value is 1.

Table-1: Comparison of Accuracy Results

	Precision	Recall	F1-Score	Sensitivity	Specificity	Accuracy (%)
LR	0.99	0.96	0.98	0.96	0.75	95.33
SVMe	0.96	1	0.98	1	0	95.85
KNN	0.97	0.98	0.98	0.98	0.37	95.85

Table 1 compares the Accuracy results of the Logistic Regression, Support Vector Machine, K-Nearest Neighbour, and Multilayer perceptron algorithms. From the calculated values, we can observe that the Support Vector Machine and Multilayer Perceptron has comparatively better results. And from further comparison, we can conclude that Multilayer Perceptron gives the highest accuracy percentage value.

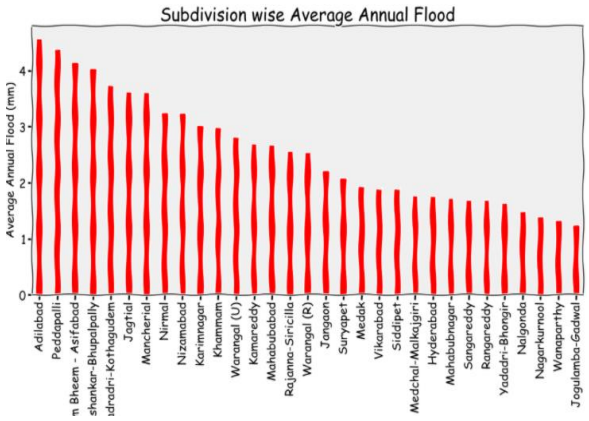

Fig2 Sub Division wise Average Flood

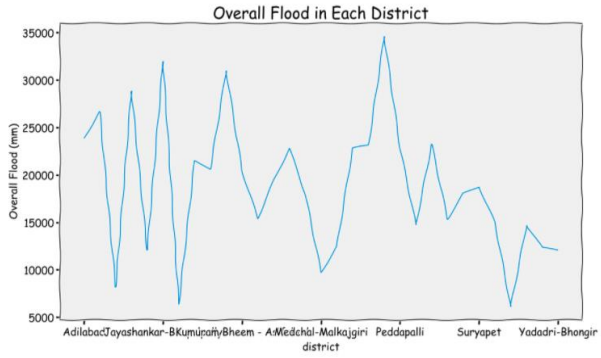

Fig3 Overall Floods in Each District

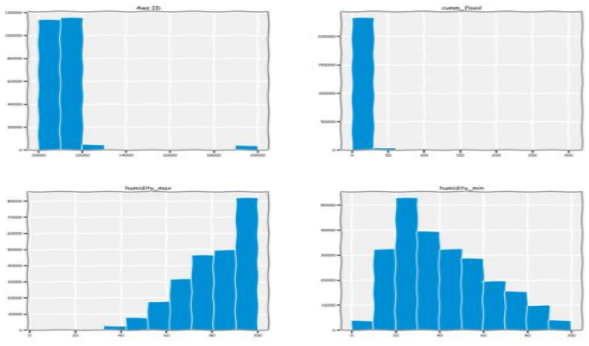

Fig4 Histogram of categories of columns heads in dataset

5. Conclusion

Damages that occur due to flash flood to living and non-living are very large. In this paper, flash flood prediction model is built. District wise Indian rainfall data collected between the periods of 1901 to 2015 is used for analysis. The pre-processed rainfall data was split into 70% training data and 30% testing data. The dataset is trained with Support Vector Machine, Logistic regression, K-nearest neighbor, and Multi-Layer Perceptron. The performance factors like precision, recall, F1 score, sensitivity, specificity was calculated for each technique. Confusion matrix with TP, TN, FP and FN were calculated. The classification accuracy achieved by LR is 95.3%, SVM is 95.85%, KNN is 95.85%, and MLP is 97.40%. Among the four techniques MLP performed with highest accuracy. The MLP flash flood prediction model predicts whether "flood may happen or not" based on the rainfall range for particular locations. This prediction model can be used by disaster management department to forecast flash flood. In future, we aim to use other artificial intelligence techniques to improve the prediction accuracy. The process can be automated by displaying the result of prediction in webpage or desktop application.

REFERENCES

1. Akshya J and Priyadarsini K, "A Hybrid Machine Learning Approach for Classifying Aerial Images of Flood-Hit Areas," in Second International Conference on Computational Intelligence in Data Science, 2019.

2. Amir Mosavi., Pinar Ozturk and Kwok-wing Chau, "Flood Prediction Using Machine Learning Models: Literature Review," MDPI, p. 40, 2018.

3. Amitkumar B and Durge P, "Different Techniques of Flood Forecasting and their Applications," 2018.

4. Analyn N., Charmaine C., Arnold C., Glenn O., JoseAngelo C., JanRalley P., and Jesse Dave S, "Real-Time Flood Water Level Monitoring System with SMS Notification," 2017.

5. Bipendra Basnyat., Nirmala Roy.,and Aryya Gangopadhyay., "A Flash Flood Categorization System using Scene-Text Recognition," in International Conference on Smart Computing, 2018.

6. Fazlina Ahmat Ruslan.,Abd Manan Samad and Ramli Adnan, "4 Hours NNARX Flood Prediction Model Using "traingd" and "trainoss" Training Function: A Comparative Study," in 2018 IEEE 14th International Colloquium on Signal Processing & its Applications (CSPA 2018), 9 -10 March 2018, Penang, Malaysia, 2018.

7. Fazlina Ahmat Ruslan., Khadijah Haron., Abd Manan Samad and Ramli Adnan, "Multiple Input Single Output (MISO) ARX and ARMAX Model of Flood Prediction System: Case Study Pahang," in IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA 2017), Penang, Malaysia, 2017.

8. Febus Reidj G., Matthew G., Marlou Ryan G., and Francis Aldrine A., "Flood Prediction Using Multi-Layer Artificial Neural Network in Monitoring System with Rain Gauge, Water Level, Soil Moisture Sensors," in Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference, (Jeju,Korea, 2018).

9. Gurleen Kaur and Anju Bala, "An Efficient Automated Hybrid Algorithm to Predict Floods in Cloud Environment," in 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), 2019.

10. Homar Baez., Idalides Vergara-Laurens., Luz Torres-Molina., and Luis G., Miguel A. Labrador, "A Real-Time Flood Alert System for Parking Lots," 2017.