FLOOD PREDICTION ANALYSIS USING SUPERVISED MACHINE LEARNING TECHNIQUES

Tripti Sharma[a], Sanjeev Kumar Prasad[b] and Priya Mishra[c]
[a]*Department of Computer Science and Engineering(Data Science), Inderprastha Engineering College Ghaziabad, Uttar Pradesh, India.*
[b]Department of Computer Application, Galgotias University, Greater Noida, India.
[c]*Department Computer Application, ABES Engineering College, Ghaziabad, U.P, India.*

Abstract— Floods also known as Cataracts have come the most well- known and murderous cataclysmic events of this century. Absence of a successful deluge soothsaying frame has brought about grave loss of mortal actuality and structure. This has reiterated the significance of having in place a deluge vaccination system. This paper looks at developing the most effective deluge determining model. AI computations and a hearty, productive and precise deluge anticipation frame will give all the abecedarian aid and backing demanded to the residers and government. Hence, the Decision Tree Model is being erected. This model actualizes colorful computations on datasets with a compass of delicacy. The model utilizes an AI computation which predicts Floods, transferring cautions to the original and government authorities using an Android Operation. The comparison of the results hahas been performed on three Machine Learning Algorithms that are Decision Tree, Random Forest and Gradient Boost. This model focuses on perfecting the rate of vaccination by dealing with further intricate information and a high- position algorithm.

Keywords—Machine learning approaches, Flood analysis, Decision Tree, Random Forest, KNN.

I. INTRODUCTION

Flood is a pervasive natural hazard al over the world. The water level rising above the riverbank causes a river flood. Floods have both direct and indirect detrimental consequences on human life, the environment, ecosystems, transportation, infrastructure, agriculture, cultural heritage, economics, and so on. They also play an important role in supplying nutrients and enriching soil [1]. There is a transition from 'flood control' to 'flood risk management,' with a focus on India's flood damage figures. Many countries waste billions of dollars every time flood hazards are analysed in terms of cost rather than prevention using structural solutions [2].

A flood happens when water submerges land that is normally dry, which can happen in an enormous number of ways. Brisk liquefying of ice, outlandish rainfall or a burst dam, can overwhelm a river, spreading over the contiguous land. Ocean front flooding happens when a colossal storm or tsunami makes the ocean flood inland. Floods are considered as the most common natural disaster on Earth, second only to the forest fires.

According to the Organization for Economic Cooperation and Development, floods cause damages of more than $40 billion worldwide every year. Most nations actually do not have successful flood cautioning frameworks. According to the Central Water Commission, 20% of flood fatalities occur in India. Bihar is the most noticeably awful influenced state, with practically 73% of its complete surface territory getting overwhelmed every year. The cost of damage to infrastructure, crops, and public utilities all over India was reported to be as much as 3% of India's gross domestic product in 2018.

There are different ways that can be undertaken to forestall floods, quite possibly the best and simplest early warning system is using AI algorithms for the forecast of floods because of substantial rains and flooding of various water bodies. With the approach of sensor innovation, different attributes have been recorded to anticipate floods. A wide scope of datasets is now available that can be utilized to create expectation frameworks. Machine Learning would guarantee vigorous, proficient, and precise predictions.

II. METHODOLOGY

A. K-Nearest Neighbors Algorithm

K-Nearest Neighbour (KNN) is a supervised Machine Learning Model. It is a direct and efficient model that can be applied to classification tasks. The K-NN model assumes that similar items can be found nearby. That is, it operates on the basic principle that "similar things are closer to each other." The distance can be calculated using a method known as "Euclidean distance".

Pandas, Matplotlib, and Numpy library has been used for implementing the K-Nearest Neighbor. The data has been processed first according to the requirements of K-Nearest Neighbors, than data has been fitted into the model, after that a comparison between predicted and actual value is done, to check the accuracy. Further Recall Score, and ROC are also calculated.

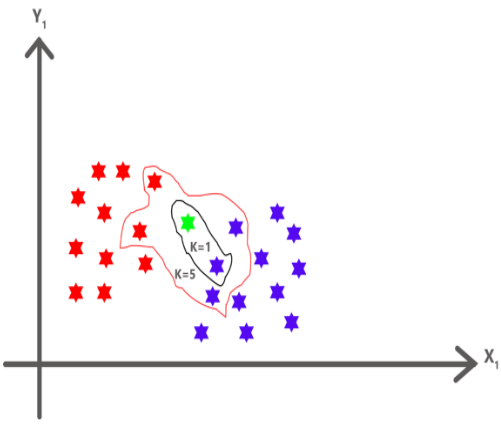

Fig. 1.K- Nearest Neighbor

B. Support Vector Machine

Support Vector Machines are a collection of Supervised models, that are used for various purposes like Regression and Classification. Support Vector Machines are always preferred in high dimensional spaces SVM uses a hyperplane that classifies data into different classes. SVM can have multiple hyperplanes. So it becomes very important to choose the correct hyperplane.

There are two types of SVM:

1. Linear SVM - When the dataset can be divided into different classes by a single line, it can be termed as Linear SVM. Fig 2 shows Linear SVM clearly.
2. Non-Linear SVM - When a single line hyperplane can't determine different classes accurately from the dataset, that will be called Non-Linear SVM.

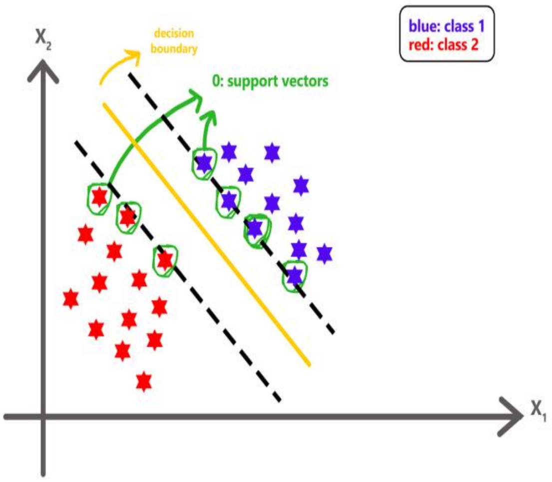

Fig. 2. Support Vector Machine

The greatest example of a separating hyperplane equation is: (4), where w and b are model parameters that determine the hyperplane's direction and distance from the origin. To fit the maximum margin hyperplane in a higher dimensions plane, the SVM employs a kernel trick. "Kernel functions allow them to work in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, instead computing the inner products between all pairs of data. This operation is frequently less computationally expensive than explicit coordinate computation" [32, 33].

C. Decision Trees

Decision Tree is a supervised machine learning technique that is widely used for classification. But it can be also used for Regression problems, although it is not recommended to implement decision trees on regression problems. A Decision tree is a graphical solution to a decision based certain conditions.

Entropy defines the randomness in the data. It's just a metric which measures the impurity. It is the first step in Decision tree. Entropy is defined as:

$$\sum_{i=1}^{k} (value_i)\, log_2(P(value_i))$$

where k represents the numbers of elements present in the dataset, P is the probability of an element.

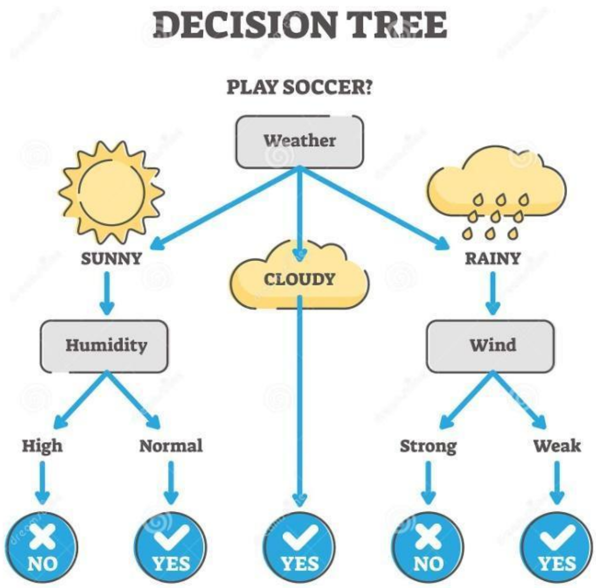

Fig. 3. Decision Tree Example

D. Logistic Regression

Logistic Regression comes under the category of supervised learning model. It is used for solving classification problems. It is used when the output is necessary to be present in the 0 or 1, Yes or No, True or False, High or Low. This algorithm works based on the equation below:

$$\text{Log } [\ /1-y\ ] = b0 + b1x1+b2x2 + \cdots + bnxn$$

Fig-4 shows the curve of logistic regression.

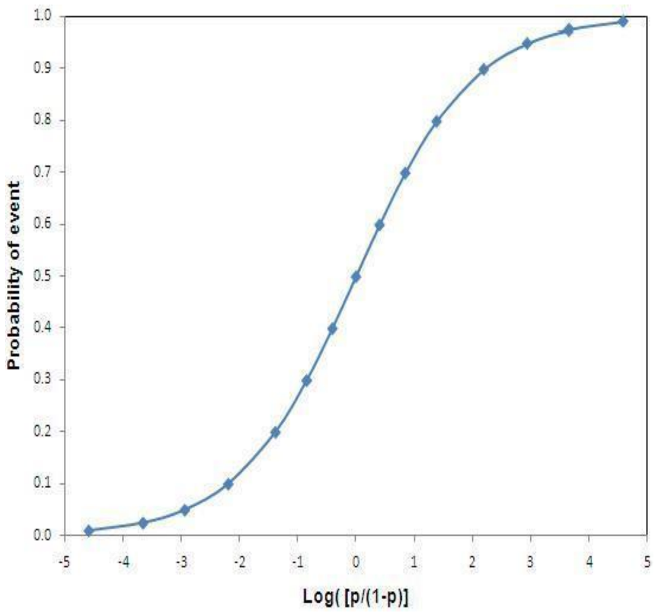

Fig. 4. Logistic Regression Curve

E. Random Forest Classifier

Random Forest algorithm is a machine learning method that is built on the notion of decision tree algorithms. The random forest method generates several decision trees. The more trees there are, the more accurate the detection. The bootstrap technique is used to generate trees. The characteristics and samples of the dataset are randomly picked using a replacement in the bootstrap approach to form a single tree. Random forest algorithm, like decision tree algorithm, will identify the best splitter for classification from randomly selected characteristics. Random forest algorithm uses gain index and information gain methods to discover the best splitter. This will continue until the random forest has produced n trees. The method will compute votes for each projected target once each tree in the forest forecasts the target value. Finally, the random forest algorithm uses the target with the most votes as the final splitter.

III. DATASETS

The dataset has been collected for more than 100 years for different regions of India i.e., Saurashtra, Kerala and more. This dataset consists of columns showing daily and monthly rainfall, various groups of monthly rainfall, and the percentage of floods for that particular year. The forecast is based on the monthly rainfall for that particular year. An examination of average monthly precipitation from 1901 to 2017 is shown in the form of a bar graph, with the peak and lowest rainfall months highlighted. The most rainfall occurs in June and July.

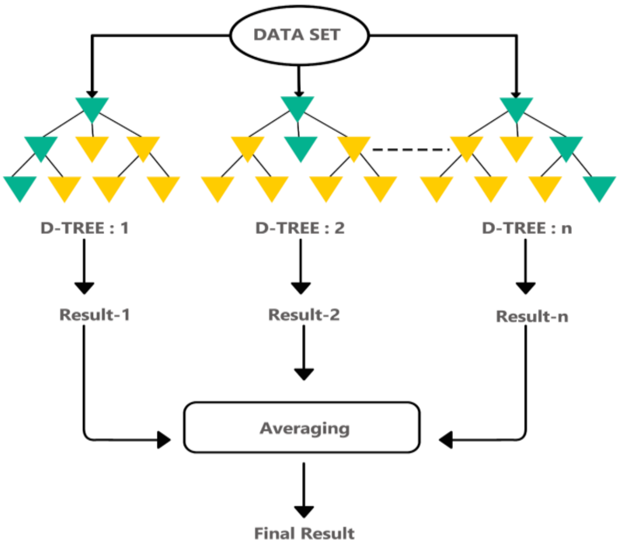

Fig. 5. Random Forest Classifier

Fig-5 Shows the rainfall dataset for Kerala state using bar graph, which shows that the peak rainfall usually comes in july and august month.

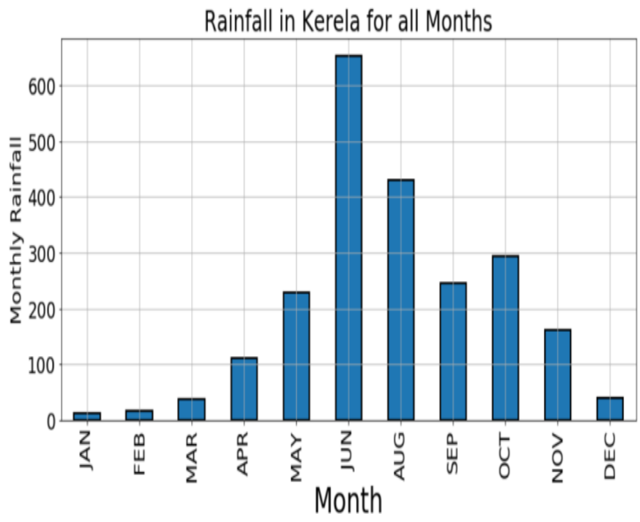

Fig. 5. Kerala Rainfall Chart

IV. RESULTS AND FINDINGS

The accuracy, recall, and ROC score of each prediction model will be assessed. Table I illustrates the prediction's performance based on a 75 percent training and 25% test dataset. Table I shows that logistic regression has an accuracy of 0.87, a recall score of o.80, indicating that there are very few chances of incorrectly predicting a positive value, and a ROC score of o.90, indicating that the precision and recall scores are well balanced, implying that the overall performance of Logistic Regression for flood prediction is very good. The other model DT, has the lowest accuracy (0.62), the least recall (0.60), and ROC-scores of 0.60 and 0.63. When we look at the metric scores of the remaining two models, the RFC and support vector machine do not demonstrate the predicted efficiency. The Logistic Regression has decisively outperformed the remaining four machine learning models in the aforementioned study, making it the best recommendable machine learning model for reliable flood prediction. As a result, it will be used to assess the significance of features. Although these values can fluctuate depending on other states and conditions.

Model	Accuracy	Recall	ROC
KNN	0.79	0.73	0.81
LR	0.87	0.80	0.90
SVC	0.75	0.60	0.80
DT	0.62	0.60	0.63
RF	0.70	0.66	0.72

Table I: Prediction results on test datasets

Fig-6 shows the prediction charts of different state with different conditions in which also Logistic Regression has highest accuracy.

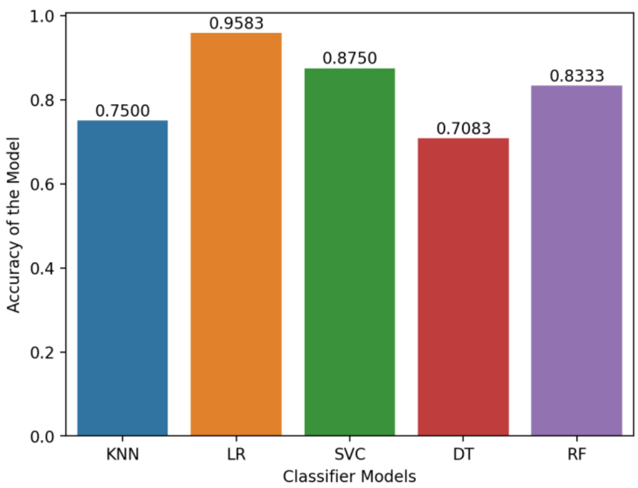

Fig. 6. Prediction Model Result

V. CONCLUSION AND DISCUSSION

The study discusses the need for a flood prediction model that is based on machine learning. The accuracy, precision, recall, and ROC-score of five different machine learning models, including KNN, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, were compared. The best model with the greatest metric score is Logistic Regression, according to the results. The work's future focus will be on deep learning models and human-machine interaction, with the goal of enabling users to find a solution that can help predict flooding in the following years. Further upgradation can be built a new system that sends out warnings and alerts of an incoming flood to the citizens and helps save the lives of civilians and if possible, the infrastructure. The system also helps the government save money in rescue operations and helps them start the relocation operations before the flood hits the town.

For models, there is room for progress and advancement. Some strategies for improving models include the use of data decomposition techniques to improve the quality of datasets and the use of an ensemble of methods to improve model generalization and reduce prediction uncertainty. Additionally, add-on optimizers can help increase the quality of machine learning algorithms.

For future efforts, conducting a survey on spatial flood prediction using machine learning algorithms is strongly recommended. Increasing the databases for flood location could potentially be a future project.

REFERENCES

[1] Chhuonvoch Koem , Sarintip Tantanee, “Flood disaster studies: A review of remote sensing perspective in Cambodia”

[2] Snehil, Ruchi Goel “Flood Damage Analysis Using Machine Learning Techniques Flood Damage Analysis Using Machine Learning Techniques”.

[3] Ruhhee, Tabbussum , Abdul Qayoom Dar “Performance evaluation of artificial intelligence
paradigms—artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction”.

[4] Li-Chiu Chang , Hung-Yu Shen , Yi-Fung Wang , Jing-Yu Huang , Yen-Tso Lin "Clustering-based hybrid inundation model for forecasting flood inundation depths".
[5] Ho Ju Keum , kun Yeun Han , and Hyun I Kim "Real-Time Flood Disaster Prediction System by Applying Machine Learning Technique".
[6] Rabindra K. Panda , Niranjan Pramanik, Biplab Bala "Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model
[7] Mohammed Khalaf , Abir Jaafar Hussain , Dhiya Al-Jumeily , Thar Baker , Robert Keight , Paulo Lisboa, Paul Fergus , Ala S. Al Kafri "A Data Science Methodology Based on Machine
Learning Algorithms for Flood Severity Prediction"
[8] Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article "Assessment of economic flood damage"
[9] Wagenaar, D., de Jong, J., and Bouwer, L. M. " Multi-variable flood damage modelling with limited data using supervised learning approaches"
[10] Hanghyun Choi, Jeonghwan Kim, Jongsung Kim, Donghyun KimYounghye Bae,2 and HungSooKim "Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data".
[11] Tripathi, Prakash " Flood Disaster in India: An Analysis of trend and Preparedness"
[12] Abebe Y, Kabir G, Tesfamariam S "Assessing urban areas vulnerability to pluvial flooding using GIS applications and Bayesian Belief Network model".
[13] Adamowski J, Karapataki C "Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN Learning algorithms".
[14] Alvisi S, Mascellani G, Franchini M, Bárdossy A "Water level forecasting through Fuzzy Logic and Artificial Neural Network approaches".
[15] Anusree K, Varghese KO "Streamflow Prediction of Karuvannur River Basin Using ANFIS, ANN and MNLR Models".
[16] Arabameri A, Saha S, Chen W, Roy J, Pradhan B, Bui DT " Flash flood susceptibility modelling using functional tree and hybrid ensemble techniques".
[17] Bates, P.D., Horritt, M.S., Aronica, G., Beven, K. "Bayesian updating of flood inundation likelihoods conditioned on flood extent data".
[18] Chang, F.J., Liang, J.M., Chen, Y.C. "Flood forecasting using radial basis function neural networks".
[19] Chaves, P., Chang, F.J. "Intelligent reservoir operation system based on evolving artificial neural networks".
[20] Dastorani, M.T., Wright, N.G. "Application of artificial neural networks for ungauged catchment flood prediction".
[21] Anmala J, Zhang B, Govindaraju RS "Comparison of ANNs and empirical approaches for predicting watershed runoff".
[22] Chang LC, Shen HY, Chang FJ "Regional flood inundation nowcast using hybrid SOM and dynamic neural networks".
[23] Han D, Kwong T, Li S "Uncertainties in real-time flood forecasting with neural networks".
[24] Cigizoglu, H.K. "Incorporation of ARMA models into flow forecasting by artificial neural networks".
[25] Bruen, M., Yang, J. "Functional networks in real time flood forecasting"
[26] Gunn, R.R. "Support vector machines for classification and regression, Technical Report".

[27] Hsu, K., Gupta, H.V., Sorooshian, S. “Artificial neural network modeling of the rainfall-runoff process”.
[28] M. A. Friedl and C. E. Brodley, “Brodley, "Decision tree classification of land cover
from remotely sensed data," .
[29] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning.
[30] L. Breiman, "Random forests," Machine learning.
[31] V. Lempitsky, M. Verhoek, J. A. Noble, and A. Blake, "Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography"
[32] Zongyan Huang, Matthew England, David Wilson, James H. Davenport, Lawrence C. Paulson & James Bridge, “Applying Machine Learning to the Problem of Choosing a Heuristic to Select the Variable Ordering for Cylindrical Algebraic Decomposition”.
[33] T. Sharma, A. Pal, A. Kaushik, A. Yadav and A. Chitragupta, "A Survey on Flood Prediction analysis based on ML Algorithm using Data Science Methodology," *2022 IEEE Delhi Section Conference (DELCON)*, 2022, pp. 1-8, doi: 10.1109/DELCON54057.2022.9753396.