JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 5419-5426
https://publishoa.com
ISSN: 1309-3452

Flood Prediction Analysis Using Supervised Machine Learning Techniques

[1]E.Indra,[2]P.R.Jayanthi

[1]Associate Professor, Department of Computer Science and Engineering,

Mailam Engineering College, Mailam

indracse@mailamengg.com

[2]Associate Professor, Department of Computer Science and Engineering,

Mailam Engineering College, Mailam

prjayanthi@gmail.com

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

***Abstract*—**FloodsalsoknownasCataractshavecomethemostwell-knownandmurderous cataclysmic events of this century. Absence of a successful deluge soothsayingframe has brought about grave loss of mortal actuality and structure. This has reiterated thesignificance of having in place a deluge vaccination system. This paper looks at developingthe most effective deluge determining model. AI computations and a hearty, productive andprecisedelugeanticipationframewillgivealltheabecedarianaidandbackingdemandedto the residers and government. Hence, the Decision Tree Model is being erected. Thismodel actualizes colorful computations on datasets with a compass of delicacy. The modelutilizes an AI computation which predicts Floods, transferring cautions to the original andgovernment authorities using an Android Operation. The comparison of the results hahasbeen performed on three Machine Learning Algorithms that are Decision Tree, RandomForest and Gradient Boost. This model focuses on perfecting the rate of vaccination bydealingwith furtherintricate information andahigh-position algorithm.

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

I. INTRODUCTION

Flood is a pervasive natural hazard al over the world. The water level rising above theriverbankcausesariverflood.Floodshavebothdirectandindirectdetrimentalconsequences on human life, the environment, ecosystems, transportation, infrastructure,agriculture, cultural heritage, economics, and so on. They also play an important role insupplying nutrients and enriching soil [1]. There is a transition from 'flood control' to 'floodrisk management,' with a focus on India's flood damage figures. Many countries wastebillionsof dollarsevery time floodhazards are analysedintermsof costrather thanprevention usingstructural solutions [2].

A flood happens when water submerges land that is normally dry, which can happen in anenormous number of ways. Brisk liquefying of ice, outlandish rainfall or a burst dam, canoverwhelm a river, spreading over the contiguous land. Ocean front flooding happens whena colossal storm or tsunami makes the ocean flood inland. Floodsare considered as themostcommon natural disasteron Earth, second onlyto theforest fires.

According to the Organization for Economic Cooperation and Development, floods causedamages of more than $40 billion worldwide every year. Most nations actually do not havesuccessful flood cautioning frameworks. According to the Central Water Commission, 20%of flood fatalities occur in India. Bihar is the most noticeably awful influenced state, withpractically73%ofitscompletesurfaceterritorygettingoverwhelmedeveryyear.Thecostof damage to infrastructure, crops, and public utilities all over India was reported to be asmuchas 3% of India'sgross domesticproduct in 2018.

There are different ways that can be undertaken to forestall floods, quite possibly the bestand simplest early warning system is using AI algorithms for the forecast of floods becauseof substantial rains and flooding of various water bodies. With the approach of sensorinnovation, different attributes have been recorded to anticipate floods. A wide scope ofdatasets is now available that can be utilized to create expectation frameworks.

MachineLearningwouldguaranteevigorous,proficient,and precise predictions.

II. METHODOLOGY

A. K-NearestNeighborsAlgorithm

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

Pandas, Matplotlib, and Numpy library has been used for implementing the K-NearestNeighbor. The data has been processed first according to the requirements of K-NearestNeighbors, than data hasbeenfittedintothe model,after thata comparisonbetweenpredicted and actual value is done, to check the accuracy. Further Recall Score, and ROCarealsocalculated.

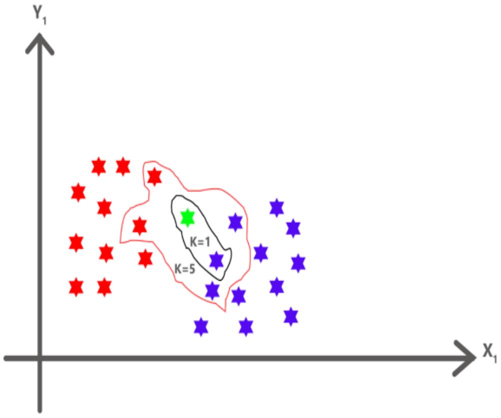

Fig.1.K-NearestNeighbor

B. SupportVectorMachine

Support Vector Machines are a collection of Supervised models, that are used for variouspurposes like Regression and Classification. Support Vector Machines are always preferredinhighdimensionalspacesSVMusesahyperplanethatclassifiesdataintodifferentclasses. SVM can have multiple hyperplanes. So it becomes very important to choose thecorrecthyperplane.

Therearetwotypes of SVM:

1. LinearSVM-Whenthedatasetcanbedividedintodifferentclassesbyasingleline,it canbetermed asLinearSVM. Fig2 showsLinear SVMclearly.
2. Non-LinearSVM-Whenasinglelinehyperplanecan’tdeterminedifferentclassesaccuratelyfrom thedataset, that will becalled Non-Linear SVM.

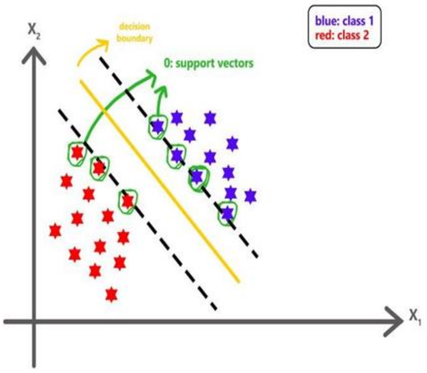

Fig.2.SupportVectorMachine

The greatest example of a separating hyperplane equation is: (4), where w and b are modelparameters that determine the hyperplane's direction and distance from the origin. To fit themaximum margin hyperplane in a higher dimensions plane, the SVM employs a kerneltrick. “Kernel functions allow them to work in a high-dimensional, implicit feature spacewithout ever computing the coordinates of the data in that space, instead computing theinner products between all pairs of data. This operation is frequently less computationallyexpensivethan explicit coordinate computation”[32, 33].

C. *DecisionTrees*

DecisionTreeisasupervisedmachinelearningtechniquethatiswidelyusedforclassification.Butitcanbealsousedfor Regression problems,althoughitisnotrecommended to implement decision trees on regression problems. A Decision tree is agraphicalsolution to a decision based certain conditions.

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

$$\sum \quad {}^{k}_{i=1}(value)log(P_i(value_2)) \quad {}_{i}$$

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

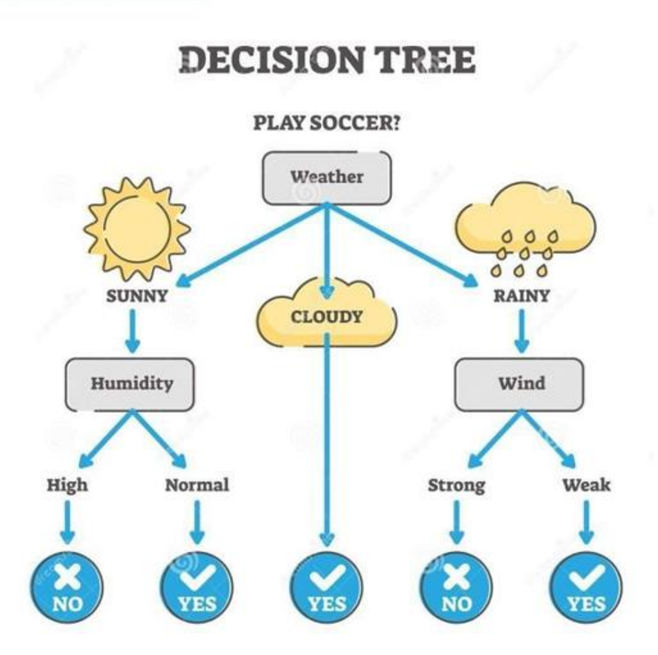

Fig.3.DecisionTreeExample

D. *LogisticRegression*

Logistic Regression comes under the category of supervised learning model. It is used forsolvingclassificationproblems.Itisusedwhentheoutputisnecessarytobepresentinthe0 or 1, Yes or No, True or False, High or Low. This algorithm works based on the equationbelow:

Log[/1−y] =$b0$ +$b1x1$+$b2x2$+⋯+$bnxn$

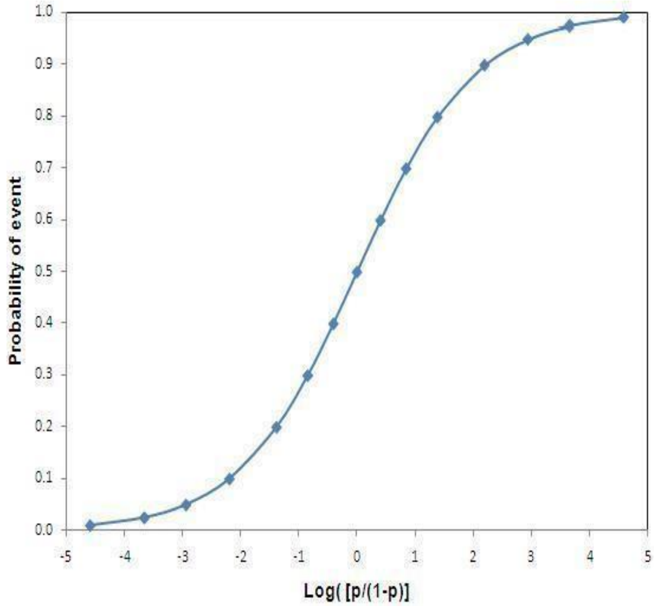

Fig.4.LogisticRegressionCurve

E. ***RandomForestClassifier***

Random Forest algorithm is a machine learning method that is built on the notion ofdecision tree algorithms. The random forest method generates several decision trees. Themore trees there are, the more accurate the detection. The bootstrap technique is used togenerate trees. The characteristics and samples of the dataset are randomly picked using areplacement in the bootstrap approach to form a single tree. Random forest algorithm, likedecision tree algorithm, will identify the bestsplitter for classification from randomlyselected characteristics. Random forest algorithm uses gain index and information gainmethodstodiscoverthebestsplitter.Thiswillcontinueuntiltherandomforesthasproducedntrees.Themethodwillcomputevotesforeachprojectedtargetonceeachtreein the forest forecasts the target value. Finally, the random forest algorithm uses the targetwiththemost 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 monthlyrainfall, various groups of monthly rainfall, and the percentage of floods for that particularyear. The forecast is based on the monthly rainfall for that particular year. An examinationof 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 andJuly.

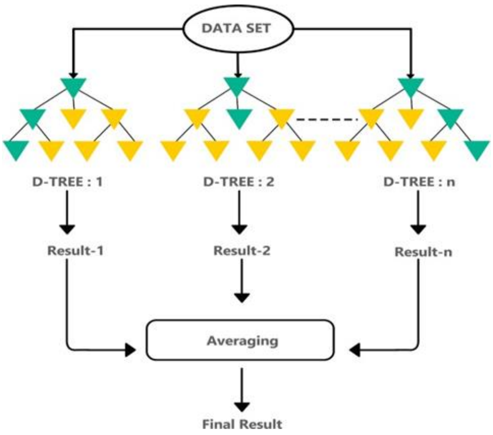

Fig.5.RandomForestClassifier

Fig-5 Shows the rainfall dataset for Kerala state using bar graph, which shows that the peakrainfallusuallycomes injulyand august month.

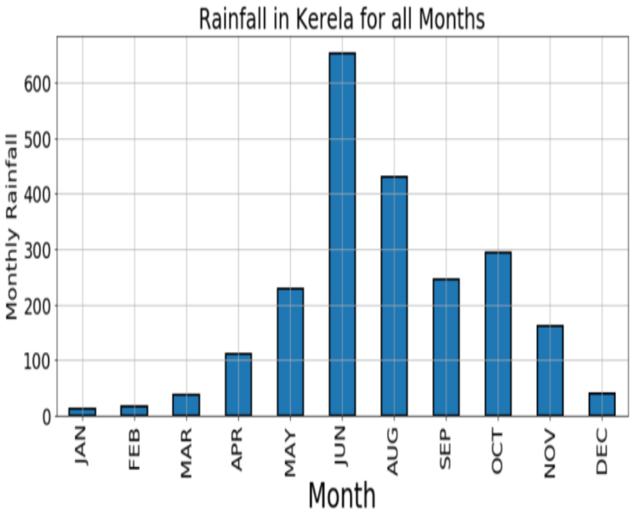

Fig.5. KeralaRainfallChart

IV. RESULTSANDFINDINGS

The accuracy, recall, and ROC score of each prediction model will be assessed. Table Iillustrates 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 aROCscoreofo.90,indicating thattheprecisionandrecallscoresarewellbalanced,implying that the overall performance of Logistic Regression for flood prediction is verygood. The other model DT, has the lowest accuracy (0.62), the least recall (0.60), andROC-scores of 0.60 and 0.63. When we look at the metric scores of the remaining twomodels, the RFC and support vector machine do not demonstrate the predicted efficiency.The Logistic Regression has decisively outperformed the remaining four machine learningmodels in the aforementioned study, making it the best recommendable machine learningmodel for reliable flood prediction. As a result, it will be used to assess the significance offeatures.Although these values canfluctuatedependingon otherstates 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

TableI:Predictionresultsontestdatasets

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

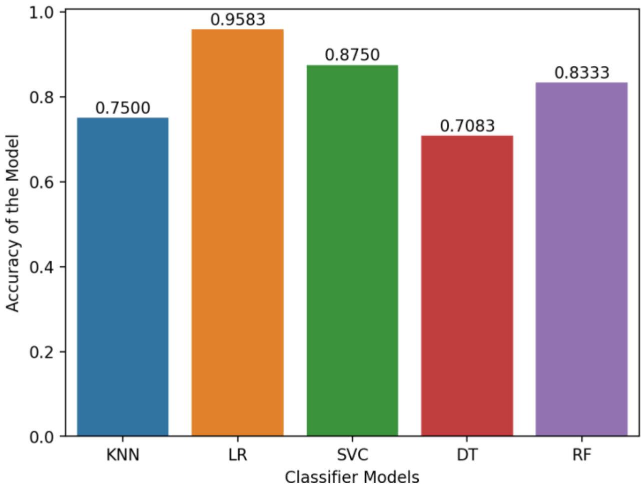

Fig.6.PredictionModelResult

V. CONCLUSIONANDDISCUSSION

Thestudydiscussestheneedforafloodpredictionmodelthatisbasedonmachinelearning. The accuracy, precision, recall, and ROC-score of five different machine learningmodels, including KNN, Logistic Regression, Decision Tree, Random Forest, and SupportVector Machine, were compared. The best model with the greatest metric score is LogisticRegression, according to the results. The work's future focus will be on deep learningmodels and human-machine interaction, with the goal of enabling users to find a solutionthatcanhelppredictfloodinginthefollowingyears.Furtherupgradationcanbebuiltanewsystemthatsendsoutwarningsandalertsofanincomingfloodtothecitizensandhelps save the lives of civilians and if possible, the infrastructure. The system also helps thegovernment save money in rescue operations and helps them start the relocation operationsbeforetheflood hits thetown.

For models, there is room for progress and advancement. Some strategies for improvingmodels include the use of data decomposition techniques to improve the quality of datasetsandtheuseofanensembleofmethodstoimprovemodelgeneralizationandreduceprediction uncertainty. Additionally, add-on optimizers can help increase the quality ofmachinelearningalgorithms.

For future efforts, conducting a survey on spatial flood prediction using machine learningalgorithms is strongly recommended. Increasing the databases for flood location couldpotentiallybeafutureproject.

REFERENCES

1. ChhuonvochKoem,SarintipTantanee,“Flooddisasterstudies:Areviewof remotesensingperspectivein Cambodia”
2. Snehil,RuchiGoel“FloodDamageAnalysisUsingMachineLearningTechniquesFloodDamageAnalysisUsingMachine LearningTechniques”.
3. Ruhhee,Tabbussum,AbdulQayoomDar“Performanceevaluationofartificialintelligence
4. paradigms—artificialneuralnetworks,fuzzylogic,andadaptiveneuro-fuzzyinferencesystemfor flood prediction”.
5. Li-Chiu Chang , Hung-Yu Shen , Yi-Fung Wang , Jing-Yu Huang , Yen-Tso Lin“Clustering-basedhybridinundationmodelforforecastingfloodinundationdepths”.
6. HoJuKeum,kunYeunHan,andHyunIKim“Real-TimeFloodDisasterPredictionSystembyApplyingMachineLearningTechnique”.
7. Rabindra K. Panda ,NiranjanPramanik, BiplabBala “Simulation of river stageusingartificial neural network andMIKE 11 hydrodynamicmodel
8. Mohammed Khalaf ,AbirJaafar Hussain , Dhiya Al-Jumeily , Thar Baker , RobertKeight , Paulo Lisboa, Paul Fergus , Ala S. Al Kafri “A Data Science MethodologyBasedon Machine
9. LearningAlgorithmsforFloodSeverityPrediction”
10. Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article “Assessmentofeconomicflood damage”
11. Wagenaar,D.,deJong,J., andBouwer,L.M.“Multi-variableflooddamagemodellingwithlimited datausingsupervised learningapproaches”

12. Hanghyun Choi, Jeonghwan Kim, Jongsung Kim, DonghyunKimYounghyeBae,2 and HungSooKim "Development of Heavy Rain Damage Prediction ModelUsingMachineLearning Based onBigData".
13. Tripathi,Prakash"FloodDisasterinIndia:AnAnalysisoftrendandPreparedness"
14. Dr. T.PriyaRadhikaDevi "Android Application ForspontaneousSoilconstant Monitoring And Controlling Systemusing Raspberry Pi"Journal Of Critical Review Vol 7 Issue 16.
15. Murali. D, Prasanna. S, Mathavan. V,Priyaradhikadevi. T" Linear Regression And Neural Networks Algorithm To Predicting The Real-Time Parameters Of Temperature And Humidity" Journal of Critical Review