ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

A FLOOD PREDICTION SYSTEM DEVELOPED USING VARIOUS MACHINE LEARNING ALGORITHMS

Mr. K. SUBHASH CHANDRA[1], [1] Assistant Professor, Dept of Computer Science and Engineering, AMRITASAI INSTITUTE OF SCIENCE AND TECHNOLOGY[Autonomous], Andhra Pradesh, India
P.GAYATHRI graduate Student, Dept of Computer Science and Engineering AMRITASAI INSTITUTE OF SCIENCE AND TECHNOLOGY[Autonomous], Andhra Pradesh, India
V SAI JAHNAVI[3] Graduate Student, Dept of Computer Science and Engineering AMRITASAI INSTITUTE OF SCIENCE AND TECHNOLOGY[Autonomous], Andhra Pradesh, India
V SAI KRISHNA[4]Graduate Student, Dept of Computer Science and Engineering AMRITASAI INSTITUTE OF SCIENCE AND TECHNOLOGY[Autonomous], Andhra Pradesh, India
P ESWAR RAO[5]Graduate Student, Dept of Computer Science and Engineering AMRITASAI INSTITUTE OF SCIENCE AND TECHNOLOGY[Autonomous], Andhra Pradesh, India

ABSTRACT
The unusual rainfall and global climate change has led to floods in different parts of the world. Floods are one of the worst affecting natural phenomena which causes heavy damage to property, infrastructure and most importantly human life. To prevent such disasters Machine learning model is created to predict the floods that can occur in the future. It's hard to create a predictive model because of its complexity. In this system the rainfall data is fed into five different Machine Learning models prior to this process, the data is cleaned and preprocessed, the dataset for training is split into Training set and Test set in the ratio of 7:3.

KEYWORDS Flood forecasting · Machine learning · Gradient boost · Decision tree · Random Forest · Android

1 INTRODUCTION
Disaster prevention and prediction Flood prediction using machine learning approach. .Machine Learning algorithms to predict the chances of Flood in the state of Kerala using the Kerela flood dataset. . This Model uses 5 Machine Learning Algorithms namely KNN Classification, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest to get the best possible model to predict the floods using Kerela Rainfall Data. . A dataset with the amount of rainfall and if a flood had occured in a particular area/state/city, in the previous years, will be used. The dataset will have the rainfall data for a duration of 3 months approx. Using this dataset, we take average rainfall for every 10 days and plot it on a graph to visualize it. We take this average data of rainfall, as input to our machine learning model and if it causes a flood or not as the output labels. We train our model and save it.(depending on some threshold value of average rainfall in the dataset).Given the input data, for consecutive 10 days, we give this data as an input, and let the model predict, if whether there is a possibility of flooding or not, by setting some threshold in the training data. Our basic approach for this problem is binaryclassification, using basic machine learning algorithms (linear regression or logistic regression).This approach can be made real time prediction and accuracy can be improved with adding more features such as the type of land in that area, the location of the area etc.

2 LITERATURE SURVEY
Yovan Felix and T. Sasipraba [2] have developed a flood warning system for which data has been collected from remote sensing satellites and ground application. The parameters considered are the amount of rainfall and water-level of nearby water bodies. Gradient Boost Algorithm [3] is used to obtain a non-linear relationship between the total sum of rainfall and runoff, thus reducing the mean-

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

squared error. For the datasets which were not a part of the training dataset, the Decision Tree Algorithm is used to predict floods on datasets which were not used in the training dataset. The architecture is divided into three layers which are interconnected. The three layers are the Physical Layer, Network Layer and Application Layer. Physical Layer collects the data from sensors and the Application Layer is the user interface which is the Mobile/Desktop Application. The historical dataset was collected from the Meteorological Department and Flood Forecasting Commission. At an interval of an hour, the realtime data is collected and sent to the Network Layer using the GSM Module and stored on the Cloud Server to be passed to the Machine Learning Model.Ding et al. [4] forecasted floods in the region of the Lech river basin in Europe. Using a Spatio-Temporal Attention Long Short-Term Memory (STA-LSTM) Model, floods were forecasted. This is a data-driven model which sets up a connection between verifiable hydrological features and the runoff. LTSM shapes the fundamental piece of the neural network framework and is the adjustment of the Recurrent Neural Organization (RNN). Using the attention model takes out the issue of the impact of the equivalent hydrological highlights on different floods. It likewise analyzes the precision of the STA-LSTM with theSupportVectorMachin(SVM)

Fully Connected Network (FCN) and the original LSTM [5]. In the STA-LSTM Model, the skewness calculated for the training dataset is 0.58 and for the test dataset is - 0.0651. At the point when the correlation at T+3 was done, the FCN model fared better compared to the SVM, LSTM and STA-LSTM network model. Yet, at T+6 and T+9, the STALSTM Model performs comparatively the best and FCN performs out the least.Using the Sparse Bayes Model, Yirui Wu, Yukai Ding and Jun Feng [6] carried out flood prediction experiments inChanghua River. SMOTE [7] algorithm eliminates the issue of lopsided sample distribution; a Sparse Bayesian model is trained using AdaBoost Methodology which improves the model's performance in over-fitting. Using a group of Sparse Bayesian models accomplishes a high accuracy when compared with a single Sparse Bayesian Model. It was seen that the model performs better compared to the single model. It was likewise presumed that the testing limit plays a significant part in deciding the performance of the model. The dataset for "Sparse Bayesian Flood Forecasting Model based on SMOTEBoost" is the yearly summer flood information of Changhua river basin from 1998 to 2010. The real-time data is recorded each hour. The data credits consist of the Changhua stream and rainfall, and the rainfall of the stations in the upper ranges of Changhua.An urban flood estimating and checking platform created as a part of a UK Newton Fund project in Malaysia by Karyotis et al. [8] utilizes a hybrid Deep Learning (DL) and Fuzzy Logic (FL) based algorithm. This model uses low-cost sensors to gather real-time data. DL [9] utilizes Artificial Neural Organizations (ANN) for both supervised and unsupervised training, giving a dependable arrangement in time forecasting issues. FL, which depends on the idea of fuzzy sets, can deal with "fractional truth". The most ideal size of the data window is 200 data points for the Deep Learning Model. It was additionally seen that if the rain intensity is high, storm span is high and soil absorption is exceptionally low at that point likelihood of flood is high.Dola et. al. [10] have developed a system to predict the occurrence of floods by using Machine Learning models. The data of rainfall from previously available data is used to predict the rainfall for the next month. Forecasting can be done for both short-term and long-term rainfall. Data has been gathered from the Indian Meteorological Department. Two distinctive datasets that comprise normal rainfall data from 1951-2000 for each month and district; the subsequent information is of 19012015, which comprises average rainfall data for each state. This Low Cost IoT based Flood Monitoring System employs IoT to figure out the time it would take for the flood to reach land. Severity of the rainfall is estimated using ML algorithms. The algorithms applied for the same are Linear Regression, Support Vector Machine and ANN. The various IoT devicesused are rain-drop sensors, water-float sensors and IoT Gecko. As and when the water level rises, a buzzer beeps and an alert is sent of an approaching flood in the area. For the linear regression model, the dataset of the last three months is taken to predict the rainfall for the next month.

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

It is the same for SVM. For ANN, CNN 1-D [11] strategy is applied. The mean absolute error for the linear regression algorithm is 40.2467874. The SVM model gave a mean absolute error of 90.606787. A mean absolute error of 21.8097545 was obtained for ANN.

3 EXISTING SYSTEM

The Flood Detection and Warning System (FLoWS) [12] is the progression taken by the Malaysian government to help prevent the genuine damage caused to houses, streets, organizations, public offices and individuals by the annual floods. It helps in monitoring and managing this critical circumstance by giving crucial data like flood conditions, plan and preparation, etc. to the general population and the local authorities at the affected territory. The system is able to measure the water level and alarm people in general and the local authorities by sending a warning through SMS and MMS in regards to the flood conditions. The system additionally empowers general society and the local authorities to see the live graph data of the water level using an Android application. It uses an ultrasonic sensor which quantifies the water level. The Raspberry Pi 3 goes about as a server to process and store every output from the microcontroller and use this data to trigger the Raspberry Pi camera to capture the image of the flood situation. The microcontroller gathers the data including water distance, temperature, humidity, and flood level. Then, the GSM SIM 900/900A is liable for sending the data from the microcontroller to the server utilizing AT command. The GSM is likewise responsible for sending warning messages, flood levels that have been measured from the sensor and image to the targeted mobile phone.Global Flood Monitoring System (GFMS) is a computer tool which can be utilized for mapping flood conditions around the world. Created by Robert Adler and Huan Wu of the University of Maryland, it is utilized by zooming into an area of interest on the system's global interactive guide to see whether the water is at flood stage, subsiding, or rising. It can also be utilized to figure out whether there is a rain event upstream, regardless of whether the rain is finished, and how the water is moving downstream. GFMS works 24x7, in any event, when there is cloud cover or other impedance. It depends on precipitation data from NASA's Earth observing satellites. Precipitation data from GFMS is joined with a land surface model that fuses vegetation cover, soil type, and terrain to decide how much water is absorbing and what amount is feeding the
streamflow, water profundity, and flooding every 3 hours. Users can likewise zoom in further to see inundation maps as fine as 1 km goal.

Dis advantages of Existing system

- The ml approach proved to be successful in the detection of floods under the cornfields where radar waves do not penetrate through the crop down to the flooded grounds - where the other two methods failedSince no technique is employed in identifying the rainfall lot of damage is happening for a region if it comes

PROPOSED SYSTEM

The aim of this project is to get all the rainfall data of India and from a dataset containing yearly rainfall data. By providing real time input to different models of machine learning, those are Logistic Regression, Support Vector Machine, K-Nearest Neighbors and Decision Tree Classifier. The input provided to models are pre-processed and patterns are extracted by getting maximum accuracy. The data provided is split into a Training set and Test set. It is split in the ratio of 7:3. The all four models are used to predict and by comparing all the results of model and considering the confusion matrix of all the models the accuracy is determined. The best model is chosen by comparing the accuracy of each mode

Advantages Proposed system

- Policy suggestion.
- Minimization of the loss of human life.
- Risk reduction.
- Rapid flood mapping.

DOI10.36893.JK.2023.V13I04N16.001-0003

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

- Reduction of property damage accociated with floods

4 DATASET

This Decision Tree model helps predict floods in the districts of Bihar and OrissainIndia.Thedistricts. The data collected is from the year 1992-2002. The data points collected during this period are 2640. The dataset was collected from a verified website called the India Water Portal [13]. The dataset had to be downloaded in a district-wise manner in a CSV format. The dataset collected has been gathered on a monthly basis

	A	B	C	D	E	F	G	H	I	J
1	LOCATION	YEAR	MONTH	MIN_TEMP	MAX_TEMP	RAINFALL	CLOUD_COVER	WET_DAY_FREQ	DIURNAL_TEMP	FLOOD OCCURRENCE
2	BIHAR-PATNA	1992	JANUARY	8.067	22.456	20.727	22.672	1.5696	14.389	NO
3	BIHAR-PATNA	1992	FEBRUARY	9.776	24.752	5.904	22.846	1	14.953	NO
4	BIHAR-PATNA	1992	MARCH	17.734	34.124	1.126	29.132	0.9778	16.364	NO
5	BIHAR-PATNA	1992	APRIL	22.842	38.889	4.345	25.191	1	16.021	NO
6	BIHAR-PATNA	1992	MAY	24.716	39.157	8.737	34.006	1.5548	14.394	NO
7	BIHAR-PATNA	1992	JUNE	27.115	37.999	33.17	58.664	3.6736	10.871	NO
8	BIHAR-PATNA	1992	JULY	26.981	34.403	191.148	73.11	11.17	7.423	NO
9	BIHAR-PATNA	1992	AUGUST	25.733	32.456	154.266	68.575	9.8425	6.721	NO
10	BIHAR-PATNA	1992	SEPTEMBER	25.44	33.177	103.491	57.506	7.2156	7.725	NO
11	BIHAR-PATNA	1992	OCTOBER	21.875	32.476	66.13	34.832	3.4364	10.598	NO
12	BIHAR-PATNA	1992	NOVEMBER	15.592	29.46	0.875	20.209	0.5023	13.845	NO
13	BIHAR-PATNA	1992	DECEMBER	10.132	24.75	0	22.926	0	14.62	NO
14	BIHAR-PATNA	1993	JANUARY	10.31	24.721	10.156	22.672	1.0301	14.389	NO
15	BIHAR-PATNA	1993	FEBRUARY	13.85	28.803	0.2	22.602	0.2	14.953	NO
16	BIHAR-PATNA	1993	MARCH	16.233	32.62	10.422	29.132	1.2841	16.364	NO
17	BIHAR-PATNA	1993	APRIL	21.817	37.861	4.268	25.191	1.002	16.021	NO
18	BIHAR-PATNA	1993	MAY	26.417	40.812	24.864	34.144	2.3305	14.394	NO
19	BIHAR-PATNA	1993	JUNE	28.39	39.262	37.736	58.715	4.0813	10.871	NO

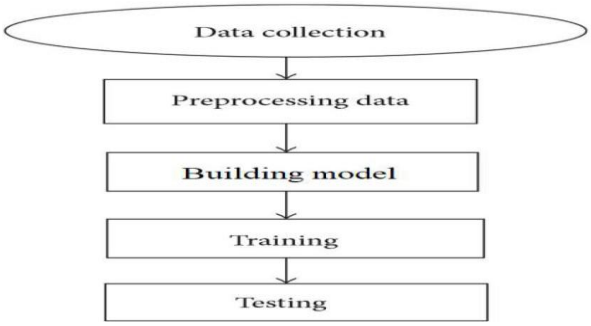
All the districts then had to be compiled together

METHODOLOGY

using data validation, and the results are compared to get accuracy. The accuracy of the training dataset, accuracy of the testing dataset, false-positive rate, specification, precision, and recall are calculated by comparing algorithms using python code.

The steps involved are:

- Define a problem
- Preparing data
- Evaluating algorithms
- Predicting results
- Predicting results



ALGORITHMS AND TECHNIQUES

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it Logistic Regression may be a machine learning algorithm that predicts the probability of a categorical variable. It is a statistical way of analyzing a group of knowledge that comprises quite one experimental variable that determines the result. dichotomous variable

DOI10.36893.JK.2023.V13I04N16.001-0003

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

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

Random forests:
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

4 RESULTS:

Accuracy is a measurement unit used to evaluate the Machine Learning Algorithm. It indicates the percentage. The Decision Tree Algorithm gave an accuracy of 94.4%. The Gradient Boost Algorithm gave an accuracy of 87.9% whereas the Random Forest Algorithm gave an accuracy of 92.4%. Hence, the Decision Tree Algorithm was chosen for the model.The scatter plot is a kind of mathematical graph in which two particular variables are mapped along the x-axis and y-axis. The resulting pattern reveals the correlation present between the two variables. For this particular model, the two variables chosen are Temperature and Rainfall. Temperature has been plotted along the y-axis and Rainfall has been plotted along the x-axis. The prediction of flood is heavily dependent on these two variables. Temperature has been calculated as an average of the minimum and maximum temperature mentioned in the dataset. The red points inside the red region indicate that the prediction of flood occurrence of those points at that particular temperature and rainfall has been done correctly by the model. The green points inside the red region are the points which have been predicted wrong. The model predicted 'NO' for thesegreen points inside the red region when the model should have predicted 'YES'. The green points inside the green region have been predicted correctly for the flood occurrence and the red points inside the green region have been predicted incorrectly.

Fig 1: Snapshot of Dataset

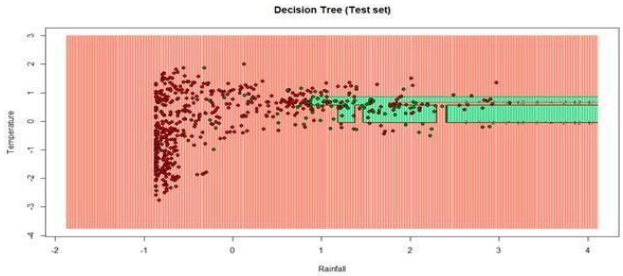


Fig. 3: Result of Decision Tree Algorithm

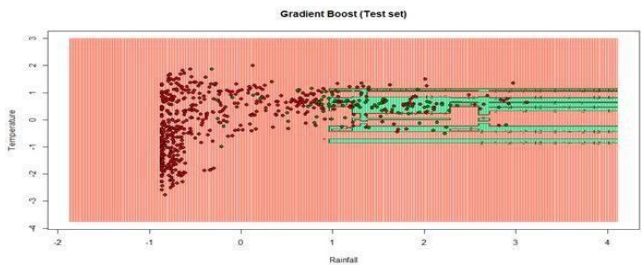


Fig. 4: Result of Gradient Boost Algorithm

DOI10.36893.JK.2023.V13I04N16.001-0003

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

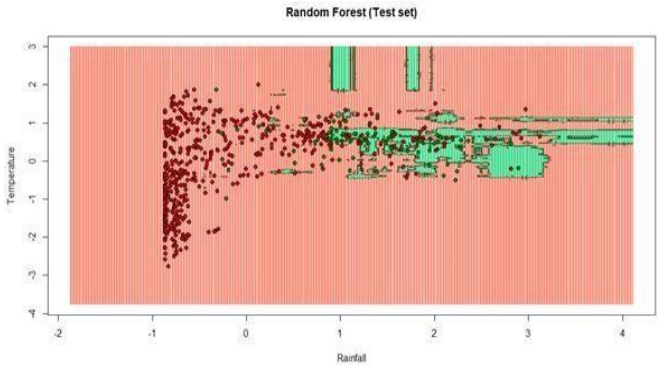


Fig. 5: Result of Random Forest Algorithm

The Decision Tree Algorithm has the highest accuracy. Figure
3 shows the scatter plot of the Decision Tree Algorithm depicting the accuracy. Figure 4 depicts the scatter plot of Gradient Boost Algorithm. The Confusion Matrix of the Gradien Boost Algorithm s False Negatives (FN), 36 False Positives (FP) and 50 True Negatives (TN). Figure 5 depict the scatter plot of the Random Forest Algorithm. The Confusion Matrix of the Random Forest Algorithm showed 553 True Positives (TP), 17 False Negatives (FN), 33 False Positives (FP) and 57 True Negatives (TN).The other existing model which uses Linear Regression, SVM and ANN gave a mean absolute error of 40.24, 90.61 and 21.81 respectively for the three algorithms. The Decision Tree Model explained in this paper gave a mean absolute error of 0.05606. Mean Absolute Error is another way of calculating accuracy for a Machine Learning Model. The lower the value of mean absolute error, the better the model's performance. Hence, the Decision Tree Model fared better when compared to the other proposed systemThe Android Application helps notify citizens of an imminent danger. It also helps the government predict floods and initiate rescue and relocation operations. The various activities included in the Android Application are Menu Activity, Flood Emergency Steps Activity, Issue Common Alerts, Issue Government Activity, Issue Alerts Menu Activity and Issue Government Alert Activity.

5 CONCLUSION AND FUTURE SCOPE

In the Decision Tree Machine Learning Algorithm, the parameters collected using an architectural set-up allows seamless integration of data. This data is then fed onto a Machine Learning model which is then able to predict the chances of flood. The proposed framework performs analysis with a high and satisfactory fault-tolerant accuracy.The system has also been built according to the conditions prevalent in a country like India. The system 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 townn the future, a collaboration between the forecast of rainfall and flood can be achieved. Using satellite imaging, the civilians can also be informed of safe places that they can relocate to and guide them towards the rehabilitation camps set up by the government.

REFERENCES

[1] Flood prediction using machine learningmodels: literature review amir mosavi 1,*, pinar ozturk 1 and kwok-wing chau 21 department of computer science (IDI), norwegian university of science and technology(NTNU), trondheim, NO-7491, norway2 department of civil and environmental engineering, hong kong polytechnic university,hong kong, china;
[2] dr.Kwok- wing.Chau@polyu.Edu.Hk.
[3] Https://un-spider.Org/links-and-resources/data- sources/daotm-floods-ml

DOI10.36893.JK.2023.V13I04N16.001-0003