AI- BASED FLOOD RESCUE SYSTEM

ABSTRACT

Floods are among the most destructive natural disasters, which are highly complex to model. In this paper we are mainly focused on flood prediction in kerala regarding risk reduction, policy suggestion, minimization of the loss of human life, and reduction the property damage associated with floods. machine learning (ML) methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions .So we propose to build and develop ML systems for timely and accurate flood prediction.

INTRODUCTION

Among the natural disasters, floods are the most destructive, causing massive damage to human life, infrastructure, agriculture, and the socioeconomic system. Governments, therefore, are under pressure to develop reliable and accurate maps of flood risk areas and further plan for sustainable flood risk management focusing on prevention, protection, and preparedness . Flood prediction models are of significant importance for hazard assessment and extreme event management.

Mainly 3 types of Flood is there,

1.M**inor flooding:** It is defined to have minimal or no property damage, but possibly some public threat.

2.Moderate flooding : It is defined to have some inundation of structures and roads near the stream. Some evacuations of people and/or transfer of property to higher elevations may be necessary.

3.Major flooding : It is defined to have extensive inundation of structures and roads. Significant evacuations of people and/or transfer of property to higher elevations are necessary

DATA PREPARATION

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data

Kerala State has an average annual precipitation of about 3000 mm. The rainfall in the State is controlled by the South-west and North-east monsoons. About 90% of the rainfall occurs during six monsoon months. The high intensity storms prevailing during the monsoon months result in heavy discharges in all the rivers. The continuous and heavy precipitation that occurs in the steep and undulating terrain finds its way into the main rivers through innumerable streams and water courses. Kerala experienced an abnormally high rainfall from 1 June 2018 to 19 August 2018. This resulted in severe flooding in 13 out of 14 districts in the State. As per IMD data, Kerala received 2346.6 mm of rainfall from 1 June 2018 to 19 August 2018 in contrast to an expected 1649.5 mm of rainfall. This rainfall was about 42% above the normal. Further, the rainfall over Kerala during June, July and 1st to 19th of August was 15%, 18% and 164% respectively, above normal.

So mainly in this paper taking 4 major area of kerala which was most effected by flood

They are,

1. IDUKKI
2. MUVATTUPUZHA
3. PERIYAR
4. KALLADA

From the respective area collected the data regarding temperature , humidity ,rainfall and from the latitude and longitude just calculate the altitude

Mainly the flood is effected due to the following KPI's (KEY PERFORMANCE INDICATORS)

1. Temperature
2. Humidity
3. Altitude
4. Rainfall

DATA COLLECTION

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Fig 1

year	month	station	Temperature(c)	Humidity(gm/ml3)	Altitude(m)	Rain(mm)
2018	jan	MUVATTUPUZHA	29.05	90.13	24	1579.21
2018	feb	MUVATTUPUZHA	28.84	78 .28	24	336.05
2018	mar	MUVATTUPUZHA	31.97	80.21	24	517.58
2018	apr	MUVATTUPUZHA	25.9	66.43	24	59
2018	may	MUVATTUPUZHA	27.32	59.66	24	145.44
2018	jun	MUVATTUPUZHA	29.61	69.27	24	323.91
2018	july	MUVATTUPUZHA	28.98	61.4	24	424.99
2018	aug	MUVATTUPUZHA	28.21	56.22	24	126.85
2018	sep	MUVATTUPUZHA	26.91	84.52	24	851.74
2018	oct	MUVATTUPUZHA	26.21	62.62	24	146.83
2018	nov	MUVATTUPUZHA	26.25	84.04	24	932.41
2018	dec	MUVATTUPUZHA	25.88	66.72	24	328.69

The data here we are using is time series data (a time **series** is a sequence taken at successive equally spaced points in time and use of a model to predict future values based on previously observed values

PROPOSED SYSTEM

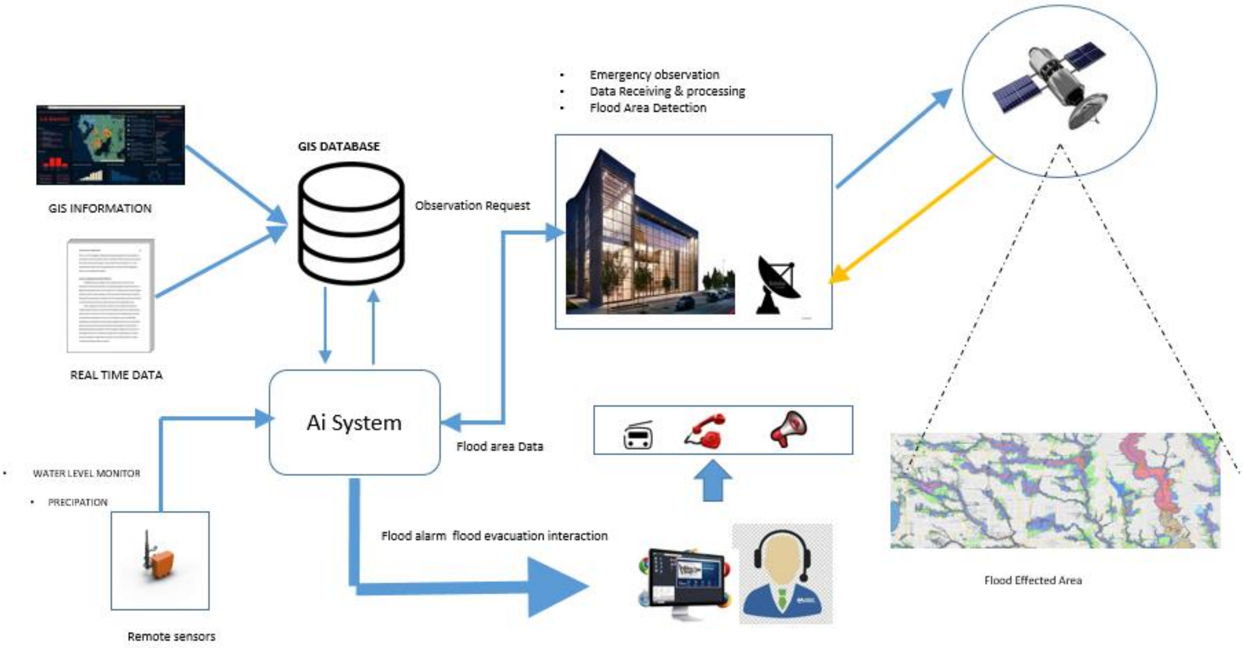
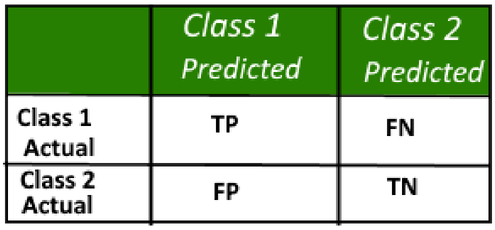


Fig 2

BASIC TERMS

CONFUSION MATRIX:

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Here,

- Class 1 : Positive
- Class 2 : Negative

Definition of the Terms:

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Classification Rate/Accuracy:

Classification Rate or Accuracy is given by the relation:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall:

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN).

Precision:

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP).

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure:

Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.

The F-Measure will always be nearer to the smaller value of Precision or Recall.

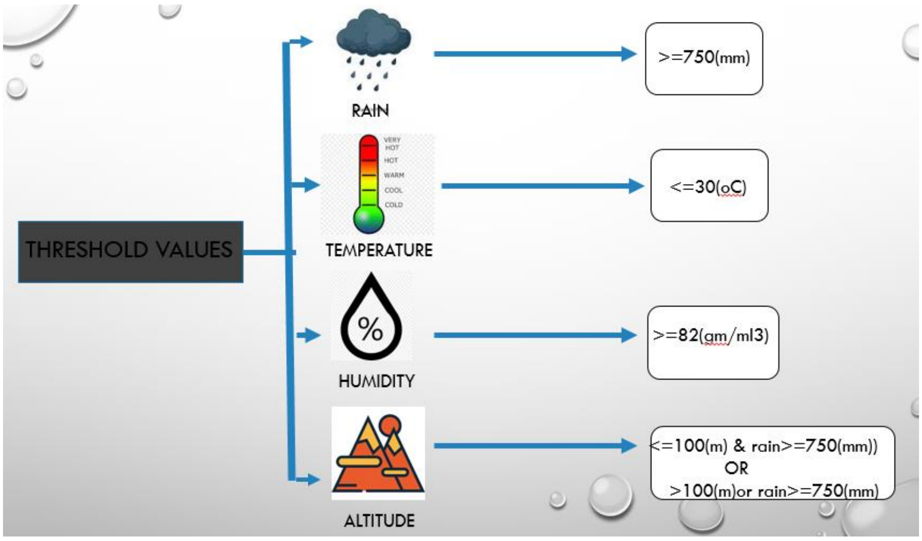
$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

ROC CURVE

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or *probability of detection* in machine learning. The false-positive rate is also known as *probability of false alarm* and can be calculated as (1 − specificity). It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting

the cumulative distribution function (area under the probability distribution from to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.

THRESHOLD VALUES

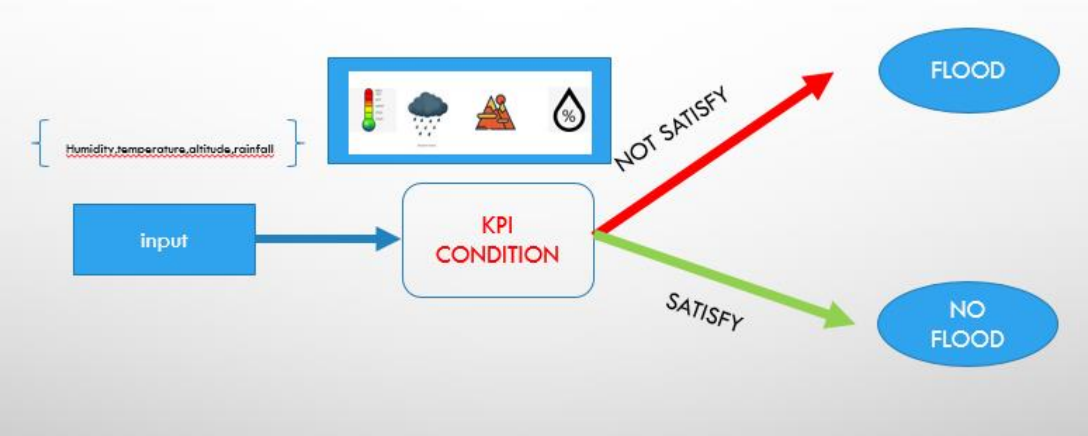


METHODS

1. LOGISITIC REGRESSION

This type of statistical analysis (also known as *logit model*) is often used for predictive analytics and modeling, and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite

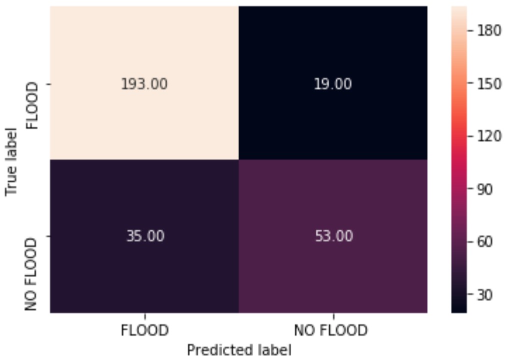
options A, B, C or D (multinomial regression). It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.
In this specific problem taking humidity,temperature,rain and altitude as independenet variable
And predict there is a flood or not as show below



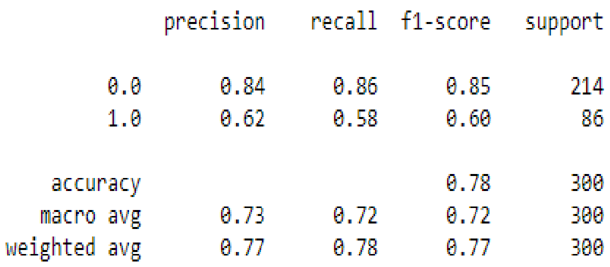
2. MULTILAYER PERCEPTRON

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). ... MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Below is the result of applying Multilayer perception



Confusion matrix



```
              precision    recall  f1-score   support

         0.0       0.84      0.86      0.85       214
         1.0       0.62      0.58      0.60        86

    accuracy                           0.78       300
   macro avg       0.73      0.72      0.72       300
weighted avg       0.77      0.78      0.77       300
```

classification report

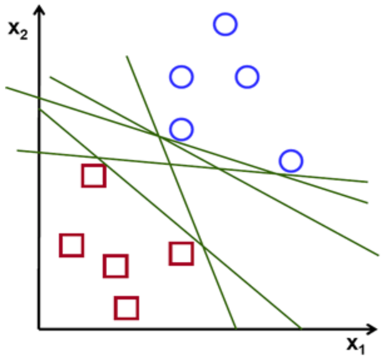
3.SUPPORT VECTOR MACHINE

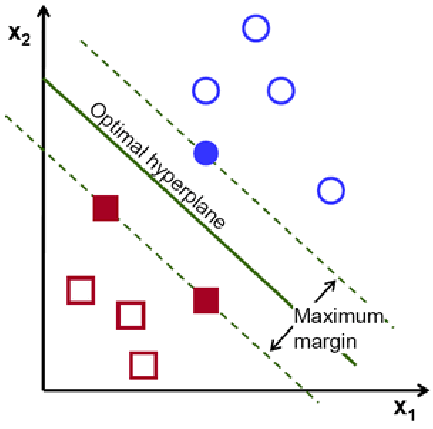
A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis. ... SVMs are used in text categorization, image classification, handwriting recognition and in the sciences. A support vector machine is also known as a support vector network (SVN)

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.
To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.





● ■ Indicates the support vectors

○ Indicates FLOOD

□ Indicates NO-FLOOD

4.K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors

K-means is a clustering algorithm that splits a dataset as to minimize the euclidean distance between each point and a central measure of its cluster. Typically, Knn works this way: ... Select the k training cases that have the smallest distance and look at their classification. These are the k Nearest Neighbors, or Knn

In this flood prediction system the effect on increasing the neighbors(k) to the accuracy listed below,

Number of neighbors	ACCUARCY
K=1	0.97
K=2	0.968
K=3	0.962
K=4	0.961
K=5	0.945

5.NAÏVE BAYES-GAUSSIAN

Gaussian Naive Bayes is an algorithm having a Probabilistic Approach. It involves prior and posterior probability calculation of the classes in the dataset and the test data given a class respectively.

In machine learning we are often interested in selecting the best hypothesis (h) given data (d).

In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (d).

One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem. Bayes’ Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.

Bayes’ Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

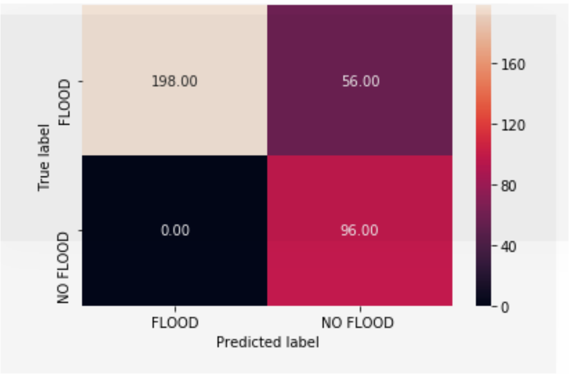
Where

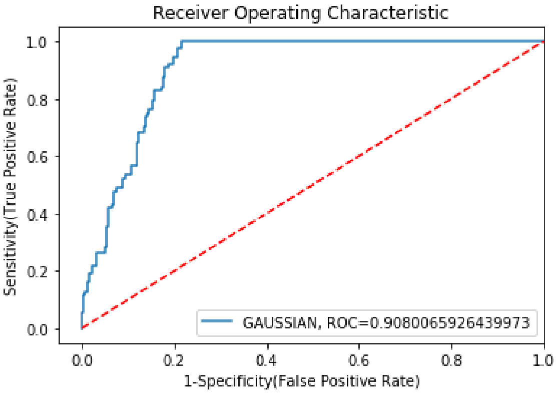
- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.

- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis).

You can see that we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h).

After calculating the posterior probability for a number of different hypotheses, you can select the hypothesis with the highest probability.





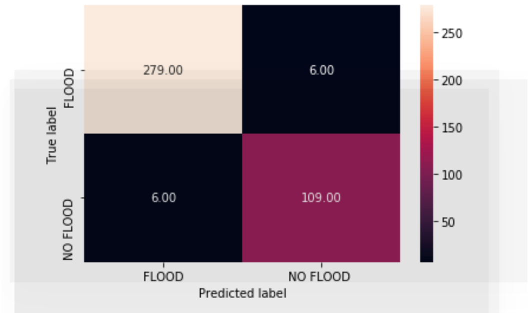
Confusion matrix

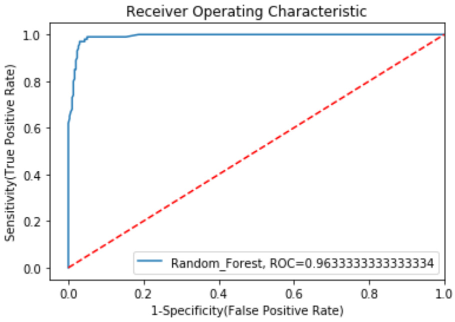
classification report

6.RANDOM FOREST

A random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree





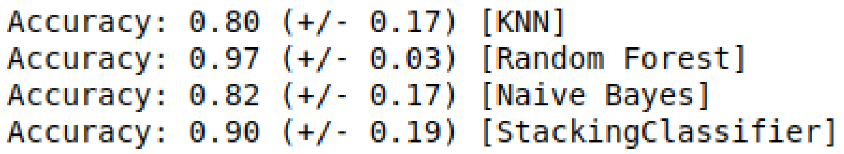
Confusion matrix

classification report

STACKING

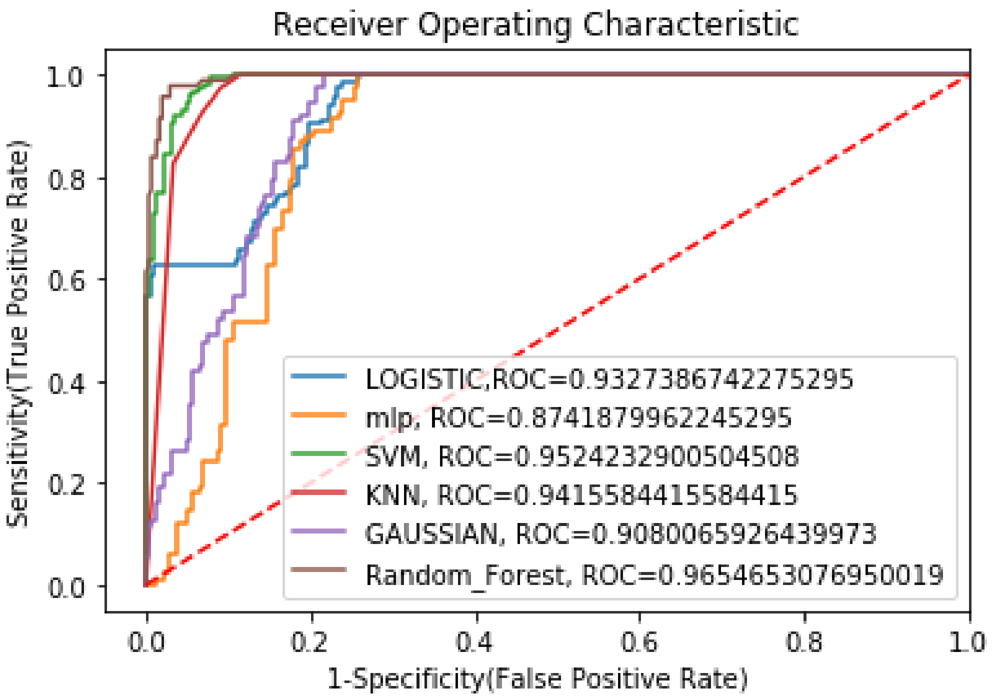
Stacking. Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features

So for our problem here the meta classifier used is Logistic Regression and the variation in accuracy shown below

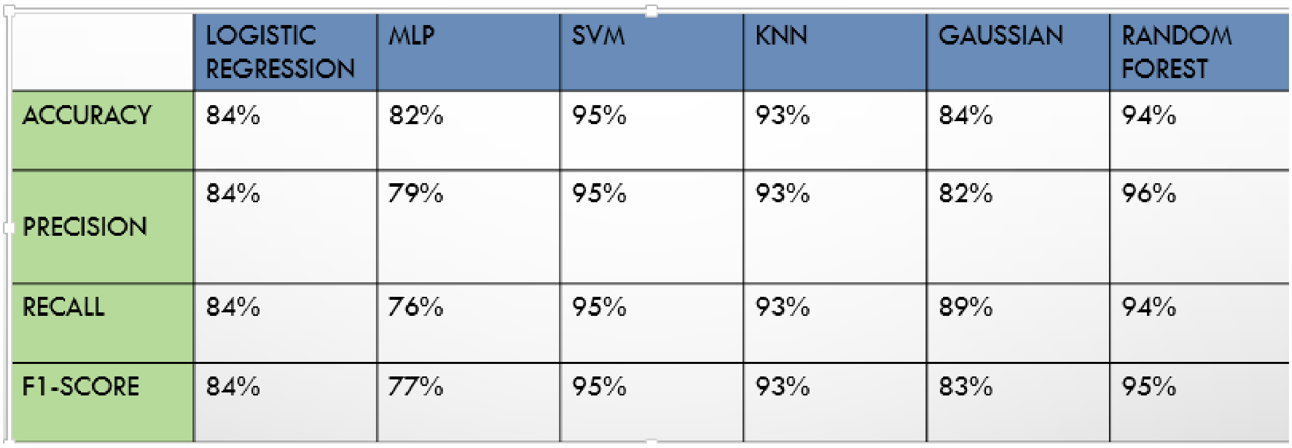


```
Accuracy: 0.80 (+/- 0.17) [KNN]
Accuracy: 0.97 (+/- 0.03) [Random Forest]
Accuracy: 0.82 (+/- 0.17) [Naive Bayes]
Accuracy: 0.90 (+/- 0.19) [StackingClassifier]
```

1.COMPARSION OF ROC CURVE



2.CLASSIFICATION REPORT



	LOGISTIC REGRESSION	MLP	SVM	KNN	GAUSSIAN	RANDOM FOREST
ACCURACY	84%	82%	95%	93%	84%	94%
PRECISION	84%	79%	95%	93%	82%	96%
RECALL	84%	76%	95%	93%	89%	94%
F1-SCORE	84%	77%	95%	93%	83%	95%

CONCLUSION

The flood prediction we get great accuracy when using Random forest essemble method.So predictive performance can compete with the best supervised learning algorithm it provide a reliable feature importance estimate and offer efficient estimates of the test error without incurring the cost of repeated model training associated with cross-validation

REFERENCE

1. Flood Prediction Using Machine Learning
(https://www.researchgate.net/publication/328562202_Flood_Prediction_Using_Machine_Learning_Models_Literature_Review)

2. Urban Pluvial Flood Forecasting using Open Data with Machine Learning Techniques in Pattani Basin
(https://www.sciencedirect.com/science/article/pii/S1877050917323979)

3. Flood modelling using Artificial Neural Network
(https://ieeexplore.ieee.org/abstract/document/6653287)

4.Smart flood disaster prediction system using IoT & neural networks
(https://ieeexplore.ieee.org/document/8358367)