IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Flood Forecasting In Kerala Using Machine Learning Techniques

Reena.S [1], Lecturer in Computer Engineering, Government Polytechnic College Nedumangad

Dipu Jose [2],Lecturer in Computer Engineering, Government Polytechnic College Nedumangad

**Abstract**

Flood forecasting is a critical component of disaster management, especially in regions like Kerala, which are prone to recurring floods. This study explores the application of machine learning techniques to enhance the accuracy and timeliness of flood predictions. By utilizing historical rainfall patterns various machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Linear Regression (LR) are developed and evaluated. The models are tested for their ability to predict flood events under diverse rainfall conditions, ranging from light to heavy rainfall scenarios. The results demonstrate the potential of machine learning to provide reliable and accurate flood forecasts, offering valuable insights for disaster preparedness and mitigation in Kerala.

Keywords: Flood forecasting, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Linear Regression (LR), Confusion Matrix

## 1.Introduction

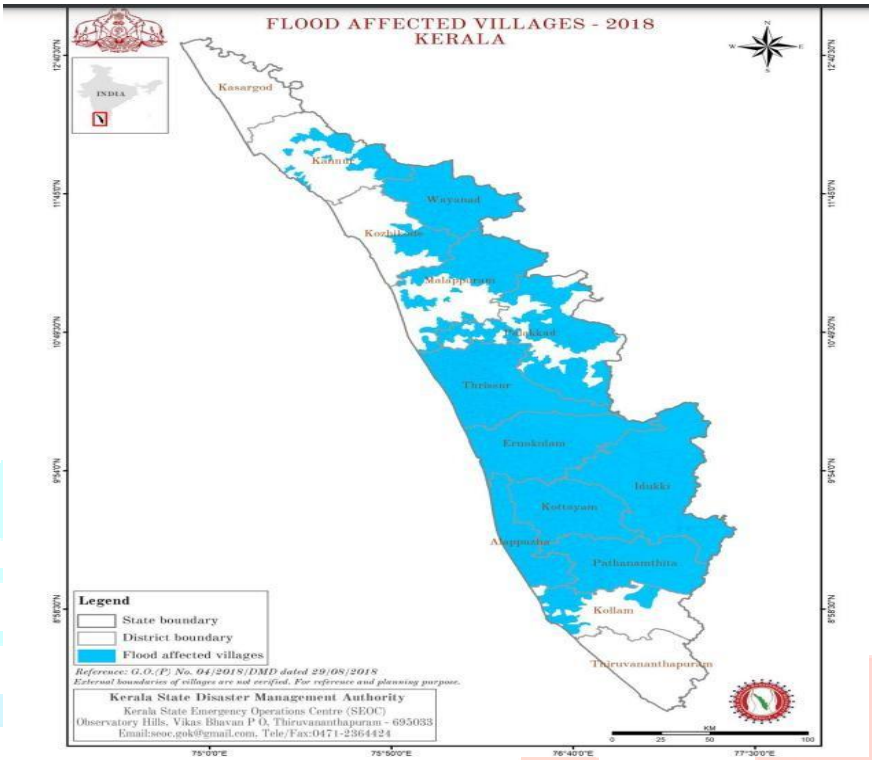
Flood forecasting plays a critical role in mitigating the devastating impacts of floods by enabling timely warnings and effective disaster management strategies. Kerala, a region prone to recurrent and severe flooding, necessitates the implementation of advanced predictive models to safeguard lives, infrastructure, and natural resources. Traditional methods of flood prediction often rely on hydrological and statistical models, which may lack the ability to efficiently handle large datasets and complex nonlinear relationships between climatic and hydrological variables.

Machine learning (ML) techniques offer a promising alternative, leveraging their ability to process vast amounts of data, identify patterns, and make accurate predictions. By integrating historical climatic records, rainfall patterns, streamflow data, and topographical information, machine learning models can significantly enhance the accuracy and reliability of flood forecasts. Techniques such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and deep learning frameworks like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been widely adopted for flood forecasting in various regions.

This study focuses on the application of machine learning models for flood forecasting in Kerala, with an emphasis on addressing the unique climatic and geographical challenges of the region. By exploring the

potential of these advanced techniques, the research aims to contribute to the development of efficient and accurate flood prediction systems tailored to Kerala's needs.

## 2. Motivation




Kerala is quite prone to floods because of the erratic and long monsoon season. In light of these it is important that kerala develop a strong flood forecasting system to prevent loss of life and property. Around 1259 Villages from 14 Districts of Kerala has been declared as Flood affected during the South West Monsoon 2018.Rainfall in August 2018 in Kerala was 96 per cent above the long-term average in the State. There were 433 deaths owing to the Kerala floods and the total economic losses were estimated at Rs. 31,000 crore.

## 3. Objectives

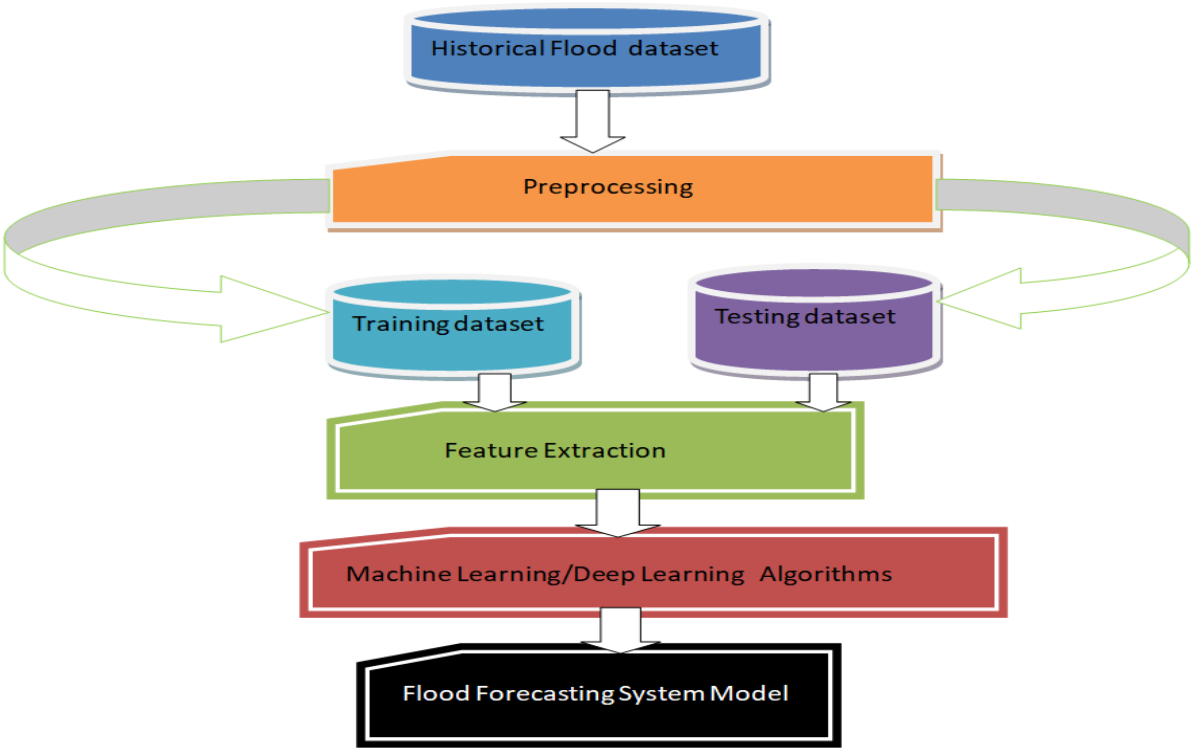
- To identify and develop suitable flood forecasting models with more accuracy by analyzing historical flood data using machine learning and deep learning algorithms
- To develop models for improving the disadvantages of traditional flood prediction methods
- To analyze the results with the present day methods and a case study if possible
- Forecast the chances of flood and communicate the same to general public so that ade quate measures can be taken by authorities.

## 4.Literature Review

Support Vector Machines (SVMs) are utilized for both classification and regression tasks by identifying a hyperplane that optimally separates the data into distinct classes. SVMs demonstrated diverse responses to varying rainfall inputs, with lighter rainfall yielding noticeably different results compared to heavier rainfall [1]. Integrating SVM with Kernel Principal Component Analysis (KPCA) and a boosting algorithm in a flood forecasting model can substantially improve prediction accuracy [2]. Flood events are classified or predicted

by comparing new observations with the knearest neighbors in the dataset. To improve flood prediction accuracy, various correlation coefficients are employed for feature selection, in combination with the k-nearest neighbors (k-NN) algorithm [3]. Regression analysis was performed to examine the relationship between weighted maximum rainfall and maximum streamflow during flood events in the River Basin. Equations were derived using annual maximum daily rainfall, streamflow, and catchment area to assign flood prioritization ranks to each catchment [4]. Additionally, an SMS-based flash flood warning system was implemented, delivering advanced alerts with predictive information on rising water levels and flow velocity [5].

## 5.System Model



## 6. Technologies used

- The entire analysis and Machine Learning model building is done using Python Programming Language in Jupyter Notebook IDE.

**Library of Python used are**

- **Pandas**

  Pandas is a software library written for the Python programming language for data manipulation and analysis.Pandas is a popular open-source data manipulation and analysis library for the Python programming language. It provides easy-to-use data structures and functions needed to manipulate structured data

- **Numpy**

NumPy is a powerful numerical computing library in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these elements.

- **Matplotlib**

Matplotlib is a popular 2D plotting library for Python. It enables the creation of static, animated, and interactive visualizations in Python. Matplotlib provides a wide range of plotting functions and customization options, making it a versatile tool for data visualization

- **Scikit Learn**

Scikit-learn, often referred to as sklearn, is a widely-used machine learning library in Python. It is built on NumPy, SciPy, and Matplotlib, and it provides simple and efficient tools for data analysis and modeling. Scikit-learn includes a variety of machine learning algorithms for tasks such as classification, regression, clustering

- **Seaborn**

Seaborn is a data visualization library built on top of Matplotlib in Python. It provides a high-level interface for creating attractive and informative statistical graphics. Seaborn is particularly well-suited for visualizing complex datasets with multiple variables.

## 7. DATASET

The dataset has in total 118 data points from year 1901 to 2018, which contain rainfall data of 12 months and annual rainfall together. Apart from this it also contains the past flood records.

**Historical Rain fall dataset of Kerala**

| | SUBDIVISION | YEAR | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | \ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | KERALA | 1901 | 28.7 | 44.7 | 51.6 | 160.0 | 174.7 | 824.6 | 743.0 | 357.5 | |
| 1 | KERALA | 1902 | 6.7 | 2.6 | 57.3 | 83.9 | 134.5 | 390.9 | 1205.0 | 315.8 | |
| 2 | KERALA | 1903 | 3.2 | 18.6 | 3.1 | 83.6 | 249.7 | 558.6 | 1022.5 | 420.2 | |
| 3 | KERALA | 1904 | 23.7 | 3.0 | 32.2 | 71.5 | 235.7 | 1098.2 | 725.5 | 351.8 | |
| 4 | KERALA | 1905 | 1.2 | 22.3 | 9.4 | 105.9 | 263.3 | 850.2 | 520.5 | 293.6 | |
| .. | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 113 | KERALA | 2014 | 4.6 | 10.3 | 17.9 | 95.7 | 251.0 | 454.4 | 677.8 | 733.9 | |
| 114 | KERALA | 2015 | 3.1 | 5.8 | 50.1 | 214.1 | 201.8 | 563.6 | 406.0 | 252.2 | |
| 115 | KERALA | 2016 | 2.4 | 3.8 | 35.9 | 143.0 | 186.4 | 522.2 | 412.3 | 325.5 | |
| 116 | KERALA | 2017 | 1.9 | 6.8 | 8.9 | 43.6 | 173.5 | 498.5 | 319.6 | 531.8 | |
| 117 | KERALA | 2018 | 29.1 | 52.1 | 48.6 | 116.4 | 183.8 | 625.4 | 1048.5 | 1398.9 | |

| | SEP | OCT | NOV | DEC | ANNUAL RAINFALL | FLOODS |
|---|---|---|---|---|---|---|
| 0 | 197.7 | 266.9 | 350.8 | 48.4 | 3248.6 | YES |
| 1 | 491.6 | 358.4 | 158.3 | 121.5 | 3326.6 | YES |
| 2 | 341.8 | 354.1 | 157.0 | 59.0 | 3271.2 | YES |
| 3 | 222.7 | 328.1 | 33.9 | 3.3 | 3129.7 | YES |
| 4 | 217.2 | 383.5 | 74.4 | 0.2 | 2741.6 | NO |
| .. | ... | ... | ... | ... | ... | ... |
| 113 | 298.8 | 355.5 | 99.5 | 47.2 | 3046.4 | YES |
| 114 | 292.9 | 308.1 | 223.6 | 79.4 | 2600.6 | NO |
| 115 | 173.2 | 225.9 | 125.4 | 23.6 | 2176.6 | NO |
| 116 | 209.5 | 192.4 | 92.5 | 38.1 | 2117.1 | NO |
| 117 | 423.6 | 356.1 | 125.4 | 65.1 | 4473.0 | YES |

## 8. Preprocessing dataset

### 1. Check for missing values

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset

- checking if any colomns is left empty or not.
- Data has no null or NaN value

### 2. Replacing the flood results of yes/no by 1/0

Label Encoding- Label encoding is a technique used in machine learning and data preprocessing to convert categorical labels into numerical representations. In this process, each unique label or category is assigned a unique integer or numerical value. This is particularly useful when working with algorithms that require numerical input

## 9. Feature extraction

### 1. Normalization

This technique is used to scale and normalize the values of a feature to a specific range, usually between 0 and 1.

**Formula:**

The formula for Min-Max Scaling is given by:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

where $X$ is the original feature, $\min(X)$ is the minimum value of the feature, and $\max(X)$ is the maximum value of the feature.

### 2. Standardization

The Standard Scaler is a method of feature scaling or normalization used in machine learning to standardize the range of independent variables or features of the dataset.

The Standard Scaler transforms the data by subtracting the mean and dividing by the standard deviation of each feature. The formula for standardization is given by:

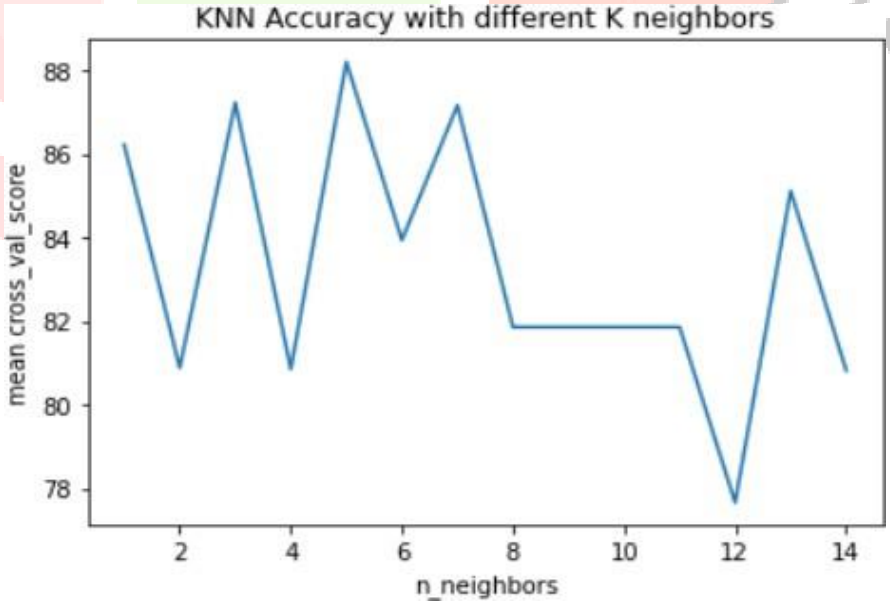
where X is the original feature, deviation mean(X) is the mean of the feature, and std is the standard of the feature.

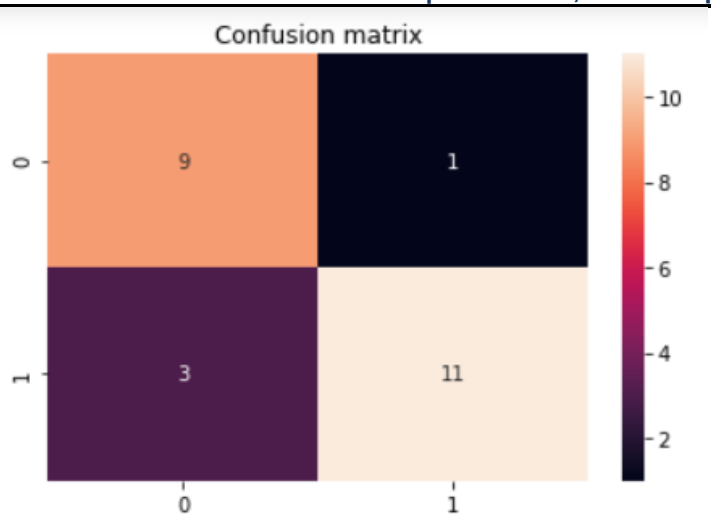
$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

**10.Machine learning methods**

**K-Nearest Neighbour(KNN)**

The kkk-Nearest Neighbors (kkk-NN) algorithm is a simple yet effective machine learning technique for flood forecasting. It classifies or predicts flood events by comparing new observations with the kkk closest data points in the dataset based on a specified distance metric, such as Euclidean or Manhattan distance. kkk-NN is particularly suitable for flood forecasting as it does not make assumptions about the underlying data distribution and can adapt to complex, nonlinear patterns inherent in hydrological and meteorological data. Feature selection methods, often utilizing correlation coefficients, are combined with the kkk-NN algorithm to identify the most relevant variables, enhancing the model's accuracy and robustness. By leveraging historical flood data and rainfall patterns, kkk-NN offers a straightforward yet powerful approach to predicting flood events and assisting in disaster preparedness






**KNN Results**

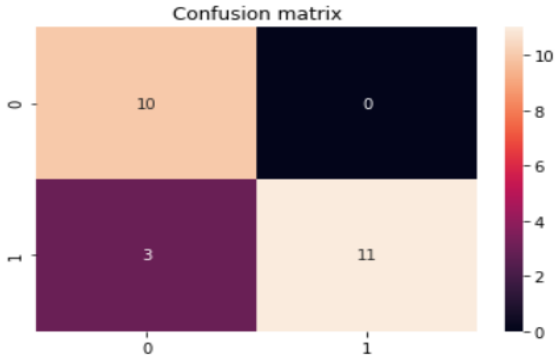
| **Terms** | values |
|---|---|
| Accuracy | 91.667 |
| Recall | 0.9 |
| prcision | 0.75 |

**Support Vector Machines(SVM)**

Support Vector Machines (SVM) have proven to be an effective machine learning technique for flood forecasting due to their ability to handle both linear and nonlinear relationships within data. By identifying an optimal hyperplane that separates different data classes, SVM is capable of accurately predicting flood events based on various climatic and hydrological inputs. The performance of SVM can be enhanced by integrating it with techniques such as Kernel Principal Component Analysis (KPCA) for dimensionality reduction and boosting algorithms for improved model performance. These enhancements enable SVM to process large and complex datasets, distinguishing between subtle variations in rainfall, streamflow, and other factors influencing floods. Additionally, SVM's robustness in managing outliers and its adaptability to diverse environmental conditions make it a reliable choice for developing predictive models tailored to specific regions, such as Kerala, where flood events are influenced by dynamic and interconnected climatic factors.

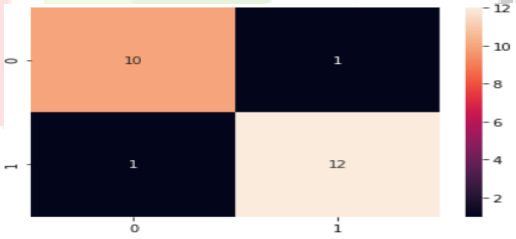
**SVM Results**

| **Terms** | values |
|---|---|
| Accuracy | 83.33 |
| Recall | 0.769 |
| Precision | 1.0 |



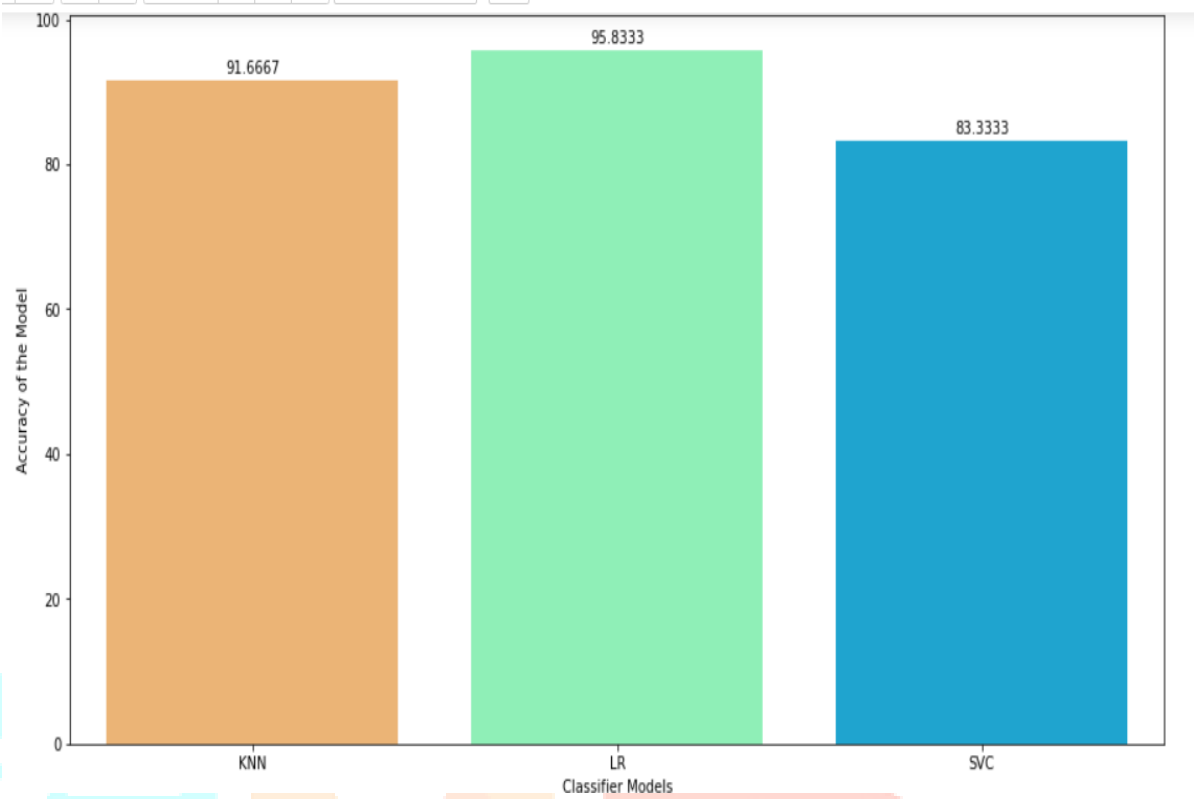

**Linear Regression**

Linear regression is a fundamental statistical method commonly used for flood forecasting to establish a relationship between predictor variables, such as rainfall, temperature, and catchment characteristics, and the target variable, such as streamflow or flood magnitude. By fitting a linear equation to observed data, this method can estimate future flood events based on historical patterns. Despite its simplicity, linear regression is effective in scenarios where the relationships between variables are predominantly linear. However, its predictive accuracy diminishes when dealing with complex, nonlinear relationships or interactions among multiple factors influencing floods. To address this, linear regression models are often enhanced by incorporating feature engineering techniques or combining them with other machine learning methods, improving their ability to capture the intricacies of flood dynamics.




| **Terms** | values |
|---|---|
| Accuracy | 95.833 |
| Recall | 0.909 |
| prcision | 0.909 |

## 11.Accuracy Comparison




## 12.Conclusion

Using monthly rainfall data for flood forecasting in Kerala, this study compared the performance of K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Linear Regression. Linear Regression demonstrated superior accuracy compared to KNN and SVR, making it the most effective algorithm for this dataset. The findings highlight the importance of selecting appropriate algorithms based on data characteristics for reliable flood forecasting models, thereby contributing to enhanced disaster preparedness and risk management strategies.

## 13.References

[1] Dawei Han, L. Chan,” Flood forecasting using support vector machines”, October 2007Journal of Hydroinformatics 9(4),DOI:10.2166/hydro.2007.027

[2] S. Li, K. Ma, Z. Jin and Y. Zhu, "A new flood forecasting model based on SVM and boosting learning algorithms," 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada, 2016, pp. 1343-1348, doi: 10.1109/CEC.2016.7743944

[3] Gauhar, Noushin & Das, Sunanda & Moury, Khadiza. ,“Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm”,.January 2021,DOI: 10.1109/ICREST51555.2021.9331199.

[4] P. Supriya,M. Krishnaveni,” Regression Analysis of Annual Maximum Daily Rainfall and Stream Flow for Flood Forecasting in Vellar River Basin”, Aquatic Procedia,Volume 4, 2015, Pages 957-963, https://doi.org/10.1016/j.aqpro.2015.02.120

[5] Joel T. de Castro, Gabriel Salistre," Flash Flood Prediction Model based on Multiple Regression Analysis for Decision Support System",Research gate,October 2013, https://www.researchgate.net/publication/289269355