IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116132

Flood Prediction and Rainfall Analysis Using Machine Learning

Nagashri K S[1], Nitin Kashyap[2], Shravan P Rao[3], Sumukha R Kashyap[4], Karthik K C[5]

Student, NIE Institute of Technology Mysore, Karnataka[1-4]

Assistant professor, NIE Institute of Technology, Mysore, Karnataka[5]

Abstract: Flooding is one of the most devastating natural events that can occur. The ability to forecast this occurrence has a significant impact on the well-being of humans and other natural beings. According to a brief history of weather study, the ancient Mayans were able to predict floods using planetary motions that were not very accurate. With the advancement of technology and the increasing reliance on computers, people are now able to collect vast amounts of data of various types, such as planetary positions using mathematical models, weather data using rain gauges, and wind turbines. It is very difficult to analyse this data and provide an output, but with the help of a machine learning algorithm, one can gain higher accuracy to forecast floods and inform the region ahead of time, avoiding priceless losses.

Keywords: Gaussian Naïve Bayes, K- Nearest Neighbours, Support Vector machine, Logistic Regression, Decision Tree Classification.

I. INTRODUCTION

This study explains how the Support Vector Classification Algorithm can help predict floods before they happen and compares its accuracy to other machine learning algorithms that are commonly employed today. This paper also discusses several machine learning techniques used in prediction, and how this research might be used to improve flood prediction ability.

II. MOTIVATION

Floods are extremely devastating and can occur as a result of a variety of reasons such as snow, wind, and even other natural catastrophes such as hurricanes. We wish to discover a solution to help with this natural calamity because floods wreak so much damage to human life, animals, and property. With a more precise prognosis, the damage can be lessened, allowing counties to conserve money for disaster relief in certain areas. In theory, anyone may use this programme, but county officials, as well as disaster and prevention relief groups, would use it to predict flooding in their areas of interest and to alert people who are in a position to implement flood-protection measures. Politicians, city planners, disaster relief organisations, and government officials dealing with weather-related concerns are among those who may be interested in this application. The application is suitable for generating more knowledge, in addition to current flood detection and prevention approaches, in order to protect portions of the country from inevitable floods.

III. MACHINE LEARNING TECHNIQUES

1. K-Nearest Neighbors: The KNN model uses a data set of characteristics such as temperature, humidity, and pressure to determine the model's behavior in predicting the disaster in chronological order using the weighted moving average technique. This model's prediction accuracy was determined to be approximately 90%, however, it does not consider numerous other characteristics such as wind direction, water level gauges, inland rainfall, and so on, which can play a crucial variable in forecasting the flood with more accuracy. [1]

2. Random Forest: The random forest model can consider larger datasets for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the classification or mean prediction where a relatively small number of sample data sets can get better accuracy, with more samples may not improve accuracy but can maintain accuracy when a large portion of the sample data is missing relative to the computational cost of the model. [5]

3. Gaussian Naïve Bayes: The developed models revealed that the Gaussian Naïve Bayes model having Standardized Precipitation Index (SPI) values as one of its predictors performed better than the model without SPI values, with correctly classified instances of 92.9% and 65.7%, respectively. [3]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116132

4. Decision Tree: Decision tree is a prediction algorithm that continuously splits data based on particular data criteria. It is a type of supervised learning in which regression and classification problems are tackled using a non-parametric method. By studying the decision criteria, the model initially targets the variables to forecast the value of the variables from the given data. This decision model determines the accuracy and output in this manner. [4]

5. Logistic Regression: When the dependent variable is binary in nature, logistic regression is applied. It is used to address problems involving binary categorization. This approach outperforms linear regression since linear regression proved ineffective at predicting the value of a binary variable. It will only forecast numbers that are outside of a certain range, such as 0 and 1. A linear relationship is not essential in logistic regression. This method has no multicollinearity. [4]

IV. TOOLS USED

- **Sklearn:** The Sklearn is a library for python which feature algorithm like SVM, Random Forest etc for machine learning analysis. It is used to build models.

- **NumPy:** In python we used NumPy library for scientific computing. It is a core library which provides tools and high performance for a given array objects.

- **Panda:** Panda library is a open source library that is used to make analysis of data and to use easily. It provides high performance and easy to use data structure.

- **Plotly:** The Plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

V. PROPOSED METHODOLOGY

The predictive model is used for the prediction of precipitation. The first step is converting data into the correct format to conduct experiments then making a good analysis of the data and observing variation in the patterns of rainfall. We predict the rainfall by separating the dataset into the training set and testing set then we apply different machine learning approaches (KNN, SVC, LR, etc.) and statistical techniques and compare and draw analyses over various approaches used. With the help of numerous approaches, we attempt to minimize the error. A flood detection model is also created in the application that detects floods. The data from all over India is collected and the model is built using various machine learning algorithms.

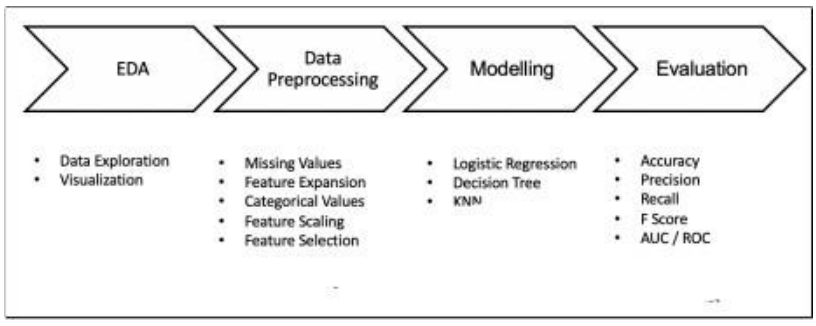

Fig 1 Model building process

The two models created are integrated into the web application which is a user-friendly application. The parameters of rainfall are given as an input to the system which predicts the rainfall based on the trained model. The flood prediction model is built when the various months of rainfall are given then the flood is predicted.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116132

A. RAINFALL PREDICTION MODULE

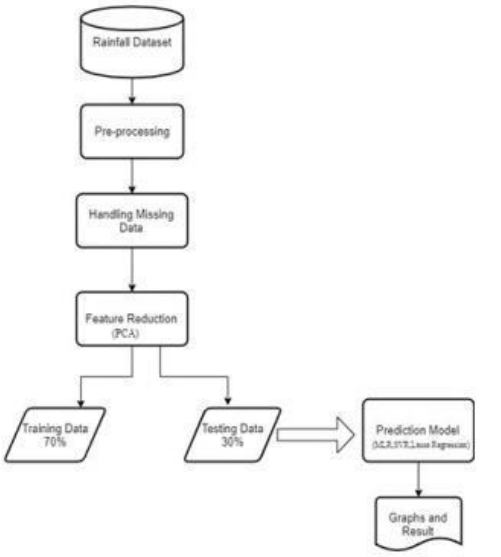

Fig. 2 Rainfall prediction model

1. DATA COLLECTION

The dataset consists of the meteorological data from year 2012-17 for each state.

2. DATA PREPROCESSING

It is a data mining approach that entails converting raw data into a usable format. Real-world data is frequently inadequate, inconsistent, and/or lacking in specific behaviors or trends, and it is rife with inaccuracies. We completed the preprocessing processes listed below. We discovered a few occurrences with null values during our EDA process. As a result, this becomes one of the critical steps. To fill in the missing values, we will sort our instances by location and date and replace the null values with their mean values. The date feature can be expanded to include Day, Month, and Year, and these newly formed features can then be utilized for additional preprocessing stages.

A categorical feature has two or more categories, but no inherent ordering of the categories. We have a few category features such as Wind, Gust, Direction, and so on. Because the models are built on mathematical equations and calculations, it is now more difficult for machines to interpret and process texts rather than numbers.

3.MODEL BUILDING

Linear Regression is a machine learning algorithm based on supervised learning**.** It performs a regression task**.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

4. EVALUATION

This is the integral part as it helps in finding the best model that represents data and how well the prediction is achieved. Input values are passed to the model and the output is the predicted values. The output is compared with the testing data to calculate accuracy and the RMSE values.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116132

B. FLOOD PREDICTION MODULE

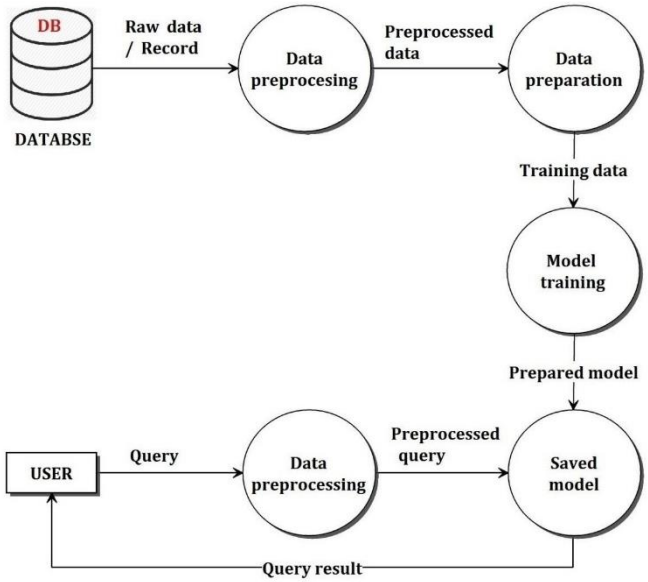

Fig. 3 Flood prediction model

1. DATA COLLECTION

The dataset consists of the measurement of rainfall from year 1901-2017 for each state.

- Data consists of 20 attributes
- The attributes are the amount of rainfall measured in mm.

2. DATA PREPROCESSING

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. We have carried out the below-pre-processing steps.

- **Missing Values:** As per our EDA step, we learned that we have a few instances with null values. Hence, this becomes one of the important steps. To impute the missing values, we will group our instances based on the location and date and thereby replace the null values with their respective mean values.
- **Categorical Values**: Categorical feature is one that has two or more categories, but there is no intrinsic ordering to the categories. We have a few categorical features – Floods classification with 2 unique values
- **Categorical data** is handled using a label encoder.

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is valuable to machine learning problems since it allows to get closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts [4]. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and researchers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.116132

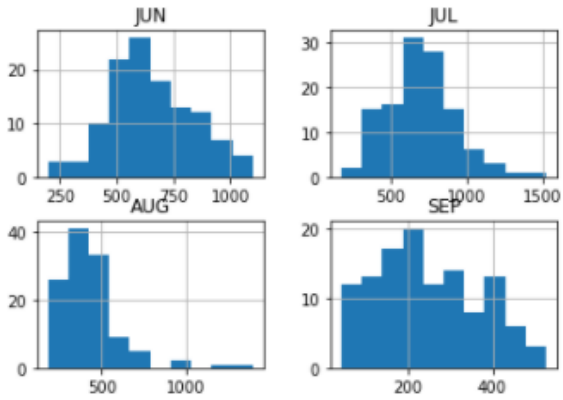

Fig. 4 EDA of flood dataset

4. MODEL BUILDING

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification problems.

The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may easily place fresh data points in the correct category in the future. A hyperplane is the optimal choice boundary. SVM selects the extreme points/vectors that aid in the creation of the hyperplane. These extreme examples are referred to as support vectors, and the method is known as a Support Vector Machine.

4. EVALUATION

This is the integral part as it helps in finding the best model that represents data and how well the prediction is achieved. Input values are passed to the model and the output is the predicted values. The output is compared with the testing data to calculate accuracy and the RMSE values.

VI. CONCLUSION

This paper discusses the flood forecasting model using the SVM algorithm implemented using SK learn library where this can be implemented along with the existing architecture for scalability in account of Moore's law where the system must be dynamically scalable

REFERENCES

[1]. [1]. C. Nikhil Binoy, N. Arjun, C. Keerthi, S. Sreerag and A. H. Nair, "Flood Prediction Using Flow and Depth Measurement with Artificial Neural Network in Canals," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019.

[2]. Zainudin, Suhaila & Jasim, Dalia & Abu Bakar, Azuraliza. (2016). Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction. International Journal on Advanced Science, Engineering and Information Technology. 6. 1148. 10.18517/ijaseit.6.6.1487.

[3]. Oluwatobi Aiyelokun, Gbenga Ogunsanwo."Gaussian Naïve Bayes Classification Algorithm for Drought and Flood Risk Reduction", In Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation 2021.

[4]. Naveed Ahamed, S.Asha, "Flood prediction forecasting using machine Learning Algorithms", In International Journal of Scientific & Engineering Research 2020.