ISSN No:-2456-2165

Prediction of Flood with Real-Time Data Integrating Machine Learning Models and Scraping Techniques

Shiju E (1), Mr. D. Justin Jose (2)
Department of CSE, MACET,

Abstract:- Floods are quite possibly of the most harming regular disappointment, which can be perceptibly mind boggling to demonstrate. The examinations on the improvement of flood expectation designs added to peril decrease, strategy thought, minimization of the deficiency of human life, and markdown the effects hurt connected with floods. to copy the convoluted numerical articulations of substantial strategies of floods, for the beyond two quite a while, brain local area techniques contributed rather inside the improvement of expectation frameworks offering better execution and practical arrangements. To save you this problem to foresee regardless of whether a flood happens through precipitation dataset it looks at the brain network-based procedures. The investigation of the dataset with the guide of Multi-Layer Perceptron Classifier (MLP) to catch various data like variable personality, missing cost cures, insights approval, and realities cleaning/planning may be finished at the total given dataset. To generally speaking execution in forecast of flood occur or presently not by exactness estimation with appraisal type record, find the disarray grid and the aftereffect of this shows that the viability of the GUI basically based programming utilizing given ascribes. Notwithstanding the above model, we increment the presentation by adding a component that gets the constant information from the live information through the web and the outcome would be a continuous expectation of flood in some random region.

***Keywords**:- Dataset, Python, Preprocessing, MLP Classifier, Web Scrapping.*

I. INTRODUCTION

AI expects to anticipate the future from past information. AI (ML) is a sort of man-made consciousness (AI) that permits PCs to learn without being expressly modified. AI centers around creating PC programs that can change when presented to new information and the rudiments of AI and carries out basic AI calculations utilizing Python. The preparation and forecast process includes the utilization of exceptional calculations. Preparing information is shipped off the calculation, which utilizes this preparing information to make expectations about new test information. There are three classes of Machine learning, in particular regulated learning, solo learning, and support learning. Managed learning programs get both information, and legitimate naming of learning information should be pre-marked by people. Solo learning isn't a name. Given to the learning calculation. This calculation needs to figure out the grouping of info information. At long last, support learning interfaces powerfully with its current circumstance and gets positive or negative input to further develop execution. Notwithstanding the model above, you can further develop execution by adding the capacity to recover ongoing information from live information over the Internet, bringing about the constant forecast of floods in unambiguous regions.

II. RELATED WORK

A. *Prediction of Flood Using Radial Basis Function (RBF) using Internet of Things (IoT)*

ANN had been prepared with the information of water levels and information of precipitation, this is utilized to foresee water level and day to day precipitation of the following month. The boundaries used to get the least blunder in the forecast course of level of water, precipitation with the best outspread premise capability brain network utilize multiple times cycles and utilize the learning rate that is equivalent to 0. 00007. Here the Radial Basis Function is been utilized to anticipate the flood. The information was gotten from Citarum River Hall. The outcome from Radial Basis Function Neural Network is shipped off an android application that shows the chance of flooding. Involving age however much 700 gives 0.027 as the mistake worth of TMA and 0.002 as the blunder worth of CH, a learning pace of 0.00007 gives 0.286 as the mistake worth of TMA and 0.002 as the blunder esteem CH, and a secret neuron of 2 gives 0.6483 as mistake worth of TMA and 15.999 as the blunder

B. *Predicting flood with the use of Multi-Layer ANN in Monitoring System Along With The Rain Gauge, Sensor of Soil Moisture*

This research requires the implementation of a real-time monitoring system capable of measuring parameters such as rainfall intensity, soil moisture, water level and rate of water rise. Various sensors are integrated into the system to record and store data. A prediction model based on multilayer artificial neural networks was developed and tested in a real-world setup. In this study, we examined the response of hierarchical network models. The flood prediction model showed an RMSD of 2.2648, slightly off from the actual water level. This was a big problem in the Philippines as it caused property damage, infrastructure damage and damage and even loss of life. Current systems address problem solving and prevent catastrophic flood disasters. A multilayer artificial neural network using MATLAB was used to develop the predictive model. The network fit very well in training, testing, validation, and the overall data set. Specifically, it was 0.99889 for the training dataset, 0.99362 for the test

ISSN No:-2456-2165

dataset, 0.99764 for the validation dataset, and 0.97952 for all data in the dataset.

C. *Technique of Optimal Web scraping.*

Data entry is one of the most tedious tasks requiring a lot of human resources to create structured data from inputs. The large amount of data entering the system can contradict the original data and cause confusion. This is especially true if you need to collect data from image files. In this paper, we propose a text recognition system that can be used to automatically recognize text from images and update it with a target file. The proposed method accepts a web URL as input and uses web scraping techniques to retrieve text or images. The system extracts text data from user-defined ranges. Additionally, the extracted text is classified using a support vector machine (SVM) and a simple Bayesian classifier. Output is saved in Google Sheets, CSV, PDF, Text, or Excel format, depending on user selection. State-of-the-art text recognition models such as PyTesseract, PyOCR, and TesseOCR are compared using metrics such as accuracy, precision, and execution speed. Experimental results show that PyTesseract provides 83.45% accuracy and 75.55% accuracy. The performance of support vector machines (SVMs) and naive Bayesian classifiers are compared. 92.08% accuracy. The recall of this is 90.148% and for the classifier algorithm Naïve Bayes classifier.

III. METHODS

A. *Data Validation, Cleaning or Preparing*

Loads the specified data set and imports the library package. Analyze variable identification according to data format and data type and evaluate missing and duplicate values. A validation dataset is a sample of data retained from training a model and used to estimate the capabilities of the model while optimizing the model and procedure. This can be used to test datasets and validate models as they are evaluated and used optimally. Data cleaning/preparation such as renaming the given dataset and removing columns. Analyze univariate, bi-variate, and multi-variate processes. Data cleansing techniques vary depending on the data set. The purpose of data cleansing is to identify and extract errors and anomalies to increase the value of data for analysis and decision making.

➢ *Data processing:*

Data preprocessing refers to transformations applied to data before it is sent to an algorithm. Data preprocessing is a technique used to transform raw data into a clean data set. This means that the data collected from various sources are collected in raw form and are not useful for analysis. To get better results from models applied by machine learning techniques, data must be in the right format. Some specific machine learning models require specific forms of information. For example, the random forest algorithm does not support null values. Therefore, to run the random forest algorithm, we need to manage null values from the original raw data set.

B. *Creating a predicted variable using the range of rainfall:*

A validation dataset is a sample of data retained during model training and used to estimate model features during model tuning, by analyzing univariate, bivariate, and multivariate processes. It is validated and tested when the model is prepared for evaluation or data. B. Rename specific datasets and remove columns. Data cleansing procedures and techniques vary from dataset to dataset. The main purpose of data cleaning is to detect and eliminate errors and anomalies in order to increase the value of data in analysis and decision making. Data visualization provides an important set of tools for qualitative understanding. This is useful for exploring and training datasets, identifying patterns, corrupted data, outliers, and more. With a little bit of expertise, data visualization can be used to represent and illustrate key relationships in charts and graphs that are more intuitive and relevant than measuring relevance and importance. Data visualization and exploratory data analysis are separate areas, and some of the books listed at the end are worth reading.

Data may not make sense until it can be presented in a visual form. With charts and numbers. Being able to quickly visualize things like data samples is a skill in applied statistics or machine learning. Discover the different types of charts you need to know about when visualizing data in Python and how you can use them to better understand your data.

C. *Performance measurement of ML algorithm*

➢ *Logistic Regression*

It is a statistical technique that is used to analyze a data set using one or more independent features to influence the outcome. Outcomes are measured using dichotomous variables (only two possible outcomes). The goal of logistic regression is to find the best model to describe the relationship between a dichotomous characteristic of interest (dependent variable directly responsible for the response or outcome variable) and a set of independent variables (predictor or explanatory variables). Logistic regression is a machine learning classification algorithm used to predict the probability of categorical dependent variables. In logistic regression, the dependent variable is a binary variable with data encoded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

➢ *Support Vector Machines*

A classifier that classifies a data set by determining the best hyperplane between the data. We chose this classifier because the number of applicable kerning functions is very diverse and this model can achieve high predictability. Support vector machines are probably one of the most popular and discussed machine learning algorithms he. These were very popular when they were developed in the 1990s and remain a method of selecting powerful algorithms with only minor tweaks.

- How to resolve multiple names used to refer to support vector machines.
- Representation used by SVM when actually saving the model to disk.
- Predict new data using the trained SVM model representation.

- How does the model learns using the training data.
- Prepare data that are optimal for the SVM algorithm.
- Where to find more information about SVM.

➢ *K-Nearest Neighbor (KNN)*

The KNN method is a supervised machine learning algorithm that stores all instances corresponding to training data points in n-dimensional space. When it receives unknown discrete data, it analyzes the nearest number of k-nearest neighbors (nearest neighbors) and returns the most common class as a prediction, and for real-valued data, it returns the average of k-nearest neighbors. The distance-weighted nearest neighbor algorithm uses the following query to weight each contribution in the k-nearest neighbors according to the distance. This gives a large weight to the nearest neighbors.

ANNs typically average over the k nearest neighbors, so they are robust to noisy data. The k-nearest neighbor algorithm is a classification algorithm and is supervised. Get a set of labeled points and use them to learn how other points are labeled.

To label a new point, find the closest labeled points (nearest neighbors) to the new point and have those points vote so that the label with the most neighbors becomes the label for the new single point . (where 'K' is the number of neighbors checked). Use the entire training set to make predictions on the validation set. ANN makes predictions about new instances by finding the k "closest" instances across the set. "Accessibility" is determined using approximate (Euclidean) measurements of all features.

D. Performance of the MLP classifier

MLPs (multilayer perceptrons) are a class of feed-forward artificial neural networks (ANNs). The term MLP refers either to feed-forward ANNs or strictly to multi-level ("threshold-activated") networks of perceptron. used loosely and sometimes loosely. Perceptrons are sometimes colloquially referred to as neural networks, especially when there is only one hidden layer. Layered perceptron can only learn linear functions, Multilayer perceptrons can also learn nonlinear functions. An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Each node, with the exception of the input node, is a neuron with a nonlinear activation function. MLP uses a supervised learning technique called backpropagation for training. Its complexity and nonlinear activation distinguish MLPs from linear perceptrons. You can distinguish between data that are not linearly separable.

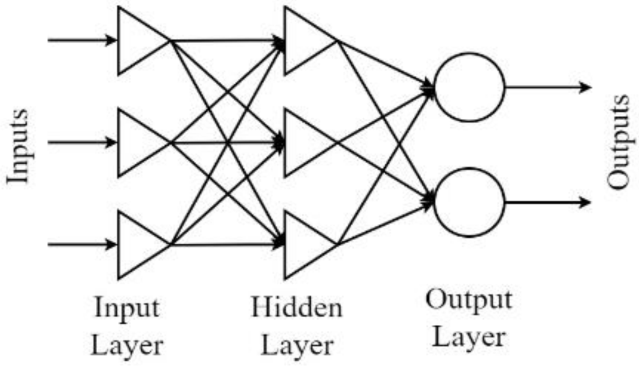

Fig 1: MLP Classifier

A perceptron is a very simple learning machine. I might get some input. Each input has a weight that indicates its importance, producing a '0' or '1' output decision. However, it combines with many other perceptrons to form artificial neural networks. Given enough training data and computational power, a neural network can theoretically answer any question.

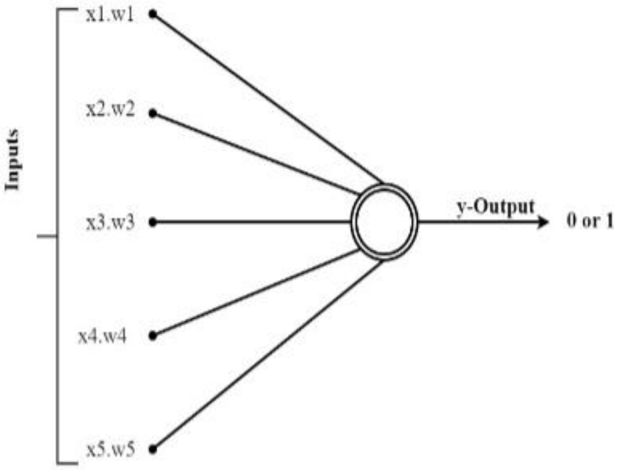

Fig 2: Input and Output of perceptron

A multi-layer perceptron (MLP) is a perceptron that solves complex problems in combination with additional perceptrons stacked in multiple layers. Each perceptron in the first layer (input layer) on the left sends an output to all perceptrons in the second layer (hidden layer), and all perceptrons in the second layer send output to the last layer (output) on the right. Send output to I will send it to you. layer). Each perceptron sends multiple signals, one to each perceptron in the next layer.

E. Web-Scraping of weather data

Extract of data from the Internet or websites using automated processes. Data on the website is unstructured. Web scraping helps collect this unstructured data and store it in a structured format. There are many ways to scrape a website, including online services, APIs, and writing your own code. This article describes how to implement web scraping in Python.

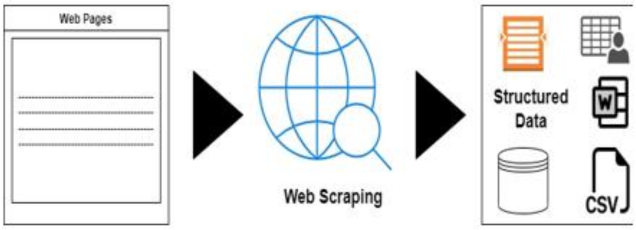

Fig 3: Web Scraping

IV. CONCLUSION

The analytical process began with data cleansing and processing, missing values, exploratory analysis, and finally modeling and evaluation. Finally, it uses machine learning algorithms to predict flash floods and produce different results. Therefore, the best result is the MLP algorithm (97.40%). This gives us the following insights about flood forecasting: Specifically, the main purpose of this project is to extract data from live weather data, use web scraping techniques to extract daily precipitation from weather data on a weather website, and then extract that specific It is to extract rainfall. day of use. Designed to predict floods using volume data.

In 3.3, you can predict flooding using multiple different algorithms like SVM, Logistic Regression, and KNN algorithms. Each algorithms mentioned above gives different accuracy levels.

In 3.4 We can see that the MLP classifier gives the highest prediction accuracy. Therefore, MLP is the best algorithm for flood forecasting. Also in 3.5 you can see information about web scraping. It can be used to extract live precipitation data from forecasts and weather websites. As a further improvement, it can also be automated on a regular basis so that the model automatically extracts data from websites and makes predictions.

REFERENCES

[1]. Febus Reidj G. Cruz , Matthew G. Binag , Marlou Ryan G. Ga , Francis Aldrine A. Uy. Flood Prediction Using Multi-Layer Artificial Neural Network in Monitoring System with Rain Gauge, Water Level, Soil Moisture Sensors, IEEE,28-31Oct.2018,DOI: 10.1109/TENCON.2018.8650387

[2]. Roopesh N, Akarsh M S, C. Narendra Babu. An Optimal Data Entry Method, Using Web Scraping and Text Recognition 2021 International Conference on Information Technology (ICIT) | 978-1-6654-2870-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT52682.2021.9491643

[3]. H. Hartenstein and L. P. Laberteaux, "A tutorial survey on vehicular ad hoc networks," IEEE Commun. Mag., vol. 46, no. 6, pp. 164–171, Jun. 2008.

[4]. B. Parno and A. Perrig, "Challenges in securing vehicular networks," in Proc. Workshop Hot Topics Netw. (HotNets-IV), MD, USA, Nov. 2005, pp. 1–6.

[5]. F. Dötzer, "Privacy issues in vehicular ad hoc networks," in Proc. Int. Workshop Privacy Enhancing Technol., May 2005, pp. 197–209.

[6]. J. R. Douceur, "The sybil attack," in Proc. Int. Workshop Peer-to-Peer Syst., 2002, pp. 251–260.

[7]. M. Mousa, X. Zhang, and C. Claudel, "Flash flood detection in urban cities using ultrasonic and infrared sensors," IEEE Sensors Journal, vol. 16, no. 19, pp. 7204–7216, 2016.