International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 8 Aug. 2022, pp: 547-551 www.ijaem.net ISSN: 2395-5252

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

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

Submitted: 01-08-2022 Revised: 11-08-2022 Accepted: 14-08-2022

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.

Index Terms—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

2.1: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 worth of CH ,can be utilized to foresee the flooding.

2.2:Flood Prediction Using Multi-Layer Artificial Neural Network in Monitoring System with Rain Gauge, Water Level, Soil Moisture Sensors

This study requires the implementation of a real-time monitoring system that can measure parameters such as precipitation intensity, soil moisture, water level, and rate of water level rise. Various sensors have been integrated into the system where data is logged and stored. A predictive model based on a multi-layer artificial neural network was developed and tested in a real setup. In this study, we examined the response of the hierarchical network model. The flood prediction model had a slight deviation from the actual water level, with an RMSD of 2.2648. This was a big problem in the Philippines, as it caused property damage, infrastructure damage, and even loss of life. Current systems are problem-solving compliant to prevent catastrophic flood disasters. A multi-layered artificial neural network using MATLAB was used to develop the predictive model. In training, testing, validation, and the overall dataset, the network showed very good fit. Specifically, it was 0.99889 for the training dataset, 0.99362 for the test dataset, 0.99764 for the validation dataset, and 0.97952 for all the data in the dataset. The network was then programmed and integrated into the system during the actual setup.

2.3:An Optimal Data Entry Method, Using Web Scraping and Text Recognition

Data entry is one ofthemost tedious tasks that requires a lot of human resources on the to create structured data from the inputs. The large amount of data that is input to the system can be inconsistent with the original data and can be confusing. This is especially true if you need to collect data from image files. This paper proposes a text recognition system that can be used to automatically recognize text from an image and update it to a target file. The proposed method accepts a web URL as input and uses web scraping techniques to retrieve the text or image. The system extracts text data from a user-specified area. In addition, the extracted text is categorized using a Support Vector Machine (SVM) and a naive Bayes classifier. The output is saved in Google Sheets, CSV, PDF, text, or Excel format, depending on the user's choice. The latest text recognition models , such as PyTesseract, PyOCR, and TesserOCR, are compared based on metrics such as accuracy, accuracy, and execution speed. Experimental results show that PyTesseract provides 83.45% accuracy and 75.55% accuracy. The performance of the support vector machine (SVM) and the naive Bayes classifier is compared. 92.08% accuracy, 90.148% recall, 90 naive Bayes classifier.

III. METHODS

3.1: Data Validation, Cleaning or Preparing

Importing the library packages with loading the given dataset. Analyzing the variable identification by data shape, and data type evaluating the missing values, and duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures which can be used to make the best use of test datasets and their validation when evaluating your models. Data cleaning/preparing by renaming the given dataset and dropping the column etc. To analyze the univariate, bivariate, and multi-variate processes. The techniques for data cleaning vary based on the dataset. The goal of cleaning the data is to identify and pull-out errors and anomalies in order to increase the data's value in analytics and decision-making.

Preprocessing of Data:

Data Pre-processing refers to the transformations that are applied to the data before it is sent to the algorithm. Data pre-processing is a technique used to transform raw data into a clean dataset. This means that whenever data is collected from different sources, it is collected in raw form and is not useful for analysis. To get better results from the model applied by machine learning techniques, the data must be in the proper format. Some specific machine learning models require a specific form of information. For example,Random Forest algorithm doesn’t support the null values. Therefore, to run the Random Forest algorithm, you need to manage the null values from the original raw dataset.

3.2:Creation ofpredicted variable by rainfall range:

A validation dataset is a sample of the data that is retained during model training, is used to estimate the model's capabilities when tuning the model, and is validated and tested when evaluating the model for optimal use. It can be used to serve a dataset.Cleaning or preparing data by analyzing univariate, bivariate, and multivariate processes, such as renaming specified datasets and deleting columns. Data cleansing procedures and techniques vary from dataset to dataset. The main goal of data cleaning is to detect and eliminate errors and anomalies to increase the value of the data in analysis and decision making. Data visualization provides an important suite of tools for qualitative understanding. This is useful for exploring and training datasets, and for identifying patterns, corrupted data, outliers, and more. With a little expertise, data visualization can be used to represent and demonstrate important relationships in charts and graphs that are more visceral and relevant than measuring relevance or importance. Data visualization and exploratory data analysis are separate areas, and it is worth digging into some of the books listed at the end.

Data may not make sense until it can be viewed in a visual format.With charts and figures. Being able to quickly visualize data samples etc. is a skill in either applied statistics or applied machine learning. Discover the different types of graphs you need to know when visualizing your data in Python and how to use them to better understand your data.

3.3: Performance measurement of ML algorithm

3.3.1: Logistic Regression

This is a statistical method for analyzing datasets that have one or more independent variables that determine the results. Results are measured using dichotomous variables (only two possible results). The goal of logistic regression is to find the best model for describing the relationship between the dichotomous feature of interest (dependent variable = response or outcome variable) and a set of independents (predictive or explanatory variables). Logistic regression is a machine learning classification algorithm used to predict the probabilities of categorical dependent variables. In logistic regression, the dependent variable is a binary variable that contains data encoded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

3.3.2: Support Vector Machines

A classifier that classifies a dataset by determining the optimal hyperplane between the data. We chose this classifier because the number of different kerning functions that can be applied is very diverse and this model can achieve high predictability. Support vector machines are probably one of the most popular and most discussed machine learning algorithms. These were very popular when they were developed in the 1990s and remain a way to choose powerful algorithms with a little tweaking.

- How to solve many names used to refer to support vector machines.
- The representation used by the SVM when actually saving the model to disk.
- How to predict new data using the learned SVM model representation.
- How SVM model learns from training data.
- How to optimally prepare data for the SVM algorithm.
- Where to get more information about SVMs.

3.3.3: K-Nearest Neighbor (KNN)

The K-nearest neighbor 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.

Normally, ANN is robust for noisy data because it averages the k-nearest neighbors. The k-nearest neighbor algorithm is a classificationalgorithm and is monitored. Get a set of labeled points and use them to learn how other points are labeled. To label a new point, look up the labeled points closest to the new point (closest neighbors) and have those neighbors vote so that the label with the most neighbors becomes the label for the new single dot. ("K" is the number of neighbors checked). Use the entire training set to make predictions about the validation set. KNN makes predictions about new instances by searching the entire set for the k "closest" instances. "Accessibility" is determined using proximity (Euclidean) measurements across all features.

3.4: Performance MLP classifier

MLP (Multilayer Perceptron is a class of feedforward artificial neural networks (ANN). The term MLP is used ambiguously, sometimes loosely to refer to feedforward ANN, or strictly to refer to a multi-tiered network of perceptrons (" With threshold activation). Multilayer perceptrons are sometimes colloquially referred to as neural networks, especially if they have a single hidden layer.

A multi-layer perceptron or multi-layer neural network contains one or more hidden layers (apart from the input and output layers). Single-layer perceptrons can only learn linear functions, while multi-layer perceptrons can also learn non-linear functions. The MLP consists of at least three node layers: the input layer, the hidden layer, and the output layer. With the exception of the input node, each node is a neuron that uses a nonlinear activation function. A supervised learning technique called backpropagation is used by MLP for training. Its complexity and non-linear activation distinguish MLP from linear perceptrons. You can distinguish data that is not linearly separable.

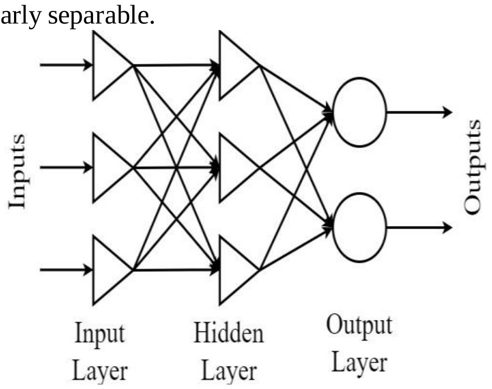

Fig 1: MLP Classifier

Perceptron is a very simple learning machine. You can receive some input. Each input has a weight that indicates how important it is and produces a "0" or "1" output decision. However, it combines with many other perceptrons to form artificial neural networks. Neural networks can theoretically answer any question with sufficient training data and computational power.

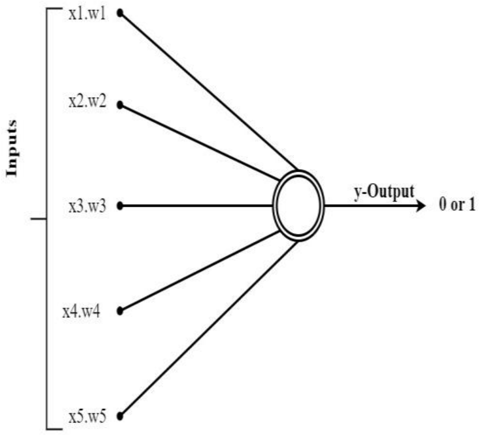

Fig 2: Input and Output of perceptron

Multilayer 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 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. I will send. layer). Each perceptron sends multiple signals, one signal to each perceptron on the next layer.

3.4: Web-Scraping of weather data

An automated method is used to extract data from the internet or websites. The data on the website is unstructured. Web scraping helps you collect this unstructured data and store it in a structured format. There are many ways to scrape your website, including online services, APIs, and writing your own code. This article describes how to implement web scraping using Python.

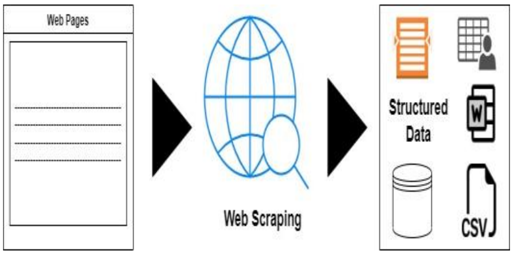

Fig 3: Web Scraping

IV. CONCLUSION

The analysis process began with cleansing and processing the data, missing values, exploratory analysis, and finally model building and evaluation. Finally, it uses machine learning algorithms to predict flash floods and produce a variety of results. Therefore, the best result is the MLP algorithm (97.40%). This provides some of the following insights into flood forecasting: Specifically, the main purpose of this project is to take data from live meteorological data, use web scraping technology to extract daily precipitation from meteorological data on metrological websites, and then use precipitation for that particular day. It is to predict floods using quantity data.

In 3.3, you can predict flooding using various algorithms such as logistic regression, help vector machines, and k-nearest neighbor algorithms. Each of the above algorithms produces different levels of accuracy.

In 3.4 you can see the MLP classifier givesthe highest accuracy for the prediction, so MLP is the best algorithm when it comes to Flood prediction. And in 3.5 you can see the information about the web scraping that you are able to use in order to extract live rainfall data from weather reporting or metrological websites. As a further

Enhancement, you can also make it automated for regular intervals of time so that the model automatically extracts data from websites and predicts.

REFERENCES

[1]. FebusReidj 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.