ABSTRACT

Floods are among the most destructive natural disasters, which are highly complex to model. The research on the advancement of flood prediction models contributed to risk reduction, policy suggestion, minimization of the loss of human life, and reduction of the property damage associated with floods. To mimic the complex mathematical expressions of physical processes of floods, during the past two decades, machine learning (ML) methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions. Due to the vast benefits and potential of ML, its popularity dramatically increased among hydrologists. Researchers through introducing novel ML methods and hybridizing of the existing ones aim at discovering more accurate and efficient prediction models. The main contribution of this project is to demonstrate the state of the art of ML models in flood prediction and to give insight into the most suitable models. This project presents the Flood prediction and Rainfall analysis using Machine Learning. The main goal of employing this application is to prevent impacts of flood. This application can be easily used by the common people or government to predict the occurrence of flood beforehand. Among many ML techniques, classification is a widely used one. We use multiple algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors and Decision Tree Classifier.

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

The unusual rainfall and global climate change has led to floods in different parts of the world. Floods are one of the worst affecting natural phenomena which causes heavy damage to property, infrastructure and most importantly human life. This paper presents the Flood prediction and Rainfall analysis using Machine Learning. The main goal of employing this application is to prevent immediate impacts of flood. This application can be easily used by the common people or government to predict the occurrence of flood beforehand. Nowadays, machine learning (ML) methods are highly contributed in the advancement of prediction systems. These methods are providing better performance as well as cost effective solutions. The advancement in the information technology, the need for easy accessibility of large cloud storage and processing power is available. Data mining technologies helps us to provide reference for decision makers as summarized information even from the large amount of data. Among many data mining techniques, classification is a widely used one. Past studies proposed many techniques that could be applied to classification, such as decision trees, neural networks, Bayesian classifiers. We use multiple algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors and Decision Tree Classifier using Python environment.

PYTHON:

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code.

Python is a programming language that lets you work quickly and integrate systems more efficiently.

It is used for:

- web development (server-side),
- software development,

- mathematics,
- System scripting.

What can Python do?

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

Why Python?

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way or a functional way.

Python Syntax compared to other programming languages

- Python was designed to for readability, and has some similarities to the English language with influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

- Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

1.1 OUTLINE OF THE PROJECT:

Floods are among the most destructive natural disasters, which are highly complex to model. The main contribution of this project is to demonstrate the state of the art of ML models in flood prediction and to give insight into the most suitable models. This project presents the Flood prediction and Rainfall analysis using Machine Learning. The main goal of employing this application is to prevent impacts of flood. This application can be easily used by the common people or government to predict the occurrence of flood beforehand. Among many ML techniques, classification is a widely used one. We use multiple algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbors and Decision Tree Classifier.

1.2 **PROBLEM STATEMENT:**

- Flood must be predicted beforehand
- Use Machine Learning technique
- Programming language- Python
- Algorithms- 4

CHAPTER-2

LITERATURE SURVEY

- Chowdhury . approached with finding the collinearity of Sea-surface temperature (SST) with a floodaffected area (FAA). They found the statistical relationship using principal component analysis. To construct a model to predict flood,

multiple regression analysis was done. The statistical model uses SST, rainfall, and streamflow in Bangladesh to serve as predictors.

- Shafizadeh-Moghadam have implemented eight different machine learning and statistical model and seven ensemble models for assessing flood susceptibility. Their dataset consists of 201 flood incidents in the Haraz watershed. They recognized eleven factors that contribute to flooding events. According to their evaluation parameters, the eight machine learning models: Artificial neural networks (ANN), Classification and regression trees (CART), Flexible discriminant analysis (FDA), Generalized linear model (GLM), Generalized additive model (GAM), Boosted regression trees (BRT), Multivariate adaptive regression splines (MARS) and Maximum entropy (MaxEnt) produce the Area under the ROC curve (AUC) scores of 0.920, 0.643, 0.822. 0.971, 0.962, 0.975, 0.941 and 0.971 respectively.
- Han applied SVM models to forecast floods. They experimented with different kernel tricks and various input combinations. To assess the effectiveness of new models, they compared it with the Naive model, Trend model, and Transfer function (TF) model. For their study, they used the gamma values of 0.001, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.3, 0.5, 0.7 and 0.9. They applied 5-fold cross-validation for the training and testing phase. It was found that SVM predicted flood with higher accuracy in comparison to the TF model. Their observation showed that the linear function performs better with unknown future rainfall input. Extremely large rainfall data that is not scalable makes the model quite unstable.
- Noymanee attempted to forecast pluvial floods using machine learning techniques. For this purpose, they used open data of the basin of Pattani River, Thailand. They used neural network (NN), bayesian linear (BL), boosted decision tree (BDT), decision forest (DF) and linear regression (LR). They split their dataset into two parts. They used the 80% of dataset for training the model and 20% dataset to test them. For the purpose of examining their models, they used RMSE, MAE and efficiency index (EI). They concluded that BL method provided the most efficient and effective result for a long-time lag. The other models also provided good results that were only acceptable for short time lag prediction.
- Khosravi did a comparative study on decision trees algorithm to create flash flood susceptibility model. The mapping was done at the Haraz Watershed, Iran.

Logistic model trees (LMT), reduced error pruning trees (REPT), Naive Bayes Trees (NBT) and Alternating Decision Trees (ADT) were used for the study. The historical data of 201 flash floods were used. For feature selection, they used information gain ratio and multicollinearity diagnostics method. 70% of the dataset was used for training the model and 30% of it was used to test the model. Statistical evaluation measures, the ROC curve and Freidman and Wilcoxon signed-rank tests were used to evaluate the performances of the four models. After analyzing the results, it was seen that ADT model performed best.

- Duncan proposed a Machine Learning-Based Early Warning System for Urban Flood Management. It works with drainage systems and the use of overall rainfall data concurrently, in order to predict flooding of multiple urban zones using a single multioutput ANN. In their study, there are three input time-series used: cumulative rainfall (mm), rainfall intensity (mm/hour), and the new antecedent precipitation index value (meters). The principal goal of their research is to find the true positive rates for sensitivity analysis. The research findings show that the predictive ability of the system depends on actual rainfall.
- Li established an ANN model to predict the trend of storm flooding in China. The ANN model was used to predict the storm flooding and built to simulate the historical storm surge during typhoon periods. The results showed that the ANN model provides stronger results than other models for flood prediction.
- Xia, and Rao designed a new application that utilized a Basis Prediction artificial (BP- artificial) neural network model to predict the flood level in lakes, rivers, and reservoirs. The errors between predicted and actual value are feedback in the process of learning. Their research primarily compares the data calculated by the model to actual monitoring data from the monitoring station from the city of Chaoan; with the result showing that the algorithm can obtain a better prediction. They compared the values calculated by the algorithm to actual monitoring data from monitoring control centre; the result shows that the BP- artificial neural network model can achieve better prediction in terms of accuracy and credibility, in comparison with other neural network architectures.
- Trombe P. have discussed about the Automatic Classification of Offshore Wind Regimes with Weather Radar Observations. In this paper, the automatic classification of offshore and wind regimes (i.e., wind fluctuations with specific frequency and amplitude) with the help of reflectivity observations from single

weather radar system has been addressed. Spatial continuity, motion of precipitation echoes on the images and global intensity are described from the above attributes. Finally, classification tree was used to find the relationship between wind regimes and precipitation attributes.

- Chawla G. have discussed about the big data analysis which will help in getting a view of the given dataset. In this paper the researchers focuses on Big Data visualization, its challenges, various tools. Researchers have explored a new way to visualize and analyse complex and dynamic datasets using virtual reality. They have also founded that how virtual reality has extremely changed the world of Big Data Visualisation.
- Mariana Belgiu used a technique that compared unsupervised multiresolution segmentation approaches with that of supervised approaches by extracting buildings from images. The classification yielded an overall accuracy that ranged between 82% to 86%. Particle Swarm Optimization was used in dynamic clustering for image segmentation. The approach was called as DCPSO which identified the best possible number of clusters from the tested images.
- Krishna Kant Singh identified flooded area from satellite images using a classifier performance analysis was done by computing the error matrix. Several methods have been proposed by various researchers for classification of only satellite images but not for aerial images.

●

CHAPTER-3

METHODOLOGY

Initially, the data was collected from Kaggle and we filtered and checked the precipitation values for null values and other common errors. Now the data is fed to the machine learning algorithms,

(i)Logistic Regression

(ii)Naive Bayes

(iii)K-Nearest Neighbors

(iv)Decision Tree Classifier and prediction of the occurrence of flood for the particular date and area is calculated based on the risk levels.

The Risk levels decide how bad the flood will affect and we use them for the forecasting of floods. The low level is defined as level 0 flood. Level 1 and level 2 have a moderate and medium-high risk for the flood event and level 3 is the High alert and level 4 is for the very high flood risk.

level 0 =Low

level 1 =Medium

level 2 =Medium-High

level 3 =High

level 4 =Very High

Libraries used:

NumPy - NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

Pandas - Pandas stands for “Python Data Analysis Library”. What’s cool about Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software. This is so much easier to work with in comparison to working with lists and/or dictionaries through for loops or list comprehension.

Matplotlib - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.Pyplot is a Matplotlib module which provides a MATLAB-like interface. Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open-source.

Date time - The datetime module supplies classes for manipulating dates and times. While date and time arithmetic is supported, the focus of the implementation is on efficient attribute extraction for output formatting and manipulation. Date and time objects may be categorized as “aware” or “naive” depending on whether or not they include time zone information. With sufficient knowledge of applicable algorithmic and political time adjustments, such as time zone and daylight saving time information, an aware object can locate itself relative to other aware objects. An aware object represents a specific moment in time that is not open to interpretation. A naive object does not contain enough information to unambiguously locate itself relative to other date/time objects. Whether a naive object represents Coordinated Universal Time (UTC), local time, or time in some other time zone is purely up to the program, just like it is up to the program whether a particular number represents meters, miles, or mass.

Naive objects are easy to understand and to work with, at the cost of ignoring some aspects of reality.

Seaborn - Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

Keras - Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Keras contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code. Keras Models provide pre-trained models, which could easily integrated into projects.

Sklearn - sklearn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. sklearn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance.