Rainfall prediction using Extreme Gradient Boosting

**Introduction:**

The agricultural practices and crop yields of India are heavily dependent on the climatic factors like rainfall and water resources. Out of 142 million ha cultivated land in India, 92 million ha (i.e. about 65%) are under the influence of rain fed agriculture. Unlike irrigated agriculture, rain fed farming is usually diverse and risk prone. The monsoon season is the principal rain bearing season and in fact a substantial part of the annual rainfall over a large part of the country occurs in this season. Small variations in the timing and the quantity of monsoon rainfall have the potential to impact on agricultural output. Rainfall is the most important climatic element that influences agriculture. Monthly rainfall forecasting plays an important role in the planning and management of agricultural scheme and water resources systems. The main objective of the present study is to develop a valid stochastic model to simulate monthly rainfall. Rainfall is a seasonal phenomenon with twelve months period, but most probably depends on monsoon. Seasonal time series are often modeled by different techniques. In recent times, many researchers modeled monthly rainfall using SARIMA methods and Box-Jenkins ARIMA Methods. In present study, rainfall modeling and forecasting is tried for many traditional Algorithms like ARMA, ARIMA, SARIMA etc, but no model gives the good fit for this data. After the rigorous search researcher preferred machine earning technique called extreme gradient boosting algorithms for forecasting.

**Aim:**

This study aims to build a multivariate rainfall prediction model using the best performing technique to date namely the Extreme Gradient Boosting.

**Objective:**

* Rainfall greatly affects human life in various sectors including agriculture, transportation, etc. and also can affect natural disasters such as drought, floods, and landslides.
* This situation prompts us to build an accurate rainfall prediction model so that prescriptive measures can be made
* As a significant part of human existence and a key supply of fresh water, rainfall prediction is a topic we will attempt to cover in this project.
* Fresh water is always a vital resource for human survival. It is used for farming, washing, and many more functions in addition to drinking.

## **Problem statement:**

* Climate is a important aspect of human life. So, the Prediction should accurate as much as possible. In this Project we will try to deal with the prediction of the rainfall which is also a major aspect of human life and which provide the major resource of human life which is Fresh Water.
* Making a good prediction of climate is always a major task now a day because of the climate change.
* A bad rainfall prediction can affect the agriculture mostly framers as their whole crop is depend on the rainfall and agriculture is always an important part of every economy. So, making an accurate prediction of the rainfall somewhat good.

**Literature Review :**

* Potdar et al. (2019) [29] used regression analysis, MK-test and Sen’s analysis to analyse the long-term time series data for prediction of rainfall. This explains that the momentum of spatial variability shows much impact on the behaviour of the rainfall pattern.
* Lata, K et al. (2020) [18] constructed SARIMA model for prediction of rainfall behaviour. This model explains, stochastic modeming is also best one, to predict the rainfall.
* Shuni Qian et al. (2020) build the conventional statistical models based on GCMs and SST dipoles with bias and without bias correction to prediction of monsoon rainfall in the Yangtze River basin. These models are the most appropriate dipoles for the forecaster and it creates a functional relationship between the SST dipoles and monsoon rainfall.
* Prediction of rainfall by using extreme gradient boost (XG boost) in Vishakapattanam area, Andhra Pradesh, Dr. Kesavulu Poola and P Hema Sekhar,(2021) This paper describes, how XG Boost Analysis predicts the rainfall in Monthly Scales. In the paper we recommend that XG-Boost Model is best fit to forecast the rainfall up to 3 to 5 years with 95% accuracy by the evidence of Analyzing last 30 years of data (1987-2017) in Vishakapattanam region.
* Machine learning techniques to predict daily rainfall amount, Chalachew Muluken Liyew\* and Haileyesus Amsaya Melese, e Journal of Big Data (2021), The main objective of this study is to identify the relevant atmospheric features that cause rainfall and predict the intensity of daily rainfall using machine learning techniques. The Pearson correlation technique was used to select relevant environmental variables which were used as an input for the machine learning model.
* Rainfall prediction using Extreme Gradient Boosting, M T Anwar, E Winarno , W Hadikurniawati1 and M Novita, Conference Series 1869 (2021), . This study aims to build a multivariate rainfall prediction model using the best performing technique to date namely the Extreme Gradient Boosting. This model is built based on 7 years of historical weather data collected by the weather station.

**Methodology:**

**DATA COLLECTION:**

* To study and analyze the selected machine learning algorithms, row data is collected from the UCI.
* The data is collected form the meteorology station includes input features and output features.
* Since the data is row data which is collected from the measuring devices in the station, the data contains missing values and it is not arranged in the appropriate format for the experiment.

**DATA PREPROCESSING:**

The data preprocessing step includes the data conversion, manage missing values, categorical encoding and splitting dataset for training and testing dataset. A total of 20 years data is collected from the metrology office. Data conversion is the process of converting the data into the appropriate data format for the experiment. The row data is collected from meteorology station arranged in year based and the attributes in rows that need to combine and rearrange the features in column. Finally the data is converted from excel data to CSV data.

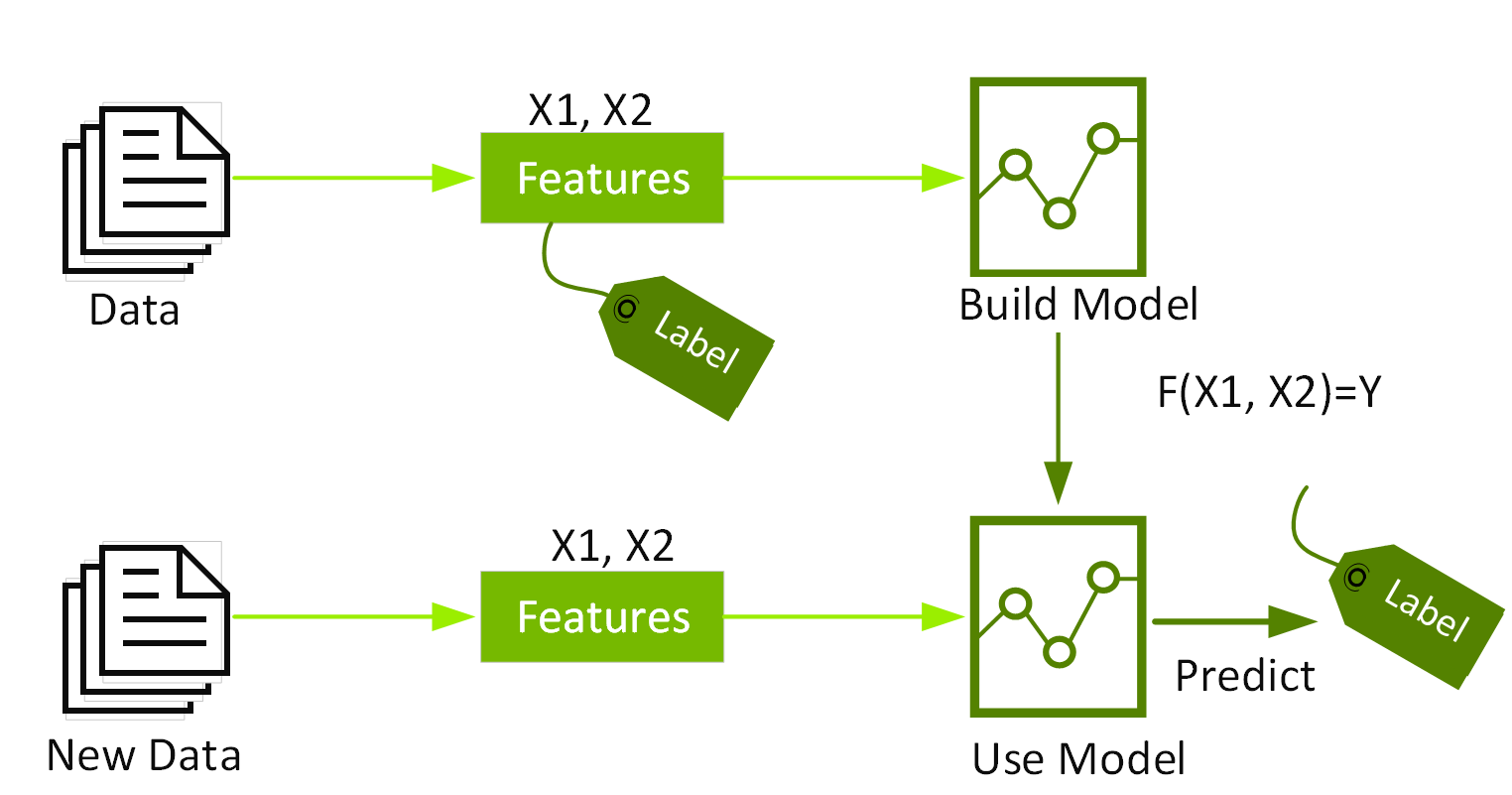
The dataset collected at the meteorology office is with missing values so that the missing value of the target variable is removed and the other features are filled using mean of the data. Encoding the dataset is performed and then the dataset is prepared for experiment. The important features for rainfall prediction are selected and the dataset is splitting as 80% for training and 20% for testing considered as an input for the model.

**MODEL:**

**Extreme Gradient Boosting:**

 XG boost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.



**Software Requirement:**

* Python
* Anaconda Distribution
* Jupyter Lab (IDE)
* Streamlit

**System Design and Flow :**

Rainfall Data Set

Preprocessing

Feature Selection

Prediction And Result

Model Training

Train Data

Test Data

**Algorithm and Steps:**

Step1: Import the rainfall data set csv file.

Step2: Fill the missing values with mean value of the data.

Step3: Scaling the features- scaling the data to a fixed scale.

Step4: The data is divided into training set (70%) and testing set (30%).

Step5: Train data using XG boost Algorithm.

Step7: Performance and metrices.

Step8: Display the results

**References:**

1. Verma A P and Chakraborty B S 2020 Performance Estimation of ARIMA Model for Orographic Rainfall Region 2020 URSI Regional Conference on Radio Science (URSIRCRS)
2. Hartomo K D, Prasetyo S Y J, Anwar M T and Purnomo H D 2019 Rainfall Prediction Model Using Exponential Smoothing Seasonal Planting Index (ESSPI) For Determination of Crop Planting Pattern Computational Intelligence in the Internet of Things (IGI Global) pp
3. Anwar M T, Nugrohadi S, Tantriyati V and Windarni V A 2020 Rain Prediction Using RuleBased Machine Learning Approach Adv. Sustain. Sci. Eng. Technol.
4. Ramli I, Rusdiana S, Basri H, Munawar A A and others 2019 Predicted Rainfall and discharge Using Vector Autoregressive Models in Water Resources Management in the High Hill Annual Conference on Science and Technology (ANCOSET 2020) Journal of Physics: Conference Series 1869 (2021) 012078 IOP Publishing doi:10.1088/1742-6596/1869/1/012078 5 Takengon IOP Conference Series: Earth and Environmental Science vol 273 p 12009
5. Sivajothi R and Karthikeyan K 2019 Forecasting of Rainfall, Average Temperature, Vapor Pressure and Cloud Cover Using Vector Autoregression Model J. Comput. Theor. Nanosci.
6. Pham B T, Le L M, Le T-T, Bui K-T T, Le V M, Ly H-B and Prakash I 2020 Development of advanced artificial intelligence models for daily rainfall prediction Atmos. Res. 237
7. Parida B R, Behera S N, Bakimchandra O, Pandey A C and Singh N 2017 Evaluation of satellite-derived rainfall estimates for an extreme rainfall event over Uttarakhand, Western Himalayas Hydrology 4
8. Anwar M T, Pumomo H D, Prasetyo S Y J and Hartomo K D 2018 Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (IEEE) pp 409–15
9. Chen T, He T, Benesty M, Khotilovich V and Tang Y 2015 Xgboost: extreme gradient boosting R Packag. version
10. Gumus M and Kiran M S 2017 Crude oil price forecasting using XGBoost 2017 International Conference on Computer Science and Engineering (UBMK) pp 1100–3
11. Ogunleye A A and Qing-Guo W 2019 XGBoost model for chronic kidney disease diagnosis IEEE/ACM Trans. Comput. Biol. Bioinforma.
12. ] Shi X, Li Q, Qi Y, Huang T and Li J 2017 An accident prediction approach based on XGBoost 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) pp 1–7
13. Jain R and Nayyar A 2018 Predicting employee attrition using xgboost machine learning approach 2018 International Conference on System Modeling & Advancement in Research Trends (SMART) pp 113–20
14. Pan B 2018 Application of XGBoost algorithm in hourly PM2. 5 concentration prediction IOP Conference Series: Earth and Environmental Science vol 113 p 12127
15. Dhaliwal S S, Nahid A-A and Abbas R 2018 Effective intrusion detection system using XGBoost Information 9 149
16. Le V M, Pham B T, Le T-T, Ly H-B and Le L M 2020 Daily Rainfall Prediction Using Nonlinear Autoregressive Neural Network Micro-Electronics and Telecommunication Engineering (Springer) pp 213–21