iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53577

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Rainfall Prediction using Extreme Gradient Boosting

Abhinav Kulkarni[1], Shaikh Adil[2], Pingalkar Venkatesh[3], Rajas Uchagaonkar[4], Prof. S.S. kolte[5]
[1, 2, 3, 4]*UG Student,* [5]*Assistant Professor, Department of computer Engineerimng AISSMS, Pune*

Abstract: ***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. Previous research on rainfall prediction uses models that have their limitations and thus produce poor performance. 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.***

I. 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. The world's welfare is agriculture. The achievement of agriculture is dependent on rainfall. It also helps with water resources. Rainfall information in the past helps farmers better manage their crops, leading to economic growth in the country. Prediction of precipitation is beneficial to prevent flooding that saves people's lives and property. Fluctuation in the timing of precipitation and its amount makes forecasting of rainfall a problem for meteorological scientists. To overcome these problems, a machine learning technology is used which is predictive analysis that is a branch of data mining which predicts the future probabilities and trends. Prediction is the phenomenon of knowing what may happen to a system in the near future. Since rainfall is the major causes of calamities like floods and typhoons, predicting the occurrence of rainfall will help us to be prepared for these calamities. The basic procedures involved are first identifying an initial model, second repeatedly changing the model by removing a predictor variable based on a criteria and then terminating the process when a model which fits the data well. Here, various rainfall prediction projects were developed using multiple linear regression and other models. This proposed method uses Australian meteorological dataset to predict the rain fall. Usually machine learning algorithms are classified into two major categories i.e. unsupervised learning and supervised learning. All of the clustering algorithms come under the supervised machine learning. Even though many models have developed, it is necessary for doing research using machine learning algorithms to get accurate predictions. The error free prediction provides better planning in the agriculture and other industries, Henceforth we have used the CatBoost model for faster and accurate prediction.

A. 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.

B. Proposed Work

The Proposed system consists of using the CatBoost Algorithm. It performs very well giving an AUC (Area under curve) score 0.8 and ROC(Receiver operating characteristic curve) score as 89.ROC is evaluating curve and AUC presents degree or measure of separability as this model is capable of distinguishing between classes. An Exploratory data analysis is done to examine data distribution, outliers and provides tools or visualizing and understanding the data through graphical representation.

II. LITERATURE REVIEW

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) pp 1–4 This situation prompts us to build an accurate rainfall prediction model so that prescriptive measures can be made. Previous research on rainfall prediction uses models that have their limitations and thus produce poor performance. 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. The result had demonstrated that the model is capable of producing accurate predictions for daily rainfall estimates with training RMSE of 2.7 mm and the testing MAE of 8.8 mm

Recent Development in Econometric Modeling and ForecastingGang LI, Stephen Witt Eighty-four post-1990 empirical studies of international tourism demand modeling and forecasting using econometric approaches are reviewed. New developments are identified, and it is shown that applications of advanced econometric methods improve the understanding of international tourism demand. An examination of the 22 studies that compare forecasting performance suggests that no single forecasting method can outperform the alternatives in all cases.

The time-varying parameter (TVP) model and structural time-series model with causal variables, however, perform consistently well.

Anwar M T, Nugrohadi S, Tantriyati V and Windarni V A 2020 Rain Prediction Using RuleBased Machine Learning Approach Adv. Sustain. Sci. Eng. Technol. 2 Rain prediction is an important topic that continues to gain attention throughout the world. The rain has a big impact on various aspects of human life both socially and economically, for example in agriculture, health, transportation, etc. Rain also affects natural disasters such as landslides and floods. The various impact of rain on human life prompts us to build a model to understand and predict rain to provide early warning in various fields/needs such as agriculture, transportation, etc. This research aims to build a rain prediction model using a rule-based Machine Learning approach by utilizing historical meteorological data. The experiment using the J48 method resulted in up to 77.8% accuracy in the training model and gave accurate prediction results of 86% when tested against actual weather data in 2020.

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), This study aims to predicted rainfall and discharge by using Vector Autoregressive (VAR), rainfall and discharge data in 2008-2015 is used to predicted rainfall for the next five years 2016-2020. While the data for 2016- 2017 is used as a comparison for the predictive data obtained in this study. The results of the actual data accuracy test with predictive data from the VAR (Vector Autoregressive) model, which is done with the NSE (Nash-Sutcliffe efficiency) worth 0.9522. Predicted rainfall value with actual rainfall based on R2 value (coefficient of determination) is 0.6584 or 66%. While the debit prediction test with actual discharge has a R 2 coefficient of determination of 0.0691 or 6.91%. It is because there are other factors that affect discharge other than rainfall. Rainfall in the Takengon plateau which is predicted to be unstable in the future. This condition is the same as the rainfall in the previous periods which fluctuated at the same point. Rainfall prediction results until 2020, has the highest peak season in December.

Sivajothi R and Karthikeyan K 2019 Forecasting of Rainfall, Average Temperature, Vapor Pressure and Cloud Cover Using Vector Autoregression Model J. Comput. Theor. Nanosci. 16 1862–9 The major problem is how to study the past, present, and future scenario of the essential climatic variables. In this article, the present study was to develop a suitable vector autoregression (VAR) model for forecasting monthly rainfall, average temperature, vapor pressure and cloud cover of Idukki district in Kerala state in India. The test for stationarity of the climatic variables has been confirmed with augmented Dickey-Fuller (ADF). The order of the Vector autoregression model was selected using Hannan-Quinn information criteria, Akaike information criterion, and Schwarz information criteria. The least square method has been used to estimate the parameters of the Vector autoregression model. The Structural analyses were well performed using impulse response function (IRF) and forecast error variance decomposition (FEV).

These structural analyses revealed that the rainfall, average temperature, vapor pressure, and cloud cover would be interrelated. Finally, Rainfall average temperature, vapor pressure, and cloud cover were forecasted for the next twenty years from January 2003 to December 2022 using the best-selected model VAR. The forecasted values showed an upward trend in average temperature and vapor pressure fluctuated in rainfall downward trend in cloud cover. This is a shocking situation for the environment so we should start off to control and save our entire world environment.

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 104845 , In this study, the main objective is to develop and compare several advanced Artificial Intelligent (AI) models namely Adaptive Network based Fuzzy Inference System optimized with Particle Swarm Optimization (PSOANFIS), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for the prediction of daily rainfall in Hoa Binh province, Vietnam. For this, meteorological variable parameters such as maximum temperature, minimum temperature, wind speed, relative humidity and solar radiation were collected and used as input parameters and daily rainfall as an output parameter in the models. Validation of the developed models was achieved using various quality assessment criteria such as correlation coefficient (R) and Mean Absolute Error (MAE), Skill Score (SS), Probability of Detection (POD), Critical Success Index (CSI), and False Alarm Ratio (FAR). The results showed that all the AI models provided reasonable predictions of daily rainfall but the SVM was found to be the best method for predicting rainfall. This method was also found to be the most robust and efficient prediction model while taking into account of input variability using the Monte Carlo approach. This AI based study would be helpful in quick and accurate prediction of daily rainfall.

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 22 This unique study was conducted to evaluate three satellite based rainfall products (i.e., TMPA-3B42, Global Satellite Mapping of Precipitation (GSMaP), and NOAA CPC Morphing Technique (CMORPH)) against the observed rain gauge-based India Meteorological Department (IMD) gridded dataset for this rainfall episode. The results from this comprehensive study confirmed that the magnitude of precipitation and peak rainfall intensity were underestimated in TMPA-3B42 and CMORPH against gauge-based IMD data, while GSMaP showed dual trends with under- and over-predictions. From the results of the statistical approach on the determination of error statistic metrics (MAE (mean absolute error), NRMSE (normalized root mean square error), PBIAS (percent bias), and NSE (Nash-Sutcliffe efficiency)) of respective satellite products, it was revealed that TMPA-3B42 predictions were more relevant and accurate compared to predictions from the other two satellite products for this major event. The TMPA- 3B42-based rainfall was negatively biased by 18%. Despite these caveats, this study concludes that TMPA-3B42 rainfall was useful for monitoring extreme rainfall event in the region, where rain- gauges are sparse.

This is an introductory document of using the xgboost package in R. xgboost is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by (Friedman, 2001) (Friedman et al., 2000). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. The package is made to be extendible, so that users are also allowed to define their own objectives easily. It has several features: 1. Speed: xgboost can automatically do parallel computation on Windows and Linux, with openmp. It is generally over 10 times faster than gbm. 2. Input Type: xgboost takes several types of input data: • Dense Matrix: R's dense matrix, i.e. matrix • Sparse Matrix: R's sparse matrix Matrix::dgCMatrix • Data File: Local data files • xgb.DMatrix: xgboost's own class. Recommended. 3. Sparsity: xgboost accepts sparse input for both tree booster and linear booster, and is optimized for sparse input. 4. Customization: xgboost supports customized objective function and evaluation function 5. Performance: xgboost has better performance on several different datasets

A. Objectives

1) The proposed method improves results in terms of accuracy and prediction by forecasting the rainfall for the dataset using the XG boost algorithm.

2) This model was created using historical weather information gathered by the weather station over a period of seven years.

3) In this study, an XGBoost-based model for predicting rainfall is proposed.

III. RESEARCH METHODOLOGY

Machine learning is defined as an automated process that mines patterns from a dataset. The art of developing and using models that make predictions based on patterns extracted from data form the past is called predictive data analytics.

A. Algorithm

1) Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.

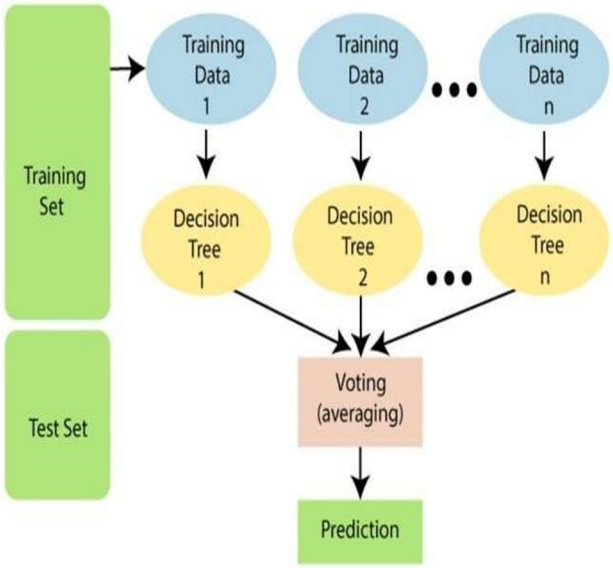


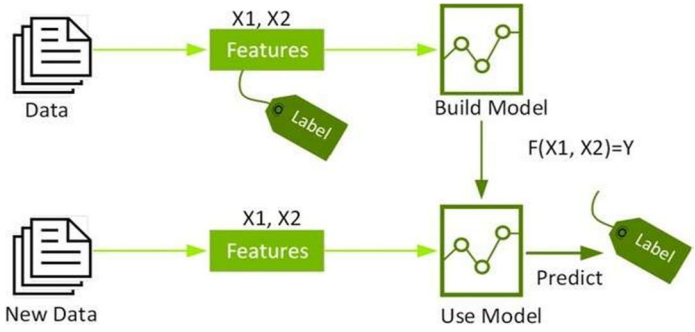
Figure1.1: Random Forest Algorithm Assumptions for Random Forest:

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.The predictions from each tree must have very low correlations

2) 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.



3) Boosting

Boosting is an ensemble modelling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

4) Gradient Boosting

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

5) XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. XGBoost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

B. Training algorithm

For Rainfall prediction, there are several different optimization algorithms used models development.

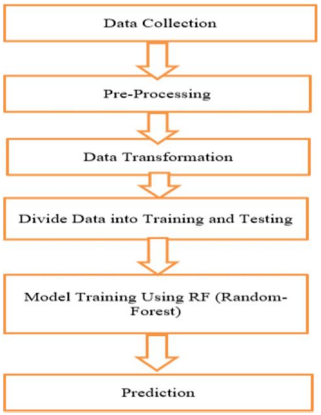


Fig 1.2: The Training Architecture Construction of models

1) Data Preprocessing

a) Noise Removal

It is very important for making the data useful because noisy data can lead to poor results. In telecom dataset, there are a lot of missing values, incorrect values like ''Null'' and imbalance attributes in the dataset.

b) Feature selection

Feature Selection is a crucial step for selecting the relevant features from a dataset based on domain knowledge. A number of techniques exist in the literature for feature selection in the context of churn predictions.

2) Training the Network

The primary goal of training is to minimize an error using error techniques. In this Project We will training machine learning model using random forest algorithm. this is classifier algorithm .

3) Testing

Through this process, the RF fined the predicted and compares it with the input values using data that was not used in training or validation process. At this stage no adjustment occurs to weights.

C. Performance Evaluation Matrix

In this study, the proposed churn prediction model is evaluated using accuracy, precision, and recall, f- measure, and ROC area. Equation 1 calculates the accuracy metric. It identifies a number of instances that were correctly classified.

Accuracy = (TP + TN) (TP + TN + FP + FN) (1)

Here ‘‘TN’’ stands for True Negative, ‘‘TP’’ stands for True Positive,
“FN’’ stands for False Negative and ‘‘FP’’ stands for False Positive.
TP Rate is also known as sensitivity. It tells us what portion of the data is correctly classified as positive.
For any classifier, the TP rate must be high. TP rate is calculated by using Equation 2.
TP Rate = True Positives Actual Positives (2)
FP Rate tells us which part of the data is incorrectly classified as positive. The result of the FP rate must be low for any classifier. It is calculated by using Equation3.
FP Rate = False Positives Actual Negatives (3)
Accuracy, also known as Positive Predictive Value (PPV), indicates which part of the prediction data is positive. It is calculated by using Equation 4.
Precision = True Positive (True Positive + False Positive) (4)
The recall is another measure for completeness i.e. the true hit of the algorithm. It is the probability that all the relevant instances are selected by the system. The low value of recall means many false negatives. It is calculated by using Equation 5.
Recall = (True Positive) (True Positive + False Negative) (5)
The F-measure value is a trade-off between correctly classifying all the data points and ensuring that each class contains points of only one class. It is calculated by using Equation 6.
$F - measure = 2 * Precision * Recall Precision + Recall$ (6)
ROC area denotes the average performance against all possible cost ratios between FP and FN. If the ROC area value is equal to 1.0, this is a perfect prediction. Similarly, the values 0.5, 0.6, 0.7, 0.8 and 0.9 represent random prediction, bad, moderate, good and superior respectively. Values of ROC areas other than these indicate something is wrong.

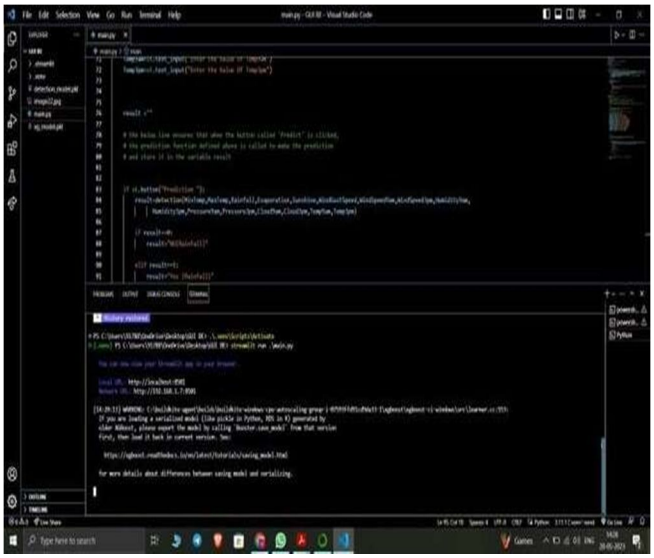
D. Expected Outcome

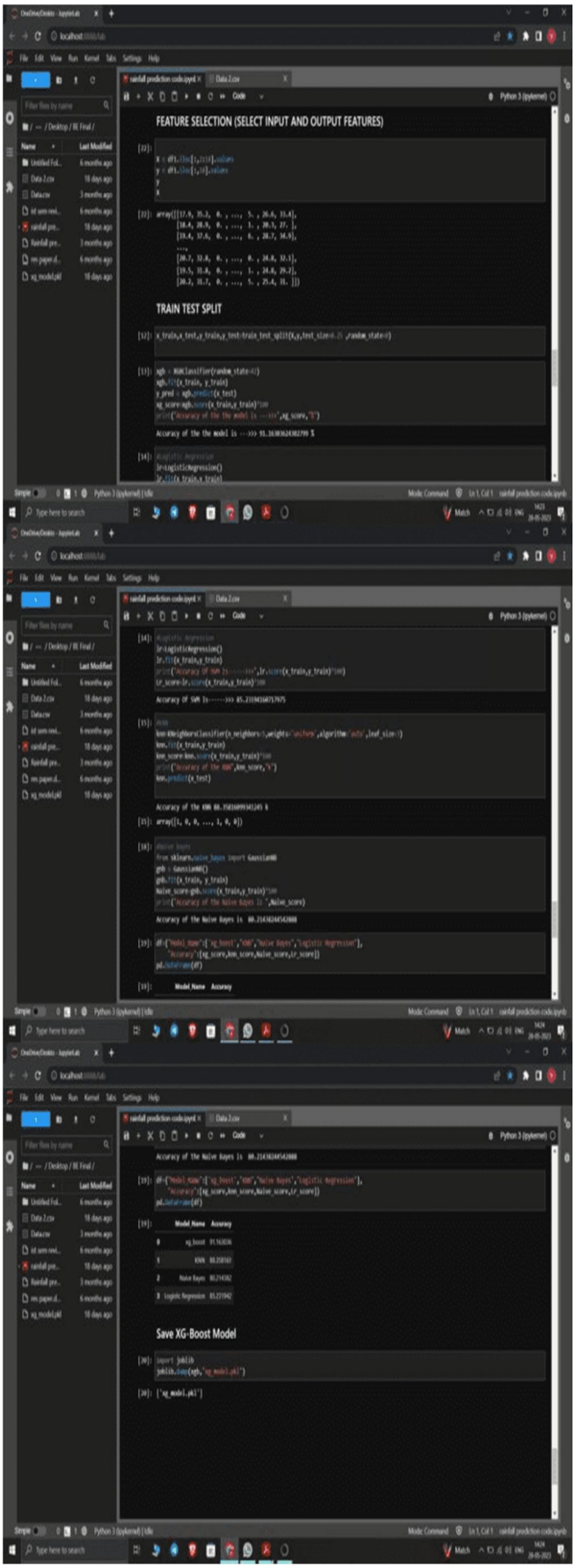
1) This work proposes an XGBoost-based non- linear multivariate rainfall forecasting model.
2) The model will be built based on historical weather data collected by the weather station. Whether it rains tomorrow or not, this project has an expected outcome.

E. Future Scope

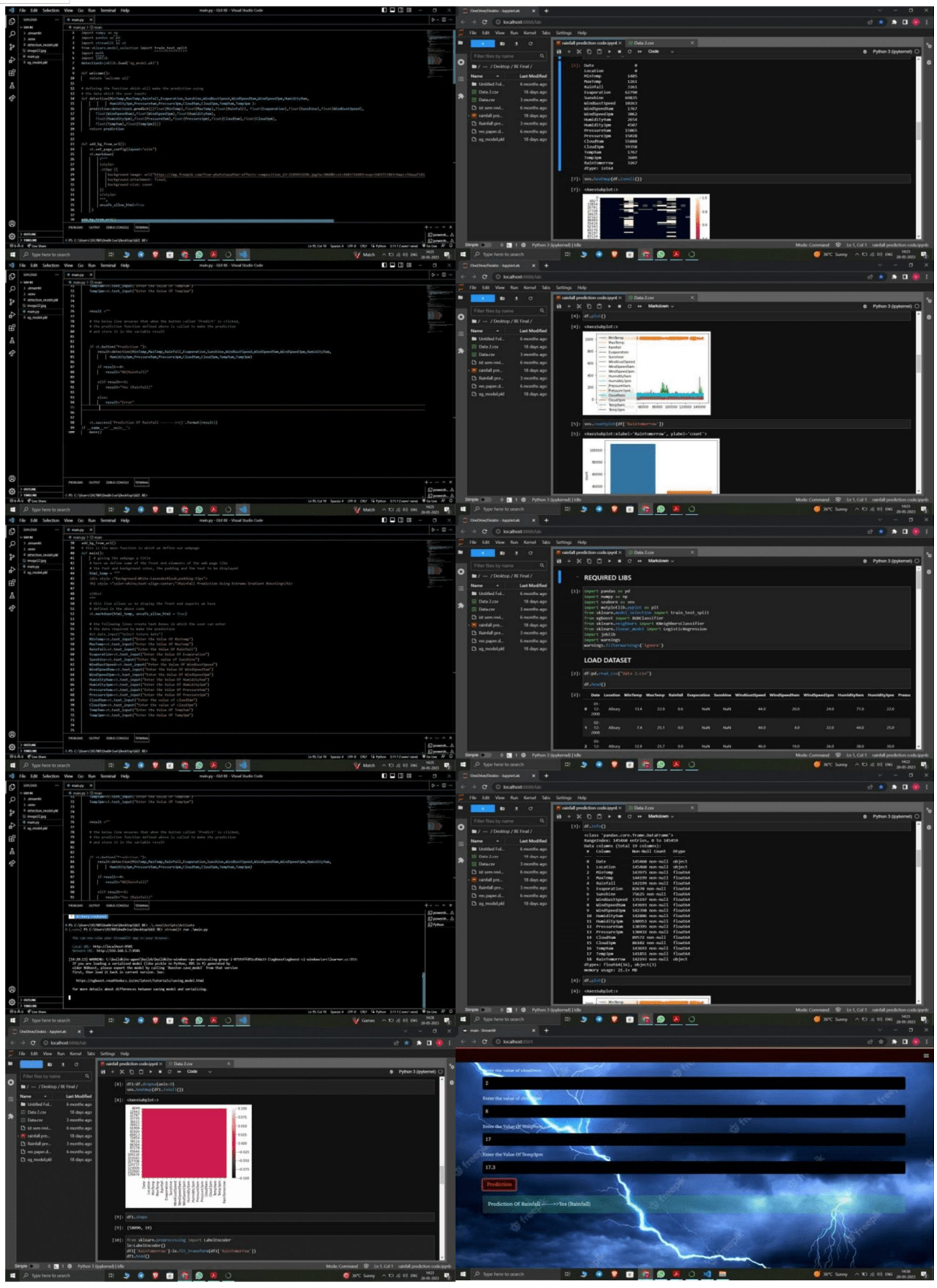
1) As rainfall is dependent on the various parameters it is also required to study how other meteorological parameters affect the Rainfall prediction.
2) We can also perform the same exercise on hourly data using various parameters to forecast next hour rainfall.
3) A study can also be done using more observations for particular region or area, and design this kind of model on big data framework so that computation can be faster with higher accuracy.

IV. IMPLEMENTATION



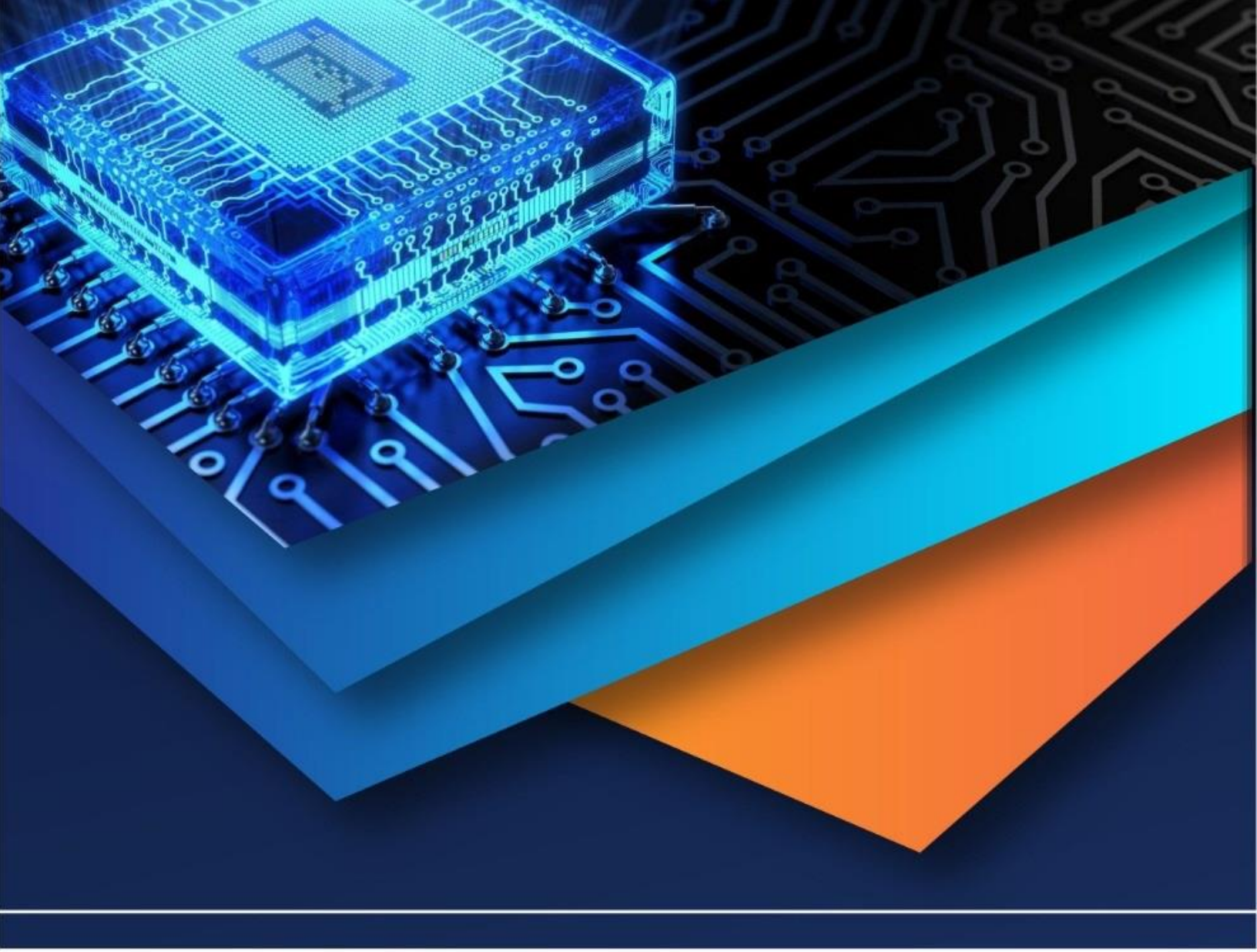

FEATURE SELECTION (SELECT INPUT AND OUTPUT FEATURES)
TRAIN TEST SPLIT
Save XG-Boost Model

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com



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) pp 1–4
[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 234–55
[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. 2
[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
[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. 16 1862–9
[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 104845
[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 22
[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 0.4-2 1–4
[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
[17] Yang Liu, Qingzhi Zhao, Wanqiang Yao, Xiongwei Ma, Yibin Yao and Lilong Liu, "Short-term rainfall forecast model based on the improved BP–NN algorithm",2019.Available:https://www.natur e.com/articles/s41598-019-56452-5[Submitted on 24 December 2019]
[18] Vishal Morde. "XGBoost Algorithm: Long May SheReign!",2019.Available:https://towardsda tascience.c om/https-medium-com- vishalmorde-xgboost-algorithmlongshe-may- rein-edd9f99be63d [Submitted on 8 April 2019]
[19] Ramya Bhaskar Sundaram, "An End-to-End Guide to Understand the Math behind XGBoost", 2018.Available:https://www.analyticsvidhya. com/blog/2018/09/an-endto-end-guide-to- understand-the-mathbehind xgboost/[Submitted on 6 September 2018]
[20] Available:https://www.kaggle.com/prashant1 11/ca tboost-classifier-in-python



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY