Rainfall Prediction Using SVM and XGBoost Algorithms

[1]Vijayasharmila S, [2]Praveen Kumar K, [3]Ranjith Kumar R, [4]Surya Prakash O S

[1]Assistant Professor, [2,3,4] Student
[1,2,3,4]Department of Information Technology,
K.L.N College of Engineering, Sivagangai, Tamil Nadu, India

Abstract: Accurate rainfall prediction has become very complicated in recent times due to climate change and variability. Rainfall prediction is one of the challenging tasks in weather forecasting. Accurate and timely rainfall prediction can be very helpful to take effective security measures in advance regarding on-going construction projects, transportation activities, agricultural tasks, flight operations and flood situation, etc. Data mining techniques can effectively predict the rainfall by extracting the hidden patterns among available features of past weather data. This research contributes by providing a critical analysis and review of latest data mining techniques, used for rainfall prediction. In our proposed system we propose an efficiency of classification algorithms in rainfall prediction has flourished. The study contributes to using various classification algorithms for rainfall prediction in the different ecological zones of Ghana. The classification algorithms include Support Vector Machine (SVM) and XGBoost (XGB). The classification result based on accuracy, precision, recall, f1-score, sensitivity, and specificity. Rain fall prediction Data Mining field concentrate on Prediction more often as compared to generate exact results for future purpose. *(Abstract)*

Index Terms: Rainfall prediction, Classification algorithms, Support Vector Machine, XGBoost, Data mining.

I. INTRODUCTION:

All Accurate and timely rainfall prediction is expected to inject a new intervention phase to the affected sectors accosted with the negative propensities of rainfall extremes. These critical sectors include but are not limited to energy, agriculture, and others, which are greatly affected by rainfall. Weather forecasting ensures the sustainable development of society and economy. Therefore, the interest in forecasts has started since 650 BC, where Babylonians tried to predict weather based on observations of clouds (observed patterns). Then, multiple philosophers proposed various forecasting theories. Over time, it was noticed that these theories were not adequate. Consequently, it was perceived that there is a need to understand the weather from a broader perspective. With the invention of new instruments, measurement of the atmosphere was undertaken. Various instruments, such as the telegraph and radios one, allowed better monitoring of weather conditions. Nowadays, these instruments are used to record weather conditions. For modern rainfall forecasting, forecasts were produced before the invention of the computer, where Lewis Fry Richardson used arithmetic equations to predict weather after World War I (1922). Consequently, scientists introduced new methods that were developed along with the vast spread of technology. Nowadays, scientists use different methods to apply forecasts. Because to its relevance to human life and needs, weather forecasting is applied everywhere in the world. The increasing availability of climate data during the last decades (observational records, radar and satellite maps, observations from ship and aircraft, proxy data, etc.) makes it important to find an effective and accurate tools to analyze and extract hidden knowledge from this huge data. Meteorological data mining is a form of Data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Useful knowledge can play important role in understanding the climate variability and climate prediction.

II. RELATED WORKS:

[1] D. T. Bui, P. Tsangaratos, P.-T.-T. Ngo, T. D. Pham, and B. T. Pham, ``Flash flood susceptibility modeling using an optimized fuzzy rule-based feature selection technique and tree-based ensemble methods," Sci. Total Environ., vol. 668, pp. 1038_1054, Jun. 2019.

The main objective of the present study was to provide a novel methodological approach for flash flood susceptibility modeling based on a feature selection method (FSM) and tree-based ensemble methods. The FSM used a fuzzy rule-based algorithm FURIA, as attribute evaluator, whereas GA were used as the search method, in order to obtain optimal set of variables used in flood susceptibility modeling assessments. The novel FURIA-GA was combined with Logit Boost, Bagging and AdaBoost ensemble algorithms. The performance of the developed methodology was evaluated at the Bao Yen district and the Bac Ha district of Lao Cai Province in the Northeast region of Vietnam. For the case study, 654 floods and twelve geo-environmental variables were used. The predictive performance of each model was estimated through the calculation of the classification accuracy, the sensitivity, the specificity, the success and predictive rate curve and the area under the curves (AUC). The FURIA-GA FSM compared to a conventional rule-based method gave more accurate predictive results. Also, the FURIA-GA based models, presented higher learning and predictive ability compared to the ensemble models that had not undergone a FSM. Based on the predictive classification accuracy, FURIA-GA-Bagging (93.37%) outperformed FURIA-GA-Logit Boost (92.35%) and FURIA-GA-AdaBoost (89.03%). FURIA-GA-Bagging also showed the highest sensitivity (96.94%) and specificity (89.80%). On the other hand, the FURIA-GA-Logit Boost showed the lowest percentage in very high susceptible zone and the highest relative flash-flood density, whereas the FURIA-GA-AdaBoost achieved the highest prediction AUC value (0.9740), based on the prediction rate curve, followed by FURIA-GA-Bagging (0.9566), and FURIA-GA-Logit Boost (0.8955). It can be concluded that the usage of different statistical metrics, provides different outcomes concerning the best prediction model, which mainly could be attributed to sites specific settings. The proposed models could be considered as a novel alternative investigation tool appropriate for flash flood susceptibility mapping.

[2] N. Oswal, ``Predicting rainfall using machine learning techniques,'' 2019, *arXiv:1910.13827*.

Rainfall prediction is one of the challenging and uncertain tasks which has a significant impact on human society. Timely and accurate predictions can help to proactively reduce human and financial loss. This study presents a set of experiments which involve the use of prevalent machine learning techniques to build models to predict whether it is going to rain tomorrow or not based on weather data for that day in major cities of Australia. This comparative study is conducted concentrating on three aspects: modeling inputs, modeling methods, and pre-processing techniques. The results provide a comparison of various evaluation metrics of these machine learning techniques and their reliability to predict the rainfall by analyzing the weather data.

[3] C. Kyei-Mensah, R. Kyerematen, and S. Adu-Acheampong, ``Impact of rainfall variability on crop production within the Worobong ecological area of Fanteakwa district, Ghana,'' Adv. Agricult., vol. 2019, May 2019, Art. no. 7930127.

Crop production in the Fanteakwa District is predominantly rainfed, exposing this major livelihood activity to the variability or change in rainfall pattern. The net potential effect of severe changes in rainfall pattern is the disruption in crop production leading to food insecurity, joblessness, and poverty. As a major concern to food production in Ghana, this study seeks to show the relationship between the production of major crops and rainfall distribution pattern in the Worobong Agroecological Area (WAA) relative to food security in the face of climate change. The study analysed the variability in local rainfall data, examining the interseasonal (main and minor) rainfall distribution using the precipitation concentration index (PCI), and determined the pattern, availability of water, and the strength of correlation with crop production in the WAA. Data from the Ghana Meteorological Agency (GMet) spanning a 30-year period and grouped into 3 decades of 10 years each was used. Selected crop data for 1993-2014 was also obtained from the Ministry of Food and Agriculture's District office and analyzed for trends in crop yield over the period and established relationship between the crop data and the rainfall data. Part of the result revealed that rainfall variability within the major seasons in the 3 groups was lower than the minor seasons. It further showed that yields of three crops have declined over the period. Among the strategies to sustain crop production is to make the findings serve as useful reference to inform discussions and policy on adaptive agricultural production methodologies for the area in the face of changing climate.

[4] P. A. Williams, O. Crespo, C. J. Atkinson, and G. O. Essegbey, ``Impact of climate variability on pineapple production in Ghana,'' Agricult. Food Secur., vol. 6, no. 1, pp. 1_14, Dec. 2017.

Background Climate variations have a considerable impact on crop production. For pineapple, variable temperatures and rainfall patterns are implicated, yet there is limited knowledge of the conditions and consequences of such variations. Pineapple production plays a major role in Ghana, primarily via socioeconomic impacts and the export economy. The aims of this study were to assess the impact of current climatic trends and variations in four pineapple growing districts in Ghana to provide stakeholders, particularly farmers, with improved knowledge for guidance in adapting to changing climate. Results Trend analysis, standardized anomaly, correlation analysis as well as focus group discussions were employed to describe climate and yields as well as assess the relationship between climate and pineapple production from 1995 to 2014. The results revealed that, relative to Ga district, temperature (minimum and maximum) in the study areas was increasing over this period at a rate of up to 0.05 °C. Rainfall trends increased in all but Nsawam Adoagyiri district. Rainfall and temperature had different impacts on production, and pineapple was particularly sensitive to minimum temperature as accounting for up to 82% of yield variability. Despite consistent report of rainfall impact on growth stages later affecting quantity and quality of fruits, minimal statistical significance was found between rainfall and yield. Conclusions With continuously increasing stresses imposed by a changing climate, the sustainability of pineapple production in Ghana is challenged. This subsequently has detrimental impacts on national employment and exports capacity resulting in increased poverty. Further research to explore short- and long-term adaption options in response to challenging conditions in the pineapple industry in Ghana is suggested.

[5] K. Owusu and N. Klutse, ``Simulation of the rainfall regime over Ghana from CORDEX,'' Int. J. Geosci., vol. 4, no. 4, pp.785_791,2013, doi:10.4236/ijg.2013.44072.

This paper investigates how well the rainfall regime on which many livelihoods depend, in Ghana is well represented by the Coordinated Regional Climate Downscaling Experiment (CORDEX). The objective of the study is to demonstrate how well the ten CORDEX models can capture the spatial and temporal rainfall seasonality over the southern and northern sub-sections of Ghana. The choice of the sub-sections is based on the fact that south of 8°N experiences a bi-modal rainfall regime while the north has a uni-modal regime. The results indicate that the rainfall over Ghana is associated with high levels of variability at the inter-annual time scale. Particularly over the southern part of Ghana, all the models follow the same trend as represented over Ghana with similar rainfall values as the observation. Over the northern part of Ghana, models record relatively low rainfall agreeing with the observation. However, most of the models overestimate the northern region rainfall as it is in the case of the southern Ghana. CORDEX as shown in this analysis could be useful in providing Ghana with at least 10 different model outputs for impact analysis. Caution is however given that, since individual models give different performance and the fact that models in general have their inherent deficiencies, an ensemble mean of the models could provide a better result.

III. METHODOLOGY:

The methodology of the project is to predict the rainfalls. The predicted feature has a value of 1, then will be a rainy day; if the value is 0, then it will be no rainfall. The dataset is divided into two parts: 80% of the data is reserved for training and 20% of the data is reserved for testing. The machine learning techniques are used to detect the rain occurrence will be tomorrow or not. The proposed hybrid algorithms are XGBoost and Support Vector Machine is used to extracts the more features and efficiently classify the dataset. Finally, it will generate the metrics in terms of accuracy, precision, recall and f1-score.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data. artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

IV. Implementation and Analysis:

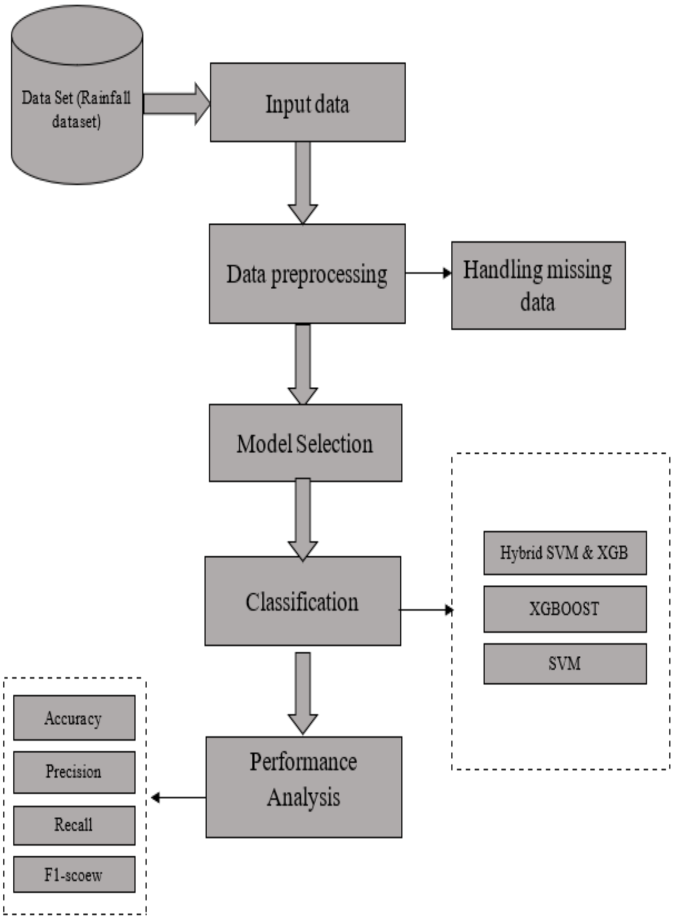

Figure 1: Solution Architecture

A. Module:

- Data Preprocessing.
- Classification.
- Performance and Analysis

B. Module Description:

Data Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data preprocessing task. The data collected to build this model is measure of rainfall and various environmental factors supporting it. Data pre-processing is done via missing data removal and encoding categorial data. In missing data removal, the rows containing null / empty values will be discarded. For categorical data's, encoding is done such as one-hot encoding to convert the categorical data into an integer value. The data is splitted in such a way that 70% is used as training data and remaining 30% is used as testing data.

Classification:

SVM is a linear model used for both classification and regression problem. The algorithm created a line or a hyperplane which separates the data into classes. After creating the optimal hyperplane, it finds the support vectors and then creates a decision boundary in such a way that the separation between two classes must be maximum. It can also support non-linear data too. XGB stands for Extreme Gradient Boosting, it provides parallel tree boosting. This helps when we have many observations in training data or when number of features is minimal than the observation made in training data. After implementation of SVM and XGBoost classification is done. All the data are mapped, and an optimal hyperplane is found. Then a maximized margin is created in such a way that the distance between various classes must be maximum. Any weak links are combined into forming a strong link which can enhance the result done via XGB. XGBoost will also help in reducing the training time and for tree pruning.

PERFORMANCE AND ANALYSIS:

Performance is measured using via the following methods,

- Accuracy $$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ number\ of\ Input\ Samples}$$
- Precision $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$
- Recall $$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
- F1 Score $$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

V. RESULT AND CONCLUSION:

In this process, the Machine learning regression is analyzing the rainfall. First Step the input data was applied into pre-processing method and standardize the range of independent variables or features of data by using scalar method. In classification predict the rainfalls using machine learning algorithms like, Hybrid SVM and XGBoost is implemented and classify the data and it will generate the result based on accuracy, precision, recall and f1-score.

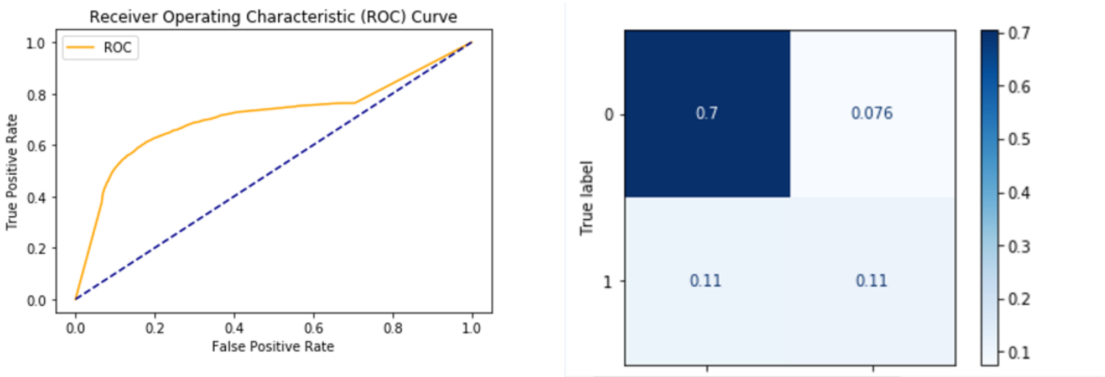

Figure 2: Evaluation of SVM

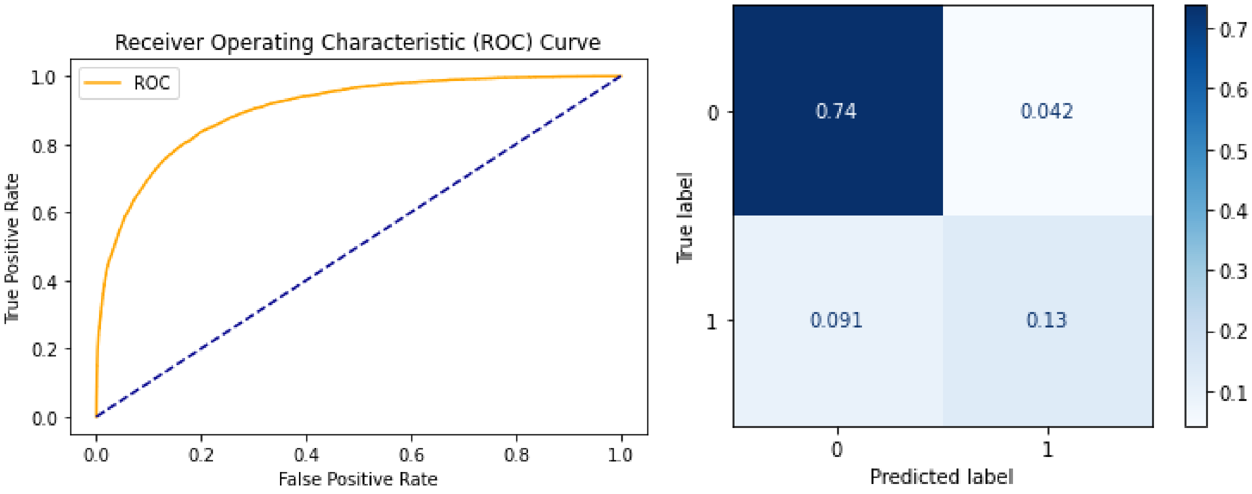

Figure 3: Evaluation for SVM and XGBoost

VI. REFERENCES:

1. D. T. Bui, P. Tsangaratos, P.-T.-T. Ngo, T. D. Pham, and B. T. Pham, ``Flash flood susceptibility modeling using an optimized fuzzy rule-based feature selection technique and tree-based ensemble methods," Sci. Total Environ., vol. 668, pp. 1038_1054, Jun. 2019.
2. C. Kyei-Mensah, R. Kyerematen, and S. Adu-Acheampong, ``Impact of rainfall variability on crop production within the Worobong ecological area of Fanteakwa district, Ghana," Adv. Agricult., vol. 2019, May 2019, Art. no. 7930127.
3. H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, ``Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," Environ. Model. Softw., vol. 101, pp. 1_9, Mar. 2018.

4. R. C. Deo, S. Salcedo-Sanz, L. Carro-Calvo, and B. Saavedra-Moreno, ``Drought prediction with standardized precipitation and evapotranspiration index and support vector regression models,'' in Integrating Disaster Science and Management. Amsterdam, The Netherlands: Elsevier, 2018, pp. 151_174.
5. P. A. Williams, O. Crespo, C. J. Atkinson, and G. O. Essegbey, ``Impact of climate variability on pineapple production in Ghana,'' Agricult. Food Secur., vol. 6, no. 1, pp. 1_14, Dec. 2017.
6. K. Owusu and N. Klutse, ``Simulation of the rainfall regime over Ghana from CORDEX,'' Int. J. Geosci., vol. 4, no. 4, pp.785_791,2013, doi:10.4236/ijg.2013.44072
7. I. Yabi and F. Afouda, ``Extreme rainfall years in Benin (West Africa),'' Quaternary Int., vol. 262, pp. 39_43, Jun. 2012.
8. K. Owusu and N. Klutse, ``Simulation of the rainfall regime over Ghana from CORDEX,'' Int. J. Geosci., vol. 4, no. 4, pp.785_791,2013, doi:10.4236/ijg.2013.44072.
9. G. Di Baldassarre, A. Montanari, H. Lins, D. Koutsoyiannis, L. Brandimarte, and G. Blöschl, ``Flood fatalities in Africa: From diagnosis to mitigation,'' Geophys. Res. Lett., vol. 37, no. 22, pp. 529_546, Nov. 2010.
10. N. K. Karley, ``Flooding and physical planning in urban areas in West Africa: Situational analysis of Accra, Ghana,'' Theor. Empirical Res. Urban Manage., vol. 4, no. 4, pp. 25_41, 2009.