Title Page:

**Improving the efficiency of anticipating the rainfall using XGBoost in comparison with logistic regression.**

Shaik Towhid Hussain1, Dr.C.Sivasankar2

Shaik Towhid Hussain1, Research Scholar,

Department of Computer Science and Engineering, Saveetha School of Engineering,

Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India – 602105

192111049.sse@saveetha.com

Dr.C.Sivasankar2  
professor,  
Department of Computer Science and Engineering,  
Saveetha School of Engineering,  
Saveetha Institute of Medical and Technical Sciences,   
Saveetha University,  Chennai, Tamil Nadu, India, Pincode: 602105  
E-Mail:sivasankarc.sse@saveetha.com

**Keywords:** XG Boost algorithm , logistic regression , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# ABSTRACT

**Aim**: The objective of the work is to evaluate the accuracy and in predicting the rainfall using machine learning algorithms XGBoost (XGB) classification and Logistic Regression (LR) algorithms. **Materials and Methods:**  XGBoost classifier is applied on a weatherAUS dataset that consists of 145461 records. A framework for rainfall prediction machine learning algorithms comparing XGBoost and Logistic Regression classifiers has been proposed and developed. The sample size was measured as 10 per group. Sample size was calculated using clinical analysis, with alpha and beta values 0.05 and 0.5, 95% confidence, pretest power 80% and enrolment ratio 1. The significance value (p) obtained for both accuracy and precision is 0.019, which is less than 0.05. The accuracy and the precision of the classifiers were evaluated and recorded. **Results:** The rainfall on the used dataset is predicted by the machine learning algorithm XG Boost classifier with 97% accuracy, whereas the LR predicts the same event with 94% accuracy**. Conclusion:** According to the study, the XG Boost algorithm predicts rainfall more accurately than LR method.

**Keywords:** XG Boost algorithm , logistic regression , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# INTRODUCTION

The primary occupation of India has been agriculture so the economy of the country highly depends on it. Therefore, prediction of rainfall is important for the economic growth of the country. Rainfall prediction has been one of the most challenging and difficult tasks in the world. India has a population of over a billion people, agriculture is the primary income of more than 60% of the population (Brownlee 2020). The advancement of technology and resources in today's world is increasing day by day. Weather of a particular location can be predicted using the applications of science and technology. Many attempts were made to predict the rainfall in the past to help the farmers to do agriculture . Informally these attempts were started way back in the eighteenth century and formally started in the nineteenth century (Dhyani 2020). Weather forecasting basically refers to the analyzing and predicting the climatic conditions of a specific region in the future. Rainfall prediction comes under the category of weather forecasting (Yuan and Forshay 2021). Interest in weather prediction started in the earliest days to help farmers for effective agriculture, and many rainfall prediction techniques were introduced to effectively predict the rainfall. Over time all those techniques evolved and each of the techniques has its own accuracy and efficiency (Kuradusenge, Kumaran, and Zennaro 2020). The proposed study helps in effectively predicting the rainfall that helps the overall development of the country.

There are 105 research articles published on the rainfall prediction in IEEE xplore and 165 articles on google scholar and 34 articles were found in sciencedirect. (Fadilah, Wigena, and Djuraidah 2020) presents a new predictor algorithm based on Bayesian Enhanced Approach (BEA) for long-term chaotic time series using Artificial Neural Networks (ANN). In many cases, the simplest representations of prior information in forecasting models are hard to surpass; the Bayes technique gives another approach to include prior knowledge in forecasting models. Predicting conditions, either because previous knowledge isn't available or because prior know ledge isn't useful. As a result, the situation appears to be stable. (Wen et al. 2018) Compared various types of rainfall models which have been used in modeling rainfall occurrence, rainfall amount or combination of both rainfall occurrence and rainfall amount. (Narejo et al. 2021) Proposed a multiple linear regression model in order to predict the rate of precipitation (PRCP). It is based on some weather parameters, such as temperature, wind speed, and dew point. The data used in this research has been provided from the website of the National Climatic Data Center. A Python code using the Pytorch library has been written to develop the model, which applies to Artificial Neural Networks. (Mahmood 2017) Proposed an algorithm for rainfall prediction through the random forest. It achieves an accuracy of 94.00% and outperforms all other feature representations, including decision tree and support vector machine classification algorithms using the same classifier on the same dataset.

The survey revealed a research gap: while many techniques have been proposed for rainfall prediction, most of them have low accuracy rates. It has been shown in numerous studies that LR performs poorly and predicts rainfall with less accuracy. In a study published in 2020, Mohammed et al. evaluate the of several mining classification algorithms in forecasting rainfall. Analysing and contrasting them is the best way to find out which classification algorithm offers the highest accuracy. Thus, for accurate rainfall, the accuracy of the LR and XGBoost algorithm ms are compared in this research paper.

# MATERIALS AND METHODS

The data analytics lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, where this study was conducted, has incredibly well-configured systems that aid in producing accurate results. There were two groups in total that were taken into consideration for the research: group 1 had ten sample sizes, and group 2 also had ten sample sizes. G-power 0.95, alpha value 0.005, beta value 0.95, and confidence interval 95% are used in the computation. The research dataset was obtained from the Kaggle website (Kerneler 2019)..

## XGBoost (XGB) - Group 1

Inputs: WeatherAUS dataset Output: Accuracy

1. Load the WeatherAUS.csv file
2. Divide the whole data records randomly into testing (20%) and training (80%)
3. Identify the output variable
4. Develop XG Boost classification design using training dataset
5. Train the developed classification
6. Project the testing dataset with respect to the training data file
7. Examine the developed classification model.
8. Show the accuracy value obtained.

In this study, sklearn.linear\_model library of XG Boost class was used for training the data. Read weatherAUS.csv file and load the same file. Data records of the dataset are distributed randomly, 80% of the dataset is taken for training and the remaining 20% of the dataset is taken for testing. The output variable is defined then, 3 of 10 XG Boost classifier are developed using the training dataset. Testing dataset is estimated using training dataset. The XG Boost classification design is tested and then accuracy is produced.

## Logistic Regression (LR) - Group 2

Input: weatherAUS dataset Output: Accuracy

1. Load and read weatherAus dataset
2. Choose properties randomly from dataset
3. Develop DT classification criteria as a variable for evaluation.
4. Gini was considered as an argument.
5. Design a classification model using RF classifiers and guess the output for each sample.
6. Random selection process was performed for each predicted output.
7. Most voted outputs were considered as final results.
8. Return the precision value obtained.

The data in this study were trained using the sklearn ensemble library of the LR class. The dataset's data records are dispersed at random, with 80% of the dataset being used for training and the remaining 20% being used for testing. It selects samples at random from the dataset, and decision trees are gathered to forecast the result. All possible outcomes were put to a vote, and the winner was determined by selecting the most votes. A LR Classifier is used by the algorithm (LR).

Google Collab and SPSS software were used to conduct this study. An Intel i3 processor, a 50GB hard drive, 4GB of RAM, and a Windows operating system are required for the hardware and software specifications of the system for evaluation.

# STATISTICAL ANALYSIS

In addition to experimental analysis, the work is statistically evaluated using the SPSS tool. The research objectives were the mean, standard deviation, accuracy, and standard error mean. Accuracy is the dependent variable, and the independent research variables are temperature, humidity, pressure, and wind speed (Dattalo 2013). An independent sample T-Test was used to compare the XGB and LR algorithms.

# RESULTS

Table 1 shows the comparison between the accuracy obtained while evaluating the XG Boost and LR for analysis with various iterations.

Table 2 depicts the various parameters of both groups. The accuracy, Recall, F1 Score and support has been calculated for XGB and LR. The analysis of two groups presents XGB as having better accuracy (97%) and compared to LR.

Table 3 shows the statistical analysis of XG Boost and LR with different test datasets. The table depicts XG Boost algorithm produces better accuracy compared with LR.

Table 4 depicts the statistical analysis of Significant levels for both groups. There is a negligible Significant difference 0.001 for accuracy among the two groups. Hence XG Boost is better than LR.

Figure. 3 inferred the mean accuracy of XG Boost (XGB) and Logistic Regression (LR). The results showed that the XGB has better accuracy (97%) compared to the accuracy of Logistic Regression .

# DISCUSSION

Agriculture has been the backbone of India and prediction of rainfall is a major issue in the agriculture sector. Experimental work was done among two groups: Logistic Regression (LR) and novel tree specific XGBoost(XGB) by varying the test size. From the experimental results (Figure 3 & 4) done in SPSS, the accuracy of 97.00% XGB , whereas Logistic Regression provides the accuracy of 94% . This depicts that XGB is better than Logistic Regression.

There are still certain research limitations even though the suggested methodology produced results that were satisfactory. On large datasets, accuracy evaluation might not yield results that are satisfactory. Additionally, the mean error in XGB is higher than in LR, which is not what is wanted result .

Reducing the average error significantly enhances the present study. One approach to lower mean error and raise accuracy is to optimise the algorithms using algorithmic techniques. Using feature selection algorithms before dataset classification is an additional way to improve classifier accuracy and yield better outcomes.

# CONCLUSION

XGBoosting is a classification technique that uses averaging to improve the accuracy and precision. The work shows that the accuracy and precision for rainfall prediction using XGBoost(XGB) appears to be better than the Logistic Regression(LR). It is found that XGB performs significantly better than LR in predicting the rainfall accurately, but the mean error in XGB is found to be little higher than LR. Hence, it is concluded that the XGBoost (XGB) classifier results in acceptable accuracy(97.00%) than Logistic Regression(LR) accuracy(94.00%)

# DECLARATIONS

## Conflicts of Interests

No conflicts of interest present in this manuscript.

## Author Contributions

Author Towhid played a key role in collecting and analysing data as well as writing the manuscript. Additionally, S.Towhid contributed significantly to conceptualization, data validation and providing critical feedback during manuscript reviews.

## Acknowledgements

The authors would like to thank the management, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the infrastructure needed to successfully complete this project.

**Funding:** We extend our gratitude to the following organisations for the financial support they have provided which helped us to conduct and complete the research.

* 1. Saveetha University.
  2. Saveetha Institute of Medical And Technical Sciences.
  3. Saveetha School of Engineering.

# REFERENCES

Alexandridis, Antonios K., and Achilleas D. Zapranis. 2014. Wavelet Neural Networks: With Applications in Financial Engineering, Chaos, and Classification. John Wiley & Sons. Anwar, M. T., E. Winarno, W. Hadikurniawati, and M. Novita. 2021. “Rainfall Prediction Using Extreme Gradient Boosting.” Journal of Physics. Conference Series 1869 (1): 012078. Brownlee, Jason. 2020. “How to Use XGBoost for Time Series Forecasting.” August 4, 2020. https://machinelearningmastery.com/xgboost-for-time-series-forecasting/. Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi. 2019. “Predictability Assessment of Northeast Monsoon Rainfall in India Using Sea Surface Temperature Anomaly through Statistical and Machine https://doi.org/10.1002/env.2533. Learning Techniques.” Environmetrics. Dattalo, Patrick. 2013. Analysis of Multiple Dependent Variables. Oxford University Press. Dhyani, Sandeep. 2020. “Predicting Rainfall for Agriculture in India Using Regression.” Dublin Business School. https://esource.dbs.ie/handle/10788/4230. Fadilah, Putri Maulidina, Aji Hamim Wigena, and Anik Djuraidah. 2020. “Extreme Rainfall Prediction Using Bayesian Binary Quantile Regression.” International Journal of Scientific and Research Publications (IJSRP). https://doi.org/10.29322/ijsrp.10.11.2020.p10702. kerneler. 2019. “Starter: weatherAUS 536c1115-4.” Kaggle. November 29, 2019. https://kaggle.com/kerneler/starter-weatheraus-536c1115-4. Kuradusenge, Martin, Santhi Kumaran, and Marco Zennaro. 2020. “Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda.” International Journal of Environmental Research and Public Health 17 (11). https://doi.org/10.3390/ijerph17114147. Mahmood, Arshad. 2017. Rainfall Prediction in Australia: Clusterwise Linear Regression Approach. Mohammed, Moulana, Roshitha Kolapalli, N. V. V. Golla, and S. Maturi. 2020. “Prediction Of Rainfall Using Machine Learning Techniques.” https://www.semanticscholar.org/paper/Prediction-Of-Rainfall-Using-Machine-Learning-M ohammed-Kolapalli/9bf0a85ff6323559663cb45c89ae1e51684401dc. Narejo, Sanam, Muhammad Moazzam Jawaid, Shahnawaz Talpur, Rizwan Baloch, and Eros Gian Alessandro Pasero. 2021. “Multi-Step Rainfall Forecasting Using Deep Learning Approach.” PeerJ Computer Science 7. https://doi.org/10.7717/peerj-cs.514. Osman, Ahmedbahaaaldin Ibrahem Ahmed, Ali Najah Ahmed, M. F. Chow, Yuk Feng Huang, and A. El-Shafie. 2021. “Extreme Gradient Boosting (Xgboost) Model to Predict the Groundwater Levels in Selangor Malaysia.” https://doi.org/10.1016/J.ASEJ.2020.11.011. Wen, Huaiyu, Sufang Li, Wei Li, Jianping Li, and Chang Yin. 2018. “Comparision of Four Machine Learning Techniques for the Prediction of Prostate Cancer Survivability.” 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). https://doi.org/10.1109/iccwamtip.2018.8632577. Yen, Meng-Hua, Ding-Wei Liu, Yi-Chia Hsin, Chu-En Lin, and Chii-Chang Chen. 2019. “Application of the Deep Learning for the Prediction of Rainfall in Southern Taiwan.” Scientific Reports 9 (1): 1–9. Yuan, Lifeng, and Kenneth J. Forshay. 2021. “Enhanced Streamflow Prediction with SWAT Using Support Vector Regression for Spatial Calibration: A Case Study in the Illinois River Watershed, U.S.” PloS One 16 (4): e0248489.

# TABLES AND FIGURES

# Table 1. Accuracy values of XG Boost and Logistic Regression algorithms obtained for each iteration while evaluating the dataset for various test sizes.

|  |  |
| --- | --- |
| **GROUP** | **ACCURACY** |
| XGB | 97 |
| XGB | 95 |
| XGB | 96 |
| XGB | 98 |
| XGB | 93 |
| XGB | 97 |
| XGB | 92 |
| XGB | 96 |
| XGB | 98 |
| XGB | 93 |
| LR | 94 |
| LR | 91 |
| LR | 85 |
| LR | 85 |
| LR | 87 |
| LR | 91 |
| LR | 83 |
| LR | 84 |
| LR | 85 |
| LR | 86 |

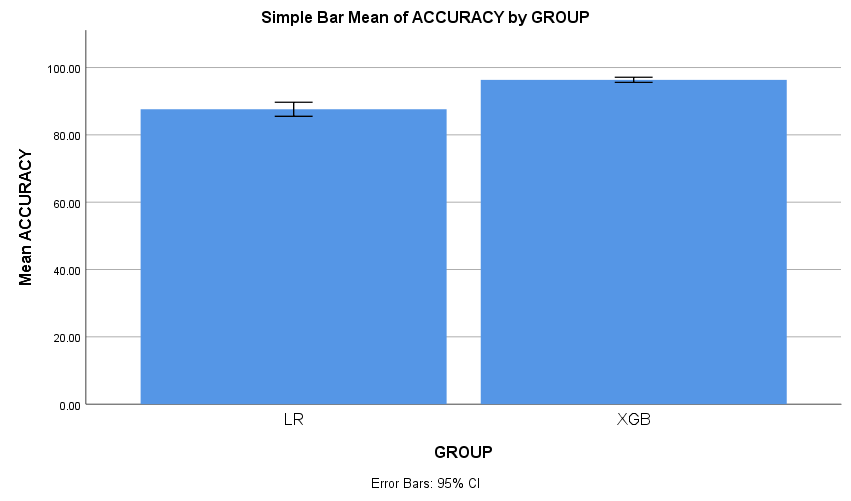
**Table 2**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Statistics** | | | | | |
|  | GROUP | N | Mean | Std. Deviation | Std. Error Mean |
| ACCURACY | XGB | 20 | 96.3500 | 1.63111 | .36473 |
| LR | 20 | 87.6000 | 4.45327 | .99578 |

**Table 3**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Levene’s Test for Equality of Variances | | T-test for Equality of means | | | | | | |
| F | Sig. | t | df | Sig( 2-ta iled  ) | Mean Difference | Std. Error Difference | 95% confidence interval of the Difference | |
| Lower | Upper |
| Accuracy Equal  variances  Assumed | 33.502 | .000 | 8.251 | 38 | .000 | 8.7500 | 1.06047 | 6.60318 | 10.89682 |
| Equal  Variances  Not assumed |  |  | 8.251 | 24.008 | .000 | 8.7500 | 1.06047 | 6.58113 | 10.93867 |

GRAPH :



|  |
| --- |
|  |