## Title Page:

## The foremost technique for enhancing Rainfall forecasting using a XGBoost classifier over decision tree classifier to improve the accuracy of forecasts.

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 , Decision tree , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# ABSTRACT

**Aim**: The machine learning algorithms XG Boost (XGB) and Decision tree (DT) are used in this study to assess the accuracy of rainfall prediction. **Materials and Methods:** The weatherAUS dataset, comprising 145461 records, is subjected to the XG Boost algorithm. This work presents and develops a standard module for rainfall prediction and compares the machine learning techniques DT and XG Boost. Ten thousand sample records were taken from each group for assessment. Clinical analysis was used to measure the sample records; the evaluation's alpha and beta values are 0.5 and 0.05, respectively, and its enrollment ratio is 1. Its confidence percentage is 95%, its pretest power percentage is 80%, and its pretest power percentage is 80%. For accuracy, the obtained significance value (p), which is less than 0.05, is 0.001. Both methods final accuracy were computed and reported.. **Results:** The rainfall on the used dataset is predicted by the machine learning algorithm XG Boost classifier with 97% accuracy, whereas the DT predicts the same event with 87% accuracy**. Conclusion:** According to the study, the XG Boost algorithm predicts rainfall more accurately than DT method.

**Keywords:** XG Boost algorithm , Decision Tree , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# INTRODUCTION

Almost every state's main industry has historically been agriculture, which makes it a major contributor to the US economy. Predicting rainfall is therefore more important for the nation's economic development. Over the years, estimating the amount of rain that will fall on Earth has shown to be the most difficult task. Over 60% of the population in India, a country of over a billion people, depends on agriculture as their main source of income (Kharwal 2020). In the modern world, resources and technological advancements are growing every day. Scientific and technological applications can be used to predict the weather in a specific location. In the past, numerous attempts have been made to forecast rainfall in order to assist farmers with their farming operations. These initiatives were first formally proposed in the nineteenth century, after being suggested informally in the eighteenth (Oswal 2019). In essence, weather forecasting is the study and projection of future climate conditions for a particular area. Predicting rainfall is one type of weather forecasting (Yen et al. 2019). Since the dawn of time, people have been interested in weather prediction, and numerous techniques for predicting rainfall have been developed. Those methods are unique in terms of accuracy and power because they have all developed over time (Shardoor and Rao 2018). The proposed study aids in the precise forecasting of rainfall, which advances the country's overall development.

About rainfall prediction, there are 105 research articles published in IEEE Xplore, 165 research papers on Google Scholar, and 34 articles found on ScienceDirect. (Rudrappa 2021) introduces a novel algorithm that uses artificial neural networks and the Bayesian Enhanced Approach to predict long-term rainfall through time series analysis. Since it is often difficult to outperform the best representation of weather forecast information, the Bayes technique provides an additional way to incorporate past knowledge into forecasting models. forecasting the situation when past information is either unreliable or useless. It seems that things are stable as a result. (2018, Shardoor and Rao) examined various rainfall prediction classifiers that aid in determining the frequency of annual precipitation, its total amount, and the combination of the frequency and total amount of annual precipitation. [(Narejo et al. 2021)](https://paperpile.com/c/1kWtcS/dooI) Proposed a multiple linear regression design for predicting the participation rate. The suggested design considers various weather parameters, including temperature, pressure, wind direction, and speed. The National Climatic Data Centre is the website where the study's data is collected. An algorithm was proposed by (Reddy, 2021) to forecast rainfall using DT. This feature representation outperforms all others and reaches an accuracy of 87.00%.

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 DT performs poorly and predicts rainfall with less accuracy. In a study published in 2020, Mohammed et al. evaluate the precision 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 DT and XGBoost algorithms 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.

## Decision Tree (DT) - 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 DT 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 DT Classifier is used by the algorithm (DT).

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 DT algorithms.

# RESULTS

Table 1 shows the comparison between the accuracy obtained while evaluating the XG Boost and DT 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 DT. The analysis of two groups presents XGB as having better accuracy (97%) and compared to DT.

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

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 DT.

Figure. 3 inferred the mean accuracy of XG Boost (XGB) and Decision tree (DT). The results showed that the XGB has better accuracy (97%) compared to the accuracy of Decision Tree.

# DISCUSSION

### India's economy has historically relied heavily on agriculture, and forecasting rainfall is a big problem for the industry. By altering the test size, two groups—XG Boost (XGB) and Decision Tree(DT) performed experimental work. Based on the SPSS experimental results (Figure 3), XGB has an accuracy of 97.00%, while DT offers an accuracy of 87.00%. This illustrates why XG Boost performs superior to DT. According to the SPSS, the developed XG Boost algorithmic classification model outperformed the DT in terms of accuracy comparison (87.00%). One of the most important factors of rainfall prediction is effectiveness. In a Research by [Rainfall Prediction Using Catboost Machine Learning Algorithm](https://ieeexplore.ieee.org/document/10276336/) [Vempaty Prashanthi](https://ieeexplore.ieee.org/author/37085565227);[Srinivas Kanakala](https://ieeexplore.ieee.org/author/37089254383);[Deepika Borgaonkar](https://ieeexplore.ieee.org/author/37090043935);[D. Suresh Babu](https://ieeexplore.ieee.org/author/37062109200)

[2023 International Conference on Network, Multimedia and Information Technology (NMITCON)](https://ieeexplore.ieee.org/xpl/conhome/10275783/proceeding)

the XGBoost learning accuracy was found to be 100% during teams learning and production phase, while its prediction accuracy was found to be 95.60% and 93.08%, respectively for the same phases. Similarly, the learning accuracy of the DNN was found to be 89.26% and 81.23%, while its prediction accuracy was found to be 80.50% and 77.36%, during the two phases. In a research [Journal of Physics: Conference Series](https://iopscience.iop.org/journal/1742-6596), [Volume 1324](https://iopscience.iop.org/volume/1742-6596/1324), [The Second International Conference on Physics, Mathematics and Statistics 22–24 May 2019, Hangzhou, China](https://iopscience.iop.org/issue/1742-6596/1324/1). And its prediction accuracy was found to be 93.40% and 91.80%, respectively for the same phases. The research from [(Zhang et al. 2020)](https://paperpile.com/c/1kWtcS/omyJ) proves that the XGBoost is better than Decision tree algorithm in predicting the effective and accurate rainfall.

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 DT, 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

A machine learning classifier called XG Boost makes use of decision trees to increase accuracy. The study demonstrates that, in comparison to DT, the XG Boost (XGB) algorithm appears to have a higher accuracy rate for rainfall prediction. It is discovered that when it comes to effective rainfall prediction, XGB outperforms DT by a significant margin. As a result, the study came to the conclusion that the XG Boost (XGB) algorithm yields higher accuracy (97%) than the Decision Tree (DT) algorithm (87.00%).

# 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

[Narejo, Sanam, Muhammad Moazzam Jawaid, Shahnawaz Talpur, Rizwan Baloch, and Eros](http://paperpile.com/b/1kWtcS/dooI)

[Yen, Meng-Hua, Ding-Wei Liu, Yi-Chia Hsin, Chu-En Lin, and Chii-Chang Chen. 2019.](http://paperpile.com/b/1kWtcS/Rtuw) [“Application of the Deep Learning for the Prediction of Rainfall in Southern Taiwan.”](http://paperpile.com/b/1kWtcS/Rtuw)

[Zhang, Pengcheng, Yangyang Jia, J. Gao, Wei Song, and H. Leung. 2020. “Short-Term](http://paperpile.com/b/1kWtcS/omyJ) [Rainfall Forecasting Using Multi-Layer Perceptron.”](http://paperpile.com/b/1kWtcS/omyJ) [http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8468083](http://ieeexplore.ieee.org/stamp/stamp.jsp?tp&arnumber=8468083)[.](http://paperpile.com/b/1kWtcS/omyJ)[Alexandridis, Antonios K., and Achilleas D. Zapranis. 2014. *Wavelet Neural Networks: With*](http://paperpile.com/b/1kWtcS/CIIe)[*Applications in Financial Engineering, Chaos, and Classification*. John Wiley & Sons.](http://paperpile.com/b/1kWtcS/CIIe)

[Gian Alessandro Pasero. 2021. “Multi-Step Rainfall Forecasting Using Deep Learning](http://paperpile.com/b/1kWtcS/dooI) [Approach.” *PeerJ Computer Science* 7. https://doi.org/](http://paperpile.com/b/1kWtcS/dooI)[10.7717/peerj-cs.514](http://dx.doi.org/10.7717/peerj-cs.514)[.](http://paperpile.com/b/1kWtcS/dooI)

[Rudrappa, Gujanatti. 2021. “Machine Learning Models Applied for Rainfall Prediction.”](http://paperpile.com/b/1kWtcS/g9Fr) [*Revista Gestão Inovação E Tecnologias*.](http://paperpile.com/b/1kWtcS/g9Fr) [https://doi.org/](http://paperpile.com/b/1kWtcS/g9Fr)[10.47059/revistageintec.v11i3.1926](http://dx.doi.org/10.47059/revistageintec.v11i3.1926)[.](http://paperpile.com/b/1kWtcS/g9Fr)

[Dattalo, Patrick. 2013. *Analysis of Multiple Dependent Variables*. Oxford University Press.](http://paperpile.com/b/1kWtcS/nPq0) [Ingsrisawang, L., S. Ingsriswang, S. Somchit, Prasert Aungsuratana, and](http://paperpile.com/b/1kWtcS/MiTV)

[WarawutKhantiyanan. 2008. “Machine Learning Techniques for Short-Term Rain](http://paperpile.com/b/1kWtcS/MiTV) [Forecasting System in the Northeastern Part of Thailand.”](http://paperpile.com/b/1kWtcS/MiTV) [https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-R](https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-Rain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d) [ain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d](https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-Rain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d)[.](http://paperpile.com/b/1kWtcS/MiTV)

[kerneler. 2019. “Starter: weatherAUS 536c1115-4.” Kaggle. November 29, 2019.](http://paperpile.com/b/1kWtcS/6ur6) <https://kaggle.com/kerneler/starter-weatheraus-536c1115-4>[.](http://paperpile.com/b/1kWtcS/6ur6)

[*Scientific Reports* 9 (1): 1–9.](http://paperpile.com/b/1kWtcS/Rtuw)

[Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi. 2019. “Predictability Assessment](http://paperpile.com/b/1kWtcS/67qd) [of Northeast Monsoon Rainfall in India Using Sea Surface Temperature Anomaly](http://paperpile.com/b/1kWtcS/67qd) [through Statistical and Machine Learning Techniques.” *Environmetrics*.](http://paperpile.com/b/1kWtcS/67qd) [https://doi.org/](http://paperpile.com/b/1kWtcS/67qd)[10.1002/env.2533](http://dx.doi.org/10.1002/env.2533)[.](http://paperpile.com/b/1kWtcS/67qd)

# TABLES AND FIGURES

**Table 1.** Accuracy values of XG Boost and Decision Tree algorithms obtained for each iteration while evaluating the dataset for various test sizes.

|  |  |
| --- | --- |
| **GROUP** | **ACCURACY** |
| XGB | 97 |
| XGB | 92 |
| XGB | 96 |
| XGB | 97 |
| XGB | 92 |
| XGB | 94 |
| XGB | 91 |
| XGB | 96 |
| XGB | 98 |
| XGB | 93 |
| DT | 87 |
| DT | 84 |
| DT | 86 |
| DT | 81 |
| DT | 82 |
| DT | 85 |
| DT | 83 |
| DT | 84 |
| DT | 79 |
| DT | 76 |

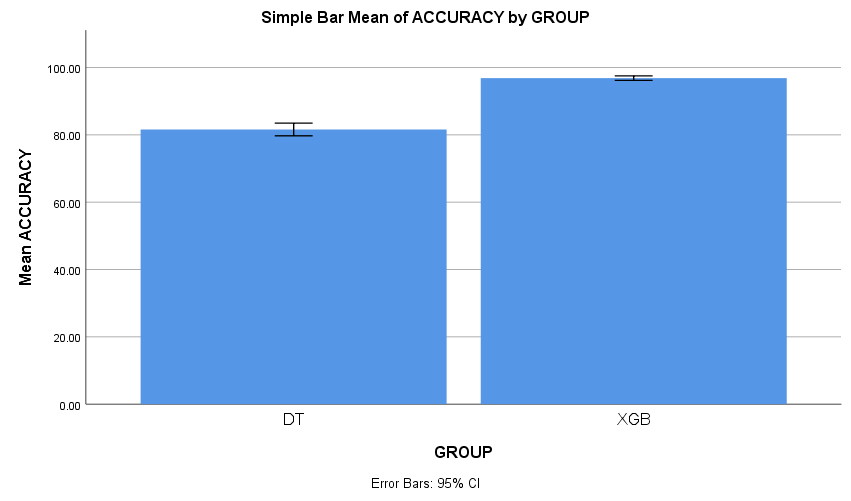
**Table 2**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Statistics** | | | | | |
|  | GROUP | N | Mean | Std. Deviation | Std. Error Mean |
| ACCURACY | XGB | 20 | 96.8500 | 1.42441 | .31851 |
| DT | 20 | 81.6000 | 4.01838 | .89854 |

**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 | 13.758 | .001 | 15.997 | 38 | .000 | 15.2500 | 0.95332 | 13.3201 | 17.1742 |
| Equal  Variances  Not assumed |  |  | 15.997 | 23.7 | .000 | 15.2500 | 0.95332 | 13.2830 | 17.17 |

GRAPH :



|  |
| --- |
|  |