Original Article

Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost over Random Forest Algorithm

P Jeyaprakaash[1], Sashi rekha K[2*]

[1]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. Pincode: 602105.
[2*]Project Guide, Corresponding Author, Department of Computer Science and Engineering , Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

Abstract

Aim: To increase the customer churn prediction model accuracy in the telecom industry using Adaboost over Random Forest Algorithm. **Materials and methods:** Adaboost algorithm and Random Forest algorithm with sample size (N=10) is executed with varying training and testing splits for predicting the accuracy for customer churn prediction and achieved the G power of 75% and threshold 0.000 and confidence interval 95%. The performance of the model is calculated based on their accuracy rate using the customer churn dataset. **Results and Discussion:** The customer churn prediction model has attained an accuracy of 90% using Novel Adaboost algorithm and 81% using Random Forest algorithm. There exists a statistical difference between Novel Adaboost and Random Forest (p=0.023) where p <0.005 . **Conclusion:** Prediction of customer churn using the Novel Adaboost algorithm results significantly greater than the Random Forest algorithm with improved accuracy.

Keywords: Customer churn, Novel Adaboost Algorithm, Random Forest algorithm, Machine Learning, Telecom Industry, Data Analytics

DOI:10.47750/pnr.2022.13.S04.178

INTRODUCTION

Customers are the valuable asset of the telecom company because they fetch more profit and benefits to the telecom industry. Customer churn is a rate of attrition or the percentage of customers that stops their subscription or service in a given period. In many telecom companies, customer churn is the biggest problem for their company. The solution to the problem is predicting the customers who are likely at risk of churning which was discussed in the paper (Geetha et al. 2020). The importance of customer churn prediction is accurately finding the future churn to ameliorate their business to gain more profit. The telecom industry is also able to improve the accuracy in the areas where customer service is lacking (Kassem et al. 2020). The applications of the customer churn prediction are to improve the accuracy and regulate the fields where the improvement is necessary for the telecom companies for instance Jio, Airtel, BSNL, VI, etc. The customer churn prediction is also useful in various industrial sectors such as banking, insurance, and mobile phone company.(Lariviere and Vandenpoel 2005) (Xia, Wang, and Jiang 2016) these two papers have convincing points about the applications of customer churn prediction.

Recently, a lot of researchers have done a variety of customer churn prediction in telecommunications using data analytics as it is one of the applications of data science and machine learning algorithms for customer churn prediction. About 482 articles had been published in ScienceDirect in the past 5 years and nearly 127 articles were published in IEEE Xplore. (Huang, Kechadi, and Buckley 2012) This work introduced another arrangement of elements for the client stir expectation in the media transmission, including the collected call subtleties, Henley division, account data, bill data, dial types, line data, installment data, grumble data, administration data, etc.(Huang, Kechadi, and Buckley 2012; Keramati et al. 2014) From this work, classification techniques using data from various sources of a dataset. Artificial Neural Network (ANN) significantly exceeds the other three algorithms, namely K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM). (Vafeiadis et al. 2015)Here, we discovered the effect of the utilization of boosting to the related classifiers utilizing the AdaBoost. (Lu et al. 2014) Rather than most agitated expectation models, our model takes into consideration an "Execution Zone" where clients with the most noteworthy stir affinity can be tended to for maintenance activities. From all the papers, we can arrive at a solution that the algorithm Ada-boost has the highest efficiency

of 84% when compared with other algorithms (Lalwani et al. 2021).

Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020). From the literature survey, it is inferred that the Adaboost algorithm has been widely used to predict the accuracy of the customer churn rate. Predicting the output as the improved accuracy to promote the telecommunication industry services in order to increase the customer rate which is mentioned in the paper (Gold 2020).So, the research focuses on improving the previous study accuracy with respect to the customer churn rate.

MATERIALS AND METHODS

The study setting of proposed analysis is done in Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (SIMATS). There are two groups of classification algorithms that have been used in this study. Group 1 is the Novel Adaboost Algorithm and Group 2 is the Random Forest algorithm. Using clinical analysis (Kane, Phar, and BCPS n.d.) 95% confidence and pretest power 80% have been attained by using 10 sample sizes for our study.

The dataset for training and testing of the model were collected from kaggle.com (Telecom Customer Churn), one of the well-known online communities for data scientists and machine learning practitioners to search and gather data to analyze using data analytics.The data sets consist of several data to train the system. Table 1 refers to datasets that are prepared by preprocessing and analysis.

Adaboost Algorithm

$$H(x) = sign\ (\sum_{t=1}^{T} \alpha_t h_t(x) \quad (1)$$

$$\alpha_t = 0.5 * ln\ ((1-E)/E) \quad (2)$$

Random Forest Algorithm

The Random forest Algorithm consists of different decision trees, each decision tree having the same nodes, but uses different data that leads to different leaves. It mergs all the decisions of the previous decision trees to obtain the result, which is given by the average of all these decision trees. It uses the divide and conquer approach. During this methodology, The training of the decision trees is done by choosing a random sample of attributes from the predictor set. Every tree matures by supporting the most extended parameters gift. The resulting call tree is made for the prediction of primarily supported weighted averages which has been mentioned in Table 3. This model has the power to handle thousands of input parameters without deleting. It also can handle the null values within the dataset for coaching the prophetic model.

The Random Forests performs based on classification data, Gini index value is expressed in equation 3 used to split the dataset.

$$\boldsymbol{Gind} = \mathbf{1} - \sum_{i=1}^{c} (\boldsymbol{pi})^{\mathbf{2}} \quad (3)$$

In equation(3), this formula uses the probability and class to determine which branch is expecting to occur. Here, Pi is the relative frequency of the class. C is the number of classes. The following formula equation (4) is used to calculate the accuracy value of the given models. Pseudocode is given in Table .3.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \quad (4)$$

Statistical Analysis

The algorithms are run in 64-bit OS, 8GB RAM Laptop and software specification includes Windows 10 along with Google collab software. The independent variable is Device Protection in the dataset and other 20 attributes such as Customer ID, gender, Monthly charges, churn etc are dependent variables for our study for customer churn.In order to compare the performances of the algorithm, an independent sample named T-Test was carried out. In SPSS, the dataset uses a sample size of 10 for the Adaboost and Random Forest algorithm. Group ID is given as grouping Accuracy and Loss is given as the testing variable. Group ID is given as 1 for Adaboost and 2 for Random Forest.

RESULTS

The input for the customer churn prediction using adaboost algorithm uses churn attribute data from the dataset and produces output with an accuracy value around 90% from Adaboost algorithm. The accuracy values of customer churn are shown in Table 4. The Random Forest Algorithm also takes input of the churn attribute from the dataset and produces output with an accuracy value around 81% this is shown in Table 5. Whereas the Table 6 shows the accuracy values of both Adaboost Classifier and Random Forest Classifier and the comparison between the classifiers.
From Table 6, It is observed that Adaboost algorithm proved with increased and higher accuracy of 90% than Random Forest algorithm (81%). From Table 7. It is observed that Adaboost algorithm has greater significance than Random Forest algorithm with the value of p = 0.023. From Fig. 2, the boxplot graph shows the G-graph which represents the comparison of Adaboost classifier and the Random Forest classifier and shows the resultant value of accuracy and loss of both the algorithms in the form of bar charts in the resultant graph in the end.From Fig. 1, the graph shows the rate of customers churned and those who are not.

DISCUSSION

In this study, it is observed that the accuracy of customer churn rate using Adaboost algorithm results in remarkable increase compared to the Random Forest algorithm. The Adaboost classifier shows a significant difference in the accuracy score, performance and speed when compared to the Random Forest classifier. In the Comparison between each algorithm's performance in terms of accuracy, f1 score and ROC AUC, Adaboost algorithm performs better overall performance than the Random Forest algorithm as discussed in Table 8.

(Idris, Iftikhar, and Rehman 2019) this work shows the accuracy prediction of customer churn using Adaboost and data used are orange cells which performed the prediction using balanced data and produced accuracy of 31% and 87% in first and second iterations respectively and those predictions was also mentioned in the word(Wu and Meng 2016).

(Xie et al. 2009) This work showed the accuracy improvement of customer churn using Random Forest Algorithm compared with Artificial Neural Network and Decision Tree with accuracies of 78% and 61% respectively but Random Forest has outplayed the other two algorithms.One of the work shows that Adaboost algorithm has boosted the performance of the prediction in the negative correlation of the customer churn (Rodan et al. 2015) this work clearly shows that Adaboost is better .

The limitations of our system are that the model is overfitted so prediction accuracy gets reduced while doing so. Adaboost appears to boost the performance by increasing the train speed.In the future, the accuracy of this model can be improved by data optimization in an unbalanced data in customer churn prediction so that the prediction can be boosted immersively.

CONCLUSION

In this research, customer churn prediction is performed using data analytics methods from the customer churn dataset gathered from Kaggle and those were described in Table .1 and it was performed using Adaboost and Random Forest algorithm.The accuracy of Adaboost algorithm is 91% whereas the accuracy value for Random Forest algorithm is about 81%. The accuracy of customer churn rate using Adaboost algorithm results in remarkable increase compared to the Random Forest algorithm.

DECLARATIONS

Conflict of interests
No conflict of interest in this manuscript.

Authors Contributions
Author PJ was involved in data collection, data analysis, and manuscript writing. Author SRK was involved in conceptualization, data validation, and critical review of manuscript.
Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Soft Square Solutions, Palavakkam, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
2. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
3. Geetha, V., A. Punitha, A. Nandhini, T. Nandhini, S. Shakila, and R. Sushmitha. 2020. "Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier." *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*. https://doi.org/10.1109/icscan49426.2020.9262288.
4. Gold, Carl S. 2020. *Fighting Churn with Data: The Science and Strategy of Customer Retention*. Manning Publications.
5. Huang, Bingquan, Mohand Tahar Kechadi, and Brian Buckley. 2012. "Customer Churn Prediction in Telecommunications." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2011.08.024.
6. Idris, Adnan, Aksam Iftikhar, and Zia ur Rehman. 2019. "Intelligent Churn Prediction for Telecom Using GP-AdaBoost Learning and PSO Undersampling." *Cluster Computing*. https://doi.org/10.1007/s10586-017-1154-3.
7. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
8. Kane, Sean P., Phar, and BCPS. n.d. "Sample Size Calculator." Accessed October 9, 2021. https://clincalc.com/stats/samplesize.aspx.
9. Kassem, Essam Abou el, Essam Abou el Kassem, Shereen Ali, Alaa Mostafa, and Fahad Kamal. 2020. "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention Based on User Generated Content." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2020.0110567.
10. Keramati, A., R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi. 2014. "Improved Churn Prediction in Telecommunication Industry Using Data Mining Techniques." *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2014.08.041.
11. Lalwani, Praveen, Manas Kumar Mishra, Jasroop Singh Chadha, and Pratyush Sethi. 2021. "Customer Churn Prediction System: A Machine Learning Approach." *Computing*. https://doi.org/10.1007/s00607-021-00908-y.
12. Lariviere, B., and D. Vandenpoel. 2005. "Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2005.04.043.
13. Lu, Ning, Hua Lin, Jie Lu, and Guangquan Zhang. 2014. "A Customer Churn Prediction Model in Telecom Industry Using Boosting." *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/tii.2012.2224355.
14. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
15. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
16. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
17. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
18. Rodan, Ali, Ayham Fayyoumi, Hossam Faris, Jamal Alsakran, and Omar Al-Kadi. 2015. "Negative Correlation Learning for Customer Churn Prediction: A Comparison Study." *The Scientific World Journal*. https://doi.org/10.1155/2015/473283.
19. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
20. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
21. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
22. Vafeiadis, T., K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas. 2015. "A Comparison of Machine Learning Techniques for Customer Churn Prediction." *Simulation Modelling Practice and Theory*. https://doi.org/10.1016/j.simpat.2015.03.003.
23. Wu, Xiaojun, and Sufang Meng. 2016. "E-Commerce Customer Churn Prediction Based on Improved SMOTE and AdaBoost." *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*. https://doi.org/10.1109/icsssm.2016.7538581.
24. Xia, Guo-En, Hui Wang, and Yilin Jiang. 2016. "Application of Customer Churn Prediction Based on Weighted Selective Ensembles." *2016 3rd International Conference on Systems and Informatics (ICSAI)*. https://doi.org/10.1109/icsai.2016.7811009.
25. Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying. 2009. "Customer Churn Prediction Using Improved Balanced Random Forests." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2008.06.121.

TABLES AND FIGURES

Table 1. Description of Customer Churn Dataset

Column	Values (For categorical variables)	Type
Tech Support	1 , 0	Numeric
Streaming TV	1, 0	Numeric
Streaming Movies	1 , 0	Numeric
Online security	1, 0	Numeric
Contract	1, 0	Numeric
Monthly Charges	1 , 0	Numeric
Total Charges	1 , 0	Numericl
Tenure	Multiple Months	Numeric
Internet Service	Multiple service	String
Online backup	1 , 0	Numericl
Phone Service	1 , 0	Numeric

Table 2. Adaboost Algorithm

Input - Telecom Customer churn dataset
1. Initialization / Selection of dataset
2. First, Create the First Base Learner
3. Calculate the Total Error (TE)
4. Calculate Performance of the Stump
5. Then, Update the Weights
6. Split the data as Training Data and Testing Data
Output - Customer Churn Predictions

Table 3. Algorithm for Random Forest classifier

Input - Telecom Customer churn dataset
1. Initialization /Selection of dataset
2. Firstly, a decision tree is made for each sample.
3. For each decision tree, a prediction result is obtained
4. There exists a poll for every decision tree
5. Finally, Select the most voted prediction result as the final result.
Output - Customer Churn Predictions

Table 4. Accuracy and Loss results for Adaboost classifier (N=5)

Iteration	**Accuracy (%)**	**Loss(%)**
1	90.04	9.96
2	90.05	9.95
3	90.21	9.79
4	90.40	9.60
5	90.11	9.89

Table 5. Accuracy and Loss results N=5 for Random forest

Iteration	Accuracy (%)	Loss(%)
1	80.80	19.20
2	80.74	19.26
3	80.32	19.68
4	80.55	19.45
5	80.45	19.55

Table 6. Group Statistics of T-Test with Mean, Std.Deviation and Std.Error Mean and Confidence = 95%

	Groups	N	Mean	Std Deviation	Std Error Mean
Accuracy	Random Forest	5	80.5720	.19942	.08919
	Adaboost	5	90.1620	.14923	.06674
Loss	Random Forest	5	19.4280	.19942	.08919
	Adaboost	5	9.8380	.14923	.06674

Table 7. Independent Sample T-Test is applied for the data set fixing confidence interval as 95% and level of significance as 0.05

		f	sig	t	df	Sig (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
Accuracy	Equal Variance assumed	0.676	0.435	-86.093	8	.0023	-9.5900	.11139	-9.84687	-9.33313

	Equal variance not assumed			-86.093	7.410	.0023	-9.5900	.11139	-9.85047	-9.32953
Loss	Equal Variance assumed	0.676	0.435	86.093	8	.0023	-9.5900	.11139	9.33313	9.84687
	Equal variance not assumed			86.093	7.410	.0023	-9.5900	.11139	9.32953	9.85047

Table 8. Comparison of the Adaboost and Random Forest algorithm with their accuracy

Classifiers	**Accuracy**	**Loss**
Adaboost	91%	9%
Random Forest	81%	19%

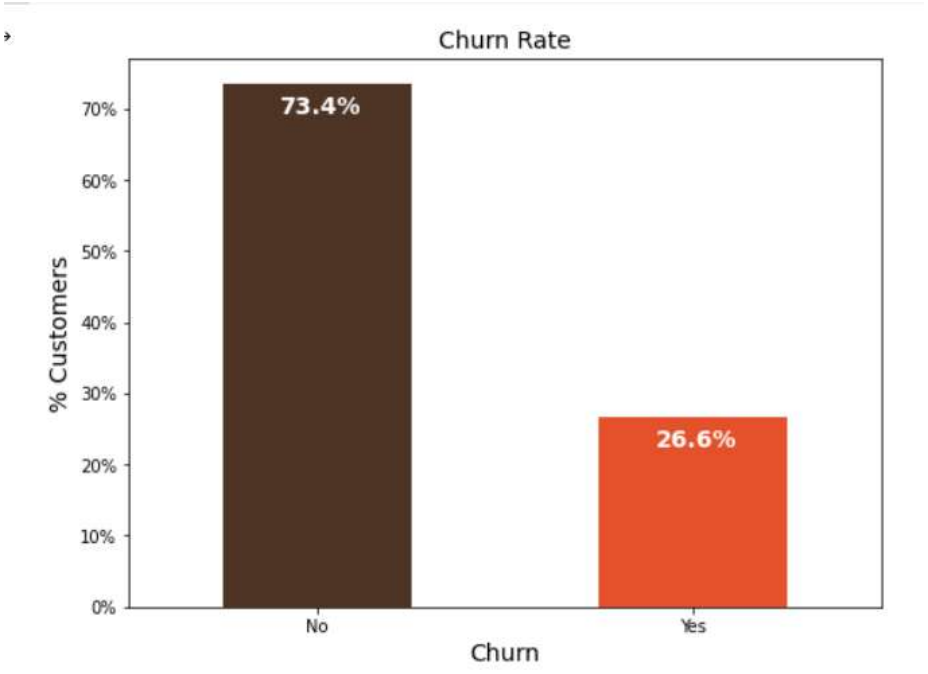

Fig. 1. Shows the difference between the rate of customers who were churn and those who were not churn.From the above bar chart we can see that only 26.6% of customers were churned from the dataset we have taken and 73.4% of customers were not churned .

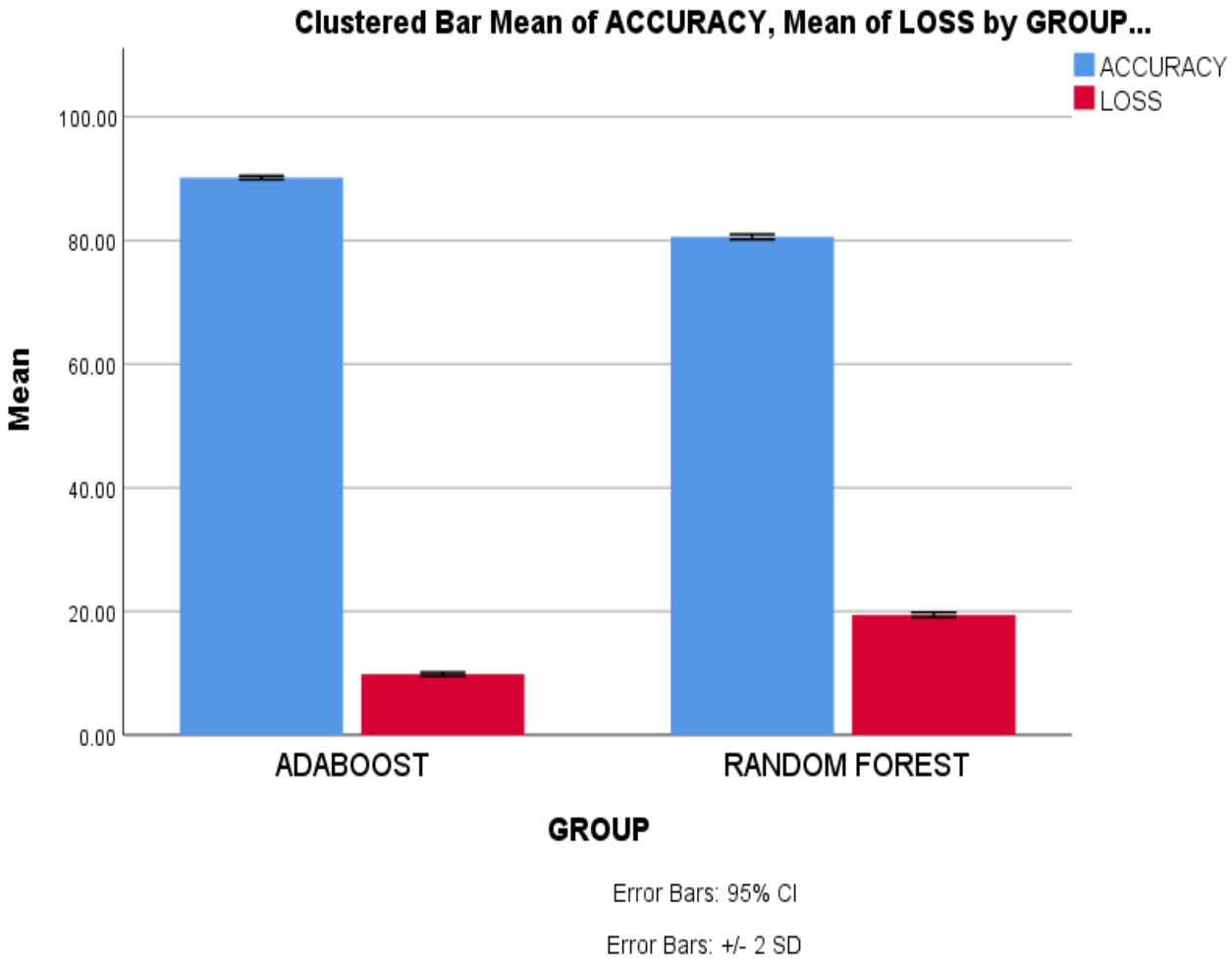

Fig. 2. Comparison of Adaboost algorithm and Random Forest in terms of mean accuracy. The mean accuracy of Adaboost is better than Random Forest and the standard deviation of Adaboost is slightly better than RF. X axis: Adaboost vs RF algorithm Y axis: Mean accuracy of prediction ± 2 SD.