Original Article

Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost over K Nearest Neighbor 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 enhance and predict the accuracy rate of customer churning in the telecommunication industry using Adaboost algorithm over K Nearest Neighbor algorithm. **Materials and methods:** Both Adaboost algorithm and K Nearest Neighbor algorithm with a sample size of (N=10) is executed with multiple training along with testing splits for predicting the accuracy of customer churning shows g-power as 75% and threshold value as 0.000 and confidence interval as 95%. The performance of these algorithms are calculated based on the rate of accuracy using customer churn dataset. **Results and Discussion:** The accuracy of predicting customer churn using Adaboost algorithm(90%) and K Nearest Neighbor algorithm(75%) is obtained. There was an analytical difference between Adaboost and K Nearest Neighbor ($p<0.005$). **Conclusion:** Prediction of customer churn using Adaboost algorithm seems to be comparatively better than the K-Nearest-Neighbor algorithm with improved accuracy.

Keywords: Customer Churn, Novel Adaboost Algorithm, K Nearest Neighbor algorithm, Machine Learning,Telecom Industry, Data Analytics

DOI:10.47750/pnr.2022.13.S04.180

INTRODUCTION

Customer attrition rate is known as the rate of consumers being transferred from a particular telecom service to choose another service or the percentage of subscribers to a service that discontinues their subscription to that service in a given period. In many telecom companies, customer churn is the biggest problem for their company because customers contribute nearly the entire profits and values of a company by paying for their services, so they are the asset of the company. The remedy to the drawback of consumers moving in the telecom industry who are at risk of churning which was referred to in the paper (Labhsetwar 2020). The relevance of customer churn prediction is accurately predicting future churn so that businesses can improve and make more money. The telecom company is also able to improve the accuracy in the areas where customer service is lacking (Hadaschik 2017) .Some of the examples in which customer churn has happened in some telecom industry such as Idea,Vodophone to telecom companies like Airtel, Jio, etc because of their poor service to customers,these are some applications of telecom customer churn. The customer churn prediction is also useful in various industrial sectors such as banking, insurance, and mobile phone companies. (Kassem et al. 2020) These two papers have ample points about the applications of customer churn prediction which provides evidence about the usage of churn prediction using machine learning techniques in the field of telecom industry as well as other fields like banking sector (Lu et al. 2014) .

Recently, a lot of researchers have done a variety of customer churn prediction in telecommunications using data analytics as it is the part of data science and ML algorithms for customer churn prediction. There were about 482 articles published in ScienceDirect in recent years and about 128 articles were published in IEEE Xplore journal.(Zhang et al. 2007)(Nawab, Sapuan Sapuan, and Shaker 2021)This work introduced another arrangement of elements for the client stir expectation in the media transmission, some of those are, the collected call types, data of account, bill ,line , installment data, grumble data, administration data, etc. (Huang, Kechadi, and Buckley

2012; Keramati et al. 2014) From this work,we gained our best classification techniques using data from various sources of a dataset. Artificial Neural Network (ANN) outplayed the other three algorithms used along with this . (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 (J, Rahul, and T. 2011).

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 found that the Adaboost ML 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. So, the research focuses on improving the previous study accuracy with respect to the customer churn rate.

Materials and Methods

The proposed work is conducted in the Image Processing Lab , Department of Computer science and Engineering at Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (SIMATS). Totally two groups are used for the study of classification algorithms. The 1st group is the Adaboost Algorithm and the latter group is the k Nearest Neighbor algorithm. Using clinical analysis (Kane, Phar, and BCPS n.d.) along with sample of sizes 10 have been carried out for our study, with a confidence of 95% and pretest power result as 80%

The dataset for our analysis work as input is collected from kaggle.com (Telecom Customer Churn) , which is one among the familiar 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 as referred in Table 1 and it prepared by preprocessing and analysis.

Adaboost Algorithm

The Novel AdaBoost algorithm, short for Adaptive Boosting, is an ensemble method used in Machine Learning to boost the performance. This algorithm uses a technique which iteratively learns from the previous weak classifiers and rectifies the mistakes. AdaBoost can be used with any classifiers, it will rectify the weak one and build a strong predictive model. It is a supervised machine learning approach. In supervised learning, boosting is mainly used in reducing the bias and variation. It generates a fixed number of decision trees and calculates the total error during the data training period. The record in the first model is improperly categorized and is given priority as the first decision tree/model is constructed which is mentioned in the Table 2. Pseudocode is given in Table 2. In equation (1) Here, ht(x) is the result of weak classifier t. X is the input of the classifier. αt is the weight given to the model expressed in equation 2 and those accuracy and loss percentage were mentioned in Table 4. This is the equation (1) and (2) used in Adaboost Classifier.

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

The output of weak classifier for input is defined as x $H(x)$ which is expressed in equation 1. α_t is weight assigned to the classifier as shown in equation 2. Pseudocode is given in Table 2.

$$\alpha_t = 0.5 * ln\left((1 - E) / E\right) \qquad (2)$$

K Nearest Neighbor

K-Nearest Neighbor comes under supervised learning technique. KNN performs an action on the dataset by storing the dataset during the time of classification . At the training phase, KNN stores the dataset then it classifies that data into a category that is much similar to the new data when it gets new data. In equation 3 Algorithms for KNN, continuous variables Euclidean distance function is used and it is referred in Table 3

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (3)$$

The two points in euclidean space are referred as p,q and the euclidean vectors as qi,pi expressed in equation 3. The following formula equation(4) is the hamming distance used for categorical variables.

$$D_H = \sum_{i=1}^{n} |p_i - q_i| \qquad (4)$$

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. The independent sample T-Test was performed to compare the performances of the algorithm. In SPSS, the dataset is developed using 10 samples for the Adaboost and K Nearest Neighbor Algorithm. Grouping Accuracy represents Group ID and Testing variable inferred in place of Loss. Group ID for Adaboost is 1 and for KNN , it is 2.

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 Novel Adaboost algorithm. The accuracy values of customer churn are shown in Table 4. The K Nearest Neighbor Algorithm also takes input of the churn attribute from the dataset and produces output with an accuracy value around 75% this is shown in Table 5 . Whereas Table 6, shows the accuracy values of both Novel Adaboost Classifier and KNN Classifier and the comparison between the classifiers. Table 8 shows the comparison between the accuracy and loss values of Adaboost Classifier and K Nearest Neighbor Classifier. It can be seen that Adamost has the best accuracy. Table 7 shows the values of independent Sample T-Test with confidence interval percentage of 95% and with significance of 0.05.
From Table 6, It is observed that the Novel Adaboost algorithm proved with greater significant accuracy (90%) than the KNN algorithm (75%). In Table 5, we can see that the Adaboost algorithm has better performance than the K Nearest Neighbor algorithm with the value of p = 0.000. The boxplot graph shown in the Fig. 2 indicates the G-graph which represents the comparison of Adaboost classifier and the KNN classifier and shows the output value of accuracy and the 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.

Discussion

In this study, it is observed that the Novel Adaboost algorithm seems to be significantly superior than the K Nearest Neighbor algorithm with improved accuracy.The Adaboost classifier shows a significant difference in the accuracy percentage, performance and speed when compared to the K Nearest Neighbor classifier. Comparison between each algorithm's resources such as accuracy, f1 score and ROC AUC. Novel Adaboost algorithm performs better overall performance than the K Nearest Neighbor algorithm as discussed in Table 4.
An amalgamation approach for constructing a binary classifier done by (Zhang et al. 2007).This work was handled with KNN classifier of one dimension and Logistic regression method. Finally, the result will be along with the application of customer churn for real world problems.In this work, blended KNN-LR classifier boost the performance of the LR.
This research work aims at comparing the performance of variants of decision trees using SPSS statistics software to work on complex data manipulation and analysis with simple commands. (Ruta, Nauck, and Azvine 2006) this paper exposed a new direction on customer attrition rate prediction and by using KNN, it highlighted the significance of analysis for events prediction
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 KNN algorithm.The accuracy of Adaboost algorithm is 90% whereas the accuracy value for K Nearest Neighbor algorithm is about 75%. The accuracy of customer churn rate using Adaboost algorithm appears to be comparatively superior than the K Nearest Neighbor 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. Hadaschik, Anne-Sophie. 2017. *Predicting Customer Churn in the Telecommunication Industry: An Analysis of Customer Switching Behavior Using the Example of a German Telecommunication Provider*.
4. 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.
5. J, Rahul, J. Rahul, and Usharani T. 2011. "Churn Prediction in Telecommunication Using Data Mining Technology." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2011.020204.
6. 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.
7. Kane, Sean P., Phar, and BCPS. n.d. "Sample Size Calculator." Accessed October 9, 2021. https://clincalc.com/stats/samplesize.aspx.
8. 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.
9. 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.
10. Labhsetwar, Shreyas Rajesh. 2020. "PREDICTIVE ANALYSIS OF CUSTOMER CHURN IN TELECOM INDUSTRY USING SUPERVISED LEARNING." *ICTACT Journal on Soft Computing*. https://doi.org/10.21917/ijsc.2020.0291.
11. 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.
12. 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.
13. Nawab, Yasir, S. M. Sapuan Sapuan, and Khubab Shaker. 2021. *Composite Solutions for Ballistics*. Woodhead Publishing.
14. 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.
15. 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.
16. 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.
17. Ruta, Dymitr, Detlef Nauck, and Ben Azvine. 2006. "K Nearest Sequence Method and Its Application to Churn Prediction." *Intelligent Data Engineering and Automated Learning – IDEAL 2006*. https://doi.org/10.1007/11875581_25.
18. 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.
19. 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.
20. 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.
21. 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.
22. Zhang, Yangming, Jiayin Qi, Huaying Shu, and Jiantong Cao. 2007. "A Hybrid KNN-LR Classifier and Its Application in Customer Churn Prediction." *2007 IEEE International Conference on Systems, Man and Cybernetics*. https://doi.org/10.1109/icsmc.2007.4414197.

Tables and Figures

Table 1. Customer Churn dataset description

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

Table 2. Algorithm for Adaboost classifier

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. Pseudocode for K Nearest Neighbor

Input- Telecom Customer churn dataset
1. Initialize all the input elements
2. Place the less weighted element of the dataset at the beginning of the tree.
3. Split the training model into different group. Each group should contain data with a same value for an attribute.
4. Repeat step 1 and step 2 on each group until leaf nodes in all of the branches of the tree are found.
5. Customer churn predicted result will the output

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

Iteration	**Accuracy (%)**	**Loss(%)**
1	90.48	9.52
2	90.22	9.78
3	90.11	9.89

4	90.38	9.62
5	90.55	9.45

Table 5. K Nearest Neighbor algorithm Accuracy and Loss results for N=5

Iteration	**Accuracy (%)**	**Loss(%)**
1	75.4	24.6
2	75.2	24.8
3	75.9	24.1
4	75.10	24.9
5	75.25	24.75

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

	Groups	**N**	**Mean**	**Std Deviation**	**Std Error Mean**
Accuracy	KNN	5	75.2200	.17378	.07772
	Adaboost	5	90.3660	.56265	.25163
Loss	KNN	5	24.7800	.17378	.07772
	Adaboost	5	13.6340	5.95622	2.66370

Table 7. Independent Sample T-Test is applied for the data set fixing confidence interval as 95% and significance as p=0.00 (p< 0.05)(2-tailed).

		f	sig	t	df	sig(2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
Accuracy	Equal Variance	3.492	0.100	-57.512	8	.000	-15.14600	.26336	-15.75330	-14.53870

	assumed									
	Equal variance not assumed			-57.512	4.756	.000	-15.14600	.26336	-15.83354	-14.45846
Loss	Equal Variance assumed	84.947	0.000	4.183	8	.000	11.14600	2.66484	5.00087	17.29113
	Equal variance not assumed			4.183	4.756	.000	11.14600	2.66484	3.75218	18.53982

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

Classifiers	**Accuracy**
Adaboost	90%
K Nearest Neighbor	75%

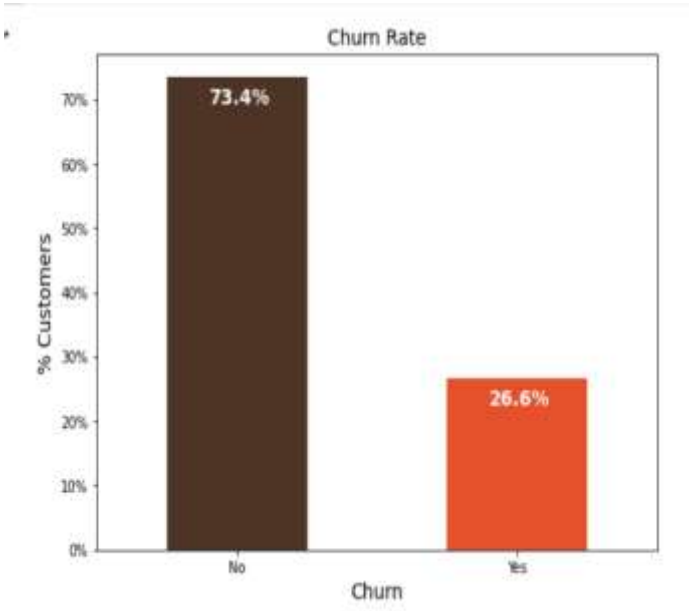

Fig. 1. Shows the difference among the rate of customers who were churn and those who were not churn. From the above bar chart we can see that only 26.6% of customers were churned from the dataset we have taken and 73.4% of customers were not churned.

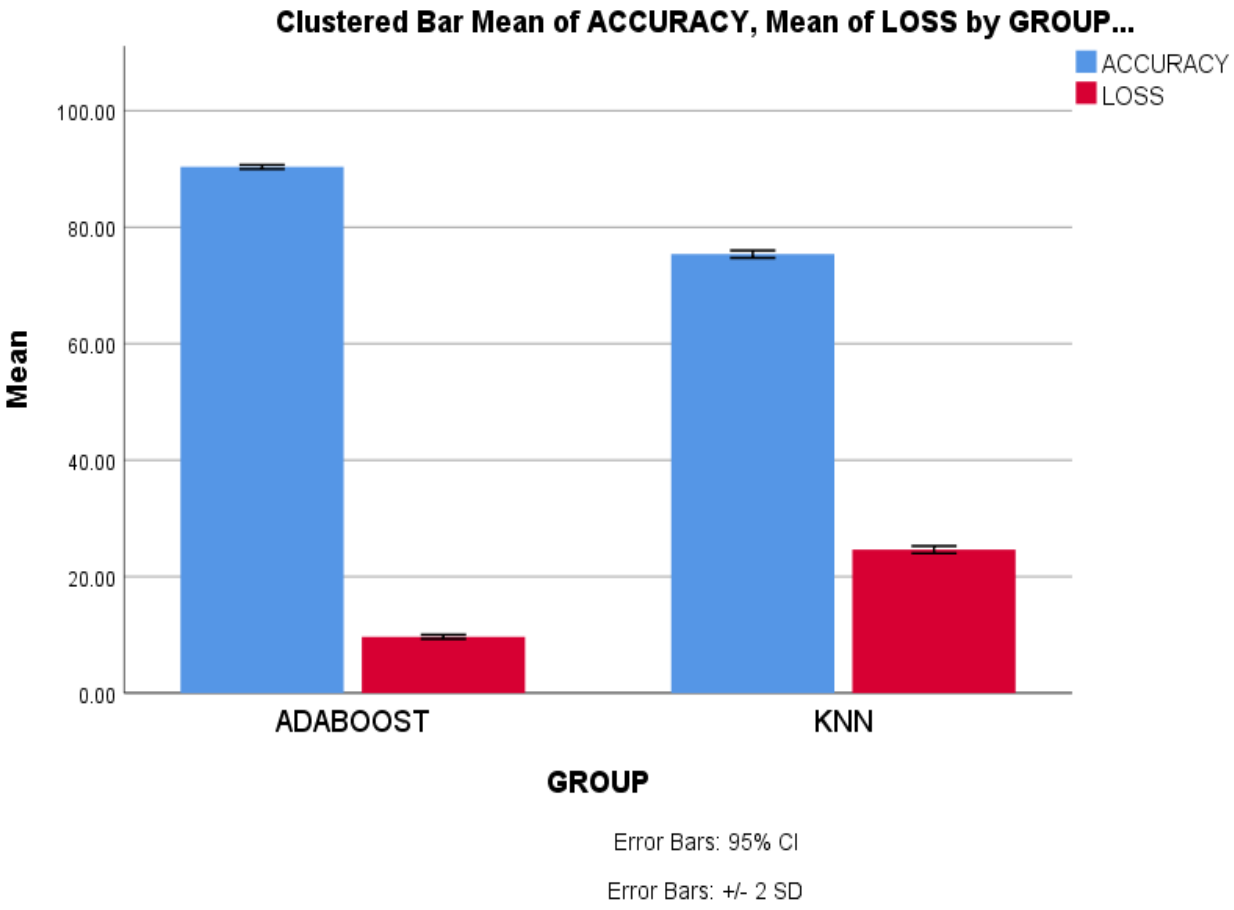

Fig. 2. Comparison of Adaboost algorithm and KNN in case of average accuracy. The mean accuracy of Adaboost is greater than KNN. X axis:(Groups) Adaboost algorithm vs KNN algorithm Y axis: Mean accuracy with ± 2 SD.