Study comparing classification algorithms for loan approval predictability (Logistic Regression, XG boost, Random Forest, Decision Tree)

K. Reddy Prathap[1], Dr.Bhavani R[2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

[2]*Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

ABSTRACT

Aim: The primary objective of this study is to perform novel loan predictions for approval using different machine learning algorithms and obtain better accuracy. **Materials and Methods:** Machine learning algorithms are applied on Loan Prediction Problem dataset that consists of train file and testing file. Four groups were used namely Logistic regression, Decision tree, Random Forest and XGBoost. The sample size was measured as 35 per group using Gpower of 80%. **Results:** The accuracy is maximum when loan approval prediction is done using Logistic Regression (83.24%) when compared with Random Forest, XGBoost and Decision Tree and there exists a statistically significant difference of ($p<0.05$) between the classifiers. The minimum accuracy is obtained in the Decision tree (70.34%). **Conclusion:** The study shows that Logistic Regression Algorithm exhibits better accuracy than other algorithms such as Decision Tree, Random Forest and XGBoost in Loan approval prediction.

Keywords: Decision Tree, Logistic Regression, Machine Learning, Novel loan prediction, Random Forest, XGBoost

INTRODUCTION

Nowadays money has become the center of everything, so bank loans are used to meet the needs of people. Loan application rates have increased at a very rapid rate over the past years (Ramachandra et al. 2021). Risks are always involved in approving loans. Bank officers are very aware of the payment of loans of customers (Zhu et al. 2019). Loan approval decisions are not always correct . Many precautions should be taken while analyzing loan applicant data (Behera et al. 2021). It is necessary to automate the process of loan approval which helps to reduce the damages and risks for banks.Machine learning algorithms can be used to perform novel loan predictions (Sheikh, Goel, and Kumar 2020). This can be used to know if a loan applicant's application can be approved or not (Gupta et al. 2020). It can be used in various banking Sector applications and online credit applications and can be used to evaluate the credit risk based on the applicant's information.

Most frequently cited works that are similar to this work have been investigated.Several research papers are

available on IEEE Xplore and ScienceDirect . From IEEE Xplore 458 articles are identified and 481 articles on ScienceDirect that relate relevant research. (Alaradi and Hilal 2020) utilized Random forest and the Decision Tree to compare accuracy for loan approval prediction. Decision tree out performs Random Forest in terms of accuracy. Applicant's personal details are considered such as age, credit risk, and their objectives are used in predicting novel loan approval (Dosalwar et al. 2021). Nature and background of the applicant is also verified and considered for prediction (Madaan et al. 2021). For prediction of credit risk the Random forest has high accuracy. The performance of algorithms were analyzed using metrics like f1 score etc. However , the metrics were meant for individual classes, not the whole data (Wu, Huang, and Duan 2019). The predictive model , Decision tree and SVM was fully used and well utilized to classify loan approval (Nafees and Rehman 2021). (Qureshi et al. 2021) has performed the prediction of missing clinical appointments using machine learning, XG boost shows best accuracy.(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)

The research gap identified from the literature survey is that classification models adopting Decision Trees are unstable.While using decision trees a minute change in the data can affect the structure of the optimal decision tree in a big way ,which is relatively inaccurate. The existing approaches have poor accuracy. There are other predictors that perform better with similar data such as Logistic Regression, Random Forest and XGBoost . The primary objective of this study is to compare these machine learning algorithms and find an efficient method for novel loan prediction .

MATERIALS AND METHODS

The study setting for the proposed work was conducted at Web Ontology Lab in Saveetha School of Engineering using Google Colab, and the required hardware and software configurations are intel i5 7th gen processor, 250 GB HDD, 8GB Ram, and windows 10 OS and Matlab. Machine algorithms such as Decision Tree , Logistic Regression , Random Forest and gradient boosting were applied on Loan prediction dataset. Five iterations per each group were performed to achieve better accuracy. The Study uses a loan prediction problem dataset downloaded from kaggle. The proposed methodology compares Logistic Regression, Decision Tree, Random Forest and XGBoost for implementing novel loan prediction. The sample size was calculated as 35 in each group using G Power (Vaidya 2017) .

Logistic Regression

A strategy for defining data and explaining the relationship between one or more independent factors and one or more dependent binary variables is logistic regression. In Logistic Regression when the dependent variable is categorical, then like all other regression analyses, those variables can be used as a prediction tool. The logistic regression statistical model is a widely used statistical model for binary classification, or predictions of the type this or that, yes or no, A or B, and so forth. It's one of the most used binary classification machine learning

approaches. The pseudocode for the Logistic Regression is given below.

Input: Loan prediction problem Dataset.
Output: Accuracy.

1. Pre-processing of data .
2. Fitting the Training set with Logistic Regression.
3. By using the above Training set, predict the test result.
4. The accuracy of the result is to be tested .
5. Finally the test result set is visualized .

Random Forest

This is a supervised learning technique where a multiple number of decision trees are created at the time of training and outputs. All these trees are combined to form a single tree and it can be used for parameters such as to evaluate the performance of a model and to find accuracy of that model. This type of application has several applications and can be used in classification and regression types of analyses. Instead of working on a single decision tree this random forest algorithm gathers data from all the decision trees independently and predicts on majority number votes of prediction which can be used to predict the end output. The accuracy will be greater if the number of decision trees is more. The advantages of proposed algorithm over other machine learning algorithms are as follows:

- This is based on ensemble learning , so complex or large databases can run very effectively.
- The errors can be fixed easily and this algorithm can prevent the overfitting problem ;
- Even with a large amount of missing data we can maintain better accuracy and support an effective way for managing the missing data.

The Pseudocode of proposed algorithm :

Input: Loan prediction problem Dataset.

Output: Accuracy.

These are the steps followed while implementing the random forest algorithm :

1. Pre-processing of the data.
2. The training set is fitted using Random forest .
3. Testing results are used to predict the credit risk .
4. The testing results are used to find accuracy.
5. Visualizing the test result and accuracy.

XG Boost

XGBoost is an ensemble learning technique which combines various models like classifiers and functions which are strategically generated for solving intelligent problems.This method is generally used to improve a certain process of classification or prediction and the execution speed is also high .This XGBoosting is an extension of gradient boosting which is also a method in ensemble learning which is different from bagging classifier. Generally these bagging and boating are commonly used and they can be applied to different kinds of models which are statistical and decision trees that have been the most popular. Gradient boosting is commonly implemented with XGBoost.

Pseudocode for XGBoost

Input: Loan prediction problem Dataset.

Output: Accuracy.

These are the steps followed while implementing XGBoost :

1. Install the XGBoost in python.
2. Upload the loan prediction dataset
3. Train the dataset using the above model.
4. Evaluate the model by using the above predictions.
5. Combine the above prediction for final output .

Decision Tree

In order for the decision tree's based algorithm to work, all attributes or features must be discretized.The feature which gathers the most information is to be used in this algorithm. To express the knowledge provided in a decision tree, IF-THEN rules can be employed. (Kirandeep et al. 2018) These decision trees can mimic human decision making by creating a tree-like structure, which is presented graphically with simple interpretations which works for both nominal and input and output variables which are continuous. By identifying the instances , the algorithm creates a tree-like structure. It's a strong tool for facilitating deciding sequential decision problems.The properties of the decision tree are

- These decisions are perceptibly interpreted and easily understood, they can mimic human decision making very simply.
- It can handle both numeric and categorical data and makes no assumptions about the distribution of attributes / predictors.
- These can be further added using bagging , boosting and Random forest classifiers.

Input: Loan prediction problem Dataset.

Output: Accuracy.

1. Importing the dataset into the model.
2. Few random features are selected from the above dataset.
3. Decision Tree Classifier is generated as a parameter.
4. Analyze the variables which are independent and dependent variables in the dataset.
5. It predicts the final output.
6. Most predicted results were selected as final output.

Statistical Analysis

For statistical analysis research the Statistical Package software (SPSS) is used. Group statistics and independent sample t-tests and Anova testing are performed. The experimental results and the graph were built for four groups with two parameters which we have used for the classification of algorithms under study. The independent variables are ApplicantIncome, CoapplicantIncome, LoanAmount, LoanAmountTerm, Gender, Status, Employment, CreditHistory, LoanStatus and the dependent variables are Dependents, Education, PropertyArea (Vaidya 2017).

RESULTS

The proposed algorithms (Logistic Regression, Random Forest, XGBoost) and the existing algorithm (Decision Tree) were compared with SPSS statistical software at same time.The sample sets for each algorithm we iterated for five times and executed with different lengths and the accuracy of the algorithms are noted in Table 1. In Table 2 we can find statistical analysis such as Mean, Standard Deviation, and Standard Error and Accuracy of the algorithms that were performed. The significant difference

between the algorithms is also less when compared with other algorithms. Logistic Regression is having the higher Accuracy (83.24%) and the decision tree has the lowest Accuracy (70.34%). In Table 3 One-way ANOVA testing of metrics was performed. It compares the study groups of algorithms . In Table 4 Multiple comparisons of the algorithms are performed with a significance value of p = 0.05.The Figure 1 shows a bar chart which compares the accuracies of all the algorithms with mean accuracy on y-axis and and the proposed algorithms on x-axis. It also produces error bars for the proposed algorithms. The standard errors appear less in the XGBoost algorithm. Logistic Regression produces higher accuracy with better performance.

DISCUSSION

The analysis and prediction is conducted between four groups of machine algorithms by varying sample sizes. By performing the analysis it is found that the proposed Logistic regression algorithm performed better in terms of Loan approval prediction by achieving better accuracy (83.24%) when compared to the other machine learning algorithms.

There are similar findings which proves Logistic Regression is better for prediction and has better accuracy. The loan approval prediction may not depend on a single attribute for loan prediction but may depend on multiple attributes for decision making (G. 2016). The Logistic regression Accuracy is 83.24% and the work implemented by (Sheikh, Goel, and Kumar 2020) has attained 80.94% of accuracy. (Yamashita, Noe, and Bailer 2012) In this paper they proposed a logistic regression method which is tree-based, it is used for identifying fall risk factors and the interaction effects that are possible because of the risk factors and the logistic regression is proven to be best.(Dosalwar et al. 2021)In this paper the analysis of medical research using machine learning techniques is performed and Logistic Regression proved to be best among the algorithms. There are opposite findings, (Goetz 2011) implemented new strategies for thinking about health and showed Decision Tree as best over Logistic Regression. (Kirandeep et al. 2018) proposed placement prediction and proved Decision Tree is best compared to Logistic Regression.

The key constraint of Logistic Regression is to assume the linearity in between the variables which may be dependent or independent. The non linear issues cannot be tackled by the surface of linear decision. Linearly separable data is uncommon in real world circumstances. In the future by comparing all the variables from the dataset the model can predict the target variable with less difficulty and can be used for the test data, and this target variable can be used to predict the accuracy.

CONCLUSION

By performing the analysis of machine learning algorithms such as Logistic Regression ,Random forest, XGBoost we can say logistic regression is best . From Fig 1 it proves that Logistic Regression has better accuracy than the other algorithms which is 83.24% and that it is very consistent. All the other algorithms performed better ,but Logistic Regression outperformed the other algorithms with higher accuracy.

DECLARATIONS

Conflicts of Interests

No conflicts of interest in this manuscript.

Author Contributions

Author KRP was involved in data collections, data analysis, algorithm framing, implementation and manuscript writing. Author RB was involved in designing the workflow, guidance, and reviewing the manuscript.

Acknowledgements

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. Qbec Infosol.
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Alaradi, Mohamed, and Sawsan Hilal. 2020. "Tree-Based Methods for Loan Approval." In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. IEEE. https://doi.org/10.1109/icdabi51230.2020.9325614.
2. Behera, Aiswarya Priyadarsini, Siddharth Swarup Rautaray, Manjusha Pandey, and Mahendra Kumar Gourisaria. 2021. "Predictive Loan Approval Model Using Logistic Regression." *Smart Computing Techniques and Applications*. https://doi.org/10.1007/978-981-16-0878-0_69.
3. 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.
4. 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.
5. Dosalwar, Sharayu, Ketki Kinkar, Rahul Sannat, and Dr Nitin Pise. 2021. "Analysis of Loan Availability Using Machine Learning Techniques." *International Journal of Advanced Research in Science, Communication and Technology*, September, 15–20.
6. Goetz, Thomas. 2011. *The Decision Tree: How to Make Better Choices and Take Control of Your Health*. Rodale.
7. G., Sudhamathy. 2016. "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R." *International Journal of Engineering and Technology* 8 (5): 1954–66.
8. Gupta, Anshika, Vinay Pant, Sudhanshu Kumar, and Pravesh Kumar Bansal. 2020. "Bank Loan Prediction System Using Machine Learning." In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE. https://doi.org/10.1109/smart50582.2020.9336801.
9. 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.

10. Kirandeep, Kirandeep Kirandeep, Neena Madan, G.N.D.U, Regional Campus, Jalandhar, Punjab, and India. 2018. “Deployment of ID3 Decision Tree Algorithm for Placement Prediction.” *International Journal of Trend in Scientific Research and Development*. https://doi.org/10.31142/ijtsrd11073.
11. Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. 2021. “Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study.” *IOP Conference Series. Materials Science and Engineering* 1022 (January): 012042.
12. Nafees, Sadeem Ahmad, and Faisal Asad Ur Rehman. 2021. “Machine Learning Based Approval Prediction for Enhancement Reports.” *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. https://doi.org/10.1109/ibcast51254.2021.9393180.
13. 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.
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. Qureshi, Zeeshan, Ayesha Maqbool, Alina Mirza, Muhammad Zubair Iqbal, Farkhanda Afzal, Deborah Dormah Kanubala, Tauseef Rana, Mir Yasir Umair, Abdul Wakeel, and Said Khalid Shah. 2021. “Efficient Prediction of Missed Clinical Appointment Using Machine Learning.” *Computational and Mathematical Methods in Medicine* 2021 (October): 2376391.
18. Ramachandra, H. V., G. Balaraju, R. Divyashree, and Harish Patil. 2021. “Design and Simulation of Loan Approval Prediction Model Using AWS Platform.” In *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE. https://doi.org/10.1109/esci50559.2021

.9397049.

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. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. 2020. "An Approach for Prediction of Loan Approval Using Machine Learning Algorithm." In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. https://doi.org/10.1109/icesc48915.2020.9155614.
21. 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.
22. 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.
23. Vaidya, Ashlesha. 2017. "Predictive and Probabilistic Approach Using Logistic Regression: Application to Prediction of Loan Approval." *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt.2017.8203946.
24. Wu, Mingli, Yafei Huang, and Jianyong Duan. 2019. "Investigations on Classification Methods for Loan Application Based on Machine Learning." *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*. https://doi.org/10.1109/icmlc48188.2019.8949252.
25. Yamashita, Takashi, Douglas A. Noe, and A. John Bailer. 2012. "Risk Factors of Falls in Community-Dwelling Older Adults: Logistic Regression Tree Analysis." *The Gerontologist* 52 (6): 822–32.
26. Zhu, Lin, Dafeng Qiu, Daji Ergu, Cai Ying, and Kuiyi Liu. 2019. "A Study on Predicting Loan Default Based on the Random Forest Algorithm." *Procedia Computer Science* 162: 503–13.

TABLES AND FIGURES

Table 1. Comparison for Accuracy achieved during the evaluation of LR, DF, RF and XG for novel loan prediction with different iterations.

Samples	Accuracy (%)			
	LR	**RF**	**XG**	**DF**
1	83.24	78.27	82.16	70.01

2	85.76	75.23	81.70	69.91
3	80.24	84.33	77.80	75.60
4	80.45	75.77	81.90	64.70
5	86.55	77.00	80.92	71.49

Table 2. Statistical Analysis of Mean, Standard Deviation, and Standard Error and Accuracy of LR, RT, XG, and DT algorithms. . Logistic Regression had the higher Accuracy (83.24%). Decision Tree had the lowest Accuracy (70.34%).The Standard error is also less in XGBoost when compared with other algorithms.

	Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
						Upper	**Lower**		
Accuracy	LR	5	83.2480	2.91934	1.30557	79.623	86.8728	80.24	86.55
	RF	5	78.1200	3.66455	1.63883	73.569	82.6701	75.23	84.33
	XG	5	80.8960	1.79150	.80118	78.6716	83.1204	77.80	82.16
	DT	5	70.3420	3.80912	1.74776	65.4894	75.1946	64.70	75.60
	Total	20	78.1515	5.77753	1.29189	75.5575	80.8555	64.70	86.55

Table 3. One Way ANOVA Test for the Accuracy metric between groups and within groups

Accuracy	**Sum of squares**	**df**	**Mean square**	**F**	**sig.**
Between Groups	472.479	3	157.493	15.580	.000
Within Groups	161.737	16	10.109		
Total	634.217	19			

Table 4. Multiple comparisons are performed between one algorithm with other remaining algorithms and the significance level is compared with value p = 0.05 and with a 95 % confidence interval with Post Hoc Tests .

Dependent Variable	**(I) Algorithm**	**(J) Algorithm**	**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	**Upper Bound**

Accuracy	LR	RF XG DT	5.12800 2.35200 12.90600	2.01083 2.01083 2.01083	.128 1.000 .000	-9212 -3.6972 6.8568	11.1772 8.4012 18.9552
	RF	LR XG DT	-5.12800 -2.77600 7.77800	2.01083 2.01083 2.01083	.128 1.000 .008	-11.1772 -8.8252 1.7288	.9212 3.2732 13.8272
	XG	LR RF DT	-2.35200 2.77600 10.55400	2.01083 2.01083 2.01083	1.000 1.000 .000	-8.4012 -3.2732 4.5048	3.6972 8.8252 16.6032
	DT	LR RF XG	-12.90600 -7.77800 -1055400	2.01083 2.01083 2.01083	.000 .008 .000	-18.9552 -13.8272 -16.6032	-6.8563 -1.7288 -4.5048

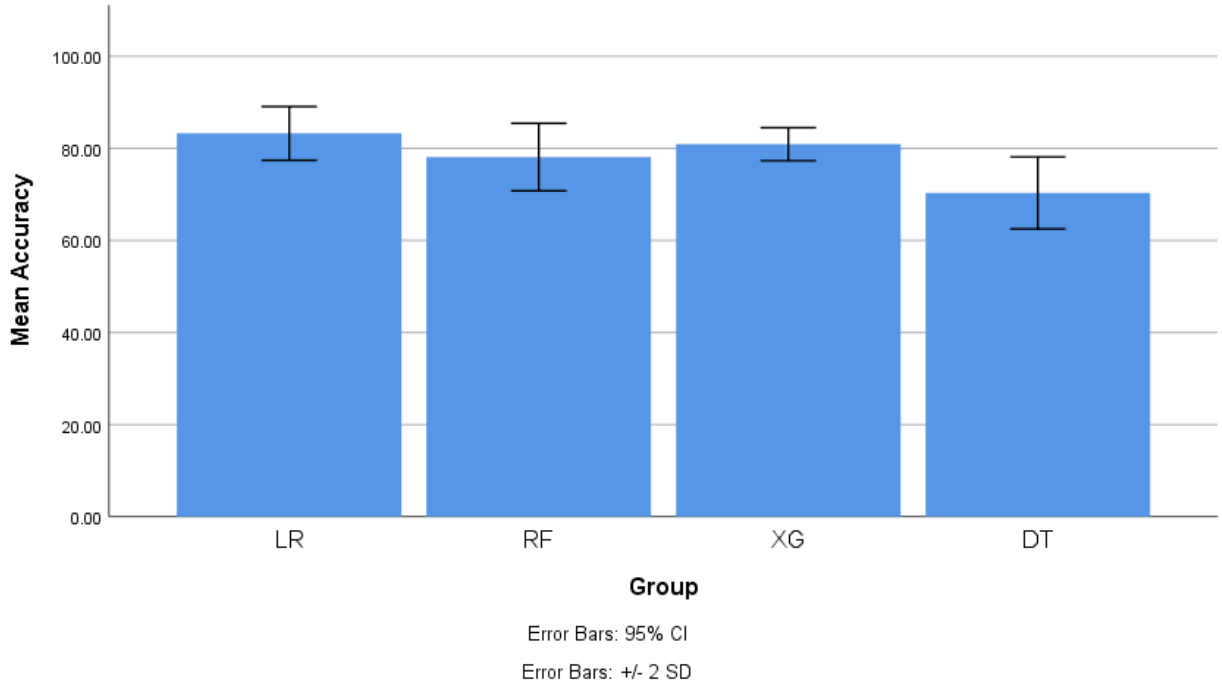

Fig. 1. A bar Graph which compares the mean accuracy of Logistic Regression, Random Forest, XGBoost and Decision Tree algorithms with error bars. The standard errors appear to be less in XG Boost algorithm. Logistic Regression has better performance and higher accuracy. X-axis: Logistic Regression vs Random Forest vs XGBoost vs Decision Tree. Y-axis: Mean Accuracy of detection +/-2 SD.