Comparison of the Support Vector Classifier algorithm with the Decision Tree algorithm for Credit Card Fraud Detection with the Goal of Improving Accuracy

S. Tharun Sainatha Reddy[1], Sriramya P[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: Evaluating the work's accuracy is its main goal in predicting the fraud using Decision Tree (DT) and Support Vector Classifier (SVC) algorithms. **Materials and Methods:** On a credit card, a novel decision tree classifier is used for 50646 records in the csv dataset.A system for detecting credit card fraud that compares the novel tree-specific decision tree algorithm (Group 1) and Support vector classifier (Group 2) has been conceived of and created. The number of samples was calculated to be 21 for each of the two groups. Data on phishing attacks and the G-power value of 80% were gathered from various web sources using a threshold of 0.05% and a confidence interval of 95% along with mean and standard deviation. The classifier's accuracy was assessed and noted. **Results:** The novel decision tree classifier produces 94.86% accuracy with the significant value of two tailed tests is 0.000 ($p<0.05$) in detection of credit card fraud dataset whereas the another machine learning algorithm Support vector classifier predicts the same at the rate of 98.59% accuracy. **Conclusion:** This research proves that Support vector classifier algorithm produces better accuracy than novel decision tree classifier algorithm in detection of credit card fraud.

Keywords: Novel Decision tree, Support Vector Classifier, Fraud Detection, Credit Card, Accuracy, Machine Learning Algorithm.

INTRODUCTION

Fraud detection refers to a set of measures for preventing money or property from being obtained through deception. In our daily lives, we are subjected to a variety of types of deceptions. Credit card fraud is the most common and widespread of all fraud (Dal Pozzolo et al. 2018). Credit card fraud is an easy and enticing target. Many online sites, including e-commerce, have expanded their payment alternatives, increasing the risk of online fraud. Fraudulent actions employing credit card payment technology are on the rise as credit cards become the most frequent form of payment, particularly in the online sector (Singh and Davidson 2015). As technology has advanced and new e-service payment methods, such as e-commerce and mobile payments, have arisen, credit card transactions have become more popular. Because cashless transactions are so common, fraudsters are more likely to perpetrate fraud and alter

their strategies frequently to evade detection(Rapp 1991). In order to prevent customers from being charged for items they did not buy, identification of fraudulent credit card transactions is essential for credit card companies. Such problems can be resolved with data science, which must be used in conjunction with machine learning(Shimp et al. 2016). The proposed work aims in the efficient detection of credit card fraud, which benefits the nation's overall development.

There are 45 research articles published on credit card fraud detection in IEEE xplore and 80 articles on google scholar and 20 articles were found in sciencedirect.(Padvekar et al. 2016) provides a new machine learning-based detection technique for long-term chaotic time series based on neural networks. The easiest and simplest presentation of advanced information in detection models is difficult to defeat in various circumstances; the neural network technique provides an alternative way to include prior knowledge into detection models in situations where background information isn't available or isn't useful. Finally, it appears that the situation has returned to normal. (Kumar et al. 2017) proposes a machine learning-based system for collecting data and categorizing it into different categories based on the risk estimate prior to deployment. The data is examined, and the final result is compared to the calculated risk. Finally, the estimated and final risks are compared to determine which algorithm produces the most accurate findings. (C. and C. 2020) presents a novel tree-specific random forest algorithm. Nationality, time, place, person name, and personal characteristics of the individual are some of the elements evaluated in the detection of credit fraud. For the analysis, they used data from IDFC First Bank. The java code, which makes use of scala modules, was written and developed for Bayesian Enhanced Approach study (BEA). (Saia and Carta 2017)Presents an algorithm for credit card fraud detection using logistic regression. It produces an accuracy of 98.87% and stands out of all other featured algorithms, including linear regression and random forest algorithm with the same classifier on the similar dataset.Our team has extensive knowledge and research experience that has translate into high quality publications(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

The research gap identified from the survey is that there are many methods already proposed for the detection of credit card fraud but all the algorithms produce less accuracy rate. Several works have demonstrated that the performance of decision trees is poor and provide lesser accuracy in detection of credit card fraud. A study by (Lucas 2019) to detect credit card fraud analyzes the accuracy of various machine learning classification systems. It is required to compare the algorithms that provide superior accuracy for identifying the best algorithm among all the algorithms. Hence, the work aims at comparing the accuracy of novel decision trees and support vector classifier

algorithms for detection of credit card fraud.

MATERIALS AND METHODS

The study was conducted in the Artificial Intelligence Lab at the SIMATS Saveetha School of Engineering's Department of Computer Science and Engineering (Saveetha Institute of Medical and Technical Sciences). The hardware configuration needed for this work was an Intel i5 processor, 1TB HDD, and 8GB of RAM, and the software configuration needed was Windows OS. Kaggle and SPSS were used to implement this work. the Kaggle website, where the dataset was downloaded(Jurgovsky 2019). There are 31 columns and 50646 rows in the dataset. By dividing the data into two groups, it was possible to determine how accurate the detection of credit card fraud was. To improve accuracy, each group underwent a total of 10 iterations. The number of recent transactions, the card issuer, the location where it is used, the card owner, the card limit, etc. are some of the crucial factors considered for the experimental setup.

Decision Tree(DT) :

Inputs: creditcard.csv data set
Output: Accuracy and Selected features.

Pseudocode for DT algorithm:

1.Activate the dataset.
2. Randomly divide dataset into a training dataset (80%) and a testing dataset (20%).
3. Configure the target variable.
4. Based on the training set, create the decision tree classifier.
5. Utilize the max depth kernel parameter to train the classifier.
6. Using the training dataset, predict the testing set.
7. Evaluate the classifier.
8. Return Accuracy.

Decision trees are supervised learning techniques that are frequently developed to address classification problems, even though they can be used to address regression and classification problems. Internal nodes hold dataset attributes, branches represent decision rules, and each leaf node provides classification in this tree-structured classifier. The decision tree in this study was trained using the decision tree class from the sklearn.model selection library. Load the creditcard.csv dataset by importing it. The dataset is randomly divided into training sets (80%) and testing sets (20%). The chosen target variable. The training data is then used to create 3 of 15 decision tree classifiers. The kernel parameter's value was set to max depth. On the basis of the training set, the testing set is predicted. The accuracy of the innovative tree-specific decision tree classifier is assessed.

Support Vector Classifier(SVC)

Input: creditcard.csv dataset
Output: Accuracy

Pseudocode for SVC algorithm:

1. Import the dataset, then read it
2. Pick random features from the dataset.
3. Create a parameter for the SVC classifier criteria.
4. The value for the parameter was a trade off.
5. Create a decision tree using SVC classifiers, and for each sample, forecast the outcome.
6. Voting was done for each predicted outcome.
7. The final result was chosen based on the results of the most popular predictions.
8. Return accuracy.

The Support vector classifier class from the sklearn.svm library is utilized in this work. Training is allocated 80% of the dataset, and testing is allocated 20%. It chooses samples at random, collecting decision trees for each sample to forecast the outcome. Every predicted outcome was put up for vote, and the outcomes with the highest number of votes were chosen as the final outcome. Support vector classifier technology is employed by the algorithm (SVC).

The various parameters for the analysis can be calculated as follows:

Equation (1) - Accuracy :The amount of instances that were appropriately categorized is counted.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

Here "TN" means True Negative, "TP" means True Positive, "FP" means False Positive and "FN" means False Negative.

Equation (2) -Recall, which determines the pertinent instances that are chosen, is also known as sensitivity.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Equation (3) - On a dataset, the F-measure assesses model accuracy.

$$F - measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (3)$$

Statistical Analysis

The Statistical Package for Social Sciences was used to perform statistical analysis in addition to experimental analysis to evaluate the study(Aldrich 2018) version 26. The analysis led to the calculation of mean, standard deviation, and standard error mean. The parameters between the two groups were compared using the independent variable T-Test. The transactions completed to date, the card issuer, the location where it is used, the card owner, and the card limit are the independent variables. The dependent variables are fraud detection and accuracy.

RESULTS

Table 1 shows the comparison of accuracy achieved during the evaluation of novel decision trees and support vector classifier models for classification with different iterations. Table 2 depicts the various parameters of both groups. The accuracy, Precision, Recall, F1 Score and support has been calculated for DT and SVC. The analysis of two groups shows that SVC has higher accuracy (98.59%) compared to the decision tree(94.86%). From Fig.1, the Learning curve shows the training and cross validation score of the decision tree classification model at various classification thresholds. From Fig.2, the Learning curve shows the training score and cross validation score of the support vector classifier model at various classification thresholds. Table 3 shows the statistical analysis of DT and SVC with different test datasets. Mean accuracy of the SVC model appears higher than the DT model. The performance of the SVC algorithm is superior to the DT algorithm. Table 4 depicts the statistical analysis of significant levels for both groups. There is a negligible Significant difference (0.000)

among the two groups. Hence SVC is better than DT. Fig.3 inferred the mean accuracy of novel decision trees and Support vector classifiers. Two independent groups' statistical analysis reveals that SVC has a higher accuracy mean than DT. The decision tree's mean error is slightly lower than that of the Support vector classifier.

DISCUSSION

A data science team will examine the data and create a model to identify and stop suspicious transactions as part of the machine learning credit card fraud detection process. Experimental work was done among two groups: novel decision tree(DT) and Support vector classifier(SVC) by varying the test size. From the experimental results (Fig. 3) done in SPSS, the accuracy of SVC is 98.59, whereas novel decision trees provide the accuracy 94.86. This depicts that SVC is better than DT. TP rate, FP rate, Precision, Recall, and F-measures are just a few of the various parameters that are compared. The proposed Support vector classifier classifier outperforms the decision tree algorithm on the SPSS graph in terms of accuracy. The analysis shows that the mean error of SVC is slightly higher than that of logistic regression and must be reduced.

The most important factor in detection of credit card fraud is accuracy and precision. In the paper by (Yazici 2020) gradient boost algorithm and random forest algorithms have been used on the Creditcardfraud.csv dataset to find the accuracy. The results of the research shows that the gradient boost algorithm is producing outstanding results than the random forest algorithm. In the article by (Porkess and Mason 2012) support vector, decision tree and linear regression classifiers were contrasted and examined.. The results show that the decision tree algorithm is producing better accuracy than the support vector and linear regression algorithms. In the article by (Everett 2003) multiple machine learning algorithms were used to detect credit card fraud. Among all the considered algorithms , the Support vector classifier stands out by producing the highest accuracy. In the work by (Baesens, Verbeke, and Van Vlasselaer 2015) naive bayes and Support vector classifier algorithms are employed to spot credit card fraud.The naive bayes algorithm produces less accuracy(93.26%) than the Support vector classifier(98.62%). (Li et al. 2021) has done research on online credit card fraud detection using the Support vector classifier, linear regression and logistic regression algorithms. The research results proved that the Support vector classifier algorithm produces better results than the linear and logistic regression algorithms. The research from (Pumplun et al. 2021) proves that the Support vector classifier algorithm is better than the decision tree algorithm in detecting effective and accurate credit card fraud.

Although the suggested methodology produced satisfactory results, the work does have some limitations. On larger data sets, the accuracy evaluation cannot produce a better result. Additionally, the mean error in SVC seems to be higher than DT. It would be preferable if the mean error could be significantly decreased. However, by using optimization algorithm techniques, the work can be improved in order to get better accuracy and a lower

mean error. Prior to classification, feature selection algorithms can be used to increase classifier accuracy. While we didn't achieve our aim of 100 percent accuracy in fraud detection, we did create a system that can get extremely close to a result within a short amount of time and make good improvements for this project. Due to the nature of the project, it is possible to combine different algorithms as modules and combine their results to increase the accuracy of the output. This model can be enhanced further by the addition of additional algorithms. This greatly increases the project's modularity and adaptability.

CONCLUSION

A new data point is classified using the SVC method based on how similar it is to previously saved data, which saves all available data. This suggests that the SVC method can quickly classify fresh data into a clearly defined category. The work shows that the accuracy for credit card fraud detection using SVC appears to be better than the DT. The accurate detection of credit card fraud is found to be significantly better with SVC than DT, but the mean error is found to be slightly higher with SVC. Thus, it can be said that SVC classifiers produce results with acceptable accuracy compared to DT.

DECLARATIONS

Conflicts of interests

No conflicts of interest in this manuscript.

Author Contributions

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

Acknowledgements

The Authors would like to impress their graduates towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formally 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. Wells Fargo, Hyderabad, India..
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Aldrich, James O. 2018. *Using IBM SPSS Statistics: An Interactive Hands-On Approach*. SAGE Publications.
2. Baesens, Bart, Wouter Verbeke, and Veronique Van Vlasselaer. 2015. *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. John Wiley & Sons.
3. C., Dr Victoria Priscill, and Victoria Priscill C. 2020. "Analysis of Performance on Classification Algorithms for Credit Card Fraud Detection." *Journal of Advanced Research in Dynamical and Control Systems*. https://doi.org/10.5373/jardcs/v12sp3/20201391.
4. Dal Pozzolo, Andrea, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. 2018. "Credit Card Fraud Detection: A Realistic Modeling and a Novel

Learning Strategy." *IEEE Transactions on Neural Networks and Learning Systems* 29 (8): 3784–97.
5. Everett, Catherine. 2003. "Credit Card Fraud Funds Terrorism." *Computer Fraud & Security*. https://doi.org/10.1016/s1361-3723(03)05001-2.
6. Jurgovsky, Johannes. 2019. *Context-Aware Credit Card Fraud Detection*.
7. Kumar, S. Senthil, S. Senthil Kumar, Ms D. Nivya, Department of Commerce with Computer Applications, SNS Rajalakshmi College of Arts and Science, Coimbatore, and TamilNadu. 2017. "Credit Card Fraud Detection Using Fire Fly Algorithm." *International Journal of Trend in Scientific Research and Development*. https://doi.org/10.31142/ijtsrd4672.
8. Li, Yuening, Zhengzhang Chen, Daochen Zha, Kaixiong Zhou, Haifeng Jin, Haifeng Chen, and Xia Hu. 2021. "Automated Anomaly Detection via Curiosity-Guided Search and Self-Imitation Learning." *IEEE Transactions on Neural Networks and Learning Systems* PP (September). https://doi.org/10.1109/TNNLS.2021.3105636.
9. Lucas, Yvan. 2019. *Credit Card Fraud Detection Using Machine Learning with Integration of Contextual Knowledge*.
10. Padvekar, Suchita Anant, Atharva College of Engineering Department of Computer Science University of Mumbai, Pragati Madan Kangane, and Komal Vikas Jadhav. 2016. "Credit Card Fraud Detection System." *International Journal Of Engineering And Computer Science*. https://doi.org/10.18535/ijecs/v5i4.22.
11. Porkess, Roger, and Stephen Mason. 2012. "Looking at Debit and Credit Card Fraud." *Teaching Statistics*. https://doi.org/10.1111/j.1467-9639.2010.00437.x.
12. Pumplun, Luisa, Mariska Fecho, Nihal Wahl, Felix Peters, and Peter Buxmann. 2021. "Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study." *Journal of Medical Internet Research* 23 (10): e29301.
13. Rapp, Burt. 1991. *Credit Card Fraud*. Loompanics Unltd.
14. Saia, Roberto, and Salvatore Carta. 2017. "A Frequency-Domain-Based Pattern Mining for Credit Card Fraud Detection." *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security*. https://doi.org/10.5220/0006361403860391.
15. Shimp, Priya Ravindra, Terna Engineering College, Navi Mumbai Nerul, India, and Vijayalaxmi Kadroli. 2016. "'Banking Expert System'With Credit Card Fraud Detection Using HMM Algorithm." *International Journal Of Engineering And Computer Science*. https://doi.org/10.18535/ijecs/v4i12.12.
16. Singh, Dueep J., and John Davidson. 2015. *Introduction to Internet Scams and Fraud - Credit Card Theft, Work-At-Home Scams and Lottery Scams*. Mendon Cottage Books.
17. Yazici, Yusuf. 2020. "Approaches to Fraud Detection on Credit Card Transactions Using Artificial Intelligence Methods." *Computer Science & Information Technology*. https://doi.org/10.5121/csit.2020.101018.

TABLES AND FIGURES

Table 1. Comparison of accuracy achieved during the evaluation of novel decision tree and Support vector classifier models for classification with different iterations.

Iterations	Sample Size	Accuracy	
		DT	SVC
1	600	94.82	98.55
2	650	94.83	98.56
3	700	94.84	98.57
4	750	94.85	98.58
5	800	94.86	98.59
6	850	94.87	98.60
7	900	94.88	98.61
8	950	94.89	98.62
9	1000	94.90	98.63
10	1050	94.91	98.64

Table 2. Experimental analysis in Kaggle for Accuracy, Precision, Recall, F1 Score and support for DT and SVC. SVC provides better Accuracy (98.59%) than DT(94.86%).

Model	Accuracy	Precision	Recall	F1 Score	Support
DT	94.86	96	96	96	23
SVC	98.59	100	96	98	23

Table 3. Statistical Analysis of Mean, Standard deviation and Standard Error of Accuracy of DT and SVC algorithms. There is a statistically significant difference in accuracy values between the algorithms. SVC had the highest accuracy (98.59%) compared with DT accuracy (94.86%). But the standard error mean is higher in SVC in comparison with DT.

GROUP		N	Mean	Std. Deviation	Std.Error Mean
Accuracy	DT	10	94.86	0.03028	0.00957
	SVC	10	98.59	0.03028	0.00957

Table 4. Comparison of the significance level for DT and SVC algorithms with value $p < 0.05$. Both DT and SVC have a significance level less than 0.05 in terms of accuracy with a 95% confidence interval.

	Levene's Test for Equality of Variances	T-test for Equality of means

	F	Sig.	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
								Lower	Upper
Accuracy	0.000	1.000	-275.479	18	0.000	-3.73000	0.01354	-3.75845	-3.70155
			-275.479	18.00	0.000	-3.73000	0.01354	-3.75845	-3.70155

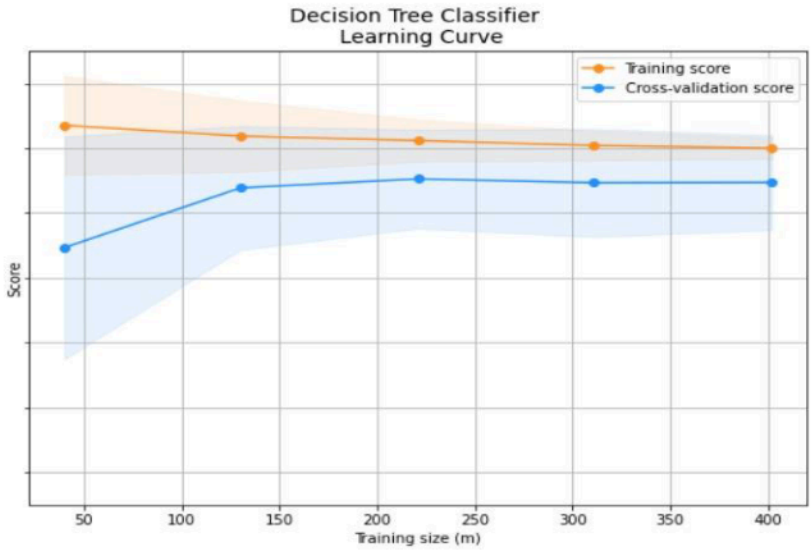

Fig.1. Decision Tree classifier learning curve
(The learning curve shows the training score and cross validation score of the decision tree classification model at various classification thresholds.)

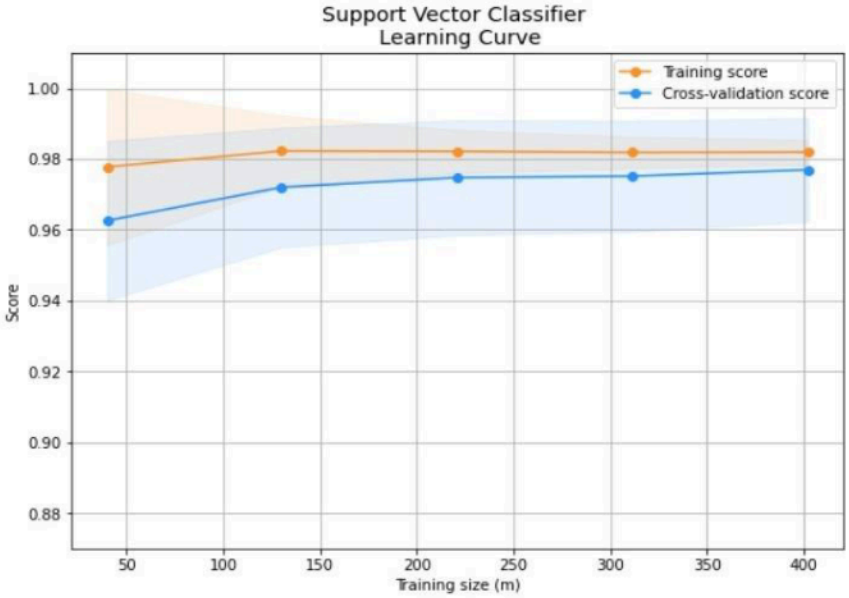

Fig.2. Support vector classifier learning curve
(The learning curve shows the training score and cross validation score of the Support vector classifier model at various classification thresholds)

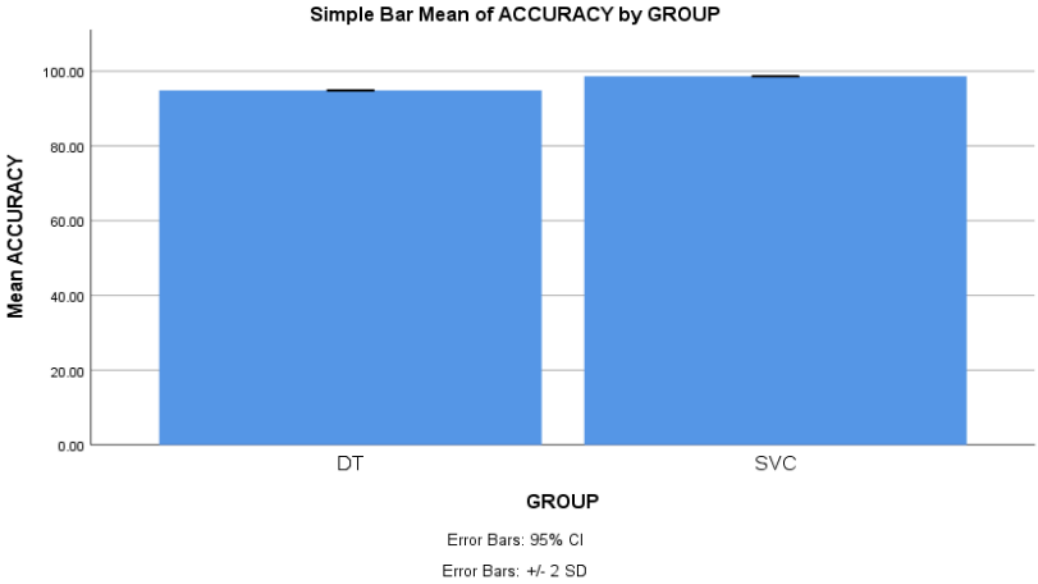

Fig.3. Bar Chart representing the comparison of mean accuracy of credit card fraud detection using DT and SVC algorithms. SVC produces better accuracy and more consistent results than DT. X-axis: DT vs SVC. Y-axis: Mean Accuracy ± 2 SD.