Title Page:

Prediction of credit card approval using support vector machine over KNN algorithm

Seeram Balu[1] , Dr.K.Somasundaram[2]

Seeram Balu[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, Pin: 602105.
seerambalu18@gmail.com

Dr.K.Somasundaram[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, Tamil Nadu, India, Pin :602105.
subbiahs.sse@saveetha.com

Keywords: Tensor, Preprocessing data, Credit Card approval, KNN Algorithm, Support vector machine, Data Manipulation, *Reduction*.

ABSTRACT

Aim: The goal of this research is to create a model that can predict if a financial institution will be able to approve credit cards for its customers. This model can assist an organisation in making an informed decision regarding whether to accept or refuse a card in order to prevent fraud, which can cause financial institutions to lose money. **Materials and methods:**Classification is performed by a support vector machine(n=10) over KNN(n=10) for false rate detection. The statistical test difference between g-power 0.08 and alpha value is α=0.05. **Result and Discussion:** The analysis of the result shows that the Support vector machine has a high accuracy compared to KNN. The mean value is 75.4910 and mean accuracy of detection is +/-1SD and significant value is p=0.001, (p<0.05) from independent sample T size.This indicates that there is a statistically significant difference between the two algorithms**.** **Conclusion:** The model is used to forecast and analyse whether a financial institution will provide a credit card or not to its customers. Our studies show that the four most significant elements that any financial business host would consider whether to provide a credit card are prior default, years worked, credit score, and debt.

Keywords: Tensor, Preprocessing data, Credit Card approval, KNN Algorithm, Support vector machine, Data Manipulation, *Reduction*.

INTRODUCTION

Credit card usage has increased significantly as a result of the internet's expansion. It is currently one of the most widely utilised payment options(United Nations Publications 2019). As the global economy grows, credit card fraud grows at an alarming rate. Credit card defaulters have also climbed dramatically. A reduction in credit score is done if the client is overdue on their loan repayment. As a consequence, credit card companies are becoming more conservative when granting credit cards to customers(Dauletova and Rahimova 2022). Furthermore, the financial institutions' decline in the United States and Europe during the US subprime mortgage crisis and the European sovereign debt crisis has highlighted worries about adequate risk management(Dauletova and Rahimova 2022). As a result, intellectuals and philosophers have paid close attention to these issues. A plethora of machine learning and statistical techniques have been developed to address credit card challenges.(Sadegh and Alizadeh 2021).A reduction in credit score is done if the client is overdue on their loan repayment.

Google Scholar has published 17,400 articles on credit card approval using machine learning in the last four years, with 1,773 publications available on ScienceDirect.(Dauletova and Rahimova 2022). One of the most well-known approaches is to use a credit card. The majority of consumers prefer to pay with credit cards since they are more convenient(Mamirbaeva and Dawletmuratova 2022). A reduction in credit score is done if the client is unwilling to repay their loan repayment. In order to maintain a low credit utilization rate, consider reducition your spending or making periodic bill payments throughout your billing cycle. A range of financial organisations, including national and commercial banks, use consumer information such as essential information, living standards, wage, yearly and monthly returns, and current livelihood income source to make decisions(Muhammadiyeva 2022). Before an application is reviewed, all of this data is analysed(International Centre for Engineering Education and UNESCO 2021).

The objective of a credit scoring model is to differentiate credit applicants into two groups: "great credit" applicants who are likely to repay their debts and "poor credit" applicants who should be gainsay credit due to a high chance of missing payments.(Dauletova and Rahimova 2022). The previous study's input data set was made up of incorrect datasets. Positive classes get a lot of attention(Shamuratov and Alimbaev 2022). The number of tuples obtained after cleaning the dataset is quite little(Shah and Clarke 2009). As a result, the input dataset is extremely small. The current study's reliability is really low.(Seytniyazova and Nizomiddinova 2022).

MATERIALS AND METHODS

SIMATS Saveetha School of Engineering's computer science and engineering department carried out the study. Each group was given a total of n=10 iterations in order to increase accuracy. The data set taken from the kaggle website. 438557 rows and 18 columns make up the credit card approval dataset. With 95% confidence and 80% pretest power, the experimental arrangement is maintained.

The dataset which proposed work used in this paper is credit card approval prediction dataset. The Dataset was collected from the open source Kaggle platform. The Hardware configuration were HP i5 processor with   a RAM size of 12GB was used.The system type used was 64-bit, OS, x64 based processor with HDD of 917 GB. Windows was used as the operating system, and the Python programming language was employed with the Jupyter Lab tool.

Support vector Machine

In a high-dimensional setting, the SVM's primary objective is to determine an ideal hyperplane for a variety of scenarios(Schalley and Springer 2009). To construct this model, you'll need more than one hyperplane(Srinivasa, Siddesh). This method employs the blister vector, which contains the information that is closest to the closed surface and coordinates with the best choice surface. It divides data into categories by projecting input vectors into a high-dimensional space and creating a hyperplane. This method is most commonly used to solve quadratic programming and non-convex, unconstrained minimization issues(Leskovec, Rajaraman, and Ullman 2014). The SVM approach is the most successful in the classification procedure. The tensor transforms a data from a low-dimensional to a high-dimensional state.(Tang et al. 2009).There are numerous kernels in tensor, and they are employed in accordance with the data. SVM algorithm convert low-dimensional to high-dimensional but reduction to low-dimensional is not possible. Distributing data is simple with Tensor. Before fitting into model data manipulation is a necessary step to be followed. Missing data is filled in, undesired data is removed, and categorical data is encoded during data manipulation. employing a single hot encoder to do preprocessing on categorical feature data. Preprocessing data of categorical features is necessary because the model will not take object data types. Transform data into an integer or float type during preprocessing data(Mani Sekhar 2021). Tensor in svm is made algorithm simple because it distributes data in such a way that hyperplane accurately separates classes.

Pseudocode for support vector algorithm

```
Input: dataset for credit card approval prediction.
Step-1:Declare all the required library files.
Step-2:Call the datasets  and assign it to a variable.
Step-3: Data manipulation
Step-4:Display the attributes that are present in the dataset.
Step-5: Preprocessing data of the categorical variables
Step-6:Test and train the SVM model using test and train dataset.
Step-7:Display the accuracy  of the algorithms.
Step-8:Plot the graph for the accuracy obtained.
Output: Accuracy in %
```

KNN algorithm

The k-nearest neighbours algorithm, abbreviated as KNN or k-NN, is a non-parametric, supervised learning classifier that exploits vicinity to make classifications or predictions about the categorization of data points(Barkalov, Shtanyuk, and Sysoyev 2022). Reduction of K value in KNN algorithm can lead to overrated or underrated. While it can be used to address both regression and classification problems, it is most typically used as a classification method, based on the idea that similar points can be found close together(An et al. 2022). Since the model only accepts integer or float data types and does not accept object data types, preprocessing of categorical features is necessary. As a result, categorical data are transferred to integer or float data types through preprocessing data (Brownlee 2016). Data processing must be done before fitting into a model. In data manipulation, missing data is filled in, undesired data is eliminated, and categorical data is encoded (Cabbri et al. 2021).

Pseudocode for KNN algorithm:

```
Input: dataset for credit card approval prediction.
Step-1:Declare all the required library files.
Step-2:Call the datasets and assign it to a variable.
Step-3: Data manipulation
```

Step-4:Display the attributes that are present in the dataset.
Step-5: Preprocessing data of the categorical variables
Step-6:Test and train the KNN model using test and train dataset.
Step-7:Plot the graph for the accuracy obtained.
Output: Accuracy in %

Statistical Analysis

A t-test is a form of presumed steady used to see if there is a significant difference in the means of two groups that are integrated in some way. The output for the grouped statistics was obtained using the spss tool (Bhattacherjee 2012). The t-test is one of many procedures used in statistics for speculation testing.The IBM SPSS version was the statistical programme employed in this study (Cooksey 2020). For this study, dependent variables include attributes like update and transaction class. In SPSS, the datasets are prepared using sample size as 32 for the support vector machine and KNN algorithm(DeCoursey 2003). The Groupid for the support vector machine algorithm is 1 and the Groupid for the KNN algorithm is 2. The Groupid is a grouping variable and the accuracy is a testing variable (Privitera 2011).

RESULTS:

In the spss the sample size is given as 10.This data is used for the analysis of the KNN algorithm and Support vector machine algorithm. These are also used to produce statistical values that can be employed for comparison, along with their loss, for each methodology. In Table 1, it is shown that the accuracy of two algorithms, KNN algorithm and Support vector machine algorithm for different n values. The group statistics table displays the total number of samples gathered along with the mean and standard deviation computed for the accuracy.
Table 3 represents the outcome of the analysis of the Independent samples test which has been performed for the KNN algorithm and the Support vector machine algorithm. From Table 3, the significance value for the one tailed test is found to be .462, two-tailed is 0.001 and it is found that the Independent samples test has been carried out at Confidence Interval of 95%.

From Table 2, the group statistics values along with the mean, standard deviation and the standard error mean for the two algorithms are also specified. The Independent sample T test is applied for the data set fixing confidence interval as 95%. Figure 1 shows the comparison of accuracy between KNN algorithm and Support vector machine algorithm. Table 3 shows the independent t sample test calculation for Accuracy and Loss for KNN algorithm and Support vector machine algorithm. Specifies mean difference and standard error difference and the comparative accuracy analysis, mean of loss between the two algorithms are specified. Figure 1 compares the mean accuracy and mean loss of the KNN algorithm and the Support vector machine technique.

DISCUSSION:

From the given study the accuracy of the KNN algorithm is 68.6990% when compared to the accuracy of the Support vector machine is 75.4910%. Statistics have been done for both the KNN algorithm and the Support vector machine algorithm in order to compare both the algorithms to find the better analysis algorithm in detection of credit card approval. For the given group, accuracy has been calculated. The mean, standard deviation and the standard error mean values for the KNN algorithm are 68.6990, 1.20455 and .38091 respectively. Similarly for Support vector machine, the mean, standard deviation and the standard error mean values are 75.4910, 1.47678 and .46700 respectively from Table 2.

Compared to antecedent analysis of KNN algorithm and Support vector machine algorithm for credit card approval,our research has got better accuracy in detecting the credit card approval analysis(Alizadeh). Previously the KNN algorithm had an accuracy of 65.6% and for the Support vector machine the accuracy was 72.9%. But in our analysis we got the accuracy of the KNN algorithm is 68.6990% and for the Support vector machine algorithm we got the accuracy of 75.4910%. Datasets have been expelled from various resources and these datasets may contain several independent and unwanted attributes are there,this should be removed in order to get the best accuracy(Sadegh and Alizadeh 2021). Hence,the data is tested and trained to get the best output accuracy. Testing takes a long time. Training for this process takes a long duration of time(Nizomiddinova 2022). Because of the random nature of the datasets, implementation time is likewise lengthy. The training data can be changed but it is a time-intensive process and if data is trained low, the accuracy will be reduced (Dauletova).

Testing and preparation of datasets for both the KNN Algorithm and Support vector machine algorithm is possible by grouping the datasets. Cleaning process of information can be efficiently increased and less time consuming in implementation of the process. The course of time utilization in preparing the dataset can be diminished.

CONCLUSION:

The KNN Algorithm(68.6990%) and the Support vector machine Algorithm (75.4910%) were used to create a model which makes predictions to approve credit cards or not. According to the findings of the experiments, the Support vector machine Algorithm method outperforms the KNN Algorithm.

DECLARATIONS

Conflict of Interests

No Conflict of Interest in this manuscript.

Authors Contributions

Author SB was involved in data collection, data analysis, and manuscript writing. Author KS was involved in conceptualization, data validation, and critical review of 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. Manac Infotech (P) Limited.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

An, Shuang, Qinghua Hu, Changzhong Wang, Ge Guo, and Piyu Li. 2022. “Data Reduction Based on NN-kNN Measure for NN Classification and Regression.” *International Journal of Machine Learning and Cybernetics*. https://doi.org/10.1007/s13042-021-01327-3.

Barkalov, Konstantin, Anton Shtanyuk, and Alexander Sysoyev. 2022. “A Fast kNN Algorithm Using Multiple Space-Filling Curves.” *Entropy* 24 (6). https://doi.org/10.3390/e24060767.

Bhattacherjee, Anol. 2012. *Social Science Research: Principles, Methods, and Practices*. CreateSpace.

Brownlee, Jason. 2016. *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Machine Learning Mastery.

Cabbri, Riccardo, Enea Ferlizza, Elisa Bellei, Giulia Andreani, Roberta Galuppi, and Gloria Isani. 2021. “A Machine Learning Approach to Study Demographic Alterations in Honeybee Colonies Using SDS-PAGE Fingerprinting.” *Animals : An Open Access Journal from MDPI* 11 (6). https://doi.org/10.3390/ani11061823.

Cooksey, Ray W. 2020. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*. Springer Nature.

Dauletova, D., and D. Rahimova. 2022. “The Importance of Vocabulary in Language Learning.” https://doi.org/10.47689/innovations-in-edu-vol-iss1-pp173-174.

DeCoursey, William. 2003. *Statistics and Probability for Engineering Applications*. Elsevier.

International Centre for Engineering Education, and UNESCO. 2021. *Engineering for Sustainable Development*. UNESCO Publishing.

Jumamuratova, N. 2022. “The Importance of Teaching English as a Second Language in Uzbekistan.”

https://doi.org/10.47689/innovations-in-edu-vol-iss1-pp176-177.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets*. Cambridge University Press.

Mamirbaeva, D., and D. Dawletmuratova. 2022. “Impact of Globalization on Higher Education.” https://doi.org/10.47689/innovations-in-edu-vol-iss1-pp149-150.

Muhammadiyeva, O. 2022. “Use of Language Elements in the Process of Social Communication.” *Zamonaviy Lingvistik Tadqiqotlar: Xorijiy Tajribalar, Istiqbolli Izlanishlar va Tillarni O‘qitishning Innovatsion Usullari*. https://doi.org/10.47689/linguistic-research-vol-iss1-pp321-323.

Privitera, Gregory J. 2011. *Statistics for the Behavioral Sciences*. SAGE.

Sadegh, Mojtaba, and Mohammad Reza Alizadeh. 2021. “Have You Noticed a Surge in Change?” *TheScienceBreaker*. https://doi.org/10.25250/thescbr.brk567.

Schalley, Christoph A., and Andreas Springer. 2009. *Mass Spectrometry of Non-Covalent Complexes: Supramolecular Chemistry in the Gas Phase*. John Wiley & Sons.

Seytniyazova, G., and Sh Nizamiddinova. 2022. “Innovative Ways of Teaching Research for Students.”https://doi.org/10.47689/innovations-in-edu-vol-iss1-pp189-191.

Shah, Mahmood, and Steve Clarke. 2009. *E-Banking Management: Issues, Solutions, and*

Strategies: Issues, Solutions, and Strategies. IGI Global.

Shamuratov, J., and M. Alimbaev. 2022. “The Significance of Technology in Teaching l2.” https://doi.org/10.47689/innovations-in-edu-vol-iss1-pp214-216.

Srinivasa, K. G., G. M. Siddesh, and S. R. Mani Sekhar. 2021. *Artificial Intelligence for Information Management: A Healthcare Perspective*. Springer Nature.

Tang, Li-Juan, Wen Du, Hai-Yan Fu, Jian-Hui Jiang, Hai-Long Wu, Guo-Li Shen, and Ru-Qin Yu. 2009. “New Variable Selection Method Using Interval Segmentation Purity with Application to Blockwise Kernel Transform Support Vector Machine Classification of High-Dimensional Microarray Data.” *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci900032q.

United Nations Publications. 2019. *Digital Economy Report 2019: Value Creation and Capture - Implications for Developing Countries*.

Tables and Figures

Table 1. Comparison between KNN algorithm and Support vector machine algorithms with N=10 samples of the dataset with the highest accuracy of respectively 70.76% and 77.56% using the 80% of training and 20% of testing dataset

n	**KNN accuracy in %**	**Support Vector Machine accuracy in%**
1	70.67	77.56
2	70.10	77.56
3	69.45	76.54
4	69.30	76.16
5	68.90	75.7
6	68.47	75.19
7	68.10	74.7
8	67.69	74.31
9	67.31	73.89
10	67.00	73.30

Table 2. Mean, Standard Deviation and Standard Error mean for Support vector machine algorithm and KNN algorithm are given below.

Accuracy	Groups	N	Mean	Std. Deviation	Std. Error Mean
	KNN	10	68.6990	1.20455	.38091
	SVM	10	75.4910	1.47678	.46700

Table 3: We find the mean and variance values by using Levene's test for equality of variance and t-test for equality means. By assuming equal variance and unequal variance values. And accuracy for both of them.

		F	**sig.**	**t**	**df**	**Sig(2-tailed)**	**Mean difference**	**std.Error difference**	**Lower**	**Upper**
Accuracy	Equal variance assumed	.566	.462	11.270	18	.001	6.79200	.60265	5.52589	8.05811
	Equal variance not assumed			11.270	17.301	.001	6.79200	.60265	5.5222q	8.06179

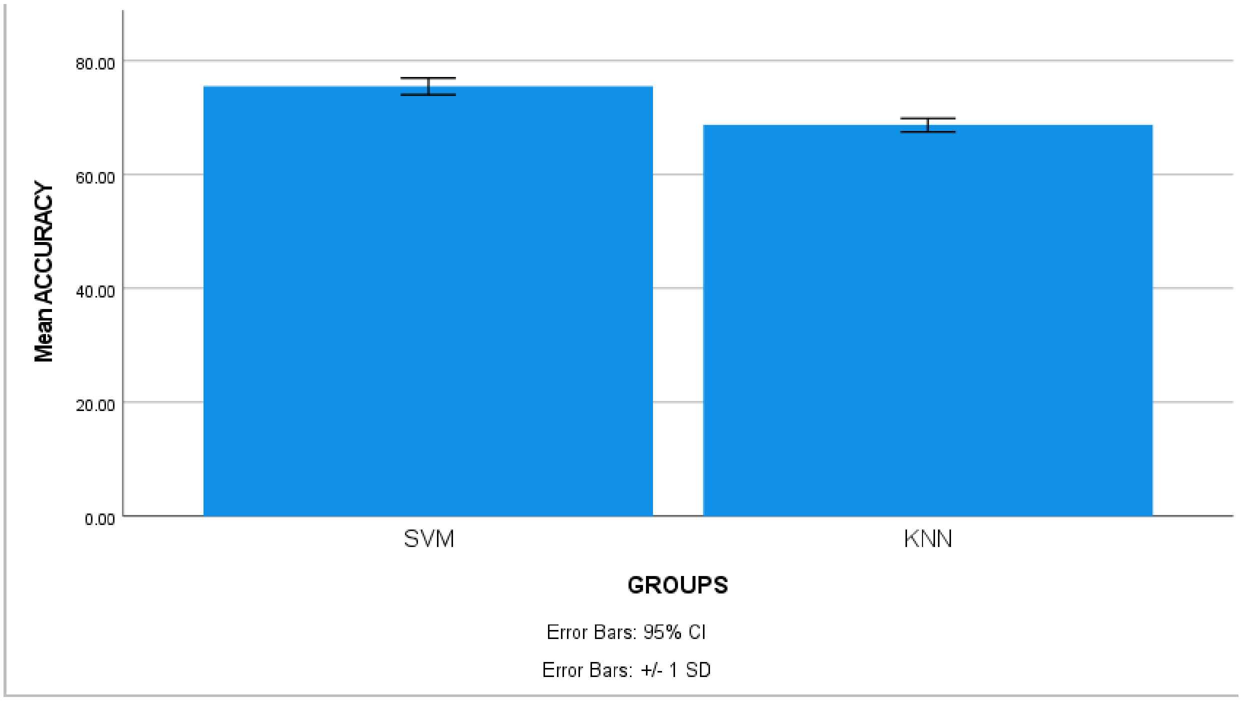

Fig 1. Comparison of KNN algorithm and Support vector machine algorithm in terms of mean accuracy. The mean accuracy of the Support vector machine algorithm is better than the KNN algorithm . X-Axis: KNN algorithm Vs :Support vector machine algorithm. Y-Axis:Mean Accuracy of Detection +/-1SD.