ILKOM Jurnal Ilmiah Vol. 15, No. 2, August 2023, pp.373-381
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Decision Tree C4.5 Performance Improvement using Synthetic Minority Oversampling Technique (SMOTE) and K-Nearest Neighbor for Debtor Eligibility Evaluation

Edi Priyanto [a,1]; Enny Itje Sela [a,2]; Luther Alexander Latumakulita [b,3,*]; Noourul Islam [c,4]
[a] Universitas Teknologi Yogyakarta, Jl. Siliwangi Jl. Ring Road Utara, Sleman, 55000, Indonesia
[b] Universitas Sam Ratulangi, Bahu, Kec. Malalayang, Manado, 95115, Indonesia
[c] Kanpur Institute of Technology, A1, UPSIDC Industrial Area, Chakeri Ward, Rooma, Kanpur, 208001, India
[1] edi.priyanto@student.uty.ac.id; [2] ennysela@uty.ac.id; [3] latumakulitala@unsrat.ac.id
** Corresponding author*

Article history: Received March 08, 2023; Revised March 27, 2023; Accepted July 21, 2023; Available online August 16, 2023

Abstract

Nowadays, information technology especially machine learning has been used to evaluate the feasibility of debtors. One of the challenges in this classification model is the occurrence of imbalanced datasets, especially in the German Credit Dataset. Another challenge is developing an optimal model for evaluating debtor eligibility. Based on these challenges, this study aims to develop an optimal model for evaluating debtor eligibility on the German Credit Dataset, using the decision trees, k-Nearest Neighbor (k-NN) and Synthetic Minority Oversampling Technique (SMOTE). SMOTE and k-NN is used to overcome challenges regarding imbalanced datasets. While the decision tree are applied to produce a debtor classification model. In general, the steps taken are preparing datasets, pre-processing data, dividing datasets, oversampling with SMOTE, and classification models using decision trees, and testing. Model performance evaluation is represented by accuracy values obtained from the confusion matrix and area under curve (AUC) values generated by the Receiver Operating Characteristic (ROC). Based on the tests that have been carried out, the best accuracy value in the test is obtained at 73.00% and the AUC value is 0.708, in parameters k = 3 and Max-Depth = 25. Based on the analysis produced, the proposed model can improve performance compared to if the dataset is not applied SMOTE.

Keywords: Debtor Eligibility; Decision Tree; Imbalanced Dataset; KNN; Machine Learning; SMOTE.

Introduction

Credit is the biggest activity in the banking world. The big problem for banks and financial institutions in general is bad loans. Although credit is the biggest contributor to profits, it is also one of the factors in unhealthy banking [1]. Credit risk must be managed properly, because if it is not managed properly, it will lead to a higher proportion of non-performing loans, which will affect banking conditions [2]. Currently, the use of information technology, especially in the field of machine learning, has been used to assess the eligibility of debtors [3]-[5], because the 5C approach (character, capital, collateral, capacity, condition) is not able to process a large number of applications every day, even though the number of debtors is increasing every day [6].

The government has launched various credit programmers aimed directly at helping the community to improve the economy on a small, medium and large scale. However, the lender still faces obstacles in distributing credit, especially in selecting potential creditors who are eligible for credit so that bad credit problems do not occur. This research aims to assess the eligibility of prospective creditors using intelligent system methods, in particular machine learning. This process can help the lender to make decisions about potential creditors, whether they are worthy of credit or not.

There are two challenges that need to be addressed in this research. First, there is data imbalance in the dataset. This describes the general state of the debtor eligibility evaluation process, namely the number of debtors who have the potential to default is less than the number of debtors who are currently in default [7]. The imbalance in the number of records in the dataset class is called an imbalanced dataset [2]. In an imbalanced dataset, machine learning algorithms produce biased predictions and have invalid accuracy [8]. The second challenge is to develop an optimal classification model.

http://dx.doi.org/10.33096/ilkom.v15i2.1676.373-381

For the first challenge related to imbalanced datasets, some researchers have implemented sampling techniques [8]-[10]. These techniques consist of under sampling and oversampling. SMOTE is an oversampling technique that creates replicates from minority data. SMOTE can increase efficiency and produce better results than under sampling methods [11], [12], [13]. What is done to overcome the second challenge is the implementation of machine learning algorithms. To produce an optimal machine learning model, the parameters of the algorithm are given a variety of values [14], [15].

The solution of each research to the problem of imbalanced datasets is very diverse. Siringoringo's research [10] combined SMOTE with k-means on a credit card fraud dataset. With 10 cross-fold parameters, the average G-mean is 81.0% and the average F-measure is 81.8%. Similar research was also carried out by Astuti & Lenti [9] using SMOTE and KNN on a car valuation dataset. The research showed that SMOTE and KNN were able to provide an accuracy value of 93.11% at k=3. In Sutoyo & Fadlurrahman's research [8], Artificial Neural Network (ANN) and SMOTE were used to classify the performance rating of television commercials. The experimental results showed that the performance of the combination of ANN and SMOTE achieved an accuracy of 87.06%. The use of SMOTE and Naïve Bayes Classifier (NBC) was used by Kurniawati [16] to study tuberculosis in children. The results show that the use of SMOTE and NBC can provide an increase in performance, namely f-measure of 3.2% and ROC area of 17.9%.

Specific research on credit scoring has also been carried out by several researchers. Researchers Yani et al [17] processed the German credit dataset without SMOTE. The method implemented is an artificial neural network with an architecture of 24 neurons in the input layer, 20 neurons in 1 hidden layer, 2 neurons in 1 output layer, 10000 iterations, activation function using binary sigmoid and learning rate of 0.1. The results of the backpropogation network analysis give an accuracy value of 0.7133 (71.333%) and an AUC value of 0.72. Thus, the research results lead to a fair model. Researchers [2], [18] used decision tree techniques and Support Vector Machine to process the German credit dataset without oversampling. While researchers [4], [19] used a decision tree algorithm to evaluate credit using primary data, with fewer attributes and records compared to the German Credit Dataset. The results of these studies show significant differences. Researchers [2], [18] produced fair models, while researchers [4], [19] produced very good models. The use of other primary data for credit evaluation was also carried out by researcher [3], who used Tsukamoto fuzzy to evaluate credit in a company. The weakness of using fuzzy techniques is the determination of the membership function that fits the data. This method is also used for ranking, not classification.

Research [8]-[10], [16] has concluded that the use of SMOTE is able to produce better values when compared to not using SMOTE with different datasets. This is because the characteristics of the dataset have a variety of attributes, data values, and many rows, which can cause significant differences in the results. Therefore, it is still possible to conduct research using the same method with different data sets. Significant differences in credit scoring model results among researchers [2], [4], [18], [19] also motivated researchers to conduct a study on the application of SMOTE to the German credit dataset. The aim of this study is to develop an optimal model for performing credit scoring on the German Credit Dataset using SMOTE, KNN, and decision trees. In SMOTE, the KNN concept is used to increase the amount of data in the minority class, which is done by generating new data based on k nearest neighbors [8]. According to researchers [14], [15], decision trees are suitable for credit scoring because the algorithm has a low computational value and produces a high-performance value. The performance model indicators used in this study are the confusion matrix and the Receiver Operating Characteristic (ROC) curve.

Method

In general, the steps taken in this research consist of 2 parts, namely classification without oversampling and classification with oversampling. This process includes the stages of preparing the dataset, preprocessing, dividing the dataset, sampling process with SMOTE, building a classification model using a decision tree, and testing, as shown in **Figure 1**.

A. Dataset Preparation

The dataset used in this research is the German Credit Dataset provided by Professor Dr Hans Hofmann Institut F'ur Statistik und "Okonometrie Universit" in Hamburg in the UCI Machine Learning Repository. The dataset consists of 1,000 rows with 21 attributes and 2 class labels.

B. Data Pre-processing

The initial data preparation stage includes attribute selection and class distribution analysis of the dataset. The selection of attributes used as determining parameters in the debtor eligibility evaluation process is done by paying attention to the value of missing value, correlation, ID-ness, stability and text-ness in each attribute.

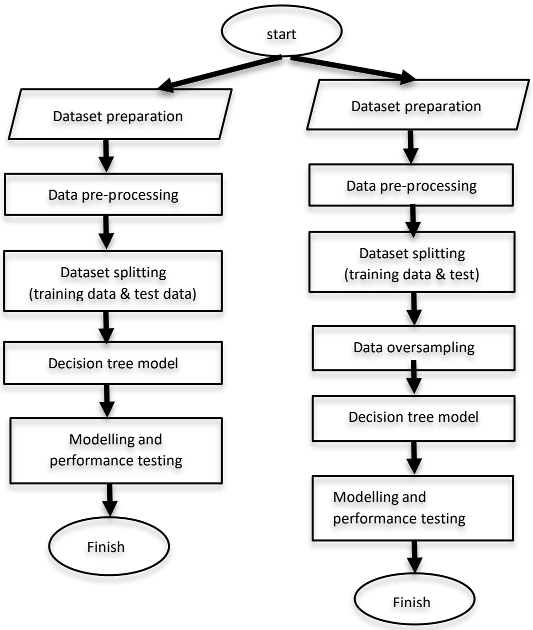
start
Dataset preparation
Data pre-processing
Dataset splitting (training data & test data)
Decision tree model
Modelling and performance testing
Finish
Dataset preparation
Data pre-processing
Dataset splitting (training data & test)
Data oversampling
Decision tree model
Modelling and performance testing
Finish

Figure 1. Flow of research steps

The selection of attributes is carried out by measuring the quality of each attribute as shown in **Figure 2**, paying attention to the statistical conditions of the training dataset in **Table 1**.

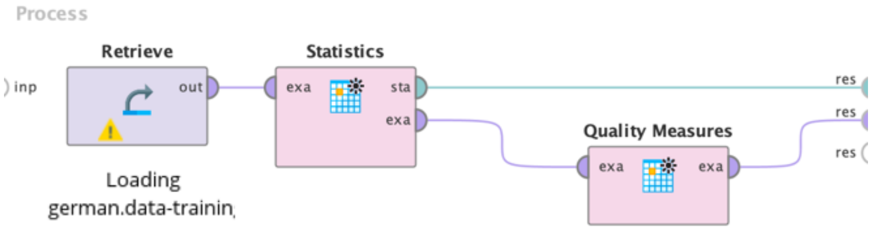

Figure 2. Process Attribute Quality Measurement

Table 1. Table statistics of initial training dataset setup

Attribute	C	I	S	M	T
checking_acc	1,12%	0,57%	39,71%	0%	1,52%
Duration	4,90%	4,29%	18,71%	0%	0,00%
credit_hist	3,49%	0,71%	52,57%	0%	1,57%
credit_purpose	0,24%	1,43%	27,14%	0%	1,82%
credit_amount	2,21%	93,71%	0,29%	0%	0,00%
saving_acc	0,00%	0,71%	62,86%	0%	1,57%
present_employment	1,93%	0,71%	2,29%	0%	1,57%
installment_rate	0,71%	0,57%	49,57%	0%	0,00%
status_sex	0,43%	0,57%	54,86%	0%	1,52%
Guarantors	0,00%	0,43%	91,57%	0%	1,92%
present_residence	0,00%	0,57%	41,71%	0%	0,00%
Property	0,00%	0,57%	32,71%	0%	1,97%
Age	0,59%	7,57%	5,43%	0%	0,00%
installment_plan	1,85%	0,43%	82,43%	0%	1,92%
Housing	1,08%	0,43%	70,86%	0%	1,92%
existing_credit	0,22%	0,57%	64,14%	0%	0,00%
Job	0,12%	0,57%	64,14%	0%	1,97%
maintenance_resources	0,00%	0,29%	84,43%	0%	0,00%
Phone	0,62%	0,29%	58,86%	0%	1,87%

Attribute	C	I	S	M	T
foreign_worker	0,52%	0,29%	96,57%	0%	1,87%

C : Correlation; I : ID-ness; S : Stability; M: missing value; T : Text-ness

Based on **Table 1**, the statistics of the dataset of attribute quality measurement results can be explained as follows:

1. All attributes in the training dataset have a low missing value of 0%, indicating that there is no null data in each attribute.
2. The majority of attributes have low ID-ness values, with the exception of the credit_amount attribute which has a very high ID-ness value of 93.71%; attributes with low ID-ness values indicate that the data for these attributes have little variation in value.
3. There are 2 attributes with very high stability values above 90%, namely the guarantors and foreign_worker attributes. Attributes with a high stability value indicate the degree of dominance of a value on that attribute.
4. There are 7 attributes with correlation values above 1%, namely duration, credit_hist, credit_amount, present_employment, installment_plan, checking_acc, and housing. The correlation value indicates the level of linear correlation of the attribute to the class label.

In general, the selected attributes should have low values for the parameters missing value, stability and ID-ness, as well as attributes with high correlation [20]. A high percentage of text-ness is usually avoided because too much text slows down the machine learning process. By paying attention to the stability value, the guarantors and foreign_worker attributes are not used as attributes to determine debtor eligibility. The credit_amount attribute, despite having a high id-ness value (93.71%), is still used as an attribute to determine debtor eligibility in this study, considering that the credit_amount attribute has low stability and missing values, and is the attribute with the third highest correlation value of all attributes in the training dataset. Thus, the number of attributes used in the following process is 18.

C. Dataset Splitting

The class distribution in the initial dataset used is that 70% (700 rows) of the data is class 1 data (feasible) and 30% (300 rows) of the data is class 2 data (not feasible), as shown in **Figure 3**.

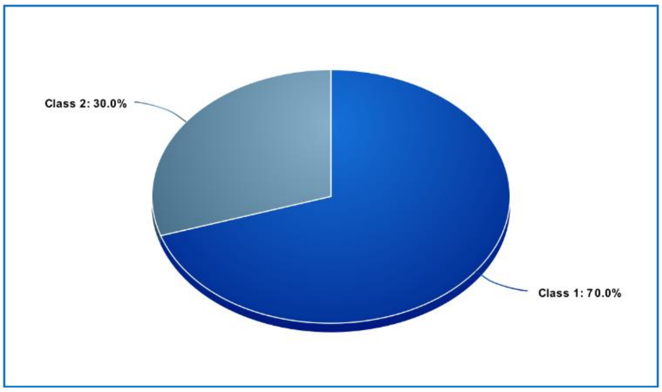

Figure 3. Initial Dataset Class Distribution

This stage is used to divide the dataset randomly (without using specific criteria) into 2 parts, namely 70% as a training dataset and 30% as a test dataset. The final result of this stage is the formation of 2 data sets, namely the training data set and the test data set. The composition of the training data distribution is class 1 as much as 490 data and class 2 as much as 210 data, while the distribution of the test data set is 210 data in class 1 and 90 data in class 2. The data set is then processed using a decision tree without oversampling and a decision tree with oversampling.

D. Data Oversampling

This stage is only performed on classification with an oversampling process, which is the process of applying oversampling techniques to an imbalanced training dataset using SMOTE. SMOTE works by using the KNN algorithm to create synthetic data in the minority class by oversampling. SMOTE randomly selects the minority data, then the KNN value is set by the user. Synthetic data is generated between the random data and the KNN value [21]. The final result at this stage is the formation of a new training dataset with balanced class distribution conditions, namely 490 class 1 data and 490 class 2 data, so that the total training data is 980 data. In this research, the variation of the k parameter value starts from 2 to 10. These values are combined with the decision tree parameters to get a good model. Each k parameter combined with the decision tree parameters will be tested and its performance will be seen. Thus, the best k value is selected based on the combination of the two parameters. To simplify the process, the selection of k

values is chosen from the smallest value of 2 (according to the number of data classes) and then increased by 1 step until the value of $k = 10$. These k values have also given good performance in [9] and [10].

E. Building Decision Tree Models

This stage is the process of forming 2 decision tree models, namely models formed using training datasets with imbalanced dataset conditions and models formed using training datasets with balanced dataset conditions.
C4.5 is one of the algorithms that can be used to create decision trees, which is the evolution of the ID3 algorithm [22]. In general, the C4.5 algorithm for building decision trees is as follows:

a. Select the attribute as the root
b. Create branches for each value
c. Split the cases in the branch
d. Repeat the process for each branch until all the cases on the branch have the same class.

The final result of this stage is the formation of 2 decision tree models. To generate the best value, the Max-Depth parameter variation is performed. These values are generated by simulating Max-Depth parameters ranging from 2 to 25. The combination of k and Max-Depth values is tested and the accuracy value is recorded. From the combination of k and Max-Depth values shown in this article, the performance accuracy is better than 0.500.

The difference between the decision tree stages of the model without oversampling and the model with oversampling is the data that becomes the input to the decision tree. The data in the model without oversampling is the result of initial data processing, while the data in the model with oversampling is the result of oversampling. In the oversampled model, the value of k and Max-Depth are combined to give the best performance.

F. Model and Performance Testing

This stage is the process of testing the two decision tree models formed in the model-building stage. The model testing process uses the test dataset formed in the dataset splitting process. Model testing is also performed by varying the depth parameter in each model. Measuring model performance using Confusion Matrix and ROC.

Confusion matrix is a method that uses a model as shown in **Table 2**, where the calculation of the accuracy value is based on the True Positive (TP) and True Negative (TN) parameters and the sum of all data. ROC is a graph that describes the relationship between false positives (horizontal axis) and true positives (vertical axis) [1].

From the confusion matrix the percentage accuracy value is obtained and from the ROC curve the area under the curve (AUC) value can be generated. The expected percentage accuracy value is close to 100%, while from **Table 3** the expected AUC value is close to 1.

Table 2. Confusion Matrix

		Actual Class	
		Positive	***Negative***
Predicted Class	***Positive***	TP	FP
	Negative	FN	TN

The interpretation of the AUC values is shown in **Table 3** [1].

Table 3. AUC Value Interpretation

AUC	**Interpretation**
1.0 (100%)	Perfect Model
0.9 – 0.99 (90 – 99%)	Excellent Model
0.8 – 0.89 (80 – 89%)	Very Good Model
0.7 – 0.79 (70 – 79%)	Fair Model
0.51 – 0.69%)	Poor Model
< 0.5 (50%)	Worthless

Results and Discussion

The results of the model accuracy with variations in parameter k and Max-Depth can be seen in **Table 4**. From **Table 4**, it can be seen that the highest model accuracy value from the results of testing parameter values using the training dataset is 0.704 or 70.4%. Testing is done with the value of $k = 2$ and Max-Depth = 2 and then the accuracy is checked. If the accuracy trend increases, the value of k and Max-Depth are increased. The highest accuracy value is obtained in the model test with the Max-Depth parameter value of 25 and the parameter k of 3, so it can be concluded that the optimum value for the Max-Depth parameter is 25 and the parameter k is 3. At this stage, as the value of k increases, the accuracy value decreases, so no additional Max-Depth is done.

Table 4. SMOTE Parameter Training Accuracy

	D=2	**D=4**	**D=7**	**D=10**	**D=15**	**D =25**
k=2	0,509	0,530	0,643	0,656	0,676	0,681
k =3	0,510	0,513	0,641	0,689	0,698	0,704
k =4	0,508	0,514	0,633	0,660	0,696	0,693
k =5	0,510	0,516	0,577	0,622	0,677	0,682
k =6	0,512	0,527	0,640	0,642	0,674	0,662
k =7	0,507	0,511	0,575	0,633	0,628	0,630
k =8	0,512	0,515	0,591	0,674	0,690	0,698
k =9	0,507	0,513	0,575	0,617	0,682	0,684
k =10	0,509	0,512	0,598	0,669	0,698	0,698

k : KNN parameters in SMOTE technique
D: Max-Depth Parameter in Decision Tree

This research also compares training without the use of SMOTE. The test results are shown in **Table 5**.

Table 5. Parameter Training Accuracy Table Without SMOTE

Max-Depth	***Accuracy***
2	0,694
4	0,693
7	0,699
10	0,700
15	0,706
25	0,709
30	0,702
35	0,690

From **Table 5** it can be seen that the highest model accuracy value from testing parameter values using the training dataset is 0.709 or 70.9%. The highest accuracy value is obtained when testing the model with the Max-Depth parameter value of 25, so it can be concluded that the optimal value for the Max-Depth parameter of the decision tree model without applying the SMOTE oversampling technique is 25, because the addition of the Max-Depth = 30 and Max-Depth = 35 values causes the accuracy value to decrease.

Model testing was performed using the test data set. **Table 6** shows the confusion matrix of testing with SMOTE and Decision Tree with parameter value $k = 3$ and Max-Depth = 25, which achieved an accuracy of 73.00%. While the test without SMOTE with the parameter Max-Depth = 25 gave an accuracy value of 69.00%. This is an increase in accuracy of 4.00%. **Figure 4** shows the ROC test with SMOTE, which has an AUC value of 70.80%. In **Figure 5** the AUC without SMOTE is 66.70%. This also supports the research finding that there is a 4.10% increase in performance when testing with SMOTE.

This research has successfully applied SMOTE and decision trees to the German Credit Dataset. SMOTE and KNN are used to deal with class imbalance in the dataset, while decision trees are used to perform the classification.

Table 6. Confusion Matrix Table of Model Testing

Model with SMOTE		**Actual Class**	
		positive	***negative***
Predicted Class	***positive***	176	47
	negative	34	43
Accuracy: 73.00%; Classification Error: 27.00%			

Model without SMOTE		**Actual Class**	
		positive	***negative***
Predicted Class	***positive***	203	86
	negative	7	4
Accuracy: 69.00%; Classification Error: 31.00%			

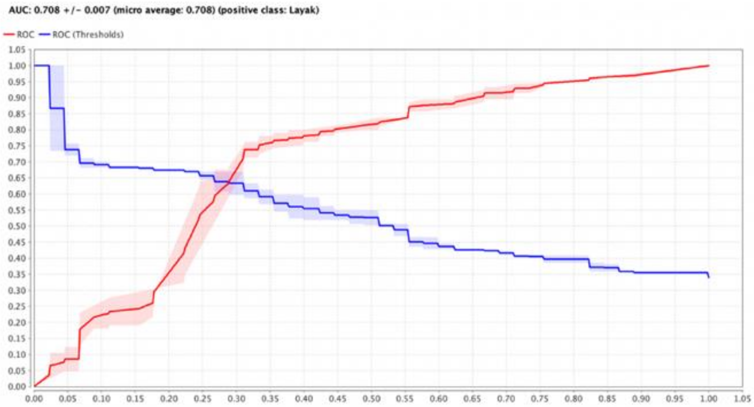

Figure 4. ROC Model with SMOTE

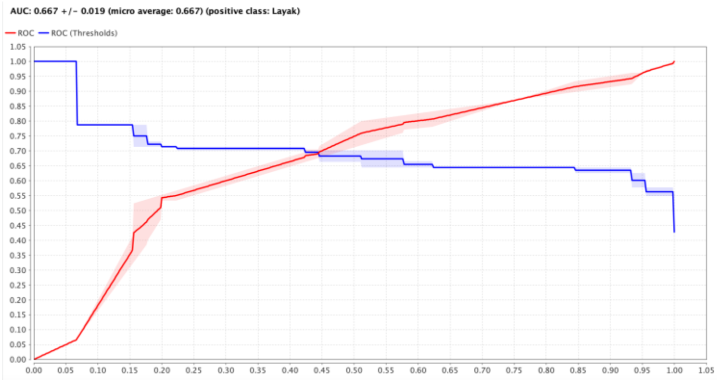

Figure 5. ROC Model without SMOTE

Improved model performance is obtained by applying the optimal parameter value for Max-Depth of 25 and the k-parameter for the SMOTE oversampling technique of 3.

The results of the model testing show that the application of oversampling techniques successfully improves the performance of the decision tree model with an accuracy value of 73.00% and an AUC value of 70.80% when compared to the performance of the decision tree model without the application of oversampling techniques with an accuracy value of 69.00% and an AUC value of 66.70%. **Table 7** shows the comparison of the results of this research with previous relevant research.

Table 7. Comparison of Oversampling Research Performance

Ref	**Domain**	**SMOTE**	**Classification**	**Accuracy (%)**
Siringoringo [10]	Credit card Fraud	Yes	KNN	81
Astuti [9]	Car Evolution	Yes	KNN	93,11
Sutoyo [8]	TV Advertisement Performance Rating	Yes	ANN	87,06
Yani [17]	German Credit dataset	No	ANN-BP	71,33
Putra [18]	German Credit Dataset	No	SVM	74
Muslim [2]	German Credit Dataset	No	Decision Tree	75,1
This Research	German Credit Dataset	Yes	Decision Tree	73

From **Table 7**, it can be seen that research using SMOTE [8]-[10] can produce better average accuracy values compared to research without SMOTE [2], [17], [18].

From the comparison with research using the German Credit Dataset, the results of this study are included in the fair model. This category is similar to studies [2], [17], [18]. Based on the accuracy value, this research is higher than research [17], but slightly lower than research [2], [18] Research [2] improved the improvement by machine learning feature selection, while research [18] by selecting the appropriate kernel.

In the implementation process, this research needs to be followed up by testing the primary data held by the lender. To facilitate this testing, it is necessary to develop an application that makes it easier for lenders to know the eligibility of prospective creditors, whether they are in a class worthy or unworthy of being granted credit.

Conclusion

Based on the performance of accuracy and AUC values, the use of oversampling techniques (SMOTE, KNN) and decision trees can produce better results than the decision tree method without oversampling. Specifically, the proposed model can produce an accuracy value of 73.00% and the AUC value obtained is 70.80%. The increase in the accuracy value with oversampling is 4% (from 69.00% to 73.00%). The AUC value also increased by 4.10% (from 66.70% to 70.80%). Therefore, the results of this study are included in the fair model category. This research still needs to be evaluated regarding the increase in accuracy value. The proposed solution is to optimize using machine learning methods for classification with parameter variations.

Acknowledgement

This research was funded by the Master's Thesis Grant Programmer of the Directorate of Research, Technology and Community Service, Directorate General of Higher Education, Research and Technology of the Ministry of Education, Culture, Research and Technology.

References

[1] M. I. Sari, S. Siregar, and I. Harahap, “Manajemen Risiko Kredit bagi Bank Umum,” in *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 2020, pp. 553–557. [Online]. Available: https://prosiding.seminar-id.com/index.php/sainteks

[2] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo, “Improving Accuracy of C4.5 Algorithm using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction,” in *2018 International Conference on Information and Communications Technology (ICOIACT)*, Mar. 2018, pp. 141–145.

[3] S. Marentika Br Tarigan, H. Jaya, and I. Santoso, “Sistem Pendukung Keputusan Untuk Menentukan Kelayakan Calon Kreditur Pada PT.ITC Finance SM Raja Medan Dengan Menggunakan Metode Fuzzy Tsukamoto,” *Jurnal CyberTech*, vol. 4, no. 1, pp. 1–10, 2021, [Online]. Available: https://ojs.trigunadharma.ac.id/

[4] I. Rahmianti, “Analisis Kelayakan Pemberian Kredit Koperasi Dengan Metode Data Mining Decision Tree,” *Jurnal Informatika & Rekayasa Elektronika)*, vol. 5, no. 2, 2022, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

[5] S. Wahyuningsih and D. Retno Utari, “Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit,” in *Konferensi Nasional Sistem Informasi 2018*, 2018, pp. 8–9.

[6] X. Dastile, T. Celik, and M. Potsane, “Statistical and machine learning models in credit scoring: A systematic literature survey,” *Applied Soft Computing Journal*, vol. 91, p. 106263, 2020, doi: 10.1016/j.asoc.2020.106263.

[7] S H. Pramesti, I. Indahwati, and U. D. Syafitri, “Analisis Regresi Logistik dan Cart untuk Credit Scoring dengan Penanganan Kelas Tak Seimbang,” *Xplore: Journal of Statistics*, vol. 11, no. 3, pp. 226–237, Sep. 2022, doi: 10.29244/xplore.v11i3.1015.

[8] E. Sutoyo, M. Asri Fadlurrahman, J. Telekomunikasi Jl Terusan Buah Batu, K. Dayeuhkolot, K. Bandung, and J. Barat, “Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network,” *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 6, no. 3, pp. 379–385, Dec. 2020.

[9] F. Dwi Astuti and F. Nova Lenti, “Implementasi SMOTE untuk mengatasi Imbalance Class pada Klasifikasi Car Evolution menggunakan KNN,” *Jurnal JUPITER*, vol. 13, no. 1, pp. 89–98, 2021.

[10] R. Siringoringo, “Klasifikasi Data Tidak Seimbang menggunakan Algoritma SMOTE dan k-Nearest Neighbor,” 2018.

[11] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, “Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk,” *Complex and Intelligent Systems*, 2022, doi: 10.1007/s40747-021-00614-4.

[12] Y. E. Ardiningtyas and P. H. P. Rosa, “Analisis Balancing Data Untuk Mningkatkan Akurasi Dalam Klasifikasi,” in *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, Mar. 2021, pp. 24–28.

[13] H. Hairani, K. E. Saputro, and S. Fadli, “K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes,” *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[14] W. Liu, H. Fan, and M. Xia, “Step-wise multi-grained augmented gradient boosting decision trees for credit scoring,” *Eng Appl Artif Intell*, vol. 97, Jan. 2021, doi: 10.1016/j.engappai.2020.104036.

[15] X. Zhou *et al.*, “A state of the art survey of data mining-based fraud detection and credit scoring,” in *MATEC Web of Conferences*, Aug. 2018, vol. 189, pp. 1–15. doi: 10.1051/matecconf/201818903002.

[16] Y. E. Kurniawati, “Class Imbalanced Learning Menggunakan Algoritma Synthetic Minority Over-sampling Technique-Nominal (SMOTE-N) pada Dataset Tuberculosis Anak,” 2019.

[17] A. Yani and E. Hegarini, “Analisa Kelayakan Kredit Menggunakan Artifcial Neural Network dan Backpropogation (Studi Kasus German Credit Data),” *Jurnal Ilmiah KOMPUTASI*, vol. 18, no. 4, pp. 385–390, Dec. 2019.

[18] D. P. Putra, D. Bheta, A. Wardijono, J. Bri, R. Dalam, and J. Selatan, “Analisis Akurasi Penerapan Algoritma Support Vector Machine Menggunakan Kernel Radial Basis Function pada Penentuan Kelayakan Kredit (Studi Kasus German Kredit Data),” *Jurnal Ilmiliah KOMPUTASI*, vol. 19, no. 2, pp. 175–180, 2020, doi: 10.32409/jikstik.19.2.2786.

[19] R. Setiawan, “Analisis Kelayakan Pemberian Kredit Nasabah Koperasi Menggunakan Algoritma C4.5,” *Techno Xplore Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 5, no. 2, pp. 75–78, 2020.

[20] I. B. K. Manuaba, I. Sutedja, and R. Bahana, “The evaluation of supervised classifier models to develop a machine learning API for predicting cardiovascular disease risk,” *ICIC Express Letters*, vol. 14, no. 3, pp. 219–226, 2020, doi: 10.24507/icicel.14.03.219.

[21] W. Suci and S. Samsudin, “Algoritma K-Nearest Neighbors dan Synthetic Minority Oversampling Technique dalam Prediksi Pemesanan Tiket Pesawat,” *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, p. 1775, Jul. 2022, doi: 10.30865/mib.v6i3.4374.

[22] E. I. Sela and R. Pulungan, “Osteoporosis identification based on the validated trabecular area on digital dental radiographic images,” *Procedia Comput Sci*, vol. 157, pp. 282–289, 2019, doi: 10.1016/j.procs.2019.08.168.