OPEN ACCESS
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 9, No. 2, Dec 2023. pp. 150-161
doi: doi.org/10.21108/ijoict.v9i2.860

Reducing Lending Risk: SVM Model Development with SMOTE for Unbalanced Credit Data

Josya Ryan Alexandro Purba [1], Qilbaaini Effendi Muftikhali [2], Bony Parulian Josaphat [3*]

[1] *BPS Kabupaten Nias Selatan*
Telukdalam, Kabupaten Nias Selatan, Sumatera Utara, Indonesia

[2] *Program Studi Sistem Informasi, Universitas Telkom*
Jakarta, Indonesia

[3] *Program Studi Komputasi Statistik, Politeknik Statistika STIS*
Jakarta, Indonesia
[*] bonyp@stis.ac.id

Abstract

Lending is an important activity for banks in managing available funds. However, lending is also an activity that has a high risk, because not all customers who borrow funds can fulfill their responsibilities for existing agreements. Therefore, it is necessary to have a method that can predict the feasibility of granting credit to customers in order to minimize the risks that arise. This research uses a machine learning method, namely a Support Vector Machine (SVM) in predicting creditworthiness. This method is applied and compared before and after using the Synthetic Minority Oversampling Technique (SMOTE) on historical bank credit data BPR NBP 16 Rantau Prapat, North Sumatra which is useful for balancing data and finally, finding the best parameters with grid search. Based on the analysis results based on the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), SVM with SMOTE shows better results, which is 96%, compared to SVM without SMOTE, which is 56%.

Keywords: Lending, Risk, Machine Learning, Support Vector Machine, SMOTE

I. INTRODUCTION

Credit activities involve the provision of goods, services, or money by the creditor or lender to the debtor or loan recipient. It is based on trust between the two parties, where the debtor promises to pay back to the creditor according to an agreed schedule [1]. Lending is an important mode of operation for banks in terms of managing available funds. However, lending is also a risky activity for banks, because not all customers who apply for loans are able to fulfill their agreed obligations.

The difference in financial conditions between one customer and another is something that is important to analyze for creditworthiness [2]. Knowing whether or not a customer is given credit, a method is needed that can minimize the risks that will arise later. Creditworthiness assessment or analysis is an evaluation of companies or individuals applying for credit. This is similar to a feasibility study that is conducted to assess whether a credit applicant or individual is eligible for credit. Lending analysis can make a significant

Received on 25 Okt 2023, Revised on 29 Nov 2023. Accepted and Published on 30 Dec 2023.

contribution to the performance of bank directors or employees in assessing credit applications from debtors. Determining whether or not a debtor is eligible to receive credit is not an easy thing, because it generally requires an assessment of factors such as capacity, character, capital, collateral, and the condition of the debtor. These factors become important references in the decision-making process related to granting credit. The ability to correctly predict consumer creditworthiness will help reduce the risk of bad debts and improve the efficiency of the decision-making process. Machine learning is one of the effective methods in analyzing credit risk and can be used to predict creditworthiness [3].

Machine learning is one of the branches in the field of data science in which the development of algorithms has the ability to learn from existing data. The algorithm can be used to make decisions or predictions based on the information contained in the data [4, 5]. Machine learning has the ability to learn and recognize patterns from large and complex data. There are various machine learning models that can be used, but in this study, the focus is on using the Support Vector Machine (SVM) model. The model has often been used for research on creditworthiness analysis, but in this study, it uses variables from the bank under study.

SVM is a machine learning model used to predict data labels by finding the best line that can separate two classes well. The line is referred to as a hyperplane. SVM can be used to predict classes or values in data by building models from training data. SVM uses a data transformation technique called kernel trick that transforms separable data and non-separable data into separable data so that it can be easily separated by the best line [4]. SVM is included in the supervised classification method because in the training process, SVM requires predetermined learning objectives [6]. One of the promising classification techniques in the field of machine learning is SVM. SVM works on the basic principle of binary classification using maximum margins, and can be developed to solve non-separable cases. SVM has been successfully utilized in various aspects, such as face recognition, handwriting recognition, iris recognition, and information retrieval systems [7]. Based on research titled Credit scoring with a data mining approach based on support vector machines, the vector machines, the SVM-based approach credit scoring model can correctly classify applications as accepted or rejected, thus minimizing creditor risk and is helpful in future savings [8].

Data imbalance is a condition where the number of samples in a dataset is very unbalanced between certain classes or labels [9, 10]. This data imbalance can lead to biased machine learning models, where the models tend to predict the majority class very accurately while the minority class is often ignored. To address the problem of data imbalance, SMOTE (*Synthetic Minority* Over-*Sampling Technique*) aims to create a synthetic sample of minority classes so that their numbers become more balanced with the majority classes.

The difference between this research and previous research is that this research uses historical credit data of BPR NBP 16 Rantau Prapat, North Sumatra. The data will be analyzed and checked. This research will use several preprocessing techniques, such as categorical data encoding using One Hot Encoding, data standardization with Robust Scaler, and finally comparison before and after data balancing using SMOTE. Furthermore, the data will be trained and tested using machine learning techniques, namely Support Vector Machine.

II. LITERATURE REVIEW

Research on imbalanced data on credit data has been done a lot. such as research conducted by Manvar et al [9]. This study uses three credit-scoring datasets that discuss oversampling techniques, namely SMOTE. All datasets represent varying amounts of imbalance ratios. It has been found that the imbalance ratio of each dataset can seriously affect the results and make them biased.

Another study was conducted by Alam et al [11], who discussed the Investigation of Credit Card Default Prediction in Unbalanced Data Sets. there are several oversampling methods performed, namely random oversampling, ADAYSN, SMOTE, and others. The use of smote in this study reaches AUC-ROC with a value of 90%.

Furthermore, there is research conducted by Doko et al [12] which discusses the Credit Risk Model Based on Central Bank Credit Registry Data. This study aims to assess various machine learning models in an effort

to create an accurate credit risk assessment model using data from the real credit registry data set from the Central Bank of the Republic of North Macedonia. This study compares five machine learning models to classify credit risk data, namely logistic regression, decision trees, random forest, support vector machines (SVM), and artificial neural networks. In this study, the SMOTE technique was carried out for data balance. The results show that the best accuracy is by using a decision tree that works on unbalanced data with and without scaling, followed by random forest and linear regression.

Then there is research from Baroka et al [13], the research discusses how to improve credit risk prediction in the online P2P lending industry. In the context of online P2P lending, there is a risk that borrowers may default on their loans, and this can have an impact on the financial performance of lenders and investors. the sampling techniques used are SMOTE and random undersampling.

III. RESEARCH METHOD

A flowchart illustrating the research methodology can be seen in Fig 1. dataset from Bank BPR NBP 16 Rantau Prapat, North Sumatra. Then preprocessing such as categorical data encoding and numerical data standardization, applying SMOTE, providing training data and testing data for SVM, evaluating SMOTE (before and after use, and finally conducting analysis.

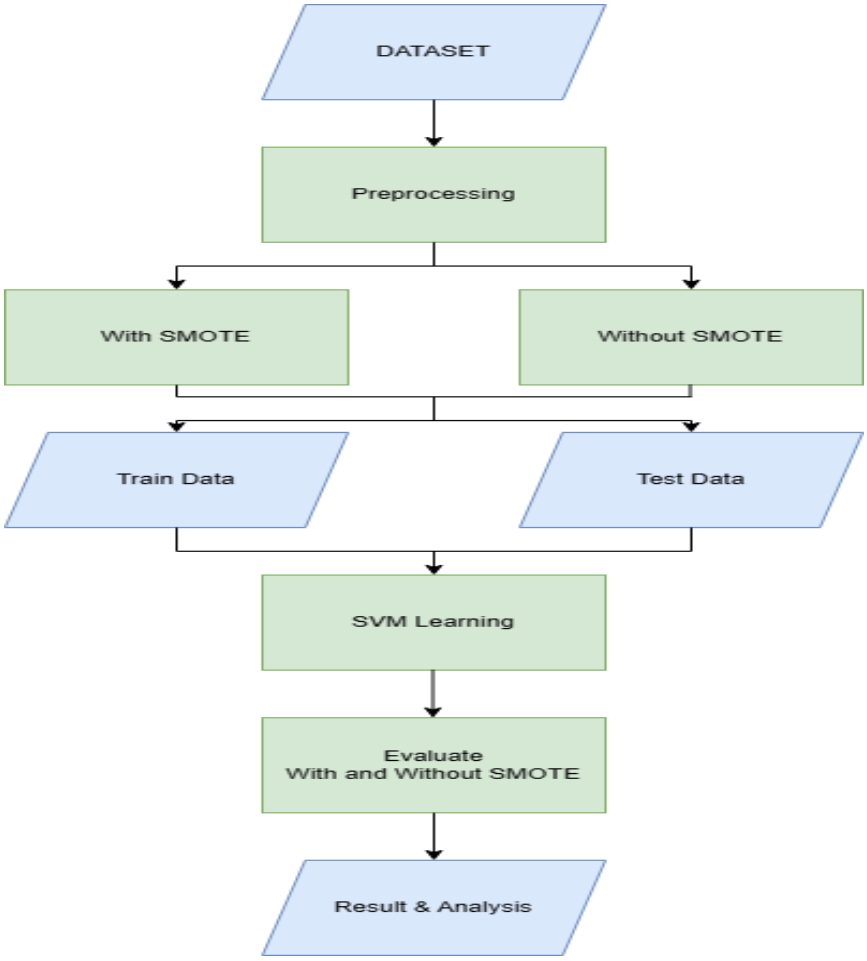

Figure 1. Research methodology flowchart

A. Dataset

The data used for this research is secondary data, namely the credit report of BPR Bank NBP 16 Rantau Prapat, North Sumatra taken from January 2022 to December 2022. The variables used in this study can be seen in Table 1 (V14) between feasible and not feasible has been determined by the bank.

TABLE I.
VARIABLE DESCRIPTION

Attribute	Description	Variable
Intended use	The reason why the debtor wants to apply for credit	V1
Plafond	Maximum amount of credit that can be received by the debtor	V2
Marital status	Debtor is married or not	V3
Number of dependents	Number of people for whom the debtor is responsible	V4
Gender	Male or Female	V5
Education	Last education of the debtor	V6
Age	Age of the debtor at the time of applying for credit	V7
Loan term	The length of time the debtor needs to repay the loan provided by the creditor	V8
Interest (%)	The interest rate of the proposed credit or additional fees that must be paid by the debtor	V9
Collateral	Goods that guarantee credit repayment	V10
Appraised value (NT)	Collateral value of the goods submitted by the debtor	V11
Income	The amount of the debtor's income	V12
Work	The work done by the debtor	V13
Target	Feasible prediction target=1 or not feasible=0	V14

B. Preprocessing

Data preprocessing is one of the most important steps in ML and can affect the efficiency and accuracy of the resulting model. Choosing the right preprocessing method can improve data quality and speed up the model-building process [14]. Some of the preprocessing techniques carried out in this study, namely

1. Looking at missing values, The purpose is to check whether there are missing or empty values in the dataset. This is important because the presence of missing values in the data can lead to inaccurate analysis results and even errors in decision-making.
2. Data exploration and visualization, to gain insight and understand the data better, to identify patterns, relationships, and outliers, and to generate hypotheses for further investigation.
3. Categorical data encoding is a technique to convert categorical variables into numerical variables in order to be processed by machine learning models. The technique used is One-Hot Encoding.
4. Standardization or normalization, aims to transform numeric data to have a uniform or comparable scale, making it easier for data analysis and modeling. The technique used is a robust scaler.

C. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is utilized to overcome the class imbalance in datasets by generating synthetic samples in minority classes. The steps of the SMOTE method [15], namely:

1. For each sample x_i in the minority class dataset X, calculate the Euclidean distance of that sample to all other samples in the dataset, and select its k nearest neighbors, denoted as y_j (j = 1, 2, ..., k).

2. The oversampling rate is determined based on the data inequality ratio to determine the sample enlargement. For each sample x_i, n numbers are randomly selected from their nearest neighbors, and new data can be constructed using the following formula:

$$x_{new} = x_i + rand(0,1) \times \left(y_{j\_}x_i\right) \quad (1)$$

x_j= 1, 2, ..., n, and rand(0,1) represent a random number between 0 and 1.

Briefly, the SMOTE process works by randomly selecting one data from the minority class x_i, and then selecting another data y_j from a different minority class than. Then, a synthetic data x_{new} is created by adding some of the differences between y_j and x_i, multiplied by a random number [9].

D. Support Vector Machine (SVM)

SVM is a classification method used in machine learning to create a model that can distinguish between two or more classes [4]. SVM is used in various fields such as face recognition, speech recognition, pattern recognition, genetic data analysis, and medical data analysis. SVM tries to find the best line or hyperplane to separate two different classes.

SVM has advantages and disadvantages [16], among others:
Advantages

1. SVM is able to process non-linear data by using kernel tricks.
2. SVM is very time-efficient in training and testing models.
3. SVM can handle data with many features without overfitting.

Weaknesses

1. SVM is very sensitive in choosing the right parameters and kernel, so it needs to be adjusted to achieve optimal results.
2. SVM has poor performance on data with multiple classes, as the model is only able to create a dividing line for two classes.
3. SVM also performs poorly on data with many outliers, as the algorithm only focuses on the most important points.

Next, SVM will perform hyperparameter tuning in Table II. There are several parameters that can be used to get the best SVM model [17], such as

1. Kernel is a parameter used to transform data into a higher dimensional space. There are several kernels in SVM, namely Linear Kernel for linearly separable data and Radial Basis Function kernel used for data transformation into infinite dimensional space.
2. C is a parameter used to control the balance between smooth decision boundaries and correct training samples.
3. Gamma is a control parameter in the non-linear SVM model that regulates the level of variation or complexity of the model.

TABLE II.
HYPERPARAMETER SVM

Model	Hyperparameter	Scale
Support Vector Machine	Kernel	linear, RBF
	C	0.1, 1, 10, 100
	Gamma	0.1, 1, 10, 100

E. Model Evaluation

Confusion matrix is a tool used to evaluate the model in this study The confusion matrix is a table used to evaluate the performance of the classification model described in Table III. This matrix shows how the model predicts the class of data [18].

TABLE III
CONFUSION MATRIX

		Predicted Class	
		+	-
Actual Class	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

Several model evaluation calculations are obtained [19, 20], such as:

1. Accuracy is the proportion of a classification predicting a condition or the ratio between the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP + TN}{N} \quad (2)$$

2. Sensitivity is the ability of the model to predict positive cases. Also commonly known as recall, sensitivity is the ratio between the number of true positives and the number of total positives.

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

3. Precision is a measure of model performance evaluation that measures how many positive outcomes are correctly identified by the model.

$$Specificity = \frac{TP}{TP + FP} \quad (4)$$

4. F-Score, a metric that takes into account precision and sensitivity in the same way or the ratio between the Harmonic mean of Precision and Recall, by combining information from precision and recall into one unified score.

$$F\text{-}Score = \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (5)$$

5. Area Under the Curve (AUC), is a measure used to compare different classifiers. One of the metrics used to measure the performance of binary classification models. AUC takes into account the model's ability to distinguish between two classes (positive and negative) and plots the classification results on a ROC (Receiver Operating Characteristic) graph. The ROC graph shows the relationship between the sensitivity (True Positive Rate) and Specificity (True Negative Rate) of the classification model.

$$\text{AUC} = \frac{Sensitivity * Specificity}{2} \quad (6)$$

IV. RESULTS AND DISCUSSION

A. Preprocessing Result

- Missing value checking shows that this dataset consists of 620 rows and 14 columns. The data types contained in this dataset include object, integer, and float. Each column has a non-null count that indicates there is no empty data in the column. This information is important for further data processing and analysis.

- Data description, then an explanation of the numerical data data is explained in Table IV, such as

    1. Plafond, shows considerable variation, with a minimum value of 1,200,000 and a maximum value of 1,000,000,000. The high standard deviation also indicates a significant difference between the values in this variable. It should be noted that the values in this variable have a large scale.
    2. Number of Dependents, shows fairly limited variation, with a minimum value of 0 and a maximum value of 6. The relatively low standard deviation indicates that most of the data tends to fall within a narrower range.
    3. Age, shows a fairly wide variation, with a range of values between 20 and 63 years. The relatively low standard deviation indicates that most of the data tends to fall within a range closer to the mean.
    4. Loan term, shows a fairly wide variation, with values ranging from 6 to 48 months. The relatively high standard deviation indicates that there are significant differences between the values in this variable.
    5. Interest, shows a fairly wide variation, with values ranging from 6.25 to 42. The relatively high standard deviation indicates a significant difference between the values in this variable.
    6. Appraised Value, shows considerable variation, with a minimum value of 400,000 and a maximum value of 1,580,000,000. The high standard deviation also indicates a significant difference between the values in this variable. It should be noted that the values in this variable also have a large scale.
    7. Income, shows considerable variation, with a minimum value of 1,100,000 and a maximum value of 16,918,600. The high standard deviation also indicates a significant difference between the values in this variable. It should be noted that the values in this variable have a large scale.

TABLE IV
NUMERICAL DATA DESCRIPTION

Var Num	Count	Mean	Stand. Deviation	Min	Max
Plafond	620	52.219.677,41	84.237.938,25	1.200.000	1.000.000.000
Number of Dependents	620	3,95	1,31	0	6
Age	620	40,6	8,59	20	63
Loan Term	620	27,43	11,33	6	48
Interest	620	21,53	6,39	6.25	42
Appraised Value	620	124.606.854,83	169.082.518,17	400.000	1.580.000.000
Income	620	6.670.244,51	3.037.203,41	1.100.000	16.918.600

- Data exploration and visualization
  In this process, outliers in the numerical data are checked. In Figure 2, the data for the Number of Dependents and Loan Term variables have no outliers as seen from the absence of dots on the box-plot graph. The Plafond, Age, Appraisal Value, and Income variables have quite a lot of outliers scattered in the data, while the Interest variable does not have a lot of outliers..

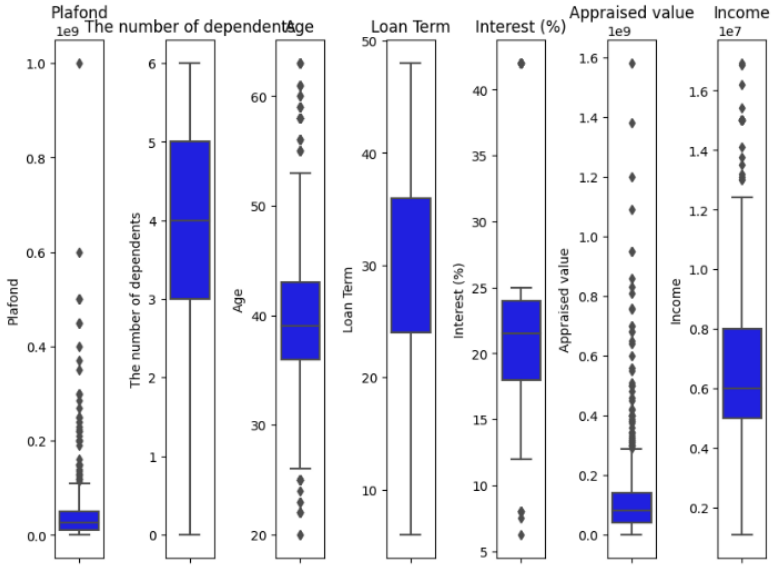

Figure 2. Distribution of Numerical Data Outliers

- The numerical data visualization in Figure 3 is carried out to be able to help in understanding the patterns and characteristics of the data used. In the Plafond variable, the data accumulates the most from 0-200 million, while the Number of Dependents is around 5-6 people, for Age the data accumulates around 35-41 years, then for the Term the data accumulates around 20-28 months and 32-40 months, then the variable Interest data accumulates around 20-25%, after that the Estimated Value of the data also ranges from 0-200 million, for Income the data is stockpiled around 5-10 million, finally the Purpose of the data is most feasible to be given credit than not.

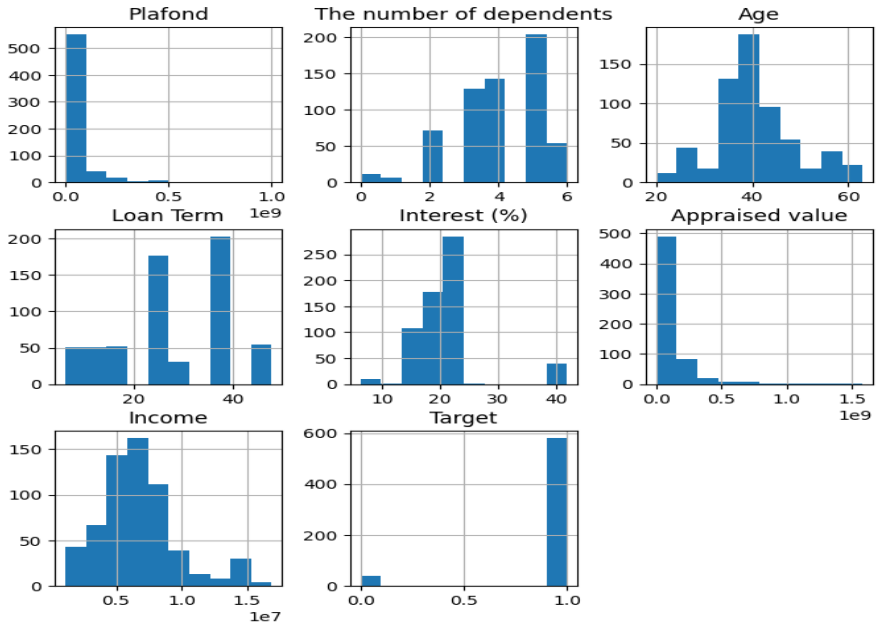

Figure 3. Numerical Data Visualization

- Encoding categorical data, After performing One-Hot Encoding (OHE) on categorical variables consisting of ['Inteded Use', 'Marital Status', 'Gender', 'Education', 'Insurance', 'Employment'], the results generated new columns that represent unique combinations of values in each categorical variable. Here is an example of some of the columns in Table V generated after OHE: 'Intended use_Increase business capital': This column indicates if the intended use of the credit request is to increase business capital. If yes, the value will be 1, if no, the value will be 0, as well as other OHE outcome categorical variables.

TABLE V.
CATEGORICAL DATA ENCODING RESULTS

No.	Intended use_Add business capital	Work_mechanical	...	Work_trader
1.	0	0	...	0
2.	0	0	...	0
3.	0	0	...	1

- Standardization, After performing robust scaling on numeric variables, in Table VI it can be confirmed that the scale of the numeric variables has been changed to a scale that is more in line with the data distribution and has better resistance or stability to outliers in the data. For example, the Ceiling values have been changed to remove the impact of outliers and make the distribution of values more consistent. By using Robust Scaler, values that are very far from the median will be transformed so as to improve the interpretation of the model, especially if there are some extreme values.

TABLE VI.
NUMERICAL DATA STANDARDIZATION

No.	Plafond	The number of dependents	Age	Loan Term	Interest	Appraised value	Income
1.	0.03125	0.0	2.714286	1.0	-0.25	-0.3	-0.3
2.	-0.28125	0.5	0.285714	1.0	0.08	-0.5	-0.8
…	…	….	…	…	…	…	…
620.	4.34375	-0.5	-0.571429	-1.5	0.58	2.12	2.0

B. SMOTE

Subsequent to the application of SMOTE, there was a significant increase in the number of minority class samples (class 0) after the application of the SMOTE method which can be seen in Table VII. This can be helpful, because

1. Before SMOTE, the data had a significant imbalance between the number of majority and minority class samples, with the majority class (1) dominating the data. After SMOTE, the number of minority class samples increased significantly, achieving a better level of balance between the two classes.
2. With the increase in the number of minority class samples after SMOTE, the model to be built has better access to the data from the minority class. This can improve the model's ability to recognize patterns and learn characteristics of the minority class, which in turn can improve the performance and prediction accuracy of the class.

TABLE VII.
SMOTE RESULTS

Before SMOTE		**After SMOTE**	
Class	**Total**	**Class**	**Total**
0	40	0	580
1	580	1	580

C. SVM and Hyperparameter Tuning

In the next stage, modeling using the selected ML is SVM. Then, data that has undergone preprocessing is divided into training data (75%) and test data (25%). Based on the results of the best SVM parameter tuning without and with SMOTE, it is shown in Table VIII.

TABLE VIII.
SELECTED SVM PARAMETERS

Model	Hyperparameter	Scale
SVM Without SMOTE	C	1
	Gamma	0.1
	Kernel	Linear
SVM With SMOTE	C	1
	Gamma	0.1
	Kernel	RBF

Then, to ensure that the model produces the best parameters, hyperparameter tuning is performed using a grid search technique with 10-fold cross-validation to ensure that the data is not overfitting so that optimal analysis is obtained. The accuracy value of SVM without SMOTE ranges from 0.91 to 0.96, with an average of about 0.94. SVM without SMOTE shows a relatively stable performance with a not-too-large variation in accuracy value between different folds. The highest performance of SVM occurs in folds 2, 3, 6, and 7 with an accuracy of 0.96, while the lowest performance occurs in fold 4 with an accuracy of 0.91. Meanwhile, the accuracy value of SVM with SMOTE ranges from 0.93 to 0.99, with an average of about 0.96. SVM with SMOTE shows a relatively stable performance with a not-too-large variation in accuracy value between different folds. The highest performance of SVM occurs on fold 8 with an accuracy of 0.99, while the lowest performance occurs on fold 1 with an accuracy of 0.93.

TABLE IX.
ACCURACY BASED ON 10-FOLD CROSS-VALIDATION

Fold	1	2	3	4	5	6	7	8	9	10
SVM Without SMOTE	0.94	0.96	0.96	0.91	0.91	0.96	0.96	0.98	0.93	0.93
SVM With SMOTE	0.93	0.95	0.94	0.97	0.94	0.97	0.98	0.99	0.98	0.95

D. Model Evaluation

Finally, the results of the model evaluation comparison can be seen in Table X based on accuracy, F-1 Score, recall, precision, and AUC performed.

1. The SVM model with SMOTE has a higher accuracy (0.96) than the model without SMOTE (0.94). This shows that the model with SMOTE is able to classify the data 96% better overall.
2. F1-Score is a measure of the balance between precision and recall. Both models have the same F1-Score (0.97), indicating that 97% of the balance between precision and recall is almost similar.
3. The model with SMOTE has a high recall (0.97), meaning that the model's ability to identify samples from minority classes is very good. Whereas the model without SMOTE has a higher recall (0.98), indicating that it is 98% better at identifying samples from minority classes.
4. Precision is almost the same for both models (0.96). This indicates that both models have 96% ability to reduce false positives (misclassification of minority classes) to a similar extent.
5. The model with SMOTE has a much better AUC (0.96) compared to the model without SMOTE (0.56), indicating that the model with SMOTE performs 96% better in distinguishing between classes.

TABLE X.
ACCURACY COMPARISON OF EVALUATION MODEL

Model Evaluation	Accuracy	F-1 Score	Recall	Precision	AUC
SVM without SMOTE	0.94	0.97	0.98	0.96	0.56
SVM with SMOTE	0.96	0.97	0.97	0.96	0.96

V. CONCLUSION

Based on the results and discussion in the previous section, one hot-encoding has converted the categorical variables into 0 and 1, and the robust scaler has standardized the numerical data for easier understanding of the model. SMOTE is able to balance target data that was previously much different by replicating. The SVM model with SMOTE performed better in terms of accuracy, AUC, and overall classification ability. However, the SVM model without SMOTE has a higher recall, which indicates a better ability to identify samples from minority classes. By applying SMOTE, the SVM model will be better at handling unbalanced datasets and can provide more accurate predictions of loan risk. A higher AUC also indicates an improved ability of the model to separate risk classes, which can be interpreted as a reduction in loan risk.

REFERENCES

[1] Bambang Sudiyatno. (2013). Pengaruh Risiko Kredit Dan Efisiensi Operasional Terhadap Kinerja Bank (Studi Empirik pada Bank yang Terdaftar di Bursa Efek Indonesia). *Jurnal Organisasi Dan Manajemen*, *9*(1), 73–86. https://doi.org/10.33830/jom.v9i1.39.2013

[2] Yang, P.-A. F.-F., Kelancaran, M., Nurani, P., Syari'ati Pramono, N., Manajemen, D., Ekonomi, F., Manajemen, D., Pertanian, I., Kampus, B., Bogor, D., & Permanasari, Y. (2016). Analisis Faktor-faktor yang Memengaruhi Kelancaran Kredit dan Penilaian Kesehatan Keuangan pada Amartha Microfinance. *Jurnal Manajemen Dan Organisasi*, *7*(1), 1–16. https://doi.org/10.29244/JMO.V7I1.14065

[3] Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). ScienceDirect ScienceDirect A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

[4] Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning second edition*. MIT Press

[5] Kubat, M. (2021). An Introduction to Machine Learning. *An Introduction to Machine Learning*, 1–458. https://doi.org/10.1007/978-3-030-81935-4/COVER

[6] Nurachim, R. I. (2019). Pemilihan Model Prediksi Indeks Harga Saham Yang Dikembangkan Berdasarkan Algoritma Support Vector Machine(Svm) Atau Multilayer Perceptron(Mlp) Studi Kasus : Saham Pt Telekomunikasi Indonesia Tbk. *Jurnal Teknologi Informatika & Komputer |*, *5*(1).

[7] Lusiyanti, D., & Nacong, D. N. (2018). Sistem Sederhana Untuk Memprediksi Risiko Pemberian Kredit. *Jurnal Ilmiah Matematika Dan Terapan*, *15*(2), 248–255. https://doi.org/10.22487/2540766X.2018.V15.I2.11360

[8] Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

[9] Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). *Credit risk prediction in an imbalanced social lending environment*. Kubat, M. (2021). An Introduction to Machine Learning. *An Introduction to Machine Learning*, 1–458. https://doi.org/10.1007/978-3-030-81935-4/COVER

[10] Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 677–690. https://doi.org/10.30812/matrik.v21i3.1726

[11] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*, *8*, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

[12] Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit Risk Model Based on Central Bank Credit Registry Data. *Journal of Risk and Financial Management*, *14*(3). https://doi.org/10.3390/jrfm14030138

[13] Boiko Ferreira, L. E., Barddal, J. P., Gomes, H. M., & Enembreck, F. (2018). Improving credit risk

prediction in online peer-To-peer (P2P) lending using imbalanced learning techniques. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2017-November*, 175–181. https://doi.org/10.1109/ICTAI.2017.00037

[14] Frye, M., Mohren, J., & Schmitt, R. H. (2021). Benchmarking of Data Preprocessing Methods for Machine Learning-Applications in Production. *Procedia CIRP*, *104*, 50–55. https://doi.org/10.1016/j.procir.2021.11.009

[15] Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A., & Yao, Y. (2020). Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. *Energies*, *13*(8). https://doi.org/10.3390/en13082039

[16] Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). *Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1*. http://asnugroho.net

[17] Tran, N., Schneider, J. G., Weber, I., & Qin, A. K. (2020). Hyper-parameter optimization in classification: To-do or not-to-do. *Pattern Recognition*, *103*, 107245. https://doi.org/10.1016/J.PATCOG.2020.107245

[18] M, H., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

[19] Valero-Carreras, D., Alcaraz, J., & Landete, M. (2022). *Comparing two SVM models through different metrics based on the confusion matrix*. https://doi.org/10.1016/j.cor.2022.106131

[20] Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *5*(2), 697–711. http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369