Data Preprocessing to Improve Accuracy in Classification Methods (Case Study: Credit Risk Analysis Dataset Classification)

Baiq Nurul Azmi
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Jl. Siliwangi, Jombor Lor, Kec. Mlati, Kab. Sleman, Daerah Istimewa Yogyakarta 55285

Arief Hermawan
Master Program of Information Technology
Universitas Teknologi Yogyakarta
Jl. Siliwangi, Jombor Lor, Kec. Mlati, Kab. Sleman, Daerah Istimewa Yogyakarta 55285

ABSTRACT

This research analyzes the use of various data pre-processing methods in the context of credit risk analysis with Support Vector Machine (SVM) classification models. The background of this research details the complexity of challenges faced in the banking industry regarding credit risk evaluation and how data pre-processing to improve model accuracy. The research method includes four experimental scenarios that consider various combinations of data pre-processing methods. Each scenario is designed to evaluate the performance of SVM models on credit risk datasets. The method steps include data preparation, Missing Data handling with Remove Features for features that have more than 50% Missing Data rate and MICE imputation for features that have less than 50% Missing Data, feature selection based on Correlation Matrix to overcome High Dimensional Data, and data resampling with SMOTE to overcome class imbalance. The test results show that using a combination of data pre-processing methods can significantly improve the accuracy of SVM models on credit risk datasets. The highest accuracy is obtained in the pre-processing scenario when overcoming Missing Data with remove features and MICE imputation with a value of 99.4%.

General Terms

Machine Learning, Data Minning, Classification

Keywords

Credit Risk Analysis, MICE, SMOTE, Support Vector Machine, Correlation Matrix

1. INTRODUCTION

In the financial world, especially in the banking industry, the use of technology in decision making is increasingly important [1]. One of the uses of technology that can be done in the banking industry is to utilize data mining or data analytics [2]. One example of the application of data mining in the banking industry is credit risk analysis in credit assessment.

Credit scoring is necessary because in recent years there has been an increase in financial fraud due to the growth of technology and paradigms such as the e-commerce sector and financial technology (FinTech) [3]. In processing credit assessments, data mining can help obtain accurate and complete information about creditor profiles, so that it can be used as a basis for decision making [4].

Challenges in data processing are high data complexity due to data diversity or heterogeneity [5], inconsistent data, noise, Missing Data, High Dimensional Data, and data imbalance [6]. Challenges in data processing can be overcome by preparing data before processing and is commonly referred to as data preprocessing [7].

Data preprocessing involves a series of techniques to clean, transform, and prepare raw data before it is analyzed using classification methods [8]. The purpose of data pre-processing is to improve data quality, reduce noise, eliminate irrelevant data, and normalize variables so that they can be optimally used in the classification process so that accuracy results increase [9]. Data pre-processing is very important to do, as it will result in a decrease in classification performance if not done. Choosing the wrong pre-processing technique can lead to incorrect predictions.

There have been many studies on data pre-processing, but those that discuss credit data, especially the Credit Risk Analysis Dataset, are still few and the classification results are not optimal as in the research by [10] which discusses credit risk classification by handling data balance with 64% accuracy results and has not discussed how the combined results of several pre-processing techniques affect the accuracy value of the classification method.

This research will apply the Removing Features pre-processing technique to overcome Missing Data with more than 50% missing data and Multivariate Imputation by Chained Equation (MICE) to overcome low Missing Data, because from previous research MICE can be used to overcome Missing Data on numeric and categorical data [11] besides that compared to several other methods MICE has the smallest error [12].

The feature selection pre-processing technique with Correlation Matrix will be used to overcome High Dimensional Data because in research by [13] and [14] with feature selection successfully reducing nuisance features and increasing accuracy values. SMOTE pre-processing technique will be used to overcome imbalanced data as in previous research by [15] stated that SVM-SMOTE got the best accuracy.

These techniques are used because based on several previous studies that have been mentioned each has the best results. Research will be conducted by comparing the classification results of models formed from pre-processing handling Missing Data, with pre-processing handling High Dimensional Data, with pre-processing handling imbalanced data, and by combining the three. Each classification process will be pre-processed for Min-max normalization on each attribute. The classification method used is Support Vector Machine (SVM) because from several previous studies that discuss pre-

processing before classification, the SVM algorithm has the best results.

2. LITERATURE REVIEW

2.1 Previous Study

Data pre-processing is a very important stage in data mining [16]. In credit risk analysis data, pre-processing is very important to improve the accuracy of classification results on borrowers who are considered as problem loans or not. Efforts to improve accuracy through the pre-processing stage can be done by dealing with Missing Data, High Dimensional Data, and imbalanced data.

Research that discusses Missing Data in credit data is research by [10] where the research uses the median technique to fill in Missing Data. This research also overcomes challenges for data imbalance with several methods, namely SMOTE, Random Oversampling (ROS), and Randon Under sampling (RUS). The classification methods used are Random Forest (RF), Logistic Regression (LR), and Multilayer Perceptron (MLP). The classification results show that the highest accuracy value is in the RF-RUS algorithm with an accuracy of 64.00%.

High dimensional data can be overcome by feature selection and feature extraction [17]. Research by [18] discussed credit score classification with the pre-processing done is feature selection. The study proposed a new approach to feature selection by combining results from five different algorithms and using a new voting method called "if_any". The results showed the proposed hybrid ensemble credit score model based on combining five feature selection algorithms combined with three different types of voting and eight ensemble models combined with a soft voting approach outperformed other models in terms of accuracy, type I error rate, sensitivity, and F-measure. The accuracy of this model is 82.015%.

Imbalanced class is one of the main problems that arise in anomaly detection on real time datasets. A dataset is considered unbalanced if one of its classes has a very large dominance compared to other classes [19].

Research about imbalance data in credit data by [20] which analyzes resampling methods (Random Under sampling, Random Oversampling, and SMOTE) with decision tree algorithms and SVM for credit card fraud detection with unbalanced datasets where data with legitimate transactions is more than data with unauthorized transactions, so that the most appropriate resampling method can be found for credit card fraud datasets. The results show that the SMOTE method has the best results in classification using SVM and the quality of the dataset is an important factor in classification performance. Research by [21] which applies the SMOTE method for resampling Credit Card Fraud data and evaluated using machine learning methods such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Decision Tree (DT), and Extra Tree (ET), then the results show that the use of AdaBoost has a positive impact on the performance of the proposed method and provides better results compared to existing methods.

Data imbalance can be resolved by resampling methods. The SMOTE resampling method is one method that is often used to overcome data imbalance. SMOTE (Synthetic Minority Over-sampling Technique) can handle unbalanced data by replicating minority data, the result of which is known as synthetic data. In numerical data, SMOTE will work by finding the k nearest neighbors for each data in the minority class using Euclidean distance [22]

Research by [23] SMOTE was used to handle class imbalance in the classification of diabetes, SMOTE can also be used to overcome class imbalance in the classification of Television Advertisement Performance Rating [24]. Both studies show that using SMOTE can overcome the problem of data imbalance and get good classification results, where in [23] the use of SMOTE with the SVM classification method has the best accuracy and sensitivity, which is 82%, while in [24] ANN performance with SMOTE achieved an accuracy of 87.06% compared to ANN without SMOTE which was only 86.35%.

2.2 Multivariate Imputation by Chained Equation (MICE)

Multivariate Imputation by Chained Equation is one of the multiple imputation techniques that can be used to handle missing data. MICE is based on filling the missing data on a variable-by-variable basis and imputing the missing data multiple times in the dataset through an iterative procedure [25]. MICE start by filling in missing values in numerical data using Linear Regression algorithm, while for categorical data filling in Logistic Regression [11]. Then, the missing values are imputed sequentially based on the columns in the data or sequentially according to predefined settings.

2.3 Correlation Matrix

Correlation Matrix is a matrix that shows the level of correlation between variables in a datasets [26]. The correlation value ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no linear correlation between the two variables [27]. The correlation matrix is obtained by calculating the correlation between each pair of variables in the dataset. The matrix is a square of size n x n, where n is the number of variables in the dataset. The main diagonal of the correlation matrix contains the value 1, because the correlation between a variable and itself is one.

2.4 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a resampling technique that generates synthetic data on minority classes based on a combination of existing data. To perform oversampling in SMOTE, the algorithm will take instances from the minority class and find the k-nearest neighbor of each instance, then generate synthetic instances instead of replicating the minority class instances. This avoids the problem of overfitting [28].

2.5 Support Vector Machines

Support Vector Machines (SVM) are machine learning algorithms used to analyze data and classify or regress data in a supervised manner. It seeks an optimal hyperplane in the classification space with the highest margin between different classes and has been shown to provide better results than other classifiers [29]. SVMs are an extension of support vector classification and are obtained by specifically expanding the feature space using a kernel. This allows SVMs to handle non-linearly separable problems well and can perform well in larger feature spaces [30].

3. RESEARCH METHODOLOGY

The research methodology used in writing this research can be seen in Fig 1.

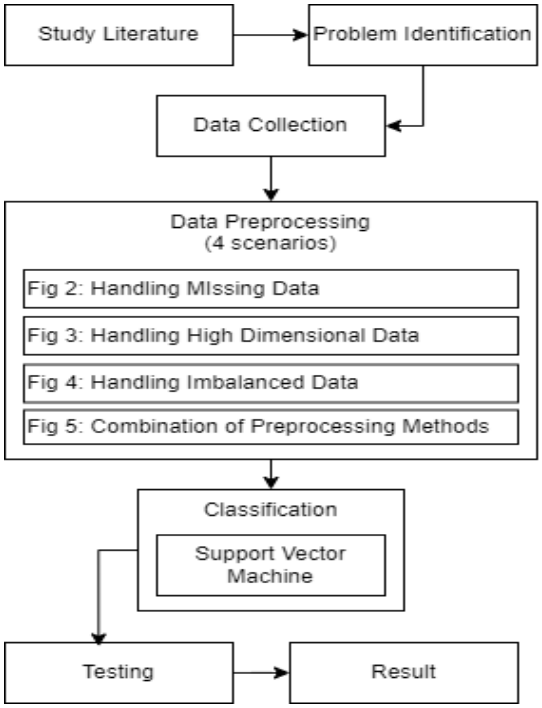

Fig 1: Research Stages

3.1 Literature Study

Research conducted by studying literature including journals and materials related to the research topic. By conducting a literature study, journals were obtained regarding the classification of credit analysis data, pre-processing data on a dataset and machine learning classification algorithms.

3.2 Problem Identification

This process looks for problems that occur related to the right data pre-processing techniques to increase the accuracy value of credit analysis data and the implications if these problems are not immediately resolved.

3.3 Data Collection

This research uses secondary data from the Kaggle dataset collection site entitled Credit Risk Analysis [31]. The Credit Risk Analysis Dataset contains complete loan data for all loans granted from 2007 to 2015. The dataset consists of 855,969 data rows and 73 attributes including the target variable. This dataset has various types of features such as categorical and numerical.

3.4 Data Preprocessing

Data preprocessing is done after collecting data. Four scenarios of the preprocessing stage were carried out, namely, the first scenario handling missing data (can be seen in Fig 2), the second scenario handling missing data and handling high dimensional data data (can be seen in Fig 3), the third scenario handling missing data and handling imbalanced datasets data (can be seen in Fig 4), and the fourth scenario handling missing data, handling high dimensional data and handling imbalanced datasets data (can be seen in Fig 5).

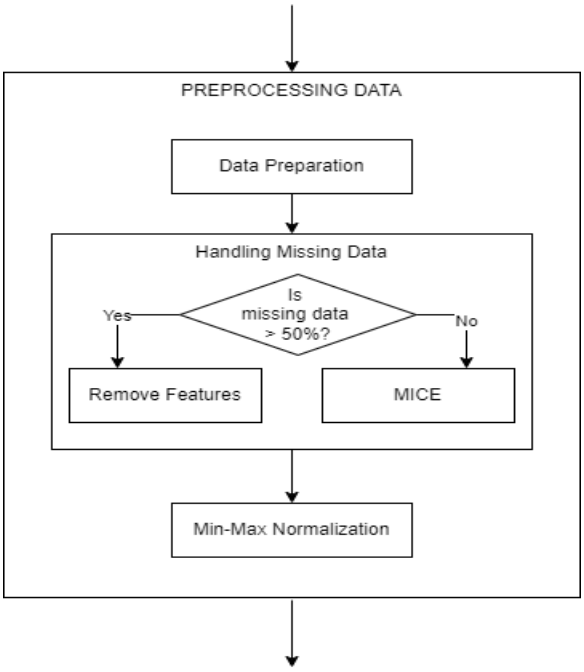

Fig 2: Preprocessing - Handling Missing Data

Fig 2 shows the preprocessing steps for handling missing data. First, data preparation is performed, then the missing data handling method is selected, and then normalization is performed.

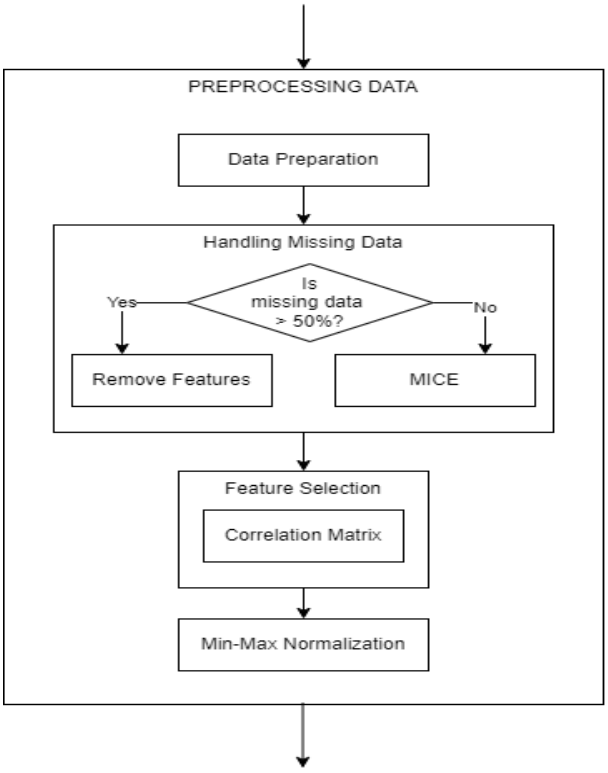

Fig 3: Preprocessing - Handling High Dimensional Data

Fig 3 shows the preprocessing steps for handling high dimensional data. First, the data is prepared, then the method of handling missing data is selected, then the features are selected based on the correlation matrix and normalization is performed.

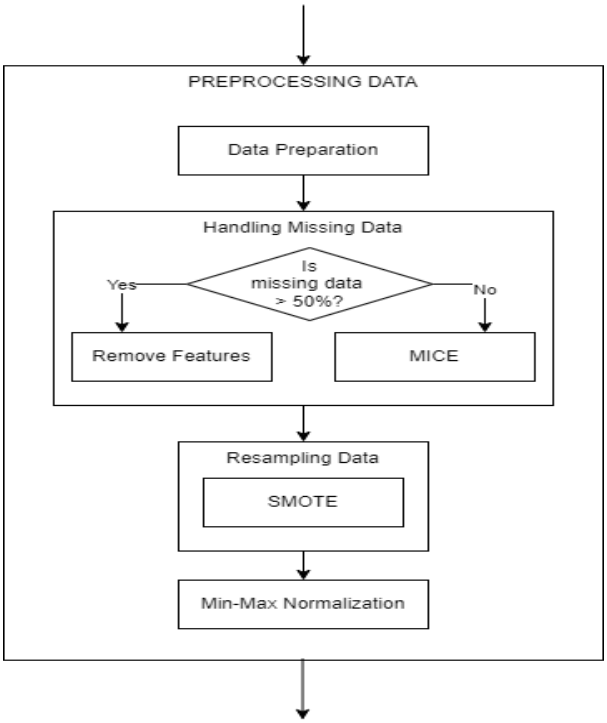

Fig 4: Preprocessing - Handling Imbalanced Data

Fig 4 shows the preprocessing steps for handling imbalanced data. First, data preparation is carried out, then selecting the method of handling missing data, after that resampling the data with the SMOTE method and then normalizing.

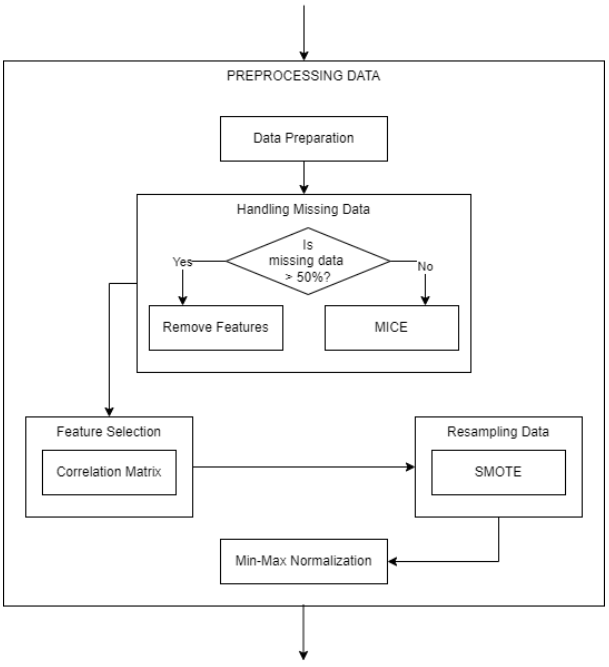

Fig 5: Combination of Preprocessing Methods

Fig 5 shows the preprocessing steps for handling missing data, high dimensional data, and imbalanced data. First, data preparation is carried out, then the method of handling missing data is selected, after that the feature selection is based on the correlation matrix, the resulting data will be resampled with the SMOTE method, and the last step will be normalized.

The following is an explanation of each type of preprocessing performed:

3.4.1 Data Preparation

Data preparation is done by labeling categorical data that is still in the form of strings so that it can be processed using the Support Vector Machine classification algorithm. Other data preparation is also done by selecting attributes that will be removed because they are unique and irrelevant to the analysis process and do not have predictive value for Credit Risk Analysis classification such as id, zip code, address code and so on.

3.4.2 Missing Data Handling

Remove Features will be performed in this study if there is Missing Data with more than 50% of the features, while data imputation in this study is used to overcome Missing Data with less than 50% of the features. The MICE technique is used in this study to fill in data with numerical and categorical types.

3.4.3 Feature Selection

This research uses the Correlation Matrix method to analyze and identify the most relevant features that have a significant impact on the classification of credit data. Features that have a high correlation will be removed.

3.4.4 Data Resampling

Data resampling aims to handle class imbalance in credit data. The resampling method or technique that will be used is SMOTE.

3.4.5 Min-max Normalization

This research uses in-Max Normalization to rescale numeric features into a specific range where the minimum value of the feature is mapped to 0 and maximum value is mapped to 1.

3.5 Classification Process

The classification method used in this research is Support Vector Machine. The classification process is carried out with four scenarios, namely the classification process after handling missing values, the classification process after handling missing values and feature selection, the classification process after handling missing values and resampling data, and the classification process after handling missing values, feature selection, and resampling data.

3.6 Testing

After the classification process is carried out, the next step is to test the classification model obtained. Testing is done with Confusion Matrix to see the value of accuracy, precision, and recall.

4. RESULTS AND DISCUSSION

4.1 Data preparation results

Data preparation is done by labeling categorical data that is still in the form of strings. This is done because SVM classification cannot be processed when there are still data types other than numbers. Another data preparation is to remove unique attributes that are not relevant to the Credit Risk Analysis classification process. The attributes that are removed are 'id', 'member_id', 'emp_title', 'desc', 'title', 'zip_code', 'addr_state', 'policy_code', 'issue_d', 'earliest_cr_line', 'last_pymnt_d', 'next_pymnt_d', 'last_credit_pull_d'. The number of attributes in the Credit Risk Analysis Dataset is 73 attributes, and as many as 13 attributes are removed because they are unique and

irrelevant for the classification of this research, so that 60 attributes remain for the next process.

4.2 SVM Classification After Handling Missing Data

Missing data handling in this study is divided into two, namely Remove Features to handle Missing Data that is more than 50% and data imputation with MICE technique to handle Missing Data that is less than 50%. The attributes to be processed at this stage amounted to 60 based on the results of the data preparation process. Attributes with more than 50% Missing Data in the dataset were found as many as 20 attributes, these attributes were removed from the dataset, so that of the 60 attributes of the data preparation results, 40 attributes remained. Missing data handling for attributes that have less than 50% missing data is done by imputation using the MICE technique. SVM classification after the Missing Data process was carried out with 5 different split data with 40 attributes and no missing data. SVM classification results after Missing Data handling can be seen in Table 1.

Table 1. SVM Classification Results After Missing Value Handling

Train:Test	Accuracy	Precision	Recall
85%:15%	0.9947811095	0.994770278	0.994781109
80%:20%	0.9944997900	0.994497498	0.994499790
75%:25%	0.9940492316	0.994050893	0.994049231
70%:30%	0.9937212858	0.993719476	0.993721285
65%:35%	0.9936272758	0.993631947	0.993627275

The graph in Fig 6 will illustrate the results based on Table 1.

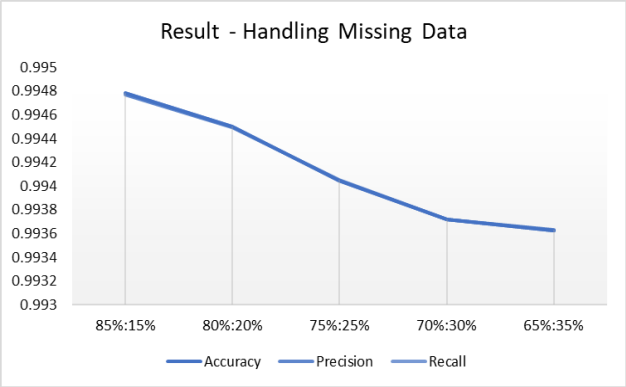

Fig 6: Result After Missing Value Handling

Table 1 and Fig 6 shows that the classification results show a very high level of accuracy, reaching a value of around 99.4% for various combinations of training and testing data, and has the highest accuracy in the combination of 85%:15% training and testing data with an accuracy of 99.478%. This shows that the preprocessing method used successfully improves the ability of the SVM model to classify credit risk data. The use of MICE (Multivariate Imputation by Chained Equation) to impute Missing Data values contributes positively to model accuracy.

4.3 SVM Classification After Handling High Dimensional Data

This process involves removing highly correlated features, with the aim of reducing the dimensionality of the data and retaining features that contribute significantly to the variability of the dataset. The dataset which initially consisted of 40 features as a result of Missing Data handling had several features with high correlation above 50%.

After the feature selection process based on Correlation Matrix, the 40 features were successfully reduced to 25 features. Data process was carried out with 5 different split data with 25 attributes and no missing data. SVM classification results after High Dimensional Data handling can be seen in Table 2.

Table 2. SVM Classification Results After High Dimensional Data Handling

Train:Test	Accuracy	Precision	Recall
85%:15%	0.9603798806	0.961974003	0.960379880
80%:20%	0.9603997946	0.961992343	0.960399794
75%:25%	0.9600553217	0.961675286	0.960055321
70%:30%	0.9603411447	0.961937900	0.960341144
65%:35%	0.9604701104	0.962056377	0.960470110

The graph in Fig 7 will illustrate the results based on Table 2.

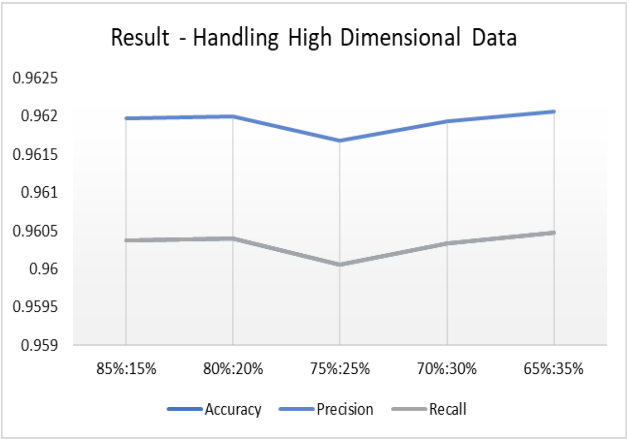

Fig 7: Result After High Dimensional Data Handling

The classification results in the second scenario show that although the resulting accuracy remains high, there is a 2-3% decrease compared to the first scenario, where the highest accuracy is 96.047% in the 65%:35% training and testing data combination. This decrease may be due to the reduction of features that could contain important information for classification. The lowest accuracy is found in the composition of training data and testing data 75%: 25% with an accuracy of 96.00%. However, the use of Correlation Matrix is effective in reducing the dimensionality of the data without significantly sacrificing performance.

4.4 SVM Classification After Handling Imbalanced Dataset

The imbalanced dataset in this study is handled with the use of SMOTE. This process produces a new dataset with the number of data rows almost doubled from the previous number of data

rows, where the number of datasets before pre-processing with SMOTE is 855,969 data rows, and after processing with SMOTE the number of data rows becomes 1,619,004 data rows. SVM classification results after Imbalanced Dataset handling can be seen in Table 3.

Table 3. SVM Classification Results Imbalanced Dataset Handling

Train:Test	Accuracy	Precision	Recall
85%:15%	0.9687747962	0.968775124	0.96877479
80%:20%	0.9687473959	0.968747483	0.96874739
75%:25%	0.9690109272	0.969010994	0.96901092
70%:30%	0.9690845089	0.969084553	0.96908450
65%:35%	0.9691131547	0.969113165	0.96911315

The graph in Fig 8 will illustrate the results based on Table 3.

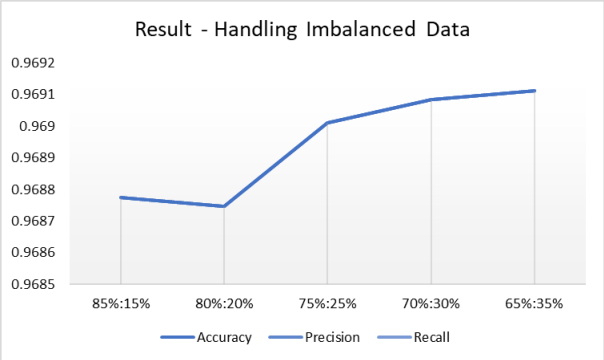

Fig 8: Result After Imbalanced Data Handling

The results in Table 3 show that the use of SMOTE successfully improves the accuracy of SVM models on imbalanced datasets. The resulting accuracy remains high and shows the stability of the model's performance against variations in training and testing data with the highest accuracy in the 65%:35% combination with a value of 96.911%. This step is important to avoid bias in the model due to class imbalance in the dataset by expanding the minority class, the model can train and produce better predictions for both classes.

4.5 SVM Classification After Missing Data, High Dimensional Data and Imbalanced Dataset

The combination of several preprocessing methods performed in this study are Removing Features for Missing Data, MICE imputation, feature selection using Correlation Matrix, and SMOTE to handle dataset imbalance. The preprocessing results involved 25 features and more than 1,619,004 rows of data. The next process involves classification using the SVM method with a variety of training and testing data. The SVM classification results after a combination of data preprocessing is performed are shown in Table 4.

Table 4. SVM Classification Results After Missing Data, High Dimensional Data, and Imbalanced Dataset Handling

Train:Test	Accuracy	Precision	Recall
85%:15%	0.8615157533	0.869649715	0.86151575
80%:20%	0.8614580182	0.869558142	0.86145801
75%:25%	0.8620047702	0.869730470	0.86200477
70%:30%	0.8617594477	0.869527645	0.86175944
65%:35%	0.8627672772	0.870675913	0.86276727

The graph in Fig 9 will illustrate the results based on Table 4.

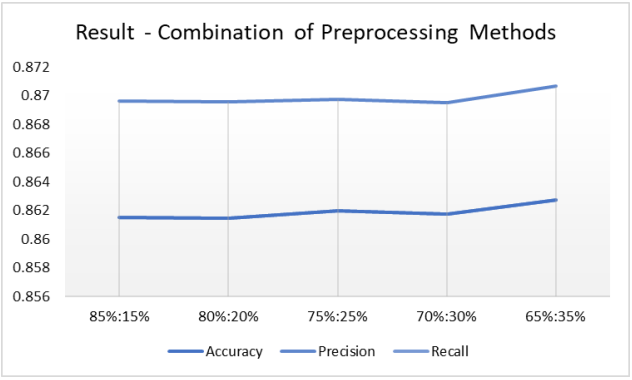

Fig 9: Result Of Combination of Preprocessing Methods

The classification results shown in Table 4 still show a high level of accuracy, even though the dataset becomes more complex after the application of some preprocessing methods. However, there is a decrease in accuracy compared to the previous scenario. This could be due to the additional complexity in the dataset after the preprocessing combination. Although the accuracy decreases compared to some of the previous scenarios, this preprocessing combination brings benefits in dealing with High Dimensional Data, Imbalanced Dataset and Missing Data problems.

The research was conducted by running four experimental scenarios involving various data preprocessing methods and the use of the Support Vector Machine (SVM) classification method on the Credit Risk Analysis dataset. From four scenarios, the highest accuracy obtained is around 99.4% while the lowest accuracy obtained is 86.14%. In previous research by [10] for the classification of Credit Risk Analysis using pre-processing carried out is the median technique for filling Missing Data. and overcoming challenges for data imbalance with several methods, namely SMOTE, Random Oversampling (ROS), and Randon Under sampling (RUS) resulted in the highest accuracy of 64.00%. This shows that the preprocessing step performed affects the resulting accuracy value.

In the first scenario, two stages of preprocessing were performed, involving the removal of features with more than 50% missing data and imputation using MICE for features with less than 50% missing data. The results showed very high accuracy, reaching around 99.4%. This success can be attributed to the management of Missing Data with Remove Features and the use of MICE to retain the necessary feature information. However, despite the high accuracy, other preprocessing should be done as the number of features is still very high and the classes are still very unbalanced, which may

cause bias towards the majority class, the need for feature selection processing to reduce dimensionality, and the need for an appropriate resampling method to deal with class imbalance.

The second scenario involves preprocessing with feature selection using Correlation Matrix. Data dimension reduction is done by removing features that have high correlation. Although the accuracy is still high, there is a decrease of about 2-3% compared to the first scenario with the highest accuracy of 96.04%. This shows that there is a trade-off between dimensionality reduction and preservation of important information, and that feature removal based on correlation alone does not always improve model performance.

In the third scenario, preprocessing was performed to handle dataset imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). The results show a significant improvement in the performance of the SVM model, especially in dealing with class imbalance. Accuracy remained high and stable across a variety of training and testing data with the highest accuracy of 96.91%. The application of SMOTE proved its effectiveness in improving model performance on imbalanced datasets.

The fourth scenario involved a combination of various preprocessing methods, namely Removing Features, MICE imputation, feature selection using Correlation Matrix, and the application of SMOTE. Although the dataset became more complex, the classification results showed an accuracy of about 86.14%-86.27%, which was a decrease from the previous scenario. This combination provides a trade-off between dimensionality reduction and model complexity.

The overall results showed that the first and third scenarios showed superiority in accuracy, with the third scenario being particularly effective in handling class imbalance. The highest accuracy was obtained in the scenario with Removing Features and MICE preprocessing to address Missing Data with an accuracy of 99.4%.

5. CONCLUSION

This research details the exploration of the use of various data pre-processing methods in the context of credit risk analysis using SVM classification models. The aspects considered involve data preparation, Missing Data handling, feature selection, and data resampling. This research provides a deeper understanding of how pre-processing stages affect model performance.

SVM classification is performed with four preprocessing scenarios namely Preprocessing with Removing Features and MICE Imputation, Preprocessing with Feature Selection using Correlation Matrix, Preprocessing with SMOTE to overcome Imbalanced Dataset and a combination of preprocessing Removing Features, MICE imputation, Feature Selection, and SMOTE. The experimental results show that using a combination of preprocessing methods can improve the accuracy of SVM models on credit risk datasets. The highest accuracy is obtained in the Preprocessing scenario with Removing Features and MICE Imputation with 99.4% accuracy. Although there is some trade-off between complexity and accuracy, these findings provide valuable insights to improve the effectiveness of credit risk analysis in the context of the banking industry.

6. REFERENCES

[1] A. Ansori, “Sistem Informasi Perbankan Syariah,” *J. Banq.*, vol. 4, no. 1, pp. 183–204, 2018.

[2] J. A. Ginting, “Data Mining Untuk Analisa Pengajuan Kredit Dengan Menggunakan Metode Logistik Regresi,” *J. Algoritm. Log. dan Komputasi*, vol. 2, no. 2, pp. 164–169, 2019, doi: 10.30813/j-alu.v2i2.1845.

[3] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, “Real-time Credit Card Fraud Detection Using Machine Learning,” *Proc. 13th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2023*, pp. 488–493, 2023.

[4] A. P. Nawary and Kurniati, “Penerapan Data Mining Dalam Memprediksi Kelancaran Kredit Nasabah Menggunakan Algoritma C4.5 (Studi Kasus Pada Pt. Astra International (Auto 2000 Plaju),” *Bina Darma Conf. Comput. Sci.*, vol. 5, pp. 1041–1047, 2021.

[5] A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera, “Handling incomplete heterogeneous data using VAEs,” *Pattern Recognit.*, vol. 107, no. 11, p. 107501, Nov. 2020, doi: 10.1016/j.patcog.2020.107501.

[6] H. Benhar, A. Idri, and J. L. Fernández-Alemán, “Data preprocessing for heart disease classification: A systematic literature review,” *Comput. Methods Programs Biomed.*, vol. 195, no. 10, p. 105635, Oct. 2020, doi: 10.1016/j.cmpb.2020.105635.

[7] B. Nugroho and A. Denih, “Perbandingan Kinerja Metode Pra-Pemrosesan Dalam Pengklasifikasian Otomatis Dokumen Paten,” *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 17, no. 2, pp. 381–387, 2020, doi: 10.33751/komputasi.v17i2.2148.

[8] E. Etriyanti, D. Syamsuar, and N. Kunang, “Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4 . 5 untuk Memprediksi Kelulusan Mahasiswa,” *Telematika*, vol. 13, no. 1, pp. 56–67, 2020, doi: http://dx.doi.org/10.35671/telematika.v13i1.881.

[9] A. P. Joshi and B. V Patel, “Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process,” *Orient. J. Comput. Sci. Technol.*, vol. 13, no. 0203, pp. 78–81, Jan. 2021, doi: 10.13005/ojcst13.0203.03.

[10] V. Moscato, A. Picariello, and G. Sperlí, “A benchmark of machine learning approaches for credit score prediction,” *Expert Syst. Appl.*, vol. 165, no. May 2020, p. 113986, 2021, doi: 10.1016/j.eswa.2020.113986.

[11] S. I. Khan and A. S. M. L. Hoque, “SICE: an improved missing data imputation technique,” *J. Big Data*, vol. 7, no. 1, p. 37, Dec. 2020, doi: 10.1186/s40537-020-00313-w.

[12] L. Li, C. G. Prato, and Y. Wang, “Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier,” *Accid. Anal. Prev.*, vol. 146, no. July, p. 105744, Oct. 2020, doi: 10.1016/j.aap.2020.105744.

[13] E. N. R. Khakim, A. Hermawan, and D. Avianto, “Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine,” *JIKO (Jurnal Inform. dan Komputer)*, vol. 7, no. 1, p. 158, 2023, doi: 10.26798/jiko.v7i1.771.

[14] A. Hermawan and A. P. Wibowo, "Implementasi Korelasi untuk Seleksi Fitur pada Klasifikasi Jamur Beracun Menggunakan Jaringan Syaraf Tiruan," *INTEK J. Inform. Dan …*, vol. 5, no. 1, pp. 63–67, 2022.

[15] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[16] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.

[17] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: a review," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3473–3515, Jun. 2021, doi: 10.1007/s10462-020-09928-0.

[18] J. Nalić, G. Martinović, and D. Žagar, "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers," *Adv. Eng. Informatics*, vol. 45, no. February 2019, p. 101130, 2020, doi: 10.1016/j.aei.2020.101130.

[19] T. Hapsari, R, K. Indriyani, "Implementasi Algoritma SMOTE Sebagai Penyelesaian Imbalance Hight Dimensional Datasets," in *Prosiding Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika*, 2022, pp. 427–432, doi: 10.31284/p.snestik.2022.2868.

[20] M. Anis, M. Ali, S. A. Mirza, and M. M. Munir, "Analysis of Resampling Techniques on Predictive Performance of Credit Card Classification," *Mod. Appl. Sci.*, vol. 14, no. 7, p. 92, 2020, doi: 10.5539/mas.v14n7p92.

[21] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[22] Islahulhaq, W. Wibowo, and I. D. Ratih, "Classification of non-performing financing using logistic regression and synthetic minority over-sampling technique-nominal continuous (SMOTE-NC)," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 115–128, 2021, doi: 10.15849/ijasca.211128.09.

[23] 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," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[24] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[25] R. K. Kim *et al.*, "Data integration of National Dose Registry and survey data using multivariate imputation by chained equations," *PLoS One*, vol. 17, no. 6, pp. 1–14, 2022, doi: 10.1371/journal.pone.0261534.

[26] R. Kurniawan, P. Pizaini, and F. Insani, "Penerapan Algoritma K-Means Clustering dan Correlation Matrix Untuk Menganalisis Risiko Penyebaran Demam Berdarah di Kota Pekanbaru," *JIMP (Jurnal Inform. Merdeka Pasuruan)*, vol. 6, no. 3, pp. 1–6, 2021, doi: http://dx.doi.org/10.37438/jimp.v6i3.353.

[27] J. Daemen and V. Rijmen, *The Design of Rijndael*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020.

[28] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 1, pp. 1–15, 2022, doi: 10.1109/TNNLS.2021.3136503.

[29] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *J. Data Anal. Inf. Process.*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[30] R. Dietrich, M. Opper, and H. Sompolinsky, "Statistical Mechanics of Support Vector Networks," *Phys. Rev. Lett.*, vol. 82, no. 14, pp. 2975–2978, Apr. 1999, doi: 10.1103/PhysRevLett.82.2975.

[31] R. Mehta, "Credit Risk Analysis," *kaggle.com*, 2021. [Online]. Available: https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis.