hAccredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 1 (2024) 54 - 61 ISSN Media Electronic: 2580-0760

Kmeans-SMOTE Integration for Handling Imbalance Data in Classifying Financial Distress Companies using SVM and Naïve Bayes

Didit Johar Maulana[1], Siti Saadah[2], Prasti Eko Yunanto[3]
[1,2,3]Department of Informatics, Informatics, Telkom University, Bandung, Indonesia
[1]djmaulana@student.telkomuniversity.ac.id, [2]sitisaadah@telkomuniversity.ac.id, [3]gppras@telkomuniversity.ac.id

Abstract

Imbalanced data presents significant challenges in machine learning, leading to biased classification outcomes favouring the majority class. This issue is especially pronounced in financial distress classification, where data imbalance is common due to the scarcity of such instances in real-world datasets. This study aims to mitigate data imbalance in financial distress companies using the Kmeans-SMOTE method approach by combining K-means clustering and the Synthetic Minority Oversampling Technique (SMOTE). Various classification approaches, including Naïve Bayes and Support Vector Machine (SVM), are implemented on a financial distress dataset from Kaggle to evaluate the effectiveness of Kmeans-SMOTE. Experimental results show that SVM outperforms Naïve Bayes with impressive accuracy (99.1%), f1-score (99.1%), Area Under Precision-Recall (AUPRC) (99.1%), and Geometric-mean (Gmean) (98.1%). Based on these results, Kmeans-SMOTE can balance the data effectively, leading to a quite significant improvement in performance.

Keywords: *K-means-SMOTE; Data Imbalance; Financial Distress*

How to Cite: D. J. Maulana, Siti Saadah, and Prasti Eko Yunanto, "Kmeans-SMOTE Integration for Handling Imbalance Data in Classifying Financial Distress Companies using SVM and Naïve Bayes", J. RESTI (Rekayasa Sist. Teknol. Inf.), vol. 8, no. 1, pp. 54 - 61, Feb. 2024.
DOI: https://doi.org/10.29207/resti.v8i1.5150

1. Introduction

The issue of imbalanced data can impact the effectiveness of machine learning models, particularly for classification methods that assume a balanced data sample size across each class. Imbalance data occurs in datasets that have classes with more majority classes than minority classes. Data imbalance can introduce bias in the classification algorithm, leading to lower accuracy when it comes to classifying the minority class. The unequal distribution of data between classes can cause the algorithm to favour the majority class, as it has more samples for training. As a result, the minority class may receive less attention and maybe misclassified more frequently [1]. This is a challenge for handling problems in the classification algorithm, especially data on companies that are experiencing financial distress. In the real world, a problem may not often occur, such as the data problem for a company experiencing financial distress, therefore causing a classification error which results in a high cost to overcome it [2]. So, this imbalance needs to be addressed, to ensure fairness and accurate classification, potentially through techniques like oversampling or undersampling to balance the dataset and improve performance for the minority class, especially in the case of financial distress classification.

Financial distress refers to a situation in which a company is unable to meet its obligations to creditors. It's also described as a state in which a company faces financial challenges. This circumstance arises before a company is officially declared bankrupt [3]. Financial distress classification has been conducted by research [1]. In that study, classification was performed using tree-based models, specifically the Decision Tree. The dataset used consisted of data from a selected group of business units accounting in the double-entry bookkeeping system for 3 periods, namely in the years 2016, 2018, and 2019, located in the Republic of Slovakia. The dataset comprised 599 companies, with 27 companies classified as experiencing financial distress and 532 companies classified as healthy. One of the findings from the research scenario was the

Received: 05-06-2023 | Accepted: 29-01-2024 | Published Online: 07-02-2024

accuracy of 99.47% for the healthy company class and 29.41% for the financial distress company class.

In research [2], prediction on companies experiencing financial distress with imbalance data was performed using SMOTE. The dataset consisted of 2628 samples, with a ratio of 2190 normal companies and 438 companies experiencing financial distress. The sample data was collected from companies listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange in China. SMOTE was applied to balance the data between companies experiencing financial distress and normal companies, addressing the data imbalance issue. The results of the research demonstrated that the data balancing process significantly improved the performance of the model for companies experiencing financial distress.

Several studies have been conducted on the classification of financial distress using SVM and Naive Bayes algorithms. The study [4] utilized the SVM algorithm without hyperparameter tuning for the classification of financial distress. The results of the research showed an accuracy of 81.06%, an error rate of 18.94%, a precision of 89.09%, and a recall of 59.04%. In a study [5], the prediction of financial distress was performed using SVM with hyperparameter tuning. The research yielded an accuracy of 92%, a sensitivity of 93%, a Matthews Correlation Coefficient (MCC) of 85%, and a precision of 90%. In a similar field of research, a study [6] conducted bankruptcy prediction using the Naïve Bayes algorithm. The best results obtained from the research showed an accuracy of 92.47%, an error rate of 7.52%, and a model-building time of 0.13 seconds. Therefore, researchers utilize SVM and Naive Bayes algorithms as methods for classifying companies experiencing financial distress.

Based on the explanation above, a crucial extension of the data imbalance issue is the adverse impact that affects the performance of the classification model, particularly in cases of financial distress companies. When accuracy decreases in the minority group, the risk of the potential financial crisis going undetected within the company also increases. The Kmeans-SMOTE approach has not been widely used in real-world problems, and it is necessary to find optimal hyperparameters. So, this research proposes to implement the Kmeans-SMOTE approach to the imbalanced data of financial distress companies to see how this approach influences classification using SVM and Naïve Bayes. Although extensive research has been conducted on SMOTE, Kmeans-SMOTE has not been widely adopted and observed. Therefore, Kmeans-SMOTE is used to solve the imbalance problem. So that after the imbalance data has been solved, data is obtained in good condition for the classification process [7]. Hence, Kmeans-SMOTE will be used to balance the data classes, especially in the financial distress problem and emerging algorithms that will be used to complete the research objectives.

The next section will discuss the research methods employed in this study, encompassing data preprocessing, handling imbalanced data, model building, and model evaluation (Section 2). Subsequently, Section 3 will present a discussion of the obtained results. Finally, Section 4 will delve into the conclusions drawn from this research.

2. Research Methods

The general design of this research is to study problems regarding data imbalance in financial distress company datasets.

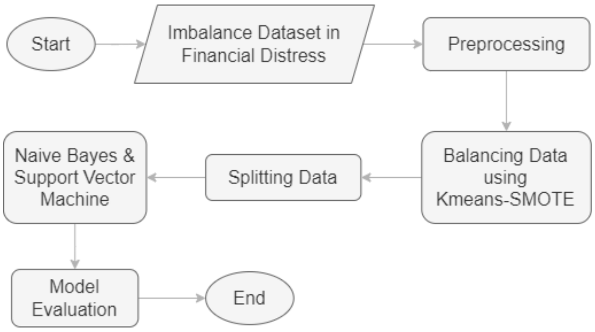

Start
Imbalance Dataset in Financial Distress
Preprocessing
Balancing Data using Kmeans-SMOTE
Splitting Data
Naive Bayes & Support Vector Machine
Model Evaluation
End

Figure 1. System Design Flow

Based on Figure 1, this study comprises several stages. Firstly, preprocessing is performed on the imbalanced data of companies experiencing financial distress. During the preprocessing stage, empty and duplicate data are eliminated, data labelling is conducted, and data scaling is performed. Next, the data is split in a ratio of 70:30 to help the model disregard or underestimate the minority class, which could result in biased predictions. Subsequently, the classification process is executed using the Naïve Bayes and SVM algorithms. Finally, the performance of the classification model on the financial distress dataset is evaluated to assess its effectiveness.

2.1 Imbalance Dataset in Financial Distress

Financial distress refers to a state where a company's financial conditions have declined, occurring before the company reaches a state of bankruptcy [8]. This will happen if the company cannot show an excellent financial performance. As a result, the company will be indicated to be experiencing a condition of financial distress [9]. According to Pan [10] states, there are three reasons for a company experiencing financial difficulties, namely significant expenses to pay off the interest on loans, relatively low operational capabilities compared to other companies, and declining industrial conditions. So, financial distress is a condition where a company is unable to pay its obligations to creditors because it is experiencing financial difficulties.

The dataset utilized in this study consists of data on financial distress companies sourced from Kaggle[1], the data used is the result of calculating financial distress scores from 422 sample companies. The dataset has a total of 86 features, as can be seen in Table 1, in the first feature is a company that represents a sample of companies, the second feature is a time that indicates different periods between 1 to 14, and the third feature is the target variable. The fourth feature (x1 to x83) is some financial and non-financial characteristics of the sample companies. The dataset showcases an imbalance ratio of 3.8%, with 136 instances representing the minority class (healthy companies) and 3536 instances representing the majority class (financial distress companies).

Table 1. Example of Financial Distress Dataset[1]

Company	Time	FD	X1	…	X83
1	1	0.010	1.2810	…	49
1	2	-0.45	1.2700	…	50
1	3	-0.32	1.0529	…	51
1	4	-0.56	1.1131	…	52
2	1	1.35	1.0623	…	27

2.2 Preprocessing

During this stage, various steps are taken to ensure the data is well-prepared for the modeling process. These steps involve removing any empty or duplicate data entries, scaling the data to a range of 0 to 1 using a min-max scaler, and assigning appropriate labels to the data. The labelling of the data is particularly important in the subsequent clustering process. By assigning labels, the data instances can be categorized and grouped based on their shared characteristics, which is essential for effective cluster formation. This becomes especially relevant in the data balancing process using Kmeans-SMOTE, which will be discussed in the next section. As an example, the preprocessing of the data presented in Table 1 is displayed in Table 2.

Table 2. Example of The Preprocessing Results

FD	X1	X2	X3	…	Class
0.126136	0.046486	0.557850	1.803817	…	0
0.119326	0.046062	0.525135	1.690617	…	0
0.121232	0.037693	0.394448	1.904430	…	0
0.117712	0.040014	0.482091	1.770910	…	1
0.145791	0.038055	0.724771	1.677862	…	0

2.3 Balancing Data using Kmeans-SMOTE

Kmeans-SMOTE is an algorithm that combines the Kmeans and SMOTE algorithms. Kmeans aims to minimize the Sum of Squared Error (SSE) (Equation 1) between data objects with several k centroids. Kmeans itself is used for the clustering process in the input space, while SMOTE is used in the oversampling process to solve data imbalance problems [7].

$$SSE = \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

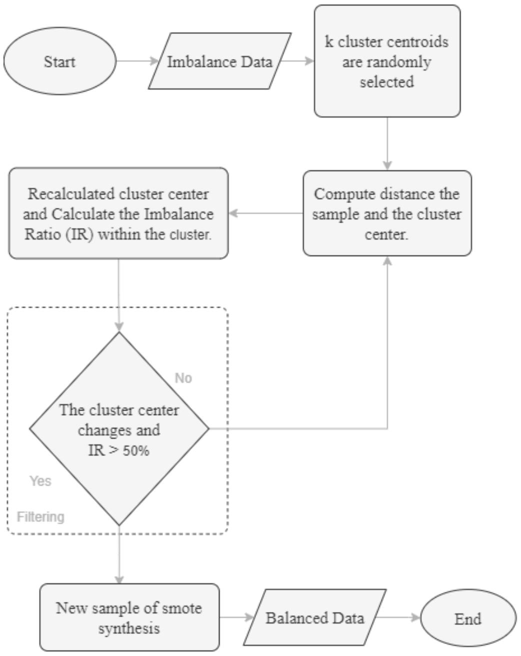

Start
Imbalance Data
k cluster centroids are randomly selected
Compute distance the sample and the cluster center.
Recalculated cluster center and Calculate the Imbalance Ratio (IR) within the cluster.
No
The cluster center changes and IR > 50%
Yes
Filtering
New sample of smote synthesis
Balanced Data
End

Figure 2. Kmeans-SMOTE Flow

Based on Figure 2, the data balancing process begins by randomly selecting k centroids to initialize the clusters. After selecting the centroids, the Euclidean distance is employed to calculate the distances between the samples and centroids. Then, the centroids are recalculated, and the imbalance ratio is calculated for each cluster to evaluate the relative proportions of the majority and minority classes. A centroid is chosen when its position becomes stable/unchanged, and the cluster contains at least 50% of the minority class. This criterion is important to enhance the performances of the minority class representation within the selected cluster. Then, SMOTE is applied to the selected clusters, generating new samples through a three-stage process. Firstly, a random sample, denoted as $\vec{a}$, is chosen from the minority class. Next, one of its k nearest neighbors from the minority class, represented by $\vec{b}$, is selected. Finally, a new sample $\vec{x}$ is generated by interpolating between $\vec{a}$ and $\vec{b}$ using a random weight, w, ranging from 0 to 1: $\vec{x} = \vec{a} + w \times (\vec{b} - \vec{a})$. Once this process is completed, the resulting balanced data can be used for classification modelling.

2.4 Splitting Data

In this research, data splitting plays a crucial role in the data preparation process. Its purpose is to divide the available dataset into two distinct sets: the training set and the testing set. The training set is utilized for constructing the classification model, while the testing set is employed to assess the model's performance on unseen data.

[1] https://www.kaggle.com/code/saurabhharsh/financial-distress-prediction

Several studies, such as those conducted by researchers [11] and [12], have utilized a data splitting ratio of 70:30, which has resulted in favourable model performance. According to those studies, this study adopted the same ratio, with 70% of the data allocated to the training set and the remaining 30% assigned to the testing set. This allocation ensures a sizable training set for effective model training, while also providing a significant testing set for rigorous performance evaluation. By employing this ratio, the study aims to achieve a balance between model training and evaluation, ultimately enabling accurate assessments of the model's performance.

2.5 Classifying using Naïve Bayes

Naive Bayes is a supervised learning classification algorithm and is based on Bayes' theorem [13]. In Bayes' theorem, a conditional possibility can be seen in Equation 2.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Based on Equation (2), P(A|B) is the probability of the hypothesis A for data B also known as the posterior probability. P(A|B) is the probability that the data at B are true for hypothesis A (posterior B). Meanwhile, P(A) is the prior probability of the A hypothesis and P(B) is the prior probability of B data.

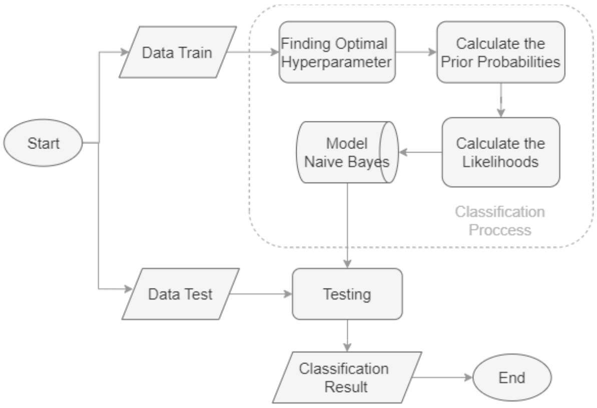


Figure 3. Naïve Bayes Model Flow

Based on Figure 3, the training and testing process on the Naïve Bayes model begins with the search for optimal hyperparameters in the train data. After finding the optimal hyperparameter, it goes into the classification process using the Gaussian Naïve Bayes Classifier. Furthermore, the Naïve Bayes model is generated from the training on financial distress data. Then testing is conducted using test data based on the data that has been trained earlier. Finally, at this stage, a classification of financial distress is produced.

2.6 Classifying using SVM

SVM is a supervised learning algorithm that can be applied to classification problems and regression analysis [14]. This SVM, in general, is categorized into 2, namely SVC (Support Vector Classification) and SVR (Support Vector Regression). SVC is a term in the SVM model that is used for the classification process where SVC has been widely used and has succeeded in categorizing 2 or more classes in the dataset.

$$w^t x + b = 0 \tag{3}$$

Based on Equation 3, w represents an n-dimensional vector and b represents a bias condition. Meanwhile, x represents the training data set from $x - i$ to n [15].

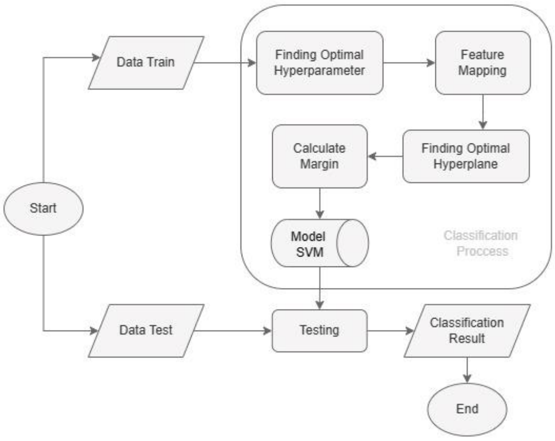
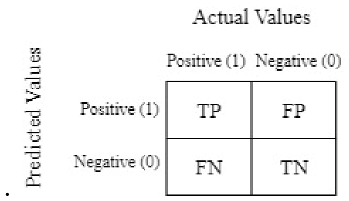


Figure 4. SVM Model Flow

Based on Figure 4, the train data in the SVM model is searched for optimal hyperparameters using the Grid Search method. Next, a classification process is performed utilizing a Support Vector Classifier (SVC) to generate a trained SVM model. Then, the testing process is conducted using test data where the model that has been trained will be tested. Finally, this stage produces a classification of financial distress.

2.7 Model Evaluation

In this research, measurement of the performance of the classification model that has been made is done by evaluating it using a confusion matrix. The confusion matrix is a table that provides a comprehensive summary of the classifier's predictions, including the number of correct and incorrect predictions.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 5. Confusion Matrix

Based on Figure 5, True Positive (TP) occurs when an actual positive value is correctly predicted as positive. On the other hand, a False Positive (FP) happens when an actual negative value is incorrectly predicted as positive. Furthermore, FN (False Negative) is a condition when an actual negative value is predicted to be positive, while TN (True Negative) is a condition where an actual negative value is predicted to be negative [16]. From this confusion matrix, some measurements can be taken to gain further understanding and analysis of the classification model that has been built, such as accuracy which focuses on calculating how often the classifier predicts the correct value. Accuracy is a metric that measures the proportion of correct predictions about the total number of predictions made by a model as can be seen in Equation 4. The next measurement is the f1-score to calculate the average harmonic value of recall (Equation 5) and precision. Other measurements namely Gmean which is the root of recall multiplied by specificity (Equation 7) as can be seen in Equation (8). The last measurement AUPRC is a measurement involving areas of under-precision recall.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$f1 - score = \frac{TP}{TP+{}^{1}/_{2}(FP+FN)} \tag{5}$$

$$recall = \frac{TP}{TP+FN} \tag{6}$$

$$specificity = \frac{TN}{TN+FP} \tag{7}$$

$$gmean = \sqrt{recall * specificity} \tag{8}$$

3. Results and Discussions

In this study, several scenarios were conducted to examine the impact of oversampling using Kmeans-SMOTE. The purpose was to investigate the effects and outcomes of applying the Kmeans-SMOTE oversampling technique: the Naïve Bayes and SVM models as classifiers to classify imbalanced data. Furthermore, the Naïve Bayes and SVM models classify balanced data using Kmeans-SMOTE. In each scenario, training data and testing data are used in a ratio of 70:30 as previously explained in the data splitting section.

3.1 Handling of Data Imbalance

At this stage, the researcher uses Kmeans-SMOTE as a method for solving unbalanced data. Kmeans-SMOTE conducts an oversampling procedure on the minority class to achieve a balanced dataset, aligning it with the number of instances in the majority class. As can be seen in Figure 6, the class comparison between financially distressed and healthy companies experienced data imbalance with the number of classes labelled financial distress as many as 136 and healthy companies as many as 3,536.

Based on Figure 6 and Figure 7, the number of minority data instances, which was originally 136, has now increased to 3,536 after applying Kmeans-SMOTE. This indicates that Kmeans-SMOTE has effectively generated synthetic samples that align with the majority of class data. As a result, the financial distress dataset is now prepared for utilization in the machine learning model development stage.

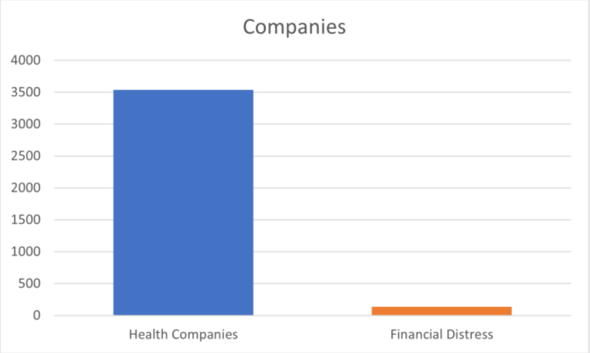


Figure 6. Imbalance Data

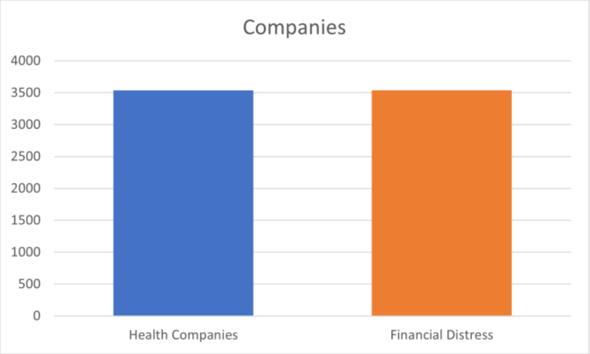


Figure 7. Balance Data

3.2 Testing Scenarios

Testing scenarios can be seen in Table 3. This testing involved conducting the modeling process using both SVM and Naïve Bayes algorithms. The testing process is conducted using both imbalanced and balanced datasets.

Table 3. Testing Scenarios

Testing	Data	Method
1	Imbalanced Data without Kmeans-SMOTE	Naïve Bayes, Support Vector Machine
2	Balanced Data with Kmeans-SMOTE	Naïve Bayes, Support Vector Machine

3.3 Testing Result

During the training of the models, hyperparameter tuning is performed for each model to be constructed. The optimal hyperparameters for each model are obtained and presented in Table 4.

Kmeans-SMOTE integration for the oversampling method on financial distress datasets that experience imbalance has a positive effect on the classifier which is in line with research [17] - [19]. As presented in Table 5, the SVM model with imbalanced data shows a high accuracy of 97.6%, but the f1-score (64.9%), AUPRC (49.3%), and Gmean (49.0%) are relatively low. This shows that the model can identify most of the data correctly, but the model experiences identification errors in the minority class classification [20] - [22].

Table 4. Hyperparameter Tuning Results

Method	Data	Parameter	Value
SVM	Imbalanced Data without Kmeans-SMOTE	C	100
		Gamma	0.001
		Kernel	Linear
	Balanced Data with Kmeans-SMOTE	C	10
		Gamma	1
		Kernel	RBF
Naïve Bayes	Imbalanced Data without Kmeans-SMOTE	Var_smoothing	0.81
	Balanced Data with Kmeans-SMOTE	Var_smoothing	0.01

In the SVM model with balanced data, it shows a significant increase in accuracy, namely 99.1%, f1-score (99.1%), AUPRC (99.1%), and Gmean (98.1%). Based on this, Kmean-SMOTE succeeded in significantly increasing the performance of the SVM model in the financial distress dataset. Meanwhile, the Naïve Bayes model with an imbalanced dataset has performance with accuracy (96.5%), f1-score (9.50%), AUPRC (5.90%), and Gmean (5.30%). Even though it has a high accuracy value, Naïve Bayes has not been able to correctly identify minority data as well as in the SVM model. Based on this performance, Naïve Bayes is not good enough to classify imbalanced data. However, the Naïve Bayes model on a balanced dataset has improved performance with accuracy (87.0%), f1-score (88.2%), AUPRC (79.8%), and Gmean (85.1%). This shows that the role of Kmeans-SMOTE in the Naïve Bayes model has a significant effect on performance, even though the results are not as good as the SVM model.

Before addressing data imbalance, two distinct machine learning models, Naïve Bayes and Support Vector Machine (SVM), were assessed for their accuracy performance across the 'Healthy Company' and 'Financial Distress Company' categories. The Naïve Bayes model demonstrated impressive accuracy, achieving a perfect 100% accuracy within the 'Healthy Company' class (class 0), while yielding a 0% accuracy within the 'Financial Distress Company' (class 1). Conversely, the SVM model showcased notable variations. It achieved a high accuracy rate of 99.90% within the 'Healthy Company' class but experienced a significant drop to 52.83% within the 'Financial Distress Company' class. These contrasting accuracy outcomes highlight the models' differential classification capabilities within the given categories. The disparity in accuracy for the 'Financial Distress Company' class emphasizes the challenge posed by class imbalance, prompting the need for mitigation to enhance predictive performance in both scenarios.

Subsequently, following the implementation of data imbalance handling using Kmeans-SMOTE, the Naïve Bayes model demonstrated improved performance, achieving an accuracy of 77.65% for the 'Healthy Company' class and 96.33% for the 'Financial Distress Company' class. Similarly, the SVM model exhibited enhanced accuracy after data handling, with 99.8% accuracy within the 'Healthy Company' class and 98.3% accuracy within the 'Financial Distress Company' class. These outcomes underline the effectiveness of data balancing techniques in rectifying the initial accuracy discrepancies, leading to improved predictive capabilities across both categories.

Table 5. Model Performances

Method	Data	Evaluation	Skor
SVM	Imbalanced Data without Kmeans-SMOTE	Accuracy	97.6
		F1-Score	64.9
		AUPRC	49.3
		Gmean	49.0
	Balanced Data with Kmeans-SMOTE	Accuracy	**99.1**
		F1-Score	**99.1**
		AUPRC	**98.1**
		Gmean	**99.1**
Naïve Bayes	Imbalanced Data without Kmeans-SMOTE	Accuracy	96.5
		F1-Score	9.50
		AUPRC	5.90
		Gmean	5.30
	Balanced Data with Kmeans-SMOTE	Accuracy	87.0
		F1-Score	88.2
		AUPRC	79. 8
		Gmean	85.1

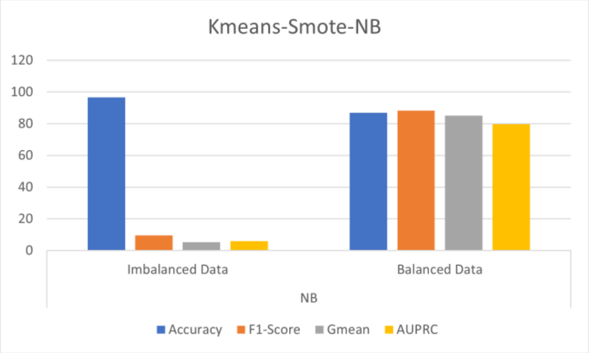


Figure 8. Model Performance Kmeans-SMOTE-NB

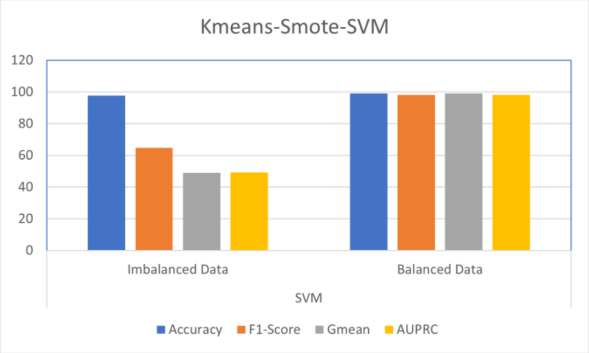


Figure 9. Model Performance Kmeans-SMOTE-SVM

The performance of each scenario can be observed through a comparison presented in Figure 8 and Figure 9. In both the Naïve Bayes and SVM models, when dealing with a dataset that hasn't been processed using Kmeans-SMOTE, the performance is not satisfactory. This is clear from Figure 6, where there's an uneven distribution of data between the healthy company class and the financially distressed company class. This result supports the findings of studies [1] and [2] which all suggest that when the imbalanced data issue in dataset classes isn't addressed, the performance suffers. Both the Naïve Bayes and SVM models perform very well when applied to a dataset that has been balanced using Kmeans-SMOTE. This observation aligns with the conclusions of research [2] and [8], indicating that addressing class imbalance in a dataset significantly improves performance compared to scenarios where no balancing is carried out at all. So, the application of Kmeans-SMOTE as a technique for handling imbalanced data has been substantiated as an effective approach to enhance model performance on financial distress datasets. This serves as evidence that the data balancing process has successfully improved the model's performance. This is important because there are not many companies experiencing financial distress, so misclassification will have significant consequences for the company. The use of hyperparameter tuning in each model also affects the performance of the model. The findings of this study reveal an enhancement in the performance of each model; however, it should be noted that this improvement may not apply to all datasets. This is due to the unique complexity and characteristics of each dataset. Therefore, the selection of an appropriate approach to address data imbalance and the choice of classification model should depend on the specific complexity of the dataset in question.

4. Conclusions

Based on this research, it can be concluded that the implementation of Kmeans-SMOTE on imbalanced financial distress company data significantly improves the handling of data imbalance. The evaluation of model performance using the confusion matrix shows a substantial improvement after applying Kmeans-SMOTE in addressing data imbalance. Specifically in enhancing accuracy within both the 'Healthy Company' and 'Financial Distress Company' categories, where the accuracy for the minority class was previously poor, it has now improved. These findings indicate the effectiveness of Kmeans-SMOTE in enhancing the classification model's performance on imbalanced financial distress data. The implications of this study underscore the importance of employing oversampling techniques such as Kmeans-SMOTE to address data imbalance issues, particularly in the context of financial distress data, to enhance the accuracy and effectiveness of classification models. The researchers recommend further exploration of alternative classification algorithms. By employing different algorithms, models can benefit from diverse learning techniques, potentially capturing unique patterns within the data. This approach may offer additional insights and improve the performance of classification results in the context of imbalanced financial distress data.

Acknowledgements

The author is grateful to Telkom University which has facilitated the completion of this research.

References

[1] P. Kr, “FINANCIAL DISTRESS CLASSIFICATION Mária Stachová – Pavol Kráľ,” pp. 977–988, 2021.

[2] J. Sun, H. Li, H. Fujita, B. Fu, and W. Ai, “Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting,” *Inf. Fusion*, vol. 54, no. December 2018, pp. 128–144, 2020, doi: 10.1016/j.inffus.2019.07.006.

[3] M. D. Costa, H. J. Huang, B. U. Bhuiyan, and L. Sun, “Determinants and consequences of financial distress : a review of the empirical literature”, doi: 10.1111/acfi.12400.

[4] N. W. D. Ayuni, N. N. Lasmini, and A. A. Putrawan, “Support Vector Machine (SVM) as Financial Distress Model Prediction in Property and Real Estate Companies,” *Proc. Int. Conf. Appl. Sci. Technol. Soc. Sci. 2022 (iCAST-SS 2022)*, pp. 397–402, 2022, doi: 10.2991/978-2-494069-83-1_72.

[5] S. Doğan, D. Koçak, and M. Atan, “Financial Distress Prediction Using Support Vector Machines and Logistic Regression,” *Contrib. to Econ.*, no. May, pp. 429–452, 2022, doi: 10.1007/978-3-030-85254-2_26.

[6] P. Patel, A. Shrivastava, and S. Nagar, “Bankruptcy Prediction Model Using Naïve Bayes Algorithms,” vol. 59, no. 01, p. 83, 2019, [Online]. Available: https://archive.ics.uci.edu/ml/machine-

[7] G. Douzas, F. Bacao, and F. Last, “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[8] A. Zhafirah and & Majidah, “Analisis Determinan Financial Distress (Studi Empiris Pada Perusahaan Subsektor Tekstil dan Garmen Periode 2013-2017),” *J. Ris. Akunt. dan Keuang.*, vol. 7, no. 1, pp. 195–202, 2019, doi: 10.17509/jrak.v7i1.15497.

[9] W. Setyowati and N. R. Sari Nanda, “Pengaruh Likuiditas, Operating Capacity, Ukuran Perusahaan Dan Pertumbuhan Penjualan Terhadap Financial Distress (Studi Pada Perusahaan Manufaktur Yang Terdaftar Di Bei Tahun 2016-2017),” *J. Magisma*, vol. 4, no. 2, pp. 618–624, 2019.

[10] I. Setyawati and R. Amelia, “The Role of Current Ratio, Operating Cash Flow and Inflation Rate in Predicting Financial Distress: Indonesia Stock Exchange,” *J. Din. Manaj.*, vol. 9, no. 2, pp. 140–148, 2018, doi: 10.15294/jdm.v9i2.14195.

[11] K. Andrić and D. Kalpić, “An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment,” vol. 16, no. 1, pp. 155–178.

[12] R. Abdillah, “The Effect of Class Imbalance Against LVQ Classification,” no. October, pp. 42–45, 2018.

[13] D. Berrar, “Bayes’ theorem and naive bayes classifier,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.

[14] J. Zhou *et al.*, “Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate,” *Eng. Appl. Artif. Intell.*, vol. 97, no. October 2020, p. 104015, 2021, doi: 10.1016/j.engappai.2020.104015.

[15] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, no. xxxx, pp. 189–215, 2020, doi:

10.1016/j.neucom.2019.10.118.

[16] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, pp. 56–62, 2018, doi: 10.1016/j.beproc.2018.01.004.

[17] Aditya Gumilar, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "Performance Analysis of Hybrid Machine Learning Methods on Imbalanced Data (Rainfall Classification)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3, pp. 481–490, 2022, doi: 10.29207/resti.v6i3.4142.

[18] O. Barukab, A. Ahmad, T. Khan, and M. R. Thayyil Kunhumuhammed, "Analysis of Parkinson's Disease Using an Imbalanced-Speech Dataset by Employing Decision Tree Ensemble Methods," *Diagnostics*, vol. 12, no. 12, pp. 1–21, 2022, doi: 10.3390/diagnostics12123000.

[19] M. Hayaty, S. Muthmainah, and S. M. Ghufran, "Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification," *Int. J. Artif. Intell. Res.*, vol. 4, no. 2, p. 86, 2021, doi: 10.29099/ijair.v4i2.152.

[20] K. Fithriasari, I. Hariastuti, and K. S. Wening, "Handling Imbalance Data in Classification Model with Nominal Predictors," *Int. J. Comput. Sci. Appl. Math.*, vol. 6, no. 1, p. 33, 2020, doi: 10.12962/j24775401.v6i1.6643.

[21] A. Indrawati, H. Subagyo, A. Sihombing, W. Wagiyah, and S. Afandi, "Analyzing the Impact of Resampling Method for Imbalanced Data Text in Indonesian Scientific Articles Categorization," *Baca J. Dokumentasi Dan Inf.*, vol. 41, no. 2, p. 133, 2020, doi: 10.14203/j.baca.v41i2.702.

[22] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," vol. 22, no. 2, pp. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.