IJAI (Indonesian Journal of Applied Informatics)
Vol. 9 No. 1 Tahun 2024 pISSN: 2548-3846, eISSN: 2598-5981

A Hybrid Model to Enhance The Performance of Classifier in Financial Distress Prediction

Mukti Ratna Dewi[1]*, Destri Susilaningrum[1]
[1]Department of Business Statistics, Institut Teknologi Sepuluh Nopember
*Email: mukti_ratna@its.ac.id

Info Artikel

Keywords :
financial distress, hybrid classifiers, bankruptcy prediction, *K*-means, SVM

Article Date
Sent : 1 November 2024
Revised : 15 November 2024
Accepted : 17 November 2024

Abstract

Accurately predicting financial distress is a critical issue in financial decision-making. Financial distress must be detected as early as possible as an important determining factor in decision-making for internal companies and financial institutions related to financing or loan decisions. Various studies on financial distress prediction in Indonesia have been carried out, ranging from traditional statistical approaches to machine learning. However, the performance of the two methods is still not optimal. Therefore, this study tries to develop machine learning techniques by combining cluster analysis and classification in a hybrid model to improve the prediction model's performance. The case study adopted in this study is the prediction of financial distress in non-financial companies listed on the IDX from 2018-2021 by combining *k*-means clustering and Support Vector Machine. The analysis results show that the hybrid classifier has an accuracy value of 92.7%, which is higher than the accuracy of the single classifier, which is 88.6%.

1. INTRODUCTION

Business failure detection plays a vital role in the professional field. The risk of business failure has become one of the crucial factors in business decisions. For years, professionals have tried to detect potential business failures to reduce the impact caused by bankruptcy [1]–[3]. Company failures are generally preceded by financial distress, which can be seen through the company's performance in the last one or several years [4]. A company is vulnerable to financial distress if it often lacks cash and small income streams; hence, it cannot pay off maturing debts [5], [6]. The criteria for a company experiencing financial distress is if the company has a negative financial record or business losses for three consecutive years [7], [8]. Financial distress must be detected as early as possible as an important determining factor in decision-making, not only for internal companies but also for financial institutions related to financing or loan decisions.

Various techniques have been developed over the years to analyze and make decisions with practical methods for predicting financial distress based on various financial ratios and mathematical models, including linear and logistic regression, Multivariate Adaptive Regression Splines (MARS), survival analysis, linear, quadratic and multi-criteria programming [9]–[11]. Most of these techniques usually rely on assumptions of linear separability, multivariate normality, and independence between explanatory variables [12]. However, this condition is often not met in real situations. As an alternative, a machine learning approach has begun to be applied to predict financial distress. Several studies show machine learning has superior predictive results to traditional statistical methods [13]–[17].

In Indonesia, Support Vector Machine (SVM) is one of the most popular machine learning techniques for predicting financial distress. It is considered better in its application to the classification model because of its ability to describe the complexity of the model [18]. The hyperplane between classes formulated from the training data is limited to the distribution of linear lines or planes and can be described as parametric radials or three-dimensional surfaces. Several studies on the application of SVM in predicting financial distress show a fairly good performance [19] and even better results when compared to Linear Discriminant Analysis (LDA) [20], [21]. Although SVM is generally superior to LDA, the accuracy obtained from previous studies is still less than optimal.

One technique that can be used to improve classifier performance is to apply ensemble learning that combines several machine learning techniques into a hybrid model, such as supervised and unsupervised learning. Unsupervised learning can reveal the underlying structure of the sample space and the intrinsic relationships between samples. This information provides an additional description of the data and increases the classification effectiveness of supervised learning [22]. Several studies have shown that a hybrid classifier performs better than a single classifier [23]–[25].

Using 20 financial ratios from various studies, this study introduces a new approach that has never been applied before in predicting financial distress for non-financial companies listed on the IDX. A new ensemble learning approach will be applied by combining *k*-means clustering with SVM. This study used *k*-means for pre-classification to arrange meaningful groups based on similarity measures. At the same time, this can also be used to filter out outliers to reduce prediction errors during building a classification model [26]. The results of clustering in the form of data groups are then used to build a classifier using SVM.

2. MATERIAL AND METHODOLOGIES

2.1 Classification

Classification is a grouping of data where the data used has a label or target class so that the algorithms to solve this problem are categorized into supervised learning. The purpose of supervised learning is that data labels or targets play the role of a 'supervisor' or 'teacher' who oversees the learning process to achieve a certain level of accuracy or precision. This study uses the classification technique Support Vector Machine (SVM).

SVM is a method developed by [27] to improve classification accuracy performance. This method is an excellent method to overcome the problem of high-dimensional classification. The fundamental concept of SVM is to find a hyperplane in *N*-dimensional space that can classify data [27].

Suppose a classification problem of m points in R^n is represented by an $\mathbf{A}$ matrix of size $m \times n$ where each $\mathbf{A}_i$ is a $\mathbf{D}$ matrix of size $m \times m$ with a value of 1 or -1 on the main diagonal. The SVM formulation for nonlinear problems is shown in equation **Error! Reference source not found.**.

$$\min_{\mathbf{u}\in R^m} \ \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{D}K\left(\mathbf{A},\mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} - \mathbf{e}^{\mathrm{T}}\mathbf{u} \qquad \text{s.t.} \quad \mathbf{e}^{\mathrm{T}}\mathbf{D}\mathbf{u} = 0,\ 0 \le \mathbf{u} \le \nu\mathbf{e} \tag{1}$$

where $K(\mathbf{A}, \mathbf{A}^{\mathrm{T}})$ is a kernel function that maps data to higher dimensions. After that, the hyperplane is created using linear SVM to optimally separate the two classes. The resulting hyperplane can be seen in equation **Error! Reference source not found.**.

$$K\left(\mathbf{x}^{\mathrm{T}},\mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} = \gamma \tag{2}$$

where

$$\gamma = K\left(\mathbf{A}_i,\mathbf{A}^{\mathrm{T}}\right)\mathbf{D}\mathbf{u} - D_{ii},\ i \in \mathbf{I} := \left\{ j \mid 0 < u_j < \nu \right\} \tag{3}$$

This study uses a linear function as a kernel function in SVM denoted as $K(\mathbf{A}, \mathbf{A}^{\mathrm{T}}) = \mathbf{A}^{\mathrm{T}}\mathbf{A}$.

2.2. Clustering

Cluster analysis is a multivariate technique with the primary goal of grouping objects based on their characteristics. It classifies objects so that each object with similar properties will be grouped into the same cluster [28]. A clustering technique used in this study is *k*-means.

K-means attempts to find the assignment of observations to a fixed number of clusters K that minimize the sum over all clusters of the sum of squares within clusters:

$$\sum_{k=1}^{K} \sum_{i:x_i \in C_k} \left(x_i - \bar{x}_{C_k}\right)^2 \tag{4}$$

where $\bar{x}_{C_k}$ is the average of all the points belonging to cluster k.

The most common algorithm uses an iterative refinement technique, often known as Lloyd's algorithm:

1. Begin with K starting centres.
2. Assign each observation to the cluster with the closest centre.

3. Re-calculate the cluster centres by finding the centroids of each cluster's assigned observations.
4. Iterate steps 2 and 3 until convergence.
5. The final cluster centres are considered the best representative of the clusters.

2.3. Hybrid Classifier

A hybrid classifier algorithm is developed by merging two or more heterogonous machine learning approaches, such as clustering and classification techniques, as shown in Figure 1 [29].

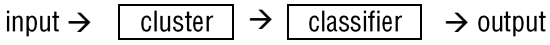

Figure 1. Architecture of a Hybrid Classifier

At first, clustering can be used as a pre-processing stage to identify pattern classes for subsequent supervised classification [28]. Therefore, the clustering result can be used to identify major populations of a given dataset or pre-classify unlabeled collections. The cluster technique can also detect outliers so that only representative data is used for the classification stage.

After the clustering process, the next step is to build a classifier. Results from the previous clustering process become the training set to train a classifier. After the classifier is trained, it can classify new instances.

2.4. Model Evaluation

The classifier's performance is evaluated using the area under the Receiver Operating Characteristics (ROC) curve that is appropriate for imbalance data [30]. Area Under Curve (AUC) is calculated using equation **Error! Reference source not found.**.

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{5}$$

where TP_{rate} and FP_{rate} can be obtained using equation **Error! Reference source not found.** and **Error! Reference source not found.**.

$$TP_{rate} = \frac{TP}{TP + FN} \tag{6}$$

$$FP_{rate} = \frac{FP}{FP + FN} \tag{7}$$

The value of TP, FP, TN, and TP is obtained from confusion matrix in Table 1.

Table 1. Confussion Matrix

		Prediction	
		Positive (0)	Negative (1)
Actual	Positive (0)	True Positive (TP)	False Negative (FN)
	Negative (1)	False Positive (FP)	True Negative (TN)

2.5. Datasets

The data used in this study were taken from companies' financial statements on the Indonesia Stock Exchange from 2020 to 2023. The companies involved in this study were non-financial companies with complete financial reports from 2020-2023. The sample will be classified into two categories: 0: financially healthy companies and 1: financially distressed companies. A company is said to be in financial distress if the company has a negative financial record or business losses for three consecutive years [7], [8]. Based on research [21], the variables that will be used as independent variables to form the prediction model are presented in Table 2.

Table 2. Financial Ratios Indexes

Variable	Definition
X1	*EBIT/ Total Asset*
X2	*Sales/ Total Asset*

X3	*Sales/ Fixed Asset*
X4	*Earning/ Debt*
X5	*Current Ratio*
X6	*Working Capital/ Total Asset*
X7	ROE
X8	*Retained Earning/ Total Asset*
X9	*Gross Profit Ratio*
X10	*Operating Profit Ratio*
X11	*Net Profit Ratio*
X12	EBIT/ *Sales*
X13	ROI
X14	*Working Capital/ Long Term Debt*
X15	*Debt to Equity*
X16	*Book Equity/ Total Capital*
X17	*Market Value Equity/ Total Capital*
X18	*Market Value Equity/ Liability*
X19	PER
X20	PBV

2.6. Methodology

The financial distress prediction model is determined based on testing the company's financial performance in year (t-1) against the actual financial condition that occurs one year later (t). Therefore, the first step is to divide the data into training and testing data using hold-out method. The training data covers companies in the non-financial sector in 2022 (t-1), while the testing data covers companies in the non-financial sector in 2023 (t).

After pre-processing the data, including outlier removal, the analysis is proceeded to build a single classifier using SVM algorithm and a hybrid classifier that combines k-means clustering with SVM. The value of k was initially set to three. As for SVM, we used grid search for hyperparameter tuning. After that, we built single and hybrid classifier. Lastly, the performance of those two classifiers will be compared to choose the best classifier for predicting the financial condition of non-financial listed companies in Indonesia.

3. RESULTS AND DISCUSSION

3.1 Financial Conditions of Non-Financial Sector Companies

This study uses financial reports from 510 non-financial companies in Indonesia. The proportions of financial conditions in 2022 and 2023 can be seen in Figure 2, where the number of financially healthy companies is around eight times that of those with financially distressed conditions. A company in 2022 is said to be financially distressed if it consecutively has a negative net income for 2019, 2020, and 2021. If the same problem occurs consecutively in 2020, 2021, and 2022, then the company also experience financial distress in 2023.

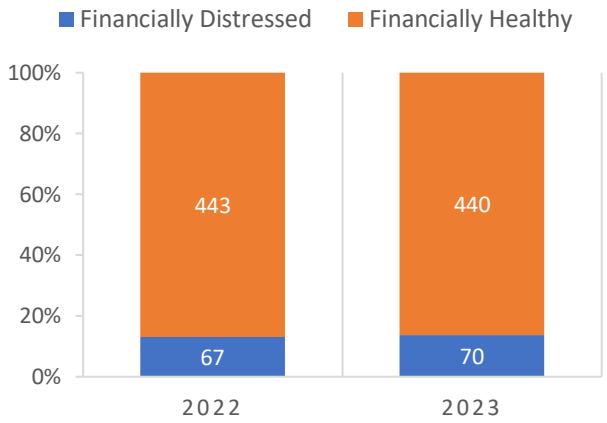

Figure 2. The Proportion of Companies Experiencing Financial Distress

3.2 Cluster Analysis

The analysis begins with cluster analysis using *k*-means for financial data 2022 as the initial stage of building a hybrid classifier and filtering outliers. Figure 3 shows three groups formed from clustering after the outliers are removed. It can be seen that Cluster 1 has the most members, followed by Cluster 2 and finally Cluster 3. The data, which initially consisted of 510 companies, is reduced to 397 companies, of which 88.2% have healthy financial conditions while the rest are financially distressed. This data is then used to build the classifier.

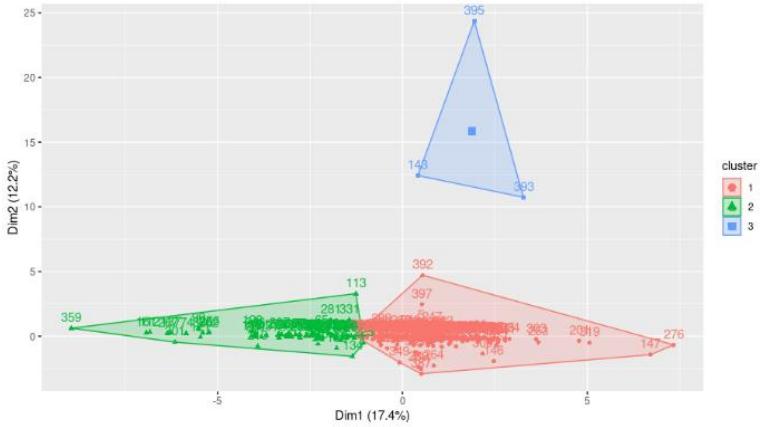

Figure 3. Clusters Formed After Outlier Removal

3.3 Building Classifier

In this study, the prediction model for financial condition was built using two approaches: single classifier and hybrid classifier. SVM is implemented to build a single classifier, whereas for a hybrid classifier, the result of clustering from *k*-means analysis is added as a predictor to SVM model. As mentioned before, the prediction model is built from 2022 data, whereas the data from 2023 is used for evaluation.

Table 3. Performance Evaluation

Methods	Accuracy	Sensitivity	Specificity	AUC
Single Classifier (SVM)	88.6%	89.0%	66.7%	0.56
Hybrid Classifier (*k*-means + SVM)	92.7%	94.7%	73.7%	0.78

The performance evaluation of single classifiers and hybrid classifiers is presented in Table 3. It is apparent that the financial condition of Indonesia listed company can be better predicted by the hybrid classifier combining *k*-means clustering and SVM. The addition of clustering at the pre-processing stage can increase the accuracy of the

classification model by 8.1% compared to a single classifier using SVM. The hybrid model can accurately predict the company's financial condition by 92.7%. Specifically, the model can accurately predict financially healthy companies by 94.7% and companies that experience financial distress by 73.7%. However, the proposed hybrid classifier is only included in a "fair classification" category, which might be due to imbalanced data. The performance of predicted models is significantly impacted when the dataset is highly imbalanced and the sample size grows [31]. In the case of imbalanced data, the classifier will tend to be better at predicting the class that has a larger proportion of data, as we can see in Table 3, where the sensitivity of both models is higher than the specificity, which reflects the model's ability to classify companies experiencing financial distressed.

4. CONCLUSION

In this study, combining unsupervised learning (*k*-means) techniques with supervised learning (SVM) in building a classifier has been proven to improve the performance of the classification model. Nevertheless, the ability of the hybrid model to classify and predict the company's financial condition in the future is still categorized as fair. This is probably caused by unbalanced data conditions where the number of companies with healthy financial conditions is about eight times that of those experiencing financial distress, which causes the model's ability to classify a financially distressed company to be low. Therefore, in further research, the problem of data imbalance can be overcome before the process of building a classification and prediction model.

ACKNOWLEDGEMENTS

The authors acknowledge and are grateful for the financial support for this work from Institut Teknologi Sepuluh Nopember through a research grant for a beginner researcher program.

REFERENCES

[1] E. J. Balleisen, *Navigating failure: bankruptcy and commercial society in antebellum America*. Univ of North Carolina Press, 2001.

[2] D. R. Henderson, *Concise Encyclopedia of Economics*. Liberty Fund, 2008.

[3] D. Liang, C. C. Lu, C. F. Tsai, and G. A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 561–572, Jul. 2016, doi: 10.1016/J.EJOR.2016.01.012.

[4] H. D. Piatt and M. B. Piatt, "Predicting corporate financial distress: Reflections on choice-based sample bias," *J. Econ. Financ. 2002 262*, vol. 26, no. 2, pp. 184–199, 2002, doi: 10.1007/BF02755985.

[5] W. H. Beaver, M. Correia, and M. F. McNichols, "Financial Statement Analysis and the Prediction of Financial Distress," *Found. Trends® Account.*, vol. 5, no. 2, pp. 99–173, 2011, doi: 10.1561/1400000018.

[6] M. S. Jahur, S. M. N. Quadir, and others, "Financial distress in small and medium enterprises (SMES) of Bangladesh: Determinants and remedial measures," *Econ. Ser. Manag.*, vol. 15, no. 1, pp. 46–61, 2012.

[7] P. Jantadej, "Using the combinations of cash flow components to predict financial distress," Jan. 2006. Accessed: Apr. 18, 2022. [Online]. Available: https://digitalcommons.unl.edu/dissertations/AAI3216429

[8] G. Kordestani, V. Biglari, and M. Bakhtiari, "Ability of combinations of cash flow components to predict financial distress," *Bus. Theory Pract.*, vol. 12, no. 3, pp. 277–285, Sep. 2011, doi: 10.3846/BTP.2011.28.

[9] M. Ezzamel, C. Mar-Molinero, and A. Beech, "On the Distributional Properties of Financial Ratios," *J. Bus. Financ. Account.*, vol. 14, no. 4, pp. 463–481, Dec. 1987, doi: 10.1111/J.1468-5957.1987.TB00107.X.

[10] G. V. Karels and A. J. Prakash, "Multivariate Normality and Forecasting of Business Bankruptcy," *J. Bus. Financ. Account.*, vol. 14, no. 4, pp. 573–593, Dec. 1987, doi: 10.1111/J.1468-5957.1987.TB00113.X.

[11] V. Ravi, H. Kurniawan, P. N. K. Thai, and P. R. Kumar, "Soft computing system for bank performance prediction," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 305–315, Jan. 2008, doi: 10.1016/J.ASOC.2007.02.001.

[12] L. Cleofas-Sánchez, V. García, A. I. Marqués, and J. S. Sánchez, "Financial distress prediction using the hybrid associative memory with translation," *Appl. Soft Comput.*, vol. 44, pp. 144–152, Jul. 2016, doi: 10.1016/J.ASOC.2016.04.005.

[13] K. Lee, D. Booth, and P. Alam, "A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms," *Expert Syst. Appl.*, vol. 29, no. 1, pp. 1–16, Jul. 2005, doi: 10.1016/J.ESWA.2005.01.004.

[14] T. Lensberg, A. Eilifsen, and T. E. McKee, "Bankruptcy theory development and classification via genetic programming," *Eur. J. Oper. Res.*, vol. 169, no. 2, pp. 677–697, Mar. 2006, doi: 10.1016/J.EJOR.2004.06.013.
[15] C. S. Ong, J. J. Huang, and G. H. Tzeng, "Building credit scoring models using genetic programming," *Expert Syst. Appl.*, vol. 29, no. 1, pp. 41–47, Jul. 2005, doi: 10.1016/J.ESWA.2005.01.003.
[16] A. Vellido, P. J. G. Lisboa, and J. Vaughan, "Neural networks in business: a survey of applications (1992–1998)," *Expert Syst. Appl.*, vol. 17, no. 1, pp. 51–70, Jul. 1999, doi: 10.1016/S0957-4174(99)00016-0.
[17] B. K. Wong and Y. Selvi, "Neural network applications in finance: A review and analysis of literature (1990–1996)," *Inf. Manag.*, vol. 34, no. 3, pp. 129–139, Oct. 1998, doi: 10.1016/S0378-7206(98)00050-0.
[18] M. R. Dewi, "Klasifikasi akses internet oleh anak-anak dan remaja dewasa di Jawa Timur menggunakan support vector machine," *J. Ris. dan Apl. Mat.*, vol. 4, no. 1, pp. 17–27, 2020.
[19] T. Oribel and D. Hanggraeni, "An Application of Machine Learning in Financial Distress Prediction Cases in Indonesia," *Int. J. Bus. Technol. Manag.*, vol. 3, no. 2, pp. 98–110, 2021.
[20] U. Z. Nisa, B. Santosa, and S. E. Wiratno, "Model Prediksi Finansial Distress Pada Perusahaan Manufaktur Go Public di Indonesia," in *Prosiding Seminar Nasional Manajemen Teknologi*, 2013, vol. 18, pp. 1–8.
[21] N. Santoso and W. Wibowo, "Financial distress prediction using linear discriminant analysis and support vector machine," in *Journal of Physics: Conference Series*, 2018, vol. 979, no. 1, p. 12089.
[22] S. Gupta, B. Parekh, and A. Jivani, "A Hybrid Model of Clustering and Classification to Enhance the Performance of a Classifier," 2019, pp. 383–396. doi: 10.1007/978-981-15-0111-1_34.
[23] S. Cui, Y. Wang, Y. Yin, T. C. E. Cheng, D. Wang, and M. Zhai, "A cluster-based intelligence ensemble learning method for classification problems," *Inf. Sci. (Ny).*, vol. 560, pp. 386–409, Jun. 2021, doi: 10.1016/J.INS.2021.01.061.
[24] H. Gan, N. Sang, R. Huang, X. Tong, and Z. Dan, "Using clustering analysis to improve semi-supervised classification," *Neurocomputing*, vol. 101, pp. 290–298, Feb. 2013, doi: 10.1016/J.NEUCOM.2012.08.020.
[25] J. Xiao, Y. Tian, L. Xie, X. Jiang, and J. Huang, "A Hybrid Classification Framework Based on Clustering," *IEEE Trans. Ind. Informatics*, vol. 16, no. 4, pp. 2177–2188, Apr. 2020, doi: 10.1109/TII.2019.2933675.
[26] N.-C. Hsieh, "Hybrid mining approach in the design of credit scoring models," *Expert Syst. Appl.*, vol. 28, no. 4, pp. 655–665, 2005.
[27] C. Cortes, V. Vapnik, and L. Saitta, "Support-vector networks," *Mach. Learn. 1995 203*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[28] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999, doi: 10.1145/331499.331504.
[29] C. F. Tsai and M. L. Chen, "Credit rating by hybrid machine learning techniques," *Appl. Soft Comput.*, vol. 10, no. 2, pp. 374–380, Mar. 2010, doi: 10.1016/J.ASOC.2009.08.003.
[30] M. Bekkar, D. Kheliouane Djemaa, and D. Akrouf Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," vol. 3, no. 10, 2013, Accessed: Oct. 21, 2022. [Online]. Available: www.iiste.org
[31] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," in *IOP conference series: materials science and engineering*, 2021, vol. 1099, no. 1, p. 12077.