Business Failure Prediction by Hybrid Data Mining Approach: A Case of Thailand Agribusiness

Jeerawadee Pumjaroen[1,*]

Received: 29 January 2023; Revised: 30 April 2023; Accepted: 16 May 2023

Abstract

The failure of a business could significantly impact private companies, the government, and the whole economy. Therefore, predicting business failure is always one major research problem in business and economics. Many methods, such as theoretical models, statistical models, and data mining techniques, were applied to predict business failures. This research developed a business failure prediction model to classify failed and non-failed companies from one to three years before the failure by a hybrid data mining technique. The interest of this research is to integrate clustering and classification techniques to predict business failure, which can be beneficial for further research related to business failure prediction or early warning models. The study involved 3,118 agribusiness companies that submitted their financial statements from 2016 to 2020 in Thailand. Based on the data of financial statements, a single classifier, including decision tree (DT), logistic regression (LR), and neural network (NN), was compared with a hybrid data mining technique—clustering and classification. The results showed that applying the hybrid method, k-mean and DT, helped to improve the business failure prediction performance.

Keywords: data mining, business failure, prediction, early warning model, classification

[*] *Corresponding Author Email : jeerawadee_p@rmutt.ac.th*
[1] Applied Statistics, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT)

1. Introduction

Business failure prediction has been a topic of interest and tends to gain more attention because of its importance to many parties—the private companies, the government, and the whole economy (Tsai, 2014; Wanke, Barros, & Faria, 2015). The failure of a business can cause social costs and the economy involving entrepreneurs, workers, lenders, suppliers, and clients, which are related to increased debt. Rising household debt possibly negatively affected the nation's economy (Pumjaroen, 2019; Pumjaroen, Vichitthamaros, & Sethapramote, 2020).

Both private and government sectors will benefit if the prediction is reliable (Geng, Bose, & Chen, 2015; Klepáč & Hampel, 2017; Lin & McClean, 2001; Pumjaroen & Sethapramote, 2023). Since business failure does not suddenly happen just in one day, the loss is a gradual process that evolves over a considerable period (Altman, 1968; Lincoln, 1984; Webb, 2003). Hence, the firms and the policymakers notice the early sign of business failure; they could initiate remedial measures to avoid deterioration before the collapse. The firm could make a strategic plan to reform itself; the investors have information to better manage their investment portfolios by reallocating the stocks' funds and avoiding companies that are about to fail. Financial sectors related to investment and lending would better assess the default risk Gepp, Kumar, and Bhattacharya (2010); (Tsai, 2014). The government policymaker will have a longer time to find a strategy for healing the economy. Therefore, accurate business failure prediction models would help the economy stabilize and increase for all involved.

Many empirical studies apply business failure as the outcome of financial distress (Geng et al., 2015). Since if the firm's financial health is weakened, it deals with financial distress, developing into a financial crisis and ending as a failure. Because financial analysis informs the company's standing, it helps identify scenarios leading to its financial health and management's quality decisions (Klepáč & Hampel, 2017).

There are many aliases in the area of business failure prediction, such as financial distress prediction, bankruptcy prediction, and business failure prediction. However, they are all related to the model aiming to predict the failure of a business before it actually happens. Some models apply only the available financial information associated with the company. Many studies include other attributes, such as industrial and economic indicators (Geng et al., 2015).

Since agribusiness is one of the driving sectors of Thailand's economy, this research aimed to develop its business failure prediction model to classify failed and non-failed companies. Data mining techniques have become popular alternatives in business failure prediction (Li, Sun, & Wu, 2010). Most of the business failure predictions in Thailand were conducted by a single classifier (Buanak, 2016; Narungsri, 2005). However, many recent studies indicate that a hybrid method, which combines multiple approaches, outperforms a single process (Alapati & Sindhu, 2016; Jader & Aminifar, 2022; Shalaby, Belal, & Omar, 2021). Moreover, some related studies of business failure prediction found that a hybrid method helps improve the model performance over using only a single process (Hsieh, 2005; West, Dellana, & Qian, 2005). Hence, this research focused on comparing the prediction performance of single and hybrid classified methods with different previous years' early warnings before the business failure. The interest of this research is to integrate the clustering and classification techniques to predict Thailand's agribusiness failure, which can be beneficial for further research related to the model of business failure prediction or early warning model. The result showed whether applying the clustering technique before classification helped improve Thailand's agribusiness business failure prediction.

The remainder of the article is organized as follows—Section Two reviews related studies on a business failure description and prediction model using data mining techniques. The research steps and empirical data collected for agribusiness failure prediction are described in Section Three. The results are discussed in Section Four. Finally, Section Five presents the conclusions of the empirical results and their implications.

2. Conceptual Background

2.1 Description of Business Failure

Many terms are used to describe when firms face financial difficulties, such as financial distress, business bankruptcy, firm default, and business failure (Geng et al., 2015). As for business failure, Dimitras, Zanakis, and Zopounidis (1996) defined it as a situation in which the firms could not continually operate.

The financial statement has long been applied to predict business failure. Since it is the accounting report assessing a firm's information focusing on the financial situation and the results of its operations, the financial statement helps provide useful information for making economic decisions. Altman (1968) was among the first researchers who used financial ratios to predict corporate bankruptcy.

In Thailand, a registered juristic person must report a financial statement to the Department of Business Development (DBD), Ministry of Commerce every year (Development, 2020). If the firm does not submit the financial report for consecutively three years, the officer will label its status as an unoccupied firm. In case the company cannot operate the business, it must report to the department for liquidation.

As for those mentioned above, the firm that registers the liquidation or does not report the financial statements for three consecutive years will be defined as a business failure for this research.

2.2 Business Failure Prediction Models

The business failure prediction model could be classified into three groups: statistical models (such as discriminant and regression analysis), data mining techniques (such as neural networks and decision trees), and theoretical models (such as expert evaluation and market risk models) (Klepáč & Hampel, 2017). Nowadays, many studies have applied hybrid methods to enhance prediction accuracy (Alapati & Sindhu, 2016; Klepáč & Hampel, 2017; Tsai & Chen, 2010). The procedure combines more than one model from statistical and machine learning techniques to improve prediction performance. Recent studies of business failure prediction have also used a hybrid method and confirmed that combining multiple approaches outperforms a single process (Hsieh, 2005; West et al., 2005). More specifically, pre-classification by clustering technique before classification helps improve predictive performance (Hsieh, 2005; Tsai, 2014).

2.2.1 Classification

Classification is considered a supervised technique since it learns by a set of examples, acting as a supervisor. The supervisor has the expertise of the environment, represented by a group of input-output examples. As for classifying unknown patterns, the process tries to generate a classifier or a model from a training sample set for each class during the learning task. The learning process tries to map between the input-output examples and correctly label the training set. Consequently, the model obtained from the training process is applied to classify an unknown pattern into each class. The classification techniques used frequently for business failure are decision tree (DT), logistic regression (LR), and neural network (NN) (Kumar & Ravi, 2007; Tsai, 2014).

The first statistical model for business failure prediction was published in the 1960s by Beaver (1966), which was a univariate model. As for the multivariate technique, Altman (1968) proposed the application of discriminant analysis. However, the method was popular in business failure prediction before the 1980s (Geng et al., 2015) because of its stick assumptions, such as linear separability, multivariate normality, and equal covariances. LR was suggested in the business failure area to overcome the discriminant limitations. Martin (1977) pioneered the logistic model to forecast bank failure. LR is mainly applied to forecast binary or multi-class response attributes. The model cannot be generated by linear regression directly since the dependent variable is a nominal type. Instead of predicting a point estimate of the event, it predicts the odds of the event. The logistic model has been widely used for classification problems; however, its restrictive assumptions make data mining techniques more popular in real-world prediction problems (Olson, Delen, & Meng, 2012).

Data mining techniques have influenced the area of business failure prediction since the 1990s. NN and DT are the popular technique for this area (Klepáč & Hampel, 2017).

NN is a nonlinear mathematical approach to the data mining technique. The method comprises a group of neural nodes linking with the weighted nodes. Each node can simulate a neuron of creatures, and the connection among these nodes is equal to the synaptic that connects among the neurons. Generally, neural networks consist of input, hidden, and output layers. Each layer is connected, of which input units will be connected to hidden units, and the hidden units will later be joined to the output units.

DT is a non-parametric prediction method as a recursive partitioning using a divide and conquers technique to distinguish instances. Its structure comprises a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. The decision tree aims to divide the unknown data recursively until every data belongs to a specific class. The algorithm for the decision tree is generally implemented in two phases. Tree-building is the first phase; in that phase, the tree is divided until all the data have its class in a top-down fashion. The second phase is tree pruning, where predictions and accuracy are improved bottom-up.

2.2.2 Clustering

Clustering is an unsupervised technique that differs from classification. The approach aims to arrange a given data of unlabelled input patterns into meaningful clusters based on a measure of similarity (Jain, Murty, & Flynn, 1999). Pattern clustering results in some well-separated groups in the feature space, which summarise and visualize data in the given collection. Algorithms in the clustering technique can be grouped into two categories-- hierarchical and non-hierarchical (or partitional) clustering. The hierarchical clustering algorithm generates a hierarchy of clusters, in which a distinct singleton cluster will be combined one by one until satisfied some thresholds. The result will produce a series of arborescence partitions. The work of Beranová, Basovníková, and Martinovičová (2013) applied this method to cluster agricultural enterprises. Some might prefer partitional clustering, especially in business problems (Dissayarungkun, 2021; Olafsson, Li, & Wu, 2008). K-means is one of the well-known partitional clustering algorithms since it procedures with a simple algorithm but gives efficient clustering results (Tsai & Chen, 2010). The method aims to partition the data set into k clusters in which each observation belongs to the cluster with the nearest mean (centroid). The algorithm computes centroids and repeats until the optimal centroid is found. The process starts with selecting the k point as the initial centroids. Each observation in the data set will be assigned to the closest centroid and then recomputed the new centroids of each cluster. The process reassigns each observation regarding their nearest new centroids. Repeatedly the process until some convergence criterion is satisfied or the assignment is not changed.

3. Research Methodology

The research follows the important steps of the CRoss Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000), which is the standard data mining process. This study focuses on data understanding, preparation, modeling, and evaluation. RapidMiner Studio Version 9.10 was used for the data analysis processing.

3.1 Data Understanding and Preparation

The research focused on predicting agribusiness failure in Thailand based on the annual financial statement collected from the Department of Business Development database. Regarding The Thailand Standard Industrial Classification (TSIC), agribusiness is classified in A--Agriculture, Hunting, Forestry, and Fishing. This research studied 3,118 companies that submitted their financial statements from 2016 to 2020 (The recent financial statement data collected was in 2019). The input variables were related to 12 kinds of financial statements and firm information attributes, including business size, company age, two consecutive years of profit or loss, recently registered capital, working capital, corporate income tax, revenue, authorized capital stock, total assets, current asset, liability, and operating profit margin.

3.2 Number of Failed Business Status

Since the research aimed to predict business failure in 2020, the target attribute was a nominal type--the "Failure" or "Non-Failure" status in 2020. There were 100 failed companies in the study, including those registered for liquidation or labeled as unoccupied firms. The rest of them are labeled as non-failure firms, including 3,018 companies.

3.3 The Time Span of the Data Set

The previous studies in Thailand used the financial statements data obtained during the last three years before the failure, which achieved the accuracy of training data of about 70% to 80%. Buanak (2016) forecasted bankruptcies of small and medium-sized enterprises (SMEs) in Thailand ahead of one to three years before failure happened. An imbalanced sample of 604 firms from 2008 to 2015, including 71 failed companies and 532 active enterprises, was applied in the training model. The study compared the accuracy of Altman, Zmijewski, and discriminant models. The finding showed that discriminant analysis gave the most accuracy of training data with 77.45%,75.80%, and 81.16% for one, two, and three years ahead. Narungsri (2005) applied logistic regression to predict financial failure for small and medium companies from one to three years ahead. The study estimated the model by the financial statement from 1999 to 2001. The sample size was 671 firms containing 321 failed companies and 350 active firms. The selected model was applied to predict the business failure of testing data, including 55 companies for each group of failed and non-failed. The result showed that the one-year-ahead prediction was better than two and three years ahead, which achieved the accuracy of training data at 72.70%, 69.27%, 67.79%, and for testing data at 71.82%, 68.87%, and 66.98% consecutively.

This research also implemented the model based on the financial statement obtained for three previous years before the companies were labeled as business failures in 2020. For example, in an early warning for one year (t-1), the study used financial data by 2019 to predict whether the firm was marked as a failure in 2020. As for the two-year prediction (t-2) of 2020, the financial statement before 2018 was applied to the model. The model used only the financial information before 2017 to predict a three-year earlier failure (t-3) in 2020.

3.4 Data Preparation

As the initial data set, the business status in 2020 and the financial statement were combined. The missing values were remedied by interpolation. Stratified random sampling was applied to partition the data into training and testing data sets. The training data set was applied to generate the learning models, while the testing data set was used to test the models' predictive ability. The random partitions of the training and testing data sets might impact the analysis; hence, we conduct both single and multiple splits.

Regarding the single split, to avoid undertraining or overtraining, we applied the training ratios at 0.6, 0.7, and 0.8. As for the training data set, it worked with a balanced sample size. For instance, of 0.6 split ratios, this training data set contained 120 observations, including 60 failed firms and 60 non-failed firms, since the observations labeled as failed firms totally in the data set were 100 firms. As for the testing data set, the number of data was 2,998 firms—40 failed firms and 2,958 non-failed firms, which the data did not use for generating the model.

The research also conducted multiple splits or cross-validation, a resampling procedure, to evaluate the models with ten folds. Cross-validation aims to reduce the bias associated with the random sampling of the training and holdout data samples. The analysis randomly selected 100 failed firms and 100 non-failed firms regarding a balanced training data set. This input data was partitioned into ten subsets of equal size at 20 observations. A single subset was retained as the test data set (20 observations) of those ten subsets. The remaining nine subsets worked as a training data set (180 observations). The cross-validation process was then iterated ten times, with each of the ten subsets used exactly once as the test data. The ten results from the ten iterations were combined to produce the estimation.

This training data set of this study might be quite small for general problems using the data mining technique; however, as for the limitation of data in the fields of business failure prediction, this training sample size was acceptable. As seen in many works related to business failure prediction, such as the work of Geng et al. (2015); Klepáč and Hampel (2017), who applied data mining models such as LR, NN, and DT, also used around the same size of this research for the training data set.

3.5 Modelling

In the first step of data modeling, we generated single-classified models. LR, NN, and DT, well-known and widely used classification methods in business failure prediction, were conducted. Regarding the RapidMiner operator, as for LR, the logistic regression operator was used. This operator applied an S-shaped curve formed by the logit transformation representing the probability of an event—Failure Non-Failur. The decision tree operator was applied for DT, in which the algorithm collects nodes to identify a decision on values affiliation to a class of business status. Each node represents a splitting rule for one variable. The new nodes were repeatedly created until meeting the stopping criterion. A prediction of a business status was identified based on the majority of Examples that reached this leaf during generation. The Neural Net operator was conducted for NN, which learns a model through a feed-forward neural network trained by a multi-layer perceptron. Hereafter, the outperform model will be selected regarding the predictive performance (Figure 1).

Secondly, the research conducted the hybrid classifier by combining the clustering method in the selected classifier to improve business failure prediction (Figure 2). The k-mean clustering is one of which the well-known clustering in business fields (Tsai & Chen, 2010) was applied for this task. Then we compared the predictive performance of the selected model from the first step and the hybrid model from the second step.

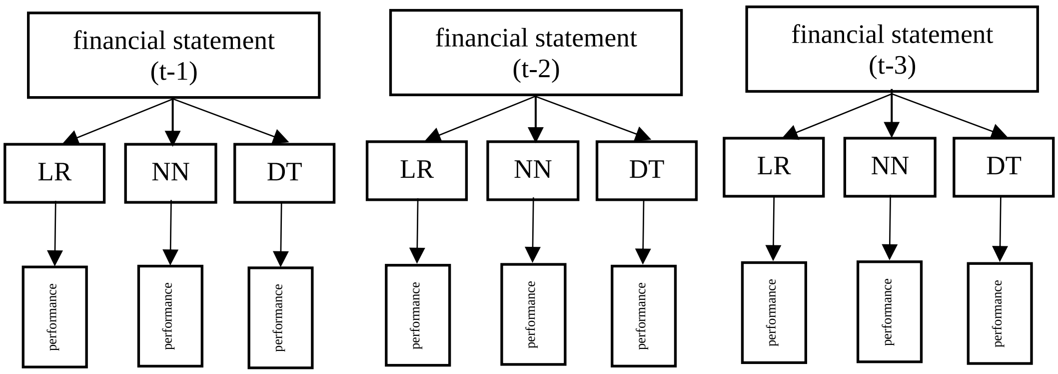

Figure 1: A single method for business failure prediction

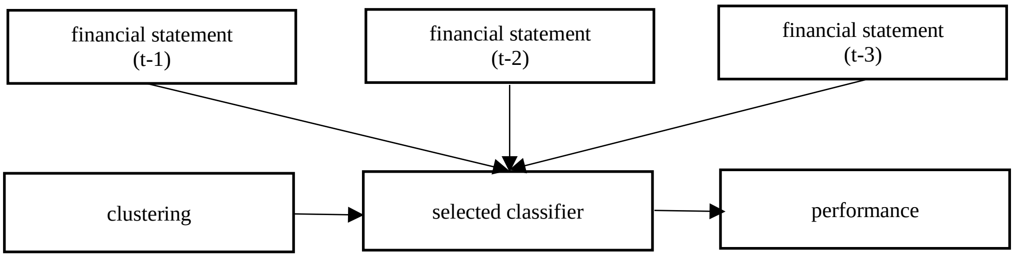

Figure 2: A hybrid method of clustering combined with classification for business failure

3.6 Evaluation

The research evaluated the model performance in terms of accuracy, recall, and precision, the common evaluation metrics of machine learning. (Davis & Goadrich, 2006; Geng et al., 2015).

The accuracy is the ratio of correctly predicted observations, including failure and non-failure, to the total number of observations. Whereas recall is the ratio of correctly predicted failure observations to the total number of actual failure observations, precision is the ratio of correctly predicted failure observations to the total number of predicted failure observations. The confusion matrix displays the classification prediction performance in Table 1.

Table 1: Confusion matrix for business failure prediction

		Actual	
		T (failure)	F (non-failure)
Predict	T (failure)	TP	FP
	F(non-failure)	FN	TN

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

4. Results and Discussion

The predictive performance of LR, DT, and NN are shown in Tables 2 and 3. The tables display the accuracy, recall, and precision based on the training ratios of 0.6, 0.7, 0.8, and ten-fold cross-validation of the early warning for t-1, t-2, and t-3.

Table 2 shows the single classifier performances of LR, NN, and DT to predict business failure. The results varied according to the performance evaluation criterion, the training ratios or the cross-validation, the early warning period, and the training or testing data set. For example, the accuracy of cross-validation for t-1 in the training data set, the result of LR was outstanding, gaining at 87.67%; however, in the testing data, NN was outperformed at 70.00%. Hence, the research considered the number of models achieving better performance in the last section of Table 2. However, please note that an accuracy rate is generally applied to evaluate the correctly classified sample number to the total sample number of a balanced data set. However, the data set of business failure is highly class-imbalanced; the accuracy rate measure cannot evaluate model performance for the minority class well (Sun, Li, Fujita, Fu, & Ai, 2020). The purpose of the business failure prediction is to capture the failure of a business (Klepáč & Hampel, 2017), which is the minority class of the data set. Hence, this research focused on the model with the highest recall and precision rates because they concentrate on correctly predicted failure business.

Figure 3 shows the recall and precision of each single technique, whereas the last row of Table 2 focuses on the number of models gaining better recall and precision. The results indicated that DT achieved better performance in both training and testing data sets (except for t-2, DT was equal to LR for training, and DT was not different from NN for the testing data set). Many research suggested applying black-box algorithms such as NN and the support vector machines method (SVM) to predict business failure; however, those techniques are less comprehensible by human users than DT (Olson et al., 2012). Our results also were consistent with Klepáč and Hampel (2017). They indicated that DT offers better accuracy for agriculture companies in the EU than LR and SVM. In addition, the study of Olson et al. (2012) related to bankruptcy prediction found that DT was relatively more accurate than NN and SVM.

Since this research aimed to improve business failure prediction by applying the hybrid classifier--segmentation before classification, the k-mean clustering was applied. K-mean separated the data into subgroups based on the company's age and the two consecutive years of profit or loss before conducting the DT classification method to improve the business failure prediction performance. However, k-mean is the non-hierarchical method; identifying the number of groups or k is needed before using the technique. Therefore, the research applied the hierarchical method with agglomerative algorithms to roughly set the number of clusters. The study decided to apply k=3 for the k-mean clustering. In addition, as the suggestion by Hair, Black, Babin, and Anderson (2014), the cluster solution was validated by considering the variable not included in the cluster analysis—business size (small, medium, and large) to ensure the practical significance of the cluster solutions. The results showed the relevant reasons. For example, the largest group of the clustering analysis trended to be the companies that had not conducted business for a long time and had two consecutive years of loss; most of the companies in the group were small-size businesses.

Table 3 displays the comparison of the predictive performance of DT and hybrid DT to predict business failure. For the same reason mentioned above, this part also considered only recall and precision. As for the training data set, clustering before classification helped improve classification performance for t-1 and t-2; however, the performance did not differ for t-3. As for the testing data set, clustering before classifying helped improve classification performance for t-1 and t-3, but as for t-2, it gained lower performance (The last row of Table 3 and Figure 4). Moreover, the result found that long-time forecasting seemed to decrease the model prediction performance, which was according to the previous studies of Klepáč and Hampel (2017); Narungsri (2005).

Table 2: Model performance of DT, LR, NN

		Training data									Testing data								
	Splits	t-1			t-2			t-3			t-1			t-2			t-3		
		LR	NN	DT	LR	NN	DT	LR	NN	DT	LR	NN	DT	LR	NN	DT	LR	NN	DT
Accuracy	0.6	**100.00**	88.33	88.33	**95.83**	90.83	88.33	**90.83**	86.67	85.83	58.30	**67.20**	54.37	49.72	**55.33**	50.12	58.54	**77.23**	50.12
	0.7	**97.14**	90.00	83.57	**88.57**	86.43	77.14	**83.57**	**83.57**	72.86	64.71	**76.04**	50.48	**63.32**	54.01	44.49	63.96	**75.72**	37.11
	0.8	85.00	**86.25**	83.75	81.25	**84.38**	80.00	80.63	**82.50**	67.50	65.06	**76.12**	55.93	**66.19**	62.02	55.93	**70.51**	**70.51**	29.33
	Cross-validation	**87.67**	82.11	83.44	**81.56**	79.28	79.72	79.67	78.83	**80.06**	64.50	**70.00**	67.00	61.00	**66.00**	65.00	**68.50**	68.00	65.00
Recall	0.6	**100.00**	86.67	**100.00**	96.67	95.00	**98.33**	91.67	81.67	**98.33**	57.50	60.00	**87.50**	55.00	**77.50**	75.00	52.50	55.00	**72.50**
	0.7	98.57	82.86	**100.00**	87.14	95.71	**100.00**	84.29	78.57	**100.00**	60.00	46.67	**73.33**	50.00	83.33	**93.33**	50.00	46.67	**93.33**
	0.8	83.75	82.50	**98.75**	82.50	93.75	**96.25**	77.50	83.75	**100.00**	50.00	45.00	**85.00**	55.00	80.00	**95.00**	50.00	55.00	**95.00**
	Cross-validation	87.44	85.56	**97.00**	85.00	84.22	**96.11**	85.22	84.56	**95.44**	71.00	76.00	**79.00**	68.00	74.00	**81.00**	75.00	75.00	**82.00**
Precision	0.6	**100.00**	89.66	81.08	**95.08**	87.69	81.94	90.16	**90.74**	78.67	4.37	5.76	**5.84**	3.49	**5.35**	4.67	4.05	**7.64**	4.53
	0.7	95.83	**96.67**	75.27	**89.71**	80.72	68.63	83.10	**87.30**	64.81	5.36	**6.31**	4.61	4.37	**5.56**	5.14	4.45	**6.22**	4.56
	0.8	85.90	**89.19**	75.96	**80.49**	78.95	72.64	**82.67**	81.71	60.61	4.59	**6.12**	5.88	5.16	6.43	**6.48**	5.43	**5.91**	4.14
	Cross-validation	**87.83**	80.04	76.31	**79.52**	76.64	72.38	**76.70**	75.87	72.98	62.83	**67.86**	63.71	59.65	**63.79**	61.36	**66.37**	65.79	61.19
number of models gaining better performance	Accuracy	**3**	1	0	**3**	1	0	**2**	**2**	1	0	**4**	0	**2**	**2**	0	2	**3**	0
	Recall	1	0	**4**	0	0	**4**	0	0	**4**	0	0	**4**	0	1	**3**	0	0	**4**
	Precision	**2**	**2**	0	**4**	0	0	**2**	**2**	0	0	**3**	1	0	**3**	1	1	**3**	0
	Recall and Precision	3	2	**4**	**4**	0	**4**	2	2	**4**	0	3	**5**	0	**4**	**4**	1	3	**4**

Note: The highlight is the better performanc

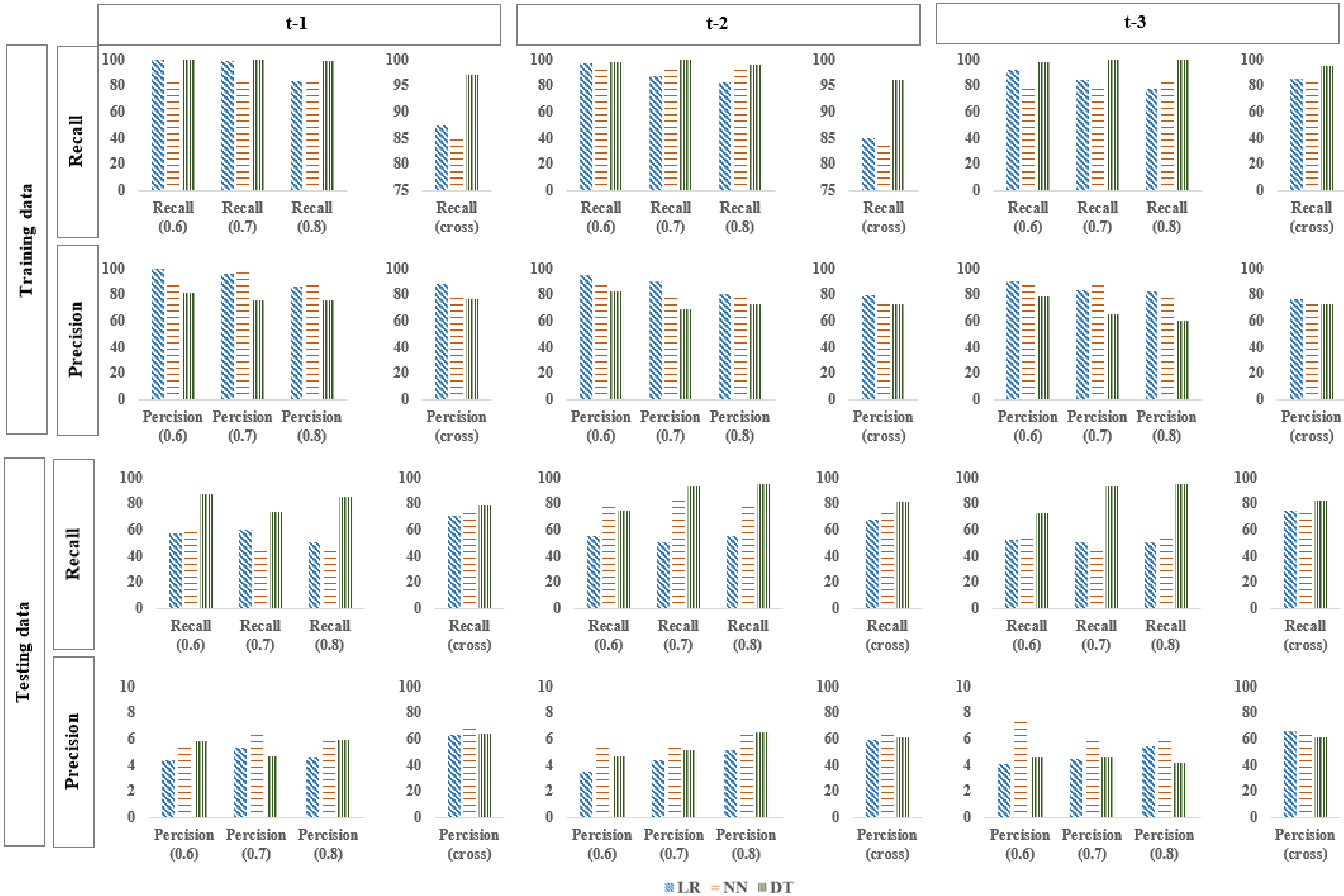

Figure 3: Recall and precision of DT, LR, NN
Note: () Training ratio or cross-validation

Table 3: Comparing the predictive performance of the single classifier and hybrid classifier of DT

	Splits	t-1		t-2		t-3		t-1		t-2		t-3	
		Single	Hybrid	Single	Hybrid	Single	Hybrid	Single	Hybrid	Single	Hybrid	Single	Hybrid
Accuracy	0.6	88.33	**89.17**	**88.33**	86.67	**85.83**	82.50	54.37	**57.74**	50.12	**54.45**	**50.12**	47.15
	0.7	83.57	**88.03**	77.14	**83.80**	72.86	**78.87**	50.48	**53.96**	44.49	**53.10**	37.11	**44.22**
	0.8	83.75	**88.13**	80.00	**83.75**	67.50	**79.38**	55.93	**60.03**	55.93	**56.66**	29.33	**58.91**
	Cross-validation	83.44	**88.22**	79.72	**86.44**	**80.06**	77.28	67.00	**68.50**	65.00	**70.50**	**65.00**	57.50
Recall	0.6	**100.00**	**100.00**	98.33	**100.00**	98.33	**100.00**	**87.50**	85.00	75.00	**87.50**	72.50	**85.00**
	0.7	**100.00**	**100.00**	**100.00**	97.18	**100.00**	98.59	73.33	**89.66**	**93.33**	75.86	**93.33**	86.21
	0.8	**98.75**	**98.75**	96.25	**98.75**	**100.00**	91.25	**85.00**	**85.00**	**95.00**	80.00	**95.00**	70.00
	Cross-validation	97.00	**97.44**	**96.11**	94.22	95.44	**99.89**	**79.00**	**79.00**	**81.00**	77.00	**82.00**	**82.00**
Precision	0.6	81.08	**82.19**	**81.94**	78.95	**78.67**	74.07	5.84	**6.13**	4.67	**5.85**	4.53	**4.95**
	0.7	75.27	**80.68**	68.63	**76.67**	64.81	**70.71**	4.61	**5.74**	**5.14**	4.86	4.56	**4.61**
	0.8	75.96	**81.44**	72.64	**75.96**	60.61	**73.74**	5.88	**6.46**	**6.48**	5.67	4.14	**5.30**
	Cross-validation	76.31	**82.27**	72.38	**81.54**	**72.98**	68.78	63.71	**65.29**	61.36	**68.14**	**61.19**	55.03
number of models gaining better performance	Accuracy	0	**4**	1	**3**	2	2	0	**4**	0	**4**	2	2
	Recall	3	**4**	2	2	2	2	3	3	**3**	1	**3**	2
	Precision	0	**4**	1	**3**	2	2	0	**4**	2	2	1	**4**
	Recall and Precision	3	**8**	3	**5**	4	4	3	**7**	**5**	3	4	**6**

Note: The highlight is the better performance

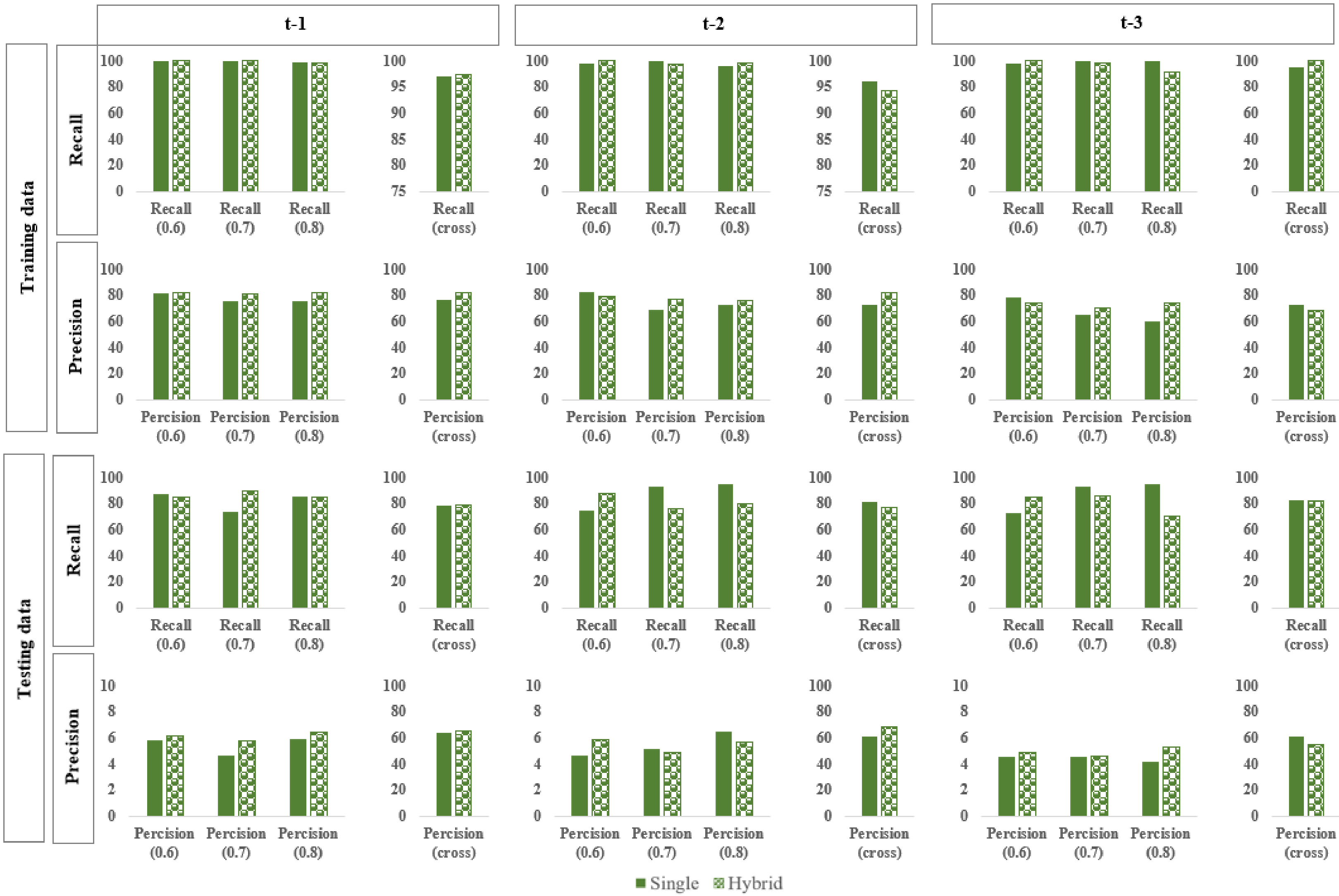

Figure 4: Recall and precision of the single classifier and hybrid classifier of DT
Note: () Training raio or cross-validation

5. Conclusions

Business failure can significantly impact many parties—the private companies, the government, and the whole economy. Therefore, predicting business failure is always one major research problem in business and economics. This research aimed to improve business failure prediction for Thailand agribusiness using hybrid classifiers. This case studied 3,118 companies submitting their financial statements from 2016 to 2020, and there were 100 failed companies in the study, including those registered for liquidation or labeled as unoccupied firms.

As for the single classifier, focusing on recall and precision, the empirical result indicated that DT performs better than LR and NN. This result was consistent with Klepáč and Hampel (2017), who showed DT offers better accuracy for agriculture companies in the EU than LR and the SVM. Moreover, Olson et al. (2012), whose study related to bankruptcy prediction, suggest that DT was relatively more accurate than NN and SVM. Although much research concludes NN obtained more accuracy than DT, it is less comprehensible by human users than DT (Olson et al., 2012).

Since the research focused on comparing the prediction performance of single and hybrid classified methods, segmentation and classification were applied to improve the business failure prediction. The k-mean clustering technique was applied to separate the data into three groups before using the DT classification method. The results of the hybrid classifier of DT were not consensus for all t-1, t-2, and t-3. As for the testing data set, the hybrid helped improve classification performance for t-1 and t-3, but for t-2, it gained lower performance than a single DT. However, when considering the overall of all periods of t-1, t-2, and t-3, The results indicated that hybrid classifiers helped improve classification performance. Clustering before classification was beneficially supported by the studies of Alapati and Sindhu (2016) and Tsai (2014). Nevertheless, the results indicated that increasing the previous year of early warning before the failure will decrease the accuracy of the business failure prediction model, which is consistent with earlier studies by Klepáč and Hampel (2017) and Narungsri (2005).

In conclusion, in this case, the results suggest applying the hybrid method of clustering and classification to the work of business failure prediction. Even though, in numerous instances, clustering can help improve classification performance; however, it does not always improve the quality of classification (Piernik & Morzy, 2021).

For future work, I suggest considering these issues. First, more rule nodes were desired to improve the NN performance; however, the compensation of this task is the more complex model. Second, several clustering techniques can be pre-processed before classification for comparisons. Third, other advanced classification methods could be regarded as the hybrid method in the same way. Forth, different data sets with higher failed firms should be applied for this work.

References

Alapati, K. Y., & Sindhu, K. (2016). Combining Clustering with Classification: A Technique to Improve Classification Accuracy. **International Journal of Computer Science Engineering (IJCSE)**, 5(6).

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **The journal of finance**, 23(4), 589-609.

Beaver, W. H. (1966). Financial ratios as predictors of failure. **Journal of accounting research**, 71-111.

Beranová, M., Basovníková, M., & Martinovičová, D. (2013). Clustering of agricultural enterprises. **Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis**, 61(2), 289-296.

Buanak, S. (2016). **A Bankruptcy Test for Small and Medium-Size Enterprises (SMEs) in Thailand**. Thammasat University, Thailand.

Davis, J., & Goadrich, M. (2006). **The relationship between Precision-Recall and ROC curves**. Paper presented at the Proceedings of the 23rd international conference on Machine learning.

Development, D. o. B. (2020). **Manual for Submitting Financial Statements**. DBD. Retrieved from https://www.dbd.go.th/download/document_file/balance/manual_finacial2563.pdf.

Dimitras, A. I., Zanakis, S. H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. **European Journal of Operational Research**, 90(3), 487-513.

Dissayarungkun, P. (2021). Cluster Analysis of Small Enterprises According to a Risk of Business. **Journal of Applied Statistics and Information Technology**, 6(2), 1-16.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. **European Journal of Operational Research**, 241(1), 236-247.

Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. **Journal of forecasting**, 29(6), 536-555.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis: Pearson new international edition. **Essex: Pearson Education Limited**, 1(2).

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. **Expert Systems with Applications**, 28(4), 655-665.

Jader, R., & Aminifar, S. (2022). Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach. **Applied Computational Intelligence and Soft Computing**, 2022.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. **ACM computing surveys (CSUR)**, 31(3), 264-323.

Klepáč, V., & Hampel, D. (2017). Predicting financial distress of agriculture companies in EU. **Agricultural Economics-Zemedelska Ekonomika**.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. **European Journal of Operational Research**, 180(1), 1-28.

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. **Expert Systems with Applications**, 37(8), 5895-5904.

Lin, F. Y., & McClean, S. (2001). **A data mining approach to the prediction of corporate failure**. In Applications and Innovations in Intelligent Systems VIII, 93-106.

Lincoln, M. (1984). An empirical study of the usefulness of accounting ratios to describe levels of insolvency risk. **Journal of banking & finance**, 8(2), 321-340.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. **Journal of banking & finance**, 1(3), 249-276.

Narungsri, K. (2005). **Financial Failure Prediction for Small and Medium Enterprises in Thailand**. Suranaree University of Technology, Thailand.

Olafsson, S., Li, X., & Wu, S. (2008). Operations research and data mining. **European Journal of Operational Research**, 187(3), 1429-1448.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. **Decision Support Systems**, 52(2), 464-473.

Piernik, M., & Morzy, T. (2021). A study on using data clustering for feature extraction to improve the quality of classification. **Knowledge and Information Systems**, 63(7), 1771-1805.

Pumjaroen, J. (2019). **Early warning system for real economy: a case study of Thailand**. (Doctoral Dissertation). National Institute of Development Administration,

Pumjaroen, J., & Sethapramote, Y. (2023). Evaluating The Robustness And Implementing Double One-Sided Hodrick-Prescott For Cycle Extraction. **ABAC Journal**, 43(2), 143-160.

Pumjaroen, J., Vichitthamaros, P., & Sethapramote, Y. (2020). Forecasting Economic Cycle with a Structural Equation Model: Evidence from Thailand. **International Journal of Economics and Financial Issues**, 10(3), 47-57.

Shalaby, M., Belal, N. A., & Omar, Y. (2021). Data clustering improves siamese neural networks classification of Parkinson's disease. **Complexity**, 2021, 1-9.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, 5(4), 13-22.

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. **Information Fusion**, 54, 128-144.

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. **Information Fusion**, 16, 46-58.

Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. **Applied soft computing**, 10(2), 374-380.

Wanke, P., Barros, C. P., & Faria, J. R. (2015). Financial distress drivers in Brazilian banks: A dynamic slacks approach. **European Journal of Operational Research**, 240(1), 258-268.

Webb, A. R. (2003). **Statistical pattern recognition**. (2^{nd} ed.). John Wiley & Sons.
West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. **Computers & operations research**, 32(10), 2543-2559.