ICRMAT 2021 Proceedings of the International Conference on Research in Management & Technovation pp. 175–184
DOI: 10.15439/2021KM30
ISSN 2300-5963 ACSIS, Vol. 28

Data Mining for Bankruptcy Prediction: An Experiment in Vietnam

Dang Ngoc Hung
Faculty of Accounting and Auditing
Hanoi University of Industry
298, Cau Dien street, Bac Tu Liem, Hanoi, Vietnam
dangngochung@haui.edu.vn

Vu Thi Thanh Binh
Faculty of Accounting and Auditing
Hanoi University of Industry
298, Cau Dien street, Bac Tu Liem, Hanoi, Vietnam
vuthithanhbinh@haui.edu.vn

***Abstract*—In the history of the world economy, the bankruptcy of some large companies has caused global financial crises. The study aimed to postulate a model of bankruptcy prediction for listed companies on Vietnam's stock market. The research used six popular algorithms in data mining to predict bankruptcy risk with data collected from 4693 observations in the period 2009-2020. The research results showed that Logistic algorithms, Artificial Neural Network, Decision Tree have a high level of predicting bankruptcy with an accuracy of 98%. The study identified the three most important indicators: inventory turnover ratio, debt to equity ratio, and debt ratio that affect the corporate bankruptcy prediction. The study showed the threshold points of 10-indicators to avoid bankruptcy likelihood. These results recommended that the model could be applied in practice to reduce risks for businesses and investors in the Vietnamese market.**

***Index Terms*—Bankruptcy prediction, data mining, Artificial Intelligence, Decision Tree, Z-Altman index.**

I. Introduction

THE REPORT of The World Bank [1] indicates that Vietnam is an active emerging economy with speedy economic growth in the East Asia area. Besides the development of business, the economy has many potential risks. The context is that the global economic growth outlook is somewhat bleak in the face of uncertain potential risks such as the US-China Trade War, Brexit, inflationary trends due to unpredictable price changes. The epidemic of Coronavirus (Covid-19 epidemic) strongly affects people's psychology in general and stock investors in particular. Measuring the health of enterprises in Vietnam is now extremely urgent. There are several models to predict corporate bankruptcy, for example, the model of market approach [2] and the model of accounting approach [3]. Employing models in practice in Vietnam's stock market is essential because of the difficulty of qualitative predictability in the increasingly unpredictable environment. The models support how to measure the bankruptcy prediction from potential business risks. In Vietnam, the quality of accounting information is not too excellent [4] and companies listed or unlisted on the stock market report losses leading to a high risk of bankruptcy. To ensure the rights and benefit of enterprises and creditors, the Law of Vietnam in which the Law on Bankruptcy 2014 and the Law on Securities 2010 (the latest being the Law on Securities 2019 takes effect from January 1, 2021) have issued and concretized these regulations.

Previous studies have given diverse criteria as financial ratios in predicting corporate bankruptcy. Some studies show that the Z-Score model has a strong practical application of financial status to the prediction of bankruptcy as studied by Liang, Lu, Tsai, Shih [5], Barboza, et al. [6], Chou, et al. [7], Antunes, et al. [8], Le, et al. [9], Le, et al. [10], Veganzones and Séverin [11], Mai, et al. [12], Son, et al. [13], Chen, et al. [14]. However, previous studies were mainly used in developed countries to predict bankruptcy and few studies applied data mining in predicting bankruptcy, especially in emerging security markets such as Vietnam.

This study uses several data mining techniques to predict corporate bankruptcy for a Vietnamese case study. The main contributions of this study are as follows*: (i)* building a framework model for predicting bankruptcy, *(ii)* Collecting Vietnam's data sets for the past twelve years for the bankruptcy prediction, *(iii)* testing to compare technical performance for predicting bankruptcy on the Vietnamese dataset; and *(iv)* Combining Bagging and Boosting methods, the test results show the best overall accuracy of 98% to improve forecasting bankruptcy.

Adopting and combining new techniques to improve the accuracy in forecasting corporate bankruptcy is encouraged by researchers and practitioners. The results help to reinforce and enhance the bankrupting prediction model.

II. Literature Review

Research on predicting the financial downturn of companies through Z-score and Zeta models Altman [15], this is a handbook that presents the quantitative techniques commonly used in research papers. empirical finance research along with real, modern research examples. By referring to this handbook, the author has understood and applied it to the study of the Z-score model. Konglai and Jingjing [16] compiled a sample of failed managed groups and normally managed groups that contained 130 listed companies from Shanghai and Shenzhen exchanges in 2009. Using the MDA discriminant analysis model and the logistic model, the author chooses 5 financial factors: profitability, debt repayment ability, operating ability, growth ability, and capital structure. Ohlson [17] was the first to apply the logistic regression model in the study to predict the probability of default of enterprises. Some related studies such as Meeampol, et al. [18] in the Thai stock market. Research by Kumar and Rao [19], on a new method to estimate internal credit risk and predict bankruptcy under the Basel II regime. The results of the study showed that the Z-score could predict bankruptcy with 98.6% accuracy compared with 93.5% according to Altman's score.

Researchers use various algorithms of intelligent techniques to solve the problem of corporate bankruptcy [20]. According to Serrano-Cinca [21] and Fletcher and Goss [22], neural networks (NNs) are the most commonly used technique. And the data mining algorithms used to predict bankruptcy risk include decision trees (DT) and support vector machines (SVM) [23]. A decision tree is a structured hierarchical tree used to classify objects based on a series of rules. When given data about objects containing attributes along with their classes, the decision tree will generate rules to predict the class of the unknown objects (unseen data). Support vector machines (SVM) is a supervised machine learning model used to analyze and classify data. SVM takes incoming data and classifies them into two different classes. Many studies have used data mining techniques in predicting bankruptcy. Some studies related to predicting bankruptcy using data mining techniques are listed in Table 1.

TABLE 1
STUDIES USING DATA MINING FOR BANKRUPTCY PREDICTION

Author	Datasets	Algorithms*	Evaluation metrics	Period
Liang, et al. [5]	Taiwan	SVM, KNN, CART, ANN, NB	Accuracy 82%	1999-2009
Barboza, et al. [6]	USA, Canada	LDA, Logit, NN, SVM, Bagging, Boosting, and RF	Accuracy 87%	1985-2013
Chou, et al. [7]	Taiwan	Fuzzy clustering, BPNN	Accuracy 95.25%	
Antunes, et al. [8]	France	GP, SVM, Logit	Accuracy 94%	
Le, et al. [9]	Korea	RF, DT, MLP, SVM	84.2% (AUC)	2016-2017
Le, et al. [10]	Korea	Cluster-based Boosting, GMBoost, DT, RF, MLP, AdaBoost.	86.8% (AUC)	2016-2017
Veganzones and Séverin [11]	France	LDA, Logit , ANN, SVM, RF	81.1% (Sensitivity)	2013-2014
Mai, et al. [12]	USA	Deep learning embedding S, CNN, SVM, Logit, RF	78.4% (AUC)	1995-2014
Son, et al. [13]	Korea	Logit, RF, XGBoost, LightBM, ANN	88% (AUC)	2011-2016
Chen, et al. [14]	UCI, LibSVM	Bagged-pSVM; Boosted-pSVM	Accuracy 84.42%	

III. METHODOLOGY

A. Measuring Variables

There are many measures for predicting bankruptcy; however, each measure has both advantages and disadvantages. Ghazali, et al. [24] state that the Altman Z-Score is probably the most popular measure of a company's financial health and has been used to determine bankruptcy prediction in numerous studies. This study will determine the bankruptcy prediction based on the Z-score approach of Altman [3]. Altman's Z-score gives a calculation of the Z-score based on the following formula:

$$Z = 0.717 * A1 + 0.847 * A2 + 0.107 * A3 + 0.420 * A4 + 0.998 * A5$$

In which: A1- Current assets minus current liabilities, then divided by total assets; A2 - Retained profit divided by total assets; A3 - Profit before tax and interest divided by total assets; A4 - Book value of equity divided by total liabilities; A5 - Revenue divided by total assets.

If the Z-index < 1.81, the company is in the bankruptcy prediction zone that the likelihood of bankruptcy will be assigned a value of 1. Otherwise, it will be assigned a value of 0.

This study uses 30 attributes of financial indicators including liquidity ratios, capital budgeting ratios, profitability ratios, efficiency ratios (activities ratios), market ratios, and debt ratios (leverage ratios). The properties are briefly described in Appendix 1.

B. Applying Data Mining Algorithms

Data mining has many different expressions. It is the process of automatically extracting valuable information which is predictive information hidden in the huge amount of data in reality. Data mining emphasizes automated and predictive aspects. This study uses Logistic Regression, Bayesian Network, K-nearest neighbor, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Tree that is commonly used to predict bankruptcy.

1) Logistic Regression: The Logistic regression model introduced by Berkson [25] is a commonly used tool in data analysis with binary variables. Some developments by Altman, et al. [26] and Flitman [27] are used in multivariate regression analysis, discriminant analysis. From this binary dependent variable, a procedure will be used to predict the probability of the event occurring according to the rule if the predicted probability is greater than 0.5 (default cut-off point) then the prediction result will be “yes” occurs, otherwise, the predicted result will be given as "no". The Binary Logistic regression model is as follows:

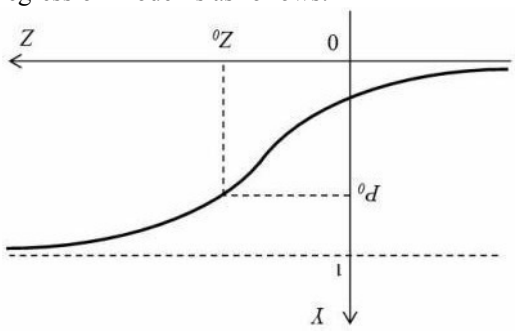

Fig. 1 Binary logistic regression model

P is the probability that Y = 1 when the independent variables take on a particular value. Accordingly, the probability that the event does not occur is:

$$1-P=Prob(Y=0)=1-\frac{e^z}{1+e^z}=\frac{1}{1+e^z}$$

The regression coefficients were estimated by the method of Maximum Likelihood (ML). The logit regression model can be used to estimate the $\log_{(odds)}$ ratio for each independent variable of the model of Ohlson [17]. The parameters β_n were estimated by the method of ML.

2) Bayesian: Bayesian Network is applied for classification based on a probabilistic graphical model and the probability of the Bayesian Network is a value from 0 to 1. Bayesian Network is a set of variables and their conditional dependencies that are linked together by a probability association. According to Carlin and Louis [28], the Bayesian method is more about statistics than regression. For fraud detection, a Bayesian network will be built with Bayesian rule along with the condition P(Y=1) + P(Y=0) = 1 written as follows:

$$P(Y=1 \mid X) = [P(X \mid Y=1)P(Y=1)]/P(X)$$
$$P(Y=0 \mid X) = [P(X \mid Y=0)P(Y=0)]/P(X)$$
$$P(Y=0 \mid X) = [P(X \mid Y=0)P(Y=0)]/P(X)$$

In which:

$$P(X) = P(Y=1)P(X \mid Y=1)+P(Y=0)$$
$$P(X \mid Y=0)$$

The components are calculated as follows: P(Y=1) is the error rate of the sample used to run the model, assuming the variables are independent.

3) K-Nearest Neighbors (K-NN): K-Nearest Neighbors algorithm is used in data mining. K-NN is a method to classify objects based on query points and all the objects in the training data. An object is classified based on its K neighbors. K is a positive integer that is determined before performing algorithms. Euclidean distance is often used to calculate the distance between objects.

4) Artificial Neural Network (ANN): Artificial Neural Network is an information processing model that is simulated based on the activity of the nervous system of an organism. A neural network can consist of one or more neurons that each neuron is an information processing unit and the connections between neurons form a network structure. A neural network is a computational model defined by parameters: Neuron type, connection architecture, and learning algorithms. The neurons are connected by a weight matrix. The typical structure of a neural network consists of three layers: input, hidden, and output [29] (see Fig. 2).

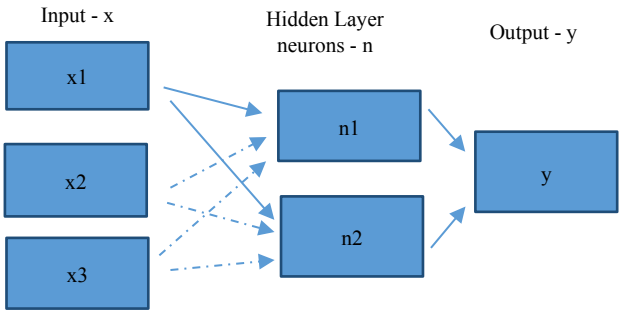

Fig. 2 Artificial Neural Network model

5) Support Vector Machine (SVM): A support vector machine (SVM) is a classical algorithm that solves problems of big data classification [30]. SVM takes input and classifies them into two different classes. With a given set of training examples belonging to two given categories, SVM builds an SVM model to classify other examples into those two categories. SVM learns a hyperplane to classify the data set into two separate classes by constructing a hyperplane or a set of hyperplanes in a multi-dimensional or infinite-dimensional space. For the best classification, it is necessary to determine the optimal hyperplane located as far away from the data points of all classes as possible.

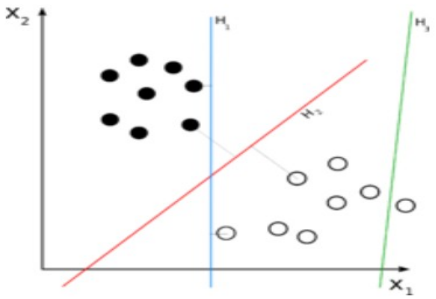

Fig. 3 Support vector machine

Fig. 3 depicts the SVM algorithm: Given a training set represented in a vector space where each document is a point, this method finds a decision hyperplane h that can best divide the points on this space into two separate layers, respectively, the layer containing the data containing the feature simulated by the black dot and the layer containing the data containing the feature simulated by the white dot. The quality of this hyperplane is determined by the boundary of the nearest data point of each layer to this plane. The purpose of the SVM algorithm is to find the maximum boundary distance.

6) Decision Tree: A Decision Tree is a classification model introduced by Belson [31], widely used in many different fields. After the introduction of the machine learning method system, the Decision Tree was further developed with the C4.5 algorithm by Quinlan [32] and the ID3 algorithm by Quinlan [33]. A Decision Tree is a structured classification tree that classifies objects based on sequences of rules. To determine which variable to use classification first, which variable to use later, the information weight (entropy) for each variable is calculated, the higher entropy, the more categorical information the variable carries.

C. Combining Techniques for Data Mining

For improving the accuracy of the method of hybridization of models in the classification problem, this research employed Boosting and Bagging to improve the accuracy of the classification algorithms.

Bagging comes from two abbreviations, Bootstrap and Aggregation [34]. Bagging is a combination of independent base models that leads to a significant reduction in errors. Therefore, the goal is to get as many base models as independent as possible. Bagging generates classifiers from subsets that revert to the Bootstrap samples and a machine learning algorithm, each of which generates a basic classifier. The classifiers will be combined by the majority voting method. That is, when there is an example that needs to be classified, each classifier will produce a result. And the result that appears the most will be taken as the result of the

combiner. The Bagging generates N-selected training sets with iterations from the original training data set.

Boosting is a method of building a set of weak classifiers to improve the efficiency of these classifiers. After each iteration, the weak classifier will focus on learning on elements that were misclassified in previous iterations. To classify newly arrived data, people use the majority voting rule from the classification results of each weak classification model [35].

D. Evaluating The Model

Confusion Matrix is commonly used in model evaluation. This study employs a calculation of indices of Confusion Matrix as shown in Table 2.

TABLE 2
CONFUSION MATRIX

		Prediction	
		Positive	Negative
Reality	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The effectiveness of the opinion classification model is evaluated based on 4 indexes: Accuracy, Precision, Recall, and Harmonized Mean (F1-score). In which:

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

$$\Pr ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2}{1/\Pr ecision + 1/\mathrm{Re}\, call}$$

E. Collecting Data

This study uses data collected from the Vietnamese stock exchange in the period 2009 - 2020. Data is collected from audited financial statements of listed companies after excluding companies in the field of listed companies. banking, securities, and insurance sectors. After determining the indicators, the data used to perform analysis and forecasting is 4693 observations, presented in Table 3 by year and by field.

The study objectives are to use data mining algorithms including Logistic Regression, Bayesian Network, K-nearest neighbor, Artificial Neural Network, Support Vector Machine, and Decision Tree for predicting bankruptcy and to determine the accuracy of these data mining algorithms. The data are randomly divided into 2 parts to build and test the model: Training data is used for building the research model and testing data is used to test the predictive likelihood of the model. The description of indicator characteristics in the research model is presented in Appendix 1. Out of 4693 observations, 2395 observations are at risk of bankruptcy, accounting for 51.03% and vice versa 48.97% is normal. Thus,

TABLE 3
DEMOGRAPHIC DESCRIPTION

Panel A: Data by year			Panel B: Data by field		
Year	Number	%	Industry	Number	%
2009	213	4.5%	Real estate - construction	1,707	36.4%
2010	306	6.5%	Technology	134	2.9%
2011	398	8.5%	Industry	544	11.6%
2012	404	8.6%	Service	528	11.3%
2013	426	9.1%	Consumer goods	395	8.4%
2014	422	9.0%	Energy	366	7.8%
2015	454	9.7%	Agriculture	402	8.6%
2016	470	10.0%	Materials	473	10.1%
2017	477	10.2%	Medical-pharmacy	144	3.1%
2018	475	10.1%			
2019	413	8.8%			
2020	235	5.0%			
Total	**4693**	**100.0%**	**Total**	**4693**	**100.0%**

the data on the number of normal enterprises and the bankruptcy likelihood is quite balanced.

Appendix 1 reveals a testing result of the difference in the mean value of 30 indicators in the research model between the normal enterprise group and the bankruptcy likelihood group. 27/30 indicators that have a difference between the two groups and are statistically significant, only 3 of the indicators of growth have no difference between the two normal groups and the bankruptcy likelihood group including X18-Operating profit growth, X19-Net profit growth, and X20-Equity growth.

IV. RESULTS AND DISCUSSIONS

To achieve the objective of the study on the question of commonly used classification algorithms in Data mining, which algorithm gives the best predictive results, Weka software is applied to research data to conduct experiments. Logistic regression and ANN algorithms give a high probability of bankruptcy prediction (accuracy over 97%).

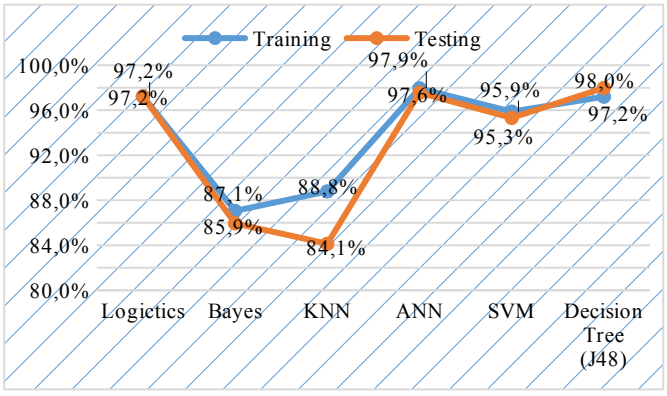

Fig. 4 Accuracy of algorithms in research data

To improve the accuracy of the method of hybridization of models in the classification, Boosting and Bagging methods are employed. The results presented in Figure 5, Figure 6, and Figure 7 show the accuracy. Bankruptcy prediction results of Bagging and Boosting methods have improved over the original basic methods.

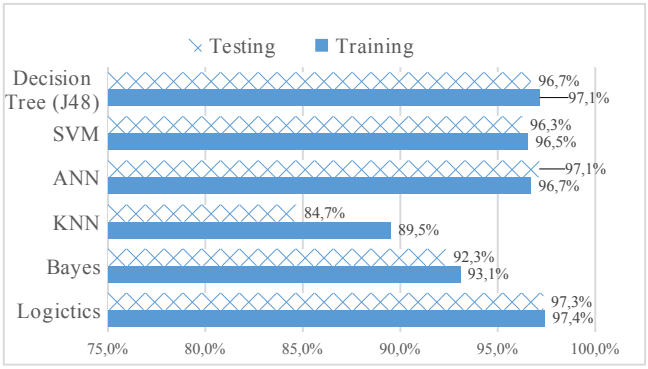

Fig. 5 Accuracy of methods according to Bagging

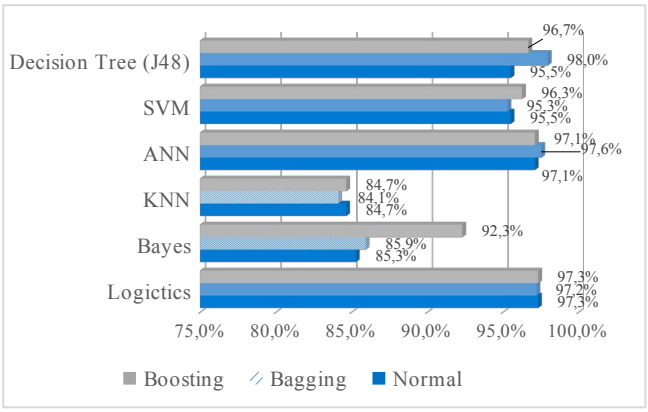

Fig. 6 Accuracy of methods according to Boosting

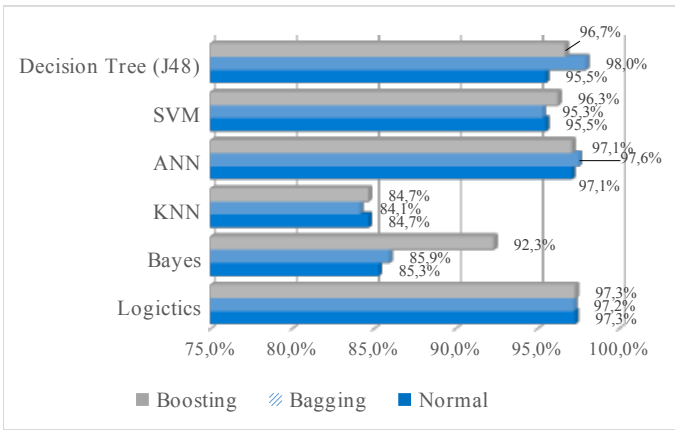

Fig. 7 Comparison of accuracy among methods

With 30 ratios used for forecasting in the model, which ratio is the most important and has the most predictive significance? When using Weka software to identify significant variables in bankruptcy prediction. Figure 8 shows 10 ratios that have the greatest impact on corporate bankruptcy prediction. In which X13 - Total asset turnover is the most important indicator, followed by X5 -Debt to equity ratio, and X24 -Debt ratio.

We select the 10 most significant indicators as a set of important indicators in predicting bankruptcy including X13 - Total assets turnover ratio, X5 - Debt to equity ratio, X24 - Debt ratio, X3 - Receivables turnover ratio, X26 – Receivables conversion period, X27 – Payables conversion period, X7 - Operating cash flows ratio, X1 - Current ratio, X14 - Inventory conversion period), and X2 - Quick ratio. To test again whether the financial ratios are the most important in predicting bankruptcy, we use the dataset with a set of 10-ratio replaces the set of 30-ratio.

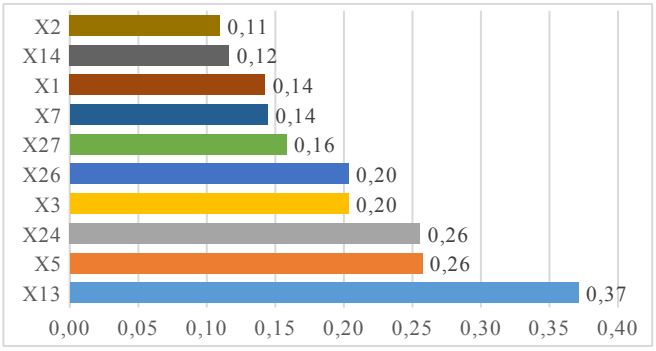

Fig. 8 The most important indicators

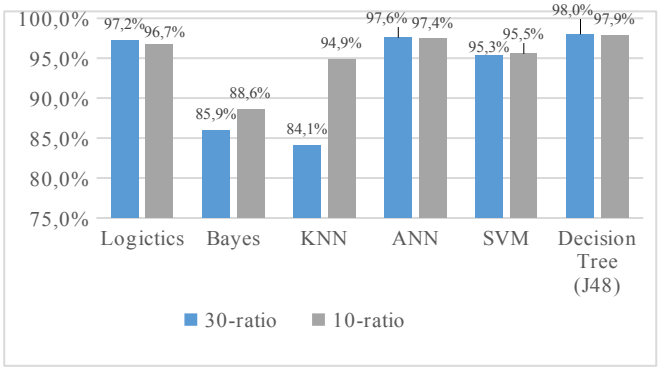

Fig. 9 Accuracy of algorithms with two datasets

The results of this study are consistent, similar, and have higher accuracy than those of Liang, et al. [5], Barboza, et al. [6], Chou, et al. [7], Antunes, et al. [8], Chen, et al. [14].

The results show that the efficiency when using the reduced data set with 10-ratio has the same accuracy as when using the data set of 30-ratio for algorithms with high prediction accuracy rates such as Logistics, ANN, and DT. Even in the Bayesian and KNN algorithms, the accuracy of the prediction is far superior to that of the dataset with full indicators. From these results, this research suggests choosing a set of the most important indicators for predicting bankrupt that saves resources in forecasting at high accuracy.

The research continues to study using the Decision Tree algorithm (J48) after removing the ratio that has no influence or little importance to perform the analysis. The results show that the Decision Tree algorithms predict bankruptcy with an accuracy of 97.9%, implying that it is appropriate to use the Decision Tree model to predict bankruptcy for Vietnamese enterprises. Appendix 7 depicts the Decision Tree results of the 10 most important indicators which lead to the corporate bankruptcy risk. At level 1, X13-Total asset turnover ratio is the most important ratio to predict bankruptcy risk for businesses that when asset turnover is less than 1.4654 then the business is forecasted to be at risk of bankruptcy. The next most important metric for bankruptcy is X5-Debt to equity ratio, at level 2 with a threshold of 0.511 will lead to bankruptcy.

V. Conclusions and Recommendations

This study uses data mining to predict corporate bankruptcy. The sample is companies that have been listed in

Vietnam in the period 2009-2020. This study is to evaluate whether data mining algorithms can be used to predict the bankruptcy of companies in Vietnam accurately or not, which financial indicators are the most effective ratios to predict. To achieve the research objectives, the research has in turn used algorithms including Logistic Regression, Bayesian Network, K-nearest neighbor, Artificial Neural Network (ANN), Support Vector Machine, and Decision Tree. Based on the research results, it can be seen that all 6 methods are accurate in predicting the status, normal, or risk of bankruptcy, of the companies in the sample. In addition, we recommend the use of Decision Tree, ANN that will give the highest prediction accuracy. It can be concluded that these models are suitable for predicting bankruptcy for Vietnamese enterprises in the current period. Moreover, the research has shown 10 financial ratios that are most important in predicting bankruptcy risk. From the above research results, the research has some recommendations for businesses and investors as well as practical suggestions for listed companies to minimize bankruptcy risk.

Total assets turnover ratios, debt to equity ratio, and debt ratio are the three most important indicators in predicting the bankruptcy risk of a business. The results also show that Vietnamese enterprises during the study period are at risk of bankruptcy due to improper implementation of investment decisions stemming from the use of excessive financial leverage and inefficient business activities. The evidence of this research is an important scientific basis for financial managers when planning strategies.

It is necessary to carry out the process to improve the health of the business on the existing foundation, the process of making fundamental changes in the business to increase the ability to operate more efficiently, and create a better "new normal" environment for the business to achieve the strategies and goals. The research postulates some recommendations: Prepare financial statements under the current regulations of the Ministry of Finance; Financial statements must be audited by reputable auditing agencies; In addition to cultivating knowledge about management and law, the listed companies need to regularly improve their knowledge in corporate finance, especially financial ratios to measure business health.

The results show that financial managers need to be careful with regulations on mobilizing funding sources, fully exploiting internal capital sources, especially from retained earnings to reduce the cost of using corporate capital and to limit the use of debt, especially short-term debt. Moreover, the financial managers need to increase the exploitation of highly liquid assets to improve investment efficiency. Furthermore, the financial managers need to regularly re-check the investment regulations so that the business plan can be adjusted in time.

However, the study still has certain limitations. The factors which impact corporate bankruptcy are not only in financial ratios but also come from human behavior. This study has not mentioned the intervening factors such as human behavior, crowd psychology, and speculation affecting the increase or decrease in bankruptcy risk of listed companies in Vietnam.

References

[1] The World Bank. (2021). *The World Bank in Vietnam: Overview* [webpage]. Available: https://www.worldbank.org/en/country/vietnam/overview#1

[2] R. C. Merton, "On the pricing of corporate debt: The risk structure of interest rates," *The Journal of Finance,* vol. 29, no. 2, pp. 449-470, 1974.

[3] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance,* vol. 23, no. 4, pp. 589-609, 1968.

[4] V. T. T. Binh, N.-M. Tran, D. M. Thanh, and H.-H. Pham, "Firm size, business sector and quality of accounting information systems: Evidence from Vietnam," *Accounting,* vol. 6, no. 3, pp. 327-334, 2020.

[5] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *European Journal of Operational Research* vol. 252, no. 2, pp. 561–572, 2016.

[6] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Systems with Applications,* vol. 83, pp. 405-417, 2017.

[7] C.-H. Chou, S.-C. Hsieh, and C.-J. Qiu, "Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction," *Applied Soft Computing,* vol. 56, pp. 298-316, 2017.

[8] F. Antunes, B. Ribeiro, and F. Pereira, "Probabilistic modeling and visualization for bankruptcy prediction," *Applied Soft Computing,* vol. 60, pp. 831-843, 2017.

[9] T. Le, M. Y. Lee, J. R. Park, and S. W. Baik, "Oversampling techniques for bankruptcy prediction: novel features from a transaction dataset," *Symmetry,* vol. 10, no. 4, p. 79, 2018.

[10] T. Le, H. Le Son, M. T. Vo, M. Y. Lee, and S. W. Baik, "A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset," *Symmetry,* vol. 10, no. 7, p. 250, 2018.

[11] D. Veganzones and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets," *Decision Support Systems,* vol. 112, pp. 111-124, 2018.

[12] F. Mai, S. Tian, C. Lee, and L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures," *European journal of operational research,* vol. 274, no. 2, pp. 743-758, 2019.

[13] H. Son, C. Hyun, D. Phan, and H. J. Hwang, "Data analytic approach for bankruptcy prediction," *Expert Systems with Applications,* vol. 138, p. 112816, 2019.

[14] Z. Chen, W. Chen, and Y. Shi, "Ensemble learning with label proportions for bankruptcy prediction," *Expert Systems with Applications,* vol. 146, p. 113155, 2020.

[15] E. I. Altman, "Predicting financial distress of companies: revisiting the Z-score and ZETA® models," in *Handbook of research methods and applications in empirical finance*: Edward Elgar Publishing, 2013, pp. 428–456.

[16] Z. H. U. Konglai and L. I. Jingjing, "Studies of discriminant analysis and logistic regression model application in credit risk for China's listed companies," *Management Science and Engineering,* vol. 4, no. 4, pp. 24-32, 2011.

[17] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research,* vol. 18, no. 1, pp. 109-131, 1980.

[18] S. Meeampol, P. Lerskullawat, A. Wongsorntham, P. Srinammuang, V. Rodpetch, and R. Noonoi, "Applying emerging market Z-score model to predict bankruptcy: A case study of listed companies in the stock exchange of Thailand (Set)," in *Management, Knowledge And Learning International Conference*, 2014, pp. 25-27.

[19] M. N. Kumar and V. S. H. Rao, "A new methodology for estimating internal credit risk and bankruptcy prediction under Basel II Regime," *Computational Economics,* vol. 46, no. 1, pp. 83-102, 2015.

[20] P. R. Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review," *European journal of operational research,* vol. 180, no. 1, pp. 1-28, 2007.

[21] C. Serrano-Cinca, "Self organizing neural networks for financial diagnosis," *Decision support systems,* vol. 17, no. 3, pp. 227-238, 1996.

[22] D. Fletcher and E. Goss, "Forecasting with neural networks: an application using bankruptcy data," *Information & Management,* vol. 24, no. 3, pp. 159-167, 1993.

[23] A. Fan and M. Palaniswami, "Selecting bankruptcy predictors using a support vector machine approach," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN*

2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, vol. 6, pp. 354-359: IEEE.

[24] A. W. Ghazali, N. A. Shafie, and Z. M. Sanusi, "Earnings management: An analysis of opportunistic behaviour, monitoring mechanism and financial distress," *Procedia Economics and Finance,* vol. 28, pp. 190-201, 2015.

[25] J. Berkson, "Application of the logistic function to bio-assay," *Journal of the American statistical association,* vol. 39, no. 227, pp. 357-365, 1944.

[26] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," *Journal of Banking & Finance,* vol. 18, no. 3, pp. 505-529, 1994.

[27] A. M. Flitman, "Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis," *Computers & Operations Research,* vol. 24, no. 4, pp. 367-377, 1997.

[28] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC, 2000.

[29] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice (2nd ed)*. Melbourne, Australia: OTexts, 2018.

[30] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*: Springer, 2016, pp. 207-235.

[31] W. A. Belson, "Matching and prediction on the principle of biological classification," *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 8, no. 2, pp. 65-75, 1959.

[32] J. R. Quinlan, "Improved use of continuous attributes in C4. 5," *Journal of artificial intelligence research,* vol. 4, pp. 77-90, 1996.[33] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, no. 1, pp. 81-106, 1986.

[34] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[35] R. Schapire and Y. Freund, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Second European Conference on Computational Learning Theory*, 1995, pp. 23-37.

APPENDIX 1

ATTRIBUTE OF INDICATORS AND DESCRIPTIVE STATISTICS OF RESEARCH VARIABLES

Indicators		Measure	Mean	Std.	1%	25%	50%	75%	99%	Normal group (2298; 48.97%)	Bankruptcy risk (2395; 51.03%)	t	P
Current ratio	X1	Current assets/Current liabilities	2.0	1.7	0.5	1.1	1.5	2.2	9.7	2.571	1.493	22.795	0.0000
Quick ratio	X2	(Current assets - Inventories)/ Current liabilities	1.4	1.5	0.2	0.6	1.0	1.6	8.7	1.885	0.983	20.9102	0.0000
Receivables turnover ratio	X3	Net sales/Receivables	7.7	10.3	0.4	2.2	4.5	8.6	54.0	11.048	4.435	23.2008	0.0000
Current operating cash flows ratio	X4	Operating cash flows/Current liabilities	0.3	0.6	-0.9	0.0	0.1	0.4	2.6	0.396	0.128	16.067	0.0000
Debt to equity ratio (D/E)	X5	Total liabilities/Shareholder's equity	1.5	1.4	0.1	0.5	1.1	2.0	7.0	0.876	2.139	-33.1772	0.0000
Fixed assets to long-term capital ratio	X6	Fixed assets/(Total liabilities and owners' equity – Current liabilities)	0.3	0.3	0.0	0.1	0.3	0.5	1.1	0.320	0.362	-5.0179	0.0000
Operating cash flows ratio	X7	Operating cash flows/Total debts	0.2	0.5	-0.7	0.0	0.1	0.3	2.2	0.336	0.052	22.1178	0.0000
Operating ratio	X8	Net profit/Net sales	0.1	0.1	-0.1	0.0	0.0	0.1	0.5	0.079	0.071	2.6419	0.0083
Operating profit margin ratio	X9	Operating profit/Net sales	0.1	0.1	-0.1	0.0	0.1	0.1	0.5	0.096	0.087	2.4434	0.0146
Net profit to book value of equity ratio	X10	Net profit/Book value of equity	0.1	0.1	-0.1	0.0	0.1	0.2	0.4	0.151	0.094	19.8108	0.0000
Return on equity ratio	X11	Net profit/Equity	0.1	0.1	-0.1	0.0	0.1	0.2	0.4	0.149	0.090	20.7952	0.0000
Operating income to book value of shareholder's equity ratio	X12	Operating incomes/ Book value of shareholder's equity	0.2	0.1	-0.1	0.1	0.1	0.2	0.5	0.187	0.118	19.9933	0.0000
Total asset turnover ratio	X13	Net sales/Total assets	1.2	0.9	0.1	0.5	0.9	1.5	4.6	1.673	0.683	45.1438	0.0000
Inventory conversion period	X14	Inventories x 365days/Cost of goods sold	-167.6	377.2	-2001.1	-147.9	-78.1	-34.2	0.0	-73.660	-257.680	17.2236	0.0000
Fixed asset turnover ratio	X15	Nets sales/Fixed assets	21.5	50.4	0.3	3.3	7.5	17.1	280.1	26.764	16.441	7.0549	0.0000
Revenue growth	X16	Net sales year t/Net sales year t-1	1.2	0.6	0.4	0.9	1.1	1.2	3.7	1.139	1.177	-2.2745	0.0230
Assets growth	X17	Total assets year t/Total assets year t-1	1.1	0.3	0.7	1.0	1.1	1.2	2.4	1.095	1.156	-6.781	0.0000
Operating profit growth	X18	Operating profit year t/Operating profit year t-1	1.2	2.0	-2.7	0.7	1.0	1.3	8.9	1.218	1.155	1.0887	0.2763
Net profit growth	X19	Net profit year t/Net profit year t-1	1.2	2.6	-3.4	0.7	1.0	1.3	10.8	1.243	1.203	0.5261	0.5988
Equity growth	X20	Equity year t/Equity year t-1	1.1	0.3	0.8	1.0	1.0	1.1	2.3	1.121	1.133	-1.2252	0.2206
Stock price trend	X21	Share price year t/Share price year t-1	1.2	0.6	0.4	0.8	1.0	1.4	3.3	1.204	1.149	3.3644	0.0008
Earnings per share (EPS)	X22	Net profit/Number of shares	2264.8	2278.3	-1511.0	656.0	1667.0	3257.0	9957.0	2889.994	1664.953	19.115	0.0000
General solvency	X23	Total assets/Current liabilities	3.7	3.5	1.2	1.8	2.5	4.2	19.1	4.383	3.079	13.0359	0.0000
Debt ratio	X24	Total debts/Total assets	0.5	0.2	0.1	0.3	0.5	0.7	0.9	0.387	0.607	-42.8424	0.0000
Firm size	X25	Ln(total assets)	27.2	1.4	24.1	26.2	27.2	28.1	30.7	26.688	27.623	-24.0555	0.0000
Receivables conversion period	X26	Receivables x 365days/Net sales	136.2	176.8	6.8	42.4	81.8	165.6	847.3	76.747	193.238	-23.9024	0.0000
Payables conversion period	X27	Payables x 365 days/Net sales	45.4	52.1	1.0	16.2	31.0	55.2	264.7	26.677	63.298	-25.6933	0.0000
Book value	X28	Shareholder's equity/ Number of shares	18253	8237	7206	12721	16010	21274	50973	19200.590	17344.450	7.7654	0.0000
State ownership ratio	X29	Proportion of State holding shares	0.2	0.2	0.0	0.0	0.2	0.5	0.8	0.264	0.220	6.219	0.0000
Foreign ownership ratio	X30	Proportion of foreign holdings of shares	0.7	0.3	0.1	0.5	0.7	0.9	1.0	0.628	0.698	-9.4886	0.0000

APPENDIX 2
STATISTIC RESULTS OF BANKRUPTCY PREDICTION

Algorithms	Training						Testing					
	Accuracy	Precision	Recall	F-Measure	ROC Area	PRC Area	Accuracy	Precision	Recall	F-Measure	ROC Area	PRC Area
Logistics	97.42%	0.974	0.974	0.974	0.998	0.998	97.34%	0.973	0.973	0.973	0.997	0.998
Bayes	87.00%	0.87	0.87	0.87	0.93	0.918	85.30%	0.85	0.85	0.85	0.91	0.89
KNN	89.82%	0.898	0.898	0.898	0.958	0.949	84.66%	0.85	0.85	0.85	0.92	0.90
ANN	96.70%	0.967	0.967	0.967	0.996	0.997	97.12%	0.97	0.97	0.97	1.00	1.00
SVM	96.00%	0.96	0.96	0.96	0.96	0.942	95.53%	0.96	0.96	0.96	0.95	0.94
Decision Tree	95.98%	0.96	0.96	0.96	0.963	0.949	95.53%	0.96	0.96	0.96	0.95	0.94

APPENDIX 3
STATISTIC RESULTS OF BANKRUPTCY PREDICTION BY BAGGING METHOD

Algorithms	Training						Testing				
	Accuracy	Precision	Recall	F-Measure	ROC Area	PRC Area	Accuracy	Precision	Recall	F-Measure	ROC Area
Logistics	97.23%	0.973	0.973	0.973	0.998	0.998	97.23%	0.972	0.972	0.972	0.998	0.998
Bayes	87.05%	0.871	0.871	0.871	0.937	0.928	85.94%	0.862	0.859	0.859	0.917	0.902
KNN	88.79%	0.893	0.892	0.892	0.947	0.94	84.13%	0.841	0.841	0.841	0.9	0.884
ANN	97.95%	0.979	0.979	0.979	0.999	0.999	97.55%	0.976	0.976	0.976	0.998	0.998
SVM	95.87%	0.959	0.959	0.959	0.977	0.967	95.31%	0.953	0.953	0.953	0.973	0.962
Decision Tree	97.20%	0.972	0.972	0.972	0.995	0.994	97.98%	0.98	0.98	0.98	0.997	0.997

APPENDIX 4
STATISTIC RESULTS OF BANKRUPTCY PREDICTION BY BOOSTING METHOD

Algorithms	Training						Testing					
	Accuracy	Precision	Recall	F-Measure	ROC Area	PRC Area	Accuracy	Precision	Recall	F-Measure	ROC Area	PRC Area
Logistics	97.42%	0.974	0.974	0.974	0.984	0.979	97.34%	0.973	0.973	0.973	0.978	0.968
Bayes	93.07%	0.931	0.931	0.931	0.983	0.984	92.33%	0.924	0.923	0.923	0.985	0.984
KNN	89.50%	0.895	0.895	0.895	0.927	0.911	84.66%	0.847	0.847	0.847	0.885	0.848
ANN	96.70%	0.967	0.967	0.967	0.979	0.977	97.12%	0.971	0.971	0.971	0.971	0.958
SVM	96.54%	0.965	0.965	0.965	0.995	0.995	96.27%	0.963	0.963	0.963	0.995	0.994
Decision Tree	97.15%	0.971	0.971	0.971	0.996	0.995	96.70%	0.967	0.967	0.967	0.997	0.997

APPENDIX 5
STATISTIC RESULTS OF BANKRUPTCY PREDICTION BY DECISION TREE

```
X13 <= 1.4654
|  X5 <= 0.511
|  |  X5 <= 0.3402
|  |  |  X13 <= 0.3798
|  |  |  |  X24 <= 0.2247: Normal (53.0)
|  |  |  |  X24 > 0.2247
|  |  |  |  |  X1 <= 3.1641: Bankruptcy (6.0)
|  |  |  |  |  X1 > 3.1641: Normal (8.0)
|  |  |  X13 > 0.3798: Normal (305.0)
|  |  X5 > 0.3402
|  |  |  X13 <= 0.6335
|  |  |  |  X13 <= 0.4234
|  |  |  |  |  X1 <= 3.594: Bankruptcy (51.0)
|  |  |  |  |  X1 > 3.594
|  |  |  |  |  |  X5 <= 0.4175: Normal (3.0)
|  |  |  |  |  |  X5 > 0.4175: Bankruptcy (2.0)
|  |  |  |  X13 > 0.4234
|  |  |  |  |  X1 <= 2.4853
|  |  |  |  |  |  X5 <= 0.4205
|  |  |  |  |  |  |  X1 <= 1.5824: Bankruptcy (4.0)
|  |  |  |  |  |  |  X1 > 1.5824: Normal (7.0/1.0)
|  |  |  |  |  |  X5 > 0.4205: Bankruptcy (19.0)
|  |  |  |  |  X1 > 2.4853
|  |  |  |  |  |  X5 <= 0.4332: Normal (13.0)
|  |  |  |  |  |  X5 > 0.4332
|  |  |  |  |  |  |  X2 <= 1.9596: Bankruptcy (2.0)
|  |  |  |  |  |  |  X2 > 1.9596: Normal (3.0)
|  |  |  X13 > 0.6335
|  |  |  |  X1 <= 1.7516
|  |  |  |  |  X13 <= 0.8173
|  |  |  |  |  |  X5 <= 0.4472: Normal (3.0)
|  |  |  |  |  |  X5 > 0.4472: Bankruptcy (6.0)
|  |  |  |  |  X13 > 0.8173: Normal (22.0)
|  |  |  |  X1 > 1.7516: Normal (119.0)
|  X5 > 0.511
|  |  X13 <= 0.9186
|  |  |  X5 <= 0.6592
|  |  |  |  X13 <= 0.6262: Bankruptcy (105.0/2.0)
|  |  |  |  X13 > 0.6262
|  |  |  |  |  X1 <= 1.8404: Bankruptcy (18.0)
|  |  |  |  |  X1 > 1.8404
|  |  |  |  |  |  X27 <= 13.2254: Normal (8.0)
|  |  |  |  |  |  X27 > 13.2254
|  |  |  |  |  |  |  X2 <= 1.409: Normal (5.0)
|  |  |  |  |  |  |  X2 > 1.409
|  |  |  |  |  |  |  |  X1 <= 1.9549: Bankruptcy (5.0)
|  |  |  |  |  |  |  |  X1 > 1.9549
|  |  |  |  |  |  |  |  |  X13 <= 0.7761: Bankruptcy (6.0/1.0)
|  |  |  |  |  |  |  |  |  X13 > 0.7761: Normal (9.0/1.0)
|  |  |  X5 > 0.6592: Bankruptcy (1072.0)
|  |  X13 > 0.9186
|  |  |  X5 <= 1.0544
|  |  |  |  X1 <= 1.5984
|  |  |  |  |  X13 <= 1.1831
|  |  |  |  |  |  X5 <= 0.7651
|  |  |  |  |  |  |  X14 <= -81.6781: Normal (10.0)
|  |  |  |  |  |  |  X14 > -81.6781: Bankruptcy (5.0)
|  |  |  |  |  |  X5 > 0.7651: Bankruptcy (23.0/1.0)
|  |  |  |  |  X13 > 1.1831
|  |  |  |  |  |  X2 <= 0.5561
|  |  |  |  |  |  |  X2 <= 0.4764: Normal (4.0)
|  |  |  |  |  |  |  X2 > 0.4764: Bankruptcy (4.0/1.0)
|  |  |  |  |  |  X2 > 0.5561: Normal (18.0)
|  |  |  |  X1 > 1.5984
|  |  |  |  |  X13 <= 1.0733
|  |  |  |  |  |  X5 <= 0.9134
|  |  |  |  |  |  |  X14 <= -4.27: Normal (26.0/1.0)
|  |  |  |  |  |  |  X14 > -4.27: Bankruptcy (3.0/1.0)
|  |  |  |  |  |  X5 > 0.9134: Bankruptcy (6.0/1.0)
|  |  |  |  |  X13 > 1.0733: Normal (65.0)
|  |  |  X5 > 1.0544
|  |  |  |  X1 <= 1.2168: Bankruptcy (277.0/1.0)
|  |  |  |  X1 > 1.2168
|  |  |  |  |  X13 <= 1.2471
|  |  |  |  |  |  X1 <= 1.807: Bankruptcy (109.0/1.0)
|  |  |  |  |  |  X1 > 1.807
|  |  |  |  |  |  |  X5 <= 1.617: Normal (6.0/1.0)
|  |  |  |  |  |  |  X5 > 1.617: Bankruptcy (3.0)
|  |  |  |  |  X13 > 1.2471
|  |  |  |  |  |  X1 <= 1.5
|  |  |  |  |  |  |  X13 <= 1.427
|  |  |  |  |  |  |  |  X5 <= 1.484
|  |  |  |  |  |  |  |  |  X13 <= 1.3288: Bankruptcy (5.0/1.0)
|  |  |  |  |  |  |  |  |  X13 > 1.3288: Normal (4.0)
|  |  |  |  |  |  |  |  X5 > 1.484: Bankruptcy (19.0/1.0)
|  |  |  |  |  |  |  X13 > 1.427: Normal (6.0)
|  |  |  |  |  |  X1 > 1.5: Normal (14.0)
X13 > 1.4654
|  X5 <= 2.275: Normal (770.0/1.0)
|  X5 > 2.275
|  |  X13 <= 1.6915
|  |  |  X1 <= 1.1012: Bankruptcy (13.0)
|  |  |  X1 > 1.1012
|  |  |  |  X5 <= 2.8411: Normal (11.0/1.0)
|  |  |  |  X5 > 2.8411: Bankruptcy (3.0)
|  |  X13 > 1.6915: Normal (121.0/2.0)
```