Classification of Dynamic Financial Distress Manufacturing Company Listed in Indonesia Stock Exchange Using Binary Logistic Regression and Classification Analysis Regression Tree

Nugroho Priyo Negoro[1,*] Aldia Wira Trispantia[1]

[1]*Department Business Management, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[*]*Corresponding author. Email: npriyo.negoro@gmail.com*

ABSTRACT
Financial Distress (FD) is a large concept consists several situations where the enterprise facing financial difficulty. Some manufacturing enterprise that are experiencing financial difficulties trying to overcome these problems by make loans and merger, or otherwise shut down their company. If the condition of FD is known, it is expected to take action to improve the situation so that the enterprise will not experience more severe difficulties such as bankruptcy or liquidation. The purpose of this study is to determine the factors that allegedly significantly influence for financial distress, and knowing classification model of financial distress. Thus, this study will try to classify the dynamic financial distress companies that listed on stock exchanges in Indonesia 2012-2014 using binary logistic regression and classification and regression tree analysis (CART). Before analyzing the binary logistic regression and CART, first the data must be balanced using SMOTE. Binary logistic regression developed the model of FD to see the factors that allegedly significantly influence for financial distress, while CART method is to solve the classification. Experimental results show that binary logistic regression has two predictor variables that significantly influence for FD condition for manufacturing enterprise and that variable is liquidity ratio and the activity ratio. CART classification method produces that maximum classification tree is equal to the optimum classification tree, with the primary node is variable solvency ratio, and the value of a classification. CART method outperforms the binary logistic regression with the classification's accuracy has more than ten percent.
Keywords: *financial distress, classification manufacturing company, binary logistic regression, CART*

1. INTRODUCTION

The global financial crisis has not only impacted the real sector, but also severely hit the financial sector. Sectors affected by the global financial crisis are all of the sectors. But the most visible symptom is the manufacturing sector. The state of the manufacturing industry is currently experiencing a decline in export performance a trade balance deficit in 2011-2013 [1]. Some manufacturing companies that are encounter financial problems try to overcome these problems by making loans and business combinations, or conversely some are closing their businesses.

Financial distress is a comprehensive concept that consists of several situations where a company is facing financial difficulties. Common terms to describe the situation are bankruptcy, failure, inability to pay off debt, and default [2]. Insolvency in bankruptcy shows negative net worth. If the condition of financial distress is known, it is expected that measurement can be taken to improve the situation so that the company will not get into difficult stage such as bankruptcy or liquidation. Bankruptcy of a company can be observed and measured through financial statements [2].

According to Sin & Li (2011), most of all financial distress prediction only focus on static models to predict models, where forecasting models are formed from data samples at specific times. Later, the static model cannot predict financial distress in changes in the economic environment or changes in company operations properly and effectively [3]. Hence, to overcome the dynamic operational environment of the company, according to time changes, research on financial distress prediction models must be done by changing the assumptions that existed in previous studies, namely that the number of data samples does not change or only in certain time periods. The concept of dynamic financial distress prediction consists of instance selection, financial distress prediction modeling, and future prediction

Logistic regression is a classification method with a parametric approach, the advantage is an odds ratio value that shows how much influence the predictor variable of a preference category on a response variable [4]. Classification Analysis and Regression Trees (CART) is one of the methods or algorithms of data exploration techniques, namely the decision tree technique. According to Breiman et al. (1993), CART is a statistical methodology with a nonparametric approach, developed for classification analysis that able to overcome the limitations of assumptions [5]. There are several advantages possessed by the CART method that is able to work on large data dimensions and complex data structures, can find out the interactions between predictor variables and the classification results obtained are more easily understood and interpreted. Therefore, the parametric approach is used with the binary logistic regression method and the nonparametric approach is using the CART method. The purpose of this study is to find out the factors that are suspected to have a significant effect on financial distress, and to know the financial distress classification model, whether included in financial distress or non-financial distress. This paper was designed as follows. Chapter I Introduction, overview the research background including research motivation, objective, and also research structure. Chapter II explanation about Logistic Regression and CART. Chapter III developing the model, source data, variables and also evaluation criteria. In chapter IV showed the analysis process and result. Lastly in chapter V the study was concluded and discussed about the future research.

2. LITERATURE REVIEW

2.1 Binary Logistic Regression

The general form of the logistic regression model is as follows.

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} \tag{1}$$

1. Parameter Estimation

The Maximum Likelihood Estimation (MLE) function obtained is:

$$l(\beta) = \prod_{i=1}^{p} \pi(x_i)^{y_i} \left[1-\pi(x_i)\right]^{1-y_i} \tag{2}$$

Where,

i : 1,2,...p

x_i : observation i-th

π_i : probability i-th

2. Parameter Testing

The hypothesis for simultaneous parameters testing is as follows.

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_i = 0$

H_1 : at least one $\beta_i \neq 0$ where i = 1, 2, ..., p

The statistic test used G test as follows.

$$G = -2\ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\sum_{i=1}^{n} \hat{\pi}_i^{\,y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}} \tag{3}$$

G statistics test is a *Likelihood Ratio Test* where *G* value follows Chi-Square distributions. Therefore, H_0 is rejected if $G > \chi^2_{(v,\alpha)}$ where *v* is degree of freedom obtained by the number of parameter model without β_0.

Partial parameter testing obtained by using Wald test statistics. The hypothesis is:

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$ where $i = 1, 2, ..., p$

The statistic test used W test as follows.

$$W_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \tag{4}$$

W statistics test is following normal distribution, so H_0 is rejected if $|W| > Z_{\alpha/2}$ [6].

2.2 Classification Analysis & Regression Tree (CART)

CART is a nonparametric statistical methodology for the type of topic classification analysis, both for categorical and continuous response variables. CART produces a classification tree if the response variable is categorical, and produces a regression tree if the response variable is continuous. The main purpose of CART is to get an accurate data group as a characteristic of a classification. CART method can be applied to a large number of data sets, very large variables and mixed-scale variables through binary sorting procedures. The steps in implementing CART algorithm are as follows.

1. Establishment Classification Tree
   The process of establishment classification tree consists three stages. First, sorting selection or classifier, process where the selection of sorters depends on the type of tree or depending on the type of response variable. Second, determined the terminal node and the last is labeling the class.
2. Pruning Classification Tree
   This step done by pruning the less important parts of the tree so that an optimal tree is obtained. The pruning size used to obtain a decent tree size is the minimum cost complexity. The measurement of cost complexity as follows.

$$R_\alpha(t) = R(T) + \alpha\,|\tilde{T}| \tag{8}$$

Where

R(T) : Resubstitution Estimate (Error proportion in sub tree)

α : Complexity parameter

$\tilde{T}$: Number of terminal tree nodes T

3. Determination of Optimal Classification Tree
   The large classification tree provides the smallest surrogate estimator value, so this tree tends to be chosen to estimate the response value. Large tree size will cause a high complexity value because the data structure tends to be complex, so it is necessary to choose an optimal tree that is simple in size but

provides a replacement value of a relatively small substitute.

3. MATERIALS & MODEL DEVELOPMENT

3.1 Data Source

The data used in this study are secondary data derived from the company's annual report for the years 2012-2014. The population in this study is a listed company on the Indonesia Stock Exchange (IDX) and the research sample is a listed manufacturing company on the IDX. Data is obtained from IDX, OK Shares, and ICMD (Indonesia Capital Market Directory). The variables in this study are shown in Table 1.

3.2 Data Analysis

The following are the stages carried out in the data analysis in this study:

1. Collecting secondary data, data on the financial statements of manufacturing companies listed on the Indonesia Stock Exchange in 2012-2014.
2. Conduct initial classification of data for financial distress and non-financial distress criteria.

Table 1. Identification Variables

Variabel	Variable Description	Measurement	Scale
x_1	Liquidity Ratio	Working Capital divide by Total Assets	Ratio
x_2	Profitability Ratio	Retained Earning divide by total Assets	Ratio
x_3	Rasio Rentabilitas	Earning Before Interest and Taxes divide by Total Assets	Ratio
x_4	Solvency Ratio	Market Value Equity divide by Value Of Total Debt	Ratio
x_5	Activity Ratio	Sales divide by Total Assets	Ratio
x_6	ROI	(Total sale-Investation) divide by Investation	Ratio
x_7	ROE	Profit Before Income Tax divide by Total Equity	Ratio
Y	Financial	1 if financial distress, 0 if	Nominal

3. Divide data into training data and testing data.
4. Perform descriptive statistical analysis to determine the characteristics of the data in each class, financial distress and non-financial distress.
5. Conduct SMOTE analysis for imbalance data. The stages are as follows:
    a. Determine minor and major data.
    b. Replicate each minor data by finding the neighbor k-nearest value.
    c. Calculate synthetic data.
6. Perform binary logistic regression analysis. The stages are as follows:
    a. Test the assumption of independence.
    b. Parameter Estimation
    c. Perform simultaneous testing of parameters using G-test statistics.
    d. Perform partial testing of parameters with Wald-test statistics.
    e. Model the company's financial condition based on influential predictor variables.
    f. Conduct a classification accuracy test.
    g. Interpretation of results
7. Conduct CART analysis. The stages are as follows.
    a. Form a classification tree by determining the selection of sorters, determination of terminal nodes, and labeling of class labels.
    b. Pruning classification trees.
    c. Determine the optimal classification tree.
8. Comparing the results of binary logistic regression analysis and CART methods by looking at the chance of misclassification.
9. Conduct conclusions and suggestions.

4. ANALYSIS AND RESULTS

4.1 Descriptive Statistics

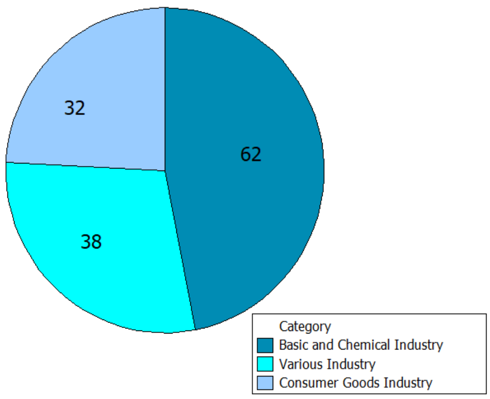

Figure 1. Pie Chart Manufacturing Sector

Based on Figure 1 the manufacturing industry sector consists of 3 sectors, that is basic and chemical industries 62 companies, various industrial sectors 38 companies and consumer goods industry 32 companies.

Table 2. Cross-Tab between Manufacturing Sector with Financial Distress Condition in 2012-2013

	Financial distress	Non-financial distress	total
sector 1	49	13	62
sector 2	31	7	38
sector 3	29	3	32
total	109	23	132

Table 3. CrossTab between Manufacturing Sector with Financial Distress Condition in 2013-2014

	Financial distress	Non-financial distress	total
sector 1	48	14	62
sector 2	31	7	38
sector 3	31	1	32
total	110	22	132

Based on Table 2. information can be obtained that in sector 1 namely the basic industrial and chemical sectors there are 13 manufacturing companies are experiencing financial problems. Then for the various industrial sectors there are 7 manufacturing companies and the consumer goods industry sector by 3 companies. So that there is a total of 23 manufacturing industry companies are experiencing financial problems.
Based on Table 3, information can be obtained that in sector 1 namely the basic industrial and chemical sectors there are 14 manufacturing companies are experiencing financial problems. Then for the various industrial sectors there are 7 manufacturing companies and the consumer goods industry sector by 1 company. So that there is a total of 22 manufacturing industry companies are experiencing financial problems.

4.2 Pre-processing Imbalanced Data

The method for pre-processing imbalanced data is Synthetic Minority Oversampling Technique (SMOTE), that is a sampling technique to increase the amount of data in the minority class by replicating the amount of data in the random minority class so that the amount is the same as the data in the majority class.

Table 4. Simulation Data Distribution Before and After Using SMOTE for Data in 2013-2014

Major Class	Minor Class	Replication	Major Class	New Minor Class
109(83%)	23 (17%)	4	109(49%)	115(51%)

Table 5. Simulation Data Distribution Before and After Using SMOTE for Data in 2012-2013

Major Class	Minor Class	Replication	Major Class	New Minor Class
110(83%)	22 (17%)	4	110(50%)	110(50%)

Tables 4 and 5 shows the results of the distribution of SMOTE simulation data where the number of class 1, data which originally amounted to 23, after being replicated 4 times will become 115 data for data in 2013-2014 and the data of class 1, which originally amounted to 22, after being replicated 4 times will be 110 data for 2012-2013.

4.3 Binary Logistic Regression

1. Independence Test
Hypothesis
H_0: There is no relationship between X_1-X_7 with the financial distress
H_1: There is a relationship between X_1-X_7 with financial distress
Critical area: Reject H_0 if $\alpha < 0.05$
Test Statistics

Table 6. Correlation between Predictor and Response Variables (2012-2013)

	y	X_1	X_2	X_3	X_4	X_5	X_6
X_1	-0.333						
	0.000						
X_2	0.224	-0.692					
	0.001	0.000					
X_3	-0.258	0.151	0.139				
	0.000	0.026	0.039				
X_4	0.235	-0.597	0.171	-0.244			
	0.000	0.000	0.011	0.000			
X_5	-0.341	0.011	-0.127	0.120	-0.020		
	0.000	0.875	0.061	0.076	0.770		
X_6	-0.061	0.170	0.088	0.742	-0.204	-0.110	
	0.366	0.012	0.195	0.000	0.002	0.102	
X_7	-0.044	0.154	-0.004	0.666	-0.091	-0.074	0.896
	0.517	0.022	0.958	0.000	0.178	0.276	0.000
Desc: liquidity ratio (X_1), profitability ratio (X_2), profitability ratio (X_3) solvency ratio (X_4), activity ratio (X_5), ROI (X_6), ROE (X_7)							

Based on Table 6, there is a correlation between the variable liquidity ratio (X1) to the activity ratio (X5) with the financial distress variable (y) which is known from the P-value is less than α (0.05). While the variable ROI (X6) and

ROE (X7) doesn't have a correlation between financial distress variables (y). High correlation between variables can indicate a case of multicollinearity and can cause insignificant parameter testing. For this reason, multicollinearity detection is carried out to determine whether multicollinearity cases occur by the VIF values shown in Table 7.

Table 7 VIF Values

Predictor Variables	VIF
Liquidity ratio (X_1)	3.844
Profitability ratio (X_2)	2.783
Profitability ratio (X_3	2.643
Solvency ratio (X_4	1.972
Activity ratio (X_5)	1.175
ROI (X_6)	7.046
[illegible]	[illegible]

2. Parameter Testing

Hypothesis

H_0 : $\beta_1 = \beta_2 = \ldots = 0$

H_1 : at least one $\beta_1 = \beta_2 = \ldots \neq 0$

Critical area: Reject H0 if $\alpha < 0.05$

Test Statistics

Table 8 Simultaneous Testing 2012-2013

Step 4	Step	8.539	1	0.003
	Block	81.822	4	0.000
	Model	81.822	4	0.000

Based on Table 8, the P-value in the model is 0,000, compared with an α value of 0.05, a decision to reject H0 is obtained, which means there is at least one predictor variable that affects the financial distress of manufacturing companies.

Hypothesis

H_0 : $\beta_i = 0$

H_1 : $\beta_i \neq 0$; $i = 1, 2, \ldots, p$

Critical area: Reject H_0 if $\alpha < 0.05$

Test Statistics

Table 9. Partial Testing 2012-2013

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4[d]	x1	-2.671	0.662	16.274	1	0.000	0.069
	x3	-15.896	4.889	10.573	1	0.001	0.000
	x5	-1.485	0.384	14.981	1	0.000	0.226
	x6	0.083	0.031	7.108	1	0.008	1.087
	Constant	3.159	0.539	34.345	1	0.000	23.536

$$\pi(x) = \frac{\exp(3{,}159 - 2{,}671x_1 - 15{,}896x_3 - 1{,}485x_5 + 0{,}083x_6)}{1 + \exp(3{,}159 - 2{,}671x_1 - 15{,}896x_3 - 1{,}485x_5 + 0{,}083x_6)}$$

Based on Table 9, the parameter significance test has been carried out using a subset and the best model. Variables that significantly influence the financial distress of manufacturing companies are liquidity ratio (X_1), profitability ratio (X_3), activity ratio (X_5), and ROI (X_6). This can be seen from the P-value less than α 5%.

3. Classification Accuracy

Table 10 Classification Accuracy Table

Classification Table									
	Observed		Predicted						
			Data *Training*			Data *Testing*			
			y		Percentage Correct	y		Percentage Correct	
			0	1		0	1		
Step 4	Y	0	79	31	71.8	61	48	56.0	
		1	26	84	76.4	35	80	69.6	
	Overall Percentage				74.1			62.9	

Based on Table 10, the classification accuracy (accuracy rate) on the overall training data is 73.6% and the proportion of correct (accuracy rate) on the overall testing data is 60.7%.

4.4 Financial Distress Classification using CART

Table 11 Importance Variables for Developing Maximum Classification Tree

Variable	Score
Liquidity ratio (X_1)	100.00
Profitability ratio (X_3)	98.66
Solvency ratio (X_4)	94.49
ROI (X_6)	92.43
Activity ratio (X_5)	72.26
Profitability ratio (X_2)	49.91
ROE (X_7)	49.74

Based on the Table 11 it is known that the X1 variable has the highest variable score, 100.00, it can be said that the variable liquidity ratio (X1) is the most important factor in classifying the financial distress of manufacturing companies. In addition, there are several other variables that also have a big influence on the formation of classification trees, namely the profitability ratio (X3), solvency ratio (X4), ROI (X6) and activity ratio (X5).

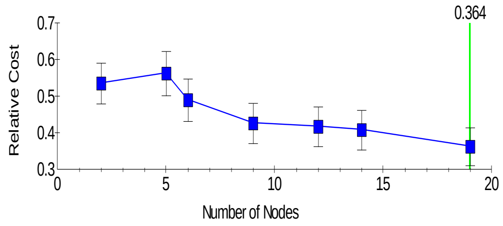

Figure 2 Relative Cost and Terminal Node Plot

From relative cost plot of the classification tree with a terminal node of 19 nodes indicated optimal result by the resulting relative cost value of 0.364, marked with a green line. Whereas, for the error cost (cross-validation relative cost) it produces the minimum value that is equal to 0.364 ± 0.064 or ranges from 0.400 to 0.300 with a complexity value of 0,000.

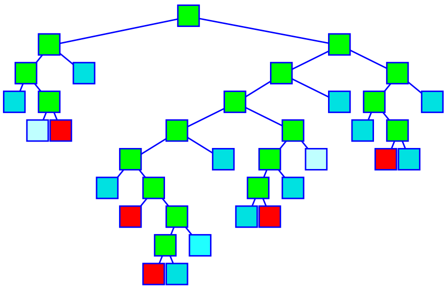

Figure 3 Classification Tree Topology Optimum with Gini Index

The solvency ratio variable (X4) divides the main node (node 1) into the left node and the right node. The node will sort left if the solvency ratio variable (X4) ≤ 1.020 (node 5) and right if the opposite (node 5). A total of 74 data is ≤ 1,020 and became the left node (node 2) and 146 > 1,020 became members of the right node (node 5). Node 2 consists of 74 manufacturing companies with a value of solvency ratio ≤ 1.020, then sorted into new vertices left and right according to the activity ratio (X5). If the value of the activity ratio ≤ 0.599 will be sorted out as members of the new left node (node 3), but if the value of the activity ratio > 0.599 then the manufacturing company will be sorted into the new right node (terminal 4 node). Among 74 companies that want to become a member of node 2, the results show that there are 15 manufacturing companies that are members of node 3 with the characteristic value of solvency ratio (X4) ≤ 1.020 and the value of the activity ratio ≤ 0.599. The remaining 59 manufacturing companies are members of terminal 4 node with the characteristic value of solvency ratio (X4) ≤ 1,020 and the value of the activity ratio > 0.599. Node 3 consisting of 15 manufacturing companies with the characteristics of the value of the solvency ratio (X4) ≤ 1.020 and the value of the activity ratio ≤ 0.599, then sorted into new vertices left and right according to the profitability ratio (X2). If the value of profitability ratio (X2) ≤ 0.155, then the manufacturing company will be sorted into the new left node (terminal node 1). Meanwhile, if the value of the profitability ratio (X2)> 0.155, it will be sorted into the new right node (node 4).

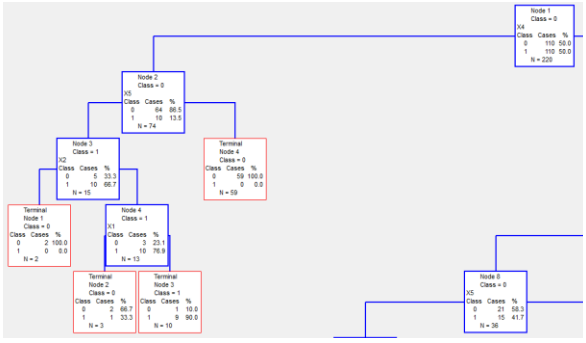

Figure 4 Highlight Optimum Classification Tree

The characteristics of each class on the response variable can be determined by tracing the optimal classification tree that has been formed based on the rules of sorting the Gini index. Strong characteristics are found at the terminal node which has the largest percentage of class (100%) on class labeling. Following are the characteristics of each class presented in Table 12.

Table 12. Class Characteristic Manufacturing Company Based on Higher Class Percentage

Class	Karakteristik
0 Non-Financial Distress	Activity ratio ≤ 0,599, and Solvency ratio ≤ 1,020
(1) Financial Distress	Liquidity ratio ≤ 0,634, Profitability ratio ≤ 0,155, Activity ratio ≤ 0,599, and Solvency ratio ≤ 1,020

Following is a comparison of the results of the accuracy of the optimal tree classification shown in Table 13.

Table 13. Comparison of Optimum Accuracy Classification Tree

Classification Tree	Accuracy (%)	
	Testing	*Training*
Optimum Tree	81,8	82,1

Based on Table 13 above, information can be obtained that for testing data the optimal tree accuracy is smaller than the training data which is 81.8% greater than 82.1%.

4.5 Result Comparison

Table 14 Accuracy Classification Comparison

Method	Accuracy (%)	
	Testing	*Training*
Binary Logistic Regression	62.9	74.1
CART	81,8	82,1

According to Hosmer and Lemeshow [10] one of measurement the goodness of the model is if it has a minimal chance of misclassification and the accuracy of prediction of the maximum model. Based on Table 14 for the Testing data obtained the accuracy of the classification of logistic regression analysis is 62.9% and the accuracy of the classification of CART is 81.8%. This shows that the CART method has a greater classification accuracy value 18.9% than the CART method. However, it can be said that both methods are good enough in predicting the response variable, in this case is financial condition of manufacturing companies.

The difference in the level of accuracy of predictions can be caused by differences in the results of the classification. In logistic regression the predictor variables selected as variables that influence the response variable are the liquidity ratio, profitability ratio, solvency ratio, and ROI. Whereas in the CART method the selected variables are profitability ratios, profitability ratios, solvency ratios, activity ratios, and ROI. Variable profitability ratios, solvency ratios, and ROI are variables that consistently influence logistic regression analysis and CART methods.

5. CONCLUSION AND SUGGESTION

The results of the comparison of the two methods namely for Testing data obtained the accuracy of the classification of logistic regression analysis is 62.9% and the accuracy of the CART method is 81.8%. This shows that the CART method has a greater classification accuracy value of 18.9% than the binary logistic regression method. However, it can be said that both methods are good enough in predicting the response variable in this case the financial condition of manufacturing companies.

Suggestions for future researchers, the addition of predictor variables with nominal scale is highly recommended if using the same method for reasons of ease in reading the output for class characteristics. Also, the data used should be added so that each class / group of data is represented and not hampered in the analysis process.

REFERENCES

[1] Kompasiana. 2013. http://www.kompasiana.com/virginiafladian/dampak-krisis-global-terhadap-ketenagakerjaan-diindonesia_55282 627f17e61a01d8b4603. Accessed 21 November 2015.

[2] R. K. Brahmana, "Identifying Financial distress Condition in Indonesia Manufacture Industry", Birmingham Business School, University of Birmingham United Kingdom, 2007.

[3] J. Sun and H. Li, "Dynamic financial distress prediction using instance selection for the disposal of concept drift", *Expert Systems with Applications*, vol. 38, no. 3, pp. 2566-2576, 2011.

[4] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*. New York: Chapman Hall, 1993.

[5] D. Hosmer and S. Lameshow, *Applied Logistic Regression*. New York: Wiley, 1989.

[6] Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley dan Sons, Inc