Comparison of Discriminant Analysis and Adaptive Boosting Classification and Regression Trees on Data with Unbalanced Class

EVA FADILAH RAMADHANI[1*], ADJI ACHMAD RINALDO FERNANDES[2], NI WAYAN SURYA WARDHANI[3]
Department of Statistics, University of Brawijaya
St. Veteran, Malang, East Java 65145
INDONESIA

Abstract: - This study aims to determine the best classification results among discriminant analysis, CART, and Adaboost CART on Bank X's Home Ownership Credit (KPR) customers. This study uses secondary data which contains notes on the 5C assessment (Collateral, Character, Capacity, Condition, Capital) and collectibility of current and non-current loans. The sample used in this study was from 2000 debtors. Comparison of classifications based on model accuracy, sensitivity, and overall specificity shows that Adaboost CART is the best method for classifying credit collectibility at Bank X. This is due to the class imbalance in the data. This study compares the classification results between parametric statistics, namely discriminant analysis and non-parametric statistics, namely CART and Adaboost CART. The results of the research can be used as material for consideration and evaluation for banks in determining the policy for providing credit to prospective borrowers from the classification results of KPR Bank X consumers.

Key-Words: - Classification, Discriminant Analysis, CART, Adaboost CART, Unbalanced Class, Credit Scoring Model

Received: June 10, 2021. Revised: November 12, 2021. Accepted: November 26, 2021. Published: December 14, 2021.

1 Introduction

In this era, the existence of a bank is very important for human life. The number of customers who come to make transactions is increasing from time to time. The number of customers must be balanced with new products and services provided by each bank to prevent a decrease in the number of customers. One of the services offered by banks for one of the primary needs is a Home Ownership Credit (KPR).

KPR is one of the credit services offered by banks to customers who apply for special credit to fulfill the need for a residence. Before a bank gives credit to a debtor, it is necessary to have an assessment from the bank to measure whether the debtor is able to pay his obligations in credit or not. One of the credit problems is the existence of debtors who have non-current credit so that it can harm the bank. From these problems, of course there needs to be supervision in terms of credit, namely by grouping which is used to determine the characteristics of debtors who fulfill credit obligations or not. Statistical analysis that can be used to deal with these problems is to use discriminant analysis and CART.

Discriminant analysis is a multivariate analysis method that aims to find a differentiating function in two or more response groups [1]. Another benefit of the discriminant function, in addition to being used to explain differences between groups, can also be used for classification. Research for credit scoring classification using discriminant analysis has been conducted by Azkya et al [2], Mukid and Widiharih [3], and Rahmadeni [4].

Apart from using discriminant analysis for the classification of KPR Bank X consumers, this study also uses CART to compare the level of classification accuracy. CART is a supervised learning method, each data has its own class label. This method was developed by Breiman, Friedman, Olshen and Stone in 1984. CART is a nonparametric statistical method that can be used for classification for both categorical and continuous scale response variables [5].

Credit collectability data at almost all banks are class unbalanced. The challenge in class imbalance has been a concern of academics and researchers in recent years. Class imbalance is a condition in which there is an unequal number of classes contained in a data set (uneven data distribution). Class imbalance can also be interpreted as a condition in a data set where there are large classes while other classes are only represented by a few objects [6]. In the case of bank credit, customers with current credit have a much larger proportion

than the non-current class. This resulted in inaccurate classification results using CART.

Research related to credit collectability classification has been conducted by Rofitanur [7] with the title "Application of Integration of Hybrid Mutual Clustering and Discriminant Analysis on Collectability of Bank X Malang City". This research uses hybrid mutual clustering which is integrated with discriminant analysis in the collectability grouping of prospective Bank X debtors based on 5C (Collateral, Character, Capacity, Condition, Capital).

Komarudin [8] conducted a study to compare the accuracy level of classification between the discriminant analysis method and CART with the title "Comparison of the Discriminant Analysis Classification Method and CART". The data used in this research is the result of the semulation data. Research shows that the CART method is relatively better than the discriminant analysis method. This can be seen from the percentage of the CART misclassification rate is smaller than the discriminant analysis and results in a more consistent classification.

Meanwhile, research related to classification with unbalanced class has been conducted by Efendi et al. [9] entitled "Ensemble Adaboost in Classification and Regression Trees to Overcome Class Imbalance in Credit Status of Bank Customers". This study aims to determine the classification results using the CART and Adaptive Boosting (Adaboost) CART method on bank loan data or credit collectibility where there is a class imbalance. The results showed that the Adaboost CART had higher classification accuracy than the classic CART.

Based on the description above, this study aims to compare the classification results of the collectability of KPR type credit at Bank X where the data obtained are classified as unbalanced class using discriminant analysis methods, CART, and Adaboost CART. Comparison of the results of the tresbut classification using 3 (three) criteria, namely accuracy, sensitivity and specificity. The data used are the 5C (Collateral, Character, Capacity, Condition, Capital) assessment data and credit collectability at the Bank which are used to assess the feasibility of providing credit.

2 Literature Review

2.1 Discriminant Analysis

Discriminant analysis is a multivariate analysis method that aims to separate different objects of observation and allocate new objects of observation into defined groups [1]. According to Solimun et al. [10], one of the uses of the function in discriminant analysis is to predict alternatives to the response variable category.

Discriminant analysis is appropriate when the response variable is a categorical variable (nominal or non-metric) and the predictor variable is a metric variable. In many cases, the response variable consists of two groups or classifications, for example, men versus women or high versus low. When the response variable consists of two classifications, this technique is referred to as two-group discriminant analysis [11].

The model of discriminant analysis is an equation that shows a linear combination of various predictor variables:

$$Y_i = a_0 + a_1 X_{1i} + a_2 X_{2i} + ... + a_p X_{pi} \quad (1)$$

Y_i : response variables (categorical data)

a_p : coefficient of the discriminant function on the p-predictor variable

X_{pi} : p predictor variable in the i object

p : number of variables, $p = 1,2,3,..., w$

i : number of objects, $i = 1,2,3,..., n$

2.1.1 Discriminant Analysis Asumption

a. Multivariate Normal Assumptions

The multivariate normal assumption test is carried out to determine whether the sample taken comes from a normal distribution or not. According to Johnson and Winchern [1], testing the normality assumption can be done using the mahalanobis distance value for the *i-th* observation (d_i^2) obtained by the following equation:

$$d_i^2 = \left(\mathbf{x_i} - \bar{\mathbf{x}}\right)' \mathbf{S}^{-1} \left(\mathbf{x_i} - \bar{\mathbf{x}}\right) \quad (2)$$

d_i^2 : Mahalanobis distance value for the i-th object

$\mathbf{x_i}$: the value vector of the i-th object

$\bar{\mathbf{x}}$: vector mean value of each variable

$\mathbf{S}^{-1}$: variance matrix

i : number of objects, dimana $i = 1,2,3,...,n$

Hypothesis:

H_0 : Multivariate normal distribution data

H_1: Data were not normally distributed multivariate

Next make a plot between the mahalanobis distance with the quantile value of chi square. When plots are formed they tend to form straight lines and there are more than 50% of the total number of observations that have a value $d_j^2 < \chi^2_{p,(0.5)}$, then the multivariate normal assumptions are fulfilled.

b. Assumption of Homogeneity of Variance Matrix

One of the assumptions that must be met when comparing two or more vector means of multivariates is that the variance of variance matrices of different populations are the same. One way to test the similarity of the variance matrix is the Box's M. test (Rancher, 2002).
Hypothesis:

H0: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma$

H1: there is at least 1 different group $\Sigma_g \neq \Sigma$

Testing Criteria:

H_0 rejected if $C > \chi^2_{p(p+1)(g-1)}$ or the p-value <α, which means that the variance matrix between groups is not homogeneous.

2.2 CART

This method utilizes a decision tree algorithm developed by Breiman, Friedman, Olshen and Stone in 1984. CART is a nonparametric statistical method used to perform classification analysis, both for categorical and continuous response variables [5]. CART results depend on the scale of the response variable. If the response variable has a continuous scale, the resulting tree model is regression trees. Meanwhile, if the response variable is categorical in scale, the resulting tree is classification trees [5].

CART analysis is known as Binary Recursive Partitioning. The process is called binary because each parent node is split into exactly two child nodes. Meanwhile, recursive means that the binary process can be applied many times. This solving process will continue until there is no more opportunity to do the next solution. The term partitioning means that the data set is divided into smaller parts or partitions [12].

2.3 Adaboost CART

Boosting is one of the popular methods used in machine learning. It is designed for problems related to classification and is applied to weak classifiers. Adaptive Boosting (Adaboost) is an enhancement algorithm developed with classifiers [13]. Adaboost can improve the accuracy of various classification methods such as Decision stumps, Decision tree, Multi-Layer perceptron, and Support Vector Machines (SVM). Adaboost is a method that incorporates a weak classifier, which is generated repeatedly from a weighted sampling sample, with adaptively adjusted weights at each step to provide added weight in cases of misclassifying the previous step.

2.4 Accuracy, Sensitivity and Specificity

Accuracy is the accuracy of the model created. Sensitivity measures the proportion of true positives identified correctly, specificity measures the proportion of true negatives that are correctly identified. False positives, known as type I errors, occur when a case that should have been classified as negative is classified as positive. False negatives, known as type II errors, occur when cases that are supposed to be classified as positive are classified as negative [14]. The value of accuracy, sensitivity and specificity can be calculated using the confusion matrix as presented in Table 1.

Table 1. Confusion Matrix

Actual	Prediction	
	Performing loan	Non performing loan
Performing loan	*a*	*b*

Non performing loan	*c*	*d*

Using the confusion matrix in Table 1, the accuracy, sensitivity and specificity formulas are:

$$Accuration = \frac{a+d}{a+b+c+d} \quad (3)$$

$$Sensitivity = \frac{a}{a+c} \quad (4)$$

$$Specificity = \frac{d}{b+d} \quad (5)$$

The criteria for selecting the best method are based on the method that has the greatest percentage of accuracy, sensitivity and specificity.

3 Methods

3.1 Data

The study was conducted using secondary data which contains notes on the 5C assessment and collectability of performing loans and non-performing loans. The sample used in this study was from 2000 debtor from Bank X.

3.2 Steps

a) Discriminant Analysis Steps:

1. Testing Multivariate Data Assumptions using the Mahalanobis distance.
2. Testing the assumption of homogeneity of the variance matrix using Box's M test.
3. Creating a Discriminant Analysis Model
4. Accuracy, Sensitivity and Specificity

b) CART steps:

1. Dividing data into two parts, training data and testing data; randomly by 80%: 20%
2. Determining the best sorter that provide the highest level of impurity based on the goodness of split criteria
3. Class labeling and performing validation using K-fold cross-validation
4. Accuracy, sensitivity and specificity

c) Adaboost CART Steps:

1. Dividing data into two parts, training data and testing data; randomly by 80%: 20%
2. Initialization of weight of training data , for all
3. Sampling of N data from training data with resampling bootstrap
4. Determining classification tree with CART method
5. Calculating classification error and determining weighting vote
6. Updating weight
7. Doing step 3 until 6 as many of T (T = 1.000).
8. Determining final classifier
9. Accuracy, sensitivity and specificity

4 Results and Discussion

4.1 Discriminant Analysis

4.1.1. Normal Multivariate Asumotion

Testing the assumption of multivariate normality is done with a Q-Q plot and a graph is obtained as shown in Figure 1.

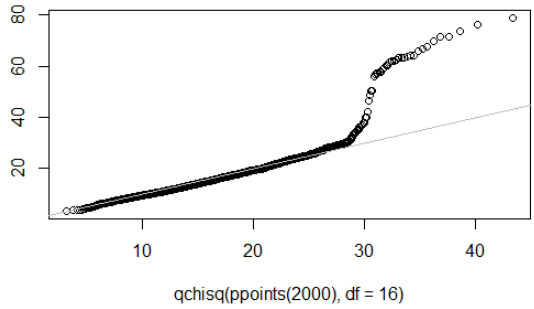

Fig. 1: Mahalanobis Distance Plot and Chi Square Quantile

From figure 1. It can be seen that the plot between the mahalanobis distance and the chi square quantile does not follow a straight line, and there are several data sets that spread very far from the straight line which indicates that there are outliers, which means that the data cannot be approached with a multivariate normal distribution [15].

4.1.2. Assumption of Homogeneity of Variance Matrix

The assumption test for the homogenity of the variance matrix in this study uses the Box'M test. Obtained p value of 0.8518, which means that the variance matrix between groups is the same. Thus, the assumption of homogeneity of the variance matrix is fulfilled so that it can be continued for linear discriminant analysis.

4.1.3. Discriminan Analysis Model

The discriminant analysis model is used as a comparative model of classification using the discriminant analysis method with CART. Comparisons are made by looking at the percentage of accuracy, sensitivity and specificity. The following models and results of accuracy, sensitivity and specificity are presented in Equation 6 and Table 2.

$$\begin{aligned} y_i = & -0.628x_{11.1i} - 0.077x_{11.2i} - 0.066x_{12i} - \\ & 0.123x_{21i} + 0.386x_{22i} - 0.009x_{23i} + \\ & 0.345x_{24.1i} + 0.065x_{24.2i} + 0.665x_{25.1i} + \\ & 0.312x_{25.2i} + 0.090x_{31i} - 0.568x_{32.1i} - \\ & 0.527x_{32.2i} - 0.453x_{32.3i} + 1.342x_{33i} + \\ & 0.967x_{34i} - 0.665x_{35.1i} - 0.278x_{35.2i} - \\ & 0.181x_{35.3i} - 0.616x_{35.4i} + 0.005x_{36i} - \\ & 0.118x_{37i} - 0.206x_{4.1i} + 0.056x_{4.2i} - 0.658x_{5i} \end{aligned} \quad (6)$$

Table 2. Accuracy, Sensitivity and Specificity of Discriminant Analysis

Criteria	Results (%)
Accuracy	73.90
Sensitivity	88.13
Specificity	0.00

4.2 Classification and Regression Trees

CART is a supervised learning classification method. Before the data is analyzed using CART, the data is divided into training data and testing data. Data sharing uses the Pareto principle which states that for many events, about 80% of the effect is due to 20% of the causes. So in this study the data is divided into training data and testing data of 80:20. Training data is used to form a classification tree, while testing data is used to validate the classification tree which is useful for determining the ability of the classification tree to predict new data.

The first step in CART analysis is solving the root node which can be seen in Figure 2. The same process continues on other nodes. The recurring sorting process will stop if it is no longer possible to do the sorting process because at the end of the classification tree there is an end node that has the same class members (homogeneous). The maximum classification tree is shown in Figure 2.

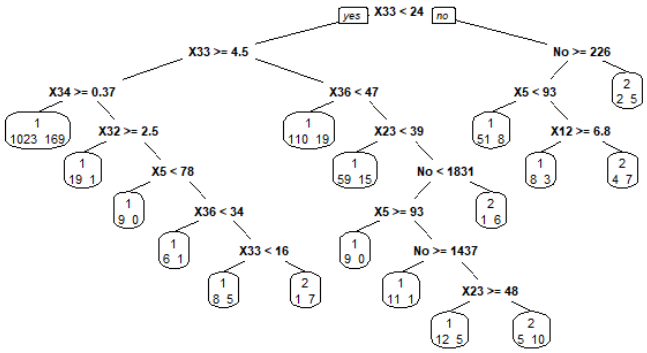

Fig. 2: Classification Trees

It can be seen in Figure 2 that the resulting classification tree can be interpreted easily. This causes the pruning process to be unnecessary which will cause some information to be lost, therefore the classification tree pruning stage is not carried out. From the results of the classification tree, the credit term variable is the most important independent variable so that it becomes the best sorting from the root node.

The classification accuracy can be seen using confusion matrix. From the confusion matrix, accuracy, sensitivity and specificity can be calculated from the classification results presented in Table 3.

Table 3. Accuracy, Sensitivity and Specificity of CART

Criteria	Results (%)
Accuracy	78.25
Sensitivity	81.72
Specificity	2.17

4.3 Adaboost CART

Adaboost is a method that combines classifiers iteratively created from weighted training data, with

weights adjusted adaptively at each step to give increased weight to cases that had misclassification in the previous step. The classifier used in the Adaboost method is the CART classification tree.

The first step in Adaboost CART is almost the same as the classic CART, which is dividing the data into 80:20 training and testing data. From the results of adaboost CART, the accuracy, sensitivity, and specificity are obtained as shown in the table below.

Table 4. Accuracy, Sensitivity and Specificity of Adaboost CART

Criteria	**Results (%)**
Accuracy	84.25
Sensitivity	97.97
Specificity	0.00

4.4 Classification Efficiency Comparison

Based on the discussion, the results of the classification accuracy were obtained based on the criteria of accuracy, sensitivity and specificity. The figure below presents the classification efficiency comparison between discriminant analysis, CART and Adaboost CART.

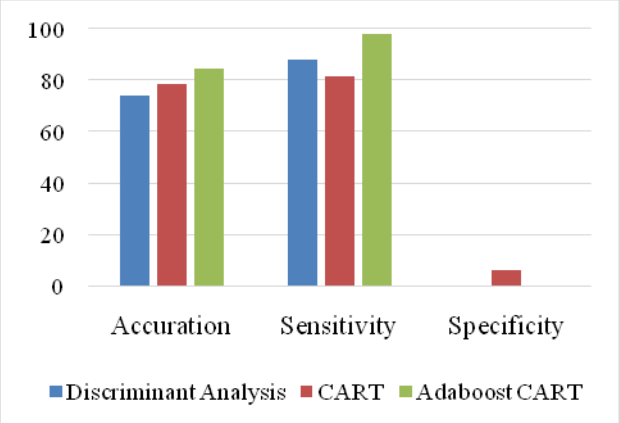

Fig. 3: Comparison of classification efficiency

Figure 3 shows that the classification accuracy using discriminant analysis is not higher than CART and Adaboost CART on all criteria. This is due to the violated multivariate normal assumptions, so that nonparametric statistics can be considered for use. This is evident from the comparison of the classification accuracy above, which shows that the results of the accuracy and sensitivity of Adaboost CART are greater than the other two methods. This is because the data used in this study are credit collectibility data at Bank X which has an unbalanced number of members in the two classes. Thus it can be concluded that Adaboost CART is a better method for classifying credit collectability at Bank X compared to discriminant analysis and CART .

5 Conclusion

5.1 Conclusion

The conclusions that can be drawn based on the results of the analysis are:

1. The classification results of discriminant analysis produce a model that is able to classify customers based on the collectability of KPR Bank X with an accuracy of 73.90%, 88.13% sensitivity and 0.00% specificity.
2. The classification results of CART resulted in the variable of credit period as the most important variable and was able to classify customers based on the collectability of KPR Bank X with an accuracy of 78.25%, 81.72% sensitivity and 5.88% specificity on the sample size.
3. The classification results of Adaboost CART are able to classify customers based on the collectability of KPR Bank X with an accuracy of 84.25%, 97.97% sensitivity and 0.00% specificity.
4. Comparison of classification accuracy based on the overall shows that Adaboost CART is the best method in classifying credit collectibility at Bank X with unbalanced data.

5.2 Recommendation

1. Further research can examine the credit scoring model with 3 categories of credit collectability as the response variable.
2. Bank X is advised to use the Adaboost CART for the classification of customer credit collectibility which can be used as a reference.

Reference:

[1] Johnson, R. A. and Wichern, D. W. 2007. *Applied Multivariate. Analysis*. Upper Saddle River, NJ: Prentice Hall.

[2] Azkiya, M., Mukid, M.A. and Ispriyanti, D., 2015. Klasifikasi Nasabah Kredit Bank "X" Di Provinsi Lampung Menggunakan Analisis *Diskriminan Kernel. Jurnal Gaussian*, *4*(4), pp. 937-946.

[3] Mukid, M.A. and Widiharih, T. 2016. Model Penilaian Kredit Menggunakan Analisis

Diskriminan dengan Variabel Bebas Campuran Biner dan Kontinu. *Media Statistika, 9*(2), pp. 107-117.

[4] Rahmadeni, R., 2019, November. Analisis Diskriminan Fisher Untuk Klasifikasi Risiko Kredit. In Seminar Nasional Teknologi Informasi Komunikasi dan Industri (pp. 478-481).

[5] Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. 1984. *Classification and Regression Trees*. US: CRC press.

[6] Sun, Y., Wong, A.K. and Kamel, M.S., 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), pp. 687-719.

[7] Rofitanur, N. 2020. "Penerapan Integrasi Hybrid Mutual Clustering dan Analisis Diskriminan pada Kolektibilitas Bank X Kota Malang".

[8] Komarudin, K. 2010. Perbandingan Metode Klasifikasi Analisis Diskriminan dan Classification and Regression Trees (CART).

[9] Efendi, A., Fitriani, R., Naufal, H. I., and Rahayudi, B. 2020. Ensemble Adaboost in Classification and Regression Trees to Overcome Class Imbalance in Credit Status of Bank Customers. *Journal of Theoretical and Applied Information Technology*, *98*(17), pp. 3428-3427.

[10] Solimun, Fernandes, A. A. R. and Nurjannah. 2017. *Metode Statistika Multivariat Pemodelan Persamaan Struktural (SEM) Pendekatan WarpPLS*. Malang: UB Press.

[11] Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2014. *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

[12] Lewis, R. J. 2000. An introduction to classification and regression tree (CART) analysis. In Annual meeting of the society for academic emergency medicine in San Francisco, California.

[13] Schapire, R.E. and Freund, Y., 2012. Foundations of machine learning.

[14] Bramer, M. 2013. *Pronciple of Data Mining*. London: Springer

[15] Fernandes A.A.R, Solimun, Nurjannah, Hutahayan B. 2020. Comparison of use of linkage in integrated cluster with discriminal analysis approach. *International Journal of Advanced Science and Technology*, 29(3), 5654 - 5668

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Eva Fadilah Ramadhani: Conceptualization, Formal analysis, Methodology, Writing - original draft

Adji Achmad Rinaldo Fernandes: Data curation, Project administration, Writing - review & editing

Ni Wayan Surya Wardhani: Methodology, Supervision, Writing - review & editing

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This study did not receive any funding in any form.