

Comparison of Distance and Linkage in Integrated Cluster Analysis with Multiple Discriminant Analysis on Home Ownership Credit Bank in Indonesia

Ni Made Ayu Astari Badung, Adji Achmad Rinaldo Fernandes*, Waego Hadi Nugroho

Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University, Indonesia

Received August 30, 2021; Revised October 21, 2021; Accepted November 11, 2021

Cite This Paper in the following Citation Styles

(a): [1] Ni Made Ayu Astari Badung, Adji Achmad Rinaldo Fernandes, Waego Hadi Nugroho, "Comparison of Distance and Linkage in Integrated Cluster Analysis with Multiple Discriminant Analysis on Home Ownership Credit Bank in Indonesia," *Mathematics and Statistics*, Vol. 9, No. 6, pp. 958 - 975, 2021. DOI: 10.13189/ms.2021.090612.

(b): Ni Made Ayu Astari Badung, Adji Achmad Rinaldo Fernandes, Waego Hadi Nugroho (2021). *Comparison of Distance and Linkage in Integrated Cluster Analysis with Multiple Discriminant Analysis on Home Ownership Credit Bank in Indonesia*. *Mathematics and Statistics*, 9(6), 958 - 975. DOI: 10.13189/ms.2021.090612.

Copyright©2021 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract This study aims to compare the size of distance (Euclidean distance, Manhattan distance, and Mahalanobis distance) and linkage (average linkage, single linkage, and complete linkage) in integrated cluster analysis with Multiple Discriminant Analysis on Home Ownership Credit Bank consumers in Indonesia. The data used are secondary data from the 5C assessment on Bank consumers in Indonesia. The data contain notes on the 5 C assessment as well as 3 credit collectability (current, special mention, and substandard) from Home Ownership Credit customers. The population in this study were all Home Ownership Credit customers in all banks in Indonesia. The sampling technique used was purposive random sampling. The sample size is 300 customers from customer data at three branches of Bank in Indonesia. This research is a quantitative study using cluster analysis integrated with multiple discriminant analysis. The best method for classifying Home Ownership Credit Bank customers based on the 5C variable assessment is an integrated cluster analysis with Multiple Discriminant Analysis based on the Mahalanobis distance with 2 clusters, namely the high cluster and the low cluster. Use of an integrated cluster with Multiple Discriminant Analysis to compare distance and linkage measures. In addition, the objects used are Home Ownership Credit Bank customers in Indonesia.

Keywords Multiple Discriminant Analysis, Cluster Analysis, Distance Measures, Linkage, Home Ownership

Credit

1. Introduction

One type of credit provided by the bank is a Home Ownership Loan (HOC). HOC is one of the financing products provided by banks for home buyers with a financing scheme of up to 90% of the house price. Debtors who have non-current credit are one of the credit problems that can harm the bank. Before the bank gives credit to the debtor, it is necessary to measure by the bank whether the debtor can carry out his obligations under the credit or not. From these problems, there is a need for supervision in terms of providing HOC, one of the statistical tools that can be used in this problem is cluster analysis which is integrated with MDA.

Multiple Discriminant Analysis (MDA) is one part of discriminant analysis. MDA and discriminant analysis have differences that lie in the number of categories of response variables. In discriminant analysis, the number of response variable categories is only one to two, while MDA has more than two categories of response variables. Discriminant analysis is a multivariate analysis that functions to model the relationship between a categorical response variable and one or more quantitative explanatory variables [3]. MDA can be used as a clustering method

because it produces a function that can distinguish between clusters. The function is formed by maximizing the distance between clusters.

MDA has several drawbacks. One of the drawbacks of MDA is that the sample to be tested has members in each sub-sample that have the same characteristics. If this happens, it will result in low model accuracy that results in discriminant analysis. Therefore, before conducting discriminant analysis, a more in-depth analysis is needed first. One solution is to use cluster analysis which is integrated with MDA. Integrating cluster analysis with MDA can be done by using dummy variables obtained from cluster results. Many clusters formed are used as categories, then used as dummy variables. The integration model of cluster analysis with MDA with a dummy variable approach is the same as the general model of multiple linear regression analysis with dummy variables. The principle of the integration model of cluster analysis with MDA is to use research variables and also dummy variables that have been multiplied by each research variable and then differentiated based on the clusters formed so that they will form the MDA model as described by [4].

Cluster analysis is an analysis of multiple variables (multivariate) which is included in the interdependence method, namely the independent or explanatory variables are not distinguished from the dependent or response variables [1]. According to [2], in research, there are two analyses that are often used, namely univariate analysis and multivariate analysis. If the research variables are observed together, it is appropriate to use multivariate analysis. Cluster analysis aims to group objects into several clusters, where the clusters have different properties. In general, there are two methods in cluster analysis, namely the hierarchical method and the non-hierarchical method. The hierarchical method consists of several methods, namely the Single Linkage method, the Average Linkage method, the Complete Linkage method, the Central Linkage method, and the Ward method (Ward's Method). The method included in the non-hierarchical method is the K-Means method.

The hierarchical method and the non-hierarchical method have differences that lie in the determination of the number of clusters. In the hierarchical method, the number of clusters has not been determined, while the non-hierarchical method has determined the number of clusters. Hierarchical methods have advantages over non-hierarchical methods. The advantage of the hierarchical method is that it is easier to study all the clusters that are formed and more informative because the hierarchical method of grouping stages is presented in the form of a dendrogram or tree diagram.

In cluster analysis, one measure of similarity used is as distance. The distance measure is a measure of similarity, the higher the distance value, the lower the similarity between objects. There are several methods of measuring

distance, including Euclidean, Manhattan/City Block, Mahalanobis, Correlation, Angle-based, Squared Euclidean. In cluster analysis, several linkages can be used to form clusters. According to [19], the linkage method consists of single linkage, complete linkage, and average linkage. This study examines the application of integrated clusters in discriminant analysis with three different linkage methods (single linkage, average linkage, and complete linkage) and three different distance measures (euclidean, manhattan, and Mahalanobis).

This research was using integrated cluster and Multiple Discriminant Analysis (MDA) with three linkage methods (single, complete, and average) and three distance measures (Euclidean, Manhattan, and Mahalanobis). The size of the distance and the linkage method used can determine the results of the number of clusters formed. Therefore, this study wanted to obtain the best distance measure to maximize the measure of accuracy, sensitivity, and specificity when an integrated cluster MDA approach was carried out.

Previous research on linkage comparison has been carried out by [20]. This study aims to classify districts/cities in West Kalimantan based on HDI indicators. This study compares single, complete, and average linkage. The results showed that the average linkage method was the best in classifying districts/cities in West Kalimantan. This study has not compared the combination of the use of multiple distances and linkage.

Research conducted by [21] also conducted a study comparing several linkages. This study compares the average, complete, and ward linkage in cluster analysis. The results showed that average linkage is the best method in classifying districts/cities in Central Java based on HDI.

Nishom [22] conducted a study on the comparison of distance measures in cluster analysis that aims to Comparison of the Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance on the Chi-Square-based K-Means Clustering Algorithm. This study compares the Euclidean, Minkowski, and Manhattan distances. This study found that the Manhattan distance was the best.

In addition, [23] also researched the comparison of several distance measures in cluster analysis. This research is Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics. The results showed that the Minkowski distance was the best.

Fernandes et al. [6] conducted a study using discriminant analysis integrated with cluster analysis to obtain the best linkage method in classifying HOC Bank X customers. However, this study only uses credit collectability with two categories, and one measure of distance, namely the Euclidean distance. The problem experienced in this study is the data that does not meet the assumption of homogeneity of the variance matrix.

Research on MDA was also conducted by [24]. This study uses MDA to classify the potential for bankruptcy of

Islamic commercial banks in Indonesia. The results showed that MDA can classify healthy and unhealthy customers correctly.

Based on the description above, this study aims to compare the distance measures (euclidean, manhattan, and Mahalanobis) in integrated cluster analysis with MDA for bank mortgage consumers. The data used are secondary data, namely the assessment of the 5C variable at the bank. The benefits of this study are reconstructing MDA modelling by adding cluster analysis to the model and providing information regarding the comparison of the value of classification accuracy, sensitivity, and specificity of integrated cluster analysis with distance and linkage-based MDA on big data-based simulation data. The originality of this study is the use of integrated clusters with MDA to compare distance and linkage measures. In addition, the object used is Bank KPR customers in Indonesia.

2. Literature Review

2.1. Cluster Analysis

According to [8], cluster analysis is a multiple variable analysis that aims to group n objects into k clusters with $k < n$ based on p variables, so that each unit object in one cluster has more homogeneous characteristics than the object units in other clusters.

2.1.1. Hierarchical Cluster Analysis

In the hierarchical cluster analysis, it is assumed that at first, each object is a separate cluster, then the two closest objects or clusters are combined to form one smaller cluster [9]. Hierarchical methods have advantages over non-hierarchical methods. The advantage of the hierarchical method is that it is easier to study all clusters that are formed and more informative because, in the hierarchical method, the stages of grouping are presented in the form of a dendrogram or tree diagram.

2.1.2. Average Linkage Method

The average linkage method is one part of the agglomerative method. In the average linkage method, the distance between two clusters is considered as the average distance between all members in one cluster and all members of the other clusters. According to [9], the grouping algorithm using the average linkage method is as follows:

1. Start by making all objects their own cluster, so that the number of clusters equals the number of objects.
2. Use a distance measure to join two objects that are closest to each other. Suppose there are objects A and B, these objects are combined to form a new cluster, namely AB.

3. Calculate the distance between AB clusters and others using the formula in equation (1).

$$d_{(AB)C} = \frac{\sum_i \sum_k d_{ik}}{N_{(AB)} N_C} \quad (1)$$

Where:

$d_{(AB)C}$: the average distance between cluster AB and cluster C

d_{ik} : the distance between object i in a cluster (AB) and object k in cluster C

$N_{(AB)}$: the number of objects in the cluster (AB)

N_C : the number of objects in cluster C

i : the number of objects, where $i : 1, 2, 3, \dots, n$

4. Repeat step (3) until all objects have become a single cluster.

2.1.3. Single Linkage Method

Determining the distance between clusters using the single linkage method can be done by looking at the distance between the two existing clusters, then selecting the closest distance or the nearest neighbor rule. According to [9], the grouping algorithm using the single linkage method is as follows:

1. Start by making all objects their own cluster, so that the number of clusters equals the number of objects.
2. Use a distance measure to join two objects that are closest to each other. Suppose there are objects A and B, these objects are combined to form a new cluster, namely AB.
3. Calculate the distance between AB clusters and others using the formula in equation (2).

$$d_{(AB)C} = \max(d_{AC}, d_{BC}) \quad (2)$$

Where:

$d_{(AB)C}$: the average distance between cluster AB and cluster C

d_{AB} : the distance between object A in cluster C and object B in cluster C

$N_{(AB)}$: the number of objects in the cluster (AB)

N_C : the number of objects in cluster C

4. Repeat step (3) until all objects have become a single cluster.

2.1.4. Average Linkage Method

In the average linkage method, the distance between two clusters is considered as the average distance between all members in one cluster and all members of the other clusters. According to [9], the grouping algorithm using the average linkage method is as follows:

1. Start by making all objects their own cluster, so that the number of clusters equals the number of objects.
2. Use a distance measure to join two objects that are closest to each other. Suppose there are objects A and B, these objects are combined to form a new cluster, namely AB.

- Calculate the distance between AB clusters and others using the formula in equation (3).

$$d_{(AB)C} = \frac{\sum_A \sum_B d_{AB}}{N_{(AB)}N_C} \quad (3)$$

Where:

$d_{(AB)C}$: the average distance between cluster AB and cluster C

d_{AB} : the distance between object A in cluster C and object B in cluster C

$N_{(AB)}$: number of objects in the cluster (AB)

N_C : the number of objects in cluster C

- Repeat step (3) until all objects have become a single cluster.

2.1.5. Cluster Analysis Distance Calculation

According to [10] the concept of similarity is important in cluster analysis because the principle of cluster analysis is to group objects that have the same characteristics. This study uses three measures of distance, including:

1. Euclidean distance

Euclidean distance is the geometric distance between two data objects [4]. The Euclidean distance between two points can be calculated using equation (4).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^p (x_{ri} - x_{rj})^2} \quad (4)$$

Where:

$d(x_i, x_j)$: distance between i and j

x_{ri} : the value of the variable r in observation i

x_{rj} : the value of the variable r in observation j

p : lots of data variables

2. Manhattan distance

Prasetyo [11] states that the Manhattan distance is perfect for detecting outliers in the data. The Manhattan distance between two points can be calculated using equation (5).

$$d(x_i, x_j) = \sum_{r=1}^p |x_{ri} - x_{rj}| \quad (5)$$

Where:

$d(x_i, x_j)$: distance between i and j

x_{ri} : the value of the variable r in an observation i

x_{rj} : the value of the variable r in observation j

p : lots of data variables

3. Mahalanobis distance

According to [12], Mahalanobis distance can not only solve the problem of differences in data scales but also consider the effect of correlation between variables. The Mahalanobis distance between two points can be calculated using equation (6).

$$D_{ij}^2 = \frac{1}{(1-r^2)} \left[\frac{(x_{i1}-x_{j1})^2}{s_1^2} - 2r \frac{(x_{i1}-x_{j1})(x_{i1}-x_{j1})}{s_1 s_1} + \frac{(x_{i2}-x_{j2})^2}{s_2^2} \right] \quad (6)$$

Where:

D_{ij}^2 : observation distance to-i and j-th

S_1^2 : variant for variable number 1

S_2^2 : variant for variable 2nd

$S_1 S_2$: covariance variable 1st and 2nd

2.1.6. Dendrogram

Dendrogram is a mathematical and visual representation of the grouping procedure performed using hierarchical cluster analysis. The shape of the dendrogram is identical to that of a tree diagram. The points on the dendrogram represent the clusters, while the length of the bars represents the distance at which the objects are combined in the cluster. An example of a dendrogram is presented in Figure 1.

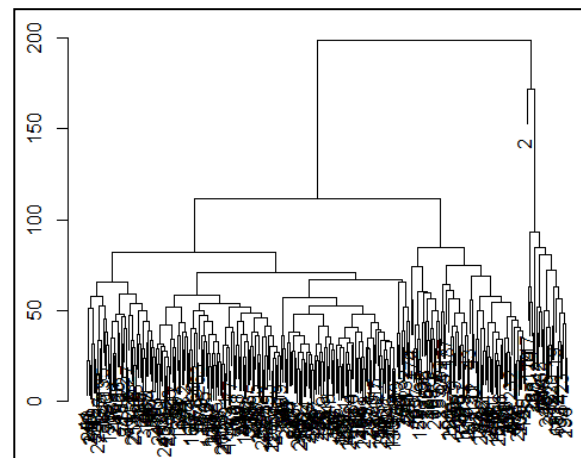


Figure 1. Dendrogram

2.2. Multiple Discriminant Analysis (MDA)

MDA is a part of discriminant analysis. The difference between MDA and discriminant analysis is in the number of categories of response variables. In the discriminant analysis, the number of response variable categories is only one to two, while MDA has more than two categories of response variables. Discriminant analysis is a multivariate analysis that functions to model the relationship between a categorical response variable with one or more quantitative explanatory variables [3].

According to [9], MDA is included in the multivariate dependence method. The model can be written as in equation (7).

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} \quad (7)$$

Where:

Y_i : response variable, namely categorical or nominal data in the unit of observations

X_{pi} : explanatory variable to-p in the i-th unit of observation

β_p : coefficient of p-th discriminant function

i : 1,2,3,...,n

2.3. Assumption Multiple Discriminant Analysis

2.3.1. Multivariate Normal Assumptions

According to [9] normality testing can be done using the Mahalanobis distance value for the i-th observation (d_i^2) obtained by the following formula:

Hypothesis:

H0: The data are multivariate normally distributed

H1: The data are not normally distributed multivariate

$$d_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \quad (8)$$

Where:

\mathbf{X}_i : vector of observed value to-i

$\bar{\mathbf{X}}$: vector average value of each variable

\mathbf{S} : sample variance matrix

Then make a plot between the Mahalanobis distance and the quantile value of chi-square. If the plot formed tends to form a straight line and there are more than 50% of the total number of observations that have valued $d_i^2 < \chi_{p;0,05}^2$, where p is many variables. Then the data can be approximated by a multivariate normal distribution [9].

2.3.2. Assumption of Homogeneity of Variance Matrix

One of the assumptions made when comparing two or more vector means of multi-variates is that the variance of variance matrices of different populations is the same. One way to test the similarity of the variance matrix is the Box's M test. According to [13], the distribution of the M test statistic can be examined with the F distribution which has the following hypothesis and test statistics:

Hypothesis:

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

H1: there is at least one different cluster ($\Sigma_g \neq \Sigma$)

Box's M Test Stats:

$$C = (1 - u) M \quad (9)$$

Where:

$$u = \left[\frac{\Sigma_g}{(n_g - 1)} - \frac{1}{\Sigma_g(n_g - 1)} \right] \left[\frac{2p^2 + 3p - 1}{6(p+1)(g-1)} \right] \quad (10)$$

$$M = \left[\Sigma_g(n_g - 1) \ln |\mathbf{S}_{pooled}| - \Sigma_g(n_g - 1) \ln |\mathbf{S}_g| \right] \quad (11)$$

$$S_{pooled} =$$

$$= \sum_i \frac{1}{(n_i - 1)} \{ (n_1 - 1)S_1 + (n_2 - 1)S_2 + \dots + (n_g - 1)S_g \} \quad (12)$$

Σ_g : variance matrix on the cluster to g

p : many explanatory variables

g: many categories of response variables

S_{pooled} : compound variance matrix in cluster

S_g : variance matrix sample at cluster to g

Testing Criteria

Reject H0 if $C > \chi_{p(p+1)(g-1)}^2$ or p-value $< \alpha$ which means that the variance matrix between groups is not homogeneous.

2.4. Integration of Cluster Analysis with Multiple Discriminant Analysis with Dummy Variable Approach

Integration of Cluster Analysis with MDA. Dummy variable approach in this study combines cluster analysis with MDA. Integrating cluster analysis with MDA can be done by using dummy variables obtained from the cluster results. The principle of the cluster analysis integration model with MDA is to use research variables and also dummy variables that have been multiplied by each research variable and differentiated based on the cluster formed so that it will form a multiple discriminant model as described by [4].

The integrated cluster model with MDA can be written in equation (13).

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + D_1 \beta_{p+1} x_{1i} + D_1 \beta_{p+2} x_{2i} + \dots + D_1 \beta_{p+q} x_{pi} + D_2 \beta_{p+q+1} x_{1i} + D_2 \beta_{p+q+2} x_{2i} + \dots + D_2 \beta_{p+2q} x_{pi} + \dots + D_q \beta_{p+q+1} x_{1i} + D_q \beta_{p+q+2} x_{2i} + \dots + D_q \beta_{p+q} x_{pi} \quad (13)$$

Where:

y_i : the response variable in the unit of observation

x_{pi} : the p-explanatory variable in the unit of observation

β_p : coefficient of the discriminant function

D_q : variable dummy to-q

p : many explanatory variables

q: the number of clusters formed minus 1

i: 1,2,3,..., n

2.5. Classification Accuracy, Specificity, and Sensitivity

2.5.1. Classification Accuracy

According to [10], classification accuracy is said to be good if it is equal to or more than the chance of classification is added by a quarter. The greater the classification accuracy value, the better the method used. The value of the classification accuracy value can be calculated using equation (14).

$$\text{classification accuracy} = \frac{\sum_{i=1}^g n_{ii}}{N} \quad (14)$$

Where:

N: many observations in all clusters

n_{ii} : the number of observations that are appropriately classified from cluster actual i-th on cluster prediction

g: many clusters

2.5.2. Specificity and Sensitivity

According to Yerushelmy in [14], sensitivity is the ability to correctly diagnose people who are sick, meaning that the test results are positive and they are indeed sick, while specificity is the ability to correctly diagnose people who are not sick means that the test results are negative and indeed not sick. The greater the sensitivity and

specificity, the better the method used. The sensitivity and specificity formulas are described in equation (15).

$$S_e = \frac{a}{a+c} ; S_p = \frac{d}{b+d} \quad (15)$$

Where:

S_e : sensitivity value

S_p : specificity value

a: the number of observations from cluster positives right classified to cluster positive (true positive)

b: the number of observations from cluster negatives that are classified to cluster positive (false positive)

c: the number of observations from cluster positives that are classified to cluster negative (false negative)

d: the number of observations from cluster negatives are right classified to cluster negative (true negative)

2.6. Credit Principles

The principle of credit or what is also known as the 5C principle serves to provide information about the willingness to pay and the ability to pay customers in repaying loans and their interest [15]. The principle of credit with the 5C formula consists of character, capacity, capital, collateral, and conditions.

a Character

Character is a person's nature or character. The character of the prospective debtor can be seen from the background of the prospective debtor. An analysis of the character of a prospective customer needs to be carried out by the bank to find out that the prospective customer has the desire to fulfill the obligation to pay back the financing that has been carried out until it is paid off. The character can be used as a measure of customer willingness to pay [16].

b Capacity

Capacity is the customer's ability to run their business to earn a profit so that they can pay off loans and interest. Analysis of the capacity of prospective customers aims to measure the extent to which customers can pay off loans in a timely manner from the results of their businesses [17]. Banks need to know the financial capacity of a prospective customer because it is the main source of payment.

c Capital

Capital is the amount of capital or personal funds owned by a prospective customer. The greater the personal funds, the higher the seriousness of the prospective customer in running his business, and the bank will feel more confident in extending credit. This capital ability is manifested in the form of an obligation to provide self-finance, which should be greater than the credit requested from the bank [18].

d Collateral

Collateral is an item that is guaranteed by a prospective customer against the financing received. The collateral assessment conducted by the bank aims to determine the extent of the risk of the prospective customer's financial obligations to the bank. The bank will not provide financing that exceeds the value of the guarantee, except for certain financing guaranteed by certain parties.

e Condition

Condition is a credit assessment based on current economic, social, and political conditions and predictions for the future. Assessment of the condition of the business sector being financed should have really good prospects so that the possibility of credit problems being relatively small [16]. The assessment is directed at the surrounding conditions that affect the prospective customer's business.

f Credit Collections

According to the decision letter of the board of directors of Bank Indonesia No. 30/267 / KEP / DIR, dated 27 February 1998 regarding product quality and reserve formation, five categories of credit collectability were determined, namely current credit, special mention credit, substandard credit, doubtful credit, and bad credit. In this study, only three credit collectability were used, consisting of the current credit, special mention credit, and substandard credit.

1. Current, which is a condition where there are no arrears in principal installments, interest arrears, or overdrafts due to credit withdrawals. The current credit is also defined as a debtor making payments on the debt as well as interest on a timely basis.
2. With special attention, namely the condition of arrears of principal installments that have not exceeded 3 months, there is an overdraft due to withdrawals but the period has not exceeded 15 working days.
3. Substandard, namely principal arrears that have exceeded 3 months but have not exceeded 6 months, there are overdrafts that have exceeded 15 working days but have not exceeded 30 working days.

2.7. Research Methods

The data used are secondary data from the 5C assessment on bank data. The data contain notes on the 5 C assessment as well as 3 credit collectability (current, special mention, and substandard) from Home Ownership Credit customers. The population in this study were all Home Ownership Credit customers in all banks in Indonesia. The sampling technique used was purposive random sampling. The sample size is 300 customers from customer data at three branches of the Bank. This research is a quantitative study using cluster analysis integrated with multiple discriminant analysis.

The steps in this research are as follows: (1) Testing the multivariate normal assumptions and the homogeneity of the variance matrix; (2) Classifying each data set into clusters and creating a dummy variable matrix; (3) Creating a model of multiple discriminant analysis with an integrated cluster approach with various linkages; (4) Calculating the value of classification accuracy, sensitivity and specificity; and (5) Comparing the hit results of the integrated cluster ratio and discriminant analysis.

3. Results and Discussion

3.1. Model Parameter Estimation

Destination the main part of Fisher's discriminant analysis is to separate the population, besides that it can also be used to classify. In this method, we assume that the covariance variant matrix is homogeneous. Where, $\sum 1 = \sum 2 = \dots = \sum g = \sum$, and $\bar{\mu}$ is the average vector of the population combination, then the between groups sums of the cross product can be described as in equation (16) below.

$$B_{\mu} = \sum_{i=1}^g (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})' \tag{16}$$

Where,

$$\bar{\mu} = \frac{1}{g} \sum_{i=1}^g \mu_i$$

Fisher groups an observation based on the score calculated from a linear combination as follows:

$$Y = a'X$$

$$E(Y) = a'E(X|\pi_i) = a'\mu_i$$

$$\text{Var}(Y) = a'Cov(X)a = a'\Sigma a$$

$$\bar{\mu}_Y = \frac{1}{g} \sum_{i=1}^g \mu_{iY} = \frac{1}{g} \sum_{i=1}^g a'\mu_i = a' \left(\frac{1}{g} \sum_{i=1}^g \mu_i \right)$$

The fisher linear discriminant function minimizes the function according to equation (17) below:

Total kuadrat jarak dari populasi ke rata – rata keseluruhan Y

$$\begin{aligned} & \frac{\text{Ragam Y}}{\sigma_Y^2} \\ &= \frac{\sum_{i=1}^g (\mu_{iY} - \bar{\mu}_Y)^2}{\sigma_Y^2} \\ &= \frac{\sum_{i=1}^g (a'\mu_i - a'\bar{\mu})^2}{a'\Sigma a} \\ &= \frac{a'(\sum_{i=1}^g (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})')a}{a'\Sigma a} \\ &= \frac{a'B_{\mu}a}{a'\Sigma a} \end{aligned} \tag{17}$$

In reality \sum and μ_i not available, so we can use the training data set which consists of sample data as much as n_i from the population, $i = 1, 2, \dots, g$. Thus, the sample mean vector is described as follows: π_i

$$\begin{aligned} \bar{x}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \\ \bar{x} &= \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_i \end{aligned}$$

Furthermore, the matrix for the sample is described in equation (18) as follows: B_{μ}

$$B = \sum_{i=1}^g (x_i - \bar{x})(x_i - \bar{x})' \tag{18}$$

In addition, the estimation \sum is based on the matrix in the sample group as in equation (19) as follows:

$$W = \sum_{i=1}^g (n_i - 1) S_i = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_i - \bar{x})(x_i - \bar{x})' \tag{19}$$

Thus, an estimate for \sum are as follows:

$$S_{\text{pooled}} = \frac{W}{(n_1 + n_1 + \dots + n_g - g)}$$

In order to separate groups maximally, the Y discriminant function is assumed by the criterion of maximizing the variability between groups. The coefficient which causes maximum, also maximizes.

$$a' \frac{a'Ba}{a'S_{\text{pooled}}a} \frac{a'Ba}{a'Wa}$$

Next, the vector of the coefficients which can impose the ratio in equation (20) follows: \hat{a}

$$\frac{\hat{a}'Ba}{\hat{a}'Wa} = \frac{\hat{a}'(\sum_{i=1}^g (x_i - \bar{x})(x_i - \bar{x})')\hat{a}}{\hat{a}'(\sum_{i=1}^g \sum_{j=1}^{n_i} (x_i - \bar{x})(x_i - \bar{x})')\hat{a}} \tag{20}$$

Fisher's discriminant classification rule is to place y into a population if the square of the distance from y to for $i \neq k$. $\pi_k \mu_{iY}$

Suppose that r is used as a discriminant function, then allocate x to the population if: π_k

$$\sum_{j=1}^s (x_j - \bar{x}_{kj})^2 \leq \sum_{j=1}^r (a'_j(x - \bar{x}_i))^2, i \neq k \tag{21}$$

With,

$$\begin{aligned} \sum_{j=1}^s (x_j - \bar{x}_{kj})^2 &= \sum_{j=1}^r (a'_j(x - \bar{x}_i))^2 \\ j &= \text{number of functions formed } (j = 1, 2, \dots, s) \\ k &= \text{number of groups} \end{aligned}$$

Thus, allocate x into the population where k returns the minimum value of $\pi_k \sum_{j=1}^s (x_j - \bar{x}_{kj})^2$.

3.2. Discriminant Analysis Assumption Testing

3.2.1. Assumption of Normally Distributed Multivariate Normal Data

Testing the assumption of multivariate normality is done by using the qq plot test and a graph is obtained as shown in Figure 2.

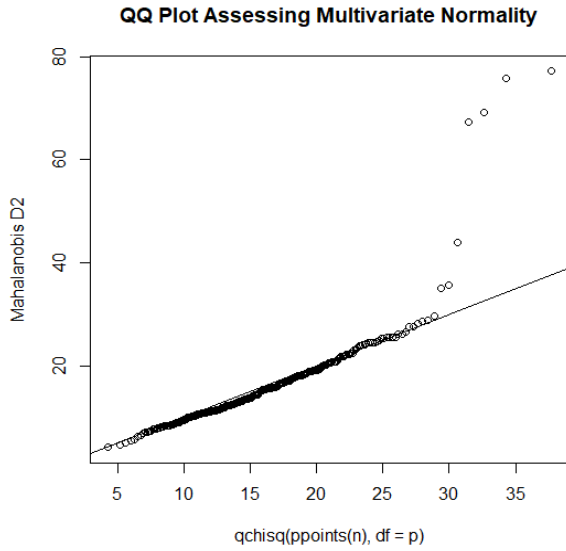


Figure 2. Mahalanobis Distance Plot and Chi-Square Quantile

Based on Figure 2, it can be seen that the plot between the Mahalanobis distance and the chi-square quantile has a straight-line pattern, but there are several data sets that spread very far from the straight line which indicates that there are outliers. If the value of the Mahalanobis distance is compared to the value $\chi^2_{p;0,05}$. It is found that the

value of the Mahalanobis distance from the data is less than $\chi^2_{p;0,05}$ as much as 98% of the total sample used. This value has exceeded the Mahalanobis distance limit value which must be less than $\chi^2_{p;0,05}$ amounted to 50%, which means that the data can be approached with multivariate normal distribution.

3.2.2. Assumption of Homogeneity of Variance Matrix

Testing the assumption of homogeneity of the variance matrix was carried out using the Box's M test and obtained a p-value of 0. Based on the p-value, it can be concluded that the variance matrix between clusters is not homogeneous. Tao Li et al. [25] stated that discriminant analysis often produces better classification results even though the assumptions of multivariate normality and homogeneity of the variance matrix are violated. In this study, the assumptions that were violated were not handled.

3.3. Cluster Analysis

3.3.1. Average Linkage Method

Dendrogram The average linkage method with Euclidean, Manhattan, and Mahalanobis distances is presented in Figure 3.

Comparison of Distance and Linkage in Integrated Cluster Analysis with Multiple Discriminant Analysis on Home Ownership Credit Bank in Indonesia

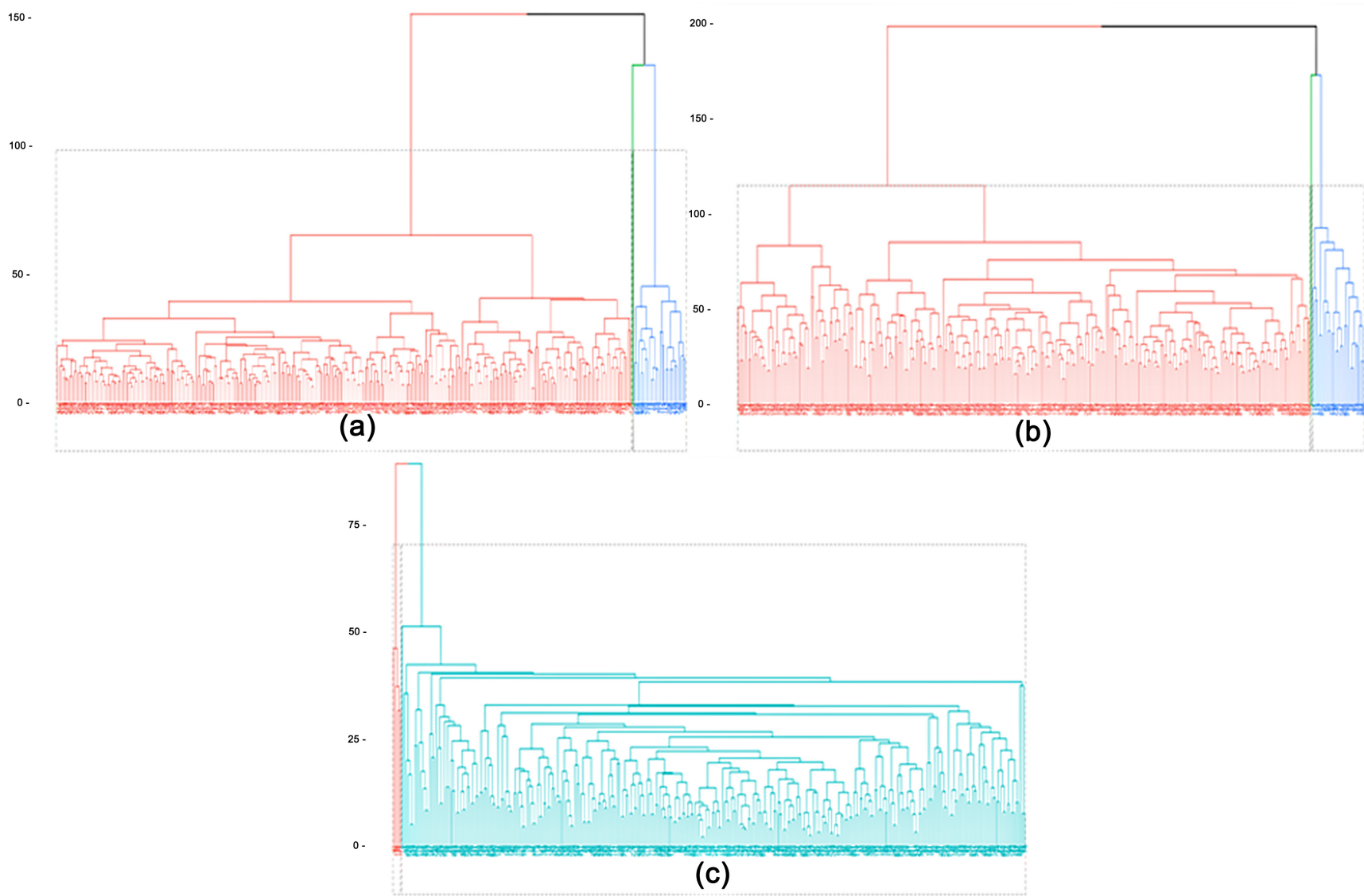


Figure 3. Dendrogram of the Average Linkage Method with Euclidean Distance (a), Manhattan Distance (b), and Mahalanobis Distance (c)

Based on Figure 3, it can be seen that there are no clusters that have the same stem length, so the cluster results are obtained with significantly different groupings. Cluster cutting is based on the length of the longest stem. Thus, the optimum number of clusters with the average linkage method with Euclidean and Manhattan distances is 3 clusters, while the Mahalanobis distance produces 2 clusters. The table of the number of members in each cluster is presented in Table 1.

Table 1. Many Members of Each - Each Cluster Average Linkage Method

Cluster	Many Cluster Members		
	Euclidean distance	Manhattan distance	Mahalanobis distance
1	274	274	274
2	1	1	26
3	25	25	-

As seen from Table 1, the number of members of the average linkage method with a distance of euclidean and manhattan in cluster 1 is 274 customers, cluster 2 is 1 customer, and cluster 3 is 25 customers. If the average indicator of each cluster is compared, it is found that the average linkage method using both euclidean and Manhattan distances produces the same number of cluster members. Meanwhile, the average linkage method with a distance of Mahalanobis has 274 customers in cluster 1 and 26 customers in cluster 2.

In the average linkage method with Euclidean and Manhattan distances, most of cluster 2 has the largest average indicator value compared to the others, so cluster 2 is a high cluster. While cluster 1 has the lowest average indicator value compared to the others, so cluster 1 is a low cluster, so cluster 3 is a medium cluster. High clusters are customers with the highest education level, the highest credit terms, the highest work experience, and the highest loan to value. The medium cluster is the customer with the longest residence time, the highest age, and the highest instalment income ratio, while the low cluster is the customer with the highest number of family dependents.

On the other hand, the average linkage method with a Mahalanobis distance. Most of cluster 2 has the largest average indicator value compared to the others, so that cluster 2 is a high cluster. While cluster 1 has the lowest average indicator value compared to the others, so cluster 1 is a low cluster. High clusters are customers with the longest residence time, highest age, highest credit period, highest instalment income ratio, highest work experience, and highest loan to value. In the low cluster are customers with the highest education level, and the highest number of family dependents.

3.3.2. Single Linkage Method

Dendogram The single linkage method with Euclidean, Manhattan, and Mahalanobis distances is presented in Figure 4.

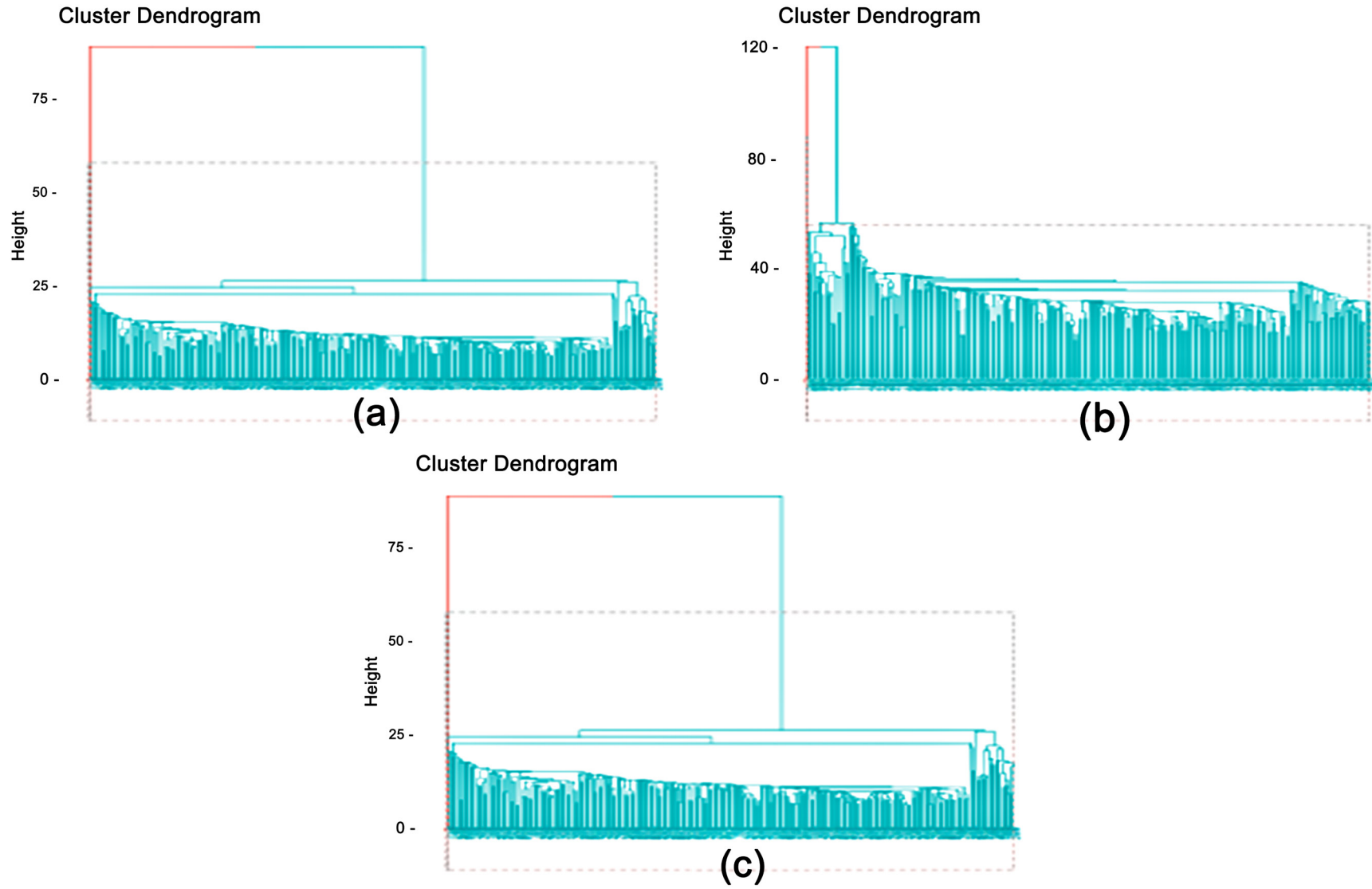


Figure 4. Single Linkage Method Dendrogram with Manhattan Distance, Euclidean Distance (a), Manhattan Distance (b), and Mahalanobis Distance (c)

Based on Figure 4, it can be seen that there are no clusters that have the same stem length, so the cluster results are obtained with significantly different groupings. Cluster cutting is based on the length of the longest stem. Thus, the optimum number of clusters with the single linkage method with a distance of euclidean, Manhattan, and Mahalanobis is 2 clusters. The table of the number of members in each cluster is presented in Table 2.

Table 2. Many members of each single linkage method cluster

Cluster	Many Cluster Members		
	Euclidean distance	Manhattan distance	Mahalanobis distance
1	299	299	298
2	1	1	2

It can be seen from Table 2. that there are 299 members of the single linkage method with Euclidean distance, and 299 in clusters 1 and 1 in cluster 2. Meanwhile, there are 298 members of the Mahalanobis distance in cluster 1 and 2 customers in cluster 2. If the average indicator for each cluster is compared, it is found that the single linkage method for Euclidean and Manhattan distances produces the same number of cluster members.

In the single linkage method with Euclidean, Manhattan, and Mahalanobis distances, most of cluster 1 has the largest average indicator value compared to the others, so that cluster 1 is a high cluster. While cluster 2 has the lowest average indicator value compared to the others, so cluster 2 is a low cluster. High clusters are customers with the longest residence time, the highest age, the highest credit period, the highest installment income ratio, the highest work experience, the highest loan to value, the highest education level, and the highest number of family dependents.

3.3.3. Complete Linkage Method

Dendogram The complete linkage method with

distances of Euclidean, Manhattan, and Mahalanobis is presented in Figure 5.

Based on Figure 5, it can be seen that there are no clusters that have the same stem length, so the cluster results are obtained with significantly different groupings. Cluster cutting is based on the length of the longest stem. Thus, the optimum number of clusters in the complete linkage method with a distance of Euclidean, Manhattan, and Mahalanobis is 2 clusters. The table of the number of members in each cluster is presented in Table 3.

Table 3. Many Members of Each Cluster Complete Linkage Method

Cluster	Many Cluster Members		
	Euclidean distance	Manhattan distance	Mahalanobis distance
1	299	299	298
2	1	1	2

As seen from Table 3. It can be seen that the number of members of the complete linkage method with a distance of euclidean and Manhattan in cluster 1 is 299 customers and cluster 2 is 1 customer. Meanwhile, the distance of Mahalanobis in cluster 1 is 298 customers and cluster 2 is 1 customer. If the average indicator for each cluster is compared, it is found that the complete linkage method for Euclidean and Manhattan distances produces the same number of cluster members.

In the complete linkage method with Euclidean, Manhattan, and Mahalanobis distances, most of cluster 1 has the largest average indicator value compared to the others, so that cluster 1 is a high cluster. While cluster 2 has the lowest average indicator value compared to the others, so cluster 2 is a low cluster. High clusters are customers with the longest residence time, the highest age, the highest credit period, the highest instalment income ratio, the highest work experience, the highest loan to value, the highest education level, and the highest number of family dependents.

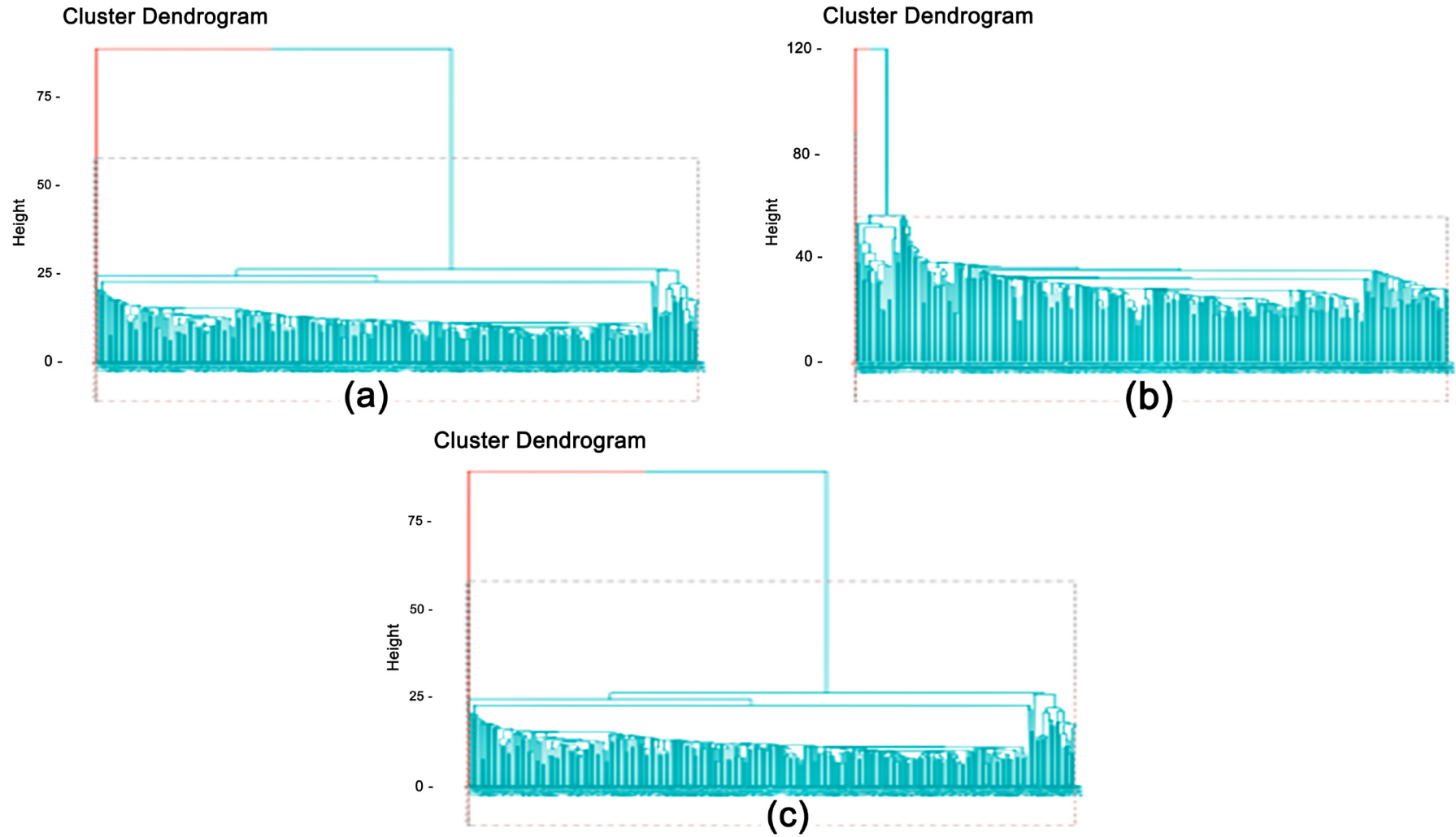


Figure 5. Complete linkage method dendrogram with Manhattan distance, Euclidean distance (a), Manhattan distance (b), and Mahalanobis distance (c)

3.4. Multiple Discriminant Analysis Model

The MDA model is used as a comparison model with the integrated cluster model with the MDA. Following are the results of classification accuracy, sensitivity, and specificity in the MDA model presented in Table 4.

Table 4. Results of Classification Accuracy, Sensitivity, and Specificity in the MDA Model

Model Goodness	Score
Classification accuracy	85.67%
Sensitivity	62.50%
Specificity	89.30%

Based on Table 4, the results of classification accuracy are 85.67%, which means that the correct model classifies as many as 61 customers out of 300 customers. A sensitivity of 62.50% means that customers classified as a current category can be classified correctly by the model as many as 171 out of 273 customers. Specificity of 89.30% means that customers belonging to the category of concern can be classified correctly by the model as many as 23 out of 26 customers.

3.5. Integrated Cluster Model on Multiple Discriminant Analysis Average Linkage Method

The average linkage method produces 3 clusters for euclidean and Manhattan distances and 2 clusters for Mahalanobis distances that optimally separate each data set. Following are the results of classification accuracy, sensitivity, and specificity in the integrated cluster model with the MDA average linkage method are presented in Table 5.

Table 5. Results of Classification Accuracy, Sensitivity, and Specificity in the Integrated Cluster Model with the MDA Average Linkage Method

Model Goodness	Euclidean distance	Manhattan distance	Mahalanobis distance
Classification accuracy	87.67%	87.67%	87.73%
Sensitivity	62.50%	62.50%	65.52%
Specificity	89.30%	89.30%	90.57%

Based on Table 5, it is found that the average linkage method with Euclidean, Manhattan, and Mahalanobis distances has a classification accuracy of 87.67%, which means that the model correctly classifies as many as 263 customers out of 300 customers. A sensitivity of 62.50% means that customers classified as a current category can be classified correctly by the model as many as 166 out of 266 customers. Specificity of 89.30% means that customers belonging to the category of concern can be classified correctly by the model as many as 26 out of 29 customers.

The average linkage method with a Mahalanobis

distance has a classification accuracy of 87.73%, which means that the model correctly classifies as many as 263 customers out of 300 customers. A sensitivity of 65.52% means that customers belonging to the current category can be classified correctly by the model as many as 174 out of 265 customers. Specificity of 90.57% means that customers belonging to the category of concern can be classified correctly by the model as many as 28 out of 31 customers.

3.6. Integrated Cluster Model on Multiple Discriminant Analysis Single Linkage Method

Method single linkage resulting in 2 clusters for Euclidean, Manhattan, and Mahalanobis distances. Following are the results of classification accuracy, sensitivity, and specificity in the integrated cluster model with the single linkage MDA method presented in Table 6.

Table 6. Results of Classification Accuracy, Sensitivity, and Specificity in the Integrated Cluster Model with the MDA Single Linkage Method

Model Goodness	Euclidean distance	Manhattan distance	Mahalanobis distance
Classification accuracy	86.00%	86.00%	86.00%
Sensitivity	64.00%	64.00%	64.00%
Specificity	89.63%	89.63%	89.63%

Based on Table 6, it is found that the single linkage method with a distance of Euclidean, Manhattan, and Mahalanobis has a classification accuracy of 86.00%, which means that the correct model classifies as many as 258 customers out of 300 customers. A sensitivity of 64.00% means that customers belonging to the current category can be classified correctly by the model as many as 191 out of 299 customers.

3.7. Integrated Cluster Model in Multiple Discriminant Analysis Complete Linkage Method

Table 7. Results of Classification Accuracy, Sensitivity, and Specificity in the Integrated Cluster Model with MDA Complete Linkage Method

Model Goodness	Euclidean distance	Manhattan distance	Mahalanobis distance
Classification accuracy	86.00%	86.00%	86.00%
Sensitivity	64.00%	64.00%	64.00%
Specificity	89.63%	89.63%	89.63%

The complete linkage method produces 2 clusters for the distances of Euclidean, Manhattan, and Mahalanobis. Following are the results of classification accuracy, sensitivity, and specificity in the integrated cluster model with the MDA complete linkage method presented in Table 7.

Based on Table 7, it is found that the complete linkage

method with a distance of Euclidean, Manhattan, and Mahalanobis has a classification accuracy of 86.00%, which means that the correct model classifies as many as 258 customers out of 300 customers. A sensitivity of 64.00% means that customers belonging to the current category can be classified correctly by the model as many as 191 out of 299 customers.

3.8. Comparison of the Accuracy Value of Multiple Discriminant Analysis and Integrated Clusters with Various Distance and Linkage Measures

Based on the results of the accuracy of the model that has been obtained from MDA and the integrated cluster on MDA, the summary results of the classification accuracy, sensitivity, and specificity values for various distances and linkage sizes are as in Table 8.

Based on Table 8, the sensitivity value of the average linkage distance Mahalanobis method is the most accurate measure of distance in detecting customers who have current credit with an accuracy value of 65.52%. As for the specificity of the integrated cluster on MDA, the Mahalanobis distance measure is the most accurate method in detecting customers who have credit in attention with an accuracy of 90.57%. The use of an integrated cluster in MDA, the Mahalanobis distance measure, is more recommended in classifying credit at the bank, this is because all results of the accuracy of the integrated cluster classification on MDA, the Mahalanobis

distance measure is better than MDA with euclidean and manhattan distance measures.

3.9. Best Model

Accuracy of classification for different measures of distance and linkage has similar results, differing only by a few decimal places. So that the integrated cluster on MDA with the Mahalanobis distance is the best method in classifying Home Ownership Credit Bank customers based on the 5C variable assessment because it has the highest value of classification accuracy, sensitivity, and specificity. In an integrated cluster on MDA with a distance of Mahalanobis, there are 2 clusters with cluster 1 totalling 274 customers and cluster 2 with 26 customers. If the average indicator for each cluster is compared, it is found that most of cluster 2 has the largest average indicator value compared to the others, so cluster 2 is a high cluster. While cluster 1 has the lowest average indicator value compared to the others.

In this study, the determination of customers in the current credit category, with special attention, and substandard credit is based on the discriminant score obtained by substituting the value of the data set into the formed discriminant function. After obtaining the discriminant score, it is compared with the centroid group value for each model. If the discriminant score is near the value of the centroid group, then the customer is classified as current collectability and vice versa bad credit.

Table 8. Calculation Results of Classification Accuracy, Sensitivity, and Specificity at Various Distance and Linkage Measures

Linkage	Distance	Classification Agreement	Sensitivity	Specificity
Average Linkage	Euclidean	87.67%	62.50%	89.30%
	Manhattan	87.67%	62.50%	89.30%
	Mahalanobis	87.73%	65.52%	90.57%
Single Linkage	Euclidean	86.00%	64.00%	89.63%
	Manhattan	86.00%	64.00%	89.63%
	Mahalanobis	86.00%	64.00%	89.63%
Complete Linkage	Euclidean	86.00%	64.00%	89.63%
	Manhattan	86.00%	64.00%	89.63%
	Mahalanobis	86.00%	64.00%	89.63%

Model 1 of each cluster can be seen in Equations (22) and equation (23). For model 1 in the low cluster as in equation (22) below.

$$\begin{aligned}
 y_i = & 0.244x_{11.1i} + 0.418x_{11.2i} - 0.043x_{12i} + 0.076x_{21i} - \\
 & 0.352x_{22i} - 0.011x_{23i} - 0.668x_{24.1i} - 0.283x_{24.2i} - \\
 & 0.302x_{25.1i} + 0.363x_{25.2i} - 0.235x_{31i} - 0.453x_{32.1i} - \\
 & 0.317x_{32.2i} - 0.140x_{32.3i} - 1.550x_{33i} - 0.220x_{34i} + \quad (22) \\
 & 0.090x_{35.1i} + 0.336x_{35.2i} - 0.065x_{35.3i} - 0.039x_{35.4i} - \\
 & 0.788x_{36i} + 0.327x_{37i} + 0.216x_{41.1i} + 0.341x_{41.2i} + \\
 & 0.272x_{5i}
 \end{aligned}$$

Based on the equation (16), it can be interpreted that the coefficient of X11 (Guarantee Document), X11.2 (Guarantee Document), X21 (City Size), X25.2 (Marital Status), X35.1 (Job), X35.2 (Job), X37 (Total family dependents), X41.1 (Savings Ownership), X41.2 (Savings Ownership), X5 (Loan to Value) are positive, meaning the higher the value of X11.1 (Guarantee Document), X11.2 (Guarantee Document), X21 (City Size), X25.2 (Marital Status), X35.1 (Occupation), X35.2 (Occupation), X37 (Number of Family Dependents), X41.1 (Savings Ownership), X41.2 (Ownership Savings), X5 (Loan to Value), it will increase the likelihood that customers in low clusters will have current credit collectability. On the other hand, X12 (Length of Residence), X22 (Education), X23 (Age), X24.1 (Collectability Status), X24.2 (Collectability Status), X25.1 (Marital Status), X31 (Joint Income), X32.1 (Business Entity Form), X32.2 (Business Entity Form), X32.3 (Business Entity Form), X33 (Credit Period), X34 (Instalment Income Ratio), X35.3 (Occupation), X35.4 (Occupation), and X36 (Work Experience) have negative coefficients, so if this value increases, it will increase the likelihood of customers having substandard credit collectability. The indicator that most influences credit collectability in the low cluster is Work Experience which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in the low cluster is Work Experience which has the largest discriminant coefficient value.

Model 1 in the high cluster as in equation (23) below.

$$\begin{aligned}
 y_i = & 0.809x_{11.1i} + 1.533x_{11.2i} - 0.403x_{12i} + 0.047x_{21i} - \\
 & 0.917x_{22i} + 0.146x_{23i} - 3.406x_{24.1i} + 1.593x_{24.2i} - \\
 & 0.525x_{25.1i} + 2.721x_{25.2i} - 0.022x_{31i} - 3.340x_{32.1i} - \\
 & 2.607x_{32.2i} - 2.014x_{32.3i} - 5.552x_{33i} - 0.006x_{34i} - \quad (23) \\
 & 0.018x_{35.1i} + 1.012x_{35.2i} - 0.668x_{35.3i} - 0.492x_{35.4i} - \\
 & 0.795x_{36i} - 0.280x_{37i} + 0.731x_{41.1i} + 0.705x_{41.2i} + \\
 & 0.820x_{5i}
 \end{aligned}$$

Based on the equation (17), it can be interpreted that the coefficient of X11.1 (Guarantee Document), X11.2 (Guarantee Document), X21 (City Size), X23 (Age), X24.2 (Collectability Status), X25.2 (Marital Status), X35.2 (Occupation), X41.1 (Savings Ownership), X41.2 (Savings Ownership), X5 (Loan to Value) are positive, meaning the higher the value of X11.1 (Collateral Document), X11.2 (Document Guarantee), X21 (City Size), X23 (Age), X24.2 (Collectability Status), X25.2 (Marital Status), X35.2 (Occupation), X41.1 (Savings Ownership), X41.2 (Ownership Savings), X5 (Loan to Value), it will increase the likelihood that customers in high clusters will have current credit collectability. Conversely, X12 (Old in Residence), X22 (Education), X24.1 (Collectability Status), X25.1 (Marital Status), X31 (Joint Income), X32.1 (Business Entity), X32.2 (Business Entity Form), X32.3 (Business Entity), X33 (Credit Period), X34 (Instalment Income Ratio), X35.1 (Job), X35.3 (Job), X35.4 (Job), X36 (Work Experience), and X37 (Number of Family Dependents) has a negative coefficient, so that if this value increases, it will increase the likelihood of the customer having substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value.

Model 2 each cluster can be seen in equation (24), and equation (25). For model 2 in the low cluster as in equation (24) below.

$$\begin{aligned}
 y_i = & -0.143x_{11.1i} - 0.026x_{11.2i} - 0.164x_{12i} + 0.175x_{21i} + \\
 & 0.319x_{22i} + 0.031x_{23i} + 0.394x_{24.1i} + 0.128x_{24.2i} + \\
 & 0.018x_{25.1i} - 0.489x_{25.2i} + 0.097x_{31i} + 0.173x_{32.1i} + \\
 & 0.17x_{32.2i} + 0.275x_{32.3i} + 0.667x_{33i} - 0.057x_{34i} - \quad (24) \\
 & 0.476x_{35.1i} - 0.184x_{35.2i} - 0.128x_{35.3i} + 0.108x_{35.4i} + \\
 & 0.572x_{36i} + 0.394x_{37i} - 0.130x_{41.1i} + 0.186x_{41.2i} - \\
 & 0.086x_{5i}
 \end{aligned}$$

Based on equation (2), it will increase the likelihood that customers in lower clusters will have current credit collectability. On the other hand, X11.1 (Guarantee Document), X11.2 (Guarantee Document), X12 (Length of Residence), X25.2 (Marital Status), X34 (Instalment Income Ratio), X35.1 (Employment), X35.2 (Occupation), X35.3 (Occupation), X41.1 (Savings Ownership), and X5 (Loan to Value) have negative coefficients, so that if this value increases it will increase the likelihood of customers having substandard credit

collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value. X12 (Length of Residence), X25.2 (Marital Status), X34 (Instalment Income Ratio), X35.1 (Employment), X35.2 (Occupation), X35.3 (Occupation), X41.1 (Savings Ownership), and X5 (Loan to Value) has a negative coefficient, so that if this value increases, it will increase the likelihood of the customer having substandard credit collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value. X12 (Old Residence), X25.2 (Marital Status), X34 (Instalment Income Ratio), X35.1 (Employment), X35.2 (Occupation), X35.3 (Occupation), X41.1 (Savings Ownership), and X5 (Loan to Value) has a negative coefficient, so that if this value increases, it will increase the likelihood of the customer having substandard credit collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value. So that if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in the low cluster is the Credit Period which has the largest discriminant coefficient value.

Model 2 in the high cluster as in equation (25) below.

$$\begin{aligned}
 y_i = & -0.532x_{11.1i} - 0.86x_{11.2i} + 0.139x_{12i} + 0.125x_{21i} + \\
 & 0.676x_{22i} - 0.082x_{23i} + 2.265x_{24.1i} + 1.417x_{24.2i} + \\
 & 0.226x_{25.1i} - 2.010x_{25.2i} - 0.028x_{31i} + 2.168x_{32.1i} + \\
 & 1.733x_{32.2i} + 1.487x_{32.3i} + 3.494x_{33i} - 0.138x_{34i} - \\
 & 0.274x_{35.1i} - 0.655x_{35.2i} + 0.333x_{35.3i} + 0.389x_{35.4i} + \\
 & 0.565x_{36i} + 0.615x_{37i} - 0.484x_{41.1i} - 0.930x_{41.2i} - \\
 & 0.487x_{5i}
 \end{aligned} \quad (25)$$

Based on equation (4.24), it can be interpreted that the coefficient of and X37 (Number of Family Dependents), will increase the likelihood that customers in high clusters will have current credit collectability. On the other hand, X11.1 (Guarantee Document), X11.2 (Guarantee Document), X23 (Age), X25.2 (Marital Status), X31 (Joint Income), X34 (Instalment Income Ratio), X35.1 (Job), X35.2 (Occupation), X41.1 (Savings Ownership), X5 (Loan to Value), and X41.2 (Savings Ownership) have negative coefficients, so if this value increases, it will increase the likelihood of customers having substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value. X31 (Joint Income), X34 (Instalment Income Ratio), X35.1 (Occupation), X35.2 (Occupation), X41.1 (Savings

Ownership), X5 (Loan to Value), and X41.2 (Savings Ownership) has a negative coefficient, so if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value. X31 (Joint Income), X34 (Instalment Income Ratio), X35.1 (Occupation), X35.2 (Occupation), X41.1 (Savings Ownership), X5 (Loan to Value), and X41.2 (Savings Ownership) has a negative coefficient, so if this value increases, it will increase the likelihood that the customer will have substandard credit collectability. The indicator that most influences credit collectability in high clusters is the Credit Period which has the largest discriminant coefficient value.

4. Conclusions and Suggestions

4.1. Conclusions

The conclusion that can be given is based on the results of the analysis, namely the best method for classifying HOME OWNERSHIP CREDIT Bank X customers based on the 5C variable assessment is an integrated cluster analysis with MDA based on the Mahalanobis distance with 2 clusters, namely the high cluster and the low cluster. High clusters are customers with a long time in residence (X12) oldest, age (X23) the highest, the credit period (X33) the highest, the instalment income ratio (X34) highest, work experience (X36) the highest, and loan to value (X5) the highest. In the low cluster are customers with a long education (X22) the highest, and the number of family dependents (X37) the highest. The results of classification accuracy, sensitivity, and specificity in the euclidean, Manhattan, and Mahalanobis distance-based integrated cluster were better than MDA.

4.2. Suggestions

Some suggestions that can be given based on the results of the integrated cluster in this discriminant analysis, namely (1) Further research can use simulation data to obtain more accurate comparison results because, in this study, the values of classification accuracy, sensitivity, and specificity between models were not significantly different; (2) Further research may use quadratic discriminant analysis or logistic regression to overcome the assumptions that are violated in this study; (3) Further research can examine the credit scoring model with 5 categories of credit collectability as the response variable.

Acknowledgements

We are very grateful to experts for their appropriate and constructive suggestions to improve this paper.

REFERENCES

- [1] Nugroho, S. 2008. *Applied Multivariate Statistics*. Bengkulu: UNIB Press.
- [2] Solimun, Fernandes, A.A.R., dan Nurjannah. 2017. *Multivariate Statistical Method Structural Equation Modeling (SEM) WarpPLS Approach*. Universitas Brawijaya Press.
- [3] Tatham, R.L., Hair, J.F, Anderson, R.E., dan Black, W.C., 1998, *Multivariate Data Analysis*. New Jersey: Prentice-Hall.
- [4] Johnson, N. dan Winchern, D. 2002. *Applied Multivariate Statistical Analysis, fifth edition*. USA: Prentice-Hall Englewood Cliffs, N.J.
- [5] Afriana, W., dan Kuswanto, A. 2012. *Factors Affecting Collectability of Onion Farmers' Credit Payments at Bri Bank Brebes Branch*.
- [6] Fernandes, A. A. R., Solimun, Nurjannah, dan Hutahayan, B. 2020. Comparison of the use of Linkage in Integrated Cluster with Discriminal Analysis Approach. *International Journal of Advanced Science and Technology*. 29(3), 5654-5668.
- [7] Rofitanur, N. 2020. *Implementation of Hybrid Mutual Clustering Integration and Discriminant Analysis on Credit Collectability of Bank X Malang City*. Unpublished.
- [8] Siswadi dan B. Suharjo. 1998. *Multiple Variable Data Exploration Analysis*. Final Project, Unpublished. Bogor: Jurusan Matematika Fakultas MIPA IPB, Bogor.
- [9] Johnson, R. A. dan Wichern, D. W. 1992. *Applied Multivariate. Analysis, Third Edition*, New Jersey: Prentice Hall Inc.
- [10] Hair, J. F., Anderson, R. E., Tatham, R. L., dan Black, W. C. 2006. *Multivariate Data Analysis. Fifth edition*. Jakarta: Gramedia. Pustaka Utama.
- [11] Prasetyo, E. 2012. *Data Mining-Concepts and Applications Using MATLAB*. Yogyakarta: Andi Offset.
- [12] Seber, A. F. 1983. *Multivariate Observations*. New York: Auckland.
- [13] Rancher, A. C. 2002, *Methods of Multivariate Analysis Second Edition*. Canada: John Wiley & Sons.
- [14] Budiarto, E. 2001. *Medical Research Methodology. EGC*. Jakarta
- [15] Astiko dan Sunardi. 1996. *Introduction to Credit Management. First Edition*. Yogyakarta: ANDI.
- [16] Kasmir. 2013. *Banking Fundamentals (Revised Edition)*. Jakarta: Rajawali Pers.
- [17] Asiyah, B. N. 2015. *Islamic bank financing management*. Yogyakarta: Kalimedia.
- [18] Suharno, 2003. *Credit Analysis: Equipped with Case Examples*, Jakarta: Djambatan.
- [19] Supranto, 2004. *Multivariate Analysis of Meaning and Interpretation*, Jakarta: PT. Rineka Cipta.
- [20] Asiska, N., Satyahadewi, N. and Perdana, H., 2019. Optimum Cluster Search For Single Linkage, Complete Linkage And Average Linkage. *BIMASTER*, 8(3).
- [21] Widodo, E., Mashita, S.N. and Prasetyowati, Y.G., 2020. Comparison of Average Linkage, Complete Linkage, and Ward'S Methods in the Grouping of Regencies/Cities in Central Java Province Based on Human Development Index Indicators. *Exact Factor*, 13(2), pp.81-87.
- [22] Nishom, M., 2019. Comparison of the Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance on the Chi-Square-based K-Means Clustering Algorithm. *J. Inform*, 4(01).
- [23] Gueorguieva, N., Valova, I. and Georgiev, G., 2017. M&MFCM: fuzzy c-means clustering with mahalanobis and minkowski distance metrics. *Procedia computer science*, 114, pp.224-233.
- [24] Jati, R.P. and Prasetyo, A., 2018. Analysis of the Bankruptcy Potential of Islamic Commercial Banks in Indonesia in the 2012-2016 Period Using the Multiple Discriminant Analysis Method. *Journal of Theoretical and Applied Islamic Economics*, 5(11), pp. 941-958.
- [25] Li, T., Zhu, S., & Ogihara, M. 2006. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, 10(4), 453-472.
- [26] Fernandes, A. A. R., Solimun, F. U., Aryandani, A., Chairunissa, A., Alifa, A., Krisnawati, E., ... & Rasyidah12, F. L. N. 2021. Comparison Of Cluster Validity Index Using Integrated Cluster Analysis With Structural Equation Modeling the War-PLS Approach. *Journal of Theoretical and Applied Information Technology*, 99(18).