p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
11(2) 2023: 31-43

K-MEDOIDS ALGORITHM CLUSTERING WITH PRINCIPAL COMPONENT ANALYSIS (PCA) (CASE STUDY: DISTRICTS/CITIES ON THE BORNEO ISLAND BASED ON POVERTY INDICATORS)

Muhammad Yafi[1*], Rito Goejantoro[2], Andrea Tri Rian Dani[3]
[1,2,3] Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia

***e-mail**: yyafimuhammad@gmail.com

Article Info:
Received: July 26, 2023
Accepted: November 21, 2023
Available Online: November 30, 2023

Keywords:
K-Medoids; PCA; Poverty; Silhouette Coefficient

Abstract: Cluster analysis is a technique in data mining that aims to group data (object) based on the information in the data. This research is used a non-hierarchical grouping named K-Medoids algorithm to group districts/cities in Borneo island based on poverty indicators and Principal Component Analysis (PCA) method to reduce research variable. So the groups can be obtained that contain districts/cities with the same poverty problems. This research is also do a cluster validity test to see how many cluster there are has the best grouping result using Silhouette Coefficient (SC) method. Based on the results of the analysis there is 3 optimal Principal Component (PC) were obtained with eigen value criteria of greater than or equal to 1. Furthermore, districts/cities on Borneo island were grouped based on the PC that formed and obtained 2 optimal clusters with an SC value of 0.61. The K-Medoids algorithm obtain 2 cluster, cluster 1 consisting of 49 districts/cities and cluster 2 consisting of 7 cities.

1. INTRODUCTION

Data Mining is the process of collecting new information by looking for certain patterns or rules from a large amount of data (Big Data) so that new information or knowledge can be obtained. Data Mining is often referred to as Knowledge Discovery in Database (KDD). KDD is an activity that includes collecting and using data to find the regularity of a pattern or relationship in large data (Big Data) [1].

One of the techniques known in data mining is cluster analysis. Cluster analysis is a method that aims to group data (objects) based on data information. Cluster analysis is divided into two, namely hierarchical clustering and non-hierarchical clustering. The principle of hierarchical clustering is that starting from a single data grouped into one group, two or more small groups that can merge into a large group and so on [2], [4].

The advantage of non-hierarchical clustering is that it is faster in the computational process when grouping a large number of objects or observations and has a better level of time efficiency than hierarchical clustering because non-hierarchical clustering first determines the number of clusters at the beginning. Then the objects will be distributed to the clusters that have been determined based on the similarity or similarity of their characteristics [5], [15].

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

As information technology develops, the diversity of information is described using variables. The more variables used, the greater the possibility of complexity problems in the clustering process. Thus, the author is interested in reducing the variables without reducing the amount of information from the variables. This can be done using the Principal Component Analysis (PCA) method. PCA is a statistical method that linearly deforms a variable into a set of smaller, uncorrelated variables that can represent the information of the original variable [8].

The principle of PCA is to find eigenvalues and eigenvectors that are used to simplify or reduce dimensions. This is done by eliminating the correlation between variables through the process of transforming variables into a new variable that is not correlated at all or what is commonly called the principal component (PC). The magnitude of eigenvectors and eigenvalues has the most important role because eigenvectors and eigenvalues that are very small will be discarded so that data can be reduced without losing much important information [7].

Cluster analysis has been widely used in various fields such as poverty, health, education, employment and so on. Therefore, the purpose of this research is to see the problems that occur in districts/cities on the island of Borneo based on poverty indicators which include health, education and decent living standards [4].

The measure of poverty is not only based on the fulfillment of food needs and low income levels, but also in terms of health, education and fair treatment before the law and so on. So that poverty has a large number of variables or data and varies [3]. Therefore, the k-medoids algorithm is suitable for grouping districts / cities on the island of Borneo based on poverty indicators because it is not sensitive to outliers and is good for large amounts of data.

2. LITERATURE REVIEW

2.1. Principal Component Analysis

The literature review provides a brief and straightforward overview of theories, statements, or anything related to and supports the problem posed either from formal literature (books, journals, written scientific reports) or real conditions that can be proven/observed.

Principal Component Analysis (PCA) is a multivariate analysis technique in statistics that serves to reduce dimension and detect multicollinearity [14]. The principle of PCA is to transform the original variables that are likely to be correlated between dimensions into new uncorrelated variables. PCA gives good results when applied to correlated variables. Therefore, PCA is a multivariable data selection technique that transforms an original data matrix into a smaller set of homogeneous combinations (reduction) but absorbs a large amount of variance from the original data [9]. In general, the variable reduction steps using the PCA method as follows:

1. Data standardization

   The standardization used Z_{score} is standardization. The formula for calculating standardization presented in Equation (1) as follows:

$$Z_{ap} = \frac{x_{ap} - \bar{x}_p}{S_p} \quad (1)$$

   with the average for each variable presented in Equation (2) as follows:

$$\bar{x}_p = \frac{1}{n}\sum_{a=1}^{n} x_{ap}, a = 1, 2, 3, ..., n \text{ dan } p = 1, 2, 3, ..., j \quad (2)$$

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

Deviation standard calculation presented in Equation (3) as follows:

$$S_p = \sqrt{\frac{1}{n-1}\sum_{a=1}^{n}(X_{ap} - \overline{X}_p)^2} \tag{3}$$

2. Calculating the variance-covariance matrix
   The formula calculates the variance-covariance matrix as follows:

$$\sum{}_z = \begin{bmatrix} S_{11}{}^2 & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22}{}^2 & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{p10}{}^2 \end{bmatrix}$$

   The variance calculation formula is presented in Equation (4) as follows:

$$S_p^2 = \frac{1}{n-1}\sum_{a=1}^{n}(x_{ap} - \overline{x}_p)^2 \tag{4}$$

   The covariance calculation formula is presented in Equation (5) as follows:

$$S_p^2 = \frac{1}{n-1}\sum_{a=1}^{n}(x_{ap} - \overline{x}_p)^2 \tag{5}$$

3. Calculating the correlation matrix
   The formula for calculating the correlation matrix is presented in Equation (6) as follows:

$$\mathbf{R_z} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

   with

$$\rho_{x_p x_{p+1}} = \frac{S_{x_p x_{p+1}}}{S_{x_p} S_{x_{p+1}}} \tag{6}$$

4. alculating eigenvalues and eigenvectors
   Calculate the eigenvalue (λ) and eigenvector ($\vec{v}$) of the correlation matrix (**R**z) with the formula shown presented in Equation (7) as follows:

$$\left|\lambda\mathbf{I} - \mathbf{R_z}\right| = 0 \tag{7}$$

5. In order to be an eigenvalue, there must be one non-zero solution of $\left|\lambda\mathbf{I} - \mathbf{R_z}\right| = 0$. This is fulfilled if and only if it fulfills the formula in Equation (8) as follows:

$$\det\left|\lambda \mathrm{I} - \mathbf{R_z}\right| = 0 \tag{8}$$

   Where $\mathbf{I}$ is the identity matrix of the eigenvector ($\vec{v}$).
6. Determining the number of Principal Component (PC) that may be formed by looking at the eigenvalue criteria that are greater than or equal to 1. Eigenvalues that are less than

one are considered to have less contribution. This method can only be applied to the correlation matrix.

7. Form a correlation matrix component to show the correlation of variables to the component scores formed using the formula in Equation (9) as follows:

$$r_{x_p,PC_t} = \vec{v}_{at}\sqrt{\lambda_t} \tag{9}$$

8. Form new reduced variables.
9. Form new variables based on linear combination with the formula in Equation (10) as follows:

$$\mathrm{PC}_{a,t} = \sum_{p=10} \vec{v}_p z_p \tag{10}$$

with $\vec{v}$ is the eigenvector.

2.2. Cluster Analysis

Cluster analysis is a method that aims to group data (objects) based on data information. This information can explain the characteristics of the object. The purpose of cluster analysis is to group objects into one cluster, so that they have similarities with each other (Homogeneity) in one cluster. But different from objects that are in other clusters (Heterogeneity) [2]. Cluster analysis is divided into two, namely hierarchical clustering and non-hierarchical clustering [13].

2.3. Similarity Measurement

There are several distance measurements that can be used, one of which is Euclidean Distance. Euclidean Distance is an efficient distance to use on numerical data. Euclidean Distance is also considered to work efficiently to calculate similarity in grouping objects based on the similarity or dissimilarity of these objects [10]. The formula for calculating Euclid distance is as follows:

$$d(x_u, y_v) = \sqrt{\sum_{u_1}^{n} \left(x_u - y_v\right)^2} \quad u, v = 1, 2, \ldots, n \tag{11}$$

2.4. K-Medoids Algorithm

K-Medoids or Partitioning Around Medoids (PAM) is a partition-based method of grouping objects into clusters. The objects selected to represent a cluster are called medoids. K-Medoids is a development of the K-Means algorithm that is sensitive to outliers. Besides being able to handle outliers, K-Medoids is also a flexible algorithm because it can work on almost every type of matrix data and able to group a large number of objects with fast computation time [11]. The stages of the K-Medoids algorithm are detailed as follows:

1. Randomly select k objects as representative objects (medoids) o_{mj} (medoids).
2. Calculate the Euclidean distance for each object against each medoids as shown in Equation (12) as follows:

$$d\left(x_{ap}, o_{mp}\right) = \sqrt{\left(x_{a1} - o_{m1}\right)^2 + \left(x_{a2} - o_{m2}\right)^2 + \ldots + \left(x_{aj} - o_{mj}\right)^2} \tag{12}$$

with $d(x_{ap}, o_{mp})$ is the distance of the data-*a* in the variable *p* against *m* medoids, where *a = 1,2,...,n* and *p = 1,2,...,j* also *m = 1,2,...,k*.

3. Assign each object to the cluster corresponding to the closest medoids and calculate the objective function which is the sum of the dissimilarities of all objects to the closest medoids based on the distance between the objects to each medoids that is the minimum.
4. Randomly select unrepresentative objects o_{hj} (non-medoids).
5. Calculate the Euclidean distance for each object against each non-medoids as shown in Equation (13) as follows:

$$d\left(x_{ap}, o_{hp}\right) = \sqrt{\left(x_{a1} - o_{h1}\right)^2 + \left(x_{a2} - o_{h2}\right)^2 + \ldots + \left(x_{aj} - o_{hj}\right)^2} \tag{13}$$

with $d(x_{ap} - o_{hp})$ is the distance of the data-*a* in the variable *p* against *h* medoids, where *a = 1,2,...,n* and *p = 1,2,...,j* also *m = 1,2,...,n-k*.

6. Assign each object to the cluster corresponding to the closest non-medoids and calculate the objective function which is the sum of the dissimilarities of all objects to the closest non-medoids based on the distance between the objects to each medoids that is the minimum.
7. Calculating the difference of the objective function by subtracting the non-medoids objective function from the medoids objective function.
8. Repeating steps (4-7) until there is no more change in the representative object and the analysis is complete when there is no change in the representative object.

2.5. Silhouette Coefficient

One of the evaluation methods that can be used to see the quality and strength of clusters is the silhouette coefficient method. The results of the SC value calculation can be varied between -1 to 1. SC value of 1 or close to 1 means that the object is already in the right cluster. If the SC value is 0, the object is between two clusters so that the object is not known whether it should be included in cluster A or cluster B. Meanwhile, if the SC is -1, it means that there are many objects that are not in the right cluster [12].

3. METHODOLOGY

3.1. Data and Source Data

The data used in this study is data from 56 districts/cities on the Borneo island based on poverty indicators. The data is taken from the official websites of BPS and the ministry of health. Variable data for this study are detailed in Table 1 as follows:

Table 1. Research Variable

Variable	Unit	Notation	Dimension	Source
Mean Years of Schooling	Year	X_1	Education	Central Statistic Agency of East Borneo
Expected Years of Schooling	Year	X_2	Education	Central Statistic Agency of East Borneo

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

Variable	Unit	Notation	Dimension	Source
Purchasing Power Parity	Rupiah (Rp)	x_3	Decent Standard Of Living	Central Statistic Agency of East Borneo
Percentage of Poor Population	Percentage (%)	x_4	Decent Standard Of Living	Central Statistic Agency of East Borneo
Percentage of Households With Access To Proper Sanitation	Percentage (%)	x_5	Decent Standard Of Living	Central Statistic Agency of East Borneo
Percentage of Households With Access To Proper Drinking/Clean Water	Percentage (%)	x_6	Decent Standard Of Living	Central Statistic Agency of East Borneo
Life Expectancy	Year	x_7	Health	Central Statistic Agency of East Borneo
Population Density	People/Km	x_8	Decent Standard Of Living	Central Statistic Agency of East Borneo
Unemployment Rate	Percentage (%)	x_9	Decent Standard Of Living	Central Statistic Agency of East Borneo
Prevalence of Stunting In Children	Percentage (%)	x_{10}	Health	Indonesia Ministry of Health

3.1. Steps of Analysis

The analysis steps in this study are as follows:

1. Conducted descriptive statistics to see the structure or pattern of research data.
2. Reducing variables using the Principal Component Analysis (PCA) method.
3. Clustering the reduced data with PCA using k-medoids algorithm.
4. Conducted cluster validity tests for each k tested using the Silhouette Coefficient (SC) method.
5. Profilization and interpretation of best clustering results.

4. RESULT AND DISCUSSION

4.1. Descriptive Statistics

In this chapter conducted a descriptive statistics process to see the characteristics of the research data. The results of descriptive statistics for poverty indicators in 56 districts/cities on the Borneo island are shown in Table 2.

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

Table 2. Descriptive Statistics

Variable of Data	Amount of Data	Minimum	Maximum	Average
x_1	56	6.02	11.53	8.40
x_2	56	11.17	15.09	12.78
x_3	56	7.06	16.76	10.75
x_4	56	2.89	12.01	6.29
x_5	56	49.23	97.51	79.71
x_6	56	48.85	99.92	77.15
x_7	56	64.10	74.76	70.76
x_8	56	1.70	9198.89	407.35
x_9	56	2.30	12.38	4.95
x_{10}	56	14.20	35.90	24.37

4.2. Descriptive Statistics

Principal Component Analysis (PCA) is a powerful technique for extracting structure from a data set with many dimensions. PCA can reduce the large dimensions of the observed data into smaller dimensions without losing significant information in describing the entire data. The steps in the PCA method are as follows:

1. Data Standardization
   In this research, Z_{score} standardization is an important thing to do. Because the data used in this study have different ranges. Therefore, by using Z_{score} standardization the data will be in the same range where the average is 0 and the standard deviation is 1. The results of data standardization using are shown in Table 3 as follows:

Table 3. Data Standardization Result

Districts/cities	x_1	x_2	...	x_{10}
Paser	0.31	0.54	...	-0.15
Kutai Barat	0.23	0.28	...	-1.78
Kutai Kartanegara	0.67	0.94	...	0.42
Kutai Timur	0.83	0.14	...	0.65
⋮	⋮	⋮	...	⋮
Tarakan	1.28	1.44	...	0.31

2. Calculating the correlation matrix
   The correlation matrix will be the same size as the variance-covariance matrix of the data of 56 districts/cities in Borneo Island based on standardized poverty indicators. So with the help of R studio software, the correlation matrix that will be formed is as follows:

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

$$\mathbf{R_z} = \begin{pmatrix} 1 & 0,77 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & -0,03 \\ 0,77 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & -0,19 \\ 0,70 & 0,60 & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & -0,07 \\ -0,27 & -0,27 & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & -0,02 \\ 0,49 & 0,48 & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots & -0,24 \\ 0,37 & 0,54 & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & \cdots & -0,10 \\ 0,40 & 0,35 & \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & \cdots & -0,21 \\ 0,38 & 0,43 & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \cdots & -0,13 \\ 0,48 & 0,61 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & -0,35 \\ -0,03 & -0,13 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 \end{pmatrix}_{(10\text{x}10)}$$

3. Calculating eigenvalues and eigenvectors

   Calculate the eigenvalue (λ) and eigenvector ($\vec{v}$) by using R studio software with the results as follows:

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \\ \lambda_8 \\ \lambda_9 \\ \lambda_{10} \end{bmatrix} = \begin{bmatrix} 4,35 \\ 1,50 \\ 0,99 \\ 0,76 \\ 0,71 \\ 0,59 \\ 0,41 \\ 0,28 \\ 0,23 \\ 0,13 \end{bmatrix}_{(10x1)}$$

   Eigenvector as follow:

$$\begin{pmatrix} -0,388 & 0,037 & -0,429 & 0,135 & 0,259 & -0,030 & 0,199 & -0,182 & 0,035 & 0,708 \\ -0,411 & -0,010 & -0,159 & 0,069 & 0,059 & 0,105 & 0,600 & 0,066 & 0,384 & -0,522 \\ -0,369 & 0,308 & -0,088 & 0,148 & 0,313 & -0,221 & -0,141 & 0,200 & -0,663 & -0,301 \\ 0,195 & -0,602 & -0,245 & -0,177 & -0,123 & -0,348 & 0,380 & 0,323 & -0,344 & 0,048 \\ -0,320 & -0,180 & -0,060 & 0,377 & -0,415 & -0,568 & -0,372 & 0,001 & 0,285 & -0,055 \\ -0,323 & 0,111 & 0,160 & -0,027 & -0,733 & 0,300 & 0,215 & -0,129 & -0,384 & 0,125 \\ -0,205 & -0,570 & -0,304 & -0,046 & 0,090 & 0,449 & -0,398 & -0,311 & -0,131 & -0,233 \\ -0,298 & 0,011 & 0,185 & -0,769 & 0,074 & -0,375 & -0,029 & -0,365 & 0,047 & -0,036 \\ -0,384 & -0,143 & 0,248 & -0,239 & 0,064 & 0,242 & -0,221 & 0,722 & 0,175 & 0,220 \\ 0,135 & 0,386 & -0,710 & -0,357 & -0,294 & 0,052 & -0,202 & 0,220 & 0,113 & -0,089 \end{pmatrix}$$

   The number of eigenvalues that are ≥ 1 is directly proportional to the number of PCs that will be formed. So in this study, three eigenvalues were obtained that met these criteria, namely $\lambda_1 = 4,35$ and $\lambda_2 = 1,50$ and $\lambda_3 = 0,99$ with the value of λ_3 rounded to 1.

4. Determine the number of principal component (PC) that may be formed

   In the principal component analysis (PCA) method, there are several methods to see how many PC are formed, one of which is by looking at the eigenvalue. Based on one of the criteria in the PCA method. Many PC will be formed and considered to have represented the original data when the resulting eigenvalue is ≥ 1 So, in this study three eigenvalues were obtained that met these criteria, which are λ1 = 4,35, λ2 = 1,50 and λ3 = 0,99 rounded to 1.

5. Forms a correlation matrix component that shows the correlation between variables and component scores are shown in Table 4 as follows:

Table 4. Correlation Matrix Components

	$\mathbf{r}_{\mathbf{xp,PC_1}}$	$\mathbf{r}_{\mathbf{xp,PC_2}}$	$\mathbf{r}_{\mathbf{xp,PC_3}}$
x_1	**-0.81**	-0.04	-0.42
x_2	**-0.85**	0.01	-0.15
x_3	**-0.77**	-0.37	-0.08
x_4	-0.40	**0.74**	-0.24
x_5	**0.66**	0.22	-0.06
x_6	**0.67**	-0.13	0.16
x_7	0.42	**0.70**	-0.30
x_8	**0.62**	-0.01	0.18
x_9	**0.80**	0.17	0.24
x_{10}	0.28	-0.47	**-0.70**

Based on Table 4, we can find out that in Mean Years of Schooling (x_1), Expected Years of Schooling (x_2), Purchasing Power Parity (x_3), Percentage of Households With Access To Proper Sanitation (x_5), Percentage of Households With Access To Proper Drinking/Clean Water (x_6), Population Density (x_8) and Unemployment Rate (x_9) most of the data is distributed to PC_1. while in Percentage of Poor Population (x_4) and Life Expectancy (x_7) most of the data is distributed to PC_2. Then on the Prevalence of Stunting In Children (x_{10}) most of the data is distributed to PC_3.

6. Forming New Variables from Reduction Results
   Based on the calculation results, the equation for the principal component (PC) that is formed and the new data set transformation obtained is shown in Table 5 as follows:

Tabel 5. New Variable Reduction Result Using PCA

Districts/cities	**PC$_1$**	**PC$_2$**	**PC$_3$**
Paser	-0.02	-1.65	-0.86
Kutai Barat	-0.15	-2.58	0.25
Kutai Kartanegara	-0.83	-0.78	-1.15
Kutai Timur	-0.47	-1.37	-1.57
⋮	⋮	⋮	⋮
Tarakan	-1.91	-0.44	-1.16

4.3. K-Medoids Algoritm

The K-Medoids algorithm is one of the non-hierarchical clustering in which, it is necessary to determine many clusters at the beginning and also the center points (medoids) to group the research objects.

4.4. Validation of Clustering Results with K-Medoids Algorithm

In this research, the local Silhouette Coefficient value is calculated. Then the global Silhouette Coefficient (SC_{global}) value will be calculated to determine the quality of each clustering result. The results of Cluster Validation Based on SC_{global} Value are shown in Table 6 as follows:

Tabel 6. Comparison of Cluster Validation Results Based on SC_{global} Value

Number of Cluster	SC_{global}
2	0,61
3	0,31
4	0,30
5	0,28

Based on Table 6 shows that the SC_{global} value for validation of data on the results of clustering districts/cities in Borneo Island based on poverty indicators using the K-Medoids algorithm has a variety of SC_{global} values. The largest SC_{global} value is 0.61. Therefore, it can be decided that the most optimal clustering in clustering districts / cities on Borneo Island with the K-Medoids algorithm $k = 2$.

4.5. K-Medoids Algorithm Clustering Results For *k = 2*

Performing clustering using the K-Medoids algorithm with 2 clusters on 56 districts/cities on Borneo Island based on poverty indicators. The clustering results are shown in Table 7.

Table 7. Cluster Members With K-Medoids Algorithm For *k* = 2

Number	Districts/cities	Cluster
1	Paser	1
2	Kutai Barat	1
3	Kutai Kartanegara	1
4	Kutai Timur	1
5	Berau	1
6	Penajam Paser Utara	1
7	Mahakam ulu	1
8	Balikpapan	2
9	Samarinda	2
⋮	⋮	⋮
56	Tarakan	1

4.5. Profilization and Interpretation of Best Clustering Results

In this study, the profilization of the most optimal cluster was obtained and by using the Silhouette Coefficient validity test, the most optimal cluster was obtained and the cluster was 2. The profilization of the best clustering results is shown in Figure 1 as follows:

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

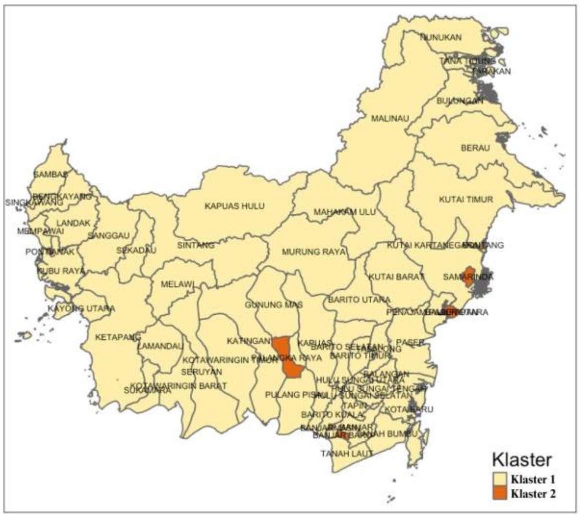

Fig 1. Profilization Map of Best Clustering Results

4. CONCLUSION

Based on the results of the research and discussion, the following conclusions can be drawn:

1. Principal components that are formed from the results of variable reduction using the PCA method by considering eigenvalues that are greater than or equal to one are 3 PC.
2. Based on the silhouette coefficient method, the most optimal k value for clustering districts/cities in Borneo Island based on poverty indicators is to use 2 clusters (k=2) with an value of 0.61.
3. The results of grouping districts/cities in Borneo Island based on poverty indicators resulted in 2 clusters. Cluster 1 consists of 49 districts/cities, while cluster 2 consists of 7 districts/cities.

REFERENCES

[1] Santosa, B. (2007). *Data Mining Techniques for Utilizing Data for Business Purposes.* Yogyakarta: Graha Science.

[2] Prasetyo, E. (2012). *Data Mining: Concepts and Applications Using MATLAB.* Yogyakarta: ANDI.

[3] Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., & Nooraeni, R. (2018). *Data Mining with R: Concepts and Implementation.* Bogor: IN MEDIA.

[4] Han, J., Micheline, K., & Pei, J. (2006). *Data Mining: Concept and Techniques.* San Fransisco: Morgan Kauffman.

[5] Singh, N., & Singh, D. (2012). Performance Evaluation of K-Means And Hierarchy Clustering In Terms of Accuracy And Running Time. *International Journal in Computer, 3*, 4119-4121.

[6] Umar, H. B. (2009). Principal Component Analysis (PCA) and its Application with SPSS. *Journal of Health, 3*, 97-101.

[7] Smith, L. I. (2002). A Tutorial on Principal Component Analysis. *Computer Science Technical Report, 1*, 1-26.

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]

[8] Nasution, M. Z. (2019). Application of Principal Component Analysis (PCA) in Determining Dominant Factors Affecting Student Learning Achievement. *Journal of Information Technology, 3*, 41-48.

[9] Ghaisani, S. Y., Hikmah, N., Prasetyo, A. H., & Widodo, E. (2018). Hierarchical Cluster Analysis for Grouping Provinces in Indonesia Based on Indonesian Democracy Indicators in 2016. 1-11.

[10] Mohammed, N. N., & Abdulazeez, A. M. (2007). Evaluation of Partitioning Around Medoids Algorithm with Various Distances on Microarray Data. *IEEE International*, 1011-1016.

[11] Kaufman, L., & Rousseeuw, P. R. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis.* New York: John Wiley.

[12] Pramesti, D. F., Furqon, M. T., & Dewi, C. (2017). Implementation of K-Medoids Clustering Method for Grouping Potential Forest/Land Fire Data Based on Hotspot Distribution. *Journal of Information Technology and Computer Science Development*, 723-732.

[13] Rachmatin, D. (2014). Application of Agglomerative Methods in Cluster Analysis on Air Pollution Level Data. *Scientific Journal of Mathematics Study Program STKIP, 3*, 133-149.

[14] Soemartini. (2008). Principal Component Analysis (PCA) as a Method to Solve Multicollinearity Problem. *Journal of Technology and Information, 6*, 1-9.

[15] Afira, N. (2019). Poverty Cluster Analysis of Provinces in Indonesia in 2019 using Partitioning and Hierarchical Methods. *Journal of Computer System, 10*, 101-109.

[DOI: 10.14710/JSUNIMUS.11.2.2023.31-43]