Application of the K-Means algorithm to determine poverty status in Hulu Sungai Tengah

by Dewi Anggraini

Submission date: 22-Aug-2022 09:52AM (UTC-0400)
Submission ID: 1885496742
File name: algorithm_to_determine_poverty_status_in_Hulu_Sungai_Tengah.pdf (475.22K)
Word count: 3729
Character count: 17481

Journal of Physics: Conference Series

Application of the K-Means algorithm to determine poverty status in Hulu Sungai Tengah

Application of the K-Means algorithm to determine poverty status in Hulu Sungai Tengah

N Istiqamah, O Soesanto, and D Anggraini

Statistics Study Program, Faculty of Mathematics and Science, Universitas Lambung Mangkurat, Indonesia

istiqamahnurul528@gmail.com

Abstract. Poverty is a condition of living in an inability to meet the minimum needs of life or basic needs. In Indonesia, poverty is one of the main problems that still need an optimal solution. Several government programs to address the problem of poverty have been carried out, but not infrequently the implementation is not right on target. The importance of this assistance is expected to improve the welfare of the community so it is very unfortunate if the assistance has not been right on target. This study aims to determine the status of poverty in Hulu Sungai Tengah Regency. By observing a problem above, it can be necessary to use a grouping method in determining poverty status. so that in this study using the cluster method, namely K-Means in clustering population data. Based on the results of data analysis using 353 head of family in the population data of HST Regency, it can be concluded that there are three poverty status clusters, namely low-level poverty (cluster 3) with a total of 130 head of family, medium-level poverty (cluster 2) with a total of 130 head of family. 111 head of family, and high poverty level (cluster 1) with a total of 112 head of family.

1. Introduction

Poverty is a condition of living in an inability to meet the minimum needs of life or basic needs. Poverty occurs not only because of an income but also because of limited household facilities and infrastructure. In Indonesia, poverty is one of the main problems that still need an optimal solution [1].

Several government programs to overcome the problem of poverty have been carried out, however, not infrequently the implementation is not right on target. There are several main factors causing the inaccuracy of government program targets/targets in overcoming poverty, namely the accuracy of the data and the accuracy of the data analysis used to determine the poverty status of the population [1]. For example, the results of the information obtained in the Mata Banua article in 2020 that in Hulu Sungai Tengah (HST) Regency there are still delays in the process of delivering social assistance in the Regency. The results of this delay information are due to one of the reasons for the non-optimal process of data collection and distribution at the local government level [2] So looking at the problems above, a grouping method is needed so that the data reading process is faster.

Clustering is a method that can be used to group data objects that have the same characteristics into one cluster, data with different characteristics will be grouped into other groups. In a previous study conducted by Aras in 2016 the clustering method using the K-Means algorithm was used to determine the priority of the recipients of home surgery assistance. Then in 2019 and 2020, this method was also used to determine poverty status clusters based on population data in South Jambi District, West Java Province, and Banten Province. From the conclusion of previous studies that the K-Means algorithm is suitable for poverty data. [1,3,4].

Based on the description above, the research was carried out using the Clustering method, namely the K-Means Algorithm to determine poverty status clusters based on population data in HST Regency with the title "Application of the K-Means Algorithm to Determine Poverty Status in Hulu Sungai Tengah Regency".

2. Literature Review

2.1. Descriptive Statistics

Descriptive statistics is a method used in collecting data, processing, presenting, and calculating other measures. In addition, to make the data easier to understand, it can be done in the form of tabulations, diagrams or graphs [5].

2.2. Data Minig

Data mining is a process that uses statistical, mathematical, and machine learning techniques in extracting and identifying useful information and related knowledge from various data [6]

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. [7]

2.3. Clustering

Cluster analysis is an analysis that has the purpose of grouping data objects on the condition that they have similar characteristics and then becomes one cluster, but if objects with different characteristics will be grouped in different clusters [8]. Clustering is divided into two clustering are hierarchical and non-hierarchical. Hierarchy is a method that creates a level of data or objects in a structured manner based onthe similarity of its nature and the desired cluster and the number of unknowns is unknown, while non-hierarchical is used for grouping data or objects in which the number of clusters to be formed can be determined in advance. [9]

2.4. K-Means Clustering

K-Mean Clustering Algorithm is one of the methods of part cluster non-hierarchical grouping to partition existing data into one or more clusters, so that data with the same characteristics will be grouped into one cluster and data with different characteristics will be grouped into other groups [10]. In K-means, each data must be entered into a certain cluster, but it is possible for each data to be entered into a certain cluster at one stage of the process, in the next step, moving to another cluster [11]. The procedure of the K-Means algorithm:

1) Determine k as the number of clusters that you want to form.
2) Determine the initial centroid randomly / randomly as many as k are formed. To calculate the next centroid, the following formula is used:

$$c_j = \frac{1}{p}\sum_{i=1}^{p} x_{ij} \quad (1)$$

3) Determine the initial centroid randomly / randomly as many as k are formed.
4) Calculate the distance of each data to the centroid of each cluster. To calculate the distance between the data and the centroid, you can use *Euclidian Distance*.

$$d_{ij} = \sqrt{\sum_{i=1}^{n} (x_i - c_j)^2} \quad (2)$$

5) Repeating Steps c to e, until a convergent condition is reached, i.e. no one moves clusters or the result of the cluster position in the last iteration is the same as the position of the previous iteration.

2.5. Elbow Method

The Elbow method is a method used to determine the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form like an elbow at a point [12]. To get the results of the comparison by calculating the SSE (Sum of Square Error) from each distance. The following is the equation of the SSE formula:

$$SSE_i = \sum_{i=1}^{j} (x_{ij} - c_j)^2 \quad (3)$$

After calculating the SSE for each cluster, the more clusters there will be a decrease in the SSE value. Elbow values are obtained from a drastic decrease in SSE and the subsequent decrease slowly. So that if the SSE value is represented in a graphic form, then the elbow value will form an angle like an elbow in the number of clusters 3. In Figure 1, the SSE value is given with the number of clusters.

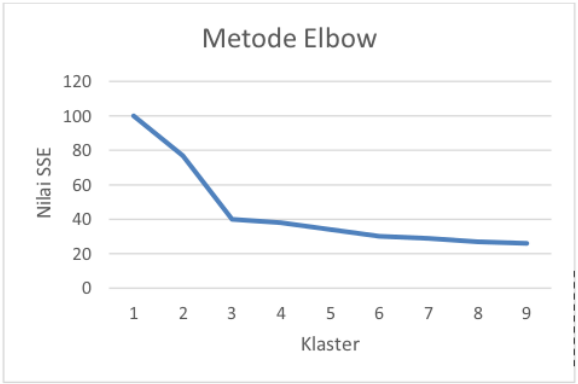

Figure 1. SSE charts that make up the Elbow

2.5.1 Decision tree.

Decision tree is a flowchart structure that resembles a tree, where each internal node represents a test on a variable, each branch represents the test result, and a leaf node represents a class [13]. Basic Concepts of Decision Tree is turning data into a decision tree and decision rules, where each node represents attribute, branch represents the value of the attribute, and leaves represents the class [14]. The root of the Decision tree can be seen in Figure 2 below.

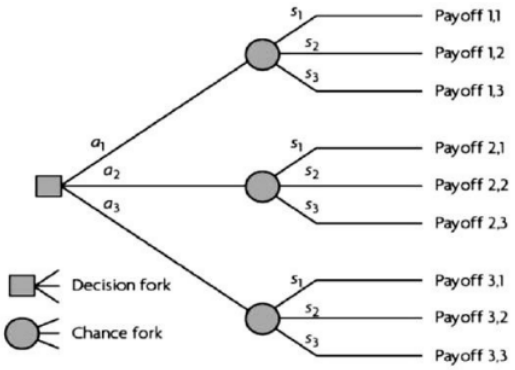

Figure 2. General Decision Tree Forms

In making a tree Decision tree can determine a tree root, the root will be taken from the selected variable by calculating the gain value of each, the highest gain value will be the first root [15].

3. Method

3.1 Data Sources

The research material used in this study is population data with a total sample of 353 household heads in 2019 in Hulu Sungai Tengah Regency, where this data was obtained from the Office of Social Affairs, Family Planning Population Control, Women Empowerment and Child Protection, Hulu Sungai Tengah (Dinas Sosial, PPKB, PPPA).

3.2 Research Variable

In this study, there were 10 variables used, namely building status, number of household members, floor type, floor area, wall type, drinking water source, power, cooking fuel, facilities for final disposal of feces, and last education.

3.3 Analysis Steps

(1) Data preprocessing
(2) Descriptive statistical analysis
(3) The process of determining poverty status using the K-Means algorithm
(4) Determining the optimal number of poverty status using the elbow method

(5) Perform interpretations based on the results of the analysis.
(6) Conclusions based on the results of the analysis that has been done.

4. Results and Discussion

4.1. Descriptive Analysis

In this section, information will be described or described regarding the variable number of household members, building status, floor area, type of floor, type of wall, source of drinking water, electricity, cooking fuel, and defecation facilities with 353 head of family. The following is a graph that represents the distribution of data from each variable used in this study.

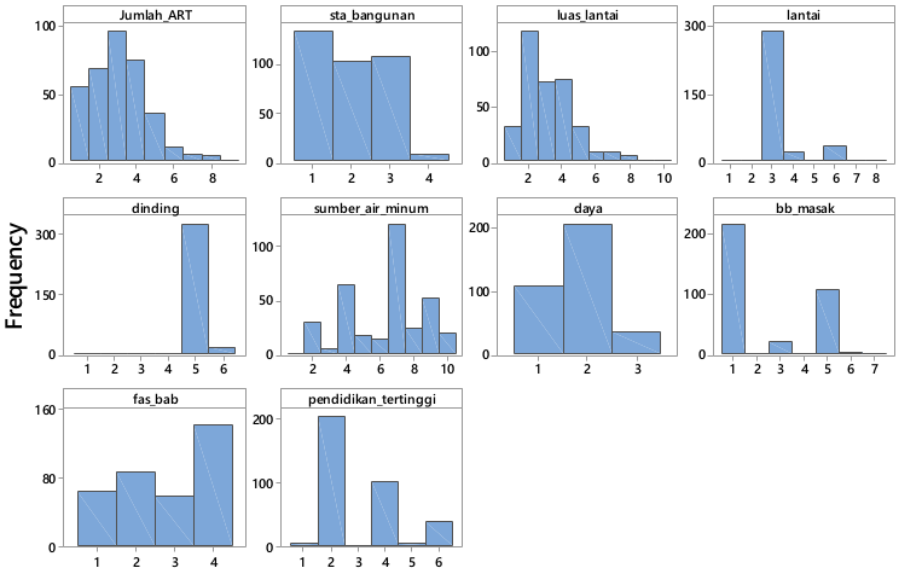

Figure 3. Distribution of Research Variable Data

From the picture above shows that the distribution of variable data from the study, so it can be seen the attributes of the dominant research variables are as follows:

1) The number of household members is 3 people, namely 97 families (27.6%).
2) The status of the building is still in the status of lease or contract as many as 134 families (38.1%).
3) The floor area is 15 – 26 M2 as many as 119 families (33.8%).
4) The type of floor is using low-quality wood as much as 290 families (82.4%).
5) The type of wall is using wood as much as 328 families (93.2%).
6) The source of drinking water is using drilled wells or pumps as many as 121 families (34.4%).
7) Electrical power is using 900 watts of electrical power as much as 208 families (59.1%).
8) The fuel is using firewood as much as 217 families (61.6%).
9) Ownership of defecation facilities is using their own as many as 142 families (40.3%).

4.2. Determining poverty status using *the* K-Means algorithm

In determining poverty status, it can be determined the number of clusters or the best number of statuses in determining poverty status. The following is the result of processing the elbow method to determine by calculating the SSE value in each cluster 2 to 9 which is shown in Table 1.

Table 1. SSE Value with Number of Clusters 1-9

Cluster	SSE	Difference
2	5174,937963944857	
3	4261,164533214533	913.7734
4	3959,239419204425	301.9251
5	3679,7822801537086	279.4571
6	3508,269247199046	171.513
7	3317,760861628314	190.5084
8	3198,9036199095026	118.8572
9	3047,3505182255667	151.5531

So that the SSE value in Table 1 be represented in graphical form, the following Figure 4 is a graph of the elbow. So that the SSE value in Table 1 can be represented in graphical form, the following Figure 4 is a graph of the elbow.

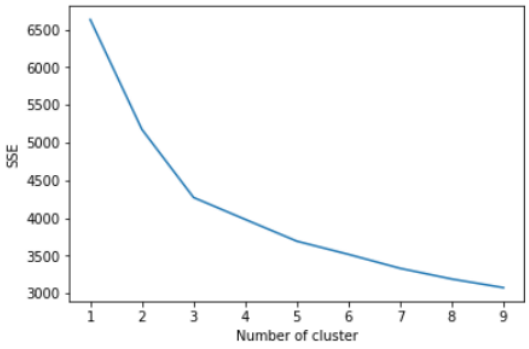

Figure 4. SSE of Value Elbow Chart

Based on the results from Table 1 and Figure 4, it shows that the largest decrease in SSE value occurs in cluster 3 so that in cluster 3 gives an angle like an elbow in the graph and then there is a stable decrease in SSE value. In accordance with the concept of the elbow method, the cluster value used is 3 clusters.

In the K-Means algorithm process, the first step is to determine the number of groups or clusters of 3 clusters from 353 head of family. After determining the number of clusters, the centroid value is determined for each cluster for each variable. The centroid value in the first iteration (first time calculation) is given randomly. In the next iteration, the centroid value (1st iteration up to the normal position/maximum iteration) is given by calculating the average value of the data in each cluster. If the old centroid value is not the same as the new centroid value, then the iteration process is continued until the value is the same or up

to the maximum iteration value that has been previously set (eg 50). For example, if the second centroid is the same as the first centroid, the grouping process stops.

Table 2. Early of Centroids

Centroid	Friday ART	Building Station	Floor area	Floor type	Wall type	Drinking Water Sum	Power	BB of cook	chapter	Pen. highest
C1	4	3	8	3	5	7	2	5	3	2
C2	5	3	4	3	5	9	2	1	2	4
C3	3	3	3	3	5	7	3	5	3	2

The centroid value in Table 2 will be used to calculate the distance between the data and the centroid. The following is an example of calculating the Euclidien distance equation.

$$d_{1,1} = \sqrt{(3-4)^2 + (1-3)^2 + (1-8)^2 + (3-3)^2 + (5-5)^2 + (10-7)^2 + (2-2)^2 + (5-5)^2 + (3-3)^2 + (6-2)^2}$$

$$d_{1,1} = 8{,}89$$

$$d_{1,2} = \sqrt{(3-5)^2 + (1-3)^2 + (1-4)^2 + (3-3)^2 + (5-5)^2 + (10-9)^2 + (2-2)^2 + (5-1)^2 + (3-2)^2 + (6-4)^2}$$

$$d_{1,2} = 6{,}24$$

$$d_{1,3} = \sqrt{(3-3)^2 + (1-3)^2 + (1-3)^2 + (3-3)^2 + (5-5)^2 + (10-7)^2 + (2-3)^2 + (5-5)^2 + (3-3)^2 + (6-2)^2}$$

$$d_{1,3} = 5{,}83$$

After the second centroid value is obtained, the second centroid value will be compared with the first centroid value. If there is a difference between the values of the two centroids, then continue the process of calculating the distance for each data using the second centroid value. Because the calculation results show that there is no difference in the value of the second centroid and the value of the first centroid, the calculation of the distance between each data is continued with the value of the second centroid. Then repeat the third step, which is to determine the distance between the data and the cluster center. In this study, the centroid value experienced no difference between the sixth and seventh centroid values. It can be concluded that all processes in K-Means were completed in the 6th iteration. 6th iteration centroid.

Table 3. Value of Centroid (last/iteration 6)

C1	2.88	2.06	2.88	3.40	4.98	3.50	1.72	1.70	2.54	3.20
C2	3.12	2.21	3.26	3.35	5.01	7.49	1.70	1.04	2.52	2.81
C3	3.36	1.60	3.23	3.44	5.07	7.85	1.98	4.78	3.36	3.08

Based on the results of the K-Means calculation, the cluster consists of 3 clusters of the poverty status is 353 families, so that there are 111 family heads in the first cluster, 130 family heads in the second cluster, and 112 family heads in the third cluster.

4.3. Process Decision tree

Decision tree is a flowchart structure that resembles a tree (Tree), which is used to represent data resulting from the K-Means clustering process. There are several stages of the Decision tree process:

1) Calculate the entropy value, to determine the gain value of each variable. The entropy value is also useful for being a condition or branch in the root of a Decision tree. It is known that the poverty status of cluster 1 is 112 families, cluster is 130 families, cluster 3 is 111 families. The following is a calculation of the total entropy value for cluster 1, cluster 2 and cluster 3.

$$S = \left(-\frac{112}{353}\right) * log_2\left(\frac{112}{353}\right) + \left(-\frac{130}{353}\right) * log_2\left(\frac{130}{353}\right) + \left(-\frac{111}{353}\right) * log_2\left(\frac{111}{353}\right)$$
$$= 1.58$$

2) After knowing each entropy, the next process is to calculate the gain value for each variable and determine the highest gain value. The variable with the highest gain value will be used as the main root in the decision tree root. For example, the gain value for the building status variable is as follows using Equation 2.7:

$$S, A = 1.51 - \left(\left(\frac{134}{353}\right) * 1.48 + \left(\frac{103}{353}\right) * 1.48 + \left(\frac{108}{353}\right) * 1.51 + \left(\frac{8}{353}\right) * 1.56\right)$$
$$= 0.09$$

3) The calculation of the entropy and gain values for all variables is carried out to obtain the highest gain value which will be used as the root. The following Table 4 provides the results of the calculation of the gain value for all variables.

Table 4. Gain of Value

Variable	Value Gain
ART member	0.025
Building Status	0.09
Floor area	0.04
Floor Type	0.02
Wall Type	0.024
Drinking Water Source	0.90
Electrical power	0.03
Cooking Fuel	0.719
BAB Facilities	0.108

4) Based on the results of the calculation, the results of the analysis show that there are 3 clusters namely poor, medium and rich for determining the status of poverty in the community in HST district. So the results of the Decision tree show that the significant variables from this research are the source of drinking water, cooking fuel and floor area from the root, a rule is obtained so that it can assist in categorizing a KKT into 3 clusters, namely cluster 1 (Poor), cluster 2 (Medium) and cluster 3 (Rich) including the categories of Poor, Medium, and Rich.
   a. Cluster 1: Sources of low-level drinking water such as rainwater, rivers, earthen wells, while low-level fuel sources such as firewood and kerosene: and have the number of household members=3.
   b. Cluster 2: Medium level drinking water sources such as protected springs: fuel used unless low level is not used such as firewood, and kerosene.
   c. Cluster 3: Sources of high-level drinking water such as protected springs and bottled drinking water: fuel used is at least >3 kg of gas to use electricity: and has a small number of household members = 2.

5. Conclusion

K-Means Clustering is one method that can assist in determining poverty status so that it can help the government to more easily overcome delays in the process of determining aid, based on the results of the formation of 3 poverty statuses in Hulu Sungai Tengah Regency, namely high poverty levels are members who are in cluster 1: medium level is a member who is in cluster 2: and low level is a member of low level cluster.

References

[1] Sunia D 2019 *J. Ilm. Mhs. Tek. Inform.* **1** 121–34
[2] Banua M 2020 Wabup HST Serahkan Dana BST dan BLT untuk 1.305 KPM/KK *Mata banua*
[3] Febianto N I and Palasara N 2019 *J. Sisfokom (Sistem Inf. dan Komputer)* **8** 130
[4] Sari Y R, Sudewa A, Lestari D A and Jaya T I 2020 *CESS (Journal Comput. Eng. Syst. Sci.* **5** 192
[5] Pradana M and Reventiary A 2016 *J. Managemen* **6** 1–10
[6] Maulida L 2018 s *JISKA (Jurnal Inform. Sunan Kalijaga)* **2** 167
[7] Marnoto T 2014 *Int. J. Sci. Eng.* **7** 155–60
[8] Nagari S S and Inayati L 2020 *J. Biometrika dan Kependud.* **9** 62
[9] Talakua M W, Leleury Z A and Talluta A W 2017 *Barekeng J. Ilmu Mat. dan Terap.* **11** 119–28
[10] Darmi Y and Setiawan A 2016 *J. Media Infotama Univ. Muhammadiyah Bengkulu* **12** 148–57
[11] Sarasvananda I B G, Wardoyo R and Sari A K 2019 *IJCCS (Indonesian J. Comput. Cybern. Syst.* **13** 313
[12] Putu N, Merliana E and Santoso A J Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means 978–9
[13] Kasih P 2019 *Innov. Res. Informatics* **2** 63–9
[14] Utama T D, Sihwi S W and Doewes A 2014 *ITSMART: J. Teknologi dan Informasi* ***3*** 74-83
[15] Nasrullah A H 2018 *Ilk. J. Ilm.* **10** 244–50

Application of the K-Means algorithm to determine poverty status in Hulu Sungai Tengah

ORIGINALITY REPORT

15%	9%	10%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	m.iopscience.iop.org Internet Source	1%
2	pdfs.semanticscholar.org Internet Source	1%
3	Ma'shum Abdul Jabbar, Suharjito Suharjito. "Fraud Detection Call Detail Record Using Machine Learning in Telecommunications Company", Advances in Science, Technology and Engineering Systems Journal, 2020 Publication	1%
4	repository.unmuhjember.ac.id Internet Source	1%
5	Submitted to Louisiana State University Student Paper	1%
6	Submitted to Universitas Prima Indonesia Student Paper	1%
7	V Acioly, T Paiva, G Azevedo, T Rocha, R Picoreti, A C F Santos. "Shedding synchrotron	1%

light on teacher training", Physics Education, 2021
Publication

8	Maghfirah Dinsyah Febriana, Zahir Zainuddin, Ingrid Nurtanio. "School zoning system using K-Means algorithm for high school students in Makassar City", 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019 Publication	1 %
9	Submitted to Udayana University Student Paper	1 %
10	Submitted to President University Student Paper	1 %
11	Vatti, Rambabu A., and A.N. Gaikwad. "Throughput Improvement of Randomly Deployed Wireless Personal Area Networks", IERI Procedia, 2014. Publication	1 %
12	karyailmiah.unisba.ac.id Internet Source	<1 %
13	Solmin Paembonan, Abdul Rachman Manga, Jusmidah, Dedy Atmajaya, Ayu Vina Waluyantari, Wistiani Astuti, St. Hajrah Mansyur. "Combination of K-Means and Profile Matching for Drag Substitution", 2018	<1 %

2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2018
Publication

14	Agyztia Premana, Akhmad Pandhu Wijaya, Moch Arief Soeleman. "Image segmentation using Gabor filter and K-means clustering method", 2017 International Seminar on Application for Technology of Information and Communication (iSemantic), 2017 Publication	<1 %
15	Dhendra Marutho, Sunarna Hendra Handaka, Ekaprana Wijaya, Muljono. "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News", 2018 International Seminar on Application for Technology of Information and Communication, 2018 Publication	<1 %
16	repositorio.unesp.br Internet Source	<1 %
17	www.research-collection.ethz.ch Internet Source	<1 %
18	Wahyudi Setiawan, Agus Purnama. "Tobacco Leaf Images Clustering using DarkNet19 and K-Means", 2020 6th Information Technology International Seminar (ITIS), 2020 Publication	<1 %

19	astronomy.byu.edu Internet Source	<1 %
20	journals.univ-danubius.ro Internet Source	<1 %
21	ojs3.unpatti.ac.id Internet Source	<1 %
22	Ina Maryani, Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, Eva Argarini Pratama. "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm", 2018 Third International Conference on Informatics and Computing (ICIC), 2018 Publication	<1 %
23	Guohua Geng. "Application of fuzzy cluster analysis for medical image data mining", IEEE International Conference Mechatronics and Automation 2005 ICMA-05, 2005 Publication	<1 %
24	Rusdiansyah Rusdiansyah, Hendra Supendar, Tuslaela Tuslaela. "Data Mining using K-means method for feasibility selection of Non-cash food Assistance recipients in the Era of Covid-19", SinkrOn, 2021 Publication	<1 %
25	eprints.ipdn.ac.id Internet Source	<1 %

26	businessdocbox.com Internet Source	<1 %
27	doczz.fr Internet Source	<1 %
28	mafiadoc.com Internet Source	<1 %
29	onlinelibrary.wiley.com Internet Source	<1 %
30	Monica Natalia Bangun, Open Darnius, Sutarman Sutarman. "OPTIMIZATION MODEL IN CLUSTERING THE HAZARD ZONE AFTER AN EARTHQUAKE DISASTER", SinkrOn, 2022 Publication	<1 %
31	Sunarmo, Achmad Affandi, Surya Sumpeno. "Clustering Spatial Temporal Distribution of Fishing Vessel Based lOn VMS Data Using K-Means", 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		