International Journal of Information System & Technology
Akreditasi No. 36/E/KPT/2019 | Vol. 3, No. 2, (2020), pp.315-322

Implementation of Data Mining Algorithms for Grouping Poverty Lines by District/City in North Sumatra

Mhd Ali Hanafiah[1], Anjar Wanto[2]*
[1]Politeknik Bisnis Indonesia, Pematangsiantar, Indonesia
[2]STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

[1]ikh.alie84@gmail.com, [2]anjarwanto@amiktunasbangsa.ac.id*

Abstract

The poverty line is useful as an economic tool that can be used to measure the poor and consider socio-economic reforms, such as welfare programs and unemployment insurance to reduce poverty. Therefore, this study aims to classify poverty lines according to regencies/cities in North Sumatra Province, so that it is known which districts/cities have high or low poverty lines. The grouping algorithm used is K-Means data mining. By using this algorithm, the data will be grouped into several parts, where the process of implementing K-Means data mining uses Rapid Miner. The data used is the poverty line data by district/city (rupiah/capita/month) in the province of North Sumatra in 2017-2019. Data sourced from the North Sumatra Central Statistics Agency. The grouping is divided into 3 clusters: high category poverty line, medium category poverty line, and the low category poverty line. The results for the high category consisted of 5 districts/cities, the medium category consisted of 18 districts/cities and the medium category consisted of 10 districts/cities. This can provide input and information for the North Sumatra government to further maximize efforts to overcome the poverty line in the area

Keywords: *Implementation, Data Mining, Grouping, Poverty Line, North Sumatra*

1. Introduction

The problem of poverty is complex and multidimensional in nature, so it becomes a development priority [1]. The Central Bureau of Statistics (BPS) calculates the poverty rate using the poverty line from household per capita expenditure data from the national socio-economic survey (Susenas) [2]. The poverty line or poverty threshold is the minimum level of income that is deemed necessary to be met to obtain a sufficient standard of living in a country [3]. The poverty line is obtained from the sum of the food poverty line and the non-food poverty line [4]. The poverty line is used as a limit to classify the population as poor or non-poor [5]. In practice, the public's official or general understanding of the poverty line (as well as the definition of poverty) is higher in developed countries than in developing countries. Almost every area has people living in poverty. The poverty line is useful as an economic tool that can be used to measure the poor and consider socio-economic reforms, for example, such as welfare improvement programs and unemployment insurance to reduce poverty [6].

The purpose of this research is to classify poverty lines according to regencies/cities in North Sumatra Province, so that it is known which districts/cities have high or low poverty lines. This can be used as input and information for the North Sumatra government to further maximize efforts to overcome the poverty line problem in the area. The method of grouping in this study uses the K-Means data mining algorithm. Apart from grouping, data mining is also often used for data classification problems [7]–[11].

Previous studies that became the reference for this study include research that discusses the grouping of rice plants in Indonesia based on 34 provinces. This study resulted in grouping data on rice plants divided into 3 groups, namely high consisting of 3 provinces, normal consisting of 23 provinces and low group consisting of 8 provinces [12]. Next is research on the grouping of disaster-prone areas according to provinces in

ISSN : 2580-7250

Indonesia. In this study, the grouping of disaster-prone areas consisted of 4 provinces with high clusters, 14 provinces included in the medium cluster and 16 provinces with low clusters [13]. The next reference research is to classify population density, human development index, open unemployment rate and school enrollment rates by province in Indonesia. There are 5 clusters from the results of this study including 12 provinces in the category of cluster 1, 6 provinces in the category of cluster 2, 1 province in the category of cluster 3, 6 provinces in the category of cluster 4 and 9 provinces in the category of cluster 5 [14]. These related studies are the background for conducting research to classify poverty lines according to districts/cities in North Sumatra which are expected to be useful information for the North Sumatra provincial government to further maximize efforts in overcoming the poverty line.

2. Research Methodology

2.1. Data collection

Data collection using quantitative methods. The research data is in the form of Poverty Line data by Regency/City (rupiah/capita/month) in North Sumatra province in 2017-2019 which is presented in table 1.

Table 1. Poverty Line by Regency/City (rupiah/capita/month), 2017-2019

No	Regency/City	Poverty Line by Regency/City (rupiah/capita/month)		
		2019	2018	2017
1	Asahan	330.460,00	315.420,00	305.868,00
2	Batu Bara	408.417,00	381.651,00	363.741,00
3	Binjai	403.798,00	380.792,00	371.387,00
4	Dairi	341.511,00	325.176,00	310.836,00
5	Deli Serdang	390.440,00	381.173,00	363.371,00
6	Gunungsitoli	339.671,00	327.303,00	318.585,00
7	Humbang Hasundutan	336.500,00	329.189,00	313.545,00
8	Karo	460.870,00	437.702,00	423.663,00
9	Labuanbatu Utara	422.063,00	395.696,00	378.024,00
10	Labuhan Batu	389.402,00	368.357,00	352.622,00
11	Labuhanbatu Selatan	368.205,00	355.517,00	346.305,00
12	Langkat	392.050,00	382.536,00	364.517,00
13	Mandailing Natal	356.058,00	336.820,00	319.777,00
14	Medan	532.055,00	518.420,00	491.496,00
15	Nias	361.698,00	353.141,00	346.374,00
16	Nias Barat	393.450,00	386.431,00	361.397,00
17	Nias Selatan	279.468,00	261.104,00	249.225,00
18	Nias Utara	390.564,00	383.552,00	381.696,00
19	Padang Lawas	332.350,00	310.569,00	281.464,00
20	Padang Lawas Utara	342.885,00	321.076,00	291.036,00
21	Padangsidimpuan	382.884,00	363.468,00	348.074,00
22	Pakpak Bharat	287.654,00	283.258,00	256.781,00
23	Pematangsiantar	502.726,00	474.084,00	464.794,00
24	Samosir	315.825,00	299.640,00	287.857,00
25	Serdang Bedagai	382.283,00	361.623,00	350.892,00
26	Sibolga	425.236,00	415.478,00	413.454,00
27	Simalungun	359.540,00	342.477,00	331.860,00
28	Tanjungbalai	421.671,00	397.647,00	374.442,00
29	Tapanuli Selatan	364.798,00	343.407,00	340.065,00
30	Tapanuli Tengah	376.474,00	369.471,00	367.687,00
31	Tapanuli Utara	377.948,00	357.464,00	344.644,00
32	Tebing Tinggi	460.533,00	426.469,00	415.307,00
33	Toba Samosir	373.020,00	352.860,00	345.591,00

Source: Central Bureau of Statistics of North Sumatra Province [15]

2.2. Research Flowchart

The following will present a research flowchart of the K-Means Clustering algorithm.

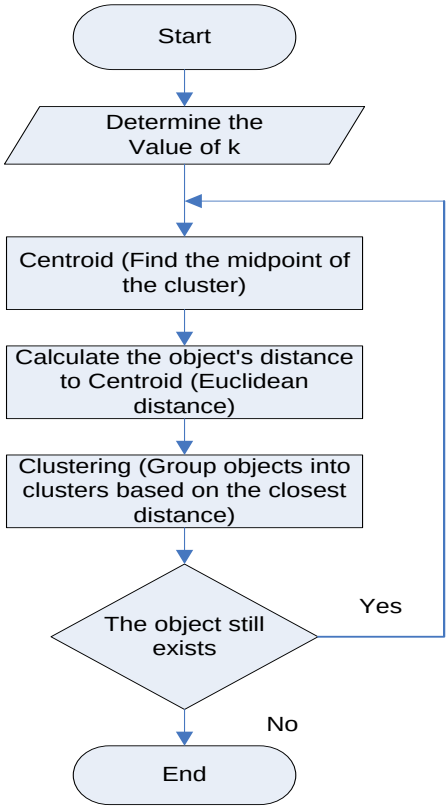

Figure 1. Research Flowchart [16][17]

The steps of the K-Means algorithm can be explained as follows [18]–[22]:

1. Determine the number of clusters (k) in the data set.
2. Determine the center value (Centroid)
   Determination of the centroid value at the initial stage is carried out randomly, while in the iteration stage the formula is used as in equation (1) below.

$$V_{ij} = \frac{1}{Ni} \sum_{k=0}^{Ni} Xkj \quad (1)$$

3. On each record, calculate the closest distance to Centroid.
   There are several ways that can be used to measure the distance of data to the center of the group, including Euclidean, Manhattan/City Block, and Minkowsky. Each method has advantages and disadvantages of each. For writing in this chapter, the Centroid distance used is Euclidean Distance, with the following formula.

$$De = \sqrt{(xi - si)^2 + (yi - ti)^2} \quad (2)$$

4. Group objects by distance to the nearest Centroid
5. Repeat step 3 to step 4, iterating until Centroid is optimal.

3. Results and Discussion

3.1. Centroid Data

The midpoint value on Centroid is determined based on the desired grouping. In this paper, the poverty line is divided into 3 criteria, namely the high group, the medium group and the low group. So based on the data in table 1, the high Centroid value (C1) is taken from the maximum data, while for the moderate Centroid value (C2) is taken from the average value and the value for low Centroid value (C3) is taken from the minimum value. The cluster point value can be seen in table 2 below (Based on table 1).

Table 3. Initial Centroid Value

Centroid	C1 (Max)	532.055,00	518.420,00	491.496,00
	C2 (Ave)	381.894,15	364.817,30	350.799,30
	C3 (Min)	279.468,00	261.104,00	249.225,00

3.2. Process Grouping with Rapidminer

In this paper, the poverty line grouping according to districts/cities in North Sumatra is divided into 3 groups, namely: the high group, the medium group and the low group. The following is the grouping process and the results of the K-Means algorithm which is carried out with Rapidminer.

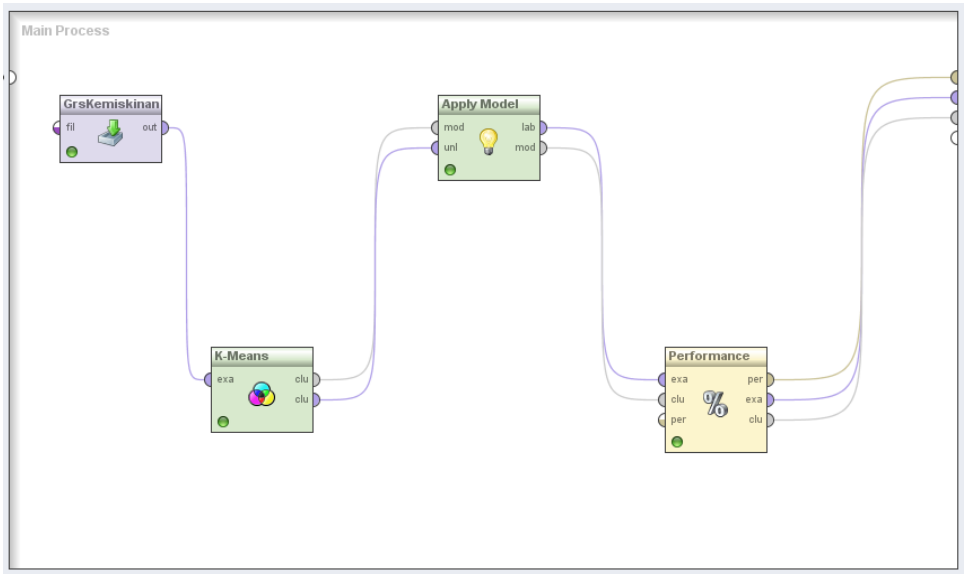

Figure 2. Process K-Means with Rapidminer (K Value = 3)

Figure 2 describes the process of grouping or clustering the K-means algorithm using Rapidminer which begins with importing poverty line excel data according to districts / cities in North Sumatra, then continues with the selection of the K-means algorithm operator for the clustering. Value k = 3, the measure types used are MixedMeasures. After that, it is connected to the Apply Model operator to apply the model that has been learned or trained. The aim is to obtain predictions on unlabeled data (testing data) that do not have a label. The next stage is to connect to the Performance operator to evaluate the performance of the model which provides a list of performance criteria values automatically according to the assigned task. The results can be seen in figure 3, figure 4, figure 5, figure 6 and figure 7 below.

Cluster Model

```
Cluster 0: 5 items
Cluster 1: 10 items
Cluster 2: 18 items
Total number of items: 33
```

Figure 3. Cluster Results (Grouping)

Based on Figure 3 it can be explained that of the three resulting clusters there are 5 items for Cluster_0, 10 items for Cluster_1 and 18 items for Cluster_2. For the final results, the Centroid table can be seen in Figure 4.While the results of the plot view of the poverty line cluster according to districts/cities in North Sumatra are presented in Figure 4.

Attribute	cluster_0	cluster_1	cluster_2
No	20.600	13.300	18.056
2019	476284	326238.200	386594.722
2018	454430.600	310955.500	369847.944
2017	441742.800	293497.400	357371.611

Figure 4. Centroid Table

The results of the description of the poverty line clustering (grouping) according to districts/cities in North Sumatra can be seen on the option button presented in Figure 5.

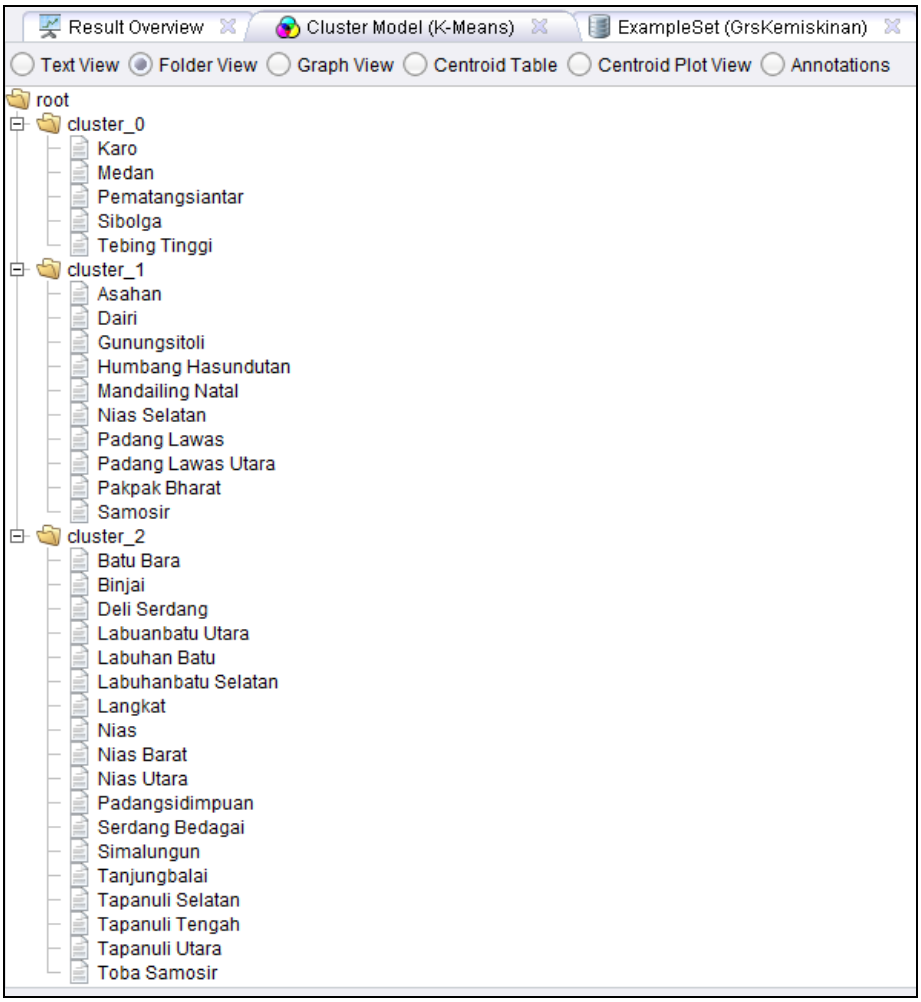

Figure 5. Poverty Line View Folder of Districts/Cities in North Sumatra

Based on number 5, it can be believed that Culster_0 consists of 5 districts/cities (Karo, Medan, Pematangsiantar, Sibolga and Tebing Tinggi). Cluster_1 consists of 10 districts/cities (Asahan, Dairi, Gunungsitoli, Humbang Hasundutan, mandailing Natal, South Nias, Padang Lawas, North Padang Lawas, Pakpak Bharat and Samosir). Cluster 2 consists of Coal, Binjai, Deli Serdang, North Labuanbatu, Labuhan Batu, South Labuanbatu, Langkat, Nias, West Nias, North Nias, Padangsidempuan, Serdang Bedagai, Simalungun, Tanjungbalai, South Tapanuli, Tapanuli Tengah, North Tapanuli and Toba Samosir

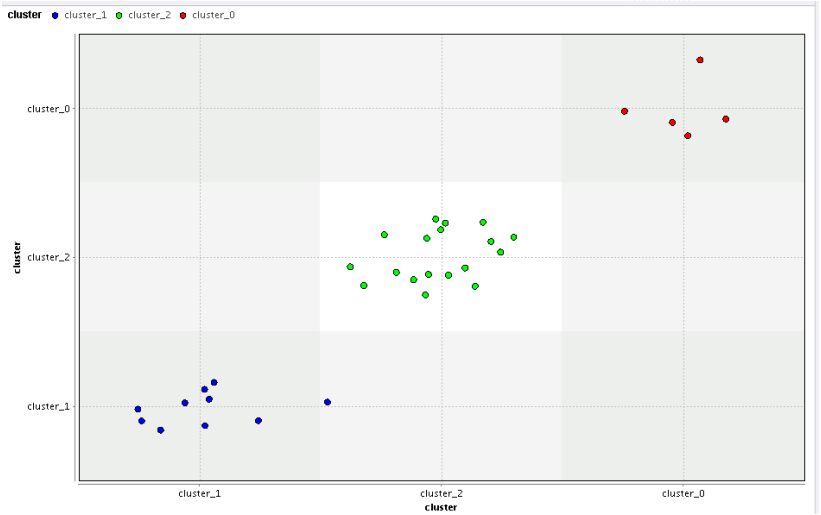

Figure 6. Plot View of Poverty Line Districts/Cities in North Sumatra

Figure 6 shows that the red dots represent the Cluster_0 group which consists of 5 districts/cities. The green dots are the Cluster_2 group consisting of 18 districts/cities and the blue dots are the Cluster_3 group which consists of 10 districts/cities.

3.4. Performance Vector K-Means

Seeing the performance of K-Means using the Rapid Miner tool is to add the% Performance operator which can evaluate the performance of the centroid-based clustering algorithm. This operator provides a list of performance criteria values based on cluster centroid. % Performance measurement parameters are Avg._within_centroid_distance and Davies Bouldin. In this study the results of Avg._within_centroid_distance = -1491908671.321 and Davies_Bouldin = -0.554. The results of K-Means performance using the Rapid Miner tools can be seen in Figure 7 below.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -1491908671.321
Avg. within centroid distance_cluster_0: -3758759024.640
Avg. within centroid distance_cluster_1: -1616529551.460
Avg. within centroid distance_cluster_2: -792994195.321
Davies Bouldin: -0.554
```

Figure 7. Performance Vector K-Means with Rapidminer

4. Conclusion

The use of the K-means algorithm can be used to group poverty lines for districts / cities in the province of North Sumatra. Based on the results of the K-means algorithm analysis with the help of Rapidminer, the description of the District / City Poverty Line groups in North Sumatra province is as follows: The poverty line group that is included in the high cluster consists of 5 districts / cities including Karo, Medan, Pematangsiantar, Sibolga and Tebing High. The poverty line group that is included in the moderate cluster consists of 18 districts / cities including Batu Bara, Binjai, Deli Serdang, Labuanbatu Utara, Labuhan Batu, South Labuanbatu, Langkat, Nias, West Nias, North Nias, Padangsidempuan, Serdang Bedagai, Simalungun, Tanjungbalai , Tapanuli Selatan, Tapanuli Tengah, Tapanuli Utara and Toba Samosir. Whereas the low cluster consists of 10 districts / cities including Asahan, Dairi, Gunungsitoli, Humbang Hasundutan, mandailing Natal, South Nias, Padang Lawas, North Padang Lawas, Pakpak Bharat and Samosir.

References

[1] D. V. Ferezagia, "Analisis Tingkat Kemiskinan di Indonesia Jurnal Sosial Humaniora Terapan," *Jurnal Sosial Humaniora Terapan*, vol. 1, no. 1, pp. 1–6, 2018.

[2] B. S. D. Bappeda DIY, "Laporan Akhir Analisis Kriteria Dan Indikator Kemiskinan Multidimensi Untuk Diagnostik Kemajuan Daerah Di Daerah Istimewa Yogyakarta," *Kerjasama Balai Statistik Daerah Bappeda DIY dengan Badan Pusat Statistik (BPS) Provinsi DIY 2017*, pp. 1–97, 2017.

[3] BPS, "II. Garis Kemiskinan (GK)," *Badan Pusat Statistik Indonesia*, 2020. [Online]. Available: https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html. [Accessed: 01-Apr-2020].

[4] A. Soleh, "Analisis dan Strategi Pengentasan Kemiskinan Di Provinsi Jambi," *Eksis: Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 9, no. 1, p. 79, 2018.

[5] A. Firdaus, S. Martha, and N. Imro'ah, "Penentuan Garis Kemiskinan Provinsi Menggunakan Metode Multiple Classification Analysis," *Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)*, vol. 08, no. 4, pp. 789–798, 2019.

[6] Wikipedia, "Garis Kemiskinan," 2020. [Online]. Available: https://id.wikipedia.org/wiki/Garis_kemiskinan. [Accessed: 01-Apr-2020].

[7] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm," *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–6, Aug. 2019.

[8] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject," *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–7, 2019.

[9] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. Perdana Windarto, "Application of Classification Method C4.5 on Selection of Exemplary Teachers," *Journal of Physics: Conference Series*, vol. 1235, no. 1, pp. 1–7, Jun. 2019.

[10] I. Parlina *et al.*, "Naive Bayes Algorithm Analysis to Determine the Percentage Level of visitors the Most Dominant Zoo Visit by Age Category," *Journal of Physics: Conference Series*, vol. 1255, no. 1, pp. 1–5, 2019.

[11] D. Hartama, A. Perdana Windarto, and A. Wanto, "The Application of Data Mining in Determining Patterns of Interest of High School Graduates," *Journal of Physics: Conference Series*, vol. 1339, no. 1, pp. 1–6, 2019.

[12] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," *IOP Conference Series: Materials Science and Engineering*, vol. 420, no. 1, pp. 1–8, 2018.

[13] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, "Classification of Natural Disaster Prone Areas in Indonesia using K-Means," *International Journal of Grid and Distributed Computing*, vol. 11, no. 8, pp. 87–98, 2018.

[14] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, "Using K-Means Clustering to Cluster Provinces in Indonesia," *Journal of Physics: Conference Series*, vol. 1028, no. 1, pp. 1–6, 2018.

[15] BPS, "Garis Kemiskinan Menurut Kabupaten/Kota (rupiah/kapita/bulan), 2017-2019," *Badan Pusat Statistik Provinsi Sumatera Utara*, 2020. [Online]. Available: https://sumut.bps.go.id/indicator/23/115/1/garis-kemiskinan-menurut-kabupaten-kota.html. [Accessed: 01-Apr-2020].

[16] Z. S. Younus *et al.*, "Content-based image retrieval using PSO and k-means clustering algorithm," *Arabian Journal of Geosciences*, vol. 8, no. 8, pp. 6211–6224, 2015.

[17] A. Wanto *et al.*, *Data Mining : Algoritma dan Implementasi*. Yayasan Kita Menulis, 2020.
[18] E. Prasetyo, *Data Mining: Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset, 2012.
[19] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*. New Jersey: John Wiley & Sons, 2005.
[20] R. Primartha, *Belajar Machine Learning Teori dan Praktik*. Bandung: Informatika Bandung, 2018.
[21] T. Khotimah, "Pengelompokan Surat dalam Al Qur'an menggunakan Algortima K-Means," *Jurnal Simetris*, vol. 5, no. 1, pp. 83–88, 2014.
[22] I. Parlina, A. P. Windarto, A. Wanto, and M. R. Lubis, "Memanfaatkan Algoritma K-Means dalam Menentukan Pegawai yang Layak Mengikuti Asessment Center untuk Clustering Program SDP," *CESS (Journal of Computer Engineering System and Science)*, vol. 3, no. 1, pp. 87–93, 2018.

Authors

1st Author
Mhd Ali Hanafiah
Lecturer of Politeknik Bisnis Indonesia, Pematangsiantar, Indonesia
ikh.alie84@gmail.com

2nd Author
Anjar Wanto
Lecturer of STIKOM Tunas Bangsa, Pematangsiantar, Indonesia.
anjarwanto@amiktunasbangsa.ac.id