Journal Mantik , 6 (3) (2022), ISSN 2685-4236 (Online)
Published by:Institute of Computer Science (IOCS)
Jurnal Mantik
Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Analysis of Silhouette Coefficient Evaluation with Euclidean Distance in the Clustering Method (Case Study: Number of Public Schools in Indonesia)

Dedy Hartama[1], Mawaddah Anjelita[2]
[1]Informatics Technique, STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra, Indonesia
[2]Information System, STIKOM Tunas Bangsa, Pematangsiantar, North Sumatra, Indonesia

ARTICLEINFO

Article history:

Received Oct 19, 2022
Revised Oct 26, 2022
Accepted Nov 16, 2022

Keywords:

Datamining,
Clustering,
K-Means,
Silhouette Coefficient,
Euclidean Distance

ABSTRACT

This study aims to find out how much the silhouette coefficient value is obtained from calculating the distance from the data to the centroid using the euclidean distance in the k-means method and to provide input in science for further research in developing the k-means method. To solve this problem, researchers use the k-means method with silhouette coefficient evaluation. Where the data source in this study took data directly from the Indonesian Central Bureau of Statistics (BPS) in the form of a softcopy entitled "Statistics of Indonesia 2021" with the URL: https://www.bps.go.id. The data used in this study uses 2020 data which consists of 34 provinces. The data will be processed using the k-means method with the silhouette coefficient using the euclidean distance. The results obtained are cluster = 4 which is the best cluster for classifying the number of public schools in Indonesia by province in 2020 with a silhouette coefficient of -0.9944. By doing research, it can provide input in science for further research in developing clustering methods, especially k-means.

Corresponding Author:

Dedy Hartama,
Informatics Technique,
STIKOM Tunas Bangsa,
Jl. Jendral Sudirman Blok A, No.1,2 & 3 Pematangsiantar Sumatera Utara – Indonesia
Email: dedyhartama@amiktunasbangsa.ac.id

1. Introduction

One of the efforts so that the next generation has high knowledge and skills that can function in social life requires what is called education where education starts from elementary school to university. Learning facilities are very important in the learning process in supporting teaching activities and can generate interest and attention from students to facilitate the delivery of learning activities [1]. The number of schools that are appropriate and good can facilitate students in the process of learning activities so that they can achieve maximum results. In reality, there are still many regions in Indonesia that do not yet have an adequate number of schools with the required capacity, especially state schools. Therefore, we need a technique that can cluster the number of public schools in Indonesia.

Artificial Intelligence (AI) has many methods that can be applied to everyday life, including decision support systems [2], Fuzzy Logic [3], data mining [4], and artificial neural networks [5]. One of the problem-solving methods used in this research is data mining. The main goal of data mining is to utilize data in the database by processing it to produce useful new information [6].

Datamining has several tasks that can be performed in the process of solving problems and finding new knowledge including clustering, estimation, classification, prediction, and association. The clustering method is a method of grouping several data into certain data groups (clusters) [7]. This study uses a clustering technique with the k-means method. K-means partitions data into groups so that data with the same characteristics are put into the same group and data with different characteristics are grouped into another group [8]. The advantages of the k-means method include the simplest and most common clustering method, due to the ability of k-means to group large amounts of data with a relatively fast and efficient processing time [9]. Behind the advantages of k-means, there are several weaknesses including determining the initial center of the cluster. The cluster results formed from the k-means method are highly dependent on the initiation of the initial cluster centre value given [10]. Basically, k-means do not have a certain standard to determine how many clusters will be formed. Determining the number of clusters currently usually only uses 2 clusters, namely high and low, or 3 clusters, namely high, normal and low.

One technique that can be used to maximize the k-means method in forming or determining the number of clusters is the silhouette coefficient. The advantages of the silhouette coefficient include evaluating and validating the quality of clustering by testing how far apart the clusters are and how dense the clusters are [11], measuring how close or far a relationship is between separate objects and other clusters [12], calculating the value of objects that are in a cluster [13]. With this silhouette coefficient, the k-means method can have a value for evaluating the number of clusters that will later be formed. The highest silhouette coefficient value can be used for clusters. This method has 4 distance measures in terms of comparison, namely Euclidean distance, minkowski distance, jaccard, and cosine distance.

This study used the silhouette coefficient method with euclidean distance using k-means. Silhouette coefficient using euclidean distance will be used as a solution to determine the distance to the centroid value which is better in iterations and to find out how much the silhouette coefficient value is obtained from calculating the data distance to the centroid using euclidean distance. The silhouette coefficient was chosen because it has the ability to evaluate the value of the number of clusters that will later be formed. Thus, the use of the silhouette coefficient with the euclidean distance in classifying the number of public schools in Indonesia is able to produce the best clusters. It is hoped that this research will become input in science to further research in developing clustering methods, especially k-means

Datamining is useful for getting useful information from large database warehouses. Data mining can be referred to as the process of finding correlations or patterns of hundreds or thousands of fields from a large relational database. One part of the Knowledge Discovery in Databases (KDD) process is where data mining oversees extracting patterns or models from data using a specific method [14]. Based on the tasks that can be performed, data mining is divided into 6 namely descriptions, estimation, prediction, classification, cluster, and association [15].

Clustering analyzes a set of objects and classifies objects into a cluster based on similarity values. The approach used in clustering is an unsupervised approach (unsupervised learning) where this approach does not require an output target. There are two methods that belong to clustering, namely hierarchical and non-hierarchical. The hierarchical method generated from the cluster will form a hierarchy starting from the most similar data to dissimilar data. The non-hierarchical method of the number of

groups is determined at the beginning while the hierarchical method is determined at the final stage of the analysis [16].

K-means method, The k-means method is the method most often used in grouping data. K-means divides data into several groups where data that has the same characteristics are grouped into the same cluster and data that has different characteristics are grouped into other groups.

The silhouette coefficient is an evaluation method to calculate cluster quality. This method combines cohesion and separation methods. Cohesion is used to measure the closeness of the relationship between one data and another in one group. While separation is used to measure how far the relationship between data is in one group [17]. The following is the process carried out in the silhouette coefficient method [13]

2. RESEARCH METHOD

2.1 Location and Time o Research

Location and Time of Research, This research was conducted in Indonesia by taking data directly from the Indonesian Central Bureau of Statistics (BPS) in softcopy entitled "Statistik Indonesia 2021" with the URL: https://www.bps.go.id and the data collection time was carried out for 5 days, ie, from March 15th to March 20th, 2021

2.2 Research Layouts

At this stage, there are several steps that will be carried out. The research design is shown in Figure 1

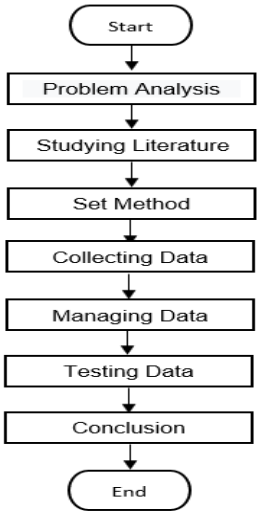

Figure 1. Research Layout

The figure explains the design of the research conducted to find the silhouette coefficient value for the number of public schools in Indonesia using k-means consisting of:

a. Problem analysis. Problems related to finding silhouette coefficient values to maximize the number of clusters to form the number of public schools in Indonesia. Values that are close to number 1 are the best values for the silhouette coefficient. The data used in this research is from 2020.
b. Study Literature. To obtain information in this study, researchers must obtain referrals.

c. Set Method. To solve the problem in this study the method used is k-means with the silhouette coefficient using the euclidean distance.
d. Collecting data. Researchers collected data for 5 days, from March 15th to March 20th, 2021. The Indonesian Central Bureau of Statistics (BPS) in softcopy entitled "Statistics of Indonesia 2021" with the URL link: https://www.bps.go.id
e. Manage Data. Perform data processing using the k-means method with the silhouette coefficient using the euclidean distance. Process of data management using Microsoft Excel 2010.
f. Test Data. Data testing in this study was carried out using the RapidMiner application version 5.3.
g. Conclusion. The results provided by this study can provide input in science for further research in developing clustering methods, especially k-means.

3. RESULTS AND DISCUSSIONS

The data used in this study is data on the number of public schools in Indonesia by 34 provinces in 2020. In this study, an analysis was first carried out on the number of clusters which would later be grouped using the k-means method. The data set for the number of public schools in Indonesia by province for 2020 can be seen in the following table:

Table 1. Dataset of the number of public schools in Indonesia

No	province	SD	Junior High School	Senior High School	SMK	University
1	Aceh	3336	895	395	151	7
2	North Sumatra	8282	1326	427	268	3
3	West Sumatra	3981	676	236	114	5
4	Riau	3200	854	303	126	2
5	Jambi	2314	556	161	104	1
6	South Sumatra	4292	901	328	114	2
7	Bengkulu	1304	381	109	64	2
8	Lampung	4356	706	238	110	3
9	Bangka Belitung	760	161	44	36	2
10	Riau islands	683	233	91	35	2
11	DKI Jakarta	1451	293	117	73	4
12	West Java	17492	1940	511	288	12
13	Central Java	17658	1769	360	237	9
14	In Yogyakarta	1427	214	69	50	4
15	East Java	17197	1726	423	297	17
16	Banten	3954	566	152	81	2
17	Bali	2306	272	83	53	4
18	West Nusa Tenggara	3011	604	154	99	1
19	East Nusa Tenggara	3348	1331	350	145	4
20	West Kalimantan	4130	1013	266	107	4
21	Central Kalimantan	2417	705	181	92	1
22	South Kalimantan	2771	522	137	63	3
23	East Kalimantan	1653	442	142	87	5
24	North Kalimantan	435	150	42	18	1
25	North Sulawesi	1361	473	121	90	4
26	Central Sulawesi	2669	726	175	106	1
27	South	6085	1265	335	168	4

No	province	SD	Junior High School	Senior High School	SMK	University
28	Sulawesi Southeast Sulawesi	2253	690	240	99	2
29	Gorontalo	896	313	60	40	1
30	West Sulawesi	1298	314	75	59	1
31	Maluku	1261	529	209	81	3
32	North Maluku	1102	355	138	63	1
33	Papuan	677	224	77	32	2
34	West Papua	1612	499	140	78	3

3.1 Dataset Normalization

Normalization is done to change a value on a scale of 0 – 1. From the dataset used on the number of public schools in Indonesia new information is obtained, namely in table 3 below:

Table 2. Maximum and Minimum Attribute Values

Attributes	Minimum Value	Maximum Value
Sd	435	17658
Junior High School	150	1940
Senior High School	42	511
Smk	18	297
University	1	17

For Aceh data with attributes:

SD = 3336, SMP = 895, SMA = 395, SMK = 151, University = 7

Aceh SD $=\frac{3336-435}{17658-435}$ =0.1684

Aceh High School $=\frac{395-42}{511-42}$ =0.7527

Aceh Middle School $=\frac{895-150}{1940-150}$ =0.4162

Aceh SMK $=\frac{151-18}{297-18}$ =0.4767

Aceh University $=\frac{7-1}{17-1}$ =0.3750

From the normalization calculations for the Aceh data, then the data is entered into the table and the next data normalization calculation is carried out in the same way. The normalization results of the data can be seen in table 4 below:

Table 3. Dataset Normalization

No	Province	SD	Junior High School	Senior High School	SMK	University
1	Aceh	0.1684	0.4162	0.7527	0.4767	0.3750
2	North Sumatra	0.4556	0.6570	0.8209	0.8961	0.1250
3	West Sumatra	0.2059	0.2939	0.4136	0.3441	0.2500
4	Riau	0.1605	0.3933	0.5565	0.3871	0.0625
5	Jambi	0.1091	0.2268	0.2537	0.3082	0
6	South Sumatra	0.2239	0.4196	0.6098	0.3441	0.0625
7	Bengkulu	0.0505	0.1291	0.1429	0.1649	0.0625
8	Lampung	0.2277	0.3106	0.4179	0.3297	0.1250
9	Bangka Belitung	0.0189	0.0061	0.0043	0.0645	0.0625
10	Riau Islands	0.0144	0.0464	0.1045	0.0609	0.0625
11	DKI Jakarta	0.0590	0.0799	0.1599	0.1971	0.1875
12	West Java	0.9904	1	1	0.9677	0.6875
13	Central Java	1	0.9045	0.6780	0.7849	0.5000
14	In Yogyakarta	0.0576	0.0358	0.0576	0.1147	0.1875
15	East Java	0.9732	0.8804	0.8124	1	1

No	Province	SD	Junior High School	Senior High School	SMK	University
16	Banten	0.2043	0.2324	0.2345	0.2258	0.0625
17	Bali	0.1086	0.0682	0.0874	0.1254	0.1875
18	West Nusa Tenggara	0.1496	0.2536	0.2388	0.2903	0.0000
19	East Nusa Tenggara	0.1691	0.6598	0.6567	0.4552	0.1875
20	West Kalimantan	0.2145	0.4821	0.4776	0.3190	0.1875
21	Central Kalimantan	0.1151	0.3101	0.2964	0.2652	0
22	South Kalimantan	0.1356	0.2078	0.2026	0.1613	0.1250
23	East Kalimantan	0.0707	0.1631	0.2132	0.2473	0.2500
24	North Kalimantan	0	0	0	0	0
25	North Sulawesi	0.0538	0.1804	0.1684	0.2581	0.1875
26	Central Sulawesi	0.1297	0.3218	0.2836	0.3154	0
27	South Sulawesi	0.3280	0.6229	0.6247	0.5376	0.1875
28	Southeast Sulawesi	0.1056	0.3017	0.4222	0.2903	0.0625
29	Gorontalo	0.0268	0.0911	0.0384	0.0789	0
30	West Sulawesi	0.0501	0.0916	0.0704	0.1470	0
31	Maluku	0.0480	0.2117	0.3561	0.2258	0.1250
32	North Maluku	0.0387	0.1145	0.2047	0.1613	0
33	Papuan	0.0141	0.0413	0.0746	0.0502	0.0625
34	West Papua	0.0683	0.1950	0.2090	0.2151	0.1250

3.2 Clustering with Euclidean Distance

First determine 6 random numbers from the data that has been normalized, then obtained:

*Clusters*1 : 1 *Clusters*2 : 0.1250 *Clusters*3 : 0.3441

*Clusters*4 : 0.6780 *Clusters*5 : 0.8804 *Clusters*6 : 0

Calculate the distance of the data to the centroid by minimizing the distance through iterations. For Aceh data with cluster 1:

Table 4. Number of public schools in Aceh

Province	Sd	Junior High School	Senior High School	Smk	University
Aceh	0.1684	0.4162	0.7527	0.4767	0.3750

$$= \sqrt{\sum_{i=1}^{5}(1-0{,}1684)^2 + (1-0{,}4162)^2 + (1-0{,}7527)^2 + (1-0{,}4767)^2 + (1-0{,}3750)^2} = 1.3259$$

For Aceh data with cluster 2:

$$= \sqrt{\sum_{i=1}^{5} (0{,}125-0{,}1684)^2 + (0{,}125-0{,}4162)^2 + (0{,}125-0{,}7527)^2 + (0{,}125-0{,}4767)^2 + (0{,}125-0{,}3750)^2} = 0.8166$$

For Aceh data with cluster 3:

$$= \sqrt{\sum_{i=1}^{5} (0{,}3441-0{,}1684)^2 + (0{,}3441-0{,}4162)^2 + (0{,}3441-0{,}7527)^2 + (0{,}3441-0{,}4767)^2 + (0{,}3441-0{,}3750)^2} = 0.4707$$

For Aceh data with cluster 4:

$$= \sqrt{\sum_{i=1}^{5} (0{,}6780-0{,}1684)^2 + (0{,}6780-0{,}4162)^2 + (0{,}6780-0{,}7527)^2 + (0{,}6780-0{,}4767)^2 + (0{,}6780-0{,}3750)^2} = 0.6828$$

For Aceh data with cluster 5:

$$= \sqrt{\sum_{i=1}^{5} \begin{matrix}(0{,}8804-0{,}1684)^2+(0{,}8804-0{,}4162)^2+(0{,}8804-0{,}7527)^2 \\ +(0{,}8804-0{,}4767)^2+(0{,}8804-0{,}3750)^2\end{matrix}} = 1.0758$$

For Aceh data with cluster 6:

$$= \sqrt{\sum_{i=1}^{5}(0-0{,}1684)^2+(0-0{,}4162)^2+(0-0{,}7527)^2+(0-0{,}4767)^2+(0-0{,}3750)^2} = 1.0658$$

The calculation is carried out until the West Papua cluster 6 data. The calculation results obtained the distance between the data and the clusters and is entered in table 6 below:

Table 5. Euclidean Distance Calculation Results

No	province	C1	C2	C3	C4	C5	C6
1	aceh	1.3259	0.8166	0.4707	0.6828	1.0758	1.0658
2	North Sumatra	1.1057	1.2129	0.8309	0.6509	0.8972	1.4600
3	West Sumatra	1.5703	0.4266	0.1879	0.8574	1.3047	0.6934
4	Riau	1.5887	0.5762	0.4030	0.9095	1.3315	0.8024
5	Jambi	1.8512	0.2764	0.4436	1.1418	1.5867	0.4720
6	South Sumatra	1.5494	0.6193	0.4123	0.8764	1.2936	0.8487
7	Bengkulu	1.9928	0.1067	0.5333	1.2743	1.7259	0.2659
8	Lampung	1.6203	0.4156	0.2614	0.9125	1.3560	0.6688
9	Bangka Belitung	2.1670	0.2180	0.7021	1.4475	1.8998	0.0921
10	Riau islands	2.1080	0.1639	0.6436	1.3886	1.8408	0.1445
11	DKI Jakarta	1.9346	0.1293	0.4807	1.2171	1.6679	0.3308
12	West Java	0.3143	1.8184	1.3360	0.6236	0.2924	2.0952
13	Central Java	0.6396	1.5017	1.0366	0.4450	0.4579	1.7731
14	In Yogyakarta	2.0372	0.1451	0.5800	1.3193	1.7704	0.2371
15	East Java	0.2241	1.8149	1.3279	0.5945	0.2045	2.0934
16	Banten	1.8129	0.2095	0.3705	1.0969	1.5466	0.4535
17	Bali	1.9801	0.0939	0.5194	1.2613	1.7131	0.2738
18	West Nusa Tenggara	1.8339	0.2703	0.4224	1.1236	1.5691	0.4775
19	East Nusa Tenggara	1.3715	0.8268	0.5147	0.7417	1.1250	1.0666
20	West Kalimantan	1.5106	0.5490	0.2807	0.8142	1.2489	0.8022
21	Central Kalimantan	1.8150	0.3146	0.4248	1.1082	1.5511	0.5173
22	South Kalimantan	1.8654	0.1196	0.4043	1.1464	1.5983	0.3798
23	East Kalimantan	1.8199	0.2068	0.3779	1.1040	1.5536	0.4480
24	North Kalimantan	2.2361	0.2795	0.7694	1.5161	1.9687	0
25	North Sulawesi	1.8626	0.1779	0.4169	1.1463	1.5962	0.4069
26	Central Sulawesi	1.7887	0.3403	0.4115	1.0837	1.5253	0.5480
27	South Sulawesi	1.2683	0.8444	0.4677	0.6235	1.0171	1.1001
28	Southeast Sulawesi	1.7332	0.3888	0.3834	1.0315	1.4706	0.6071
29	Gorontalo	2.1323	0.1899	0.6685	1.4130	1.8651	0.1292
30	West Sulawesi	2.0783	0.1607	0.6183	1.3599	1.8114	0.1935
31	Maluku	1.8186	0.2775	0.4091	1.1084	1.5538	0.4904
32	North Maluku	2.0110	0.1756	0.5632	1.2950	1.7447	0.2873
33	Papuan	2.1280	0.1771	0.6625	1.4084	1.8608	0.1179
34	West Papua	0.3850	2.6046	1.4682	0.7009	1.1704	2.0462

3.3 Silhouette Coefficient with Euclidean Distance

The steps for calculating the silhouette coefficient are as follows: Calculate the average distance to all other objects in the cluster, then called ai.

$$a(i) = \frac{\sum D\,(i,j)}{|A|-1} \quad (5)$$

A = Total clusters

a(1) =(1.3259 + 1.1057 + 1.5703 + 1.5887 + 1.8512 + 1.5494 + 1.9928 + 1.6203 + 2.1670 + 2.1080 + 1.9346 + 0.3143 + 0 .6396 + 2.0372 + 0.2241 + 1.8129 + 1.9801 + 1.8339 + 1.3715 + 1.5106 + 1.8150 + 1.8654 + 1.8199 + 2.2361 + 1.8626 + 1.7887 + 1.2683 + 1.7332 + 2.1323 + 2.0783 + 1.8186 + 2.0110 + 2.1280 + 0.3850) / 0.2241

= 247.5813

a(2)= 196.4640

a(3)= 105.3836

a(4)= 79.5984
a(5)= 235.6451
a(6)= 23.3886

Calculate the average distance to all other objects in another cluster, which is then called bi. Calculate the silhouette coefficient

$$s(1) = \frac{0{,}2241 - 247{,}5813}{247{,}5813} = -0.9991 \qquad s(4) = \frac{0{,}4450 - 79{,}5984}{79{,}5984} = -0.9944$$

$$s(2) = \frac{0{,}0939 - 196{,}4640}{196{,}4640} = -0.9995 \qquad s(5) = \frac{0{,}2045 - 235{,}6451}{235{,}6451} = -0.9991$$

$$s(3) = \frac{0{,}1879 - 105{,}3836}{105{,}3836} = -0.9982 \qquad s(6) = \frac{0 - 23{,}3886}{23{,}3886} = -1{,}000$$

The average value of the silhouette coefficient for each point is the best, that is, where it is closer to number 1.

Table 6. The silhouette coefficient value of clustering results

CLUSTER	SC Value	Description
s(1)	-0.9991	*No. Structure*
s(2)	-0.9995	*No. Structure*
s(3)	-0.9982	*No. Structure*
s(4)	-0.9944	*No. Structure*
s(5)	-0.9991	*No. Structure*
s(6)	-1.0000	*No. Structure*

Based on table 7 it can be seen that cluster 4 is the best cluster in classifying the number of public schools in Indonesia.

3.4 Test results

RapidMiner has a lot of modelling that can be used to test especially the methods included in data mining. In this study the model used is k-means. Next, we will select the model to be used. On the modelling, menu select clustering and segmentation, then select k-means. connect or connect the read excel model with clustering modelling (k-means) then at the top right, in option k select 4 then run by clicking the play button (blue) on the top centre. More details can be seen in Figure 2 below:

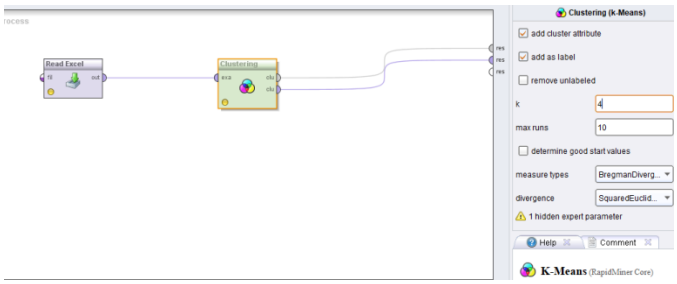

Figure 2. New Process Final Display

So that the results of grouping the number of public schools in Indonesia in 2020 according to provinces with 4 clusters can be seen in figure 3

ExampleSet (34 examples, 2 special attributes, 6 regular attributes)

Row No.	Provinsi	label	No	SD	SMP	SMA	SMK	Universitas
1	ACEH	cluster_2	1	3336	895	395	151	7
2	SUMATERA UTARA	cluster_3	2	8282	1326	427	268	3
3	SUMATERA BARAT	cluster_2	3	3981	676	236	114	5
4	RIAU	cluster_2	4	3200	854	303	126	2
5	JAMBI	cluster_2	5	2314	556	161	104	1
6	SUMATERA SELATAN	cluster_2	6	4292	901	328	114	2
7	BENGKULU	cluster_0	7	1304	381	109	64	2
8	LAMPUNG	cluster_2	8	4356	706	238	110	3
9	BANGKA BELITUNG	cluster_0	9	760	161	44	36	2
10	KEPULAUAN RIAU	cluster_0	10	683	233	91	35	2
11	DKI JAKARTA	cluster_0	11	1451	293	117	73	4
12	JAWA BARAT	cluster_1	12	17492	1940	511	288	12
13	JAWA TENGAH	cluster_1	13	17658	1769	360	237	9
14	DI YOGYAKARTA	cluster_0	14	1427	214	69	50	4
15	JAWA TIMUR	cluster_1	15	17197	1726	423	297	17
16	BANTEN	cluster_2	16	3954	566	152	81	2
17	BALI	cluster_2	17	2306	272	83	53	4
18	NUSA TENGGARA BARAT	cluster_2	18	3011	604	154	99	1
19	NUSA TENGGARA TIMUR	cluster_2	19	3348	1331	350	145	4

Figure 3. Clustering results with RapidMiner part 1

The following is a continuation of the results of grouping the number of public schools in Indonesia in 2020 according to provinces with 4 clusters.

Row No.	Provinsi	label	No	SD	SMP	SMA	SMK	Universitas
16	BANTEN	cluster_2	16	3954	566	152	81	2
17	BALI	cluster_2	17	2306	272	83	53	4
18	NUSA TENGGARA BARAT	cluster_2	18	3011	604	154	99	1
19	NUSA TENGGARA TIMUR	cluster_2	19	3348	1331	350	145	4
20	KALIMANTAN BARAT	cluster_2	20	4130	1013	266	107	4
21	KALIMANTAN TENGAH	cluster_2	21	2417	705	181	92	1
22	KALIMANTAN SELATAN	cluster_2	22	2771	522	137	63	3
23	KALIMANTAN TIMUR	cluster_0	23	1653	442	142	87	5
24	KALIMANTAN UTARA	cluster_0	24	435	150	42	18	1
25	SULAWESI UTARA	cluster_0	25	1361	473	121	90	4
26	SULAWESI TENGAH	cluster_2	26	2669	726	175	106	1
27	SULAWESI SELATAN	cluster_3	27	6085	1265	335	168	4
28	SULAWESI TENGGARA	cluster_2	28	2253	690	240	99	2
29	GORONTALO	cluster_0	29	896	313	60	40	1
30	SULAWESI BARAT	cluster_0	30	1298	314	75	59	1
31	MALUKU	cluster_0	31	1261	529	209	81	3
32	MALUKU UTARA	cluster_0	32	1102	355	138	63	1
33	PAPUA	cluster_0	33	677	224	77	32	2
34	PAPUA BARAT	cluster_0	34	1612	499	140	78	3

Figure 4. Clustering results with RapidMiner part 2

The cluster model can be seen in Figure 5 below:

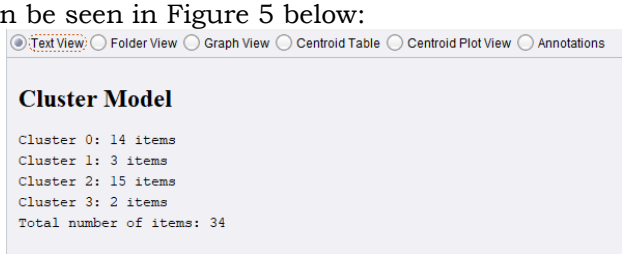

Figure 5. Clustering results with RapidMiner part 3

The grouping results in cluster 1 are Bengkulu, Bangka Belitung, Riau Islands, DKI Jakarta, Yogyakarta Special Region, East Kalimantan, North Kalimantan, North Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, Papua and West Papua. For cluster 2, namely the provinces of West Java, Central Java and East Java. For cluster 3 there are the provinces of Aceh, West Sumatra, Riau, Jambi, South Sumatra, Lampung, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, Central Sulawesi and Southeast Sulawesi. For cluster 4, there are provisions for North Sumatra and South Sulawesi.

4. CONCLUSION

Based on the discussions that have been done it can be concluded that: The silhouette coefficient with euclidean distance can be applied to evaluate the k-means clustering method in the case of the number of public schools in Indonesia in 2019-2020 by province. Based on the calculation results on the silhouette coefficient with the Euclidean distance of the 6 clusters used, it is found that cluster = 4 is the best cluster for grouping data in the case of the number of public schools in Indonesia in 2020 by province with a silhouette coefficient value of -0.9944. Testing using the RapidMiner software version 5.3 using cluster = 4 to group data from 34 provinces, the results obtained are cluster 1 of 14 provinces, cluster 2 of 3 provinces, cluster 3 of 15 provinces, and cluster 4 of 2 provinces.

REFERENCES

[1] L. Maftuhatin and HZ Rosyid, "The Influence of Parents' Socioeconomic Status and Students' Perceptions of School Facilities on Student Learning Outcomes," Vol. 3, No. 1, 2019.
[2] T. Imandasari, A. Wanto, And AP Windarto, "Analysis of Decision Making in Determining Pkl Students Using the Promethee Method," J. Ris. Comput., Vol. 5, No. 3, pp. 234–239, 2018.
[3] DA Silitonga, M. Anjelita, And AP Windarto, "Fuzzy Inference System on Prediction of Pertamax Fuel Purchases at gas stations in Pematangsiantar City," Syntax J. Inform., Vol. 8, No. 2, P. 75, 2019, Doi: 10.35706/Syji.V8i2.1841.
[4] K. Handoko, "Application of Data Mining in Improving the Quality of Learning in Higher Education Institutions Using the K - Means Clustering Method (Case Study in the Tkj Study Program at the South Solok Community Academy)," Teknosi, Vol. 2, No. 3, pp. 31–40, 2016.
[5] AP Windarto, MR Lubis, And Solikhun, "Neural Network Architectural Models with Backpropogation in Total Profit Prediction," Gathered. J. Computing Science., Vol. 5, No. 2, pp. 147–158, 2018.
[6] MRL Iin Parlina, Agus Perdana Windarto, Anjar Wanto, "Utilizing the K-Means Algorithm in Determining Eligible Employees to Take the Assessment Center," Utilizing the Algorithm. K-Means In Determining Eligible Employees To Follow Asessment Cent. For Clusts. program SDP, Vol. 3, No. 1, pp. 87–93, 2018.
[7] A. Aditya, I. Jovian, and BN Sari, "Implementation of K-Means Clustering for National Junior High School Examinations in Indonesia in 2018/2019," J. Media Inform. Budidarma, Vol. 4, pp. 51–58, 2020, Doi: 10.30865/Mib.V4i1.1784.
[8] MG Sadewo, AP Windarto, And D. Hartama, "Application of Datamining in Broiler Populations in Indonesia by Province Using K-Means Clustering," Infotekjar (Jurnal Nas. Inform. And Teknol. Network), Vol. 2, No. 1, pp. 60–67, 2017, Doi: 10.30743/Infotekjar.V2i1.164.
[9] H. Sulastri And AI Gufroni, "Application of Data Mining in Grouping Patients with Thalassemia," J. Nas. Technol. And Sist. Inf., Vol. 3, No. 2, pp. 299–305, 2017, Doi: 10.25077/Teknosi.V3i2.2017.299-305.
[10] S. Butsianto And NT Mayangwulan, "Application of Data Mining to Predict Car Sales Using the K-Means Clustering Method," J. Nas. Computing And Technology. Inf., Vol. 3, No. 3, pp. 187–201, 2020.
[11] AP Tiratana, B. Mulyawan, And MD Lauro, "Development of Web-Based Customer Relationship Management Applications Using the K-Means Method," J. Computer Science. And Sist. Inf., Pp. 179–184, 2019.
[12] S. Asmiatun, N. Wakhidah, AN Putri, U. Semarang, and K. Jalan, "Application of the K-Medoids Method for Classifying Road Conditions in the City of Semarang 1.2," J. Tek. inform. And Sist. Inf., Vol. 6, No. 2, 2020.
[13] EF Sirat, BD Setiawan, And F. Ramdani, "A Comparative Analysis of the K-Means and Isodata Algorithms for Clustering Data on Hot Spots in the Sumatra Region in 2001 to 2014," J. Pemmb. Technol. inf. And Computing Science., Vol. 2, No. 11, pp. 5105–5112, 2018.
[14] N. Rahmawati, YN Nasution, And FDT Amijaya, "Dataminingmarket Basket Analysis Application to Find Purchase Patterns at the Main Metro Stores in Balikpapan," J. Exponential, Vol. 8, No. 2012, pp. 1–8, 2017.
[15] R. Setiawan, "Application of Data Mining Using the K-Means Clustering Algorithm to Determine New Student Promotion Strategies (Case Study: Polytechnic Lp3i Jakarta)," J. Lentera Ict, Vol. 3, No. 1, pp. 76–92, 2016.

[16] F. Selva And D. Pratama, "Identification of Clusters of Working Age Population in South Sumatra Province Using K-Modes," J. Politek. Caltex Riau, Vol. 4, No. 1, pp. 1–9, 2018.
[17] DP Sinaga, PP Adikara, And YA Sari, "Clusterization of Hot Spot Data Using the Self Organizing Map Method in the Java Region," J. Pemmb. Technol. inf. And Computing Science., Vol. 3, No. 10, pp. 9543–9551, 2019.
[18] K. Sari, "Analysis of Evaluating Distance Calculations for Silhouette Coefficient Values in the K-Means Algorithm," Thesis, 2020.
[19] M. Nishom, "Comparison of the Accuracy of Euclidean Distance, Minkowski Distance, and Manhattan Distance in the Chi-Square-Based K-Means Clustering Algorithm," Vol. 4, No. 1, pp. 20–24, 2019, Doi: 10.30591/Jpit.V4i1.1253.