Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia

Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia

Aang Munawar[1], Gen Gen Gendalasari[2], I Made Gede Ariestova Kurniawan[3], D Purnomo[4], Nur Haris Ependi[5], Rulinawaty[6], Muhammad Isa Indrawan[7], Muhammad Sadri[8]

[1]Institut Bisnis dan Informatika Kesatuan, Indonesia. Email: aangmun@yahoo.com
[2]Institut Bisnis dan Informatika Kesatuan, Indonesia. Email: gen2_sari@yahoo.co.id
[3]Universitas teknologi indonesia, Indonesia. Email: kurniawanariestova@gmail.com
[4]Politeknik LP3I Jakarta, Jakarta, Indonesia. Email: djumatpurnomo@gmail.com
[5]Politeknik LP3I Jakarta, Jakarta, Indonesia. Email: hrsbox@gmail.com
[6]Universitas Terbuka, Indonesia. Email: ruly@ecampus.ut.ac.id
[7]Universitas Pembangunan Panca Budi, Medan, Indonesia. Email: isaindrawan@dosen.pancabudi.ac.id
[8]STKIP Al Maksum, Indonesia. Email: soedrytheone@gmail.com

Abstract. The aim of this study is to develop a grouping model in order to determine the means of trade and accommodation according to the regions in Indonesia. Research can be a reference for the government to increasing the income of each region in Indonesia equally. Research data were taken from a website that provides government statistical data, namely BPS (*Badan Pusat Statistik*)-www.bps.go.id. The solution is to use data mining techniques with clustering methods. The data test process uses the Rapid Miner software. Three clusters of mapping labels are used, namely the high cluster (K1), the normal cluster (K2) and the low cluster (K3). The results of the rapidminer processing were obtained from the centroid data for high clusters, namely ((1527), (810.4), (5865), (6655.3), (323), (315.1); the medium cluster, namely ((286), (199.591), (1327), (2240.227), (93.227), (140.955)); and the low cluster, namely (139.25), (122.5), (508.833), (919.222), (64.417), (94.444)). The cluster results show that 5 provinces are classified as high in clusters; 13 provinces are classified as medium clusters; and 16 provinces are classified as low clusters. Out of the results of the study, some 47% of areas in Indonesia still have low trade and housing facilities. With this analysis, it is hoped that the government will be able to pay more attention to regions whose revenues are still below average.

Keywords: Datamining, Clustering, K-Means, Accommodation Facilities, Trading Facilities, Rapid Miner, Indonesia.

1. Introduction

Times are growing and each region competes to make its territory one of the best in Indonesia, not only in terms of size and small area of the region, but also in terms of how the region progresses and develops from a variety of aspects, including trade facilities and accommodation. These two facilities are one of the factors that will help the area to increase the income of its people. The greater the means of trade and accommodation, the higher the income from the region. These facilities can be made up

of shops, mini markets, inns and hotels. By building a number of facilities in an area, the area can develop better. In terms of the tourism industry, Indonesia is a country with a lot of natural beauty. Therefore, in order to anticipate the arrival of domestic and foreign tourists to Indonesia for business activities, it requires support facilities such as accommodation for lodging, one of which is a star hotel providing complete facilities for business meetings as well as recreation areas. The importance of trade and accommodation means that the government must be able to maximize the potential that exists in each region and then add trade and accommodation facilities that have an impact on increasing the income of the region. The aim of this research is to map areas in Indonesia with commercial and housing facilities.

There are several computer science methods that can be adopted, such as the utilization of data mining techniques [1]–[4]. This is a technique for gathering information through statistical models. Using mapping in research for commercial and housing facilities can help identify geographical conditions [5]–[8]. Various clustering techniques are known for their extraordinarily rapid computational capability [9] [10] such as the k-means method [11]. The K-means method is more effective with a limited number of inputs [6]. There have been previous studies using the k-means method to solve problems regarding the water delivery system [5]. This study showed that the k-means method can be used to decide a customer's class if 70% of Indonesians are still do not know the need for clean water. Based on this, it is hoped that the results of the research will provide a regional cluster mapping so that the results of the cluster can help the government to identify areas that still have low trade and housing facilities so that they can be monitored.

2. Methodology

Research data that were obtained from https://www.bps.go.id is needed for the usage of clustering techniques in area-based mapping of trade and accommodation facilities. The data used in this study are data on trade and accommodation facilities in 34 provinces in Indonesia. There are six attributes used, having a mini market, a restaurant, a tavern, a shop, a hotel, and a lodging. The data was analysed using a tool called cluster analysis. A flowchart of the cluster k-means methods flow is shown in Figure 1 above.

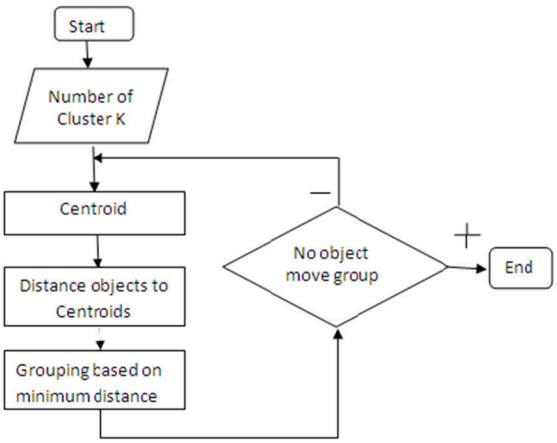

Figure 1. K-Means Clustering Flowchart

2.1. Data Mining

Data mining is [12] the process of extracting raw data into useful information. Companies can learn more about their customers by using software to look for patterns from large data sets. This allows business to develop more effective marketing strategies, increase sales and reduce costs. Data mining [13] requires an effective data collection, storage and processing. Data mining processes are used to build learning models that power applications, including Internet search technology and targeted Internet recommendations [14].

2.2. Clustering

Clustering is an important aspect of data mining. This is a method for different groups of similar data to be arranged by their measures of similarity. They are much more similar to each other than data items in other clusters. Unsupervised learning is an example of clustering [15].

2.3. K-Means Method

K-Means is [16]one of the most widely used clustering algorithms. This procedure helps to partition a set of n objects into k distinct groups so that the intra-group similarity is high and inter-group similarity is low [5], [17], [18]. This key idea is to extract centroids from clusters. To get these best results, the centroids should be placed in a cunning way. The better choice would be to push them away from each other [19]. K-means function of the algorithm is as follows: 1) Initialization by setting the initial coordinate system centers to infinity; 2). Group or cluster each data point of the given k clusters; 3). One method involves assigning data or objects to the nearest cluster centroid as a function of distance. When all objects have been assigned, re-calculate or adjust the position of the objects. Keep repeating this method until the centroids no longer move [9].

3. Results and Discussion

For cluster mapping, the processed data will be calculated using the k-means method by taking the average value of the six predetermined attributes, namely Mini Markets, Restaurants, Food and Beverage Shops, Shops, Hotels and Lodging, to make it easier to process as shown in Table 1 below.

Table 1. Average value of each attribute

Province	Trade Facility (Territory)				Accommodation Facilities (Territory)	
	Mini Markets	Restaurants	Food and Beverage Shops	Shops	Hotels	Lodging
Aceh	298	151	4193.5	5224.5	80	148.5
North Sumatra	575.5	572.5	4505.5	5224.5	227	269.5
West Sumatra	343.5	381	1129	1169.5	102.5	164.5
Riau	303	243	1406.5	1798	87.5	157.5
Jambi	182	179.5	1026	1515.5	67.5	58
South Sumatra	426.5	353.5	2086	3182	108.5	131
Bengkulu	106	155	808	1478	56.5	57.5
Lampung	508.5	320	2046.5	2631	95.5	105
Kep. Bangka Belitung	59.5	88.5	343	382.5	43.5	43
Kep. Riau	114	114	300.5	405	56	61.5
DKI Jakarta	259	227.5	264.5	264.5	135	83
West Java	2229.5	1299	5078	5894	408	441
Central Java	2143	997.5	7343.5	8508	476.5	327
In Yogyakarta	229	139	412.5	436.5	73.5	105.5
East Java	2389	1032	8204.5	8425.5	423.5	389.5
Banten	578	248.5	1132.5	1516	98	82
Bali	351	198	702.5	709.5	163.5	269
West Nusa Tenggara	245	146	739.5	1111.5	92	107
East Nusa Tenggara	100.5	94.5	416.5	2749.5	109.5	107
West Kalimantan	215	172.5	1057	2032	61	167
Central Kalimantan	74.5	68	785.5	1478.5	45.5	107
South Borneo	150	143.5	1752	1949.5	83	93
East Kalimantan	169.5	116.5	692	989.5	88	183.5
North Kalimantan	35.5	19.5	134	355	22.5	42
North Sulawesi	230	184	1196	1731.5	79	133
Central Sulawesi	86.5	140.5	1017.5	1897	61.5	202
South Sulawesi	465.5	216	1702.5	2924	149	236.5
Southeast Sulawesi	82.5	79.5	784	2232	93	136.5
Gorontalo	72.5	89	557	704.5	29.5	48.5
West Sulawesi	25	43	288.5	605.5	17	45
Maluku	46	62.5	206	978	30.5	96
North Maluku	58.5	64	238.5	1097.5	30.5	89
West Papua	44	31.5	184.5	937.5	39	68.5
Papua	92	82.5	347.5	1927.5	67	71.5

After processing the average value of each attribute as shown in Table 1, the next step is to process the data using the RapidMiner software using three cluster labels, namely high, medium and low clusters.

3.1. Centroid Data

In the k-means method, the determination of the cluster value is based on needs which are divided into three cluster labels, namely high (K1), medium (K2) and low (K3) labels. The K1 value is obtained from the maximum value of a record, the K2 value is obtained from the average value of a record, and the K3 value is obtained from the minimum value of a record. The following is the initial value of the cluster process using the k-means method as shown in Table 2 below:

Table 2. Initial Value of Centroid

	A	B	C	D	E	F
C1	2389	1299	8204.5	8508	476.500	441
C2	390.809	248.603	1561.206	2190.147	111.765	141.941
C3	25	19.5	134	264.5	17.000	42

Information :
A = Mini Market
B = Restaurant
C = Food and Beverage Shop
D = Shop
E = Hotel
F = Lodging

3.2. Clustering data

After the determination of 3 cluster labels from 3 different centroid values, the first iteration process will continue and continue until the result of the last cluster value = the previous cluster value. In this study, the iteration process ended in the fifth iteration with the final cluster and the centroid results as shown in the table below:

Table 3. Calculation of the cluster center distance in the last iteration

Province	high (K1)	medium (K2)	low (K3)	the Euclidean distance
Aceh	2621.670	4264.143	5760.114	2621.670
North Sumatra	2206.507	4517.810	6000.835	2206.507
West Sumatra	7360.710	1018.119	806.495	806.495
Riau	6735.874	410.295	1360.280	410.295
Jambi	7223.012	684.940	884.849	684.940
South Sumatra	5276.781	1353.904	2877.746	1353.904
Bengkulu	7413.761	821.322	735.953	735.953
Lampung	5670.067	984.163	2427.649	984.163
Kep. Bangka Belitung	8524.515	2001.549	471.224	471.224
Kep. Riau	8523.324	1994.051	456.170	456.170
DKI Jakarta	8616.538	2128.274	622.503	622.503
West Java	1397.948	5835.743	7268.269	1397.948
Central Java	2461.021	9058.795	10542.423	2461.021
In Yogyakarta	8405.058	1908.466	400.255	400.255
East Java	3068.331	9660.506	11114.368	3068.331
Banten	7080.272	727.679	1054.942	727.679
Bali	7986.834	1553.936	385.030	385.030
West Nusa Tenggara	7693.249	1158.465	402.666	402.666
East Nusa Tenggara	6897.661	1021.092	1932.971	1021.092
West Kalimantan	6829.764	225.196	1345.338	225.196
Central Kalimantan	7441.895	844.918	729.298	729.298
South Borneo	6442.758	578.578	1697.176	578.578
East Kalimantan	7826.753	1294.549	282.953	282.953
North Kalimantan	8692.130	2133.092	604.694	604.694
North Sulawesi	6943.389	431.590	1161.707	431.590
Central Sulawesi	6981.737	391.807	1207.496	391.807
South Sulawesi	5724.120	929.789	2459.932	929.789
Southeast Sulawesi	6934.532	497.370	1446.553	497.370
Gorontalo	8147.398	1620.814	165.102	165.102
West Sulawesi	8408.860	1841.658	330.314	330.314
Maluku	8194.040	1584.714	342.617	342.617
North Maluku	8086.837	1473.927	387.586	387.586
West Papua	8240.677	1632.330	352.194	352.194
Papua	7450.425	935.339	1119.264	935.339

Table 4. Mapping results are in the form of clusters in the last iteration

Province	high (K1)	medium (K2)	low (K3)
Aceh	✓		
North Sumatra	✓		
West Sumatra			✓
Riau		✓	
Jambi		✓	
South Sumatra		✓	
Bengkulu			✓
Lampung		✓	
Kep. Bangka Belitung			✓
Kep. Riau			✓
DKI Jakarta			✓
West Java	✓		
Central Java	✓		
In Yogyakarta			✓
East Java	✓		
Banten		✓	
Bali			✓
West Nusa Tenggara			✓
East Nusa Tenggara		✓	
West Kalimantan		✓	
Central Kalimantan			✓
South Borneo		✓	
East Kalimantan			✓
North Kalimantan			✓
North Sulawesi		✓	
Central Sulawesi		✓	
South Sulawesi		✓	
Southeast Sulawesi		✓	
Gorontalo			✓
West Sulawesi			✓
Maluku			✓
North Maluku			✓
West Papua			✓
Papua		✓	

3.3. Data Analysis

The results of the calculations using the k-means method were obtained from 34 provinces in which 5 provinces belong to the high cluster; 13 provinces are classified as medium clusters; and 16 provinces belong to the low cluster. According to research, around 47% of areas in Indonesia still have low trade and housing facilities. With this analysis, it is hoped that the government will be able to pay more attention to regions whose revenues are still below average.

4. Conclusion

The results of the research carried out state that the k-means method can be applied and implemented in the case of mapping areas with trade and accommodation facilities. The results of the mapping show that there are still many areas in Indonesia where the potential for trade and housing must be maximized. Given that Indonesia is a country of natural beauty. In addition, for further research, it is necessary to review the results of the cluster formed by using other clustering methods to see the performance of the cluster formed. To get the maximum results, make comparisons.

References

[1] I. G. I. Sudipa, C. Astria, K. F. Irnanda, and A. Perdana, "Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses . Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses .," 2020.

[2] N. Nasution *et al.*, "Application of ELECTRE Algorithm in Skincare Product Selection," *J.*

Phys. Conf. Ser., vol. 1471, no. 1, 2020.
[3] H. Pratiwi *et al.*, "Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020.
[4] W. M. Sari *et al.*, "Improving the Quality of Management with the Concept of Decision Support Systems in Determining Factors for Choosing a Cafe based on Consumers," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020.
[5] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019.
[6] F. Rahman, I. I. Ridho, M. Muflih, and S. Pratama, "Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination C," 2020.
[7] 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," *J. Phys. Conf. Ser.*, vol. 1255, no. 012005, pp. 1–7, 2019.
[8] S. Mujanah, I. Wahyurini, and Murgianto, "The effect of perceived organizational support, self-efficacy, emotional intelligence on employee performance with interpersonal communication as intervening variable," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 5, pp. 612–622, 2019.
[9] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, "Classification of natural disaster prone areas in Indonesia using K-means," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.
[10] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method," *Int. J. Artif. Intell. Res.*, vol. 1, no. 2, pp. 26–33, 2017.
[11] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara," *Semin. Nas. Teknol. Komput. Sains*, pp. 762–767, 2019.
[12] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020.
[13] Z. R. S. Elsi *et al.*, "Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia," *J. Phys. Conf. Ser.*, vol. 1594, no. 1, 2020.
[14] W. T. Ji, Q. J. Guo, and S. Zhong, "The improvement of K-medoids clustering algorithm under semantic web," *Appl. Mech. Mater.*, vol. 380–384, no. Iccsee, pp. 1286–1289, 2013.
[15] H. Nguyen, X. N. Bui, Q. H. Tran, and N. L. Mai, "A new soft computing model for estimating and controlling blast-produced ground vibration based on Hierarchical K-means clustering and Cubist algorithms," *Appl. Soft Comput. J.*, vol. 77, pp. 376–386, 2019.
[16] D. B. Bisandu, R. Prasad, and M. M. Liman, "Data clustering using efficient similarity measures," *J. Stat. Manag. Syst.*, vol. 22, no. 5, pp. 901–922, 2019.
[17] O. Koren, carina A. Hallin, nir Perel, and D. Bendet, "Decision-Making Enhancement in a Big Data Environment: Application of the K-Means Algorithm to Mixed Data," *J. Artif. Intell. Soft Comput. Res.*, vol. 9, no. 4, pp. 293–302, 2019.
[18] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 2, pp. 521–526, 2019.
[19] S. K. Majhi and S. Biswal, "Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer," *Karbala Int. J. Mod. Sci.*, vol. 4, no. 4, pp. 347–360, 2018.