Model of Data mining Clustering Rules on Population Determination of Trade and Accommodation Facilities in Indonesia with K-Means

Rino Subekti

Informatics Study Program, Institut bisnis dan informatika Kosgoro 1957

*rino.subekti@ibi-k57.ac.id, rino.subekti@gmail.com*

Abstract

The research aims to conduct a mapping model in the form of the grouping of residents of trade and accommodation facilities according to regions in Indonesia using data mining techniques. This research is a reference specifically for the role of the government in increasing regional income in Indonesia evenly. The data source is obtained from the government statistical data provider website, namely the Central Statistics Agency (BPS) with the URL address www.bps.go.id. The mapping method used is K-Mens and tested with the Rapid Miner software. There are 3 clusters used in mapping the area to the population of trade and accommodation facilities, namely the high (C1), medium (C2), and low (C3) clusters. The results obtained are cluster C1 centroid data, namely ((1527), (810.4), (5865), (6655.3), (323), (315.1)); cluster C2, namely ((286), (199,591), (1327), (2240,227), (93,227), (140,955)); and cluster C3, namely ((139,25), (122,5), (508,833), (919,222), (64,417), (94,444)). The results of the mapping show that in cluster C3, there are 16 provinces with a low population of trade and accommodation facilities..

Keywords: *Data mining, Cluster, K-means, Population Determination of Trade and Accommodation Facilities dan Indonesia*

1. Introduction

Times are increasingly developing, and each region or region is competing so that its territory becomes one of the best areas in Indonesia not only from the size of the area but also how the region progresses and develops from various aspects and is entirely related to the buildings and facilities in it. Including trade and accommodation facilities because these two facilities are also one of the factors a region is said to be developing and the more trade and accommodation facilities are, the higher the income from the region. These facilities include food stalls, mini markets, shops, inns and hotels. By building many of these facilities in an area, it indirectly makes the area develop better because there are many opportunities for tourists or people who want to go on vacation to the region, without bothering to think about where to live and eat because there are already many available trade and accommodation facilities that have been built. Many methods can perform cluster mapping in the form of [1], [2]. One of them is data mining [3], [4]. Data mining is starting to rapidly develop into an important field of research and is needed to enhance scientific studies in the field of artificial intelligence and statistics [5]. Clustering is a part of data mining which is a method of grouping data where it works to find the grouping of a series of patterns, points, objects and documents [6], [7].

One of the well-known clustering methods is K-Means which is an algorithm that plays an important role in the field of data mining because this method is simple to implement and run [8], [9]. With the advantages of the K-Means method, it is hoped that it can assist in conducting cluster mapping of the population of trading facilities and accommodation based on regions in Indonesia. The purpose of the research substance is:
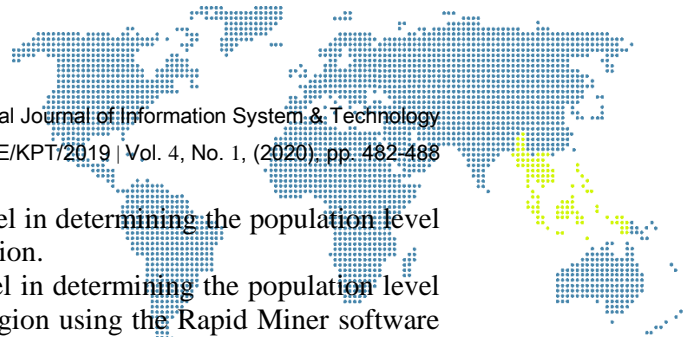
a) Analyze the K-Means data mining rule model in determining the population level of trade and accommodation facilities by region.
b) Testing the K-Means data mining rule model in determining the population level of trade and accommodation facilities by region using the Rapid Miner software tools.

2. Research Methodology

In mapping the population groupings of trade and accommodation facilities according to regions in Indonesia using data mining techniques where the research data source is obtained from data collected based on documents produced by the https://www.bps.go.id site. The data used in the study is population data for trade and accommodation facilities by region in Indonesia, 2014-2018, which consists of 34 provinces. There are 6 variables used, namely mini markets, restaurants / houses, stalls / stalls, shops / grocery stalls, hotels, lodgings which are the average number of each variable. The data is processed by using the k-means method on the population of trade facilities and accommodation according to regions in Indonesia and grouping them into three clusters, namely the high cluster (C1), medium cluster (C2) and low cluster (C3). The following is a flowchart using the K-Means clustering method.
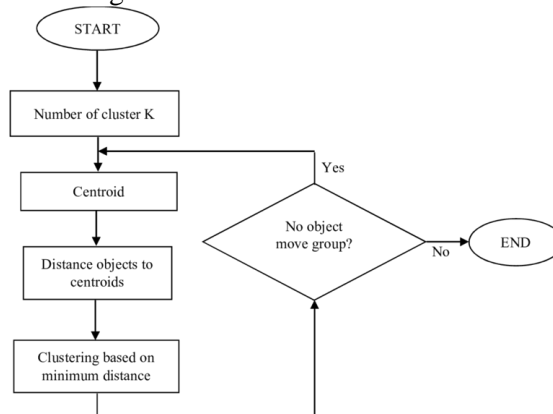
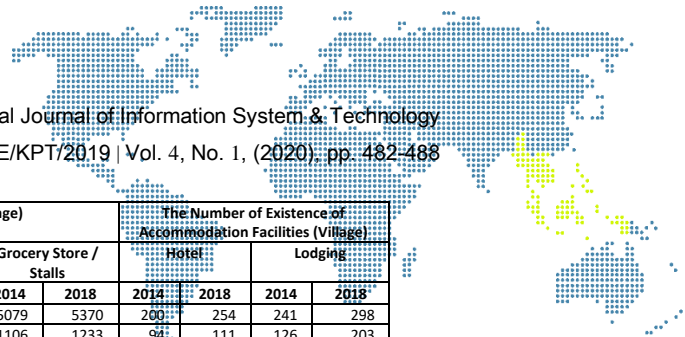**Figure 1. Process flowchart on the K-Means method**

a) Prepare training data in vector form.
b) Set the K cluster value.
c) Set the initial value for centroids.
d) Calculate the distance between data and centroid using the Euclidean Distance formula.
e) Data partition based on minimum values.
f) Then do iteration as long as the data partition is still moving (no more objects are moving to another partition), if so then go to point c).
g) If the current data group is the same as the previous data group, then stop the iteration.
h) Data has been partitioned according to the final centroid value.

3. Result and Discussion

In conducting clustering, the data used is the average value for each variable so that it is easier to process data as can be seen in table 1 and table 2 below:

Table 1. Trade & Accommodation Facilities by Province

Province	Number of Existence of Trading Facilities (Village)								The Number of Existence of Accommodation Facilities (Village)			
	Mini Market		Restaurants		Food Stalls / Beverages		Grocery Store / Stalls		Hotel		Lodging	
	2014	2018	2014	2018	2014	2018	2014	2018	2014	2018	2014	2018
Aceh	267	329	148	154	4527	3860	5178	5271	71	89	132	165

Province	Number of Existence of Trading Facilities (Village)								The Number of Existence of Accommodation Facilities (Village)			
	Mini Market		Restaurants		Food Stalls / Beverages		Grocery Store / Stalls		Hotel		Lodging	
	2014	2018	2014	2018	2014	2018	2014	2018	2014	2018	2014	2018
North Sumatra	504	647	511	634	4638	4373	5079	5370	200	254	241	298
West Sumatra	290	397	348	414	1065	1193	1106	1233	94	111	126	203
Riau	235	371	223	263	1406	1407	1749	1847	81	94	143	172
Jambi	162	202	132	227	1038	1014	1490	1541	61	74	37	79
South Sumatra	383	470	368	339	2110	2062	3144	3220	90	127	113	149
Bengkulu	88	124	166	144	762	854	1485	1471	53	60	48	67
Lampung	435	582	326	314	1872	2221	2616	2646	84	107	88	122
Kep. Bangka Belitung	58	61	74	103	341	345	375	390	41	46	35	51
Kep. Riau	111	117	110	118	301	300	402	408	53	59	57	66
DKI Jakarta	261	257	219	236	267	262	266	263	128	142	67	99
West Java	2008	2451	1189	1409	5248	4908	5858	5930	373	443	367	515
Central Java	1740	2546	856	1139	7368	7319	8476	8540	445	508	216	438
DI Yogyakarta	216	242	111	167	413	412	437	436	63	84	78	133
East Java	2034	2744	877	1187	8192	8217	8393	8458	380	467	283	496
Banten	511	645	219	278	1131	1134	1494	1538	85	111	72	92
Bali	310	392	173	223	707	698	703	716	152	175	238	300
West Nusa Tenggara	197	293	144	148	690	789	1097	1126	81	103	90	124
East Nusa Tenggara	90	111	81	108	379	454	2563	2936	95	124	92	122
West Kalimantan	175	255	118	227	934	1180	1974	2090	51	71	149	185
Central Kalimantan	59	90	65	71	805	766	1436	1521	41	50	100	114
South Borneo	117	183	129	158	1827	1677	1958	1941	82	84	82	104
East Kalimantan	155	184	100	133	681	703	972	1007	77	99	167	200
North Kalimantan	34	37	15	24	122	146	335	375	20	25	36	48
North Sulawesi	165	295	156	212	1247	1145	1693	1770	80	78	124	142
Central Sulawesi	80	93	112	169	1021	1014	1838	1956	55	68	176	228
South Sulawesi	399	532	166	266	1626	1779	2886	2962	135	163	211	262
Southeast Sulawesi	63	102	72	87	762	806	2178	2286	90	96	121	152
Gorontalo	51	94	77	101	542	572	693	716	27	32	40	57
West Sulawesi	21	29	34	52	280	297	591	620	18	16	42	48
Maluku	46	46	48	77	195	217	868	1088	29	32	84	108
North Maluku	60	57	54	74	225	252	1076	1119	26	35	87	91
West Papua	44	44	17	46	168	201	844	1031	39	39	55	82
Papua	99	85	67	98	334	361	1592	2263	64	70	61	82

Table 2. Conversion Data

Province	Number of Existence of Trading Facilities (Village)				The Number of Existence of Accommodation Facilities (Village)	
	Mini Market	Restaurants	Food Stalls / Beverages	Grocery Store / Stalls	Hotel	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
DI 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
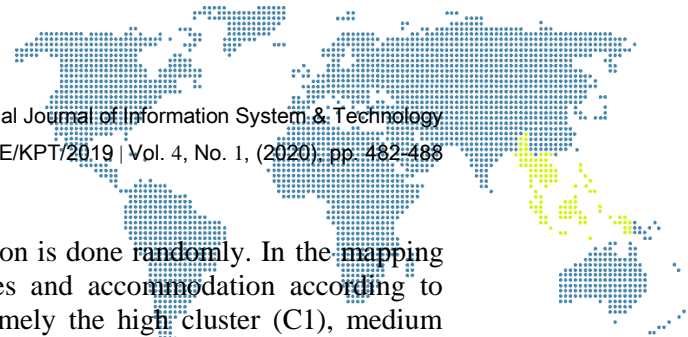
3.1. Centroid Data

In the K-Means method, the centroid determination is done randomly. In the mapping case, a cluster of the population of trade facilities and accommodation according to regions in Indonesia is divided into 3 clusters, namely the high cluster (C1), medium cluster (C2) and low cluster (C3). In cluster C1, the value is obtained from the maximum value of a data, cluster C2, the value is obtained from the average value of a data and cluster C3, the value is obtained from a minimum value as shown in Table 3 below:

Table 3. *Centroid* Data

	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/ Restaurant
C = Food and Beverage Stalls
D = Grocery Store / Stalls
E = Hotel
F = Lodging

3.2. Data Clustering

After obtaining three centroid points from the existing data, the iteration process is carried out using the initial centroid data. Following are the results of the first iteration as shown in Tables 4 and 5 below:

### Table 4. Calculation of cluster center distance

Province	A	B	C	D	E	F	C1	C2	C3	Distance
Aceh	298	151	4193.5	5224.5	80	148.5	5727.342	4019.386	6417.809	4019.386
North Sumatra	575.5	572.5	4505.5	5224.5	227	269.5	5326.558	4247.910	6664.558	4247.910
West Sumatra	343.5	381	1129	1169.5	102.5	164.5	10447.938	1117.535	1436.486	1117.535
Riau	303	243	1406.5	1798	87.5	157.5	9845.564	431.608	2028.898	431.608
Jambi	182	179.5	1026	1515.5	67.5	58	10337.505	893.854	1553.615	893.854
South Sumatra	426.5	353.5	2086	3182	108.5	131	8413.039	1127.647	3551.220	1127.647
Bengkulu	106	155	808	1478	56.5	57.5	10534.392	1083.759	1397.705	1083.759
Lampung	508.5	320	2046.5	2631	95.5	105	8787.085	671.148	3097.125	671.148
Kep. Bangka Belitung	59.5	88.5	343	382.5	43.5	43	11621.734	2213.924	253.495	253.495
Kep. Riau	114	114	300.5	405	56	61.5	11620.349	2209.174	257.323	257.323
DKI Jakarta	259	227.5	264.5	264.5	135	83	11701.653	2326.241	361.463	361.463
West Java	2229.5	1299	5078	5894	408	441	4078.986	5545.059	7933.689	4078.986
Central Java	2143	997.5	7343.5	8508	476.5	327	951.701	8783.421	11210.127	951.701
DI Yogyakarta	229	139	412.5	436.5	73.5	105.5	11495.675	2106.132	412.634	412.634
East Java	2389	1032	8204.5	8425.5	423.5	389.5	289.061	9368.987	11774.317	289.061
Banten	578	248.5	1132.5	1516	98	82	10176.279	822.852	1711.627	822.852
Bali	351	198	702.5	709.5	163.5	269	11072.028	1718.330	855.773	855.773
West Nusa Tenggara	245	146	739.5	1111.5	92	107	10799.073	1368.239	1076.240	1076.240
East Nusa Tenggara	100.5	94.5	416.5	2749.5	109.5	107	10037.309	1316.237	2505.821	1316.237
West Kalimantan	215	172.5	1057	2032	61	167	9963.349	564.924	2013.220	564.924
Central Kalimantan	74.5	68	785.5	1478.5	45.5	107	10565.274	1116.442	1381.336	1116.442
South Borneo	150	143.5	1752	1949.5	83	93	9553.687	408.132	2344.164	408.132
East Kalimantan	169.5	116.5	692	989.5	88	183.5	10931.936	1505.255	944.638	944.638
North Kalimantan	35.5	19.5	134	355	22.5	42	11796.007	2366.721	91.273	91.273
North Sulawesi	230	184	1196	1731.5	79	133	10059.692	612.306	1833.342	612.306
Central Sulawesi	86.5	140.5	1017.5	1897	61.5	202	10110.948	701.409	1868.591	701.409
South Sulawesi	465.5	216	1702.5	2924	149	236.5	8858.830	758.600	3133.854	758.600
Southeast Sulawesi	82.5	79.5	784	2232	93	136.5	10074.667	854.302	2077.298	854.302
Gorontalo	72.5	89	557	704.5	29.5	48.5	11250.022	1832.449	616.291	616.291
West Sulawesi	25	43	288.5	605.5	17	45	11517.228	2079.748	375.117	375.117
Maluku	46	62.5	206	978	30.5	96	11314.302	1862.290	720.871	720.871
North Maluku	58.5	64	238.5	1097.5	30.5	89	11209.458	1759.927	842.795	842.795
West Papua	44	31.5	184.5	937.5	39	68.5	11360.837	1908.544	676.144	676.144
Papua	92	82.5	347.5	1927.5	67	71.5	10587.527	1290.701	1680.172	1290.701
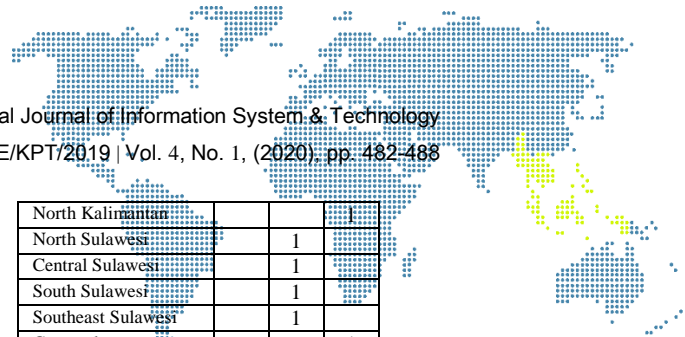
Table 5. Results of the first iteration grouping

Province	C1	C2	C3
Aceh		1	
North Sumatra		1	
West Sumatra		1	
Riau		1	
Jambi		1	
South Sumatra		1	
Bengkulu		1	
Lampung		1	
Kep. Bangka Belitung			1
Kep. Riau			1
DKI Jakarta			1
West Java	1		
Central Java	1		
DI Yogyakarta			1
East Java	1		
Banten		1	
Bali			1
West Nusa Tenggara			1
East Nusa Tenggara		1	
West Kalimantan		1	
Central Kalimantan		1	
South Borneo		1	
East Kalimantan			1
North Kalimantan			1
North Sulawesi		1	
Central Sulawesi		1	
South Sulawesi		1	
Southeast Sulawesi		1	
Gorontalo			1
West Sulawesi			1
Maluku			1
North Maluku			1
West Papua			1
Papua		1	
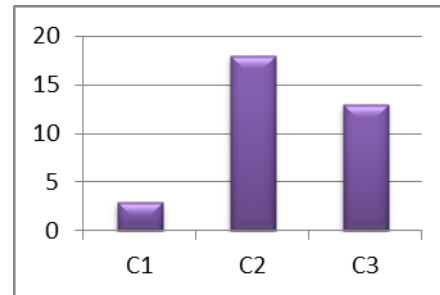
Figure 2. Bar graph of the first iteration process

In this case the iteration process continues to the second iteration and stops if the last iteration has the same data group as the previous iteration. The following is the last centroid data on the final centroid process as shown in Table 6 below:

Table 6. Latest Centroid Data

	A	B	C	D	E	F
C1	1527,000	810,400	5865,000	6655,300	323,000	315,100
C2	286,000	199,591	1327,000	2240,227	93,227	140,955
C3	139,250	122,500	508,833	919,222	64,417	94,444

Table 7. Calculation of the distance to the center of the last iteration cluster

Province	A	B	C	D	E	F	C1	C2	C3	Distance
Aceh	298	151	4193.5	5224.5	80	148.5	2621.670	4264.143	5760.114	2621.670
North Sumatra	575.5	572.5	4505.5	5224.5	227	269.5	2206.507	4517.810	6000.835	2206.507
West Sumatra	343.5	381	1129	1169.5	102.5	164.5	7360.710	1018.119	806.495	806.495
Riau	303	243	1406.5	1798	87.5	157.5	6735.874	410.295	1360.280	410.295
Jambi	182	179.5	1026	1515.5	67.5	58	7223.012	684.940	884.849	684.940
South Sumatra	426.5	353.5	2086	3182	108.5	131	5276.781	1353.904	2877.746	1353.904
Bengkulu	106	155	808	1478	56.5	57.5	7413.761	821.322	735.953	735.953
Lampung	508.5	320	2046.5	2631	95.5	105	5670.067	984.163	2427.649	984.163
Kep. Bangka Belitung	59.5	88.5	343	382.5	43.5	43	8524.515	2001.549	471.224	471.224
Kep. Riau	114	114	300.5	405	56	61.5	8523.324	1994.051	456.170	456.170
DKI Jakarta	259	227.5	264.5	264.5	135	83	8616.538	2128.274	622.503	622.503
West Java	2229.5	1299	5078	5894	408	441	1397.948	5835.743	7268.269	1397.948
Central Java	2143	997.5	7343.5	8508	476.5	327	2461.021	9058.795	10542.423	2461.021
DI Yogyakarta	229	139	412.5	436.5	73.5	105.5	8405.058	1908.466	400.255	400.255
East Java	2389	1032	8204.5	8425.5	423.5	389.5	3068.331	9660.506	11114.368	3068.331
Banten	578	248.5	1132.5	1516	98	82	7080.272	727.679	1054.942	727.679
Bali	351	198	702.5	709.5	163.5	269	7986.834	1553.936	385.030	385.030
West Nusa Tenggara	245	146	739.5	1111.5	92	107	7693.249	1158.465	402.666	402.666
East Nusa Tenggara	100.5	94.5	416.5	2749.5	109.5	107	6897.661	1021.092	1932.971	1021.092
West Kalimantan	215	172.5	1057	2032	61	167	6829.764	225.196	1345.338	225.196
Central Kalimantan	74.5	68	785.5	1478.5	45.5	107	7441.895	844.918	729.298	729.298
South Borneo	150	143.5	1752	1949.5	83	93	6442.758	578.578	1697.176	578.578
East Kalimantan	169.5	116.5	692	989.5	88	183.5	7826.753	1294.549	282.953	282.953
North Kalimantan	35.5	19.5	134	355	22.5	42	8692.130	2133.092	604.694	604.694
North Sulawesi	230	184	1196	1731.5	79	133	6943.389	431.590	1161.707	431.590
Central Sulawesi	86.5	140.5	1017.5	1897	61.5	202	6981.737	391.807	1207.496	391.807
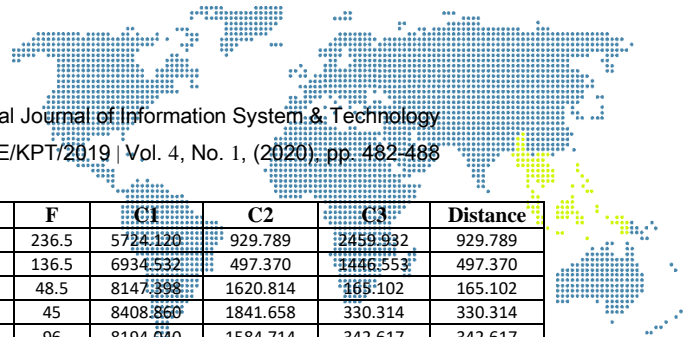
Province	A	B	C	D	E	F	C1	C2	C3	Distance
South Sulawesi	465.5	216	1702.5	2924	149	236.5	5724.120	929.789	2459.932	929.789
Southeast Sulawesi	82.5	79.5	784	2232	93	136.5	6934.532	497.370	1446.553	497.370
Gorontalo	72.5	89	557	704.5	29.5	48.5	8147.398	1620.814	165.102	165.102
West Sulawesi	25	43	288.5	605.5	17	45	8408.860	1841.658	330.314	330.314
Maluku	46	62.5	206	978	30.5	96	8194.040	1584.714	342.617	342.617
North Maluku	58.5	64	238.5	1097.5	30.5	89	8086.837	1473.927	387.586	387.586
West Papua	44	31.5	184.5	937.5	39	68.5	8240.677	1632.330	352.194	352.194
Papua	92	82.5	347.5	1927.5	67	71.5	7450.425	935.339	1119.264	935.339

Table 8. Results of the last iteration grouping

No	Province	C1	C2	C3
1	Aceh	1		
2	North Sumatra	1		
3	West Sumatra			1
4	Riau		1	
5	Jambi		1	
6	South Sumatra		1	
7	Bengkulu			1
8	Lampung		1	
9	Kep. Bangka Belitung			1
10	Kep. Riau			1
11	DKI Jakarta			1
12	West Java	1		
13	Central Java	1		
14	DI Yogyakarta			1
15	East Java	1		
16	Banten		1	
17	Bali			1
18	West Nusa Tenggara			1
19	East Nusa Tenggara		1	
20	West Kalimantan		1	
21	Central Kalimantan			1
22	South Borneo		1	
23	East Kalimantan			1
24	North Kalimantan			1
25	North Sulawesi		1	
26	Central Sulawesi		1	
27	South Sulawesi		1	
28	Southeast Sulawesi		1	
29	Gorontalo			1
30	West Sulawesi			1
31	Maluku			1
32	North Maluku			1
33	West Papua			1
34	Papua		1	
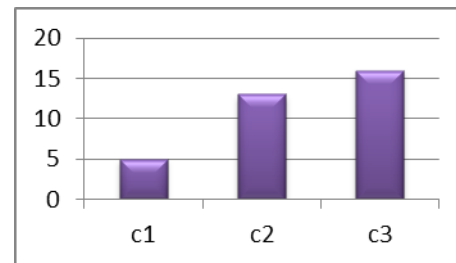
Figure 3. Bar graph of the last iteration

From this process, it was found that from 34 provinces, there were five provinces belonging to the population of high level of trade facilities & accommodation (C1), namely Aceh, North Sumatra, West Java, Central Java, and East Java. Thirteen provinces are classified in the population level of medium trade and accommodation facilities (C2) and 16 other provinces are classified as low population (C3).

4. Conclusion

The application of trade & accommodation facilities based on provinces using clustering with the k-means method can be applied with the results of mapping in the form of clusters, namely five provinces belonging to the population at the high level of trade & accommodation facilities (C1), 13 provinces belonging to the population level of medium trade & accommodation facilities (C2) and 16 other provinces are classified as low population (C3). The results of the research can be used as information for entrepreneurs who want to establish trade facilities and accommodation in the competitive world of industry based on the clustering process that has been carried out.

References

[1]   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.

[2]   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, doi: 10.1088/1742-6596/1255/1/012001.

[3]   A. P. Windarto, "Penerapan Data Mining Pada Ekspor Buah-Buahan Menurut

Negara Tujuan Menggunakan K-Means Clustering," *Techno COM*, vol. 16, no. 4, pp. 348–357, 2017.

[4]   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.

[5]   I. Kamila, U. Khairunnisa, and Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, pp. 119–125, 2019.

[6]   B. Wira, A. E. Budianto, and A. S. Wiguna, "Implementasi Metode K-Medoids Clustering Untuk Mengetahui Pola Pemilihan Program Studi Mahasiwa Baru Tahun 2018 Di Universitas Kanjuruhan Malang," *Rainstek*, vol. 1, no. 3, pp. 54–69, 2019.

[7]   S. Harikumar and P. V. Surya, "K-Medoid Clustering for Heterogeneous DataSets," *Procedia Comput. Sci.*, vol. 70, pp. 226–237, 2015, doi: 10.1016/j.procs.2015.10.077.

[8]   J. Jamal and D. Yanto, "Analisis RFM dan Algoritma K-Means untuk Clustering Loyalitas Customer," *Energy*, vol. 9, no. 1, pp. 0–8, 2019.

[9]   P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

Authors

1st **Author**
Rino Subekti
Informatics Study Program, Institut bisnis dan informatika Kosgoro 1957