Educational Data Mining with Clustering technique on the Distribution of Civil Servant Teachers in Indonesia

Moh. Mahbub[1], Dewi Purnamawati[2], Maslamah[3], Samel Sopakua[4], Mohammad Fauziddin[5]

[1]*Institut Agama Islam Negeri Surakarta, Indonesia.*
[2]*Poltekkes Kemenkes Mataram, Indonesia.*
[3]*Institut Agama Islam Negeri Surakarta, Indonesia.*
[4]*Institut Agama Kristen Negeri Ambon, Indonesia.*
[5]*Universitas Pahlawan Tuanku Tambusai, Indonesia.*

Abstract-The teacher is a professional educator with the main task of educating, teaching, guiding, directing, training, assessing, and evaluating students in early childhood education through formal education, basic education and secondary education. The purpose of this research is to analyze the distribution of teachers in Indonesia by utilizing data mining techniques. Research output in the form of mapping results in the form of clusters of regions in Indonesia. The data source was obtained by the Ministry of Education and Culture data in 2019 consisting of 35 data records. The technique used is K-Means which is part of clustering. Attributes on the study are the name of the province and the number of teachers distribution of Civil Servants (abbreviated as PNS). The calculation and analysis process uses the RapidMiner 5.3 software as a tool. To maximize the cluster, the Davies-Bouldin Index calculation result is done with the number of clusters (k = 2): 0.126. The results state that the high cluster label (C1) consists of 3 provinces namely West Java, Central Java, East Java. The rest are in the lower cluster (C2) consisting of 32 provinces (91%). This result shows that almost all regions in Indonesia the teacher distribution process is uneven and is only centered on the island of Java.

Keywords: Teacher Distribution, Data Mining, K-Means, Clustering, Indonesia

Introduction

According to Law No. 14 of 2005, teachers are professional educators with the main task of educating, teaching, guiding, directing, training, evaluating, and gathering students in early childhood education through formal education, basic education and secondary education. In Indonesia, the government continues to make priority policies to strengthen the role of teachers issued by the government and to arrange the needs of teachers; Improvement of academic qualifications; completion of teacher certification; Competency improvement based on professional work groups; also the awarding, welfare and protection. Because the main role of civil servant teachers (abbreviated as PNS) plays a very important role. However, what is most important is the uneven distribution of teachers. According to the Ministry of Education and Culture, the low national teacher ratio shows that Indonesia does not lack teachers but an unequal distribution of teachers. Based on these problems, it is necessary to have an intelligent system that can map regions in Indonesia concerning special teachers of civil servants in Indonesia.

Many branches of computer science can do mapping in the form of clusters. One of them is data mining [1]–[4]. Data mining is a technique used to explore by taking patterns in the data to be processed and then the output in the form of very important information [5], [6]. One of the most popular data mining techniques in clustering is k-means [7]. The k-means method is one of the well-known simple and easy learning methods for solving the problem of grouping from a dataset [8]. Besides this method is commonly used because it is relatively fast and easy to adapt[9]. Many previous studies have used the k-means method as a solution to the problem. One of them was done [8]with the title Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province. The results of the study state that the k-means method can be applied in classifying the number of clean water customers by province (1995-2015) into 3 clusters. Based on these advantages, it is expected that the results of the study can provide information in the form of cluster mapping of the distribution area of PNS teachers in Indonesia so that the government can follow up to immediately conduct teacher distribution in Indonesia.

DOI: 10.5373/JARDCS/V12I6/S20201171
**Corresponding Author: Moh. Mahbub, Email: Moh.Mahbub1@gmail.com*
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

Methodology

1.1. Data

Source of research data obtained from data from the Ministry of Education and Culture in 2019 about the distribution of civil servant teachers (abbreviated as PNS) as many as 35 data records. The attributes used in the study are the name of the province and the number of PNS teacher distribution. Here are complete data about the attributes used in the study.

Table 1. Research attributes

Role	Name	Type	Statistics	Range	Missings
id	Region Name	polynominal	mode = Aceh (1), least = Aceh (1)	Aceh (1), Bali (1), Bangka Belitung (1), Banten (1), Bengkulu (1), In Yogyakarta (1), DKI Jakarta (1), Gorontalo (1), Jambi (1), West Java (1), Central Java (1), East Java (1), West Kalimantan (1), South Borneo (1), Central Kalimantan (1), East Kalimantan (1), North Kalimantan (1), Riau islands (1), Lampung (1), Overseas (1), Maluku (1), North Maluku (1), NTB (1), NTT (1), Papua (1), West Papua (1), Riau (1), West Sulawesi (1), South Sulawesi (1), Central Sulawesi (1), Southeast Sulawesi (1), North Sulawesi (1), West Sumatra (1), South Sumatra (1), North Sumatra (1)	0.0
cluster	cluster	nominal	mode = cluster_1 (32), least = cluster_0 (3)	cluster_1 (32), cluster_0 (3)	0.0
regular	Number of Teachers for Civil Servants	integer	avg = 43438.686 +/- 47839.238	[12.000 ; 190787.000]	0.0

The preprocessing stage is still carried out before the data is used. This aims to maximize the results of clustering. This process uses the help of Microsoft Excel software to see missing data. The following research data are used as shown in the following table:

Table 2. Research Data

No	Region Name	Number of Teachers for Civil Servants
1	Aceh	50545
2	Bali	30011
3	Bangka Belitung	10272
4	Banten	40694
5	Bengkulu	18799
6	DI Yogyakarta	24377
7	DKI Jakarta	29918
8	Gorontalo	9827
9	Jambi	28259
10	West Java	180349
11	Central Java	181070
12	East Java	190787
13	West Kalimantan	36619
14	South Borneo	30923
15	Central Kalimantan	27291
16	East Kalimantan	23971
17	North Kalimantan	5862
18	Riau islands	9590
19	Lampung	51061

No	Region Name	Number of Teachers for Civil Servants
20	Overseas	12
21	Maluku	23755
22	North Maluku	12692
23	NTB	32671
24	NTT	44244
25	Papua	17975
26	West Papua	8517
27	Riau	40627
28	West Sulawesi	10362
29	South Sulawesi	70199
30	Central Sulawesi	27398
31	Southeast Sulawesi	25456
32	North Sulawesi	23308
33	West Sumatra	52363
34	South Sumatra	53034
35	North Sumatra	97516

Source: Ministry of Education and Culture in 2019

K-Means Method

The K-Means method is an algorithm for unsupervised training that starts with cluster formation at the beginning and then iteratively cluster will be repaired until there are no significant changes in the cluster [8]. In general, the completion steps of the k-means method are as follows [10], [11]:

Determine the number of clusters;
Randomly allocate data into clusters;
Calculate the centroids of the data in each cluster;
Allocate each data to the nearest centroid;
Return to Step c, if there is still data that moves the cluster and the process will stop if there is no change between clusters.

Results and Discussion

1.2. Application of the K-Means Method

This stage is where the K-Means method analysis process is carried out on the distribution of PNS teachers in Indonesia. The results of the analysis will be proven by RapidMiner 5.3 software testing. Before calculating, determining the number of clusters is done using RapidMiner 5.3 software. This is used to determine the number of the best cluster that is very influential on cluster results. By calculating the Davies Bouldin Index (abbreviated as DBI), the best cluster will be obtained. The following test results are carried out on clusters k = 2 to k = 4 as shown in the following table:

Table3. Comparison of Davies Bouldin Index Results

Cluster	K-Means
k=2	0,126
k=3	0,446
k=4	0,456

Based on the test results in table 3, the results obtained k = 2 to reference the best clustering cluster in the distribution of civil servant teachers in Indonesia. Following are the initial centroids used in the distribution of PNS teachers in Indonesia using the k-means method:

Table4. Preliminary Centroid results

Cluster	Centroid value
C1: high cluster in the distribution area of PNS teachers	190787
C2: low cluster in the distribution area of PNS teachers	12

In table 4 it can be explained, the cluster labels used in the distribution of PNS teachers by the k-means method are C1: high cluster in the distribution area of PNS and C2: low cluster in the distribution area of PNS teachers. The process of determining centroids is done by taking the maximum value from research data for the

DOI: 10.5373/JARDCS/V12I6/S20201171
*Corresponding Author: Moh. Mahbub, Email: Moh.Mahbub1@gmail.com
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

high cluster label (C1) and the minimum cluster label low value (C2). Following are the results of the first iteration calculation as shown in the following table:

Table 5. The results of the first iteration

No	Region Name	Number of Teachers for Civil Servants	C1	C2	Shortest Distance
1	Aceh	50545	140242	50533	50533
2	Bali	30011	160776	29999	29999
3	Bangka Belitung	10272	180515	10260	10260
4	Banten	40694	150093	40682	40682
5	Bengkulu	18799	171988	18787	18787
6	DI Yogyakarta	24377	166410	24365	24365
7	DKI Jakarta	29918	160869	29906	29906
8	Gorontalo	9827	180960	9815	9815
9	Jambi	28259	162528	28247	28247
10	West Java	180349	10438	180337	10438
11	Central Java	181070	9717	181058	9717
12	East Java	190787	0	190775	0
13	West Kalimantan	36619	154168	36607	36607
14	South Borneo	30923	159864	30911	30911
15	Central Kalimantan	27291	163496	27279	27279
16	East Kalimantan	23971	166816	23959	23959
17	North Kalimantan	5862	184925	5850	5850
18	Riau islands	9590	181197	9578	9578
19	Lampung	51061	139726	51049	51049
20	Overseas	12	190775	0	0
21	Maluku	23755	167032	23743	23743
22	North Maluku	12692	178095	12680	12680
23	NTB	32671	158116	32659	32659
24	NTT	44244	146543	44232	44232
25	Papua	17975	172812	17963	17963
26	West Papua	8517	182270	8505	8505
27	Riau	40627	150160	40615	40615
28	West Sulawesi	10362	180425	10350	10350
29	South Sulawesi	70199	120588	70187	70187
30	Central Sulawesi	27398	163389	27386	27386
31	Southeast Sulawesi	25456	165331	25444	25444
32	North Sulawesi	23308	167479	23296	23296
33	West Sumatra	52363	138424	52351	52351
34	South Sumatra	53034	137753	53022	53022
35	North Sumatra	97516	93271	97504	93271

In table 5, cluster determination is seen from the shortest distance from each cluster to obtain the following cluster results:

Table 6. The results of the first iteration

No	Region Name	Number of Teachers for Civil Servants	C1	C2
1	Aceh	50545		1
2	Bali	30011		1
3	Bangka Belitung	10272		1
4	Banten	40694		1
5	Bengkulu	18799		1
6	DI Yogyakarta	24377		1
7	DKI Jakarta	29918		1
8	Gorontalo	9827		1
9	Jambi	28259		1
10	West Java	180349	1	
11	Central Java	181070	1	
12	East Java	190787	1	

13	West Kalimantan	36619	1	
14	South Borneo	30923		1
15	Central Kalimantan	27291		1
16	East Kalimantan	23971		1
17	North Kalimantan	5862		1
18	Riau islands	9590		1
19	Lampung	51061		1
20	Overseas	12		1
21	Maluku	23755		1
22	North Maluku	12692		1
23	NTB	32671		1
24	NTT	44244		1
25	Papua	17975		1
26	West Papua	8517		1
27	Riau	40627		1
28	West Sulawesi	10362		1
29	South Sulawesi	70199		1
30	Central Sulawesi	27398		1
31	Southeast Sulawesi	25456		1
32	North Sulawesi	23308		1
33	West Sumatra	52363		1
34	South Sumatra	53034		1
35	North Sumatra	97516	1	

In table 6 which is the first iteration result, where 5 provinces are in the high cluster in the distribution area of civil servants (C1) and 30 provinces are in the low cluster in the distribution area of PNS teachers (C2). The iteration process continues to do recalculation for centroid values. In the new centroid process, it is done based on the results of grouping on the first iteration. The iteration process stops if the results of the last iteration are the same as the results of the previous iteration. In this study, the process stops at the fourth iteration. This means that the results of data clustering in the fourth iteration have the same value as the results of the third iteration clustering. The following calculation results are as follows:

Table7. Final centroid results

Cluster	Centroid value
C1: high cluster in the distribution area of PNS teachers	184068,67
C2: low cluster in the distribution area of PNS teachers	30595,125

Table 8. The results of the last iteration

No	Region Name	Number of Teachers for Civil Servants	C1	C2	Shortest Distance	Cluster results on C1	Cluster results on C2
1	Aceh	50545	133524	20290	20290		1
2	Bali	30011	154058	244	244		1
3	Bangka Belitung	10272	173797	19983	19983		1
4	Banten	40694	143375	10439	10439		1
5	Bengkulu	18799	165270	11456	11456		1
6	DI Yogyakarta	24377	159692	5878	5878		1
7	DKI Jakarta	29918	154151	337	337		1
8	Gorontalo	9827	174242	20428	20428		1
9	Jambi	28259	155810	1996	1996		1
10	West Java	180349	3720	150094	3720	1	
11	Central Java	181070	2999	150815	2999	1	
12	East Java	190787	6718	160532	6718	1	
13	West Kalimantan	36619	147450	6364	6364		1
14	South Borneo	30923	153146	668	668		1
15	Central Kalimantan	27291	156778	2964	2964		1
16	East Kalimantan	23971	160098	6284	6284		1
17	North Kalimantan	5862	178207	24393	24393		1

DOI: 10.5373/JARDCS/V12I6/S20201171
*Corresponding Author: Moh. Mahbub, Email: Moh.Mahbub1@gmail.com
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

No	Region Name	Number of Teachers for Civil Servants	C1	C2	Shortest Distance	Cluster results on C1	Cluster results on C2
18	Riau islands	9590	174479	20665	20665		1
19	Lampung	51061	133008	20806	20806		1
20	Overseas	12	184057	30243	30243		1
21	Maluku	23755	160314	6500	6500		1
22	North Maluku	12692	171377	17563	17563		1
23	NTB	32671	151398	2416	2416		1
24	NTT	44244	139825	13989	13989		1
25	Papua	17975	166094	12280	12280		1
26	West Papua	8517	175552	21738	21738		1
27	Riau	40627	143442	10372	10372		1
28	West Sulawesi	10362	173707	19893	19893		1
29	South Sulawesi	70199	113870	39944	39944		1
30	Central Sulawesi	27398	156671	2857	2857		1
31	Southeast Sulawesi	25456	158613	4799	4799		1
32	North Sulawesi	23308	160761	6947	6947		1
33	West Sumatra	52363	131706	22108	22108		1
34	South Sumatra	53034	131035	22779	22779		1
35	North Sumatra	97516	86553	67261	67261		1

In table 8, the final results of the mapping are in the form of regional clusters for the distribution of civil servant teachers in Indonesia where C1 cluster consists of 3 regions (West Java, Central Java, East Java) and C2 cluster consists of 32 regions (Aceh, Bali, Bangka Belitung, Banten, Bengkulu, DI Yogyakarta, DKI Jakarta, Gorontalo, Jambi, West Kalimantan, South Borneo, Central Kalimantan, East Kalimantan, North Kalimantan, Riau islands, Lampung, Overseas, Maluku, North Maluku, NTB, NTT, Papua, West Papua, Riau, West Sulawesi, South Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, West Sumatra, South Sumatra, North Sumatra).

3.2. Testing with RapidMiner software 5.3

After the manual calculation process, testing with RapidMiner is carried out to see the results obtained. Based on the test results obtained the same results to the manual calculations performed. Following are the complete results of testing with RapidMiner 5.3 as shown in the following image:

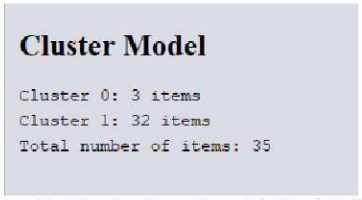

Figure 1. Cluster Results with RapidMiner

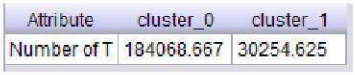

Figure 2. Final centroid results on RapidMiner

The final cluster and centroid results on the Rapidminer test have the same results as shown in table 7 and table 8. The following is a diagram of the distribution of civil servant teachers using plot view scatter from RapidMiner 5.3.

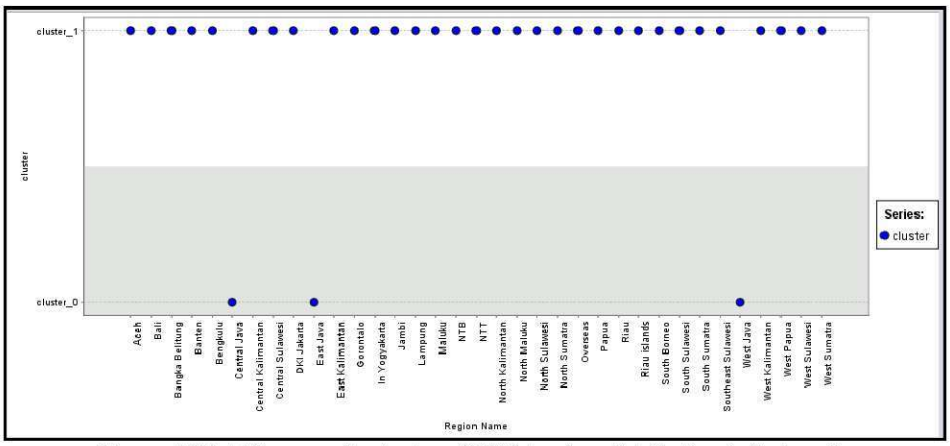

Figure 3. Plot View results cluster of PNS teacher distribution in Indonesia

Conclusion

The results obtained indicate that the distribution of civil servant teachers in Indonesia using k-means techniques can be applied. The resulting output is in the form of a mapping of cluster provinces in Indonesia to the distribution of civil servant teachers. By using the RapidMiner 5.3 software as a tool, the same results are obtained from the manual calculation process. The results of cluster mapping stated that 32 provinces were in cluster C2 (91%). And only 3 provinces are in cluster C1. This proves that teacher distribution is uneven and is only concentrated on Java islands such as West Java, Central Java, East Java. While in the western and eastern parts of Indonesia there is still a shortage of teachers.

References

[1] 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, doi: 10.1088/1742-6596/1255/1/012005.

[2] D. Hartama, A. Perdana Windarto, and A. Wanto, "The Application of Data Mining in Determining Patterns of Interest of High School Graduates," *J. Phys. Conf. Ser.*, vol. 1339, no. 1, 2019, doi: 10.1088/1742-6596/1339/1/012042.

[3] 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 Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.

[4] R. W. Sari, A. Wanto, and A. P. Windarto, "Implementasi Rapidminer Dengan Metode K-Means (Study Kasus: Imunisasi Campak Pada Balita Berdasarkan Provinsi)," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 224–230, 2018, doi: 10.30865/komik.v2i1.930.

[5] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch," *J. Phys. Conf. Ser.*, vol. 1255, no. 012002, pp. 1–6, 2019, doi: 10.1088/1742-6596/1255/1/012002.

[6] M. G. Sadewo, A. Eriza, A. P. Windarto, and D. Hartama, "Algoritma K-Means Dalam Mengelompokkan Desa / Kelurahan Menurut Keberadaan Keluarga Pengguna Listrik dan Sumber Penerangan Jalan Utama Berdasarkan Provinsi," *Semin. Nas. Teknol. Komput. Sains SAINTEKS 2019*, pp. 754–761, 2019.

[7] M. G. Sadewo, A. P. Windarto, and A. Wanto, "Penerapan Algoritma Clustering Dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/ Mitigasi Bencana Alam Menurut Provinsi Dengan K-Means," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 311–319, 2018, doi: 10.30865/komik.v2i1.943.

[8] 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.

[9] 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.

[10] 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 Conf. Ser. Mater. Sci. Eng.*, vol. 420, p. 012089, 2018, doi: 10.1088/1757-899X/420/1/012089.

[11] 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.

DOI: 10.5373/JARDCS/V12I6/S20201171
*Corresponding Author: Moh. Mahbub, Email: Moh.Mahbub1@gmail.com
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020