Mapping study using the unsupervised learning clustering approach

Mapping study using the unsupervised learning clustering approach

Nur Azis[1*], Yusmar Ali[2], Rino Subekti[3], Pilifus Junianto[4], Lispridona Diner[5], S Suhendra[2], Siti Aisyah[2,7], Agus Perdana Windarto[6]

[1] Universitas Krisnadwipayana, Indonesia
[2] Politeknik Negeri Media Kreatif, Indonesia
[3] Institut Bisnis dan Informatika Kosgoro 1957, Indonesia
[4] Sekolah Tinggi Ilmu Ekonomi Bentara Persada Batam, Indonesia
[5] Universitas Negeri Semarang, Indonesia
[6] STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
[7] Business School, Universiti Kuala lumpur Malaysia, Malaysia

*nuraziz@unkris.ac.id

Abstract. The aim of this research is to analyze the number of criminal cases in Indonesia by utilizing unsupervised learning techniques. The unsupervised learning technique used is data mining by mapping clusters of regions in Indonesia. Sources of data were obtained from the Operations Control Bureau, National Police Headquarters of the Republic of Indonesia through processed data from the Central Statistics Agency (abbreviated as BPS) with data url: https://www.bps.go.id. The data mining method used to map the form of calcter is k-medoid. The data used is data on the number of crimes according to the regional police (2017-2019) which consists of 34 records. The attribute used is the number of crimes in the past three years based on the regional police for each province. The mapping label used is the high cluster (D1) and the low cluster (D2) on the number of criminal acts in Indonesia. The mapping analysis process uses the help of Rapid Miner software. In determining the amount of calcter (k = 2) is done using the Davies Bouldin Index (DBI) parameter with a value of 0.876 (the smaller the better). The results showed that six provinces were in the high cluster (D1) and twenty-eight provinces were in the low cluster (D2). The final centroid in each cluster is 16,008; 21,498; 21,616 (cluster_0 / D1) and 6,994; 7,311; 6,785 (cluster_1 / D2). The six provinces in the high cluster of criminal cases are North Sumatra, South Sumatra, Metro Jaya, West Java, East Java and South Sulawesi. The results of the research are expected to provide information for the government to reduce the number of criminal acts in Indonesia based on the number of clusters that exist.

1. Introduction

Data mining is one of the techniques of unregulated learning, in which anybody can not know the predicted results [1], [2]. The results shown are calculated only by the weight value which has been obtained at the start of the device construction and the classification of similarly coveted items in a specific region or area [3], [4]. In other words, data mining is a form of learning that can be used to find or classify a pattern for several related objects not totally identical [5]. One approach is k-means k-medoids, which is a very common data mining method for business, academia and industry [6]–[11].

As seen in the following image, the following is the unchecked learning technique with clustering technique:

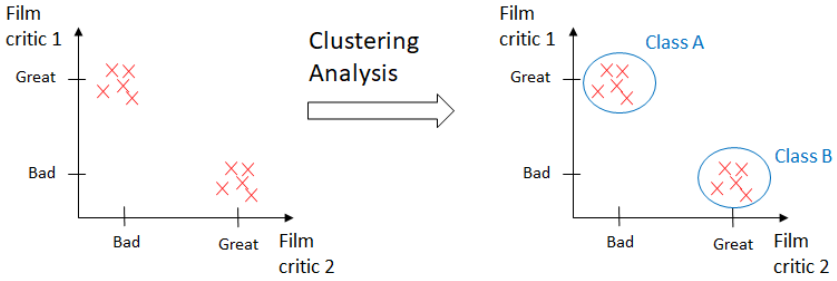

Figure 1. Unsupervised Learning techniques in data mining clustering

Criminal activity or also called crime is a community phenomenon because space and time cannot be isolated. It is therefore the responsibility of the State to provide protection to its people. The effort to achieve and build a sense of community security is a strategic move which also influences the achievement of national growth [12]. One is to avoid the occurrence in Indonesia of crime and to reduce the crime rate [13]. There are fluctuating and growing numbers of criminal offenses in Indonesia. As can be seen in Figure 2 of the data collected from the BPS.

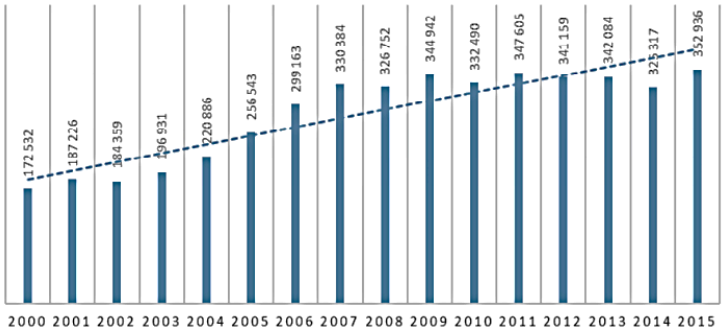

Figure 2. Graph of the Development of the Number of Criminal Acts in Indonesia 2000-2015
Source: BPS

As can be seen in the above curve, the number of criminal acts fluctuated and increased in Indonesia during 2000-2015. The number of criminal offenses in Indonesia rose in 2007 to 330,384 and then continued to fluctuate until 2014. The study focuses on data on the number of criminal actions handled by the Central Statistical Bureau (abbreviated as BPS) in which the data is collect by unregulated clustering techniques, which map regions to the number of Indonesian criminal cases based on provincial regional police. In this research, the clustering approach is k-medoids which is well known for its superiority [14]. The development of the K-means method [15] is one of the advantages of this strategy. A number of previous studies have been carried out to solve clustering issues, apart from the advantages of the k-medoid technique. As Zulhipni Reno Saputra Elsi (2020) [16] was also doing during the Covid-19 pandemic in Indonesia, with respect to the use of data mining techniques in national food security. This paper introduces the k-medoid approach for mapping the national food safety of Indonesia during the Pandemic Covid-19. The results of the calculations show that 19 provinces in C1 (high) and 14 provinces in C2 (low) or about 42% of Indonesians are still under-food. The results of research will hopefully increase data mining awareness and provide information on mapping in the form of regional clusters with a number of cases of criminal activity in Indonesia. In order to enable the Government to reduce the number of criminal cases in Indonesia in order to establish a fair and stable society.

2. Methodology

2.1. Data Mining

The method of data mining reveals meaningful patterns, interconnections and new trends through the filtering out of vast quantities of data contained in previously unknown stores [8], [10], [17]–[19]. Methods include the selection, classifying, associating and estimating of the data [3], [20]–[22]. The first phase of the data mining process is often clustering. Many clustering algorithms, such as k-means, strengthens k-means k-medoids (PAM), fuzzy c-means, and others have been used by previous researchers [23].

2.2. K-Medoids Method

The distinction between the k-medoids algorithm and the k-means algorithm is that the k-medoids method uses artifacts as a (medoid) cluster center for the cluster, and a median value for the cluster center for the k-means method. Furthermore, the k-medoid approach is better suited than the k-median method to community data [24], [25]**.**

2.3. Data

Daten was collected through data processed from the Central Statistic Agency (abbreviated as the BPS) with data Url: https:/www.bps.go.id from the Bureau of Control of Operations, National Policing Headquarters, Republic of Indonesia. For mapping areas, the data mining approach is k-medoid. Data from Regional Police (2017-2019), which contains 34 reports, was used as data on the number of crimes. Using the Rapid Miner app, clean data is analyzed using clustering methods. As shown in the following table, the following raw data and processed data:

Table 1. Research data (source: BPS)

Regional Police	2019	2018	2017
Aceh	7,483	8,758	8,885
North Sumatra	30,831	32,922	39,867
West Sumatra	11,064	12,953	13,205
Riau	6,570	7,246	6,869
Jambi	6,848	6,313	9,531
South Sumatra	12,861	13,558	15,728
Bengkulu	3,453	3,389	4,867
Lampung	8,534	8,963	11,089
Kep. Bangka Belitung	1,953	2,048	1,931
Kep. Riau	3,159	3,409	3,673
Metro Jaya	31,934	34,655	34,767
West Java	13,145	16,209	25,183
Central Java	10,317	9,127	12,033
In Yogyakarta	6,650	6,731	7,251
East Java	26,985	26,295	34,598
Banten	3,287	3,623	3,692
Bali	3,047	3,212	3,589
West Nusa Tenggara	8,185	6,451	8,132
East Nusa Tenggara	5,865	6,257	6,729
West Kalimantan	4,721	5,814	6,020
Central Kalimantan	2,444	2,667	2,699
South Borneo	5,375	5,699	6,578
East Kalimantan	4,417	6,287	9,149
North Kalimantan	876	396	0
North Sulawesi	7,425	10,247	7,981
Central Sulawesi	6,265	9,379	10,240
South Sulawesi	16,008	21,498	21,616
Southeast Sulawesi	1,213	1,263	2,866
Gorontalo	2,367	2,836	3,099
West Sulawesi	1,863	1,817	1,841
Maluku	3,495	2,751	3,086
North Maluku	718	722	789
West Papua	2,972	3,475	2,284
Papua	6,994	7,311	6,785

3. Results and Discussion

Data in Table 1 is processed at this stage with the k-medoids method. The mapping label used for clustering is two-class, namely the high cluster (C1) and the low cluster (C2). This is a model design with Rapid Miner clustering.

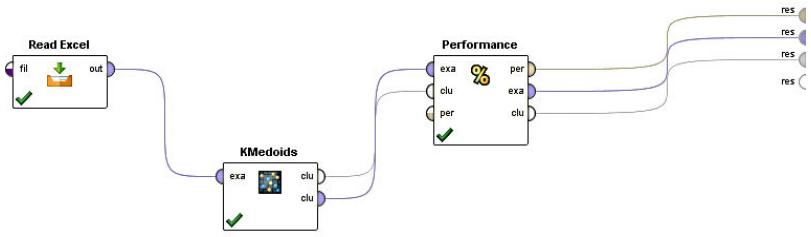

Figure 3. Mapping modeling the form of clusters in cases of the number of crimes with Rapid miner

In Figure 3, the data entry process is clarified by using the read excel method to enter data prepared as in Table 1. The K-medoids model is included for its duties and roles, namely cluster mapping of the number of criminal cases based on regions in Indonesia. Besides, performance tools (Davies Bouldin Index (DBI)) are used to assess the strength of the formed clusters. This study uses 2 clusters, namely the high cluster (C1) and the low cluster (C2), in Indonesia's number of criminal cases.

Cluster Model

```
Cluster 0: 6 items
Cluster 1: 28 items
Total number of items: 34
```

Figure 4. The results of the k-medoid mapping

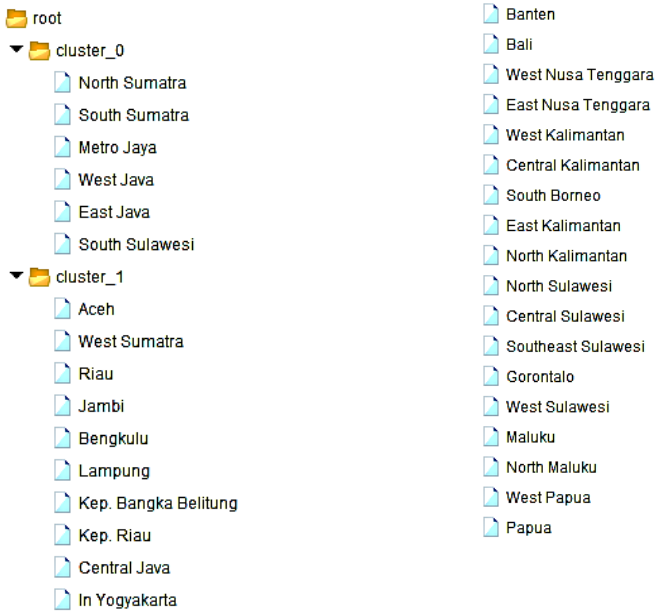

Figure 5. The high cluster (cluster_0) and the low cluster (cluster_1)

Figures 4 and 5 explain the mapping results in clusters for the number of criminal cases in Indonesia based on regions where the results for high clusters (C1) are approximately 18% (6 provinces) and 82% (28 provinces) in low clusters (C2). The six provinces are North Sumatra, South Sumatra, Metro Jaya, West Java, East Java, and South Sulawesi. Here are the last values for the high (cluster 0) and low (cluster 1) clusters shown below:

Attribute	cluster_0	cluster_1
2019	16008	6994
2018	21498	7311
2017	21616	6785

Figure 6. The final centroid results

The following is a mapping image in the form of scattered plots by region in the number of criminal cases in Indonesia as shown in the following figure:

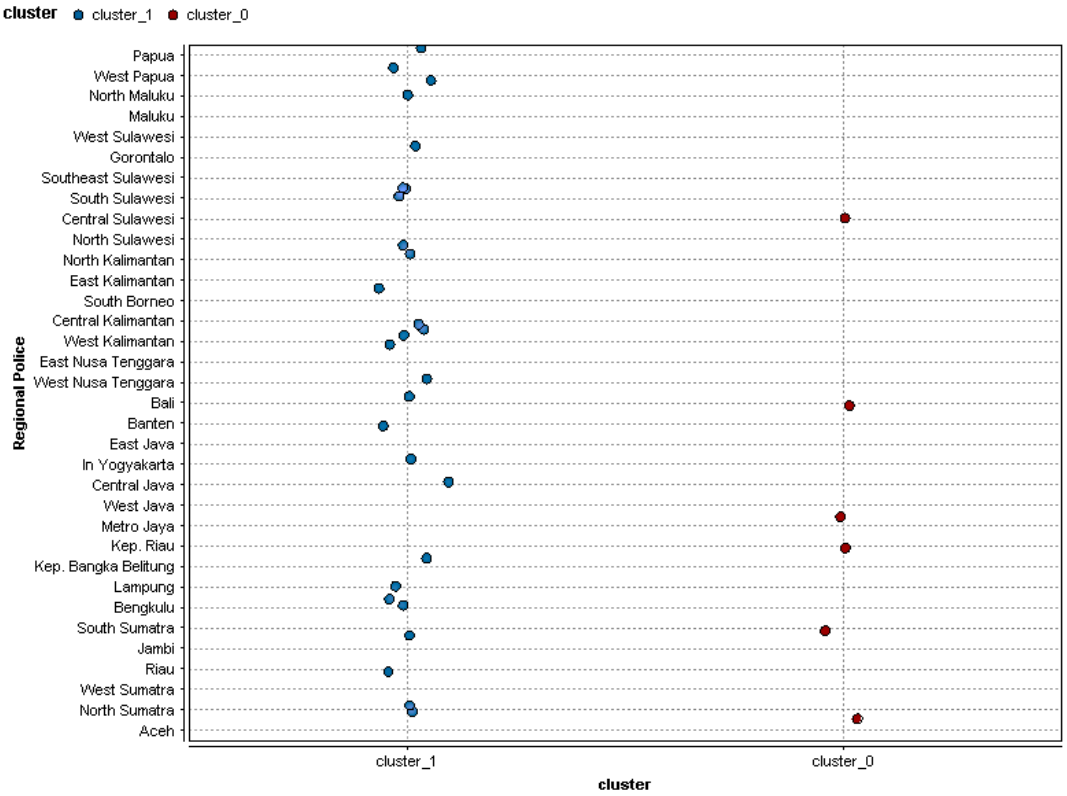

Figure 7. Clustering visualization with a scatter plotter

Davies-Bouldin Index (DBI) is an internal evaluation tool that tests cluster evaluation using a clustering method. The smaller the DBI value obtained (non-negative>=0), the stronger the cluster (k) obtained from the k-medoid group used. In the experiments carried out on 34 datasets and the number of clusters 2 (k = 2), the average value of the Davies-Bouldin Index (DBI) for 34 datasets was 0.876. The results of the DBI using the RapidMiner program are shown below.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -83480223.088
Avg. within centroid distance_cluster_0: -291928104.333
Avg. within centroid distance_cluster_1: -38812819.964
Davies Bouldin: -0.876
```

Figure 8. Performance Vector Results

4. Conclusion

Based on the results obtained, a mapping of the number of criminal cases in Indonesia using the k-medoids method can be used in clusters. This is demonstrated by developing two optimal clusters (k = 2) using the Davies-Bouldin Index (DBI) parameter. The mapping of the number of crimes in Indonesia must be a concern for the government to build a sense of security in the community.

References

[1] 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.
[2] Y. Elgimati, "Weighted Bagging in Decision Trees: Data Mining," *JINAV J. Inf. Vis.*, 2020.
[3] 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.
[4] 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.
[5] E. H. S. Atmaja, "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta," *Int. J. Appl. Sci. Smart Technol.*, vol. 1, no. 1, pp. 33–44, 2019.
[6] K. F. Irnanda, F. N. Arifah, M. R. Raharjo, A. Arifin, and A. P. Windarto, "The selection of Calcium Milk Products that are appropriate for advanced age using PROMETHEE II Algorithm," *J. Phys. Conf. Ser.*, vol. 1381, no. 1, 2019.
[7] 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.
[8] N. Rofiqo, A. P. Windarto, and D. Hartama, "Penerapan Clustering Pada Penduduk Yang Mempunyai Keluhan Kesehatan Dengan Datamining K-Means," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 216–223, 2018.
[9] 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.
[10] 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.
[11] 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.
[12] R. G. A. Obe and A. Masyhar, "Position Of The Victim In Criminal Acts Illegal Logging," *J. Law Leg. Reform*, vol. 1, no. 3, pp. 507–518, 2020.
[13] M. S. K. Madari, "Penyesuaian Batasan Tindak Pidana Ringan dan Jumlah Denda Dalam KUHP Terhadap Perkara Tindak Pidana Pencurian," *J. Cita Huk.*, vol. 1, no. 2, pp. 335–350, 2013.
[14] 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.
[15] S. Defiyanti, M. Jajuli, and N. Rohmawati, "K-Medoid Algorithm in Clustering Student Scholarship Applicants," *Sci. J. Informatics*, vol. 4, no. 1, pp. 27–33, 2017.
[16] 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.
[17] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, "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.
[18] 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.
[19] 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.
[20] 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.
[21] M. G. Sadewo, A. P. Windarto, and D. Hartama, "PENERAPAN DATAMINING PADA POPULASI DAGING AYAM RAS PEDAGING DI INDONESIA BERDASARKAN PROVINSI MENGGUNAKAN K-MEANS CLUSTERING," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 2, no. 1, pp. 60–67, 2017.
[22] 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.
[23] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.
[24] 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.
[25] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, 2013. .