Educational Data Mining on Higher Education Level Education Costs using Clustering Techniques in Indonesia

Article *in* Journal of Advanced Research in Dynamical and Control Systems · January 2020

DOI: 10.5373/JARDCS/V12I6/S20201169

5 authors, including:

Imam Makruf
State Institute For Islamic Studies, Surakarta
17 PUBLICATIONS **28** CITATIONS

SEE PROFILE

Educational Data Mining on Higher Education Level Education Costs using Clustering Techniques in Indonesia

Imam Makruf[1], Lubna[2], Khasanah[3], Ridawati Sulaeman[4], Dahrul Aman Harahap[5]

[1]Institut Agama Islam Negeri Surakarta, Indonesia.
[2]Universitas Islam Negeri Mataram, Indonesia.
[3]STMIK Indonesia Jakarta, Indonesia.
[4]Poltekkes Kemenkes Mataram, Indonesia.
[5]Universitas Labuhanbatu, Indonesia

Abstract-The purpose of this research is to utilize data mining techniques in classifying the cost of tertiary education in Indonesia by region. The data source was obtained from the National Socio-Economic Survey (Susenas) Module of Socio-Culture and Education which was managed by the Central Statistics Agency (abbreviated as BPS) in the 2017/2018 school year consisting of 35 regions in Indonesia. The variable used is the average tuition. The method used is K-Medoids which is part of clustering. The data processing is assisted with the RapidMiner 5.3 application. Cluster label determination using Cluster Distance Performance tools with Davies Bouldin performance = 0.428. So that the cluster label used 3 is C1: high cluster, C2: normal cluster and C3: medium cluster. The results of the study stated that 6 regions were in the high cluster (C1) for the cost of education at the tertiary level (Bali, Banten, DI Yogyakarta, DKI Jakarta, West Java, South Sulawesi); 18 areas are in the normal cluster (C2) and 11 areas are in the lower cluster (C3). It is expected that regions that have higher education costs at the tertiary level can be of particular concern to the government, because to improve the quality of human resources in Indonesia, it still requires a significant amount of funding.
Keywords: Data Mining, Klastering, K-Medoids Method, Education Costs, Higher Education

Introduction

Education is one of the most important aspects in human life to become a quality human being by carrying out the process of self-development from time to time so as to be able to compete in the current era of the industrial revolution. In Indonesia education especially at the tertiary level still views class differences in terms of tuition fees. So the quality is greatly influenced by the cost of education at all levels of education. This problem causes people who are felt to be unable to receive education in higher education after their children graduate from high school / vocational high school. As a result, poor people can only get less quality education in ordinary educational institutions. Quality education in Indonesia should apply to all citizens without exception, not just the upper classes. According to data from the Ministry of Education and Culture, the quality of education in Indonesia ranks 12th out of 12 countries in Asia, Indonesia ranks below Vietnam. This is also due to several fundamental factors including the low quality of education and weak public awareness of the underdevelopment of education in Indonesia due to the high cost of education. The purpose of this study is to classify the cost of tertiary education in Indonesia so that the results of the grouping can be of particular concern to the government towards the high cost of tertiary education in each region in Indonesia.

Many solutions are offered to solve these problems. One of them is by utilizing data mining clustering techniques [1]. The purpose of clustering is to group a series of points, patterns, documents that have similarities between one object and that distinguishes from other objects [2]. There are several clustering techniques used in data mining such as k-medoids (Partitioning Around Medoids) and k-means [3], [4]. Each method has its advantages [5]. K-Medoids is a development variant of k-means that can handle sensitive outliers due to an object with a large value [6] and can work on any type of data matrix [7]. This was proven by several previous studies such as [8]with the title K-Medoid Clustering for Heterogeneous DataSets. The results show that the new clustering algorithm with new similarity measures outperforms the k-means clustering for mixed datasets. Next [9]with the title Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta. The K-Medoids method can be applied to grouping crime patterns with the results obtained in this study were three crimes. Based on this, it is expected that research results can provide information in the form of cluster mapping of the costs of tertiary education in Indonesia. So that the government can make decisions based on the results of cluster mapping in order to improve the quality of education in Indonesia.

DOI: 10.5373/JARDCS/V12I6/S20201169
**Corresponding Author: Imam Makruf*
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

Methodology

1.1. Data

The data source used in this study is the National Socio-Economic Survey (Susenas) Module for Socio-Culture and Education which is managed by a statistical central body (abbreviated as BPS) by url: https://www.bps.go.id/. The data used is the average total cost of national education at tertiary level in the academic year 2017/2018 consisting of 35 regions in Indonesia with variable data is the average total cost of education. Following are the research data used as shown in table 1 below:

Table 1. Research data

No	Region Name	Average Cost of Study (Rupiah)
1	Aceh	Rp9.200.000
2	Bali	Rp18.620.000
3	Bangka Belitung	Rp12.430.000
4	Banten	Rp22.260.000
5	Bengkulu	Rp11.360.000
6	In Yogyakarta	Rp18.580.000
7	DKI Jakarta	Rp20.730.000
8	Gorontalo	Rp8.610.000
9	Jambi	Rp10.770.000
10	West Java	Rp20.220.000
11	Central Java	Rp14.460.000
12	East Java	Rp14.450.000
13	West Kalimantan	Rp12.740.000
14	South Borneo	Rp15.540.000
15	Central Kalimantan	Rp13.510.000
16	East Kalimantan	Rp15.020.000
17	North Kalimantan	Rp8.840.000
18	Riau islands	Rp15.330.000
19	Lampung	Rp10.470.000
20	Maluku	Rp9.410.000
21	North Maluku	Rp6.980.000
22	West Nusa Tenggara	Rp9.090.000
23	East Nusa Tenggara	Rp9.070.000
24	Papua	Rp12.760.000
25	West Papua	Rp7.000.000
26	National average	Rp15.330.000
27	Riau	Rp11.610.000
28	West Sulawesi	Rp6.580.000
29	South Sulawesi	Rp20.530.000
30	Central Sulawesi	Rp9.100.000
31	Southeast Sulawesi	Rp7.130.000
32	North Sulawesi	Rp12.470.000
33	West Sumatra	Rp11.530.000
34	South Sumatra	Rp14.790.000
35	North Sumatra	Rp12.660.000

source: National Socio-Economic Survey (Susenas)

Based on the introduction described, this research through several stages, the following stages of the framework in the preparation of research as shown in the following figure:

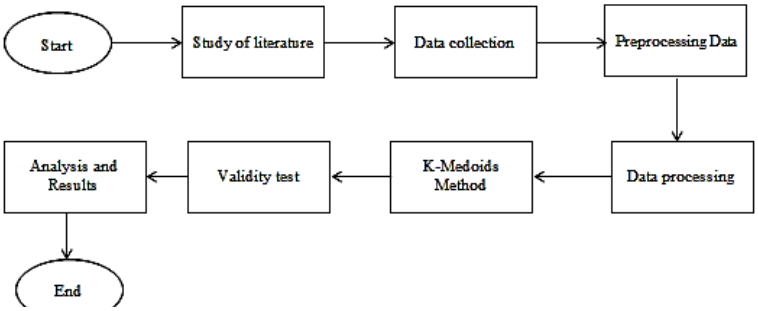

Figure 1. Research methodology

1.2. Data Mining

Data mining is a process that uses various techniques and data analysis tools to find hidden relationships and patterns that cannot be found using simple query analysis by extracting information that is reasonable and previously unknown [10]–[12]. There are several data mining techniques including: clustering, classification, estimation and association [13], [14].

K-Medoids Method (Partitioning Around Medoids)

The K-Medoids (Partitioning Around Medoids) method is a method similar to k-means because both of these methods break down the dataset into groups that have similarities [15]. Some advantages are can handle sensitive outliers due to an object with a large value [6] and can work on any type of data matrix [7]. The K-medoids method is more suitable for grouping data than the K-Means method [16]. Some steps to complete the K-Medoids method[17]:

a) Initialise the cluster center as many as the number of clusters (k).
b) Each data or object is allocated to the nearest cluster using the Euclidian Distance equation:

$$d(x,y)= |x-y| = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

Where:

d = the distance between x and y
x = cluster data center
y = data on the attribute
i = every data
n = amount of data
xi = data in the cluster center
yi = data on each data

c) Choose objects on each cluster randomly as new medoid candidates.
d) Calculate the distance of each object contained in each cluster with the new medoid candidate.
e) Calculate the total deviation (S) by calculating the total new distance value - the total old distance. If S <0 is obtained, swap objects with data
f) Cluster to create a new set of k objects as medoids.
g) Repeat steps c through e until there is no change in the medoid, so that clusters and cluster members are obtained.

Results and Discussion

This section explains how to use the K-Medoids method in the analysis of the cost of tertiary education in Indonesia supported by RapidMiner 5.3 as a tool. Data collection obtained from the National Socio-Economic Survey (Susenas) Module for Socio-Culture and Education which is managed by a statistical central body (abbreviated as BPS) is used as a research reference. The data consists of 35 records consisting of 2 attributes namely the name of the region and the average cost of higher education which is a requirement of the clustering process. Data cleaning is still done with Microsoft Excel to analyze and reduce noise that can affect calculation results.

The process of determining the amount of clustering in this study uses the Davies-Bouldin Index (DBI). By using DBI, a cluster is considered to have an optimal clustering scheme if it has a minimum DBI[18].Following are the results of the Davies-Bouldin Index (DBI) calculation using Rapidminer 5.3 as shown in the following figure:

DOI: 10.5373/JARDCS/V12I6/S20201169
*Corresponding Author: Imam Makruf
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

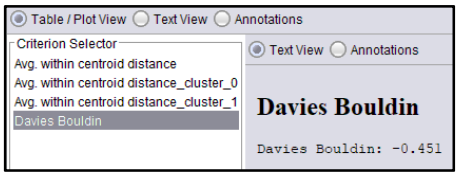

Figure 2. Davies-Bouldin Index (DBI) results for k = 2

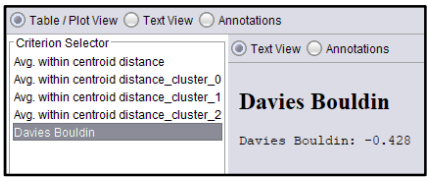

Figure 3. Davies-Bouldin Index (DBI) results for k = 3

Based on figures 2 and 3, the best value for Davies-Bouldin Index (DBI) is the number of k = 3. So that the clustering process in this study uses 3 cluster labels namely: CI = high cluster label, C2 = normal cluster label and C3 = low cluster label.

After the process of determining the number of clusters is done, the data prepared in table 1 will be processed with RapidMiner 5.3 by importing data through the available Data Import Wizard tool, as shown in the following figure:

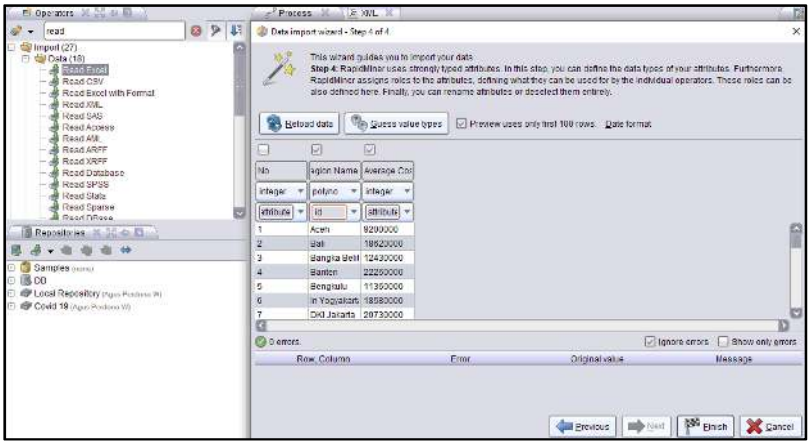

Figure 4. RapidMiner 5.3 import data

The process of importing data in RapidMiner 5.3 can not only be done with file extensions (xls) but can be done with many file extensions such as csv, xml, aml, ARFF, XRFF, access, spss and others. In Figure 4, the attribute of the region name becomes the output of clustering and the attribute of the average cost of education is included in the clustering process at the tertiary level education.On the main process sheet, clastering modeling with the k-medoids method is made by adjusting the number of meters as shown in the following figure:

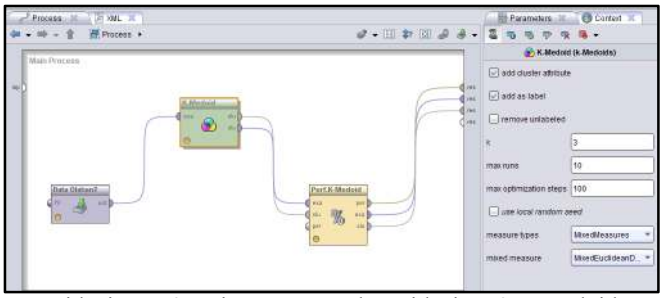

Figure 5. RapidMiner 5.3 main process and RapidMiner 3 K-medoids parameter

Using the RapidMiner 5.3 software the K-Medoids calculation process is carried out with the final centroid result on C1 (high cluster) is Rp. 20.530.000; C2 (normal cluster) is Rp. 12.660.000 and C3 (low cluster) is Rp. 7.130.000 as shown in the following table:

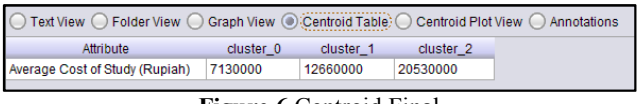

Figure 6. Centroid Final

The final centroid process shown in table 6 is the final result of clustering on the grouping of tertiary education costs in Indonesia based on regions where 6 regions are obtained for higher clusters (C1: cluster_2) which have higher education costs; 18 areas for normal clusters (C2: cluster_1) that have normal education costs and 11 regions for low cluster (C3: cluster_0) which have low education costs as shown in the following figure:

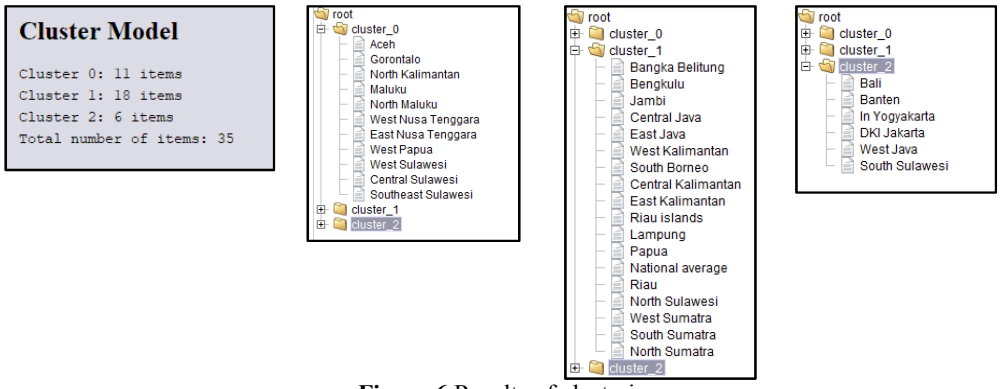

Figure 6. Results of clustering

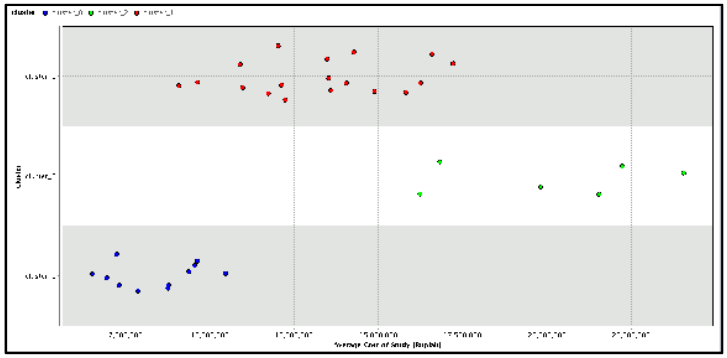

Figure 7. Visualization of clustering results with scatter plotter

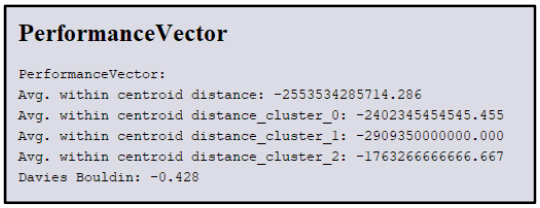

Figure 8. Performance Vector Results

In figure 8, the validity test is performed with the Davies-Bouldin Index (DBI). By using DBI, the cluster results obtained are more optimal by maximizing the number of clusters (k). for the number of k = 3 the result is 0.428 which has been previously compared with the number of k = 2 the result is 0.451. So that in the process of clustering the cost of education at the tertiary level in Indonesia by region using 3 clusters.

Conclusion

Based on the results of research at the tertiary level of education in Indonesia, the K-Medoids method can be applied and implemented by producing mapping in the form of clustering where the number of clusters used is 3 labels namely C1: high cluster label consisting of 6 regions (Bali, Banten, DI Yogyakarta, DKI Jakarta, West Java, South Sulawesi); C2: normal cluster label consisting of 18 regions (Bangka Belitung, Bengkulu, Jambi, Central Java, East Java, West Kalimantan, South Borneo, Central Kalimantan, East Kalimantan, Riau islands, Lampung, Papua, National average, Riau, North Sulawesi, West Sumatra, South Sumatra, North Sumatra) and C3: low cluster labels consisting of 11 regions (Aceh, Gorontalo, North Kalimantan, Maluku, North Maluku, West Nusa Tenggara, East Nusa Tenggara, West Papua, West Sulawesi, Central Sulawesi, Southeast Sulawesi Sulawesi). The final centroid results obtained for cluster C1 are 20.530.000; C2 is 12.660.000 and C3 is 7.130.000 with the best Davies Bouldin (DBI) 0,428.

DOI: 10.5373/JARDCS/V12I6/S20201169
*Corresponding Author: Imam Makruf
Article History: Received: Mar 28, 2020, Accepted: June 30, 2020

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] 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.

[3] D. Sun, H. Fei, and Q. Li, "A Bisecting K-Medoids clustering Algorithm Based on Cloud Model," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 308–315, 2018, doi: 10.1016/j.ifacol.2018.08.301.

[4] H. Fei, N. Meskens, and C. H. Moreau, "Clustering of patient trajectories with an auto-stopped bisecting K-medoids algorithm," *IFAC Proc. Vol.*, vol. 13, no. PART 1, pp. 355–360, 2009, doi: 10.3182/20090603-3-RU-2001.0281.

[5] E. M. Rangel, W. Hendrix, A. Agrawal, W. K. Liao, and A. Choudhary, "AGORAS: A fast algorithm for estimating medoids in large datasets," *Procedia Comput. Sci.*, vol. 80, pp. 1159–1169, 2016, doi: 10.1016/j.procs.2016.05.446.

[6] 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.

[7] F. R. Senduk and F. Nhita, "Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm," *Ind. J. Comput.*, vol. 4, no. 2016, pp. 65–76, 2019, doi: 10.21108/indojc.2019.4.3.359.

[8] 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.

[9] 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, doi: 10.24071/ijasst.v1i1.1859.

[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] N. Aggarwal, K. Aggarwal, and K. Gupta, "Comparative Analysis of K-means and Enhanced K-means Clustering Algorithm for Data Mining," in *International Jouranl of Scientific & Engineering Research*, 2012, vol. 3, no. 3.

[12] S. Mujiasih, "Utilization Of Data Mining For Weather Forecasting`," *J. Meteorol. dan Geofis.*, vol. 12, no. 2, pp. 189–195, 2011.

[13] Yuhefizar, B. Santosa, I. K. E. Purnama, and Y. K. Suprapto, "Combination of cluster method for segmentation of web visitors," *Telkomnika*, vol. 11, no. 1, pp. 207–214, 2013, doi: 10.12928/TELKOMNIKA.v11i1.825.

[14] E. Sugiharti and M. A. Muslim, "On-line clustering of lecturers performance of computer science department of semarang state university using K-MeansAlgorithm," *J. Theor. Appl. Inf. Technol.*, vol. 83, no. 1, pp. 64–71, 2016.

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

[16] 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, doi: 10.24014/coreit.v4i2.4498.

[17] A. Wanto *et al.*, *Data Mining : Algoritma dan Implementasi*. Medan: Yayasan Kita Menulis, 2020.

[18] 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.