Journal of Computer Science Advancements, 1(4) - August 2023 204-214

Application of K-Means Clustering Algorithm to Obtain Recommendations for Strategies to Increase the Number of Students in the Information Systems Study Program at ITB Ahmad Dahlan Jakarta

Diana Yusuf [1], Xie Guilin[2], Deng Jiao[3]
[1]*Institut Teknologi dan Bisnis Ahmad Dahlan Jakarta, Indonesia*
[2]*University of Science and Technology of Hanoi, Vietnam*
[3]*University Sains Malaysia, Malaysia*

Corresponding Author: Diana Yusuf, E-mail; dianayusuf@gmail.com

Article Information:
Received August 10, 2023
Revised August 19, 2023
Accepted August 25, 2023

ABSTRACT

The rapid development of technology today has almost touched all sectors of life such as the economy, health and education. The technology currently used produces a lot of data every day, one of which is in the field of education. Data mining is a group of methods used to investigate and reveal complex relationships in very large data sets. Data here means information organized in a tabular format, as is often used in relational database management. This research uses data from the academic section of ITB Ahmad Dahlan, namely data on students of the Information Systems study program from 2019 to 2022. The attributes that will be used for this research are student gender, student employment status and student achievement index. Recommendations for promotional strategies to increase the number of new students are to conduct visits to high schools or vocational schools. Not only that, the new student admission team can also promote to companies or offices.

Keywords: *Application, Clustering, Information*

Journal Homepage https://journal.ypidathu.or.id/index.php/jcsa

How to cite: Yusuf, D., Guilin, X., & Jiao, D. (2023). Application of K-Means Clustering Algorithm to Obtain Recommendations for Strategies to Increase the Number of Students in the Information Systems Study Program at ITB Ahmad Dahlan Jakarta. *Journal of Computer Science Advancements*, 1(4). 204-214. https://doi.org/10.70177/jsca.v1i4.581
Published by: Yayasan Pendidikan Islam Daarut Thufulah

INTRODUCTION

The rapid development of technology today has almost touched all sectors of life such as the economy, health and education (Agus Triansyah dkk., 2023). The technology currently used produces a lot of data every day, one of which is in the field of education (Oztemel & Gursev, 2020). There is a lot of data in the field of education

that can be obtained due to the impact of the application and use of technology, for example data on a college.

Every college today must utilize technology in its operational activities. Every year the college will conduct graduation and new student admissions, this will indirectly increase the amount of data in the college continuously (Belanche dkk., 2020). The process of admitting new students is of course one of the routine activities carried out by universities, there is a lot of data from prospective students who register (Archibald dkk., 2019). Then the data will change and increase when the student has been accepted to the college, for example there will be data on name, number, major, and score (Safiri dkk., 2020). This data will increase until the student completes his education.

Based on this explanation, it can be seen that there is an abundance of data generated by a university, starting from the beginning of the new student admission process until the student graduates (Thangaramya dkk., 2019). So if the myriad of data is processed by the college, it will provide benefits for the institution itself. The process of processing student data will provide new knowledge for institutions, for example, such as universities can gain knowledge based on the grouping of students who work or do not work (Ben-Daya dkk., 2019). This can be utilized to become one of the promotional strategies for universities to obtain new students.

Based on data obtained from ITB Ahmad Dahlan academics, especially for the Information Systems study program, the number of students from the class of 2019 to 2022 is 292 students (Y. Wang dkk., 2019). When viewed from this data, the number of students in the Information Systems study program is still quite small, because indeed this study program is still quite new, namely starting to accept students in 2019.

Therefore (Hüllermeier & Waegeman, 2021), this research will conduct clustering of Ahmad Dahlan ITB Information Systems study program students (Chaabouni dkk., 2019). The attributes that will be used in the research are gender, employment status, and IP (Achievement Index) of students.

The objectives of the research that will be carried out are:

1. Grouping data of Information Systems study program students based on their employment status and academic potential.
2. Obtaining recommendations for study program promotion strategies for each cluster obtained.

THEORETICAL FOUNDATION

Data mining is a group of methods used to investigate and reveal complex relationships in very large data sets (Giordani dkk., 2019). By data, we mean information organized in a tabular format, as is often used in relational database management. However, data mining techniques can also be implemented on various other types of data representations, including spatial domain data, text, and multimedia (Arhami & Nasir, 2020).

Clustering is a process of grouping a set of objects into similar classes (Rodriguez dkk., 2019). A cluster is a set of data objects that are similar in the same cluster, but

different from objects in other clusters. A set of data objects can be interpreted as a group and can also be considered a form of data compression.

The k-means clustering algorithm requires an input parameter usually called k, and partitions n objects into k clusters aiming to achieve a high degree of similarity within each cluster and low similarity between clusters (Zhang dkk., 2019). The degree of similarity is measured based on the average value of objects within each cluster, which can be considered as the cluster centroid.

The steps of the k-means algorithm are as follows. First, k objects are randomly selected that initially represent the average or cluster centroid (Tang dkk., 2019). Then, each remaining object is placed into the cluster that is most similar to it based on the distance between it and the cluster average. After that, a new average is calculated for each cluster (Shahapure & Nicholas, 2020). This process is repeated repeatedly until the criterion function reaches the point of convergence (Rodriguez dkk., 2019). According to (Prasetyo, 2012) the K-Means Clustering algorithm has several completion steps, including:

1. Determine the value of k as the number of clusters to be formed.
2. Initializing the k cluster centers can be done in various ways, but most often by randomly drawing from existing data.
3. Calculate the distance from each input datum to each centroid using the Euclidean distance formula until the closest distance from each datum to the centroid is found. Here is the Euclidean distance equation:

$$De = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2}$$

4. Classify each value above based on its proximity to the centroid (minimum distance).
5. Update the centroid value. The new centroid value is taken from the cluster mean mentioned by formula :

$$v_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj}$$

6. Repeat from steps 2 to 5, until the members of each cluster have not changed.

RESEARCH METHODOLOGY

This research methodology consists of steps that will be taken to solve the problems in this study (Modi & Dunbrack, 2019). The research steps must be clear and structured, these steps include the following:

Data Collection

In this step, researchers collect data that will be processed using 3 (three) methods, namely observation, interviews and literature studies (Peng & Liu, 2019). Observations and interviews were conducted directly with the academic section of ITB Ahmad Dahlan.

Needs Analysis

After all the data is collected, the next step is that the researcher will analyze the data. This is useful so that the expected results match the needs and provide accurate results.

Application of the K-Means Algorithm

In this step (Dong dkk., 2020), researchers will apply the k-means clustering algorithm to student data that has been collected previously (Kang dkk., 2019). The application of the k-means algorithm can be done using the Rapidminer application and also manual calculations.

Results

The results of data processing using the k-means algorithm are said to be successful if they provide new knowledge or knowledge that can be useful for determining the promotion strategy for new student admissions to the information systems study program at ITB Ahmad Dahlan.

RESULT AND DISCUSSION

This research uses data from the academic section of ITB Ahmad Dahlan, namely data on students of the Information Systems study program from 2019 to 2022 (Zhan dkk., 2019). The attributes that will be used for this research are student gender, student employment status and student achievement index (Jain dkk., 2019). Clustering is done to group Information Systems study program students based on their employment status in order to obtain a pattern of future promotion strategies.

Tabel 1. Data Cleaning

Student	IP	Working Status
MUHAMMAD ARIFIN AL FATIR	2,19	Not Working
DICKY NOVRIANTO	3,46	Working
OMAN ROHMAN	2,99	Not Working
ELSANISSA FASYAH	3,19	Working
AGUNG DWI PRIYANTO	3,34	Working
RASYID BUDI RAMADHAN	3,11	Working
ARDIANSYAH PRABOWO	3,58	Working
MUHAMMAD KHAIRULLAH	2,87	Not Working
MUHAMMAD NAUFAL AB`ROR	2,79	Working
FARHAN FARISI	2,87	Not Working
….	….	....
….	….	....
MUHAMMAD FAHLEFY	3,61	Working

The next stage is to transform nominal data such as gender and employment status into numeric form by initialization (C. Wang dkk., 2020). Initialization can be seen as in the table below:

Table 2. Gender Data Initialization

Student	Initialize
MUHAMMAD ARIFIN AL FATIR	1
DICKY NOVRIANTO	1
OMAN ROHMAN	1
ELSANISSA FASYAH	2
AGUNG DWI PRIYANTO	1
RASYID BUDI RAMADHAN	1
ARDIANSYAH PRABOWO	1
MUHAMMAD KHAIRULLAH	1
MUHAMMAD NAUFAL AB`ROR	1
FARHAN FARISI	1
….	….
….	….
MUHAMMAD FAHLEFY	1

Table 3 Job Status Data Initialization

Student	Employment Status	Initialize
MUHAMMAD ARIFIN AL FATIR	Not Working	1
DICKY NOVRIANTO	Working	2
OMAN ROHMAN	Not Working	1
ELSANISSA FASYAH	Working	2
AGUNG DWI PRIYANTO	Working	2
RASYID BUDI RAMADHAN	Working	2
ARDIANSYAH PRABOWO	Working	2
MUHAMMAD KHAIRULLAH	Not Working	1
MUHAMMAD NAUFAL AB`ROR	Working	2
FARHAN FARISI	Not Working	1
….	....	….
….	....	….
MUHAMMAD FAHLEFY	Working	2

The initialized data above can be used and processed using the k-means clustering algorithm (Al-Fraihat dkk., 2020). The following dataset will be processed with clustering techniques:

Table 4. Student dataset

JK	**IP**	Employment Status
1	2,19	1
1	3,46	2
1	2,99	1
2	3,19	2
1	3,34	2
1	3,11	2
1	3,58	2
1	2,87	1
1	2,79	2
1	2,87	1
….	….	….
….	….	….
1	3,61	2

The above dataset that has been transformed into numbers is ready to be grouped using the k-means clustering algorithm (Coman dkk., 2020). But to do the algorithm process needs to be done first, namely:

1. Determine the number of clusters to be formed, in this study the data will be grouped into two clusters.
2. Determine the initial center point (centroid) for each cluster. The initial centroid can be determined randomly (Abbasi dkk., 2019). The initial center point in this study can be seen in the table below.

Table 5. Initial Centroid for Each Cluster

Cluster	JK	IP	Employment Status
Cluster 1	1	3,58	2
Cluster 2	1	2,87	1

The next step after determining the number of clusters and initial centroid is to calculate the closest distance to the initial centroid of each cluster. The calculation can use the euclidean distance formula (Wortham dkk., 2020). The calculated distance starts from the first student data to the first cluster center to the last data. For example, the calculation below is the first data to the first cluster center point:

$$D(1,1) = \sqrt{(1 - 1)^2 + (2{,}19 - 3{,}58)^2 + (1 - 2)^2} = 1{,}71233$$

$$D(1,2) = \sqrt{(1 - 1)^2 + (3{,}46 - 3{,}58)^2 + (2 - 2)^2} = 0{,}12$$

$$D(1,3) = \sqrt{(1 - 1)^2 + (2{,}99 - 3{,}58)^2 + (1 - 2)^2} = 1{,}1610$$

Calculation of the closest distance is carried out until all data, namely 292 data to the cluster that has been determined (Albrecht & Chin, 2020). The results of the calculation of the closest distance of iteration 1 are as in the following table.

Table 6. Results of Data Calculation to Each Cluster

No	JK	IP	Employment Status	Distance to		Closest Distance to Cluster
				C1	C2	
1	1	2,19	1	1,7123	0,68	C2
2	1	3,46	2	0,12	1,1610	C1
3	1	2,99	1	1,1610	0,12	C2
4	2	3,19	2	1,0733	1,4499	C1
5	1	3,34	2	0,24	1,1049	C1
6	1	3,11	2	0,47	1,0283	C1
7	1	3,58	2	0	1,2264	C1
8	1	2,87	1	1,2264	0	C2
9	1	2,79	2	0,79	1,0031	C1
10	1	2,87	1	1,2264	0	C2
….	….	….	….			….
….	….	….	….			….
292	1	3,61	2	0,03	1,2440	C1

It can be seen in the table above that all data has been placed into the nearest cluster, then the next step is to determine the new center point based on the average members in the cluster (Reed dkk., 2019). Determining the new center point can be done using the calculation below:

$$C1\,(JK) = \frac{1+2+1+1+1+1+2+1+2+2+1+1+1+1+1+1}{15}$$

Here is the new cluster center point determined using the formula above. The new center point will be used to calculate the closest distance to the cluster in the 2nd iteration.

Table 7. New Center Point

Cluster	JK	IP	Employment Status
Cluster 1	1,2666	3,3013	2
Cluster 2	1,2727	3,3990	0,9545

The next step is to recalculate the closest distance from each data to the existing cluster using the newly formed center point above. The calculation results can be seen in the table below:

Table 8. Data Calculation Results 2nd Iteration

No	JK	IP	Employment Status	Distance to		Closest Distance to Cluster
				C1	C2	
1	1	2,19	1	1,5186	1,2403	C2
2	1	3,46	2	0,3103	1,0821	C1
3	1	2,99	1	1,0807	0,4937	C2
4	2	3,19	2	0,7417	1,2905	C1
5	1	3,34	2	0,2694	1,0820	C1
6	1	3,11	2	0,3282	1,1184	C1
7	1	3,58	2	0,3857	1,0954	C1
8	1	2,87	1	1,1212	0,5969	C2
9	1	2,79	2	0,5766	1,2403	C1
10	1	2,87	1	1,1212	0,5969	C2
….	….	….	….	…..	….	….
….	….	….	….	….	….	….
292	1	3,61	2	0,4079	1,1008	C1

In this study, clustering iterations stopped at iteration 2. This is because at iteration 2 the center point of each cluster has not changed and there is no movement of data from one cluster to another. The results of cluster 1 show that it is dominated by male students with working status.

Table 9. Clustering Analysis Results

Cluster 1 Result	Cluster 2 Result
Cluster 1 consists of 112 students, which consists of: LK = 92, P = 20 In addition, in cluster 1 all students are working students with an average Grade Point Average (IP) of 3.301.	Cluster 2 consists of 180 students, which consists of: LK = 150, P = 30 Students in cluster 2 are students who do not work with an average Grade Point Average (IP) of 3, 40

CONCLUSION

After clustering student data with attributes of employment status and academic potential using the k-means clustering algorithm, two clusters were formed. Where cluster 1 consists of 112 students where all students work with an average Presentation Index of 3.30 which is dominated by male students. While cluster 2 obtained a total of 180 students with a non-working employment status and an average Presentation Index of 3.40 which is dominated by male students.

Recommendations for promotional strategies to increase the number of new students are to conduct visits to high schools or vocational schools. Not only that, the new student admission team can also promote to companies or offices.

ACKNOWLEDGEMENT

The author would like to thank all parties who have been involved in the process of completing this journal. And thank you to the academic department of ITB Ahmad

Dahlan for being willing to provide the data and information that the author needs to complete this scientific work.

REFERENCES

Abbasi, S., Keshavarzi, B., Moore, F., Turner, A., Kelly, F. J., Dominguez, A. O., & Jaafarzadeh, N. (2019). Distribution and potential health impacts of microplastics and microrubbers in air and street dusts from Asaluyeh County, Iran. *Environmental Pollution*, *244*, 153–164. https://doi.org/10.1016/j.envpol.2018.10.039

Agus Triansyah, F., Hejin, W., & Stefania, S. (2023). Factors Affecting Employee Performance: A Systematic Review. *Journal Markcount Finance*, *1*(2), 118–127. https://doi.org/10.55849/jmf.v1i2.102

Albrecht, E., & Chin, K. J. (2020). Advances in regional anaesthesia and acute pain management: A narrative review. *Anaesthesia*, *75*(S1). https://doi.org/10.1111/anae.14868

Al-Fraihat, D., Joy, M., Masa'deh, R., & Sinclair, J. (2020). Evaluating E-learning systems success: An empirical study. *Computers in Human Behavior*, *102*, 67–86. https://doi.org/10.1016/j.chb.2019.08.004

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*, 160940691987459. https://doi.org/10.1177/1609406919874596

Belanche, D., Casaló, L. V., Flavián, C., & Schepers, J. (2020). Service robot implementation: A theoretical framework and research agenda. *The Service Industries Journal*, *40*(3–4), 203–225. https://doi.org/10.1080/02642069.2019.1672666

Ben-Daya, M., Hassini, E., & Bahroun, Z. (2019). Internet of things and supply chain management: A literature review. *International Journal of Production Research*, *57*(15–16), 4719–4742. https://doi.org/10.1080/00207543.2017.1402140

Chaabouni, N., Mosbah, M., Zemmari, A., Sauvignac, C., & Faruki, P. (2019). Network Intrusion Detection for IoT Security Based on Learning Techniques. *IEEE Communications Surveys & Tutorials*, *21*(3), 2671–2701. https://doi.org/10.1109/COMST.2019.2896380

Coman, C., Țîru, L. G., Meseșan-Schmitz, L., Stanciu, C., & Bularca, M. C. (2020). Online Teaching and Learning in Higher Education during the Coronavirus Pandemic: Students' Perspective. *Sustainability*, *12*(24), 10367. https://doi.org/10.3390/su122410367

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Giordani, M., Polese, M., Roy, A., Castor, D., & Zorzi, M. (2019). A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies. *IEEE Communications Surveys & Tutorials*, *21*(1), 173–196. https://doi.org/10.1109/COMST.2018.2869411

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Jain, N., Brock, J. L., Malik, A. T., Phillips, F. M., & Khan, S. N. (2019). Prediction of Complications, Readmission, and Revision Surgery Based on Duration of Preoperative Opioid Use: Analysis of Major Joint Replacement and Lumbar Fusion. *Journal of Bone and Joint Surgery*, *101*(5), 384–391. https://doi.org/10.2106/JBJS.18.00502

Kang, Z., Wen, L., Chen, W., & Xu, Z. (2019). Low-rank kernel learning for graph-based clustering. *Knowledge-Based Systems*, *163*, 510–517. https://doi.org/10.1016/j.knosys.2018.09.009

Modi, V., & Dunbrack, R. L. (2019). Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences*, *116*(14), 6818–6827. https://doi.org/10.1073/pnas.1814279116

Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, *31*(1), 127–182. https://doi.org/10.1007/s10845-018-1433-8

Peng, X., & Liu, L. (2019). Information measures for *q* -rung orthopair fuzzy sets. *International Journal of Intelligent Systems*, *34*(8), 1795–1834. https://doi.org/10.1002/int.22115

Reed, G. M., First, M. B., Kogan, C. S., Hyman, S. E., Gureje, O., Gaebel, W., Maj, M., Stein, D. J., Maercker, A., Tyrer, P., Claudino, A., Garralda, E., Salvador-Carulla, L., Ray, R., Saunders, J. B., Dua, T., Poznyak, V., Medina-Mora, M. E., Pike, K. M., … Saxena, S. (2019). Innovations and changes in the ICD-11 classification of mental, behavioural and neurodevelopmental disorders. *World Psychiatry*, *18*(1), 3–19. https://doi.org/10.1002/wps.20611

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Safiri, S., Kolahi, A.-A., Smith, E., Hill, C., Bettampadi, D., Mansournia, M. A., Hoy, D., Ashrafi-Asgarabad, A., Sepidarkish, M., Almasi-Hashiani, A., Collins, G., Kaufman, J., Qorbani, M., Moradi-Lakeh, M., Woolf, A. D., Guillemin, F., March, L., & Cross, M. (2020). Global, regional and national burden of osteoarthritis 1990-2017: A systematic analysis of the Global Burden of Disease Study 2017. *Annals of the Rheumatic Diseases*, *79*(6), 819–828. https://doi.org/10.1136/annrheumdis-2019-216515

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Tang, C., Zhu, X., Liu, X., Li, M., Wang, P., Zhang, C., & Wang, L. (2019). Learning a Joint Affinity Graph for Multiview Subspace Clustering. *IEEE Transactions on Multimedia*, *21*(7), 1724–1736. https://doi.org/10.1109/TMM.2018.2889560

Thangaramya, K., Kulothungan, K., Logambigai, R., Selvi, M., Ganapathy, S., & Kannan, A. (2019). Energy aware cluster and neuro-fuzzy based routing algorithm for wireless sensor networks in IoT. *Computer Networks*, *151*, 211–223. https://doi.org/10.1016/j.comnet.2019.01.024

Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., & Ho, R. C. (2020). Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General

Population in China. *International Journal of Environmental Research and Public Health*, *17*(5), 1729. https://doi.org/10.3390/ijerph17051729

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2019). Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Transactions on Smart Grid*, *10*(3), 3125–3148. https://doi.org/10.1109/TSG.2018.2818167

Wortham, J. M., Lee, J. T., Althomsons, S., Latash, J., Davidson, A., Guerra, K., Murray, K., McGibbon, E., Pichardo, C., Toro, B., Li, L., Paladini, M., Eddy, M. L., Reilly, K. H., McHugh, L., Thomas, D., Tsai, S., Ojo, M., Rolland, S., … Reagan-Steiner, S. (2020). Characteristics of Persons Who Died with COVID-19—United States, February 12–May 18, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*(28), 923–929. https://doi.org/10.15585/mmwr.mm6928e1

Zhan, K., Nie, F., Wang, J., & Yang, Y. (2019). Multiview Consensus Graph Clustering. *IEEE Transactions on Image Processing*, *28*(3), 1261–1270. https://doi.org/10.1109/TIP.2018.2877335

Zhang, R., Chen, Z., Chen, S., Zheng, J., Büyüköztürk, O., & Sun, H. (2019). Deep long short-term memory networks for nonlinear structural seismic response prediction. *Computers & Structures*, *220*, 55–68. https://doi.org/10.1016/j.compstruc.2019.05.006