International Journal of Cyber and IT Service Management (IJCITSM) p-ISSN: 2797-1325
Vol. 3 No. 1 April 2023 e-ISSN: 2808-554X

Data Mining Methods: K-Means Clustering Algorithms

Mohammad Annas[1], Siti Norida Wahab[2]
Faculty of Business, Multimedia Nusantara University[1]
Faculty of Business Management, Universiti Teknologi MARA[2]
Indonesia[1], Malaysia[2]
e-mail: mohammad.annas@umn.ac.id[1], sitinorida@ucsiuniversity.edu.my[2]

Author Notification
03 March 2023
Final Revised
13 March 2023
Published
29 March 2023

To cite this document:
Annas, M. ., & Wahab, S. N. . (2023). Data Mining Methods: K-Means Clustering Algorithms. International Journal of Cyber and IT Service Management, 3(1), 40–47. *Retrieved from https://iiast.iaic-publisher.org/ijcitsm/index.php/IJCITSM/article/view/122*

DOI:
https://doi.org/10.34306/ijcitsm.v3i1.122

Abstract

A data warehouse is a straightforward definition of a database. Data mining technology can be used to process mountains of data in databases to uncover new, fascinating, and useful information.Clustering is an approach to data gathering. As one technique for grouping data into clusters or groups, the K-Means Clustering Algorithm algorithm divides the data into those that share the cluster's traits and those that don't. data into groups, and data into groups, so that data into groups, and data into groups, so that data has the same traits is grouped in the same cluster. Other clusters are formed from data and clusters with distinct properties. additional categories. The knowledge/information gathered in the groups or clusters is helpful to policy consumers in the decision-making process.mact of making decisions.

Keywords*: Data Mining, Clustering, K-Means Clustering Algorithm*

1. Introduction

A large amount of data has been generated by the advancement of information technology, which is now becoming more complex. enormous amounts of information [1]. What can be done with all of that data will be a pressing question brought on by the explosion of data. To address this query Data mining, a database technique, can be used to find the solution to this query [2]. You can use data mining to glean knowledge from a data set that cannot be manually discovered, adding value to the data set. that has not previously been manually known [3].

Data mining employs a variety of approaches, clustering being one of them. Hierarchical clustering and non-hierarchical clustering are the two types of clustering techniques used in data grouping [4]. The data are divided into one or more clusters using the non-hierarchical data clustering approach known as K-means clustering . data is organized into groups or clusters, with similar features being put together into one cluster and differing attributes being sorted into various groups. Users of policies can use the groupings or clusters of knowledge/information to help them make decisions. decision-makers should involve policy users [5], [6].

2. Research Method

Mind Mapping/Concept Mapping

The bottom-level directive's purpose is to make it easier for us to understand the subject matter that is covered in this matrix [7]. As an example, consider the following macrocosmic principle:

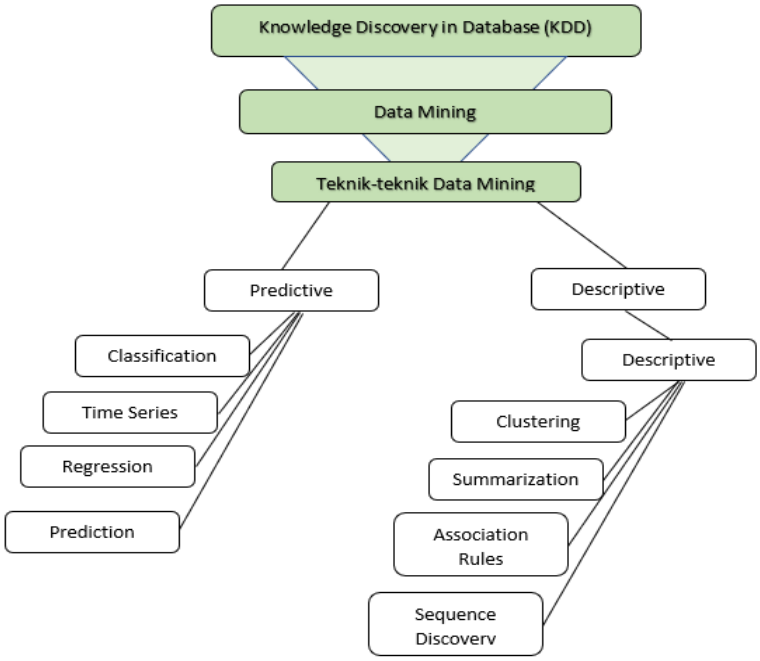

Figure 1. Mind Mapp/Concept Map Paper

Knowledge Discovery in Database (KDD)

Data mining is a term that refers to the process of extracting knowledge from databases (KDD), which is also known as large-scale knowledge discovery [8]. It is frequently referred to as "data mining" to extract useful information from huge amounts of data. KDD is a systematic method for extracting relevant patterns from vast and complicated data sets that are useful and understandable. The KDD method' primary component is data mining, which uses inferring algorithms to sift through data, create models, and spot patterns that weren't there before. known. These models are employed to evaluate, predict, and comprehend occurrences from the data [9][10], [11]. Figure 2 below provides an illustration of the Knowledge Discovery in Databases (KDD) process:

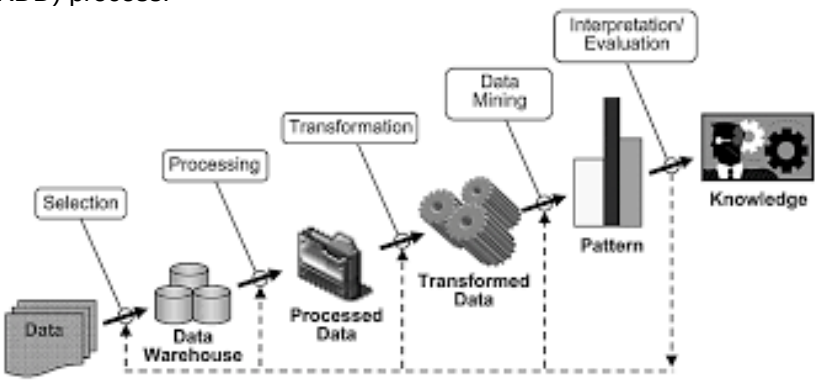

Figure 2. Knowledge Discovery Process in the Database

2.2 Literature Review

By rearranging data objects into a collection of connected classes, or clusters, Madhu Yedha defines clustering [12], [13]. Unsupervised classification example. Data items are categorized via a process called classification. Because the clustering is unsupervised, it is not reliant on training or class norms.

Deka asserts that one data mining approach used to obtain groups of items is clustering. if the data is vast enough, find groupings of items that share attributes.traits in sufficient amounts

of data. The clustering method's major objective is to organize various data or objects into clusters or groups so that each cluster will have data that is similar to the data in the other clusters.There will be as much similarity as possible among the data in the cluster [14]. Clustering is a type of unsupervised learning strategy since it groups data based on similarities between items.as a strategy for unsupervised learning [15]. Hierarchical clustering and non-hierarchical clustering are two ways that clustering might be divided, in Oyelade's opinion.

A method for grouping two or more related objects together is the first step in the hierarchical clustering approach for data clustering [16]. the closest similarity between two or more things is grouped. Once another object has a second proximity, the procedure moves on to it. Continue in this manner until the cluster resembles a tree with a distinct hierarchy (level) between objects, starting with the closest match and moving down to the second [17], [18]. order of similarity between items, from most to least. Everything will eventually group together into a single cluster, according to logic [19]. The hierarchical approach is typically made clearer via dendrograms.

Unlike the hierarchical clustering approach, which begins by deciding on the desired number of groups, the non-hierarchical clustering method actually does the opposite (two clusters, three clusters, or so on) [20], [21]. The clustering procedure is carried out without using a hierarchical structure after the number of clusters is known. K-Means Clustering is the name given to this technique[22], [23].

3. Findings

3.1 Problem

Algoritma K-means Clustering

A non-hierarchical cluster analysis technique called K-means clustering aims to divide existing items into one or more clusters. aims to combine objects with similar qualities together into one or more clusters or groups of objects, in order to achieve the goal of grouping similar objects together. Clusters are formed from things with similar properties and sub-clusters from objects with differing characteristics. are categorized into additional groups according to their different traits.

Table 1. Student data

No	Name	Department	Hometown	IPK
1	Ade firdaus	SI	Jakarta	3,16
2	Husnul Khotimah	SK	Semarang	3,22
3	Ahmad Fahrizki	TI	Bekasi	3,29
4	Seila Ananda	BD	Jakarta	2,83
5	Istianti Dewi	MM	Jakarta	3,15
6	Farhan Mustofa	BD	Banda Aceh	3,25
7	Adi Friandi	BD	Bogor	3,43
8	Putri Andini	MM	Bekasi	3,06
9	Deswita Putri	TI	Banda Aceh	3,36
10	Solihihudin	SK	Bandung	3,28

Data Transformation

The aforementioned data must be initialized before it can be processed with the k-means clustering technique. The initialization of data in the form of numbers must be done for items like hometown and major initially.

Table 2. Home City Region Data Initialization

Region	Frequency	Initials
Jakarta	84	1
Jawa Barat	82	2
Sumatera Utara	28	3

Sulawesi	14	4
Jawa Timur	13	5
Sumatera Selatan	13	6
Bali	8	7
Kalimantan	1	8

Table 3. Department Data Initialization

Department	abbreviation	Frequency	Initial
Sistem Informasi	SI	46	1
Sistem Komputer	SK	37	2
Teknik Informatika	TI	35	3
Bisnis Digital	BD	28	4
Multi Media	MM	23	5
Industrial Engineering	IE	20	6
Informasi Technology	IT	18	7
Accounting	ACC	12	8
International Relation	IR	9	9
Public Relation	PR	6	10
Visual Communication Design	VCD	4	11
Electrical Engineering	EE	2	12
Business Administration	BA	1	13
Management, concentration in Human Resources Management	HRM	1	14
Management	MGT	1	15

3.2 Research Implementation

Data processing

The K-Means Clustering technique can now be used to integrate all student data that has been translated to angular format. There are multiple stages that must be followed in order to divide the data into various clusters, including:

1. Decide how many clusters you want. Existing data for this study will be divided into three groupings.
2. The initial center of each cluster should be identified. The initial focal point of this investigation is As can be seen in table 2.4, each cluster's center point was chosen at random and determined.

Table 4. Initial Center Point of Each Cluster

Center Point Originally	Name	Department	Hometown	IPK
Cluster 1	Dally Teguh Sesario	9	3	2,94
Cluster 2	Hervina Juliana	1	1	3,18
Cluster 3	Pascal Muhammadi	1	2	3,15

3. Cluster every piece of data. Each piece of data is assigned into a cluster in this study using the hard k-means approach, which places the data in the cluster with the closest proximity to the point. each cluster's geographic centroid. Calculating the distance between each data point and each cluster's center point is important to determine which cluster is closest to the data. As an illustration, we will figure out how far the first student's data are from the first cluster's center:

$$D(1,1) = \sqrt{(14-9)^2 + (1-3)^2 + (3{,}16-2{,}94)^2} = 5{,}390$$

From the above calculation results, it is found that the distance of the first student data with the center of the first cluster is 5.390.
The distance of the first student data to the second cluster center:

$$D(1,2) = \sqrt{(14-1)^2 + (1-1)^2 + (3{,}16-3{,}18)^2} = 13{,}000$$

From the above calculation results, it is found that the distance of the first student data to the center of the second cluster is 13.
Distance of the first student data to the third cluster center:

$$D(1,3) = \sqrt{(14-1)^2 + (1-2)^2 + (3{,}16-3{,}15)^2} = 13{,}038$$

From the above calculation results, it is found that the distance of the first student data with the center of the third cluster is 13.038.
Based on the results of the three calculations above, it can be concluded that the distance of the first student data is the closest to cluster 1. data is closest to cluster 1, so the first student data is put into cluster 1. The complete calculation results for the first 5 student data can be found in the following table calculation results for the first 5 student data can be seen in table 2.5.

Table 5. Example of Calculation Results of Each Data to Each Cluster

No	Name	Department	Hometown	IPK	Distance to			Distance closest to Cluster
					C1	C2	C3	
1	Ade firdaus	14	1	3,16	5,390	13,000	13,038	1
2	Husnul Khotimah	1	5	3,22	8,251	4,000	3,001	3
3	Ahmad Fahrizki	4	2	3,29	5,111	3,164	3,003	3
4	Seila Ananda	2	1	3,83	7,281	1,059	1,450	2
5	Istianti Dewi	3	1	3,15	6,328	2,000	2,236	2

4. After all the data is placed into the closest cluster, then recalculate the new cluster center based on the average of the members in the cluster center based on the average of the members in the cluster.
5. After obtaining the new center point of each cluster, do it again from step three until the center point of each cluster does not change anymore and there are no clusters. step three until the center point of each cluster no longer changes and there is no more data moving from one cluster to another. data moving from one cluster to another.

In this study, iteration of student data clustering occurred 7 times. In this 7th iteration, the center point of each cluster has not changed and there is no more data moving from one cluster to another. data moving from one cluster to another.

From the results of cluster 1, it can be seen that the characteristics of students in cluster 1 are dominated by students who come from Information Technology and Marketing majors. Meanwhile, based on the city of origin, it is dominated by students who come from the Jakarta and West Java, so it can be concluded that the average student in cluster 1 who comes from the city of origin of DKI Jakarta and West Java majors in Information Technology and Marketing. West Java majoring in Information Technology and Marketing.

Table 6. Clustering Analysis Results

Cluster 1 Result	Cluster 2 Result	Cluster 3 Result
Cluster 1 consists of 70 people, who come from majors IT = 19 people MKT = 15 people VCD = 12 people HTM = 9 people EE = 6 people BA = 4 people IR = 2 people MGT = 1 person IS = 1 person HRM = 1 person	Cluster 2 consists of 132 people, who came from activists ACC = 39 people IB = 30 people BF = 22 people PR = 21 people IE = 20 people	Cluster 3 consists of 41 people, who come from the department: PR = 14 people ACC = 7 people IB = 7 people BF = 6 people E-3 = 3 people MKT = 3 people TI = 1 people
And come from the Region: DKI Jakarta = 30 people West Java = 20 people North Sumatra = 12 people Sulawesi = 2 people East Java = 2 people South Sumatra = 2 people Bali = 1 person Kalimantan = 1 person With an average GPA of 3.2	And came from the Region: West Java = 62 people DKI Jakarta = 54 people North Sum atra = 16 people	And came from the Region: Sulawesi = 12 people. East Java = 11 people South Sumatra = 11 people Bali = 7 people
With an average GPA of 3.2	With an average GPA of 3.25	With an average GPA of 3.31

Then, from the results of cluster 2 above, it can be seen that the characteristics of students in cluster 2 are dominated by students who come from Accounting and Business majors. International Business. Meanwhile, based on the city of origin, it is dominated by students who Jakarta and West Java, so it can be concluded that the average student in cluster 2 who comes from the city of origin of DKI Jakarta and West Java. that the average student in cluster 2 who comes from the hometown area of DKI Jakarta and West Java majors in Information Technology. and West Java majored in Information Technology and Marketing.

Meanwhile, from the results of cluster 3 above, it can be seen that the characteristics of students in cluster 3 are dominated by students who come from Public Relations, Accounting, and Public Relations majors. and International Business. Meanwhile, based on the city of origin, it is dominated by students who come from the cities of Sulawesi, East Java and South Sumatra, so that it can be concluded that the average student in cluster 3 who comes from it can be concluded that the average student in cluster 3 who comes from the city of Sulawesi, East Java and South Sumatra major in Public Relations, Accounting and International Business.

4. Conclusion

A non-hierarchical cluster analysis technique called K-means clustering aims to divide existing objects into one or more clusters or groups of objects. to organize or cluster existing objects according to their properties. based on their traits, with the goal of grouping objects with similar traits together into clusters. items with similar traits are placed in one cluster, whereas objects with differing traits are grouped into multiple clusters. items with similar traits are placed

into one cluster, whereas objects with differing traits are grouped into multiple clusters. When employed in decision support, the resulting clusters may offer fresh and intriguing information.

References

[1] S. Rahayu and J. J. Purnama, "Klasifikasi Konsumsi Energi Industri Baja Menggunakan Teknik Data Mining," *Jurnal Teknoinfo*, vol. 16, no. 2, p. 395, 2022, doi: 10.33365/jti.v16i2.1984.

[2] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[3] D. F. al Husaeni and A. B. D. Nandiyanto, "Mapping visualization analysis of computer science research data in 2017-2021 on the google scholar database with vosviewer," *International Journal of Informatics Information System …*, vol. 3, no. 1, pp. 1–18, 2022.

[4] Mulyati, S. Zebua, M. H. R. Chakim, and Khairul, "Effect of Human Resources Quality, Performance Evaluation, and Incentives on Employee Productivity at Raharja High School," *APTISI Transactions on Management (ATM)*, vol. 7, no. 1, pp. 1–7, 2022, doi: 10.33050/atm.v7i1.1732.

[5] M. S. Yang and K. P. Sinaga, "A feature-reduction multi-view k-means clustering algorithm," *IEEE Access*, vol. 7, pp. 114472–114486, 2019, doi: 10.1109/ACCESS.2019.2934179.

[6] A. Cahyono and Y. D. Nurcahyanie, "Identification and Evaluation of Logistics Operational Risk Using the FMEA Method at PT . XZY," vol. 5, no. 1, pp. 1–10, 2023.

[7] S. Purnama, Q. Aini, U. Rahardja, N. P. L. Santoso, and S. Millah, "Design of Educational Learning Management Cloud Process with Blockchain 4.0 based E-Portfolio," *Journal of Education Technology*, vol. 5, no. 4, p. 628, 2021, doi: 10.23887/jet.v5i4.40557.

[8] D. Shoxboz, "the Essence of Teaching Engineering Computer," *European Journal of Research and Reflection in …*, vol. 7, no. 12, pp. 18–23, 2019, [Online]. Available: http://www.idpublications.org/wp-content/uploads/2020/01/Full-Paper-THE-ESSENCE-OF-TEACHING-ENGINEERING-COMPUTER-GRAPHICS-AS-A-GENERAL-TECHNICAL.pdf

[9] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J (Basel)*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[10] T. C. Handayanti, A. P. B. Prasetyo, and P. Widianingrum, "Tingkat Kepuasan Dan Hasil Belajar Biologi Dalam Penerapan Media Interaktif Quipper School," *Bioma : Jurnal Ilmiah Biologi*, vol. 9, no. 1, pp. 1–12, 2020, doi: 10.26877/bioma.v9i1.6030.

[11] J. Hilton *et al.*, "Identifying Student Perceptions of Different Instantiations of Open Pedagogy," *International Review of Research in Open and Distance Learning*, vol. 21, no. 4, pp. 1–14, 2020, doi: 10.19173/IRRODL.V21I4.4895.

[12] M. Savić, M. Ivanović, I. Luković, B. Delibašić, J. Protić, and D. Janković, "Students' preferences in selection of computer science and informatics studies a comprehensive empirical case study," *Computer Science and Information Systems*, vol. 18, no. 1, pp. 251–283, 2020, doi: 10.2298/CSIS200901054S.

[13] A. N. Halimah and H. Abdullah, "" Student preference towards the utilization of Edmodo as a learning platform to develop responsible learning environments " study," vol. 1, no. 1, pp. 53–58, 2022.

[14] F. M. Javed Mehedi Shamrat, Z. Tasnim, I. Mahmud, N. Jahan, and N. I. Nobel, "Application of k-means clustering algorithm to determine the density of demand of different kinds of jobs," *International Journal of Scientific and Technology Research*, vol. 9, no. 2, pp. 2550–2557, 2020.

[15] P. A. Sunarya, M. Ilmu, M. Administrasi, and K. Tangerang, "The Impact of Gamification on IDU (ILearning Instruction) in Expanding Understudy Learning Inspiration," vol. 1, no. 1, pp. 59–67, 2022.

[16] A. Nur Khormarudin, "Teknik Data Mining: Algoritma K-Means Clustering," *Jurnal Ilmu Komputer*, pp. 1–12, 2016, [Online]. Available: https://ilmukomputer.org/category/datamining/

[17] A. Y. Pratama, “Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk Sman 99 Jakarta Untuk Siswa / Siswi Smpn 9 Jakarta Menggunakan Decision Tree,” *Jurnal TEDC*, pp. 49–54, 2015, [Online]. Available: http://ejournal.poltektedc.ac.id/index.php/tedc/article/download/240/185

[18] V. S. Moertini, “Data Mining Sebagai Solusi Bisnis,” *Integral*, vol. 7, no. 1, pp. 44–56, 2002.

[19] N. Lutfiani, L. Meria, U. Raharja, and U. E. Unggul, “Utilization of Big Data in Educational Technology,” vol. 1, no. 1, pp. 73–83, 2022.

[20] N. Ramadhona, A. A. Putri, D. Sri, and S. Wuisan, “Students ’ Opinions of the Use of Quipper School as an Online Learning Platform for Teaching English,” vol. 1, no. 1, pp. 35–41, 2022.

[21] G. Alfarsi and A. bin Mohd Yusof, “Virtual reality applications in education domain,” *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*, vol. 1, no. 1, pp. 68–72, 2020, doi: 10.1109/ACIT50332.2020.9300056.

[22] C. S. Bangun, S. Purnama, and A. S. Panjaitan, “Analysis of New Business Opportunities from Online Informal Education Mediamorphosis Through Digital Platforms,” vol. 1, no. 1, pp. 42–52, 2022.

[23] N. N. Azizah and T. Mariyanti, “Education and Technology Management Policies and Practices in Madarasah,” vol. 1, no. 1, pp. 29–34, 2022.