JITCoS : Journal of Information Technology and Computer System
Vol. 1, No. 1; June 2025
e-ISSN: XXXX-XXXX
Published by CV. Multimedia Teknologi Kreatif

Application of K-Means Cluster Algorithm to Determine Student Achievement

Rafly Auliya Yahya[1], Reza Abdillah Siregar[2], Boy Arnol Sitompul[3]

[1]Universitas Medan Area; raflyauliayahya@gmail.com
[2]Universitas Muhammadiyah Sumatera Utara; rezasiregar98@gmail.com
[3]Universitas Prima Indonesia; boyarnolsitompul17@gmail.com

ABSTRACT

The application of the K-Means Algorithm by dividing the data into one or more groups, with data included in one group representing similarities and other groups representing differences. To assess whether or not student motivation is superior, the data from the student achievement evaluation shows the average value of each topic. Analysis, design, coding, and system testing are all included in the research steps. Evaluation is carried out to be used as a basis for the characteristics used in the calculation to ensure higher values. Information systems can achieve successful clustering classification results by including the k-means clustering method. This technique rotates the centroid distance at each iteration, forming cluster points and reducing clustering time.

Keywords : *Algoritma K-Means; Clustering; Data*

Corresponding Author:
Rafly Auliya Yahya
Universitas Medan Area
Email : raflyauliayahya@gmail.com

1. INTRODUCTION

Today's technology has developed more rapidly than ever before, making it easier for us to utilize some of its advanced features to obtain information faster. Similar to the use of artificial intelligence in computers and mobile phones, artificial intelligence simplifies our lives while drawing attention to the consequences it will cause. In addition, because alignment errors will affect students' final grades later, we use this to construct the application of the K-Means technique to determine student achievement through the grades received. Therefore, this system is needed to enable the right flow of data and information determined by the process. IT implementation in the academic environment is proven to increase the efficiency and speed of student data management (Irwan et al., 2024)

Large amounts of data can be processed using data mining techniques into meaningful information, which is often stored in a knowledge discovery database (KDD) (Darlinda et al., 2022). Clustering approaches are used to group exceptional children's data. In large data sets, clustering can be used to identify groups of comparable items as well as patterns of distribution and connection. The most important step in the clustering process is to collect patterns into relevant groups so that similarities and differences can be identified and in-depth conclusions can be drawn (Priyatman et al., 2019).

The k-means algorithm is used in the clustering approach to student achievement data in groups. One potential option for categorizing object attributes is the k-means cluster analysis method. Because of its relatively high accuracy to object size, the k-means method is relatively more scalable and effective when dealing with large numbers of objects. In addition, the order of objects has nothing to do with the k-means method (Aranda et al., 2016).

This research was conducted at related educational institutions where samples were taken from UINSU students. Educational institutions can identify groups of students who need more support and can accept further challenges in their programs by having a better understanding of the differences in these groups. Improving the overall level of education can be done by using effective learning techniques. It is hoped that the application of clustering algorithms can help in sorting and further developing levels of effective learning.

2. LITERATURE REVIEW

2.1. K-Means Clustering

K-Means The clustering process involves dividing a set of items into groups (a number of positive integers) based on shared properties or characteristics. A cluster is characterized by its mass, which is short for cluster mean. K-means is a widely used cluster analysis in data mining. This quantization technique is vector-based (Melpa et al., 2015).

Based on the similarity of each piece of data in the existing grouping, clustering divides the data into several groups (clusters). Group structure, data membership in groups, and data cohesion in groups are three factors that distinguish data grouping. Two types of groupings identified by structure are partitioning and hierarchy. A piece of data can be viewed as a group of two or more in a hierarchy. Each piece of data is partitioned into members of one group. Data membership in a group is separated into two categories: overlapping and exclusive. Data can be guaranteed to belong exclusively to one group and not to another group in an exclusive category. On the other hand, data can be grouped into many groups using overlapping categories. Clustering is separated into two categories, full and partial, based on the compactness category. It can be said that all data will be compressed into one group if all of them can be integrated into one. Data is considered to show deviant behavior if one or two of them are not included in the majority group. Deviant data can be referred to as noise, uninteresting background, or outliers (Asroni et al., 2016).

K-Means Using a non-hierarchical approach, K-means divides the current data into two or more categories. This technique divides the data into groups, placing identical traits in one group and different traits in different groups. This data is grouped to maximize the variance between groups and minimize the objective function specified in each group. The k-means technique attempts to divide the available data into several groups, each of which has unique characteristics from the other groups. The following is the basic k-means algorithm :

1. Determine the number of clusters you want to build.
2. Set k to be the randomly generated centroid at the beginning.
3. Using the Euclidean Distance equation as follows, determine the distance between each data point and each centroid :

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

4. Sort each dataset based on how close it is to its center of mass.
5. Verify the location of the new centroid (k).
6. If the position of the new centroid is different from the previous centroid, return to step 3 (Asroni et al., 2018).

To boost learning motivation and achieve more, a data processing approach based on the K-means clustering algorithm from student achievement results is proposed in this study.

2.2. Student Achievement

Student learning achievement is one of the indicators that determine the quality of education in higher education. Learning achievement acts as a detector of the extent of student development after completing their studies (Yuli Alam, 2018). Achievement will never be achieved without effort, either in the form of knowledge or skills. Achievement states the results that have been achieved, done, worked on, and so on, with results that are pleasing to the heart and obtained through hard work (Goa Wea et al., 2018). In addition, in various previous studies it was found that academic achievement is influenced by the learning motivation of students (Lutfiwati et al., 2020). Education is a learning process that aims to develop individual potential to become better and of higher quality (Asril et al., 2023).

In general, students have three strategic functions, namely as conveyors of truth, agents of change, and future generations (Betie et al., 2013). Students have the drive or motivation to carry out learning activities in college to achieve their desired learning goals (Masni, 2015). Students who have a disciplined attitude in learning are likely to apply it in their daily lives without deviating from the rules and obligations imposed on them (Winata et al., 2021). With students having a disciplined attitude in learning, students are committed to continuing to study according to lecture times, like

face-to-face lectures, and continuing to carry out the assigned learning assignments. This is because having discipline in learning will educate them to like the rules or schedules that have been set so as to achieve satisfactory results and can also motivate them to achieve their desired goals (Wirantasa, 2017). Students can also ensure maximum learning achievement because discipline is one of the factors that influences learning outcomes (Rahmawati et al., 2021) and one of the keys to achieving success and success in learning (Anas et al., 2019).

3. METHOD

This research methodology explains the steps involved in applying the research framework methodically. The graphical framework is used to illustrate the many steps, which can be understood from the needs analysis procedure to the research findings. The steps in conducting the research include determining the data analysis needs, collecting data, analyzing it using the K-Means method, processing it using RapidMiner, and reporting the findings. The steps in this research will be discussed later based on the framework, which will be clarified by Figure 1.

Figure 1. Research Framework

3.1. Data Collection Technique

1. Observation Method

   The observation method is used for data collection. Data is processed and taken from academic portals, and data cleaning is carried out.

2. Literature Study

   A literature study is conducted to obtain data that is appropriate for research. This method is carried out by searching for and understanding literature from several journals related to the process of creating a clustering model.

3.2. K-Means Clustering

One of the techniques used in non-hierarchical clustering is K-means clustering, which involves grouping identical and related items together. Grouped data shows a higher degree of similarity and a higher degree of difference from other groups. In essence, clustering is a technique for classifying or organizing a collection of items based on features or qualities that are the same as other data (Rahmalinda et al., 2022). Clustering is a data mining technique whose algorithmic work process is unsupervised. This means that without a teacher, training is no longer needed, and output is also not needed. Hierarchical clustering and non-hierarchical clustering are two types of clustering techniques used in data mining for grouping data (Sulistiyawati et al., 2021).

3.3. Design Stage

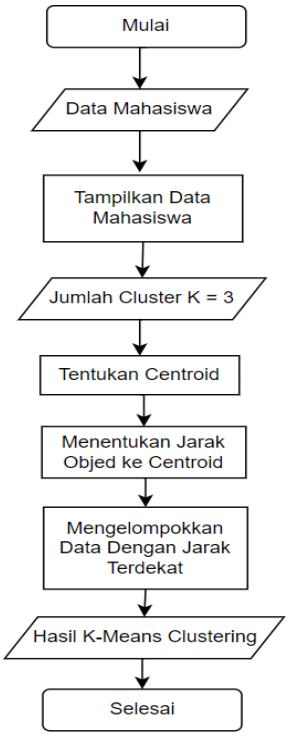

Figure 2. Flowchart K-Means

The first thing to do is to start by inputting student data into a program by determining each cluster that has been selected. From each cluster result, the centroid distance is taken from all clusters and grouped into the closest distance. After that, the results will be obtained through the program command.

4. RESULTS AND DISCUSSION

4.1 Application of K-Means Clustering

In using data grouping techniques, the first step is to determine the sample dataset to be used. The sample dataset used is 8 student data and 2 attributes, namely semester average value and number of credits per semester. From these 8 student data, they will be grouped into 3 clusters, namely not feasible (C1), feasible (C2), and very feasible (C3).

Table 1. Student Data

No	Students	Semester Average Grade	Number of Credits per Semester
1	M1	75	15
2	M2	82	18
3	M3	90	20
4	M4	65	12
5	M5	78	16
6	M6	88	19
7	M7	72	14
8	M8	95	22

After determining the sample dataset, it is necessary to determine the initial central centroid randomly. In this study, the initial central centroid uses Centroid 1, which is taken from the 1st data.

Centroid 2, which is taken from the 4th data. And Centroid 3, which is taken from the 8th data in the table below.

Table 2. Initial Centroid

Initail Centoid	Semester Average Grade	Number of Credits per Semester
Centroid 1	75	15
Centroid 2	65	12
Centroid 3	95	22

After determining the initial central centroid, the next step is to calculate using the formula below:

$$D(ij) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2}$$

Calculate Distance and Cluster Update:

1. Student M1:
   Distance to Cluster 1: 0
   Distance to Cluster 2: √((75-65)² + (15-12)²) ≈ 10.05 Distance to Cluster 3: √((75-95)² + (15-22)²) ≈ 21.54
   M1 remains in Cluster 1 because it has the shortest distance.
2. Student M2:
   Distance to Cluster 1: √((82-75)² + (18-15)²) ≈ 7.07 Distance to Cluster 2: √((82-65)² + (18-12)²) ≈ 17 Distance to Cluster 3: √((82-95)² + (18-22)²) ≈ 14.21
   M2 moves to Cluster 1 because it has the shortest distance.
3. Student M3:
   Distance to Cluster 1: √((90-75)² + (20-15)²) ≈ 15.81 Distance to Cluster 2: √((90-65)² + (20-12)²) ≈ 28.60 Distance to Cluster 3: √((90-95)² + (20-22)²) ≈ 7.07
   M3 moves to Cluster 3 because it has the shortest distance.
4. and so on for all students.

From the calculation of the nearest centroid distance above, the calculation results are obtained in the table below.

Table 3. Cluster Results

No	Students	Semester Average Grade	Number of Credits per Semester	Cluster
1	M1	75	15	1
2	M2	82	18	1
3	M3	90	20	3
4	M4	65	12	2
5	M5	78	16	3
6	M6	88	19	3
7	M7	72	14	2
8	M8	95	22	3

4.2 Implementation of K-Means Clustering in Python (Jupyter Notebook)

4.21 Import Data

The first step in utilizing the Python programming language for data analysis is to import the data. The Pandas library, which is widely used in data analysis, is used to import the data. Make sure

that the Jupyter Notebook has Pandas installed. After that, import the library by creating a code that is compatible with Jupyter Notebook so that Jupyter Notebook can use it properly. The data can be processed after the import is complete. The student dataset consists of 8 data points consisting of 2 attributes semester average grade and number of credits per semester in CSV format at the import stage data transformation.

The requirements of the data modeling process are met by this data transformation stage. The process of changing data into numeric data, categorization, and other forms is called data transformation. The results of the transformed data are shown below.

No.	Mahasiswa	Nilai Rata Rata Semester	Jumlah Kredit Per Semester
1	M1	75	15
2	M2	82	18
3	M3	90	20
4	M4	65	12
5	M5	78	16
6	M6	88	19
7	M7	72	14
8	M8	95	22

Figure 3. Initial Student Dataset

4.22 Data Selection

The process of selecting and preparing data according to the purpose of data analysis is called the data selection stage. Select the variables to be used in the grouping procedure during the data selection stage. The results below show the attribute results after the data selection process.

	No	Nilai Rata Rata Semester	Jumlah Kredit Per Semester
0	1	75	15
1	2	82	18
2	3	90	20
3	4	65	12
4	5	78	16
5	6	88	19
6	7	72	14
7	8	95	22

Figure 4. Data Selection Results

Based on the results of the data selection above, it can be seen that the number of clusters for the K-Means method is 3 clusters. This is because the number of clusters is 3. 3 of the clusters are blue, green, and red. The cluster results can be seen in Figure 4.

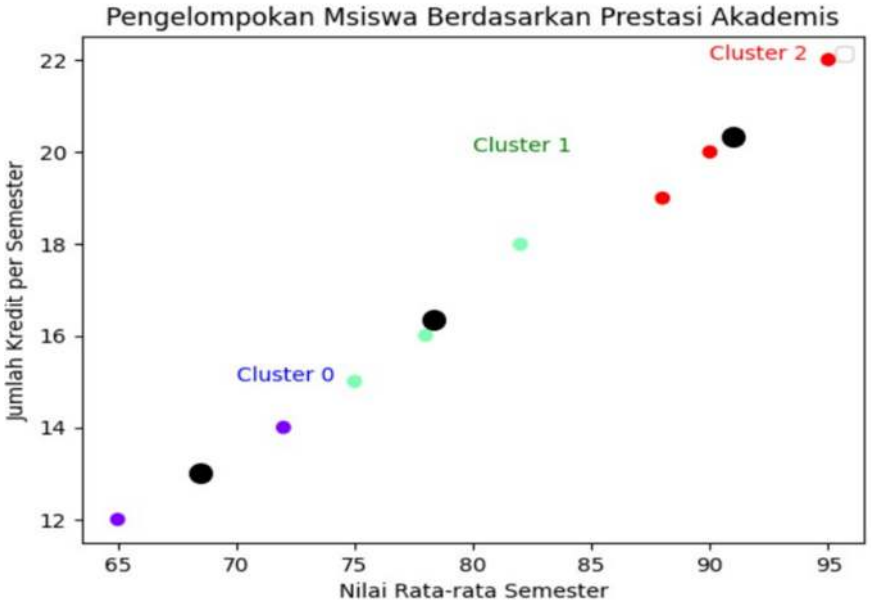

Figure 5. Clustering Results

Three distinct clusters, each represented by a different color, have been formed from the collected data. An explanation of each cluster is presented below:

1. Cluster 0 (Blue): This cluster, located near the bottom of the graph, consists of students who, compared to the other clusters, have relatively lower average grades and fewer credits. Students in this cluster often earn between 12 and 15 credits, with an average grade between 65 and 75. To improve their academic performance, this group may need additional assistance.
2. Cluster 1 (Green): Students in this cluster, located in the middle of the graph, have average grades and credits that are in the middle range. Students in this cluster often earn between 75 and 85, with a total of between 14 and 18 credits. This indicates that this group of students is able to effectively manage their academic performance and credit load.
3. Cluster 2 (Red): Students in this cluster, which is at the top of the graph, have the highest average grade points and total credits. Students in this cluster typically earn between 19 and 22 credits, with an average grade point average of 85 and 95. Students in this cluster demonstrate superior academic achievement and the capacity to manage a higher course load.

By having a better understanding of the differences between these groups, educational institutions can identify groups of students who can benefit from more involvement in their programs and those who need more support. Improving the overall level of education can be done by using effective learning techniques.

5. CONCLUSION

Based on the data processing justification using the aforementioned software tools, the students were grouped into three distinct clusters. Group 1 consists of students with average scores ranging from 65 to 75 and total credits between 12 and 15. These students may require additional support in managing their time and study routines. Group 2 includes students whose average scores fall between 75 and 85, with total credits ranging from 14 to 18. This group demonstrates the ability to maintain a balance between a demanding academic schedule and excellent academic performance. Group 3 comprises students with the highest academic performance, having average scores between 85 and 95 and total credits ranging from 19 to 22. These students exhibit outstanding academic achievements and are capable of handling heavier course loads. In addition to the results obtained through manual processing, which were comparable to those generated automatically, the K-means clustering method provided grouped insights into student achievement scores, including both high and low performers.

REFRENCES

Anas, Aswar, and A. Fitriani. 2019. "Dampak Media E-Learning Terhadap Kedisiplinan Dalam Mengerjakan Tugas Dan Motivasi Belajar Mahasiswa Universitas Cokroaminoto Palopo." *Pedagogy* 4(1):74–101.

Aranda, Josi, Wirda Astari, and Galvani Natasya. 2016. "Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk." *Seminar Nasional Teknologi Informasi Dan Multimedia 2016* 4.2-1-4.2-6.

Asril, Jaenam, Syahrizal, Armalena, and Yuherman. 2023. "Peningkatan Nilai-Nilai Demokrasi Dan Nasionalisme Pada Mahasiswa Melalui Pembelajaran Pendidikan Pancasila Dan Kewarganegaraan." *JIM: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah* 8(3):1300–1309. https://doi.org/10.24815/jimps.v8i3.25109

Asroni, Asroni, and Ronald Adrian. 2016. "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang." *Semesta Teknika* 18(1):76–82. doi: 10.18196/st.v18i1.708.

Asroni, Asroni, Hidayatul Fitri, and Eko Prasetyo. 2018. "Penerapan Metode Clustering Dengan Algoritma K-Means Pada Pengelompokkan Data Calon Mahasiswa Baru Di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran Dan Ilmu Kesehatan, Dan Fakultas Ilmu Sosial Dan Ilmu Politik)." *Semesta Teknika* 21(1):60–64. doi: 10.18196/st.211211.

Betie Febriana, Amriyatun, Luky Winanti, Sandra Amelia. 2013. "Hubungan Antara Keaktifan Organisasi Dengan Prestasi Belajar (Indeks Prestasi) Mahasiswa Fakultas Ilmu Keperawatan Universitas Indonesia." *Jurnal Basicedu* 5(1):94–100.

Darlinda, Darlinda, and Joy Nashar Utamajaya. 2022. "Sistem Pendukung Keputusan Penerima Beasiswa Program Indonesia Pintar Menggunakan Metode Algoritma K-Means Clustering." *JURIKOM (Jurnal Riset Komputer)* 9(2):167. doi: 10.30865/jurikom.v9i2.3971.

Goa Wea, Antonius, and Ignatius Adiwidjaja. 2018. "Pengaruh Beasiswa Terhadap Motivasi Dan Prestasi Belajar Mahasiswa Universitas Tribhuwana Tunggadewi Malang." *Jisip* 7(1):21. https://doi.org/10.33366/jisip.v7i1.1439

Irwan, Muhammad, and Padli Nasution. 2024. "Journal of Sharia Economics Scholar (JoSES) Meningkatkan Kualitas Informasi Melalui Strategi Pengolahan Data Yang Efektif Journal of Sharia Economics Scholar (JoSES)." 2(2):91–96. https://doi.org/10.5281/zenodo.12608571

Lutfiwati, Sri, and Psikologi Pendidikan Islam. 2020. "Motivasi Belajar Dan Prestasi Akademik." 10 https://doi.org/10.24042/alidarah.v10i1.5642

Masni, Harbeng. 2015. "Strategi Meningkatkan Motivasi Belajar Mahasiswa." *Dikdaya* 5(1):34–45. http://dx.doi.org/10.33087/dikdaya.v5i1.64

Melpa Metisen, Benri, and Herlina Latipa Sari. 2015. "Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila." *Jurnal Media Infotama* 11(2):110–18.

Priyatman, Hendro, Fahmi Sajid, and Dannis Haldivany. 2019. "JEPIN (Jurnal Edukasi Dan Penelitian Informatika) Klasterisasi Menggunakan Algoritma K-Means Clustering Untuk Memprediksi Waktu Kelulusan Mahasiswa." *Jurnal Edukasi Dan Penelitian Informatika* 5(1):62–66. https://doi.org/10.26418/jp.v5i1

Rahmalinda, Nanda Ayu, and Arief Jananto. 2022. "Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru." *Jurnal Tekno Kompak* 16(2):163. doi: 10.33365/jtk.v16i2.1971.

Rahmawati, Devi, Annas Setiawan Prabowo, and Riyadi Purwanto. 2021. "Implementasi Model Waterfall Pada Pengembangan Sistem Informasi Monitoring Prestasi Mahasiswa." *Journal of Innovation Information Technology and Application (JINITA)* 3(1):82–93. doi: 10.35970/jinita.v3i1.678.

Sulistiyawati, Ari, and Eko Supriyanto. 2021. "Implementasi Algoritma K-Means Clustring Dalam Penetuan Siswa Kelas Unggulan." *Jurnal Tekno Kompak* 15(2):25. doi: 10.33365/jtk.v15i2.1162.

Winata, Rahmat, Rizki Nurhana Friantini, and Robia Astuti. 2021. "Kemandirian Belajar Dan Kedisipilinan Belajar Terhadap Prestasi Mahasiswa Pada Perkuliahan Daring." *JURNAL E-DuMath* 7(1):18–26. doi: 10.52657/je.v7i1.1343.

Wirantasa, Umar. 2017. "Pengaruh Kedisiplinan Siswa Terhadap Prestasi." *Jurnal Formatif* 7(1):83–95. http://dx.doi.org/10.30998/formatif.v7i1.1272

Yuli Alam. 2018. "Kompetensi Dosen, Motivasi Belajar Mahasiswa Dan Dampaknya Terhadap Prestasi Mahasiswa Dalam Pembelajaran Pengantar Ekonomi (Studi Pada Mahasiswa Program Studi Manajemen Informatika AMIK Bina Sriwijaya Palembang)." 8.