Clustering Of Informatics Students Based On Understanding The Material Using The K-Means Method

Meiselina Irmayanti[1], Naufal Ibra Prasetyo[2], Dionisia Kasilda Bria[3], Jeki Bani Paratu[4], Pauline Hanfiro[5]

[1] Department of Communication Science, Bengkulu, Bengkulu University, Indonesia
[2,3,4] Department of Informatics Engineering, University of Widyagama Malang, Jl. Borobudur No. 35 Malang, Indonesia
[5] University of the South Pacific , Fiji, Fiji

Article Info

Article history:

Received April 01, 2024
Revised May 10, 2024
Accepted June 28, 2024

Keywords:

K-means Clustering
Student Understanding
Academic Performance
Data Analysis
Personalized Learning

ABSTRACT

The level of student understanding in coursework is a crucial determinant of academic success, reflecting both teaching quality and the effectiveness of applied learning methods. In the context of Informatics, challenges often stem from the complexity of subjects such as algorithms, programming, and data analysis, which require analytical and in-depth comprehension. However, differences in learning abilities, backgrounds, and styles often result in varying levels of understanding among students. This study investigates the application of k-means clustering as an innovative method to analyze academic data and classify students based on their understanding of course materials. By utilizing data such as exam scores, quiz results, and classroom engagement, k-means clustering identifies patterns in students' comprehension levels, offering educators insights to tailor teaching strategies effectively. The findings of this study are expected to aid educators in designing targeted interventions, enhance learning processes, and support an inclusive and effective academic environment.

Corresponding Author:

Dionisia Kasilda Bria
Department of Informatics Engineering
Faculty of Engineering, University of Widyagama Malang
Jl. Borobudur No. 35 Malang, Jawa Timur
Email: dionisiakasildab@gmail.com

1. INTRODUCTION

Understanding of lecture material is one of the important factors that influence students' academic success. The level of students' understanding of the lecture material not only reflects the quality of teaching, but also illustrates the effectiveness of the learning methods applied. In the context of the Informatics Study Program, the challenges faced often involve the complexity of materials such as algorithms, programming, and data analysis, which require in-depth understanding and analytical skills. However, variations in students' learning abilities, backgrounds, and learning styles are often factors that cause differences in the level of understanding between individuals [1].

Based on previous research, many students have difficulty understanding the basic concepts that are the foundation of Informatics courses, which ultimately affect their academic results. Data from an internal survey at one university showed that more than 40% of Informatics students faced difficulties in courses that focused on algorithmic logic and data structures. This highlights the importance of developing a system that can effectively analyze students' level of understanding to support more targeted learning and personalized teaching strategies [2].

One innovative approach to addressing this challenge is to use the k-means clustering method. This technique, which is part of data mining, can group students based on their level of understanding of the course material. By grouping student data into clusters based on certain characteristics such as test scores, quiz results, and engagement in class discussions, this method allows educators to understand the underlying patterns of variation in understanding among students [3].

Previous studies have shown the effectiveness of k-means clustering in educational data analysis. For example, this method has been used to identify student performance patterns and design learning programs tailored to individual needs [4]. In the context of the Informatics Study Program, the application of k-means clustering can provide deeper insight into the factors that influence students' level of understanding of the course material.

This study aims to examine the application of the k-means clustering method in analyzing academic data of Informatics Study Program students. By grouping data from assessment results, surveys of understanding of the material, and student participation in the learning process, this study is expected to provide relevant information for lecturers to design more effective teaching methods. The results of this analysis can also be used to identify groups of students who need special attention so that appropriate interventions can be carried out. Through this data-based approach, this research seeks to contribute to improving the quality of the learning process in academic environments, supporting the development of student potential to the maximum, and creating a more inclusive and effective educational ecosystem [5].

2. METHOD

This study uses the k-means clustering method to analyze the level of student understanding of the Informatics Study Program course. The steps taken are as follows:

1. Data Collection
   Student academic data is collected, including assessment results such as exam scores, quiz scores, and participation in class discussions. This data is used to determine student characteristics in the grouping.
2. Initial Centroid Determination
   The initial centroid is randomly determined for each cluster. Each centroid represents a specific cluster, with initial data including courses such as Computer Organization & Architecture, Intelligent Systems, and others.
3. Distance Calculation
   Using the Euclidean distance formula, the distance between each student data and the initial centroid is calculated. The formula is :

   $$D(S_1, C_1) = \sqrt{(S_{1a} - C_{1a})^2 + (S_{1b} - C_{1b})^2 + (S_{1c} - C_{1c})^2 + \ \dots\ (S_{1n} - C_{1n})^2}$$

   Student data is grouped into clusters with the closest distance.
4. New Centroid Calculation
   After the data is grouped, a new centroid is calculated based on the average value in each cluster.
5. Iteration
   Steps 3 and 4 are repeated until there is no significant change in the data grouping (centroid stability). In this study, iteration was carried out until the third iteration, where the results showed no change in the centroid.
6. Clustering Results
   Students are grouped into four clusters based on their level of understanding of the course. Each cluster reflects the category:
   1. Very Good
   2. Good
   3. Enough
   4. Not Good
7. Result Analysis
   Based on the clustering results, the distribution of students' level of understanding is analyzed to identify courses that require special attention. The percentage of students in each category is calculated to provide insight to educators.

3. RESULTS AND DISCUSSION

In this section, the research results are explained and the discussion is given. After the data has been successfully collected, the data can be presented in the form of a table to provide a clear and organized picture of the information analyzed. Furthermore, the data in the table will be used in the calculation process to obtain optimal grouping results.

3.1. Data Collection

After the data has been successfully collected, the data can be presented in the form of a table to provide a clear and organized picture of the information analyzed.

Table 1. Data on the Results of Students Understanding of the Course

NO	MATA KULIAH	EVALUASI MATA KULIAH (BERDASARKAN KUISIONER)													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Organisasi & Arsitektur Komputer	4,00	1,11	2,89	1,95	1,32	1,68	2,84	2,74	2,84	1,11	3,79	3,42	2,42	3,05
2	Rekayasa Perangkat Lunak	4,00	3,00	2,00	1,00	2,00	3,00	1,08	3,88	2,08	1,04	4,00	2,00	2,00	3,00
3	Manajemen Proyek TI	4,00	3,00	2,70	1,80	2,05	2,60	2,40	2,70	2,10	2,55	3,50	2,40	2,40	1,45
4	Analisa & Perancangan Sistem	4,00	3,70	4,00	2,70	2,00	2,50	2,40	2,60	2,50	2,35	3,00	2,70	2,90	3,00
5	Analisis Algorithma	4,00	3,91	3,40	3,60	3,20	3,00	3,20	3,60	3,00	3,10	3,00	3,20	3,82	3,55
6	Mobile Computing	4,00	2,84	2,53	2,42	2,58	2,63	2,68	2,63	2,63	2,32	2,58	2,74	2,79	2,74
7	Sistem Pakar	4,00	2,13	2,30	2,04	1,87	2,39	2,43	2,13	2,35	2,39	3,65	3,52	2,74	3,26
8	Data Warehouse & Data Mining	4,00	3,00	3,70	3,96	3,20	3,00	3,00	3,23	3,00	3,08	3,40	3,92	3,92	3,08
9	Sistem Cerdas	4,00	3,10	3,90	2,60	2,80	3,70	2,90	3,60	3,40	3,30	3,60	2,40	3,40	3,90
10	Komunikasi Data	4,00	2,62	3,61	2,57	2,43	2,76	2,14	2,29	2,19	2,62	2,76	2,95	2,86	2,90
11	Teori Bahasa Automata	4,00	2,68	2,86	2,91	2,91	2,95	2,86	2,55	2,68	2,68	2,95	2,68	2,73	3,05
12	Sistem Operasi	4,00	1,70	2,96	2,00	2,17	2,39	2,13	2,52	2,35	2,17	3,83	3,39	2,57	3,48
13	Basis Data	4,00	3,00	3,14	3,66	3,48	3,86	3,72	3,86	3,90	3,38	3,86	2,93	3,90	3,97
14	Riset Operasional	4,00	3,05	3,00	3,05	2,80	3,00	3,05	2,55	2,80	2,85	3,05	2,95	2,95	2,70
15	Pengembangan Berbasis Platform	4,00	1,32	2,89	1,95	1,47	1,79	2,84	2,74	2,84	1,26	3,79	3,42	2,42	3,05
16	Jaringan Komputer	4,00	3,93	3,93	2,67	3,41	3,52	4,00	3,96	2,85	3,59	2,30	3,00	2,19	3,00
17	Logika Fuzzy	4,00	3,00	3,80	2,40	2,90	3,20	3,80	4,00	2,30	3,00	4,00	2,00	2,00	2,00

3.2. Initial Centroid Determination

The initial cluster in this analysis is determined randomly to start the data clustering process. The initial cluster selection is done by selecting a number of data points as the initial center (centroid) for each cluster. This process aims to provide an initial basis for the k-means method to calculate the distance between the data and the predetermined centroid.

Table 2. Initial Center Point of Each Cluster

Course	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Organisasi & Arsitektur Komputer	4	1,11	2,89	1,95	1,32	1,68	2,84	2,74	2,84	1,11	3,79	3,42	2,42	3,05
Sistem Cerdas	4	2,68	2,86	2,91	2,91	2,95	2,86	2,55	2,68	2,68	2,95	2,68	2,73	3,05
Teori Bahasa Automata	4	3,1	3,9	2,6	2,8	3,7	2,9	3,6	3,3	3,3	3,6	2,4	3,4	3,9
Logika Fuzzy	4	3	3,8	2,4	2,9	3,2	3,8	4	3	3	4	2	2	2

3.3. Distance Calculation

Furthermore, this initial centroid is used in the first iteration to calculate the Euclidean distance of each data to each centroid, which then becomes the basis for the initial clustering process. Using the Euclidean distance formula, the distance between each student data and the initial centroid is calculated. The formula is :

$$D(S_1, C_1) = \sqrt{(S_{1a} - C_{1a})^2 + (S_{1b} - C_{1b})^2 + (S_{1c} - C_{1c})^2 + \ \dots \ (S_{1n} - C_{1n})^2}$$

With the calculation above, the results of the calculation of the distance of the cluster center for iteration 1 are obtained.

Table 3. Results of Iteration Distance Calculation 1

Data	C1	C2	C3	C4
1	0	3,377	4,504	4,373
2	3,833	3,925	4,543	4,284
3	3,393	2,421	3,767	2,724
4	3,589	1,95	2,633	3,099
5	6,408	4,179	3,065	4,35
6	3,089	0,998	2,93	3,083
7	2,19	2,136	3,579	3,895
8	5,643	3,738	2,919	3,95
9	4,504	2,368	0	2,912
10	3,171	1,203	3,089	3,453
11	3,377	0	2,368	2,782
12	1,926	2,248	3,37	3,727
13	5,25	3,072	1,949	3,728
14	3,633	0,724	2,332	2,693
15	0,318	3,101	4,243	4,13
16	5,361	2,882	2,674	2,634
17	4,373	2,782	2,912	0

After determining the initial centroid, the next step is to compare the distance of each course to each of the predetermined centroids. This distance calculation is done using the Euclidean distance formula, where the resulting distance reflects the relative proximity between the course and the cluster center. The courses are then grouped based on the closest distance to a particular centroid. The grouping results are shown in the following table:

Table 4. Cluster 1 Grouping Results

Cluster	Course	Amount
C1	1 , 2 , 12 , 15	4
C2	3 , 4 , 6 , 7 , 10 , 11 , 14	7
C3	5 , 8 , 9 , 13	4
C4	16 , 17	2

The results of the clustering analysis show that Cluster C2 has the largest number of members, namely 7 courses. This indicates that Cluster C2 is the cluster with the highest data density, which illustrates the strong similarity of characteristics between the courses included in this cluster. Meanwhile, Cluster C1 and Cluster C3 each have 4 members. This number shows that both clusters are moderate in size, with a fairly significant level of similarity between data, although not as dense as Cluster C2. This provides insight that the courses in these two clusters have more varied patterns compared to Cluster C2. On the other hand, Cluster C4 is the smallest cluster, only including 2 courses. This small number of members reflects that Cluster C4 groups courses with more unique or specific characteristics compared to other clusters.

3.4. New Centroid Calculation

Then, determine the new cluster center based on the previous calculation results. This process is done to identify a more representative midpoint of each cluster, which will later be used in the next iteration to improve the clustering results.

Table 5. New Cluster Centers

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C1	4	1,78	2,69	1,73	1,74	2,22	2,22	2,97	2,53	1,4	3,85	3,06	2,35	3,15
C2	4	2,86	2,85	2,5	2,38	2,69	2,69	2,49	2,46	2,54	3,07	2,85	2,77	2,73
C3	4	3,25	3,76	3,56	3,57	3,88	3,88	3,87	3,82	3,44	3,87	3,31	3,76	3,36
C4	4	3,47	3,87	2,54	3,16	3,36	3,36	3,98	2,58	3,3	3,15	2,5	2,1	2,5

3.5. Iteration

Then, recalculate in the same way as before. The grouping results obtained after this iteration are as follows:

Tabel 6. Cluster 2 Grouping Results

Cluster	Course	Amount
C1	1 , 2 , 7 , 12 , 15	5
C2	3 , 4 , 6 , 10 , 11 , 14	6
C3	5 , 8 , 9 , 13	4
C4	16 , 17	2

Since the grouping results in the first and second iterations do not produce the same number, a third iteration needs to be carried out by following the same procedure. The results found in the third iteration are as follows:

Table 7. Cluster 3 Grouping Results

Cluster	Matkul	Amount
C1	1 , 2 , 7 , 12 , 15	5
C2	3 , 4 , 6 , 10 , 11 , 14	6
C3	5 , 8 , 9 , 13	4
C4	16 , 17	2

3.6. Clustering Results

In the second and third iteration tests, there was no change in the centroid value, indicating that the clustering results in that iteration were stable and the same as the centroid in the previous iteration. This indicates that the clustering process has reached convergence, where changes in the position of the cluster center are no longer significant. Thus, the iteration process can be considered complete, and the final result is the formation of 4 clusters that have been identified in 3 iterations. This process shows efficiency and consistency in the formation of optimal clusters.

Table 8. Results of K-Means Clustering Analysis

Cluster	No.	Amount	Course
1	1 , 2 , 7 , 12 , 15	5	Organisasi & Arsitektur Komputer, Rekayasa Perangkat Lunak , Sistem Pakar , Sistem Operasi , Pengembangan Berbasis Platform
2	3 , 4 , 6 , 10 , 11 , 14	6	Manajemen Proyek TI , Analisa & Perancangan Sistem , Mobile Computing , Komunikasi Data , Teori Bahasa Automata , Riset Operasional
3	5 , 8 , 9 , 13	4	Analisis Algorithma , Data Warehouse & Data Mining , Sistem Cerdas , Basis Data
4	16 , 17	2	Jaringan Komputer , Logika Fuzzy

3.7. Result Analysis

From the table above, it can be seen that the understanding of the course material "Very Good" is 29% consisting of 5 courses, namely Computer Organization & Architecture, Software Engineering, Expert Systems, Operating Systems and Platform-Based Development. The understanding of the course material "Good" is 35% consisting of 6 courses, namely IT Project Management, System Analysis & Design, Mobile Computing, Data Communication, Automata Language Theory, Operational Research, the understanding of the course material "Quite Good" is 24% consisting of 4 courses, namely Algorithm Analysis, Data Warehouse & Data Mining, Intelligent Systems and Databases. While the understanding of the course material "Less Good" is 12% consisting of 2 courses, namely Computer Networks and Fuzzy Logic.

4. CONCLUSION

Based on the results of the research conducted, it can be concluded that the application of the K-Means method has proven effective in "Clustering Of Informatics Program Students Based On Understanding The Material Using The K-Means Method". By using this method, different patterns of understanding among students can be identified, which can then be used as a basis for more targeted planning in the learning process. In addition, the results of this clustering can provide deeper insight for policy makers in optimizing the assessment system and designing educational strategies that are more adaptive to student needs, in order to improve the quality of education in the academic environment.

ACKNOWLEDGEMENTS

Throughout the course of this study, from the initial planning to the research process and finalization of this journal, our research team encountered numerous challenges. We are profoundly grateful to God Almighty for His unwavering guidance and blessings, which have been instrumental in completing this research.

We would also like to express our sincere gratitude to Mrs. Fitri Marisa, S.Kom., M.Pd., Ph.D., for her exceptional mentorship and unwavering support throughout this study. Her invaluable guidance, expertise, and dedication have greatly enriched our research and the preparation of this journal. Her contributions have inspired and motivated us at every step.

Furthermore, we extend our heartfelt thanks to our fellow researchers and all individuals who have played a crucial role in the data collection, analysis, and the preparation of this journal. Their collaborative efforts, hard work, and contributions have been vital to the successful completion of this project. Without their assistance, this research would not have been possible.

REFERENCES

[1] Mellasanti Ayuwardani, "Pemahaman Materi Terhadap Hasil Belajar Mahasiswa Pada Matakuliah Praktek," J. Ekon. Bisnis Dan Manaj., vol. 1, no. 2, pp. 213–221, 2023, doi: 10.59024/jise.v1i2.130.

[2] A. Isroqmi, "Kemampuan Mahasiswa Memahami Logika Pemrograman," J. Pendidik. Mat., no. May, 2020.

[3] B. J. M. Putra and D. A. F. Yuniarti, "Analisis Hasil Belajar Mahasiswa Dengan Clustering Menggunakan Metode K-Means," Poros Tek., vol. 12, no. 2, pp. 49–58, 2020, [Online]. Available: https://ejurnal.poliban.ac.id/index.php/porosteknik/article/view/927

[4] S. Fitri, N. Nurjanah, and W. Astuti, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa (Studi Kasus: Umtas)," Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 9, no. 1, pp. 633–640, 2018, doi: 10.24176/simet.v9i1.2002.

[5] B. Poerwanto and R. Y. Fa'rifah, "Analisis Cluster K-Means Dalam Pengelompokan Kemampuan Mahasiswa," Indones. J. Fundam. Sci., vol. 2, no. 2, pp. 92–96, 2016, [Online]. Available: https://ojs.unm.ac.id/pinisi/article/view/2434