Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2024. Vol. 4. No. 1; e-ISSN: 2808-4519

Classification of Students Based on Factors That Affect Student Learning Achievement Using The K-Means Clustering Algorithm (Case Study: STMIK Kaputama Binjai)

Dea Puspita[1], Relita Buaton[2], Husnul Khair[3]

[1, 2, 3]*STMIK Kaputama*

deapuspitavivo0406@gmail.com[1*], bbcbuaton@gmail.com[2], husnul.khair@gmail.com[3]

Abstract

In the world of education, students are the main object of every educational implementation that always prioritizes disciplines that are beneficial to the students themselves. However, in lecture activities there are students who are diligent in participating in lecture activities and there are also those who rarely participate in lecture activities, this can be caused by internal and external factors, so that there can be significant variations in student learning achievements, with some achieving high grades, while others face difficulties in achieving the same achievements. Based on the description of the problem, the researcher conducted a study that aimed to group students based on factors that affect student learning achievement using the k-means clustering algorithm. The results of the research conducted produced 3 clusters with cluster 1 there were 5 data, the group of students with a very satisfactory predicate GPA (3.50-4.00), supported by both internal and external factors (interval 3.1-4). Cluster 2 has 3 data, the group of students with a satisfactory predicate GPA (3.00-3.49), supported by both internal and external factors (interval 2.1-3), and cluster 3 has 5 data, the group of students with a satisfactory predicate GPA (3.00-3.49), supported by both internal and external factors (interval 3.1-4).

Keywords: *Learning Achievement, Data Mining, K-Means*

1. Introduction

Education is a learning process towards self-development in order to be able to compete in accordance with the demands of today's times. Advances in science and technology are necessary to improve the quality of education. Improving the quality of education can be achieved through improvement, change, and renewal of factors that affect the success of education. The success of education can be seen from the achievement of learning achievements obtained by students after carrying out their learning activities which are expressed in the form of numerical values or letters. The achievement of learning achievement is influenced by two factors, namely internal factors are factors that exist in individuals such as health, interests, talents, motivation, attitude and intelligence level. And external factors are factors that exist outside the individual such as family support, association and learning environment [1].

In the world of education, students are the main object of every educational implementation that always prioritizes disciplines that are beneficial for the students themselves or others. However, in lecture activities there are students who are diligent in participating in lecture activities and there are also those who rarely participate in lecture activities which can be caused by internal and external factors, so that there can be significant variations in student learning achievements [2].

Based on the description of the problem, this study aims to group student data based on factors that affect student learning achievement by using the k-means clustering algorithm which is to find out and form student clusters that have the same characteristics.

2. Literature Review

2.1. Data Mining

Data mining is an integral part of Knowledge Discovery in Database (KDD) which is the overall process of converting raw data into useful information. The KDD process consists of a series of transformation stages, from the data preprocessing process and the data postprocessing process where the data preprocessing process aims to convert the raw data into a suitable format for further analysis. The steps taken include fixing dirty or duplicate data, and selecting records and features that are relevant to the next data management process [3], [4].

2.2. Clustering

Clustering is a data analysis technique used to group objects or data into similar groups based on their characteristics or attributes. The main purpose of clustering is to find hidden structures or patterns in the data without labels and previous information about existing classes or groups [5], [6].

2.3. K-Means

The K-Means algorithm is a data grouping technique that involves a number of k clusters. This approach divides data into clusters based on the degree of similarity between data in a cluster and dissimilarities grouped into different clusters. The center of the cluster is the average of the cluster member values, called the centroid or center of gravity [7].

The following are the stages in clustering or grouping with the K-Means algorithm:
1) Determine the number of clusters (k)
2) Specifies the centroid (the midpoint coordinates of each cluster), for the first iteration it is randomly retrieved
3) Calculate the distance of an object to the centroid by using the dEuclidean or Manhattan distance formula
4) Determining the distance of each object to the midpoint coordinates
5) Group these objects based on their closest distance .

2.4. Learning Achievement

Achievement reflects the achievements obtained after carrying out learning activities. To understand the extent of success in the learning process, measurement or evaluation of learning is carried out. The results of this evaluation indicate the achievements that have been achieved in participating in a special learning process. Learning achievement is defined as the result of an assessment of knowledge, skills, and attitudes expressed in the form of grades [8].

3. Research Methods

3.1. Research Methods

Research methodology is a way to investigate and trace a problem by using careful and meticulous scientific methods to collect, process, and analyze data, as well as draw conclusions systematically and objectively to solve a problem or to reach a hypothesis, with the aim of obtaining knowledge that is useful for human life. The stages of research methodology in completing this research are [9]:

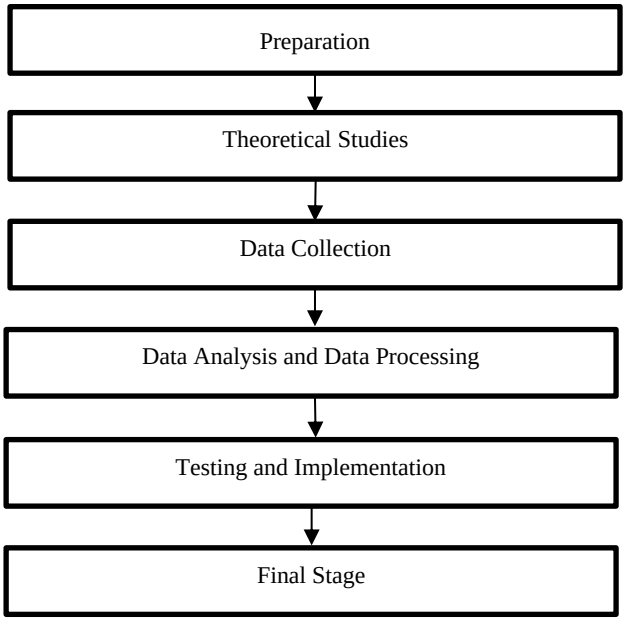

Fig. 1: Research Workflow

1. Preparation
   This stage is the initial stage of the research process which will be carried out by determining the background of the problem, by looking for a problem and then formulating what problems are and how the process of solving them.
2. Theoretical Studies
   The collection of theories related to the subject matter such as data *mining, clustering, K-Means* algorithms, and factors that affect student learning achievement as well as case studies collected from various sources such as journals, books, and other references.
3. Data Collection
   At this stage, the data needed to make the thesis was collected, namely student data obtained from a questionnaire filled out by STMIK Kaputama Binjai students.
4. Data Analysis and Data Processing

At this stage, the supporting data that has been obtained in the previous stage will be analyzed, by conducting an analysis using *the K-Means Clustering* algorithm to obtain the results of student grouping based on factors that affect student learning achievement.

5. Testing and Implementation
   At this stage, the results obtained from the data that have been processed or implemented into the Matlab R2014a application will be tested.
6. Final Stage
   At this stage, it is the stage of drawing conclusions and suggestions from the research that has been carried out. With the conclusion, the results of the entire thesis will be known and suggestions for improvement and benefits for others are expected.

4. Results And Discussion

4.1. Results

In analyzing data in a study, supporting data is needed so that a research can run as expected. The data collection process was taken from a questionnaire that was shared with the number of questions related to the GPA range as well as Internal Factors and External Factors, each of which had 25 questions and the data to be grouped was the number of questionnaire answer results. The supporting data for the research in this study are:

Table 1: Internal Factor Data (Ability, Interest, Motivation, and Health)

No	Name	Internal Factor																									Average
		Ability						Interest								Motivation							Health				
		A	B	C	D	E	F	A	B	C	D	E	F	G	H	A	B	C	D	E	F	A	B	C	D	E	
1	Ronauli Silaban	4	4	4	4	4	4	3	4	1	4	4	4	4	2	4	5	4	4	2	4	1	4	1	3	2	4
2	Abdullah Husein	4	3	3	4	3	3	4	4	1	3	4	4	3	4	3	5	4	4	2	3	2	4	2	3	3	4
3	Elsa Risqi Amalia	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	2	2	4	2	3
4	Widya Natasya	4	4	4	3	4	4	4	4	2	5	4	5	4	4	4	5	5	5	2	3	2	4	2	3	2	4
5	Sabina Eis Z.	5	3	3	3	3	5	5	5	3	5	3	5	5	2	5	5	5	5	4	5	4	3	1	2	2	4
6	Tia Permata Sari	4	4	4	4	1	4	4	4	1	5	4	5	5	2	4	5	5	4	1	4	1	2	1	3	2	4
7	Febby Ria	4	3	4	4	3	3	4	4	1	4	4	5	5	4	4	5	4	4	1	5	1	5	1	2	2	4
8	Ajisro Siringoringo	3	3	4	2	3	2	3	2	2	3	2	2	2	3	2	2	2	2	1	1	1	1	5	1	1	2
9	Melani Puspita Sari	3	2	3	4	4	3	3	3	4	3	3	2	3	3	4	2	2	3	3	4	4	2	4	3	3	4
10	Nadya Amalia	5	2	2	2	5	2	2	2	3	2	2	5	2	4	2	2	2	2	4	2	4	2	4	3	3	3

Table 2: External Factor Data (Teaching Quality, Learning Facilities, Student Activities, and Associations)

No	Name	External Factor																									Average
		Teaching Quality						Learning Facilities					Student Activities								associations						
		A	B	C	D	E	F	A	B	C	D	E	A	B	C	D	E	F	G	H	A	B	C	D	E	F	
1	Ronauli Silaban	3	3	3	3	3	3	3	3	3	2	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3
2	Abdullah Husein	2	3	3	2	4	2	3	2	2	2	2	1	4	2	2	3	1	3	1	2	2	2	2	2	2	3
3	Elsa Risqi Amalia	3	3	3	3	3	3	3	4	4	3	3	2	2	3	3	3	3	3	3	2	2	2	2	2	2	3
4	Widya Natasya	4	4	5	4	5	5	4	4	4	5	4	1	2	2	2	2	2	2	2	2	2	2	2	2	2	3
5	Sabina Eis Z.	4	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	5	5	5	3	5	5	5	5	5	4
6	Tia Permata Sari	4	4	4	4	4	4	3	4	4	4	4	2	5	5	5	5	5	5	5	3	2	3	2	2	3	4
7	Febby Ria	4	4	4	4	4	4	4	4	4	3	3	2	2	5	5	5	5	5	5	3	2	5	5	5	5	4
8	Ajisro Siringoringo	1	1	1	1	1	1	3	5	5	5	4	2	2	1	1	1	1	2	2	2	3	1	2	2	3	3
9	Melani Puspita Sari	2	2	3	2	2	3	2	3	3	4	3	1	4	3	3	4	3	3	3	3	3	3	3	3	3	3
10	Nadya Amalia	2	2	2	2	2	2	3	3	3	2	5	1	3	2	2	3	3	3	5	3	2	2	2	2	3	3

Table 3: Temporary GPA Range Data Transformation

GPA Range	Transformation Value
1,00 - 1,99	1
2,00 - 2,99	2
3,00 - 3,49	3
3,50 - 4,00	4

Table 4: Transformation of Average Answers for Internal Factors and External Factors

Average	Transformation Value	Description
0-1	1	Very Bad
1,1-2	2	Poor
2,1-3	3	Good Enough
3,1-4	4	Good
4,1-5	5	Excellent

The existing data is then transformed to make it easier to make calculations, here are the data that have been changed with the transformation value, namely:

Table 5: Altered Data With Transform Values

No	Name	GPA Range (X)	Internal Factor (Y)	External Factor (Z)
1	Ronauli Silaban	4	4	3
2	Abdullah Husein	3	4	3
3	Elsa Risqi Amalia	3	3	3
4	Widya Natasya	3	4	3
5	Sabina Eis Z.	3	4	4
6	Tia Permata Sari	3	4	4
7	Febby Ria	4	4	4
8	Ajisro Siringoringo	3	2	3
9	Melani Puspita Sari	3	4	3
10	Nadya Amalia	2	3	3

After the data is transformed, the next thing is to determine the number of clusters, determine the centroid randomly and then perform the calculation with the K-Means Clustering algorithm with the following Euclidean Distance formula:

$$D(ij) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (X3i - X3j)^2}.$$

The cluster center (Centroid) used is 3, which is as follows:
Centroid 1 is taken from the 1st data i.e. C1 = (4, 4, 3)
Centroid 2 is taken from the 8th data i.e. C2 = (3, 2, 3)
Centroid 3 is taken from the 9th data i.e. C3 = (3, 4, 3)

And the results of the first iteration that has been carried out can be seen in the following table:

Table 6: Group 1 Determination Results

No	Name	X	Y	Z	Distance from C1	Distance from C2	Distance from C3	Group
1	Ronauli Silaban	4	4	3	0,00	2,24	1,00	1
2	Abdullah Husein	3	4	3	1,00	2,00	0,00	3
3	Elsa Risqi Amalia	3	3	3	1,41	1,00	1,00	2
4	Widya Natasya	3	4	3	1,00	2,00	0,00	3
5	Sabina Eis Z.	3	4	4	1,41	2,24	1,00	3
6	Tia Permata Sari	3	4	4	1,41	2,24	1,00	3
7	Febby Ria	4	4	4	1,00	2,45	1,41	1
8	Ajisro Siringoringo	3	2	3	2,24	0,00	2,00	2
9	Melani Puspita Sari	3	4	3	1,00	2,00	0,00	3
10	Nadya Amalia	2	3	3	2,24	1,41	1,41	2

After the calculation is carried out using the existing formula, the *group* based on the nearest *centroid* distance is:
Old group = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
New group = {1, 3, 2, 3, 3, 3, 1, 2, 3, 2}

If there is a change in the group, it will be continued to the next iteration, namely iteration II. Before counting iteration II, it is necessary to create a new centroid center from all three clusters. Here are the 3 new centroids:
C1 (4,00; 4,00; 3,50)
C2 (2,67; ,2,67; 3,00)
C3 (3,00; 4,00; 3,40)

The results of the calculation of *the Euclidean Distance* value in iteration II can be seen in the following table:

Table 7: Group 1 Determination Results

No	Name	X	Y	Z	Distance from C1	Distance from C2	Distance from C3	Group
1	Ronauli Silaban	4	4	3	0,50	1,89	1,08	1
2	Abdullah Husein	3	4	3	1,12	1,37	0,40	3
3	Elsa Risqi Amalia	3	3	3	1,50	0,47	1,08	2
4	Widya Natasya	3	4	3	1,12	1,37	0,40	3
5	Sabina Eis Z.	3	4	4	1,12	1,70	0,60	3
6	Tia Permata Sari	3	4	4	1,12	1,70	0,60	3
7	Febby Ria	4	4	4	0,50	2,13	1,17	1
8	Ajisro Siringoringo	3	2	3	2,29	0,75	2,04	2
9	Melani Puspita Sari	3	4	3	1,12	1,37	0,40	3
10	Nadya Amalia	2	3	3	2,29	0,75	1,47	2

By using iteration II of the group 1 results, the group results based on the minimum distance to the nearest centroid are obtained:
Old group = {1, 3, 2, 3, 3, 3, 1, 2, 3, 2}
New group = {1, 3, 2, 3, 3, 3, 1, 2, 3, 2}

4.2. Discussion

The following are the results *of clustering* using the Matlab programming application, the following are the results of *clustering* data grouping:

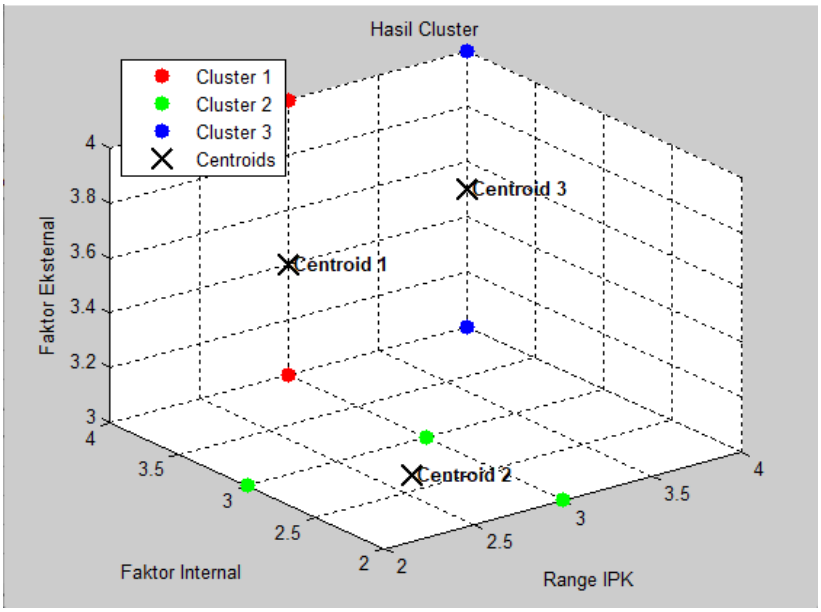

Fig. 2: Cluster Results from Data on Factors Affecting Student Learning Achievment

Description:

1) Cluster 1 (4.00; 4.00; 3.50)
   It can be seen that in cluster 1, the group of students with a predicate GPA is very satisfactory (3.50-4.00), supported by good internal and external factors (interval 3.1-4).
2) Cluster 2 (2.67; 2.67; 3.00)
   It can be seen that in cluster 2, the group of students with a satisfactory predicate GPA (3.00-3.49), supported by internal factors and external factors that are quite good (interval 2.1-3).
3) Cluster 3 (3.00; 4.00; 3.40)
   It can be seen that in cluster 3, the group of students with a satisfactory predicate GPA (3.00-3.49), supported by good internal and external factors (interval 3.1-4).

5. Conclusions And Suggestions

5.1. Conclusions

Based on the research that has been conducted, it can be concluded that the *K-Means Clustering* algorithm is able to group students based on factors that affect student learning achievement. The results of the research conducted produced 3 clusters with cluster 1 with 2 data, cluster 2 with 3 data, and cluster 3 with 5 data. The results of the study also show that a combination of internal factors such as ability, interest, motivation, and health, as well as external factors such as the quality of teaching, learning facilities, student activities, and excellent associations can encourage the achievement of student learning achievement. And the most dominant factor affecting student learning achievement is the internal factor with an average sum of 3,196.

5.2. Suggestions

Based on the research that has been carried out, the author outlines several suggestions that are expected to be input for further research, the suggestions are as follows:

1. In the next research with the same topic, it is recommended to use other *clustering* methods or by expanding the variables studied to get more comprehensive results regarding the factors that affect student learning achievement.
2. It is expected to add input data so that the grouping results carried out by the system can be maximized, because a lot of data can affect the grouping results.
3. It is hoped that this research can help the campus in developing academic support programs such as workshops.

References

[1] Buaton, R., Zarlis, M., Efendi, S., & Yasin, V. (2019). Time Series Data Mining (Publish WADE, Ed.). WADE Group,
[2] FAKTOR-FAKTOR YANG MEMPENGARUHI PRESTASI BELAJAR SISWA KELAS X DALAM PROGRAM. (n.d.),
[3] Mona, S., & Yunita, P. (n.d.). FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN PRESTASI BELAJAR MAHASISWA FACTORS RELATED TO ACHIEVEMENT STUDENT LEARNING,
[4] Muharmi, Y. (2016). Pengelompokan Siswa Berdasarkan Faktor-faktor Yang Mempengaruhi Keberhasilan Siswa dalam Belajar Menggunakan Metode Clustering K-Means. Jurnal Teknologi Informasi dan Pendidikan, 9(1), 94-101.
[5] F. Ningsih, Y. Maulita, and M. Sihombing, "Classification Of Population Data On Status In The Family Based On Last Education And Work Using The Clustering Method (Case Study: Sei Prison Village Office)", *j. of artif. intell. and eng. appl.*, vol. 3, no. 1, pp. 93–101, Oct. 2023.
[6] S. Dwi Pratiwi, A. Fauzi, and I. G. Prahmana, "Grouping Number of Library Members For Determining the Location of Socialization Using Clustering Method", *j. of artif. intell. and eng. appl.*, vol. 3, no. 1, pp. 120–127, Oct. 2023.
[7] A. R. Arianti, N. Novriyenni, and I. Ambarita, "The Grouping Of Types Of Tax Revenue In Binjai City Uses The K-Means Clustering Method Algorithm", *j. of artif. intell. and eng. appl.*, vol. 3, no. 1, pp. 197–202, Oct. 2023.
[8] Borkat parulian pohan, "Data Mining Grouping Areas Of Diarrhea Disease In The City Medan Using K-Means (Clustering) Algorithm In Service Environment In Medan City", *j. of artif. intell. and eng. appl.*, vol. 3, no. 1, pp. 448–453, Oct. 2023.