Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

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

Clustering Students Level of Understanding of Programming Language Courses Using the K-Means Algorithm

Noval Ramadana[1], Yani Maulita[2], Hermansyah Sembiring[3]

[1, 2 ,3]STMIK KAPUTAMA

novalramadana11@gmail.com[1*], yani.maulita@gmail.com[2], hermansyahsembiring2@gmail.com[3]

Abstract

This study aims to categorize students based on their level of understanding of programming language courses using the K-Means algorithm. Students often experience difficulties in understanding the basic concepts of programming languages, which can affect their ability to solve programming problems. Using data obtained from questionnaires filled out by STMIK Kaputama Binjai students, this study analyzed variables such as attendance rate, learning interest, and level of understanding. The analysis results show the existence of patterns and relationships between these variables, which can be used to identify groups of students who have a good, poor, or no understanding of the course. This research is expected to provide input for educational institutions in designing learning strategies that are more effective and attractive to students.

Keywords: *K-Means, Clustering, Student Understanding, Programming Language, Data Mining*

1. Introduction

A programming language is a notation used to write programs on a computer. Programming language is one of the important courses in computer-based universities. Programming language courses are taught both in theory and practice. As is known, programming languages have different levels of difficulty so that many students still have difficulty understanding them. [1]

In programming languages, there are many concepts that must be understood by students, making it difficult for students to solve programming problems effectively. Just like the students of STMIK Kaputama Binjai, who still find it difficult to understand programming language courses such as difficulties in understanding basic programming language concepts such as variables, data types, logical operations and program flow control, in the process of writing program codes using certain programming languages, and difficulties in designing algorithms and implementing them in program code.

Based on the description of these problems, the author will conduct research that aims to cluster students based on their level of understanding of programming language courses using the K-Means algorithm. Clustering or grouping is done to form clusters of students' level of understanding of various programming language courses and find out how many students understand, understand less and do not understand programming language courses.

2. Literature Review

2.1. Data Mining

Data mining is the process of discovering meaningful relationships, patterns, and trends by analyzing large amounts of stored data, using pattern recognition technologies such as statistics and mathematics. Data mining can also be defined as the process of finding patterns in data. mining is the process of analyzing large amounts of data to find unexpected relationships and summarize data so that it can be understood and utilized [2].

2.2. Clustering

Cluster is a group of objects that have certain similarities. Clustering is one of the methods in unsupervised machine learning that is used to group objects into groups or clusters based on their similarity with each other. Since this method belongs to the unsupervised category, the dataset used for the clustering model does not have a label. [3]

2.3. K-Means Algorithm

K-Means algorithm is an Unsupervised Learning method with an iterative process, where the dataset is grouped into k predetermined number of non-overlapping clusters or subgroups. The algorithm attempts to keep the points in the cluster close to each other, while the clusters themselves are in different spaces. The goal of this algorithm is to allocate data points to clusters so that the sum of the squared distances between the data points and the cluster centroids is minimal. At this point, the centroid of the cluster is the average value of the data points in the cluster.

The stages of the K-Means algorithm in general are as follows:
1. Determine the number of clusters
2. Selecting clusters randomly and grouping other data into these clusters based on their closest distance
3. Calculate the distance to the centroid. Each data will be determined the closest centroid with the Euclidean formula:

$$D(x,y) = \sqrt{(X_{1x} - X_{1y})^2 + (X_{1x} - X_{1y})^2 + \cdots + (X_{kx} - X_{ky})^2}$$

Description:
D(x,y) = Distance of xth data to cluster center y
x = Data to x
y = Centroid data to x
Xkx = Data to i at the kth data attribute
Xky = Center point to j at the kth attribute

4. Reallocate each data into the nearest centroid / cluster average
5. Repeat Step 3, if there is still data that moves clusters so that it causes a change in the cluster centroid value. [4]

2.4. Programming Languages

A computer program or often referred to as a program is a set of instructions written to perform a specific function on a computer. A program usually has a certain form of execution model so that it can be directly executed by the computer. [5]

A programming language, also known as a computer language or computer programming language, is a standardized set of instructions used to control a computer. It consists of syntax and semantic rules used to define computer programs.

3. Research Methods

3.1. Research Methods

Research methodology is a field of study that examines the various methods and techniques used in research to collect, analyze, and interpret data. Its main purpose is to provide clear and structured guidance in carrying out research, so that the results obtained can be trusted, valid, and scientifically accountable. The stages of research methodology in the preparation of this thesis are as follows:

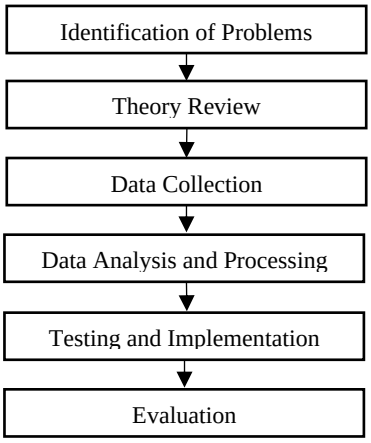

Fig. 1: Research Workflow

1. Problem Identification
   This stage aims to understand and formulate the problem to be studied. Furthermore, it determines the research objectives, problem boundaries and research benefits.
2. Theory Review
   Reviewing literature and theories related to the subject matter such as data mining, clustering, K-Means algorithm, and programming languages.
3. Data Collection
   At this stage, data collection is carried out in accordance with the predetermined clustering variables, namely attendance level, interest in learning and level of understanding. The data is obtained from filling out a questionnaire filled out by STMIK Kaputama Binjai students.
4. Data Analysis and Data Processing
   The data that has been collected is analyzed and processed to find patterns and relationships between variables using the k-means clustering algorithm.
5. Testing and Implementation.
   At this stage, data testing will be carried out using the k-means algorithm and then implemented into the Matlab R2014a application.
6. Evaluation
   At this stage is the stage of taking conclusions and suggestions from the research that has been done.

4. Results and Discussion

4.1. Result

In analyzing data in a study, supporting data is needed so that the research can run as expected. From the research conducted at STMIK Kaputama Binjai, data is obtained that will be used to analyze the grouping of students' level of understanding in subjects. These data are as follows:

Table. 1: Supporting Research Data to be Processed

No	alternative	Attendance Rate	Learning Interest							Comprehension Level												
			A	B	C	D	E	F	G	A	B	C	D	E	F	G	H	I	J	K	L	M
1	A1	5	5	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	A2	5	5	5	5	5	5	5	1	5	1	1	1	5	1	5	5	1	1	5	1	5
3	A3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
4	A4	5	5	5	5	5	5	5	5	5	1	1	1	5	5	5	5	5	5	5	5	5
5	A5	5	5	5	1	5	1	1	5	5	1	1	1	5	5	5	1	5	1	5	1	5
6	A6	5	5	1	5	5	5	5	5	5	5	1	1	5	1	5	5	5	5	5	1	5
7	A7	1	5	1	5	1	1	1	1	5	1	1	1	1	1	5	5	1	1	1	1	1
8	A8	5	5	5	5	1	1	1	1	5	5	1	1	5	1	5	5	1	5	5	5	5
9	A9	1	5	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	A10	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
11	A11	5	5	5	5	5	1	5	1	5	5	1	1	1	1	5	5	5	1	5	1	1
12	A12	5	5	1	5	1	1	1	1	5	1	1	5	5	1	5	5	1	1	5	1	5
13	A13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	A14	5	5	5	5	5	1	1	1	5	1	1	1	1	1	1	5	1	1	1	1	1
15	A15	1	1	1	5	1	1	5	1	1	5	5	5	1	1	1	1	5	5	1	1	5
16	A16	1	5	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	A17	5	5	5	5	5	5	5	5	5	5	1	1	5	1	5	5	5	5	5	5	5
18	A18	5	5	5	5	5	5	5	1	5	5	1	1	1	1	5	5	5	5	5	5	5
19	A19	1	5	5	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	A20	1	5	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table. 2: Data that has been transformed

No	Alternative	Attendance rate (X)	Learning Interest (Y)	Comprehension Level (Z)
1	A1	5	3	3
2	A2	5	3	2
3	A3	5	3	3
4	A4	5	3	3
5	A5	5	2	2
6	A6	5	3	3
7	A7	1	2	1
8	A8	5	2	3
9	A9	1	2	1
10	A10	5	3	3
11	A11	5	3	2
12	A12	5	2	2
13	A13	1	1	1
14	A14	5	2	1
15	A15	1	2	2

16	A16	1	2	1
17	A17	5	3	3
18	A18	5	3	3
19	A19	1	2	1
20	A20	1	2	1

After the data is transformed, the next step is to determine the number of clusters, which in this study is determined to be 3 clusters. Then perform calculations to measure the distance between two points or X, Y, and Z. Distance measurement using the Euclidean Distence equation as follows: $\sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (X3i - X3j)^2}$

The calculation process is as follows:
In iteration 1, the determination of the initial centroid is taken randomly or randomly
Centroid 1 = (5, 3, 3) is taken randomly from the 1st data
Centroid 2 = (5, 3, 2) is taken randomly from the 2nd data
Centroid 3 = (1, 2, 1) randomly drawn from the 7th data

Section A1 (5, 3, 3)
K=3, C1 = (5, 3, 3), C2 = (5, 3, 2), C3 = (1, 2, 1)
Distance from C1 (X) = $\sqrt{(5-5)^2 + (3-3)^2 + (3-3)^2}$ = 0,00

Distance from C2 (Y) = $\sqrt{(5-5)^2 + (3-3)^2 + (3-2)^2}$ = 1,00

Distance from C3 (Z) = $\sqrt{(5-1)^2 + (3-3)^2 + (3-1)^2}$ = 4,58

The calculation is continued until A20 and the results obtained in iteration 1 are as follows:

Table. 3: Results of Group Determination in Iteration

No	Alternative	X	Y	Z	Distance from C1	Distance fro C2	Distance fro C3	Group
1	A1	5	3	3	0,00	1,00	4,58	1
2	A2	5	3	2	1,00	0,00	4,24	2
3	A3	5	3	3	0,00	1,00	4,58	1
4	A4	5	3	3	0,00	1,00	4,58	1
5	A5	5	2	2	1,41	1,00	4,12	2
6	A6	5	3	3	0,00	1,00	4,58	1
7	A7	1	2	1	4,58	4,24	0,00	3
8	A8	5	2	3	1,00	1,41	4,47	1
9	A9	1	2	1	4,58	4,24	0,00	3
10	A10	5	3	3	0,00	1,00	4,58	1
11	A11	5	3	2	1,00	0,00	4,24	2
12	A12	5	2	2	1,41	1,00	4,12	2
13	A13	1	1	1	4,90	4,58	1,00	3
14	A14	5	2	1	2,24	1,41	4,00	2
15	A15	1	2	2	4,24	4,12	1,00	3
16	A16	1	2	1	4,58	4,24	0,00	3
17	A17	5	3	3	0,00	1,00	4,58	1
18	A18	5	3	3	0,00	1,00	4,58	1
19	A19	1	2	1	4,58	4,24	0,00	3
20	A20	1	2	1	4,58	4,24	0,00	3

Old group = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
New group = {1, 2, 1, 1, 2, 1, 3, 1, 3, 1, 2, 2, 3, 2, 3, 3, 1, 1, 3, 3}
There is a change in the group, so it will continue to iteration 2 and a new centroid is needed. The following is the new centroid:
C1 (5,00; 2,88; 3,00)
C2 (5,00; 2,40; 1,80)
C3 (1,00; 1,86; 1,14)

After the calculation process in iteration 2, the group results are obtained in the following table:

Table. 4: Results of Group Determination in Iteration II

No	Alternative	X	Y	Z	Distance from C1	Distance from C2	Distance from C3	Group
1	A1	5	3	3	0,13	1,34	4,56	1
2	A2	5	3	2	1,01	0,63	4,25	2
3	A3	5	3	3	0,13	1,34	4,56	1
4	A4	5	3	3	0,13	1,34	4,56	1
5	A5	5	2	2	1,33	0,45	4,09	2
6	A6	5	3	3	0,13	1,34	4,56	1
7	A7	1	2	1	4,56	4,10	0,20	3

No	Alternative	X	Y	Z	Distance from C1	Distance from C2	Distance from C3	Group
8	A8	5	2	3	0,88	1,26	4,41	1
9	A9	1	2	1	4,56	4,10	0,20	3
10	A10	5	3	3	0,13	1,34	4,56	1
11	A11	5	3	2	1,01	0,63	4,25	2
12	A12	5	2	2	1,33	0,45	4,09	2
13	A13	1	1	1	4,85	4,31	0,87	3
14	A14	5	2	1	2,18	0,89	4,01	2
15	A15	1	2	2	4,21	4,02	0,87	3
16	A16	1	2	1	4,56	4,10	0,20	3
17	A17	5	3	3	0,13	1,34	4,56	1
18	A18	5	3	3	0,13	1,34	4,56	1
19	A19	1	2	1	4,56	4,10	0,20	3
20	A20	1	2	1	4,56	4,10	0,20	3

Old group = {1, 2, 1, 1, 2, 1, 3, 1, 3, 1, 2, 2, 3, 2, 3, 3, 1, 1, 3, 3}
New group = {1, 2, 1, 1, 2, 1, 3, 1, 3, 1, 2, 2, 3, 2, 3, 3, 1, 1, 3, 3}

Because in the 2nd iteration there is no group change or there is equality in the group, the iteration calculation is stopped.

4.2. Discussion

The following are the clustering results using the Matlab application, namely:

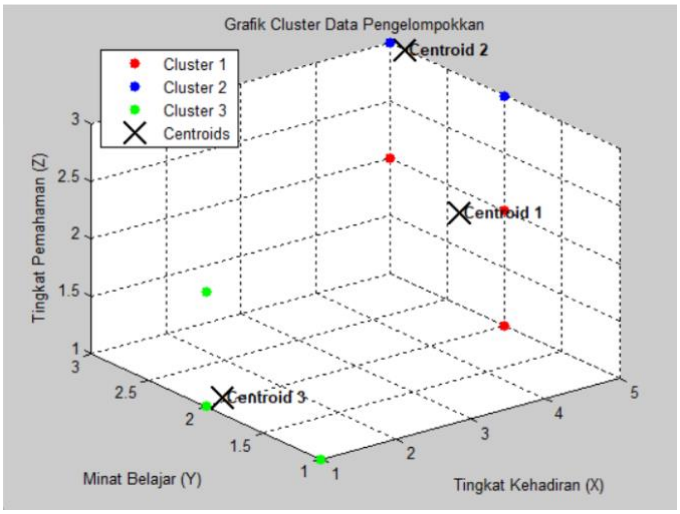

Fig. 2: Cluster Results in Matlab Application

Description:

1. Cluster 1 with 8 data (5.00; 2.88; 3.00); Cluster 1: Students with high attendance, moderate interest in learning, and moderate level of understanding. Students in this cluster have a high attendance rate (>75%) with moderate interest in learning, and have a moderate level of understanding of programming languages.
2. Cluster 2 with 5 data (5.00; 2.40; 1.80); Cluster 1: Students with high attendance, moderate interest in learning, and low level of understanding. Students in this cluster have a high attendance rate (>75%) with moderate interest in learning, and a moderate level of understanding of programming language courses.
3. Cluster 3 with 7 data (1.00; 1.86; 1.14); Cluster 1: Students with low attendance, moderate interest in learning, and low level of understanding. Students in this cluster have a low attendance rate (<75%) with moderate interest in learning, and a low level of understanding in programming language courses.

5. Conclusions and Suddestions

5.1. Conclusions

In this section you should present the conclusion of the paper. Conclusions must focus on the novelty and exceptional results you acquired. Allow a sufficient space in the article for conclusions. Do not repeat the contents of Introduction or the Abstract. Focus on the essential things of your article.

5.2. Suggestions

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

1. In future research with the same topic, it can add or change with other clustering variables in order to obtain a more comprehensive clustering.
2. It is recommended to add input data so that the clustering results performed by the system in the Matlab application can be maximized, because a lot of input data can affect the clustering results.
3. It is hoped that this research can help the campus to create a strategy to increase student interest in learning, especially those in the calculus with low understanding, by providing additional courses or learning activities that are more interactive and interesting.

References

[1] Halimah, D., Ridwan, M., Stikom, L., Bangsa, T., & Saputra, W. (2022). Algoritma C4.5 Untuk Menentukan Klasifikasi Tingkat Pemahaman Mahasiswa Pada Matakuliah Bahasa Pemrograman. *Jurnal Teknik Mesin, Industri, Elektro Dan Informatika (JTMEI)*, *1*(3).
[2] Abdurrahman, G. (n.d.). Clustering Data Ujian Tengah Semester (UTS) Data Mining Menggunakan Algoritma K-Means.
[3] Ishak, R. (2022). Clustering Tingkat Pemahaman Dasar Mahasiswa Pada Pra-Perkuliahan Probabilitas Statistika Dengan Metode K-Means. 4.
[4] Sagala, R. M. (n.d.). Prediksi Kelulusan Mahasiswa Menggunakan Data mining Algoritma K-means.
[5] Dan, P., Pemrograman, B., & Saragih, R. R. (n.d.). STMIK-STIE Mikroskil. https://www.researchgate.net/publication/329885312