Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

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

Clustering of Student Expertise Fields Using the K-Means Algorithm (Case Study: STMIK Kaputama Binjai)

Damai Aulia Br Karo[1], Yani Maulita[2], I Gusti Prahmana[3]

[1, 2, 3]STMIK Kaputama

damaiaulia5@gmail.com[1], yani.maulita@gmail.com[2] , igustiprahmana4@gmail.com[3]*

Abstract

Grouping students' fields of expertise in higher education is an important issue that can provide significant benefits for students and educational institutions. STMIK Kaputama is one of the universities that has students with various fields of expertise, but the absence of data that informs the field of expertise of students is very unfortunate. Research data was obtained through questionnaires distributed to students, which included information about study programs, Grade Point Average (GPA), and areas of expertise. Clustering analysis was conducted using Matlab software to validate and implement the clustering results. The results show that the K-Means algorithm is effective in grouping students into clusters that have similar characteristics. The first cluster consists of students with expertise in programming and database, the second cluster focuses on students with networking expertise, and the third cluster includes students with various combinations of expertise.This study also found a tendency that students with certain expertise have a higher GPA than students with other expertise.

Keywords: *Area of Expertise, K-Means, Data Mining, Clustering*

1. Introduction

In higher education, the clustering of students' areas of expertise is an important issue. Each student has diverse interests and expertise, and proper clustering can provide significant benefits for both students and universities. Through accurate clustering, universities can improve teaching efficiency, facilitate collaboration between students with similar areas of expertise.

STMIK Kaputama is one of the universities that has students with diverse fields of expertise, but the absence of data that informs the field of expertise of students is a very unfortunate thing. Therefore, the researcher created a questionnaire that can determine the field of expertise of students. After students fill out the questionnaire that researchers have distributed, many of their fields of expertise are different and there are also students who have more than one field of expertise. This study also aims to see if there is a tendency that students with expertise in certain fields have a higher Ipk compared to students with expertise in other fields.

This research is strengthened by the journal "Application of Data Mining for Grouping BPJS Employment Participants Based on the Program Taken Using the Clustering Method" using the k-means method to group participant data based on income, contributions, and programs taken. The results show three clusters: cluster 1 (370 data), cluster 2 (359 data), and cluster 3 (271 data). The cluster centroid is taken from data A1 (1,1,3), A7 (3,3,6), and A20 (4,5,4). The k-means method proved effective for this clustering [1].

The research "Data Mining Grouping of Student Areas of Expertise Using the K-Means Algorithm (Case Study: Cic Cirebon University)" applies the k-means algorithm to group students based on related courses to determine their areas of expertise. From 10 data for 3 clusters, 4 students were found in the Programmer field, 2 in Database Administrator, and 4 in Networking [2].

1.1. Data Mining

Data mining is the process of finding and extracting information from large sets of data. The main process of data mining is to identify patterns or relationships that are not directly visible in piles of data. The main goal is to generate new information that provides deeper understanding or insight [3], [4].

1.2. Clustering

Clustering is one of the data mining techniques used to obtain groups of objects that have common characteristics in large enough data. The main purpose of the clustering method is to group a number of data or objects into clusters or groups so that each cluster will contain data that is as similar as possible. Clustering performs data grouping based on similarities between objects [5], [6].

1.3. K-Means

The K-Means algorithm is a nonhierarchical data clustering method that aims to partition data into two or more groups based on similar characteristics. In the K-Means method, data is partitioned so that objects with similar characteristics are grouped together in one group, while objects with different characteristics are grouped in different groups. The main objective of this data clustering is to minimize a specified objective function, which generally seeks to reduce variation within each group and increase variation between groups [7], [8].

2. Research Methods

2.1. Research Methods

In the research method, it is carried out to find something systematic by using scientific methods and applicable sources. With this research method process, it can provide good and precise research results. There are several stages of research methods carried out in problem solving. These stages are as follows [9]:

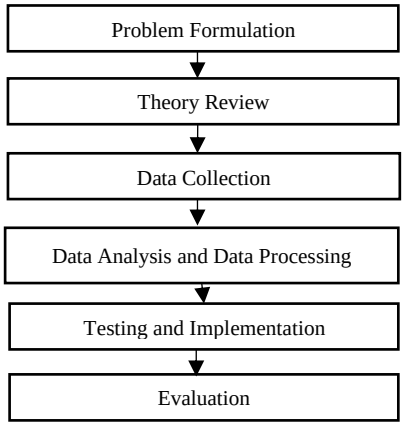

Fig. 1: Research Workflow

2.2. Research Supporting Data

Table 1: Data to be processed

No	Alternative	Study Program	Grade	Expertise
1	A1	SI	3,51 - 4,00	Programming and Databases
2	A2	SI	3,51 - 4,00	Programming
3	A3	TI	3,01 - 3,50	Artificial Intelligence
4	A4	SI	3,01 - 3,50	Networking
5	A5	SI	3,01 - 3,50	Programming and Databases
6	A6	SI	3,51 - 4,00	Database
7	A7	MI	3,01 - 3,50	Networking
8	A8	TI	3,01 - 3,50	Artificial Intelligence
9	A9	TI	3,51 - 4,00	Network Programming, and Artificial Intelligence
10	A10	SI	3,51 - 4,00	Programming, Databases, and Networks
11	A11	TI	3,51 - 4,00	Programming, Databases, Networks, and Artificial Intelligence
12	A12	SI	3,01 - 3,50	Programming
13	A13	SI	3,51 - 4,00	Networking
14	A14	TI	3,51 - 4,00	Programming and Networking
15	A15	SI	3,01 - 3,50	Networking
16	A16	SI	3,01 - 3,50	Programming, Databases, and Networking
17	A17	SI	3,01 - 3,50	Database and Networking
18	A18	SI	3,51 - 4,00	Programming and Databases
19	A19	SI	3,01 - 3,50	Programming and Databases
20	A20	SI	3,01 - 3,50	Networking

3. Results and Discussion

3.1. Results

In the analysis of testing the Clustering method on the data grouping system, data is needed as input for processing and analysis. After conducting research at STMIK Kaputama Binjai by distributing questionnaires, the author obtained data regarding the fields of expertise of STMIK Kaputama Binjai students. The data that has been transformed is as follows:

Table 2: Data to be processed

No	Alternative	X	Y	Z
1	A1	1	4	5
2	A2	1	4	1
3	A3	2	3	4
4	A4	1	3	3
5	A5	1	3	5
6	A6	1	4	2
7	A7	3	3	3
8	A8	2	3	4
9	A9	2	4	13
10	A10	1	4	11
11	A11	2	4	15
12	A12	1	3	1
13	A13	1	4	3
14	A14	2	4	6
15	A15	1	3	3
16	A16	1	3	11
17	A17	1	3	8
18	A18	1	4	5
19	A19	1	3	5
20	A20	1	3	3

Calculation:
The number of clusters (K) is 3 clusters.
The cluster center (Centroid) used is 3, which is as follows:
Centroid 1 = (1,4,5) taken randomly from data 1
Centroid 2 = (1,4,1) taken randomly from data 2
Centroid 3 = (1,4,11) taken randomly from data 10
To determine the group of an object, the first thing to do is to measure the distance, the Euclidean distance between two points or objects or X,Y, and Z which is defined as follows:

Euclidean $\sqrt{\sum(X1-X2)^2+(Y1-Y2)^2+(Z1-Z1)^2}$

The table below is a table that shows the results of determining the group at iteration I as follows:

Table 3: Results of Group Determination in Iteration I

No	Alternative	Study Program (X)	Grade (Y)	Expertise (Z)	Distance From C1	Distance From C2	Distance From C3	Groups
1	A1	1	4	5	0,00	4,00	6,00	1
2	A2	1	4	1	4,00	0,00	10,00	2
3	A3	2	3	4	1,73	3,32	7,14	1
4	A4	1	3	3	2,24	2,24	8,06	1
5	A5	1	3	5	1,00	4,12	6,08	1
6	A6	1	4	2	3,00	1,00	9,00	2
7	A7	3	3	3	3,00	3,00	8,31	1
8	A8	2	3	4	1,73	3,32	7,14	1
9	A9	2	4	13	8,06	12,04	2,24	3
10	A10	1	4	11	6,00	10,00	0,00	3
11	A11	2	4	15	10,05	14,04	4,12	3
12	A12	1	3	1	4,12	1,00	10,05	2
13	A13	1	4	3	2,00	2,00	8,00	1
14	A14	2	4	6	1,41	5,10	5,10	1
15	A15	1	3	3	2,24	2,24	8,06	1
16	A16	1	3	11	6,08	10,05	1,00	3
17	A17	1	3	8	3,16	7,07	3,16	1
18	A18	1	4	5	0,00	4,00	6,00	1
19	A19	1	3	5	1,00	4,12	6,08	1
20	A20	1	3	3	2,24	2,24	8,06	1

Group based on the minimum distance to the nearest centroid which is:
If the shortest distance is in C1 then the data is included in group 1
If the shortest distance is in C2 then the data is included in group 2
If the shortest distance is in C3 then the data is included in group 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 = {1,2,1,1,1,2,1,1,3,3,3,2,1,1,1,3,1,1,1,1}
There is a group change, so proceed to the next iteration,
For group 1 there are 13 data, namely:
C1 = (1+2+1+1+3+2+1+2+1+1+1+1+1+1+1+1+1)/13 = 1,38
C2 = (4+3+3+3+3+3+4+4+3+3+4+3+3+3+3+3)/13= 3,30
C3 = (5+4+3+5+3+4+3+6+3+8+5+5+3)/13 = 4,38

For group 2 there are 3 data, namely:
C1 = (1+1+1)/3 = 1,00
C2 = (4+4+3)/3 = 3,66

C3 = (1+2+1)/3 = 1,33

For group 3 there are 4 data, namely:
C1 = (2+1+2+1)/4 = 1,50
C2 = (4+4+4+3)/4 = 3,75
C3 = (13+11+15+11)/4 = 12,50
The table below is a table that shows the results of determining the group in iteration II as follows:

Table 4: Results of Group Determination in Iteration II

No	Alternative	Study Program (X)	Grade (Y)	Expertise (Z)	Distance From C1	Distance From C2	Distance From C3	Groups
1	A1	1	4	5	1,01	3,69	7,52	1
2	A2	1	4	1	3,47	0,47	11,51	2
3	A3	2	3	4	0,79	2,93	8,55	1
4	A4	1	3	3	1,46	1,80	9,54	1
5	A5	1	3	5	0,79	3,73	7,55	1
6	A6	1	4	2	2,51	0,75	10,51	2
7	A7	3	3	3	2,15	2,69	9,65	1
8	A8	2	3	4	0,79	2,93	8,55	1
9	A9	2	4	13	8,67	11,72	0,75	3
10	A10	1	4	11	6,67	9,68	1,60	3
11	A11	2	4	15	10,66	13,71	2,56	3
12	A12	1	3	1	3,41	0,74	11,54	2
13	A13	1	4	3	1,59	1,70	9,52	1
14	A14	2	4	6	1,87	4,79	6,52	1
15	A15	1	3	3	1,46	1,80	9,54	1
16	A16	1	3	11	6,64	9,69	1,75	3
17	A17	1	3	8	3,65	6,70	4,59	1
18	A18	1	4	5	1,01	3,69	7,52	1
19	A19	1	3	5	0,79	3,73	7,55	1
20	A20	1	3	3	1,46	1,80	9,54	1

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 = {1,2,1,1,1,2,1,1,3,3,3,2,1,1,1,3,1,1,1,1}
New Group = {1,2,1,1,1,2,1,1,3,3,3,2,1,1,1,1,3,1,1,1,1}

3.2. Discussion

From the results of the discussion of the old group from iteration 1, the results of the discussion of the new group are also obtained as listed above, because the results of the 1st iteration and the 2nd iteration do not change or there are similarities, then the iteration calculation is stopped, here are the results of the Clustering graph using the Matlab programming application:

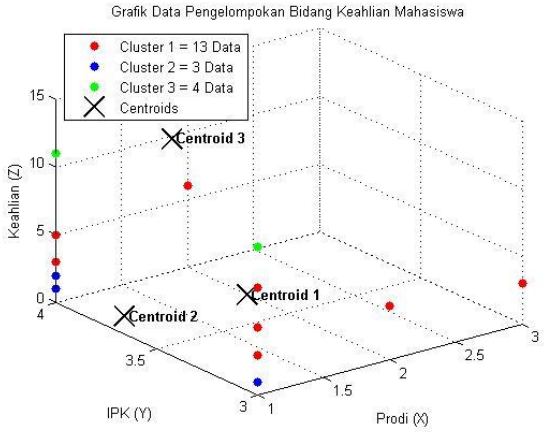

Fig. 2: Clustering Chart

Description:
X Y Z
Sentroid 1 = (1.38; 3.30; 4.38)
Centroid 2 = (1.00; 3.66; 1.33)
Sentroid 3 = (1.50; 3.75; 12.50)
The explanation of the above results is:
Of the 20 data there are 3 groups 1 there are 13 data, group 2 there are 3 data and group 3 there are 4 data, The explanation of the 3 groups is as follows:

1. Cluster 1 There are 13 data (1.38; 3.30; 4.38)
   Data can be grouped based on the "Information Systems" Study Program with Ipk "3.01-3.50 (Very Satisfactory)" and "Programming and Database" expertise,

2. Cluster 2 There are 3 data (1.00; 3.66; 1.33)
   The data can be grouped based on the "Information Systems" Study Program with Ipk "3.51-4.00 (Complimentary)" and "Programming" expertise,
3. Cluster 3 There are 4 data (1.50; 3.75; 12.50)
   Data can be grouped based on "Information Systems" Study Program with Ipk "3.01-3.50 (Satisfactory)" and expertise "Programming, Database & Network".

From the explanation above, it can be stated that there is a tendency that students with satisfactory Grade (Ipk) only have 1 field of expertise while students with very satisfactory Ipk have 2 or 3 fields of expertise.

4. Conclusions

Based on the research that has been done on grouping students' fields of expertise using the K-Means algorithm at STMIK Kaputama Binjai, some conclusions that can be drawn are as follows:

1. K-Means algorithm is effective in grouping students based on their field of expertise. This grouping is based on data collected through questionnaires that have been distributed to students. The results show the existence of several groups or clusters that group students according to their interests and expertise.
2. The iteration process is performed to measure the Euclidean distance between the data points and the centroid, then determine the group of each object based on the shortest distance to the nearest centroid. This iteration process is repeated until there is no significant group change.
3. Based on the iterations performed, it is found that student data is distributed in three groups or clusters with the following centroids, group 1 is dominated by students with expertise in Programming and Database, group 2 students with expertise in Programming and Artificial Intelligence, group 3 students with a more complex combination of expertise, including Programming, Networking, and Artificial Intelligence.
4. This research also shows that there is a tendency for students with expertise in certain fields to have different GPAs compared to students in other fields of expertise. For example, students with expertise in Programming and Database tend to have higher GPAs compared to other groups.
5. Proper grouping can provide significant benefits to the college. By knowing students' areas of expertise, STMIK Kaputama can improve teaching efficiency, facilitate collaboration between students with similar skills, and direct them to career paths that match their expertise. It also allows the college to craft a curriculum that better suits students' interests and needs.

5. Suggestions

The suggestions put forward in this study are:

1. STMIK Kaputama Binjai is advised to develop a decision support system that uses the K-Means algorithm to routinely classify students. This system can assist campus management in making decisions related to curriculum development, internship placement, and student career guidance.
2. Future research can expand the variables used in clustering. In addition to Prodi, GPA, and Expertise, additional variables such as interests, extracurricular activities, work experience, and certifications can also be considered. More comprehensive data will result in more accurate and relevant groupings.
3. It is recommended to organize trainings and workshops for students in the fields that are the main expertise of each cluster. This will help students to further explore and develop their expertise, and better prepare them to enter the workforce.

References

[1] Buaton, R,, Sundari, Y,, & Maulita, Y, (2016), Clustering Tindak Kekerasan Pada Anak Menggunakan Algoritma K-Means Dengan Perbandingan Jarak Kedekatan Manhattan City Dan Euclidean, MEANS (Media Informasi Analisa Dan Sistem), 47–53, https://doi,org/10,54367/means,v1i2,8

[2] Nas, C, (2020b), Data Mining Pengelompokan Bidang Keahlian Mahasiswa Menggunakan Algoritma K-Means (Studi Kasus : Universitas Cic Cirebon), In Syntax: Jurnal Informatika (Vol, 09, Issue 1),

[3] Nur Khomarudin, A, (2003), Teknik Data Mining : Algoritma K-Means Clustering, https://agusnkhom,wordpress,com

[4] Maulita, Y,, & Arliana, L, (2022), Penerapan Data Mining Pengelompokan Peserta Bpjs Ketenagakerjaan Berdasarkan Program Yang Diambil Menggunakan Metode Clustering, Jurnal Sistem Informasi Kaputama (Jsik), 6(2),

[5] Mayona, Y,, Buaton, R,, & Simanjutak, M, (2022a), Data Mining Clustering Tingkat Kejahatan Dengan Metode Algoritma K-Means (Studi Kasus : Kejaksaan Negeri Binjai), Agustus, 6(3).

[6] 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.

[7] 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.

[8] 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.

[9] 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.