Emerging Information Science and Technology
Vol. 1, No. 3, (2020), pp. 75-83
DOI: 10.18196/eist.v1i3.13172

The Implementation of Clustering Method With K-Means Algorithm In Grouping Data of Students' Course Scores at Universitas Muhammadiyah Yogyakarta

Dita Kurniasari[1], Asroni[1*] and Slamet Aprilia Kurniasari[1]

[1]*Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia*

Corresponding author: asroni@umy.ac.id

Abstract

Student grades can be a reference. A large number of student grade data in a university causes data accumulation; thus, data are grouped with data mining. This study aims to classify student grade data in the second semester. Grouping student grade data was performed using the clustering method with the K-means algorithm. The research data were derived from the database of Universitas Muhammadiyah Yogyakarta. The data were students' grades in the academic years of 2010/2011, 2011/2012, 2012/2013, 2013/2014, and 2014/2015. The analysis process was carried out using WEKA software, SQL Server 2014 Management Studio and Microsoft Excel. The clustering method could be applied to group student grade data. Clustering with K-means formed three clusters, with cluster 0 comprising 72 students, cluster 1 consisting of 190 students, and cluster 2 totaling 133 students. A cluster with the lowest average score could be used as a consideration in updating the learning methods to optimize students' score acquisition.

Keywords: *Data mining, clustering, K-Means, WEKA*

1. Introduction

Education plays an essential role for a nation. The future of a nation depends on the quality of its human resources and students' ability to master science and technology. Without education, a nation will be left behind by other nations. Education in Indonesia has several levels, such as primary, secondary and higher education, in which each level has its learning method.

Indonesian public awareness toward education is relatively high. Parents try to send their children to universities with excellent accreditation and learning methods. Learning methods are essential to help students absorb knowledge optimally. Besides, learning methods are also applicable for daily life and the world of work.

Universitas Muhammadiyah Yogyakarta (UMY) is a private university accredited by the Higher Education Accreditation Board (BAN-PT), included in the top 50 best universities in Indonesia (DIKTI version). It is inseparable from how the teaching and learning system implemented by the university affects the quality of students and increases their academic abilities.

Students are one of the critical factors for universities in improving their education quality. Course score achievement for each student can be a reference. With many students and courses in each department, more data will be entered into the database server at the

Information Systems Bureau (BSI) of UMY. The amount of incoming data certainly causes data to accumulate over a long time. These data can be processed to find useful information for the university.

Many companies and educational institutions have implemented data mining. Several studies had investigated the application of data mining to extract information from a student database as reference material. In research using the Decision Tree algorithm, Bayes classifier, logistic models, rule-based learning and random forest, monitoring and student support in the first year were stated to be crucial. Students majoring in Electrical Engineering at the University of Eindhoven who quit their studies in the first year accounted for up to 40%. The difficult curriculum was considered as one of the reasons for the high number of students dropping out. Moreover, grades, achievements, personality, and social background also played a role in student academic success [1].

In a study entitled Analysis and Prediction of Student Performance Using Data Mining Technique, GPA was one requirement to get a job in a company. Besides, student achievement was one of the determining factors for the quality of higher education. As evidenced in the department accreditation forms in the standard 3A book 3 Students and Graduates, one of the factors is the GPA. With the higher average GPA of students, more than 3.00, the assessment on the forms would also receive the maximum score of four, resulting in a better quality of the department [2].

A study explained that the most influential factors in determining student academic performance classification were the GPA, semester GPA of the first and fourth semesters, and gender. This study utilized the C4.5 algorithm in determining graduation predictions based on gender attributes from high school and social studies from the first to the sixth semesters [3].

A previous study classified student academic performance based on the GPA using K-means clustering. It aimed to develop software to classify students based on their academic performance. The success of grouping student data was calculated based on the standard error score of the repeated grouping process. A total of 100 data samples from the fifth-semester students of the Computer Science Faculty of Universitas Sriwijaya were collected as test data to be grouped. The data used as a parameter of academic performance was the student GPA data from the first to the fourth semesters. The data were then grouped 50 times [4].

A study entitled Application of the K-means Method for Student Clustering Based on Academic Values using WEKA Interface (A Case Study at the Informatics Engineering Department of Universitas Muhammadiyah Magelang) tested the existing warehouse data to gather five students majoring in Informatics Engineering to participate in the Cyberjawara competition organized by the Indonesian Security Incident Response Team on the Internet Infrastructure (ID SIRTII) of the Communication and Information Ministry of the Republic of Indonesia. The data consisted of five attributes: student number, algorithm and programming-1 course scores, basic physics course scores, calculus-1 course scores, and GPA with 124 instances. This research employed WEKA software to compare the results with theoretical calculations with those obtained at WEKA Interface. The distance calculation utilized the Euclidean equation. It resulted in four cluster groups: cluster 0 with GPA = 0.5167 comprising nine students (7%), cluster 1 with GPA = 3.4143 consisting of 28 students (23%), cluster 2 with GPA = 3.3092 totaling 40 students (32%), and cluster 3 with GPA = 3.8991 amounted to 47 students (38%). Thus, cluster 1 with the highest GPA was used to select five students to join the competition [5].

In this study, the authors applied the clustering method with the K-means algorithm in grouping Informatics Engineering Department students based on weighted score data for

courses and semester GPA in the second semester. This clustering technique classified students based on the similarities of the weighted score data of the courses and their IPS. The results of this study are expected to provide information to determine policies in improving learning methods in the Informatics Engineering Department.

Data mining was employed to process student score data at Universitas Muhammadiyah Yogyakarta (UMY). It can perform abundant data processing to dig up hidden information in an extensive database, usually called knowledge discovery in a database (KDD). The clustering method was utilized to group student score data. There are two data clustering types, hierarchical and non-hierarchical. This study applied the non-hierarchical clustering called the K-means algorithm.

2. Method

This research was conducted at Universitas Muhammadiyah Yogyakarta (UMY) from September 2019 to October 2020 using the clustering method with the K-means algorithm. Figure 1 displays the process flow on K-means. In the K-means process, the first step taken was determining k as the number of clusters desired, followed by randomly determining the starting centroid (center) points. After determining the centroid points, the calculation was performed on the distance between the objects and the centroid points. The objects were then grouped based on the minimum distance. If there was a change in the centroid position, the calculation process was repeated. If nothing changed, the K-means process was complete.

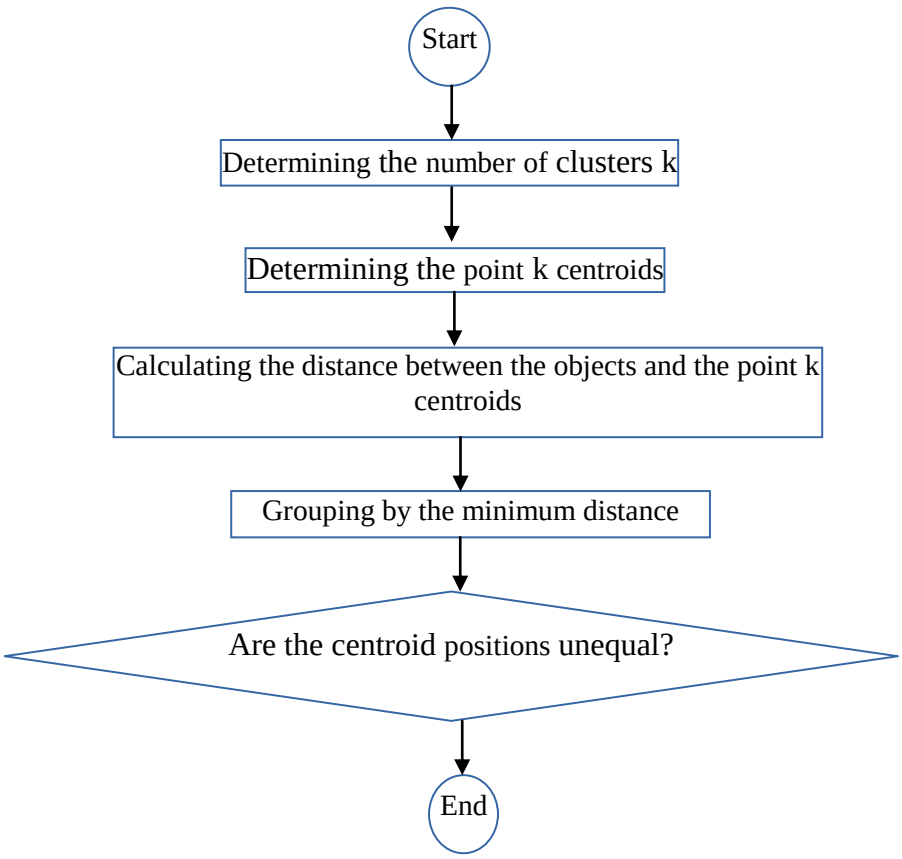

Figure 1. Flowchart K-means

3. Results

The authors used the student scores of the Informatics Engineering Department of the Engineering Faculty for five years, the academic years of 2010/2011 to 2014/2015, processed using the SQL Server 2014 Management Studio [6]. These data were taken as they could be accurate or valid over a specific period, for example, five to ten years. Access permission was obtained from the Information System Bureau (BSI) of UMY to collect the research data. The dataset used is presented in Figure 2 [7] [8].

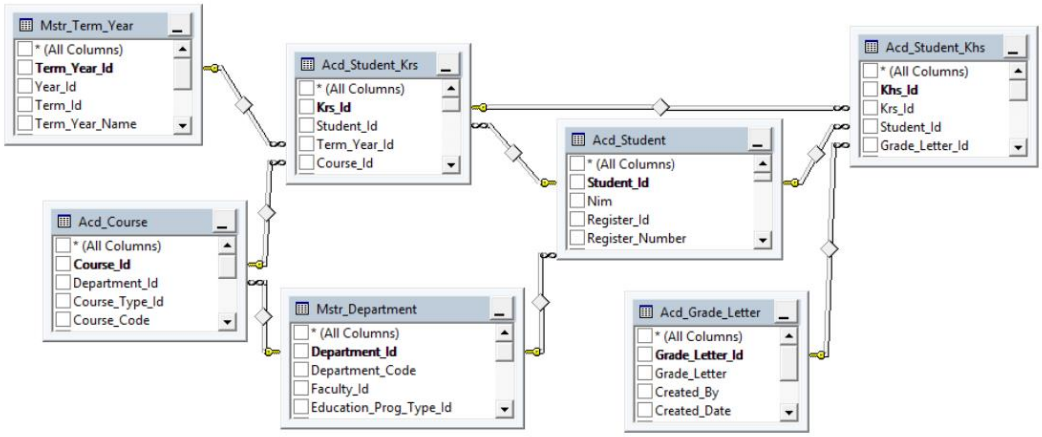

Figure 2. Dataset display

Data mining is an activity that describes an analysis process that occurs iteratively in an extensive database, intending to extract accurate and potentially useful information and knowledge for workers related to decision making and problem-solving (Vercellis, 2009). Data mining is often called knowledge discovery in a database (KDD). KDD is the whole non-trivial process to look for and identify patterns in data, where the patterns found are valid, new, useful and understandable [9].

The terms data mining and KDD are frequently used interchangeably to describe the process of extracting hidden information in an extensive database. These two terms have different concepts but are related to one another. Moreover, one of the stages in the whole KDD process is data mining, as shown in Figure 2. The KDD process, in general, can be explained as follows [10].

After preprocessing the data, the Elbow method was applied to determine the best number of clusters. The data tested amounted to 395. The number of clusters tested was from K = 2 to K = 10. Table 1 demonstrates the Sum of Square Error (SSE) results.

Table 1. Sum of Square Error Results

Cluster	Sum of Square Error	Difference
k= 2	159,7077822	159,7077822
k = 3	139,5344582	20,17332395
k = 4	128,1204876	11,41397064
k = 5	112,5599592	15,5605284
k = 6	106,6607634	5,899195849
k = 7	102,40885	4,251913346
k = 8	100,368056	2,04079403
k = 9	93,19791638	7,170139596
k = 10	89,52486005	3,673056331

The following figure displays the graph of the SSE results.

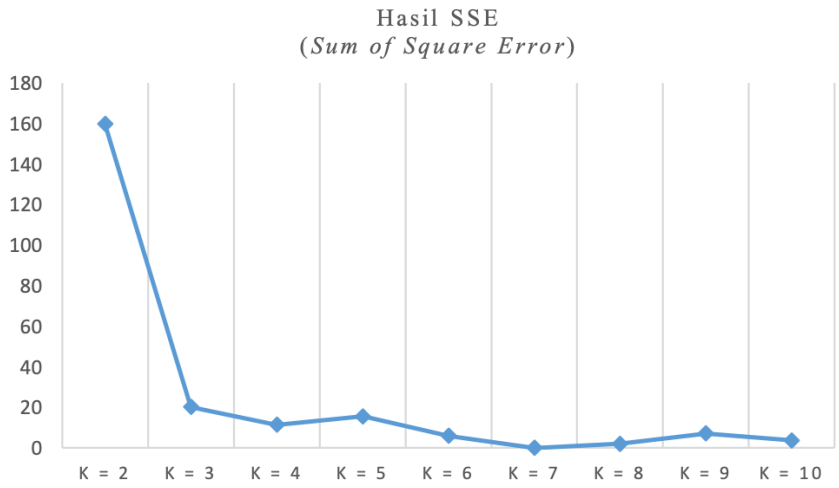

Figure 3. Graph of the SSE results

The graph shows that K = 3 is the optimal number of clusters because the elbow is formed at K = 3.

Figure 4 exhibits the calculation results using the K-means algorithm with 3 clusters functioning in grouping data on course scores. The results of cluster centroids using WEKA can be seen in Figure 4. The WEKA calculation resulted in 19 iterations [6].

Final cluster centroids:

Attribute	Full Data (395.0)	Cluster# 0 (72.0)	1 (190.0)	2 (133.0)
Agama Islam 2 (3)	3.7747	3.6667	3.9053	3.6466
Bahasa Inggris 2 (2)	3.1241	2.4722	3.6105	2.782
Implementasi Basis Data (3)	2.9418	2.5208	3.3789	2.5451
Logika dan Teknik Pemrograman(2)	2.8608	2.5139	3.1447	2.6429
Pemrograman Berorientasi Obyek(2)	3.5038	2.3819	3.7947	3.6955
Pengembangan Aplikasi Windows(2)	3.4975	2.7292	3.8553	3.4023
Perancangan Basis Data(2)	3.1797	2.6597	3.4184	3.1203
Praktikum Implementasi Basis Data(1)	2.9443	2.5208	3.3789	2.5526
Praktikum Pemrograman Berorientasi Obyek(1)	3.5013	2.3681	3.7947	3.6955
Praktikum Pengembangan Aplikasi Windows(1)	3.5025	2.7431	3.8553	3.4098
Indeks Prestasi	3.2854	2.7214	3.6061	3.1327

Figure 4. Results of cluster centroids

The results of cluster instances using WEKA are as follows [5]: Cluster 0 (the lowest score group) consisted of 72 students out of 395 student data (18%). Cluster 1 (the highest score group) contained 190 students from 395 student data (48%). Cluster 3 (the moderate score group) comprised 133 students from 395 student data (34%). Figure 5 demonstrates the result visualization of the cluster instances carried out using WEKA.

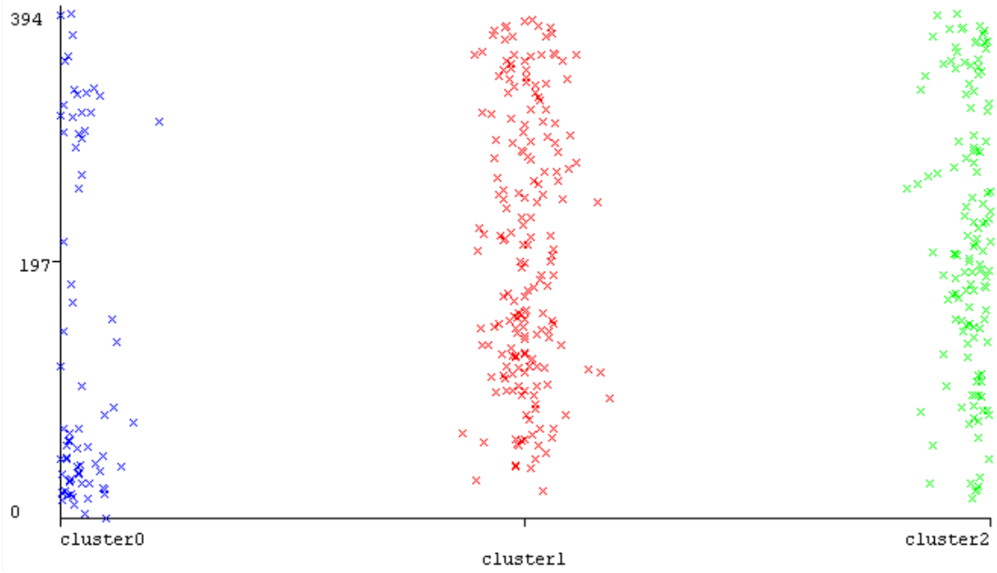

Figure 5. Visualization of cluster results

The calculation using the K-means algorithm was repeated until it reached stability. The testing on student grade data was performed through the following stages. Determining the number of clusters and the center point coordinates of the clusters. Three cluster groups were made. The score data were taken from each attribute and then randomly determined to discover the center of the clusters. Determining the cluster scores used as a reference in

calculating the distance of the objects to the centroids. The distance calculation refers to the Euclidean formula run using Microsoft Excel.

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

Equation 1

After the distance between centroids was calculated using the Euclidean distance formula, the centroids were grouped based on the distance between centroids. The distance calculation results were used to determine the clustering groups. The grouping determination is as follows.

If the distance on centroid 1 is smaller than the distance between centroids 2 and 3, it is included in centroid 1.

If the distance on centroid 2 is smaller than the distance between centroids 3 and 1, it is included in centroid 2.

If the distance on centroid 3 is smaller than the distance between centroids 2 and 1, it is included in centroid 3.

The attribute names were written as X and followed by the numbers to facilitate calculation, as shown in Table 2.

Table 2. Results Names of attributes

Islam Religion 2	X1
English 2	X2
Database Implementation	X3
Logic and Programming Techniques	X4
Object-Oriented Programming	X5
Windows Application Development	X6
Database Design	X7
Database Implementation Practice	X8
Object-Oriented Programming Practice	X9
Windows Application Development Practice	X10
Grade Point Average (GPA)	X11

The data were then grouped into three clusters. The cluster centroid results are exhibited in Table 3.

Table 3. Cluster centroids

	Cluster		
	1	2	3
X_1	3,66	3,90	3,64
X_2	2,47	3,61	2,7
X_3	2,52	3,37	2,54
X_4	2,51	3,14	2,64
X_5	2,38	3,79	3,69
X_6	2,72	3,85	3,40
X_7	2,65	3,41	3,12
X_8	2,52	3,37	2,55
X_9	2,36	3,79	3,69
X_{10}	2,74	3,85	3,40
X_{11}	2,72	3,60	3,13

Table 4. Object distance to the centroids

Centroid 1	Centroid 2	Centroid 3
1.56	2.84	1.96
2.09	2.10	1.94
1.56	2.67	1.71
2.55	2.89	2.76
2.20	2.70	2.30
3.02	3.19	3.40
4.36	1.43	3.02
4.58	1.51	3.17
1.52	1.95	1.42
2.28	3.16	2.71
2.26	2.92	2.74
2.36	2.35	1.81
2.50	4.84	3.89
2.62	4.68	3.26
3.20	6.06	4.64
1.56	2.67	1.71
2.68	4.64	3.27
1.40	3.19	2.68
2.77	3.08	3.28

Table 4 presents the results of the distance from the objects to the centroids. Centroid groups is chosen based on the minimum value between Centroid 1, Centroid 2, and Centroid 3. The chosen group is marked with green color in Table 4.

4. Conclusions

After calculating using the K-means algorithm to classify students based on data on student scores majoring in Informatics Engineering using 11 attributes, several conclusions were drawn as follows: (1) Cluster 0 consisted of 72 students with the characteristics of low course scores of the Islamic Religion 2, English 2, Database Implementation, Logic and Programming Techniques, Object-Oriented Programming, Windows Application Development, Database Design, Database Implementation Practice, Object-Oriented Programming Practice, and Windows Application Development Practice, and low GPA of 2.72. (2) Cluster 1 comprised 190 students with the characteristics of high course scores of Islamic Religion 2, English 2, Database Implementation, Logic and Programming Techniques, Object-Oriented Programming, Windows Application Development, Database Design, Database Implementation Practice, Object-Oriented Programming Practice, and Windows Application Development Practice, and high GPA of 3.60. Students included in cluster 2 were those with course scores and GPA above the average. (3) Cluster 2 consisted of 133 students with the characteristics of high course scores of Islamic Religion 2, English 2, Database Implementation, Logic and Programming Techniques, Object-Oriented Programming, Windows Application Development, Database Design, Database Implementation Practice, Object-Oriented Programming Practice, and Windows Application Development Practice, and GPA of 3.13, considered high, but still under cluster 2. (4) Improving the quality of learning in the Informatics Engineering Department was carried out by considering students' course scores and GPA. Cluster 0, dominated by students with low scores, could be used as an evaluation of the department to update the learning methods to optimize students' scores.

References

[1] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop Out: A Case Study.," International Working Group on Educational Data Mining, 2009.

[2] S. Defiyanti, "Analisis dan Prediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining," Syntax: Jurnal Informatika, vol. 2, no. 01, 2013.

[3] M. Ridwan and H. Suyono, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," vol. 7, no. 1, p. 6, 2013.

[4] R. Z. Alberto, W. K. Sari, and A. Primanita, "PENGELOMPOKKAN PERFORMA AKADEMIK MAHASISWA BERDASARKAN INDEKS PRESTASI MENGGUNAKAN K-MEANS CLUSTERING," KNTIA, vol. 4, 2017.

[5] A. Asroni and R. Adrian, "Penerapan metode K-means untuk clustering mahasiswa berdasarkan nilai akademik dengan Weka Interface studi kasus pada jurusan Teknik Informatika UMM Magelang," Semesta Teknika, vol. 18, no. 1, pp. 76–82, 2015.

[6] J. García-Tobar, "Study of Indoor Radon Using Data Mining Models Based on OLAP Cubes," Physical Science International Journal, pp. 53–61, 2020.

[7] J. Hunker, A. A. Scheidler, and M. Rabe, "A systematic classification of database solutions for data mining to support tasks in supply chains," 2020, pp. 395–425.

[8] F. Meskine and S. Nait-Bahloul, "A support architecture to MDA contribution for data mining," International Journal of Data Mining, Modelling and Management, vol. 12, no. 2, pp. 207–236, 2020.

[9] C. Vercellis, Business intelligence: data mining and optimization for decision making. Wiley Online Library, 2009.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework.," 1996, vol. 96, pp. 82–88.