Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.8 No.1 January 2023, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Clustering Courses Based On Student Grades Using K-Means Algorithm With Elbow Method For Centroid Determination

Muhammad Al Ghifari[1], Wahyuningdiah Trisari Harsanti Putri[2]

[1,2]*Informatics Engineering Department, Paramadina University, Indonesia*

[1]muhammad.ghifari@student.paramadina.ac.id

[2]wahyuningdiah.trisari@paramadina.ac.id(*)

Received: 2022-05-31; Accepted: 2022-10-31; Published: 2023-10-27

Abstract— Students who have taken courses will receive grades from a performance index with a weight of 0 to 4. The amount of historical student data, particularly on course grades, has the potential to discover new insights. Still, course grades are closed data and are only for academic and management purposes. The research aims to a grouping of courses with high average grades. In this research, the clustering of courses using the k-means clustering algorithm using the elbow method to determine the centroid. Based on the Sum of Squares calculation, the optimal number of clusters with k=2 was obtained. The clustering results produced cluster 1 with a centroid value of 2.686 and 15 members and cluster 2 with a centroid value of 3.245 and 40 members. It can be concluded from this research that the members of cluster 2 are a group of courses with high average grades.

Keywords—K-Means, Elbow Method, Student Course, Clustering, Data Mining

I. INTRODUCTION

Students participate in and learn from classes measured in Semester Credit System (SCS) units, which can be earned after each semester following the number of SCS earned in that particular semester. The grades the pupils receive come from a performance index that can be either numbers or letters and ranges from 0 to 4 in terms of weight value. In 1785, Yale University first implemented these four index grades as part of a ranking system known as Optimi students, second Optimi, Inferiores, and Pejores [1]. At the time, grades were still referred to and did not employ numerics. Then, in 1813, Yale University began using a numerical grading system by introducing a minus or plus sign into the assessment process.

With the rapid advancement of technology and the rise in the amount of information that is being stored, massive amounts of data can be analyzed using data mining [2], also known as Knowledge Discovery in Databases (KDD) [3]. In 1989 [4,] the concept of data mining was initially presented; nevertheless, at that time, only a limited number of data mining algorithms were simple to implement. Clustering is one of the strategies used in data mining, which is a field that analyses enormous data sets and searches for hidden potentially relevant information using several different ways [5]. Finding a center point that can be used as a reference to create groups from a data source that will be divided into several clusters is the goal of the clustering technique [6]. Additionally, the clustering technique seeks to find cluster structures in data distinguished by similarities in a data value. There is a wide variety of approach types available regarding clustering algorithms. The technique that was utilized in this investigation is known as k-means [7]. One way to accomplish this goal is by using university student data as a basis for applying the k-means algorithm with large datasets [8].

The amount of historical student data has the potential to discover new insights and be useful for the university and students, especially at Paramadina University. However, course grade data, particularly data on the computer engineering program, is closed and only processed internally by academics for management purposes. Still, as a student, the author is curious about this internal data to find out how to identify groups of computer engineering courses with high average student grades. Therefore, access to the data is granted by the academic side for this research.

To identify groups of computer engineering courses with high average student grades, we use the clustering method to determine the number of centroids using the elbow method.

II. LITERATURE REVIEW

Previously, there have been several studies that have used the k-means algorithm in the academic scope of the campus, such as the application of the k-means method for GPA performance [9]. The GPA clustering with the k-means method integrated with SQL [10], academic evaluation clustering using k-means [11], and the application of k-means for clustering Bidikmisi scholarship students [12]. From these studies, the selection of the number of clusters is still randomly taken with the provision that the number of clusters is smaller than the number of data (k <n) to find the optimal cluster value. In this research, the selection of the number of clusters uses the elbow method (elbow effect). The data used is student course grade data.

A. Elbow Method

The Elbow approach is a statistical technique that examines the proportion of the variance that can be attributed to the effect of the cluster size [13]. This method is used to find the optimal number of clusters to be used in the K-Means algorithm [14], where it works by adding clusters, the distortion value will

DOI : http://dx.doi.org/10.25139/inform.v8i1.4519

decrease rapidly based on the recording of the Sum Square Error (SSE) [15], using Equation (1). Where the y_i variable is the predicted value, and the $\bar{y}$ variable is the actual value.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (1)$$

The calculation is carried out multiple times, with the outcome of each squared distance and the total number of n clusters growing with each iteration of the calculation. After iteration, the number of clusters will affect the SSE value, which will continue to decrease until it is finished on iteration n. A graph is used to illustrate the outcomes that were acquired from the study. The point of the SSE result that has the shortest angle or point from the beginning of the landing is the one that is selected. This is done because this point represents the best compromise between the size of the SSE and the clusters.

B. K-Means Algorithm

The K-Means algorithm is one of the simple and easy-to-use unsupervised machine learning algorithms [16], and is an algorithm that works based on grouping and division (partition) [17]. This algorithm is chosen because of its popularity and simple repetition to find a minimal solution [18]. The way the k-means algorithm works is as follows::

- It randomly determines the initial number and value of centroids from the obtained data.
- Calculate each data's distance to the cluster center using the Eucledian Distance in Equation (2). The variable d is the distance, y is the data value, and x is the centroid data.

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

- Grouping clusters based on the nearest distance to the cluster center.
- Repeating until there are no changes in cluster membership.

Although the k-means algorithm is easy to implement and popular, there are shortcomings in the iteration time required by the algorithm [19]. The average complexity given by the k-means algorithm is *O(k n T)*, with the value of n being the number of samples and T being the number of iterations. In contrast, for the worst case, the complexity given is exponential $O(n^{(k+2/p)})$, with p being the number of features.

C. Scikit-learn

There are many auxiliary tools or libraries from programming languages to facilitate the clustering process, one of which is scikit-learn, which can simplify the calculation of K-Means with accurate results. Scikit-learn is a library from the Python programming language that integrates various machine-learning algorithms for supervised and unsupervised problems [20]. This library focuses on implementing machine learning as a high-level programming language to make it easy for non-specialists to use.

This library is built from several APIs (Application Programming Interfaces), which consist of 3 complementary ones that can be used and interconnected, namely Estimators for building and adjusting models, Predictors for making predictions, and Transformers for transforming data into results [21].

III. RESEARCH METHODOLOGY

The research method in this paper uses quantitative research methods using an experimental research design carried out according to Figure 1.

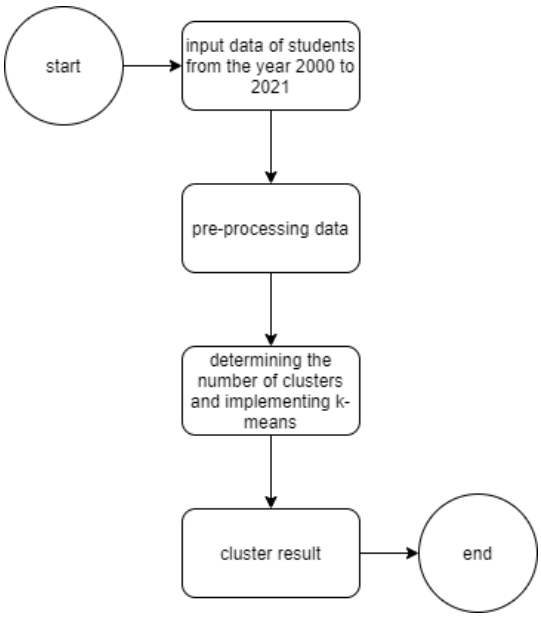

Figure 1 Stages Of Data Processing

A. Input Data

The data collection was obtained from the academic department in the form of SQL format data containing 66,211 academic grade data for students between 2000 and 2021. The data contains the final grade attribute of the course with float data type and the name of the course with the string data type. The student data is then entered into a MySQL database for data pre-processing.

B. Data Pre-processing

The student course grade data stored in MySQL is calculated for each course's average using Structured Query Language (SQL) using Equation (3). Where the m_i variable is the average of the course, the $\sum x_j$ variable is the total of the student course grades and the n_i variable is the total number of students who took the course.

$$m_i = \frac{\sum x_j}{n_i} \quad (3)$$

By selecting computer science courses and calculating each course's average, we obtained a total of 55 courses, which can be seen in Table I.

DOI : http://dx.doi.org/10.25139/inform.v8i1.4519

TABLE I
AVERAGE RESULTS OF COURSE GRADES

Name of Course	Average Grade
Algoritma dan Pemrograman I	2,82
Algoritma dan Pemrograman II	2,84
Algoritma dan Pemrograman III	2,76
Analisis dan Perancangan Objek	3,11
Animasi	3,31
Arsitektur Komputer	3,12
Audit Sistem Informasi	3,07
Basis Data 1	2,76
Basis Data 2	3,01
Data Mining dan Perolehan Informasi	3,31
Desain Web	2,67
Forensik Digital	3,33
Game Enterprise	3,35
Game Komputer	3,41
Interaksi Manusia dan Komputer	2,62
Jaringan Komputer	3,17
Kapita Selekta	3,27
Kecerdasan Buatan	2,72
Komputasi Awan	3,04
Komputer dan Masyarakat	2,92
Komunikasi Data	3,26
Logika Pemrograman	3,01
Manajemen Proyek Sistem Informasi	3,25
Matematika Diskrit	3,02
Matematika Komputer 1	2,85
Matematika Komputer 2	3,13
Metode Numerik	2,67
Metodologi Penelitian	3,25
Pembelajaran Mesin	3,65
Pemrograman Game	3,25
Pemrograman Game Lanjut	3,20
Pemrograman IoT	3,70
Pemrograman Jaringan	3,16
Pemrograman Web I	2,99
Pemrograman Web II	2,47
Pengantar Teknologi Informasi	3,27
Perancangan Game	3,31
Perancangan Web	3,25
Praktikum Algoritma dan Pemrograman I	3,07
Praktikum Algoritma dan Pemrograman II	3,08
Praktikum Algoritma dan Pemrograman III	2,62
Praktek Basis Data 1	3,74
Praktek Kerja Lapangan	2,51
Praktikum Teknologi Informasi dan Komunikasi	2,33
Rekayasa Perangkat Lunak	2,82
Rekayasa Proses Bisnis	3,23
Sistem Informasi Geografis	3,15
Sistem Informasi Manajemen	3,21
Sistem Keamanan	3,49
Sistem Operasi 1	3,46
Sistem Operasi 2	3,43
Statistika	3,04
Struktur Data	2,97
Teknologi Informasi dan Komunikasi	3,35
Teori Bahasa dan Automata	3,58

C. Determining the Number of Clusters and Implementing K-Means

The average grades data is then uploaded to Google Colaboratory, and the optimal number of clusters is determined using the elbow method elaborated in Code I.

CODE I
IMPLEMENTATION OF ELBOW METHOD ON SCIKIT-LEARN

```
distortions = []
K = range(1,10)
  for k in K:
  kmean_model = KMeans(n_clusters=k)
  kmean_model.fit(data)
  distortions.append(kmean_model.inertia_)

plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('Optimal Klaster Metode Elbow’)
plt.show()
```

The initial number of clusters entered is 10. The k-means calculation iteration is performed using the data in Table 1 stored in the distortion (SSE) array to be displayed on the graph. In the K-Means function, the init parameter is filled with 'k_means ++' for the best method of initializing the initial cluster centers to accelerate convergence. Then the n_cluster is filled with the value obtained from the elbow method or SSE calculation. Then random_state is given the value 0 so that the resulting clustering results remain unchanged and do not change. The code can be seen in Code II.

CODE II
APPLICATION OF K-MEANS IN SCIKIT-LEARN

```
kmeans_model= Kmeans
(init='k means++',n_clusters=N_CLUSTERS,
random_state=0)
kmeans_model.fit(data)
kmeans_predict = kmeans_model.predict(data)
print(kmeans_model.cluster_centers_)

for cluster in range(N_CLUSTERS):
   print('klaster: ', cluster)
   print(labels[np.where(kmeans_predict ==
cluster)])
```

After running k-means clustering, the result is an array of predicted cluster assignments for the data. This array is then compared with the "labels" variable, which contains the true class labels for the data. The program matches the cluster assignments with the correct class labels to determine which

DOI : http://dx.doi.org/10.25139/inform.v8i1.4519

data points belong to which cluster. This information can be used to analyze further and understand the formed clusters.

IV. Result and Discussion

The results obtained from the calculation of SSE or the elbow method can be seen in Table II.

TABLE II
RESULT OF ELBOW METHOD CALCULATION

K Value	Distortion (SSE)	Distortion Difference
K = 1	6.877073333333335	0
K = 2	**2.367434920634921**	**4.5096384127**
K = 3	1.083633115079365	1.28380180556
K = 4	0.6648532828282829	0.41877983225
K = 5	0.35395357142857137	0.3108997114
K = 6	0.25413842780026974	0.09981514362
K = 7	0.18396319444444437	0.07017523335
K = 8	0.13625730769230768	0.04770588675
K = 9	0.10282333333333335	0.03343397435

The data in Table II suggest that using 2 clusters results in the largest reduction in distortion (as measured by the difference in SSE between the 2 clusters and the previous cluster). This can be visualized more clearly by plotting the data in Table II using a tool like matplotlib, as shown in Figure 2.

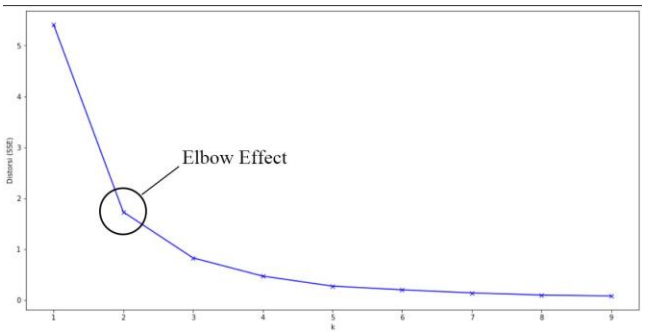

Figure 2. Elbow Graph

Figure 2 is a line graph used to determine the optimal number of clusters for k-means using the elbow method. The horizontal x-axis represents the number of clusters, and the vertical y-axis represents the SSE (sum of squared errors) value. The elbow effect pattern is found at point k=2, considered the optimal point as it has the smallest angle and relatively low SSE value. The number of clusters is determined using the elbow method, scikit-learn, and matplotlib. After obtaining the optimal number of clusters, k-means clustering is performed on the data in Table I, resulting in the cluster assignments seen in Table III.

TABLE III
RESULT OF COURSE CLUSTERING

Cluster 1	Cluster 2
Algoritma dan Pemrograman I	*Analisis dan Perancangan Berbasis Objek*
Algoritma dan Pemrograman II	*Animasi*
Algoritma dan Pemrograman III	*Arsitektur Komputer*
Basis Data I	*Audit Sistem Informasi*
Desain Web	*Basis Data II*
Interaksi Manusia dan Komputer	*Data Mining dan Perolehan Informasi*
Kecerdasan Buatan	*Forensik Digital*
Komputer dan Masyarakat	*Game Enterprise*
Matematika Komputer	*Game Komputer*
Metode Numerik	*Jaringan Komputer*
Pemrograman Web II	*Kapita Selekta*
Praktek Kerja Lapangan	*Komputasi Awan*
Praktikum Algoritma dan Pemrograman III	*Komunikasi Data*
Praktikum Teknologi Informasi dan Komunikasi	*Logika Pemrograman*
Rekayasa Perangkat Lunak	*Manajemen Proyek Sistem Informasi*
	Matematika Diskrit
	Matematika Komputer II
	Metodologi Penelitian
	Pembejalaran Mesin
	Pemrograman Game
	Pemrograman Game Lanjut
	Pemrograman IoT
	Pemrograman Jaringan
	Pemrograman Web
	Pengantar Teknologi Informasi
	Perancangan Game
	Perancangan Web
	Praktek Basis Data I
	Praktikum Algoritma dan Pemrograman I
	Praktikum Algoritma dan Pemrograman II
	Rekayasa Proses Bisnis
	Sistem Informasi Geografis
	Sistem Informasi Manajemen
	Sistem Keamanan
	Sistem Operasi
	Sistem Operasi II
	Statistik
	Struktur Data
	Teknologi Informasi dan Komunikasi
	Teori Bahasa dan Automata

The results of the k-means algorithm with 2 clusters show that cluster 1 has a total of 15 members and cluster 2 has a larger number of members with a total of 40. Based on the centroid value of each cluster, the members of cluster 1 have a similarity in value and distance to the cluster 1 centroid (2.686), and cluster 2 has a similarity in value and distance to the cluster 2 centroid (3.245).

Although the optimal cluster value produced by the elbow method shows the number 2, it turns out that if the clustering is repeated with a cluster number of 3, the results are still relevant with the cluster 1 centroid at the value of 2.686 and a shift in the centroid of cluster 2 with a value of 3.175. The centroid of cluster 3 with a value of 3.57, and the number of members in each cluster is 7, 15, and 33.

V. Conclusion

The elbow method to determine the optimal number of clusters and k-means as the clustering algorithm shows that courses with high average values are in cluster 2, including animation, computer architecture, game programming, and so

DOI : http://dx.doi.org/10.25139/inform.v8i1.4519

on. The result of this study is expected to be an evaluation for the academic division or management, especially the Computer Engineering program, regarding the low value of course grades and the potential for further research to cluster courses not only based on value, but also method and substance of the studied courses, such as seeing the trend of grades or grade patterns in courses with low average values and making special efforts to improve students' understanding of these courses.

REFERENCES

[1] Lester H. Hunt, *Grade Inflation: Academic Standards in Higher Education*. SUNY Press, 2008.

[2] V. N. Budiyasari, P. Studi, T. Informatika, F. Teknik, U. Nusantara, and P. Kediri, "Implementasi Data Mining Pada Penjualan kacamata Dengan Menggunakan Algoritma Apriori," *Indones. J. Comput. Inf. Technol.*, vol. 2, no. 2, pp. 31–39, 2017.

[3] R. S. J. Baker, "Data Mining for Education," *Encycl. Data Warehous. Min.*, 2011, doi: 10.4018/978-1-59140-557-3.

[4] J. He, "Advances in data mining: History and future," *3rd Int. Symp. Intell. Inf. Technol. Appl. IITA 2009*, vol. 1, pp. 634–636, 2009, doi: 10.1109/IITA.2009.204.

[5] F. Coenen, "Data mining: Past, present and future," *Knowl. Eng. Rev.*, vol. 26, no. 1, pp. 25–29, 2011, doi: 10.1017/S0269888910000378.

[6] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[7] M. Z. Rodriguez *et al.*, *Clustering algorithms: A comparative approach*, vol. 14, no. 1. 2019. doi: 10.1371/journal.pone.0210236.

[8] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[9] J. J. Manoharan, S. H. Ganesh, M. L. P. Felciah, and A. K. S. Banu, "Discovering students' academic performance based on GPA using K-means clustering algorithm," *Proc. - 2014 World Congr. Comput. Commun. Technol. WCCCT 2014*, pp. 200–202, 2014, doi: 10.1109/WCCCT.2014.75.

[10] I. Arwani, "Integrasi Algoritma K-Means Dengan Bahasa SQL Untuk Klasterisasi IPK Mahasiswa (Studi Kasus: Fakultas Ilmu Komputer Universitas Brawijaya)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, p. 143, 2015, doi: 10.25126/jtiik.201522148.

[11] Y. E. Fadrial, "Klasterisasi Hasil Evaluasi Akademik Menggunakan Metode K-Means (Studi Kasus Fakultas Ilmu Komputer UNILAK)," *Semin. Nas. Teknol. Inf. Ilmu Komput.*, vol. 1, no. 1, pp. 53–65, 2020.

[12] A. E. Rahayu, K. Hikmah, N. Yustia, and A. C. Fauzan, "Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 1, no. 2, pp. 82–86, 2019, doi: 10.28926/ilkomnika.v1i2.23.

[13] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 975–8887, 2014.

[14] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.

[15] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 533–538, 2018, doi: 10.1109/ISEMANTIC.2018.8549751.

[16] H. Islam and M. Haque, "An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 8, pp. 146–149, 2012, doi: 10.14569/ijacsa.2012.030824.

[17] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015, doi: 10.1007/s40745-015-0040-1.

[18] T. Kanungo, D. M. Mount., N. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," *Am. J. Clin. Nutr.*, vol. 56, no. 2, pp. 385–393, 1992, doi: 10.1093/ajcn/56.2.385.

[19] D. Arthur and S. Vassilvitskii, "How Slow is the k-Means method ? General Terms :," *Construction*, pp. 144–153, 2006.

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Environ. Health Perspect.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[21] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," pp. 1–15, 2013, [Online]. Available: http://arxiv.org/abs/1309.0238

DOI : http://dx.doi.org/10.25139/inform.v8i1.4519