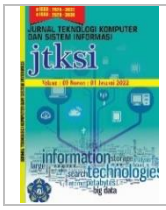

JTKSI (Jurnal Teknologi Komputer dan Sistem Informasi)

JTKSI, Volume 5, Nomor 1, Januari 2022
E ISSN: 2620-3030; P ISSN: 2620-3022, pp.61-66
http://ojs.stmikpringsewu.ac.id/index.php/jtksi

Received: 4 Agustus 2021; Revised: 14 Desember 2021; Accepted: 29 Desember 2021

Drop Out Student Clusterization Using the k-Medoids Algorithm

Mohammad Guntara[1], Totok Suprawoto[2]
[1]Informatics Department, Universitas Teknologi Digital Indonesia (UTDI), Yogyakarta, Indonesia
[2]Information System Department, Universitas Teknologi Digital Indonesia (UTDI), Yogyakarta, Indonesia
[1,2]Jl. Raya Janti Karang Jambe No. 143, Yogyakarta, Indonesia
E-mail : guntara@utdi.ac.id, totok@utdi.ac.id
*Corresponding author E-mail: guntara@utdi.ac.id

Abstract

Student dropout (resign) is a problem that needs to be addressed as early as possible. The number of students dropping out will decrease the quality of the performance of a university, as well as reduce it as much as possible because it will have an impact on the public appreciation. As a first step to reducing it, it requires the clustering of students who experience this. Based on this cluster, a pattern of student tendency to drop out can be identified. The parameters used in this study were the GPA, the study period, the number of credits received, and the number of semesters inactive. To compile a cluster, the k-Medoids algorithm is used with 3 types of clusters. Based on the results of the clustering, it can be seen that the dominance of dropout students is due to GPA <2.00 as much as 38.2% and due to not being active in college as much as 52.2%. To measure the cluster quality, the Silhouette coefficient algorithm is used and the resulting coefficient value is 0.3, meaning that the cluster separation rate weak structure.

Keywords: drop out, student, k-Medoids, GPA, Silhouette

I. INTRADUCTION

Drop out (withdrawal) is a situation where a student is declared not to be registered as a student [1]. The resignation process was carried out after going through various comprehensive evaluations. Students resign at STMIK AKAKOM around 30% of students enrolled in a batch (SK Chairman of STMIK AKAKOM on Student Resignation, 2015-2019). This situation needs to be observed considering the high number of students who resigned.

To take strategic steps so that student resignations can be reduced, analysis is needed by making clustering based on data from students that have been accumulated over several periods. This clustering is needed to identify the tendency of students to resign whether it is due to the grade point factor, the number of credits taken (study period), or other factors that do not meet the requirements. The clustering process of student withdrawals was carried out using the k-Medoids method. The results of this clustering are then analyzed and used for decision-making to prevent student resignation.

II. LITERATURE REVIEW

2.1. Teory Drop Out

Drop out is "to not do something that you were going to do, or to stop doing something before you have finished" (https://dictionary.cambridge.org, 2020), or quit the association (http://dictionary.belajarbahasa.id, 2020). In its implementation, the term "resigned" is used (SK Chairman of STMIK AKAKOM, 2015-2019). Students can resign based on [1] article 17 paragraph 2c, which states that the student is considered to have resigned if:

1. Not active in the first semester.
2. Have a GPA of 0.00 in the first year.
3. Inactive for 4 consecutive semesters.
4. Passing the predetermined study period limit.
5. No level progression 2 times in a row for Diploma Three Program students.
6. Involved in per criminal act declared by the High School and/or the authorities.

Students are declared to resign officially when a decree has been issued by the head of the high school.

1) The k-Medoids Method

The K-Medoids method is part of partitioning clustering. K-method. Medoids are quite efficient in small datasets. The initial step of K-Medoids is to find the most representative points (medoids) in the dataset by calculating the distance from the group in all possible combinations of medoids so that the distance between points in a cluster is small while the distance between points in clusters is large[2].

The k-Medoids method or often referred to as the PAM (Partitioning Around k-Medoids) algorithm was developed by Leonard Kaufman and Peter J. Rousseeuw, which is an algorithm similar to k-Means because these two algorithms function as partitions that break the data set into groups. The difference between the k-Means algorithm and the k-Medoids algorithm lies in determining the center of the cluster, where the k-Means algorithm uses the mean value of each cluster as the center of the cluster, while the k-Medoids algorithm uses data objects as representatives (k-Medoids). as the center of the cluster[3]. The K-Medoids algorithm applies objects as representatives (medoid) for each cluster. The application of the K-Medoids algorithm takes longer than K-Means because it takes about 2 minutes on Rapidminer, while the K-Means method only takes about 1 second [4]. K-Medoids have better cluster quality as measured using the Silhouette Coefficient than k-Means for the medium structure category[5].

The initial value of an object as the center of the cluster, freely selected. Through an interactive process, representative objects are replaced with non-representative objects, so that the results of the cluster are better. To determine whether a non-representative object (O random) can replace the current representative object (Oj), there are 4 cases as follows for a non-representative object (p), as follows [6].

Case 1: p now has a representative object Oj, if Oj is replaced by O random as a representative object, and p is closer to one representative object Oi, with i <> j, then p is moved to oi.

Case 2: p now has representative object Oj, if Oj is replaced by O random as representative object, and p is closer to O random, with i <> j, then p is moved to O random.

Case 3: p currently has a representative object Oj, with i <> j, if Oj is replaced by O random as representative object, and p is closer to O random, with i <> j, then p is moved to O O random

Illustration of the center point transition of the k-Medoids algorithm as in Figure 1.

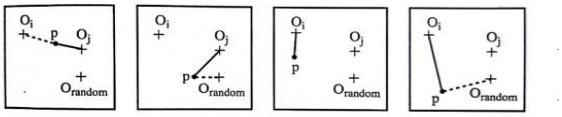

Figure 1. Illustration of the k-Medoids method center point transition

The k-Medois algorithm is as follows[6]

1. Initialize k cluster centers (number of clusters)
2. Allocate each data (object) to the nearest cluster using the Euclidian Distance measurement equation with the equation:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2};i=1,2,3,...,n \quad (1)$$

3. Randomly select objects in each cluster as candidates for the new medoid
4. Randomly select objects in each cluster as candidates for the new medoid.
5. Calculate the distance of each object in each cluster with the new medoid candidate.
6. Calculate the total deviation (S) by calculating the value of the new total distance - the old total distance.

Silouette Coefficient/Index

Silhouette Coefficient is a method used to see the quality and strength of clusters and how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. Cohesion is the degree of closeness or strength between objects in one cluster, while separation is the degree to which one cluster is separated from another[7].

The stages of calculating the Silhouette Coefficient are as follows:

1. Calculate the average distance from a document for example i with all other documents that are in one cluster $(i) = 1\ |A|-1\ j\ ,j\neq i\ (i,j)$ (4) with j is another document in one cluster A and d(i,j) is the distance between document i and j.
2. Calculate the average distance from document i to all documents in other clusters, and take the smallest value. $(i,\ ) = 1\ |A| \in (i,j)$ (5) where d(i,C) is the average.

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters[9]. The following are the criteria for measuring the Silhouette Coefficient value in table I [8].

Table 1. Criteria of Silhouette Coefficient

Silhouette Coefficient (SC)	Criteria
0.7<SC<=1	Strong Structure
0.5 < SC ≤ 0.7	Medium Structure
0.25<SC<=0.5	Weak Structure
<=0,25	No Stucture

2.2. Literature Review

The potential clustering of dropout students was studied by [7] using the C4.5 Algorithm which resulted in 9 (nine) rules and the accuracy level generated by this method was 96.15% and the AUC (Area Under the ROC Curve) value was 0.998

According to [8] predicting student dropout using outlier groups using the k-Means or k-Medoids algorithm because there is not enough data. To

overcome this, it is necessary to apply the dendrogram method to our datasets for hierarchical clustering.

Research of [10] compiled a study on the optimization of the k-Medoids method for clustering student applicants for scholarships and formed three groupings for scholarship application data: recipients, considered, and not receiving. Based on the results of the evaluation by calculating the value of the Cubic Clustering Criterion, it was found that the data set with the overall codification of the data occupied the best predicate in grouping uniformity with a value of 2.245.

Whereas [11] made a study by grouping students who had the potential to drop out using the K-Means Clustering Method. The results of the cluster are: class 2014 is in cluster 0 totaling 4 students or 30.77% of 13 samples, class 2015 is in cluster 1 totaling 4 students and cluster 2 is 2 students or 66.7% from 9 samples, class 2016 in cluster0 there are 2 students and cluster 1 totaling 10 students or 50% of 24 samples, and class 2017 in cluster 2 totaling 4 students or 22.22% of 18 samples and class 2017 there are 4 students who have the potential to drop out.

The other researcher like [12] discussed clustering students prone to dropout with the k-Means method. There are 3 clusters where each cluster has a percent. Like cluster 0 has 15 data (38%), cluster 1 has 12 data (30%), and cluster 2 has 13 data (33%). Quality of clustering used for Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN .The results showed that K-Means had the best level of validity than K-Medoids and DBSCAN, where the Davies-Bouldin Index yield was 0.33009058, and the Silhouette Index yield was 0.912671056 [4].

The results of the silhouette test in this study reached 85%, which indicates that the clustering method to determine the characteristics of students who are likely to graduate or drop out (DO) in the management department of the National University, Jakarta, can be applied as an effort to overcome the high dropout rate [13].

III. METHODOLOGY

The research method used is

1. Preparation and data processing
   a. Data preparation
      Drop out student data is taken from the list of students who have resigned based on the Decree of the Head of STMIK AKAKOM. From the data source then normalized. Normalization is used to neutralize the scale of data attributes into a specific range[[14]. Data normalization is done by selecting the attributes used in the clustering process. The attributes used are: NIM, GPA, Number of SKS, batch, number of semesters inactive. Data is collected from different files so that it is the attributes are not uniform, then the sequence of attributes is standardized. Valid data from each file are combined into 1 file that is ready to be processed.
   b. Preparation and implementation of the clustering process
      Preparation and implementation that are: running the tools (Rapid Miner) the clustering, reading files containing DO data to the system, select the column or attribute used for clustering, determine the column format or attribute and role, executes the application program and exports cluster results to Excel format, and Analyze the results of clustering
   c. Clusters quality test using Silhouette Coefficient

2. The tools and materials used
   Tools used in this research is Rapid Miner Software [15]. This tool is used to process data and produce labels, namely clusters for each object (dropout students). Excel software, used to calculate the frequency of each parameter in a cluster. The material processed by Rapid Miner is DO data in row-column form and packaged in a spreadsheet format

IV. RESULT AND DISCUSSION

4.1. Result

The data source is in the form of an Excel file which is data on dropout students from 2015 to 2019 as follows.

Table 2. Initial Source Data

No	Program Studi	Jenjang	NIM	Nama Mahasiswa	Jml NA (sem)	SKS	IPK
1	TEKNIK KOMPUTER	DIPLOMA 3	113310030	GILANG SURYA SUMIRAT	3	25	0,48
2	TEKNIK KOMPUTER	DIPLOMA 3	113310033	WISNU PRAYOGA PANGESTU	4	47	0,72
3	TEKNIK KOMPUTER	DIPLOMA 3	123310004	ENDRI PUJI PRABOWO	4	70	1,4
4	TEKNIK KOMPUTER	DIPLOMA 3	123310017	ABDI ANUGRAH PRASETYA	3	13	1,54
5	TEKNIK KOMPUTER	DIPLOMA 3	133310018	DANIEL OBET WANMA	4	21	0,33
.....	.....	....	....	....	....	..	..
955	TEKNIK INFORMATIKA	STRATA 1	155410126	VENI SUCIANINGSIH	4	22	2,64
956	TEKNIK INFORMATIKA	STRATA 1	155410132	RIDHO BAKHRUL RAIS	5	72	3,39
957	TEKNIK INFORMATIKA	STRATA 1	155410137	RIDHO FAHMI ABDULLAH	4	22	3,82
958	TEKNIK INFORMATIKA	STRATA 1	155410183	RAJIF SETYAWAN	4	42	1,38
959	TEKNIK INFORMATIKA	STRATA 1	165410093	KELVIN ELFIAN	1	15	0
960	TEKNIK INFORMATIKA	STRATA 1	165410187	RIYAN HIDAYAT	1	22	0

Based on the data source above, it is filtered, formatted, and validated into data that is ready to be processed as follows.

1) Data Input

Table 3. Data from the filtering, formatting, and validation processes

id	nim	jumna	sks	ipk	ang
1	113310030	3	25	0,48	11
2	113310033	4	47	0,72	11
3	123310004	4	70	1,4	12
4	123310017	3	13	1,54	12
5	133310018	4	21	0,33	13
....	.....	.....	....	...	...
950	155410126	4	22	2,64	15
951	155410132	5	72	3,39	15
952	155410137	4	22	3,82	15
953	155410183	4	42	1,38	15
954	165410093	1	15	0,00	16
955	165410187	1	22	0,00	16

Processing Result

the results of clustering in tabulated form

Table 4. Clustering Results using k-Medois

1	113310030	3	25	0,48	11	cluster_1
2	113310033	4	47	0,72	11	cluster_1
3	123310004	4	70	1,40	12	cluster_0
4	123310017	3	13	1,54	12	cluster_1
5	133310018	4	21	0,33	13	cluster_1
.....	....	....	....	...	...	...
951	155410132	5	72	3,39	15	cluster_0
952	155410137	4	22	3,82	15	cluster_1
953	155410183	4	42	1,38	15	cluster_1
954	165410093	1	15	0,00	16	cluster_1
955	165410187	1	22	0,00	16	cluster_1

Cluster Model

```
Cluster 0: 197 items
Cluster 1: 429 items
Cluster 2: 329 items
Total number of items: 955
```

Figure 2. Number of Objects per Cluster

Centroid value per cluster

Table 5. Centroid Value per Cluster

Attribute	cluster_0	cluster_1	cluster_2
jumna	5	3	1
sks	72	22	127
ipk	3.390	0	2.370
ang	4	14	2

Frequency distribution of each parameter per cluster

Table 6. Frequency Distribution of each Parameter per Cluster

	IPK			Jumlah SKS			
Cluster	<2.00	<3.00	<=4.00	<90	<120	<144	>=144
0	148	43	6	164	33	0	0
1	367	48	14	429	0	0	0
2	112	172	45	0	94	0	75

Table 7. Frequency Distribution of each Parameter per Cluster (cont..)

	Tidak aktif (semester)				Angkatan		
Cluster	1	2	3	>=4	<2010	<2015	<2020
0	7	5	18	167	34	140	23
1	17	99	40	271	2	230	197
2	7	24	17	280	143	172	14

4.2. Discussion

Based on data sources that have gone through the filtering, formatting, and validation processes with the number of objects 995 students drop out with parameters: Student Identification Number (NIM), number of inactive semesters (number), number of credits obtained (credits), Cumulative Achievement Index (IPK), and the year batch (ANGKATAN), with the number of clusters: 3 clusters, it was found that the largest cluster was cluster_1 (44.9%), followed by cluster_2 (34.4%), and then cluster_0 (20.6%).

Based on the frequency distribution of the parameters for each cluster, it is known that in cluster_0 the most dominant parameter is the number of credits <90 credits (17.1%) and not active > = 4-semester credits (17.4%). For cluster_1, it was dominated by dropout students with a GPA <2.00 (38.4%) and total credits <90 credits (52.2%), while cluster_2 parameters with the greatest frequency were inactive students 4 semesters and above (29.3%).

V. CONCLUSIONS

Based on the previous discussion, it was concluded that the clusters have been successfully created using the k-Medoids algorithm for each dropout student consisting of 3 clusters: cluster_0, cluster_1, and cluster_2.

The frequency of the parameters in cluster_1 it is known that the biggest cause of students dropping out is due to limited academic ability, seen from the GPA <2.00 reaching 38.4% and activity as a student is lacking or taking several semesters off, seen from the number of credits taken <90 credits reaching 52.2%. While cluster_0, the difference in frequency is not as extreme as cluster_1 with the parameter with the highest number of credits <90 credits (17.1%). The cluster_2 is like cluster_0, the dominance of inactive 4 semesters or more (29.3%). Based on calculations using the Silhouette Coefficient algorithm, the best quality cluster, which has separation is if the anchor cluster (a) is cluster_2 where the Silhouette Coefficient value is 0.3 include in category, weak structure.

REFERENCES

[1] STMIK_AKAKOM, "Peraturan akademik stmik akakom 2019," 2019.

[2] I. I. P. Damanik, S. Solikhun, I. S. Saragih, I. Parlina, D. Suhendro, and A. Wanto, "Algoritma K-Medoids untuk Mengelompokkan Desa yang Memiliki Fasilitas Sekolah di Indonesia," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 520, 2019, doi: 10.30645/senaris.v1i0.58.

[3] B. Riyanto, "Penerapan Algoritma K-Medoids Clustering Untuk Pengelompokkan Penyebaran Diare Di Kota Medan (Studi Kasus: Kantor Dinas Kesehatan Kota Medan)," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 562–568, 2019, doi: 10.30865/komik.v3i1.1659.

[4] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 11, no. 1, p. 32, 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[5] S. Defiyanti, M. Jajuli, and N. Rohmawati, "K-Medoid Algorithm in Clustering Student Scholarship Applicants," *Sci. J. Informatics*, vol. 4, no. 1, pp. 27–33, 2017, doi: 10.15294/sji.v4i1.8212.

[6] M. Han, Jian and Kamber, *Data Mining*. San Francisco, USA: Han, Jian & Kamber, and Michelin, 2006.

[7] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index," *J. Gaussian*, vol. 8, no. 2, pp. 161–170, 2019.

[8] R. A. Farissa, R. Mayasari, and Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient," vol. 5, no. 2, pp. 109–116, 2021.

[9] S. Kumar, *Silhouette Method — Better than*

Elbow Method to find Optimal Clusters, 2020th ed. https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891.

[10] S. Defiyanti, M. Jajuli, and N. Rohmawati, "Optimalisasi K-MEDOID dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan CUBIC CLUSTERING CRITERION," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 211–218, 2017, doi: 10.25077/teknosi.v3i1.2017.211-218.

[11] I. Vhallah, S. Sumijan, and J. Santony, "Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Metode Clustering K-Means," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 572–577, 2018, doi: 10.29207/resti.v2i2.308.

[12] M. Muhamad, A. P. Windarto, and S. Suhada, "Penerapan Algoritma C4.5 Pada Klasifikasi Potensi Siswa Drop Out," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 1–8, 2019, doi: 10.30865/komik.v3i1.1688.

[13] M. Darwis, L. H. Hasibuan, M. Firmansyah, and N. Ahady, "Implementation of K-Means C lustering A lgorithm in M apping the G roups of G raduated or D ropped-out S tudents in the Management Department of the National University," vol. 04, no. 01, pp. 1–9.

[14] S. Asmiatun, "Penerapan Metode K-Medoids Untuk Pengelompokkan Kondisi Jalan Di Kota Semarang," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 2, pp. 171–180, 2019, doi: 10.35957/jatisi.v6i2.193.

[15] M. Sari, I. R. Munthe, and I. Irmayani, "Metode Clustering K-Medoids untuk Aplikasi Pembelajaran di Masa Pandemi COVID-19," *MEANS (Media Inf. Anal. dan Sist.*, vol. 6, no. 1, pp. 101–105, 2021, doi: 10.54367/means.v6i1.1255.