Recommendation of Student Admission Priorities Using K-Means Clustering

Hidayatul Muttaqien[1], Muhammad Lutfi[2], Musliadi KH[3], Abdul Muis[4], Hazriani Zainuddin[5]
[1,2,3,4,5]Departement of Computer System, Sekolah Tinggi Manajemen Informatika dan Komputer Handayani, Indonesia, 90321

{hidayatul@student.handayani.ac.id[1], saintechlutfi@handayani.ac.id[2], musliadi.esqway165@yahoo.co.id[3], abdulmuis.160674@gmail.com[4]}

Abstract. This study aims to investigate student's characteristics based on three variables, namely grade point average (GPA), period of study, and administrative obedience in order to draw a recommendation for student admission priorities at Mulawarman University. This recommendation will be used as one of reference variable on new student recruitment. The 8.741 records of student data sourced from the university data warehouse were mined using K-Means clustering. This mining process produced three clusters, cluster-1 includes 1,758 students with centroid {0.158,0.694,0.663}, while cluster-2 embraces 4,928 students with centroid {0.970,0.700,0.675}, and cluster-3 with centroid {0.953,0554,0.386} covers 2.055 students. This result shows that cluster-2 has the best combination of centroid values, implied that new students from schools where students in cluster-2 graduated from are recommended as the high priority students to be admitted at Mulawarman University

Keywords: *Recommendation, K-Means, Clustering, Priorities, Centroid.*

1 Introduction

Mulawarman University (also called as UNMUL) is a University at Borneo-Indonesia consists of 14 (fourteen) faculty, 95 (ninety five) study program, reach to 35.000 (thirty five thousand) students and has grade „A" of institution accreditation. UNMUL implemented an integrated information system for all activities to serve students in academy proccess as well as to ease management in managing the university's roles. All data and transactions are recorded using database enterprise system. These long-time recorded data grows and become a university big data. This big data can be utilized to obtain important information or for another proccesses to support management in maintaining quality of the university.

One important factor in maintaining quality of the university is to admit best student candidate on the selection proccess. This proccess might affect on good quality student and alumnus of the university as well. We believe that good student characteristhics are influenced by some factors, including the schools where they are graduated from and the students performance themselves. Therefore, we conducted a research by mining the university big data,

ICOST 2019, May 02-03, Makassar, Indonesia

DOI 10.4108/eai.2-5-2019.2284614

especially students data in order to formulate student clusters based on three key characteristics, grade point average (GPA), period of study and administrative obedience using Algorithm K-Means, in order to draw a recommendation for student admission priorities at Mulawarman University.

Clustering has been widely applied in various cases in data mining. Clustering is a process of extracting previously unknown, valid, positional userful and hidden pattern form large data sets[1]. There are several previous researches on clustering university data using K-Means have been conducted so far. For instance, in [2] and [3] clustering process using K-Means base on GPA has been implemented as an indicator of student performance, in order to provide information that is used to make effective decisions at university. Another research on clustering process based on GPA, hometown and Field of Study is used to make marketing decision [4]. However, mutual application of GPA, period of study, and administrative obedience on deciding new students admission priorities haven't investigated yet. Therefore, this research is conducted to examine this related issue to obtain more comprehensive solution in determining new student admmission policy.

In this paper we will test existing data at data sourced from UNMUL warehouse. There are three main stages conducted to find the result. *First*, data collection phase bu collecting representative of student charactheristics consist of identity number, name, schools, study program, period of study, GPA, and academic obedience status. Student's period of study influences the institution acceditation, shorter periode of study can improve the point [5], [6]. Grade point average (GPA) is a measure of student performance at university[1],[5],[8],[9] and administrative obedience is needed to support academic activities, because the availability of funds is greatly affect in academic process. *Second*, data cleaning by clearing data from noice, because in *data mining* there is a term *„Garbage In, Gabage Out"*, means if wrong input then we get wrong information. *Third*, the clean data should be processing use Algorithm K-Means to clustering students characterisctics. And finally, the result will show list of priority schools to be admitted at UNMUL.

2 Clustering Student Characteristics Using K-Means

2.1 Data Clustering

Data clustering is unsupervised and statistical data analysis technique . It is used to clasify the same data into heteregonous group. It is used to operate on large data-set to discover hidden information and relationship to make effective decision. Cluster analysis is used to devide a large data into subsets called clsuters. The cluster is a group of data objects that are similar to one other are placed within the sama cluster but are dissimilar to object in other clusters. Typcal pattern clustering activity involves the following steps is shown below[10] :

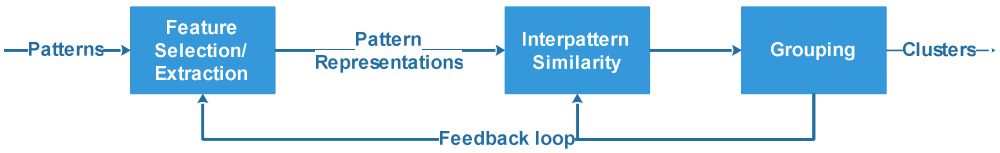

Fig. 1. Stages in clustering.

Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. *Feature selection* is the process of identifying the most effective subset of the original features to use in clustering. *Feature extractio*n is the use of one or more transformations of the input features to produce new salient features. *Pattern proximity* is usually measured by a distance function defined on pairs of patterns. The *grouping* step can be performed based on minimum distance.

2.2 K-means

In this paper, we used K-Means to clustering students based on GPA, Period of Study and Adminsitrative Obedience. The result of clusters is used to make the priorities schools. Algoritm K-Means is an algorithm for clustering processes that group data based on the nearest to centroid of data. First we calculate centroid *cj*. In the much case, centroid is generate randomly depend on count of cluster. After *k* (centroid) are formed then calculate distance every data *xi* with *centroid* to-*j until k* is notated *d(xi,cj)*. We used *Euclidean* to calculate distance between instance data, show in equation (1). The distance between instance and centroid is show the similarity between them, the smaller distance is mean the instance is very close to centroid [11]. Futhermore is calculate average of distance all of data to centroid, this value used to update new centroid, show in equation (2)[12]. This process will loop until the centroid do not change or the centroid is constant or the process will be stop until reach maximal value of iteration.

$$d(Xi, Cj) = \sqrt{\sum_{i=1}^{N}(Xi - Cj)2} \quad (1)$$

$$C_{new} = \frac{\sum_{i}^{N} Dxi}{\mathrm{N}} \quad (2)$$

Where is :

- C_{new} is new centroid
- Dxi is Distance of instance between centroid
- N is number of instance

The flowchart of K-Means algorithm that means how the K-Means work-out is shown below.

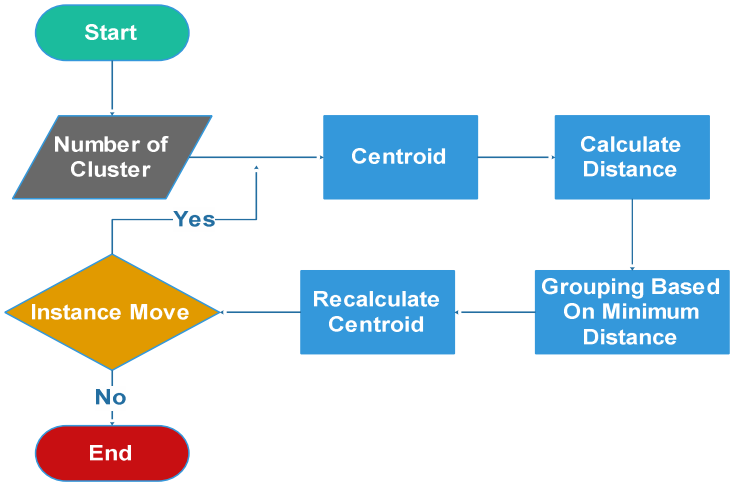

Fig. 2. Flowchart of K-Means

2.3 Pre-Processing

Preprocessing is very important process berfore we do clustering[13], [14]. It is used to cleaning of data for decreas noise of object data. That matter is used for input data it is clean form noise information that causes an miss information in the result. In this paper we used Min-Max method [15] for linier transformation of data using formula is shown in equation (3). Table 1 shows 10 samples among 8741 instances of student data with attribut GPA, administrative obedience (Ontime), and Period of Study before pre-processing phase, while processed student data is presented in Table 2.

$$A' = (\frac{A - \min value\ of\ A}{\max value\ of\ A - \min value\ of\ A} * (D - C) + C \quad (3)$$

Where is :
A‘ is result of normalisation
A is data
Min value of A is minimum value from set of data
Max value of A is maximum value from set of data
D is maximum value new data
C is minimum value new data

Table 1. Student Data with attribut GPA, Ontime, Period of Study

Id. Number	Name	Ontime	GPA	Period Of Study
1308015122	A DEBY CLARA SHATY	0.86	3.52	49.97
1305115013	A'AN DWI PURWATI	1.00	3.59	47.42
1305115056	A'YUNI AFIFAH	1.00	3.63	48.97
1402025051	A. ISNAL	0.00	3.46	45.81
1205125137	A. TUTI MARLINA	0.20	3.66	46.84
1309025006	A. WIRDAH FEBY	1.00	3.49	51.52
1405025066	A.RISMAWATI	1.00	3.39	50.65
1202105025	A.STEPHANO DECIDERY	1.00	3.22	58.74
1305085111	AAN NURSIFA HAMIZAN	1.00	3.08	61.00
1304015021	AANG GUNAIFI	0.86	3.86	50.58

Table 2. Student Data with attribut GPA, Ontime, Period of Study after Normalisation

Id. Number	Name	Ontime	GPA	Period Of Study
1308015122	A DEBY CLARA SHATY	0.86	0.75	0.68
1305115013	A'AN DWI PURWATI	1.00	0.79	0.74
1305115056	A'YUNI AFIFAH	1.00	0.81	0.71
1402025051	A. ISNAL	0.00	0.71	0.78
1205125137	A. TUTI MARLINA	0.20	0.83	0.76
1309025006	A. WIRDAH FEBY	1.00	0.73	0.64

1405025066	A.RISMAWATI	1.00	0.67	0.66
1202105025	A.STEPHANO DECIDERY	1.00	0.57	0.47
1305085111	AAN NURSIFA HAMIZAN	1.00	0.49	0.41
1304015021	AANG GUNAIFI	0.86	0.94	0.67

2.4 Implementation of K-Means Algoritma

K-Means is one of the simplest unsupervised learning algorithm used for clustering. based on the previous explanation in Figure 1. then we must set first value of centroid with minimal value of instance *Ontime* is used to *cluster-1*, minimal value of instance *GPA* is used to *cluster-2* and minimal value of instance *Period of Study* is used to *cluster-3* is shown below:

- Centroid for cluster-1: (0.00, 0.88, 0.73)
- Centroid for cluster-2: (1.00, 0.00, 0.34)
- Centroid for cluster-3: (1.00, 0.50, 0.00)

Then we calculate the distance for every instance using equation (1), the sample calculation are presented below:

$$D_{x1}k0 = \sqrt{(0.86 - 0.00)^2 + (0.75 - 0.88)^2 + (0.68 - 0.73)^2}$$

$$D_{x1}k1 = \sqrt{(0.86 - 1.00)^2 + (0.75 - 0.00)^2 + (0.68 - 0.34)^2}$$

$$D_{x1}k2 = \sqrt{(0.86 - 1.00)^2 + (0.75 - 0.50)^2 + (0.68 - 0.00)^2}$$

The results are $D_{x1}k0 = 0.871,\ D_{x1}k1 = 0.835, D_{x1}k2 = 0.738$. From the result, it can be conclude that this instance is belong to *cluster-3*. The same proces recurs (looping) until value of „N“ same as number of data object. After all the object have a group then next step is to calculate the new centroid using eqation (2). The results are shown below:

- Centroid for cluster-1: (0.212,0.702,0.672)
- Centroid for cluster-2: (0.955,0.516,0.657)
- Centroid for cluster-3: (0.975,0.681,0.565)

If the result is different with the previous centroid, then distance calculation is used to determine in which group (cluster) an object will be entered. An object will be entered into a group with minum distance of its centroid value. This process will recurs until the centroid value do not change, or the centroid value is constant, or when the proces reached maximum value of iteration.

3 Result and Discussion

From the result of clustering process using K-Means algorithm, number of instances for each cluster is preserted in Table 3 as well as figured-out in Figure 3.

Table 3. Claster's Instance Values

Cluster	Instances
Cluster 1	1,758
Cluster 2	4,928
Cluster 3	2055

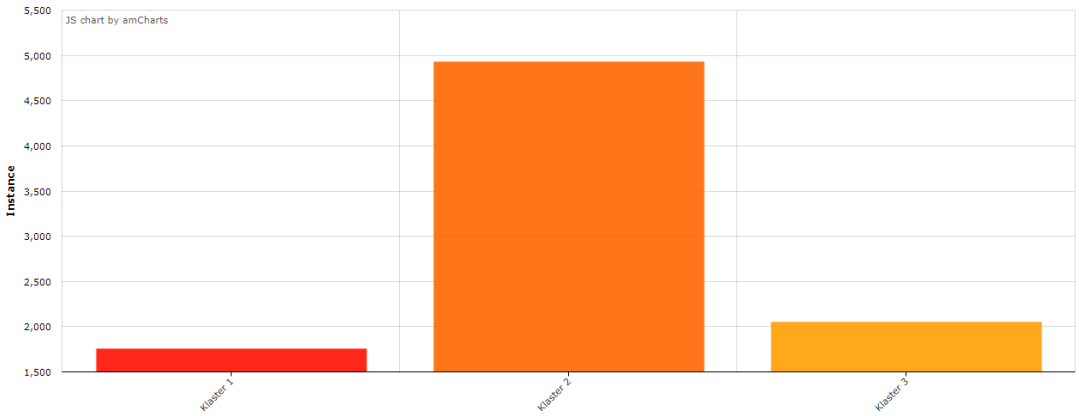

Fig. 3. Clusterization Result.

Centroid for each cluster as in Table 4. The results show that students who are included in Cluster 2 have the best administrative compliance, GPA and Study Period. These results indicate that the origin of school from students can be used as an indicator in new student admission.

Table 4. Centroid for each cluster.

Cluster	Ontime	GPA	Period Of Study
Cluster 1	0.158	0.694	0.663
Cluster 2	0.970	0.700	0.675
Cluster 3	0.953	0.554	0.386

Based on the results of the centroid in the clustering process, it reveals the information that students belong to cluster 2 are students who have administrative compliance, a high GPA, and have a good period of study. Thus, there is a correlation between school of origin and students who are able to commence their study on time in the university. This fact can be obtain by calculating the percentage of students belong to cluster 2 of the overall students who come from these schools as shown in Table 5.

Table 5. List of Prioritized Schools

No.	Name	Ontime	GPA	Period Of Study
1	SMAN 5 BALIKPAPAN	03.57	78.57	17.86

2	SMK YPK TENGGARONG	13.04	78.26	08.70
3	SMAN 1 MUARA JAWA	01.85	77.78	20.37
4	SMAN 6 BALIKPAPAN	01.85	77.78	20.37
5	SMAN 1 SAMARINDA	03.73	77.18	19.09
6	SMAN 1 TANJUNG SELOR	02.38	76.19	21.43
7	SMAN 1 SANGATTA UTARA	12.61	75.68	11.71
8	SMAN 6 SAMARINDA	17.24	74.71	08.05
9	SMAN 2 BALIKPAPAN	02.82	74.65	22.54
10	SMAN 3 BALIKPAPAN	18.52	74.07	07.41
11	SMAN 1 SAMBOJA	04.35	73.91	21.74
12	SMAN 1 BONTANG	14.04	72.81	13.16
13	SMAN 1 BALIKPAPAN	05.88	72.55	21.57
14	SMAN 8 SAMARINDA	06.59	72.53	20.88
15	SMKN 1 BALIKPAPAN	06.90	72.41	20.69
16	SMAN 3 SAMARINDA	03.77	72.17	24.06
17	SMAN 1 TENGGARONG	07.38	71.14	21.48
18	SMKN 1 TANJUNG REDEB	05.88	70.59	23.53
19	SMK (SMEA) N 3 BALIKPAPAN	06.67	70.00	23.33
20	SMAN 1 TANAH GROGOT	01.20	69.88	28.92

Prospective students who come from schools in Table 5 are at the top priority of new students admission process. The selection process which is based on students' report and some other school attributes are still inadequate to describe the actual quality. Each school has its own standard of assessment, allowing students who have higher assessment report than students with lower grade point who come from different schools cannot be used as the determinant factor. By conducting deeper investigation,more relevant information can be used to assist the decision process in accepting prospective students. This information can be used as the main consideration if the same points occur because the origin of the school can be used as the influential factor, as seen in Table 5.

4 Conclusion

Based on the research, it can be concluded that the K-Means algorithm can be applied to classify students based on the attributes of administration compliance (Ontime), GPA and Study Period. The results of the grouping can be used as the information that is useful in the selection process of new students so that it can improve the quality of institutions because it generates students who are complied to the administration, students with good GPA and are able to graduate based on time setting or even faster. Among 8,741 data, student distributions in each cluster are; cluster 1 with 1,758 student; cluster 2 with 4,928 students; and cluster 3 with 2055 students. Centroid for each cluster is cluster 1 {0,158,0,694,0,663}; Cluster 2;{0.970,0,700,0,675 }, and cluster 3 {0.953,0.554,0,386}. From these results it can be concluded that cluster 2 is the best group of students because it has the highest centroid value for these mutual attributes (ontime, GPA and Study Period). Therefore, it can be used as considerations for new student admission priority. In addition, based on this investigation top twenty (Top 20) of prioritized schools are also revealed.

References

[1] T. M. Connolly and C. E. Begg, *Database systems: a practical approach to design, implementation, and management*, 5th ed. Boston ; London: Addison-Wesley, isbn: 978-0-321-52306-8, 2010.

[2] 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, doi: 10.14569/IJACSA.2012.030824, 2012.

[3] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database systems: the complete book*, 2nd ed. Upper Saddle River, N.J: Pearson Prentice Hall,isbn: 978-0-13-187325-4, 2009.

[4] J. O. Ong, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," *J. Ilm. Tek. Ind.*, vol. 12, no. 1, p. 11, 2013.

[5] *Akreditasi Institusi Perguruan Tinggi*. Badan Akreditasi Nasional Perguruan Tinggi, 2011.

[6] R. Refianti, A. B. Mutiara, A. Juarna, and S. N. Ikhsan, "Analysis And Implementation Of Algorithm Clustering Affinity Propagation And K-Means At Data Student Based On Gpa And Duration Of Bachelor-Thesis Completion," . *Vol.*, vol. 35, p. 9, 2005.

[7] J. Oyelade, O. Oladipupo, and I. Obagbuwa, *Application of k Means Clustering algorithm for prediction of Students Academic Performance*, vol. 7. 2010.

[8] J. Jamesmanoharan, S. H. Ganesh, M. L. P. Felciah, and A. K. Shafreenbanu, "Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm," in *2014 World Congress on Computing and Communication Technologies*, Trichirappalli, India,doi: 10.1109/WCCCT.2014.75, 2014, pp. 200–202.

[9] Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, and SVCE indore, "Evaluating Student's Performance using K-Means Clustering," *Int. J. Eng. Res.*, vol. V6, no. 05,doi: 10.17577/IJERTV6IS050070, May 2017.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323,doi: 10.1145/331499.331504, Sep. 1999.

[11] B. S. Duran and P. L. Odell, *Cluster analysis: a survey*. Berlin, New York: Springer-Verlag, 1974.

[12] L. Orloci, "An Agglomerative Method for Classification of Plant Communities," *J. Ecol.*, vol. 55, no. 1, p. 193, isbn: 978-0-387-06954-8 Mar. 1967.

[13] T. Sattarov and Mircea-Serban Pavel, "On the importance of preprocessing and initialization in k-means." Unpublished, doi:10.13140/RG.2.1.3570.6727,2015.

[14] I. B. Mohamad and D. Usman, "Standardization and Its Effects on K-Means Clustering Algorithm," *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, doi: 10.19026/rjaset.6.3638, Sep. 2013.

[15] P. Berkhin, "Survey of Clustering Data Mining Techniques," p. 56.