Jurnal Inovatif : Inovasi Teknologi Informasi dan Informatika, 6 (1) (2023) 92-100
Journal homepage: http://ejournal.uika-bogor.ac.id/index.php/INOVA-TIF/index
E-ISSN : 2654-5519. P-ISSN : 2654-553 | DOI : 10.32832/inovatif
Published by: Universitas Ibn Khaldun

Application Of K-Means Clustering Algorithm Method In New Student Admissions

Tarisno Amijoyo[1*], Mel Siti Nurhaliza[2],

[1,2]Information System, Computer Science, University Of Saintek Muhammadiyah, Jl. Kelapa Dua Wetan No.17, East Jakarta, 13730, Indonesia

E-mail: ahbibadil@gmail.com[*]

Received : December 2022 Accepted : February 2023 Published : April 2023

Abstract

In the admission process of Muhammadiyah Junior High School 6 which is done repeatedly every year and the data increases continuously, thus slowing down the search for information on existing data. This study aims to classify the new student admission data in SMP Muhammadiyah 6 using Clustering techniques. The algorithm used for cluster formation is k-Means algorithm. K-Means is one of the clustering methods that can group the data of new students who have very similar characteristics grouped in the same cluster. The implementation uses RapidMiner which is used to help find more accurate values. The attributes used are school origin and national test scores. To find out the number of clusters and items in the cluster Euclidean distance calculation of the data in the can, calculated from the distance of the first student data to the center of the first cluster is 0,then the distance of the first new student data to the second cluster is 2.2 and the distance of the first new student data to the third cluster is 1.4. The results of the research carried out formed three clusters, with the first cluster totaling 31 items, the second cluster totaling 37 items and the third cluster totaling 32 items. From the cluster that we can be used as one of the basic decision-making for students who are accepted and not accepted at SMP Muhammadiyah 6.

Keywords : K-Means, Clustering, New Students.

Introduction

In today's era has started many schools that use computerized systems in the admission of new students, because with this it can facilitate the process of input and output data quickly and accurately. However, because the process of admission of new students is carried out repeatedly every year and the data will increase continuously, so this will slow down the search for information on the data. Based on the amount of new student data, it is necessary to process the data to find out important information in the form of new knowledge (Knownledge discovery). Data mining is the process of manually finding and identifying an important or interesting pattern that is not yet known from a set of data using certain techniques or methods.[1] one of the methods contained in the data mining used in this study is grouping (Clustering) where the method identifies objects that have similar characteristics as possible.

In the process of grouping new student data for the 2020/2021 school year, the Clustering technique used is to use the K-Means Clustering algorithm, the K-Means Clustering algorithm can group data in the same group and different data in different groups. So that it will be possible to see the data of new students for the 2020/2021 academic year at SMP Muhammadiyah 6 which is not structured to be structured.

Clustering is one of the known techniques in Data mining. In the group will contain data that is as similar as possible and different from the objects in the other groups.[2] there are two types of clustering methods: hierarchical clustering and partitioning. Hierarchical clustering method itself consists of complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering. While in the partition method itself consists of K-means and fuzzy k-means.[3]

The K-Means algorithm is an implementation of data mining clustering. Clustering is data that does not have a class/label so it is often referred to as unsupervised learning techniques. The purpose of clustering is to Group data into several clusters based on the degree of similarity, the degree of similarity is calculated based on the shortest distance between the data to the centroid point.[4] The K-Means algorithm is very easy to implement and has relatively little time and space difficulty. This algorithm is also a fairly efficient algorithm and gives good results if the cluster is compact, hyperspherical in shape and can separate the characteristics of the space well. According to has & Kamber, K-Means algorithm works by dividing data into k clusters that have been determined.[5]

Here is the formula for determining the data distance from each centroid :

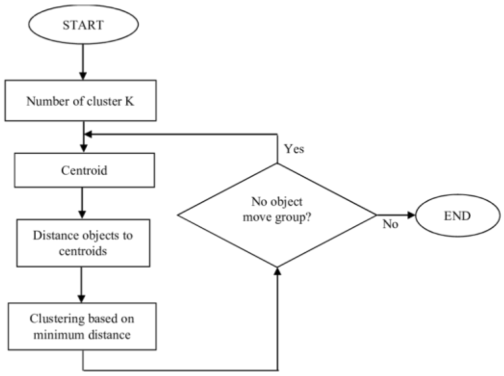
$$d(P,Q) = \sqrt{\sum_{j=1}^{p} \left(x_j(P) - x_j(Q)\right)^2}$$

Description :
D = Document Point
P = Data Record
Q = Data Center
The shortest distance between the centroid and the document determines the position of the cluster of a document. The other iteration formula is defined as follows :

$$C\ (i) = \frac{x1 + x2 + x3 + \cdots..+xn}{\sum x}$$

Description :
X_1 = value of the 1st data record
X_2 = value of the 2nd data record
$\sum x$ = number of data records
The basic calculation of the k-means algorithm is as follows :
First. Determining the number of clusters
Second. Put each document into the most suitable cluster based on proximity to the centroid. The Centroid is the data point at the center of the cluster.
Third. After all the documents go to the cluster. Recalculate the cluster centroid based on the documents that are inside that cluster.
Fourth. If the centroid does not change (with a certain treshold) then stop. If not, go back to step 2.

Method


START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

Analysis of new student admission process in this study using K-means clustering algorithm. The first step is understanding and preprocessing data, clustering data using k-means clustering. The data in this study comes from muhammadiyah School 6 where this data is secondary data consisting of new student data for the 2020/2021 school year. The number of data obtained as many as one hundred consisting of student names, School origin, and national test scores. Here is an example of the 2020/2021 new student data obtained.

Table 1. New student Data obtained

No	Name	Origin Of The School	National Exam Scores
1.	Syifa Syafiqoh U	SDN Kebon Melati 01 Pagi	280,5
2.	Sema	MI Jami’at Kheir	272,0
3.	Nisma Gita P	SDN Kebon Kacang 5	272,0
4.	Azzalea Zahra	MI Jami’at Kheir	272,0
5.	Bella Lestari	SDN Kebon Kacang 5	280,0
6.	Fitriya Natasha	MI Jami’at Kheir	260,5
7.	Tarlina Lestari	SD Muhammadiyah 56	260,5
8.	Nur Annisa K	SDN Kebon Melati 01 Pagi	270,5
9.	Najla Muthiah	SDN Petamburan 01	270,0
10.	Furqon Hidayat	SD Muhammadiyah 56	260,5
11.	Ahmad Budi M	MI Jami’at Kheir	285,5
12.	Daffa Kusnandar	SDN Petamburan 01	270,0
13.	Panca Putra	SD Muhammadiyah 56	260,5
14.	Fatam Mustaqim	MI Jami’at Kheir 01	285,5
15.	Hayu Nisa Nur S	MI Jami’at Kheir	285,5
16.	Kokom	SDN	280,0

	Komariah	Petamburan 01	
17.	Adinda Ningrum	SDN Petamburan 01	280,0
18.	Praja Aditya	SDN Kebon Melati 01 Pagi	280,5
19.	Susanto	SD Muhammadiyah 56	260,5
20.	Akhdiatul Faiz	MI Jami'at Kheir	280,0

Result

From the data of new students of muhammadiyah School 6, data transformation will be carried out to change the data, the purpose of data transformation is so that the data can be processed using the K-Mean Clustering method. The variables used in the registration of new students is the origin of school data and national exam scores. For school origin variables are grouped into 3 groups. The First Data Group with the origin of SDN schools was transformed with a value of 1, the origin of Muhammadiyah elementary schools was transformed with a value of 2, and for the origin of MI schools was transformed with a value of 3. For the national exam score variables are grouped into 2 groups, the first for the national exam score with an average of <=270 is transformed with a value of 1 and the value of >270 is transformed with a value of 2.

The results of the transfomation can be seen in the following table:

Table 2. Data Transformation Results

ID	Origin Of The School	National Exam Scores
K1	1	2
K2	3	2
K3	1	2
K4	3	2
K5	1	2
K6	3	1
K7	2	1
K8	1	2
K9	1	2
K10	2	1
K11	3	2

K12	1	2
K13	2	1
K14	3	2
K15	3	2
K16	1	2
K17	1	2
K18	1	2
K19	2	1
K20	3	2

New student data processing is done after the transformation process so that new student data can be processed using the K-Mean Clustering method. The steps of K-Mean Clustering algorithm process are as follows:

First. Done k of the number of new clusters to be formed. The cluster to be created is 3 clusters.

Second. Determine the starting center point of each cluster. The determination of the initial Center Point in this study was determined randomly and the center point obtained can be seen in the following table:

Table 3. Cluster Start Center Point

Center Point		
Centroid 1	**1**	**2**
Centroid 2	**3**	**1**
Centroid 3	**2**	**1**

Third. Calculate the distance of each data to the center of the cluster between objects to the nearest centroid. The nearest Centroid will be the cluster followed by that data. The calculation of Euclidean distance can be done with the following equation :

$$d(p,q) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2}$$

The above equation is used because the attribute used is 2. From the data in the can, it will be calculated the distance from the first student data to the center of the first cluster with the equation :

$$d(1,1) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2}$$
$$= \sqrt{(1 - 1)^2 + (2 - 2)^2}$$
$$= 0$$

From the results of the calculations above get the result that the distance of the first new student data with the first cluster is 0. The distance of the first new student data to the center of the second cluster is calculated by the equation :

$$d(1,2) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2}$$
$$= \sqrt{(1 - 3)^2 + (2 - 1)^2}$$
$$= 2,2$$

From the above calculation results obtained that the distance of the first new student data to the second cluster is 2,2. The distance of the first new student data to the center of the third cluster can be calculated using the equation :

$$d(1{,}3) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2}$$
$$= \sqrt{(1 - 2)^2 + (2 - 1)^2}$$
$$= 1{,}4$$

From the above calculation results obtained that the distance of the first new student data to the third cluster is 1,4. Based on the results of the three calculations above it can be concluded that the distance of the first new student data closest is cluster 1, so the first new students into cluster 1. For more calculation results with 20 samples of new student data can be seen in the following table :

Table 4. Calculation results of each data to each cluster iteration 1

Data	Distance To Centroid			Closest	Cluster
to-i	1	2	3	Distance	followed
K1	0	2,2	1,4	0	1
K2	2	1	1,4	1	2
K3	0	2,2	1,4	0	1
K4	2	1	1,4	1	2
K5	0	2,2	1,4	0	1
K6	2,2	0	1	0	2
K7	1,4	1	0	0	3
K8	0	2,2	1,4	0	1
K9	0	2,2	1,4	0	1
K10	1,4	1	0	0	3
K11	2	1	1,4	1	2
K12	0	2,2	1,4	0	1
K13	1,4	1	0	0	3
K14	2	1	1,4	1	2
K15	2	1	1,4	1	2
K16	0	2,2	1,4	0	1
K17	0	2,2	1,4	0	1
K18	0	2,2	1,4	0	1
K19	1,4	1	0	0	3
K20	2	1	1,4	1	2

The next step is to perform cluster grouping from the results of iteration 1. Cluster grouping can be seen in the following table :

Table 5. Clustering Results Iteration 1

Cluster Group	**Group Members**	**Amount**
C1	K1,K3,K5,K8,K9,K12,K16,K17,K18	9
C2	K2,K4,K6,K11,K14, K15, K20	7
C3	K7,K10,K13,K19	4

After obtaining members of each cluster, then the new cluster center is calculated based on each cluster member that has been obtained by using the following formula :

$$C = \frac{\sum m}{n}$$

Description :

C = centroid data
M = data members belonging to a particular centroid.
N = number of data that belong to a given centroid.
An example calculation on cluster 1 is as follows :

$$C_1 = \frac{1+1+1+1+1+1+1+1+1}{9} = 1$$

$$= \frac{2+2+2+2+2+1+2+2+2}{9} = 1{,}8$$

An example calculation on cluster 2 is as follows :

$$C_2 = \frac{3+3+3+3+3+3+3}{7} = 3$$

$$= \frac{2+2+1+2+2+2+2}{7} = 1{,}85$$

An example calculation on cluster 3 is as follows :

$$C_3 = \frac{2+2+2+2}{4} = 2$$

$$= \frac{1+1+1+1}{4} = 1$$

In grouping the data using the above formula obtained a cluster center point with the following values :

Table 6. Center point iteration 1 after cluster

Center Point		
Centroid 1	1	1,8
Centroid 2	3	1,85
Centroid 3	2	1

Since the new centroids used have not converged, the iteration must continue. The final results of clustering 20 new student data can be seen in the following table :

Table 7. Data calculation results to each cluster on iteration 2

Data to-i	Distance To Centroid			Closest Distance	Cluster followed
	1	2	3		
K1	0,2	2,01	1,41	0,2	1
K2	2,01	0,15	1,41	0,15	2
K3	0,2	2,01	1,41	0,2	1
K4	2,01	0,15	1,41	0,15	2
K5	0,2	2,01	1,41	0,2	1
K6	2,15	0,85	1	0,85	2
K7	1,28	1,31	0	0	3
K8	0,2	2,01	1,41	0,2	1
K9	0,2	2,01	1,41	0,2	1
K10	1,28	1,31	0	0	3
K11	2,01	0,15	1,41	0,15	2
K12	0,2	2,01	1,41	0,2	1
K13	1,28	1,31	0	0	3
K14	2,01	0,15	1,41	0,15	2
K15	2,01	0,15	1,41	0,15	2
K16	0,2	2,01	1,41	0,2	1

K17	0,2	2,01	1,41	0,2	1
K18	0,2	2,01	1,41	0,2	1
K19	1,28	1,31	0	0	3
K20	2,01	0,15	1,41	0,83	2

Tabel 8. Hasil Pengelompokan Iterasi 2

Cluster Group	Group Members	Amount
C1	K1,K3,K5,K8,K9,K12,K16,K17,K18,	9
C2	K2,K4,K6,K11,K14,K15, K20	7
C3	K7,K10,K13,K19	4

In testing the 2nd and 3rd iterations did not change (the same as the previous centroid), then the iteration process is completed and obtained 3 clusters with 3 iterations.

The new student data processing using K-means clustering algorithm and RepidMiner software can be seen in the following figure :

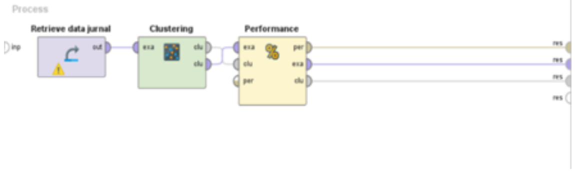
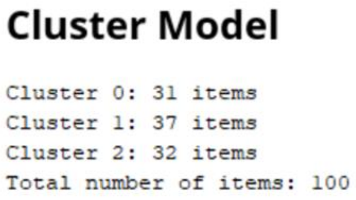


Figure 1. K-means modeling in RapidMiner

By using K-means clustering modeling as shown in Figure 2, with a total of one hundred data and the number of clusters of 3 pieces, in accordance with the definition of k values with the number of cluster_0 : 31 items, cluster_1 : 37 items, and cluster_2 : 32 items.



Cluster Model

```
Cluster 0: 31 items
Cluster 1: 37 items
Cluster 2: 32 items
Total number of items: 100
```

Figure 2. Cluster Model

The results of the distribution of cluster_0, cluster_1, and cluster_2 with a hundred data held in K-means clustering modeling using rapidminer, for 3 groups of data can be seen in the following figure :

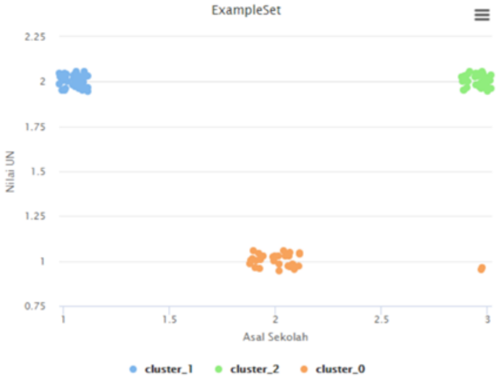


Figure 3. K-Means Clustering modeling on RapidMiner

For groups of data in the picture above consists of three data groups. The first group looks at the orange dots, the second group looks at the green dots, and the third group looks at the blue dots. The results of the analysis in Figure 3 contains the results of the grouping based on the proximity of the distance between the central points with new student data on each attribute.

Table 9. Cluster Analysis Results 1 (Cluster_0)

Cluster 1 Results	
Cluster 1 consists of students from : 31	
Comes from school :	**National Exam Scores :**
SDN : 0	<=270 : 31
SD Muhammadiyah : 29	>270 : 0
MI : 2	

Table 10. Cluster Analysis Results 2 (Cluster_1)

Cluster 1 Results	
Cluster 2 consists of students from : 37	
Comes from school :	**National Exam Scores :**
SDN : 37	<=270 : 0
SD Muhammadiyah : 0	>270 : 37
MI : 0	

Table 11. Cluster Analysis Results 3 (Cluster_2)

Cluster 1 Results	
Cluster 3 consists of students from : 32	
Comes from school :	**National Exam Scores :**
SDN : 0	<=270 : 0
SD Muhammadiyah : 0	>270 : 32
MI : 32	

From the clustering data that has been done above, it can be determined which new students are accepted or not accepted in Muhammadiyah 6 schools, according to the results of each cluster that has been formed.

Conclusion

After going through many stages undertaken in the application of k-means clustering can be concluded that : 1. the determination of the center point (centroid) at the beginning of the K-means algorithm stage has an influence on the cluster results as in the test results that have been carried out using 100 datasets with different centroids can produce different cluster results as well. 2. After clustering new student data using the K-means clustering algorithm from the data of 100 new students, three clusters were formed, namely cluster one with 31 items, cluster two with 37 items, and cluster three with 32 items. 3. Determination of new students who are accepted and not accepted in Muhammadiyah schools 6 will follow the cluster formed in accordance with the national exam scores.

Refrences

[1] A. Muhidin, P. Studi, T. Informatika, S. Tinggi, and T. Pelita, “Klasifikasi Penduduk Tidak Mampu Desa Mandiraha Wetan Menggunakan Algoritma C4.5,” *J. Pelitabangsa*, vol. 9, no. 3, pp. 13–18, 2019.

[2] 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.

[3] A. Muhidin, “Analisa Metode Hierarchical Clustering Dan K-Mean Dengan Model Lrfmp Pada Segmentasi Pelanggan,” *SIGMA (Jurnal Teknol. Pelita Bangsa)*, pp. 2407–3903, 2017.

[4] 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 Tek.*, vol. 18, no. 1, pp. 76–82, 2016, doi: 10.18196/st.v18i1.708.

[5] G. Abdurrahman, “Clustering Data Ujian Tengah Semester (UTS) Data Mining,” *J. Sist. Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 71–79, 2016.