SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The Implementation of K-Means Algorithm for Cluster Majoring to New Students in SMKN 2 of South Tangerang

Enggar Tyastiwi Munawaroh 1st
STMIK Nusa Mandiri Jakarta
raggnesayt@gmail.com

Rani Irma Handayani 2nd
STMIK Nusa Mandiri Jakarta
Rani.rih@nusamandiri.ac.id

Euis Widanengsih 3rd
Universitas Bina Sarana Informatika Karawang
Euis.ewh@bsi.ac.id

Abstract— The diversity of majors in vocational schools of SMKN 2 South Tangerang makes some students confuse their choices. Determination of majors is important because it will affect the academic activities of students. The purpose of the right majoring is so that students can learn optimally, and be able to equip themselves with competency skills according to their talents, interests, and abilities when entering the workforce. This study applies the Clustering Method with the K-means Algorithm, to help students determine their majors, also helps the school in clustering majors. Determination of these majors is based on 320 student data with attributes of the National Examination during Junior High School (Mathematics, English, Indonesian, and Science), Registration Pathways, and Gender. Calculations that occur as many as 7 iterations with the K-Means Clustering Method. The K-Means Clustering Method makes it easier to grouping new students' data. The calculation produces information on the number of students per majoring. Industrial Electronics Engineering has 49 students. Light Vehicle Engineering 95 students. Accounting major has 96 students. Multimedia major has 44 students. And Motorcycle Business Engineering has 36 students.
Keyword— Majoring, Vocational High School, Clustering, K-Means

I. INTRODUCTION

According to (Nugroho & Haryanti, 2015) "Determination of majoring carried out so far many weaknesses, such as, based on the wishes of students without considering at their value background of academic. So, the chosen majoring sometimes becomes a problem for the students in the future." The purpose of the right majoring is so that students can learn optimally, and be able to equip themselves with competency skills according to their talents, interests, and abilities when entering the workforce. "By applying this classification and clustering techniques in data mining (DM), strategic information can be extracted, then can be used to find new opportunities and strategic plans in the process of classifying majors towards

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

students." (Nugroho & Haryanti, 2015).

In this research, the determination of these majors was based on the National Examination scores (English, Indonesia, Mathematics, and Science), Registration pathways, and Gender. Through this research, it is expected to be able to direct students to the majoring that are in accordance with their potential and ability. There are 5 majors in SMKN 2 of South Tangerang: Multimedia, Accounting, Light Vehicle Engineering, Motorcycle Business Engineering, and Industrial Electronics Engineering.

According to (Baedowi, 2015) The initiators of vocational schools argue that "learning to do is most important; knowledge will somehow become into the process." The important role of vocational schools is to facilitate the construction of knowledge carried out by students through a series of field experiences (internships), contextual with the conditions and developing social environment. Because the main point of vocational schools is to improve students' skill, and as a forum for a learning process.

According to Widodo in (Metisen & Sari, 2015) "Clustering or classification is a method that used to divide data sets into several groups based on predetermined similarities." According to (Nasari, Darma, & Informasi, 2015) "K-Means Clustering Algorithm is able to cluster data in the same group and different data in different groups."

There are 5 clusters that will represent each majoring in this research: C1 = Industrial Electronics Engineering, C2 = Light Vehicle Engineering, C3 = Accounting, C4 = Multimedia, C5 = Motorcycle Business Engineering. Those data will be processed with The K-Means Algorithm, then the results are grouped into each cluster of majors. For instance: if students X's data is grouped in C4, then the right majoring for him/her is Multimedia.

II. LITERATURE REVIEW

A. KDD (Knowledge Discovery in Database)

According to (Nofriansyah, 2015) the term of data mining and knowledge discovery in database (KDD) is often used to explain the process of extracting hidden information in a large database. The fact, the two terms have different concepts but are related to each other. And one of the stages in the whole KDD process is data mining.

B. Data Mining

According to Maimon & Rokach in (Muningsih & Kiswati, 2015) Data mining is the main of the Knowledge Discovery in Database (KDD). Including allegations of algorithms that explore data, build models and find patterns that are unknown. KDD is automatic, it can be defined as organizing processes for correct identification, use and discovery of pattern from large and complex data sets.

According to Hermawati (2013) (Listriani, Setyaningrum, & M.A, 2016) Data mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically. Data mining is an iterative and interactive process for finding new patterns or valid models (perfect), useful and understandable in a large database.

C. Clustering

According to (Irwansyah & Faisal, 2015) *Clustering* or is a technique or method for grouping data. According to Tan, 2006, clustering is a process to group data into several clusters or groups, so that one cluster has a maximum level of data similarity and the data between clusters has a minimum similarity.

According to Widodo in (Mardalius, 2018), Clustering or Classification is a method used to divide data sets into gropus based on predetermined similarities. According to Santosa in (Titiani & Widiyono, 2017), Clustering is a method for finding and grouping data that has similarity characteristics (similarity) between one data with other data.

D. K-Means

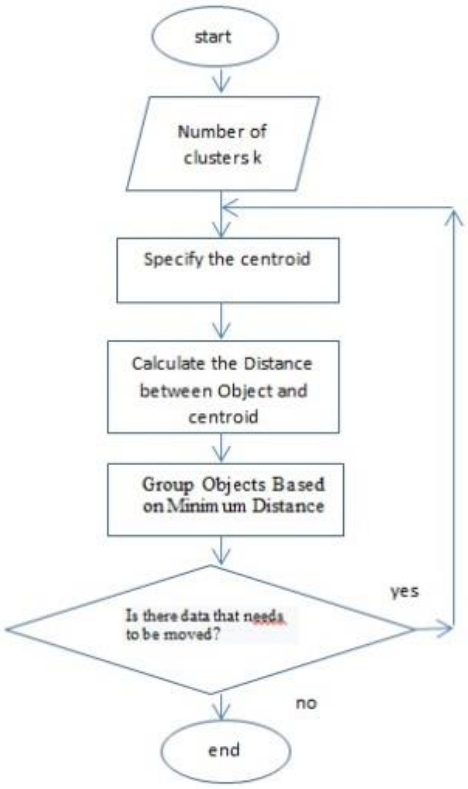
According to Larose in (Muningsih & Kiswati, 2015), K-Means is one method in clustering or grouping data. Clustering refers to observation in a case based on the similarity of the object. A cluster is a group of data that has similarity to the other, and dissimilarity to other data in another group.

"K-Means is an algorithm used in partitioning groups to separate data into different groups. This algorithm can minimize the distance between one data to its cluster." (Sari, Wanto, & Windarto, 2018).

III. PROPOSED METHOD

K-Means research steps.

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X



Source: (Sabariah, Istoningtyas, & Sharipuddin, 2018)

Picture.1. Flowchart K-Means

K-Means is a clustering algorithm with a center-based partitioning methode. According to (Sulistiyani, Soedijono, & Syahdan, 2015), the K-Means algorithm consists of several steps.

Step 1 : Determine the number of clusters.
Step 2 : Allocate data into clusters randomly.
Step 3 : Calculate centroids of the data in each cluster.
Step 4 : Allocate each data to the closest centroid.
Step 5 : Return to Step 3, if there is still data that moves within clusters or if there is a change in the centroid value.

IV. RESULT AND DISCUSSION

A. The Number of Cluster K

To determine the number of clusters K, the researcher determined it by the number of majors in SMKN 2 of South Tangerang. There are 5 clusters that will represent each major.

Table 1
Table Cluster Representing Majors

Cluster	**Major**
C1	Industrial Electronics Engineering
C2	Light Vehicle Engineering
C3	Accounting
C4	Multimedia
C5	Motorcycle Business Engineering

Table 2
Gender Transformation Table

Gender	Transformation value
Male	1
Female	2

Table 3
Table of Transformation Registration Pathway

Registration pathway	Transformation value
By Achievement	1
Public Administration	2

Table 4
Initialization table of the National Examination Score (English, Indonesian, Mathematics, and Science)

National Examination Score	Transformation value
10 – 29,9	1
30 – 49,9	2
50 – 69,9	3
70 – 89,9	4
90 - 100	5

In this research, the amount of data that we used is 320. Here is the link to 320 transformation data that can be seen and downloaded: http://bit.ly/320data

B. Specify the Centroid

According to Vulandari (2017:54), the K-Means Algorithm established the Cluster (K) values randomly, while the value becomes the center of the cluster or commonly referred to as centroid, mean or 'means'. According to (Irwansyah & Faisal, 2015), Randomly establish data k to be the initial center of the cluster location. For the initial cluster center point in this research, it's randomly determined. C1 uses the 6th data. C2 uses the 87th data, C3 uses the 131st data, C4 uses

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the 196th data, and C5 uses the 313th data.

Table 5
Cluster's Initial Center Point

C1	2	5	5	5	4	2
C2	1	4	3	3	4	2
C3	2	5	3	3	3	2
C4	1	4	3	1	2	1
C5	1	4	2	1	4	2

C. Calculate the Distance between Object and Initial Center Point

Calculate the distance of each data to the center of the cluster. Between objects to centroid by calculating Euclidean Distance. According to (Sabariah et al., 2018) the equation is:

$$(x2, x1) = \sqrt{\sum^{p} j = 1(x2j - x1j)^2}$$

That is:
P = data dimension
X1 = 1st point position
X2 = 2nd point position

$(p,q)=$

$$\sqrt{(p1-q1)^2 + (p2-q2)^2 + (p3-q3)^2 + (p4-q4)^2 + (p5-q5)^2 + (p6-q6)^2}$$

The equation happens because the data has 6 attributes, such as Gender, Indonesian, English, Mathematics, Science, and Registration pathway. For example, the distance from the 1st data will be calculated with all the clusters' center points.

Table 6
1st Data

Student name	Gender	indo	eng	MTK	sci	reg
ANNISA DINDA RAHMASARI	2	4	4	5	4	1

(1,1) $= \sqrt{(2-2)^2 + (4-5)^2 + (4-5)^2 + + (5-5)^2 + (4-4)^2 + + (1-2)^2}$
$= 1{,}732050808$

From the calculation above, it can be concluded that the distance from the first student named ANNISA DINSA RAHMASARI to the 1st cluster is 1,732

The distance of the first student to the second cluster is:
(1,1) $= \sqrt{(2-1)^2 + (4-4)^2 + (4-3)^2 + + (5-3)^2 + (4-4)^2 + + (1-2)^2}$
$= 2{,}645751311$

From the calculation above, it can be concluded that the distance from the first student named ANNISA DINSA RAHMASARI to the 2nd cluster is 2,645

The distance of the first student to the third cluster is:
(1,1) $= \sqrt{(2-2)^2 + (4-5)^2 + (4-3)^2 + + (5-3)^2 + (4-3)^2 + + (1-2)^2}$
$= 2{,}828427125$

From the calculation above, it can be concluded that the distance from the first student named ANNISA DINSA RAHMASARI to the 3rd cluster is 2,82.

The distance of the first student to the fourth cluster is:
(1,1) $= \sqrt{(2-1)^2 + (4-4)^2 + (4-3)^2 + + (5-1)^2 + (4-2)^2 + + (1-2)^2}$
$= 4{,}69041576$

From the calculation above, it can be concluded that the distance from the first student named ANNISA DINSA RAHMASARI to the 4th cluster is 4,69

The distance of the first student to the fifth cluster is:
(1,1) $= \sqrt{(2-1)^2 + (4-4)^2 + (4-2)^2 + + (5-1)^2 + (4-4)^2 + + (1-2)^2}$
$= 4{,}69041576$

From the calculation above, it can be concluded that the distance from the first student named ANNISA DINSA RAHMASARI to the 5th cluster is 4,69

Based on the calculation above, it can be concluded that the closest distance (minimum distance) of the 1st student named ANNISA DINSA RAHMASARI is to Cluster-1 (1st Cluster) with 1,732. So, this student goes into Cluster-1 (C1).

D. Group Objects Based on Minimum Distance

Table 7
Cluster Grouping Result.

Initial Center Point							total of student
C1	2	5	5	5	4	2	10
C2	1	4	3	3	4	2	141
C3	2	5	3	3	3	2	111
C4	1	4	3	1	2	1	42
C5	1	4	2	1	4	2	16
						TOTAL	320

Based on Table.7, the temporary results of students entering C1 (Industrial Electronics Engineering) are 10

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

students. C2 (Light Vehicle Engineering) total of 141 students. C3 (Accounting) there are 111 students. C4 (Multimedia) there are 42 students. And C5 (Motorcycle Business Engineering) there are 16 students.

E. Update Cluster Center Point

Recalculate the new center point by determining the average of each data that has formed its cluster. Then continue to calculate the closest distance with the next iterations in order to have data convergence. (Nur, Zarlis, & Nasution, 2015)

Table 8
Calculate the New Center Point of C1

NO	DATA-	C1					
1	1	2	4	4	5	4	1
2	3	2	4	4	4	4	1
3	6	2	5	5	5	4	2
4	7	2	4	4	4	4	2
5	8	2	5	4	4	4	2
6	10	2	4	4	4	4	2
7	52	2	4	4	5	3	2
8	98	2	4	4	4	4	1
9	103	1	4	5	4	4	2
10	105	1	4	4	5	4	2
New Centroid		1,8	4,2	4,2	4,4	3,9	1,7

The new center point of C1 is formed based on the average data that has formed by its cluster. Likewise, for the new center point of C2, calculate the average of total 141 data that formed its cluster. Then calculate the other clusters' average to get the whole new center point.

Table 9
The New Center Point (1st iteration).

C1	1,8	4,2	4,2	4,4	3,9	1,7
C2	1,11	3,70	3,23	2,83	3,23	1,98
C3	1,9	4,16	3,55	3,018	2,86	1,97
C4	1,14	3,88	3,17	1,98	2,07	1,86
C5	1,125	3,88	2	1,94	3,0625	2

After getting the new center point (1st iteration) see table 9 above, redo the C, D, and E steps. Calculate each data with the new center point. After that, re-update the center point, untill the center point of each cluster has no more change, and no more data moves from one cluster to another cluster.

Table 10
Cluster Grouping Result (1st iteration).

New Centroid for 1st iteration							Total of student
C1	1,8	4,2	4,2	4,4	3,9	1,7	32
C2	1,11	3,70	3,23	2,83	3,23	1,98	112
C3	1,9	4,16	3,55	3,018	2,86	1,97	102
C4	1,14	3,88	3,17	1,98	2,07	1,86	41
C5	1,125	3,88	2	1,94	3,0625	2	33
						TOTAL	320

Based on table 10, the temporary result os students that entering C1(Industrial Electronics Engineering) are 32 students. C2 (Light Vehicle Engineering) total of 112 students. C3 (Accounting) there are 102 students. C4 (Multimedia) there are 41 students. And C5 (Motorcycle Business Engineering) there are 33 students.

Re-update the center point for the 2nd iteration. It's because the center point isn't convergence, and the possibility of data moving from one cluster to another cluster is large.

Table. 11
The Result of the 2nd Iteration.

New Centroid for 2nd Iteration							Total of student
C1	1,625	4,09375	3,96875	4,125	3,53125	1,90625	46
C2	1,071	3,6875	3,241	2,696	3,178	1.99	90
C3	1,945	4,118	3,609	2,9818	2,945	1,954	99
C4	1,146341	3,878049	3,268293	2,073171	2,097561	1,853659	53
C5	1,06061	3,818182	2	2,333333	2,848485	2	32
						TOTAL	320

F. The Calculation is Complete, Data is No Longer Moved

In this research, 7 iterations occurred to generate the convergent center point. So, the data doesn't move from cluster to another cluster.

Table. 12
The Result of the 7th Iteration.

New Centroid for 7th iteration							Total of student
C1	1,55102	4,08163	3,653061	4,06122	3,3673	1,9388	49
C2	1	3,697917	3,260417	2,552083	3,09375	1,96875	95
C3	1,989583	4,09375	3,604167	2,791667	3,083333	1,947917	96
C4	1,186047	3,906977	3,395349	2,186047	2,023256	1,930233	44
C5	1,055556	3,722222	2	2,388889	2,861111	1,972222	36
						TOTAL	320

C1 = 1,55; 4,08; 3,65; **4,06; 3,36**; 1,93. The average result of National Examination score for Mathematics and Science on 1st Cluster is the biggest of all other clusters, then it drawn conclusion with those scores as Industrial Electronics Engineering Majors.

Center point of cluster C2 = **1**; 3,69; 3,26; **2,55**; **3,09**; 1,96. The result found that all of the students on the 2nd cluster were male. The average score of the National Examination for Mathematics and Science are quite big, it has drawn conclusion as Light Vehicle Engineering Major.

Center Point of cluster C3 = **1,98**; 4,09; 3,60; **2,79**; 3,083; 1,94. The result found that almost all of the

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

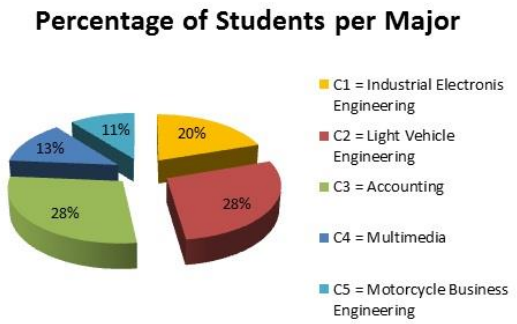
students on the 3rd cluster were female, the rest in small numbers were male. The average score of the National Examination for Mathematics was the second biggest after C1. Which is drawn as an Accounting Major.

Center Point of cluster C4 = **1,18**; **3,90**; **3,39**; 2,18; 2,02; 1,93. The result found that 36 out of 44 students were male. The average score of the National Examination for Mathematics is the lowest, but it has a big average score of Indonesian and English. Then it drew the conclusion as a Multimedia Major.

Center Point of cluster C5 = 1,055; 3,72; 2; 2,389; 2,86; 1,97. And the last center point was determined as Motorcycle Business Engineering Major. Because most of all students are male, and the average score of National Examination on all subjects are pretty good.

Tabel.13
The Final Result Cluster for 320 Data.

Cluster	Major	Total of students
C1 =	Industrial Electronics Engineering	49
C2 =	Light Vehicle Engineering	95
C3 =	Accounting	96
C4 =	Multimedia	44
C5 =	Motorcycle Business Engineering	36
	Total	320



Picture.2. Percentage of Students per Major

V. CONCLUSION AND SUGGESTION

Vocational High Schools is a solution for students who are going to enter the workspace, by preparing them with competency skills and experiences (internship). The purpose of the right majoring is, the students can learn optimally, and be able to equip themselves with competency skills according to their talents, interests, and abilities.

The conclusion and suggestion from this research, by implementing the K-Means Algorithm are:

1. The iterations that occurred in this research is 7 times. The cluster results are influenced by the center point and the amount of data usage.
2. The K-Means Clustering Method makes it easier to grouping new students' data. The calculation produces information on the number of students per majoring. Industrial Electronics Engineering has 49 students. Light Vehicle Engineering 95 students. Accounting major has 96 students. Multimedia major has 44 students. And Motorcycle Business Engineering has 36 students.
3. For further research, an application or calculation program is needed to make it easier to calculate large amounts of data. Because manual calculations need high accuracy. If there is a miscalculation in the 1st iteration, the results of the next iteration will be invalid.

VI. REFERENCE

Baedowi, A. (2015). *Potret Pendidikan Kita*. (Aisyah, Ed.) (1st ed.). South Tangerang: PT Pustaka Alvabet.

Irwansyah, E., & Faisal, M. (2015). *Advanced Clustering Teori dan Aplikasi* (1st ed.). Yogyakarta: CV BUDI UTAMA.

Listriani, D., Setyaningrum, A. H., & M.A, F. E. (2016). PENERAPAN METODE ASOSIASI MENGGUNAKAN ALGORITMA APRIORI PADA APLIKASI ANALISA POLA BELANJA KONSUMEN (Studi Kasus Toko Buku Gramedia Bintaro), *9*(2), 120–127.

Mardalius. (2018). PENGELOMPOKAN DATA PENJUALAN AKSESORIS MENGGUNAKAN ALGORITMA K-MEANS, *IV*(2), 401–411.

Metisen, B. M., & Sari, H. L. (2015). ANALISIS CLUSTERING MENGGUNAKAN METODE K-MEANS DALAM PENGELOMPOKKAN PENJUALAN PRODUK PADA SWALAYAN FADHILA, *11*(2), 110–118.

Muningsih, E., & Kiswati, S. (2015). Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang, *3*(1).

Nasari, F., Darma, S., & Informasi, S. (2015). PENERAPAN K-MEANS CLUSTERING PADA DATA PENERIMAAN MAHASISWA BARU, 6–8.

Nofriansyah, D. (2015). *Konsep Data Mining VS Sistem Pendukung Keputusan* (1st ed.). Yogyakarta: CV BUDI UTAMA.

Nugroho, Y. S., & Haryanti, S. N. (2015). Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali, *I*(1), 3–8.

Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10133
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Nur, F., Zarlis, M., & Nasution, B. B. (2015). Penerapan Algoritma K-Means Pada Siswa Baru Sekolah Menengah Kejuruan Untuk Clustering Jurusan, (9), 100–105.

Sabariah, Istoningtyas, M., & Sharipuddin. (2018). Penentuan Jurusan ke Perguruan Tinggi Menggunakan Metode Clustering di SMAN 3 Kuala Tungkal, *13*(2).

Sari, R. W., Wanto, A., & Windarto, A. P. (2018). IMPLEMENTASI RAPIDMINER DENGAN METODE K-MEANS (STUDY KASUS : IMUNISASI CAMPAK PADA BALITA BERDASARKAN PROVINSI), *2*, 224–230.

Sulistiyani, M. E., Soedijono, B., & Syahdan, S. A. (2015). SISTEM PENENTUAN JURUSAN SEKOLAH MENENGAH ATAS NEGERI, 6–8.

Titiani, T., & Widiyono, C. (2017). PENENTUAN STRATEGI PROMOSI PENERIMAAN MAHASISWA BARU DENGAN ALGORITMA CLUSTERING K-MEANS, *XII*(25), 39–44.

Vulandari, R. T. (2017). *Data Mining Teori dan Aplikasi Rapidminer* (1st ed.). Yogyakarta: Gava Media.