



## Analysis of Factors Affecting Student Graduation Using the K-Means clustering Method

Indrawan<sup>1</sup>, Agung Triayudi<sup>2\*</sup>, Novi Dian Nathasia<sup>3</sup>

Fakultas Teknologi Komunikasi dan Informatika,  
Universitas Nasional, Sawo Manila, RT.14/RW.3, Ps. Minggu, Kec. Ps. Minggu, Kota Jakarta Selatan,  
Daerah Khusus Ibukota Jakarta

Email :<sup>1</sup>ndraa9@gmail.com,<sup>2</sup>agungtriayudi@civitas.unas.ac.id,<sup>3</sup>novidiannathasia@civitas.unas.ac.id

\*Corresponding author

### ARTICLEINFO

Articlehistory:  
Received: 01/11/2020  
Revised: 16/01/2020  
Accepted: 01/02/2020

Keywords:  
Factor Analysis, Clustering,  
K-means

### ABSTRACT

*This research was conducted because the system was not yet available to analyze the graduates of FTKE students. The purpose of this study was to obtain the results of the analysis of factors affecting the graduation of the National University FTKE students. This study uses the K-means method and uses GPA, income, place of birth, organizational activity, SKPI scores as parameters. Data collection techniques in the form of quantitative data collection. The subjects of this study were alumni of the FTKE National University. From the results of research conducted cluster 3 has the highest average IPK value of 3.57, of the total birth place clusters most are in Jabodetabek, all of the clusters show that organizational inactivity is the highest value, overall income clusters have the most value in class one, in cluster 3 the SKPI score has the highest average value of 3451.*

Copyright © 2020 Journal Mantik.  
All rights reserved,

### 1. Introduction

Lack of information about the graduate students to be a an issue in this digital era. In a graduation course, is inseparable from the various aspects of the supporters. In this study, the authors analyze the factors that may affect graduation, while the factors analyzed in this study is a GPA, earning parents, place of birth, the activity of the organization, and the score SPKI.

Economic factors in the present study can be interpreted also as the financial ability of each parent, because the better the financial capability then it is likely to support the learning process a little or a lot will increase.

Activeness of the organization in this study can be interpreted also as students follow the organization on campus, in this study the organization in question is a set of students who are at FTKE, namely HIMASI and HIMTI, student participation in the activities of the organization is also a provision for interaction with other students so that from student participation process can create a good learning environment, following student organizations are trained to set patterns of thought and it is more or less affect student achievement. SPKI score companion diploma or certificate is an official document issued by the college, briefly SPKI is a track record notes that have been passed by the student when undergoing the lecture. SPKI can be obtained by depositing some evidence of a good achievement in academic and non-academic, such evidence may be a plaque, certificate and others. In this study score SPKI be used as one factor, because of the activities of the students themselves who score SPKI can be obtained. Limitation of problems in the present study is when the author has received the results of the calculation of the K-means algorithm assisted with the software RapidMiner. In research conducted by Jaroji, which applies the algorithm k-means to obtain the results of the analysis and determination of the recipient Misi on the campus of Polytechnic Bengkalis showed as many as 17 people recommended by the consideration, 24 were recommended highly feasible, 32 people recommended feasible and 56 people recommended less feasible [1].





The aim of this study is

- To obtain the results of the analysis factors affecting graduation GPA mahasiswa yaitu obtain analysis results.
- To obtain the results of the analysis factors that affect student graduation is the place of birth.
- To obtain the results of the analysis factors that affect student graduation is liveliness organization.
- To obtain the results of the analysis factors that affect graduation students are earning.
- And to get the results of the analysis factors that affect student graduation namely score SPKI.

## 2. Research methods

In this research, the stages can be seen below:

### a) Data retrieval

In this stage the process of making data available on BPSI National University, which contains the raw data from tables.

### b) Sort Data

At this stage, the authors conducted a sorting of data is needed. Like, take some research supporting variables.

### c) Preparing dataset

The study involved faculty alumni data communication and information technology obtained from the National University BPSI pass commencing from the year 2017- 2019, the data includes the name, NPM, courses, place of birth, ipk, active organization, income, and score SPKI. The amount of data used in this study, numbering as many as 350 data, which in 2017 amounted to 101 student graduation, in October 2018 amounted to 54 students, in 2019 april totaling 106 students, and in 2019 the month of September totaled 47 students. The data is the data with the .xlsx format, because the data obtained is an excel document.

**Table 1.**  
Examples of alumni data table

Nama	NPM	Program Studi	Tempat Lahir	Ipk	Keaktifan organisasi	Penghasilan	Skor SKPI
Andi	123112700650019	1	3	3.11	1	2	2777
Bayu Adi Tyarinaldi	133112700650006	1	2	3.03	2	2	2264
Ilham Maulana Abdillah	143112700650004	1	2	3.5	2	2	3701
Fahmi Aditya	163112700620006	1	1	3.47	1	2	3233
Wijayanto Adiwibowo	163112700620024	1	2	3.44	1	2	2627
Ardiansyah Ilham	163112700620088	1	1	3.57	2	2	3547
Alfian Frandika	163112700620092	1	1	3.19	2	2	3672
M. Taufan Murdiansyah	163112700620098	1	1	3.04	2	2	3716
Agung Riyadi	163112700620147	1	2	3.6	2	1	2407
Wisnu Prabowo	163112700620152	1	2	3.6	2	5	3071

### d) Pre-Process Data

Pre-process the data is a stage that is needed in the data mining process that is closely related to the preparation and manufacture of the initial data sets. Initialization of data needed to facilitate the process of clustering, the K-means algorithm the data used must be numeric, then the data manifold nominal data such as course of study, place of birth, the activity of the organization, the income must be changed beforehand in the form of numbers.

**Table 2.**  
Sample data initialization program of study.

Sistem Informasi	1
Informatika	2

In Table 2 do initialization on the course for information systems and informatics. The study program is initialed 1 information systems and informatics courses given inisial 2.





**Table 3.**

Sample data initialization place of birth.

Jakarta	1
Bogor	1
Depok	1
Tangerang	1
Bekasi	1
Jogja	2
Palembang	3

In Table 3 do inialisasi place of birth that is in the Greater Jakarta area is initialed 1, the outer bead area but still in the island of Java is initialed 2, and outside the island of Java by inisal 3.

**Table 4.**

Sample data initialization activity of the organization.

Iya	1
Tidak	2

In Table 4 do initialization data activeness of the organization, namely the data iya is initialed 1, and the data is not Tiberi value 2.

**Table 5.**

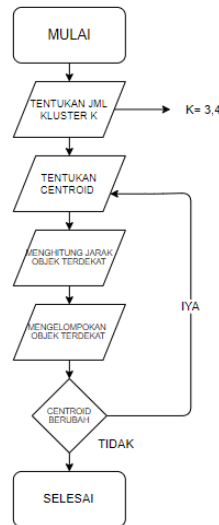
Sample data initialization income.

Tidak Berpenghasilan	1
0 s.d Rp. 250 rb	2
Rp. 250 rb - Rp 500 rb	3
Rp. 500 rb - Rp 1 Juta	4
Rp. 1 Juta - Rp 2 Juta	5
Rp. 2 Juta - Rp 3 Juta	6
Rp. 3 Juta - Rp 4 Juta	7
Rp. 4 Juta - Rp 5 Juta	8
> Rp. 5 Juta	9

In Table 5 do inialisasi on income data are on income not rated 1, at 0 up to Rp. 250K given initialization 2, at Rp. 250K - US \$ 500K by initialization 3, at Rp. 500K - US \$ 1 Million by initialization 4, at Rp. 1 Million - US \$ 2 Million by initialization 5, at Rp. 2 Million - US \$ 3 million by the initialization 6, at Rp. 3 Million - US \$ 4 million was given initialization 7, at Rp. 4 Million - US \$ 5 million by the initialization 8, at > USD. 5 Million given initialization 9.

e) K-means clustering algorithm

At this stage, the K-means clustering is used to perform an existing dataset into three kluster. K-meansklustering a non-hierarchical clustering methods that categorize the data in the form of one or more clusters / groups. Data that have the same characteristics are grouped in satukluster / groups and data that have different characteristics are grouped by cluster / group to another, so that the data are in one cluster / group has a level of variation is kecil. K-means clustering algorithm that is partitioned D into k clusters set of data. K-means clustering algorithm mengkluster all the data points in D such that the data point xi being the only k partitions. In other words, one data point only fit into a single cluster.



**Image 1.** Flowchart method of K-means algorithm.

The steps of the algorithm K-means adalah as follows:

- 1) Determine K data as a centroid, K adalah the desired number of clusters (as determined by the investigator).
- 2) Each dot (data) and then look for the nearest centroid.
- 3) Each set of points (centroid data into so-called clusters).
- 4) Recalculate the centroid of each cluster.
- 5) Repeat steps 1-4 until centroid tidak changed.

Clustering method using K-means algorithm, the size of the proximity of the data is calculated using the Euclidean distance. K-means algorithm is aimed at minimizing the total Euclidean distance between each point  $x_i$ , and the nearest cluster. Euclidean distance is determined using the following equation [5].

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \dots (1)$$

d) Evaluation

At this stage the information obtained from the data mining process needs to be displayed in the form of informative so that it can be easily understood by the parties concerned.

### 3. Results and Discussion

Further processing of the data to obtain information about graduate students FTKI calculated by means of manual calculations. The process this time using three clusters and attributes of the data used is the program of study, place of birth, ipk, active organization, income, and score SPKI.

In Table 6, the authors randomly select the data to get the centroid which amounts to 20 data.



**Table 6.**  
Sample student data.

NPM	Program Studi	Tempat Lahir	lpk	Keaktifan organisasi	Penghasilan	Skor SKPI
153112700620042	1	1	3.46	1	9	2474
153112700650043	1	1	3.63	2	9	2721
153112700640044	1	2	3.75	2	5	3701
153112700650045	1	2	3.82	2	7	3306
153112700650047	1	2	3.35	2	1	3226
153112700650048	1	2	3.74	2	7	3475
153112700640055	1	1	3.55	2	1	2702
153112700620061	1	3	3.68	2	1	3656
153112700650062	1	3	3.77	2	8	3204
153112700640074	1	3	3.54	2	7	3653
153112700620079	1	3	3.2	2	1	2571
163112700670087	1	3	3.23	2	5	2155
163112700620104	1	1	2.82	2	1	2153
163112700620141	1	1	2.94	2	1	2072
163112700620146	1	1	3.25	1	1	2110
173112700620005	1	1	3.41	2	8	3613
173112700620082	1	2	3.62	1	1	3702
173112700620124	1	1	3.28	2	7	2177
173112700620136	1	1	3.09	1	1	3673
173112700620137	1	2	3.19	2	9	2307
173112700620138	1	1	3.04	2	5	2035

The next thing to do is to choose the desired number of clusters, and determine the centroid randomly.

**Table 7.**  
Samples centroid.

C1	1	2	3.35	2	1	3226
C2	1	3	3.23	2	5	2155
C3	1	2	3.19	2	9	2307

Next is to calculate the distance of each data to the center of the cluster with the calculation of the Euclidean distance, then we will get the distance matrix calculation as follows:

$$d = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \dots (1)$$

From the calculation of the distance of the first data to the data for 20 of the centroid, thus gaining more distance calculation results are:

$$d = \sqrt{(1-1)^2 + (1-2)^2 + (3.46-3.35)^2 + (1-2)^2 + (9-1)^2 + (2474-3226)^2}$$

$$d = 752.044$$

Calculation to obtain C1 is like the above equation, the calculation result has a value equal to 8 tables.

$$d = \sqrt{(1-1)^2 + (1-3)^2 + (3.46-3.23)^2 + (1-2)^2 + (9-5)^2 + (2474-2155)^2}$$

$$d = 319.033$$

Calculation to find C2 is like the above equation, the calculation result has a value equal to 8 tables.

$$d = \sqrt{(1-1)^2 + (1-2)^2 + (3.46-3.19)^2 + (1-2)^2 + (9-9)^2 + (2474-2307)^2}$$

$$d = 167.066$$

Calculation to find the C3 is like the above equation, the calculation result has a value equal to 8 tables.





**Table 8.**

The results of one iteration.

NO	C1	C2	C3	JARAK MIN	CLUSTER
1	752.044	319.033	167.006	167.006206	3
2	505.064	566.018	414.001	414.001442	3
3	475.017	1546	1394.01	475.01701	1
4	80.2261	1151	999.002	80.2260612	1
5	0	1071.01	919.035	0	1
6	249.073	1320	1168	249.072584	1
7	524.001	547.018	395.082	395.082434	3
8	430.001	1501.01	1349.02	430.001289	1
9	23.1123	1049	897.001	23.1122565	1
10	427.043	1498	1346	427.043366	1
11	655.001	416.019	264.123	264.123078	3
12	1071.01	0	152.056	0	2
13	1073	4.91611	154.211	4.91610618	2
14	1154	83.1209	235.138	83.1209005	2
15	1116	45.2327	197.167	45.2327359	2
16	387.065	1458	1306	387.064599	1
17	476.001	1547.01	1395.02	476.001127	1
18	1049.02	22.1811	130.019	22.1811294	2
19	447.002	1518.01	1366.02	447.002313	1
20	919.035	152.056	0	0	3

Furthermore, the calculation back to get a new centroid obtained from a single iteration, the process of getting a new centroid authors calculated the average of the individual cluster members. So we get a new centroid, such as the following:

**Table 9.**

New centroid.

C1	1	2.1	3.57	1.8	4.6	3520.9
C2	1	1.4	3.10	1.8	3	2133.4
C3	1	1.6	3.40	1.8	5.8	2555

Re-calculation process is then performed using the new centroid that generate iterations of the two, such as the following:

**Table 10.**

The results of two iterations.

NO	C1	C2	C3	JARAK MIN	CLUSTER
1	1046.91	340.6542	81.06938	81.06938	3
2	799.9129	587.631	166.0322	166.0322	3
3	180.1007	1567.602	1146	180.1007	1
4	214.9137	1172.607	751.0012	214.9137	1
5	294.9221	1092.602	671.0173	294.9221	1
6	45.96356	1341.606	920.001	45.96356	1
7	818.9087	568.6039	147.0798	147.0798	3
8	135.1511	1522.602	1101.011	135.1511	1
9	316.9196	1070.613	649.0054	316.9196	1
10	132.125	1519.606	1098.002	132.125	1
11	949.9073	437.6076	16.76544	16.76544	3
12	1365.9	21.75263	400.0033	21.75263	2
13	1367.905	19.70884	402.0296	19.70884	2
14	1448.905	61.4344	483.0245	61.4344	2
15	1410.905	23.50282	445.027	23.50282	2
16	92.16966	1479.609	1058.002	92.16966	1
17	181.1376	1568.602	1147.01	181.1376	1
18	1343.903	43.78576	378.0025	43.78576	2
19	152.1494	1539.602	1118.011	152.1494	1
20	1213.908	173.7048	248.0211	173.7048	2

The calculation will be dismissed if the cluster point nothing has changed again and no data is moved from one cluster to another cluster or clusters result is stable and convergent. At this time the author calculations iterate 3 times to reach the converging point.





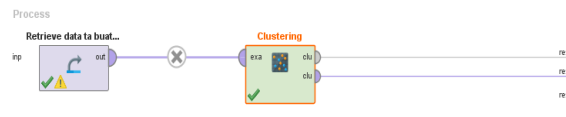
**Table 11.**

The results of the three iterations.

NO	C1	C2	C3	JARAK MI	CLUSTER
1	1046.91	311.7085	143.0588	143.0588	3
2	799.9129	558.6895	104.0785	104.0785	3
3	180.1007	1538.667	1084	180.1007	1
4	214.9137	1143.671	689.0032	214.9137	1
5	294.9221	1063.671	609.0134	294.9221	1
6	45.96356	1312.67	858.0026	45.96356	1
7	818.9087	539.6754	85.09595	85.09595	3
8	135.1511	1493.671	1039.009	135.1511	1
9	316.9196	1041.676	587.0097	316.9196	1
10	132.125	1490.671	1036.003	132.125	1
11	949.9073	408.6805	46.19935	46.19935	3
12	1365.9	7.554466	462.0026	7.554466	2
13	1367.905	9.822066	464.018	9.822066	2
14	1448.905	90.38483	545.0152	90.38483	2
15	1410.905	52.42844	507.0166	52.42844	2
16	92.16966	1450.672	996.0047	92.16966	1
17	181.1376	1539.67	1085.008	181.1376	1
18	1343.903	14.98058	440.0049	14.98058	2
19	152.1494	1510.67	1056.008	152.1494	1
20	1213.908	144.754	310.0264	144.754	2

Furthermore, the authors did clustering software RapidMiner.

Here are processing the data using RapidMiner:



**Figure 2.**Modeling K-means using RapidMiner.

By making the process of modeling the K-means as shown above, the result in clusters of about 3 clusters liking writer is, kluster\_0 there were 101 items, kluster\_1 there are 118 items, kluster\_2 there are 131 items, with the total number of items as many as 350 items.

**Table 12.**

The results of the implementation of the K-means clustering using rapid miner.

Cluster model	
Cluster_0	101 Items
Cluster_1	108 Items
Cluster_2	131 Items
Total Number of items	350 Items

Cluster distribution pattern that can be seen in view of rapid miner plot is as follows:





Figure 3. Display plot view on RapidMiner.

Inter-cluster distribution pattern not much different look for each cluster not disagree much on the number of members.

Results of cluster grouping which has been calculated using the application RapidMiner can be seen in the following table:

Table 13.  
Results of cluster grouping.

HASIL KLUSTER 1	HASIL KLUSTER 2	HASIL KLUSTER 2	HASIL KLUSTER 2
SI	32	SI	49
TI	69	TI	69
JABODETABEK	58	JABODETABEK	81
PULAU JAWA	31	PULAU JAWA	28
LUAR PULAU JAWA	12	LUAR PULAU JAWA	9
RATA-RATA IPK	3.48	RATA-RATA IPK	3.479
AKTIF ORGANISASI	IYA = 32	AKTIF ORGANISASI	IYA = 38
AKTIF ORGANISASI	TIDAK = 69	AKTIF ORGANISASI	TIDAK = 80
PENGHASILAN	GOL 1 = 27	PENGHASILAN	GOL 1 = 36
	GOL 2 = 11		GOL 2 = 11
	GOL 3 = 2		GOL 3 = 3
	GOL 4 = 4		GOL 4 = 6
	GOL 5 = 19		GOL 5 = 15
	GOL 6 = 14		GOL 6 = 13
	GOL 7 = 13		GOL 7 = 18
	GOL 8 = 6		GOL 8 = 11
	GOL 9 = 5		GOL 9 = 5
RATA-RATA SKPI	2770	RATA-RATA SKPI	2210
		RATA-RATA SKPI	3451

In cluster 1 has an average ipk at 3:48, which is composed of students majoring in information systems as many as 32 students, a student majoring in informatics as many as 69 students, housed born in Jabodetabek as many as 58 students, the island of Java in addition to the Greater Jakarta as many as 31 students, outer islands 12 students, as many as 32 active student organizations, is off by 69 student organizations, a majority-income or no-income first goal, and has an average SPKI of 2770.

In cluster 2 has an average ipk amounted to 3,479, which consisted of students majoring in information systems as many as 49 students, a student majoring in informatics as many as 69 students, housed born in Jabodetabek as many as 81 students, the island of Java in addition to the Greater Jakarta as many as 28 students, outside Java 9 as many as 38 active student organizations, is off by 80 student organizations, a majority-income or no-income first goal, and has an average SPKI of 2210.

In cluster 3 have an average ipk at 3:57, which consists of students majoring in information systems for 42 students, a student majoring in informatics as many as 89 students, housed born in Jabodetabek total of 77 students, the island of Java in addition to the Greater Jakarta were 38 students, outer islands 16 as many as 37 active student organizations, is off as much as 94 student organizations, a majority-income or no-income first goal, and has an average SPKI of 3451.

#### 4. Conclusion

Based on the clusters using K-means algorithm and software assisted by rapidminerdi above it can be concluded that:

- Cluster has an average value is the highest ipk cluster 3 that with a value of 3:57, which consists of students as much as 42 student information system, as well as informatics as many as 89 students.





- b) Furthermore, the results of the three clusters of factors cradle has a mode value or the highest value in the Greater Jakarta area.
- c) On the third cluster mode liveliness organizational value or value most in not actively organize.
- d) Later in the third cluster, factor income has a mode value or at most in group 1 or no income.
- e) In SPKI value factors into three clusters have an average value of the largest in the amount of 3451.

## 5. Reference

- [1] Jaroji, J., Danuri, D., & Putra, F. P. (2016). K-Means Untuk Menentukan Calon Penerima Beasiswa Bidik Misi Di Polbeng. *INOVTEK Polbeng - Seri Informatika*, 1(1), 87.
- [2] Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, 5(1), 62.
- [3] Widodo and D. Wahyuni, "Implementasi Algoritma K-Means Clustering Untuk Mengetahui Bidang Skripsi Mahasiswa Multimedia Pendidikan Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta," *Jurnal Pinter*, vol. 1, 2017.
- [4] Islam, H., & Haque, M. (2012). An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree. *International Journal of Advanced Computer Science and Applications*, 3(8), 146–149.
- [5] Asroni, R. A. (2015). Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang. *Ilmiah Semesta Teknika*, 18(1), 76–82.
- [6] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of Student Academic Performance By an Application of K-Means Data Mining Techniques. *Management and Artificial Intelligence*, 6(7), 110–114.
- [7] Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance, 7, 292–295.
- [8] Durairaj, M., & Vijitha, C. (2014). Educational Data mining for Prediction of Student Performance Using K-Means Clustering Algorithms. *International Journal of Computer Science and Information Technologies*, 5(4), 5987–5991.
- [9] M, B., Tomy, J., A, U., & Jacob, P. (2011). K- Means Clustering Student Data to Characterize Performance Patterns. *International Journal of Advanced Computer Science and Applications*, 1(3), 138–140.
- [10] Haris, A., & Hendrian, E. (2019). SISTEM MONITORING DAN KLASTER KETERSEDIAAN ENERGI MENGGUNAKAN METODE K-MEANS PADA PEMBANGKIT LISTRIK TENAGA SURYA, 4(2), 266–271

