JATI (Jurnal Jaringan dan Teknologi Informasi)
Vol. 3, No. 1, Juni 2023, hlm. 80-86 p-ISSN: xxxx-xxxx
DOI: 00.0000/jati e-ISSN: xxxx-xxxx

PENENTUAN JUMLAH KLASTER TERBAIK PADA K-MEANS DALAM MELIHAT POLA KLASTERING DATA MAHASISWA YANG TELAH LULUS

Aisya Basri[1], Abdul Mubarak[2], Hairil Kurniadi Siradjuddin[3], Saiful Do. Abdullah[4]

[1,2,3,4] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Khairun Jl. Jati Metro, Kota Ternate Selatan
Email: [1]aisyabasri28@gmail.com, [2]abdulmubarak029@gmail.com, [3]hairilkurniadi@gmail.com, [4]saifulabdullah12@gmail.com

(Naskah masuk: 25-05-2023, diterima untuk diterbitkan: 01-06-2023)

Abstrak

Keberhasilan *K-Means* dalam menganalisa data dapat terlihat dari pengelompokan yang terbentuk berdasarkan pada jumlah klaster yang ditentukan. Algoritma yang digunakan *K-Means* dalam menentukan banyak nya jumlah klaster dilakukan secara acak, hal ini dapat menyebabkan hasil klaster yang terbentuk tidak optimal. Untuk menentukan jumlah klaster yang optimal maka dilakukan penelitian dengan menggunakan metode *Elbow*. *Elbow* merupakan salah satu metode yang dapat digunakan dalam menentukan jumlah klaster terbaik dengan representasi grafik hasil dari perhitungan *Sum of Square Error* (SSE). Adapun penelitian yang dilakukan menggunakan dataset dengan parameter nilai IPK dan jumlah SKS dalam melakukan klastering waktu kelulusan dengan jangkauan jumlah klaster sebanyak 9 klaster. Pengelompokan data dengan jangkauan jumlah 9 klaster menggunakan *K-Means* menghasilkan data berubah-ubah klaster, bergantung pada jumlah klaster yang ditentukan. Setiap klaster yang terbentuk dari jangkauan 9 klaster digunakan dalam perhitungan SSE untuk menentukan jumlah klaster terbaik dengan representasi menggunakan grafik *Elbow*. Berdasarkan hasil perhitungan SSE maka didapatkan bahwa jumlah klaster terbaik pada penelitian ini ialah 2 klaster dengan nilai selisih SSE sebesar 4611.379920 dan berhasil membentuk garis siku pada grafik. Pengelompokan data berdasarkan jumlah klaster optimal dalam melakukan klastering waktu kelulusan terdiri dari klaster 1 sebanyak 199 data sebagai klaster tepat waktu dan klaster 2 sebanyak 41 data sebagai klaster tidak tepat waktu.

Kata kunci: klastering, k-means, elbow, sse, waktu kelulusan

DETERMINATION THE BEST NUMBER OF CLUSTERS IN K-MEANS FOR SEEING CLUSTERING PATTERN OF GRADUATES DATA

Abstract

The success of K-Means in analyzing data can be seen from the grouping formed based on the number of clusters specified. The algorithm used by K-Means in determining the number of clusters is selected randomly, this can cause the results of the clusters formed to be not optimal. To determine the optimal number of clusters, so the research was conducted using the Elbow method. Elbow is one of the methods that can be used for determining the best number of clusters by representation of the graph that results from the Sum of Square Error (SSE) calculation. The research was conducted using a dataset with parameters of GPA value and the number of credits for clustering the graduation time with a range of 9 clusters. Grouping data with a range of 9 clusters using K-Means it resulted that data keep on change the clusters, depending on the number of clusters specified. Each cluster formed from a range of 9 clusters is used on SSE calculations to determine the best number of clusters with representation using an Elbow graph. Based on the results of the SSE calculation, it obtained that the best number of clusters in this research was 2 clusters with an SSE difference value of 4611.379920 and it managed to form an Elbow line on the graph. Data grouping based on the number of optimal clusters for clustering the graduation time consists of cluster 1 as many as 199 data as a timely cluster and cluster 2 as many as 41 data as an untimely cluster.

Keywords: *clustering, k-means, elbow, sse, graduation*

1. PENDAHULUAN

K-Means adalah salah satu dari metodedalam teknik klastering yang paling sederhana[1]. Algoritma K-Means ialah sebuah algoritma non hirarki yang bertujuan untuk membagi data ke dalam satu atau lebih klaster berdasarkan kedekatan data ke titik pusat klaster[2]. Menurut Helen Stetsenko (2021) yang melakukan penelitian dengan judul "Top 10 Popular Data Mining Algorithms" dan dimuat pada situs KeyUa, K-Means termasuk dalam 10 besar algoritma yang banyak digunakan untuk diterapkan pada data mining terutama dalam teknik klastering. Dari penelitian yang dilakukan dapat disimpulkan bahwa K-Means banyak digunakan karena prosedur kerja yang mudah untuk diterapkan pada dataset dalam menerapkan data mining[3].

Prosedur dari K-Means dimulai dengan menentukan jumlah dari klaster yang akan dibentuk secara acak, selanjutnya penentuan nilai titik pusat dari tiap klaster, kemudian mengukur kedekatan objek terhadap nilai titik pusat dan mengelompokan data sesuai dengan jarak terdekat dengan titik pusat klaster[4]. Prosedur kerja K-Means bertujuan untuk menentukan berapa jumlah dari klaster yang terbentuk, namun penentuan jumlah klaster pada K-Means ditentukan secara acak. Penentuan jumlah klaster secara acak dapat menimbulkan masalah sensitivitas, pola yang didapatkan bisa buruk dan begitu pula klastering yang dibuat bisa menghasilkan informasi yang tidak optimal[5].

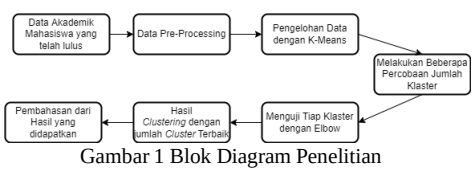
Penelitian yang dilakukan oleh Jannah dan Arifianto di tahun 2007 untuk memprediksi waktu kelulusan mahasiswa menggunakan K-Means menghasilkan akurasi sebesar 59%. Hal ini dapat disebakan oleh penentuan jumlah klaster yang dilakukan secara acak menyebabkan pengelompokan menjadi kurang optimal[6].

Berdasarkan permasalahan yang ada pada penelitian sebelumnya, maka dilakukan penelitian untuk menentukan nilai dari jumlah klaster secara sistematis dengan menggunakan metode Elbow.Hasil dari penelitian ini diharapkan tidak hanya untuk mengelompokan data berdasarkan jarak terdekatnya pada nilai *centroid*, namun juga diharapkan hasil pengelompokan dapat digunakan untuk penentuan waktu kelulusan dari mahasiswaTeknik Informatika.

2. METODE PENELITIAN

1. Alur penelitian

Pada penelitian ini akan digunakan dataset yang berasal dari mahasiswa program studi Informatika dengan menggunakan nilai IPK dan jumlah SKS sebagai parameter yang digunakan. Alur dari penelitian ini dapat dilihat pada gambar 1.

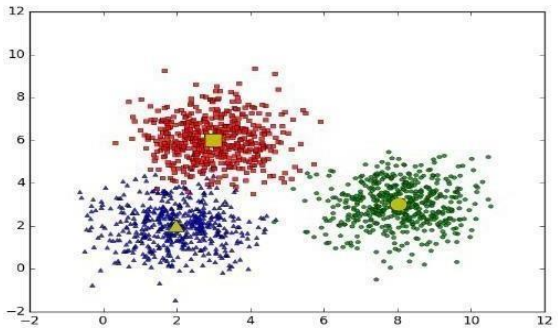


Gambar 1 Blok Diagram Penelitian

Penjelasan dari gambar diatas ialah tahapan awal yaitu pengumpulan data akademik dari Mahasiswa yang telah lulus dari Teknik Informatika melakukan klastering, dalam proses klastering akan dilakukan dalam beberapa jumlah klaster lalu akan diuji tiap jumlah dari klaster dengan menggunakan Elbow. Penelitian ini diharapkan dapat menghasilkan jumlah klaster terbaik yang hasilnya dapat dibahas untuk menemukan informasi yang berguna.

2. Klastering

Klastering merupakan sebuah teknik untuk mengelompokan data ke dalam beberapa klaster sehingga data dalam satu klaster memiliki tingkat kesamaan yang maksimum dan data antar klaster memiliki tingkat kesamaan yang minimum[7]. Klastering membagi data ke beberapa klaster tertentu berdasarkan suatu kemiripan atribut-atribut antar data tersebut. Klastering secara umum dapat dibagi menjadi dua yaitu secara hirarkis dan secara non hirarki[8].



Gambar 2 Contoh penerapan klastering partisi[9]

Pada gambar 2 menunjukkan bagaimana contoh penerapan dari prosedur dari algoritma K-Means non hirarki atau partisi. Pada metode partisi setiap klaster memiliki titik pusat klaster (centroid) dan secara umum metode ini memilki tujuan yaitu mengurangi jarak dari seluruh data ke pusat klaster masing-masing[10].

3. K-Means

K-Means merupakan salah satu metode penganalisaan data yang menggunakan proses pengelompokan data dengan sistem partisi. Metode ini mempartisi data ke dalam bentuk satu atau lebih klaster, sehingga data yang memiliki kesamaan dikelompokkan ke dalam satu klaster yang sama dan data yang memiliki perbedaan dikelompokan ke dalam klaster yang lain [11].

Adapun tahapan algoritma dari penerapan metode K-Means sebagai berikut:

1. Tentukan jumlah klaster yang akan dibentuk
2. Tentukan nilai *centroid* (titik pusat klaster) awal secara acak. Penentuan nilai *centroid* pada iterasi selanjutnya menggunakan rumus pada persamaan (1)

$$\sum^{n} x_i$$

Universitas Khairun, setelah data telah dikumpulkan akan dilakukan Preprocessing yaitu sebuah teknik yang digunakan untuk mengubah data mentah dalam format yang diperlukan dalam penelitian ini. Setelah itu data diolah menggunakan K-Means untuk v = $\frac{\text{i=1}}{n}$: i = 1,2,3, ... n (1)

Keterangan;

v : *Centroid* pada klaster

x_i : Data ke-i

n : Jumlah data per klaster

3. Hitung jarak dari setiap objek ke *centroid* dari masing-masing klaster. Untuk menghitung jarak antara objek dan *centroid* digunakan rumus Euclidean Distance yang dapat dilihat pada persamaan (2).

$$D_eC_i = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \cdots\cdots\cdots\cdots (2)$$

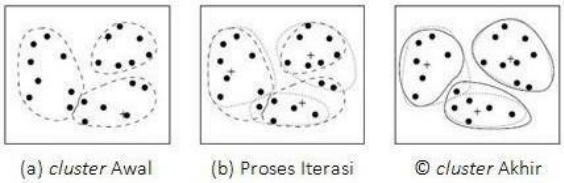
Keterangan:

(x, y) : Koordinat Data

(s, t) : Koordinat *Centroid*

4. Kelompokan tiap objek ke dalam *centroid* yang paling dekat
5. Lakukan iterasi dengan menentukan kembalinilai *centroid* dari setiap klaster dengan menggunakan rumus pada langkah ke-2 persamaan 2.1
6. Ulangi Langkah ke-3 dan ke-4 jika posisi *centroid* baru tidak sama, lakukan iterasi kembali hingga hasil *centroid* tetap dan anggota klaster tidak berpindah ke klaster lain.

Prosedur dari K-Means yang menentukan secara acak menyebabkan hasil klaster yang didapatkan tidak selalu optimal. K-Means membutuhkan jumlah klaster yang tepat karena pusat klaster di awal dapat berubah sehingga proses ini dapat mengakibatkan pengelompokan data yang tidak optimal. Untuk itu dilakukan penelitian ini untuk mengoptimasi dari prosedur kerja K-Means dalam menentukan jumlah klaster. Proses klasterisasi objek menggunakan K-Means dapatdilihat pada gambar 3.



Gambar 3 Proses Klasterisasi K-Means (Han dkk, 2012)

4. Sum of Square (SSE)

$$SSE = \sum_{K=i}^{k} \sum_{K\ i \in S\ k} |x_i - c_k|^2 \cdots\cdots\cdots\cdots (3)$$

1. Keterangan:

K = Klaster ke-i

x_i = Jarak Data ke-i

2. c_k= Pusat Klaster ke-i

Sum of Square Error (SSE) adalah sebuah persamaan yang digunakan untuk mengukur perbedaan dari nilai antar jumlah klaster. SSE sering digunakan sebagai referensi penelitian dalammenentukan klaster yang optimal. Selain untuk mengukur perbedaan antar nilai klaster, SSE juga sebuah validasi dari jumlah klaster yang ditentukan. Validasi dari SSE menggunakan prinsip semakin besar dari nilai jumlah klaster maka nilai dari SSE akan semakin kecil [12].

Untuk menghitung nilai dari SSE dapat digunakan persamaan rumus (3).

Elbow

Metode Elbow adalah metode yang umum dalam menentukan jumlah klaster terbaik, contoh penggunaan metode ialah pada K-Means dan klastering dengan algoritma hirarkis [13]. Metode Elbow merupakan suatu metode yang digunakan dengan melihat persentase dari hasil perbandingan antar nilai dari jumlah klaster yang akan membentuk siku pada suatu titik, hasil persentase yang berbeda dari setiap nilai klaster dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya[14]. Jika nilai klaster kedua dengan nilai klaster ketiga memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai klaster tersebut yang terbaik [15].

Prosedur kerja pada algoritma dari metode Elbow dalam menentukan jumlah klaster terbaik pada K-Means ditunjukkan langkah-langkah berikut ini:

1. Lakukan inisialisasi awal nilai dari jumlah klaster
2. Lakukan proses klastering dalam penelitianini dilakukan menggunakan K-Means.
3. Hitung hasil dari K-Means menggunakan Sum of Square Error. Hitung tiap nilai k.
4. Lakukan presentasi grafik dari hasil perhitungan nilai SSE
5. Tetapkan nilai K yang berhasil membentuk siku sebagai jumlah klaster terbaik.

3. HASIL DAN PEMBAHASAN

1. Analisis data

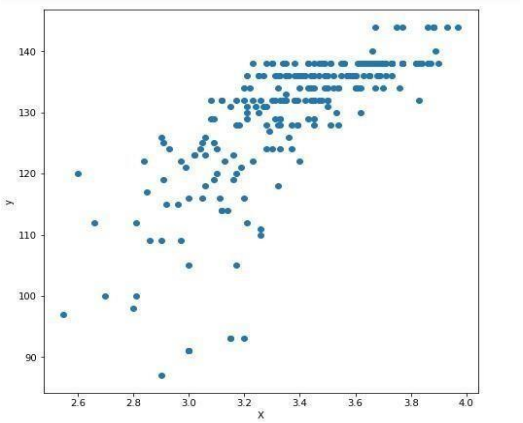
Tahapan awal yang dilakukan dalam penelitian ini ialah pengumpulan data-data yang

	No	Nama	IPK	SKS
0	1	Babullah A. Syah	3.20	134
1	2	Fitriayani Syamsul Idris	3.89	140
2	3	Kartika H Barakati	3.49	138
3	4	Jultana Jusuf	3.56	138
4	5	Ardi Kasim	3.45	134
...	...	...	...	...
235	236	MUHAMMAD RISAL PATTI	2.70	100
236	237	IWAN LA UDIN	3.88	144
237	238	MARTANIA RESKI NOYANDA	3.77	138
238	239	LATOMI	2.60	120
239	240	ABDUL HAFIZH KASTRI CINDARWASIS	3.73	138

240 rows × 4 columns

Gambar 4 *Dataset* Mahasiswa yang Telah Lulus

Data yang digunakan merupakan data yang berasal dari Tata Usaha Fakultas Teknik dengan mengambil nilai IPK dan jumlah SKS pada semester 8 saja. Berikut ini merupakan gambar dari visualisasi persebaran data yang digunakan, persebaran data dapat dilihat pada gambar 5.



Gambar 5 Persebaran Data

2. Implementasi K-Means

Penerapan K-Means pada penelitian ini dilakukan untuk melihat pola yang dihasilkan berdasarkan hasil pengelompokan dari penerapan K-Means. Berikut ini prosedur dari algoritma penerapan K-Means.

1. Penentuan Jumlah *Cluster* dan *Centroid*

Prosedur dari K-Means dalam menentukan jumlah *cluster* yaitu dilakukan dengan acak dan berdasarkan pada kebutuhan dari penelitian yang dilakukan. Dalam penelitian ini akan dibentuk 2 *cluster* dengan tujuan pelabelan akhir sesuai dengan kebutuhan. Jumlah *cluster* yang telah ditentukan selanjutnya ditentukan *centroid* dari masing-masing *cluster* secara acak. Berikut ini ialah nilai *centroid* dari 2 *cluster* yang dapat dilihat pada tabel 1.

Tabel 1 *Centroid* Tiap *Cluster*

Centroid 1	3.45	128
Centroid 2	2.66	112

Penentuan dari *centroid* dalam penelitian ini dilakukan dengan menggunakan

2. Perhitungan Jarak Euclidean

Tahap selanjutnya ialah menghitung jarak dari data ke *centroid* dari masing-masing *cluster* menggunakan perhitungan jarak Euclidean. Berikut ini persamaan untuk menghitung perhitungan jarak dari data ke-1 pada kedua *centroid*:

$$D_eC1 = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

a. Perhitungan Jarak Data ke-1 ke centroid 1

$$D_eC_i = \sqrt{(3{,}20 - 3{,}45)^2 + (134 - 128)^2} = 6{,}005206$$

b. Perhitungan Jarak Data ke-1 ke centroid 2

$$DeC_i = \sqrt{(3{,}20 - 2{,}66)2 + (134 - 112)2} = 22{,}006626$$

Lakukan langkah yang sama pada seluruh *dataset* lalu lakukan pengelompokan berdasarkan jarak terdekatnya. Hasil pengelompokan awal dapat dilihat pada tabel 2.

Tabel 2 Hasil Perhitungan Jarak dan Pengelompokan Data Awal

Data	Jarak C1	Jarak C2	*Cluster*
Data Ke-1	6.005206	22.006626	1
Data Ke-2	12.008064	28.027003	1
Data Ke-3	10.000080	26.013245	1
Data Ke-4	10.000605	26.015572	1
Data Ke-5	6.000000	22.014180	1
. . .	. . .	. . .	. . .
Data Ke-236	28.010043	12.000067	2
Data Ke-247	16.005777	32.023248	1
Data Ke-238	10.005119	26.023683	1
Data Ke-239	8.045030	8.000225	2
Data Ke-240	10.003919	26.022008	1

Jumlah data-data dari pengelompokan awal pada tiap-tiap *cluster* ialah, jumlah data pada *cluster* 1 sebanyak 203 data dan jumlah *cluster* 2 sebanyak 37 data.

3. Perhitungan *Centroid* Baru

Tahap selanjutnya pada K-Means ialah penentuan centroid baru. Proses dilakukan dengan mencari nilai rata-rata berdasarkan jumlah anggota dari tiap *cluster*. Berikut ini contoh perhitungan nilai untuk menghitung nilai *centroid* baru.

a. Perhitungan *centroid* baru untuk *cluster* 1

$$v(x) = \frac{3\,20\,3\,3\,3\,3\,77\,3\,73}{203} = 3$$

$$v(y) = \frac{3\quad 0\quad 3\qquad\qquad 3\quad 3}{203} = 133$$

b. Perhitungan *centroid* baru untuk *cluster* 2

$$v(x) = \frac{3\,3\,0\,3\,7\,3\,2\,2\,7\,2}{37} = 3$$

$$v(y) = \frac{0\qquad\qquad 00\quad 20}{37} = 1\quad 22$$

Proses dari K-Means dapat dikatakan berhenti apabila tidak ada perubahan dari tiap anggota atau data di tiap cluster. Untuk mengetahui hal itu maka dilakukan iterasi lanjutan sesuai denganprosedur dari algoritma K-Means. Proses iterasi dimulai dengan menentukan nilai centroid baru dari tiap cluster yang selanjutnya dilakukan perhitungan seperti dalam algoritma K-Means hingga tidak ditemukan lagi anggota cluster yang berpindah tempat. Untuk melihat centroid akhir dari penelitian ini maka dapat dilihat pada tabel 3.

Tabel 3 *Centroid* Akhir Tiap *Cluster*

Centroid 1	3.46	133.80402
Centroid 2	3.010244	109.17073

4. Pengelompokan Data pada Tiap Cluster Tahap akhir dari K-Means ialah

mengelompokan data berdasarkan jarak terdekat ke *centroid* dari tiap *cluster*. Berikut ini ialah hasil dari 240 data yang telah dikelompokan pada 2 cluster, tabel dari data yang telah dikelompokan dapat dilihat pada tabel 4.

Tabel 4 Pengelompokan Akhir dengan K-Means

Data	Jarak C1	Jarak C2	*Cluster*
Data Ke-1	0.325591	24.829994	1
Data Ke-2	6.210887	30.841819	1
Data Ke-3	4.196091	28.833261	1
Data Ke-4	4.197175	28.834511	1
Data Ke-5	0.196239	24.833163	1
. . .	. . .	. . .	. . .
Data Ke-236	33.812558	9.175977	2
Data Ke-247	10.204631	34.840127	1
Data Ke-238	4.207420	28.839279	1
Data Ke-239	13.830779	10.837037	2
Data Ke-240	4.204662	28.838253	1

Dalam tabel 4.4 didapatkan bahwa beberapa anggota dari cluster dengan menggunakan centroid awal berpindah cluster setelah menggunakan centroid akhir. Pengelompokan ini merupakan pengelompokan akhir dari implementasi K-Means dalam dataset yang digunakan. Hasil yang didapatkan ialah cluster 1 berisikan 199 data dan cluster 2 berisikan 41 data.

3. Pengujian K-Means

Pengujian ini dilakukan untuk menentukan jumlah *cluster* optimal pada *dataset* yang digunakan dalam menggunakan metode K-Means. Pengujian dilakukan pada salah satu *dataset* dengan menggunakan ruang lingkup jumlah *cluster* 1 *cluster* hingga 9 *cluster*. Penentuan jumlah *cluster* serta nilai *centroid* dari tiap *cluster* dalam prosedur K- Means dilakukan secara acak. Berikut ini 9 *centroid* yang digunakan dalam proses pengujian untuk ruang lingkup 9 *cluster* dapat dilihat pada tabel 5.

Tabel 5 *Centroid* Awal untuk 9 *Cluster*

Centroid 1	3.3403637	130.5091
Centroid 2	2.9561539	95.69231
Centroid 3	3.0222223	113.166664
Centroid 4	3.5482142	136
Centroid 5	3.1234782	123.91304
Centroid 6	3.6614287	138.92064
Centroid 7	3.498	133.95
Centroid 8	3.059	119.5
Centroid 9	3.3	136

Langkah selanjutnya ialah melakukan perhitungan jarak dari data ke *centroid* serta mengelompokan berdasarkan jarak terdekat. berdasarkan prosedur K-Means maka setelah melakukan pengelompokan ialah menentukan *centroid* baru dari tiap *cluster*. Berikut ini merupakan perhitungan jarak dari data ke-1 pada pengujian cluster dengan lingkup 1 hingga 9 cluster menggunakan centroid akhir dan pengelompokan data nya, dapat dilihat pada tabel 6.

Tabel 6 Percobaan pada Data ke-1

Centroid	1 *Cluster*	*Cluster*
3.3831666, 129.59584	$(\sqrt{(3{,}20 - 3.3831666)^2} + (134 - 129.59584)^2) = 4.407967$	1
Centroid	2 *Cluster*	
3.46,	$(\sqrt{(3{,}20 - 3.46)^2} +$	1
133.80402	$(134 - 133.80403)^2) = 0.325591$	
3.010244, 109.17073	$(\sqrt{(3{,}20 - 3.010244)^2} + (134 - 109.17073)^2) = 24.829994$	
Centroid	3 *Cluster*	
3.5039773, 135.09659	$(\sqrt{(3{,}20 - 3.5039773)^2} + (134 - 135.09659)^2) = 1.137940$	1
2.9475, 98.1875	$(\sqrt{(3{,}20 - 2.9475)^2} + (134 - 98{,}1875)^2) = 35.813390$	
3.0854166, 119.895836	$(\sqrt{(3{,}20 - 3.0854166)^2} + (134 - 119.89584)^2) = 14.104630$	
Centroid	4 *Cluster*	
3.3131745, 130.01587	$(\sqrt{(3{,}20 - 3.3131745)^2} + (134 - 130.01587)^2) = 3.985738$	4
2.9475, 98.1875	$(\sqrt{(3{,}20 - 2.9475)^2} + (134 - 98{,}1875)^2) = 35.813390$	
3.0839024, 118.902435	$(\sqrt{(3{,}20 - 3.0839024)^2} + (134 - 118.90243)^2) = 15.098011$	
3.58025, 137.21666	$(\sqrt{(3{,}20 - 3.58025)^2} + (134 - 137.21666)^2) = 3.239075$	
Centroid	5 *Cluster*	
3.3405356, 130.55357	$(\sqrt{(3{,}20 - 3.3405356)^2} + (134 - 130.55357)^2) = 3.449291$	4
2.9561539, 95.69231	$(\sqrt{(3{,}20 - 2.9561539)^2} + (134 - 95.69231)^2) = 38.308470$	
3.031154, 115	$(\sqrt{(3{,}20 - 3.031154)^2} + (134 - 115)^2) = 19.000750$	
3.58025, 137.21666	$(\sqrt{(3{,}20 - 3.58025)^2} + (134 - 137.21666)^2) = 3.239057$	
3.1208, 123.68	$(\sqrt{(3{,}20 - 3.1208)^2} + (134 - 123.68)^2) = 10.320304$	
Centroid	6 *Cluster*	
3.3403637, 130.5091	$(\sqrt{(3{,}20 - 3.3403637)^2} + (134 - 130.5091)^2) = 3.493727$	4
2.9822223, 99.55556	$(\sqrt{(3{,}20 - 2.9855556)^2} + (134 - 99.55556)^2) = 34.445131$	
3.0321739, 116.695656	$(\sqrt{(3{,}20 - 3.0321739)^2} + (134 - 116.69565)^2) = 17.305158$	
3.4881034, 135.2931	$(\sqrt{(3{,}20 - 3.4881034)^2} + (134 - 135.2931)^2) = 1.324812$	
3.1234782, 123.91304	$(\sqrt{(3{,}20 - 3.1234782)^2} + (134 - 123.91304)^2) = 10.087250$	
3.6614287, 138.92064	$(\sqrt{(3{,}20 - 3.6614287)^2} + (134 - 138.92064)^2) = 4.942227$	
Centroid	7 *Cluster*	
3.3403637 130.5091	$(\sqrt{(3.20 - 3.3403637)^2} + (134 - 130.5091)^2) = 3.493727$	4
2.9822223, 99.55556	$(\sqrt{(3.20 - 2.9822223)^2} + (134 - 99.55556)^2) = 34.445131$	
3.0321739, 116.695656	$(\sqrt{(3.20 - 3.0321739)^2} + (134 - 116)^2) = 17.305158$	
3.498, 133.95	$(\sqrt{(3.20 - 498)^2} + (134 - 133.95)^2) = 0.302166$	
3.1234782, 123.91304	$\sqrt{(3.20 - 3.1234783)^2} + (134 - 123.91304)^2) = 10.087250$	
3.6614287, 138.92064	$(\sqrt{(3.20 - 3.6614287)^2} + (134 - 138.92064)^2) = 4.942227$	
3.4828947, 136	$(\sqrt{(3{,}20 - 3.4828947)^2} + (134 - 136)^2) = 2.019908$	
Centroid	8 *Cluster*	
3.3403637, 130.5091	$(\sqrt{(3.20 - 3.3403637)^2} + (134 - 130.5091)^2) = 3.493727$	4
2.9822223, 99.55556	$(\sqrt{(3.20 - 2.9822223)^2} + (134 - 99.55556)^2) = 3.445131$	
3.01, 120.4	$(\sqrt{(3.20 - 3.01)^2} + (134 - 120.4)^2) = 13.60137$	
3.498, 133.95	$(\sqrt{(3.20 - 3.498)^2} + (134 - 133.95)^2) 0.302166$	
3.1234782, 123.91304	$(\sqrt{(3.20 - 3.1234782)^2} + (134 - 123.91304)^2) = 10.087250$	
3.6614287, 138.92064	$(\sqrt{(3.20 - 3.6614287)^2} + (134 - 138.92064)^2) = 4.942227$	
3.4828947, 136	$(\sqrt{(3{,}20 - 3.4828947)^2} + (134 - 136)^2) = 2.019908$	

3.0383334, 115.666664	$(\sqrt{(3.20 - 3.0383334)^2 + (134-115.666664)^2}) = 19.000714$	
Centroid	9 *Cluster*	
3.3403637, 130.5091	$(\sqrt{(3.20 - 3.4903637)^2 + (134 - 130.5091)^2}) = 3.493727$	9
2.9822223, 99.55556	$(\sqrt{(3.20 - 2.9822223)^2 + (134 - 99.55556)^2}) = 34.445131$	
3.01, 120.4	$(\sqrt{(3.20 - 3.01)^2 + (134 - 120.4)^2}) = 13.60133$	
3.601, 136	$(\sqrt{(3.20 - 3.601)^2 + (134 - 136)^2}) = 2.039804$	
3.1234782, 123.91304	$(\sqrt{(3.20 - 3.1234782)^2 + (134 - 123.91304)^2}) = 10.087250$	
3.6614287, 138.92064	$(\sqrt{(3.20 - 3.6614287)^2 + (134 - 138.92064)^2}) = 4.942227$	
3.3516667, 136	$(\sqrt{(3,20 - 3.516667)^2 + (134 - 136)^2}) = 2.005742$	
3.0383334, 115.666664	$(\sqrt{(3.20 - 3.0383334)^2 + (134-115.666664)^2}) = 18.334049$	
3.498, 133.95	$(\sqrt{(3.20 - 3.498)^2 + (134 - 133.95)^2}) = 0.302166$	

Hasil perhitungan jarak terdekat serta pengelompokan dari pengujian ini menunjukkan bahwa data ke-1 dapat berada dalam 3 *cluster* yang berbeda bergantung pada jumlah *cluster* yang digunakan. Untuk mengetahui letak optimal dari data ke-1 atau dalam hal ini untuk mengetahui letak tiap data dengan optimal maka perlu ditentukan jumlah cluster terbaik. Penentuan dari jumlah clusterdalam penelitian ini akan dilakukan dengan menggunakan metode Elbow.

4. Implementasi Elbow

Penerapan menggunakan metode Elbow dimulai setelah proses klastering dengan metode K-Means selsai diterapkan. Tujuan penggunaan Elbow ialah untuk menentukan jumlah *cluster* yang optimal dengan melakukan evaluasi pada tiap *cluster*. Berikut ini ialah algoritma penerapan Elbow pada *dataset* yang digunakan.

1. Perhitungan SSE Tiap *Cluster*

Proses evaluasi dari tiap *cluster* yang dibentuk dilakukan dengan menggunakanperhitungan Sum of Square Error (SSE). Perhitungan nilai SSE dilakukan dari *cluster* 1 hingga *cluster* 9, berikut ini perhitungan SSE pada 1 *cluster* dan 9 *cluster*.

A. Perhitungan SSE pada 1 *cluster*

$$SSE = |4{,}407967 - 8{,}259955|^2 + \cdots |8{,}411313 - 8{,}259955|^2 = 13662{,}571630$$

B. Perhitungan SSE pada 9 *cluster*

$$SSE = |0{,}302166 - 1{,}449602|2 + \cdots |4{,}056643 - 1{,}449602|2 = 802{,}826792$$

Hasil keseluruhan perhitungan nilai SSE dari lingkup jumlah *cluster* 1 hingga *cluster* 9 dapat dilihat pada tabel.

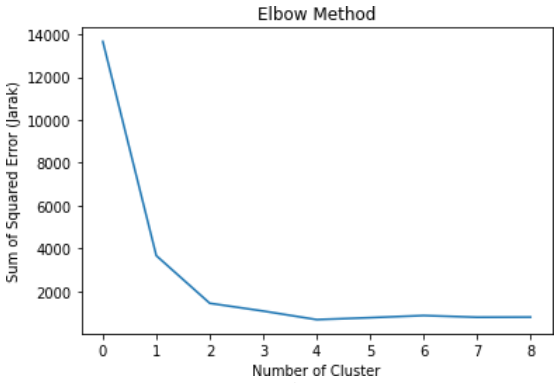
Tabel 6 Hasil Perhitungan SSE Tiap *Cluster*

Jumlah *Cluster*	Hasil Perhitungan SSE	Selisih
1	13662.57163	-
2	3670.325734	9992.245896
3	1448.527237	2221.798497
4	1084.981685	363.545552
5	685.5825526	399.399133
6	775.6868077	-90.104255
7	875.5996207	-99.912813
8	796.0027336	79.596887
9	802.8267916	-6.824058

Prosedur yang digunakan pada SSE ialah semakin kecil nilai SSE yang didapat, semakin sejenis data dari tiap-tiap *cluster* maka semakin baik *cluster* yang dihasilkan [16].

2. Grafik Metode Elbow

Grafik dari metode Elbow merupakan langkah terakhir untuk melihat jumlah *cluster* terbaik dari *dataset* yang digunakan. Grafik yangterbentuk merupakan hasil dari perhitungan nilai SSE dari tiap *cluster*, representasi grafik dapat dilihat pada gambar 6.



Gambar 6 Grafik Elbow

Garis siku yang terbentuk merupakan garis yang dihasilkan dari selisih terbesar dari perhitungan nilai SSE pada lingkup jumlah *cluster* 1 hinggan *cluster* 9. Dengan garis siku yang terbentuk maka dapat disimpulkan bahwa *cluster* yang optimal pada *dataset* ini ialah sebanyak 2 *cluster*.

4. KESIMPULAN

1. Hasil analisis dan penerapan metode menunjukkan bahwa menggunakan metode Elbow dapat digunakan untuk menentukan jumlah *cluster* terbaik pada metode K-Means dengan studi kasus *cluster*ing data mahasiswa yang telah lulus
2. Penelitian menghasilkan 2 *cluster* terbentuk berdasarkan penerapan metode pada data dengan menggunakan dua atribut yaitu, nilai IPK dan jumlah SKS.
3. Pembentukan siku pada grafik Elbow dihasilkan berdasarkan perhitungan SSE dari tiap jumlah *cluster* dengan selisih terbesar terbentuk di antara jumlah 1 *cluster* dan 2 *cluster*, dengan nilai selisih sebesar 9992.245896
4. Nilai rata-rata yang dihasilkan dari *cluster* 1 ialah sebesar 3.46 untuk atribut IPK dan

133.804020 untuk atribut SKS lalu nilai rata-rata dari *cluster* 2 untuk kedua atribut ialah 3.010244 dan 109.170732. Hasil *cluster*ing data dari penelitian ini dapat dilihat pada gambar

5. DAFTAR PUSTAKA

[1] Waworuntu, M. N. V., & Amin, M. F. (2018). Penerapan Metode K-Means Untuk Pemetaan Calon Penerima Jamkesda. KLIK-Kumpulan Jurnal Ilmu Komputer, 5(2), 190-200. (http://dx.doi.org/10.20527/klik.v5i2.157) diakses 12 januari 2022.

[2] Trayasiwi, G. P. (2017). Penerapan Metode Klastering dengan Algoritma k-Means Untuk Prediksi Kelulusan Mahasiswa Pada Program Studi Teknik Informatika Strata Satu. UDiNus Repos, 1(1), 1-11.

[3] KeyUA, 2021. "10 Popular Data Mining Algorithms". Tersedia [https://keyua.org/blog/top-10-data-mining-algorithms/] diakses 20 Desember 2021.

[4] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.(10.1088/1757-899X/336/1/012017) diakses 12 januari 2022.

[5] Dubey, A. K., Gupta, U., & Jain, S. (2018). Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology, 8(1), 18-29. (http://dx.doi.org/10.18517/ijaseit.8.1.3490) diakses tanggal 12 januari 2022.

[6] Jannah, A. R., Arifianto, D., & Kom, M. 2017. "Penerapan Metode Clustering dengan Algoritma K-Means untuk Prediksi Kelulusan Mahasiswa Jurusan Teknik Informatika di Universitas Muhammadiyah Jember". Jurnal Manajemen Sistem Informasi an Teknologi, 1(1210651237), 1–10. <http://repository.unmuhjember.ac.id/id/eprint/589>

[7] Studio, M. (2018). Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma K-Means. vol. IV, (2), 401-411.

[8] Aditya, K. B., Puspitaningrum, D., & Setiawan, Y. (2017). Sistem Informasi Geografis Pemetaan Faktor-Faktor Yang Mempengaruhi Angka Kematian Ibu (Aki) Dan Angka Kematian Bayi (AKB) Dengan Metode K-Means Clustering (Studi Kasus: Provinsi Bengkulu). Jurnal Teknik Informatika UIN Syarif Hidayatullah, 10(1), 133712. (http://dx.doi.org/10.15408/jti.v10i1.6817) diakses 21 januari 2022.

[9] Duong, M. Q., Lam, B. L. H., Tu, G. Q. H., & Hieu, N. H. (2019). Combination of K- Mean clustering and elbow technique in mitigating losses of distribution network. GMSARN International, 13, 153-158.

[10] 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. (http://dx.doi.org/10.26418/jp.v5i1.29611) diakses 21 januari 2022.

[11] Eldo, H. (2020). Penentuan Cluster TerbaikK-Means Menggunakan Algoritma Silhouette (Doctoral dissertation, Universitas SumateraUtara) http://repositori.usu.ac.id/handle/123456789/ 27537.

[12] Santoso, T., & Saftarina, F. 2020. "Klasterisasi Petani Padi Sawah di Kota Metro Provinsi Lampung MenggunakanAlgoritma K-Means Cluster dan ElbowMethod". Journal of Agribusiness andCommunity Empowerment, 3(1), 37–43. http:doi.org/2655-4526/26552965/jace/11.10.2019.

[13] Putu, N., Merliana, E., & Santoso, A. J. (2015). "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means". 978–979. https://doi.org/10.22146/ijccs.6641

[14] Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. Matrix: Jurnal Manajemen Teknologi Dan Informatika, 9 (3), 102–109 http://dx.doi.org/10.31940/matrix.v9i3.1662.

[15] Nainggolan, R., Perangin-Angin, R., "Simarmata, E., & Tarigan, A. F. 2019. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method". Journal of Physics: Conference Series, 1361(1). https://doi.org/10.1088/1742-6596/1361/1/012015.

[16] Qi, J., Yu, Y., Wang, L., & Liu, J. (2016, October). K-means: An effective and efficient K-means clustering algorithm. In 2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom) (pp. 242-249). IEEE 10.1109/BDCloud-SocialCom-SustainCom.2016.46.