

## Analisis Keberhasilan Studi Awal Mahasiswa Menggunakan Klasterisasi K-Means

Painem<sup>1</sup>, Hari Soetanto<sup>2</sup>, Achmad Solichin<sup>3</sup>

<sup>1,2,3</sup>Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia

---

### Article Info

#### Article history:

Received Aug 2, 2023

Revised Oct 04, 2023

Accepted Oct 25, 2023

---

#### Keywords:

Success analysis

Student

Clustering

K-Means

---

### ABSTRACT

Students are an important elements in studying at university. Every student who studies at a university certainly wants to graduate on time and fulfill optimal academic qualifications. Likewise, for programs study and universities, the success of student studies is an important indicator of the success of providing education in higher education. The success analysis of the student study should be carried out periodically from the start of the study to the end of the study. The results of the study success analysis can be used as a basis for decision making and evaluation of learning programs for programs study and universities. However, analyzing the success of student studies at a university with a large number of students is sometimes difficult and quite complicated. University managers and/or study programs often have difficulty preparing learning programs that are right on target for students in order to produce graduates who have academic abilities. optimal and graduate on time. To assist the director of the study program in analyzing the success of the initial studies using the K-Means clustering method. Based on the analysis of the success of students' initial studies using K-Means clustering, 22.6% or 3055 students entered cluster 0, while 69.5% or 9405 students entered cluster 1 and 9405 students entered cluster 2. are 7.9% or 1066 students. These results are useful for programs to analyze student study success.

Copyright © 2023 Universitas Indraprasta PGRI.  
All rights reserved.

---

### Corresponding Author:

Painem,

Teknik Informatika,

Universitas Budi Luhur,

Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan

Email: painem@budiluhur.ac.id

---

### 1. PENDAHULUAN

Perguruan Tinggi di Indonesia merupakan institusi pendidikan tinggi yang berfungsi sebagai pusat pengembangan sumber daya manusia dan penelitian ilmiah. Saat ini, terdapat lebih dari 4.000 perguruan tinggi di Indonesia yang terdiri dari universitas, institut, sekolah tinggi, dan politeknik. Universitas adalah jenis perguruan tinggi yang paling populer di Indonesia dan biasanya menawarkan berbagai program studi di berbagai bidang, seperti teknik informatika, sistem informasi, akuntansi, manajemen, ilmu komunikasi, arsitektur, dan lain sebagainya. Selain itu, terdapat juga perguruan tinggi swasta yang menawarkan program studi dengan biaya yang lebih terjangkau dibandingkan dengan perguruan tinggi negeri.

Meskipun terdapat banyak perguruan tinggi di Indonesia, kualitas pendidikan di beberapa perguruan tinggi masih perlu ditingkatkan. Saat ini, terdapat beberapa perguruan tinggi yang telah meraih akreditasi unggul, baik sekali, atau baik dari Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT), yang menunjukkan kualitas pendidikan dan pengajaran yang baik. Namun, masih terdapat perguruan tinggi yang belum memenuhi standar kualitas yang ditetapkan. Mahasiswa merupakan salah satu elemen penting dalam perkuliahan di perguruan tinggi. Mereka berperan sebagai pelaku utama dalam proses belajar mengajar, dengan mengikuti

perkuliahan dan mengerjakan tugas-tugas yang diberikan oleh dosen. Dengan mengoptimalkan peran mereka, mahasiswa dapat mengembangkan keterampilan, pengetahuan, dan wawasan yang dapat membantu mereka dalam mempersiapkan diri menghadapi tantangan di masa depan.

Perguruan Tinggi perlu melakukan evaluasi secara berkala untuk membantu program studi dalam memonitor kemajuan akademik mahasiswa, serta membantu mahasiswa dalam memperbaiki prestasi akademik dan meningkatkan tingkat kelulusan mahasiswa. Dengan demikian, evaluasi studi mahasiswa secara berkala sangat penting dilakukan untuk memastikan bahwa kualitas pendidikan di perguruan tinggi terus meningkat dan memenuhi standar yang ditetapkan.

Setiap mahasiswa yang menempuh kuliah di suatu perguruan tinggi tentunya menginginkan dapat lulus tepat waktu dengan memenuhi kualifikasi akademik yang optimal. Demikian juga bagi pihak program studi dan universitas, keberhasilan studi mahasiswa merupakan salah satu indikator penting dalam keberhasilan penyelenggaraan pendidikan di perguruan tinggi. Analisis keberhasilan studi mahasiswa seharusnya dilakukan secara berkala mulai dari awal studi hingga akhir studi. Hasil analisis keberhasilan studi dapat dijadikan dasar dalam pengambilan keputusan dan evaluasi program pembelajaran bagi program studi maupun universitas

Pentingnya analisis keberhasilan studi mahasiswa terletak pada kemampuannya untuk memperbaiki dan meningkatkan kualitas pendidikan dan pengajaran di perguruan tinggi. Dengan melakukan analisis yang tepat, perguruan tinggi dapat menentukan program-program pembelajaran yang sesuai dengan kebutuhan mahasiswa, meningkatkan efektivitas pengajaran, serta membantu mahasiswa dalam memperbaiki prestasi akademik mereka. Selain itu, analisis keberhasilan studi mahasiswa juga membantu perguruan tinggi dalam mengevaluasi kebijakan-kebijakan pendidikan yang telah diterapkan dan melakukan perbaikan jika diperlukan

Namun demikian, melakukan analisis keberhasilan studi mahasiswa pada sebuah perguruan tinggi dengan jumlah mahasiswa yang cukup banyak terkadang sulit dilakukan dan cukup rumit. Pengelola universitas dan/atau program studi seringkali kesulitan dalam menyusun program pembelajaran yang tepat sasaran bagi mahasiswa dalam rangka menghasilkan lulusan yang memiliki kemampuan akademik yang optimal dan lulus tepat waktu. Jika pengelola universitas atau program studi disajikan sebuah sistem yang mampu menganalisis data mahasiswa secara cepat, tepat dan efektif tentunya akan sangat membantu dalam proses pengambilan berbagai keputusan. Dalam hal ini, teknologi informasi dapat menjadi solusi untuk membantu perguruan tinggi melakukan analisis keberhasilan studi mahasiswa secara lebih efektif dan efisien. Dengan menggunakan sistem informasi yang tepat, perguruan tinggi dapat memperoleh informasi yang akurat dan cepat sehingga dapat dijadikan dasar dalam pengambilan keputusan yang berkaitan dengan peningkatan kualitas pendidikan dan pengajaran.

Untuk menyelesaikan permasalahan diatas maka dibutuhkan sebuah metode. Salah satu metode yang akan digunakan adalah Klusterisasi. Klusterisasi sekumpulan data (*object*) yang memiliki kesamaan (*similarity*) diantara setiap anggota klaster, atau ketidaksamaan (*dissimilarity*) dengan data pada klaster yang lain. Salah satu metode Klusterisasi yang akan digunakan pada penelitian ini adalah Algoritma K-Means.

Penelitian ini dapat bermanfaat bagi pengelola program studi untuk membantu menganalisa keberhasilan studi mahasiswa dari awal semester sampai akhir semester sehingga pengelola program studi mempunyai strategi awal untuk mengantisipasi perkembangan studi mahasiswa sejak awal semester dan tidak terlambat dalam melakukan penanganan. Penelitian terkait dengan klusterisasi dalam bidang pendidikan khususnya perguruan tinggi menggunakan K-means untuk klusterisasi mahasiswa *drop out* dilakukan oleh [1]. Selanjutnya klusterisasi menggunakan k-means untuk menentukan besaran uang kuliah tunggal [2]. Dalam bidang ekonomi klusterisasi untuk potensi penjualan Linkaja sudah dilakukan oleh [3]. Klustering untuk memprediksi waktu kelulusan mahasiswa dilakukan oleh [4] dan klustering untuk menentukan strategi promosi oleh [5].

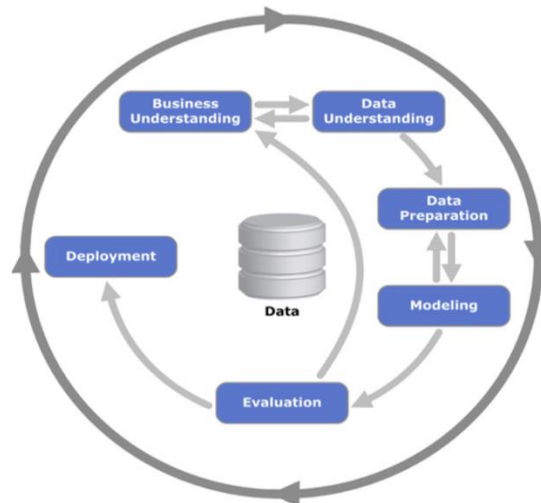
## 2. METODE

### 2.1. K-Means

Algoritma K-Means adalah salah satu algoritma *clustering* yang bersifat *iteratif* yang mencoba untuk mempartisi *dataset* menjadi subkelompok *non-overlapping* berbeda yang ditentukan oleh K (*cluster*) yang mana setiap titik data hanya dimiliki oleh satu kelompok. [6]

## 2.2. Cross-Industry Standard Process for Data Mining (CRISP-DM)

Dalam melaksanakan penelitian, diperlukan tahapan-tahapan yang terstruktur agar tujuan penelitian dapat tercapai dengan baik. Tahapan-tahapan penelitian tersebut menjadi acuan dalam melaksanakan penelitian. Pada penelitian ini akan menggunakan metode penelitian CRISP-DM. CRISP-DM adalah suatu kerangka kerja atau *framework* yang digunakan oleh para praktisi *data mining* untuk mengatasi permasalahan dalam proyek *data mining*. Kerangka kerja ini terdiri dari enam tahapan yang harus dilalui untuk menyelesaikan suatu proyek data mining secara efektif dan efisien.



Gambar 1. Tahapan CRISP-DM [7]

Berikut adalah penjelasan singkat tentang masing-masing tahap CRISP-DM [8][7]:

1. Tahap Bisnis Understanding (*Business Understanding*)  
Tahap pertama adalah memahami masalah bisnis yang akan diselesaikan dan tujuan dari penelitian data mining yang akan dilakukan. Pada tahap ini, peneliti data mining harus berdiskusi dengan stakeholder untuk memahami masalah yang dihadapi, tujuan penelitian, serta memperkirakan dampak bisnis yang akan dihasilkan. Masalah yang dihadapi pada penelitian ini adalah pengelola program studi belum menganalisa studi mahasiswa setiap semesternya. Dampak bisnis yang diharapkan dari penelitian ini adalah penelitian ini diharapkan dapat menganalisa keberhasilan mahasiswa sejak awal studi.
2. Tahap Pemahaman Data (*Data Understanding*)  
Tahap kedua adalah memahami data yang tersedia dan mengevaluasi kualitas data. Pada tahap ini, peneliti harus mempelajari data yang tersedia, mengevaluasi kualitas data, dan menentukan metode untuk mengintegrasikan data.
3. Tahap Persiapan Data (*Data Preparation*)  
Tahap ketiga adalah menyiapkan data untuk pemodelan. Pada tahap ini, peneliti harus membersihkan dan mentransformasi data menjadi format yang sesuai untuk analisis data mining.
4. Tahap Pemodelan (*Modeling*)  
Tahap keempat adalah membuat model yang dapat menyelesaikan masalah bisnis yang telah didefinisikan pada tahap bisnis understanding. Pada tahap ini, peneliti akan memilih teknik analisis data mining yang tepat dan membangun model yang dapat menghasilkan prediksi atau klasifikasi yang akurat. Metode *Data Mining* yang melakukan proses pemodelan *un-supervised learning* adalah metode K-Means. Tahap pertama pada metode K-Means adalah menentukan nilai K atau menentukan jumlah kluster (K), tahap kedua menginisialisasi centroid awal. Tahap ketiga menghitung jarak setiap titik data ke masing-masing setiap centroid. Metode yang akan digunakan untuk menghitung jarak setiap titik data ke masing – masing setiap centroid adalah *Euclidean Distance*. Perhitungan *Euclidean Distance* dengan menggunakan rumus sebagai berikut[9] :

$$d_{x,y} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Keterangan :

$d_{x,y}$  = jarak antara data x dan y

n= jumlah atribut

$x_i$  =obyek x ke - i

$y_i$  =obyek y ke - i

Tahap selanjutnya Masukkan setiap titik data ke dalam kluster berdasarkan jarak terdekat dengan pusat kluster . Dan tahap berikutnya untuk setiap kluster, tentukan nilai centroid baru berdasarkan rata-rata (*means*) dari setiap data di dalam kluster atau menghitung centroid berikutnya.

#### 5. Tahap Evaluasi (*Evaluation*)

Pada tahap evaluasi untuk menentukan optimasi jumlah cluster pada penelitian ini menggunakan salah satu metode analisis cluster yaitu metode Elbow, dengan memperhatikan nilai perbandingan (dari perhitungan SSE (Sum of Square Error) untuk setiap nilai cluster) antara jumlah cluster yang akan membentuk siku pada suatu titik, Semakin besar nilai SSE, semakin berkurang kualitas cluster, begitu sebaliknya, Semakin kecil nilai SSE, semakin baik kualitas cluster [10].

Rumus perhitungan nilai SSE sebagai berikut :

$$SSE = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2)$$

#### 6. Tahap Penyampaian (*Deployment*)

Tahap terakhir adalah mengimplementasikan model dan membuat rekomendasi berdasarkan hasil analisis *data mining*. Pada tahap ini, peneliti akan menyiapkan laporan hasil analisis data mining dan memberikan rekomendasi kepada *stakeholder*.

### 3. HASIL DAN PEMBAHASAN

#### 3.1 Tahap Bisnis Understanding (*Business Understanding*)

Masalah yang dihadapi pada penelitian ini adalah pengelola program studi belum menganalisa studi mahasiswa empat semester yaitu semester satu, semester dua, semester tiga dan semester empat. Dengan melakukan analisa selama empat semester pengelola program studi dapat memantau perkembangan studi mahasiswa sejak dini dan memberikan peringatan dini kepada mahasiswa. Penelitian ini diharapkan dapat menganalisa keberhasilan mahasiswa sejak awal studi.

#### 3.2 Tahap Pemahaman Data (*Data Understanding*)

Tahap kedua adalah memahami data yang tersedia dan mengevaluasi kualitas data. Data-data dikumpulkan dari data mahasiswa dengan atribut sebagai berikut

- 1) CNIM : NIM
- 2) (CNAMA) : Nama mahasiswa
- 3) CSTATUS : status mahasiswa ((karyawan atau reguler)
- 4) CSMTAWAL : semester (Gasal/Genap)
- 5) CTH AJARAWAL Tahun ajaran
- 6) IPS1 : Indeks Prestasi Semester 1
- 7) IPS2 : Indeks Prestasi Semester 2
- 8) IPS3 : Indeks Prestasi Semester 3
- 9) IPS4 : Indeks Prestasi Semester 4
- 10) SKS1SD4 : total SKS yang didapatkan sampai dengan semester 4
- 11) IPK : Indeks Prestasi Kumulatif

#### 3.3 Tahap Persiapan Data (*Data Preparation*)

Tahap ketiga adalah menyiapkan data untuk pemodelan. Data yang akan digunakan adalah 13530 mahasiswa. Ada beberapa tahap untuk persiapan data sebagai berikut :

- 1) Menampilkan data yang nilainya null.

CNIM	4
CNAMA	4
STATUS	4
CSMTAWAL	4
CTHAJARAWAL	4
IPS1	247
IPS2	43
IPS3	120
IPS4	152
SKS1SD4	1
IPK	1
dtype:	int64

Gambar 2. Tampilan Atribut yang nilainya Null

Pada Gambar 2 untuk atribut Cnim terdapat 4 data yang nilainya Null, Cnama ada 4 data yang nilainya Null, Cnama ada 4 data yang nilainya Null, Status ada 4 data yang nilainya Null, Csmताल ada 4 data yang

nilainya Null, Cthajarawal ada 4 data yang nilainya Null, IPS1 terdapat 247 data yang nilainya Null, IPS2 ada 43 data yang nilainya Null, IPS3 ada 120 data yang nilainya NULL, IPS4 ada 152 data yang nilainya Null dan SKS1SD4 ada 1 data yang nilainya Null dan IPK ada 1 data yang nilainya Null.

2) Menghapus Data yang nilainya Null

Pada tahap ini melakukan penghapusan data yang nilainya null. Khusus untuk atribut IPS1, IPS2, IPS3, dan IPS4 tidak perlu dihilangkan karena seorang mahasiswa memungkinkan untuk cuti/tidak mengambil mata kuliah pada semester tertentu antara semester 1 sampai dengan semester 4.

3) Penghilangan Kolom yang tidak diperlukan

Pada tahap ini, kita akan menghilangkan beberapa kolom, karena tidak terlalu diperlukan dalam proses analisis data:

1) Kolom "STATUS" yang berisi status mahasiswa apakah kelas pagi (reguler) atau malam (karyawan). Kolom ini dihapus karena bukan menjadi fokus analisis.

2) Kolom "CSMTAWAL" yang berisi semester awal masuk mahasiswa.

3) Kolom "IPK" berisi IPK akhir mahasiswa. Kolom ini tidak diperlukan karena tujuan utama analisis adalah untuk menganalisis studi awal mahasiswa

4) Cek data duplikasi

Setelah dilakukan pengecekan duplikasi maka hasil yang didapatkan adalah 0. Artinya tidak terdapat duplikasi data

5) Statistik Deskriptif Data

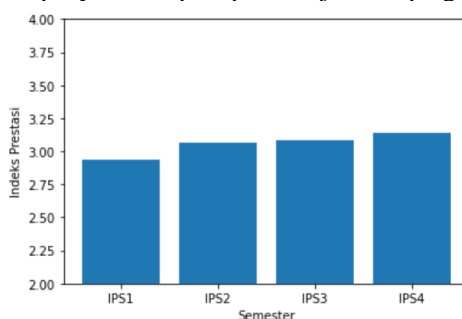
	IPS1	IPS2	IPS3	IPS4	SKS1SD4
count	13526.000000	13526.000000	13526.000000	13526.000000	13526.000000
mean	2.937608	3.065516	3.086825	3.140556	83.206417
std	0.671509	0.606884	0.666287	0.670487	11.380821
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	2.640000	2.760000	2.810000	2.880000	81.000000
50%	3.000000	3.160000	3.220000	3.290000	87.000000
75%	3.390000	3.500000	3.540000	3.600000	91.000000
max	4.000000	4.000000	4.000000	4.000000	118.000000

Gambar 3 Statistik Deskriptif Data Mahasiswa

Berdasarkan Gambar 3, terlihat untuk atribut IPS1, IPS2, IPS3, IPS4 telah VALID (tidak mengandung *outlier*), karena memiliki rentang antara 0 - 4.0. Semua data (*count*) sudah terisi atau tidak ada data kosong, Rata-rata (Mean) IPS semester 1 adalah 2,93, IPS semester 2 adalah 3,08, semester 2 adalah 3,08 dan semester 4 adalah 3,00. Pada Quartil pertama (Q1) IPS semester 1 adalah 2,65.

6) Rerata IPS Mahasiswa dalam 4 Semester Awal.

Berdasarkan data dan grafik rerata IPS mahasiswa pada Gambar 4, terlihat bahwa selama 4 (empat) semester awal, IPS mahasiswa terus mengalami peningkatan. Hal tersebut dapat dipahami karena mahasiswa baru memerlukan penyesuaian pola pembelajaran di perguruan tinggi.

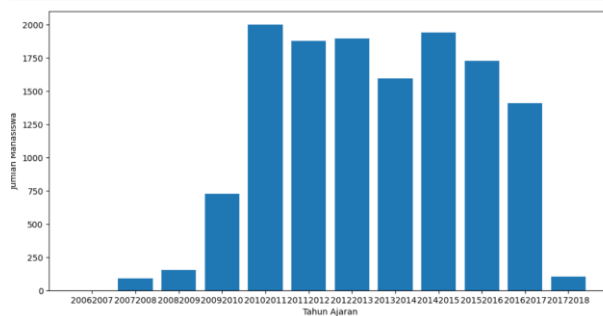


Gambar 4. Grafik Rerata IPS Mahasiswa dalam 4 semester Awal

7) Data Per Angkatan

Data mahasiswa baru pada tahun ajaran 2007/2008 sebanyak 89 mahasiswa, tahun ajaran 2008/2009 sebanyak 152 Mahasiswa, tahun ajaran 2009/2010 sebanyak 728 Mahasiswa, tahun ajaran 2010/2011 sebanyak 2002 Mahasiswa, tahun ajaran 2011/2012 sebanyak 1877 mahasiswa, tahun ajaran 2012/2013 sebanyak 1898 Mahasiswa, tahun ajaran 2013/2014 sebanyak 1597 mahasiswa, tahun ajaran 2014/2015 sebanyak 1940 mahasiswa, tahun ajaran 2015/2016 sebanyak 1727 mahasiswa, tahun 2016/2017 sebanyak 1411 mahasiswa,

tahun ajaran 2017/2018 sebanyak 104 mahasiswa. Berdasarkan data tersebut jika dibuatkan grafik seperti pada Gambar 5.



Gambar 5. Grafik data mahasiswa per angkatan

### 8) Analisa Korelasi

Analisa korelasi digunakan untuk mengetahui seberapa penting antar atribut yang akan digunakan. Dari Atribut IPS1, IPS2, IPS3, IPS4 dan SKS1SD4 maka dapat dilihat korelasi seperti pada gambar 6:

	IPS1	IPS2	IPS3	IPS4	SKS1SD4
IPS1	1.000000	0.467050	0.430552	0.379789	0.345803
IPS2	0.467050	1.000000	0.650377	0.560997	0.305139
IPS3	0.430552	0.650377	1.000000	0.612420	0.361452
IPS4	0.379789	0.560997	0.612420	1.000000	0.352101
SKS1SD4	0.345803	0.305139	0.361452	0.352101	1.000000

Gambar 6. Korelasi Antar Atribut

Pada Gambar 6 hasil output korelasi menunjukkan bahwa Kolom IPS1 memiliki hubungan kuat dengan kolom IPS2, IPS3 dan IPS4. Kolom IPS2 memiliki korelasi kuat dengan IPS1, IPS3, IPS4, dan IPS4. Kolom SKS1SD4 yang memiliki korelasi lemah dengan semua kolom lainnya. Nilai korelasi dengan hubungan kuat paling tinggi antar kolom adalah kolom IPS2 dengan IPS3

### 9) Penambahan Atribut "Skewness"

**Skewness** diartikan sebagai kemiringan distribusi data. Saat **skewness** positif maka distribusi lebih panjang ke sisi kanan kurva. Jika **skewness** antara -0,5 dan 0,5, data cukup simetris. Jika kemiringannya antara -1 dan 0,5 atau antara 0,5 dan 1, datanya agak miring. Jika kemiringannya kurang dari -1 atau lebih besar dari 1, datanya sangat miring. Pada Gambar 6 disajikan tampilan statistik deskriptif data setelah di ditambahkan atribut **skewness** :

	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS
count	13526.000000	13526.000000	13526.000000	13526.000000	13526.000000	13526.000000
mean	2.937608	3.065516	3.086825	3.140556	83.206417	-0.279227
std	0.671509	0.606884	0.666287	0.670487	11.380821	1.065711
min	0.000000	0.000000	0.000000	0.000000	3.000000	-2.000000
25%	2.640000	2.760000	2.810000	2.880000	81.000000	-1.160807
50%	3.000000	3.160000	3.220000	3.290000	87.000000	-0.265374
75%	3.390000	3.500000	3.540000	3.600000	91.000000	0.448458
max	4.000000	4.000000	4.000000	4.000000	118.000000	2.000000

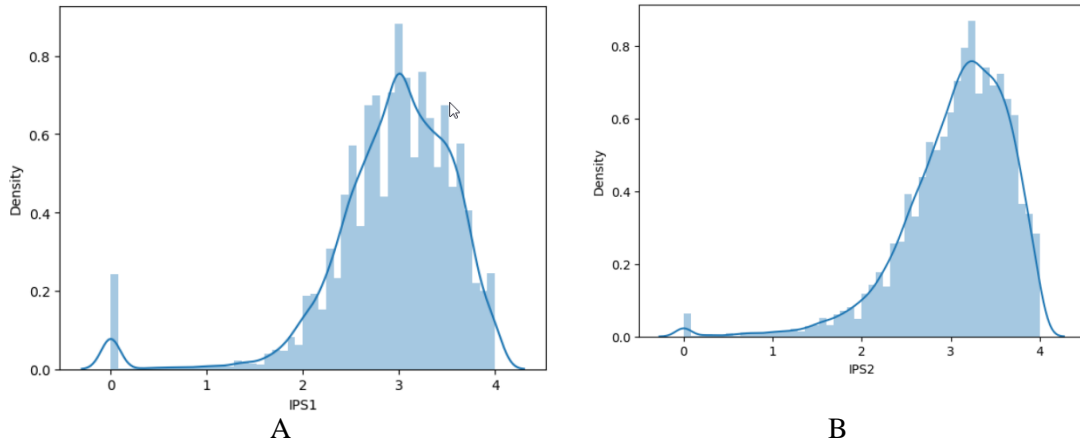
Gambar 7. Tampilan statistik deskriptif data setelah ditambahkan atribut **skewness**

### 10) Analisis **Skewness**

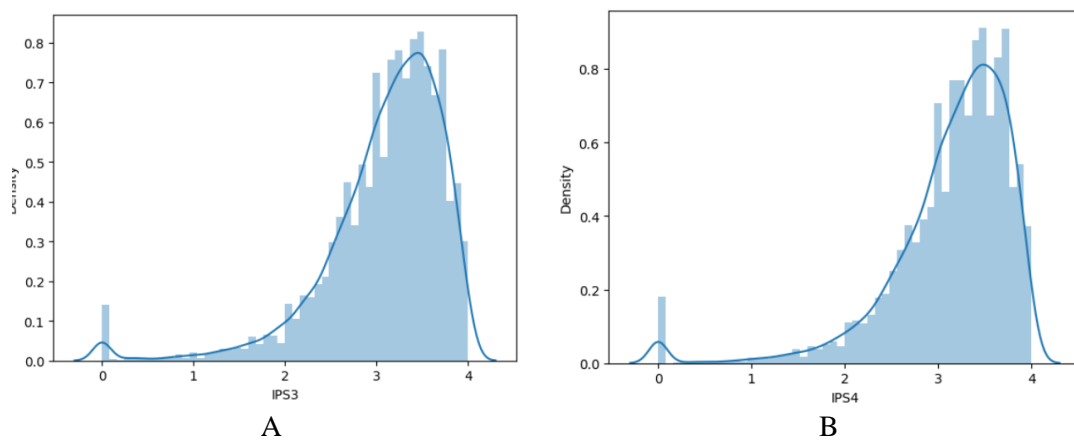
Pada Gambar 8 bagian A grafik nilai **skewness** IPS1 terlihat kemiringan kurva cenderung ke kanan artinya memiliki kemiringan positif. Sedangkan Gambar 7 bagian B. Pada Gambar 8 menunjukkan grafik nilai **skewness** IPS1 terlihat kemiringan kurva cenderung ke kanan artinya memiliki kemiringan positif.

Pada Gambar 9 A dan B menunjukkan grafik nilai **skewness** IPS1 terlihat kemiringan kurva cenderung ke kanan artinya memiliki kemiringan positif





Gambar 8. A Grafik nilai *Skewness* IPS semester 1 dan B Grafik nilai *Skewness* IPS2 (IPS Semester 2)



Gambar 9. A Grafik nilai *Skewness* IPS Semester 3 dan B Grafik nilai *Skewness* IPS Semester 4

3.4 Tahap Pemodelan (*Modeling*)

Tahap keempat adalah membuat model yang dapat menyelesaikan masalah bisnis yang telah didefinisikan pada tahap bisnis understanding. Pada tahap ini, peneliti akan memilih teknik analisis data mining yang tepat dan membangun model yang dapat menghasilkan prediksi atau klasifikasi yang akurat. Metode *Data Mining* yang melakukan proses pemodelan *un-supervised learning* adalah metode K-Means. Pada tahap pemodelan dengan menggunakan contoh data sebanyak 10 data yang disajikan pada Tabel 1:

Tabel 1. Data Mahasiswa Semester 1 sampai dengan semester 4

CNIM	CNAMA	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS
0932510999	A	3.12	3.21	2.75	2.70	63.0	0.072167
0932510999	B	3.47	3.33	4.00	3.29	58.0	1.688810
0932510999	C	3.29	3.25	3.29	3.00	68.0	-1.894787
0931511999	D	3.82	3.40	3.33	3.42	68.0	1.814372
0932511999	E	3.56	3.50	3.23	3.13	83.0	-0.125171
0932500999	F	2.79	3.00	2.71	2.38	93.0	-0.663766
0931511999	G	3.38	3.13	3.22	3.00	83.0	0.251165
1032500999	H	2.75	3.00	3.15	3.50	92.0	0.437807
1032500999	I	3.46	3.54	3.60	3.88	92.0	1.419697
1032502999	J	3.46	3.63	3.45	3.50	92.0	1.62387

Berdasarkan Tabel 1, atribut yang digunakan untuk perhitungan k-means adalah IPS1, IPS2, IPS3, IPS4, SKS1SD4 dan *SKEWNESS*. Berikut tahapan-tahapan perhitungan k-means[11] :

- a. Menentukan nilai k: Menentukan jumlah kluster (k) yang diinginkan untuk cluster yang akan dibentuk. Pada contoh perhitungan ini menggunakan nilai kluster=2 atau nilai k=2
- b. Menginisialisasi centroid awal: Memilih k titik data secara acak sebagai centroid awal untuk setiap kluster. Jumlah K kluster tidak boleh lebih dari jumlah data. Pada contoh ini menggunakan menggunakan K=2 atau ada dua kluster. Selanjutnya menentukan centroid awal, dimana untuk menentukan centroid awal dilakukan secara acak. Disini data yang akan digunakan untuk centroid awal ada data pertama dan data ke dua. Data pertama sebagai C1 dan data ke kedua sebagai C2, seperti pada Tabel 2.

Tabel 2. Nilai Centroid Awal

CENTROID	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS
C1	3.12	3.21	2.75	2.7	63	0.072
C2	3.47	3.33	4	3.29	58	1.689

Berdasarkan Tabel 2 maka C1= (3.12, 3.21, 2.75, 2.70, 63,0, 0.072167) dan C2 = (3.47, 3.33, 4.00, 3.29, 58.0, 1.688810)

- c. Menghitung jarak setiap titik data ke masing-masing setiap centroid  
Metode yang akan digunakan untuk menghitung jarak setiap titik data ke masing – masing setiap centroid adalah *Euclidean Distance*.

Contoh data yang digunakan adalah 10 data seperti pada Tabel 1 maka iterasi pertama atau iterasi(1) :

Jumlah Data (n)=10 Data mahasiswa

Nilai K=2

Centroid Cluster 1 (C1) = {**3.12, 3.21, 2.75, 2.7, 63, 0.072**}

Centroid Cluster 2 (C2) = {**3.47, 3.33, 4, 3.29, 58, 1.689**}

Untuk menghitung jarak dengan menggunakan *Euclidean Distance* dengan cara :

Perhitungan C1 :

$$d(x_1, y_1) = \sqrt{(x_{1a} - y_{1a})^2 + (x_{1b} - y_{1b})^2 + (x_{1c} - y_{1c})^2 + (x_{1d} - y_{1d})^2 + (x_{1e} - y_{1e})^2 + (x_{1f} - y_{1f})^2}$$

$$d(x_1, y_1) = \sqrt{(3.12 - 3.12)^2 + (3.21 - 3.21)^2 + (2.75 - 2.75)^2 + (2.7 - 2.7)^2 + (63 - 63)^2 + (0.072 - 0.072)^2}$$

$$d(x_1, y_1) = 0$$

Perhitungan C2 :

$$d(x_1, y_2) = \sqrt{(x_{1a} - y_{2a})^2 + (x_{1b} - y_{2b})^2 + (x_{1c} - y_{2c})^2 + (x_{1d} - y_{2d})^2 + (x_{1e} - y_{2e})^2 + (x_{1f} - y_{2f})^2}$$

$$d(x_1, y_2) = \sqrt{(3.12 - 3.47)^2 + (3.21 - 3.33)^2 + (2.75 - 4)^2 + (2.7 - 3.29)^2 + (63 - 58)^2 + (0.072 - 1.689)^2}$$

$$d(x_1, y_2) = 5.45$$

Tabel 3. Hasil perhitungan jarak iterasi 1

IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	C1	C2	JARAK TERDEKAT
3.12	3.21	2.75	2.7	63	0.072	0.000	5.45	0.000
3.47	3.33	4	3.29	58	1.689	5.446	0.00	0.000
3.29	3.25	3.29	3	68	-1.895	5.411	10.65	5.411
3.82	3.4	3.33	3.42	68	1.814	5.424	10.03	5.424
3.56	3.5	3.23	3.13	83	-0.125	20.018	25.08	20.018
2.79	3	2.71	2.38	93	-0.664	30.013	35.12	30.013
3.38	3.13	3.22	3	83	0.251	20.010	25.06	20.010
2.75	3	3.15	3.5	92	0.438	29.019	34.04	29.019
3.46	3.54	3.6	3.88	92	1.420	29.072	34.01	29.072
3.46	3.63	3.45	3.5	92	1.624	29.066	34.01	29.066

- d. Masukkan setiap titik data ke dalam kluster berdasarkan jarak terdekat dengan pusat kluster  
Untuk menentukan data yang akan masuk kluster 1 atau kluster 2 seperti pada Tabel 3 menggunakan cara sebagai berikut :

Kluster 1 :

Jika  $d(x_1, y_1) < d(x_1, y_2)$  maka data mahasiswa -1 masuk ke dalam kluster 1

Jika  $0.000 < 5.45$  maka data mahasiswa -1 masuk ke dalam kluster 1

Kluster 2 :

Jika  $5.446 < 0.000$  maka data mahasiswa-2 masuk ke dalam kluster 2

Tabel 4. Tabel Pengelompokkan data sesuai jarak terpendek

IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	C1	C2	JARAK TERDEKAT	KLASTER
3.12	3.21	2.75	2.7	63	0.072	0.000	5.45	0.000	1
3.47	3.33	4	3.29	58	1.689	5.446	0.00	0.000	2
3.29	3.25	3.29	3	68	-1.895	5.411	10.65	5.411	1
3.82	3.4	3.33	3.42	68	1.814	5.424	10.03	5.424	1
3.56	3.5	3.23	3.13	83	-0.125	20.018	25.08	20.018	1
2.79	3	2.71	2.38	93	-0.664	30.013	35.12	30.013	1
3.38	3.13	3.22	3	83	0.251	20.010	25.06	20.010	1
2.75	3	3.15	3.5	92	0.438	29.019	34.04	29.019	1
3.46	3.54	3.6	3.88	92	1.420	29.072	34.01	29.072	1
3.46	3.63	3.45	3.5	92	1.624	29.066	34.01	29.066	1



e. Untuk setiap kluster, tentukan nilai centroid baru berdasarkan rata-rata (*means*) dari setiap data di dalam kluster atau menghitung *centroid* berikutnya.

Pada tahap berikutnya *centroid* atau menentukan centroid awal dengan cara menghitung rata-rata dari hasil perhitungan data pada masing-masing kluster (*cluster*). Berdasarkan tabel 4 maka yang masuk ke dalam kluster 1 ada 9 data yaitu data1, data3, data4, data5, data6, data7, data8, data9. Berikut ini perhitungan rata-rata centroid kluster 1 untuk masing-masing atribut :

$$IPS1 = (3.12+3.29+3.82+3.56+2.79+3.38+2.75+3.46+3.46)/9 = 3.292$$

$$IPS2 = (3.12+3.25+3.4+3.5+3+3.13+3+3.54+3.63)/9 = 3.296$$

$$IPS3 = (2.75+3.29+3.33+3.23+2.71+3.22+3.15+3.6+3.45)/9 = 3.192$$

dan seterusnya sampai atribut *skewness*

Sedangkan yang masuk kedalam kluster 2 hanya satu data yaitu data2 maka nilai rata-rata kluster 2 untuk masing – masing atribut :

$$IPS1 = 3.12/1 = 3.12; IPS2 = 3.33/1 = 3.33; IPS3 = 4/1 = 4$$

dan seterusnya sampai dengan atribut *skewness*. Hasil perhitungan rata-rata *centroid* kluster 1 dan kluster 2 terlihat pada Tabel 5.

Tabel 4. Centroid baru berdasarkan mean

Centroid	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS
C1	3.292	3.296	3.192	3.168	81.556	0.326
C2	3.47	3.33	4	3.29	58	1.689

Setelah mendapatkan centroid baru berdasarkan mean atau rata-rata seperti pada Tabel 5 maka selanjutnya menghitung jarak, dan menentukan data yang akan masuk kluster 1 atau kluster 2 seperti pada Tabel 6.

Tabel 5. Pengelompokkan data berdasarkan jarak terdekat iterasi 2

IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	C1	C2	JARAK TERDEKAT	KLASTER
3.12	3.21	2.75	2.7	63	0.072	18.575	5.446	5.446	2
3.47	3.33	4	3.29	58	1.689	23.624	0.000	0.000	2
3.29	3.25	3.29	3	68	-1.895	13.738	10.652	10.652	2
3.82	3.4	3.33	3.42	68	1.814	13.651	10.030	10.030	2
3.56	3.5	3.23	3.13	83	-0.125	1.552	25.079	1.552	1
2.79	3	2.71	2.38	93	-0.664	11.549	35.123	11.549	1
3.38	3.13	3.22	3	83	0.251	1.469	25.056	1.469	1
2.75	3	3.15	3.5	92	0.438	10.469	34.043	10.469	1
3.46	3.54	3.6	3.88	92	1.420	10.546	34.009	10.546	1
3.46	3.63	3.45	3.5	92	1.624	10.543	34.006	10.543	1

f. Ulangi langkah c-e sedemikian hingga tidak ada perubahan anggota kluster.

Pada iterasi ke 3 sudah tidak ada perubahan anggota kluster seperti terlihat pada Tabel 7.

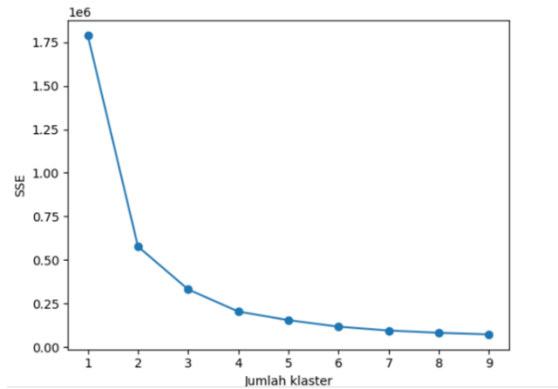
Tabel 6. Pengelompokkan data berdasarkan jarak terdekat iterasi 2

IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	C1	C2	JARAK TERDEKAT	KLASTER
3.12	3.21	2.75	2.7	63	0.072	26.20	1.52	1.52	2
3.47	3.33	4	3.29	58	1.689	31.27	6.41	6.41	2
3.29	3.25	3.29	3	68	-1.895	21.21	4.41	4.41	2
3.82	3.4	3.33	3.42	68	1.814	21.19	4.03	4.03	2
3.56	3.5	3.23	3.13	83	-0.125	6.29	18.76	6.29	1
2.79	3	2.71	2.38	93	-0.664	4.00	28.79	4.00	1
3.38	3.13	3.22	3	83	0.251	6.18	18.75	6.18	1
2.75	3	3.15	3.5	92	0.438	3.05	27.76	3.05	1
3.46	3.54	3.6	3.88	92	1.420	3.16	27.78	3.16	1
3.46	3.63	3.45	3.5	92	1.624	2.92	27.78	2.92	1

Berdasarkan Tabel 7, maka ada sebanyak empat data mahasiswa yang masuk dalam kluster dua(2) dan enam data mahasiswa yang masuk ke dalam kluster satu(1).

### 3.5 Tahap Evaluasi (*Evaluation*)

Optimalnya kita akan mencari skema kluster yang SSE semakin kecil. Berdasarkan data mahasiswa yang digunakan pada penelitian ini maka hasil pengujian nilai SSE dari kluster yang terbentuk, dapat ditarik kesimpulan bahwa nilai k yang direkomendasikan untuk digunakan adalah k=3. Sesuai Gambar 10.



Gambar 10. Nilai SSE

Berdasarkan gambar grafik maka sumbu X adalah kluster yang ada dan sumbu Y adalah SSE. Jika kita perhatikan pada grafik gambar di atas maka ada suatu titik yang mempunyai nilai sudut terkecil yaitu pada kluster 3. Kluster 3 merupakan titik yang paling optimal untuk dipilih karena nilai SSE kecil dan jumlah klasternya tidak banyak. Nilai titik pada kluster 3 adalah  $3.303968e+05$ . Fenomena ini disebut sebagai *Elbow Effect*. Metode *Elbow Effect* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik.

Setiap perhitungan jarak dikuadratkan karena dua alasan pertama agar angka SSE selalu positif dan untuk memperjelas nilai perbedaan untuk membantu para analis dalam menganalisis data. Metode Elbow untuk menentukan berapa kluster yang optimal suatu data.

### 3.6 Tahap Penyampaian (*Deployment*)

Tahap terakhir adalah mengimplementasikan model dan membuat rekomendasi berdasarkan hasil analisis data mining. Pada tahap ini, peneliti akan menyiapkan laporan hasil analisis data mining dan memberikan rekomendasi kepada stakeholder.

### 3.7 Hasil klasterisasi

Hasil klasterisasi dari data yang digunakan yang masuk ke kluster 0 adalah terlihat pada Tabel 8 sebanyak 1066 data yang masuk ke kluster 0.

Tabel 7. Hasil kluster 0

No	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	KLASTER
4	3.56	3.50	3.23	3.13	83.0	-0.125171	0
5	2.79	3.00	2.71	2.38	93.0	-0.663766	0
6	3.38	3.13	3.22	3.00	83.0	0.251165	0
7	2.75	3.00	3.15	3.50	92.0	0.437807	0
8	3.46	3.54	3.60	3.88	92.0	1.419697	0
...	...	...	...	...	...	...	...
13518	3.63	3.70	3.88	3.60	84.0	1.393327	0
13519	3.04	3.67	3.06	2.22	88.0	-0.523620	0
13520	3.00	3.49	3.67	3.32	84.0	-0.634525	0
13523	3.55	3.48	0.00	3.33	87.0	-1.983239	0
13525	3.35	2.65	2.96	2.57	88.0	0.893820	0

Sedangkan yang masuk ke kluster 1 terlihat pada Tabel 9, ada sebanyak 9045 data yang masuk ke kluster 1.

Tabel 8. Hasil kluster 1

No	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	KLASTER
0	3.12	3.21	2.75	2.70	63.0	0.072167	1
1	3.47	3.33	4.00	3.29	58.0	1.688810	1
24	3.57	3.39	2.89	3.31	65.0	-1.145572	1
28	4.00	0.00	0.00	0.00	3.0	2.000000	1
29	2.88	2.30	3.36	2.75	65.0	0.100096	1
...	...	...	...	...	...	...	...
13433	3.20	3.26	3.38	3.41	58.0	-0.228728	1
13465	3.50	3.21	3.08	2.57	56.0	-0.794865	1
13468	2.56	3.07	1.43	1.96	53.0	-0.032142	1
13471	3.38	2.51	2.92	2.61	57.0	0.994071	1
13492	3.58	3.41	0.00	0.00	50.0	0.006136	1

Dan untuk data mahasiswa yang masuk ke kluster 2 sebanyak 9045 seperti pada Tabel 10.

Tabel 9. Hasil Klaser 3

No	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS	KLASTER
2	3.29	3.25	3.29	3.00	68.0	-1.894787	2
3	3.82	3.40	3.33	3.42	68.0	1.814372	2
13	3.67	3.88	3.79	4.00	78.0	0.000000	2
32	3.24	3.36	2.80	3.00	66.0	-0.328468	2
35	3.14	3.17	1.87	3.00	70.0	-1.917194	2
...	...	...	...	...	...	...	...
13516	3.23	3.17	3.26	2.52	80.0	-1.933409	2
13517	2.63	2.04	2.55	2.20	82.0	-0.203331	2
13521	2.99	3.21	3.61	3.56	80.0	-0.434567	2
13522	2.75	3.01	2.76	2.68	80.0	1.624283	2
13524	3.35	1.57	3.24	3.11	81.0	-1.918427	2

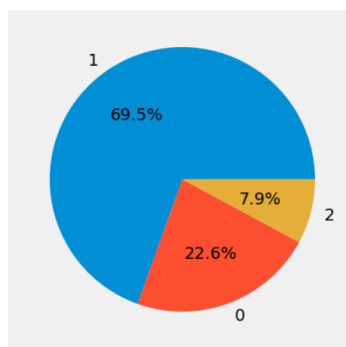
Untuk nilai means atau rerata dari centroid 0, centroid 1 dan centroid 2 terlihat pada Tabel 11.

Tabel 10. Nilai centroid setiap kluster

	IPS1	IPS2	IPS3	IPS4	SKS1SD4	SKEWNESS
Centroid 0	2.64	2.68	2.75	2.83	76.22	-0.236067
Centroid 1	2.47	2.84	2.64	2.68	52.61	-0.545125
Centroid 2	2.63	2.67	2.74	2.83	76.56	-0.232596

Berdasarkan Tabel 11, pada centroid 0 nilai means IPS1 adalah 2.64, IPS2 adalah 2.68, IPS3 adalah 2.75, IPS4 adalah 2.83, nilai SKS1SD4 adalah 76.22 dan nilai SKEWNESS adalah -0.236067. Sedangkan untuk centroid 1 nilai means IPS1 adalah 2.47, nilai IPS2 adalah 2.84, nilai IPS3 adalah 2.64, nilai IPS4 adalah 2.68, nilai SKS1SD4 adalah 52.61 dan nilai SKEWNESS adalah -0.545125. Dan untuk centroid 3 nilai rata-rata IPS1 adalah 2.63, nilai IPS2 adalah 2.67, nilai IPS3 adalah 2.74, nilai IPS4 adalah 2.83, nilai SKS1SD4 adalah 76.56, nilai SKEWNESS adalah -0.232596.

Hasil klusterisasi dari 1066 data mahasiswa yang digunakan pada penelitian dapat dilihat pada gambar 11 untuk kluster 0 yang masuk ke kluster 0 adalah terlihat pada tabel 7 sebanyak 1066 data yang masuk ke kluster 0, kluster 1 dan kluster 2.



Gambar 11. Grafik hasil klusterisasi

Berdasarkan Gambar 11 jumlah data yang masuk kedalam kluster 0 adalah data yang terbanyak yaitu 22.6 % atau 3055 mahasiswa, kluster 1 sebanyak 69.5 % atau 9405 mahasiswa sedangkan yang masuk kedalam kluster 2 adalah 7.9% atau 1066 mahasiswa.

#### 4. PENUTUP

Berdasarkan data mahasiswa sebanyak 13526 mahasiswa yang meliputi nilai IPS semester satu sampai dengan semester empat dan jumlah sks yang didapatkan dari semester satu sampai dengan semester 4 dikelompokkan ke dalam tiga kluster: Kluster 0 terdiri dari 3055 mahasiswa dengan rata-rata IPS semester pertama adalah 2.47, rata-rata IPS semester 2 adalah 2.83, rata-rata IPS semester tiga adalah 2.68, rata-rata IPS semester empat adalah 2.68 dan jumlah SKS lulus semester satu sampai dengan semester empat adalah 52.61 SKS. Kluster 0 menggambarkan mahasiswa dengan keberhasilan studi yang kurang baik, sehingga mahasiswa yang masuk dalam kluster ini perlu dilakukan pembimbingan akademik secara lebih intensif.

Kluster 1 merupakan kluster dengan jumlah mahasiswa yang paling banyak yaitu berjumlah 9045 mahasiswa. Mahasiswa yang berada di kluster 1 adalah mahasiswa dengan nilai rata-rata IPS semester satu adalah 3.09, rata-rata IPS semester dua adalah 3.22, rata-rata IPS semester tiga adalah 3.24, rata-rata IPS semester empat adalah 3.29 dan jumlah SKS lulus semester satu sampai dengan semester empat adalah 88.94 SKS. Mahasiswa yang masuk ke dalam kluster 1 merupakan mahasiswa dengan keberhasilan studi yang baik dan memiliki prestasi akademik yang baik pula. Mahasiswa dalam kluster ini dapat diberikan program khusus

untuk meningkatkan potensinya akademiknya, termasuk diikutsertakan dalam program beasiswa dan lomba-lomba akademik.

Selanjutnya, klaster 2 merupakan mahasiswa dengan nilai rata-rata IPS semester satu adalah 2.64, rata-rata IPS semester dua adalah 2.68, rata-rata IPS semester tiga adalah 2.74, rata-rata IPS semester empat adalah 2.68 dan jumlah SKS lulus semester satu sampai dengan semester empat adalah 88.94 SKS. Mahasiswa yang masuk dalam klaster 2 ini merupakan mahasiswa yang memiliki prestasi akademik menengah. Mahasiswa memiliki indeks prestasi yang cukup baik namun dengan jumlah SKS lulus yang tinggi. Program studi dapat melakukan bimbingan akademik untuk meningkatkan motivasi mahasiswa sehingga prestasi akademiknya semakin meningkat.

#### DAFTAR PUSTAKA

- [1] I. Vhallah, S. Sumijan, and J. Santony, "Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Metode Clustering K-Means," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 572–577, 2018, doi: 10.29207/resti.v2i2.308.
- [2] H. Kurniawan, S. Defit, and Sumijan, "Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 2, pp. 80–89, 2020, doi: 10.52158/jacost.v1i2.102.
- [3] I. Rahma, P. P. Arhandi, and A. T. Firdausi, "Penerapan Metode Hierarchical Clustering Dan K-Means Clustering Untuk Mengelompokkan Potensi Lokasi Penjualan Linkaja," *J. Inform. Polinema*, vol. 6, no. 1, pp. 15–22, 2020, doi: 10.33795/jip.v6i1.287.
- [4] H. Priyatman, F. Sajid, and D. Haldivany, "Klusterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 1, p. 62, 2019, doi: 10.26418/jp.v5i1.29611.
- [5] R. Dekaprasetya, U. R. Anisatur, and L. A. Muharom, "Clustering Untuk Menentukan Strategi Promosi Universitas Muhammadiyah Jember Dengan Algoritma K-Medoids," *J. Smart Teknol.*, vol. 3, no. 3, pp. 312–320, 2022.
- [6] B. A. Nugroho, I. D. Irawati, and A. L. Saonard, "Pengelompokan Transaksi Pelanggan Kartu Kredit Menggunakan Algoritma Clustering K-Means," vol. 9, no. 1, pp. 135–138, 2023.
- [7] P. Chapman *et al.*, *CRISP-DM 1.0 Step-by-step data mining guide*. 2000.
- [8] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
- [9] V. A. Ekasetya and A. Jananto, "Klusterisasi Optimal Dengan Elbow Method Untuk Pengelompokan Data Kecelakaan Lalu Lintas Di Kota Semarang," *J. Din. Inform.*, vol. 12, no. 1, pp. 20–28, 2020, doi: 10.35315/informatika.v12i1.8159.
- [10] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.
- [11] S. Handoko, F. Fauziah, and E. T. E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-Means Clustering," *J. Ilm. Teknol. dan Rekayasa*, vol. 25, no. 1, pp. 76–88, 2020, doi: 10.35760/tr.2020.v25i1.2677.