

Penerapan K-Means Clustering dalam Pengelompokan Data (Studi Kasus Profil Mahasiswa Matematika FMIPA UNM)

Ahmad Zaki, Irwan*, dan Imanuel Agung Sembe

¹Jurusan Matematika, Universitas Negeri Makassar

Corresponding Author: irwanthaha@unm.ac.id

Abstrak. Penelitian ini adalah penelitian terapan yang bertujuan untuk mengetahui cluster yang ada pada mahasiswa Jurusan Matematika FMIPA UNM menggunakan K-Means Clustering. Metode penelitian ini adalah studi literatur. Hasil penelitian diperoleh 4 cluster dimana durasi belajar mandiri dan IPK dari tertinggi ke terendah berturut-turut adalah Cluster 1, Cluster 2, Cluster 4, dan Cluster 3. Cluster 1 didominasi mahasiswa SBMPTN, Semester 3, umur rata-rata 19,20 tahun, durasi belajar mandiri 2,49 jam, 23,97 SKS, IPS 3,69, dan IPK 3,67. Cluster 2 didominasi mahasiswa SBMPTN, Semester 1, umur rata-rata 18,08 tahun, durasi belajar mandiri 2,07 jam, 22 SKS, IPS 3,64 dan IPK 3,63. Cluster 4 didominasi mahasiswa MANDIRI, Semester 5, umur rata-rata 19,78 tahun, durasi belajar mandiri 1,89 jam, 21,62 SKS, IPS 3,48, dan IPK 3,36. Cluster 3 didominasi mahasiswa SBMPTN dan SNMPTN bersama-sama, Semester 3, umur rata-rata 18,52 tahun, durasi belajar mandiri 1,29 jam, 21,87 SKS, IPS 3,13, dan IPK 3,19. Variabel yang paling berpengaruh dalam pembentukan cluster secara berturut-turut adalah Semester, Jumlah SKS, IPK, Umur, IPS, Rata-rata Durasi Belajar Mandiri, dan Jalur Masuk.

Kata Kunci: Cluster, K-Means Clustering, IPK, Durasi Belajar Mandiri.

Abstract. This research is an applied research that aims to find out the clusters that exist in students of the Mathematics Department of FMIPA UNM using K-Means Clustering. This research method is a literature study. The results of the study obtained 4 clusters where the duration of independent learning and GPA from highest to lowest were Cluster 1, Cluster 2, Cluster 4, and Cluster 3. Cluster 1 was dominated by SBMPTN students, Semester 3, average age 19.20 years, duration of self-study 2.49 hours, 23.97 credits, 3.69 IPS, and 3.67 GPA. Cluster 2 is dominated by SBMPTN students, Semester 1, average age 18.08 years, duration of independent study 2.07 hours, 22 credits, 3.64 IPS and GPA 3.63. Cluster 4 is dominated by MANDIRI students, Semester 5, average age 19.78 years, duration of independent study 1.89 hours, 21.62 credits, 3.48 IPS, and 3.36 GPA. Cluster 3 is dominated by SBMPTN and SNMPTN students together, Semester 3, average age 18.52 years, duration of independent study 1.29 hours, 21.87 credits, 3.13 IPS, and GPA 3.19. The most influential variables in the formation of clusters are Semester, Number of Credits, GPA, Age, Social Studies, Average Duration of Independent Study, and Entry Path.

Keywords: Cluster, K-Means Clustering, IPK, Duration of Independent Learning.

PENDAHULUAN

Clustering (pengelompokan) adalah salah satu teknik yang paling signifikan dalam *data mining* yang mengeksplorasi kumpulan data. Dalam beberapa decade terakhir, beberapa pendekatan *clustering* (pengelompokan) dengan kinerja yang lebih baik telah diterapkan untuk berbagai aplikasi, seperti bidang psikologi dan ilmu sosial lainnya, biologi, statistik, pengenalan pola, pencarian informasi, *machine learning*, dan *data mining* (penambangan data) (Omar et al., 2020; Tan et al., 2006). Analisis *cluster* merupakan teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya. Analisis *cluster* mengklasifikasi objek sehingga objek-objek yang paling dekat kesamaanya dengan

objek lain berada dalam *cluster* yang sama (misalnya, responden, produk, atau entitas yang lain) pada serangkaian karakteristik yang dipilih pengguna. *Cluster* yang dihasilkan harus menunjukkan homogenitas internal (*within-cluster*) yang tinggi dan heterogenitas eksternal (*between-cluster*) yang tinggi. Dengan demikian, jika klasifikasi berhasil, objek-objek dalam cluster akan berdekatan ketika diplot secara geometris, dan cluster yang berbeda akan menjadi bagian yang jauh (Hair Jr. et al., 2009; D. N. P. Sari & Sukestiyarno, 2021).

Profil Mahasiswa Matematika UNM dalam penelitian ini adalah Jalur Masuk, Semester, Rata-rata Durasi Belajar Mandiri per hari, Jumlah SKS, nilai IPS, dan nilai IPK. Variabel IPK, dan IPS adalah penilaian hasil belajar yang diperoleh oleh mahasiswa setiap akhir semesternya dan dalam mencapai hasil terbaik pada penilaian tersebut perlu didukung oleh proses belajar yang tinggi, hal ini kemudian diambil peneliti yang didefinisikan sebagai Durasi Belajar. Sebagian besar mahasiswa melakukan kesalahan dalam hal durasi belajar. Banyak mahasiswa yang lupa bahwa untuk mendapatkan hasil yang besar diperlukan usaha yang besar pula. Oleh karena itu peneliti tertarik untuk mengelompokkan mahasiswa berdasarkan kategori rata-rata durasi belajarnya per minggu dan umur mahasiswa tersebut. Selanjutnya Peneliti juga menambahkan faktor lain sebagai variabel dalam penelitian ini yakni Jalur Masuk, Semester, Umur, dan Jumlah SKS pada Tahun Akademik 2021-2022.

K-Means Cluster adalah salah satu metode dan clustering non hierarki yang berusaha mengelompokkan data ke dalam suatu *cluster* sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama (Wardono et al., 2019). *K-Means* termasuk algoritma *clustering* dengan proses berulang-ulang. Huruf K diartikan sebagai jumlah *cluster* yang hendak dibuat. Selanjutnya nilai K ditetapkan secara acak. Sedangkan *Means* adalah nilai sementara yang menjadi pusat dari *cluster* atau disebut juga dengan *centroid*. Setiap data yang ada dihitung jaraknya terhadap masing-masing *centroid* dengan memakai rumus Euclidean hingga dihasilkan jarak terdekat dari setiap data dengan *centroid* (Y. R. Sari et al., 2020).

Penelitian ini akan mengelompokkan mahasiswa berdasarkan Jalur Masuk, Semester, Umur, Rata-rata Durasi Belajar Mandiri per hari, Jumlah SKS, nilai IPS, dan nilai IPK. Metode yang digunakan adalah *K-Means Clustering*. *K-Means Cluster Analysis* merupakan salah satu metode *cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang lain. Metode *K-Means* berusaha mengelompokkan data yang ada kedalam satu kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain (D. N. P. Sari & Sukestiyarno, 2021). Pengelompokan ini bertujuan agar mahasiswa bisa melihat golongan mahasiswa yang terdapat di Jurusan Matematika FMIPA UNM berdasarkan variabel Jalur Masuk, Semester, Umur, Rata-rata Durasi Belajar Mandiri, IPS, dan IPK sehingga mahasiswa bisa lebih mengenali karakteristik belajar yang ada di lingkup jurusan matematika, khususnya saat pemberlakuan pembelajaran secara jarak jauh (PJJ) atau ketika pandemi Covid-19 belum usai.

Analisis Diskriminan diterapkan pada hasil *cluster* yang diperoleh. Dalam penelitian ini analisis diskriminan digunakan untuk mengetahui ketepatan klasifikasi. Jika nilai ketepatan klasifikasi besar, maka peng-*cluster*-an semakin baik sedangkan jika ketepatan klasifikasi kecil, maka peng-*cluster*-an kurang baik (Soemartini & Supartini, 2017). Hasil *cluster* yang diperoleh akan ditunjukkan dalam plot 2 dimensi, akibatnya perlu dilakukan reduksi variabel penelitian dari 7 variabel menjadi 2 variabel menggunakan Analisis Komponen Utama (*Principal Component Analysis*). *Principal Component Analysis* (PCA) adalah salah satu metode yang dapat digunakan untuk mereduksi suatu data berdimensi tinggi atau data dengan banyak variabel. PCA dapat di-

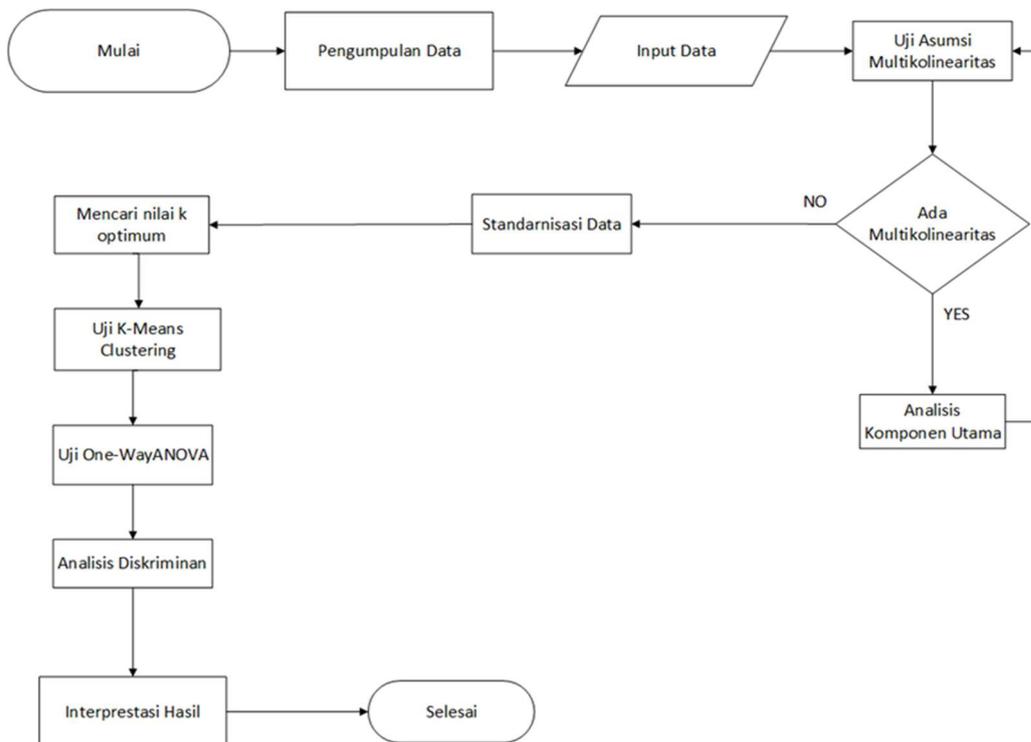
gunakan untuk mereduksi dimensi suatu data tanpa mengurangi informasi pada data tersebut secara signifikan (Toraismaya et al., 2020).

METODE PENELITIAN

Prosedur Penelitian

Dalam penelitian ini, peneliti menggunakan data hasil belajar mahasiswa yang diambil pada pihak jurusan, yang berupa variabel IPK, IPS, dan Jumlah SKS masing-masing mahasiswa. Selain itu, peneliti juga mengambil data berisi variabel durasi belajar dan umur mahasiswa melalui pemberian angket (kuesioner). Prosedur dalam penelitian ini adalah sebagai berikut:

- 1) Uji Missing Data
- 2) Uji Asumsi Multikolinearitas
- 3) Standarnisasi Data dan Uji Data Outlier
- 4) Penentuan K Optimum *Cluster*
- 5) Proses *K-Means Clustering*
- 6) Uji *One-Way ANOVA* dan *p-value*.
- 7) Interpretasi Hasil
- 8) Evaluasi *Cluster*



GAMBAR 1. Flowchart Penelitian

Instrumen Penelitian

Instrumen penelitian dalam penelitian ini adalah kuesioner yang berisi data jalur masuk, semester, umur, rata-rata durasi belajar mandiri per hari. Selain itu digunakan juga instrumen penelitian berupa data hasil belajar tiap responden yang diambil dari pihak jurusan matematika berupa jumlah SKS, nilai IPS, dan nilai IPK pada semester ganjil Tahun Akademik 2021/2022.

HASIL DAN PEMBAHASAN

Setelah serangkaian uji data telah selesai dilakukan dan data telah memenuhi asumsi pada analisis *cluster* (sampel yang mewakili dan nonmultikolinearitas) maka dilakukan penentuan jumlah optimum *cluster*. Jumlah optimum *cluster* dalam penelitian ini adalah sebanyak 4 *cluster* menurut metode Elbow. Setelah itu dilakukan prosedur *K-Means Clustering*. Langkah awal adalah menentukan jumlah *cluster*, menentukan *centroid* awal, menghitung jarak antara objek dan setiap *centroid*, pengalokasian data ke *centroid* terdekat, dan validasi (iterasi sampai setiap anggota *cluster* atau *centroid* tidak berubah).

Pengumpulan Data

Data yang diperlukan dalam penelitian ini diperoleh melalui bagian akademik Jurusan Matematika FMIPA UNM yang berupa Jumlah SKS, IPS, dan IPK pada semester ganjil Tahun Akademik 2021 – 2022. Adapun untuk Kelas, Jalur Masuk, Semester, Umur dan Durasi Belajar diperoleh dari Kuesioner yang dibagikan tiap – tiap kelas di Jurusan Matematika FMIPA UNM. Berikut ditampilkan 5 data teratas dari 323 data yang diperoleh pada Tabel 1.

TABEL 1. Data Awal Seluruh Kelas

No	Jalur Masuk	Semester	Umur	Rata-rata Durasi Belajar Mandiri per hari (Jam)	Jumlah SKS	IPS	IPK
1	MANDIRI	5	20	3,00	22	3,57	3,17
2	SNMPTN	5	20	3,14	22	3,80	3,77
3	MANDIRI	5	20	2,29	22	3,48	3,24
4	SBMPTN	5	20	2,14	22	3,59	3,34
5	SNMPTN	5	20	2,21	22	3,76	3,56
...
323	MANDIRI	1	18	3,14	22	3,82	3,82

Inisialisasi Variabel

Pada kolom Jalur Masuk memiliki data berupa data kualitatif yaitu ordinal karena semua data memiliki posisi setara dalam arti tidak ada data yang memiliki tingkat lebih atau kurang dibandingkan dengan data yang lain sehingga diperlukan inisialisasi data terlebih dahulu ke dalam bentuk angka. Adapun inisialisasi data yang peneliti lakukan ditampilkan dalam Tabel 2 dan Tabel 3 berikut.

TABEL 2. Inisialisasi Variabel Jalur Masuk

No	Jalur Masuk	Inisial
1	SNMPTN	1
2	SBMPTN	2
3	MANDIRI	3

Kemudian Tabel 2 disusun kembali dengan mendefinisikan Operasional Variabel sebagai berikut:

X_1 = Jalur Masuk Mahasiswa

X_2 = Semester Mahasiswa pada semester ganjil Tahun Akademik 2021 – 2022

X_3 = Umur Mahasiswa pada semester ganjil Tahun Akademik 2021 – 2022

X_4 = Rata – rata Durasi Belajar Mandiri Mahasiswa per Hari (Jam) pada semester ganjil Tahun Akademik 2021 – 2022

X_5 = Jumlah SKS Mahasiswa pada semester ganjil Tahun Akademik 2021 – 2022

X_6 = IPS Mahasiswa pada semester ganjil Tahun Akademik 2021 – 2022

X_7 = IPK Mahasiswa pada semester ganjil Tahun Akademik 2021 – 2022

Sehingga diperoleh tabel baru seperti pada Tabel 3 berikut.

TABEL 3. Pendefinisian Operasional Variabel dan Inisialisasi Data Ordinal

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	MANDIRI	5	20	3,00	22	3,57	3,17
2	SNMPTN	5	20	3,14	22	3,80	3,77
3	MANDIRI	5	20	2,29	22	3,48	3,24
4	SBMPTN	5	20	2,14	22	3,59	3,34
5	SNMPTN	5	20	2,21	22	3,76	3,56
...
323	MANDIRI	1	18	3,14	22	3,82	3,82

Uji Multikolinearitas

Multikolinearitas adalah suatu kondisi dimana terjadi korelasi antara variabel bebas atau antar variabel bebas tidak bersifat saling bebas. Besaran (*quality*) yang dapat digunakan untuk mendeteksi adanya multikolinearitas adalah faktor inflasi ragam (*Variance Inflation Factor/VIF*). VIF digunakan sebagai kriteria untuk mendeteksi multikolinearitas pada regresi linear yang melibatkan lebih dari dua variabel bebas (Sriningsih et al., 2018).

Terjadinya multikolinearitas apabila nilai $VIF_j \geq 10$. Jika terindikasi terjadi multikolinearitas maka harus dilakukan tindakan perbaikan multikolinearitas (Ningrat et al., 2016). Terdapat berbagai cara dalam menangani multikolinearitas yaitu, transformasi logaritma natural, metode *Partial Least Square* (PLS), metode Principal Component Analysis (PCA), dan metode Stepwise (Marcus et al., 2012; Pendi, 2021; D. N. P. Sari & Sukestiyarno, 2021; Silitonga et al., 2021; Supriyadi et al., 2017).

TABEL 4. Variance Inflation Factor

Variabel	VIF	Keterangan
X_1	8,1967	Tidak Terjadi Multikolinearitas
X_2	3,3670	Tidak Terjadi Multikolinearitas

Variabel	VIF	Keterangan
X_3	3,1546	Tidak Terjadi Multikolinearitas
X_4	1,0787	Tidak Terjadi Multikolinearitas
X_5	1,4881	Tidak Terjadi Multikolinearitas
X_6	4,0984	Tidak Terjadi Multikolinearitas
X_7	5,3191	Tidak Terjadi Multikolinearitas

Berdasarkan Tabel 4 terlihat bahwa semua variabel penelitian tidak mengalami multikolinearitas karena memiliki nilai $VIF < 10$ sehingga proses peng-cluster-an dapat dilanjutkan karena telah memenuhi asumsi nonmultikolinearitas.

Standarnisasi Data

Standarnisasi data dilakukan dengan merubah data ke dalam bentuk z score menggunakan rumus standarnisasi data. Dengan menggunakan nilai mean, dan standar deviasi pada Tabel 2 dihitung nilai z score setiap data sebagai berikut:

$$Z_{X_{1,1}} = \frac{3 - 2,06}{0,770} = 1,2208$$

$$Z_{X_{1,2}} = \frac{1 - 2,06}{0,770} = -1,3766$$

$$\vdots$$

$$Z_{X_{1,323}} = \frac{3 - 2,06}{0,770} = 1,2208$$

Dengan cara yang sama diperoleh bentuk standarnisasi data ke dalam bentuk z score seperti pada Tabel 5 berikut.

TABEL 5 Standarnisasi Data

No	ZX1	ZX2	ZX3	ZX4	ZX5	ZX6	ZX7
1	1,21877	1,40461	1,23703	0,77481	-0,23238	0,22741	-1,32239
2	-1,37967	1,40461	1,23703	0,87998	-0,23238	1,07982	1,23375
3	1,21877	1,40461	1,23703	0,24898	-0,23238	-0,10613	-1,02417
4	-0,08045	1,40461	1,23703	0,14381	-0,23238	0,30154	-0,59815
5	-1,37967	1,40461	1,23703	0,19640	-0,23238	0,93157	0,33910
...
323	1,21877	-1,04777	-0,89396	0,87998	-0,23238	1,15394	1,44676

Uji Data Outlier

Menurut Danasla, Kusuma, Tuheteru, dan Gautama (2021) pengujian data outlier dilihat dengan mengkonversi data ke dalam bentuk Z-Score. Outlier terjadi apabila nilai Z-Score berada di luar rentang -3,00 sampai +3,00. Berdasarkan Tabel 6 dapat dilihat beberapa data yang mengalami outlier seperti data nomor 12 variabel ke-5 dengan nilai -3,19, data nomor 44 variabel ke-5 dengan nilai -3,19, dan data nomor 149 variabel ke-4 dengan nilai 4,034 yang kesemua nilainya berada di luar rentang -2,5 sampai +2,5. Namun data-data tersebut tidak dihapus atau dihilangkan melainkan dipertahankan karena diperlukan

Asumsi Pada Analisis Cluster

Terdapat dua asumsi yang perlu dipenuhi sebelum melakukan proses *clustering* yakni sampel yang mewakili dan nonmultikolinearitas.

Sampel yang mewakili (Sampel Representatif)

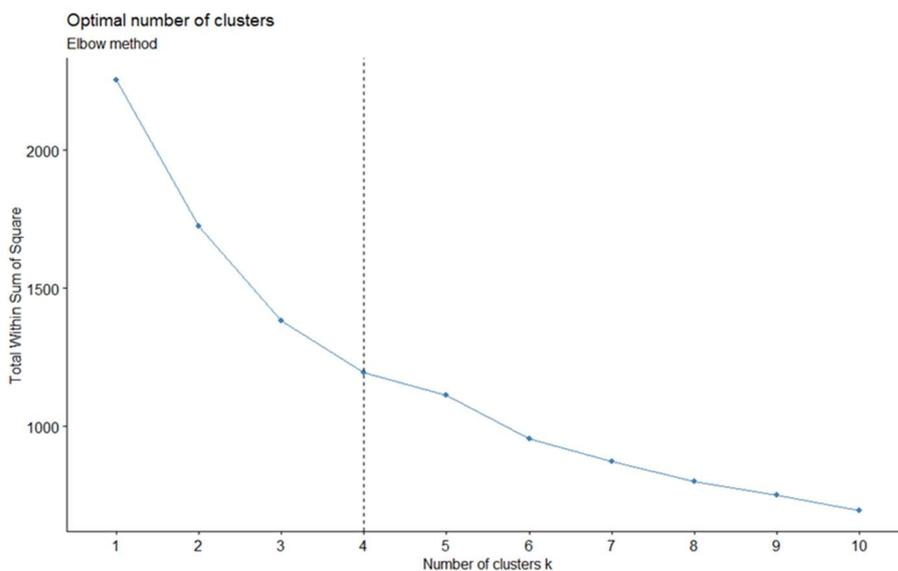
Nilai KMO sebesar $0,518 > 0,5$ sehingga disimpulkan sampel telah mewakili populasi atau sampel representative.

Multikolinearitas

Berdasarkan Tabel 5 terlihat bahwa semua variabel penelitian tidak mengalami multikolinearitas karena memiliki nilai $VIF < 10$ sehingga asumsi nonmultikolinearitas terpenuhi.

Penentuan K-Optimum Cluster

Penentuan jumlah *cluster* optimal ditentukan dengan menggunakan metode Elbow. Metode ini menggunakan nilai *Sum of Square Error* (SSE) dari masing-masing jumlah *cluster*. Semakin besar jumlah *cluster*, maka SSE akan terus mengecil.



GAMBAR 2. Plot Grafik Elbow

Gambar 2 menunjukkan jumlah kuadrat “within” atau dikenal dengan nama *within sum of square* (WSS). Tentunya diinginkan nilai WSS yang rendah, namun WSS akan bernilai 0 atau tidak ada persebaran sama sekali ketika banyaknya *cluster* sama dengan banyaknya observasi. Sehingga pada metode ini dilihat perubahan WSS yang cukup sedikit. Dari plot di atas terlihat bahwa perubahan WSS dari $k = 4$ ke $k = 5$ mulai bernilai cukup kecil. Artinya hasil reduksi WSS dari penambahan 1 *cluster* untuk $k = 4$ tidaklah terlalu berarti dan dapat diputuskan bahwa $k = 4$ adalah banyaknya *cluster* optimala pada penelitian ini.

Uji K-Means Clustering

Prosedur *K-Means Clustering* dimulai dengan menentukan jumlah *cluster*, menentukan *centroid* awal, menghitung jarak antara objek dan setiap *centroid*, pengalokasian data ke *centroid*

terdekat, dan validasi (iterasi sampai setiap anggota *cluster* atau *centroid* tidak berubah). Pemilihan nilai *centroid* awal dipilih secara random berdasarkan data yang digunakan. Hasil *Clustering* dengan 4 *cluster* ditampilkan pada Tabel 7 berikut.

TABEL 6. Hasil Cluster Akhir

No.	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	2	100	70	1
2	6	128	88	3
3	9	129	89	4
...
60	189	250	320	80
61		251	281	81
..	
89		287		298
...	
113		323		

Tabel 6 merupakan hasil akhir *K-Means Clustering*. Nomor pada setiap kolom Cluster menunjukkan nomor data dari Tabel 5. Standarnisasi Data. Dari Tabel 6 terlihat bahwa Cluster 1 diisi oleh 60 anggota, Cluster 2 diisi oleh 113 anggota, Cluster 3 diisi oleh 61 anggota, dan Cluster 4 diisi oleh 89 anggota.

Plot Cluster

Untuk membuat plot grafik *cluster* diperlukan reduksi dataset menggunakan Analisis Komponen Utama (Kassambara, 2017). Variabel penelitian yang berjumlah tujuh direduksi menjadi dua variabel baru menggunakan Analisis Komponen Utama.



GAMBAR 3. Plot Cluster

Gambar 3 menjelaskan bahwa sumbu X dan sumbu Y merupakan komponen-komponen utama (*principal component*) yang diformasi dari 7 variabel penelitian pada Tabel 5. Standarnisasi Data. Masing-masing komponen ini (Dim1 dan Dim2) menjelaskan variansi data yang terjelaskan setiap komponen, masing-masing sebesar 34% dan 25,7%. Daerah yang berwarna merah muda menunjukkan Cluster 1, daerah yang berwarna hijau menunjukkan Cluster 2, daerah yang berwarna biru menunjukkan Cluster 3, dan daerah yang berwarna ungu menunjukkan Cluster 4.

Uji One Way ANOVA

Menurut Hair, Black, Babin, dan Anderson (2009) dalam bukunya *Multivariate Data Analysis*, untuk menguji apakah ada perbedaan yang signifikan secara statistik antara cluster yang terbentuk dengan masing-masing variabel *clustering* (variabel bebas) adalah dengan menggunakan statistik *F* dari *One-Way-ANOVA* (Hair Jr. et al., 2009). Dengan membandingkan nilai *F*, dapat diketahui bahwa cluster yang diperoleh dengan menggunakan metode *K-Means* adalah cluster yang paling berbeda (Dikmen et al., 2009).

Menurut, Zancanaro, Kuflik, Boger, Goren-Bar, dan Goldwasser (2007) menyatakan bahwa dalam uji *One Way ANOVA* variabel dependent (terikat) adalah nilai-nilai pada variabel bebas (dalam penelitian ini adalah nilai-nilai pada ZX1, ZX2, ZX3, ZX4, ZX5, ZX6, ZX7, dan ZX8) sedangkan faktor adalah *cluster* K-Means. Berikut dirangkum semua nilai F_{hit} dan pada setiap variabel hasil ANOVA dalam Tabel 7.

Selain menghitung nilai *One-Way ANOVA*, peneliti juga menghitung nilai *p – value* (.sig) sebagai uji hipotesis dari masing-masing variabelnya dengan terlebih dahulu menentukan asumsi hipotesisnya sebagai berikut:

$$H_0 = \text{Ketiga cluster tidak mempunyai perbedaan signifikan}$$

$$H_1 = \text{Ketiga cluster mempunyai perbedaan signifikan}$$

dimana

Jika angka signifikan (*p – value*) > 0.05; H_0 diterima, H_1 ditolak.

Jika angka signifikan (*p – value*) ≤ 0.05; H_0 ditolak, H_1 diterima.

Uji Hipotesis ini bertujuan untuk mengetahui apakah ketiga *cluster* memiliki perbedaan yang signifikan atau tidak. Selanjutnya, diperoleh *p – value* untuk setiap variabel.

TABEL 7. Nilai F_{hitung} dan *p – value* setiap variabel

Variabel	F_{hitung}	<i>p – value</i>
ZX1	6,4789	0,000
ZX2	307,985	0,000
ZX3	131,8764	0,000
ZX4	8,8728	0,000
ZX5	238,162	0,000
ZX6	122,349	0,000
ZX7	184,937	0,000

Dari Tabel 7 terlihat bahwa variabel ZX2 adalah variabel yang memiliki nilai F_{hitung} terbesar, disusul variabel ZX5, ZX7, ZX3, ZX6, ZX4, dan ZX1. Sehingga variabel ZX2 (Semester) adalah variabel yang paling berperan dalam pembentukan *cluster*, disusul variabel ZX5 (Jumlah SKS), ZX7 (IPK), ZX3 (Umur), ZX6 (IPS), ZX4 (Rata-rata Durasi Belajar Mandiri), dan ZX1

(Jalur Masuk). Untuk nilai $p - value$ pada setiap variabel lebih kecil dari 0,05 sehingga tolak H_0 , terima H_1 atau dengan kata lain: keempat *cluster* memiliki perbedaan yang signifikan.

Analisis Diskriminan

Untuk mengetahui hasil ketepatan klasifikasi dengan *K-Means* dilakukan analisis diskriminan dengan terlebih dahulu dilakukan pengujian asumsi yang terdiri dari uji normal multivariate, uji kehomogenan matriks kovarian kedua kelompok, dan uji perbedaan vektor rata-rata. Setelah memenuhi ketiga asumsi di atas maka dilakukan perhitungan ketepatan klasifikasi *cluster*.

TABEL 8. Tabel Klasifikasi K-Means

D		Prediksi				Total
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	
Aktual	Cluster 1	59	0	0	1	60
	Cluster 2	0	109	0	4	113
	Cluster 3	0	1	60	0	61
	Cluster 4	0	1	4	84	89

Maka diperoleh persentase ketepatan klasifikasi *cluster* K-Means sebagai berikut:

$$CCR = \frac{59 + 109 + 60 + 84}{323} \times 100\% = 96,594\%$$

Selanjutnya diperoleh persentase kesalahan dalam klasifikasi *cluster* K-Means sebagai berikut:

$$APER = \frac{1 + 4 + 1 + 5}{323} \times 100\% = 3,41\%$$

Sehingga perhitungan tersebut dapat terlihat bahwa nilai ketepatan klasifikasi K-Means tinggi yakni 97% dengan tingkat kesalahan 3%.

Statistik Deskriptif Variabel

Statistik deskriptif dari hasil *cluster* yang diperoleh pada Tabel 6 ditunjukkan dalam Tabel 9 berikut.

TABEL 9. Statistik Deskriptif Variabel

Variabel		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Jalur Masuk	(SNMPTN)	22	33	3	28
	(SBMPTN)	24	47	29	31
	(MANDIRI)	14	33	29	30
Semester	1	0	108	25	1
	3	39	5	35	23
	5	21	0	1	65
Umur	Mean	19,20	18,08	18,52	19,78
Durasi Belajar Mandiri	Mean	2,49	2,07	1,29	1,89
Jumlah SKS	Mean	23,97	22,00	21,87	21,62
IPS	Mean	3,69	3,64	3,13	3,48
IPK	Mean	3,67	3,63	3,19	3,36

Cluster 1 untuk variabel Jalur Masuk diisi oleh 22 orang mahasiswa jalur SNMPTN, 24 orang mahasiswa jalur SBMPTN, dan 14 orang mahasiswa jalur MANDIRI. Variabel Semester diisi oleh 39 orang mahasiswa Semester 3, 21 orang mahasiswa semester 5, dan tidak ada seorang pun mahasiswa dari semester 1. Rata-rata mahasiswa di Cluster 1 berumur 19 tahun, memiliki durasi belajar mandiri per hari rata-rata 2,49 jam, jumlah SKS rata-rata berjumlah 24 SKS, dengan IPS rata-rata 3,69, dan IPK rata-rata 3,67.

Cluster 2 untuk variabel Jalur Masuk diisi oleh 33 orang mahasiswa jalur SNMPTN, 47 orang mahasiswa jalur SBMPTN, dan 33 orang mahasiswa jalur MANDIRI. Variabel Semester diisi oleh 108 orang mahasiswa semester 1, 5 orang mahasiswa semester 3, dan tidak ada seorang pun mahasiswa dari semester 5. Rata-rata mahasiswa di Cluster 2 berumur 18 tahun, memiliki durasi belajar mandiri per hari rata-rata 2,07 jam, jumlah SKS rata-rata 22 SKS, dengan IPS rata-rata 3,64, dan IPK rata-rata 3,63.

Cluster 3 untuk variabel Jalur Masuk diisi oleh 3 orang mahasiswa jalur SNMPTN, 29 orang mahasiswa jalur SBMPTN, dan 29 orang mahasiswa jalur MANDIRI. Variabel Semester diisi oleh 25 orang mahasiswa semester 1, 35 orang mahasiswa semester 3, dan 1 orang mahasiswa semester 5. Rata-rata mahasiswa di Cluster 3 berumur 19 tahun, memiliki durasi belajar mandiri per hari rata-rata 1,29 jam, jumlah SKS rata-rata 22 SKS, dengan IPS rata-rata 3,13, dan IPK rata-rata 3,19.

Cluster 4 untuk variabel Jalur Masuk diisi oleh 28 orang mahasiswa jalur SNMPTN, 31 orang mahasiswa jalur SBMPTN, dan 30 orang mahasiswa jalur MANDIRI. Variabel Semester diisi oleh 1 orang mahasiswa semester semester 1, 23 orang mahasiswa semester 3, dan 65 orang mahasiswa semester 5. Rata-rata mahasiswa di Cluster 4 berumur 20 tahun, memiliki durasi belajar mandiri per hari rata-rata 1,89 jam, jumlah SKS rata-rata 22 SKS, dengan IPS rata-rata 3,48, dan IPK rata-rata 3,36.

Hasil penelitian menunjukkan bahwa nilai rata-rata IPK setiap *cluster* diperoleh IPK tertinggi ke terendah berturut-turut diisi oleh Cluster 1, Cluster 2, Cluster 4, dan Cluster 3. Diperoleh juga durasi belajar mandiri per hari dari tertinggi ke terendah berturut-turut diisi oleh Cluster 1, Cluster 2, Cluster 4, dan Cluster 3. Sehingga dari keempat *cluster* di atas dapat disimpulkan bahwa mahasiswa Jurusan Matematika FMIPA UNM pada tahun akademik 2021-2022 dapat dikelompokkan menjadi 4 *cluster* dimana mahasiswa dengan IPK dan durasi belajar mandiri per hari yang ‘tinggi’ masuk ke Cluster 1. Mahasiswa dengan IPK dan durasi belajar mandiri per hari yang ‘sedang’ masuk ke Cluster 2. Mahasiswa dengan IPK dan durasi belajar mandiri yang ‘rendah’ masuk ke Cluster 4. Sedangkan mahasiswa dengan IPK dan durasi belajar mandiri yang ‘sangat rendah’ masuk ke Cluster 3.

KESIMPULAN

Hasil penelitian ini menunjukkan bahwa :

- a) Dengan menerapkan algoritma *K-Means Clustering* dalam pengelompokan data profil mahasiswa Matematika UNM diperoleh 4 *cluster*, dimana mahasiswa dengan IPK dan durasi belajar mandiri per hari yang ‘tinggi’ masuk ke Cluster 1. Mahasiswa dengan IPK dan durasi belajar mandiri per hari yang ‘sedang’ masuk ke Cluster 2. Mahasiswa dengan IPK dan durasi belajar mandiri yang ‘rendah’ masuk ke Cluster 4. Sedangkan mahasiswa dengan IPK dan durasi belajar mandiri yang ‘sangat rendah’ masuk ke Cluster 3.
- b) Mahasiswa – mahasiswa dengan durasi belajar mandiri yang ‘tinggi’ dan nilai IPK yang ‘tinggi’ berada dalam satu *cluster* yang sama begitu juga sebaliknya. Hal

ini menunjukkan durasi belajar mandiri dan IPK berbanding lurus pada sebagian besar mahasiswa Matematika FMIPA UNM.

- c) Variabel yang paling berpengaruh dalam pembentukan *cluster* adalah Semester. Mahasiswa-mahasiswa dengan semester yang sama cenderung memiliki umur dan mata kuliah yang sama sehingga cenderung akan memperoleh hasil belajar yang sama karena jenis tugas, penyelesaian, dan pengerjaannya yang besar kemungkinan dikerjakan bersama – sama.
- d) Mahasiswa pada semester 1 akan lebih banyak masuk ke dalam *cluster* yang sama dibandingkan mahasiswa pada semester lainnya. Hal ini menunjukkan saat masih berstatus mahasiswa baru, semangat dan motivasi belajar sesama mahasiswa cenderung masih sama sehingga mempengaruhi durasi belajar mandiri, IPK, IPS yang tidak akan jauh berbeda jika dibandingkan dengan mahasiswa di atas semester 1.
- e) Keempat *cluster* yang terbentuk memiliki perbedaan signifikan dan variabel yang paling berpengaruh dalam pembentukan *cluster* berturut-turut adalah Semester, Jumlah SKS, IPK, Umur, IPS, Rata-rata Durasi Belajar Mandiri, dan Jalur Masuk.
- f) Keempat *cluster* yang terbentuk memiliki nilai ketepatan klasifikasi sebesar 97% dengan persentase kesalahan sebesar 3%.

DAFTAR PUSTAKA

- Dikmen, I., Birgonul, M. T., & Budayan, C. (2009). Strategic Group Analysis in the Construction Industry. *Journal of Construction Engineering and Management*.
- Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (Seventh Ed). Pearson.
- Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R* (First). STHDA (<http://www.sthda.com>).
- Marcus, G. L., Wattimanela, H. J., & Lesnussa, Y. A. (2012). Analisis Regresi Komponen Utama Untuk Mengatasi Masalah Multikolinearitas Dalam Analisis Regresi Linear Berganda. *Barekeng-Jurnal Ilmu Matematika Dan Terapan*, Vol.6(No.1), 31–40.
- Ningrat, D. R., Maruddani, D. A. I., & Wuryandari, T. (2016). Analisis Cluster Dengan Algoritma K-Means dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi. *Jurnal Gaussian*, 5(4), 641–650.
- Omar, T., Alzahrani, A., & Zohdy, M. (2020). Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. *Journal of Data Analysis and Information Processing*, 8, 171–182.
- Pendi. (2021). Analisis Regresi Dengan Metode Komponen Utama Dalam Mengatasi Masalah Multikolinearitas. *Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster)*, 10(1), 131–138.
- Sari, D. N. P., & Sukestiyarno, Y. (2021). Analisis Cluster dengan Metode K-Means pada Persebaran Kasus COVID-19 Berdasarkan Provinsi di Indonesia. *PRISMA, Prosiding Seminar Nasional Matematika*, 602–610.
- Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer. *CESS (Journal of Computer Engineering, System and Science)*, 5(2), 192–198.
- Silitonga, Y. C., Kamid, & Multahadah, C. (2021). Perbandingan Metode Stepwise dan PCA pada Kasus Faktor-Faktor yang Mempengaruhi Pendapatan Asli Daerah di Provinsi Jambi. *Gamma-Pi: Jurnal Matematika Dan Terapan*, 3(2).
- Soemartini, & Supartini, E. (2017). Analisis K-Means Cluster Untuk Pengelompokan

- Kabupaten/Kota di Jawa Barat Berdasarkan Indikator Masyarakat. *Konferensi Nasional Penelitian Matematika Dan Pembelajarannya II (KNPMP II)*, 144–154.
- Sriningsih, M., Hatidja, D., & Prang, J. D. (2018). Penanganan Multikolinearitas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras di Provinsi Sulut. *Jurnal Ilmiah Sains*, 18, 18–24.
- Supriyadi, E., Mariani, S., & Sugiman. (2017). Perbandingan Metode Partial Least Square (PLS) dan Principal Component Regression (PCR) Untuk Mengatasi Multikolinearitas Pada Model Regresi Linear Berganda. *UNNES Journal of Mathematics*, 6(2).
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction To Data Mining*. Pearson Education, Inc.
- Toraismaya, A., Sasongko, L. R., & Rondonuwu, F. S. (2020). Principal Component Dan K-Means Cluster Analysis Untuk Data Spektrum Black Tea Grades Guna Penilaian Kualitas Alternatif. *Journal of Fundamental Mathematics and Applications (JFMA)*, 3(2), 148–157.
- Wardono, Sunarmi, & Wirawan, M. R. (2019). Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Dengan Metode K-Means Cluster. *Seminar Nasional Edusaintek*, 599–610.
- Dikmen, I., Birgonul, M. T., & Budayan, C. (2009). Strategic Group Analysis in the Construction Industry. *Journal of Construction Engineering and Management*.
- Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (Seventh Ed). Pearson.
- Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R* (First). STHDA (<http://www.sthda.com>).
- Marcus, G. L., Wattimanela, H. J., & Lesnussa, Y. A. (2012). Analisis Regresi Komponen Utama Untuk Mengatasi Masalah Multikolinearitas Dalam Analisis Regresi Linear Berganda. *Barekeng-Jurnal Ilmu Matematika Dan Terapan*, Vol.6(No.1), 31–40.
- Ningrat, D. R., Maruddani, D. A. I., & Wuryandari, T. (2016). Analisis Cluster Dengan Algoritma K-Means dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi. *Jurnal Gaussian*, 5(4), 641–650.
- Omar, T., Alzahrani, A., & Zohdy, M. (2020). Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. *Journal of Data Analysis and Information Processing*, 8, 171–182.
- Pendi. (2021). Analisis Regresi Dengan Metode Komponen Utama Dalam Mengatasi Masalah Multikolinearitas. *Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster)*, 10(1), 131–138.
- Sari, D. N. P., & Sukestiyarno, Y. (2021). Analisis Cluster dengan Metode K-Means pada Persebaran Kasus COVID-19 Berdasarkan Provinsi di Indonesia. *PRISMA, Prosiding Seminar Nasional Matematika*, 602–610.
- Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer. *CESS (Journal of Computer Engineering, System and Science)*, 5(2), 192–198.
- Silitonga, Y. C., Kamid, & Multahadah, C. (2021). Perbandingan Metode Stepwise dan PCA pada Kasus Faktor-Faktor yang Mempengaruhi Pendapatan Asli Daerah di Provinsi Jambi. *Gamma-Pi: Jurnal Matematika Dan Terapan*, 3(2).
- Soemartini, & Supartini, E. (2017). Analisis K-Means Cluster Untuk Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Indikator Masyarakat. *Konferensi Nasional Penelitian Matematika Dan Pembelajarannya II (KNPMP II)*, 144–154.

- Sriningsih, M., Hatidja, D., & Prang, J. D. (2018). Penanganan Multikolinearitas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras di Provinsi Sulut. *Jurnal Ilmiah Sains*, 18, 18–24.
- Supriyadi, E., Mariani, S., & Sugiman. (2017). Perbandingan Metode Partial Least Square (PLS) dan Principal Component Regression (PCR) Untuk Mengatasi Multikolinearitas Pada Model Regresi Linear Berganda. *UNNES Journal of Mathematics*, 6(2).
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction To Data Mining*. Pearson Education, Inc.
- Toraismaya, A., Sasongko, L. R., & Rondonuwu, F. S. (2020). Principal Component Dan K-Means Cluster Analysis Untuk Data Spektrum Black Tea Grades Guna Penilaian Kualitas Alternatif. *Journal of Fundamental Mathematics and Applications (JFMA)*, 3(2), 148–157.
- Wardono, Sunarmi, & Wirawan, M. R. (2019). Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Dengan Metode K-Means Cluster. *Seminar Nasional Edusaintek*, 599–610.