Segmentasi Mahasiswa Berdasarkan Kebiasaan Belajar dan Karakteristik Personal Menggunakan PCA dan K-Means Clustering

Aldi Pramudya[1]

[1] G6401231003
Sekolah Sains Data, Matematika dan Informatika,
Departemen Ilmu Komputer, IPB University

Abstrak :

Penelitian ini bertujuan untuk melakukan segmentasi mahasiswa berdasarkan kebiasaan belajar dan karakteristik personal menggunakan metode Principal Component Analysis (PCA) dan K-Means Clustering. Dataset yang digunakan mencakup 493 observasi dengan 16 atribut, termasuk variabel demografis, prestasi akademik, dan aktivitas digital seperti durasi bermain game. Tahapan analisis meliputi pra-pemrosesan data, reduksi dimensi dengan PCA, dan pengelompokan data menggunakan K-Means Clustering. Hasil penelitian menunjukkan bahwa mahasiswa dapat dikelompokkan ke dalam tiga klaster utama, dengan karakteristik yang berbeda dalam hal performa akademik, kebiasaan belajar, dan aktivitas bermain game. Evaluasi menggunakan Silhouette Score menunjukkan nilai 0.12, mengindikasikan pemisahan klaster yang belum optimal. Temuan ini memberikan wawasan bagi institusi pendidikan untuk merancang intervensi yang lebih terarah, khususnya bagi mahasiswa dengan intensitas bermain game tinggi dan performa akademik rendah.

Kata Kunci: K-Means Clustering, Kebiasaan Belajar, PCA, Performa Akademik, Segmentasi Mahasiswa.

1 Pendahuluan

Dalam dunia pendidikan tinggi, pemahaman terhadap karakteristik mahasiswa menjadi aspek penting untuk meningkatkan kualitas pembelajaran. Mahasiswa memiliki latar belakang, kebiasaan belajar, serta pola perilaku yang beragam, yang secara langsung ataupun tidak langsung dapat memengaruhi capaian akademik mereka Aljaffer et al. (2024). Dalam konteks ini, pengelompokan atau segmentasi mahasiswa berdasarkan atribut-atribut tersebut menjadi penting agar institusi pendidikan dapat merancang pendekatan yang lebih tepat sasaran.

Salah satu metode yang dapat digunakan untuk memahami segmentasi mahasiswa adalah *unsupervised learning*, khususnya *clustering*. Metode ini memungkinkan peneliti untuk mengidentifikasi kelompok-kelompok mahasiswa yang memiliki kemiripan dalam berbagai aspek tanpa harus mengetahui label atau kategori sebelumnya. Dalam penelitian ini, metode *K-Means Clustering* dipilih untuk mengelompokkan mahasiswa berdasarkan kebiasaan belajar, aktivitas digital (seperti waktu bermain *game*), serta performa akademik mereka.

Agar analisis menjadi lebih efisien dan mudah divisualisasikan, metode *Principal Component Analysis (PCA)* digunakan untuk mereduksi dimensi data tanpa kehilangan informasi penting. Dengan kombinasi metode ini, penelitian ini bertujuan untuk menggali pola-pola tersembunyi dalam data mahasiswa dan mengidentifikasi kelompok-kelompok mahasiswa yang memiliki ciri khas tertentu.

1.1 Rumusan Masalah

1. Bagaimana segmentasi mahasiswa dapat dibentuk berdasarkan perilaku belajar dan faktor-faktor non-akademik seperti durasi bermain game?
2. Apakah terdapat karakteristik khusus pada masing-masing segmen mahasiswa yang dapat diidentifikasi menggunakan metode clustering?

1.2 Tujuan Penelitian

1. Mengelompokkan mahasiswa berdasarkan atribut seperti nilai akademik, kebiasaan belajar, dan aktivitas digital menggunakan metode K-Means Clustering.
2. Menyederhanakan dimensi data menggunakan PCA untuk visualisasi dan pemilihan fitur yang relevan.
3. Menginterpretasikan setiap segmen mahasiswa untuk mengetahui pola perilaku yang dominan.

2 Data

Dataset yang digunakan dalam penelitian ini adalah *Student Performance Metrics Dataset* Hasan et al. (2024). Dataset ini diterbitkan tahun 2024 dan mencakup 493 observasi mahasiswa dengan 16 atribut, mencakup variabel demografis, prestasi akademik, kondisi sosial-ekonomi, serta kegiatan ekstrakurikuler. Berikut adalah atribut yang terdapat pada dataset ini:

- **Department**: Jurusan akademik mahasiswa
- **Gender**: Jenis kelamin
- **HSC**: Skor ujian tingkat menengah atas
- **SSC**: Skor ujian tingkat menengah pertama
- **Income**: Pendapatan keluarga per bulan
- **Hometown**: Jenis daerah tempat tinggal (misalnya urban/rural)
- **Computer**: Kemampuan komputer
- **Preparation**: Waktu persiapan belajar harian
- **Gaming**: Waktu bermain game harian
- **Attendance**: Tingkat kehadiran kuliah
- **Job**: Status pekerjaan paruh waktu
- **English**: Kemampuan bahasa Inggris
- **Extra**: Partisipasi dalam kegiatan ekstrakurikuler
- **Semester**: Semester berjalan
- **Last**: Prestasi semester sebelumnya
- **Overall**: Indeks Prestasi Komulatif (IPK) keseluruhan

Dataset ini memungkinkan analisis hubungan antara variabel-variabel tersebut dan kinerja akademik mahasiswa.

3 Metode Penelitian

Analisis data mahasiswa dilakukan melalui beberapa tahan terstruktur. Pertama, data mentah harus melalui tahap *pra-pemrosesan* karena data mentah rentan terhadap gangguan, kerusakan, dan tidak konsisten. Data yang buruk dapat memengaruhi keakuratan dan menyebabkan prediksi yang salah, Sehingga perlu untuk meningkatkan kualitas data dengan *pra-pemrosesan* Maharana et al. (2022). Proses *pra-pemrosesan* mencakup transformasi data berikut:

- **Pengodean variabel kategorikal ordinal**: Pengodean variabel kategorikal ordinal atau *Integer Encoding* adalah strategi paling sederhana untuk mengkonversi data kategorikal ordinal menjadi data numerik

Pargent et al. (2022). Sebuah bilangan bulat (*integer*) diberikan kepada setiap kategori, asalkan jumlah kategori yang ada diketahui. Pengodean ini tidak menambahkan kolom baru ke data, tetapi menyiratkan urutan variabel yang mungkin tidak benar-benar ada Potdar et al. (2017). Sebagai contoh, pada penelitian Prasetyawan et al. (2025), atribut "Tingkat Pekerjaan" dengan kategori "Lokal", "Nasional", dan "Internasional" diubah menjadi 0, 1, dan 2 mengikuti urutannya.

- ***One-hot encoding* variabel kategori nominal**: *One-hot encoding* adalah teknik yang umum digunakan di bidang statistik dan machine learning, terutama saat menghadapi variabel kategorikal. Proses ini melibatkan representasi setiap kategori sebagai vektor biner. Dalam proses ini, vektor biner dibuat untuk setiap kategori unik, dengan semua elemen disetel ke nol kecuali yang sesuai dengan kategori pengamatan yang diberikan, yang disetel ke satu. Hal ini menghasilkan matriks vektor biner yang mewakili variabel kategorikal dalam kumpulan data (Jamell Ivor Samuels, 2024).
- **Standarisasi fitur numerik**: Dalam Principal Component Analysis (PCA), komponen utama (principal components) dibentuk sebagai kombinasi linear dari variabel-variabel asli. Namun, ketika variabel-variabel tersebut memiliki satuan (unit) pengukuran yang berbeda-beda, PCA bisa memberikan hasil yang kurang representatif. Hal ini karena PCA berfokus pada varians (keragaman) dari data, dan varians sangat dipengaruhi oleh skala pengukuranJolliffe et al. (2016).
  Misalnya, jika satu variabel diukur dalam ribuan (seperti pendapatan), dan yang lain dalam skala kecil (seperti skor 1–5), maka variabel dengan skala besar akan otomatis memiliki varians lebih tinggi, dan lebih "menonjol" dalam pembentukan komponen utama, meskipun secara substansi belum tentu lebih penting Jolliffe et al. (2016).
  Untuk mengatasi hal ini, biasanya dilakukan standarisasi data sebelum menjalankan PCA. Standarisasi dilakukan dengan:

  1. Mengurangi setiap nilai data dengan rata-ratanya (centring), dan
  2. Membagi hasilnya dengan standar deviasi masing-masing variabel (scaling).

  Hasilnya, semua variabel memiliki rata-rata nol dan deviasi standar satu, sehingga tidak ada variabel yang "mendominasi" hanya karena perbedaan skala. Dengan demikian, PCA menjadi lebih adil dan interpretasi komponen utama lebih bermakna.

3.1 Klasterisasi K-Means

K-means clustering merupakan salah satu metode pengelompokan data (clustering) yang banyak digunakan, di mana data dibagi ke dalam sejumlah klaster berdasarkan nilai rata-rata (mean) dari objek-objek dalam klaster tersebut (Ikotun et al., 2023). Fungsi objektifnya adalah meminimalkan within-cluster sum of squares (WCSS), yaitu jumlah kuadrat jarak (biasanya Euclidean) antara setiap titik data dengan centroid (rata-rata) klasternya. Dengan kata lain, algoritma ini berusaha agar data dalam setiap klaster sedekat mungkin dengan centroid masing-masing. Pada buku yang ditulis oleh VanderPlas (2016), proses K-Means bersifat iteratif dan dapat dijelaskan dalam langkah-langkah berikut:

1. **Menentukan jumlah klaster (k)**: Tentukan nilai k sesuai jumlah klaster yang diinginkan.
2. **Inisialisasi centroid awal**: Pilih secara acak k titik data sebagai centroid awal, atau gunakan metode *K-Means++* untuk hasil inisialisasi yang lebih baik.
3. **Penugasan data ke klaster**: Hitung jarak setiap titik data ke masing-masing centroid (umumnya menggunakan jarak Euclidean), lalu tetapkan setiap titik ke klaster dengan centroid terdekat.
4. **Pembaruan centroid**: Hitung rata-rata posisi seluruh anggota setiap klaster untuk mendapatkan centroid baru.
5. **Iterasi hingga konvergen**: Ulangi langkah penugasan dan pembaruan centroid hingga tidak ada perubahan signifikan pada anggota klaster atau posisi centroid (algoritma telah konvergen).

Dalam tiap iterasi K-Means mengkalkulasi jarak dari seluruh titik data ke setiap centroid untuk menentukan

pembagian yang optimal. Algoritma ini membuat asumsi bahwa klaster-klaster berbentuk sferis (isotropik) dan distribusi data normal searah (Gaussian) sehingga penggunaan jarak Euclidean relevan. Akibatnya, K-Means kurang efektif untuk klaster yang bentuknya tidak bulat atau data dengan fitur kategorikal tanpa encoding khusus. Namun, setelah pra-pemrosesan yang tepat, K-Means populer karena kesederhanaan dan efisiensinya bagi dataset berukuran besar.

3.2 Analisis Komponen Utama (PCA)

Analisis Komponen Utama (PCA) adalah teknik reduksi dimensi yang mentransformasikan kumpulan data berdimensi tinggi menjadi sejumlah komponen utama (principal components) yang saling ortogonal. Tujuan utama PCA adalah mempertahankan sebanyak mungkin varians asli data dengan jumlah fitur yang lebih sedikit (You et al., 2020). Langkah-langkah teknis PCA meliputi:

1. **Standarisasi data**: Sebelum PCA, fitur-fitur numerik distandarisasi sehingga setiap fitur memiliki rata-rata nol dan variansi satu.
2. **Matriks kovarians**: Hitung matriks kovarians dari data terstandarisasi untuk mengukur korelasi antar fitur.
3. **Dekomposisi eigen**: Lakukan dekomposisi eigen pada matriks kovarians untuk memperoleh eigenvektor dan eigenvalue. Eigenvektor menjadi arah komponen utama, sedangkan eigenvalue mencerminkan variansi data pada arah tersebut.
4. **Pemilihan komponen**: Urutkan eigenvektor berdasarkan nilai eigenvalue-nya. Komponen utama dengan eigenvalue terbesar menangkap bagian variansi data yang paling besar. Komponen utama pertama (PC1) menjelaskan variansi terbesar, diikuti oleh komponen kedua (PC2), dan seterusnya.
5. **Proyeksi data**: Proyeksikan data asli ke subruang yang dibentuk oleh beberapa komponen utama teratas. Jika dipilih p komponen, maka setiap titik data direpresentasikan dalam ruang berdimensi p dengan mengalikan data terstandarisasi dengan matriks eigenvektor terpilih.

Dengan dekomposisi eigen ini, PCA menyederhanakan data sambil mempertahankan pola terpentingnya. Kontribusi variansi dari tiap komponen utama dapat dihitung dari rasio eigenvalue terhadap total eigenvalue (total variansi). Misalnya, jika dua komponen utama pertama menangkap lebih dari 90% variansi, data dapat divisualisasikan dalam bidang dua dimensi yang melibatkan PC1 dan PC2. Pengurangan dimensi melalui PCA sangat berguna untuk visualisasi klaster. Hasil klaster K-Means dapat diplot pada 2 komponen utama teratas sehingga pola pengelompokan menjadi lebih mudah dipahami (Decheva et al., 2018).

3.3 Evaluasi Klaster

Untuk menentukan jumlah klaster optimal dan menilai kualitas hasil klasterisasi digunakan beberapa metode evaluasi klaster berikut:

- ***Elbow Method***: pendekatan visual untuk menilai jumlah klaster (K) yang optimal dengan menganalisis nilai Sum of Squared Errors (SSE) di dalam setiap klaster. Titik K optimal diidentifikasi dengan mencari "elbow" pada kurva yang menggambarkan nilai SSE untuk berbagai nilai K. Titik elbow ini menandakan penurunan SSE yang paling signifikan, yang menunjukkan titik di mana penambahan jumlah klaster selanjutnya memberikan penurunan SSE yang semakin kecil (diminishing returns) (Sugar et al., 2003; Umargono et al., 2020).
- ***Silhouette Score***: *Silhouette* menunjukkan objek mana yang berada dengan baik di dalam klasternya, dan objek mana yang hanya berada di antara dua klaster (Rousseeuw, 1987), mengimplementasikan bahwa silhouette dapat dijadikan sebagai *metric* evaluasi dari proses klastering. *Silhouette score* adalah angka yang didapatkan dari hasil evaluasi metode *silhouette* itu sendiri. Angka tersebut mengukur seberapa mirip suatu objek dengan klasternya sendiri dibandingkan dengan kemiripannya terhadap klaster lain (Januzaj et al., 2023). *Silhouette Score* memiliki rentang nilai -1 hingga 1. Dimana *Silhouette Score* mendekati -1

menandakan klastering yang buruk, sedangkan *Silhouette Score* mendekati 1 menandakan klastering yang baik.

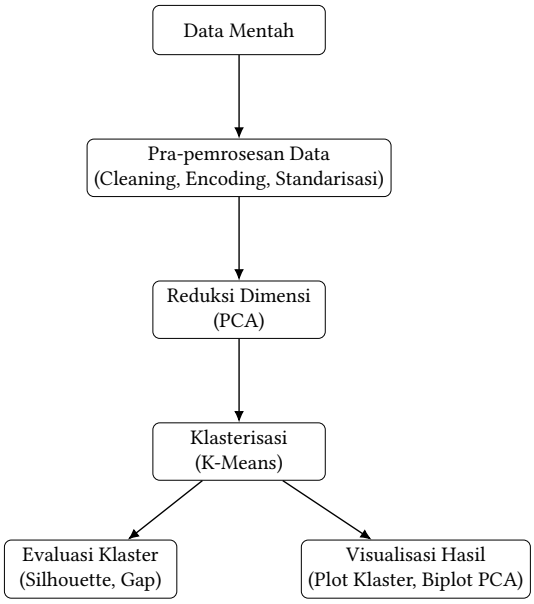
- ***Gap Statistic***: *Gap Statistic* adalah metode statistik yang digunakan untuk menentukan jumlah klaster optimal dengan membandingkan perubahan within-cluster dispersion (SSE) pada data asli dengan data acak yang tidak memiliki struktur klaster (Tibshirani et al., 2001). Prosedurnya melibatkan perhitungan SSE untuk berbagai nilai *k* pada data asli, lalu membandingkannya dengan SSE rata-rata dari beberapa dataset acak yang dihasilkan dengan distribusi serupa. Nilai *Gap* dihitung sebagai selisih logaritmik antara SSE acak dan SSE data asli. Jumlah klaster optimal dipilih pada titik di mana nilai *Gap* maksimum, yang menunjukkan bahwa struktur klaster pada data nyata jauh lebih baik daripada data acak, sehingga mengindikasikan adanya klaster yang valid.

3.4 Alur Analisis Data

Alur analisis data dalam penelitian ini dilakukan secara sistematis melalui beberapa tahapan utama sebagai berikut:

1. **Pra-pemrosesan Data:** Data mentah terlebih dahulu dibersihkan dan dipersiapkan, termasuk penanganan data hilang, pengodean variabel kategorikal (ordinal dan nominal), serta standarisasi fitur numerik agar seluruh variabel berada pada skala yang sebanding.
2. **Reduksi Dimensi (PCA):** Setelah data siap, dilakukan Analisis Komponen Utama (PCA) untuk mereduksi dimensi data dan mengekstrak fitur-fitur utama yang paling berkontribusi terhadap variasi data.
3. **Klasterisasi K-Means:** Data hasil reduksi dimensi kemudian dikelompokkan menggunakan algoritma K-Means. Penentuan jumlah klaster optimal dilakukan dengan metode Elbow.
4. **Evaluasi Klaster:** Hasil klasterisasi dievaluasi menggunakan metrik seperti Silhouette Score dan Gap Statistic untuk menilai kualitas dan validitas klaster yang terbentuk.
5. **Visualisasi Hasil:** Hasil akhir divisualisasikan menggunakan berbagai teknik visualisasi, seperti plot klaster dan biplot PCA, untuk memudahkan interpretasi pola pengelompokan dalam data.

Untuk memperjelas tahapan analisis data yang dilakukan, Tabel 1 berikut menyajikan diagram alur proses secara visual dari awal hingga akhir penelitian.



Grafik 1. Diagram alur analisis data dalam penelitian ini.

Setiap tahapan di atas diimplementasikan secara terstruktur menggunakan bahasa pemrograman R, sehingga seluruh proses dapat direplikasi dan diverifikasi oleh peneliti lain.

3.5 Implementasi dengan R

R adalah bahasa pemrograman dan lingkungan perangkat lunak yang banyak digunakan untuk analisis statistik, komputasi data, dan visualisasi (R Core Team, 2024). Dalam penelitian ini, implementasi analisis dilakukan menggunakan beberapa paket utama di R, antara lain `dplyr` untuk manipulasi data, `factoextra` dan `cluster` untuk analisis klasterisasi dan visualisasi, serta `corrplot` untuk visualisasi korelasi antar variabel. Paket-paket ini menyediakan fungsi-fungsi yang sangat mendukung proses pengolahan data, analisis statistik, serta penerapan metode machine learning seperti klasterisasi dan Principal Component Analysis (PCA) (Wickham, 2023; Szczesna, 2022).

R bersifat open source dan didukung oleh komunitas yang sangat aktif, sehingga tersedia banyak paket tambahan seperti `tidyverse`, `cluster`, `factoextra`, dan `corrplot` untuk analisis data dan visualisasi. Selain itu, penggunaan script di R memungkinkan proses analisis yang transparan, terstruktur, dan mudah direplikasi oleh peneliti lain.

3.5.1 Contoh Kode Implementasi di R Berikut adalah contoh kode R yang digunakan dalam penelitian ini untuk melakukan pra-pemrosesan data, penentuan jumlah klaster optimal, klasterisasi K-Means, analisis PCA, dan visualisasi hasil:

Listing 1: Contoh implementasi analisis klasterisasi dan PCA di R

```r
# Memuat library yang diperlukan
library(dplyr)
library(factoextra)
library(cluster)
library(corrplot)

# Pra-pemrosesan: encoding variabel ordinal & nominal
df_encoded <- df %>%
  mutate(
    Preparation_num = recode(Preparation, "0-1 Hour" = 1, "2-3 Hours" = 2, "More than 3 Hours" = 3),
    Gaming_num = recode(Gaming, "0-1 Hour" = 1, "2-3 Hours" = 2, "More than 3 Hours" = 3),
    Attendance_num = recode(Attendance, "Below 40%" = 1, "40%-59%" = 2, "60%-79%" = 3, "80%-100%" = 4),
    Semester_num = as.numeric(gsub("[^0-9]", "", Semester)),
    Income_num = recode(Income, "Low (Below 15,000)" = 1, "Lower middle (15,000-30,000)" = 2, "Upper middle (30,000-50,000)" = 3, "High (Above 50,000)" = 4),
    Job_num = ifelse(Job == "Yes", 1, 0),
    Extra_num = ifelse(Extra == "Yes", 1, 0),
    Hometown_num = ifelse(Hometown == "City", 1, 0),
    Gender_num = ifelse(Gender == "Male", 1, 0)
  )

# One-hot encoding untuk variabel Department
dept_onehot <- model.matrix(~ Department - 1, data = df)
```

```r
df_ready <- cbind(df_encoded, dept_onehot)

# Standarisasi data
df_scaled <- scale(df_ready)

# Menentukan jumlah klaster optimal dengan Elbow Method
set.seed(6969)
fviz_nbclust(df_scaled, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2) +
  labs(subtitle = "Elbow Method untuk menentukan jumlah cluster")

# Klasterisasi K-Means
kmeans_result <- kmeans(df_scaled, centers = 3, nstart = 25)

# Visualisasi hasil klaster
fviz_cluster(kmeans_result, data = df_scaled, geom = "point", ellipse.type =
    "convex", palette = "jco", ggtheme = theme_minimal())

# Analisis PCA dan visualisasi biplot
pca <- prcomp(df_scaled)
fviz_pca_biplot(pca, habillage = kmeans_result$cluster, addEllipses = TRUE,
   palette = "jco", ggtheme = theme_minimal())
```

Kode di atas menunjukkan alur utama analisis: mulai dari pra-pemrosesan data, encoding variabel, standarisasi, penentuan jumlah klaster optimal, klasterisasi K-Means, hingga visualisasi hasil klaster dan PCA. Seluruh proses dilakukan secara terstruktur dan dapat direplikasi, sehingga mendukung transparansi dan validitas hasil penelitian.

Seluruh kode R dan data yang digunakan dalam penelitian ini dapat diakses secara terbuka melalui repositori `https://github.com/stezd/paper-metkuan`. Kode tersebut didistribusikan dengan lisensi ***MIT No Attribution***, sehingga dapat digunakan, dimodifikasi, dan didistribusikan kembali secara bebas untuk keperluan apa pun tanpa batasan.

4 Hasil dan Pembahasan

Bagian ini menyajikan hasil analisis data serta interpretasi dari temuan penelitian. Hasil yang diperoleh dari proses klasterisasi, evaluasi, dan visualisasi akan dibahas secara sistematis.

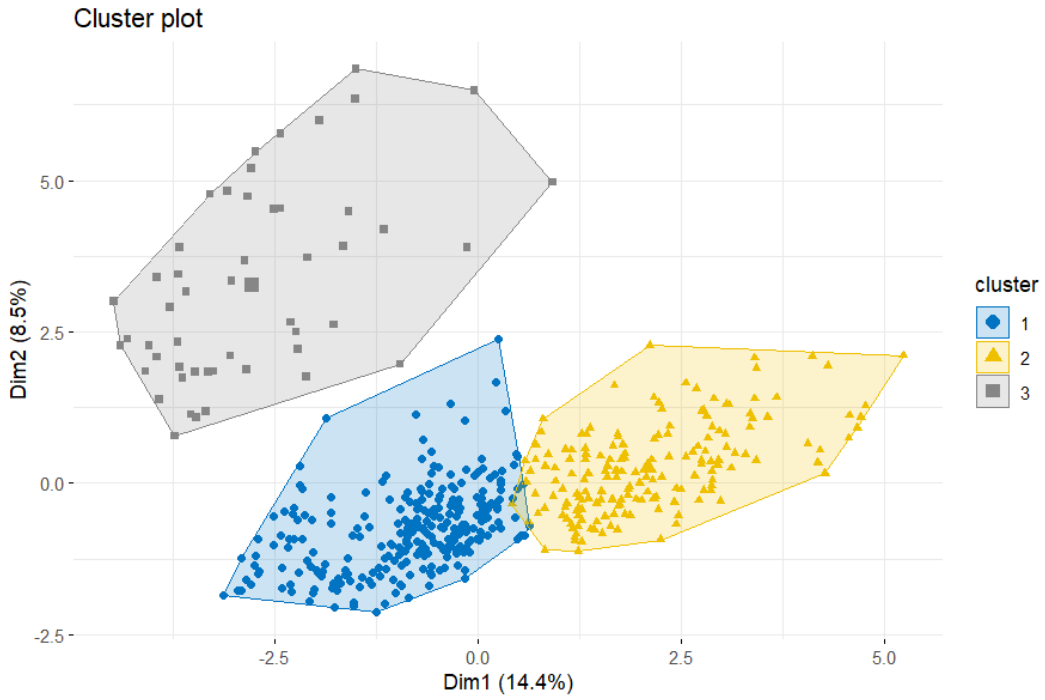
4.1 Hasil Klasterisasi

Berdasarkan *Elbow Method* (Grafik 3) dan *Gap Statistics* (Grafik 4), jumlah cluster optimal ditentukan sebanyak tiga. Walaupun terdapat elbow di angka 6 dan 9, angka 3 dipilih karena angka lebih dari 3 akan menyebabkan overfitting. Overfitting biasanya menyebabkan penurunan akurasi pada model (Webb, 2011). Setelah proses *scaling* dan penerapan *K-Means*, didapatkan hasil distribusi mahasiswa ke dalam 3 cluster sebagai berikut:

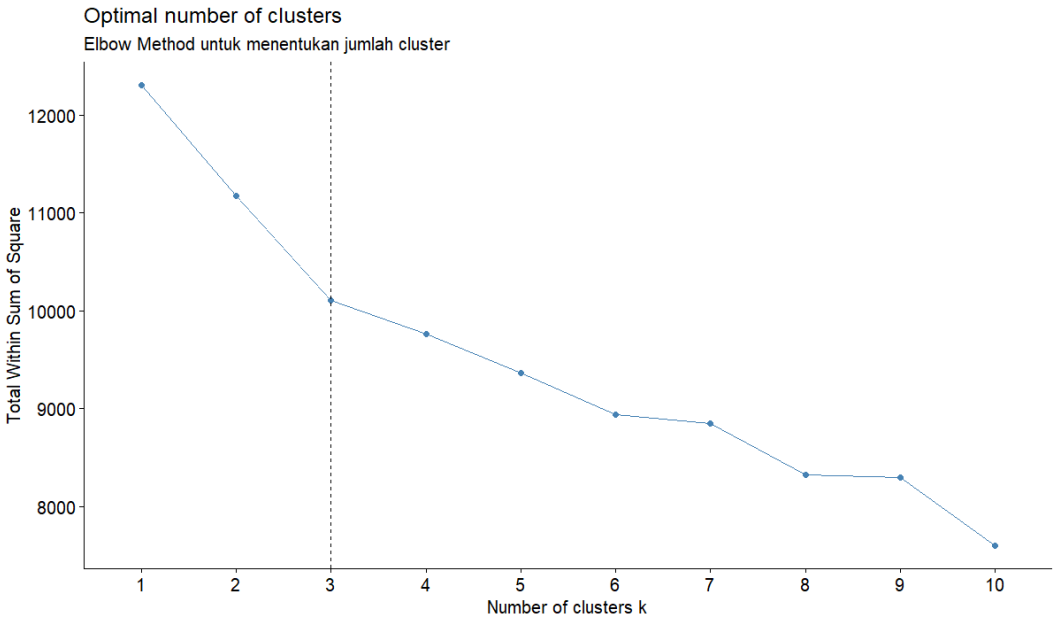
Pemilihan kolom yang pada Tabel 1 berdasarkan hasil visualisasi kontribusi tiap fitur.

Tabel 1
Distribusi Mahasiswa pada Setiap Cluster

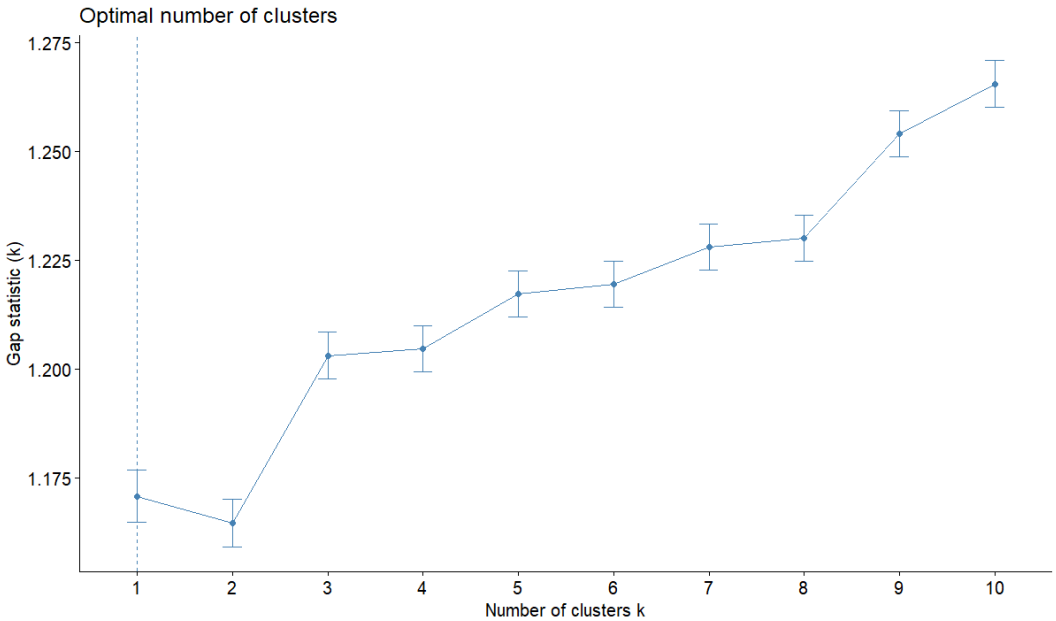
Cluster	Last	Overall	Attendance_num	Gaming_num	Preparation_num	Computer	Income_num
1	3.52	3.52	3.56	2.29	1.92	3.57	2.59
2	2.50	2.57	2.56	2.95	1.18	3.11	2.71
3	3.54	3.53	3.66	2.16	1.86	2.92	2.24



Grafik 2. Visualisasi hasil klasterisasi mahasiswa



Grafik 3. Elbow Method



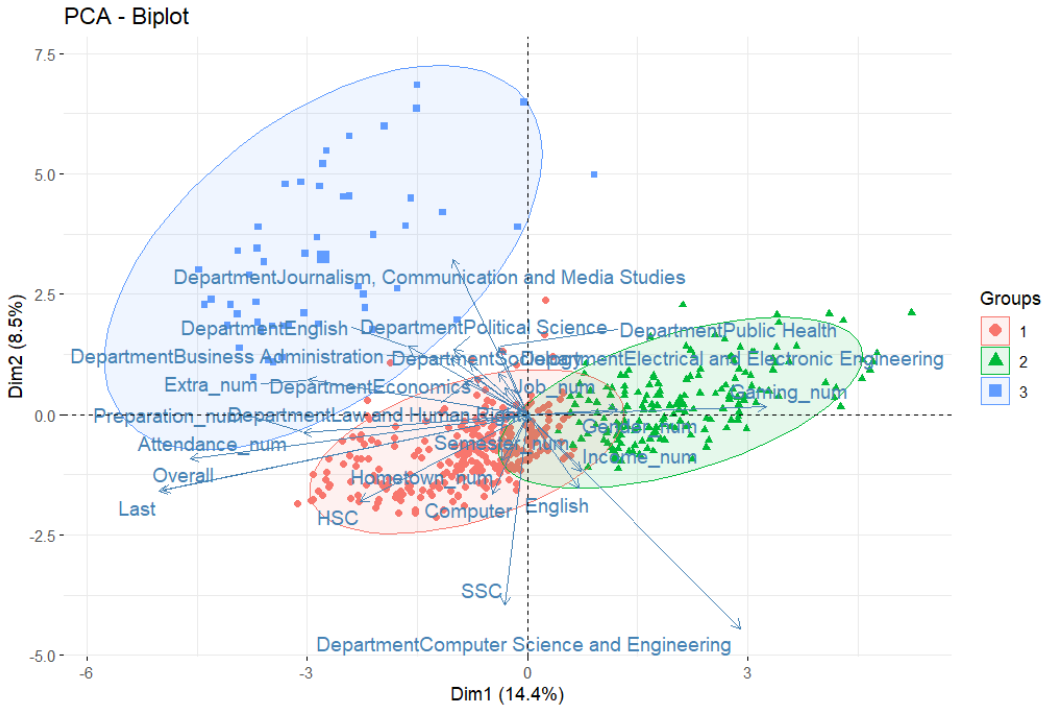
Grafik 4. Gap Statistics

4.2 PCA

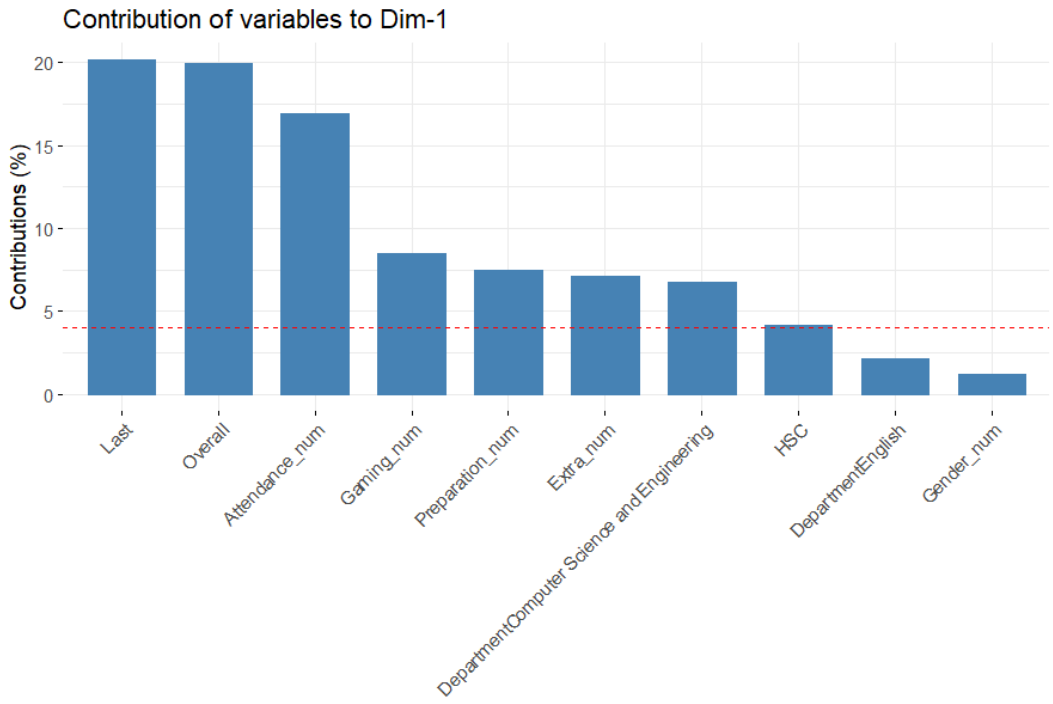
Dataset terdiri dari 16 fitur, sehingga visualisasi biplot menjadi kurang jelas dan informatif jika semua fitur ditampilkan sekaligus (lihat Grafik 5). Oleh karena itu, biplot divisualisasikan dengan memisahkan fitur-fitur berdasarkan tema yang serupa. Berikut adalah hasil visualisasinya:

- Grafik 8 menampilkan biplot yang hanya memuat fitur-fitur akademik, yaitu "HSC", "SSC", "Computer", "English", "Last", dan "Overall".
- Grafik 9 menampilkan biplot yang hanya memuat fitur-fitur kebiasaan dan demografi, yaitu "Preparation_num", "Gaming_num", "Attendance_num", "Semester_num", "Income_num", "Job_num", "Extra_num", "Hometown_num", dan "Gender_num".
- Grafik 10 menampilkan biplot berdasarkan variabel departemen masing-masing mahasiswa dalam dataset.

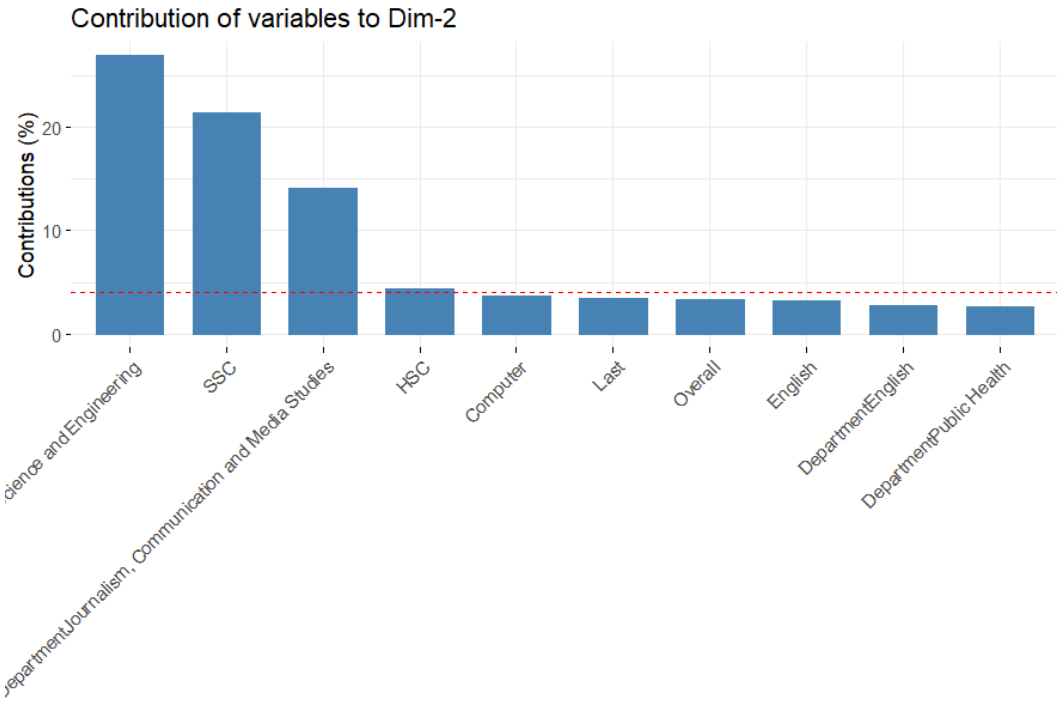
Berdasarkan Grafik 6 dan Grafik 7, dapat dilihat bahwa fitur "Last", "Overall", "Attendance_num", dan "Gaming_num" memberikan kontribusi terbesar pada Dimensi 1, yang merepresentasikan aspek performa akademik mahasiswa. Sementara itu, Dimensi 2 memisahkan mahasiswa berdasarkan latar belakang departemen, khususnya antara kelompok STEM dan non-STEM. Hal ini juga tercermin pada Grafik 5, di mana panah fitur "Overall" dan "Last" mengarah ke kiri, sedangkan "Gaming_num" mengarah ke kanan, mengindikasikan adanya hubungan negatif antara waktu bermain game dan performa akademik. Selain itu, pada Grafik 10 terlihat bahwa mahasiswa dari "Department Computer Science and Engineering" memiliki arah panah yang berbeda dengan departemen lainnya, menunjukkan adanya perbedaan karakteristik yang cukup signifikan. Ini dapat disebabkan karena jumlah mahasiswa ilmu komputer pada dataset ini sangat banyak jika dibandingkan dengan mahasiswa dari departemen lain.



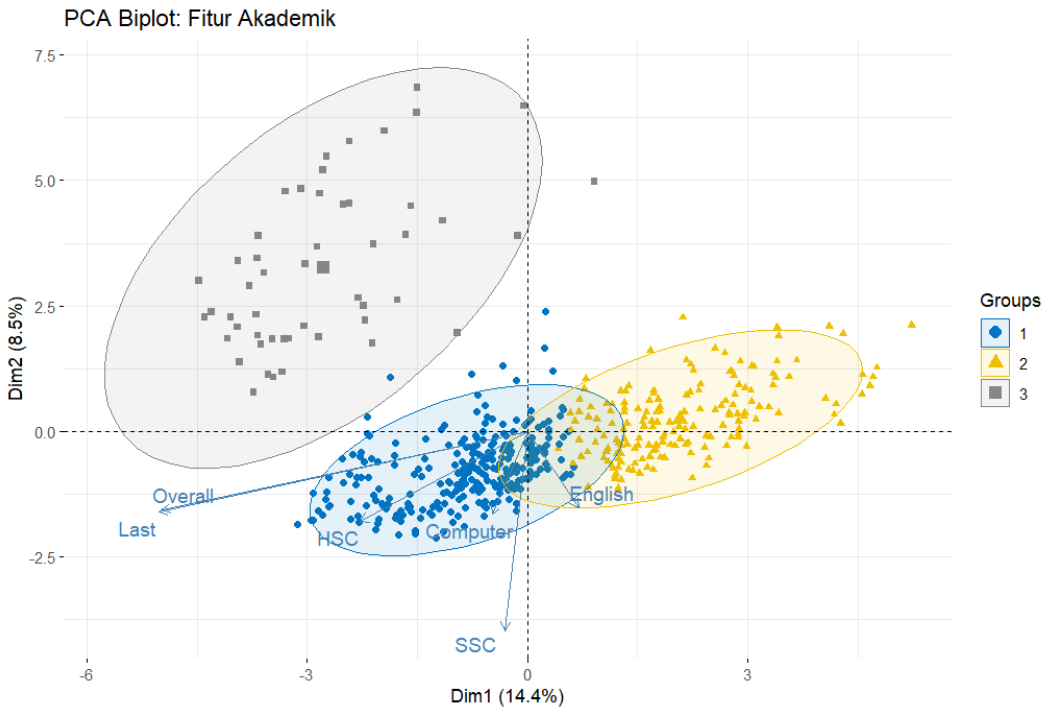
Grafik 5. Contoh biplot dengan terlalu banyak fitur sehingga tampak berantakan



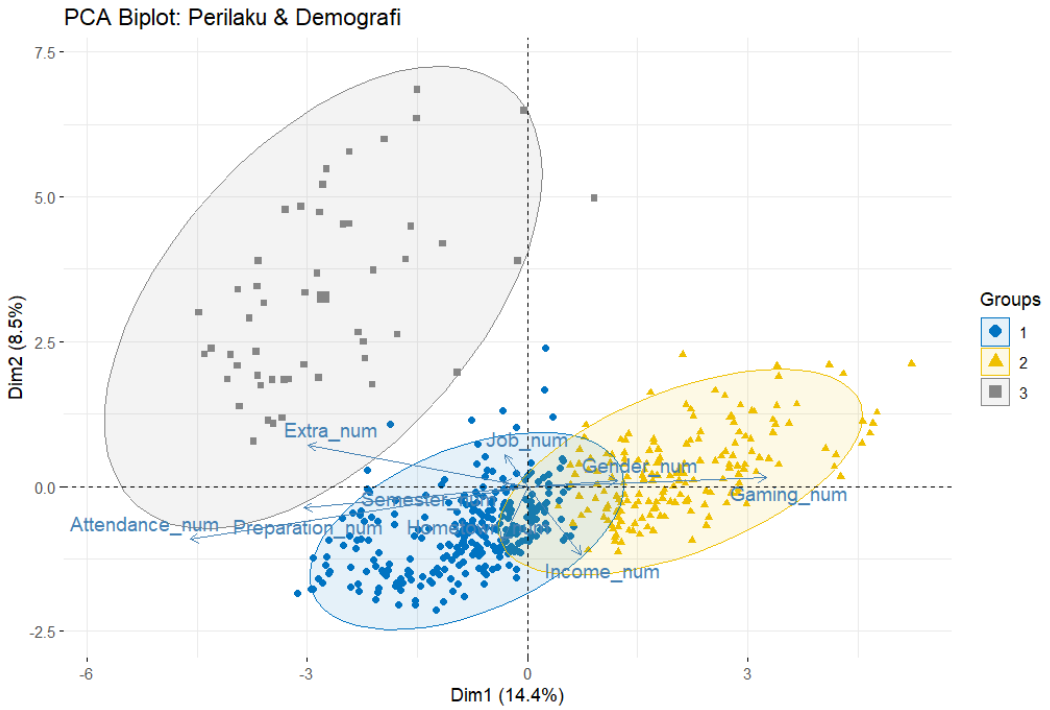
Grafik 6. Visualisasi kontribusi tiap fitur terhadap pembentukan cluster



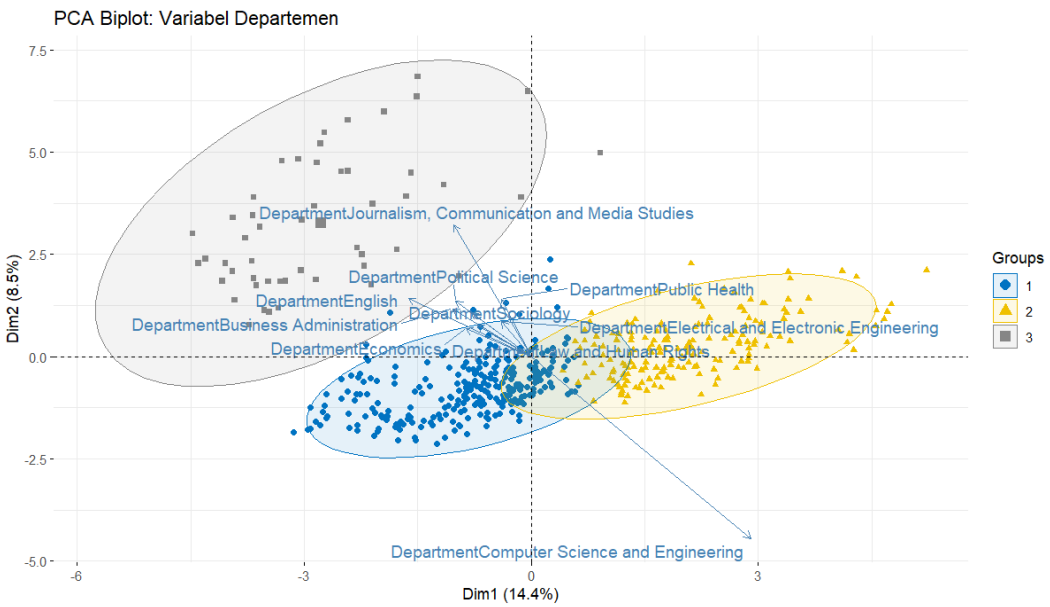
Grafik 7. Visualisasi kontribusi tiap fitur terhadap pembentukan cluster



Grafik 8. Biplot fitur akademik



Grafik 9. Biplot fitur perilaku dan demografi



Grafik 10. Biplot variabel departemen

4.3 Evaluasi Clustering

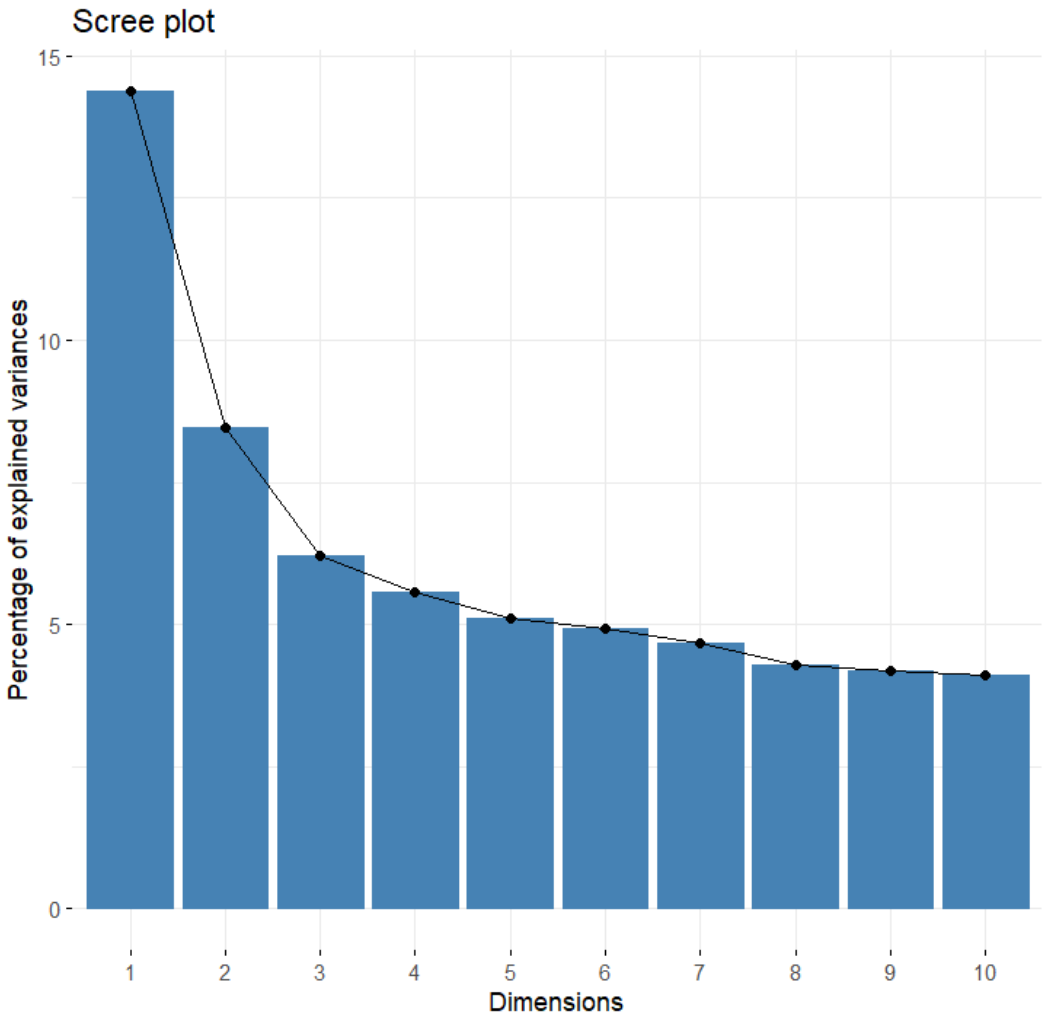
Grafik 11. Silohutte plot

Evaluasi hasil clustering dilakukan menggunakan metode Silhouette Analysis. Rata-rata nilai silhouette width yang diperoleh sebesar 0.12, menandakan pemisahan antar klaster masih kurang kuat dan terdapat potensi tumpang tindih antar anggota klaster. Hal ini mengindikasikan bahwa hasil klasterisasi K-Means belum sepenuhnya mencerminkan struktur alami data.

Meskipun demikian, berdasarkan Elbow Method dan Gap Statistic (Grafik 3 dan Grafik 4), jumlah klaster optimal tetap berada pada angka 3, sehingga K-Means masih dapat digunakan sebagai dasar segmentasi awal.

Validasi tambahan melalui visualisasi PCA biplot juga memperlihatkan adanya pemisahan klaster, meskipun tidak terlalu tegas. Untuk memperoleh hasil yang lebih baik, metode clustering lain seperti Hierarchical Clustering atau DBSCAN dapat dipertimbangkan sebagai alternatif perbandingan.

4.4 Evaluasi PCA



Grafik 12. Scree plot

Berdasarkan Grafik 12, terlihat adanya elbow pada dimensi ke-3. Namun, Tabel 2 menunjukkan bahwa diperlukan delapan komponen utama (PC1–PC8) untuk menjelaskan lebih dari 50% total variansi. Hal ini mengindikasikan bahwa tidak terdapat satu atau dua komponen yang secara dominan merepresentasikan data, sehingga variansi tersebar pada banyak dimensi. Dengan demikian, visualisasi PCA dua dimensi (PC1 vs PC2) memang kurang mampu menangkap keseluruhan struktur data, namun tetap bermanfaat untuk eksplorasi awal pola klaster dan arah kontribusi variabel.

Tabel 2
Hasil PCA: Proporsi Variansi pada Komponen Utama

Komponen Utama	Simpangan Baku	Proporsi Variansi	Akumulasi Variansi
PC1	1.8960	14.38%	14.38%
PC2	1.4549	8.47%	22.85%
PC3	1.2471	6.22%	29.07%
PC4	1.1790	5.56%	34.63%
PC5	1.1291	5.10%	39.73%
PC6	1.1098	4.93%	44.65%
PC7	1.0806	4.67%	49.33%
PC8	1.0358	4.29%	53.62%

4.5 Pembahasan

Hasil analisis menunjukkan bahwa mahasiswa dalam dataset dapat dikelompokkan ke dalam tiga klaster utama, meskipun pemisahan antar klaster masih kurang tegas sebagaimana tercermin dari nilai silhouette yang rendah (0.12). Hal ini mengindikasikan adanya kompleksitas struktur data yang belum sepenuhnya terakomodasi oleh metode K-Means, kemungkinan akibat adanya tumpang tindih antar variabel atau hubungan non-linier antar fitur.

Meskipun demikian, hasil clustering yang dihasilkan model (lihat Grafik 2) tetap dapat diinterpretasikan sebagai berikut:

- Klaster biru didominasi oleh mahasiswa dari "Department Computer Science and Engineering" dengan performa akademik tinggi. Berdasarkan Tabel 1, klaster 1 berisi mahasiswa dengan GPA $\geq$ 3,5, tingkat kehadiran baik, frekuensi bermain game rendah, persiapan belajar optimal, serta kemampuan komputer yang baik.
- Klaster kuning juga didominasi oleh mahasiswa dari "Department Computer Science and Engineering", namun dengan performa akademik lebih rendah. Klaster ini mencakup mahasiswa dengan GPA sekitar 2, tingkat kehadiran rendah, frekuensi bermain game lebih tinggi, persiapan belajar kurang, serta kemampuan komputer yang sedang.
- Klaster merah terdiri dari mahasiswa dengan performa akademik tinggi, tingkat kehadiran dan persiapan belajar yang baik, namun kemampuan komputer relatif lebih rendah dibandingkan kedua klaster lainnya. Klaster ini umumnya berasal dari departemen non-STEM.

Perlu dicatat bahwa GPA yang diamati merupakan indikator performa akademik yang dapat mengandung bias sistematis, sehingga kurang tepat jika dijadikan satu-satunya dasar untuk menilai ketimpangan penilaian. Analisis menggunakan model logistik terhadap GPA memberikan gambaran yang lebih komprehensif mengenai performa akademik. Ketika membandingkan mahasiswa dengan kemampuan akademik yang setara, ditemukan bahwa mahasiswa STEM cenderung memperoleh nilai dan GPA yang lebih rendah. Hal ini terutama disebabkan oleh standar penilaian yang lebih ketat pada mata kuliah STEM. Temuan ini sejalan dengan penelitian oleh Tomkin et al., 2022, yang menunjukkan bahwa GPA mahasiswa non-STEM cenderung lebih tinggi dibandingkan mahasiswa STEM karena mata kuliah di bidang STEM umumnya memiliki tingkat kesulitan yang lebih tinggi.

Dari visualisasi PCA dan kontribusi fitur terhadap dimensi utama, didapati bahwa variabel "Gaming_num" dan "Preparation_num" memiliki hubungan yang signifikan terhadap nilai GPA "Overall". Mahasiswa dengan waktu bermain game yang tinggi cenderung memiliki nilai GPA yang lebih rendah, sedangkan mereka yang meluangkan waktu lebih banyak untuk persiapan belajar cenderung memiliki GPA yang lebih tinggi. Pola ini terlihat konsisten pada arah panah dalam biplot, di mana "Gaming_num" berlawanan arah dengan "Overall" dan "Last", menunjukkan korelasi negatif. Sebaliknya, "Preparation_num" cenderung searah dengan variabel akademik, mengindikasikan pengaruh positif terhadap performa.

Temuan ini juga sejalan dengan hasil penelitian eksperimental oleh Weis et al., 2010, yang menunjukkan bahwa anak laki-laki yang diberikan akses ke video game cenderung menghabiskan lebih banyak waktu untuk bermain dan lebih sedikit waktu untuk aktivitas akademik setelah sekolah. Studi tersebut menemukan bahwa kelompok yang langsung menerima konsol video game memiliki skor membaca dan menulis yang lebih rendah serta lebih banyak masalah akademik menurut laporan guru dibandingkan kelompok kontrol. Selain itu, durasi bermain video game terbukti menjadi mediator antara kepemilikan video game dan penurunan hasil akademik. Hasil ini memberikan bukti bahwa video game dapat menggantikan waktu untuk aktivitas edukatif dan berpotensi menghambat perkembangan keterampilan membaca dan menulis pada anak-anak.

Temuan ini semakin menegaskan bahwa perilaku belajar dan pengelolaan waktu, khususnya terkait kebiasaan bermain game dan persiapan belajar, berperan penting dalam pencapaian akademik mahasiswa. Selain faktor akademik, variabel seperti departemen dan latar belakang demografis juga berkontribusi terhadap pembentukan klaster, meskipun pengaruhnya tidak sebesar faktor perilaku.

Walaupun hasil clustering belum sepenuhnya optimal, pendekatan ini tetap memberikan gambaran awal yang bermanfaat untuk segmentasi mahasiswa. Informasi ini dapat dimanfaatkan institusi pendidikan untuk merancang intervensi yang lebih terarah, misalnya dengan memberikan dukungan tambahan kepada kelompok mahasiswa yang berada pada klaster dengan skor akademik rendah dan intensitas bermain game yang tinggi.

Ke depannya, penerapan metode klasterisasi alternatif seperti Hierarchical Clustering atau DBSCAN, serta penggunaan teknik reduksi dimensi non-linier seperti t-SNE atau UMAP, dapat dipertimbangkan untuk memperoleh klaster yang lebih representatif dan mudah diinterpretasikan.

5 Simpulan dan Saran

5.1 Simpulan

Penelitian ini berhasil mengelompokkan mahasiswa ke dalam tiga klaster utama berdasarkan karakteristik akademik, perilaku, dan demografi menggunakan metode K-Means dan analisis PCA. Hasil analisis menunjukkan bahwa variabel seperti frekuensi bermain game, persiapan belajar, dan kehadiran memiliki pengaruh signifikan terhadap performa akademik mahasiswa. Klaster yang terbentuk memperlihatkan adanya perbedaan karakteristik yang cukup jelas, terutama antara mahasiswa dari departemen STEM dan non-STEM. Namun, nilai silhouette yang rendah mengindikasikan bahwa pemisahan antar klaster masih kurang optimal dan terdapat potensi tumpang tindih antar anggota klaster.

Selain itu, ditemukan bahwa mahasiswa dengan waktu bermain game yang tinggi cenderung memiliki nilai akademik yang lebih rendah, sedangkan mereka yang lebih banyak mempersiapkan diri untuk belajar cenderung memiliki performa akademik yang lebih baik. Temuan ini sejalan dengan penelitian sebelumnya yang menyoroti pentingnya perilaku belajar dan pengelolaan waktu dalam pencapaian akademik.

5.2 Saran

Berdasarkan hasil penelitian, beberapa saran yang dapat diberikan adalah sebagai berikut:

- Institusi pendidikan disarankan untuk memberikan perhatian khusus kepada mahasiswa yang berada pada klaster dengan performa akademik rendah dan intensitas bermain game yang tinggi, misalnya melalui program pendampingan belajar atau konseling manajemen waktu.
- Penelitian selanjutnya dapat mempertimbangkan penggunaan metode klasterisasi lain seperti Hierarchical Clustering atau DBSCAN, serta teknik reduksi dimensi non-linier seperti t-SNE atau UMAP untuk memperoleh klaster yang lebih representatif dan mudah diinterpretasikan.
- Perlu dilakukan analisis lebih lanjut terhadap faktor-faktor lain yang mungkin memengaruhi performa akademik, seperti motivasi belajar, lingkungan keluarga, atau aktivitas ekstrakurikuler.
- Pengumpulan data dengan jumlah sampel yang lebih besar dan distribusi yang lebih merata antar departemen dapat meningkatkan validitas hasil klasterisasi.

Daftar Pustaka

Aljaffer, Mohammed A. et al. (Aug. 2024). "The impact of study habits and personal factors on the academic achievement performances of medical students". In: *BMC Medical Education* 24.1. ISSN: 1472-6920. DOI: 10.1186/s12909-024-05889-y.

Decheva, Desislava and Lars Linsen (2018). "Data Aggregation and Distance Encoding for Interactive Large Multidimensional Data Visualization". In: *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, pp. 225–235. DOI: 10.5220/0006602502250235.

Hasan, Tahmid, Md.Mahmud Hasan, and Tahbib Manzoor (2024). *Student Performance Metrics Dataset*. DOI: 10.17632/5B82YTZ489.1.

Ikotun, Abiodun M. et al. (Apr. 2023). "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data". In: *Information Sciences* 622, pp. 178–210. ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.11.139.

Jamell Ivor Samuels (2024). "One-Hot Encoding and Two-Hot Encoding: An Introduction". en. In: DOI: 10.13140/RG.2.2.21459.76327.

Januzaj, Ylber, Edmond Beqiri, and Artan Luma (Apr. 2023). "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique". In: *International Journal of Online and Biomedical Engineering (iJOE)* 19.04, pp. 174–182. ISSN: 2626-8493. DOI: 10.3991/ijoe.v19i04.37059.

Jolliffe, Ian T. and Jorge Cadima (Apr. 2016). "Principal component analysis: a review and recent developments." eng. In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 374 (2065), p. 20150202.

Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade (June 2022). "A review: Data pre-processing and data augmentation techniques". In: *Global Transitions Proceedings* 3.1, pp. 91–99. ISSN: 2666-285X. DOI: 10.1016/j.gltp.2022.04.020.

Pargent, Florian et al. (Mar. 2022). "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features". In: *Computational Statistics* 37.5, pp. 2671–2692. ISSN: 1613-9658. DOI: 10.1007/s00180-022-01207-6.

Potdar, Kedar, Taher S., and Chinmay D. (Oct. 2017). "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers". In: *International Journal of Computer Applications* 175.4, pp. 7–9. ISSN: 0975-8887. DOI: 10.5120/ijca2017915495.

Prasetyawan, Daru, Agus Mulyanto, and Rahmadhan Gatra (Apr. 2025). "Pemetaan Lintasan Karir Alumni Berdasarkan Analisis Cluster: Kombinasi K-Means dan Reduksi Dimensi Autoencoder". In: *Edumatic: Jurnal Pendidikan Informatika* 9.1, pp. 198–207. DOI: 10.29408/edumatic.v9i1.29713. URL: https://e-journal.hamzanwadi.ac.id/index.php/edumatic/article/view/29713.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rousseeuw, Peter J. (Nov. 1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

Sugar, Catherine A and Gareth M James (Sept. 2003). "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach". In: *Journal of the American Statistical Association* 98.463, pp. 750–763. ISSN: 1537-274X. DOI: 10.1198/016214503000000666.

Szczesna, Karolina (2022). *K-Means Clustering in R*. Accessed: 2025-06-02. URL: https://rpubs.com/KarolinaSzczesna/862710.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (July 2001). "Estimating the Number of Clusters in a Data Set Via the Gap Statistic". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 63.2, pp. 411–423. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00293.

Tomkin, Jonathan H. and Matthew West (Mar. 2022). "STEM courses are harder: evaluating inter-course grading disparities with a calibrated GPA model". In: *International Journal of STEM Education* 9.1. ISSN: 2196-7822. DOI: 10.1186/s40594-022-00343-1.

Umargono, Edy, Jatmiko Endro Suseno, and S.K Vincensius Gunawan (2020). "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula". In: *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*. isstec-19. Atlantis Press. DOI: 10.2991/assehr.k.201010.019.

VanderPlas, Jake (2016). *Python data science handbook. Essential tools for working with data*. First edition. Beijing: O'Reilly. 1529 pp. ISBN: 9781491912140.

Webb, Geoffrey I. (2011). "Overfitting". In: *Encyclopedia of Machine Learning*. Springer US, pp. 744–744. ISBN: 9780387301648. DOI: 10.1007/978-0-387-30164-8_623.

Weis, Robert and Brittany C. Cerankosky (Feb. 2010). "Effects of Video-Game Ownership on Young Boys' Academic and Behavioral Functioning: A Randomized, Controlled Study". In: *Psychological Science* 21.4, pp. 463–470. ISSN: 1467-9280. DOI: 10.1177/0956797610362670.

Wickham, Hadley (2023). *R for data science. Import, tidy, transform, visualize, and model data*. Ed. by Mine Çetinkaya-Rundel and Garrett Grolemund. 2nd edition. Literaturangaben. - Index. Beijing: O'Reilly. 548 pp. ISBN: 9781492097402.

You, Shingchern D. and Ming-Jen Hung (June 2020). "Reducing Dimensionality of Spectro-Temporal Data by Independent Component Analysis". In: *2020 2nd International Conference on Computer Communication and the Internet (ICCCI)*. IEEE, pp. 93–97. DOI: 10.1109/iccci49374.2020.9145984.