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 :

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent porttitor arcu luctus, imperdiet urna iaculis, mattis eros. Pellentesque iaculis odio vel nisl ullamcorper, nec faucibus ipsum molestie. Sed dictum nisl non aliquet porttitor. Etiam vulputate arcu dignissim, finibus sem et, viverra nisl. Aenean luctus congue massa, ut laoreet metus ornare in. Nunc fermentum nisi imperdiet lectus tincidunt vestibulum at ac elit. Nulla mattis nisl eu malesuada suscipit. Aliquam arcu turpis, ultrices sed luctus ac, vehicula id metus. Morbi eu feugiat velit, et tempus augue. Proin ac mattis tortor. Donec tincidunt, ante rhoncus luctus semper, arcu lorem lobortis justo, nec convallis ante quam quis lectus. Aenean tincidunt sodales massa, et hendrerit tellus mattis ac. Sed non pretium nibh. Donec cursus maximus luctus. Vivamus lobortis eros et massa porta porttitor.

Kata Kunci: Keyword A, Keyword B, Keyword C.

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 Data***:

4 Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Table 1
Floating point benchmark. $\mathbf{R}_{max}$: the performance in GFlops for the largest problem run on a machine; $\mathbf{N}_{max}$: the size of the largest problem run on a machine; $\mathbf{N}_{1/2}$: the size where half the $\mathbf{R}_{max}$ execution rate is achieved; $\mathbf{R}_{peak}$: the theoretical peak performance in GFlops.

Linpack Benchmark (Full precision)	Proc. or Cores	$\mathbf{R}_{max}$ GFlops	$\mathbf{N}_{max}$ Order	$\mathbf{R}_{peak}$ GFlops
Thinking Machine CM-5	32	1,900	9216	4
Pentium 4 3.0 GHz	1	4,730	7600	6
IBM Cell BE 3.2 GHz	9	98,05	4096	204,8 (32b) 14,6 (64b)
Thinking Machine SD-3	36	1,120	6728	3

5 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

6 Conclusion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Acknowledgements

This research received support during the XXX course, instructed by Professor XXX, PhD at the School of XXX, XXX.

References

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.

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.

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.

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.