"Analisis Pengelompokan Mahasiswa Berdasarkan Kebiasaan Pendidikan Menggunakan Metode K-Means Clustering, PCA, dan Visualisasi Heatmap"

Oleh :
Muh Farid FB

DEPARTEMEN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

BAB 1: Pendahuluan

1.1 Latar Belakang

Era globalisasi dan urbanisasi membuat kehidupan masyarakat semakin kompleks dengan perubahan sosial, budaya, dan ekonomi yang pesat, yang berdampak pada kesejahteraan manusia (Wajar, *et al*., 2022). Dalam konteks ini, perguruan tinggi berperan vital dalam mengembangkan kemampuan individu untuk menjadi tenaga profesional yang kompeten, serta menyiapkan peserta didik menjadi anggota masyarakat yang mampu menerapkan, mengembangkan, dan menciptakan ilmu pengetahuan, teknologi, dan seni (Rulanggi, *et al*., 2021).

Mahasiswa, sebagai komponen penting dalam pendidikan, berperan sebagai subjek dan objek dalam proses belajar mengajar. Performa belajar mereka dipengaruhi oleh kemampuan mengendalikan diri dan beradaptasi dengan lingkungan (Fatimah, *et al.,* 2021). Performa akademik mencakup prestasi seperti nilai dan IPK, serta keterampilan yang diperoleh selama studi. Penelitian ini menggunakan metode kuantitatif untuk menganalisis variabel seperti status hubungan (punya pasangan), jam belajar mingguan, frekuensi membaca, kehadiran di kelas, kebiasaan mencatat, mendengarkan di kelas, dan IPK terakhir.

Status hubungan mempengaruhi fokus dan keseimbangan antara kehidupan pribadi dan akademik. Jam belajar mingguan dan frekuensi membaca mencerminkan usaha akademik. Kehadiran di kelas, kebiasaan mencatat, dan mendengarkan aktif menunjukkan keterlibatan dalam pembelajaran. IPK terakhir mencerminkan hasil kumulatif usaha akademik. Memahami interaksi variabel-variabel ini memberikan wawasan tentang faktor-faktor yang mempengaruhi performa akademik mahasiswa dan dapat digunakan untuk mengembangkan strategi peningkatan kualitas belajar dan kesejahteraan akademik, yang sangat relevan dalam era globalisasi dan urbanisasi saat ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, dapat dirumuskan beberapa masalah utama yang akan dipecahkan dalam penelitian ini. Adapun rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana pengelompokan mahasiswa berdasarkan punya pasangan, jam belajar mingguan, frekuensi membaca, kehadiran di kelas, mencatat saat kelas berlangsung, mendengarkan, dan IPK terakhir menggunakan metode analisis klaster K-Means?
2. Apa saja karakteristik utama yang membedakan setiap klaster mahasiswa, dan bagaimana faktor-faktor tersebut mempengaruhi performa akademik mahasiswa?
3. Bagaimana keterhubungan variabel satu dan lainnya melalui visualisasi data Heatmap?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah diatas, peneliti memiliki beberapa tujuan dalam penelitian ini. Adapun, tujuan tersebut di antaranya:

1. Mengetahui klaster atau kelompok mahasiswa yang terbentuk berdasarkan variabel punya pasangan, jam belajar mingguan, frekuensi membaca, kehadiran di kelas, mencatat saat kelas berlangsung, mendengarkan, dan IPK terakhir

2. Mengidentifikasi dan mengetahui karakteristik utama dalam perbedaan setiap klaster mahasiswa serta faktor-faktor yang mempengaruhi performa akademik mahasiswa
3. Mengetahui keterhubungan variabel satu dan lainnya pada performa mahasiswa

1.4 Manfaat Penelitian

Manfaat penelitian ini terbagi menjadi tiga: bagi institusi pendidikan, dapat merancang strategi pembelajaran lebih efektif sesuai kebutuhan mahasiswa. Bagi peneliti selanjutnya, temuan ini dapat menjadi dasar untuk penelitian lebih lanjut mengenai faktor-faktor yang mempengaruhi performa akademik mahasiswa. Sementara itu, pemerintah dapat memanfaatkan data empiris dari penelitian ini untuk merumuskan kebijakan pendidikan tinggi yang lebih baik.

BAB 2: Deskripsi dan Sumber Data

2.1 Deskripsi Data

Penelitian ini menggunakan data sekunder yang diperoleh dari situs Kaggle dengan judul "Predict students' end-of-term performances using ML techniques". Data ini terdiri dari 32 kolom dan 145 baris. Pada kolom 1-10 merupakan pertanyaan pribadi, 11-16 merupakan pertanyaan yang berhubungan dengan keluarga, dan sisanya merupakan kebiasaan dan pengalaman pelajar/mahasiswa.

2.2 Sumber Data

Data penelitian ini diperoleh dari situs Kaggle, sebuah platform online yang menyediakan berbagai macam dataset untuk keperluan pembelajaran mesin dan analisis data. Dataset judul "Predict students' end-of-term performances using ML techniques" diunggah oleh pengguna Joakim Arvidsson dua bulan yang lalu dan dapat diakses secara gratis di https://www.kaggle.com/datasets/joebeachcapital/students-performance/data.

BAB 3: Alat dan Metode Penelitian

3.1 Alat Penelitian

Penelitian ini menggunakan analisis clustering k-means untuk pengelompokan mahasiswa berdasarkan variabel punya pasangan, jam belajar mingguan, frekuensi membaca, kehadiran di kelas, mencatat saat kelas berlangsung, mendengarkan, dan IPK terakhir guna memahami pola kebiasaan belajar mahasiswa dan pengaruh faktor terhadap performa akademik mereka. Melalui perangkat lunak RStudio, penelitian ini juga memanfaatkan Principal Component Analysis (PCA) untuk memperdalam wawasan. PCA membantu mereduksi dimensi data dan mengidentifikasi komponen utama yang menjelaskan variabilitas terbesar. Hasilnya menunjukkan hubungan yang signifikan antara pola kebiasaan belajar dan performa akademik mahasiswa, dengan klaster yang memiliki kebiasaan belajar yang serupa cenderung mencapai tingkat IPK yang lebih tinggi dibandingkan dengan klaster lainnya. Selain itu, analisis PCA mengungkapkan dimensi-dimensi utama yang paling mempengaruhi performa akademik, memberikan wawasan lebih dalam tentang faktor-faktor yang penting dalam meningkatkan prestasi belajar mahasiswa.

3.2 Metode Penelitian

Teknik data mining yang digunakan pada penelitian ini adalah metode clustering K-Means dengan penerapan algoritma reduksi dimensi Principal Component Analysis (PCA). Metode

K-Means merupakan salah satu metode terbaik dan paling populer dalam algoritma clustering dimana K-Means mencari partisi yang optimal dari data dengan meminimalkan kriteria jumlah kesalahan kuadrat dengan prosedur iterasi yang optimal (Syahfitri, *et al*., 2023).

3.2.1 Definisi PCA

Principal Component Analysis (PCA) adalah teknik statistik multivariat yang digunakan untuk mengurangi dimensi data. PCA bekerja dengan mengubah sekumpulan variabel asli yang mungkin saling berkorelasi menjadi sekumpulan variabel baru yang lebih kecil dan tidak berkorelasi, yang disebut komponen utama (Wangge, 2021). Proses ini dilakukan secara linear, dimana komponen utama dihasilkan sebagai kombinasi linear dari variabel asli. Tujuan utama PCA adalah untuk menjelaskan sebanyak mungkin variasi yang ada dalam data asli dengan menggunakan sesedikit mungkin komponen utama. Dengan demikian, PCA membantu menyederhanakan kompleksitas data, memungkinkan analisis yang lebih mudah dan visualisasi yang lebih efektif, tanpa kehilangan informasi yang signifikan dari variabel asli (Sitio & Fauzan, 2020).

3.2.2 Definisi Metode Elbow

Metode elbow adalah metode di mana pada suatu titik tertentu terjadi penurunan yang signifikan dalam grafik, berbentuk lengkungan yang tajam. Nilainya kemudian akan menjadi nilai k atau banyaknya *cluster* yang baik (Abrar , *et al*., 2023). Mencari nilai k optimal dapat dilakukan dengan membandingkan nilai Sum of Square Error (SEE) yang disajikan dalam bentuk grafik. Tujuan dari metode elbow yaitu memilih nilai k yang terkecil dan mempunyai nilai internal yang rendah. Penentuan jumlah *cluster* yang optimal diidentifikasi dengan mempertimbangkan perbandingan perhitungan SEE pada setiap nilai cluster, peningkatan jumlah *cluster* akan membentuk siku, sehingga semakin besar nilai k, nilai SEE akan semakin kecil.

3.2.4 K-Mean Clustering

K-Means Clustering adalah salah satu teknik clustering yang paling umum digunakan dalam analisis data. Algoritma ini bekerja dengan cara membagi data ke dalam K kelompok yang telah ditentukan sebelumnya, di mana K merupakan jumlah kelompok yang diinginkan (Hendrastuty, 2024). Proses dimulai dengan memilih secara acak K titik pusat kelompok (*centroid*) di dalam ruang data, lalu mengelompokkan setiap titik data ke dalam kelompok yang memiliki *centroid* terdekat. Kemudian, titik pusat setiap kelompok dihitung kembali berdasarkan rata-rata titik data dalam kelompok tersebut, dan proses ini diulangi hingga tidak ada lagi perubahan dalam penempatan titik data ke dalam kelompok atau hingga batasan iterasi yang ditentukan tercapai. K-Means *Clustering* sering digunakan dalam segmentasi pelanggan, pengelompokan dokumen, analisis citra, dan bidang lainnya karena kecepatan dan skalabilitasnya, meskipun hasilnya dapat dipengaruhi oleh inisialisasi awal *centroid* (Yudhistira & Andika, 2023).

BAB 4: Pembahasan dan Hasil

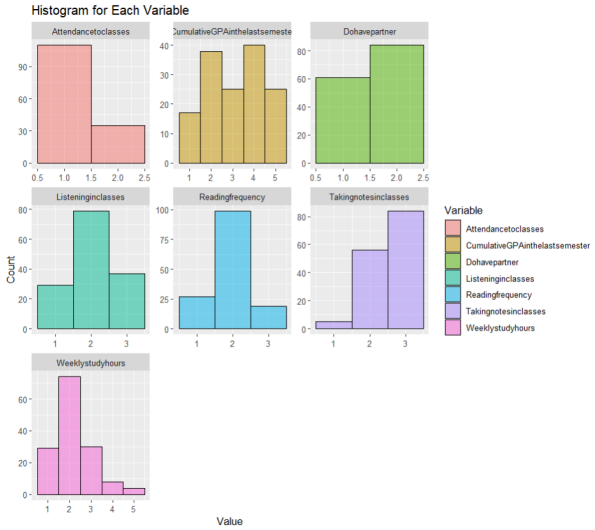
4.1 *Checking Missing Data*

```
> # Memilih hanya beberapa variabel tertentu
> data_selected <- data %>%
+   select(Dohavepartner, Weeklystudyhours, Readingfrequency, Attendancetoclasses, Takingnotesinclasses, Listeninginclasses, CumulativeGPAinthelastsemester)
>
> # Basic Data Exploration
> head(data_selected)
  Dohavepartner Weeklystudyhours Readingfrequency Attendancetoclasses Takingnotesinclasses Listeninginclasses CumulativeGPAinthelastsemester
1             2                3                 2                    1                   3                   2                             1
2             2                2                 2                    1                   3                   2                             2
3             2                2                 1                    1                   2                   2                             2
4             3                1                 1                    1                   3                   2                             3
5             1                2                 1                    1                   2                   2                             2
6             2                1                 1                    1                   1                   2                             4
> str(data_selected)
'data.frame':   145 obs. of  7 variables:
 $ Dohavepartner               : int  2 2 2 1 1 2 2 1 1 1 ...
 $ Weeklystudyhours            : int  3 2 2 3 2 1 2 1 1 2 ...
 $ Readingfrequency            : int  2 2 1 1 1 1 2 2 2 2 ...
 $ Attendancetoclasses         : int  1 1 1 1 1 1 2 1 1 2 ...
 $ Takingnotesinclasses        : int  3 3 2 3 2 1 3 3 3 2 ...
 $ Listeninginclasses          : int  2 2 2 2 2 2 3 2 2 2 ...
 $ CumulativeGPAinthelastsemester: int  1 2 2 3 2 4 4 1 4 1 ...
> summary(data_selected)
 Dohavepartner   Weeklystudyhours Readingfrequency Attendancetoclasses Takingnotesinclasses Listeninginclasses CumulativeGPAinthelastsemester
 Min.   :1.000   Min.   :1.0      Min.   :1.000     Min.   :1.000       Min.   :1.000        Min.   :1.000       Min.   :1.000
 1st Qu.:1.000   1st Qu.:2.0      1st Qu.:2.000     1st Qu.:1.000       1st Qu.:2.000        1st Qu.:2.000       1st Qu.:2.000
 Median :2.000   Median :2.0      Median :2.000     Median :1.000       Median :3.000        Median :2.000       Median :3.000
 Mean   :1.579   Mean   :2.2      Mean   :1.945     Mean   :1.241       Mean   :2.545        Mean   :2.055       Mean   :3.124
 3rd Qu.:2.000   3rd Qu.:3.0      3rd Qu.:2.000     3rd Qu.:1.000       3rd Qu.:3.000        3rd Qu.:3.000       3rd Qu.:4.000
 Max.   :2.000   Max.   :5.0      Max.   :3.000     Max.   :2.000       Max.   :3.000        Max.   :3.000       Max.   :5.000
>
> # Checking Missing Values
> colSums(is.na(data_selected))
               Dohavepartner             Weeklystudyhours             Readingfrequency          Attendancetoclasses
                           0                            0                            0                            0
        Takingnotesinclasses           Listeninginclasses CumulativeGPAinthelastsemester
                           0                            0                            0
```

Gambar 1. *Checking Missing Data*

Kode berikut menjelaskan proses membuat dan memproses data mahasiswa untuk analisis lebih lanjut. Pertama, data dimuat dari file CSV yang berisi informasi performa akademik mahasiswa. Dipilih beberapa variabel penting seperti status hubungan, jam belajar mingguan, frekuensi membaca, kehadiran di kelas, kebiasaan mencatat, mendengarkan di kelas, dan IPK semester terakhir. Selanjutnya, dilakukan eksplorasi data dasar dengan menampilkan lima baris pertama, memeriksa struktur data, dan memberikan ringkasan statistik deskriptif untuk memahami karakteristik data. Pemeriksaan nilai hilang dilakukan dengan menghitung jumlah nilai yang hilang di setiap kolom. Untuk menangani nilai yang hilang, digunakan metode imputasi dengan mengganti nilai hilang menggunakan rata-rata kolom yang bersangkutan, memastikan integritas dataset untuk analisis lebih lanjut.

4.2 Plot Setiap Variabel

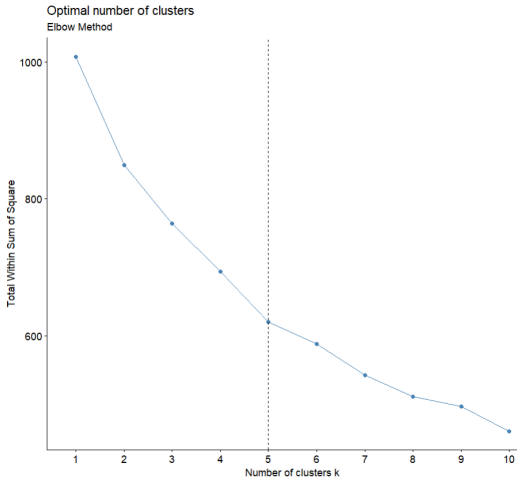


Gambar 2. *Histogram for each variable*

Gambar tersebut menunjukkan histogram dari beberapa variabel dalam dataset mahasiswa. Sebagian besar mahasiswa rutin hadir di kelas dan memiliki IPK yang baik (nilai 3-4). Mayoritas memiliki pasangan, mendengarkan dengan baik di kelas, dan membaca dengan frekuensi sedang. Banyak mahasiswa sering mencatat di kelas dan sebagian besar belajar 1-2 jam

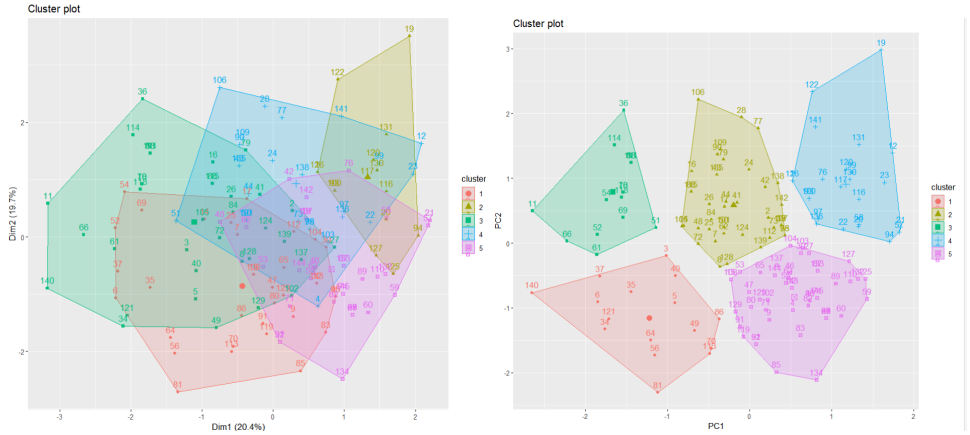
per minggu. Histogram ini memberikan gambaran tentang distribusi kebiasaan belajar dan performa akademik mahasiswa, yang penting untuk analisis lebih lanjut.

4.4 Penentuan Jumlah *Cluster* Optimum Menggunakan Metode *Elbow*



Gambar tersebut menunjukkan hasil metode Elbow untuk menentukan jumlah klaster optimal. Titik "elbow" terlihat pada k = 5, yang menunjukkan bahwa jumlah klaster optimal adalah 5, karena penurunan Total Within Sum of Squares (TWSS) mulai melambat setelah titik ini.

4.3 Clustering dengan PCA dan Tidak PCA



Gambar 4. Dengan PCA dan Tidak

Kedua plot klaster menunjukkan bahwa algoritma K-means berhasil mengidentifikasi kelompok-kelompok yang bermakna dalam data. Namun, penggunaan PCA menghasilkan representasi data yang lebih merata dan lebih mudah diinterpretasikan. Hal ini menunjukkan bahwa PCA dapat menjadi langkah pra proses yang berguna untuk algoritma clustering, terutama ketika berhadapan dengan data berdimensi tinggi.

4.5 Rataan Setiap Variabel Untuk Tiap Variabel Mahasiswa

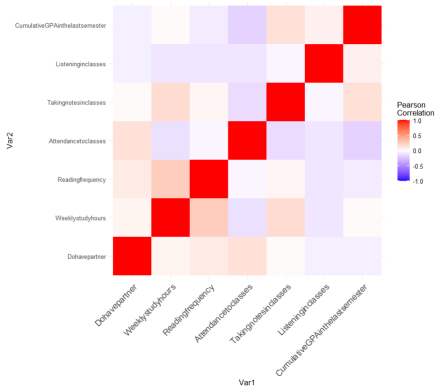
	Group.1	Dohavepartner	Weeklystudyhours	Readingfrequency	Attendancetoclasses
1	1	1.558824	1.441176	1.764706	1.235294
2	2	1.600000	3.866667	2.266667	1.066667
3	3	1.567568	1.729730	1.783784	1.405405
4	4	1.680000	2.320000	2.360000	1.280000
5	5	1.529412	2.647059	1.852941	1.117647

	Takingnotesinclasses	Listeninginclasses	CumulativeGPAinthelastsemester
1	2.500000	2.323529	3.911765
2	2.800000	2.066667	2.000000
3	2.351351	2.189189	1.621622
4	2.600000	1.520000	2.840000
5	2.647059	2.029412	4.676471

Gambar 5. Rataan Setiap Variabel Untuk Tiap Variabel Mahasiswa

Cluster 1 terdiri dari mahasiswa yang mayoritas memiliki pasangan, namun memiliki jam belajar mingguan rendah. Frekuensi membaca mereka rendah hingga sedang, namun mereka sering hadir di kelas, cukup sering mencatat, dan mendengarkan di kelas. IPK mereka berada di rentang 3.00-3.49. Cluster 2 mencakup mahasiswa dengan sedikit kecenderungan memiliki pasangan dan jam belajar mingguan cukup tinggi. Mereka memiliki frekuensi membaca sedang, kehadiran di kelas sangat tinggi, sering mencatat, dan cukup sering mendengarkan. Meskipun demikian, IPK mereka sedang. Cluster 3 terdiri dari mahasiswa yang cenderung memiliki pasangan dengan jam belajar mingguan rendah. Frekuensi membaca mereka rendah hingga sedang, kehadiran di kelas cukup tinggi, serta cukup sering mencatat dan mendengarkan. IPK mereka rendah, berada di rentang 2.00-2.49. Cluster 4 mencakup mahasiswa yang cenderung memiliki pasangan dengan jam belajar mingguan sedang. Frekuensi membaca mereka sedang, kehadiran di kelas cukup tinggi, dan mereka sangat sering mencatat serta cukup sering mendengarkan. IPK mereka cukup tinggi, berada di rentang 2.50-2.99. Cluster 5 terdiri dari mahasiswa yang cenderung memiliki pasangan dengan jam belajar mingguan sedang. Frekuensi membaca mereka sedang, kehadiran di kelas sangat tinggi, serta sangat sering mencatat dan cukup sering mendengarkan. IPK mereka sangat tinggi, di atas 3.49.

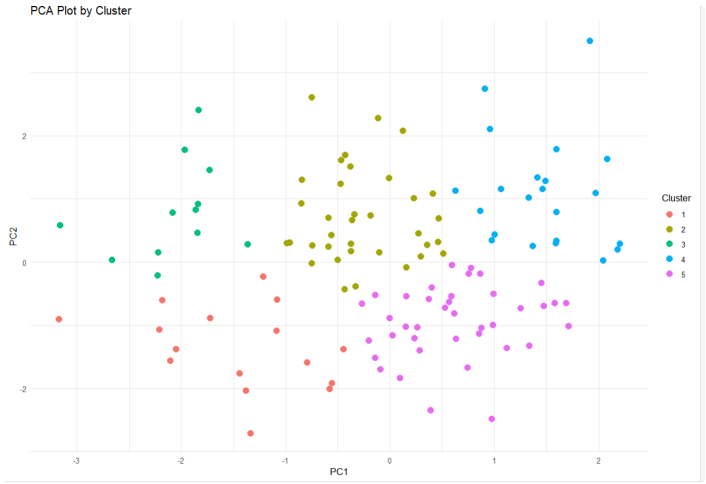
4.5 Heatmap



Gambar 6. Heatmap

Heatmap menyatakan korelasi positif yang kuat antara berbagai kebiasaan belajar dan performa akademik. Ini menunjukkan bahwa kebiasaan belajar ini kemungkinan berkontribusi terhadap hasil akademik yang lebih baik. Namun, diperlukan penelitian lanjutan untuk menetapkan hubungan kausal antara variabel-variabel ini.

4.6 PCA Plot by Cluster



Gambar 7. PCA Plot by Cluster

Mahasiswa dalam Kluster 1 memiliki waktu belajar lebih sedikit tetapi mencapai GPA tinggi berkat kehadiran, kebiasaan mencatat, dan mendengarkan yang baik, dengan partisipasi aktif dalam pembelajaran menjadi kunci kesuksesan mereka. Kluster 2, meskipun menghabiskan lebih banyak waktu belajar, hanya mencapai GPA sedang dengan kehadiran, kebiasaan mencatat, dan mendengarkan yang baik. Kluster 3 menunjukkan mahasiswa dengan kesulitan mencapai GPA tinggi karena waktu belajar, kehadiran, dan kebiasaan belajar yang rata-rata. Kluster 4 mirip dengan Kluster 2 tetapi lebih mungkin memiliki pasangan; mereka mencapai GPA sedang dengan waktu belajar moderat dan kebiasaan belajar yang baik. Mahasiswa dalam Kluster 5, meskipun memiliki pasangan dan waktu belajar moderat, mencapai GPA tertinggi karena kehadiran, kebiasaan mencatat, dan mendengarkan yang baik, yang sangat berkontribusi pada kesuksesan akademik mereka.

BAB 5: Kesimpulan dan Saran

Kesimpulan

Metode Principal Component Analysis (PCA) dan K-means clustering digunakan untuk menganalisis pola karakteristik mahasiswa dalam lima kluster. Hasilnya menunjukkan bahwa pola kebiasaan belajar dan performa akademik dapat dikelompokkan ke dalam lima klaster yang berbeda. Kluster 1 memiliki mahasiswa dengan jam belajar mingguan rendah, namun memiliki IPK yang cukup tinggi. Kluster 2 terdiri dari mahasiswa dengan jam belajar tinggi, kehadiran di kelas sangat tinggi, namun IPK sedang. Kluster 3 menunjukkan mahasiswa dengan jam belajar rendah, kehadiran di kelas cukup tinggi, dan IPK rendah. Kluster 4 dan 5 memiliki mahasiswa dengan jam belajar sedang, kehadiran di kelas tinggi, serta IPK yang berada di rentang tinggi, dengan Kluster 5 memiliki IPK tertinggi di antara kelompok. Dengan menggunakan kombinasi PCA dan K-means, analisis ini memberikan wawasan tentang pola belajar dan performa akademik mahasiswa, serta memungkinkan institusi pendidikan untuk merancang strategi pembelajaran yang lebih efektif.

Saran

Untuk penelitian sejenis yang menggunakan metode K-means dan PCA dalam analisis performa mahasiswa, beberapa saran berikut dapat dipertimbangkan: (1) memperluas Variabel Penelitian (2) Peningkatan Kualitas Data, dan (3) Segmentasi Mahasiswa yang Lebih Mendetail.

BAB 6: Daftar Pustaka

Abrar , I. N., Abdullah, A. & Sucipto, S., 2023. Liver Disease Classification Using The Elbow Method to Determine Optimal K in The K-Nearest Neighbour Alghoritm. *Jurnal Sistem Informasi dan Komputer,* pp. 10-20.

Fatimah, S., Manuardi, A. R. & Meilani, R., 2021. Tingkat Efikasi Diri Performa Akademik Mahasiswa Ditinjau Dari Perspektif Dimensi Bandura. *Professional, Empathy and Islamic Counseling Journal,* pp. 25-36.

Hendrastuty, N., 2024. Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer,* pp. 46-56.

Putri, C. P., Mayangsari, M. D. & Rusli, R., 2018. Pengaruh Stress Academic terhadap Help Seeking Pada Mahasiswa Psikologi Unlam dengan Indeks Prestasi Kumulatif Rendah. *Jurnal Kognisia,* pp. 28-37.

Rulanggi, R., Fahera, J. & Novira, N., 2021. Faktor-Faktor yang Memengaruhi Subjective Well-Being Pada Mahasiswa. pp. 406-412.

Sitio, N. M. & Fauzan, F., 2020. Analisis Komponen Utama (PCA) Untuk Mereduksi Faktor Organisasi Pembelajaran. *Jurnal Sains Manajemen,* pp. 85-95.

Syahfitri, N., Budianita, E. & Nazir, A., 2023. Pengelompokan Produk Berdasarkan Data Persediaan Barang. *Jurnal Ilmiah Informatika dan Komputer,* pp. 1668-1675.

Wajar, M. . S. A. M., Hamzah, R., Mohammad, A. M. & Andin, C., 2022. The Influencesof Mental Health and Spiritual Intelligence Towards Well Being and Academic Performance. *International Journal of Humanities Technology and Civilization,* pp. 10-21.

Wangge, M., 2021. Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktorfaktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA. *Jural Cendekia,* pp. 974-988.

Yudhistira, A. & Andika, R., 2023. Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering. *Journal of Artificial Intelligence and Technology Information,* pp. 20-28.