Optimasi K-Means++ Menggunakan Principal Component Analysis (PCA) pada Klasterisasi Profil Kelulusan Mahasiswa

Hana Solikatun[1], Herdiesel Santoso*[2]
[1,2] Program Studi Sistem Informasi STMIK El Rahma Yogyakarta
*e-mail: [1]hanasolihatun@gmail.com *[2]herdiesel.santoso@stmikelrahma.ac.id*
*Correspondence author email: **

Abstrak

Ketepatan waktu kelulusan mahasiswa merupakan pilar utama dan indikator keberhasilan institusi pendidikan tinggi. Namun, pengelompokan data rekam jejak akademik mahasiswa yang memiliki tipe data campuran (numerik dan kategorikal) menggunakan algoritma K-Means klasik sering kali menghasilkan bias jarak matematis dan noise akibat tingginya dimensi atribut (curse of dimensionality). Penelitian ini mengusulkan solusi integratif melalui optimasi algoritma K-Means++ menggunakan teknik One-Hot Encoding dan Principal Component Analysis (PCA). Evaluasi dilakukan terhadap 200 observasi data lulusan STMIK El Rahma Yogyakarta. Hasil penelitian membuktikan bahwa reduksi menjadi 2 komponen utama melalui PCA mampu meningkatkan performa klasterisasi secara signifikan. Validasi objektif menunjukkan metrik Silhouette Score meningkat dari 0,3275 menjadi 0,4979, Davies-Bouldin Index membaik dari 1,405 menjadi 0,871, dan Calinski-Harabasz Index melonjak dari 70,448 menjadi 168,035. Model ini memetakan mahasiswa menjadi dua kelompok, yaitu Profil Akademik Konsisten (157 mahasiswa) dan Profil Multi-Peran Rentan (43 mahasiswa) yang didominasi oleh mahasiswa pekerja paruh waktu. Temuan ini dapat dimanfaatkan oleh pemangku kebijakan kampus sebagai dasar perancangan Sistem Peringatan Dini (Early Warning System) akademik yang presisi.

Kata kunci*— Clustering; Data Akademik; Data Mining; K-Means++; Principal Component Analysis.*

1. PENDAHULUAN

Ketepatan waktu kelulusan mahasiswa merupakan salah satu pilar utama yang menentukan kualitas sumber daya manusia sekaligus indikator keberhasilan sebuah institusi pendidikan tinggi [1] . Institusi dituntut untuk mampu memantau persentase kelulusan tepat waktu secara presisi, mengingat hal tersebut berdampak langsung pada nilai akreditasi program studi [2]. Namun, banyak perguruan tinggi yang masih menghadapi kendala dalam mendeteksi potensi keterlambatan lulus sejak dini. Oleh karena itu, diperlukan suatu sistem analitik berbasis data mining yang mampu menggali informasi tersembunyi dari rekam jejak akademik dan profil demografi mahasiswa untuk mengantisipasi permasalahan akademik sebelum terjadi [3].

Dalam bidang data mining, algoritma *clustering* (pembelajaran tanpa pengawasan/ *unsupervised learning*) seperti *K-Means* sangat direkomendasikan untuk memetakan karakteristik mahasiswa berdasarkan kemiripan pola data[4]. Algoritma ini membagi data ke dalam beberapa kelompok agar institusi dapat mengenali segmentasi mahasiswanya [5]. Akan tetapi, penerapan *K-Means* klasik pada data pendidikan sering kali menghadapi tantangan teknis yang fundamental. Tantangan utama tersebut adalah kehadiran tipe data campuran (*mixed-data types*), di mana atribut numerik (seperti Indeks Prestasi Semester) bercampur dengan atribut kategorikal/nominal (seperti program studi, keaktifan organisasi, dan status bekerja) . Karena *K-Means* mengandalkan perhitungan jarak spasial (*Euclidean distance*), algoritma ini tidak dapat memproses data kategorikal secara langsung dan sering kali mengalami bias perhitungan matematis jika kategori tersebut hanya diubah menjadi angka berurutan (*label encoding*)[6]. Selain itu, banyaknya dimensi atribut data yang digunakan dapat memunculkan *noise* dan *outlier* yang mengaburkan batas antar klaster (*curse of dimensionality*).

Beberapa penelitian terdahulu telah berupaya menyelesaikan masalah pengelompokan data akademik mahasiswa. Penelitian [5] telah menerapkan algoritma *K-Means* untuk

memprediksi kelulusan tepat waktu dengan membagi data mahasiswa ke dalam dua kelompok, namun penelitian tersebut tidak merinci penanganan khusus pada variabel kategorikal sehingga rentan terhadap bias jarak matematis. Penelitian lain oleh [7] berupaya meningkatkan efisiensi waktu konvergensi algoritma menggunakan inisialisasi *K-Means++* untuk mengelompokkan mahasiswa berpotensi *drop out*, namun fokus fitur yang digunakan masih terbatas pada tipe data numerik homogen seperti IPK dan jumlah SKS. Sementara itu, [3] membuktikan bahwa *K-Means* sangat efektif digunakan sebagai langkah klasifikasi awal sebelum masuk ke algoritma prediksi *Random Forest* untuk memetakan kelulusan. Di sisi lain, terkait masalah dimensionalitas, [8] membuktikan bahwa penerapan reduksi dimensi menggunakan *Principal Component Analysis* (PCA) sebelum melakukan *clustering* berhasil mengurangi kompleksitas data tanpa menghilangkan informasi penting dalam evaluasi mutu perguruan tinggi.

Meskipun penelitian-penelitian di atas telah mengonfirmasi keandalan *K-Means* dan PCA, terdapat celah penelitian yang belum terselesaikan. Sebagian penelitian belum mengintegrasikan penyelesaian masalah bias jarak pada data campuran secara bersamaan dengan penanganan *noise* dimensionalitas dalam satu *pipeline* pemodelan. Berdasarkan urgensi tersebut, penelitian ini mengusulkan solusi integratif berupa optimasi algoritma *K-Means++* menggunakan teknik *One-Hot Encoding* dan *Principal Component Analysis* (PCA). Penggunaan *K-Means++* didasarkan pada keunggulannya dalam proses inisialisasi *centroid* yang mempertimbangkan sebaran data secara proporsional, sehingga mampu mengurangi risiko *local optimum*, menghasilkan klaster yang lebih stabil dan konsisten, serta mempercepat proses konvergensi dibandingkan *K-Means* dengan inisialisasi acak [6], [9]. *One-Hot Encoding* ditawarkan sebagai solusi data untuk memecah variabel kategorikal menjadi matriks biner ortogonal, sehingga jarak *Euclidean* antar kategori menjadi bernilai setara dan valid [1]. Selanjutnya, PCA diterapkan untuk mereduksi ledakan dimensi akibat *encoding* tersebut menjadi komponen utama yang lebih padat dan bebas *noise* [8], [10].

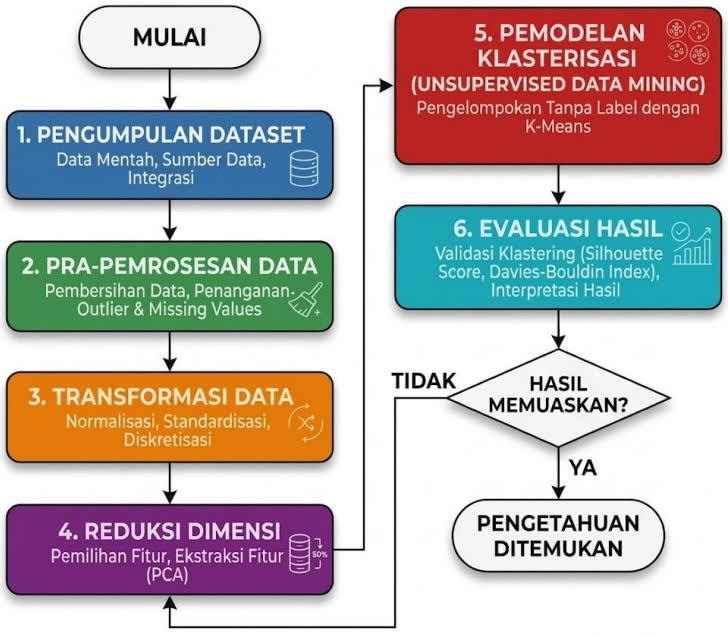
Penelitian ini bertujuan untuk menguji tingkat peningkatan performa algoritma *K-Means++* setelah dioptimasi dengan PCA dan *One-Hot Encoding* menggunakan matriks evaluasi *Silhouette Score*, *Davies-Bouldin Index* dan *Calinski-Harabasz Index*, serta memetakan profil karakteristik kelulusan mahasiswa di STMIK El Rahma Yogyakarta. Adapun kontribusi utama dari penelitian ini adalah memberikan kerangka metodelogi pemrosesan data campuran yang bebas bias untuk algoritma spasial, serta menyajikan wawasan strategis berupa profil spesifik mahasiswa (seperti penemuan pola anomali pada mahasiswa pekerja) yang dapat dimanfaatkan oleh pemangku kebijakan kampus dalam merancang Sistem Peringatan Dini (*Early Warning System*) dan pendampingan akademik yang presisi.

2. METODE PENELITIAN

Metode penelitian menggunakan pendekatan *Knowledge Discovery in Databases* (KDD) yang berfokus pada teknik Data Mining kategori *unsupervised learning*[4]. Alur penelitian ditunjukan pada Gambar 1.

Dataset Penelitian

Dataset yang digunakan berjumlah 200 data mencakup data mahasiswa lulusan dalam rentang waktu tahun 2016 hingga 2020, sehingga merepresentasikan kondisi akademik dalam periode tersebut. Data difokuskan pada mahasiswa yang telah memiliki status kelulusan (Tepat Waktu atau Terlambat). Dataset ini mencakup variabel-variabel seperti Indeks Prestasi Semester (IPS 1 hingga IPS 7), Program Studi, Kelas (Reguler/Non-Reguler), Jenis Kelamin, keaktifan berorganisasi, status pernikahan, dan status bekerja. Total dataset yang diolah berjumlah 200 baris observasi yang merepresentasikan populasi lulusan. Atribut pada dataset kelulusan mahasiswa dapat dilihat pada Tabel 1 sedangkan contoh data kelulusan mahasiswa dapat dilihat pada Tabel 2. Tabel 2 hanya menampilkan sebagian atribut, sedangkan atribut lengkap yang digunakan dalam penelitian dijelaskan pada Tabel 1.



Gambar 1. Proses dan tahapan *Knowledge Discovery in Databases* (KDD)

Tabel 1. Atribut dan tipe data kelulusan mahasiswa

No	Atribut	Tipe data	Nilai
1.	NIM	Text	Metadata
2.	Nama	Text	Metadata
3.	Prodi	Nominal	TI, SI
4.	Jenis Kelamin	Nominal	L, P
5.	Kelas	Nominal	Reguler, Non Reguler
6.	Organisasi	Nominal	Ya, Tidak
7.	Bekerja	Nominal	Ya, Tidak
8-14.	IP Semester 1-7	Numerik	0-4
15.	Status Mahasiswa	Nominal	Tepat, Terlambat

Tabel 2. Contoh data sebelum dilakukan preprocessing(sebagian atribut)

No	Nim	Nama	Prodi	Kelas	IPS4	IPS7	Status Kelulusan
1	12100918	Mhs 1	TI	Non Reguler	3.35	3.20	Terlambat
2	12111003	Mhs 2	TI	Reguler	2.83	2.83	Terlambat
3	11120048	Mhs 3	SI	Reguler	2.95	2.95	Tepat
4	11120077	Mhs 4	SI	Reguler	2.95	3.72	Tepat

Pra-pemrosesan Data (Preprocessing)

Tahap pra-pemrosesan merupakan langkah krusial untuk memastikan algoritma K-Means dapat berjalan secara optimal tanpa bias. Langkah-langkah yang dilakukan meliputi[11]:

a. Pembersihan Data: Melakukan pengecekan terhadap nilai yang kosong (*missing values*) dan duplikasi data. Baris data yang tidak lengkap atau memiliki anomali yang tidak masuk akal dihapus dari dataset. Atribut identitas personal seperti 'No', 'NIM', dan 'Nama' dihapus karena tidak memiliki nilai komputasi untuk *clustering*.
b. Pencegahan Data *Leakage*: Dalam konteks *unsupervised learning* untuk menemukan pola kelulusan, variabel "Status Kelulusan" dihapus (di-*drop*) dari matriks fitur utama sebelum model dilatih. Variabel ini disimpan secara terpisah dan hanya digunakan di tahap akhir sebagai observasi/label silang (*crosstab*) untuk menganalisis dominasi karakteristik kelulusan di masing-masing klaster yang telah terbentuk [12]. Hal ini dilakukan untuk memastikan algoritma mengelompokkan mahasiswa murni berdasarkan performa akademik dan profil, bukan karena algoritma "mengetahui" hasil akhir kelulusannya.

c. Selain pembersihan data, dilakukan juga analisis terhadap keberadaan *outlier* menggunakan pendekatan statistik berbasis *Z-Score*. Data dengan nilai di luar rentang ± 3 *standar deviasi* diidentifikasi sebagai *outlier* potensial. Hal ini dilakukan untuk menjaga representasi distribusi data asli dan menghindari bias akibat penghapusan data ekstrem. Meskipun teridentifikasi beberapa nilai ekstrem berdasarkan *Z-Score*, observasi tersebut tetap dipertahankan karena merepresentasikan kondisi nyata mahasiswa dan tidak menunjukkan anomali yang bersifat kesalahan data.

Transformasi Data

Karena *K-Means++* mengandalkan perhitungan jarak (*Euclidean distance*), algoritma ini sangat sensitif terhadap skala data dan tidak dapat memproses tipe data non-numerik secara langsung. Oleh karena itu, dilakukan dua teknik transformasi:

a. Standardisasi Numerik: Atribut Indeks Prestasi Semester (IPS 1 hingga IPS 7) dinormalisasi menggunakan *Z-Score/StandardScaler*. Teknik ini mengubah rentang nilai menjadi distribusi standar dengan nilai rata-rata (mean) 0 dan standar deviasi 1 [13]. Hal ini mencegah variabel IPS mendominasi perhitungan jarak semata-mata karena rentang angkanya. Hasil transformasi atribut pada data kelulusan mahasiswa di sajikan pada Tabel 4.
b. *One-Hot Encoding*: Variabel kategorikal/nominal seperti Program Studi, Kelas, Jenis Kelamin, Organisasi, Menikah, dan Bekerja ditransformasikan menggunakan metode *One-Hot Encoding* [1]. Metode ini memecah setiap kategori menjadi kolom biner baru (0 atau 1). Berbeda dengan *Label Encoding* yang memberikan nilai urut (0, 1, 2) yang dapat memicu bias jarak matematis pada algoritma spasial [6]. *One-Hot Encoding* merepresentasikan setiap kategori setara (jarak ortogonal), sehingga perbandingan antar atribut kategorikal menjadi valid. Contoh data yang telah melalui proses standardisasi dan One-Hot Encoding ditunjukan pada Tabel 5.

Tabel 4. Transformasi atribut / fitur pada data kelulusan mahasiswa

No	Atribut	Kode Atribut	Transformasi Nilai
1.	Prodi	Prodi	INF_YA = 0, SI_YA= 1
2.	Jenis Kelamin	Jk	L= 0, P =1
3.	Kelas	Kelas	Reguler = 0, Non Reguler=1
5.	Organisasi	Organisasi	Ya = 0, Tidak = 1
6.	Bekerja	Bekerja	Ya = 0, Tidak = 1
7.	IP Semester 1-7	ips1, ips2, ips3, ips4, ips5 ips6 ips7	Skala Standar (- s.d +)

Tabel 5. Contoh data setelah dilakukan preprocessing (hasil transformasi fitur)

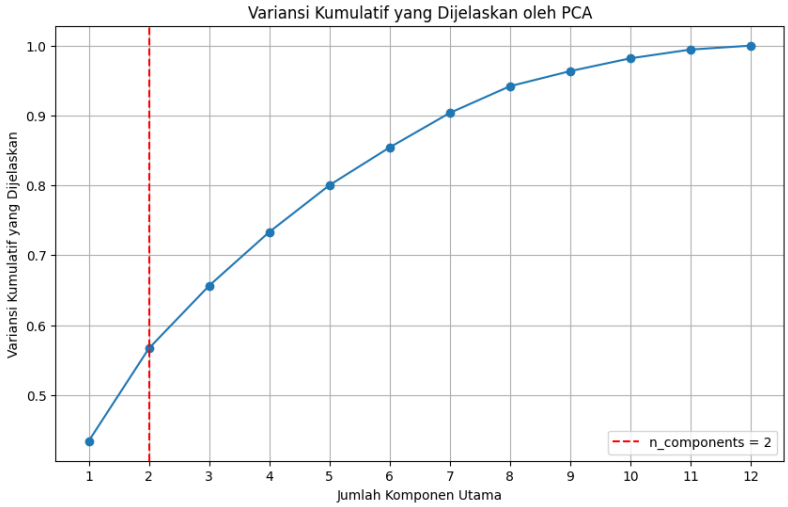
Prodi_Tenik Informatika	Kelas_Reguler	Jenis Kelamin P	Organisasi_Ya	IPS6	IPS7
1	0	0	0	0.26	-0.10
1	1	1	0	-0.75	-0.75
0	1	0	0	-0.52	-0.54
0	1	1	0	-0.52	0.80

Reduksi Dimensi dengan Principal Component Analysis (PCA)

Penggunaan *One-Hot Encoding* pada variabel kategorikal menyebabkan peningkatan jumlah fitur secara signifikan, yang memicu fenomena *curse of dimensionality*. Dimensi data yang tinggi dapat mengaburkan jarak antar titik pada algoritma berbasis jarak seperti K-Means, sehingga menurunkan kualitas pemisahan klaster [14]. Untuk mengatasi permasalahan tersebut, diterapkan teknik reduksi dimensi menggunakan *Principal Component Analysis* (PCA), yang bertujuan untuk memproyeksikan data ke dalam ruang berdimensi lebih rendah tanpa kehilangan struktur informasi utama [8].

Berdasarkan hasil analisis *cumulative explained variance* yang ditunjukkan pada Gambar 2, dua komponen utama pertama mampu menjelaskan sekitar 57% dari total variansi data.

Meskipun proporsi ini belum mencakup mayoritas absolut variansi, kedua komponen tersebut telah mampu merepresentasikan struktur utama data yang relevan untuk proses klasterisasi. Selain itu, peningkatan jumlah komponen memang berkontribusi terhadap kenaikan variansi yang dijelaskan, namun tidak secara langsung meningkatkan interpretabilitas model.



Gambar 2. Analisis *cumulative explained variance*

Dalam konteks penelitian ini yang berfokus pada analisis klaster dan interpretasi pola, pemilihan dua komponen utama juga didasarkan pada kebutuhan visualisasi dua dimensi yang memungkinkan eksplorasi struktur klaster secara intuitif. Representasi dua dimensi memberikan keunggulan dalam mengidentifikasi separasi dan pola distribusi data yang tidak dapat diamati secara langsung pada dimensi yang lebih tinggi. Oleh karena itu, penggunaan dua komponen utama dalam PCA dinilai sebagai kompromi yang optimal antara preservasi informasi, efisiensi komputasi, dan kemudahan interpretasi hasil, sehingga mendukung tujuan utama penelitian dalam menghasilkan klaster yang tidak hanya optimal secara matematis, tetapi juga bermakna analitis.

Pemodelan dengan K-Means++

Dataset hasil reduksi PCA kemudian diproses menggunakan algoritma *K-Means++ Clustering*. Metode ini secara cerdas menempatkan titik pusat (*centroid*) awal berjauhan satu sama lain, sehingga proses konvergensi menjadi lebih cepat dan menghindarkan algoritma terjebak pada *local optimum* [6]. Secara lebih rinci, tahapan komputasi penerapan algoritma *K-Means++* pada data observasi dalam penelitian ini diuraikan sebagai berikut [15] [9].

Inisialisasi Centroid Pertama: Mengambil satu titik data (x) secara acak dari keseluruhan dataset hasil reduksi $PCA(X)$ berukuran n untuk ditetapkan sebagai pusat klaster (centroid) pertama, dinotasikan sebagai c_1. Perhitungan Jarak Kuadrat (Squared Distance): Perhitungan jarak antara titik data dan centroid merupakan komponen fundamental dalam algoritma K-Means++, karena proses pengelompokan sepenuhnya bergantung pada kedekatan geometris antar data dalam ruang fitur. Secara umum, jarak antar titik dihitung menggunakan Euclidean Distance, yang mengukur jarak lurus antara dua titik dalam ruang berdimensi m. Secara matematis, jarak antara titik data xi dan centroid cj dinyatakan sebagai persamaan (1).

$$D(x_i, c_j) = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \quad (1)$$

Sebagai ilustrasi, misalkan terdapat dua titik data hasil reduksi PCA dalam ruang dua dimensi, yaitu x_1 =(1.2, −0.5) dan centroid c_1=(0.3,0.8). Maka *jarak Euclidean* antara kedua titik tersebut dapat dihitung sebagai berikut:

$$D(x_1, c_1) = \sqrt{(1.2 - 0{,}3)^2 + (-0.5 - 0{,}8)^2} = \sqrt{(0{,}9)^2 + (-1.3)^2} = \sqrt{0.81 + 1.69} = \sqrt{2.5} = 1.58$$

Meskipun Euclidean Distance digunakan dalam proses pengelompoka untuk menentukan kedekatan antar titik, pada tahap inisialisasi centroid dalam algoritma K-Means++ digunakan bentuk jarak kuadrat (squared distance) [15]. Penggunaan jarak kuadrat bertujuan untuk meningkatkan efisiensi komputasi dengan menghindari operasi akar kuadrat, tanpa mempengaruhi urutan perbandingan jarak antar titik. Jarak kuadrat minimum antara suatu titik data terhadap centroid terdekat dinyatakan sesuai persamaan (2).

$$D(x_i)^2 = \min_{c_k \in C} \sum_{j=1}^{m} \left(x_{ij} - c_{kj}\right)^2 \quad (2)$$

di mana : x_i= data ke i; c = centroid dalam himpunan C, dan m = jumlah dimensi fitur.
Pemilihan Centroid Lanjutan Berbasis Probabilitas: Memilih titik data berikutnya untuk dijadikan *centroid* baru (c_i). Titik data dipilih menggunakan distribusi probabilitas terbobot, di mana titik data x_i memiliki peluang terpilih $P(x_i)$ yang sebanding dengan jarak kuadratnya terhadap centroid terdekat menggunakan persamaan (3). Mekanisme ini memastikan penyebaran centroid awal yang maksimal dan menghindari penumpukan pada satu area (menghindari *local optimum*).

$$P(x_i) = \frac{D(x_i)}{\sum_{x_j \in X} D(x_j)} \quad (3)$$

Iterasi Inisialisasi: Mengulangi langkah 2 dan 3 secara terus-menerus hingga sejumlah k centroid awal berhasil ditentukan sesuai dengan jumlah klaster yang diuji. Alokasi Data (Pengelompokan): Setelah k centroid awal terbentuk, menghitung jarak (menggunakan Euclidean distance) dari setiap titik data ke seluruh centroid. Titik data kemudian dialokasikan ke dalam klaster yang memiliki jarak centroid terpendek. Pembaruan Posisi Centroid: Memperbarui titik pusat setiap klaster (c_i) dengan cara menghitung nilai rata-rata (mean) dengan persamaan (4) dari seluruh titik data yang telah tergabung ke dalam klaster tersebut.

$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x \quad (4)$$

di mana S_i adalah himpunan titik data dalam klaster ke-*i*, dan $|S_i|$ adalah jumlah anggotanya.
Konvergensi Model: Mengulangi langkah 5 dan 6 untuk meminimalkan nilai fungsi objektif berupa *Within-Cluster Sum of Squares* (WCSS) dengan persamaan (5) [12]

$$WCSS = \sum_{j=1}^{k} \sum_{x \in S_j} \left\| x - c_j \right\|^2 \quad (5)$$

Iterasi dihentikan apabila posisi *centroid* menjadi stabil (tidak ada lagi titik data yang berpindah klaster, atau batas maksimal iterasi telah tercapai.

Penentuan jumlah klaster optimal (*k*) merupakan tahap penting dalam pemodelan *K-Means*++. Pada penelitian ini, digunakan pendekatan validasi ganda, yaitu metode *Elbow* sebagai estimasi awal dan *Silhouette Score* sebagai validasi akhir. Metode *Elbow* dipilih karena efisien secara komputasi dan mampu memberikan visualisasi intuitif melalui penurunan nilai *Within-Cluster Sum of Squares* (WCSS), meskipun interpretasi titik "siku" cenderung subjektif [16]. Oleh karena itu, hasilnya dikonfirmasi menggunakan *Silhouette Score* yang mengukur kohesi dan separasi klaster secara kuantitatif, sehingga diperoleh jumlah klaster yang lebih optimal, objektif, dan merepresentasikan struktur data secara lebih baik.

Evaluasi Hasil

Untuk memvalidasi kualitas klasterisasi tanpa mengacu pada kebenaran eksternal (*internal validation*), digunakan dua matrik evaluasi utama yaitu *Silhouette Score*, *Davies-Bouldin Index* (DBI) dan *Calinski-Harabasz Index*. *Silhouette Score* : Mengukur seberapa kohesif sebuah objek dengan klasternya sendiri dibandingkan dengan seberapa terpisah objek tersebut dari klaster terdekat lainnya (rentang -1 hingga 1; semakin mendekati 1 semakin baik) [10].

$$s(i) = \frac{(b(i) - a(i))}{max(a(i), b(i))} \quad (6)$$

Pada metrik persamaan (6), s(i)= nilai Silhouette Score; a(i) = rata-rata jarak antara titik data i terhadap seluruh titik lain di dalam klaster yang sama (jarak intra-klaster); b(i) = rata-rata jarak dari titik i ke titik-titik pada klaster terdekat lainnya (jarak inter-klaster). Rentang nilai

indeks ini berada pada interval -1 hingga 1. Kualitas klasterisasi dievaluasi sebagai model yang baik apabila menghasilkan skor positif (ai < bi). Kondisi paling ideal tercapai saat nilai a(i) mendekati 0, yang akan mendorong Silhouette Score mencapai batas maksimumnya, yakni 1.
Davies-Bouldin Index (DBI): Mengevaluasi rasio jarak sebaran data di dalam klaster (*intra-cluster*) terhadap jarak antar pusat klaster (*inter-cluster*). Nilai DBI yang rendah (mendekati 0) mengindikasikan bahwa klaster padat dan terpisah dengan jelas. Nilai *DB index* dihitung menggunakan persaman (7).

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} Max_{j\neq i}\left(\frac{S_i+S_j}{M_{ij}}\right) \quad (7)$$

di mana *DBI* menunjukan nilai *Davies-bouldin*; *Si* = rata-rata jarak titik ke centroid cluster *i*; *Mij* = jarak antar centroid i dan j. Semakin kecil nilai *Davies Bouldin Index* menunjukan skema konfigurasi klaster telah optimal dan kualitas klaster semakin baik [11].

Calinski-Harabasz Index (CHI): Metrik evaluasi internal untuk menentukan kualitas model pengelompokan (clustering) yang juga dikenal sebagai Variance Ratio Criterion. Berbeda dengan Silhouette Score yang fokus pada jarak antar-titik, CHI mengukur rasio antara dispersi (penyebaran) antar-klaster dengan dispersi di dalam klaster. Semakin tinggi nilai CHI, maka semakin baik pemisahan klasternya. Hal ini menandakan bahwa klaster-klaster tersebut terpisah jauh satu sama lain (antar-klaster) dan sangat padat di dalamnya (dalam-klaster). Rumus untuk menghitung CHI (s) bagi kelompok k menggunakan persaman (8).

$$CH = \frac{Tr(B_k)}{TrW_k} x \frac{n-k}{k-1} \quad (8)$$

$Tr(B_k)$: merepresentasikan seberapa jauh jarak antara pusat klaster satu dengan pusat klaster lainnya, TrW_k : merepresentasikan seberapa dekat data-data di dalam satu klaster dengan pusatnya, *n*: Jumlah total sampel data, *k*: Jumlah klaster yang terbentuk.

3. HASIL DAN PEMBAHASAN

Hasil Transformasi dan Reduksi Dimensi (PCA)

Dataset yang telah melalui tahap pra-pemrosesan (penanganan *missing values* dan eksklusi variabel Status Kelulusan) kemudian ditransformasikan. Variabel kategorikal diubah menggunakan *One-Hot Encoding*, sementara variabel numerik (IPS 1-6) distandardisasi menggunakan *StandardScaler*. Proses encoding ini meningkatkan jumlah fitur secara signifikan, yang berpotensi menimbulkan noise pada perhitungan jarak *K-Means++*. Untuk mengatasi permasalahan dimensionalitas, teknik PCA diterapkan untuk mereduksi fitur-fitur tersebut menjadi 2 Komponen Utama (*Principal Components*). Hasil pengurangan dimensi menggunakan PCA yang diikuti oleh proses *clustering* ditunjukan pada Tabel 6.

Tabel 6. Pengurangan dimensi menggunakan PCA

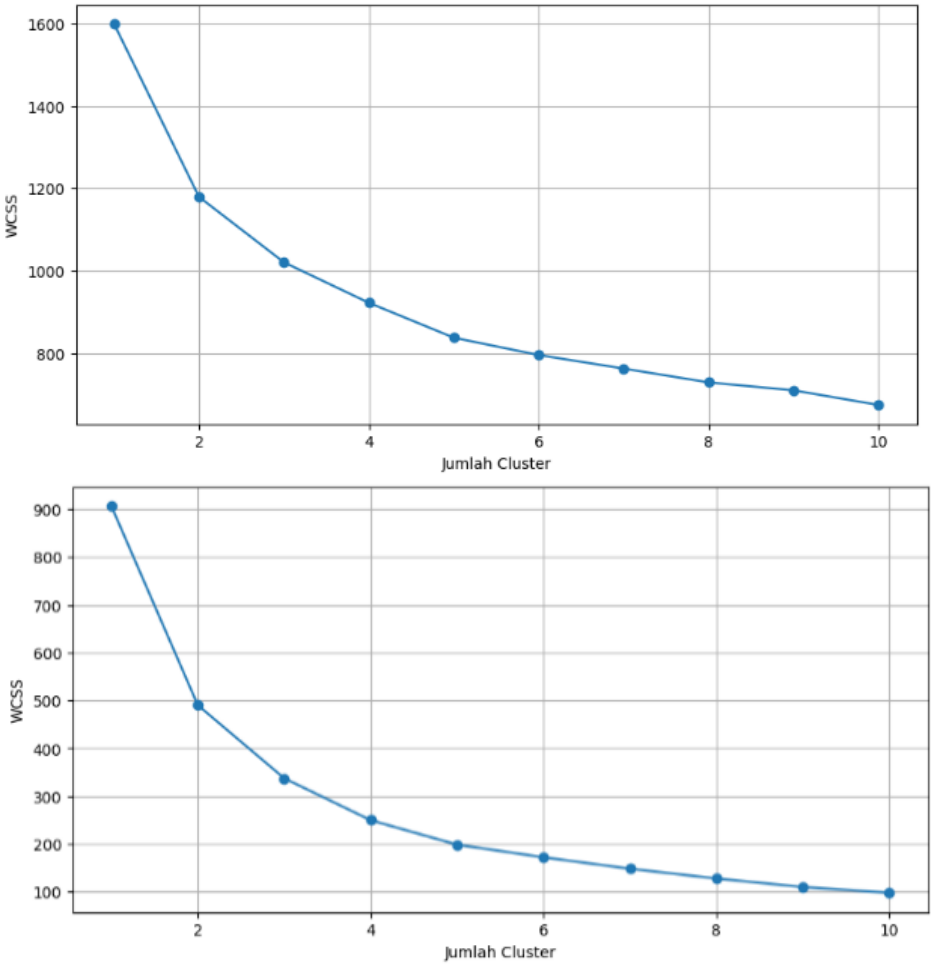
Principal Component 1	Principal Component 2	Cluster
-1.572.787	0.329325	0
-0.855722	-0.426146	1
-1.448.590	-1.340.699	0
-0.356611	0.334584	1
0.806024	-0.361326	1

Tabel 6 merepresentasikan struktur dimensi data setelah melalui tahap reduksi *Principal Component Analysis* (PCA) dan pelabelan *K-Means*. Komponen PC1 dan PC2 merupakan kombinasi linier dari seluruh atribut asli (meliputi capaian IPS dan faktor demografi) yang diproyeksikan ke dalam ruang dimensi baru. Dalam hal ini, PC1 merepresentasikan arah

variabilitas informasi yang paling dominan dari dataset, sementara PC2 secara ortogonal menangkap sisa variansi terbesar kedua. Transformasi ini berhasil memadatkan informasi esensial dan menyaring *noise* komputasional tanpa mereduksi karakteristik bawaan data. Integrasi label *Cluster* pada ruang PCA ini pada akhirnya memfasilitasi visualisasi persebaran kelompok secara dua dimensi, serta menciptakan fondasi pemodelan yang membuat algoritma K-Means bekerja jauh lebih efisien dan akurat.

Penentuan Jumlah Klaster Optimal dan Evaluasi Model

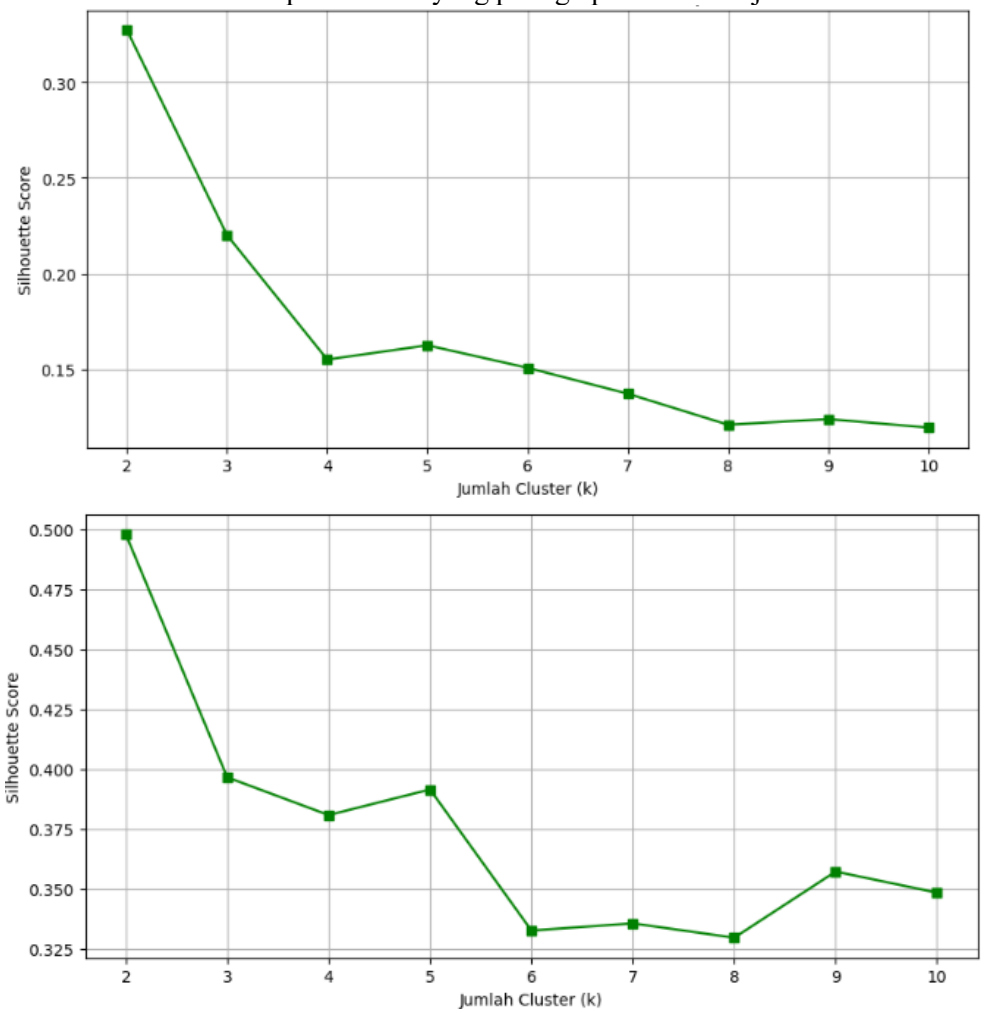
Langkah krusial dalam algoritma K-Means adalah menentukan parameter jumlah klaster (*k*) yang paling optimal. Untuk menghindari penentuan yang subjektif, penelitian ini menggunakan pendekatan kuantitatif berupa metode *Elbow* dan iterasi *Silhouette Score* untuk mencari titik terbaik.



Gambar 3. Analisis pengujian jumlah klaster (k) menggunakan Metode *Elbow* : (a) Sebelum reduksi dimensi dengan PCA; (b) Setelah reduksi dimensi dengan PCA

Gambar 3a dan Gambar 3b merepresentasikan pengujian jumlah klaster (k) menggunakan Metode Elbow berdasarkan evaluasi metrik Within-Cluster Sum of Squares (WCSS). Gambar 3a menunjukkan kurva Elbow pada data yang telah distandarisasi, sedangkan Gambar 3b menampilkan kurva pasca-reduksi dimensi menggunakan PCA. Kedua grafik tersebut menunjukkan pola pelandaian penurunan WCSS yang konsisten, di mana titik belok atau 'siku' mulai terbentuk pada rentang nilai k=2 hingga k=3. Kesamaan konfigurasi kurva sebelum dan sesudah PCA ini membuktikan bahwa proses reduksi dimensi berhasil mempertahankan struktur variansi esensial dari data asli tanpa mendistorsi persebaran klaster. Meskipun rentang k=2 dan

k=3 menjadi kandidat kuat, pendekatan Elbow secara inheren bersifat heuristik dan subjektif. Oleh karena itu, penelitian ini akan melakukan validasi lanjutan menggunakan pengujian Silhouette Score untuk menetapkan nilai k yang paling optimal dan objektif.



Gambar 4. Analisis pengujian *Silhouette Score* untuk validasi kualitas klasterisasi : (a) Sebelum reduksi dimensi dengan PCA; (b) Setelah reduksi dimensi dengan PCA

Gambar 4a dan Gambar 4b mengilustrasikan hasil pengujian metrik *Silhouette Score* untuk memvalidasi kualitas klasterisasi secara objektif pada data sebelum dan sesudah reduksi dimensi. Pada ruang data asli yang telah distandarisasi (Gambar 4a), konfigurasi optimal tercapai pada *k*=2 dengan perolehan skor sebesar 0.3275. Menariknya, setelah dilakukan ekstraksi fitur menggunakan PCA (Gambar 4b), konfigurasi optimal tetap konsisten pada *k*=2, namun mencatatkan lonjakan *Silhouette Score* yang signifikan menjadi 0.4979. Eskalasi metrik ini memberikan bukti bahwa penerapan PCA tidak hanya mereduksi beban komputasi, tetapi juga berhasil mengeliminasi noise dan redundansi fitur pada dataset profil kelulusan. Sebagai dampaknya, algoritma K-Means mampu mengidentifikasi batas antar-kelompok dengan lebih tegas, menghasilkan klaster yang jauh lebih kohesif secara internal dan terpisah secara maksimal (separasi) dibandingkan dengan pemodelan pada data asli.

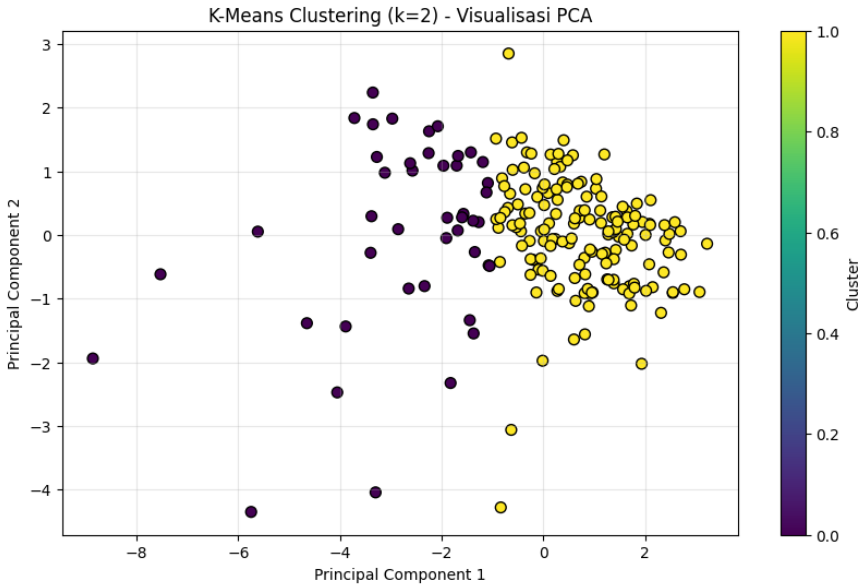
Meskipun metrik *Silhouette Score* telah memberikan indikasi kuat mengenai peningkatan kualitas separasi klaster, validasi silang menggunakan metrik evaluasi tambahan sangat direkomendasikan dalam penelitian *data mining*. Oleh karena itu, pengujian dilanjutkan dengan mengevaluasi tingkat *overlap* antarklaster dan dispersi varians menggunakan *Davies-Bouldin Index* (DBI) dan *Calinski-Harabasz Index* (CHI). Hasil komparasi performa algoritma K-Means pada data asli dibandingkan dengan data pasca-reduksi PCA disajikan pada Tabel 7.

Berdasarkan Tabel 7, metrik evaluasi lanjutan mengonfirmasi keunggulan dari penerapan PCA pada pemodelan *K-Means++*. Pada pengukuran *Davies-Bouldin Index* (DBI), terjadi perbaikan (penurunan) skor yang signifikan dari 1.405 menjadi 0.8715. Mengingat kriteria rasio DBI yang ideal adalah semakin mendekati angka nol, penurunan ini menjadi bukti bahwa transformasi PCA berhasil mereduksi tingkat tumpang-tindih (*overlap*) antar-kelompok mahasiswa, sehingga batas karakteristik setiap klaster menjadi jauh lebih spesifik dan terdefinisi dengan baik. Tren optimasi yang sama juga tervalidasi secara meyakinkan oleh kenaikan signifikan nilai *Calinski-Harabasz Index* (CHI), yang meningkat dari 70.488 menjadi 168.035. Skor CHI berbanding lurus dengan kualitas klasterisasi; semakin tinggi nilainya, semakin besar varians antar-klaster dibandingkan dengan varians di dalam klaster itu sendiri. Peningkatan skor CHI yang mencapai hampir dua kali lipat ini mengindikasikan bahwa kelompok profil kelulusan yang terbentuk pasca-reduksi dimensi memiliki tingkat kepadatan internal yang sangat tinggi dan tingkat perbedaan (separasi) eksternal yang tegas.

Tabel 7. Matriks evaluasi perbandingan clustering

Metrik	K-Means (k=2) - tanpa PCA	K-Means (k=2) - dengan PCA
Silhouette Score	0.3275	0.4978
Davies-Bouldin Index	1.405	0.8715
Calinski-Harabasz	70.488	168.035

Secara keseluruhan, kombinasi dari ketiga metrik evaluasi (*Silhouette Score*, DBI, dan CHI) menyajikan kesimpulan yang konsisten: ekstraksi 2 komponen utama melalui PCA bukan sekadar instrumen untuk meringankan beban komputasi, melainkan intervensi pemodelan yang krusial untuk mengurangi noise pada data bertipe campuran, sehingga *K-Means++* dapat mencapai performa klasterisasi yang jauh lebih akurat dan bermakna. Pembuktian separasi klaster ini dapat diamati secara visual pada Gambar 5.



Gambar 5. Visualisasi Persebaran Klaster Mahasiswa pada Ruang Reduksi PCA 2D.

Profiling dan Interpretasi Karakteristik Klaster

Berdasarkan hasil pemodelan K-Means++ pada ruang reduksi dua dimensi hasil *Principal Component Analysis* (PCA), dataset mahasiswa tersegmentasi secara jelas ke dalam dua klaster utama yang menunjukkan karakteristik akademik dan demografis yang kontras. Pemisahan ini

tidak hanya terlihat secara visual pada distribusi data, tetapi juga didukung oleh peningkatan signifikan pada metrik evaluasi klasterisasi.

Klaster pertama (Klaster 1), yang selanjutnya disebut sebagai Profil Akademik Konsisten, terdiri dari mayoritas mahasiswa dengan pola performa akademik yang stabil. Mahasiswa dalam klaster ini menunjukkan nilai Indeks Prestasi Semester (IPS) yang relatif tinggi dan konsisten dari semester awal hingga akhir. Selain itu, kelompok ini didominasi oleh mahasiswa yang tidak memiliki pekerjaan sampingan dan cenderung aktif dalam kegiatan organisasi. Kombinasi ini mengindikasikan bahwa alokasi waktu dan fokus akademik yang optimal berkontribusi terhadap kestabilan performa belajar.

Sebaliknya, Klaster kedua (Klaster 0), yang diidentifikasi sebagai Profil Multi-Peran Rentan, terdiri dari mahasiswa dengan performa akademik yang lebih fluktuatif dan cenderung lebih rendah. Karakteristik utama dari klaster ini adalah tingginya proporsi mahasiswa yang bekerja paruh waktu, yang berpotensi menyebabkan pembagian fokus antara tanggung jawab akademik dan pekerjaan. Hal ini tercermin dalam variabilitas nilai IPS antar semester yang lebih tinggi dibandingkan dengan klaster pertama.

Analisis lebih lanjut menunjukkan bahwa perbedaan antara kedua klaster tidak hanya terletak pada nilai rata-rata akademik, tetapi juga pada pola konsistensi performa. Klaster 1 menunjukkan variansi IPS yang rendah, yang mengindikasikan kestabilan akademik jangka panjang. Sebaliknya, Klaster 0 memiliki variansi IPS yang lebih tinggi, mencerminkan adanya ketidakstabilan performa yang kemungkinan dipengaruhi oleh faktor eksternal. Temuan ini menegaskan bahwa stabilitas akademik merupakan indikator penting dalam membedakan profil kelulusan mahasiswa, bukan sekadar capaian nilai rata-rata.

Meskipun terdapat kecenderungan dominasi program studi tertentu dalam masing-masing klaster, hasil visualisasi menunjukkan adanya irisan (*overlap*) antar program studi. Fenomena ini mengindikasikan bahwa performa akademik mahasiswa tidak sepenuhnya ditentukan oleh latar belakang program studi, melainkan lebih dipengaruhi oleh kombinasi faktor multidimensi seperti status pekerjaan, keterlibatan organisasi, serta pola belajar individu. Dengan demikian, pendekatan clustering berbasis data mining mampu mengungkap bahwa faktor perilaku dan kondisi sosial mahasiswa memiliki peran yang lebih signifikan dibandingkan faktor struktural semata.

Secara keseluruhan, hasil ini menunjukkan bahwa integrasi PCA dalam pipeline *K-Means++* tidak hanya meningkatkan kualitas pemisahan klaster secara kuantitatif, tetapi juga menghasilkan segmentasi yang lebih bermakna secara interpretatif. Klaster yang terbentuk mampu merepresentasikan fenomena nyata dalam konteks pendidikan tinggi, khususnya terkait dinamika mahasiswa dengan beban ganda. Temuan ini memberikan implikasi praktis bagi institusi pendidikan dalam merancang strategi intervensi berbasis data, seperti pengembangan sistem peringatan dini (*early warning system*) yang lebih adaptif terhadap karakteristik mahasiswa.

4. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa optimasi algoritma *K-Means++* yang diintegrasikan dengan teknik *One-Hot Encoding* dan *Principal Component Analysis* (PCA) mampu secara efektif mengatasi tantangan bias jarak spasial dan *curse of dimensionality* pada pemrosesan data rekam jejak akademik bertipe campuran. Hasil pengujian menunjukkan performa klasterisasi yang signifikan pasca-reduksi dimensi, di mana nilai *Silhouette Score* meningkat mencapai 0,4979, *Davies-Bouldin Index* membaik dan turun ke angka optimal 0,872, serta *Calinski-Harabasz Index* melonjak hingga 168.035. Kelebihan utama dari pemodelan ini adalah kemampuannya menyaring *noise* data sehingga dapat memetakan mahasiswa ke dalam dua karakteristik polar yang sangat tegas, yaitu kelompok Profil Akademik Konsisten berjumlah 157 mahasiswa dan kelompok Profil Multi-Peran Rentan berjumlah 43 mahasiswa. Pemisahan

yang tajam ini memberikan wawasan objektif bagi institusi untuk mendeteksi kerentanan akibat fenomena beban ganda pada mahasiswa pekerja.

5. SARAN

Hasil pengelompokan dua karakteristik mahasiswa memberikan dasar rekomendasi bagi institusi. Label klaster kemudian dianalisis dengan variabel target "Status Kelulusan", yang tidak digunakan saat pelatihan untuk menghindari bias (*data leakage*). Hubungan ini dapat dimanfaatkan sebagai sistem peringatan dini (*Early Warning System*) untuk mengidentifikasi mahasiswa berisiko, khususnya pada Klaster 0, sehingga pendampingan akademik dapat dilakukan lebih tepat sasaran. Namun, karena menggunakan pendekatan *unsupervised learning*, model ini hanya menghasilkan pengelompokan pola dan belum mampu memprediksi probabilitas kelulusan secara langsung. Meskipun model yang diusulkan menunjukkan peningkatan performa yang signifikan, penelitian ini memiliki beberapa keterbatasan. Dataset yang digunakan masih terbatas pada satu institusi pendidikan dengan jumlah sampel sebanyak 200 observasi, sehingga generalisasi hasil ke populasi yang lebih luas perlu dilakukan dengan hati-hati. Selain itu, penelitian ini belum mempertimbangkan variasi kurikulum antar institusi maupun faktor eksternal lainnya seperti kondisi sosial ekonomi mahasiswa.

DAFTAR PUSTAKA

[1] R. Swastyani and H. Santoso, "Perbandingan Algoritma Klasifikasi K-NN dengan Variasi Jarak, Naive Bayes, Logistic Regression, dan Decision Tree Untuk Prediksi Kelulusan Mahasiswa," JATI (Jurnal Mhs. Tek. Inform., vol. 9, no. 4, pp. 7057–7064, 2025, doi: https://doi.org/10.36040/jati.v9i4.14255.

[2] S. Junaidi, R. V. Anggela, and D. Kariman, "Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes , Random Forest , Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN)," J. Appl. Comput. Sci. Technol. (JACOST), vol. 5, no. 1, pp. 109–119, 2024, doi: https://doi.org/10.52158/jacost.v5i1.489.

[3] M. H. Sytar and Ermatita, "Prediksi Kelulusan Mahasiswa Tepat Waktu Dengan Metode Random Forest Berdasarkan Klasifikasi Algoritma K-Means," J. Pendidik. Mat. Judika Educ., vol. 8, no. 3, pp. 391–410, 2025, doi: https://doi.org/10.52436/1.jpti.577.

[4] U. Subagyo, A. B. Thoha, A. Syafrianto, H. Santoso, Siswaya, and R. Sanuri, "Customer Behavior Profiling in Wholesale Retail Using RFM Analysis and K-Means Clustering," 2025 4th Int. Conf. Electron. Represent. Algorithm, pp. 467–472, 2025, doi: https://doi.org/10.1109/ICERA66156.2025.11087354.

[5] T. H. Mardzuki, R. Lubis, and F. F. Adiwijaya, "Penerapan Algoritma K-Means Clustering pada Sistem Prediksi Kelulusan Tepat Waktu," Komputika J. Sist. Komput., vol. 13, no. 2, pp. 289–299, 2024, doi: https://doi.org/10.34010/komputika.v13i2.14097.

[6] D. Hosanna, N. Setiyawati, and H. D. Purnomo, "Comparison between K-Means and K-Means ++ Clustering Models Using Singular Value Decomposition (SVD) in Menu Engineering," Int. J. INFORMATICS Vis. Int. J., vol. 7, no. 3, 2023.

[7] R. P. Nugraha, G. F. Laxmi, and F. Riana, "Penerapan K-Means ++ untuk Pengelompokan Mahasiswa Berpotensi Drop Out (Studi Kasus : Universitas Ibn Khaldun Bogor)," JATI (Jurnal Mhs. Tek. Inform., vol. 8, no. 3, pp. 3493–3500, 2024, doi: https://doi.org/10.36040/jati.v8i3.9738.

[8] R. Rianti, R. Andarsyah, and R. M. Awangga, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," NUANSA Inform., vol. 18, no. 2, 2024, doi: https://doi.org/10.25134/ilkom.v18i2.211.

[9] I. M. Nur and Abdurakhman, "Application of the K-Means ++ Method for Grouping Health Services Based on Districts in West Java Province," EKSAKTA J. Sci. Data Anal., vol. 5, no. 1, pp. 96–102, 2024, doi: https://doi.org/10.20885/EKSAKTA.vol5.iss1.art11.
[10] K. Sa'diyah, K. D. Primadhieta, and H. Al Rosyid, "Clustering Wilayah Pulau Jawa Berdasarkan Indikator Sosial Ekonomi Menggunakan Metode K-Means," JATI (Jurnal Mhs. Tek. Inform., vol. 10, no. 1, pp. 1032–1036, 2026, doi: https://doi.org/10.36040/jati.v10i1.16934.
[11] M. D. Salman, N. R. Pratama, and M. N. F. A, "Comparison of K-Means and K-Medoids Clustering Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure.," Inst. Ris. dan Publ. Indones., vol. 5, no. July, pp. 797–806, 2025, doi: https://doi.org/10.57152/malcom.v5i3.1950.
[12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, And Advances In The Era Of Big Data," Inf. Sci. (Ny)., vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.
[13] S. Butsianto and A. Siswandi, "Implementasi K-Means Clustering Berbasis RapidMiner untuk Optimalisasi Segmentasi Penjualan Produk dalam Meningkatkan Efektivitas Strategi Pemasaran," J. Inf. Syst. Res., vol. 7, no. 1, pp. 200–210, 2025, doi: https://doi.org/10.47065/josh.v7i1.8439.
[14] F. D. Agustiar, B. N. Sari, and I. Maulana, "Penerapan Data Mining Untuk Pengelompokan Produk Penjualan Menggunakan Algoritma K-Means (Studi Kasus : Toko Agung Makmur Jaya)," JATI (Jurnal Mhs. Tek. Inform., vol. 9, no. 1, pp. 59–67, 2025, doi: https://doi.org/10.36040/jati.v9i1.12178.
[15] N. Nugroho and F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa," Teknika, vol. 11, no. 3, pp. 170-179., 2022, doi: https://doi.org/10.34148/teknika.v11i3.502.
[16] N. A. Maori, "Metode Elbow Dalam Optimasi Jumlah Cluster Pada K-Means Clustering," J. SIMETRIS, vol. 14, no. 2, pp. 277–287, 2023, doi: https://doi.org/10.24176/simet.v14i2.9630.