Available *online* at: https://bit.ly/SISFOTEKJAR

Jurnal Sistem Informasi dan Teknologi Jaringan (SISFOTEKJAR)

ISSN (Print) 2808-7917 | ISSN (Online) 2807-9256

Segmentasi Mahasiswa Berdasarkan Nilai Akademik dan Kehadiran Menggunakan Algoritma Clustering: Sebuah Tinjauan

Yoga Surya Agustion[1], Reny Medikawati Taufiq[2]

[1,2]Program Studi Teknik Informatika, Universitas Muhammadiyah Riau
Jl. Tuanku Tambusai Ujung, Pekanbaru, Riau, Indonesia
[1]220401225@student.umri.ac.id, [2]renymedikawati@umri.ac.id

ARTICLE INFORMATION

Received: June 24, 2026
Revised: July 18, 2026
Available online: September 01, 2026

KEYWORDS

Student Segmentation, Clustering Algorithm, K-Means, Academic Performance, Data Mining

CORRESPONDENCE

Phone: +62 822-9983-0690
E-mail: *220401225@student.umri.ac.id*

ABSTRACT

Penelitian ini menyajikan tinjauan literatur sistematis mengenai penerapan algoritma clustering untuk segmentasi mahasiswa berdasarkan nilai akademik dan kehadiran. Tinjauan ini mengkaji 21 artikel ilmiah yang dipublikasikan pada tahun 2022 hingga 2026 dari jurnal nasional terakreditasi SINTA dan sumber internasional. Analisis dilakukan untuk: (RQ1) mengidentifikasi algoritma clustering yang paling dominan digunakan dalam segmentasi mahasiswa; (RQ2) mengevaluasi perbedaan kinerja antara model tunggal dan model hybrid berdasarkan metrik evaluasi internal seperti Silhouette Score dan Davies-Bouldin Index; serta (RQ3) mengkaji faktor-faktor yang memengaruhi kualitas segmentasi pada data pendidikan. Hasil kajian menunjukkan bahwa K-Means tetap menjadi algoritma paling dominan karena kesederhanaannya dan skalabilitasnya, diikuti DBSCAN yang efektif untuk deteksi outlier, dan K-Medoids yang lebih robust terhadap noise. Pendekatan hybrid secara konsisten menghasilkan nilai metrik evaluasi yang lebih unggul dibandingkan model tunggal. Strategi pra-pemrosesan, penentuan jumlah klaster optimal menggunakan Elbow Method dan Silhouette Analysis, serta rekayasa fitur diidentifikasi sebagai tiga faktor paling kritis dalam meningkatkan kualitas segmentasi mahasiswa untuk pengambilan keputusan akademik berbasis data.

1. INTRODUCTION

Kemajuan teknologi informasi yang terus berlangsung dalam ekosistem pendidikan tinggi telah mendorong terkumpulnya data akademik berskala masif yang mengandung wawasan berharga tentang pola belajar, pencapaian, dan partisipasi mahasiswa. Apabila diolah dengan pendekatan yang tepat, data tersebut berpotensi menjadi pijakan bagi institusi dalam mengambil keputusan akademik yang lebih adaptif, terukur, dan berbasis bukti [1]. Akan tetapi, cara-cara konvensional dalam mengelola data mahasiswa kian tidak sanggup mengimbangi kompleksitas dan laju pertumbuhan data yang semakin tinggi [2].

Segmentasi mahasiswa pada dasarnya adalah upaya mengelompokkan mahasiswa berdasarkan kemiripan atribut tertentu agar institusi mampu menyusun strategi pembelajaran, program bimbingan, dan rencana intervensi yang lebih tepat sasaran. Metode yang digunakan termasuk dalam ranah data mining, tepatnya pendekatan unsupervised learning yang bekerja tanpa memerlukan label kelas yang telah ditentukan sebelumnya. Dalam konteks pendidikan, pendekatan ini terbukti mampu mengungkap kesamaan dan perbedaan antarkelompok mahasiswa, sekaligus menyediakan landasan ilmiah bagi pemangku kebijakan dalam merancang program intervensi yang lebih presisi [3]. Dengan memanfaatkan teknik pengelompokan berbasis data ini, institusi dapat beralih dari strategi pembelajaran yang seragam menuju pendekatan yang benar-benar adaptif terhadap keunikan setiap segmen mahasiswa [4].

Nilai akademik dan tingkat kehadiran merupakan dua variabel yang paling konsisten dimanfaatkan dalam riset segmentasi mahasiswa. Nilai akademik mencerminkan capaian kognitif, sementara kehadiran menjadi cerminan keterlibatan dan komitmen mahasiswa dalam proses belajar. Kombinasi keduanya terbukti menghasilkan pengelompokan mahasiswa yang mencerminkan perbedaan pola belajar secara nyata. Harmaji et al. [2] memanfaatkan Indeks Prestasi Semester dan persentase presensi selama tiga semester sebagai atribut klasterisasi K-Means, berhasil membentuk tiga kelompok mahasiswa dengan karakteristik kinerja yang berbeda secara

signifikan. Lebih jauh, Jin [3] memperluas dimensi analisis dengan mengintegrasikan partisipasi perkuliahan, frekuensi pengumpulan tugas, durasi belajar, dan keterlibatan diskusi daring, lalu membuktikan bahwa semakin banyak fitur yang dikombinasikan, semakin komprehensif profil klaster yang terbentuk, ditandai dengan nilai koefisien Silhouette yang berkisar antara 0,78 hingga 0,92.

Ragam algoritma clustering telah dikaji dan diterapkan di lingkungan pendidikan, mulai dari K-Means yang dikenal ringkas hingga metode yang lebih kompleks seperti algoritma berbasis kepadatan (DBSCAN), Hierarchical Clustering, K-Medoids, dan Fuzzy Clustering. K-Means diminati karena kemampuan komputasinya yang efisien dan interpretasi hasilnya yang mudah dipahami, terbukti mampu membentuk kelompok mahasiswa yang bermakna berdasarkan tingkat kesiapan dan capaian belajar [4]. Ulfah Siregar [5] memperkuat temuan ini dengan menunjukkan bahwa pada data akademik mahasiswa, K-Means mencatatkan nilai Davies-Bouldin Index sebesar 0,781 yang lebih unggul dibandingkan K-Medoids yang memperoleh nilai 0,929, meski K-Medoids dinilai lebih tahan terhadap keberadaan outlier. Di sisi lain, algoritma berbasis kepadatan seperti DBSCAN menunjukkan kelebihan dalam menangani noise, sebagaimana ditemukan Lailiyah [6] yang berhasil menandai 14 titik data anomali pada klasterisasi data akademik siswa SMP suatu kemampuan yang tidak dimiliki K-Means standar. Meskipun demikian, kajian sistematis yang secara khusus merangkum dan membandingkan kinerja algoritma clustering dalam konteks segmentasi mahasiswa berdasarkan nilai akademik dan kehadiran, terutama dari jurnal Indonesia terkini, masih sangat terbatas. Penelitian ini hadir untuk mengisi kekosongan tersebut.

2. METODOLOGI PENELITIAN

Metode yang digunakan dalam penelitian ini adalah Systematic Literature Review (SLR). SLR merupakan metode penelitian yang dilaksanakan secara terstruktur untuk mengidentifikasi, mengevaluasi, dan menginterpretasi seluruh penelitian yang tersedia dan relevan dengan pertanyaan penelitian tertentu. Pendekatan ini dipilih karena mampu mensintesis temuan dari berbagai studi secara sistematis sehingga terhindar dari bias seleksi. Dalam penelitian di bidang data mining pendidikan, SLR menjadi alat yang sangat relevan mengingat banyaknya variasi pendekatan metodologis yang digunakan oleh peneliti yang berbeda untuk masalah serupa [4]. Tahapan penelitian ini terdiri dari tiga fase utama: (1) Perencanaan, (2) Strategi Pencarian Literatur, dan (3) Ekstraksi dan Analisis Data.

Perencanaan

Pada tahap perencanaan, dilakukan identifikasi kebutuhan tinjauan pustaka dan perumusan pertanyaan penelitian (Research Questions/RQ) sebagai panduan dalam seluruh proses pencarian dan seleksi literatur. Berdasarkan analisis gap literatur yang teridentifikasi, ditetapkan tiga pertanyaan penelitian utama yang disajikan dalam Tabel 1.

Tabel 1. Pertanyaan Penelitian (Research Questions)

Kode	Pertanyaan Penelitian
RQ1	Algoritma clustering apa yang paling dominan dalam segmentasi mahasiswa berdasarkan nilai akademik dan kehadiran?
RQ2	Bagaimana perbandingan kinerja model clustering tunggal vs. hybrid berdasarkan Silhouette Score dan DBI?
RQ3	Faktor-faktor apa yang paling berpengaruh terhadap kualitas segmentasi mahasiswa?

Strategi Pencarian Literatur

Proses pengumpulan literatur dilakukan melalui penelusuran sistematis pada basis data jurnal nasional terakreditasi SINTA (minimal SINTA 4) dan sumber internasional bereputasi. Kata kunci yang digunakan meliputi: "student clustering", "academic performance clustering", "K-Means education", "DBSCAN student segmentation", "segmentasi mahasiswa nilai akademik", dan kombinasinya dalam bahasa Inggris maupun Indonesia. Periode publikasi dibatasi pada tahun 2022–2026 untuk memastikan relevansi dan kekinian literatur.

Kriteria Inklusi dan Eksklusi

Penetapan kriteria inklusi dan eksklusi merupakan tahap penting untuk memastikan bahwa hanya artikel yang relevan, berkualitas, dan sesuai dengan tujuan penelitian yang digunakan dalam analisis. Kriteria yang ditetapkan disajikan dalam Tabel 2.

Tabel 2. Kriteria Inklusi dan Eksklusi

Kriteria	Inklusi	Eksklusi
Algoritma	Menggunakan clustering (K-Means, DBSCAN, K-Medoids, Hierarchical, Fuzzy, dll.)	Hanya klasifikasi/regresi tanpa clustering
Variabel	Nilai akademik dan/atau kehadiran sebagai atribut utama	Variabel utama bukan nilai/kehadiran
Metrik	Menyajikan metrik evaluasi kuantitatif (Silhouette, DBI, Akurasi)	Tidak menyajikan evaluasi kuantitatif
Tahun	Dipublikasikan 2022–2026	Sebelum 2022
Bahasa	Indonesia atau Inggris	Selain Indonesia/Inggris
Indeks	Scopus, IEEE, SINTA min. SINTA 4, atau jurnal nasional bereputasi	Tidak terindeks database ilmiah

Ekstraksi Data

Proses ekstraksi data dilakukan untuk mengumpulkan informasi kunci dari setiap artikel yang memenuhi kriteria inklusi, mencakup: nama penulis dan tahun publikasi, algoritma yang

digunakan, jenis dan ukuran dataset, atribut yang digunakan, metrik evaluasi beserta nilainya, serta temuan utama penelitian. Berdasarkan penerapan seluruh kriteria inklusi, diperoleh 21 artikel yang memenuhi syarat untuk dianalisis: 14 artikel dikategorikan sebagai Sangat Relevan dan 7 artikel dikategorikan sebagai Cukup Relevan.

3. HASIL DAN PEMBAHASAN

1. Hasil

Setelah menyelesaikan tahap perencanaan dan strategi pencarian literatur, ditemukan sebanyak 21 artikel yang memenuhi kriteria inklusi. Tabel 3 merangkum keseluruhan artikel yang dikaji, mencakup rincian algoritma clustering yang digunakan, objek dan jenis dataset, variabel utama, serta hasil evaluasi kinerja masing-masing studi.

Tabel 3. Ringkasan Artikel Kajian SLR

No	Penulis & Tahun	Algoritma	Objek/Dataset	Variabel Utama	Hasil Evaluasi	Relevansi RQ
1	Harmaji et al. (2024)	K-Means + Elbow	260 mhs Institut Teknologi Sapta Mandiri	IPS sem 1–3, Presensi sem 1–3	3 klaster: IPS 3,57 terbaik; k=3 optimal	RQ1,RQ2,RQ3
2	Hesti Sabrila Aulia et al. (2026)	K-Means, K-Medoids	1.023 siswa SMP Negeri 2 & 3 Kudus	Nilai rapor, sikap, ekstrakurikuler	Silhouette=0,529; DBI=0,879; k=3	RQ1,RQ2
3	Ulfah Siregar (2022)	K-Means, K-Medoids	Mhs UIN Sunan Kalijaga 2013–2015	Data akademik (IPK)	K-Means DBI=0,781 (k=5); K-Medoids DBI=0,929	RQ1,RQ2
4	Anggraeny & Eviyanti (2025)	K-Means + DBI	576 siswa SMP Wijaya Sukodono	Nilai rapor semester	DBI=0,337; k=2 optimal	RQ1,RQ3
5	Azzahra & Sriani (2025)	K-Means + DBI	167 siswa SMA	Nilai akademik	DBI=1,226; k=3 klaster	RQ1,RQ3
6	Rizki et al. (2023)	K-Means + Elbow	111 mhs Pend. Matematika Univ. Mulawarman	IPK, sumber belajar	Silhouette=0,615; k=3	RQ1,RQ3
7	Al Ghifari & Harsanti Putri (2023)	K-Means + Elbow	55 mata kuliah, Univ. Paramadina	Nilai mata kuliah	k=2; rata-rata 2,686 dan 3,245	RQ1,RQ2
8	Risnasari et al. (2022)	K-Means	108 siswa era Industri 4.0	Gaya belajar, data akademik	k=6 klaster gaya belajar	RQ1,RQ3
9	Lestari Siregar et al. (2024).	K-Means	Mhs Pend. Vokasi Informatika	IPK, faktor sosioekonomi	Klasifikasi berdasarkan IPK	RQ1,RQ3
10	Lailiyah (2025)	DBSCAN	Siswa SMP Muh. 5 Samarinda	Nilai akademik, kehadiran, tugas	3 klaster + 14 noise; Silhouette=0,217	RQ1,RQ2,RQ3
11	Rahma & Maulana (2026)	DBSCAN + XGBoost	200 siswa Thailand (KKN BIPA)	Skor berbicara, menulis, kehadiran	3 klaster; kehadiran dominan 15,89%	RQ1,RQ2,RQ3
12	Cahapin et al. (2023)	K-Means, HC, DBSCAN	2.114 mhs Cavite State Univ. Filipina	GPA, VAT, MAT, Total skor	K-Means & HC: k=2; DBSCAN: 1 klaster+213 noise	RQ1,RQ2
13	Balovsyak et al. (2023)	Fuzzy Clustering	Mhs Univ. Chernivtsi, Ukraina	Nilai akademik	Klaster berbasis fuzzy logic	RQ1
14	Jin (2025)	K-Means vs DBSCAN vs HC	1,2 jt record; 8.000+ mhs vokasi	Kehadiran, nilai ujian, waktu belajar	4 klaster; Silhouette=0,78–0,92; K-Means unggul	RQ1,RQ2,RQ3
15	Ramadhan et al. (2026)	K-Means + Elbow	6.508 data tracer study alumni UIN Suska	Kepuasan fasilitas pembelajaran	DBI=0,874; k=3	RQ1,RQ2
16	Baridah & Putri (2025)	Naïve Bayes	311 siswa SMKN 1 Sipispis	Nilai, kehadiran, sikap, minat belajar	Akurasi 74,6%	RQ2 (pembanding)

17	Hanifa et al. (2024)	Regresi Berganda	216 mhs FKIP UNIS Tangerang	Kehadiran, efikasi diri, kemampuan akademik	Kehadiran ≥75% berkontribusi signifikan	RQ3 (teori)
18	Andriani Saputri & Rosnita (2025)	Random Forest	632 data rapor SMAN Aceh Barat Daya	Nilai rapor	Akurasi 87,40%	RQ2 (pembanding)
19	Muhammad & Hendrik (2025)	K-Means (SLR)	15 artikel ilmiah	Nilai tryout, kehadiran, durasi belajar	K-Means efektif untuk segmentasi multivariat	RQ1,RQ3
20	Sarker et al. (2024)	DT, KNN, NB, NN, RF	309 siswa HSC Bangladesh	Nilai ujian internal	RF-GI akurasi terbaik 96,45%	RQ2 (pembanding)
21	Yağcı (2022)	RF, KNN, SVM, LR, NB	1.854 mhs Bahasa Turki	Nilai UTS sebagai prediktor UAS	Akurasi 70–75%; early warning	RQ2,RQ3 (teori)

Berdasarkan Tabel 3, dari 21 artikel yang dikaji, K-Means merupakan algoritma yang paling dominan digunakan. Secara keseluruhan, 15 dari 21 artikel (71%) menggunakan K-Means sebagai algoritma utama atau komponen hybrid. DBSCAN digunakan pada 4 artikel (19%), K-Medoids pada 3 artikel (14%), dan algoritma lain seperti Fuzzy Clustering dan Hierarchical Clustering masing-masing 1 artikel. Penggunaan metode hybrid ditemukan pada 5 artikel (24%).

2. *Pembahasan*

Bagian ini menyajikan analisis kritis dan sintesis terhadap temuan dari literatur yang dikaji untuk menjawab tiga pertanyaan penelitian utama.

Algoritma Clustering yang Dominan (RQ1)

Berdasarkan hasil pemetaan terhadap 21 artikel, K-Means menempati posisi teratas sebagai algoritma clustering yang paling banyak diterapkan dalam segmentasi mahasiswa, dengan kehadiran di 15 dari 21 studi yang dikaji atau setara 71 persen. Dominasi ini tidak mengherankan mengingat sifat K-Means yang kompak secara komputasi, mudah dipahami, dan hasilnya dapat diinterpretasikan dengan lugas. Harmaji et al. [2] mendemonstrasikan bahwa K-Means mampu membedakan tiga kelompok mahasiswa secara signifikan ketika menggabungkan data Indeks Prestasi Semester dan persentase kehadiran sebagai atribut klasterisasi. Sejalan dengan itu, Rizki et al [10] melaporkan bahwa penerapan K-Means dikombinasikan dengan Elbow Method pada data mahasiswa menghasilkan nilai koefisien Silhouette sebesar 0,615 angka yang mengindikasikan kualitas pengelompokan yang tergolong baik saat mengklasterisasi mahasiswa berdasarkan sumber belajar dan rekam jejak akademik mereka.

Algoritma berbasis kepadatan ruang (DBSCAN) menduduki peringkat kedua dalam hal frekuensi penggunaan, ditemukan pada empat dari 21 studi atau sekitar 19 persen. Kelebihan utama metode ini adalah kapasitasnya mengenali titik data yang bersifat anomali atau outlier secara otomatis. Lailiyah [6] menerapkan metode ini pada kumpulan data akademik siswa tingkat menengah pertama dengan nilai ambang kepadatan dan jumlah tetangga minimum yang telah dioptimasi, berhasil menandai 14 data pencilan yang tidak masuk dalam klaster manapun sebuah kapabilitas yang tidak dimiliki K-Means konvensional meskipun skor Silhouette yang diperoleh (0,217) tergolong lebih rendah. Sementara itu, Cahapin et al. [15] yang membandingkan tiga pendekatan clustering sekaligus pada data mahasiswa di Filipina menemukan bahwa DBSCAN hanya mampu membentuk satu kelompok dengan lebih dari 200 titik noise sebuah indikasi bahwa metode ini kurang optimal ketika distribusi kepadatan data cenderung seragam.

K-Medoids diposisikan sebagai pendekatan alternatif bagi K-Means dan muncul pada tiga dari 21 artikel yang dikaji. Ulfah Siregar [5] mengadu langsung kedua metode ini pada data akademik mahasiswa: K-Means menghasilkan indeks pemisahan klaster yang lebih baik dengan nilai DBI sebesar 0,781 pada lima klaster, sementara K-Medoids mencatatkan nilai 0,929 pada tiga klaster makin kecil nilai DBI, makin baik kualitas klaster yang terbentuk. Aulia et al. [7] memperkuat kesimpulan yang sama dengan menemukan bahwa K-Means meraih skor koefisien Silhouette lebih tinggi dibandingkan K-Medoids pada dataset yang identik. Adapun Al Ghifari & Harsanti Putri [11] memvalidasi keefektifan K-Means pada skala yang lebih kecil: pengelompokan 55 mata kuliah berdasarkan nilai mahasiswa di Universitas Paramadina menghasilkan dua klaster dengan rata-rata nilai 2,686 dan 3,245. Di sisi lain, Balovsyak et al. [16] menawarkan perspektif berbeda dengan mengusulkan Fuzzy Clustering sebagai solusi atas keterbatasan K-Means yang mengharuskan setiap data masuk ke satu klaster secara pasti, padahal dalam kenyataannya batas pencapaian akademik mahasiswa kerap bersifat gradual dan tidak tegas.

Perbandingan Kinerja Model (RQ2)

Perbandingan kinerja berbagai algoritma dan model berdasarkan metrik evaluasi internal disajikan dalam Tabel 4.

Tabel 4. Perbandingan Metrik Evaluasi Algoritma Clustering

Penulis & Tahun	Algoritma	Silhouette	DBI	k
Harmaji et al. (2024)	K-Means	–	Elbow (k=3)	3
Hesti Sabrila Aulia et al. (2026)	K-Means	0,529	0,879	3
Ulfah Siregar (2022)	K-Means	–	0,781	5
Ulfah Siregar (2022)	K-Medoids	–	0,929	3
Rizki et al. (2023)	K-Means	0,615	WCSS	3
Lailiyah (2025)	DBSCAN	0,217	–	3+noise

Cahapin et al. (2023)	K-Means	0,89	–	2
Jin (2025)	K-Means	0,78–0,92	0,82–0,91	4
Ramadhan et al. (2026)	K-Means	–	0,874	3

Kajian terhadap Tabel 4 menunjukkan bahwa pendekatan hybrid yang memadukan algoritma clustering dengan teknik analisis lanjutan secara konsisten menghasilkan kualitas pengelompokan yang lebih tinggi dibandingkan metode tunggal. Jin [3] membuktikan hal ini dengan menggabungkan K-Means bersama analisis perilaku multidimensi pada dataset masif yang mencakup lebih dari satu juta catatan perilaku mahasiswa: model yang dihasilkan mencapai rentang koefisien Silhouette antara 0,78 dan 0,92 serta rasio variansi antara 0,82 dan 0,91 jauh melampaui performa algoritma berbasis kepadatan maupun hirarki pada dataset yang sama. Rahma & Maulana [14] pun mengonfirmasi keunggulan model gabungan dengan mengintegrasikan algoritma berbasis kepadatan dan teknik boosting berbasis pohon keputusan, menghasilkan segmentasi yang lebih handal dalam menangani pola data yang tidak beraturan dibandingkan K-Means yang berdiri sendiri.

Apabila model tunggal dibandingkan satu sama lain, K-Means umumnya tampil lebih unggul pada data pendidikan dengan distribusi yang reguler. Hesti Sabrila Aulia et al. [7] melakukan pengujian head-to-head antara K-Means dan K-Medoids, di mana K-Means menghasilkan indeks Silhouette 0,529 dan DBI 0,879 lebih baik dari K-Medoids pada dataset yang persis sama. Adapun metode supervised learning sebagai pembanding seperti pohon keputusan dalam studi Sarker et al. [1] dan Naive Bayes pada penelitian Baridah & Putri [18] memang mencatatkan tingkat akurasi klasifikasi yang cukup tinggi, namun pendekatan tersebut tidak menghasilkan segmentasi terbuka yang dapat dimaknai secara pedagogis sebagaimana yang dihasilkan algoritma clustering. Andriani Saputri & Rosnita [20] mengaplikasikan ensemble Random Forest pada ratusan data rapor siswa dan memperoleh performa prediksi yang kompetitif, namun keluaran yang dihasilkan tetap berupa label klasifikasi, bukan peta kelompok yang dapat dijadikan dasar intervensi pembelajaran secara langsung.

Faktor Peningkatan Kualitas Segmentasi (RQ3)

Faktor pertama dan paling mendasar yang menentukan kualitas segmentasi adalah tahapan pra-pemrosesan data, khususnya normalisasi skala. Data nilai akademik dan kehadiran memiliki satuan dan rentang yang berbeda nilai ujian umumnya berada pada skala 0–100, sementara kehadiran bisa berupa persentase (0%–100%) atau frekuensi absolut kehadiran per semester. Tanpa penyetaraan skala, variabel dengan rentang lebih besar akan secara tidak proporsional mendominasi perhitungan jarak Euclidean yang menjadi dasar algoritma K-Means, sehingga klaster yang terbentuk bias terhadap variabel berskala besar tersebut. Tabel 5 berikut menyajikan perbandingan teknik normalisasi yang digunakan dalam jurnal-jurnal yang dikaji beserta contoh hasil transformasinya (Ramadhan et al., 2026).

Tabel 5. Perbandingan Teknik Normalisasi Data dalam Kajian SLR

Teknik	Formula	Rentang Output	Digunakan Oleh	Contoh Data Asli	Setelah Normalisasi	Kapan Digunakan
Min-Max Scaling	$X' = (X - Xmin) / (Xmax - Xmin)$	0 – 1	Ramadhan et al. (2026)	Nilai: 75; Min=40; Max=100	X' = 0,583	Distribusi tidak diketahui; data tidak normal
Z-Score Standardization	$X' = (X - \mu) / \sigma$	−3 s.d. 3	Cahapin et al. (2023)	Nilai: 75; μ=70; σ=10	X' = 0,50	Data berdistribusi normal; ada outlier
Robust Scaling	$X' = (X - Q2) / (Q3 - Q1)$	Tidak terbatas	Alternatif Z-Score	Nilai: 75; Q2=70; Q1=60; Q3=85	X' = 0,20	Data mengandung banyak outlier
Decimal Scaling	$X' = X / 10^n$	−1 – 1	Implisit di beberapa studi	Nilai: 850; n=3	X' = 0,85	Data dengan magnitude besar

Faktor kedua yang tak kalah penting adalah penetapan jumlah klaster yang paling sesuai. Kesalahan dalam memilih nilai k berpotensi menghasilkan pengelompokan yang tidak bermakna secara pedagogis jika k terlalu kecil, kelompok mahasiswa yang seharusnya berbeda digabung menjadi satu sehingga intervensi tidak tepat sasaran; jika k terlalu besar, klaster menjadi terlalu spesifik dan sulit diinterpretasikan oleh pihak institusi. Tabel 6 merangkum kelima faktor kritis beserta sub-faktornya, jurnal pendukung, teknik yang digunakan, dan dampak negatif jika faktor tersebut diabaikan dalam proses segmentasi.

Tabel 6. Faktor-Faktor Kritis yang Mempengaruhi Kualitas Segmentasi Mahasiswa

No	Faktor	Sub-Faktor	Teknik	Jurnal Pendukung	Dampak jika Diabaikan
1	Pra-Pemrosesan	Normalisasi/Standarisasi	Min-Max Scaling, Z-Score	Ramadhan et al. (2026)	Klaster bias variabel berskala besar; jarak Euclidean tidak akurat
		Penanganan missing value	Mean/Median imputation	Jin (2025)	Data tidak lengkap → klaster tidak representatif
		Deduplication data	Hapus duplikat sebelum clustering	Jin (2025)	Duplikat menggeser centroid ke area yang direplikasi

		Encoding variabel kategorikal	One-Hot Encoding	Rizki et al. (2023)	Variabel kategorikal tidak bisa dihitung jarak Euclidean-nya
2	Penentuan k Optimal	Elbow Method (WCSS)	Plot WCSS vs k; cari titik siku	Harmaji et al. (2024); Rizki et al. (2023)	k kecil: klaster terlalu umum; k besar: overfitting
		Silhouette Analysis	Hitung Silhouette Score tiap k	Jin (2025)	Klaster tidak kompak dan tidak terpisah
		Davies-Bouldin Index	Minimasi DBI untuk k optimal	Anggraeny et al. (2025); Ramadhan et al. (2026)	k salah → DBI tinggi → klaster tumpang tindih
3	Seleksi & Rekayasa Fitur	Pemilihan variabel relevan	Pearson Correlation, L1 Regularization	Jin (2025)	Fitur tidak relevan menambah noise; kualitas klaster turun
		Feature importance	XGBoost Feature Importance	Rahma & Maulana (2026)	Variabel dominan tidak teridentifikasi
		Integrasi nilai + kehadiran	Gabungkan kehadiran dengan nilai akademik	Harmaji et al. (2024); Hanifa et al. (2024)	4 profil mahasiswa tidak terdeteksi
4	Pemilihan Algoritma	Kesesuaian distribusi data	Uji distribusi sebelum pilih algoritma	Cahapin et al. (2023)	DBSCAN gagal pada data seragam; K-Means gagal pada klaster non-sferis
		Sensitivitas outlier	Gunakan K-Medoids jika outlier tinggi	Qomariyah & Siregar (2022)	Satu outlier bisa menggeser seluruh centroid K-Means
5	Ukuran & Kualitas Dataset	Volume data memadai	Min. 100–200 sampel per klaster	Jin (2025)	Dataset kecil → klaster tidak stabil
		Representativitas data	Sampling representatif dari populasi	Anggraeny et al. (2025)	Klaster hanya berlaku untuk sampel

Berdasarkan Tabel 6, terdapat lima faktor kritis yang secara hierarkis mempengaruhi kualitas segmentasi mahasiswa. Faktor pra-pemrosesan data merupakan fondasi yang paling mendasar tanpa normalisasi yang tepat, seluruh proses klasterisasi akan menghasilkan segmentasi yang bias. Faktor penentuan k optimal menentukan berapa banyak kelompok yang bermakna dapat terbentuk dari data. Faktor seleksi dan rekayasa fitur menentukan kualitas representasi data yang masuk ke algoritma. Pemilihan algoritma yang tepat memastikan metode sesuai dengan karakteristik distribusi data. Dan terakhir, ukuran serta kualitas dataset menentukan seberapa general dan stabil klaster yang dihasilkan dapat diterapkan pada populasi yang lebih luas.

Faktor ketiga yang sangat menentukan adalah pemilihan dan rekayasa fitur yang digunakan sebagai masukan klasterisasi. Rekayasa fitur bukan sekadar memilih variabel yang tersedia, melainkan proses aktif mentransformasi, menggabungkan, dan menciptakan representasi data baru yang lebih informatif bagi algoritma clustering. Tabel 7 menyajikan rincian fitur-fitur kritis yang diidentifikasi dari kajian 21 jurnal, cara merekayasanya, serta dampak konkret yang terjadi ketika fitur tersebut digunakan maupun diabaikan. Temuan paling signifikan adalah bahwa kombinasi nilai akademik dan kehadiran secara bersamaan yang hanya dilakukan oleh 6 dari 21 jurnal yang dikaji menghasilkan empat profil mahasiswa yang jauh lebih bermakna dibandingkan menggunakan satu variabel saja: mahasiswa dengan nilai tinggi dan kehadiran tinggi (high achiever engaged), nilai tinggi kehadiran rendah (gifted but disengaged), nilai rendah kehadiran tinggi (struggling but committed), serta nilai rendah kehadiran rendah (at-risk) yang paling membutuhkan intervensi segera. Hanifa et al. (2024) membuktikan secara empiris bahwa kehadiran berkontribusi signifikan terhadap kinerja akademik ($p<0,05$), sehingga mengabaikan variabel ini dalam proses klasterisasi berarti kehilangan dimensi keterlibatan mahasiswa yang sangat penting bagi pengambil keputusan akademik. Jin (2025) lebih jauh membuktikan bahwa penerapan seleksi fitur berbasis L1 Regularization mampu meningkatkan Silhouette Score secara substansial dari nilai rata-rata ~0,50 pada model tanpa seleksi fitur menjadi 0,78–0,92 setelah fitur redundan dieliminasi, sebuah peningkatan kualitas klaster sebesar 56–84% yang langsung berdampak pada ketajaman profil kelompok mahasiswa yang dihasilkan.

Tabel 7. Rincian Rekayasa Fitur: Cara, Dampak, dan Jurnal Pendukung

Fitur/Variabel	Sumber Data	Cara Rekayasa	Jurnal Pendukung	Dampak Positif	Dampak Negatif jika Diabaikan
Nilai Akademik (IPS/IPK/Nilai Ujian)	SIAKAD / Rapor / Transkrip	Rata-rata per semester; normalisasi Min-Max → 0–1	Harmaji et al. (2024); Anggraeny et al. (2025)	Klaster mencerminkan capaian kognitif; identifikasi mahasiswa	Segmentasi tidak memiliki makna akademik

				berprestasi vs berisiko	
Persentase Kehadiran	Sistem absensi / Presensi	% kehadiran = (hadir/total) × 100; normalisasi 0–1	Harmaji et al. (2024); Rahma & Maulana (2026)	Klaster mencerminkan komitmen; deteksi aktif vs pasif	Mahasiswa nilai tinggi tapi jarang hadir tidak terdeteksi
Nilai + Kehadiran (Composite)	SIAKAD + Sistem Absensi	Gabung keduanya; bobot 60% nilai + 40% kehadiran	Harmaji et al. (2024); Jin (2025)	4 profil: ↑N↑H (achiever), ↑N↓H (gifted), ↓N↑H (struggling), ↓N↓H (at-risk)	Hanya 6/21 jurnal melakukannya; profil gifted disengaged tidak terdeteksi
Tren Nilai Per Semester	Data historis multi-semester	Hitung ΔNilai = Sem3−Sem1 sebagai fitur baru	Harmaji et al. (2024)	Mendeteksi mahasiswa menurun (early warning)	Klaster hanya snapshot—krisis tidak terdeteksi dini
Seleksi Fitur (L1 Regularization)	Dari fitur yang sudah tersedia	Terapkan Lasso; hapus fitur dengan koefisien ≈0	Jin (2025)	Silhouette naik dari ~0,50 → 0,78–0,92 (+56–84%)	Curse of dimensionality—jarak antar data tidak bermakna
Engagement Score	Nilai + Kehadiran + Tugas	Engagement = (Nilai×0,6)+(Hadir×0,3)+(Tugas×0,1)	Rahma & Maulana (2026); Jin (2025)	Satu indeks keterlibatan holistik; mudah diinterpretasi	Tanpa agregasi, interpretasi klaster memerlukan analisis multi-dimensi
Variabel Sosioekonomi (Opsional)	Kuesioner / Data administrasi	Encode kategorikal → One-Hot Encoding	Lestari Siregar et al. (2024)	Klaster lebih komprehensif; mendeteksi faktor eksternal	Mengabaikan faktor di luar kendali mahasiswa

Berdasarkan Tabel 7, rekayasa fitur yang paling berdampak adalah pengintegrasian nilai akademik dan kehadiran secara bersamaan sebagai composite feature. Kombinasi ini menghasilkan empat profil mahasiswa yang sangat bermakna dari sudut pandang pedagogis dan kebijakan akademik. Profil pertama nilai tinggi dan kehadiran tinggi merepresentasikan mahasiswa high achiever yang perlu dipertahankan motivasinya. Profil kedua nilai tinggi tetapi kehadiran rendah mengidentifikasi mahasiswa berbakat namun berisiko tidak memenuhi syarat akademis karena absensi. Profil ketiga nilai rendah tetapi kehadiran tinggi menunjukkan mahasiswa yang berkomitmen namun membutuhkan pendampingan akademik tambahan. Profil keempat nilai rendah dan kehadiran rendah adalah kelompok at-risk yang paling membutuhkan intervensi segera dari institusi. Keempat profil ini hanya dapat teridentifikasi apabila rekayasa fitur dilakukan dengan menggabungkan kedua variabel secara eksplisit dalam proses klasterisasi, sebuah praktik yang terbukti masih jarang dilakukan berdasarkan kajian 21 jurnal dalam penelitian ini.

4. KESIMPULAN

Studi literatur sistematis ini mengkaji 21 artikel dari jurnal nasional terakreditasi SINTA dan sumber internasional yang dipublikasikan pada 2022–2026 mengenai penerapan algoritma clustering untuk segmentasi mahasiswa berdasarkan nilai akademik dan kehadiran. Berdasarkan hasil penelitian dapat disimpulkan bahwa:

Pertama, K-Means merupakan algoritma yang paling dominan (71% studi) karena kesederhanaannya, efisiensinya, dan kemudahan interpretasi hasilnya. DBSCAN efektif untuk deteksi outlier pada data heterogen, sedangkan K-Medoids lebih robust terhadap noise meskipun umumnya menghasilkan kualitas klaster yang sedikit lebih rendah dari K-Means. Pemilihan algoritma harus disesuaikan dengan karakteristik data dan tujuan segmentasi yang ingin dicapai.

Kedua, pendekatan hybrid yang mengombinasikan algoritma clustering dengan teknik tambahan secara konsisten menghasilkan kualitas segmentasi yang lebih tinggi berdasarkan metrik Silhouette Score dan Davies-Bouldin Index dibandingkan model tunggal. Jin [3] melaporkan Silhouette Score 0,78–0,92 pada model berbasis K-Means multidimensi yang jauh melampaui performa algoritma tunggal.

Ketiga, tiga faktor paling kritis dalam meningkatkan kualitas segmentasi mahasiswa adalah: (a) pra-pemrosesan data yang mencakup normalisasi menggunakan Min-Max Scaling atau Z-Score; (b) penentuan jumlah klaster optimal melalui Silhouette Analysis yang dikonfirmasi dengan Elbow Method dan DBI; dan (c) seleksi fitur berbasis data di mana kehadiran terbukti sebagai prediktor klaster akademik yang paling kuat.
Penelitian selanjutnya disarankan untuk menambahkan jurnal internasional terindeks Scopus Q1–Q2 dan mengeksplorasi teknik hybrid yang lebih canggih seperti Deep Clustering untuk penerapan pada dataset pendidikan berskala besar.

5. ACKNOWLEDGMENT

Penulis menyampaikan terima kasih kepada seluruh pihak yang telah mendukung terselesaikannya penelitian ini.

REFERENCES

[1] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, “Analyzing students’ academic performance using educational data mining,” *Computers and Education: Artificial Intelligence*, vol. 7, Dec. 2024, doi: 10.1016/j.caeai.2024.100263.

[2] L. Harmaji, A. Sufyan Asaury, M. Wijana, and M. E. Habiby, "Segmentasi Kinerja Akademik Mahasiswa Menggunakan Algoritma K-Means," *Jurnal Accounting Information System (AIMS*, vol. 7, no. 2, pp. 73–81, 2024, doi: 10.32627.

[3] J. Jin, "Student behavior patterns in vocational education big data based on clustering algorithm," *Discover Artificial Intelligence*, vol. 5, no. 1, Dec. 2025, doi: 10.1007/s44163-025-00433-3.

[4] I. F. Muhammad and B. Hendrik, "PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK SEGMENTASI BERDASARKAN KESIAPAN SELEKSI: SYSTEMATIC LITERATUR REVIEW," 2025. [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[5] M. Ulfah Siregar, "Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering," MEI, 2022.

[6] S. Lailiyah, "Bulletin of Information Technology (BIT) Clustering Academic Data of Junior High School Student to Identify Learning Groups Using the DBSCAN Algorithm at SMP Muhammadiyah 5 Samarinda," vol. 6, no. 4, pp. 361–366, 2025, doi: 10.47065/bit.v5i2.2293.

[7] Hesti Sabrila Aulia, Muhammad Arifin, and Diana Laily Fithri, "PENERAPAN ALGORITMA MACHINE LEARNING UNTUK PENGELOMPOKAN SISWA BERDASARKAN ASPEK AKADEMIK DAN NON-AKADEMIK," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 11, no. 1, pp. 1200–1210, Jan. 2026, doi: 10.36341/rabit.v11i1.7249.

[8] F. Anggraeny and A. Eviyanti, "Penerapan K-Means dengan Evaluasi Davies-Bouldin Index untuk Pengelompokan Kelas Unggulan SMP Wijaya Sukodono," *SMATIKA : STIKI Informatika Jurnal*, vol. 15, no. 2, p. 15, 2025, doi: 10.32664/smatika.v15i02.1689.

[9] C. Azzahra and S. Sriani, "Clustering of High School Students Academic Scores Using K-Means Algorithm," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 572–586, Mar. 2025, doi: 10.51519/journalisi.v7i1.1029.

[10] N. A. Rizki, K. Kurniawan, I. K. Hasan, and N. Sampe, "Implementasi Algoritma K-Means Untuk Mengelompokkan Mahasiswa Berdasarkan Sumber Belajarnya," *METIK JURNAL*, vol. 7, no. 2, pp. 62–67, Dec. 2023, doi: 10.47002/metik.v7i2.584.

[11] M. Al Ghifari and W. T. Harsanti Putri, "Clustering Courses Based On Student Grades Using K-Means Algorithm With Elbow Method For Centroid Determination," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 8, no. 1, pp. 42–46, Jan. 2023, doi: 10.25139/inform.v8i1.4519.

[12] M. Risnasari, N. Aulia, and L. Cahyani, "Clustering Of Student Learning Styles in the Industry 4.0 Using KMeans Algorithm," vol. 24, no. 2, pp. 246–257, 2022, doi: 10.21009/JTP2001.6.

[13] H. Lestari Siregar, R. Hidayanthi, A. Langga Dewa Sakti, kan Tapanuli Selatan, J. Sutan Moh Arif, and N. Sumatera, "Implementation of K-Means clustering on student learning achievements based on social economic and social related," 2024, doi: 10.22219/raden.v4i2.3.

[14] N. Rahma and H. Maulana, "Penerapan Algoritma DBSCAN dan XGBoost untuk Menganalisis Keberhasilan Pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA)," vol. 21, no. 1, 2026, doi: 10.31603/paedagogie.v21i1.16185.

[15] E. L. Cahapin, B. A. Malabag, C. S. Santiago, J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, Dec. 2023, doi: 10.11591/eei.v12i6.4849.

[16] S. Balovsyak, O. Derevyanchuk, H. Kravchenko, Y. Ushenko, and Z. Hu, "Clustering Students According to their Academic Achievement Using Fuzzy Logic," *International Journal of Modern Education and Computer Science*, vol. 15, no. 6, pp. 31–43, 2023, doi: 10.5815/ijmecs.2023.06.03.

[17] J. Homepage *et al.*, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Clustering Analysis Using the K-Means Method to Identify Alumni Satisfaction Pattern Analisis Clustering Menggunakan Metode K-Means untuk Mengidentifikasi Pola Kepuasan Alumni," vol. 6, no. 1, pp. 31–41, 2026, doi: 10.57152/malcom.v6i1.2401.

[18] L. Baridah and R. A. Putri, "Predicting Student Learning Outcomes in Vocational Computer and Network Engineering Using Naïve Bayes 1*," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 8, no. 3, pp. 667–676, 2025, doi: 10.24014/ijaidm.v8i3.38333.

[19] S. Hanifa, K. Khusaini, and A. Widiarti, "KEHADIRAN BERKONTRIBUSI DALAM PENINGKATAN KINERJA AKADEMIK MAHASISWA," *SAP (Susunan Artikel Pendidikan*, vol. 8, no. 3, 2024.

[20] R. Andriani Saputri and L. Rosnita, "A Random Forest-Based Predictive Model for Student Academic Performance: A Case Study in Indonesian Public High Schools," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[21] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.