

PENERAPAN MODEL HYBRID K-MEANS-GAUSSIAN MIXTURE MODEL UNTUK PERSONALISASI REKOMENDASI STUDI LANJUT BERDASARKAN POTENSI AKADEMIK

Putri Efrida Marbun¹, Imran Lubis², I Gusti Prahmana³

Sistem Informasi, STMIK Kaputama Binjai, Sumatera Utara

E-mail: [*putriefrida4@gmail.com](mailto:putriefrida4@gmail.com)¹, imran.loebis.medan@gmail.com², igustiprahmana4@gmail.com³

ABSTRAK

Pemilihan jalur studi lanjut yang tidak tepat setelah lulus sekolah menengah atas sering terjadi akibat keterbatasan analisis data prestasi siswa secara objektif dan sistematis di sekolah. Meskipun institusi pendidikan mengumpulkan data akademik secara rutin, pemanfaatannya masih bersifat manual dan administratif sehingga gagal mendeteksi pola potensi siswa yang kompleks dan multidimensi. Penelitian ini bertujuan menerapkan model *clustering hybrid K-Means-Gaussian Mixture Model* (GMM) untuk melakukan segmentasi prestasi akademik siswa guna menghasilkan personalisasi rekomendasi studi lanjut yang akurat dan berbasis data. Model *hybrid* ini mengintegrasikan kekuatan inisialisasi *centroid* dari *K-Means* dengan fleksibilitas probabilitas posterior dari GMM dalam menangani ketidakpastian serta tumpang tindih data. Penelitian dilakukan menggunakan data sekunder dari 566 siswa SMA Negeri 9 Medan dengan enam variabel mata pelajaran inti. Evaluasi kualitas pengelompokan pada konfigurasi 2 hingga 6 kluster menunjukkan bahwa konfigurasi 4 kluster menghasilkan performa terbaik dengan nilai *Davies-Bouldin Index* sebesar 0,161706 dan *Silhouette Coefficient* sebesar 0,870467. Hasil akhir memetakan 566 siswa ke dalam kluster C1 (241 siswa) dan C3 (124 siswa) yang dikategorikan sebagai segmentasi Sains, Teknologi, Logika & STEM, serta kluster C2 (36 siswa) dan C4 (165 siswa) sebagai segmentasi Komunikasi, Literasi & Teknologi Digital. Hasil segmentasi ini berhasil diintegrasikan dengan basis aturan berbasis regulasi kementerian untuk memberikan rekomendasi program studi dan prospek karir yang spesifik. Penelitian ini menyimpulkan bahwa model *hybrid K-Means-GMM* sangat efektif dalam mengidentifikasi potensi akademik siswa serta memberikan kontribusi nyata bagi layanan bimbingan karir sekolah.

Kata kunci

Segmentasi Prestasi, Clustering Hybrid, K-Means, Gaussian Mixture Model, Rekomendasi Studi Lanjut.

ABSTRACT

Inappropriate choices of further study pathways after graduating high school often occur due to the limitations of objective and systematic analysis of student performance data in schools. Although educational institutions collect academic data routinely, its utilization is still manual and administrative, failing to detect complex and multidimensional student potential patterns. This study aims to apply a hybrid K-Means-Gaussian Mixture Model (GMM) clustering model to segment high school students' academic performance to generate personalized and objective recommendations for further studies. This hybrid model integrates the centroid initialization strength of K-Means with the posterior probability flexibility of GMM in handling data uncertainty and overlapping distributions. The research was conducted using a secondary dataset of 566 students from SMAN 9 Medan, encompassing six core subject variables. Clustering quality evaluation across cluster configurations ranging from 2 to 6 indicated that the 4-cluster configuration achieved the best performance, with a Davies-Bouldin Index value of 0.161706 and a Silhouette Coefficient of 0.870467. The final results distributed the 566 students into C1 (241 students) and C3 (124 students), categorized as Science, Technology, Logic & STEM segments, and C2 (36 students) and C4 (165 students), categorized as Communication, Literacy & Digital Technology segments. These segmentation results were then integrated with ministry

regulation-based rules to provide specific study program and career recommendations. This study concludes that the hybrid K-Means–GMM model is highly effective in segmenting students' academic potential, providing a concrete decision-making foundation for school career counseling services.

Keywords

Academic Performance Segmentation, Hybrid Clustering, K-Means, Gaussian Mixture Model, Further Study Recommendation.

1. PENDAHULUAN

Perkembangan teknologi informasi telah mentransformasi data akademik dari sekadar arsip administratif menjadi aset strategis dalam proses pengambilan keputusan di institusi pendidikan. Pada tingkat Sekolah Menengah Atas (SMA), pengelolaan data prestasi menjadi krusial karena siswa berada pada fase transisi menentukan arah studi lanjut dan karir. Data akademik yang dikumpulkan oleh lembaga pendidikan menyimpan informasi berharga yang tidak hanya mengevaluasi performa masa lalu, melainkan juga memiliki kekuatan prediktif terhadap keberhasilan masa depan siswa (Pu et al., 2024). Oleh karena itu, pemanfaatan *Educational Data Mining* (EDM) menjadi fenomena penting dalam mengekstrak pengetahuan tersembunyi guna mendukung personalisasi layanan pendidikan.

Namun, pada realitasnya, analisis potensi akademik siswa di banyak sekolah masih dilakukan secara konvensional dan manual. Pendekatan manual ini sering kali gagal menangkap pola-pola kompetensi yang kompleks dan multidimensi, yang pada akhirnya menyebabkan pemberian rekomendasi studi lanjut menjadi kurang objektif. Ketidaksesuaian antara pilihan jalur studi dan potensi aktual siswa berisiko menurunkan tingkat keberhasilan akademik di jenjang pendidikan tinggi. Urgensi permasalahan ini diperkuat oleh temuan empiris bahwa sekitar 81,2% guru Bimbingan dan Konseling (BK) sangat membutuhkan sistem rekomendasi karir berbasis data untuk meminimalisir subjektivitas dalam asesmen minat dan bakat (Hidayat et al., 2025a).

Penelitian terdahulu telah berupaya menyelesaikan masalah ini dengan berbagai pendekatan algoritma *clustering*. Rahayu et al. (2024) menggunakan algoritma K-Means tunggal untuk memetakan siswa berprestasi, namun masih terbatas pada identifikasi kategori tinggi, sedang, dan rendah tanpa mempertimbangkan probabilitas keanggotaan data. Di sisi lain, Nimy dan Mosia (2023) membandingkan K-Means dengan *Gaussian Mixture Model* (GMM) untuk menganalisis keterlibatan siswa dan menemukan bahwa GMM lebih unggul dalam menangkap ketidakpastian melalui pemberian probabilitas keanggotaan. Sementara itu, Saravanakumar et al. (2024) membuktikan bahwa pendekatan *hybrid* yang mengintegrasikan GMM dengan algoritma optimasi mampu meningkatkan akurasi klasterisasi secara signifikan dibandingkan penggunaan algoritma tunggal pada pengolahan data berukuran besar.

Meskipun penelitian-penelitian tersebut telah memberikan kontribusi signifikan, terdapat keterbatasan kritis dalam implementasinya di bidang pendidikan. Sebagian besar penelitian cenderung menggunakan metode klasterisasi tunggal yang kurang optimal dalam menangani distribusi data yang tumpang tindih (*overlapping*), sebagaimana sering ditemukan pada distribusi nilai siswa (Adek Maulidya et al., 2024). Selain itu, terdapat keterpisahan antara proses klasterisasi dengan sistem rekomendasi fungsional; penelitian klasterisasi sering kali berhenti pada tahap pengelompokan data tanpa memberikan interpretasi yang dapat ditindaklanjuti dalam bentuk saran studi lanjut yang spesifik. Konteks variabel yang digunakan juga sering kali terbatas pada nilai

kognitif umum tanpa mengaitkannya dengan standar mata pelajaran pendukung program studi yang relevan secara institusional.

Research gap utama dalam penelitian ini terletak pada belum adanya integrasi antara analisis klasterisasi berbasis probabilistik dengan sistem rekomendasi jalur studi lanjut yang terhubung langsung dengan standar kebijakan pendidikan. Untuk mengisi celah tersebut, penelitian ini mengusulkan *novelty* berupa penerapan model *Hybrid K-Means-GMM* yang menggabungkan efisiensi inisialisasi jarak K-Means dengan fleksibilitas probabilistik GMM. Keunggulan model ini adalah kemampuannya menyeimbangkan pendekatan *person-centered* untuk segmentasi dan pendekatan *variable-centered* untuk membangun sistem rekomendasi berbasis aturan (*rule-based*) yang aplikatif (Niu & Reeping, 2024). Pendekatan ini memastikan bahwa setiap segmen siswa mendapatkan rekomendasi studi lanjut yang personal dan objektif berdasarkan pola prestasi akademik yang heterogen.

Tujuan dari penelitian ini adalah untuk menganalisis efektivitas model *Hybrid K-Means-GMM* dalam menghasilkan segmentasi prestasi siswa yang akurat dan representatif, serta memanfaatkannya sebagai dasar sistem rekomendasi studi lanjut dan karir. Secara akademis, penelitian ini memberikan kontribusi pada pengembangan EDM melalui penerapan pendekatan *hybrid* dalam analisis data pendidikan. Secara praktis, hasil penelitian ini diharapkan dapat membantu siswa dan guru BK di SMA Negeri 9 Medan dalam melakukan pemetaan potensi secara objektif, sehingga proses pengambilan keputusan studi lanjut menjadi lebih terstruktur dan berbasis data guna meningkatkan kualitas layanan pendidikan.

2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian ini dirancang sebagai penelitian kuantitatif eksperimental yang menerapkan pendekatan *Educational Data Mining* (EDM). Desain penelitian ini mengintegrasikan pendekatan berfokus pada individu (*person-centered approach*) dengan pendekatan berfokus pada variabel (*variable-centered approach*) secara berurutan (Niu dan Reeping, 2024). Pendekatan *person-centered* diwujudkan melalui eksperimen klasterisasi hibrida menggunakan kombinasi algoritma *K-Means* dan *Gaussian Mixture Model* (GMM) untuk mengidentifikasi kelompok-kelompok laten siswa yang heterogen berdasarkan pola prestasi akademiknya. Selanjutnya, hasil segmentasi tersebut ditindaklanjuti dengan pendekatan *variable-centered* berupa perancangan sistem rekomendasi berbasis aturan (*rule-based*) untuk memetakan klaster siswa ke dalam rekomendasi program studi dan prospek karir yang fungsional. Alasan pemilihan pendekatan hibrida ini adalah untuk mengatasi keterbatasan metode klasterisasi tunggal berbasis partisi jarak seperti *K-Means* yang bersifat kaku (*hard clustering*). Dengan mengintegrasikan GMM (*soft clustering*), ketidakpastian posisi data prestasi siswa yang berada di batas wilayah antarkelompok dapat ditangani secara objektif melalui perhitungan nilai probabilitas keanggotaan posterior.

2.2 Objek, Batasan, dan Ruang Lingkup

Objek utama penelitian ini adalah pola prestasi akademik siswa sekolah menengah atas yang direpresentasikan melalui nilai akumulasi mata pelajaran. Unit analisis terdiri dari data nilai rapor individual siswa kelas XI dan XII di SMA Negeri 9 Medan. Ruang lingkup dan batasan penelitian ditetapkan sebagai berikut:

- a) Variabel analisis dibatasi pada 6 (enam) mata pelajaran inti yang relevan dengan skema seleksi nasional masuk perguruan tinggi.
- b) Eksplorasi difokuskan pada dimensi akademik, tanpa melibatkan variabel non-akademik seperti demografi sosial-ekonomi atau kondisi psikologis.

- c) Algoritma pembandingan dibatasi pada skema hibrida K-Means–GMM tanpa mengevaluasi metode pengelompokan hierarki (*Hierarchical Clustering*) atau berbasis kerapatan (*DBSCAN*).
- d) Penentuan aturan rekomendasi berbasis keputusan (*rule-based decision*) diturunkan secara langsung dari nilai rata-rata pusat (*centroid*) tiap kluster akademik yang diselaraskan dengan regulasi pendidikan nasional.

2.3 Data dan Sumber Data

Penelitian ini menggunakan data sekunder kuantitatif kontinu (skala 0–100) yang diperoleh dari arsip rekapitulasi penilaian kurikulum SMA Negeri 9 Medan semester berjalan. Populasi pengujian mencakup 566 rekaman data siswa kelas XI dan XII. Kriteria inklusi data mensyaratkan kelengkapan nilai tanpa adanya data hilang (*missing value*) pada seluruh variabel utama. Izin riset diperoleh secara resmi dari pihak sekolah, dan prinsip etika penelitian dijaga melalui anonimisasi identitas siswa menjadi kode rekapitulasi (S01, S02, ..., S566).

2.4 Operasionalisasi Variabel

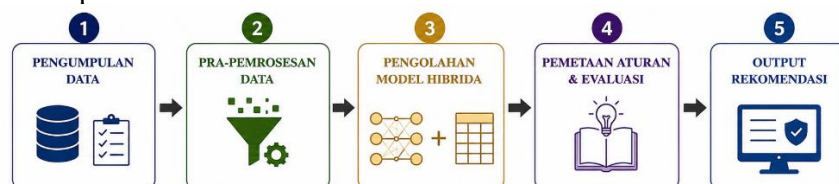
Variabel penelitian mencakup enam mata pelajaran inti yang merepresentasikan multidimensi kompetensi akademik siswa, seperti disajikan pada Tabel 1.

Tabel 1. Operasionalisasi Variabel Penelitian

No	Variabel	Definisi Operasional	Indikator	Skala	Sumber Data
1	Bahasa Indonesia (X1)	Mengukur kemampuan literasi, komunikasi verbal, dan penulisan ilmiah.	Nilai rapor Bahasa Indonesia	Rasio (0–100)	SMAN 9 Medan
2	Bahasa Inggris (X2)	Mengukur penguasaan bahasa asing dan pemahaman literasi global.	Nilai rapor Bahasa Inggris	Rasio (0–100)	SMAN 9 Medan
3	Matematika (X3)	Mengukur kemampuan penalaran logis, numerik, dan penyelesaian masalah.	Nilai rapor Matematika	Rasio (0–100)	SMAN 9 Medan
4	IPA (X4)	Mengukur kompetensi penalaran sains terapan dan eksakta.	Rata-rata nilai gabungan Fisika, Kimia, dan Biologi	Rasio (0–100)	SMAN 9 Medan
5	IPS (X5)	Mengukur pemahaman fenomena sosial, kemasyarakatan, dan dinamika ekonomi.	Rata-rata nilai gabungan Ekonomi, Geografi, Sosiologi, dan Sejarah	Rasio (0–100)	SMAN 9 Medan
6	Informatika (X6)	Mengukur kecakapan teknologi informasi, logika komputasi, dan literasi digital.	Nilai rapor Informatika	Rasio (0–100)	SMAN 9 Medan

2.5 Tahapan dan Prosedur Penelitian

Eksperimen komputasi dijalankan secara sistematis melalui alur *Educational Data Mining* seperti ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alir Komputasi Penelitian

1. Pra-pemrosesan Data (*Data Preprocessing*):
 - a. *Data Cleaning*: Memeriksa dan memastikan ketiadaan nilai kosong (*missing value*) atau data inkonsisten yang berpotensi merusak perhitungan matriks kovarians.
 - b. *Feature Engineering*: Menggabungkan sub-mata pelajaran IPA (Fisika, Kimia, Biologi) dan IPS (Ekonomi, Geografi, Sosiologi, Sejarah) menjadi indikator komposit tunggal melalui perhitungan rata-rata tertimbang.
 - c. *Normalisasi*: Menyeimbangkan rentang skala nilai seluruh variabel guna mengeliminasi bias dominasi jarak Euclidean pada variabel tertentu.
2. Permodelan Hibrida K-Means–GMM:
 - a. Tahap I (K-Means): Mengelompokkan $n = 566$ data siswa $x_i \in \mathbb{R}^6$ ke dalam K kluster secara iteratif hingga diperoleh pusat centroid (c_k) yang konvergen. Jarak Euclidean (d) dihitung dengan persamaan:

$$d(x_i, c_k) = \sqrt{\left(\sum_{j=1}^m (x_{(ij)} - c_{(kj)})^2\right)} \quad (1)$$
 Pusat kluster diperbarui berdasarkan nilai rata-rata titik data di setiap kelompok:

$$c_k = \frac{1}{|S_k|} \sum_{x_j \in S_k} x_j \quad (2)$$
 - b. Tahap II (GMM): Memanfaatkan centroid akhir K-Means (c_k) sebagai nilai inisialisasi awal ($\mu_k = c_k$) pada fungsi Distribusi Normal Multivariat:

$$N(x_i | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right) \quad (3)$$
 Probabilitas posterior (r_{ik}) tiap siswa terhadap kluster k dihitung menggunakan Teorema Bayes melalui algoritma *Expectation-Maximization* (EM):

$$r_{ik} = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_i | \mu_j, \Sigma_j)} \quad (4)$$
 Keputusan penetapan kluster akhir ditentukan berdasarkan nilai probabilitas keanggotaan tertinggi ($\arg \max_k r_{ik}$).
3. Permodelan Hibrida K-Means–GMM:

Nilai centroid akhir dari setiap kluster dianalisis berdasarkan tiga dominasi bidang kompetensi utama:

 - a. Skor STEM: $X_3 + X_4 + X_6$ (Matematika + IPA + Informatika)
 - b. Skor Komunikasi & Literasi: $X_1 + X_2$ (Bahasa Indonesia + Bahasa Inggris)
 - c. Skor Sosial: X_5 (IPS)

Segmentasi kluster dipetakan ke dalam rekomendasi program studi dan jalur karir dengan merujuk secara formal pada Keputusan Menteri Pendidikan Dasar dan Menengah Republik Indonesia Nomor 102/M/2025 tentang mata pelajaran pendukung program studi perguruan tinggi.
4. Permodelan Hibrida K-Means–GMM:

Pengujian performa pengelompokan dievaluasi pada konfigurasi $K = 2$ hingga $K = 6$ menggunakan dua metrik internal:

 - a. Davies-Bouldin Index (DBI): Mengukur rasio kesamaan antar-kluster berdasarkan dispersi internal dan jarak antar-pusat kluster. Nilai DBI yang lebih rendah mengindikasikan struktur kluster yang semakin optimal dan terpisah.
 - b. Silhouette Coefficient (S): Mengukur derajat kerapatan data di dalam klasternya sendiri dibandingkan dengan kluster terdekat. Nilai S mendekati +1 mengidentifikasi pemisahan batas kluster yang sangat baik.

2.6 Lingkungan Eksperimen Komputasi

Pengujian eksperimental diimplementasikan menggunakan perangkat keras komputer (*Multi-core Processor*, RAM 8 GB, SSD) dan perangkat lunak MATLAB R2015b memanfaatkan pustaka *Statistics and Machine Learning Toolbox*. Penanganan masukan data serta pelaporan luaran diproses melalui integrasi berkas spreadsheet (Hasil_GMM.xlsx).

3. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Dataset dan Pra-pemrosesan Data

Data penelitian bersumber dari rekapitulasi penilaian akademik siswa kelas XI dan XII SMA Negeri 9 Medan dengan populasi total $N = 566$ siswa. Variabel penelitian merepresentasikan performa kumulatif pada skala 0 hingga 100. Untuk kebutuhan verifikasi alur komputasi dan representasi sampel, Tabel 2 menampilkan sub-sampel awal sebanyak 20 siswa ($n = 20$). Pengujian komprehensif, evaluasi metrik kinerja, dan konstruksi model akhir tetap mengeksplorasi keseluruhan dataset ($N = 566$).

Tabel 2. Sampel Dataset Nilai Akademik Siswa ($n = 20$)

No	Objek	Bahasa Indonesia (X1)	Bahasa Inggris (X2)	Matematika (X3)	IPA (X4)	IPS (X5)	Informatika (X6)
1	S01	83,50	85,67	87,54	87,05	87,92	83,50
2	S02	88,00	90,00	85,42	88,28	88,42	85,00
3	S03	93,00	94,50	92,84	92,89	92,83	90,50
4	S04	86,25	93,00	88,13	62,33	43,94	80,00
5	S05	92,17	91,67	89,09	89,17	88,42	92,00
6	S06	90,50	92,25	93,88	64,17	46,75	85,00
7	S07	90,00	93,00	89,79	91,78	86,63	90,00
8	S08	88,83	89,83	94,75	89,33	87,54	91,00
9	S09	93,67	94,33	92,84	91,33	91,46	89,50
10	S10	94,17	94,17	94,88	92,61	92,13	92,00
11	S11	89,00	90,83	88,79	86,55	85,42	81,50
12	S12	87,50	89,83	86,92	88,95	83,38	85,00
13	S13	90,50	89,33	86,88	87,06	85,09	84,50
14	S14	90,50	91,17	89,09	90,22	90,29	90,50
15	S15	92,00	92,83	93,34	93,17	92,00	89,00
16	S16	93,17	93,50	91,04	91,33	90,84	89,00
17	S17	89,33	89,67	90,63	91,28	91,13	88,00
18	S18	90,00	90,67	87,42	89,28	88,54	89,50
19	S19	87,50	93,67	89,13	91,39	85,25	82,00
20	S20	93,17	94,33	94,50	92,11	90,25	92,00

Tahap pra-pemrosesan data (*data preprocessing*) dijalankan melalui tiga tingkatan utama:

- Data Cleaning*: Pengujian validitas terhadap 566 baris data menunjukkan tingkat kelengkapan 100% tanpa ditemukan nilai kosong (*missing value*) atau data duplikat. Dengan demikian, seluruh rekaman data dapat langsung dianalisis tanpa melalui prosedur imputasi.
- Feature Engineering*: Untuk mengoptimalkan dimensi data tanpa menghilangkan esensi informasi, 9 mata pelajaran dalam kurikulum direduksi menjadi 6 fitur kompetensi inti. Nilai mata pelajaran Fisika, Kimia, dan Biologi diintegrasikan menjadi fitur komposit IPA (X4). Secara paralel, nilai Ekonomi, Geografi, Sosiologi, dan Sejarah digabungkan menjadi fitur komposit IPS (X5).
- Data Normalization*: Seluruh variabel dikalibrasi ke dalam rentang terstandar guna mengeliminasi bias rentang antar-variabel dan menjamin kontribusi yang setara pada kalkulasi jarak ruang multidimensi.

3.2 Fase Partisi Awal (K-Means Clustering)

Fase pertama bertujuan untuk memperoleh inisialisasi titik pusat (*centroid*) yang konvergen dan terhindar dari krisis *local optima*. Sebagai gambaran iterasi pada konfigurasi $K = 3$, proses pengelompokan menggunakan centroid awal dari objek sampel S01, S03, dan S04 menghasilkan pola pembaharuan berikut:

- Iterasi 1: Penentuan keanggotaan berdasarkan jarak Euclidean terdekat menunjukkan distribusi awal yang belum seimbang.
- Iterasi 2: Pembaruan posisi centroid menghasilkan kondisi konvergen penuh (*zero membership shift*). Distribusi anggota klaster menetapkan Klaster 1 (C1) sebanyak 329 siswa, Klaster 2 (C2) sebanyak 166 siswa, dan Klaster 3 (C3) sebanyak 71 siswa.

Centroid konvergen dari K-Means ini selanjutnya diekstrak sebagai nilai rata-rata awal (μ_k) pada fase optimasi Gaussian Mixture Model.

3.3 Fase Optimasi Probabilistik (Gaussian Mixture Model)

Pusat klaster dari K-Means ditransisikan secara langsung sebagai estimasi parameter awal μ_k , sedangkan bobot pencampuran (mixing coefficient) diawali secara merata ($\pi_k = 1/K$). Sebagai ilustrasi kalkulasi probabilistik pada objek S01, jarak Euclidean terhadap ketiga centroid awal adalah $d(S01, C1) = 6,37$, $d(S01, C2) = 14,07$, dan $d(S01, C3) = 49,64$. Berdasarkan fungsi kepadatan peluang multivariat normal, nilai probabilitas keanggotaan posterior (π_{ik}) terbesar dimiliki oleh C1, sehingga objek S01 secara probabilistik diklasifikasikan ke dalam Klaster 1.

Namun demikian, ketika optimasi GMM dijalankan pada keseluruhan populasi ($N=566$) dengan konfigurasi $K=3$, terjadi anomali distribusi yang tidak seimbang: C1 menampung 401 siswa, C2 menampung 165 siswa, sedangkan C3 tidak memiliki anggota ($C3=0$). Fenomena klaster kosong (empty cluster) ini mengindikasikan adanya degenerasi estimasi fungsi kepadatan akibat ketidaksesuaian hipotesis $K=3$ dengan struktur dispersi data riil. Temuan ini menegaskan pentingnya eksplorasi multi-klaster ($K=2$ hingga $K=6$) guna menemukan topologi pengelompokan yang paling optimal.

3.4 Evaluasi Kualitas Klasterisasi

Evaluasi kriteria internal dilakukan secara komprehensif menggunakan dua metrik independen, yaitu *Davies-Bouldin Index* (DBI) dan *Silhouette Coefficient* (S). Hasil pengujian performa pada berbagai variasi jumlah klaster disajikan pada Tabel 3.

Tabel 3. Ringkasan Hasil Evaluasi Kualitas Klasterisasi ($N = 566$)

Jumlah Klaster (K)	Davies-Bouldin Index (DBI)	Silhouette Coefficient (S)	Keterangan Evaluasi
K = 2	0,532191	0,716529	Sub-optimal
K = 3	0,264810	0,775204	Terdegradasi (Terdapat Klaster Kosong pada GMM)
K = 4	0,161706	0,870467	Sangat Optimal (Puncak Performa)
K = 5	0,946768	0,490007	Buruk / <i>Overfitting</i>
K = 6	0,718801	0,453025	Buruk / <i>Overfitting</i>

Berdasarkan data pada Tabel 3, konfigurasi $K=4$ secara konsisten menunjukkan performa terbaik. Konfigurasi ini menghasilkan nilai DBI terendah sebesar 0,161706, yang menunjukkan tingkat kohesi internal yang tinggi serta pemisahan antar-klaster yang sangat baik. Selain itu, nilai *Silhouette Coefficient* tertinggi sebesar 0,870467 juga diperoleh pada konfigurasi ini. Nilai tersebut mendekati +1, sehingga mengonfirmasi

bahwa objek dalam setiap klaster memiliki tingkat kemiripan yang tinggi dengan kelompoknya dan memiliki pemisahan yang jelas dari klaster lainnya.

3.5 Analisis Segmentasi Potensi Akademik Siswa

Struktur pengelompokan optimal pada $K=4$ dipetakan menggunakan aturan keputusan (*rule-based*) untuk mentransformasikan nilai centroid menjadi rekomendasi fungsional. Profil karakteristik akademik, pola distribusi, rekomendasi program studi, dan prospek karier disajikan secara rinci pada Tabel 4.

Tabel 4. Ringkasan Karakteristik Klaster dan Sistem Rekomendasi ($K = 4, N = 566$)

Klaster	Jumlah Siswa	Kategori Segmentasi	Fokus Atribut Dominan	Rekomendasi Program Studi Utama	Prospek Karir Utama
C1	241 (42,6%)	Sains, Teknologi, Logika & STEM	Matematika, IPA, Informatika (Skor STEM Tinggi)	Teknik Informatika, Ilmu Komputer, Sains Data, Kedokteran, Farmasi, Teknik	<i>Software Engineer, Data Scientist, AI Specialist, Dokter, Peneliti</i>
C2	36 (6,4%)	Komunikasi, Literasi & Teknologi Digital	Bahasa Indonesia, Bahasa Inggris (Skor Komunikasi Tinggi)	Ilmu Komunikasi, Sastra Inggris/Indonesia, Hubungan Internasional, Administrasi Bisnis	<i>Public Relations, Spesialis Komunikasi, Editor, Konsultan Bisnis</i>
C3	124 (21,9%)	Sains, Teknologi, Logika & STEM	Matematika, IPA, Informatika (Skor STEM Menengah-Tinggi)	Sistem Informasi, Statistika Terapan, Pendidikan Sains, Teknik Lingkungan	<i>System Analyst, Analis Data, Tenaga Pendidik Eksakta, Spesialis K3</i>
C4	165 (29,1%)	Komunikasi, Literasi & Teknologi Digital	Bahasa Indonesia, Bahasa Inggris (Skor Komunikasi Menengah)	Pendidikan Bahasa, Administrasi Publik, Bisnis Digital, Ilmu Perpustakaan	<i>Content Writer, Staf Administrasi, Pengembang Konten Digital, Jurnalis</i>

Hasil analisis pada Tabel 4 menunjukkan bahwa dari total 566 siswa, 64,5% (365 siswa) terdistribusi pada ranah kompetensi *Science, Technology, Engineering, and Mathematics* (STEM) yang terbagi atas Klaster C1 dan C3. Di sisi lain, 35,5% (201 siswa) terkelompokkan dalam dimensi Komunikasi, Literasi, dan Teknologi Digital melalui Klaster C2 dan C4. Dominasi kelompok STEM di SMA Negeri 9 Medan mencerminkan konsentrasi pola prestasi akademik siswa yang kuat pada penalaran analitis dan komputasi.

3.6 Penyelarasan Kebijakan Pendidikan Nasional

Kebaruan (*novelty*) utama dari penelitian ini terletak pada pengikatan hasil analisis *Educational Data Mining* secara langsung dengan kebijakan regulasi nasional

teraktual, yaitu **Keputusan Menteri Pendidikan Dasar dan Menengah Republik Indonesia Nomor 102/M/2025** tentang mata pelajaran pendukung program studi perguruan tinggi. Pendekatan ini berhasil mentransformasi layanan Bimbingan Konseling (BK) di sekolah:

- a. Eliminasi Subjektivitas: Proses pengarahan studi lanjut yang selama ini cenderung intuitif kini digantikan oleh pemetaan obyektif berbasis data kuantitatif (*evidence-based counseling*).
- b. Optimalisasi Peluang SNBP: Penyelarasan indikator dominan klaster dengan mata pelajaran pendukung resmi kementerian secara signifikan meningkatkan rasio kesesuaian (*matching rate*) antara potensi riil siswa dan syarat masukan Seleksi Nasional Berdasarkan Prestasi (SNBP).

3.7 Keunggulan Metodologis Penelitian

Dibandingkan dengan studi klasterisasi pendidikan sebelumnya yang umumnya hanya mengandalkan algoritma partisi tunggal seperti K-Means (Adek Maulidya et al., 2024; Rahayu et al., 2024), penerapan model hibrida K-Means–GMM pada penelitian ini terbukti memiliki dua keunggulan metodologis utama:

- a. Inisialisasi Stabil: Eksploitasi *centroid* K-Means berhasil memitigasi kelemahan klasik GMM yang sangat sensitif terhadap inisialisasi acak.
- b. Fleksibilitas Probabilistik: Penggunaan teori Bayes dan probabilitas keanggotaan posterior pada GMM efektif menoleransi ketidakpastian posisi data di wilayah perbatasan antarklaster, sehingga menghasilkan pemisahan kelompok yang sangat rapat ($S = 0,870$).

4. KESIMPULAN

4.1 Kesimpulan

Penerapan model hibrida *K-Means–Gaussian Mixture Model* (GMM) terbukti efektif dalam melakukan segmentasi prestasi akademik siswa berbasis *Educational Data Mining*. Pengujian multi-klaster terhadap 566 data siswa SMA Negeri 9 Medan menunjukkan bahwa konfigurasi empat klaster ($K=4$) merupakan topologi yang paling optimal. Hal ini ditunjukkan oleh nilai *Davies-Bouldin Index* (DBI) terendah sebesar 0,161706 dan *Silhouette Coefficient* tertinggi sebesar 0,870467. Struktur pengelompokan tersebut memetakan 64,5% siswa ke dalam domain STEM (Klaster C1 dan C3), sedangkan 35,5% siswa berada dalam domain Komunikasi dan Literasi Digital (Klaster C2 dan C4). Integrasi probabilitas posterior GMM dengan aturan keputusan berbasis Keputusan Menteri Pendidikan Dasar dan Menengah Nomor 102/M/2025 menghasilkan rekomendasi program studi dan karier yang memiliki landasan regulatif, sekaligus mentransformasikan layanan bimbingan dan konseling konvensional menjadi bimbingan berbasis bukti (*evidence-based counseling*) yang lebih obyektif dan terukur.

4.2 Saran

Penelitian selanjutnya disarankan untuk mengintegrasikan variabel non-akademik seperti hasil tes psikometri, asesmen minat-bakat, dan prestasi ekstrakurikuler untuk meningkatkan kovergensitas sistem personalisasi. Selain itu, pengujian pada dataset longitudinal lintas sekolah diperlukan guna mengevaluasi generalisasi model hibrida. Secara teknis, pengembangan arsitektur sistem disarankan bertransisi dari lingkungan *desktop* ke platform berbasis *web* yang terintegrasi secara *real-time* dengan basis data perguruan tinggi agar dapat diakses secara lebih fleksibel oleh sekolah dan siswa.

5. DAFTAR PUSTAKA

- Adek Maulidya, K., Sitorus, Z., Siahaan, A.P.U. and Iqbal, M., 2024. Analysis Of Increasing Student Service Satisfaction Using K-Means Clustering Algorithm and Gaussian Mixture Models (GMM). *International Journal Of Computer Sciences and Mathematics Engineering*, 3(1), pp.29–35. <https://doi.org/10.61306/ijecom.v3i1.62>
- Dacwanda, D.O. and Nataliani, Y., 2021. Implementasi K-Means Clustering untuk Analisis Nilai Akademik Siswa Berdasarkan Nilai Pengetahuan dan Keterampilan. *Jurnal Teknologi Informasi*, 18(2), pp.125–138.
- Farell, G., Wahyudi, R., Novid, I., Faiza, D. and Anori, S., 2024. Analysis of Instruments for Implementing Intelligent Job Matching Models. *TEM Journal*, 13(3), pp.2186–2194. <https://doi.org/10.18421/TEM133-46>
- Hidayat, D.R., Marfu'i, L.N.R. and Nufus, N.P., 2025a. Seberapa perlukah kebutuhan layanan bimbingan karier ai-driven: studi pada guru bimbingan dan konseling di Indonesia. *Jurnal Konseling Dan Pendidikan*, 13(3), pp.387–398.
- Hutajulu, K. and Wulandhari, L.A., 2024. Prediction of student performance at polytechnic using machine learning approach. *International Journal of Electrical and Computer Engineering*, 14(5), pp.5356–5365. <https://doi.org/10.11591/ijece.v14i5.pp5356-5365>
- Kepmendikdasmen, 2025. *Keputusan Menteri Pendidikan Dasar dan Menengah Republik Indonesia Nomor 102/M/2025 tentang Mata Pelajaran Pendukung Program Studi Dalam Seleksi Nasional Berdasarkan Prestasi*. Jakarta: Kementerian Pendidikan Dasar dan Menengah.
- Nimy, E. and Mosia, M., 2023. Web-based Clustering Application for Determining and Understanding Student Engagement Levels in Virtual Learning Environments. *E-Journal of Humanities, Arts and Social Sciences*, 4(12), pp.4–19. <https://doi.org/10.38159/ehass.20234122>
- Niu, J. and Reeping, D., 2024. Humanizing Data-Driven Methods in Engineering Education Research: A Systematic Literature Review of Four Journals From 2011 to 2021. *Studies in Engineering Education*, 5(2), pp.150–174. <https://doi.org/10.21061/see.159>
- Pu, S., Yan, Y. and Zhang, B., 2024. Predicting Students' Future Success: Harnessing Clickstream Data with Wide & Deep Item Response Theory. *Journal of Educational Data Mining*, 16(2), pp.1–31. <https://doi.org/10.5281/zenodo.13627151>
- Rahayu, N.D., Anshor, A.H. and Afriantoro, I., 2024. Penerapan Data Mining untuk Pemetaan Siswa Berprestasi menggunakan Metode Clustering K-Means. *JUKI: Jurnal Komputer Dan Informatika*, 6, pp.71–83.
- Saravanakumar, R., TamilSelvi, T., Pandey, D., Pandey, B.K., Mahajan, D.A. and Lelisho, M.E., 2024. Big data processing using hybrid Gaussian Mixture Model with salp swarm algorithm. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-01015-3>