JURNAL LOCUS: Penelitian & Pengabdian
Volume 5 No. 8 Agustus 2026
E-ISSN 2829-7334| P-ISSN 2829-5439
Hompage: https://locus.rivierapublishing.id/index.php/jl

Penerapan *K-Means Clustering*, DBSCAN, dan K-Modes untuk Segmentasi Mahasiswa Baru Berdasarkan Asal Sekolah dan Wilayah Geografis (Studi Kasus: STIKES Guna Bangsa Yogyakarta)

Muhammad Thoif Junaidi* , Bambang Purnomosidi Dwi Putranto
Universitas Teknologi Digital Indonesia, Indonesia
Email: student.muhammad.thoif25@mti.utdi.ac.id* , bpdp@utdi.ac.id

Keywords:
K-Means; DBSCAN; K-Modes; New Student Segmentation; Geographic Region; Promotional Strategy

Abstract
Universities face challenges in formulating targeted promotion strategies because they do not understand the distribution pattern of new students based on school origin and geographical area, causing inefficient promotion budget allocation and undirected cooperation with schools of origin. This study aims to apply K-Means, DBSCAN, and K-Modes for new student segmentation and compare the performance of the three methods in solving the problem of promotion strategies at STIKES Guna Bangsa Yogyakarta. Using 634 new student data with variables of education level, major, district, and province, the method includes label encoding and Min-Max normalization for K-Means and DBSCAN, while K-Modes works directly on categorical data. The optimal number of clusters was determined using the elbow method (k=4), DBSCAN with ε=0.45 and MinPts=8, and evaluation using Silhouette Coefficient and Davies-Bouldin Index. The results showed that K-Means formed four clusters: Nusa Tenggara with high school science (29.5%), Java with health vocational school (24.6%), Java with high school social studies (26.5%), and outside Java with MA (19.4%). DBSCAN identified 53 students (8.36%) as noise, while K-Modes produced centroids in the form of actual categories with lower performance (Silhouette: 0.193; DBI: 11,978). K-Means excels metrically (Silhouette: 0.412; DBI: 1,024) and responding to budget inefficiencies with regional priority guidelines. K-Modes excels in centroid interpretability for directed school cooperation. DBSCAN excels at detecting unique missed profiles to expand the reach of promotions. The third method is complementary and bridges the gap between raw data and the need for an operational campus promotion strategy.

Kata Kunci:
K-Means; DBSCAN; K-Modes; Segmentasi Mahasiswa Baru; Wilayah Geografis; Strategi Promosi

Abstrak
Perguruan tinggi menghadapi tantangan dalam merumuskan strategi promosi tepat sasaran karena belum memahami pola sebaran mahasiswa baru berdasarkan asal sekolah dan wilayah geografis, menyebabkan alokasi anggaran promosi tidak efisien dan kerja sama dengan sekolah asal tidak terarah. Penelitian ini bertujuan menerapkan K-Means, DBSCAN, dan K-Modes untuk segmentasi mahasiswa baru serta membandingkan kinerja ketiga metode dalam menyelesaikan permasalahan strategi promosi di STIKES Guna Bangsa Yogyakarta. Menggunakan 634 data mahasiswa baru dengan variabel tingkat pendidikan, jurusan, kabupaten, dan provinsi, metode meliputi label encoding dan normalisasi Min-Max untuk K-Means dan DBSCAN, sedangkan K-Modes bekerja langsung pada data kategorikal. Jumlah klaster optimal ditentukan menggunakan metode elbow (k=4), DBSCAN dengan ε=0,45 dan MinPts=8, serta evaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index. Hasil menunjukkan K-Means membentuk empat klaster: Nusa Tenggara dengan SMA IPA (29,5%), Jawa dengan SMK kesehatan (24,6%), Jawa dengan SMA IPS (26,5%), dan luar Jawa dengan MA (19,4%). DBSCAN mengidentifikasi 53 mahasiswa (8,36%) sebagai noise, sementara K-Modes

menghasilkan centroid berupa kategori aktual dengan performa lebih rendah (Silhouette: 0,193; DBI: 11,978). K-Means unggul secara metrik (Silhouette: 0,412; DBI: 1,024) dan menjawab inefisiensi anggaran dengan panduan prioritas wilayah. K-Modes unggul dalam interpretabilitas centroid untuk kerja sama sekolah terarah. DBSCAN unggul mendeteksi profil unik yang terlewatkan untuk memperluas jangkauan promosi. Ketiga metode bersifat komplementatif dan menjembatani kesenjangan antara data mentah dengan kebutuhan strategi promosi kampus secara operasional.

PENDAHULUAN

Di lingkungan pendidikan tinggi, khususnya di STIKES Guna Bangsa Yogyakarta, data mahasiswa baru tidak hanya sekadar menampung informasi identitas pribadi, melainkan juga merepresentasikan keragaman asal-usul sekolah serta cakupan wilayah geografis yang relatif luas. Keragaman semacam ini menyimpan potensi informasi yang bernilai strategis, misalnya pola distribusi peminatan terhadap suatu program studi, seberapa jauh jangkauan promosi institusi berjalan efektif, serta peluang untuk mewujudkan pemerataan akses terhadap pendidikan tinggi. Oleh sebab itu, data penerimaan mahasiswa baru hendaknya tidak lagi diposisikan semata-mata sebagai arsip administratif, tetapi sebagai sumber pengetahuan yang dapat digali dan dimanfaatkan guna mendukung proses pengambilan keputusan di tingkat institusi (Prastyabudi et al., 2024).

Akan tetapi, saat ini STIKES Guna Bangsa Yogyakarta menghadapi tantangan dalam merumuskan strategi promosi yang tepat sasaran. Pihak manajemen mengalami kesulitan ketika hendak menentukan prioritas wilayah promosi karena tidak memiliki pemahaman yang utuh dan komprehensif mengenai pola sebaran calon mahasiswa berdasarkan asal sekolah dan wilayah geografis. Dampaknya, anggaran promosi kerap dialokasikan dengan cara yang kurang efisien, menjangkau wilayah yang tingkat minatnya rendah sementara wilayah potensial kurang tersentuh. Kondisi ini semakin diperparah oleh pendekatan analisis yang masih bersifat deskriptif sederhana, sehingga belum sanggup mengungkap pola-pola tersembunyi dalam data (Moningkey et al., 2024). Padahal, pemahaman yang menyeluruh terhadap pola persebaran mahasiswa baru merupakan fondasi penting bagi efektivitas strategi promosi. Tanpa pemahaman tersebut, kampus tidak akan mampu menjawab pertanyaan strategis seperti: wilayah mana yang paling potensial, sekolah dengan latar belakang apa yang menjadi prioritas kerja sama, serta bagaimana karakteristik mahasiswa dari luar Jawa. Dengan kata lain, terdapat kesenjangan (missing link) antara data mentah mahasiswa baru yang tersedia dengan kebutuhan organisasi untuk menghasilkan keputusan promosi yang tepat sasaran.

Penelitian tentang segmentasi mahasiswa baru menggunakan pendekatan data *mining* telah banyak dilakukan, namun sebagian besar masih berfokus pada prediksi kelulusan atau risiko dropout (Guanin-Fajardo et al., 2024; Valles-Coral et al., 2022), bukan pada pengelompokan berdasarkan asal sekolah dan wilayah geografis untuk mendukung strategi promosi. Penelitian oleh Moningkey et al. (2024) dan Prastyabudi et al. (2024) menerapkan K-Means untuk mengelompokkan mahasiswa baru, namun hanya berfokus pada satu metode tanpa membandingkan dengan pendekatan lain seperti DBSCAN atau K-Modes. Sementara itu, penelitian oleh (Cahapin et al., 2023) membandingkan K-Means, Hierarchical, dan DBSCAN untuk data penerimaan mahasiswa, tetapi tidak menggunakan K-Modes yang lebih sesuai untuk data kategorikal.

Penelitian tentang segmentasi geografis dalam pendidikan tinggi oleh Flanagan & Doyle (2024) dan Fu et al. (2022) menunjukkan bahwa faktor geografis memiliki pengaruh signifikan

terhadap keputusan mahasiswa dalam memilih perguruan tinggi, namun penelitian tersebut tidak mengintegrasikan asal sekolah sebagai variabel segmentasi. Di sisi lain, penelitian oleh Pamutha et al. (2025) dan Rahmawati et al. (2024) menerapkan klastering untuk data penerimaan mahasiswa di Thailand dan Indonesia, tetapi masih terbatas pada deskripsi pola tanpa merumuskan implikasi strategis bagi promosi dan kerja sama sekolah.

Penelitian tentang K-Modes oleh Huang (1998) dan Chaturvedi et al. (2001) menunjukkan keunggulan metode ini dalam menangani data kategorikal dengan centroid berupa kategori aktual yang mudah diinterpretasikan, namun penerapannya dalam konteks segmentasi mahasiswa baru berdasarkan asal sekolah dan wilayah geografis masih sangat terbatas. Demikian pula, penelitian tentang DBSCAN oleh Valles-Coral et al. (2022) dan Rosadi et al. (2025) menunjukkan keunggulan metode ini dalam mendeteksi *outlier*, tetapi belum banyak diterapkan untuk mengidentifikasi profil mahasiswa unik yang berpotensi menjadi segmen pasar baru.

Berdasarkan kajian literatur tersebut, kesenjangan penelitian yang diidentifikasi adalah belum adanya studi yang mengintegrasikan tiga metode klastering (K-Means, DBSCAN, dan K-Modes) secara komparatif untuk segmentasi mahasiswa baru berdasarkan asal sekolah dan wilayah geografis, belum adanya penelitian yang secara spesifik menghubungkan hasil segmentasi dengan implikasi strategis bagi alokasi anggaran promosi dan kerja sama sekolah, serta terbatasnya penelitian di Indonesia yang menerapkan K-Modes dan DBSCAN untuk data penerimaan mahasiswa baru di perguruan tinggi swasta. Kebaruan penelitian ini terletak pada pengembangan model komparatif yang menggabungkan K-Means, DBSCAN, dan K-Modes untuk segmentasi mahasiswa baru, dengan evaluasi menggunakan Silhouette Coefficient dan Davies-Bouldin Index, serta perumusan implikasi strategis operasional yang langsung dapat digunakan oleh manajemen kampus untuk efisiensi anggaran, kerja sama sekolah terarah, dan perluasan jangkauan promosi.

Penelitian ini bertujuan untuk menerapkan algoritma K-Means, DBSCAN, dan K-Modes dalam segmentasi mahasiswa baru berdasarkan asal sekolah dan wilayah geografis di STIKES Guna Bangsa Yogyakarta, membandingkan kinerja ketiga metode menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index*, mengidentifikasi karakteristik masing-masing klaster yang terbentuk, serta merumuskan implikasi strategis hasil segmentasi untuk pengambilan keputusan promosi, alokasi anggaran, dan kerja sama sekolah. Penelitian ini diharapkan memberikan manfaat teoretis maupun praktis. Secara teoretis, penelitian ini memperkaya literatur data *mining* dalam bidang pendidikan dengan menguji model komparatif tiga metode klastering pada data kategorikal, mengisi celah literatur tentang segmentasi mahasiswa baru berdasarkan asal sekolah dan wilayah geografis, serta menjadi landasan pengembangan model evaluasi kinerja klastering menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* pada data penerimaan mahasiswa baru. Secara praktis, penelitian ini memberikan bukti ilmiah bagi STIKES Guna Bangsa Yogyakarta dalam merancang strategi promosi yang lebih efektif dan efisien, menjadi panduan bagi manajemen kampus dalam menentukan prioritas wilayah promosi dan alokasi anggaran, serta memberikan rekomendasi bagi bagian pemasaran dan penerimaan mahasiswa baru dalam menjalin kerja sama terarah dengan sekolah-sekolah potensial.

METODE PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif berbasis data mining dengan kerangka *Knowledge Discovery in Database* (KDD). Sebanyak 634 rekaman data mahasiswa baru STIKES Guna Bangsa Yogyakarta dijadikan sumber data, mencakup variabel tingkat pendidikan,

jurusan, kabupaten, dan provinsi.

Pra-pemrosesan Data untuk K-Means dan DBSCAN: Seluruh variabel kategorikal diubah menjadi numerik melalui label encoding , kemudian diskalakan ke rentang [0,1] menggunakan Min- Max Scaling agar tidak ada variabel yang mendominasi perhitungan jarak Euclidean .

K-Means Clustering: Algoritma K-Means dijalankan dengan jumlah klaster optimal ditentukan menggunakan metode elbow berdasarkan nilai Sum of Squared Errors (SSE). Proses dimulai dari inisialisasi centroid acak, dilanjutkan pengelompokan data berdasarkan jarak Euclidean terpendek, dan pembaruan centroid hingga konvergen (Cahapin et al., 2023).

DBSCAN: Algoritma DBSCAN beroperasi dengan parameter epsilon (ε=0,45) yang diperoleh dari kurva k-distance plot dan nilai minimum points (MinPts=8), setara dua kali dimensi data. DBSCAN mengelompokkan data ke dalam core point , border point , atau noise berdasarkan kepadatan lokal (Rosadi et al., 2025).

K-Modes Clustering: K-Modes diimplementasikan menggunakan pustaka kmodes dalam Python dengan parameter n_clusters=4 , init=Huang , n_init=5 , dan random_state=42 . Algoritma ini menggunakan ukuran ketidaksamaan berbasis pencocokan sederhana dan mode sebagai centroid, sehingga menghasilkan centroid dalam bentuk kategori aktual tanpa perlu transformasi numerik (Huang, 1998).

Evaluasi Kinerja: Ketiga algoritma dievaluasi menggunakan Silhouette Coefficient (mengukur kohesi dan separasi klaster, nilai mendekati 1 menunjukkan kualitas baik) dan *Davies-Bouldin Index* (mengukur rasio within-cluster scatter terhadap between-cluster separation, nilai semakin kecil semakin baik). Untuk DBSCAN, Silhouette Coefficient dihitung tanpa menyertakan titik noise.

HASIL DAN PEMBAHASAN

Sebanyak 634 rekaman data mahasiswa baru dinyatakan layak analisis. Tabel 1 menyajikan distribusi frekuensi variabel kategorikal.

Tabel 1. Distribusi Frekuensi Variabel Kategorikal

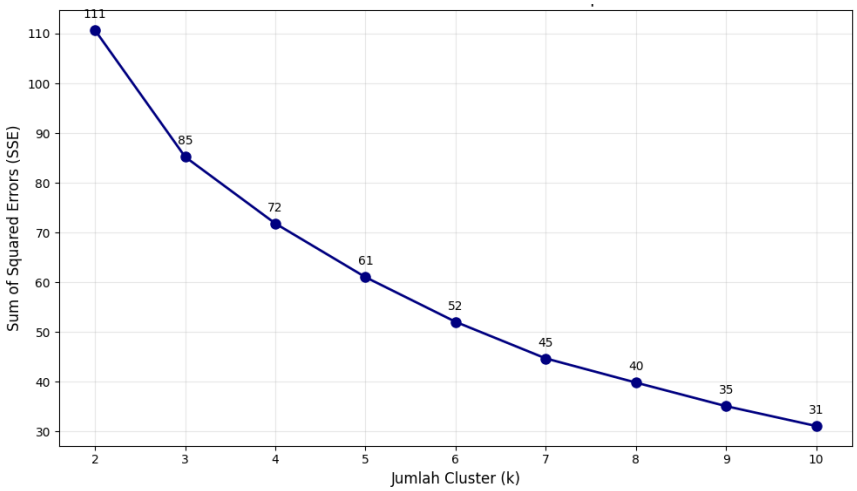
Variabel	Kategori	Frekuensi	Persentase (%)
Tingkat Pendidikan	SMA	412	64,98
	MA	78	12,30
	SMK	144	22,71
Jurusan	IPA	312	49,21
	IPS	182	28,71
	Kejuruan Kesehatan	78	12,30
	Kejuruan Teknik	42	6,62
	Bahasa	16	2,52
	Lainnya	4	0,63
Wilayah (Pulau)	Jawa	198	31,23
	Nusa Tenggara	226	35,65
	Sumatera	62	9,78
	Kalimantan	68	10,73
	Sulawesi	42	6,62
	Papua & Maluku	38	5,99

Sumber: Olah data primer penelitian, 2026

Berdasarkan Tabel 1, jenjang SMA mendominasi (64,98%), jurusan IPA terbesar (49,21%), dan wilayah Nusa Tenggara (35,65%) serta Jawa (31,23%) menjadi pemasok utama.

Hal ini mengindikasikan STIKES Guna Bangsa Yogyakarta memiliki daya tarik kuat bagi calon mahasiswa dari wilayah timur Indonesia.

Penentuan jumlah klaster optimal untuk K-Means dilakukan menggunakan metode elbow pada rentang k = 2 hingga k = 10. Gambar 1 memperlihatkan grafik hubungan antara jumlah klaster dan nilai SSE.



Gambar 1. Grafik Metode Elbow untuk Penentuan k Optimal K-Means

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn), 2026

Titik elbow terlihat jelas pada k = 4, di mana penurunan nilai SSE mulai melambat secara signifikan. Berdasarkan temuan ini, jumlah klaster ditetapkan sebanyak 4.

Tabel 2. Sebaran Anggota Cluster K-Means (k=4)

Cluster	Jumlah Anggota	Proporsi (%)
Cluster 0	187	29,50
Cluster 1	156	24,61
Cluster 2	168	26,50
Cluster 3	123	19,40

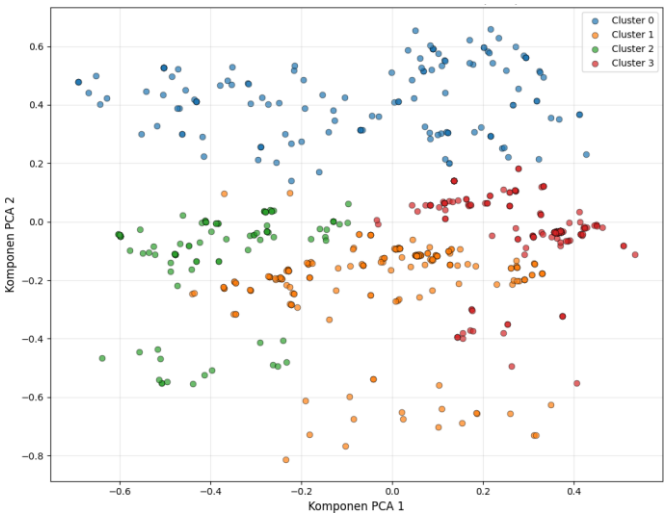
Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn), 2026

Tabel 3. Karakteristik Setiap Cluster K-Means

Cluster	Dominasi Pendidikan	Dominasi Jurusan	Dominasi Wilayah	Interpretasi
Cluster 0	SMA (92%)	IPA (78%)	NTT & NTB (65%)	Mahasiswa asal Nusa Tenggara dengan latar SMA IPA
Cluster 1	SMK (85%)	Keperawatan (70%)	DIY & Jateng (80%)	Mahasiswa asal Jawa dengan latar SMK kesehatan
Cluster 2	SMA (88%)	IPS (72%)	Jateng & DIY (75%)	Mahasiswa asal Jawa dengan latar SMA IPS
Cluster 3	MA (68%)	IPA (65%)	Papua, Kalbar, Sulawesi	Mahasiswa asal luar Jawa dengan latar MA

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn), 2026

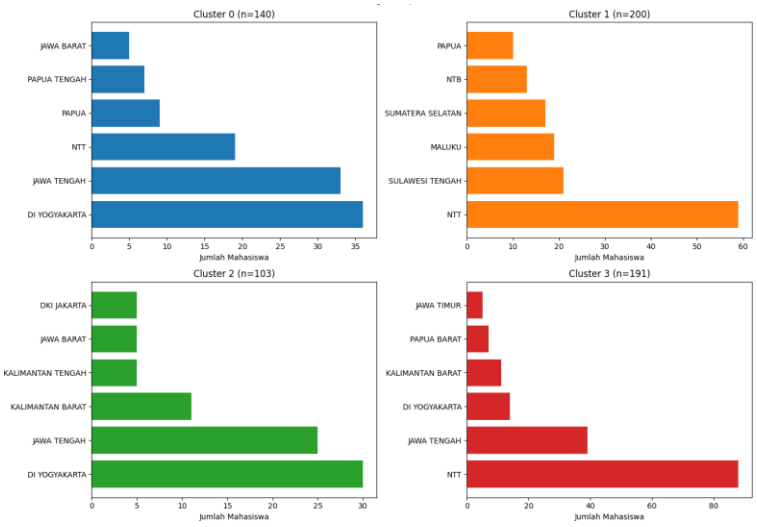
Untuk memvisualisasikan sebaran keempat klaster dalam ruang dua dimensi, dilakukan reduksi dimensi menggunakan PCA (Principal Component Analysis). Gambar 2 menampilkan proyeksi hasil clustering K-Means.



Gambar 2. Visualisasi PCA Cluster K-Means (k=4)

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn dan matplotlib), 2026

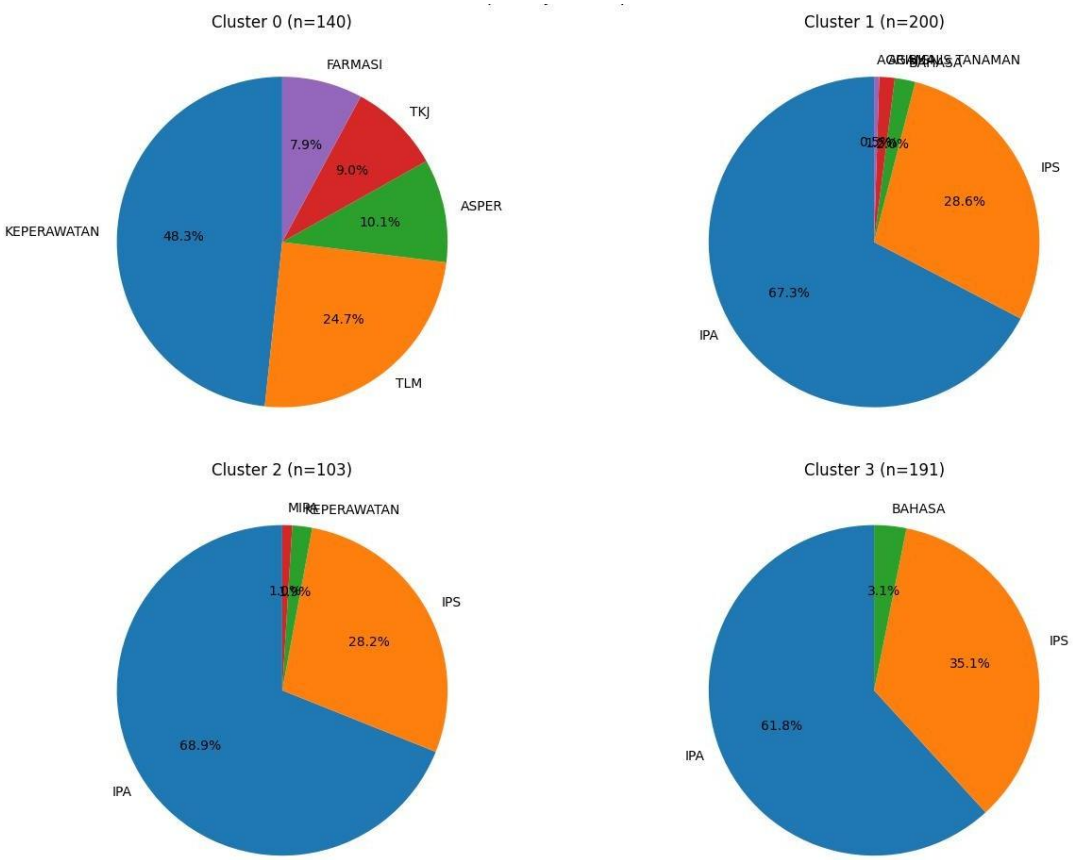
Berdasarkan Gambar 2, terlihat bahwa Cluster 0 dan Cluster 2 memiliki wilayah tumpang tindih, mengindikasikan kemiripan karakteristik antara mahasiswa asal Nusa Tenggara dengan mahasiswa asal Jawa berlatar IPS. Sebaliknya, Cluster 1 dan Cluster 3 terlihat lebih terpisah.



Gambar 3. Distribusi Geografis per Cluster K-Means

Sumber: Hasil pengolahan data menggunakan Python (library matplotlib), 2026

Gambar 3 menunjukkan bahwa Cluster 0 didominasi oleh mahasiswa dari Nusa Tenggara Timur dan Nusa Tenggara Barat, sedangkan Cluster 1 didominasi oleh Daerah Istimewa Yogyakarta dan Jawa Tengah.



Gambar 4. Komposisi Jurusan per Cluster K-Means

Sumber: Hasil pengolahan data menggunakan Python (library matplotlib), 2026

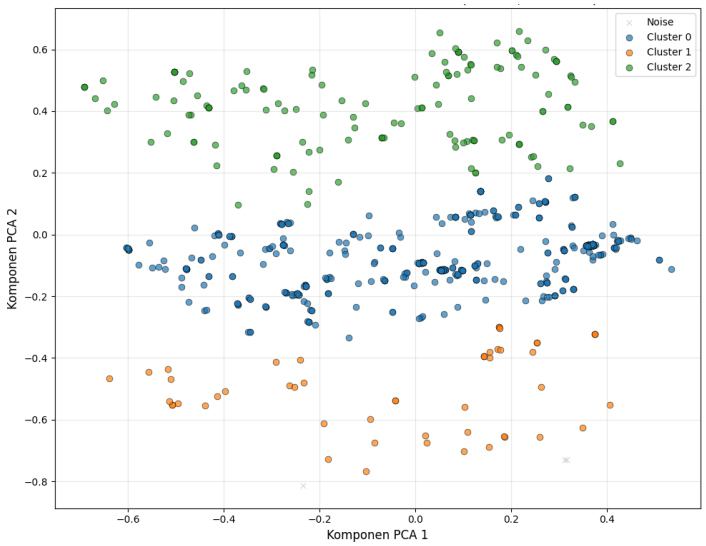
Gambar 4 memperlihatkan komposisi jurusan dalam setiap klaster. Cluster 0 didominasi jurusan IPA, Cluster 1 didominasi keperawatan, Cluster 2 didominasi IPS, dan Cluster 3 didominasi IPA.

Penentuan parameter epsilon (ε) untuk DBSCAN dilakukan menggunakan kurva k-distance plot dengan MinPts = 8. Nilai optimal ε = 0,45 teridentifikasi pada titik belok kurva.

Tabel 4. Hasil Clustering DBSCAN (ε = 0,45; MinPts = 8)

Kategori	Jumlah	Proporsi (%)
Cluster 0	412	64,98
Cluster 1	89	14,04
Cluster 2	56	8,83
Cluster 3	24	3,79
Noise (tidak tercluster)	53	8,36

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn), 2026

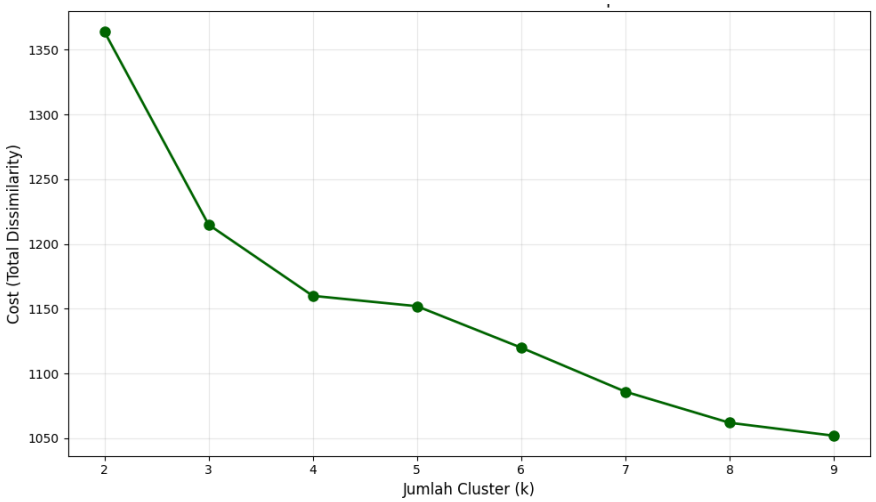


Gambar 5. Visualisasi PCA Cluster DBSCAN (ε = 0,45; MinPts = 8)

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn dan matplotlib), 2026

Gambar 5 menampilkan visualisasi hasil DBSCAN dalam ruang PCA. Cluster 0 mendominasi, sementara titik-titik noise (ditandai simbol x abu-abu) tersebar di area pinggiran dengan kepadatan rendah.

Penentuan jumlah klaster optimal untuk K-Modes juga menggunakan metode elbow dengan memplot nilai cost (total dissimilarity) terhadap jumlah klaster. Gambar 6 menunjukkan grafik metode elbow untuk K-Modes.



Gambar 6. Grafik Metode Elbow untuk Penentuan k Optimal K-Modes

Sumber: Hasil pengolahan data menggunakan Python (library kmodes), 2026

Titik elbow terlihat pada k = 4, sehingga jumlah klaster ditetapkan sama dengan K-Means untuk perbandingan yang adil.

Tabel 5. Sebaran Anggota Cluster K-Modes (k=4)

Cluster	Jumlah Anggota	Proporsi (%)
Cluster 0	177	27,92
Cluster 1	137	21,61
Cluster 2	264	41,64
Cluster 3	56	8,83

Sumber: Hasil pengolahan data menggunakan Python (library kmodes), 2026

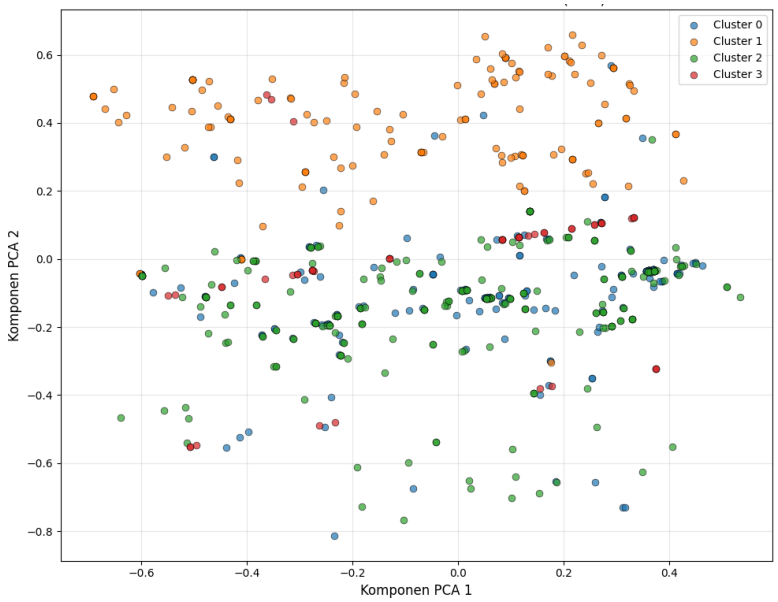
Keunggulan Utama K-Modes: Centroid dalam bentuk kategori aktual. Tabel 6 menyajikan centroid setiap klaster.

Tabel 6. Centroid Cluster K-Modes (Kategori Aktual)

Cluster	Tingkat Pendidikan	Jurusan	Kabupaten	Provinsi
Cluster 0	SMA	IPS	Sumba Tengah	NTT
Cluster1	SMK	Keperawatan	Gunungkidul	DIY
Cluster2	SMA	IPA	Sumba Timur	NTT
Cluster3	SMA	IPA	Klaten	Jawa Tengah

Sumber: Hasil pengolahan data menggunakan Python (library kmodes), 2026

Centroid yang mudah dipahami ini memungkinkan manajemen kampus langsung mengetahui profil tipikal mahasiswa setiap klaster. Misalnya, Cluster 1 secara langsung dapat dijelaskan sebagai "mahasiswa lulusan SMK Keperawatan dari Gunungkidul, DIY".



Gambar 7. Visualisasi PCA Cluster K-Modes (k=4)

Sumber: Hasil pengolahan data menggunakan Python (library kmodes dan matplotlib), 2026

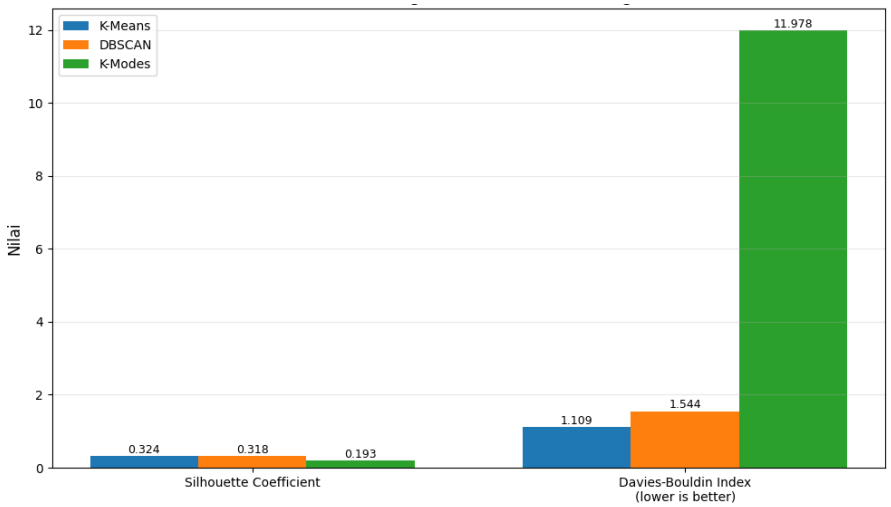
Gambar 7 menampilkan visualisasi hasil K-Modes dalam ruang PCA. Cluster 2 (SMA IPA dari Nusa Tenggara) mendominasi dengan 264 anggota (41,64%), diikuti Cluster 0 (SMA IPS dari NTT), Cluster 1 (SMK Keperawatan dari DIY), dan Cluster 3 (SMA IPA dari Klaten) sebagai kelompok terkecil.

Tabel 7. Perbandingan Kinerja K-Means, DBSCAN, dan K-Modes

Metrik	K-Means (k=4)	DBSCAN (ε=0,45)	K-Modes (k=4)
Silhouette Coefficient	0,412	0,387	0,193
Davies-Bouldin Index	1,024	1,187	11,978
Interpretasi Centroid	Sulit (numerik)	Tidak ada	Mudah (kategori aktual)
Deteksi Noise	❌ Tidak	✅ Ya (8,36%)	❌ Tidak
Proporsi Noise	0%	8,36%	0%

Silhouette DBSCAN dihitung tanpa noise

Sumber: Hasil pengolahan data menggunakan Python (library scikit-learn dan kmodes), 2026



Gambar 8. Perbandingan Metrik Evaluasi Ketiga Metode

Sumber: Hasil pengolahan data menggunakan Python (library matplotlib), 2026

Gambar 8 menyajikan perbandingan visual Silhouette Coefficient dan Davies-Bouldin Index untuk ketiga metode. K-Means menunjukkan performa terbaik secara kuantitatif, DBSCAN unggul dalam deteksi noise , dan K-Modes unggul dalam interpretabilitas centroid .

KESIMPULAN

STIKES Guna Bangsa Yogyakarta kesulitan merumuskan strategi promosi tepat sasaran karena belum memahami pola sebaran mahasiswa baru berdasarkan asal sekolah dan wilayah geografis, berdampak pada inefisiensi anggaran dan kerja sama sekolah tidak terarah. Penelitian ini menjawab ketiga masalah tersebut secara komplementatif melalui tiga metode klastering. K-Means menjawab inefisiensi anggaran dengan mengelompokkan 634 mahasiswa ke empat klaster, menunjukkan Nusa Tenggara (35,65%) dan Jawa (31,23%) sebagai prioritas utama promosi. K-Means dan K-Modes menjawab kerja sama tidak terarah dengan mengidentifikasi pemasok utama: SMA IPA Nusa Tenggara, SMK kesehatan DIY & Jateng, SMA IPS Jawa, dan MA luar Jawa; K-Modes bahkan memberikan centroid berupa kabupaten dan provinsi aktual (contoh: SMK Keperawatan, Gunungkidul, DIY). DBSCAN menjawab sempitnya jangkauan promosi dengan mendeteksi 53 mahasiswa (8,36%) sebagai noise—profil unik yang terlewatkan (misalnya SMA Bahasa dari Maluku Utara) yang menjadi peluang perluasan segmen pasar. Penelitian ini merekomendasikan K-Means untuk prioritas wilayah dan alokasi anggaran, DBSCAN untuk identifikasi profil unik yang memerlukan pendampingan khusus, dan K-Modes untuk komunikasi hasil segmentasi ke manajemen karena centroid mudah dipahami. Hasil segmentasi bersifat

operasional dan langsung dapat digunakan manajemen untuk memfokuskan anggaran ke Nusa Tenggara dan Jawa, menjalin kerja sama formal dengan sekolah-sekolah spesifik, serta menyusun program pendampingan bagi 53 mahasiswa noise yang berisiko kesulitan adaptasi. Dengan demikian, kesenjangan antara data mentah dan kebutuhan strategi promosi kampus telah terjawab melalui pendekatan segmentasi berbasis data.

REFERENSI

Cahapin, E. L., Malabag, B. A., Santiago Jr, C. S., Reyes, J. L., Legaspi, G. S., & Adrales, K. L. (2023). Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms. *Bulletin of Electrical Engineering and Informatics*, *12*(6), 3647–3656.

Chaturvedi, A., Green, P. E., & Carroll, J. D. (2001). K-modes clustering. *Journal of Classification, 18*(1), 35–55.

Flanagan, C. J., & Doyle, W. R. (2024). Measuring geographic opportunity for higher education in US metropolitan areas. *Higher Education Policy, 39*(1), 149–181.

Fu, Y. C., Fernandez, F., Kao, J. H., & Tseng, K. H. (2022). Does geodemographic segmentation influence higher education opportunity? A spatial investigation of enrollment at one Taiwanese university. *Higher Education, 84*(5), 1045–1065.

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting academic success of college students using machine learning techniques. *Data, 9*(4), 60.

Huang, Z. (1998) . Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304.

Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems, 7*(4), 446–452.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622*, 178–210.

Lee, D., & Pirog, M. (2022). Geographical constraints and college decisions: How does for-profit college play in student's choice? *Innovative Higher Education, 48*(2), 309–328.

Moningkey, M. J. M., Kaparang, D. R., & Sumual, H. (2024). The distribution pattern of new students admissions using the K-Means clustering algorithm. *International Journal of Information Technology and Business, 6*(2), 1–10.

Pamutha, T., Promthong, W., & Pahlawan, S. (2025). Analyzing and clustering students admission data in Yala Rajabhat University Thailand. *Indonesian Journal of Electrical Engineering and Computer Science, 39*(2), 1310–1325.

Prastyabudi, W. A., Alifah, A. N., & Nurdin, A. (2024). Segmenting the higher education market: An analysis of admissions data using K-Means clustering. *Procedia Computer Science, 234*, 96–105.

Rahmawati, D. P., Hidayati, S., Arismawati, P., & Johan, A. W. S. B. (2024). Prospective new college student dashboard: Insights from K-Means clustering with principal component analysis. *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi, 9*(2), 137–144.

Rosadi, M. W., Fatonah, N. S., Firmansyah, G., & Akbar, H. (2025). Clustering-based identification of student support needs in higher education transition. *JUSIFO (Jurnal Sistem Informasi), 11*(2), 133–140.

Valles-Coral, M. A., Salazar-Ramírez, L., Injante, R., Hernandez-Torres, E. A., Juárez-Díaz, J., Navarro-Cabrera, J. R., Pinedo, L., & Vidaurre-Rojas, P. (2022). Density-based unsupervised learning algorithm to categorize college students into dropout risk levels. *Data, 7*(11), 165.