Jurnal AL-AZHAR INDONESIA SERI SAINS DAN TEKNOLOGI
Vol. 11, Nomor 03, September 2026, hal. 235-250
DOI http://dx.doi.org/10.36722/sst.v11i3.5149 https://jurnal.uai.ac.id/index.php/SST

Evaluasi Multi-Metrik Kinerja K-Means, K-Medoids, dan K-Modes untuk Segmentasi dan Strategi Promosi Mahasiswa Baru Tepat Sasaran

Rachmat Firman[1*], Utomo Budiyanto[1]

[1]Program Magister Ilmu Komputer, Teknologi Informasi, Universitas Budi Luhur,
Jalan Ciledug Raya,Petukangan Utara, Jakarta Selatan 12260

Penulis untuk Korespondensi/E-mail: 2111600538@student.budiluhur.ac.id

Abstract – **Categorical student segmentation often suffers from bias due to subjective cluster (K) determination and vulnerability to local optima. This research proposes a robust comparative framework among K-Means, K-Modes, and K-Medoids to optimize university promotion strategies. Model robustness was tested via 30 independent iterations and evaluated fairly in an "Equivalent Evaluation Space" based on Hamming distance. Optimal architecture determination was automated using Multi-Criteria Decision Making (MCDM) integrating Borda Score and Z-Score Composite. Regarding computational Output, K-Means (K=10) dominated with the strongest cluster density (BetaCV=1.8705), highest initialization stability (ARI=0.5010), and maximum Borda score (12 points). As a managerial Outcome to mitigate Customer Acquisition Cost (CAC) inefficiency, the ten clusters were synthesized into three strategic macro-segments. Promotion tactics are directed towards: (1) Regional Network (25.47%) driven by Word-of-Mouth referrals; (2) Urban Premium (25.00%) via Omnichannel Ads; and (3) Cost-Sensitive Vocational (49.53%) requiring proactive school roadshows and scholarships. This integrated modeling successfully transforms intuitive mass marketing into measurable, efficient, and data-driven precision marketing.**

Abstrak – **Segmentasi mahasiswa berdimensi kategorikal sering bias akibat penentuan jumlah klaster (K) subjektif dan kerentanan algoritma terhadap *local optima*. Penelitian ini bertujuan untuk mengusulkan dan mengevaluasi kerangka komparasi robust antara algoritma K-Means, K-Modes, dan K-Medoids guna menetapkan arsitektur model klasterisasi terbaik sebagai landasan penyusunan strategi promosi perguruan tinggi yang presisi dan efisien. Ketahanan model diuji melalui 30 iterasi independen dan dievaluasi pada "Ruang Evaluasi Setara" berbasis jarak *Hamming*. Penentuan arsitektur optimal diotomatisasi melalui *Multi-Criteria Decision Making* (MCDM) mengintegrasikan *Borda Score* dan *Z-Score Composite*. Secara *Output* komputasi, K-Means (K=10) mendominasi dengan kepadatan klaster terkuat (*BetaCV*=1.8705), stabilitas inisialisasi tertinggi (ARI=0.5010), dan skor maksimal *Borda* (12 poin). Sebagai *Outcome* manajerial untuk menekan *Customer Acquisition Cost* (CAC), kesepuluh klaster tersebut disintesis menjadi tiga segmen makro strategis. Taktik promosi diarahkan pada: (1) Jaringan Regional (25,47%) berbasis rujukan *Word-of-Mouth*; (2) *Urban Premium* (25,00%) melalui kampanye *Omnichannel Ads*; dan (3) Vokasi Sensitif Biaya (49,53%) melalui intervensi *roadshow* sekolah dan opsi beasiswa. Pemodelan terintegrasi ini terbukti mampu mentransformasi pemasaran massal berbasis intuisi menjadi taktik *precision marketing* yang terukur, efisien, dan digerakkan oleh data.**

Keywords – *Comparison Algorithm, Multi-Criteria Decision Making, Promotion Strategy, Robust Evaluation, Student Clustering*

PENDAHULUAN

Penerimaan Mahasiswa Baru (PMB) yang krusial di tengah persaingan ketat menuntut institusi untuk menerapkan strategi promosi yang presisi. Pendekatan promosi konvensional (*one-size-fits-all*) berbasis asumsi subjektif terbukti memboroskan anggaran pemasaran tanpa garansi tingkat konversi

Received: 02 December 2025, Accepted: 30 July 2026, Published: 30 September 2026

yang optimal [1]. Oleh karena itu, pendekatan *Educational Data Mining* sangat dibutuhkan untuk mengekstraksi pola tersembunyi karakteristik sosio-demografi calon mahasiswa dari sekumpulan data historis [2-3].

Teknik *cluster analysis* sangat relevan dan ideal untuk tugas segmentasi pasar. Berbagai penelitian terdahulu telah sukses mengaplikasikan algoritma K-Means guna memetakan strategi dan lokasi penyebaran promosi PMB secara lebih terarah [4- 8]. Khusus untuk menangani data profil pendaftar yang didominasi tipe kategorikal, algoritma K-Modes juga telah diimplementasikan [9]. Pada evaluasi komparatif, perbandingan kinerja antara K-Means dan K-Medoids telah banyak dikaji menggunakan metrik internal seperti *Davies-Bouldin Index* (*DBI*) dan *Silhouette Coefficient* [3], [10-11].

Studi-studi terdahulu masih menyisakan dua celah metodologis yang fundamental. Pertama, dominasi penggunaan inisialisasi tunggal (*single-run initialization*) membuat model klasterisasi rentan terjebak pada *local optima* dan menghasilkan partisi yang tidak stabil [3], [10]. Kedua, sering terjadi bias perbandingan (*unfair comparison*) akibat pemaksaan metrik *Euclidean distance* pada data kategorikal murni, sebuah evaluasi yang secara matematis tidak valid [3], [11].

Guna menjembatani celah penelitian tersebut, penelitian ini mengusulkan kerangka komparasi sistematis yang *robust* antara algoritma K-Means, K-Modes, dan K-Medoids dalam memetakan profil calon mahasiswa berdimensi kategorikal. Stabilitas dan ketahanan model diuji secara ketat melalui eksekusi 30 iterasi independen (*multiple runs*) berbasis *Adjusted Rand Index* (ARI). Guna mengeliminasi bias metrik jarak, dibangun "Ruang Evaluasi Setara" (*Equivalent Evaluation Space*) menggunakan proksimasi jarak *Hamming*, lalu performa diagregasikan secara objektif melalui *Multi-Criteria Decision Making* (MCDM) berbasis integrasi *Borda Score* dan *Z-Score Composite*. Melalui pemodelan matematis terpadu ini, penentuan arsitektur klaster optimal tidak lagi bergantung pada intuisi subjektif, melainkan diarahkan untuk mentransformasi pendekatan pemasaran perguruan tinggi menjadi taktik promosi presisi berbasis data (*data-driven*), sehingga mampu memitigasi risiko pembengkakan *Customer Acquisition Cost* (CAC), sekaligus mengoptimalkan efisiensi alokasi iklan, pemetaan wilayah prioritas, serta saluran kampanye PMB.

METODE

Desain, tempat dan waktu

Penelitian ini menggunakan Desain Eksperimen Komputasional Terapan di bidang penambangan data (*Data Mining*) yang secara khusus berfokus pada komparasi performa algoritma *unsupervised learning* (K-Means, K-Modes dan K-Medoids). Penelitian dilaksanakan di Universitas Al-Azhar Indonesia pada rentang waktu Juli hingga September 2025. Data yang digunakan adalah data mahasiswa dari tahun akademik 2019/2020 sampai 2024/2025. Eksperimen komputasi menggunakan spesifikasi perangkat keras prosesor Intel Core i5, memori RAM 16 GB berbasis sistem operasi Windows 10.

Jumlah dan cara pengambilan subjek (survei) / bahan dan alat (laboratorium)

Penelitian ini murni menggunakan pendekatan komputasional terhadap data historis pendaftaran, untuk mematuhi etika penelitian dan menjamin privasi data subjek, seluruh data primer pendaftar telah dianonimkan (*anonymized*) dengan menghapus identitas personal (*Personally Identifiable Information* / PII) seperti nama lengkap, NIK dan nomor telepon sebelum tahapan pra-pemrosesan (*preprocessing*) dilakukan. Izin penggunaan data sekunder ini telah disetujui secara internal untuk kepentingan penelitian akademis. Penelitian ini menggunakan pendekatan *Educational Data Mining* (EDM) berbasis instrumen komputasi terhadap basis data sekunder (historis pendaftaran mahasiswa baru). Oleh karena itu, penelitian tidak melakukan penyebaran kuesioner survei secara primer.

Padanan logis dari teknik *purposive sampling* yang lazim digunakan pada penelitian sosial, penelitian ini menerapkan kriteria inklusi dan eksklusi yang ketat melalui mekanisme data *cleansing*. Kriteria inklusi meliputi data pendaftar yang memiliki kelengkapan atribut demografi dan asal sekolah, sedangkan kriteria eksklusi membuang data yang berpotensi *noise* (seperti nilai yang kosong/*NULL*). Melalui *tools* bahasa pemrograman Python (Pustaka Pandas dan Scikit-Learn), proses seleksi dari total populasi awal sebanyak 7.111 rekam data mentah berhasil menyaring dan menghasilkan 1.720 sampel data bersih (*Gold Dataset*). Pendekatan *purposive* berbasis aturan kebersihan data (*data rule-based*), dinilai sangat representatif untuk mewakili karakteristik populasi secara komprehensif tanpa memerlukan instrumen survei manual, serta menghasilkan keluaran (*output*) berupa matriks data siap-olah untuk pemodelan algoritma.

Jenis dan Cara Pengumpulan Data (Survei)/ Tahapan Penelitian (laboratorium)

Penelitian ini dirancang melalui sebuah jalur (*pipeline*) pemrosesan terotomatisasi untuk mengevaluasi dan membandingkan secara objektif algoritma K-Means, K-Modes dan K-Medoids pada dataset kategorikal. Tahapan penelitian komputasional ini diorkestrasikan dalam enam langkah sekuensial (Gambar 1), (1) Pra-pemrosesan dan Standardisasi data mentah melalui teknik pembersihan dan purifikasi pencocokan string berbasis NPSN, (2) Transformasi fitur menggunakan *One-Hot Encoding* untuk K-Means serta *Ordinal Encoding* untuk K-Modes dan K-Medoids, (3) Eksekusi pemodelan iteratif sebanyak 30 kali (*multiple runs*) guna mengukur stabilitas model secara ketat, (4) Evaluasi metrik internal dan eksternal di dalam "Ruang Evaluasi Setara" berbasis jarak *Hamming*, (5) Penentuan jumlah klaster optimal pada rentang K=3 hingga 10 melalui kalkulus diskrit (titik siku analitik) dan validasi probabilitas hingga (6) Penerapan *Multi-Criteria Decision Making* (MCDM) melalui *Borda Score* dan *Z-Score Composite* untuk menetapkan algoritma terbaik sekaligus mengekstraksi profil klaster final.

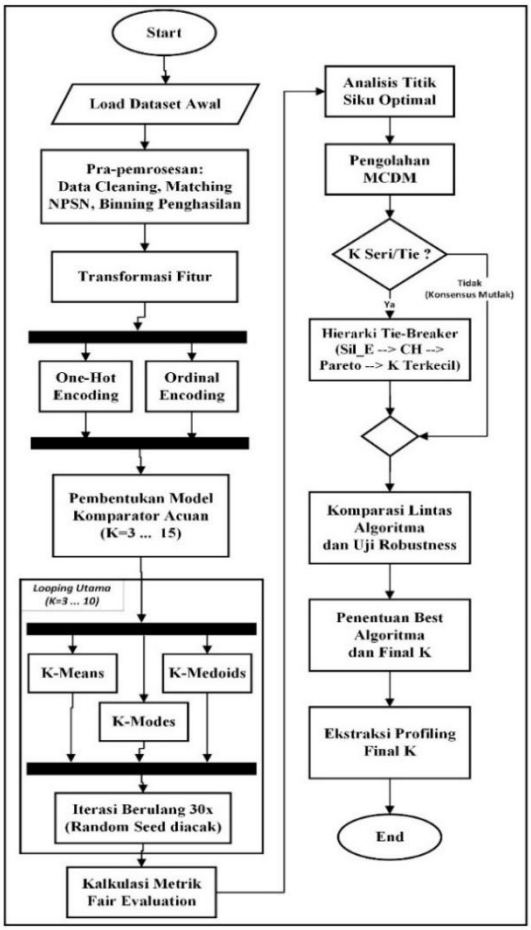
Seluruh eksperimen dieksekusi menggunakan *Python* 3.10. Pustaka *Pandas* dan *NumPy* diandalkan untuk proses manipulasi matriks. Pemodelan K-Means memanfaatkan *Scikit-Learn*, sementara K-Modes dan K-Medoids diimplementasikan menggunakan pustaka Kmodes untuk menjamin akurasi kalkulasi jarak *Hamming*. Terakhir, visualisasi analisis titik siku dikomputasi menggunakan pustaka grafis *Yellowbrick* dan *Matplotlib*.

Pengolahan dan Analisis data

Secara garis besar, tahapan pengolahan dan analisis data dalam penelitian ini mengadopsi kerangka kerja *Knowledge Discovery in Databases* (KDD) yang meliputi seleksi data, pra-pemrosesan, transformasi, *data mining* dan evaluasi [2-3]. Alur ini berfungsi sebagai peta jalan metodologis yang menjadi acuan bagi seluruh proses komputasi klasterisasi.

Pra-pemrosesan dan Standardisasi data

Sebelum tahapan pemodelan klasterisasi dieksekusi, kualitas data mentah dipastikan melalui tahapan pembersihan data (*data cleansing*) dan pengayaan (*enrichment*) untuk menangani anomali, ketidakkonsistenan format, serta nilai kosong (*missing values*) [2], [12]. Pra-pemrosesan diawali dengan ekstraksi fitur esensial dari file awal pendaftaran. Observasi dengan identitas asal sekolah yang terindikasi tidak valid segera dieliminasi. Langkah krusial berikutnya adalah standardisasi entitas asal institusi menggunakan pendekatan *string matching* [13].



Gambar 1. Diagram Alir Penelitian Komparasi Algoritma Klasterisasi

Selanjutnya, transformasi dimensi diterapkan pada atribut finansial kontinu yang didiskritisasi (*binning*) ke dalam tiga rentang kategorial untuk meredam dampak *outlier* [3], [14]. Tahapan final pra-pemrosesan berfokus pada purifikasi data menggunakan strategi penghapusan berjenjang. Proses transformasi data mentah yang tidak terstruktur menjadi terstruktur ini sejalan dengan pendekatan yang dilakukan oleh Isnanto dan Widodo [15] dalam memetakan potensi calon mahasiswa baru, sehingga menghasilkan set data bersih (*Gold Dataset*) yang dikunci sebagai basis empiris pemodelan utama [2].

Transformasi Fitur *(Feature Encoding)*

Karena penelitian ini membandingkan tiga algoritma dengan basis perhitungan metrik jarak yang berbeda, standardisasi ruang input menjadi hal mutlak. Pada algoritma K-Means, seluruh fitur profil siswa

ditransformasikan menggunakan *One-Hot Encoding* (OHE) menjadi matriks biner independen agar jarak *Euclidean* tidak salah menafsirkan urutan hierarki spasial [3].

Pada sisi lain, perlakuan ganda (*Dual Encoding*) diterapkan khusus pada algoritma K-Modes dan K-Medoids, mengingat kedua algoritma ini optimal memproses metrik jarak *Hamming* (*Simple Matching Distance*) untuk data kategorikal, fitur kategorikal utama ditransformasikan menggunakan *Ordinal Encoding* [3]. Konversi ini memungkinkan algoritma menghitung proporsi persentase ketidakcocokan (*mismatches*) kategori secara mangkus [2]. Pendekatan ini terbukti sangat efektif dalam menangani data kategorikal kompleks pada domain pendidikan, seperti latar belakang ekonomi atau demografi mahasiswa, sebagaimana divalidasi oleh Wirawan dan Prasetyawan [9], dalam pemodelan K-Modes untuk rekomendasi kebijakan kampus. Pemilihan K-Medoids dan K-Modes secara komparatif juga didasari oleh sifat algoritma K-Medoids yang lebih robust terhadap *outlier* dibandingkan K-Means [10], [11].

Pemodelan Iteratif dan Analisis Titik Siku Analitik

Algoritma klastering partisional memiliki sifat stokastik, konvergensi akhir sangat rentan terhadap sensitivitas inisialisasi awal [2]. Guna memitigasi bias tersebut, penelitian ini mengimplementasikan kerangka *robust* melalui pengulangan komputasi mandiri sebanyak 30 iterasi independen untuk setiap skenario jumlah klaster (K = 3 hingga K = 10).

Penentuan kandidat jumlah klaster (K) tidak lagi mengandalkan inspeksi visual kurva heuristik *Elbow Method* konvensional yang rentan bias dan bersifat subjektif [16-17]. Pendekatan visual ini disubstitusi dengan kalkulus diskrit matematis. Penurunan nilai *Sum of Squared Errors* (SSE) diekstraksi dari konsep selisih error antar klaster [18-19] yang secara algoritmik dihitung sebagai aproksimasi turunan pertama diskrit (1) [20].

$$\Delta SSE_k = SSE_{k-1} - SSE_k \tag{1}$$

Selanjutnya, untuk mensimulasikan pencarian titik siku secara otomatis, analisis dilanjutkan dengan mengamati perubahan lereng (*slopes*) melalui kalkulasi turunan kedua diskrit. Titik siku optimal ditentukan dengan mencari nilai puncak dari turunan kedua yang merepresentasikan tingkat percepatan perlambatan dari penurunan *error* klaster (2) [20].

$$\Delta^2 SSE_k = (SSE_{k-2} - SSE_{k-1}) - (SSE_{k-1} - SSE_k) \tag{2}$$

Arsitektur model yang ideal diproyeksikan pada kandidat K yang berhasil mencatatkan kelengkungan maksimum, yakni nilai positif paling tinggi. Lebih jauh, guna mengantisipasi ambiguitas pada kelengkungan kurva yang kompleks yang turunan kedua bisa saja menghasilkan beberapa puncak lokal akibat *noise* data, deteksi matematis ini dikalibrasi ulang menggunakan penaksiran Algoritma Geometris Kneedle yang mendeteksi titik tekuk berdasarkan jarak tegak lurus maksimum terhadap garis singgung kurva [16].

Selanjutnya, validasi statistik probabilitas acak dilakukan melalui *Gap Statistic* yang membandingkan dispersi intra-klaster data aktual dengan ekspektasi dispersi dari distribusi referensi acak (*null reference distribution*) untuk memastikan signifikansi jumlah klaster [21]. Integrasi ketiga metode ini (Turunan Kedua, Kneedle dan *Gap Statistic*) menjamin bahwa batas klaster terpilih merupakan sebuah representasi absolut dari pola data intrinsik mahasiswa pendaftar, bukan sekadar kebetulan komputasional.

Sistem Pengambilan Keputusan Multi-Kriteria (MCDM)

Dalam memastikan validitas arsitektur sebelum diagregasi, performa algoritma diukur secara kuantitatif menggunakan metrik standar evaluasi internal klastering. Dua indikator krusial yang digunakan untuk mengevaluasi kohesi dan separasi ruang Euclidean adalah *Silhouette Coefficient* dan *Davies-Bouldin Index* (DBI) [2], [22].

Silhouette Coefficient mengukur seberapa baik objek yang berada di dalam klasternya sendiri dibandingkan klaster lain, dengan rumus (3) [22]:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)))} \tag{3}$$

$a(i)$ adalah rata-rata jarak objek i ke semua objek lain dalam klaster yang sama (intra-klaster) dan $b(i)$ adalah rata-rata jarak minimum objek i ke semua objek pada klaster terdekat yang berbeda (inter-klaster). Sementara itu, DBI mengukur rasio penyebaran intra-klaster terhadap jarak separasi antar-klaster dengan persamaan (4) [11].

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \tag{4}$$

k adalah jumlah klaster, S_i dan S_j adalah ukuran dispersi dari klaster i dan j, serta M_{ij} adalah jarak pemisah antara pusat klaster i dan j. Evaluasi

Euclidean ini juga diperkuat oleh *Calinski-Harabasz Index* (*CH Index*) dan *Dunn Index* [3]. Selain menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* (DBI), kerangka kerja *Fair Evaluation Space* dalam penelitian ini memperhitungkan 3 metrik validasi internal komplementer guna menghindari bias geometri klaster, yaitu *CH Index*, *Dunn Index* dan *BetaCV Index*.

CH Index (secara matematis ekuivalen dengan metrik *Pseudo F-Statistic*) mengukur rasio dispersi antar-klaster (*between-cluster variance*) terhadap dispersi di dalam klaster (*within-cluster variance*) [23][10]. Formulasi CH didefinisikan Formula (5).

$$CH = \frac{SS_B}{SS_W} \, x \, \frac{N-K}{K-1} \tag{5}$$

SS_B adalah *Sum of Squares Between-clusters* (tingkat separasi antar-klaster), SS_W adalah *Sum of Squares Within-clusters* (tingkat kohesi internal klaster), N melambangkan total sampel data (1.720 data) dan K melambangkan jumlah klaster yang diuji ($3 \leq K \leq 10$). Nilai CH yang semakin tinggi menunjukkan struktur klasterisasi yang semakin baik dan terpisah secara tegas. Guna mengevaluasi ketajaman batas klaster pada variasi model seperti K-Modes, *Dunn Index* digunakan untuk membandingkan jarak spasial terkecil antar-klaster dengan diameter klaster terbesar [24-25]. Formulasi DI sebagai berikut (6).

$$DI = \frac{min_{1 \leq i < j \leq K} \, d(C_i, C_j)}{max_{1 \leq k \leq K} \, diam\,(C_k)} \tag{6}$$

$d(C_i, C_j)$ merupakan jarak inter-klaster minimum antara klaster C_i dan C_j, sedangkan $diam\,(C_k)$ melambangkan jarak intra-klaster maksimum (diameter) pada klaster C_k. Struktur pengelompokan dinyatakan optimal apabila nilai DI mendekati maksimal yang mencerminkan pemisahan tinggi dan kekompakan internal yang rapat.

Metrik *BetaCV Index* diintegrasikan untuk mengevaluasi konsistensi pengelompokan yang mengukur rasio margin intra-klaster pada margin inter-klaster [24-25]. Metrik ini mengevaluasi rasio rerata bobot jarak pasangan objek pada klaster yang sama terhadap rerata bobot jarak pasangan objek klaster berbeda, dengan formula (7).

$$BetaCV = \frac{W_c \,/\, N_W}{B_c \,/\, N_B} \tag{7}$$

W_c melambangkan total jarak spasial seluruh pasangan data intra-klaster, B_c melambangkan total jarak pasangan data inter-klaster, sedangkan N_W dan N_B berturut-turut menunjukkan jumlah pasangan objek intra-klaster dan inter-klaster. Secara teoritis, nilai *BetaCV* konvensional diukur dengan mengoptimalkan nilai minimum, namun pada implementasi sistem komputasi dalam penelitian ini, formula diinversi menjadi rasio jarak pemisah (inter-klaster) terhadap jarak kepadatan (intra-klaster). Oleh karena itu, target optimasi pada sistem ini adalah maksimum, nilai *BetaCV* yang semakin besar menunjukkan separasi batas antar-klaster yang lebih tegas dan kohesi internal yang lebih rapat.

Mengingat metrik evaluasi internal kerap merekomendasikan jumlah K yang kontradiktif, pengambilan keputusan diatasi dengan metode pemungutan suara lintas-metrik (*Voting Baseline*) menggunakan prinsip *Borda Score* [26]. Setiap metrik mendistribusikan peringkat absolut kepada nilai K yang menduduki titik ekstrem lokalnya. Apabila mekanisme agregasi ini memicu perolehan poin yang setara (*tie*), resolusi *cascading tie-breaker* diterapkan secara bertahap: (1) Mencari nilai rata-rata *Silhouette* tertinggi; (2) Melacak skor *CH Index* maksimum; (3) Menentukan model yang selisih jarak absolutnya paling dekat dengan titik lengkungan Δ^2 terbesar; hingga akhirnya (4) Bersandar pada prinsip *Occam's Razor* sebuah prinsip fundamental dalam *Machine Learning* yang menyatakan bahwa di antara model-model dengan tingkat validitas yang setara, model dengan kompleksitas terendah harus dipilih untuk mencegah *overfitting* dan menjaga efisiensi komputasi [2-3]. Sebagai jaring pengaman analitis ganda, sistem mensimulasikan agregasi alternatif meliputi ekstraksi komposit *Z-Score* dari seluruh rata-rata metrik standar, serta pendekatan *Pareto Optimality* untuk mencari set keseimbangan (*non-dominated solutions*) pada performa tiap model. Stabilitas persetujuan metrik pada uji tarik komputasional inilah yang menggaransi kualitas pembentukan klaster.

Ruang Evaluasi Setara dan Komparasi Lintas Algoritma

Pada tahap akhir pengerjaan, ketiga kandidat terbaik (satu model pemenang dari masing-masing algoritma K-Means, K-Modes dan K-Medoids) diperhadapkan secara paralel dalam analisis komparatif performa. Tantangan metodologis utama dalam komparasi lintas-algoritma adalah adanya perbedaan basis matriks jarak yang digunakan [10], untuk menjamin perbandingan silang (*cross-*

algorithm) yang bebas dari bias spasial, seluruh label diproyeksikan dan dievaluasi setara kategorikal dengan mengimplementasikan matriks *pairwise* jarak *Hamming*. Khusus untuk K-Means yang memproses matriks *One-Hot Encoding* (OHE), menjadi nilai probabilitas kontinu, dilakukan tahapan transformasi balik (*decoding*) menggunakan fungsi ambang batas maksimum (*argmax*) untuk mengembalikan sentroid ke bentuk diskrit kategorikal aslinya. Langkah ini krusial untuk memastikan K-Means dapat diukur secara adil, mengingat data pendaftaran mahasiswa baru didominasi oleh atribut campuran dan kategorikal [9], [27]. Pengukuran universal lintas algoritma didasarkan pada komparasi gabungan skor kedekatan *Silhouette-H*, tingkat kerenggangan *Dunn-H*, serta tingkat kekokohan distribusi label lintas-run melalui *Stability* ARI [3].

Seluruh peringkat metrik *Hamming* tersebut kemudian dikonversi menjadi perolehan *Z-Score Composite* dan *Borda Score* agregat untuk menyeimbangkan nilai komposit secara proporsional. Model tunggal lintas-algoritma yang berhasil meraih nilai tertinggi dari konsensus agregat kedua metode ini akan dikukuhkan sebagai *Best Algorithm* dan dipilih sebagai arsitektur final untuk klaster pendaftaran mahasiswa baru. Proses pengerjaan diakhiri dengan melakukan *decoding cluster* profiling untuk memetakan karakteristik fundamental dari kelompok terpilih. Atribut partisi diterjemahkan ke dalam dua format yaitu profil *Pure* (hanya memuat atribut masukan klasterisasi) dan profil *Extended* (yang menarik kembali atribut institusi tersembunyi seperti nama sekolah asal secara eksplisit) untuk memastikan penarikan wawasan bisnis yang holistik bagi penyusunan strategi promosi dan manajerial kampus [7-9].

HASIL DAN PEMBAHASAN

Hasil Eksekusi *Pipeline Cleansing* dan Purifikasi *Dataset* Final Sebelum pemodelan klasterisasi dieksekusi, dataset pendaftaran mahasiswa dipurifikasi melalui tujuh tahapan pemrosesan terotomatisasi untuk memastikan kelengkapan absolut (*strict completeness*), dari populasi awal 7.111 rekam data mentah (Tahap 1), sistem mengeksklusi baris tanpa identitas nama sekolah yang valid sehingga menyusut menjadi 5.871 baris (Tahap 2). Data ini kemudian diintegrasikan dengan data master NPSN untuk menstandarkan entitas asal sekolah (Tahap 3), disusul dengan pelaksanaan *feature engineering* pada atribut penghasilan orang tua (Tahap 4).

Purifikasi berjenjang selanjutnya mengeliminasi data dengan nilai kosong pada atribut kalkulasi penghasilan yang menyisakan 4.882 baris (Tahap 5). Penyusutan paling masif terjadi saat menyaring kelengkapan "Status Sekolah" yang memotong populasi menjadi 2.610 baris (Tahap 6). Terakhir, filter simultan diterapkan untuk membuang anomali data kosong pada "Sumber Informasi" dan "Provinsi Alamat Mahasiswa" (Tahap 7). Rangkaian purifikasi ini secara presisi mengunci populasi akhir pada 1.720 baris data. Hasil akhir yang disebut sebagai *Gold Dataset* ini terjamin bebas dari bias *missing value*, menjadikannya fondasi empiris yang kokoh untuk analisis segmentasi promosi mahasiswa baru. Alur penyusutan data dari setiap fase purifikasi ini dirangkum secara transparan pada Tabel 1.

Tabel 1. Ringkasan Filtrasi dan Purifikasi *Dataset*

Tahap	Aktivitas Pemrosesan Data (*Cleansing* & *Enrichment*)	Jumlah Data Tersisa
1	*Load* Data Awal & *Feature Selection* (9 atribut esensial)	7.111 baris
2	Eliminasi *missing value* pada atribut "Nama Sekolah - SMA"	5.871 baris
3	Standardisasi institusi, pemetaan kamus, & *enrichment* (Integrasi *Data master*)	5.871 baris
4	*Feature Engineering* (Kalkulasi estimasi & pemetaan 3 kelas penghasilan)	5.871 baris
5	Eliminasi *missing value* / nilai nol pada "Penghasilan Orang Tua"	4.882 baris
6	Eliminasi *missing value* pada atribut "Status Sekolah"	2.610 baris
7	Eliminasi *strict missing* value akhir ("Sumber Informasi" & "Provinsi Alamat")	1.720 baris

Sebagai gambaran awal (*Exploratory Data Analysis*), populasi *Gold Dataset* sebanyak 1.720 sampel ini memperlihatkan kecenderungan distribusi demografis tertentu. Berdasarkan ekstraksi awal, pendaftar didominasi oleh kelompok prasejahtera (estimasi pendapatan < Rp 4 Juta) yang mencakup hampir setengah populasi (49,52%), diikuti oleh kelas menengah dan kelas atas. Sebaran asal institusi pendidikan menunjukkan dominasi lulusan SMK/Vokasi, dengan konsentrasi domisili terpusat di kawasan DKI Jakarta dan wilayah penyangganya (Jawa Barat dan Banten). Pemahaman terhadap distribusi dasar (*baseline*) ini menjadi parameter krusial pada tahapan evaluasi selanjutnya untuk memvalidasi apakah algoritma klasterisasi benar-benar berhasil menemukan sub-pola multivariat yang tersembunyi, atau sekadar mereplikasi distribusi statistik mayoritas.

Validasi Internal dan Penentuan Jumlah Klaster (K) Optimal

Tahapan pemodelan klasterisasi pada penelitian ini diawali dengan proses pencarian jumlah kelompok (K) yang paling optimal untuk ketiga algoritma (K-Means, K-Modes dan K-Medoids). Pengujian dieksekusi secara ketat pada rentang kandidat arsitektur K=3 hingga K=10. Guna menjamin objektivitas hasil dan menghindari bias inisialisasi awal yang berisiko menjebak model pada *local optima*, setiap skenario K dieksekusi secara independen sebanyak 30 kali iterasi (30x *multiple runs*). Nilai rata-rata (*mean*) dari kinerja setiap metrik internal pada 30 iterasi tersebut kemudian diposisikan sebagai parameter kuantitatif dasar. Selain itu, kalkulus turunan kedua diskrit (Δ^2) turut diimplementasikan untuk mendeteksi titik patahan siku (Elbow) secara algoritmik otomatis, menggantikan metode penaksiran visual yang subjektif. Seluruh instrumen validasi tersebut selanjutnya dikonsolidasikan melalui sebuah kerangka agregasi khusus.

Kerangka Evaluasi Multi-Kriteria (MCDM) pada Pemodelan Klasterisasi

Dalam studi *unsupervised learning*, penentuan arsitektur atau jumlah klaster (K) sering kali terjebak pada heuristik tunggal, seperti inspeksi visual pada kurva *Elbow*. Pendekatan konvensional tersebut rentan terhadap bias subjektivitas dan sering mengabaikan multidimensionalitas kualitas partisi, terutama ketika menghadapi matriks data campuran (*sparse matrix*) berdimensi tinggi. Dalam mereduksi bias tersebut, penelitian ini mengimplementasikan kerangka *Multi-Criteria Decision Making* (MCDM) untuk menentukan nilai K optimal pada seluruh algoritma yang diuji (K-Means, K-Modes dan K-Medoids).

Proses evaluasi ini dibangun di atas 'Ruang Evaluasi Setara'. Pada algoritma K-Means, evaluasi difokuskan pada proksimasi geometris (*Euclidean*), sementara pada K-Modes dan K-Medoids, ruang evaluasi diperluas dengan mengintegrasikan metrik jarak *Hamming* (seperti *Silhouette Hamming*, *Dunn Hamming* dan *BetaCV*). Seluruh parameter performa dari 30 kali iterasi (*multiple runs*) termasuk kualitas kohesi, separasi, akselerasi kurva siku (Δ^2*Cost*/SSE) dan stabilitas stokastik (ARI) diekstraksi menjadi skor terstandardisasi menggunakan *Z-Score Composite*. Skor kualitas agregat ini kemudian dipertandingkan melalui sistem pemungutan suara mayoritas (*Borda Score*). Melalui pendekatan konsensus holistik ini, arsitektur final dari masing-masing algoritma ditetapkan secara matematis, objektif dan terhindar dari interpretasi visual parsial.

Evaluasi dan Penentuan Arsitektur Optimal Algoritma K-Means

Algoritma K-Means memiliki ketergantungan yang sangat tinggi terhadap penempatan pusat klaster (*centroid*) awal yang ditentukan secara acak. Inisialisasi tunggal (*single run*) sering kali menyebabkan model terjebak pada optimum lokal (*local optima*) dan menghasilkan partisi data yang tidak konsisten. Guna mengatasi kelemahan stokastik tersebut serta menguji ketangguhan model terhadap *curse of dimensionality* akibat ekspansi fitur *one-hot encoding*, pengujian pada penelitian ini dilakukan menggunakan skema *multiple runs* sebanyak 30 kali iterasi independen untuk setiap kandidat nilai K (dari K=3 hingga K=10).

Evaluasi kinerja model diukur secara komprehensif menggunakan ruang evaluasi multi-metrik yang mencakup *Silhouette Score*, *Davies-Bouldin Index* (DBI), *CH Index*, *Dunn Index*, serta nilai *Sum of Squared Errors* (SSE) beserta turunan kedua diskritnya (Δ^2SSE). Selain metrik kualitas partisi, konsistensi struktural geometris antar-iterasi diuji secara ketat menggunakan *Adjusted Rand Index* (ARI) sebagai parameter stabilitas model. Seluruh rekapitulasi data performa rata-rata (*mean*) dan simpangan baku (*standard deviation*) dari 30 kali eksekusi tersebut dirangkum dalam Tabel 2.

Tabel 2. Matriks Komprehensif Evaluasi Kinerja dan *Robustness* K-Means (30 Iterasi)

K	SSE / *Inertia*	Δ^2 SSE	*Silhouette Euc*	*DBI*	*CH Index*	*Dunn Index*	*Best* ARI vs Pembanding	*Z-Score*	*Borda Score*
3	5773.60 ± 75.28	-	0.12 ± 0.01	2.48 ± 0.13	201.71 ± 13.79	0.381	0.330	-1.424	20
4	5344.81 ± 68.13	-	0.13 ± 0.01	2.32 ± 0.14	191.11 ± 11.63	0.380	0.554	-0.54	20
5	5046.80 ± 115.76	130.77	0.13 ± 0.02	2.24 ± 0.16	177.20 ± 13.84	0.381	0.983	-0.729	35
6	4762.70 ± 100.92	13.92	0.13 ± 0.01	2.16 ± 0.15	170.53 ± 10.73	0.379	0.737	0.28	35
7	4601.52 ± 84.11	122.93	0.13 ± 0.01	2.18 ± 0.11	157.07 ± 8.21	0.381	0.703	-0.167	32
8	4455.90 ± 71.31	15.55	0.14 ± 0.01	2.19 ± 0.09	146.80 ± 6.95	0.384	0.782	0.736	37
9	4308.37 ± 59.41	-1.91	0.14 ± 0.01	2.17 ± 0.10	140.07 ± 6.12	0.382	0.514	0.851	31
10	4186.42 ± 48.92	25.58	0.14 ± 0.01	2.15 ± 0.08	133.60 ± 5.02	0.383	0.771	0.992	42

* Keterangan : Nilai SSE / Inertia, *Silhouette Euclidean*, *Davies - Bouldin Index*, *Calinski – Harabasz* disajikan dalam format rata-rata (*mean*) ± simpangan baku. Δ^2 SSE adalah ketajaman siku

Pemetaan metrik evaluasi internal pada Tabel 2 menunjukkan tren optimalisasi yang bervariasi dan saling berkontradiksi antar-arsitektur klaster. Kriteria berbasis kerapatan geometris merekomendasikan K=10 sebagai partisi paling solid, dibuktikan oleh puncak *Silhouette Euclidean* (0.14) dan titik minimum *Davies-Bouldin Index* (2.15). Sebaliknya, metrik *CH Index* mencatatkan skor maksimum pada K=3 (201.71) akibat fenomena degradasi varians bawaan pada matriks data renggang, sementara *Dunn Index* berfluktuasi dengan margin terlebar pada K=8 (0.384). Berdasarkan perspektif reduksi galat, kalkulus turunan kedua mendeteksi akselerasi pengereman siku (Δ^2SSE) paling tajam pada K=5 (130.77) yang turut divalidasi oleh tingkat kecocokan eksternal (*Best* ARI) tertinggi sebesar 0.983. Mengingat indikator fundamental tersebut memunculkan kandidat pemenang yang terpecah (K=3, K=5, K=8, dan K=10), penentuan jumlah kelompok tidak dapat disimpulkan secara heuristik parsial yang rentan bias subjektivitas. Oleh karena itu, seluruh performa metrik pada Tabel 2 diagregasikan secara kuantitatif menjadi nilai *Z-Score (Robust Composite Score*) dan perolehan *Borda Score*. Dua kolom agregasi terakhir tersebut selanjutnya menjadi landasan komputasi yang akan dibedah secara komprehensif melalui kerangka *Multi-Criteria Decision Making* (MCDM) guna menentukan pemenang mutlak pada matriks keputusan Tabel 3.

Berdasarkan matriks keputusan pada Tabel 3, persaingan performa arsitektur berlangsung sangat ketat dengan terpecahnya keunggulan pada K=5 (Elbow dan ARI), K=3 (CH), serta K=8 (*Dunn Index*). Mengingat perolehan metrik fundamental memicu kebuntuan antara K=5 dan K=10, penentuan arsitektur optimal harus diselesaikan melalui mekanisme *tie-breaker* multikriteria. Konfigurasi K=10 sukses memenangkan konsensus ini berkat dominasi kohesi *Silhouette* yang kokoh, sekaligus menyapu bersih agregasi pemeringkatan *Borda Score* dengan total 42 poin. Keunggulan mutlak tersebut turut divalidasi oleh capaian *Robust Composite Score* tertinggi sebesar 0.992. Berbeda dengan standardisasi agregasi biasa, kalkulasi pada *Robust Composite Score (Z-Score)* ini secara eksplisit menerapkan penalti volatilitas untuk menghukum ketidakstabilan inisialisasi acak yang menjadi kelemahan bawaan K-Means. Melalui kerangka MCDM yang tidak hanya mengukur puncak performa tetapi juga ketahanan terhadap instabilitas komputasional ini, arsitektur K=10 secara definitif dikukuhkan sebagai pemodelan yang paling tangguh dan optimal.

Tabel 3. Keputusan Konsensus MCDM untuk Penentuan K Optimal pada K-Means

Kriteria / Metrik Internal	Optimasi Target	K Pemenang	Nilai Performa Terbaik
Silhouette	Maksimum	10	0.14
DBI	Minimum	10	2.15
CH Index	Maksimum	3	201.71
Dunn Index	Maksimum	8	0.384
Δ^2SSE	Maksimum	5	130.77
Best ARI vs Pembanding	Maksimum	5	0.98
Z-Score	Maksimum	10	0.99
Borda Score	Maksimum	10	42
Total Akumulasi Voting Mayoritas	-	10	Terpilih *via* Konsensus & *Tie-Breaker*

Evaluasi dan Penentuan Arsitektur Optimal Algoritma K-Modes

Algoritma K-Modes secara khusus dirancang untuk menangani data kategorikal murni melalui optimalisasi titik pusat berbasis modus (*mode*) dan perhitungan jarak menggunakan metrik *Hamming* (ketidakcocokan atribut). Serupa dengan K-Means, algoritma ini sangat sensitif terhadap inisialisasi pusat klaster awal (algoritma inisialisasi Huang). Inisialisasi tunggal (*single run*) berisiko tinggi mengarahkan model pada *local optima*, yang menghasilkan topologi klaster yang tidak stabil. Oleh karena itu, pengujian ketangguhan model (*robustness*) dilakukan melalui skema *multiple runs* sebanyak 30 iterasi independen untuk setiap konfigurasi arsitektur K (dari K=3 hingga K=10).

Tantangan utama dalam mengevaluasi K-Modes adalah ketidaksesuaian metrik validasi internal klasik (seperti *Silhouette* dan *Davies-Bouldin*) yang pada dasarnya beroperasi di ruang *Euclidean*, untuk menciptakan "Ruang Evaluasi Setara", pengujian ini mengekstraksi metrik ke dalam dua kubu yaitu metrik yang diproyeksikan melalui *One-Hot Encoding* (*Euclidean*) dan metrik asli berbasis kategori (*Hamming*). Evaluasi kinerja diukur menggunakan rata-rata (*mean*) dan simpangan baku (*standard deviation*) dari fungsi *Cost* (pengganti SSE pada K-Modes), proksimasi *Euclidean* (*Silhouette*, DBI, CH), hingga konsistensi struktural menggunakan *Adjusted Rand Index* (ARI). Rekapitulasi performa 30 kali iterasi tersebut diuraikan pada Tabel 4.

Tabel 4. Matriks Komprehensif Evaluasi Kinerja dan *Robustness* K-Modes (30 Iterasi)

K	KModes *Cost / Error*	Δ^2 *Cost*	*Silhouette Euc*	*Silhouette Ham*	*DBI*	*CH Index*	*Dunn Index (Euc / Ham)*	*BetaCV Ham*	*Best* ARI vs Pembanding	*Z-Score*	*Borda Score*
3	4314.73 ± 96.69	-	0.086 ± 0.012	0.14 ± 0.02	2.82 ± 0.14	150.07 ± 15.23	0.38 / 0.14	1.264	0.277	-0.799	25
4	3990.83 ± 74.86	-	0.086 ± 0.013	0.14 ± 0.03	2.66 ± 0.13	140.01 ± 12.86	0.38 / 0.14	1.321	0.300	-0.436	21
5	3782.70 ± 97.57	-115.77	0.086 ± 0.009	0.13 ± 0.02	2.57 ± 0.13	127.60 ± 9.92	0.38 / 0.14	1.361	0.324	0.218	31
6	3567.43 ± 88.68	7.13	0.086 ± 0.012	0.13 ± 0.02	2.45 ± 0.13	120.51 ± 11.15	0.38 / 0.14	1.41	0.313	-0.169	22
7	3434.43 ± 89.07	-82.27	0.088 ± 0.013	0.13 ± 0.02	2.42 ± 0.13	115.59 ± 8.11	0.38 / 0.14	1.459	0.321	0.346	31
8	3317.93 ± 85.04	-11.90	0.087 ± 0.014	0.13 ± 0.02	2.38 ± 0.12	108.61 ± 8.03	0.38 / 0.14	1.489	0.309	-0.048	26
9	3181.77 ± 89.28	-7.23	0.092 ± 0.012	0.13 ± 0.02	2.35 ± 0.14	104.21 ± 6.89	0.38 / 0.14	1.529	0.325	0.359	29
10	3093.77 ± 74.96	-8.17	0.093 ± 0.013	0.13 ± 0.02	2.29 ± 0.11	100.70 ± 6.94	0.38 / 0.14	1.569	0.328	0.528	31

* Keterangan : Nilai KModes *Cost/Error*, *Silhouette Euclidean* dan *Hamming*, *Davies - Bouldin Index*, *Calinski – Harabasz* disajikan dalam format rata-rata (*mean*) ± simpangan baku. Δ^2 *Cost* adalah ketajaman siku

Evaluasi multi-metrik pada Tabel 4 memperlihatkan dikotomi performa yang unik antara evaluasi pada ruang proksimasi geometris (*Euclidean*) dan ruang natif kategorikal (*Hamming*). Metrik geometris secara konsisten mengunggulkan partisi K=10, dibuktikan oleh titik minimum *Davies-Bouldin Index* (2.29) dan kohesi *Silhouette Euclidean* tertinggi (0.093) yang pergerakannya turut divalidasi oleh puncak rasio kepadatan *BetaCV* (1.569) serta kecocokan eksternal *Best* ARI (0.328). Sebaliknya, *CH Index* memuncak secara absolut pada K=3 (150.07) seiring dengan tingginya *Silhouette Hamming* (0.14), merepresentasikan fenomena degradasi varians bawaan yang lazim terjadi pada matriks biner renggang ketika jumlah klaster bertambah. Berdasarkan perspektif reduksi galat, kalkulus turunan kedua mendeteksi akselerasi pengereman siku (Δ^2 *Cost*) paling tajam pada K=6 (7.13), sementara *Dunn Index* justru mengalami stagnasi absolut yang seragam (0.38/0.14) akibat keterbatasan pengukuran metrik jarak batas pada ruang diskrit. Mengingat indikator fundamental tersebut merekomendasikan kandidat arsitektur yang terpecah dan saling berkontradiksi (K=3, K=6, dan K=10), penentuan jumlah kelompok tidak dapat diputuskan secara heuristik parsial. Oleh karena itu, seluruh parameter performa tersebut diagregasikan secara objektif ke dalam kolom pemungutan suara *Borda Score* dan pembobotan *Z-Score Composite*. Kedua nilai agregasi inilah yang selanjutnya akan dibedah secara komprehensif melalui kerangka *Multi-Criteria Decision Making* (MCDM) guna menetapkan arsitektur pemenang mutlak pada matriks keputusan Tabel 5.

Matriks keputusan pada Tabel 5 merepresentasikan persaingan performa yang ketat, keunggulan terpecah antara K=3 pada proksimasi natif *Hamming* dan K=6 sebagai titik akselerasi penurunan galat terkuat (Δ^2*Cost* = 7.13). Di sisi lain, keterbatasan *Dunn Index* pada matriks diskrit renggang membuatnya gagal menyumbangkan suara akibat jarak inter-klaster yang konstan. Walaupun performa terfragmentasi, arsitektur K=10 berhasil mendominasi suara mayoritas dengan menguasai empat kriteria fundamental sekaligus, termasuk rasio kepadatan pemisahan kelompok (*BetaCV*) tertinggi sebesar 1.569. Keunggulan parsial ini membawa K=10 pada persaingan agregasi seri (*tie*) di peringkat *Borda Score* (31 poin) melawan arsitektur K=5 dan K=7. Guna memecahkan kebuntuan pemeringkatan tersebut, kerangka *Multi-Criteria Decision Making* (MCDM) menggunakan capaian *Z-Score Composite* sebagai *tie-breaker* yang murni diekstraksi dari rata-rata performa terstandardisasi. Berbekal perolehan skor komposit tertinggi (0.528), arsitektur K=10 divalidasi secara definitif sebagai konfigurasi yang paling tangguh dan optimal untuk memetakan data kategorikal pada algoritma K-Modes.

Tabel 5. Keputusan Konsensus MCDM untuk Penentuan K Optimal pada K-Modes

Kriteria / Metrik Internal	Optimasi Target	K Pemenang	Nilai Performa Terbaik
Silhouette (Euc)	Maksimum	10	0.093
Silhouette (Ham)	Maksimum	3	0.141
DBI	Minimum	10	2.293
CH Index	Maksimum	3	150.072
Δ^2 Cost	Maksimum	6	7.13

Kriteria / Metrik Internal	Optimasi Target	K Pemenang	Nilai Performa Terbaik
BetaCV Ham	Maksimum	10	1.569
Best ARI vs Pembanding	Maksimum	10	0.328
Z-Score	Maksimum	10	0.528
Borda Score	Maksimum	5,7,10	31
Total Akumulasi *Voting* Mayoritas	-	10	4 Kemenangan

Evaluasi dan Penentuan Arsitektur Optimal Algoritma K-Medoids

Tabel 6. Matriks Komprehensif Evaluasi Kinerja dan *Robustness* K-Medoids (30 Iterasi)

K	KMedoids *Cost / Error*	Δ^2 *Cost*	*Silhouette Euc*	*Silhouette Ham*	*DBI*	*CH Index*	*Dunn Index (Euc / Ham)*	*BetaCV Ham*	*Best* ARI vs Pembanding	*Z-Score*	*Borda Score*
3	624.50 ± 18.02	-	0.090 ± 0.013	0.149 ± 0.023	2.81 ± 0.19	149.46 ± 14.62	0.378 / 0.14	1.283	0.231	-0.762	20
4	561.05 ± 10.05	-	0.088 ± 0.012	0.141 ± 0.021	2.68 ± 0.14	136.25 ± 11.77	0.378 / 0.14	1.374	0.298	-0.006	30
5	541.75 ± 16.76	-8.429	0.091 ± 0.010	0.145 ± 0.019	2.57 ± 0.14	129.04 ± 12.44	0.378 / 0.14	1.477	0.281	0.143	30
6	510.12 ± 18.84	-8.286	0.089 ± 0.012	0.137 ± 0.022	2.45 ± 0.11	119.73 ± 8.36	0.378 / 0.14	1.451	0.259	-0.303	23
7	491.79 ± 19.21	-7.000	0.091 ± 0.009	0.137 ± 0.017	2.37 ± 0.09	117.83 ± 7.87	0.378 / 0.14	1.506	0.431	0.209	30
8	462.35 ± 9.63	1.143	0.089 ± 0.011	0.132 ± 0.021	2.36 ± 0.08	107.86 ± 8.06	0.378 / 0.14	1.496	0.269	0.073	25
9	441.68 ± 7.53	-1.286	0.091 ± 0.010	0.132 ± 0.019	2.29 ± 0.09	105.32 ± 5.43	0.378 / 0.14	1.574	0.294	0.198	27
10	423.42 ± 9.28	1.429	0.094 ± 0.007	0.134 ± 0.014	2.28 ± 0.06	101.06 ± 4.61	0.378 / 0.14	1.611	0.329	0.447	30

* Keterangan : Nilai KMedoids *Cost/Error*, *Silhouette Euclidean* dan *Hamming*, *Davies - Bouldin Index*, *Calinski – Harabasz* disajikan dalam format rata-rata (*mean*) ± simpangan baku. Δ^2 *Cost* adalah ketajaman siku

Berbeda dengan K-Means yang menghitung nilai rata-rata maya (*virtual mean*) atau K-Modes yang menggunakan modus kategorikal, algoritma K-Medoids menggunakan titik data observasi aktual (*representative objects* atau medoids) sebagai pusat klaster. Karakteristik ini membuat K-Medoids secara inheren lebih tangguh (*robust*) terhadap data pencilan (*outliers*) dan *noise*. Lebih jauh, algoritma ini memproses data kategorikal secara langsung pada ruang diskrit menggunakan metrik jarak *Hamming* tanpa memerlukan transformasi *One-Hot Encoding* pada saat pembentukan klaster.

Meskipun tangguh terhadap *outliers*, K-Medoids tetap memiliki kerentanan terhadap inisialisasi awal. Guna mengatasi masalah *local optima* dari inisialisasi algoritma K-medoids, pengujian dilakukan melalui skema *multiple runs* sebanyak 30 kali iterasi independen untuk setiap arsitektur K (dari K=3 hingga K=10). Kinerja algoritma tidak hanya dievaluasi pada ruang jarak *Hamming* bawaannya (melalui nilai *Cost* dan *Silhouette Hamming*), tetapi juga divalidasi silang pada ruang proksimasi geometris (*Euclidean*) untuk memberikan "Ruang Evaluasi Setara". Tingkat stabilitas topologi antar-iterasi diukur menggunakan *Adjusted Rand Index* (ARI). Rekapitulasi performa model dari 30 eksekusi tersebut disajikan Tabel 6.

Pemetaan performa pada Tabel 6 memperlihatkan bahwa fungsi *KMedoids Cost* tereduksi secara konsisten seiring bertambahnya jumlah klaster, dengan titik minimum dan akselerasi pengereman galat terkuat (Δ^2 *Cost*) terdeteksi selaras pada arsitektur K=10. Keunggulan K=10 ini turut divalidasi kuat di ruang proksimasi geometris melalui capaian *Silhouette Euclidean* tertinggi (0.094) dan tingkat pemisahan *Davies-Bouldin Index* paling solid (2.28), namun evaluasi di ruang natif kategorikal justru memunculkan polarisasi, kekompakan *Silhouette Hamming* (0.149) dan rasio *CH Index* (149.46) memuncak secara absolut pada K=3, yang merepresentasikan batas alamiah terluas dari matriks renggang sebelum titik-titik medoid saling berdesakan. Pada sisi lain, indikator rasio kepadatan klaster (*BetaCV*) kembali mengunggulkan K=10 (1.611), sementara metrik *Dunn Index* mengalami stagnasi akibat keterbatasan kalkulasi jarak seragam (0.378/0.14) dan validitas eksternal (*Best* ARI) secara independen memenangkan K=7 (0.431). Mengingat fragmentasi rekomendasi metrik fundamental ini memicu hasil seri yang ekstrem pada perolehan *Borda Score*, penentuan arsitektur yang paling tangguh (*robust*) tidak mungkin disimpulkan secara parsial. Oleh karena itu, seluruh nilai agregasi pemungutan suara beserta *Z-Score Composite* dari Tabel 6 ini akan didelegasikan secara komprehensif ke dalam kerangka *Multi-Criteria Decision Making* (MCDM) guna memecahkan kebuntuan pemeringkatan pada matriks keputusan Tabel 7.

Tabel 7. Keputusan Konsensus MCDM untuk Penentuan K Optimal pada K-Medoids

Kriteria / Metrik Evaluasi	Optimasi Target	K Pemenang	Nilai Performa Terbaik
Silhouette (Euc)	Maksimum	10	0.094
Silhouette (Ham)	Maksimum	3	0.149
DBI	Minimum	10	2.28

Kriteria / Metrik Evaluasi	Optimasi Target	K Pemenang	Nilai Performa Terbaik
CH Index	Maksimum	3	149.461
Δ^2 *Cost*	Maksimum	10	1.429
BetaCV (Ham)	Maksimum	10	1.611
Best ARI vs Pembanding	Maksimum	7	0.431
Z-Score	Maksimum	10	0.447
Borda Score	Maksimum	4,5,7,10	30
Total Akumulasi *Voting* Mayoritas	-	10	4 Kemenangan

Tabel 8. Komparasi Performa Lintas-Algoritma pada Ruang Evaluasi Setara (*Fair* Evaluation *Space*)

Algoritma (K-Optimal)	*Silhouette* (*Hamming*)	*Dunn* (*Hamming*)	*BetaCV* (*Hamming*)	*Stabilitas Model* (ARI)	*Borda Score*	*Z-Score Composite*
K-Means (K=10)	0.2482	0.1667	1.8705	0.5010	12	1.3753
K-Modes (K=10)	0.1196	0.1429	1.5411	0.2019	4	-0.9239
K-Medoids (K=10)	0.1667	0.1429	1.6111	0.2673	8	-0.4514

Konstelasi hasil MCDM pada Tabel 7 merepresentasikan peta persaingan yang sangat ketat, di mana keunggulan parsial sempat didominasi oleh arsitektur moderat seperti K=3 pada ruang natif *Hamming* dan K=7 pada tingkat validitas eksternal. Namun, arsitektur berdimensi tinggi (K=10) berhasil merebut suara mayoritas dengan mengamankan empat indikator fundamental sekaligus, meliputi pemisahan batas luar (DBI), kohesi geometri (*Silhouette Euclidean*), akselerasi siku (Δ^2 *Cost*), dan rasio kepadatan klaster (*BetaCV*). Fragmentasi rekomendasi metrik tersebut memicu hasil seri (*tie*) yang ekstrem pada sistem pemeringkatan *Borda Score*, di mana arsitektur K=4, K=5, K=7, dan K=10 sama-sama bersinggungan pada skor 30 poin. Guna meresolusi kebuntuan agregasi ini secara objektif, capaian *Z-Score Composite* diimplementasikan sebagai hakim pemutus (*tie-breaker*). Melalui standardisasi performa analitis tersebut, K=10 sukses membuktikan superioritas mutlaknya dengan meraih skor komposit tertinggi sebesar 0.447. Dominasi absolut pada kerangka evaluasi MCDM ini secara meyakinkan mematahkan asumsi sekunder (seperti *Gap Statistic*) yang sebelumnya merekomendasikan batas tiga klaster untuk mencegah *overfitting*. Oleh karena itu, K=10 dikukuhkan secara definitif sebagai arsitektur yang paling optimal dan tangguh (*robust*) untuk pemodelan K-Medoids pada dataset ini.

Komparasi Kinerja Lintas-Algoritma dan Pemilihan Model Final

Komparasi performa lintas-algoritma yang beroperasi pada fungsi optimasi berbeda rentan terhadap bias evaluasi apabila divalidasi menggunakan basis perhitungan matriks jarak bawaannya. Dalam menjamin perbandingan yang berkeadilan (*apple-to-apple*), penelitian ini membangun "Ruang Evaluasi Setara" yang memproyeksikan seluruh model ke dalam matriks ketidakcocokan diskrit (*Hamming distance*). Guna menstandarkan format probabilitas kontinu pada sentroid K-Means akibat transformasi *One-Hot Encoding*, sistem mengimplementasikan tahap binarisasi sentroid menggunakan fungsi *argmax* untuk mengembalikan nilai desimal menjadi representasi absolut kategorikal aslinya. Melalui tahapan kalibrasi silang ini, partisi klaster final dari masing-masing algoritma diekstraksi dan dievaluasi ulang secara seragam. Kinerja algoritma diukur secara objektif berdasarkan kohesi internal (*Silhouette Hamming*), rasio kepadatan klaster (*BetaCV Hamming*), lebar batas pemisah terluar (*Dunn Hamming*), serta tingkat kekebalan inisialisasi (*Stability* ARI). Rangkuman hasil komparasi performa pada ruang evaluasi terstandardisasi tersebut disajikan secara komprehensif pada Tabel 8.

Perlu ditekankan bahwa metrik komparasi pada Tabel 8 merupakan hasil ekstraksi dinamis di dalam ruang evaluasi setara berbasis proksimasi *Hamming* yang secara empiris mematahkan asumsi konvensional mengenai superioritas algoritma natif kategorikal. Berlawanan dengan ekspektasi umum, algoritma K-Means (K=10) yang ditransformasikan melalui *One-Hot Encoding* justru mendominasi seluruh kriteria evaluasi dengan capaian homogenitas *Silhouette* tertinggi (0.2482), kepadatan partisi *BetaCV* terkuat (1.8705), margin pemisah *Dunn Index* terlebar (0.1667), serta kekebalan absolut terhadap *local optima* melalui skor stabilitas ARI 0.5010. Meskipun pergeseran K-Medoids menuju ruang dimensi tinggi (K=10) berhasil mengantarkannya sebagai kandidat terkuat kedua yang mampu mengungguli K-Modes pada tingkat kohesi, pencapaian tersebut belum memadai untuk meruntuhkan kekokohan partisi struktur K-Means. Superioritas mutlak K-Means ini kemudian dikuantifikasi secara definitif melalui kerangka *Multi-Criteria Decision Making* (MCDM) yang model tersebut menyapu bersih poin *Borda Score*

(12) dan membukukan normalisasi Z-*Score Composite* bernilai positif ekstrem (1.3753). Secara teoritis, dominasi K-Means pada *dataset* ini didorong oleh kemampuannya memetakan kombinasi atribut kategorikal yang kompleks ke dalam ruang fitur biner ortogonal secara jauh lebih berbutir (granular), melampaui kapabilitas algoritma pesaingnya yang terlalu restriktif akibat pencarian murni berbasis modus atau letak observasi aktual. Oleh karena itu, arsitektur K-Means (K=10) secara final dikukuhkan sebagai pemodelan *Best Algorithm* yang paling tangguh untuk mengekstraksi segmentasi karakteristik mahasiswa pada tahap perumusan strategi promosi lanjutan.

Profiling Karakteristik dan Implikasi Manajerial (Strategi Promosi)

Penerapan algoritma K-Means dengan konfigurasi K=10 terbukti mampu mengekstraksi varians data secara *highly*-granular (sangat berbutir) pada ruang multidimensi, namun dalam konteks aksiabilitas praktis manajemen perguruan tinggi, terdapat batasan efisiensi operasional (*operational friction*) yang harus diperhitungkan. Mengelola dan mengeksekusi sepuluh kampanye pemasaran yang berbeda secara simultan di lapangan akan memicu fragmentasi fokus, inefisiensi pengerahan sumber daya manusia, serta lonjakan biaya akuisisi mahasiswa (*Customer Acquisition Cost*/CAC). Oleh karena itu, meskipun K=10 adalah arsitektur murni yang paling optimal secara matematis, kesepuluh klaster tersebut perlu diagregasikan kembali menjadi tiga Segmen Makro strategis agar kebijakan bauran pemasaran (*marketing mix*) institusi dapat dieksekusi secara logis dan terfokus.

Guna menjembatani dikotomi antara ketepatan perhitungan ruang data dan efisiensi implementasi bisnis, penelitian ini menerapkan pendekatan segmentasi dua lapis (*Two-Tier Segmentation*). Lapis pertama diselesaikan melalui komputasi murni K-Means (K=10) untuk menjamin purifikasi pemisahan data dengan tingkat bias sekecil mungkin. Lapis kedua dilakukan melalui sintesis reduksi makro dengan mengagregasikan kesepuluh klaster tersebut ke dalam satu Matriks Master Strategis. Agregasi ini tidak dilakukan secara heuristik subjektif, melainkan dibantu oleh analisis proksimitas menggunakan *Hierarchical Agglomerative Clustering* (HAC) dengan tautan rata-rata (*average linkage*) terhadap sepuluh sentroid klaster final. Proses ini secara algoritmik mengelompokkan klaster-klaster murni yang memiliki proksimitas kriteria esensial (ekonomi, geografi dan saluran informasi dominan) ke dalam tiga Segmen Makro agar kebijakan bauran pemasaran (*marketing mix*) institusi dapat dieksekusi secara logis dan terfokus. Matriks integrasi karakteristik, penjangkauan pasar, serta mitigasi risikonya disajikan secara utuh pada Tabel 9.

Tabel 9. Matriks Master Sintesis Karakteristik, Strategi Promosi dan Hipotesis Tantangan Lintas Segmen

Segmen Makro & Proporsi Populasi	Klaster	Profil Karakteristik Dominan	Metode Promosi yang Diusulkan	Hipotesis Tantangan Berbasis Profil
Segmen 1: Jaringan Kekerabatan & Regional *(25,47% Populasi)*	0, 5, 9	**Geografi:** Berpusat di daerah penyangga (Banten, Jawa Barat). **Finansial:** Menengah (Rp4 - Rp7 Juta). **Perilaku:** Dipengaruhi sirkel sosial (Keluarga & Teman).	**Alumni & *Student Referral*:** Insentif bagi mahasiswa yang merekrut pendaftar dari sekolah asal. **MGBK (Musyawarah Guru Bimbingan dan Konseling) Regional *Partnership*:** Kemitraan B2B melalui *workshop* guru.	**Ketergantungan Rujukan:** Karakteristik komunal mengindikasikan tingginya sensitivitas terhadap *Word-of-Mouth*. Penurunan kualitas layanan berisiko menyebar cepat sebagai sentimen negatif di wilayah ini.
Segmen 2: *Urban Premium* & *Digital-Savvy* *(25,00% Populasi)*	2, 3	**Geografi:** Urban metropolitan (DKI Jakarta). **Finansial:** Kelas atas (≥ Rp7 Juta). **Perilaku:** Literasi teknologi tinggi, pencarian via Internet, asal SMA Swasta unggulan.	***Targeted Omnichannel Ads*:** Iklan digital berbayar radius urban. ***Prestige Value Proposition*:** Fokus pada lab modern, akreditasi, & prospek kerja. ***Exclusive Open House*:** *Campus tour* interaktif.	**Persaingan Atensi Tinggi:** Tingginya literasi digital dan domisili urban mengindikasikan segmen ini memiliki banyak alternatif pilihan kampus, sehingga menuntut waktu respon admisi yang sangat cepat.

Segmen Makro & Proporsi Populasi	Klaster	Profil Karakteristik Dominan	Metode Promosi yang Diusulkan	Hipotesis Tantangan Berbasis Profil
Segmen 3: Vokasi & Sensitif Biaya *(49,53% Populasi)*	1, 4, 6, 7, 8	**Geografi:** Urban pinggiran dan kawasan industri. **Finansial:** Prasejahtera (< Rp4 Juta). **Perilaku:** Lulusan SMK/Vokasi. Pasif, bergantung pada Brosur dan anjuran Guru Sekolah.	***Affordability Campaign*:** Fokus pada beasiswa (KIP-K), opsi cicilan tanpa bunga, dan kelas karyawan. ***Direct School Roadshow*:** Penetrasi presentasi tatap muka ke SMK/vokasi target.	**Sensitivitas Pembiayaan:** Mengingat profil pendapatan di bawah Rp 4 juta, kelompok ini dihipotesiskan sangat rentan terhadap fluktuasi ekonomi mikro keluarga yang berpotensi memicu kendala pelunasan biaya awal.

Berdasarkan sintesis matriks pada Tabel 9, dapat diidentifikasi asimetri yang jelas antara karakteristik demografis dengan perilaku pencarian informasi pendaftar. Anomali ini menegaskan urgensi adopsi bauran pemasaran (*marketing mix*) yang heterogen guna mencapai optimalisasi *Customer Acquisition Cost* (CAC). Sebagai turunan taktis dari matriks tersebut, pilar analisis manajerial pertama menyoroti dinamika *Word-of-Mouth* pada segmen 1 (Jaringan Regional) yang mencakup 25,47% populasi. Pada wilayah penyangga seperti Banten dan Jawa Barat, *brand penetration* diindikasikan tidak berjalan seefektif melalui sirkel kekerabatan seperti "Keluarga" dan "Teman". Fenomena ini menunjukkan bahwa keputusan pendidikan di daerah tersebut masih bersifat sangat komunal. Oleh karena itu, institusi dapat mengoptimalkan anggaran iklan berbayar di wilayah ini dengan mengalihkannya pada program insentif rujukan organik. Meskipun demikian, institusi dihipotesiskan perlu mewaspadai ketergantungan rujukan ini, dalam komunitas yang sangat mengandalkan *Word-of-Mouth*, kualitas layanan akademik kampus yang kurang optimal berpotensi memicu sentimen negatif yang dapat memengaruhi arus pendaftar secara organik.

Berbeda halnya dengan Segmen 2 (*Urban Premium*) yang secara kuantitas hanya mencakup 25,00% populasi, namun memiliki daya beli tertinggi (≥ Rp7 Juta). Literasi teknologi yang tinggi pada kelompok ini menjadikan mereka target yang relevan untuk akuisisi digital melalui *Omnichannel Ads*. Kampanye pada segmen ini direkomendasikan untuk tidak menggunakan narasi "keterjangkauan biaya", melainkan harus berfokus pada penciptaan *Premium Value Proposition*. Tantangan hipotesis utama yang teridentifikasi dalam matriks untuk segmen ini adalah tingginya persaingan atensi. Mengingat sengitnya kompetisi antar-perguruan tinggi di ibu kota, kecepatan respon dari unit admisi kampus diestimasikan menjadi kunci utama agar preferensi kandidat mahasiswa tidak bergeser ke institusi kompetitor.

Sementara itu, temuan paling krusial terletak pada Segmen 3 (Mayoritas Pasar Vokasi) yang mencakup hampir separuh populasi, yakni 49,53%. Segmen ini didominasi oleh kelompok ekonomi berpenghasilan < Rp4 Juta dari sekolah vokasi yang perilakunya sangat bergantung pada Brosur dan anjuran Guru. Karakteristik yang pasif ini membutuhkan pendekatan jemput bola (*direct selling* / *roadshow*) karena mereka bukan tipe pendaftar yang memvalidasi institusi murni melalui pencarian internet mandiri. Strategi *Cost-Leadership* melalui kampanye beasiswa dan cicilan fleksibel adalah instrumen krusial untuk segmen ini, namun sensitivitas pembiayaan segmen ini patut diantisipasi dengan cermat. Institusi perlu mempersiapkan strategi pengelolaan arus kas (*Cash Flow*) guna merespons tingginya kebutuhan cicilan dan potensi kendala pelunasan awal akibat fluktuasi ekonomi mikro keluarga calon mahasiswa.

Pada konklusi manajerial, implementasi kerangka kerja pemodelan data terintegrasi ini secara radikal mampu mentransformasikan tata kelola alokasi modal pemasaran institusi. Dengan peta presisi dari Tabel 9, kebocoran anggaran promosi akibat salah sasaran (*budget hemorrhage*) dapat ditekan ke titik terendah. Dana kampanye *digital premium* dapat dialokasikan eksklusif untuk mengejar Segmen 2, program insentif rujukan dikunci untuk menguasai Segmen 1, sementara energi operasional *roadshow* dan subsidi beasiswa dikonsentrasikan secara penuh untuk mengamankan volume populasi di Segmen 3. Secara keseluruhan, strategi ini mengubah arah kebijakan institusi dari asumsi intuitif pemasaran massal menuju pengambilan keputusan strategis yang murni dikendalikan oleh wawasan data empiris (*data-driven decision making*).

Keterbatasan Penelitian

Meskipun kerangka evaluasi *robust clustering* berbasis *multiple runs* dan *Fair Evaluation Space* terbukti mampu menghasilkan partisi yang objektif, penelitian ini masih memiliki dua keterbatasan fundamental. Pertama, dataset pendaftaran bersifat eksklusif pada satu institusi (*institution-specific*), sehingga karakteristik *student persona* yang terbentuk rentan terhadap bias demografi lokal dan tidak dapat digeneralisasi secara langsung pada perguruan tinggi lain. Kedua, pemodelan ini murni mengandalkan atribut sosio-demografi pendaftar, tanpa mengintegrasikan variabel analitik pemasaran aktual secara *real-time* seperti *Return on Investment* (ROI) maupun *Cost per Click* (CPC). Ketiadaan metrik finansial tersebut menyebabkan rekomendasi strategi promosi yang diusulkan masih bersifat preskriptif-konseptual dan belum mampu memetakan optimasi efisiensi anggaran secara kuantitatif. Guna menjembatani kelemahan tersebut, penelitian selanjutnya sangat disarankan untuk menguji kerangka *Multi-Criteria Decision Making* (MCDM) menggunakan *dataset* lintas institusi (*multi-institutional*) untuk memvalidasi ketangguhan generalisasi eksternalnya. Selain itu, pengintegrasian data metrik biaya pemasaran digital juga menjadi sangat krusial untuk dikembangkan guna membangun pemodelan prediktif berbasis *Customer Lifetime Value* (CLV) di sektor pendidikan tinggi.

KESIMPULAN

Penelitian ini berhasil menjembatani celah metodologis komparasi algoritma klasterisasi melalui eksekusi 30 iterasi independen guna mengeliminasi bias inisialisasi awal. Evaluasi dieksekusi secara berkeadilan pada "Ruang Evaluasi Setara" menggunakan kerangka *Multi-Criteria Decision Making* (MCDM) berbasis *Borda Score* dan *Z-Score Composite*. Hasilnya, K-Means (K=10) ditetapkan secara absolut sebagai pemodelan final karena mendominasi ruang metrik *Hamming* dengan kepadatan klaster terkuat (*BetaCV* = 1.8705) dan kekebalan inisialisasi tertinggi (ARI = 0.5010).

Guna memitigasi inefisiensi operasional, kesepuluh klaster tersebut direduksi melalui Segmentasi Dua Lapis (*Two-Tier Segmentation*) menjadi tiga segmen makro strategis. Segmen 1 (Regional, 25,47%) diintervensi melalui rujukan alumni dan kemitraan MGBK. Segmen 2 (*Urban Premium*, 25,00%) yang berdaya beli tinggi ditargetkan melalui kampanye *Omnichannel Ads* bernilai prestise. Segmen 3 (Vokasi & Sensitif Biaya, 49,53%) mewajibkan pendekatan proaktif melalui *roadshow* sekolah dan opsi beasiswa fleksibel.

Meskipun komprehensif, riset ini terbatas pada dataset spesifik satu institusi dan belum melibatkan metrik biaya pemasaran riil (seperti CPC). Ke depan, riset lanjutan sangat disarankan untuk mengintegrasikan metrik finansial *real-time* dan menguji kerangka ini pada dataset multi-institusi guna melahirkan pemodelan prediktif berbasis *Customer Lifetime Value* (CLV) yang universal.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Unit Pelaksana Teknis Promosi dan Penerimaan Mahasiswa Baru Universitas Al-Azhar Indonesia atas dukungannya dalam penyediaan data mentah biodata mahasiswa baru selama kurun waktu enam tahun akademik. Ketersediaan dan kelengkapan data tersebut menjadi landasan utama yang memungkinkan terlaksananya proses analisis dan penyelesaian penelitian ini.

REFERENSI

[1] Z. P. C. M. Napitupulu, J. J. Siang, and A. G. Prabawati, “Segmentasi Wilayah Target Promosi Menggunakan Algortima K-Means Clustering pada Perguruan Tinggi,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 4, pp. 1160–1171, Oct. 2025, doi: 10.57152/malcom.v5i4.2125.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining. Concepts and Techniques*, Third Edition. Morgan Kaufmann Publishers, 2012.

[3] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, Second Edition. Pearson, 2019.

[4] N. A. Rahmalinda and A. Jananto, “Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru,” *Jurnal TEKNO KOMPAK*, vol. 16, no. 2, pp. 163–175, Aug. 2022. DOI:10.33365/jtk.v16i2.1971.

[5] Herlina, S. Mundari, E. A. Putri, and A. D. Rahmawati, “Penerapan K-Means Clustering Untuk Strategi Promosi Untuk Meningkatkan Jumlah Perolehan Mahasiswa Baru,” *JIME (Journal of Industrial and Manufacture Engineering)*, vol. 8, no. 1, pp. 98–103, May 2024, doi: 10.31289/jime.v8i1.10928.

[6] I. Irawan, U. Rizki, P. M. Jakak, M. B. Prayogi, and M. Rahman, "Penerapan Metode K-Means Clustering Dalam Pengembangan Strategi Promosi Berbasis Data Penerimaan Mahasiswa Baru (Studi Kasus :Universitas Nurul Huda)," *Jurnal Nasional Ilmu Komputer*, vol. 5, no. 1, pp. 2746–1343, Feb. 2024.

[7] I. M. Widiarta, M. N. Fietroh, and A. Ibrahim, "Analisis Strategi Promosi Mahasiswa Baru Universitas Teknologi Sumbawa Menggunakan Data Mining," *Jurnal Ilmu Sosial dan Pendidikan (JISIP)*, vol. 6, no. 2, pp. 2598–9944, Mar. 2022, doi: 10.36312/jisip.v6i1.3037.

[8] M. Farozi, "Metode K-Means Clustering Dalam Merancang Strategi Promosi Penerimaan Mahasiswa Baru Pada STIE Serelo Lahat," *Jurnal Ilmiah Informatika Global*, vol. 12, no. 2, Dec. 2021. DOI: https://doi.org/10.36982/jiig.v12i2.2191.

[9] A. Wirawan and D. Prasetyawan, "Analisis cluster data latar belakang ekonomi mahasiswa untuk rekomendasi penentuan uang kuliah tunggal dengan model K-Modes," *INFOTECH : Jurnal Informatika & Teknologi*, vol. 4, no. 2, pp. 234–246, Dec. 2023, doi: 10.37373/infotech.v4i2.898.

[10] T. Syafrudin, A. Hermawan, D. Avianto, and I. Maulana, "Analisis Komparasi Algoritma K-Means Dan K-Medoids Dalam Segmentasi Data Untuk Strategi Promosi Mahasiswa Baru Di Universitas X," *Komputika : Jurnal Sistem Komputer*, vol. 14, no. 2, pp. 203–211, Nov. 2025, doi: 10.34010/komputika.v14i2.16698.

[11] Fathurrahman, "Evaluasi Clustering K-Means Dan K-Medoid Pada Persebaran COVID-19 Di Indonesia Dengan Metode Davies Bouldin Index (DBI)," Universitas Islam Negri Maulana Malik Ibrahim, MALANG, 2023.

[12] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[13] A. Setiawan and Saprudin, "Implementasi Sistem Pendeteksian Kemiripan Judul Skripsi Dengan Algoritma Levenshtein Distance Pada Perpustakaan Universitas Pamulang," *LOGIC: Jurnal Ilmu Komputer dan Pendidikan*, vol. 2, no. 1, p. 1, 2023, [Online]. https://www.journal.mediapublikasi.id/index.php/logic/article/view/2638.

[14] U. Linarti, A. Rahmawati, A. H. S. Jones, and L. Zahrotun, "Penerapan Metode K-Medoids Guna Pengelompokan Data Usaha Mikro, Kecil dan Menengah (UMKM) Bidang Kuliner Di Kota Yogyakarta," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 37–45, Jan. 2024. DOI: https://doi.org/10.55338/jikomsi.v7i1.2194.

[15] S. Isnanto and S. Widodo, "Penerapan Data Mining Pada Penerimaan Mahasiswa Baru Dengan Algoritma K-Means Clustering," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 4, no. 2, pp. 158–167, Dec. 2021, doi: 10.37600/tekinkom.v4i2.367.

[16] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'kneedle' in a haystack: Detecting Knee Points in System Behavior," in *Proceedings - International Conference on Distributed Computing Systems*, 2011, pp. 166–171. doi: 10.1109/ICDCSW.2011.20.

[17] M. Guntara and N. Lutfi, "Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method," *JuTI "Jurnal Teknologi Informasi,"* vol. 2, no. 1, pp. 43–52, Aug. 2023, doi: 10.26798/juti.v2i1.944.

[18] N. Fajriati and B. Prasetiyo, "Optimasi Algoritma Naive Bayes dengan Diskritisasi K-Means pada Diagnosis Penyakit Jantung," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 10, no. 3, pp. 503–512, Jun. 2023, doi: 10.25126/jtiik.2023106510.

[19] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow Pada Data Pengguna Narkoba Dengan Pemrograman Python," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 5, no. 1, pp. 113–119, Jan. 2021. DOI:10.59697/jtik.v5i1.593.

[20] Stack Exchange, "Identifying k-means clusters number programmaticaly. Cross Validated." Accessed: Jun. 30, 2026. [Online]. Available: https://stats.stackexchange.com/questions/386777/identifying-k-means-clusters-number-programmaticaly

[21] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Royal Statistical Society*, pp. 411–423, 2001, doi: 10.1111/1467-9868.00293.

[22] R. Hidayati, A. Zubair, A. Hidayat Pratama, L. Indana, P. Studi Sistem Informasi, and F. Teknologi Informasi, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering," 2021. DOI: https://doi.org/10.33633/tc.v20i2.4556.

[23] S. N. Priansyah and D. U. Wutsqa, "Penerapan metode ensemble clustering pada pengelompokan kabupaten dan kota_2024," *Jurnal Statistika dan Sains Data*, vol. 2, no. 1, pp. 1–14, Oct. 2024. DOI:10.21831/jssd.v2i1.19804.

[24] D. Sugianti, H. W. Hapsoro, and W. Setianto, "Penerapan Machine Learning Untuk Penentuan Segmentasi Mahasiswa Baru Dengan Metode K-Modes," *IC-Tech*, vol. 17, no. 1, pp. 1–9, Apr. 2022, [Online]. DOI: https://doi.org/10.47775/ictech.v17i1.227.

[25] M. Sholeh and K. Aeni, "Perbandingan Evaluasi Metode Davies Bouldin, Elbow Dan Silhouette Pada Model Clustering Dengan Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 8, no. 1, pp. 56–65, Aug. 2023, [Online]. DOI: http://dx.doi.org/10.30998/string.v8i1.16388.

[26] S. Z. N. Al Hasani, Q. N. Salsabela, and E. Z. Permata Sari, "Implementasi Metode TOPSIS dan Borda Count dalam Penentuan RTLH (Rumah Tidak Layak Huni) (Studi Kasus: Kampung Inggris Pare, Kediri)," *Jurnal Matematika Integratif*, vol. 20, no. 1, pp. 35–46, Jun. 2024, doi: 10.24198/jmi.v20.n1.52227.35-46.

[27] A. F. H. Marsandy, M. N. Hayati, and M. Fauziyah, "Klasterisasi Prevalensi Stunting Menggunakan K-Prototype pada Data Campuran," *METIK JURNAL*, vol. 8, no. 2, pp. 48–54, Dec. 2024, doi: 10.47002/metik.v8i2.824.