Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 2036-2047
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Student Profiling Based on Academic Achievement and Socioeconomic Status Using K-Means Clustering Method

Profiling Siswa SMA/SMK Berdasarkan Nilai Akademik dan Status Sosial-Ekonomi Menggunakan Metode K-Means

Urip Yogantara[1*], I Gede Aris Gunadi[2], I Made Gede Sunarya[3]

[1,2,3]Program Studi Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha, Indonesia

E-Mail: [1]uripyogantara@gmail.com, [2]igedearisgunadi@undiksha.ac.id, [3]sunarya@undiksha.ac.id

Received Jun 05th 2026; Revised Jul 06th 2026; Accepted Sep 17th 2026; Available Online Oct 30th 2026
Corresponding Author: Urip Yogantara
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The diversity of academic abilities and socioeconomic backgrounds among senior high school (SMA/SMK) students demands an evidence-based differentiated learning approach. However, comprehensive student profiling that integrates academic and socioeconomic data has rarely been conducted. This study aims to profile SMA/SMK students by integrating academic score data and socioeconomic status data using the K-Means Clustering method. The dataset consisted of 45,244 student records with 11 variables and was processed using the Knowledge Discovery in Databases (KDD) framework. Feature selection used two main variables: nilai_total (the sum of scores across three subjects) and desil_kesejahteraan (welfare decile, 1–10). After removing missing data, 42,722 valid records were obtained. Data standardization was performed using StandardScaler. The optimal number of clusters (K=4) was determined through the Elbow Method with KneeLocator. The evaluation results showed a Silhouette Score of 0.3658 and a Davies-Bouldin Index (DBI) of 0.8589. Four clusters were formed: Cluster 0 (Prosperous High-Achieving Profile, n=9,870), Cluster 1 (Vulnerable Profile, n=10,896), Cluster 2 (Developing Prosperous Profile, n=12,454), and Cluster 3 (Limited-Potential Profile, n=9,502). These findings provide an empirical basis for designing well-targeted differentiated learning interventions in line with the Merdeka Curriculum policy.

Keywords: Academic Score, K-Means Clustering, Merdeka Curriculum, Socioeconomic Status, Student Profiling

Abstrak

Keberagaman kemampuan akademik dan latar belakang sosial-ekonomi siswa SMA/SMK menuntut pendekatan pembelajaran berdiferensiasi yang berbasis bukti. Namun, *profiling* siswa secara komprehensif yang mengintegrasikan data akademik dan sosial-ekonomi secara simultan belum banyak dilakukan. Penelitian ini bertujuan memetakan profil siswa SMA/SMK dengan mengintegrasikan data nilai akademik dan status sosial ekonomi menggunakan metode *K-Means Clustering*. Dataset terdiri dari 45.244 data siswa dengan 11 variabel, yang diproses melalui kerangka *Knowledge Discovery in Databases* (KDD). Seleksi fitur menggunakan dua variabel utama: nilai_total (jumlah nilai tiga mata pelajaran) dan desil_kesejahteraan (1–10). Setelah penghapusan data yang hilang, diperoleh 42.722 data valid. Standardisasi data dilakukan menggunakan *StandardScaler*. Jumlah klaster optimal (K=4) ditentukan menggunakan *Elbow Method* dengan *KneeLocator*. Hasil evaluasi menunjukkan *Silhouette Score* sebesar 0,3658 dan *Davies-Bouldin* Index (DBI) sebesar 0,8589. Empat klaster yang terbentuk: Klaster 0 (Profil Unggul Sejahtera, n=9.870), Klaster 1 (Profil Rentan, n=10.896), Klaster 2 (Profil Sejahtera Berkembang, n=12.454), dan Klaster 3 (Profil Potensial Terbatas, n=9.502). Temuan ini memberikan landasan empiris bagi perancangan intervensi pembelajaran berdiferensiasi yang tepat sasaran sesuai kebijakan Kurikulum Merdeka.

Kata Kunci: *K-Means Clustering*, Kurikulum Merdeka, Nilai Akademik, *Profiling* Siswa, Status Sosial-Ekonomi

1. PENDAHULUAN

Kebijakan Kurikulum Merdeka melalui penggantian Ujian Nasional (UN) dengan Asesmen Nasional (AN) pada tahun 2021 memprioritaskan kompetensi literasi, numerasi, dan karakter [1]. AN mengukur hasil belajar kognitif melalui Asesmen Kompetensi Minimum (AKM) serta hasil non-kognitif melalui survei karakter dan lingkungan belajar [2]. Meskipun demikian, Driana dkk. [3] mengungkapkan bahwa pemanfaatan

hasil asesmen oleh guru masih menjadi tantangan. Wargina dkk. [4] dalam evaluasinya terhadap implementasi AN di SMP Negeri Baturiti menggunakan model *Content, Input, Process, Product* (CIPP) menunjukkan bahwa meskipun AN berjalan efektif pada komponen konteks dan proses, masih diperlukan penguatan pada komponen *input* dan produk, khususnya dalam pemanfaatan hasil asesmen untuk perbaikan pembelajaran. Kebutuhan akan pemanfaatan data ini dilatarbelakangi oleh persistennya disparitas mutu hasil belajar. Data hasil UN jenjang SMA/MA tahun 2018 dan 2019 yang dipublikasikan Pusat Penilaian Pendidikan menunjukkan rerata nilai UN nasional yang cenderung menurun serta variasi capaian antarprovinsi yang lebar, dengan mata pelajaran Matematika secara konsisten menjadi yang terendah [5]. Analisis klaster terhadap hasil UN 2016–2019 oleh Janu dkk. [6] bahkan menegaskan masih terjadinya disparitas mutu luaran pendidikan antarprovinsi, khususnya antara wilayah Indonesia bagian barat dan timur. Kondisi ini diperkuat oleh capaian PISA 2022 yang menempatkan Indonesia pada peringkat ke-71 dari 81 negara pada domain membaca, matematika, dan sains, yang mengindikasikan rendahnya literasi dan numerasi peserta didik secara nasional [7]. Fakta empirik inilah yang menjadikan pemetaan profil siswa secara berbasis data menjadi kebutuhan mendesak.

Paradoks kebijakan AN terletak pada kesenjangan antara tujuan pemetaan mutu pendidikan dengan keterbatasan pemahaman terhadap profil peserta didik. Pemerintah telah mengumpulkan data akademik dan data sosial-ekonomi, namun analisis integratif kedua sumber data masih sangat terbatas. Azzizah [8] menegaskan bahwa ketimpangan sosial-ekonomi antarwilayah Indonesia tercermin dalam ketimpangan pendidikan. Saputra dkk. [9] membuktikan potensi data terpadu sosial-ekonomi sebagai basis analisis klasterisasi yang valid melalui perbandingan *Fuzzy C-Means* dan *Naïve Bayes* untuk penentuan keluarga penerima manfaat di Kabupaten Buleleng. Sejumlah penelitian terdahulu telah membuktikan efektivitas analisis klaster untuk memetakan kondisi sosial-ekonomi dan akademik. Mohamed Nafuri dkk. [10] menerapkan K-Means teroptimasi untuk mengelompokkan mahasiswa kategori B40 (kelompok berpendapatan rendah) berdasarkan kinerja akademik dan menghasilkan lima klaster yang membantu pemerintah menekan angka putus kuliah. Pada konteks pemetaan wilayah, Suraya dan Wijayanto [11] membandingkan Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means untuk pengelompokan provinsi di Indonesia dan menemukan performa K-Means yang kompetitif dengan interpretabilitas yang tinggi. Selain itu, Liu [12] yang menerapkan K-Means pada data capaian belajar siswa berbasis faktor sosial-ekonomi menegaskan bahwa klasterisasi mampu mengungkap profil siswa tersembunyi yang tidak tampak melalui analisis konvensional. Temuan-temuan tersebut memperkuat relevansi pendekatan klasterisasi bagi integrasi data akademik dan sosial-ekonomi dalam penelitian ini.

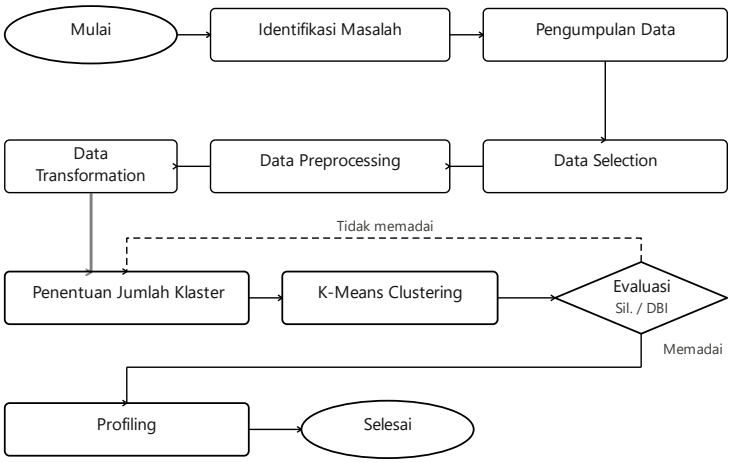
Analisis klaster menjadi pendekatan relevan untuk mengelompokkan siswa berdasarkan kemiripan karakteristik sosial-akademik. Alan & Temiz [13] mendemonstrasikan efektivitas data *mining* dalam memprofilkan siswa berdasarkan kondisi sosial-ekonomi keluarga. Dalam konteks pendidikan, Gunadi dkk. [14] menerapkan *K-Means* dan *K-Harmonic Means* untuk pengelompokan siswa di SPN Singaraja, sedangkan Ariasa dkk. [15] mengoptimasi *K-Means* dinamis untuk pengelompokan kinerja akademik mahasiswa Undiksha dengan DBI 0,168. Cahyaningsih dkk. [16] menggunakan *K-Means* untuk mengidentifikasi pola kepribadian siswa SLB (DBI=0,9095), dan Widiantara dkk. [17] mengklasifikasikan siswa berpotensi *dropout* menggunakan *Learning Vector Quantization*. Selain itu, Ananda dkk. [18] mengklasifikasikan gaya belajar siswa SD menggunakan berbagai algoritma *machine learning* dengan akurasi terbaik 95%, menunjukkan potensi besar data *mining* untuk mendukung intervensi pendidikan yang tepat sasaran. Pemilihan pendekatan klasterisasi K-Means didasarkan pada karakteristik data dan tujuan penelitian. Data integrasi nilai akademik dan status sosial-ekonomi tidak memiliki label kelas atau kategori profil yang telah ditentukan sebelumnya (ground truth), sehingga persoalan ini bersifat unsupervised learning. Algoritma klasifikasi mensyaratkan data berlabel untuk proses pelatihan, sedangkan tujuan penelitian ini justru menemukan pola pengelompokan alami yang belum diketahui, bukan memprediksi label yang sudah ada. K-Means dipilih karena sifatnya yang efisien secara komputasi untuk data berskala besar (lebih dari 42 ribu record), mudah diinterpretasikan, dan sesuai dengan asumsi jarak Euclidean pada ruang fitur numerik dua dimensi. Dengan demikian, penggunaan klasterisasi selaras dengan hakikat masalah, yaitu membangun profil siswa dari data tak berlabel sebagai basis intervensi.

Penelitian ini bertujuan untuk menerapkan *K-Means Clustering* dalam memetakan profil sosial-akademik siswa SMA/SMK di berdasarkan integrasi nilai akademik dan status sosial-ekonomi dengan menentukan jumlah klaster optimal melalui *Elbow Method*. Hasil pemetaan tersebut selanjutnya digunakan untuk mendeskripsikan karakteristik setiap klaster sebagai landasan empiris bagi perancangan pembelajaran berdiferensiasi yang tepat sasaran. Kebaruan penelitian ini terletak pada tiga aspek yang membedakannya dari penelitian terdahulu. Pertama, berbeda dengan penelitian sebelumnya yang umumnya menganalisis dimensi akademik atau sosial-ekonomi secara terpisah, penelitian ini mengintegrasikan secara simultan data akademik nasional dan data sosial-ekonomi resmi dalam satu kerangka klasterisasi. Kedua, penelitian ini beroperasi pada skala data individu siswa yang besar (42.722 *record* valid), berbeda dari studi terdahulu yang umumnya berbasis data agregat provinsi atau sampel terbatas satu sekolah. Ketiga, hasil klasterisasi tidak berhenti pada pemetaan, melainkan diterjemahkan menjadi profil siswa yang secara langsung menjadi landasan empiris

perancangan pembelajaran berdiferensiasi sesuai amanat Kurikulum Merdeka. Kombinasi integrasi dua sumber data resmi berskala nasional dan orientasi terapan pada pembelajaran berdiferensiasi inilah yang menjadi pembeda utama penelitian ini.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Database* (KDD) sebagai kerangka kerja sistematis dalam proses penggalian pengetahuan dari data siswa. KDD merupakan proses non-trivial untuk mengekstraksi pengetahuan yang valid, baru, dan potensial berguna dari sekumpulan data besar [19]. Dalam penelitian ini, kerangka KDD diterapkan secara berurutan mulai dari identifikasi masalah, pengumpulan data, seleksi fitur, data preprocessing, transformasi data, pemodelan *K-Means*, evaluasi klaster, hingga interpretasi profil klaster. Alur metodologi penelitian secara keseluruhan disajikan pada Gambar 1.



Gambar 1. Metodologi Penelitian

Gambar 1 memperlihatkan alur kerja penelitian yang terdiri atas delapan tahap KDD secara berurutan, mulai dari identifikasi masalah hingga interpretasi profil klaster, di mana keluaran setiap tahap menjadi masukan bagi tahap berikutnya.

2.1. Identifikasi Masalah

Tahap identifikasi masalah merupakan titik awal dari kerangka KDD yang bertujuan menetapkan tujuan dan ruang lingkup penelitian secara eksplisit [8]. Penelitian ini dilatarbelakangi oleh kenyataan bahwa pembelajaran di SMA/SMK belum sepenuhnya menyentuh kebutuhan individual siswa. Kebijakan Kurikulum Merdeka yang memprioritaskan pembelajaran berdiferensiasi mensyaratkan tersedianya profil siswa yang komprehensif mencakup dimensi akademik dan sosial-ekonomi secara integratif. Namun demikian, profil sosial-akademik siswa SMA/SMK yang terintegrasi belum tersedia sebagai basis perumusan intervensi pembelajaran yang tepat sasaran.

Azzizah [8] menegaskan bahwa ketimpangan sosial-ekonomi antarwilayah Indonesia tercermin dalam ketimpangan pendidikan, dengan faktor kemiskinan sebagai determinan paling berpengaruh terhadap angka partisipasi sekolah. Dalam konteks ini, identifikasi masalah penelitian diarahkan pada tiga pertanyaan utama: (1) bagaimana distribusi profil sosial-akademik siswa SMA/SMK berdasarkan nilai akademik dan status sosial-ekonomi; (2) berapa jumlah klaster optimal yang merepresentasikan profil tersebut; dan (3) apa implikasi profil klaster bagi perancangan pembelajaran berdiferensiasi sesuai Kurikulum Merdeka.

2.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari dua basis data berskala nasional yang saling diintegrasikan. Sumber data pertama adalah data nilai akademik siswa yang berasal dari pelaksanaan asesmen berskala nasional yang diselenggarakan oleh kementerian yang membidangi urusan pendidikan. Data ini merepresentasikan kemampuan akademik siswa pada tiga mata pelajaran utama, yang diukur menggunakan instrumen standar tervalidasi secara nasional [3].

Sumber data kedua adalah basis data sosial-ekonomi terpadu tingkat nasional. Basis data ini bersifat multidimensi, memuat informasi sosial, ekonomi, dan demografi rumah tangga pada tingkat individu, disusun berdasarkan pendataan langsung dan diperbarui secara berkala. Penggunaan data SSE dari basis data terpadu pemerintah untuk keperluan analisis klasterisasi telah dilakukan sebelumnya oleh Saputra dkk. [9] dalam penentuan keluarga penerima manfaat bantuan sosial di Kabupaten Buleleng menggunakan data Basis Data Terpadu (BDT).

Kedua sumber data diintegrasikan menggunakan kunci identifikasi unik peserta asesmen yang termuat dalam kedua dataset. Dataset gabungan mencakup 45.244 siswa SMA/SMK, dengan 11 variabel yang meliputi identitas siswa, data sekolah, nilai mata pelajaran, dan desil sosial ekonomi siswa.

Tabel 1. Deskripsi Kolom Dataset

No	Nama Kolom	Tipe Data	Deskripsi
1	nisn	String	Nomor Induk Siswa Nasional
2	nomor_peserta	String	Nomor Peserta yang diberikan saat mengikuti tes
3	nama_siswa	String	Nama lengkap siswa
4	nama_sekolah	String	Nama sekolah tempat siswa terdaftar
5	jenjang	String	Tingkat pendidikan sekolah (SMA atau SMK)
6	status_sekolah	String	Status kepemilikan sekolah (Negeri atau Swasta)
7	nama_kabupaten	String	Nama kabupaten/kota lokasi sekolah berada
8	nilai_bahasa_indonesia	Float	Nilai mata pelajaran Bahasa Indonesia
9	nilai_matematika	Float	Nilai mata pelajaran Matematika
10	nilai_bahasa_inggris	Float	Nilai mata pelajaran Bahasa Inggris
11	desil_kesejahteraan	Integer	Peringkat desil sosial-ekonomi siswa (1–10)

Tabel 1 merangkum kesebelas variabel penyusun *dataset* gabungan beserta tipe data dan deskripsinya, yang meliputi atribut identitas siswa, informasi sekolah, nilai tiga mata pelajaran, dan indikator desil sosial-ekonomi.

Tabel 2. Sampel Data Siswa (dari 45.244 baris)

Nama	Jenjang	Nama_sekolah	Nilai_Bahasa Indonesia	Nilai_ Matematika	Nilai_Bahasa_ Inggris	Desil Kesejahteraan
Siswa A	SMA	SMA N 1 ABC	80,25	72,50	68,30	7
Siswa B	SMA	SMA N 2 XYZ	75,40	69,80	65,20	4
Siswa C	SMK	SMK N 5 DEF	82,10	78,90	71,50	3
…	…	…	…	…	…	…

Tabel 2 menyajikan cuplikan beberapa baris data siswa sebagai ilustrasi struktur dan format dataset yang dianalisis sebelum dilakukan seleksi fitur dan preprocessing data.

2.3. Data *Selection*

Dari 11 variabel yang tersedia dalam *dataset*, dilakukan seleksi fitur untuk memilih variabel yang paling relevan dan representatif bagi tujuan klasterisasi profil sosial-akademik siswa. Seleksi fitur merupakan langkah penting dalam alur KDD untuk memastikan bahwa hanya fitur yang signifikan digunakan dalam proses pemodelan [19]. Widiantara dkk. [17] menunjukkan bahwa pemilihan fitur yang tepat berdasarkan relevansi terhadap tujuan analisis secara signifikan meningkatkan kualitas model klasifikasi maupun klasterisasi.

Dua fitur utama yang dipilih dalam penelitian ini adalah:

1. nilai_total yaitu jumlah nilai tiga mata pelajaran (nilai_bahasa_indonesia + nilai_matematika + nilai_bahasa_inggris), merepresentasikan kemampuan akademik siswa secara keseluruhan dengan rentang 15–231. Penggunaan skor total multi-mata-pelajaran sebagai representasi kemampuan akademik konsisten dengan pendekatan Oyelade dkk. [20] yang menggunakan skor rata-rata multi-kursus untuk klasterisasi kinerja akademik mahasiswa di institusi pendidikan tinggi.
2. desil_kesejahteraan yaitu indeks kesejahteraan pada skala ordinal 1–10 (1 = desil termiskin, 10 = desil terkaya) yang merepresentasikan kondisi sosial-ekonomi rumah tangga siswa. Variabel desil ini merupakan ringkasan dari puluhan indikator sosial-ekonomi multidimensi, menjadikannya representasi yang komprehensif sekaligus efisien untuk klasterisasi [8].

Pemilihan kedua fitur ini didasarkan pada tiga pertimbangan: (a) keterwakilan dimensi akademik (nilai_total) dan dimensi sosial-ekonomi (desil_kesejahteraan) secara simultan; (b) sifat numerik kedua variabel yang sesuai dengan asumsi perhitungan jarak Euclidean pada algoritma K-Means; serta (c) efisiensi dan interpretabilitas model melalui ruang fitur dua dimensi yang memudahkan visualisasi dan interpretasi profil klaster. Variabel identitas (nama, NISN, nomor peserta) dan variabel deskriptif lainnya tidak diikutsertakan dalam proses klasterisasi.

2.4. Data *Preprocessing*

Tahap *preprocessing* bertujuan memastikan kualitas dan konsistensi data sebelum dianalisis. Proses pembersihan data (data *cleaning*) merupakan langkah kritis dalam alur KDD untuk mengeliminasi data yang tidak valid, tidak lengkap, atau tidak konsisten yang dapat mengganggu kualitas hasil klasterisasi [19]. Saputra dkk. [9] menegaskan bahwa proses *cleaning* data merupakan proses analisa kualitas data dengan cara

menghapus data yang tidak lengkap atau memiliki format yang salah dalam basis data guna menghasilkan data berkualitas tinggi.

Langkah-langkah preprocessing yang dilakukan dalam penelitian ini meliputi:

1. Pembuatan fitur nilai_total: dihitung sebagai penjumlahan nilai_bahasa_indonesia + nilai_matematika + nilai_bahasa_inggris untuk setiap siswa, menghasilkan skor agregat yang merepresentasikan kemampuan akademik keseluruhan.
2. Penanganan missing value: baris yang memiliki nilai kosong (*missing value*) pada salah satu dari dua fitur utama dihapus dari *dataset* (*listwise deletion*).

2.5. Data *Transformation*

Setelah proses *preprocessing*, dilakukan transformasi data menggunakan metode standardisasi *z-score* (*Standard Scaler*) untuk menyamakan skala antara kedua fitur sebelum proses klasterisasi. Standardisasi merupakan langkah transformasi yang penting karena algoritma K-Means berbasis jarak Euclidean sangat sensitif terhadap perbedaan skala antarvariabel; tanpa standardisasi, fitur dengan rentang nilai lebih besar akan mendominasi perhitungan jarak dan menghasilkan klaster yang bias [21]. Djun dkk. [22] menerapkan normalisasi data sebagai tahap *preprocessing* ketiga dalam pipeline klasterisasi K-Means berbasis RFM, menegaskan pentingnya kesetaraan skala untuk hasil klasterisasi yang valid.

Transformasi *StandardScaler* menggunakan rumus *z-score* normalization ditunjukkan pada Persamaan 1.

$$z = (x - \mu)/\sigma \tag{1}$$

di mana x adalah nilai data asli, μ adalah rata-rata fitur, dan σ adalah standar deviasi fitur.

2.6. Algoritma *K-Means*

K-Means adalah algoritma *unsupervised learning* yang mengelompokkan data ke dalam K klaster berdasarkan kesamaan jarak terhadap *centroid* [23]. Algoritma bekerja secara iteratif melalui empat tahap: (1) inisialisasi K centroid secara acak; (2) penugasan setiap titik data ke centroid terdekat berdasarkan jarak minimum; (3) pembaruan centroid sebagai rata-rata nilai seluruh anggota klaster; dan (4) pengulangan langkah 2–3 hingga centroid tidak berubah (konvergen). Algoritma meminimalkan nilai *Within-Cluster Sum of Squares* (WCSS) atau *inertia*, ditunjukkan pada Persamaan 2.

$$WCSS = \sum_{j=1}^{K} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2 \tag{2}$$

di mana x_i adalah titik data ke-i, μ_j adalah centroid klaster ke-j, dan C_j adalah himpunan data pada klaster ke-j.

Jarak antara setiap titik data dengan *centroid* dihitung menggunakan *Euclidean Distance* ditunjukkan pada Persamaan 3.

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - c_{jk})^2} \tag{3}$$

Pemilihan *Euclidean Distance* didasarkan pada hasil studi Bimasena dkk. [24] yang membandingkan tiga metrik jarak (*Euclidean, Manhattan, Chebyshev*) pada *K-Means* dan membuktikan bahwa *Euclidean Distance* menghasilkan nilai DBI dan *Average Within Centroid Distance* terkecil, sehingga paling optimal untuk data numerik terstandarisasi. Suwirya dkk. [25] juga menerapkan *K-Means* berbasis *Euclidean Distance* untuk klasterisasi data layanan dengan hasil klasifikasi yang tervalidasi.

Penelitian ini secara sengaja memfokuskan pada algoritma *K-Means* sebagai metode tunggal berdasarkan beberapa pertimbangan metodologis. Pertama, tujuan penelitian bersifat eksploratif-deskriptif untuk memetakan profil siswa pada ruang fitur dua dimensi yang bersifat numerik dan terstandarisasi, kondisi yang secara teoretis paling sesuai dengan asumsi *K-Means* berbasis jarak *Euclidean*. Kedua, dengan ukuran data yang besar (42.722 *record*), *K-Means* menawarkan kompleksitas komputasi yang linear dan efisien dibanding metode berbasis kepadatan atau hierarki seperti DBSCAN dan *Hierarchical Clustering* yang berbiaya komputasi tinggi pada data besar. Ketiga, sejumlah studi pembanding menunjukkan *K-Means* menghasilkan partisi yang setara atau lebih baik untuk data numerik terstandarisasi. Farahdinna dkk. [26] menemukan *K-Means* mengungguli *K-Medoids* pada klasterisasi produk, sementara Budiman dkk. [27] menemukan *K-Means* memberikan nilai *Silhouette* lebih tinggi dibanding DBSCAN untuk data dengan klaster yang relatif seragam. Keempat, validitas hasil klaster tetap dijamin secara objektif melalui dua metrik evaluasi

internal, yaitu *Silhouette Score* dan *Davies-Bouldin Index*, sehingga kualitas klaster dapat dipertanggungjawabkan tanpa harus membandingkan beberapa algoritma.

2.7. Penentuan Jumlah Klaster Optimal

Penentuan nilai K optimal dalam penelitian ini menggunakan *Elbow Method* berbasis nilai WCSS (*Within-Cluster Sum of Squares*), yaitu jumlah kuadrat jarak antara setiap titik data dengan *centroid* klasternya [28]. Nilai K optimal ditentukan pada titik *elbow* kurva WCSS, yakni posisi di mana penambahan jumlah klaster tidak lagi memberikan penurunan WCSS yang signifikan [28].

2.8. Evaluasi Hasil

Kualitas hasil klasterisasi dievaluasi menggunakan dua metrik utama yaitu *Silhouette Score* dan *Davies-Bouldin Index* (DBI).

2.8.1. *Silhouette Score*

Silhouette Score mengukur seberapa baik suatu objek cocok dengan klasternya sendiri dibandingkan klaster tetangga. Nilai dihitung dengan Persamaan 4.

$$SI = (b - a) \,/\, \max(a, b) \tag{4}$$

di mana *a* adalah rata-rata jarak antara titik data i dengan seluruh anggota dalam klaster yang sama (kohesi intra-klaster), dan *b* adalah rata-rata jarak minimum antara titik data i dengan anggota klaster lain yang terdekat (separasi inter-klaster). Nilai *Silhouette Score* berkisar −1 hingga +1; nilai > 0,25 diinterpretasikan sebagai struktur klaster yang dapat diterima [29].

2.8.2. *Davies-Bouldin Index* (DBI)

Davies-Bouldin Index (DBI) mengukur rata-rata similaritas antara setiap klaster dengan klaster yang paling mirip dengannya, dihitung sebagai [30], yang diformulasikan pada Persamaan 5.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{j \neq i}\left[\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right] \tag{5}$$

di mana σ_i adalah rata-rata jarak anggota klaster *i* terhadap *centroid*-nya, dan $d(c_i, c_j)$ adalah jarak antara *centroid* klaster *i* dan *j*. Nilai DBI yang lebih rendah mengindikasikan klaster yang lebih kompak dan terpisah dengan baik; DBI < 1,0 umumnya diterima sebagai indikasi kualitas klaster yang baik [30].

2.9. *Profiling*

Profiling klaster merupakan langkah penting dalam interpretasi hasil klasterisasi, yaitu dengan menganalisis karakteristik tiap klaster melalui evaluasi nilai rata-rata, median, dan distribusi setiap fitur untuk mengidentifikasi tren yang membedakan satu klaster dengan klaster lainnya sehingga dapat diterjemahkan menjadi informasi yang dapat ditindaklanjuti [31]. Interpretasi dilakukan secara semantik berdasarkan kombinasi nilai akademik dan status sosial-ekonomi, menghasilkan label klaster yang bermakna secara pedagogis.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penerapan kerangka KDD terhadap data siswa SMA/SMK, meliputi hasil data *preprocessing*, penentuan jumlah klaster optimal, hasil klasterisasi *K-Means*, evaluasi model, serta interpretasi profil setiap klaster secara pedagogis.

3.1. Data *Preprocessing*

Dataset awal terdiri dari 45.244 siswa SMA/SMK yang setelah proses *preprocessing*, ditemukan 32 baris dengan nilai_total hilang (0,07%) dan 2.493 baris dengan desil_kesejahteraan hilang (5,51%), sehingga total 2.522 baris (5,58%) dihapus dari dataset. Dataset valid yang digunakan dalam analisis adalah 42.722 siswa (94,42%). Distribusi data hilang tersebut dirangkum pada Tabel 3, sedangkan statistik deskriptif kedua fitur utama sebelum *preprocessing* disajikan pada Tabel 4.

Tabel 3. Distribusi *Missing Value* pada Fitur Utama

Fitur	*Missing Value*	Persentase
desil_kesejahteraan	2.493	5,51%
nilai_total	32	0,07%
Total dihapus	2.522	5,58%
Data valid tersisa	42.722	94,42%

Tabel 3 menunjukkan bahwa data hilang paling banyak terdapat pada variabel desil_kesejahteraan (5,51%), sedangkan nilai_total hampir lengkap (0,07%), sehingga penghapusan *listwise* hanya mengurangi 5,58% dari keseluruhan data.

Tabel 4. Statistik Deskriptif Fitur Utama (sebelum *preprocessing*)

Statistik	Nilai_Total	Desil_Kesejahteraan
Count	45.212	42.751
Mean	124,83	5,49
Std	28,21	2,79
Min	15,00	1,00
Q1 (25%)	104,82	3,00
Median (50%)	122,97	6,00
Q3 (75%)	143,73	8,00
Max	231,38	10,00

Nilai nilai_total memiliki rata-rata 124,83 dengan standar deviasi 28,21, menunjukkan variasi kemampuan akademik yang cukup besar di antara siswa. Sementara itu, desil_kesejahteraan memiliki rata-rata 5,49 dengan distribusi yang relatif merata di seluruh desil, mengindikasikan keberagaman kondisi sosial-ekonomi yang signifikan. Konsentrasi siswa pada desil menengah-atas (desil 6–8) mencerminkan profil sosial-ekonomi dominan.

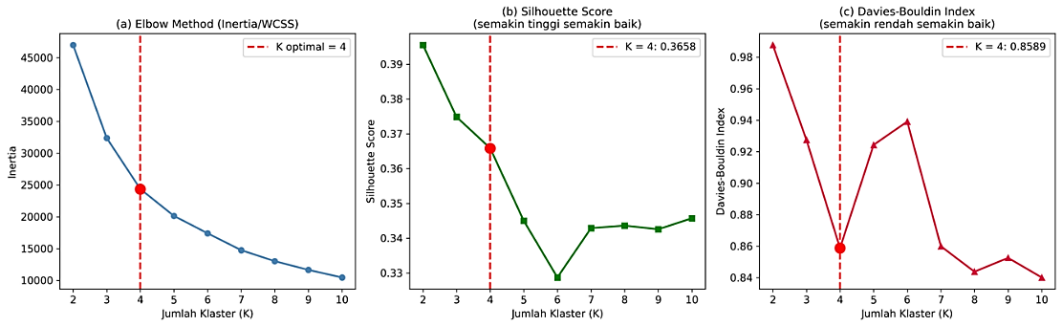
3.2. Penentuan Jumlah Klaster Optimal

Proses penentuan K optimal dilakukan dengan menghitung nilai WCSS (*inertia*), *Silhouette Score*, dan *Davies-Bouldin Index* (DBI) untuk K=2 hingga K=10. Hasil lengkap disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi Metrik untuk K=2 hingga K=10

K	Inertia (WCSS)	Silhouette Score	Davies-Bouldin Index
2	47.000,13	0,3955	0,9877
3	32.383,32	0,3748	0,9276
4 *	24.363,02	0,3658	0,8589
5	20.137,78	0,3450	0,9243
6	17.411,03	0,3287	0,9391
7	14.755,52	0,3429	0,8600
8	13.027,94	0,3436	0,8438
9	11.652,30	0,3425	0,8526
10	10.468,72	0,3457	0,8402

Tabel 5 menyajikan nilai *inertia* (WCSS), *Silhouette Score*, dan *Davies-Bouldin Index* untuk K=2 hingga K=10 yang menjadi dasar penentuan jumlah klaster optimal.



Gambar 2. Grafik *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index*

Berdasarkan analisis grafik *Elbow Method* (Gambar 2) menggunakan algoritma *KneeLocator*, penurunan WCSS yang paling signifikan terjadi hingga K=4, setelah itu mulai melandai membentuk titik siku (*knee*) pada K=4. Hal ini menunjukkan bahwa K=4 merupakan titik optimal di mana penambahan klaster berikutnya tidak memberikan peningkatan kualitas pengelompokan yang berarti. Pada K=4, nilai DBI sebesar 0,8589 merupakan nilai terendah pada rentang K=2 hingga K=6, mengindikasikan klaster yang lebih kompak dan terpisah dengan baik. Nilai *Silhouette Score* sebesar 0,3658 pada K=4 berada pada kategori klaster sedang, yang dapat diterima untuk data dengan distribusi heterogen dan dimensi tinggi seperti data sosial-akademik [29].

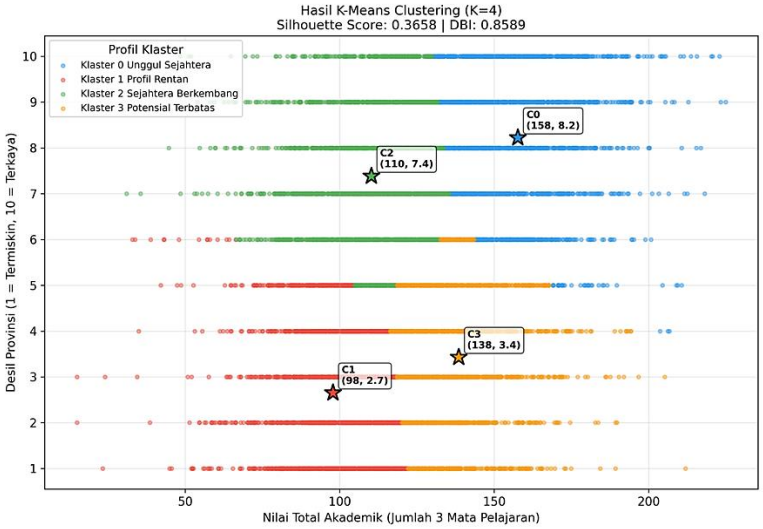
3.3. Hasil Klasterisasi *K-Means* (K=4)

Penerapan *K-Means* dengan K=4, *random_state=42*, dan *n_init=10* menghasilkan distribusi anggota klaster seperti yang disajikan pada Tabel 6, dengan profil rata-rata nilai asli (sebelum standardisasi) pada setiap klaster.

Tabel 6. Distribusi Anggota dan Profil Rata-rata Tiap Klaster

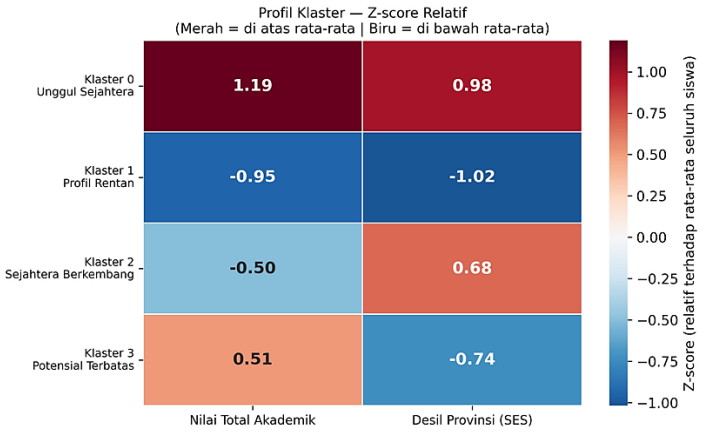
Klaster	Jumlah Siswa	Persentase	Rata-rata nilai_total	Rata-rata desil_kesejahteraan	Label Profil
0	9.870	23,1%	157,50	8,23	Profil Unggul Sejahtera
1	10.896	25,5%	97,77	2,65	Profil Rentan
2	12.454	29,2%	110,14	7,38	Profil Sejahtera Berkembang
3	9.502	22,2%	138,51	3,43	Profil Potensial Terbatas
Total	42.722	100%	124,83	5,49	

Tabel 6 merangkum jumlah dan proporsi anggota tiap klaster beserta rata-rata nilai aslinya, yang memperlihatkan perbedaan karakteristik akademik (nilai_total) dan sosial-ekonomi (desil_kesejahteraan) antarklaster sebagai dasar pelabelan profil.



Gambar 3. *Scatter Plot* Hasil *K-Means Clustering* (K=4)

Gambar 3 menampilkan sebaran siswa pada ruang dua fitur (nilai_total dan desil_kesejahteraan) yang terbagi ke dalam empat kelompok dengan tanda bintang sebagai *centroid* masing-masing klaster, memperlihatkan pemisahan antarklaster sekaligus area tumpang tindih pada batas-batasnya.



Gambar 4. Heatmap Z-score Profil Klaster

Gambar 4 menyajikan *heatmap* nilai *z-score* rata-rata tiap fitur pada masing-masing klaster relatif terhadap rata-rata seluruh siswa, nilai *z-score* positif menandakan capaian di atas rata-rata populasi, sedangkan nilai

negatif menandakan capaian di bawah rata-rata, sehingga posisi relatif tiap klaster pada kedua dimensi terlihat jelas.

3.4. Evaluasi Model

Evaluasi model *K-Means* pada K=4 menghasilkan tiga metrik kualitas sebagai berikut:

1. *Silhouette Score*: 0,3658, mengindikasikan bahwa setiap titik data secara rata-rata cukup lebih dekat ke klasternya sendiri dibandingkan klaster terdekat lainnya. Nilai ini masuk dalam kategori *acceptable* untuk data dengan distribusi heterogen [29].
2. *Davies-Bouldin Index* (DBI): 0,8589 (< 1,0), mengindikasikan bahwa klaster yang terbentuk cukup kompak secara internal dan terpisah satu sama lain, sebagaimana standar DBI < 1,0 yang umumnya diterima sebagai indikasi klaster berkualitas baik [30].

Hasil evaluasi ini konsisten dengan temuan penelitian lain yang menerapkan *K-Means* pada data pendidikan. Ariasa dkk. [15] mencapai DBI 0,168 pada klasterisasi kinerja akademik mahasiswa Undiksha. Djun dkk. [22] dan Cahyaningsih dkk. [16] melaporkan DBI di bawah 1,0 pada klasterisasi data siswa dan pelanggan, mengonfirmasi bahwa nilai DBI yang diperoleh dalam penelitian ini (0,8589) berada dalam rentang kualitas klaster yang dapat diterima [15], [22], [30].

3.5. Analisis Profil Tiap Klaster

Berdasarkan data pada Tabel 6 dan visualisasi scatter plot serta heatmap z-score, setiap klaster menunjukkan karakteristik sosial-akademik yang berbeda dan dapat diinterpretasikan secara pedagogis.

1. Klaster 0, Profil Unggul Sejahtera (n=9.870; 23,1%)
   Klaster 0 merupakan kelompok siswa dengan nilai akademik tertinggi (rata-rata *nilai_total* = 157,50; di atas rata-rata populasi 124,83) sekaligus kondisi sosial-ekonomi terbaik (rata-rata *desil_kesejahteraan* = 8,23 dari skala 10). Kelompok ini didominasi oleh siswa dari kabupaten dengan pertumbuhan ekonomi tinggi seperti Denpasar dan Badung. Dari perspektif pembelajaran berdiferensiasi, klaster ini memerlukan strategi pengayaan (*enrichment*) dan tantangan akademik yang lebih tinggi agar potensi siswa dapat berkembang secara optimal.

2. Klaster 1, Profil Rentan (n=10.896; 25,5%)
   Klaster 1 merupakan kelompok siswa yang menghadapi dua tantangan sekaligus: nilai akademik terendah (rata-rata *nilai_total* = 97,77; jauh di bawah rata-rata populasi) dan kondisi sosial-ekonomi tersulit (rata-rata *desil_kesejahteraan* = 2,65). Klaster ini merupakan kelompok siswa yang paling membutuhkan perhatian dan intervensi prioritas. Kondisi kemiskinan yang terindikasi melalui desil rendah berkontribusi pada terbatasnya akses terhadap fasilitas belajar, yang pada gilirannya berpengaruh negatif terhadap capaian akademik [8]. Strategi pembelajaran yang diperlukan mencakup remedial intensif, pemenuhan kebutuhan dasar belajar, serta dukungan psikososial.

3. Klaster 2, Profil Sejahtera Berkembang (n=12.454; 29,2%)
   Klaster 2 merupakan klaster terbesar dengan proporsi 29,2% dari total siswa. Klaster ini menunjukkan kondisi sosial-ekonomi yang relatif baik (rata-rata *desil_kesejahteraan* = 7,38), namun capaian akademik masih berada di bawah rata-rata populasi (rata-rata *nilai_total* = 110,14). Pola ini mengindikasikan bahwa kondisi sosial-ekonomi yang memadai belum secara otomatis diterjemahkan ke dalam capaian akademik yang tinggi, yang dapat disebabkan oleh faktor pedagogis, motivasi belajar, atau kualitas layanan pendidikan di sekolah. Kelompok ini berpotensi untuk ditingkatkan secara signifikan melalui intervensi pembelajaran yang tepat sasaran.

4. Klaster 3, Profil Potensial Terbatas (n=9.502; 22,2%)
   Klaster 3 menampilkan profil yang menarik: nilai akademik di atas rata-rata populasi (rata-rata *nilai_total* = 138,51), namun kondisi sosial-ekonomi di bawah rata-rata (rata-rata *desil_kesejahteraan* = 3,43). Klaster ini mencerminkan siswa yang mampu mencapai prestasi akademik yang baik meskipun berasal dari latar belakang sosial-ekonomi yang terbatas menggambarkan resiliensi akademik (*academic resilience*). Temuan ini sejalan dengan penelitian Alan & Temiz [13] yang menunjukkan bahwa profil akademik tidak selalu berkorelasi linear dengan kondisi sosial-ekonomi. Kelompok ini memerlukan dukungan berupa beasiswa, pemenuhan fasilitas belajar, dan program *mentoring* agar potensi akademiknya tidak terhalang oleh keterbatasan ekonomi.

3.6. Diskusi

Hasil klasterisasi menghasilkan empat profil siswa yang secara pedagogis bermakna dan tidak dapat direduksi menjadi dikotomi sederhana antara siswa “mampu” dan “tidak mampu”. Temuan paling penting dari penelitian ini adalah terbelahnya kelompok siswa menjadi empat kuadran pada ruang akademik–sosial-

ekonomi, di mana dua di antaranya Klaster 2 (Sejahtera Berkembang) dan Klaster 3 (Potensial Terbatas) justru menunjukkan ketidakselarasan antara kondisi ekonomi dan capaian akademik. Munculnya kedua klaster "paradoks" ini menjelaskan mengapa integrasi dua sumber data menjadi penting: jika hanya menggunakan dimensi akademik, siswa pada Klaster 2 dan Klaster 3 akan tampak serupa dengan kelompok lain, padahal kebutuhan intervensinya sangat berbeda. Pola ini muncul karena *K-Means* memisahkan data berdasarkan jarak *Euclidean* pada dua dimensi terstandarisasi secara simultan, sehingga interaksi antara status ekonomi dan prestasi tertangkap sebagai struktur klaster, bukan sebagai dua distribusi terpisah.

Temuan ini memperkuat sekaligus memperluas hasil penelitian terdahulu. Keberadaan Klaster 1 (Profil Rentan) yang menggabungkan capaian akademik terendah dengan desil kesejahteraan terendah konsisten dengan penelitian Azzizah [8] bahwa kemiskinan merupakan determinan dominan terhadap ketimpangan pendidikan, dan sejalan dengan studi Liu [12] yang menemukan pengaruh negatif status sosial-ekonomi rendah terhadap prestasi belajar. Sebaliknya, Klaster 3 (Potensial Terbatas) yang menampilkan prestasi tinggi di tengah keterbatasan ekonomi memperkuat temuan Alan & Temiz [13] bahwa profil akademik tidak selalu berkorelasi linear dengan kondisi sosial-ekonomi, dan mencerminkan konsep resiliensi akademik (*academic resilience*) yang dikenal dalam literatur pendidikan. Dari sisi kualitas metode, nilai DBI 0,8589 (< 1,0) yang diperoleh sebanding dengan rentang yang dilaporkan Cahyaningsih dkk. [16] (DBI = 0,9095) pada klasterisasi data siswa, perbedaan ini wajar mengingat penelitian ini bekerja pada data individu berskala sangat besar (42.722 *record*) dengan heterogenitas yang jauh lebih tinggi dibanding data kinerja akademik pada satu institusi.

Nilai *Silhouette Score* sebesar 0,3658 yang tergolong moderat perlu dimaknai secara kontekstual, bukan sebagai kelemahan model semata. Adanya area tumpang tindih pada batas antarklaster merupakan konsekuensi alami dari fenomena sosial-pendidikan yang bersifat kontinum, bukan kategori yang terpisah tegas. Siswa pada batas antara Klaster 2 dan Klaster 0, misalnya, merepresentasikan transisi bertahap dari "sejahtera berkembang" menuju "unggul sejahtera", sehingga tumpang tindih tersebut justru mencerminkan realitas empiris yang jujur. Dengan demikian, kombinasi *Silhouette* moderat dan DBI rendah mengindikasikan bahwa keempat klaster cukup kompak dan terpisah untuk kepentingan perumusan kebijakan, meskipun batas antarindividu tidak selalu tegas.

Implikasi utama penelitian ini terletak pada penyediaan basis empiris untuk pembelajaran berdiferensiasi yang tepat sasaran sesuai amanat Kurikulum Merdeka. Keempat profil menuntut strategi yang berbeda: Klaster 0 memerlukan pengayaan dan tantangan akademik lanjutan; Klaster 1 membutuhkan intervensi prioritas berupa remedial intensif, pemenuhan kebutuhan dasar belajar, dan dukungan psikososial; Klaster 2 berpotensi ditingkatkan melalui perbaikan mutu layanan pembelajaran dan penguatan motivasi karena hambatannya bukan pada ekonomi; sedangkan Klaster 3 memerlukan dukungan afirmatif seperti beasiswa, pemenuhan fasilitas belajar, dan program mentoring agar potensinya tidak terhambat keterbatasan ekonomi. Pemetaan ini memungkinkan alokasi sumber daya pendidikan dilakukan berbasis bukti, bukan berdasarkan asumsi tunggal bahwa kesejahteraan ekonomi selalu berbanding lurus dengan capaian akademik.

Kelebihan penelitian ini adalah integrasi dua sumber data resmi berskala nasional (nilai akademik dan status sosial-ekonomi) pada tingkat individu siswa serta orientasi terapannya yang langsung menghubungkan hasil klasterisasi dengan rancangan intervensi pembelajaran. Meskipun demikian, penelitian ini memiliki sejumlah keterbatasan. Pertama, klasterisasi hanya menggunakan dua fitur (nilai_total dan desil_kesejahteraan) sehingga dimensi lain seperti motivasi, kualitas guru, akses teknologi, dan latar belakang keluarga belum terwakili. Kedua, penggunaan *K-Means* sebagai metode tunggal belum dibandingkan secara empiris dengan algoritma klasterisasi lain seperti DBSCAN, Gaussian Mixture Models, atau K-Medoids. Ketiga, penanganan data hilang dengan *listwise deletion* berpotensi menimbulkan bias karena data yang hilang terkonsentrasi pada variabel desil kesejahteraan. Penelitian selanjutnya disarankan untuk menambah dimensi fitur, membandingkan beberapa algoritma klasterisasi, menerapkan teknik imputasi yang lebih robust, serta menguji efektivitas intervensi pembelajaran berdiferensiasi yang dirancang berdasarkan profil klaster ini melalui studi lanjutan berbasis eksperimen.

4. KESIMPULAN

Penelitian ini berhasil memetakan profil sosial-akademik siswa SMA/SMK melalui integrasi data nilai akademik dan status sosial-ekonomi menggunakan metode *K-Means Clustering* dalam kerangka *Knowledge Discovery in Database* (KDD). Dari 45.244 data awal, sebanyak 42.722 data valid (94,42%) dianalisis berdasarkan dua fitur utama, yaitu nilai_total dan desil_kesejahteraan. Jumlah klaster optimal ditetapkan sebesar K=4 melalui *Elbow Method* dengan *KneeLocator*, dengan *Silhouette Score* sebesar 0,3658 dan *Davies-Bouldin Index* sebesar 0,8589 (< 1,0), yang mengindikasikan kualitas pengelompokan yang dapat diterima.

Keempat klaster yang terbentuk merepresentasikan profil sosial-akademik yang berbeda, yaitu: (1) Profil Unggul Sejahtera (Klaster 0; n=9.870; 23,1%) dengan capaian akademik dan kesejahteraan tertinggi; (2) Profil Rentan (Klaster 1; n=10.896; 25,5%) dengan capaian akademik terendah dan kondisi sosial-ekonomi tersulit; (3) Profil Sejahtera Berkembang (Klaster 2; n=12.454; 29,2%) yang relatif sejahtera secara ekonomi namun capaian akademiknya masih di bawah rata-rata populasi; dan (4) Profil Potensial Terbatas (Klaster 3; n=9.502;

22,2%) yang menampilkan resiliensi akademik di tengah keterbatasan ekonomi. Temuan ini menegaskan bahwa kondisi sosial-ekonomi yang baik tidak serta-merta berkorelasi linear dengan capaian akademik, sehingga setiap klaster menuntut strategi pembelajaran berdiferensiasi yang berbeda mulai dari pengayaan dan tantangan akademik, intervensi remedial yang disertai dukungan psikososial, peningkatan mutu layanan pembelajaran, hingga pemberian beasiswa dan program *mentoring*.

Secara praktis, hasil *profiling* ini memberikan landasan empiris bagi guru dan pengambil kebijakan untuk merancang intervensi pembelajaran yang lebih tepat sasaran sesuai semangat Kurikulum Merdeka. Penelitian ini memiliki sejumlah keterbatasan, antara lain penggunaan dua fitur utama serta cakupan data yang terbatas pada satu wilayah dan satu periode asesmen. Penelitian selanjutnya disarankan untuk menambah dimensi fitur (misalnya komponen literasi, numerasi, dan karakter dari AKM), membandingkan *K-Means* dengan algoritma klasterisasi lain, serta menguji generalisasi model pada cakupan wilayah dan periode yang lebih luas.

REFERENCES

[1] H. Aisah, Q. Y. Zaqiah, and A. Supiana, "Implementasi Kebijakan Asesmen Kemampuan Minimum (AKM): Analisis Implementasi Kebijakan AKM)," *Jurnal Pendidikan Islam Al-Affan*, vol. 1, no. 2, pp. 128–135, Mar. 2021, doi: 10.69775/jpia.v1i2.25.

[2] N. Novita, M. Mellyzar, and H. Herizal, "Asesmen Nasional (AN): Pengetahuan dan Persepsi Calon Guru," *JISIP (Jurnal Ilmu Sosial dan Pendidikan)*, vol. 5, no. 1, Jan. 2021, doi: 10.58258/jisip.v5i1.1568.

[3] E. Driana, A. M. Ningsih, E. Ernawati, and H. Setiadi, "Teachers' and Students' Understanding and Perception of the National Assessment in Indonesian Primary Schools: Implications for Learning and Classroom Assessment," *Mimbar Sekolah Dasar*, vol. 12, no. 2, pp. 205–223, Jun. 2025, doi: 10.53400/mimbar-sd.v12i2.86500.

[4] I. M. Wargina, I. M. Candiasa, and I. G. N. Pujawan, "Evaluasi Efektivitas Pelaksanaan Asesmen Nasional Pada SMP Negeri Di Kecamatan Baturiti Tahun 2024," *Cetta: Jurnal Ilmu Pendidikan*, vol. 9, no. 1, pp. 10–24, Jan. 2026, doi: 10.37329/cetta.v9i1.4661.

[5] K. P. dan K. Pusat Penilaian Pendidikan, "Ringkasan Eksekutif Hasil Ujian Nasional 2019: Masukan untuk Pembelajaran di Sekolah SMA/MA," Pusat Penilaian Pendidikan, Kementerian Pendidikan dan Kebudayaan, Jakarta, 2019. [Online]. Available: https://repositori.kemendikdasmen.go.id/18370/

[6] M. N. P. Janu, Hartono, and Ig. A. Dwiatmoko, "Analisis Hasil Ujian Nasional Tahun 2016-2019 Tingkat Sekolah Menengah Pertama dengan Pendekatan Sains Data," in *Prosiding Seminar Nasional Ilmu Sosial dan Teknologi (SNISTEK)*, Batam: LPPM Universitas Putera Batam, 2022, pp. 370–376. [Online]. Available: https://ejournal.upbatam.ac.id/index.php/prosiding/article/view/5335

[7] OECD, "PISA 2022 Results (Volume I): The State of Learning and Equity in Education," OECD Publishing, Paris, 2023. doi: 10.1787/53f23881-en.

[8] Y. Azzizah, "Socio-Economic Factors on Indonesia Education Disparity," *International Education Studies*, vol. 8, no. 12, p. 218, Nov. 2015, doi: 10.5539/ies.v8n12p218.

[9] P. S. Saputra, I. G. A. Gunadi, and G. R. Dantes, "Perbandingan Algoritma Fuzzy C-Means dan Algoritma Naïve Bayes dalam Menentukan Keluarga Penerima Manfaat (KPM) berdasarkan Status Sosial Ekonomi (SSE) Terendah," *JST (Jurnal Sains dan Teknologi)*, vol. 10, no. 1, pp. 1–8, Mar. 2021, doi: 10.23887/jstundiksha.v10i1.23340.

[10] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Applied Sciences*, vol. 12, no. 19, p. 9467, 2022, doi: 10.3390/app12199467.

[11] G. R. Suraya and A. W. Wijayanto, "Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia According to the Special Index for Handling Stunting," *Indonesian Journal of Statistics and Its Applications*, vol. 6, no. 2, pp. 180–201, 2022, doi: 10.29244/ijsa.v6i2p180-201.

[12] J. Liu, "Implementation of K-Means Clustering on Student Learning Achievements Based on Social Economic and Social Related," *Journal of Education, Humanities and Social Sciences*, vol. 29, pp. 475–479, 2024, doi: 10.54097/9x0m0j15.

[13] M. A. Alan and M. Temiz, "A Study On Profiling Students via Data Mining," *Alphanumeric Journal*, vol. 7, no. 2, pp. 239–248, Dec. 2019, doi: 10.17093/alphanumeric.630866.

[14] I. G. A. Gunadi, D. P. S. Putri, I. N. Mistanada, G. S. Prayoga, and N. M. Y. D. Rahayu, "Analisis Cluster Pada Pengelompokan Siswa Diktuk Bintara Polri TA. 2018/2019, SPN Singaraja - Polda Bali Menggunakan K-Means dan K-Harmonic Means," *Jurnal Ilmiah SINUS*, vol. 17, no. 2, p. 13, Jul. 2019, doi: 10.30646/sinus.v17i2.421.

[15] K. Ariasa, I. G. A. Gunadi, and I. M. Candiasa, "Optimasi Algoritma Klaster Dinamis pada K-means dalam Pengelompokan Kinerja Akademik Mahasiswa (Studi Kasus: Universitas Pendidikan Ganesha)," *Jurnal Nasional Pendidikan Teknik Informatika*, vol. 9, no. 2, pp. 181–192, 2020, doi: 10.23887/janapati.v9i2.23491.

[16] A. U. G. Cahyaningsih, I. M. Candiasa, and I. G. A. Gunadi, "Identifikasi Pola Komunikasi Dan Kepribadian Siswa Sekolah Luar Biasa (SLB) Melalui Analisis Konten Media Sosial dengan Metode Anova dan K-Means," vol. 12, no. 6, pp. 2355–7699, 2025.

[17] I. G. M. W. K. Widiantara, K. Y. E. Aryanto, and I. M. G. Sunarya, "Application of the Learning Vector Quantization Algorithm for Classification of Students with the Potential to Drop Out," *Brilliance: Research of Artificial Intelligence*, vol. 3, no. 2, pp. 262–269, Nov. 2023, doi: 10.47709/brilliance.v3i2.3155.

[18] I. K. N. Ananda, N. P. N. P. Dewi, N. W. Marti, and L. J. E. Dewi, "Klasifikasi Multilabel Pada Gaya Belajar Siswa Sekolah Dasar Menggunakan Algoritma Machine Learning," *Journal of Applied Computer Science and Technology*, vol. 5, no. 2, pp. 144–154, Dec. 2024, doi: 10.52158/jacost.v5i2.940.

[19] J. C. Alejandrino, "Application of Data Mining in Knowledge Management: A Review," *International Journal of Advanced Trends in Computer Science and Engineering*, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:242455535

[20] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," 2010. [Online]. Available: http://sites.google.com/site/ijcsis/

[21] K. H. Izzuddin and A. W. Wijayanto, "Pemodelan Clustering Ward, K-Means, Diana, dan PAM dengan PCA untuk Karakterisasi Kemiskinan Indonesia Tahun 2021," *Komputika : Jurnal Sistem Komputer*, vol. 13, no. 1, pp. 41–53, Apr. 2024, doi: 10.34010/komputika.v13i1.10803.

[22] S. F. Djun, I. G. A. Gunadi, and S. Sariyasa, "Analisis Segmentasi Pelanggan pada Bisnis dengan Menggunakan Metode K-Means Clustering pada Model Data RFM," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 5, no. 4, pp. 354–364, Feb. 2024, doi: 10.35746/jtim.v5i4.434.

[23] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means Clustering Algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[24] M. I. Bimasena, I. G. A. Gunadi, and I. M. A. Wirawan, "Studi Perbandingan Algoritma Euclidean, Manhattan dan Chebysev Distance untuk Optimasi Metode K-means Clustering pada Pengelompokkan Data Penerimaan Mahasiswa Baru," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 22, no. 2, pp. 116–127, Jul. 2025, doi: 10.23887/jptk-undiksha.v22i2.98863.

[25] I. P. Suwirya, I. M. Candiasa, and G. R. Dantes, "Evaluation of ATM Location Placement Using the K-Means Clustering in BNI Denpasar Regional Office," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 4, no. 2, pp. 158–168, Jul. 2022, doi: 10.47709/cnahpc.v4i2.1580.

[26] F. Farahdinna, I. Nurdiansyah, A. Suryani, and A. Wibowo, "Perbandingan Algoritma K-Means dan K-Medoids dalam Klasterisasi Produk Asuransi Perusahaan Nasional," *Jurnal Ilmiah FIFO*, vol. 11, no. 2, p. 208, 2019, doi: 10.22441/fifo.2019.v11i2.010.

[27] S. A. D. Budiman, D. Safitri, and D. Ispriyanti, "Perbandingan Metode K-Means dan Metode DBSCAN pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang," *Jurnal Gaussian*, vol. 5, no. 4, pp. 757–762, 2016, doi: 10.14710/j.gauss.5.4.757-762.

[28] A. Kuraria, N. Jharbade, and M. Soni, "Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining," *Int. J. Sci. Res. Sci. Eng. Technol.*, pp. 190–195, Dec. 2018, doi: 10.32628/IJSRSET21841122.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[30] R. K. Dinata, Bustami, and S. Retno, "Optimizing the Evaluation of K-means Clustering Using the Weight Product," *Revue d'Intelligence Artificielle*, vol. 38, no. 4, pp. 1223–1233, Aug. 2024, doi: 10.18280/ria.380416.

[31] C. P. Ufeli, M. U. Sattar, R. Hasan, and S. Mahmood, "Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms," *Information*, vol. 16, no. 6, p. 441, May 2025, doi: 10.3390/info16060441.