

BAB I

PENDAHULUAN

1.1 Latar Belakang

Peningkatan kualitas pendidikan merupakan agenda prioritas dalam pembangunan nasional Indonesia. Sebagai salah satu langkah strategis, pemerintah mengganti Ujian Nasional (UN) dengan Asesmen Nasional (AN) pada tahun 2021. Berbeda dengan UN yang berpusat pada pencapaian nilai akhir peserta didik, AN memprioritaskan penguasaan kompetensi literasi, numerasi, dan penguatan karakter (Aisah dkk., 2021). Kebijakan ini diarahkan untuk mempersiapkan sumber daya manusia Indonesia menghadapi tantangan abad ke-21 melalui penguasaan kompetensi yang bermanfaat bagi kehidupan sehari-hari peserta didik.

Asesmen Nasional mengukur hasil belajar kognitif melalui AKM (literasi dan numerasi) serta hasil non-kognitif melalui survei karakter dan lingkungan belajar, yang dirancang sebagai alat refleksi peningkatan kualitas pembelajaran tanpa konsekuensi individual atau pemeringkatan sekolah (Driana dkk., 2025). Selanjutnya Driana dkk. (2025) mengungkapkan bahwa meskipun guru memiliki pemahaman dasar tentang AN dan memandangnya sebagai instrumen penting, masih diperlukan dukungan infrastruktur dan pedagogis, sementara siswa memandang AN sebagai tes berisiko rendah (*low-stakes test*). Implementasi AN menghadapi berbagai tantangan, Desi Rahmawati dkk. (2024) dalam evaluasinya di SMPN 216 Jakarta mengidentifikasi strategi optimalisasi meliputi perencanaan strategis, peningkatan kompetensi guru, dan pemanfaatan teknologi digital, namun studi tersebut belum menyentuh pemahaman mendalam mengenai karakteristik sosial-akademik siswa.

Paradoks kebijakan Asesmen Nasional terletak pada kesenjangan antara tujuan pemetaan mutu pendidikan secara komprehensif dengan keterbatasan pemahaman terhadap profil peserta didik itu sendiri, di mana pemerintah telah mengumpulkan data akademik melalui nilai tes kemampuan akademik dan data sosial-ekonomi yang sangat kaya, namun analisis integratif kedua sumber data untuk memahami profil sosial-akademik siswa secara utuh masih sangat terbatas. Azzizah (2015) menegaskan bahwa ketimpangan sosial-ekonomi antar wilayah Indonesia tercermin dalam ketimpangan pendidikan, dengan faktor kemiskinan sebagai determinan paling berpengaruh terhadap angka partisipasi sekolah, diikuti pertumbuhan PDRB dan Indeks Gini, sehingga pemahaman kondisi sosial-ekonomi menjadi prasyarat mutlak intervensi pendidikan yang efektif. Alan & Temiz (2019) mendemonstrasikan efektivitas *data mining* melalui algoritma Apriori dalam memprofilkan 443 siswa berdasarkan kondisi kesehatan, sumber daya keuangan, standar hidup, dan tingkat pendidikan keluarga dengan menghasilkan 72 aturan berakurasi di atas 90% yang berkontribusi pada manajemen sekolah, guru, orang tua, dan siswa. Kompleksitas data sosial-akademik yang bersifat multidimensional dengan data akademik kontinu dan data sosial-ekonomi berskala beragam menghasilkan ruang fitur berdimensi tinggi dan distribusi heterogen, sehingga analisis kluster menjadi pendekatan relevan untuk mengelompokkan siswa berdasarkan kemiripan karakteristik.

Kebutuhan akan profil sosial-akademik siswa yang utuh semakin mendesak dalam kerangka Kurikulum Merdeka, yang menempatkan pembelajaran berdiferensiasi (*differentiated instruction*) dan pengajaran sesuai tingkat capaian (*teaching at the right level*) sebagai pendekatan utama. Pembelajaran

berdiferensiasi menuntut guru menyesuaikan konten, proses, sarana, dan model pembelajaran dengan kebutuhan, kesiapan, serta karakteristik tiap siswa. Namun, pembelajaran di SMA/SMK saat ini belum berjalan optimal karena belum sepenuhnya menyentuh kebutuhan individual siswa (Driana dkk., 2025; Wargina dkk., 2026). Akar persoalannya adalah tidak tersedianya profil siswa yang komprehensif dan terintegrasi sebagai dasar perancangan intervensi. Tanpa informasi profil yang menggambarkan kondisi akademik sekaligus sosial-ekonomi siswa secara utuh, penyesuaian sarana, model, dan strategi pembelajaran berisiko tidak tepat sasaran. Oleh karena itu, ketersediaan data profil siswa yang valid menjadi prasyarat agar dalam skala mikro di kelas, pembelajaran berdiferensiasi dapat diterapkan secara efektif. Dalam skala makro, layanan pendidikan juga dapat dilakukan secara berdiferensiasi dengan mempertimbangkan kondisi sosial-ekonomi siswa, selain kemampuan akademiknya. Sebagai contoh, siswa berprestasi tinggi dari keluarga kurang mampu membutuhkan bantuan finansial dan beasiswa, sedangkan siswa berprestasi rendah dari keluarga mampu justru membutuhkan intervensi pedagogis dan bukan bantuan ekonomi. Bentuk layanan yang tepat karenanya hanya dapat ditentukan apabila kondisi akademik dan sosial-ekonomi siswa diketahui secara bersamaan.

Kompleksitas data sosial-akademik siswa meniscayakan penggunaan metode analisis yang mampu menangkap struktur multidimensional dalam data. Data akademik yang tercermin dari hasil tes kemampuan kognitif bersifat kontinu dan terukur, sementara data sosial-ekonomi mencakup beragam aspek seperti pendapatan orang tua, tingkat pendidikan orang tua, kondisi tempat tinggal, akses terhadap fasilitas belajar, dan berbagai indikator sosial lainnya yang memiliki skala

pengukuran beragam. Integrasi kedua jenis data ini menghasilkan ruang fitur berdimensi tinggi dengan karakteristik distribusi yang heterogen. Dalam konteks ini, analisis kluster menjadi pendekatan yang relevan untuk mengelompokkan siswa ke dalam profil-profil yang memiliki kemiripan karakteristik sosial-akademik.

Penelitian mengenai analisis kluster untuk pemetaan sebaran guru telah dilakukan oleh Muttaqin dkk. (2024) yang mengimplementasikan algoritma *K-Means Clustering* dalam memetakan distribusi guru di Kota Langsa berdasarkan jumlah guru, siswa, dan rombongan belajar. Studi tersebut berhasil mengelompokkan 30 sekolah ke dalam tiga kluster kekurangan guru (20 sekolah), cukup guru (7 sekolah), dan kelebihan guru (3 sekolah) dengan tingkat akurasi mencapai 92,8%. Namun demikian, studi tersebut terbatas pada data administratif pendidikan dan belum mengintegrasikan aspek sosial-ekonomi dalam analisisnya. Sementara itu, Weingärtner-Junior dkk. (2024) dalam studinya mengenai profil psikologis mahasiswa teknik di Universitas São Paulo mengombinasikan analisis statistik *multivariat* dengan *Self-Organizing Map* (SOM) untuk mengelompokkan profil psikologis mahasiswa berdasarkan *Myers-Briggs Type Indicator* (MBTI). Penelitian tersebut menunjukkan bahwa algoritma SOM mampu mengelompokkan pola-pola kepribadian berdasarkan afinitas data dan membandingkannya dengan hasil yang diperoleh melalui statistik non-parametrik *Kruskal-Wallis*. Studi ini membuktikan bahwa pendekatan berbasis kecerdasan buatan, khususnya *unsupervised learning*, dapat memberikan hasil yang sebanding bahkan melampaui metode statistik konvensional dalam analisis profil individu. Dehuri dkk. (2006) dalam studi komparatifnya mengenai algoritma klustering membandingkan kinerja *K-Means*, SOM, dan DBSCAN menggunakan *dataset* artifisial dan *dataset* IRIS.

Hasil studi menunjukkan bahwa jika kluster berbentuk *non-convex* atau *arbitrary shape*, algoritma berbasis densitas seperti DBSCAN lebih unggul, sementara jika kluster berbentuk *hyperspherical* atau *convex* dan terpisah dengan baik, maka SOM atau *K-Means* lebih disarankan. Studi ini menjadi fondasi penting dalam memahami bahwa tidak ada satu algoritma klustering yang unggul untuk seluruh kondisi data. Karakteristik data menentukan algoritma mana yang paling tepat digunakan. Lebih lanjut, Amalina & Fauzan (2024) mengimplementasikan algoritma HDBSCAN, yang merupakan pengembangan dari DBSCAN, untuk mengelompokkan potensi desa di Kabupaten Buleleng. Studi tersebut berhasil mengidentifikasi dua kluster utama dan satu kluster *noise* yang terdiri dari 24 desa yang teridentifikasi sebagai *outlier* dengan nilai *Silhouette Coefficient* sebesar 0,37. Penelitian ini membuktikan keunggulan HDBSCAN dalam menangani data dengan kepadatan bervariasi dan ketahanannya terhadap data pencilan (*noise*) serta kemampuannya mengelompokkan wilayah berdasarkan potensi desa untuk mendukung perumusan kebijakan pembangunan yang lebih terarah dan efisien.

Penelitian yang diidentifikasi dalam studi ini bersumber pada kompleksitas integrasi data nilai akademik dan sosial-ekonomi yang belum pernah dianalisis menggunakan pendekatan komparatif antar algoritma klustering dengan karakteristik berbeda. Data nilai akademik merepresentasikan kemampuan akademik siswa yang bersifat kuantitatif kontinu, sementara data sosial-ekonomi yang bersifat multidimensi dengan campuran tipe data numerik dan kategorikal. Integrasi kedua sumber data ini menghasilkan struktur data yang kompleks dengan karakteristik dimensi tinggi akibat banyaknya variabel indikator sosial-ekonomi, distribusi *non-spherical* karena data sosial-ekonomi cenderung mengelompok pada

rentang nilai tertentu, kepadatan bervariasi antar kelompok siswa, dan keberadaan data pencilan yang merepresentasikan kondisi ekstrem baik dari sisi kemampuan akademik maupun kondisi sosial-ekonomi.

Kompleksitas ini menimbulkan permasalahan fundamental dalam pemilihan algoritma klustering yang paling sesuai. Algoritma *K-Means* yang berbasis *centroid* mengasumsikan kluster berbentuk *spherical* dan berukuran serupa, serta sensitif terhadap *outlier* (Dehuri dkk., 2006). Algoritma HDBSCAN yang berbasis densitas mampu mengidentifikasi kluster dengan bentuk arbitrer dan kepadatan bervariasi serta secara eksplisit mengklasifikasikan data pencilan sebagai *noise* (Amalina & Fauzan, 2024; Stewart & Al-Khassaweneh, 2022). Sementara *Self-Organizing Map* yang berbasis topologi jaringan saraf tiruan mampu memproyeksikan data berdimensi tinggi ke dalam peta dua dimensi sambil mempertahankan struktur topologi data (Gunawan dkk., 2024; Weingärtner-Junior dkk., 2024). Ketiga algoritma ini merepresentasikan paradigma yang berbeda dalam analisis kluster: *centroid-based*, *density-based*, dan *model-based*. Pertanyaan mengenai paradigma mana yang paling sesuai untuk menangkap struktur profil sosial-akademik siswa belum terjawab dalam literatur yang ada.

Urgensi penelitian komparatif ini diperkuat oleh kebijakan Merdeka Belajar yang menekankan pembelajaran *berdiferensiasi* dan *teaching at the right level*, di mana pemetaan profil sosial-akademik siswa yang akurat menjadi prasyarat bagi guru untuk merancang pembelajaran responsif, namun Driana dkk. (2025) menegaskan bahwa pemahaman dan pemanfaatan hasil asesmen oleh guru masih menjadi tantangan sehingga intervensi berisiko tidak tepat sasaran. Agustina dkk. (2025) merekomendasikan adopsi model penilaian komprehensif seperti

pendekatan CIPP (*Context, Input, Process, Product*) disertai pengembangan profesional guru untuk mengukur nilai-nilai karakter secara efektif, menjadikan pengembangan metode klasterisasi yang valid dan reliabel sebagai kebutuhan mendesak. Alan & Temiz (2019) menegaskan bahwa analisis data siswa melalui *data mining* mampu menghasilkan informasi bermanfaat untuk membangun basis pedagogis dalam perancangan pendekatan pendidikan.

Berdasarkan penjabaran permasalahan di atas, maka akan dilakukan penelitian “Analisis Komparatif Algoritma *K-Means*, HDBSCAN, dan *Self-Organizing Map* untuk Profiling Siswa SMA/SMK berdasarkan Nilai Akademik dan Status Sosial-Ekonomi”, penelitian ini bertujuan melakukan analisis komparatif terhadap algoritma *K-Means*, HDBSCAN, dan *Self-Organizing Map* dalam klasterisasi profil sosial-akademik siswa berdasarkan integrasi data nilai akademik dan tingkat kesejahteraan sosial-ekonomi.

1.2 Identifikasi Masalah

Berdasarkan penjelasan latar belakang di atas, dapat diidentifikasi beberapa permasalahan sebagai berikut:

1. Pembelajaran di SMA/SMK belum berjalan optimal karena belum menyentuh kebutuhan individual siswa.
2. Pemberian layanan yang berdiferensiasi kepada siswa dengan kemampuan akademik dan status sosial-ekonomi (SSE) yang berbeda belum optimal.
3. Diperlukan informasi profil siswa agar guru dapat menyesuaikan sarana, model, dan strategi dalam pembelajaran berdiferensiasi.
4. Profil sosial-akademik siswa SMA/SMK yang terintegrasi antara data

akademik dan data sosial-ekonomi belum tersedia.

5. Terdapat beberapa metode pengklasteran dengan paradigma berbeda, yaitu K-Means (*centroid-based*), HDBSCAN (*density-based*), dan Self-Organizing Map (*model-based*).
6. Diperlukan perbandingan antar metode pengklasteran agar diperoleh metode yang paling efektif untuk memprofilkan siswa berdasarkan karakteristik data sosial-akademik.

1.3 Batasan Masalah

Untuk mempermudah dalam penelitian, batasan masalah dirancang sebagai berikut:

1. Proses pra-pemrosesan data meliputi penanganan *missing value*, normalisasi data numerik, dan transformasi data kategorikal ke format numerik melalui pendekatan label *encoding* dengan tetap memperhatikan implikasi terhadap interpretasi jarak.
2. Algoritma *K-Means* dilakukan melalui penentuan jumlah kluster optimal menggunakan metode *Elbow* dan *Silhouette Coefficient*. Optimalisasi algoritma HDBSCAN dilakukan melalui pencarian parameter *min_cluster_size* dan *min_samples*. Optimalisasi algoritma SOM dilakukan melalui penentuan ukuran *grid* dan jumlah iterasi pelatihan.
3. Perbandingan evaluasi kinerja antar algoritma menggunakan metrik internal *Davies-Bouldin Index* (DBI) dan nilai *Silhouette Score*, serta analisis deskriptif terhadap profil kluster yang terbentuk.

4. Variabel tingkat kesejahteraan diperlakukan sebagai variabel numerik dengan keterbatasan bahwa jarak antar desil diasumsikan ekuivalen.

1.4 Rumusan Masalah

Berdasarkan uraian pada bagian latar belakang dan identifikasi masalah, penulis merumuskan masalah sebagai berikut:

1. Bagaimana komparasi kinerja algoritma *K-Means*, HDBSCAN, dan *Self-Organizing Map* dalam klusterisasi profil sosial-akademik siswa SMA/SMK berdasarkan integrasi data nilai akademik dan tingkat kesejahteraan sosial-ekonomi?
2. Algoritma manakah yang paling efektif dalam memprofilkan siswa berdasarkan metrik evaluasi internal *Silhouette Score* dan *Davies-Bouldin Index* (DBI)?
3. Bagaimana karakteristik klaster profil sosial-akademik siswa yang terbentuk dari hasil klusterisasi algoritma terbaik?

1.5 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menganalisis dan membandingkan kinerja algoritma *K-Means*, HDBSCAN, dan *Self-Organizing Map* dalam klusterisasi profil siswa SMA/SMK berdasarkan integrasi data nilai akademik dan tingkat kesejahteraan sosial-ekonomi.

2. Menentukan algoritma yang paling efektif dalam memprofilkan siswa berdasarkan metrik evaluasi internal *Silhouette Score* dan *Davies-Bouldin Index* (DBI).
3. Mendeskripsikan karakteristik klaster profil sosial-akademik siswa yang terbentuk dari hasil klasterisasi algoritma terbaik.

1.6 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi yang bermanfaat, baik dari segi akademis maupun praktis, dengan rincian sebagai berikut.

1.6.1 Manfaat Akademik

1. Memberikan kontribusi pada pengembangan ilmu pengetahuan, khususnya dalam bidang *data mining* dan pembelajaran mesin tak terawasi (*unsupervised learning*), terkait analisis komparatif algoritma klasterisasi berbasis *centroid* (*K-Means*), berbasis densitas (HDBSCAN), dan berbasis topologi jaringan saraf tiruan (*Self-Organizing Map*) dalam menangani data sosial-akademik berdimensi tinggi dan berdistribusi heterogen.
2. Menyediakan kerangka kerja metodologis untuk integrasi data akademik dan data sosial-ekonomi dalam analisis profil peserta didik yang dapat diadaptasi pada konteks dan wilayah lain.
3. Menjadi fondasi bagi penelitian lanjutan mengenai pengembangan sistem rekomendasi intervensi pembelajaran berdiferensiasi berbasis profil sosial-akademik siswa serta penelitian mengenai prediksi risiko putus sekolah atau rendahnya capaian akademik berbasis klaster.

1.6.2 Manfaat Praktis

1. Bagi Pemerintah, penelitian ini memberikan pemetaan komprehensif mengenai profil sosial-akademik siswa yang dapat digunakan sebagai dasar perumusan kebijakan pendidikan yang lebih tepat sasaran, efisien, dan berkeadilan, termasuk dalam alokasi bantuan operasional sekolah, program intervensi pembelajaran, dan pengembangan kurikulum yang responsif terhadap keragaman siswa.
2. Bagi sekolah dan guru, hasil klasterisasi profil sosial-akademik siswa dapat dimanfaatkan untuk merancang strategi pembelajaran berdiferensiasi (*differentiated instruction*) dan pengajaran sesuai tingkat capaian (*teaching at the right level*), sehingga intervensi pedagogis dapat diberikan secara tepat kepada kelompok siswa dengan karakteristik spesifik.
3. Bagi siswa dan orang tua, penelitian ini secara tidak langsung memberikan manfaat melalui peningkatan kualitas layanan pendidikan yang lebih personal dan responsif terhadap kondisi sosial-ekonomi keluarga.
4. Bagi peneliti lain, penelitian ini menyediakan referensi empiris mengenai penerapan komparatif tiga paradigma algoritma klasterisasi pada konteks data pendidikan Indonesia yang memiliki karakteristik unik dan kompleks.