Vol 3 No. 1 Februari 2026 P-ISSN : 3047-2113 E-ISSN : 3047-2121, Hal 48 - 56

JURNAL ILMIAH
MULTIDISIPLIN ILMU
Halaman Jurnal: https://journal.smartpublisher.id/index.php/jimi
Halaman UTAMA Jurnal : https://journal.smartpublisher.id/

DOI:https://doi.org/10.69714/1kays705

PERBANDINGAN METODE K-MEANS DAN DBSCAN PADA ANALAISIS KLASTER MULTIVARIAT PROFIL SISWA

Ana Fauziah [a*]
[a] Fakultas Keguruan dan Ilmu Pendidikann/ Pendidikan Matematika; ana@ubibanyuwangi,ac.id,
Universitas Bakti Indonesia; Jl. Jember No. 40, Cempokosari Banyuwangi, Jawa Timur
[*] Penulis Korespondensi: Ana Fauziah

ABSTRACT

In educational practice, students are often treated as a homogeneous group within a heterogeneous environment. The "one-size-fits-all" approach has proven ineffective as it overlooks the segmented needs of diverse individuals. This study aims to perform student profiling through cluster analysis to identify groups requiring specialized support or enrichment. The methodology compares two clustering algorithms: K-Means (partition-based) and DBSCAN (density-based). Prior to modeling, optimal parameters were determined using the Elbow method for K-Means and the K-distance graph for DBSCAN. Model evaluation was based on the number of resulting clusters and the Silhouette Score. The results indicate that DBSCAN outperformed K-Means with a Silhouette Score of 0.4 compared to 0.3. Furthermore, DBSCAN produced a more optimal and cohesive structure with 5 clusters, making it more interpretable for student profiling than the 8-cluster solution from K-Means. The analysis successfully identified five student profiles: High Achievers, Above Average, Moderate, Below Average, and At-Risk/Underprivileged. This study concludes that multivariate cluster analysis is an effective instrument for targeted educational intervention, particularly in prioritizing economic aid and academic mentoring for vulnerable groups without relying on demographic identity as a determinant of individual ability.

Keywords: *Student Profiling; Cluster Analysis; DBSCA; K-Means; Educational Intervention*

Abstrak

Dalam praktik pendidikan, siswa sering kali diperlakukan sebagai kelompok homogen di tengah lingkungan yang heterogen. Pendekatan *one-size-fits-all* terbukti kurang efektif karena mengabaikan segmentasi kebutuhan individu yang beragam. Penelitian ini bertujuan untuk melakukan profil siswa melalui analisis klaster guna mengidentifikasi kelompok yang memerlukan dukungan khusus maupun pengayaan. Metodologi penelitian ini membandingkan dua algoritma klaster, yaitu K-Means yang berbasis partisi dan DBSCAN yang berbasis kepadatan. Sebelum dilakukan pemodelan, parameter optimal ditentukan menggunakan metode *Elbow* untuk K-Means dan grafik *K-distance* untuk DBSCAN. Evaluasi model didasarkan pada jumlah klaster yang terbentuk dan nilai *Silhouette Score*. Hasil penelitian menunjukkan bahwa DBSCAN memiliki performa yang lebih unggul dengan nilai *Silhouette Score* sebesar 0,4, dibandingkan K-Means yang hanya mencapai 0,3. Selain itu, DBSCAN mampu menghasilkan jumlah klaster yang lebih optimal dan padat, yaitu sebanyak 5 klaster, sehingga lebih mudah diinterpretasikan dalam pemetaan profil siswa dibandingkan K-Means yang menghasilkan 8 klaster. Analisis menggunakan algoritma klaster DBSCAN berhasil memetakan lima profil siswa, yaitu: *High Achievers, upper-middle class, moderat, lower-middle class,* dan *Underprivileged/At-Risk*. Penelitian ini menyimpulkan bahwa analisis klaster multivariat merupakan instrumen efektif untuk intervensi pendidikan yang tepat sasaran, khususnya dalam memprioritaskan bantuan ekonomi dan pendampingan akademik bagi kelompok rentan tanpa terjebak pada determinisme identitas demografis.

Kata Kunci: *Student Profiling*; Analisis Klaster; DBSCAN; K-Means, Intervensi Pendidikan

Naskah Masuk 13 Januari 2026; Revisi 15 Januari 2026; Diterima 19 Januari 2026; Tersedia 20 Januari 2026

1. PENDAHULUAN

Pendidikan berkualitas merupakan fondasi utama dalam pengembangan sumber daya manusia, di mana pemahaman terhadap karakteristik siswa memegang peranan penting dalam keberhasilan proses pembelajaran. Keberhasilan pendidikan tidak hanya dipengaruhi oleh faktor institusional seperti kurikulum dan metode pembelajaran, tetapi juga oleh karakteristik peserta didik yang bersifat multidimensional, meliputi latar belakang sosial, dukungan keluarga, dan kemampuan akademik. Berbagai penelitian menunjukkan bahwa faktor-faktor tersebut memiliki pengaruh signifikan terhadap capaian belajar siswa [1].

Dalam lingkungan pendidikan yang heterogen, siswa memiliki latar belakang sosial, dukungan keluarga, dan kemampuan akademik yang sangat bervariasi. Namun, dalam praktik pendidikan, siswa sering kali diperlakukan sebagai kelompok yang homogen, sehingga variasi karakteristik individual kurang mendapatkan perhatian yang memadai. Fenomena ini menciptakan tantangan bagi pengambil kebijakan pendidikan untuk memberikan intervensi yang tepat sasaran. Pendekatan "satu ukuran untuk semua" (*one-size-fits-all*) seringkali tidak efektif karena mengabaikan segmentasi kebutuhan siswa yang berbeda-beda. Oleh karena itu, diperlukan pemetaan atau profiling siswa yang akurat untuk mengidentifikasi kelompok yang membutuhkan dukungan khusus maupun pengayaan.

Dalam praktik pendidikan, keberagaman karakteristik siswa sering kali belum sepenuhnya menjadi dasar dalam perancangan pembelajaran. Pendekatan pembelajaran yang bersifat seragam (one-size-fits-all) masih banyak diterapkan, sehingga kurang mampu mengakomodasi perbedaan kebutuhan dan potensi siswa. Kondisi ini menimbulkan tantangan bagi pendidik dan pengambil kebijakan untuk merancang intervensi pendidikan yang tepat sasaran. Oleh karena itu, diperlukan pemetaan atau profiling siswa yang sistematis dan berbasis data guna mengidentifikasi kelompok siswa dengan karakteristik dan kebutuhan yang berbeda.

Analisis klaster merupakan salah satu teknik analisis multivariat yang efektif untuk mengelompokkan individu berdasarkan kemiripan karakteristik tertentu [2]. Dalam konteks pendidikan, teknik ini telah banyak digunakan untuk mengungkap pola tersembunyi dalam data siswa dan mendukung pengambilan keputusan berbasis data. Beberapa penelitian terdahulu menerapkan algoritma K-Means dan DBSCAN untuk mengelompokkan siswa atau mahasiswa berdasarkan performa akademik maupun faktor sosial, sehingga membantu institusi pendidikan dalam merancang strategi pembelajaran dan intervensi yang lebih sesuai [3,4,5].

Namun, sebagian penelitian sebelumnya mentransformasikan seluruh data kategorikal ke dalam bentuk numerik menggunakan *one-hot encoding*, termasuk data ordinal. Pendekatan tersebut berpotensi menghilangkan informasi urutan pada data ordinal yang dapat memengaruhi akurasi perhitungan jarak dalam analisis klaster. Selain itu, perbandingan efektivitas metode analisis klaster dengan karakteristik pendekatan yang berbeda, khususnya dalam menangani data multivariat dan keberadaan *outlier*, masih relatif terbatas.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk melakukan analisis klaster profil siswa menggunakan algoritma K-Means dan DBSCAN dengan mempertahankan informasi urutan pada data ordinal melalui *ordinal/ label encoding*. Penelitian ini juga membandingkan efektivitas kedua metode dalam memetakan karakteristik siswa dan mendeteksi *outlier*. Hasil penelitian diharapkan dapat memberikan kontribusi empiris dalam pemahaman profil siswa serta menjadi dasar bagi pendidik dan pengambil kebijakan dalam merancang strategi pembelajaran dan intervensi pendidikan yang lebih tepat sasaran dan berbasis data.

2. TINJAUAN PUSTAKA

2.1. Analisis Klaster Multivariat (*Multivariate Clastering*)

Analisis klaster multivariat merupakan bagian dari teknik statistik dan *machine learning* yang digunakan untuk mengelompokkan objek-objek data berdasarkan kemiripan karakteristiknya dalam ruang multidimensi. Pendekatan ini termasuk dalam *unsupervised learning* karena tidak memerlukan label kelas sebelumnya. Dalam konteks statistik multivariat, teknik ini relevan saat sejumlah besar variabel harus dianalisis secara simultan untuk memahami keterkaitan antar variabel dan pengelompokan observasi secara efektif. Salah satu metode yang paling sering digunakan adalah K-Means, disamping metode hierarkis seperti *hierarchical clastering* dan *model-based clastering* yang dapat menangani bentuk struktur data yang berbeda-beda [6]. Penerapan analisis klaster multivariat saat ini semakin luas dan beragam, dari pengelompokan wilayah

berdasarkan indikator sosial-ekonomi hingga segmentasi pelanggan atau penyusunan strategi portfolio finansial [7, 8, 9].

2.2. Algoritma K-Means

Algoritma klaster K-Means adalah metode *unsupervised learning* yang termasuk dalam kelompok *partitioning clastering*, di mana tujuan utamanya adalah mempartisi sekumpulan data ke dalam k klaster sehingga setiap objek berada dalam klaster yang memiliki *centroid* terdekat secara jarak. Metode ini bekerja iteratif melalui dua langkah utama yakni menetapkan setiap titik data ke *centroid* terdekat dan memperbarui posisi *centroid* berdasarkan rata-rata anggota klaster tersebut. Secara matematis, optimasi K-Means didefinisikan sebagai meminimalkan *within-claster sum of squares* (WCSS**)** [10].

$$\sum_{k=1}^{K}\sum_{i\epsilon C_k}\sum_{j=1}^{n}\left(x_{ij}-c_{kj}\right)^2 \qquad (1)$$

Nilai WCSS menggambarkan tingkat variasi data di dalam setiap klaster, di mana nilai yang lebih besar menunjukkan dispersi data yang lebih tinggi. Seiring dengan bertambahnya jumlah klaster, nilai WCSS umumnya akan menurun karena data terbagi ke dalam kelompok yang semakin kecil dan lebih homogen. Penurunan WCSS pada awalnya bersifat signifikan, namun akan melambat setelah jumlah klaster tertentu. Pola ini membentuk grafik *elbow*, di mana titik siku menunjukkan kondisi ketika penambahan klaster tidak lagi memberikan pengurangan WCSS yang berarti. Titik tersebut digunakan sebagai dasar penentuan jumlah klaster optimal.

2.3. Algoritma DBSCAN

Algoritma DBSCAN merupakan metode pengelompokan data yang bekerja berdasarkan tingkat kepadatan data [11]. Metode ini mengelompokkan data dengan cara mengenali kumpulan titik yang saling berdekatan, tanpa harus menentukan jumlah kelompok sejak awal. DBSCAN memandang area yang memiliki banyak data berdekatan sebagai satu kelompok, sedangkan area dengan sedikit data dianggap sebagai data pencilan (*noise*). Dengan pendekatan ini, DBSCAN mampu membentuk kelompok dengan pola yang beragam dan tidak terbatas pada bentuk tertentu, berbeda dengan metode K-Means yang cenderung menghasilkan kelompok berbentuk bulat. DBSCAN mencari kelompok data dengan cara melihat apakah ada cukup banyak tetangga (MinPts) di sekitar suatu titik dalam jarak tertentu (ε). Jika cukup banyak, maka titik-titik tersebut membentuk satu kelompok. Jika suatu titik tidak punya cukup tetangga, titik itu dianggap sebagai *noise* (pencilan).

$$N_{\epsilon\, p}\colon \{q \mid d(p,q) \le \varepsilon\} \qquad (2)$$

2.4. Evaluasi Kualitas Klaster

Penelitian ini menggunakan salah satu metode evaluasi internal yakni *silhouette score*. Metode evaluasi ini digunakan untuk menilai kualitas analisis klaster berdasarkan tingkat kedekatan data dalam satu klaster dan keterpisahannya dengan klaster lain. Nilai *silhouette* untuk setiap data dihitung menggunakan selisih antara jarak rata-rata data terhadap klaster terdekat lainnya dan jarak rata-rata data terhadap klaster tempatnya berada, yang kemudian dinormalisasi [12].

$$Silhouette\ Score = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{max(a(i),b(i))} \qquad (3)$$

Nilai *silhouette* berada pada rentang −1 hingga 1. Semakin mendekati +1 artinya semakin dekat objek tersebut dengan klaster-nya sendiri atau termasuk pengelompokan sangat baik. Sebaliknya, semakin dekat ke -1artinya semakin jauh objek tersebut dari klaster-nya atau kemungkinan objek salah masuk kelompok.

2.5. Karakteristik Siswa

Karakteristik siswa merujuk pada berbagai atribut individu yang mencerminkan profil seorang peserta didik, baik dari segi demografis, perilaku, maupun aspek akademik. Atribut-atribut ini dapat berupa jenis kelamin, usia, latar belakang keluarga, gaya belajar, prestasi akademik, motivasi, dan tingkat partisipasi dalam kegiatan pembelajaran. Dalam beberapa studi, variabel-variabel profil siswa secara signifikan digunakan untuk mengelompokkan data dan menilai pola-pola pendidikan yang relevan, seperti dalam klasterisasi siswa

berdasarkan profil pelajar Pancasila atau korelasi antara profil siswa dan nilai akademik mereka, dimana atribut profil seperti pendidikan orang tua, penghasilan, dan nilai akademik dijadikan dasar klasterisasi untuk tujuan analisis pendidikan yang lebih dalam [13].

Pentingnya karakteristik siswa dalam analisis data pendidikan juga ditunjukkan melalui peranannya dalam strategi pedagogis, seperti pembelajaran berdiferensiasi yang mempertimbangkan perbedaan gaya belajar siswa untuk meningkatkan kolaborasi dan hasil belajar. Variasi karakteristik ini menunjukkan heterogenitas dalam populasi siswa yang dapat menghasilkan klaster-klaster dengan pola yang berbeda dalam analisis klaster multivariat, sehingga memberikan wawasan yang lebih kaya untuk pengambilan keputusan pendidikan berbasis data [14].

Berdasarkan tinjauan terhadap penelitian-penelitian terdahulu, sebagian besar studi analisis klaster dalam bidang pendidikan masih berfokus pada penggunaan algoritma K-Means dan variabel akademik semata. Penggunaan algoritma DBSCAN pada data pendidikan multivariat, khususnya yang mengombinasikan karakteristik akademik dan sosial siswa, masih relatif terbatas dan belum banyak dibandingkan secara langsung dengan metode partisi. Oleh karena itu, penelitian ini memposisikan diri untuk mengisi celah tersebut dengan menerapkan dan membandingkan K-Means dan DBSCAN pada data karakteristik siswa yang heterogen, sekaligus mengevaluasi kemampuan DBSCAN dalam menangani kepadatan data dan mendeteksi *outlier*.

3. METODOLOGI PENELITIAN

3.1. Data Penelitian

Penelitian ini merupakan penelitian kuantitatif deskriptif dengan pendekatan *data mining*. Data yang digunakan dalam penelitian ini bersumber dari dataset sintetis yang dikembangkan untuk tujuan pembelajaran dan eksperimen analisis data. Dataset tersebut diperoleh dari laman Kaggle, yang disediakan secara publik melalui Royce Kimmons' Generated Data Tools [15].

Dataset ini memuat informasi mengenai siswa tingkat Sekolah Menengah Atas (SMA) di Amerika Serikat dan dirancang untuk mengeksplorasi pengaruh latar belakang orang tua, persiapan ujian, serta faktor-faktor terkait lainnya terhadap performa akademik siswa. Dataset yang digunakan terdiri atas 1000 entri siswa dengan 8 fitur utama, yang meliputi:

1. Tingkat pendidikan orang tua
2. Ketersediaan makan siang yang dipengaruhi oleh kemampuan ekonomi keluarga (*lunch)*;
3. Persiapan ujian (*test preparation*);
4. Nilai akademik siswa pada beberapa mata pelajaran, seperti matematika, membaca, dan menulis;
5. Faktor demografis tambahan, antara lain jenis kelamin dan etnis.

Pemilihan dataset sintetis dalam penelitian ini dilakukan dengan beberapa pertimbangan. Pertama, penggunaan data sintetis memungkinkan eksplorasi dan evaluasi metode analisis klaster tanpa keterbatasan akses, privasi, dan etika yang sering melekat pada data pendidikan riil. Kedua, dataset sintetis memberikan kontrol yang lebih baik terhadap struktur data, variasi karakteristik, serta keberadaan *outlier*, sehingga sesuai untuk tujuan penelitian yang berfokus pada perbandingan kinerja algoritma analisis klaster, khususnya K-Means dan DBSCAN. Dengan demikian, dataset ini dipandang memadai untuk menguji kemampuan kedua metode dalam menangani data pendidikan multivariat yang heterogen.

3.2. Pra-proses Data

Dataset terlebih dahulu melalui tahap pra-pemrosesan data sebelum proses klasterisasi dilakukan untuk meningkatkan kualitas data dan menghasilkan klaster yang lebih optimal. Hasil pemeriksaan awal menunjukkan bahwa seluruh variabel yang digunakan tidak memiliki nilai hilang, sehingga tidak diperlukan penanganan *missing value*. Selanjutnya dilakukan seleksi variabel dengan mengeluarkan variabel *gender* dan *test preparation course* karena tidak secara langsung merepresentasikan latar belakang sosial, dukungan keluarga, etnis maupun kemampuan akademik siswa yang menjadi fokus penelitian.

Variabel kategorikal nominal dikodekan menggunakan teknik *one-hot encoding*, sedangkan variabel kategorikal ordinal ditransformasikan menggunakan *label* atau *ordinal encoding* untuk mempertahankan

informasi urutan. Terakhir, seluruh variabel numerik distandarisasi menggunakan metode *StandardScaler* untuk memastikan kontribusi setiap variabel seimbang dalam perhitungan jarak serta mencegah dominasi variabel dengan skala yang lebih besar. Standarisasi bekerja dengan mentransformasikan data sehingga memiliki nilai rata-rata nol dan standar deviasi satu.

$$z = \frac{x_i - \mu}{\sigma} \quad (4)$$

3.3. Proses Analisis Klaster

3.3.1. Analisis Klaster menggunakan K-Means

Penelitian ini menggunakan Metode *Elbow* untuk mendapatkan jumlah klaster yang optimal pada algoritma K-Means dengan langkah-langkah sebagai berikut:

1. Iterasi Nilai *k*

Algoritma K-Means dijalankan secara berulang dengan rentang nilai k tertentu. Penelitian ini menggunakan $k = 2$ hingga $k = 15$.

2. Perhitungan WCSS

Pada setiap iterasi k, dihitung nilai WCSS atau inersia. WCSS mengukur total kuadrat jarak antara setiap titik data dengan pusat klaster (*centroid*) yang bersesuaian.

3. Visualisasi Kurva

Nilai WCSS yang diperoleh kemudian diplot ke dalam grafik garis, di mana sumbu *X* mewakili jumlah klaster (k) dan sumbu *Y* mewakili nilai WCSS.

4. Identifikasi Titik Siku (*Elbow Point*)

Penurunan nilai WCSS akan sangat tajam pada nilai k yang kecil dan akan melandai secara signifikan setelah mencapai titik tertentu. Titik di mana penurunan mulai melambat secara drastis (membentuk struktur menyerupai siku tangan) dipilih sebagai jumlah klaster optimal.

5. Penerapan algoritma K-Means

Algoritma K-Means diterapkan pada dataset yang telah melalui tahap pra-peroses. Hal ini untuk menghasilkan struktur klaster awal yang merepresentasikan pengelompokan siswa berdasarkan karakteristik yang dianalisis. Selanjutnya, petakan kembali ke dataset asli untuk keperluan interpretasi profil dan analisis karakteristik setiap kelompok.

3.3.2. Analisis Klaster menggunakan DBSCAN

Algoritma DBSCAN mengelompokkan data berdasarkan area yang memiliki kepadatan tinggi dan memisahkan area dengan kepadatan rendah sebagai *noise*. Langkah-langkahnya adalah sebagai berikut:

1. Penentuan Parameter DBSCAN

Pada langkah awal adalah menentukan nilai MinPts berdasarkan dimensi data dan jumlah observasi. Selanjutnya menentunkan nilai ε dengan menggunakan grafik *k-distance*, di mana nilai k disesuaikan dengan MinPts yang telah ditentukan sebelumnya.

2. Pembentukan Klaster

Berdasarkan parameter yang telah ditentukan, setiap data diklasifikasikan menjadi titik inti (*core point*), titik batas (*border point*), atau noise. Titik inti memiliki jumlah tetangga minimal MinPts dalam radius ε, sedangkan *noise* merupakan data yang tidak termasuk dalam klaster manapun. Klaster dibentuk dengan mengelompokkan titik-titik inti yang saling terhubung secara kepadatan (*density-connected*). Titik batas akan dimasukkan ke klaster terdekat sementara titik-titik yang tidak memenuhi syarat kepadatan dan tidak terjangkau oleh klaster mana pun akan diberi label sebagai -1 atau *noise*.

3.4. Evaluasi

Tahap akhir dari metodologi ini adalah evaluasi performa algoritma menggunakan *silhouette score* untuk mengukur tingkat kohesi dan separasi klaster yang dihasilkan oleh K-Means dan DBSCAN. Perbandingan dilakukan dengan menganalisis nilai rata-rata *silhouette* dari masing-masing model. Algoritma dengan skor tertinggi ditetapkan sebagai metode yang paling optimal dalam merepresentasikan profil siswa.

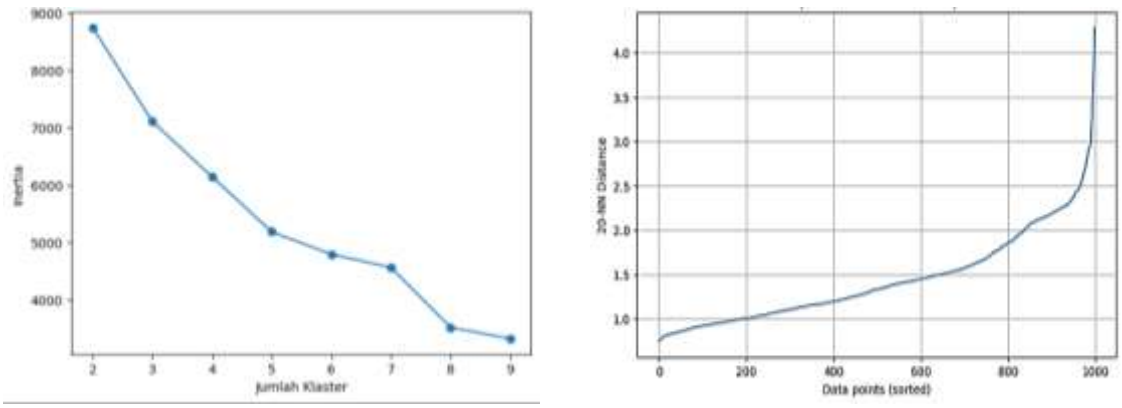
4. HASIL DAN PEMBAHASAN

Tahap pra-proses menghasilkan peningkatan jumlah variabel menjadi 10 variabel dari 6 variabel awal yang dipilih. Hal ini dikarenakan adanya variabel *dummy* yang dihasilkan dari *one-hot encoding*.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   parental level of education  1000 non-null   int64
 1   lunch                        1000 non-null   int64
 2   math score                   1000 non-null   int64
 3   reading score                1000 non-null   int64
 4   writing score                1000 non-null   int64
 5   race/ethnicity_group A       1000 non-null   bool
 6   race/ethnicity_group B       1000 non-null   bool
 7   race/ethnicity_group C       1000 non-null   bool
 8   race/ethnicity_group D       1000 non-null   bool
 9   race/ethnicity_group E       1000 non-null   bool
dtypes: bool(5), int64(5)
memory usage: 44.1 KB
```

Gambar 1. Ringkasan Data Hasil Pra-Proses

Pada metode K-Means, parameter *k* ditetapkan sebesar 8 berdasarkan titik siku pada kurva inertia yang menunjukkan jumlah klaster optimal. Sementara itu, pada metode DBSCAN, nilai MinPts ditentukan sebesar 20 atau 2 kali dari jumlah variabel, dengan perubahan kemiringan grafik *k-distance* yang signifikan terjadi pada jarak sekitar ε = 2.5. Penelitian ini menetapkan satu nilai parameter optimal untuk masing-masing metode guna menjaga fokus pada perbandingan performa antar-algoritma secara langsung. Meskipun variasi parameter khususnya pada DBSCAN dapat memberikan wawasan tambahan mengenai stabilitas model, analisis mengenai sensitivitas variabel tersebut berada di luar lingkup penelitian ini dan diposisikan sebagai penguatan bagi studi lanjutan di masa mendatang.



(a) (b)

Gambar 2. (a) Metode Elbow (*Elbow Method*) pada K-Means; (b) Grafik *k-distance* pada DBSCAN

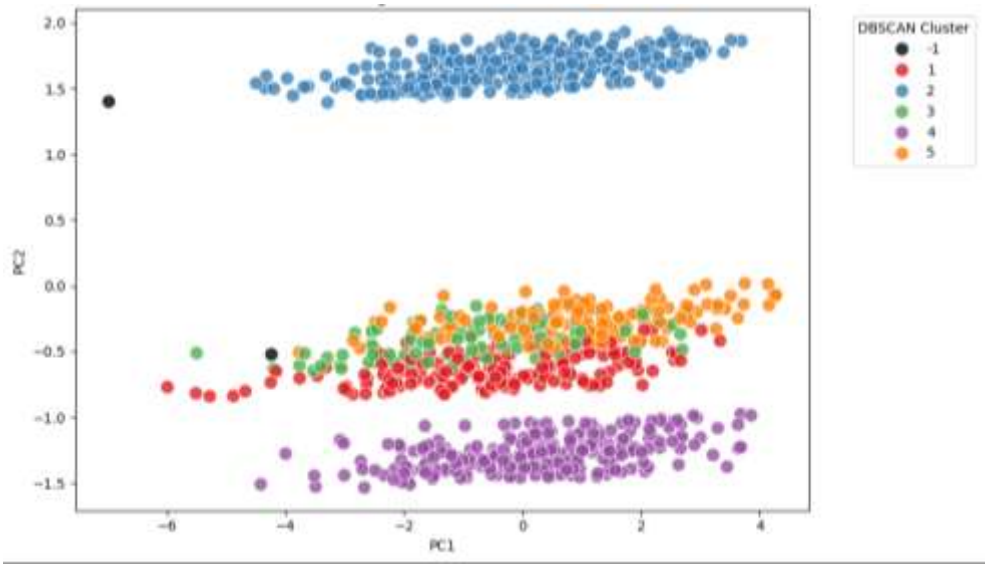
Perbandingan hasil analisis klaster siswa menggunakan metode K-Means dan DBSCAN menghasilkan jumlah klaster yang berbeda seperti yang tersaji pada Tabel 1. Meskipun metode K-Means menghasilkan jumlah klaster yang lebih banyak, hasil evaluasi menunjukkan bahwa DBSCAN memiliki struktur klaster yang lebih baik dan stabil. Hal ini ditunjukkan melalui nilai *silhouette score* yang lebih tinggi pada algoritma DBSCAN.

Tabel 1. Perbandingan Hasil Analisis Klaster

Algoritma Klaster	**n Klaster**	***Silhouette Score***
K-Means	8	0,3

DBSCAN	5	0,4

Berdasarkan perbandingan jumlah klaster dan nilai *silhouette score*, DBSCAN dipilih karena mampu menghasilkan klaster yang lebih mudah diinterpretasikan dalam pemetaan profil siswa. Jumlah klaster yang lebih sedikit memungkinkan pengelompokan tingkat kemampuan siswa yang lebih jelas dan aplikatif dalam konteks pendidikan. Selain itu, metode DBSCAN memiliki nilai tambah berupa kemampuan dalam mengidentifikasi siswa *outlier* (data pencilan) yang tidak dapat terdeteksi oleh metode K-Means.



Gambar 3. Klaster Algoritma DBSCAN

Hasil analisis klaster menggunakan algoritma DBSCAN berhasil memetakan lima profil siswa yang menunjukkan korelasi linear antara dukungan sosio-ekonomi dan capaian akademik seperti yang tersaji pada Tabel 2. Nilai pada kolom *parental level of education*, *lunch*, dan *academic* merupakan nilai rata-rata (mean) yang menggambarkan pusat massa dari masing-masing kelompok.

Tabel 2. Profil Karakteristik Rata-Rata Setiap Klaster Hasil Algoritma DBSCAN

Klaster DBSCAN	***parental level of education***	***lunch* (keamampuan ekonomi)**	***academic* (matematika, membaca, menulis)**	**Etnis Dominan**	**Jumlah Siswa**
Klaster 1	2,87	0,637	65,5	B	190
Klaster 2	3,17	0,645	67,3	C	318
Klaster 3	2,79	0,596	63,0	A	89
Klaster 4	3,12	0,637	69,2	D	262
Klaster 5	3,31	0,705	73,1	E	139

Klaster 5 merupakan kelompok dengan profil unggul dibandingkan klaster lainnya. Klaster ini dikategorikan sebagai *high achievers*, memiliki rata-rata skor akademik tertinggi dan dukungan ekonomi serta pendidikan orang tua yang paling tinggi. Kelompok ini didominasi oleh etnis E. Sebaliknya, klaster 3 merepresentasikan kategori *underprivileged / at-risk*, dengan rata-rata skor akademik, dukungan ekonomi, dan pendidikan orang tua terendah. Klaster ini memiliki jumlah anggota paling sedikit dan didominasi oleh etnis A. Sementara itu, klaster 2 menjadi representasi mayoritas siswa dengan profil *moderat* sekaligus merupakan kelompok terbesar. Klaster 1 dan klaster 4 masing-masing dapat dikategorikan sebagai kelompok *lower-middle class* dan *upper-middle class*.

Pengelompokan siswa berdasarkan karakteristik akademik dan non-akademik memungkinkan pendidik dan pengambil kebijakan untuk mengidentifikasi kelompok siswa dengan kebutuhan yang berbeda secara lebih objektif dan berbasis data. Klaster dengan capaian akademik rendah dan keterbatasan dukungan sosial dapat diprioritaskan untuk menerima pendampingan akademik intensif dan bantuan ekonomi, sementara klaster dengan capaian akademik tinggi dan dukungan yang memadai dapat difokuskan pada program pengayaan dan pengembangan potensi. Pendekatan ini membantu menghindari generalisasi perlakuan terhadap seluruh siswa serta mendukung prinsip keadilan dalam perencanaan intervensi pendidikan.

Penting untuk ditegaskan bahwa meskipun variabel etnisitas dilibatkan dalam proses pembentukan klaster, penelitian ini tidak mengasumsikan adanya hubungan kausalitas antara latar belakang rasial dengan kapasitas intelektual siswa. Kehadiran pola etnisitas yang kuat dalam klaster-klaster yang dihasilkan lebih dipandang sebagai refleksi dari ketimpangan struktural dan sosio-ekonomi yang berlangsung dalam jangka waktu yang lama. Oleh karena itu, hasil penelitian ini harus dimaknai sebagai alat diagnostik untuk mengidentifikasi kelompok siswa yang memerlukan dukungan tambahan, terlepas dari latar belakang identitas mereka, guna menciptakan lingkungan pendidikan yang lebih inklusif.

5. KESIMPULAN DAN SARAN

Penelitian ini menyimpulkan bahwa metode analisis klaster multivariat efektif untuk memetakan profil siswa. Algoritma DBSCAN menunjukkan performa yang lebih unggul dibandingkan K-Means dalam menangani karakteristik data yang kompleks serta mengidentifikasi outlier. Hasil analisis berhasil mengidentifikasi lima profil utama mulai dari *high achiever* hingga kelompok *at-risk*. Pemetaan klaster ini memberikan dasar data yang objektif mengenai struktur sosiodemografi dalam lingkungan pendidikan yang heterogen.

Kontribusi utama penelitian ini adalah menunjukkan secara empiris bahwa penerapan DBSCAN pada data pendidikan multivariat, dengan mempertahankan informasi ordinal, mampu menghasilkan segmentasi profil siswa yang lebih adaptif dan informatif dibandingkan metode klaster berbasis partisi**.**

SARAN

Hasil analisis klaster direkomendasikan sebagai instrumen intervensi pendidikan yang tepat sasaran melalui penyediaan bantuan ekonomi dan pendampingan akademik bagi kelompok rentan, tanpa menjadikan identitas demografis sebagai determinan kemampuan individu. Untuk pengembangan penelitian selanjutnya, sangat disarankan penggunaan data pendidikan riil guna memvalidasi dan memperkuat temuan yang diperoleh dalam studi ini. Selain itu, integrasi variabel non-akademik seperti motivasi belajar dan perilaku belajar, serta eksplorasi pendekatan klastering hibrida atau optimasi parameter algoritma, berpotensi meningkatkan akurasi dan ketajaman segmentasi profil siswa.

DAFTAR PUSTAKA

[1] OECD. (2023, December 5). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*.[On-line]. 1. Available: https://www.oecd.org/en/publications/pisa-2022-results-volume-i_53f23881-en.html. [Nov. 2, 2025].

[2] G. James. D. Witten. T. Hastie. and R. Tibshirani.(2021). *An Introduction to Statistical Learning with Applications in R* (*Latest edition*). [On-line]. Available : https://link.springer.com/book/10.1007/978-1-0716-1418-1. [Nov. 13, 2025]

[3] C. Azzahra and S. Sriani. "Clustering of High School Students Academic Scores Using K-Means Algorithm." *J. Inf. Syst. Informatics*, vol. 7, no. 1, pp. 572–586, 2025, https://doi.org/10.47738/ijaim.v5i3.109.

[4] A. F. M. 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," *Appl. Sci.*, vol. 12, no. 19, p. 9467, 2022, https://doi.org/10.51519/journalisi.v7i1.1029.

[5] F. Rahma and S. Z. Ulfah, "Clustering Students Based on Academic Performance and Social Factors: An Unsupervised Learning Approach to Identify Student Patterns Title," *Int. J. Appl. Inf. Manag.*, vol. 5, no. 3, pp. 139–154, 2025, https://doi.org/10.3390/app12199467.

[6] D. Saputra, A. Ardania, S. Putri, A. T. J. A. Asri, and L. Harsyiah, "Analisis Cluster untuk Pengelompokan Provinsi di Indonesia berdasarkan Tingkat Kemiskinan menggunakan Metode Average Linkage," *Indones. J. Appl. Stat. Data Sci.*, vol. 1, no. 1, pp. 20–30, 2024, https://doi.org/10.29303/ijasds.v1i1.5446.

[7] A. Maulana, A. Maulana, E. Susanti, E. S. Rahmaya, and I. A. Haris, "Penerapan Algoritma K-Means Clustering Untuk Segmentasi Wilayah Berdasarkan Indikator Ekonomi di Indonesia," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 4055–4062, 2025, https://doi.org/10.31004/riggs.v4i2.1169.

[8] F. Amanah, F. Roshafara, P. I. Lestari, S. Salsabila, and R. Maharani, "Utilizing K-Means Clustering for Constructing Black-Litterman Portfolio Models," *J. Mat. Stat. Dan Komputasi*, vol. 20, no. 3, pp.

679–679, 2024, https://doi.org/10.20956/j.v20i3.34165.
[9] B. Sinaga, "Multivariate Data Analysis for Customer Segmentation Using Principal Component Analysis and K-Means Clustering," *J. Info Sains Inform. Dan Sains*, vol. 15, no. 1, pp. 283–291, 2025, https://doi.org/10.54209/infosains.v15i01.
[10] D. S. TURAN, "A New Method for Determining the Number of Clusters Without Clustering," *Yuz. Yil Univ. J. Inst. Nat. Appl. Sci.*, vol. 30, no. 2, pp. 596–607, 2025, https://doi.org/10.53433/yyufbed.1612608.
[11] P. Jain, M. S. Bajpai, and R. Pamula, "A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality," *Int. Arab J. Inf. Technol.*, vol. 19, no. 1, pp. 23–28, 2022, https://doi.org/10.34028/IAJIT/19/1/3.
[12] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *Int. J. online Biomed. Eng.*, vol. 19, no. 04, pp. 174–182, 2023, https://doi.org/10.3991/IJOE.V19I04.37059.
[13] M. D. Cahya, Y. Pamungkas, and E. N. Faiqoh, "Analisis Karakteristik Siswa sebagai Dasar Pembelajaran Berdiferensiasi terhadap Peningkatan Kolaborasi Siswa," *Bioma J. Biol. Dan Pembelajaran Biol.*, vol. 8, no. 1, pp. 31–45, 2023, https://doi.org/10.32528/bioma.v8i1.372.
[14] A. Syahfitri, Novriyenni, and I. Gultom, "Pengelompokan Data Siswa Berdasarkan Profil Pelajar Pancasila Menggunakan Metode Clustering(Studi Kasus SMK Putra Anda Binjai)," *Indones. J. Educ. Comput. Sci.*, vol. 2, no. 2, pp. 107–120, 2024, https://doi.org/10.60076/indotech.v2i2.644.
[15] R. Kimmons. "Students Performance Dataset," Internet: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams,2023[Nov. 2, 2025].

NOMENKLATUR

K	: jumlah klaster
C_k	: *centroid* (titik pusat) di klaster k
x_{ij}	: nilai pada variabel ke-j dari data ke-i yang berada dalam klaster k
c_{ij}	: nilai pada variabel ke-j dari titik pusat (centroid) klaster k.
n	: banyaknya variabel
$N_{\in p}$	: himpunan tetangga (ε-*neighborhood*) dari titik p
p	: titik data yang sedang dievaluasi
q	: titik data lain di dalam dataset
$d(p,q)$	: jarak antara titik p dan titik q
ε	: radius atau jarak maksimum untuk menentukan tetangga
N	: jumlah total data
i	: indeks data ke-iii, dengan i=1,2,…,N
$a(i)$	: rata-rata jarak antara data ke-i dengan seluruh data lain yang berada dalam klaster sama
$b(i)$	: rata-rata jarak terkecil antara data ke-i dengan data-data pada klaster terdekat lainnya
$max(a(i),b(i))$	: nilai maksimum antara $a(i)$ dan $b(i)$
z	: nilai hasil standardisasi (Z-*score*)
x_i	: nilai data ke-i
μ	: nilai rata-rata (mean) dari seluruh data
σ	: simpangan baku (standar deviasi) data