

Klasterisasi Multidimensional Terhadap Preferensi Bidang Studi Prospektif Mahasiswa

1st Muhammad Fadhil Alviano
Program Studi SI Sains Data
Telkom University Surabaya
Surabaya, Indonesia

fadhilalviano@student.telkomuniversit
y.ac.id

2nd Rifdatun Ni'mah. S.Si., M.Si.
Program Studi SI Sains Data
Telkom University Surabaya
Surabaya, Indonesia

rifdatun@telkomuniversity.ac.id

3rd Regita Putri Permata, S.Stat.,
M.Stat.

Program Studi SI Sains Data
Telkom University Surabaya
Surabaya, Indonesia

regitapermata@telkomuniversity.ac.id

Abstrak — Fenomena salah jurusan di Indonesia masih menjadi masalah serius karena berdampak pada rendahnya motivasi belajar dan ketidaksesuaian karier, yang umumnya disebabkan oleh kurangnya pemahaman siswa terhadap minat, bakat, dan kemampuan akademik. Penelitian ini bertujuan mengelompokkan calon mahasiswa ke dalam rumpun studi yang sesuai menggunakan pendekatan klasterisasi multidimensi berbasis data. Data berasal dari 528 siswa SMA NU 1 Gresik, mencakup nilai akademik serta 24 indikator non-akademik terkait minat, bakat, dan hobi. Tahapan analisis meliputi normalisasi data, reduksi dimensi dengan Principal Component Analysis (PCA), serta penerapan tiga algoritma klasterisasi, yaitu K-Means, Hierarchical Clustering, dan DBSCAN. Evaluasi dilakukan menggunakan Silhouette Score dan Davies-Bouldin Index (DBI). Hasil menunjukkan bahwa Hierarchical Clustering memiliki kinerja terbaik (Silhouette 0.189; DBI 1.874), diikuti oleh K-Means (Silhouette 0.184; DBI 1.915), sedangkan DBSCAN kurang optimal karena menghasilkan nilai Silhouette negatif dan DBI tinggi. Secara umum, terbentuk dua klaster utama, yaitu rumpun Saintek dan Soshum, yang berpotensi menjadi dasar sistem rekomendasi pemilihan jurusan secara objektif dan personal.

Kata kunci— Klasterisasi Multidimensi, Minat Bakat, Rekomendasi Jurusan, Rumpun Studi.

I. PENDAHULUAN

Masa remaja merupakan fase penting pada tahap ini, individu mulai meng- hadapi berbagai pilihan memengaruhi masa depan mereka, termasuk dalam aspek pendidikan, karir, dan pengembangan diri. Salah satu keputusan pen- ting adalah melanjutkan pendidikan ke jenjang perguruan tinggi[1]. Meskipun idealnya proses pengambilan keputusan untuk melanjutk- an pendidikan tinggi harus didasarkan pada keselarasan antara minat, bakat, dan potensi individu, kenyataan di lapangan menunjukkan kondisi yang sangat kontradiktif. Kegagalan dalam mengintegrasikan aspek-aspek personal terse- but tercermin dalam data yang dirilis oleh Indonesia Career Centre Network (ICCN), yang mengungkapkan bahwa mayoritas mahasiswa di Indonesia— mencapai 87 persen—merasa telah mengambil keputusan yang salah dalam memilih jurusan[2].

Masalah signifikan di kalangan mahasiswa digambarkan secara nyata me- lalui banyaknya perasaan bahwa jurusan

yang dipilih tidak mencerminkan minat atau potensi mereka secara maksimal, yang berakar dari pengambil- an keputusan yang tidak tepat[3]. Permasalahan jurusan yang dipilih tidak mencerminkan minat atau potensi mahasiswa menjadikan pemahaman terhadap konsep pengelompokan rumpun studi penting sebagai landasan awal dalam proses pemilihan jurusan yang lebih terarah dan sesu- ai dengan minat serta potensi mahasiswa. Pengelompokan jurusan ke dalam rumpun-rumpun studi diperlukan untuk menciptakan sistem pendidikan yang lebih terstruktur, mempermudah pemetaan keilmuan, serta memastikan kesi- nambungan dan keterkaitan antar bidang studi dalam satu kelompok keilmuan yang serupa guna menciptakan struktur akademik yang lebih terorganisir dan mendukung pengembangan keilmuan secara berkelanjutan.

Salah satu cara melakukan pengelompokan dengan pendekatan objektif dan berbasis data adalah dengan menerapkan teknik data mining diantaranya menggunakan algoritma clustering[4]. Clustering adalah teknik dalam data mining yang digunakan untuk mengelompokkan data ber- dasarkan kesamaan karakteristik, tanpa memerlukan label atau kategori sebe- lumnya. Salah satu algoritma Clustering yang paling populer adalah K-Means, yang bekerja dengan membagi data ke dalam sejumlah klaster berdasarkan kedekatan nilai terhadap pusat klaster centroid. Pendekatan clustering dapat membantu dalam memberikan rekomendasi akademik yang lebih terarah dan mengurangi risiko salah jurusan[5]. Selain K-Means terdapat algoritma lain yang digunakan dalam pengelompokan data yaitu, Hierarchi- cal Clustering adalah metode pengelompokan data yang membentuk struktur hierarki klaster dalam bentuk pohon yang disebut dendrogram[6].

Clustering merupakan salah satu metode pengelompokan data yang ba- nyak digunakan dalam proses pengambilan keputusan berbasis data, terma- suk dalam konteks pemilihan jurusan pendidikan tinggi. Keunggulan utama dari metode clustering terletak pada kemampuannya untuk mengeksplorasi ruang data multidimensi, sehingga mampu menangkap kompleksitas struktur data dan mengidentifikasi pola-pola tersembunyi yang tidak dapat ditemuk- an apabila hanya mempertimbangkan satu atau dua dimensi saja[7]. Dalam konteks studi pendidikan, pendekatan

ini dinilai relevan karena memungkinkan terbentuknya kelompok-kelompok mahasiswa berdasarkan kesamaan minat, bakat, maupun kemampuan akademik, yang kemudian dapat digunakan untuk tujuan bimbingan karier dan studi lanjut secara lebih personal dan kontekstual. Penggunaan clustering secara multidimensi juga telah terbukti mendukung proses pengelompokan yang lebih akurat dan prediktif terhadap keberhasilan akademik mahasiswa di masa depan[8]. Berdasarkan dari studi kasus yang diangkat dan penelitian terdahulu, maka penelitian ini menggunakan model clustering secara multidimensi untuk memprediksi rumpun studi yang sesuai dan terprofil bagi siswa kelas 11 dan 12 di SMA Nahdlatul Ulama 1 Gresik.

II. KAJIAN TEORI

A. Rumpun Studi

Menurut Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 154 Tahun 2014, rumpun ilmu pengetahuan dan teknologi dibagi menjadi enam kelompok utama, yaitu: (1) rumpun ilmu agama, (2) rumpun ilmu humaniora, (3) rumpun ilmu sosial, (4) rumpun ilmu alam, (5) rumpun ilmu formal, dan (6) rumpun ilmu terapan[9]. Setiap rumpun ini menaungi sejumlah cabang ilmu atau disiplin yang memiliki karakteristik dan metode yang saling berkaitan. Pertama, rumpun ilmu agama, yang mencakup bidang-bidang seperti Studi Islam, Teologi Kristen, Pendidikan Agama Hindu dan Filsafat Keagamaan.

Kedua, rumpun ilmu humaniora, yang meliputi Sastra, Sejarah, Filsafat Umum, Bahasa, Budaya, dan Arkeologi. Ketiga, rumpun ilmu sosial, yang terdiri atas Sosiologi, Ilmu Politik, Ekonomi, Ilmu Komunikasi, Hubungan Internasional, dan Antropologi. Keempat, rumpun ilmu alam, yang mencakup Fisika, Kimia, Biologi, Geografi Fisik, dan Ilmu Kelautan. Kelima, rumpun ilmu formal, meliputi Matematika, Statistika, Ilmu Komputer, dan Ilmu Aktuaria.

Terakhir, rumpun ilmu terapan, yang terdiri atas bidang-bidang seperti Teknik Sipil, Arsitektur, Teknologi Pangan, Kedokteran, Keperawatan, Farmasi, dan Ilmu Pertanian. Pengelompokan ini penting agar mahasiswa dapat merancang jalur pendidikan dan karier yang linier dan konsisten, terutama ketika melanjutkan studi ke jenjang lebih tinggi atau mengajukan jabatan akademik[10].

B. Sampling

Purposive sampling adalah metode pemilihan sampel secara sengaja berdasarkan kriteria-kriteria tertentu yang ditetapkan oleh peneliti sesuai tujuan penelitian. Penggunaan teknik *purposive* sampling dianggap tepat karena memungkinkan peneliti untuk memperoleh data dari kelompok siswa yang relevan dan memiliki data lengkap yang dibutuhkan dalam proses klusterisasi multidimensi[11]. Selain itu, distribusi kuesioner dilakukan secara daring untuk memudahkan akses dan menjangkau seluruh siswa dari 16 kelas yang tersebar di dua jenjang tersebut. Estimasi jumlah siswa kelas 11 dan 12 kurang lebih 584 siswa.

C. Uji Validitas dan Reliabilitas

Metode yang umum digunakan untuk mengukur reliabilitas adalah teknik Cronbach's Alpha[12]. *Cronbach's Alpha* (α) merupakan koefisien reliabilitas yang menunjukkan tingkat konsistensi internal antar item dalam suatu instrumen. Nilai α berada dalam rentang 0 hingga 1, dengan interpretasi bahwa semakin mendekati 1 maka

instrumen tersebut semakin reliabel. Menurut Sugiyono[13], nilai $\alpha \geq 0,6$ dianggap cukup reliabel untuk penelitian eksploratif, sedangkan nilai di atas 0,7 lebih disarankan untuk penelitian konfirmatif.

Apabila terdapat nilai α di bawah ambang batas, maka dilakukan evaluasi terhadap kontribusi tiap item. Item yang memiliki korelasi item-total yang rendah dapat dipertimbangkan untuk direvisi atau dihilangkan agar reliabilitas instrumen meningkat dan konsistensi internal dapat terjaga.

D. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) merupakan metode reduksi dimensi yang digunakan untuk menyederhanakan jumlah variabel dalam suatu dataset tanpa menghilangkan informasi penting yang terkandung di dalamnya. PCA bekerja dengan mentransformasi variabel awal menjadi sejumlah variabel baru yang disebut *principal components*, yaitu variabel-variabel yang saling bebas (*orthogonal*) dan diurutkan berdasarkan besar kontribusi variansinya.

E. K-Means

K-Means merupakan salah satu algoritma *unsupervised learning* yang paling populer dan banyak digunakan dalam proses klusterisasi data. Algoritma ini bertujuan untuk mengelompokkan data ke dalam sejumlah klaster berdasarkan kemiripan fitur yang dimiliki oleh masing-masing entitas data[14]. Proses pengelompokan dilakukan dengan meminimalkan variasi intraklaster, yaitu jarak antara setiap titik data terhadap pusat klaster (*centroid*), sehingga data dalam satu klaster memiliki karakteristik yang relatif homogen[15].

F. Mean Vector

Mean vector (vektor rata-rata) merupakan representasi matematis yang digunakan untuk merangkum karakteristik pusat dari data multivariat. Setiap objek data direpresentasikan sebagai sebuah vektor yang terdiri dari beberapa variabel atau fitur, misalnya nilai mata pelajaran dan indikator minat siswa. *Mean vector* diperoleh dengan menghitung nilai rata-rata untuk setiap variabel secara terpisah pada seluruh objek data atau pada anggota suatu klaster tertentu.

G. Hierarchical Clustering

Hierarchical Clustering adalah metode pengelompokan yang membentuk hierarki klaster melalui pendekatan *agglomerative (bottom-up)* atau *divisive (top-down)*. Dalam penelitian ini, pendekatan *agglomerative* lebih banyak digunakan karena lebih mudah diimplementasikan dan sesuai untuk data multidimensi dari calon mahasiswa. Metode *linkage* digunakan untuk menentukan jarak antar dua klaster. Terdapat beberapa metode *linkage* yang umum digunakan[16].

H. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN merupakan salah satu metode *unsupervised clustering* yang mengelompokkan data berdasarkan kepadatan (*density*) titik-titik data dalam ruang fitur. Berbeda dengan metode seperti *K-Means* atau Hierarchical Clustering, DBSCAN tidak memerlukan penentuan jumlah

klaster di awal dan mampu menangani bentuk klaster yang tidak beraturan serta data yang mengandung noise[17].

DBSCAN sangat berguna dalam konteks analisis pendidikan karena dapat mengelompokkan calon mahasiswa berdasarkan karakteristik unik mereka (misalnya nilai akademik, minat, atau latar belakang) tanpa perlu asumsi awal jumlah kelompok. Metode ini juga memungkinkan deteksi siswa yang memiliki profil sangat berbeda dari kebanyakan (*outlier*), yang bisa menjadi perhatian khusus dalam proses seleksi atau pembinaan[17].

I. Evaluasi

Evaluasi model clustering diperlukan untuk menilai sejauh mana hasil pengelompokan merepresentasikan struktur alami dalam data. Dalam penelitian ini, dua metode evaluasi yang digunakan adalah Silhouette Score (SS) dan Davies-Bouldin Index (DBI). Kedua metode ini tidak memerlukan label atau ground truth, sehingga cocok untuk metode unsupervised seperti clustering (Hashemi et al. 2023)). Silhouette Score mengukur kualitas dari klaster yang terbentuk dengan mempertimbangkan dua aspek utama, yaitu cohesion (kekompakan) dan separation (pemisahan antar klaster).

III. METODE

Penelitian ini menggunakan pendekatan kuantitatif berbasis data-driven analysis dengan mengintegrasikan analisis sentimen dan Social Network Analysis (SNA). Data bersumber dari ulasan pengguna Google Maps terhadap 26 masjid wisata religi di Jawa Timur pada periode 2021–2025. Metodologi penelitian terdiri atas tahapan pengumpulan data, pra-pemrosesan teks, pelabelan sentimen, ekstraksi fitur, klasifikasi sentimen, pemodelan topik, konstruksi jaringan sosial, analisis sentralitas, serta visualisasi dan interpretasi hasil.

A. Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan melalui survei langsung ke SMA Nadhlatul Ulama 1 Kabupaten Gresik dengan tujuan memperoleh dua jenis data utama, yaitu data akademik dan non-akademik siswa. Data non-akademik diperoleh melalui penyebaran kuesioner yang dirancang khusus untuk mengukur dimensi minat, hobi, dan bakat siswa sebagai representasi aspek afektif. Instrumen kuesioner terdiri dari sejumlah pernyataan yang dijawab menggunakan skala Likert, sehingga memungkinkan pengukuran preferensi siswa secara terstruktur. Sementara itu, data akademik dikumpulkan dalam bentuk nilai rata-rata mata pelajaran dari semester 1 hingga semester 4 meliputi (Matematika, Biologi, Kimia, Fisika, Bahasa Indonesia, Bahasa Inggris, PPKN, Agama, Olahraga, Seni, Sejarah), yang mencerminkan performa kognitif siswa secara kuantitatif. Kedua jenis data ini kemudian digunakan secara terpadu sebagai masukan dalam proses klasterisasi untuk membentuk model pengelompokan rumpun studi berdasarkan karakteristik multidimensi siswa.

B. Kuisisioner

Pengumpulan data kuisisioner non-akademik dalam penelitian ini dilakukan melalui penyebaran kuisisioner digital menggunakan platform Google Form. Metode ini dipilih karena dinilai praktis, efisien, dan dapat menjangkau

responden dalam jumlah besar secara serentak. Tautan kuisisioner dibagikan kepada siswa melalui saluran komunikasi internal sekolah dengan dukungan dari pihak guru dan wali kelas. Penyebaran kuisisioner ditargetkan kepada siswa dari 16 kelas, yang terdiri dari 8 kelas tingkat XI dan 8 kelas tingkat XII, dengan estimasi total populasi sebanyak 584 siswa. Pemilihan dua jenjang ini bertujuan agar data yang dikumpulkan mencerminkan kondisi siswa menjelang masa penentuan jurusan kuliah, sekaligus mendapatkan keberagaman profil responden dari dua angkatan berbeda.

Instrumen kuisisioner dirancang untuk menggali aspek minat, bakat, dan hobi siswa, yang merupakan representasi dari variabel non-akademik. Pertanyaan dalam kuisisioner disusun dalam bentuk pernyataan dengan skala Likert untuk mengukur tingkat kecenderungan responden terhadap berbagai jenis aktivitas atau bidang studi tertentu. Aspek ini penting sebagai landasan dalam proses klasterisasi multidimensi yang mempertimbangkan faktor afektif dalam penentuan rumpun studi.

TABEL 1
(A)

Skor	Kategori Penilaian
5	Sangat Setuju
4	Cukup Setuju
3	Setuju
2	Kurang Setuju
1	Sangat Kurang Setuju

Dalam penyusunan kuisisioner, jumlah item yang digunakan untuk mengukur suatu konstruk (seperti minat, bakat, dan hobi) tidak harus terlalu banyak, tetapi harus cukup untuk menangkap variasi dimensi yang relevan dari konstruk tersebut. Menurut Sugiyono (2019) dalam buku Metode Penelitian Kuantitatif, Kualitatif, dan RD, instrumen pengukuran dianggap baik jika setiap variabel diwakili oleh minimal 3 hingga 10 indikator yang relevan. Dengan demikian, penyusunan sebanyak 8 butir pertanyaan untuk masing-masing topik telah memenuhi prinsip kecukupan dalam mengukur konstruk laten secara statistik dan mendukung proses analisis data lebih lanjut secara valid dan reliabel.

TABEL 1
(B)

No	Pernyataan
1	Saya tertarik untuk mempelajari lebih dalam tentang ilmu komputer dan teknologi.
2	Saya menyukai kegiatan yang berkaitan dengan analisis data atau pemecahan masalah.
3	Saya merasa senang mengikuti pelajaran yang melibatkan diskusi kelompok atau debat.
4	Saya memiliki minat tinggi terhadap mata pelajaran seni seperti musik, tari, atau seni rupa.
5	Saya tertarik mengikuti perkembangan isu sosial dan politik.
6	Saya menyukai kegiatan eksperimen atau praktikum di laboratorium.
7	Saya sering mengikuti seminar, webinar, atau
8	Saya tertarik untuk mempelajari lebih dalam tentang ilmu komputer dan teknologi.

	pelatihan yang sesuai dengan bidang yang saya sukai. Saya memiliki ketertarikan untuk melanjutkan studi di jurusan tertentu sejak awal SMA.
--	---

TABEL 1
(C)

No	Pernyataan
9	Saya memiliki kemampuan menggambar atau mendesain secara visual dengan baik.
10	Saya mampu memahami konsep matematika atau logika lebih cepat dibanding teman sebaya.
11	Saya memiliki kemampuan berbicara di depan umum dengan percaya diri.
12	Saya memiliki kemampuan dalam bermain alat musik atau menyanyi.
13	Saya memiliki keahlian dalam menggunakan aplikasi komputer atau perangkat lunak tertentu.
14	Saya mampu memimpin kelompok dan mengatur pembagian tugas secara efektif.
15	Saya menunjukkan kemampuan tinggi dalam menulis esai, puisi, atau karya tulis lainnya.
16	Saya sering menyelesaikan tugas dengan kualitas tinggi tanpa perlu banyak bimbingan.

TABEL 1
(D)

No	Pernyataan
17	Saya senang menghabiskan waktu luang dengan menonton video
18	atau membaca tentang topik tertentu.
19	Saya memiliki kegiatan rutin yang saya lakukan setiap minggu sebagai hobi.
20	Saya senang merakit, membuat, atau memodifikasi sesuatu secara manual (DIY).
21	Saya aktif dalam kegiatan fotografi,
22	videografi, atau konten digital. Saya memiliki minat tinggi dalam bermain atau membuat game. Saya mengikuti kegiatan olahraga tertentu secara teratur.
23	Saya memiliki blog, media sosial, atau catatan pribadi untuk mengekspresikan minat saya.
24	Saya berpartisipasi dalam komunitas atau ekstrakurikuler yang sesuai dengan hobi saya.

IV. HASIL DAN PEMBAHASAN

Metode pengumpulan data dilakukan dengan menggunakan platform Google Form karena dinilai praktis, efisien, dan mampu menjangkau responden dalam jumlah besar secara serentak tanpa keterbatasan ruang dan waktu. Pelaksanaan pengumpulan data di lapangan, jumlah responden yang berhasil diperoleh adalah sebanyak 248 responden, atau melebihi jumlah sampel minimum yang dibutuhkan. Karena jumlah responden aktual lebih besar daripada ukuran sampel minimum, maka tingkat keakuratan penelitian semakin meningkat. Konsekuensinya, nilai margin of error yang awalnya ditetapkan sebesar 5% mengalami penurunan menjadi sekitar 4,1%.

TABEL 2
(A)

Jenis Kelamin	Jumlah	Presentase
Positif	151	60.9%
Netral	97	39.1%

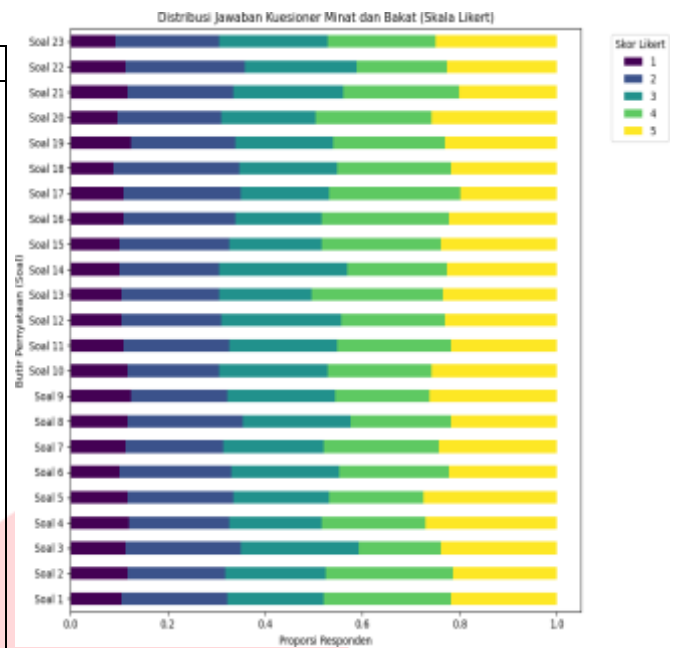
Dari Tabel 4.1 distribusi responden berdasarkan jenis kelamin, dapat dilihat bahwa mayoritas responden adalah laki-laki, yaitu sebanyak 151 orang atau 60,9% dari total sampel. Sementara itu, responden perempuan berjumlah 97 orang, yang merupakan 39,1% dari keseluruhan. Hal ini menunjukkan bahwa komposisi sampel penelitian cenderung didominasi oleh laki-laki. Perbandingan distribusi jenis kelamin ini juga dapat menjadi salah satu faktor yang memengaruhi pola segmentasi dalam analisis.

A. Uji Validitas dan Reliabilitas

Berdasarkan hasil uji validitas, hampir seluruh butir pertanyaan memenuhi kriteria ini, yang menunjukkan bahwa mayoritas instrumen mampu merefleksikan variabel yang diukur dengan baik.

TABEL 2
(B)

	R Hitung	p-Value	Validitas
1	0.472356	3.462527e-15	Valid
2	0.475165	2.257164e-15	Valid
3	0.417944	6.633126e-12	Valid
4	0.432328	1.023538e-12	Valid
5	0.479605	1.138542e-15	Valid
6	0.375339	1.023689e-09	Valid
7	0.480397	1.006745e-15	Valid
8	0.342546	3.10E-02	Valid
9	0.513669	4.24E-12	Valid
10	0.480945	9.24E-10	Valid
11	0.422563	3.675038e-12	Valid
12	0.375390	1.017970e-09	Valid
13	0.452010	6.848823e-14	Valid
14	0.439659	3.814755e-13	Valid
15	0.513327	4.495888e-18	Valid
16	0.404124	3.681234e-11	Valid
17	0.478858	1.278339e-15	Valid
18	0.423783	3.139904e-12	Valid
19	0.462791	1.443716e-14	Valid
20	0.460480	2.025022e-14	Valid
21	0.364138	3.429167e-09	Valid
22	0.439559	3.866903e-13	Valid
23	0.462010	3.462527e-37	Valid
24	0.279142	8.091179e-06	Tidak Valid



bahwa skor yang diperoleh dari responden bersifat konsisten dan dapat diandalkan untuk analisis lebih lanjut, sehingga kesimpulan yang ditarik dari data memiliki dasar yang kuat dan dapat dipercaya.

B. Principal Componen Analisis (PCA)

Proses pemrosesan data melibatkan beberapa langkah, antara lain: pemeriksaan data kosong (missing value), penghapusan kolom yang tidak relevan, dan reduksi dimensi untuk dataset menggunakan PCA.

Berdasarkan hasil uji validitas Tabel 4., hampir seluruh butir pertanyaan memenuhi kriteria ini, yang menunjukkan bahwa mayoritas instrumen mampu merefleksikan variabel yang diukur dengan baik. Namun, terdapat satu butir yang memiliki p-value di bawah 0,3 yakni pada pertanyaan ke 24, sehingga tidak memenuhi syarat validitas dan dianggap tidak relevan atau kurang konsisten dengan konstruk yang diukur. Pertanyaan ke 24 yang tidak valid ini kemudian dihapus dari instrumen penelitian untuk menjaga kualitas data dan memastikan analisis selanjutnya hanya menggunakan butir yang valid. Dengan demikian, penghapusan pertanyaan ke 24 yang tidak valid ini tidak hanya meningkatkan reliabilitas instrumen, tetapi juga meningkatkan keakuratan interpretasi hasil penelitian, karena setiap skor yang dianalisis berasal dari butir yang telah teruji validitasnya.

TABEL 2
(C)

Topik	Hasil Cronbach's Alpha	Hasil Uji
Minat	0.741	Reliabel
Bakat	0.755	Reliabel
Hobi	0.706	Reliabel

Dalam penelitian ini, reliabilitas diukur menggunakan Cronbach's Alpha, dengan ketentuan bahwa instrumen dianggap reliabel apabila nilai Cronbach's Alpha lebih dari 0,6. Hasil analisis menunjukkan bahwa seluruh butir pertanyaan memiliki nilai Cronbach's Alpha di atas ambang batas tersebut yaitu di 0,706 hingga 0,755, sehingga instrumen penelitian dapat dinyatakan reliabel. Hal ini menandakan

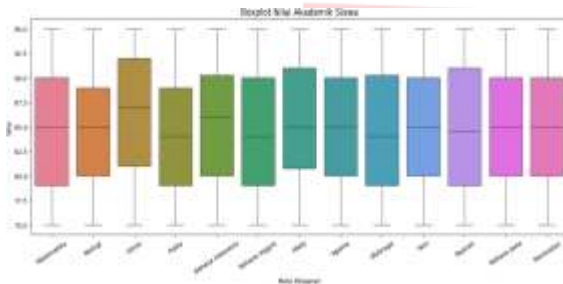
TABEL 2
(D)

Variabel	PC1	PC2	PC3	...	PC21	PC22	PC23
Matematika	0.043	0.127	0.152	...	-0.087	0.095	0.132
Biologi	-0.067	-0.180	0.410	...	0.243	0.168	0.113
Kimia	0.023	0.291	-0.027	...	0.128	-0.146	0.115
...	...	...	...	...	...	...	...
Pertanyaan 21	0.520	0.112	-0.190	...	-0.256	-0.127	-0.001
Pertanyaan 22	0.566	-0.006	-0.074	...	-0.127	0.014	-0.250
Pertanyaan 23	0.402	-0.138	0.317	...	0.114	-0.080	-0.048

Hubungan korelasi antara variabel data asli (nilai akademik dan minat) dengan Komponen Utama (PC) yang dihasilkan dari reduksi dimensi. Nilai loading menunjukkan seberapa besar kontribusi suatu variabel dalam mendefinisikan dimensi PC tersebut.

C. Karakteristik Data

Distribusi nilai akademik siswa pada seluruh mata pelajaran menunjukkan pola yang relatif homogen dengan rentang nilai utama berada pada kisaran 75 hingga 95. Median nilai pada sebagian besar mata pelajaran berada di sekitar 84–86, yang mengindikasikan bahwa secara umum capaian akademik siswa berada pada tingkat cukup tinggi dan stabil.



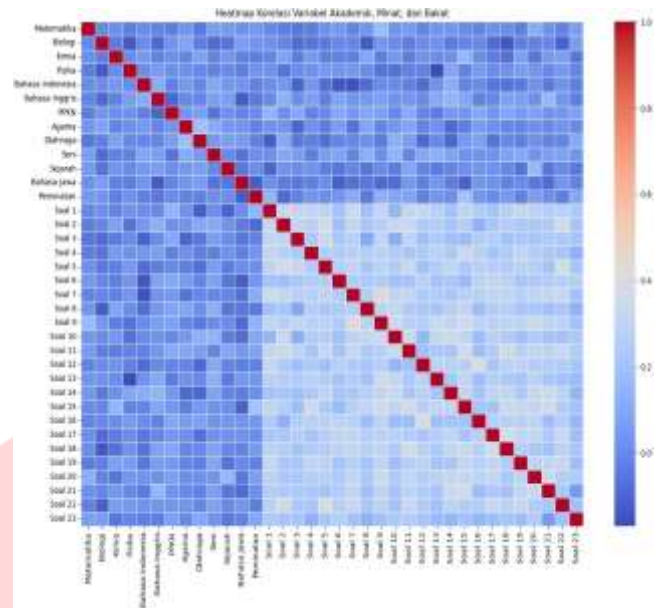
GAMBAR 3
(A)

Distribusi jawaban kuesioner minat dan bakat, terlihat bahwa pada hampir seluruh butir pernyataan (Soal 1–Soal 23) proporsi responden pada setiap kategori skala Likert (1 hingga 5) cenderung relatif merata. Tidak terdapat dominasi yang sangat kuat pada satu kategori jawaban tertentu, baik pada skor rendah (1–2) maupun skor tinggi (4–5). Penyebaran jawaban yang relatif seimbang ini menunjukkan bahwa tingkat minat, bakat, dan preferensi siswa bersifat heterogen, sehingga tidak dapat digeneralisasikan ke dalam satu kecenderungan sikap yang seragam. Kondisi tersebut mengindikasikan bahwa responden memiliki latar belakang minat dan kemampuan yang beragam, sesuai dengan tujuan penelitian yang berfokus pada pemetaan karakteristik siswa secara multidimensi.

GAMBAR 3
(C)

Berdasarkan hasil analisis korelasi yang divisualisasikan GAMBAR 3 (C), terlihat bahwa hubungan antara variabel akademik dengan variabel minat dan bakat umumnya bersifat lemah, baik positif maupun negatif. Korelasi negatif lemah, seperti pada pasangan Fisika–Soal 13 ($r = -0,1689$), Bahasa Indonesia–Soal 7 ($r = -0,1523$), Bahasa Indonesia–Soal 6 ($r = -0,1397$), Biologi–Soal 18 ($r = -0,1377$), serta Biologi–Fisika ($r = -0,1301$), menunjukkan bahwa peningkatan pada satu variabel cenderung diikuti penurunan kecil pada variabel lainnya. Namun, karena nilainya relatif rendah, hubungan ini tidak dapat diartikan sebagai hubungan yang saling bertentangan secara kuat, melainkan mencerminkan bahwa capaian akademik dan minat siswa berkembang secara relatif independen. Hal ini mengindikasikan bahwa prestasi pada mata pelajaran tertentu tidak selalu sejalan dengan minat atau kecenderungan personal yang diukur melalui kuesioner.

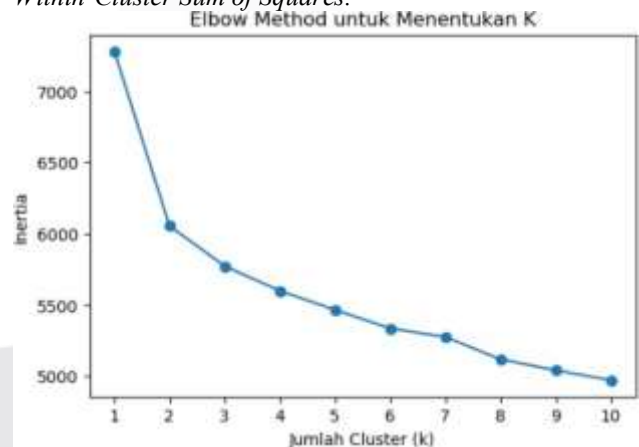
GAMBAR 3



(B)

D. K-MEANS

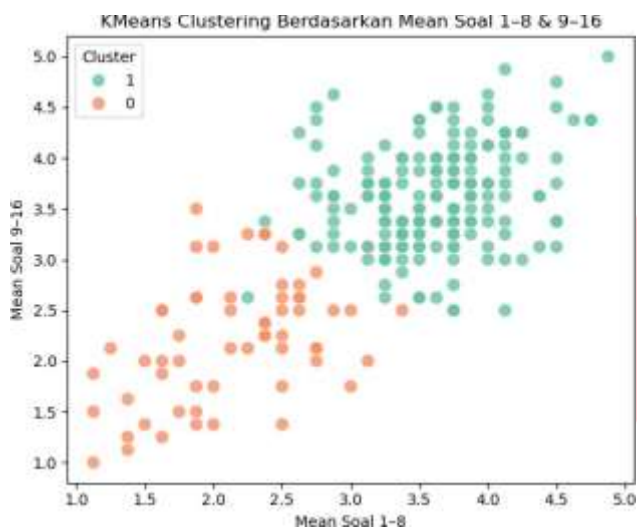
Penentuan jumlah kluster optimal pada penelitian ini dilakukan dengan menerapkan dua pendekatan, yaitu metode elbow dan metode silhouette. Metode *elbow* digunakan untuk melihat titik optimumnya berdasarkan penurunan nilai *Within-Cluster Sum of Squares*.



GAMBAR 4
(A)

Grafik Elbow Method digunakan untuk menentukan jumlah kluster (k) yang optimal dalam K-Means clustering. Metode ini memvisualisasikan hubungan antara Jumlah Kluster (k) pada sumbu X dan nilai Inersia (atau Within-Cluster Sum of Squares, WCSS) pada sumbu Y. Secara inheren, nilai Inersia akan selalu menurun seiring dengan bertambahnya k . Inersia mengukur seberapa rapat titik-titik data berada di dalam kluster mereka; nilai yang lebih rendah menunjukkan model kluster yang lebih baik. Berdasarkan analisis visual kurva, penurunan Inersia paling drastis diamati dari $k = 1$ ke $k = 2$. Kurva mulai menunjukkan pelengkungan dan penurunan Inersia mulai melambat secara nyata setelah mencapai nilai $k = 3$ dan $k = 4$. Oleh karena itu, berdasarkan prinsip Elbow Method, nilai $k = 2$ atau $k = 3$

adalah kandidat terbaik dan paling optimal untuk jumlah kluster yang akan digunakan. Pemilihan salah satu nilai k ini bertujuan untuk menyeimbangkan antara meminimalkan Inersia (mencapai kluster yang rapat) dan menghindari kompleksitas berlebihan (seperti overfitting) yang tidak memberikan manfaat tambahan signifikan terhadap interpretasi atau pemahaman struktur data yang mendasarinya.



GAMBAR 4

(B)

Visualisasi 5.5 menyajikan hasil pengelompokan data menggunakan algoritma K-Means, yang telah mengklasifikasikan titik data menjadi dua kluster Analisis spasial terhadap Cluster 0 (Oranye) mengungkapkan adanya

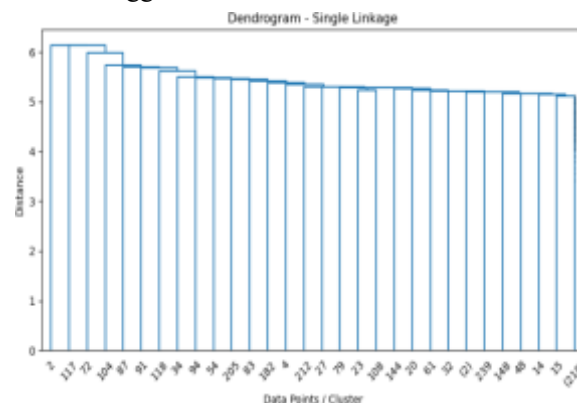


konsentrasi subjek pada kuadran kiri bawah, yang mengindikasikan bahwa kelompok ini memiliki profil performa rendah secara simultan pada kedua dimensi variabel. Secara statistik, rata-rata skor pada Mean Soal 1–8 yang berada di bawah 2.5 serta Mean Soal 9–16 di bawah 3.0 menunjukkan adanya hambatan kognitif atau kurangnya afinitas terhadap materi yang diujikan.

Secara integratif, pemisahan yang tegas antara kedua kluster ini memberikan bukti autentik mengenai adanya stratifikasi kemampuan yang linier di dalam populasi penelitian. Fenomena ini menegaskan bahwa instrumen Mean Soal 1–16 memiliki validitas diskriminan yang kuat dalam memetakan profil subjek berdasarkan tingkat kompetensinya. Integrasi data ini menyarankan pentingnya transformasi kebijakan instruksional dari pendekatan generalis menuju pendekatan berbasis data (data-driven instruction). Strategi ini krusial untuk memastikan bahwa intervensi pendidikan diberikan secara proporsional; baik melalui dukungan intensif bagi Cluster 0 untuk mencapai standar kompetensi minimum, maupun melalui penyediaan ruang eksplorasi yang lebih luas bagi Cluster 1 agar tidak terjadi stagnasi intelektual dalam proses pembelajaran.

E. HIERARCHICAL CLUSTERING

Pertama menggunakan 4 metode dalam klusterisasi.



GAMBAR 5

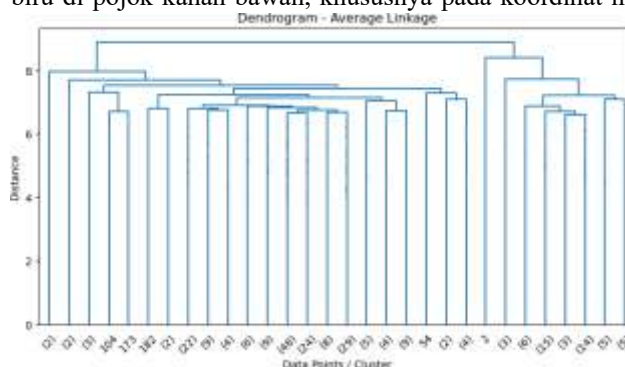
(A)

Menunjukkan hasil dari *Hierarchical Clustering* dengan metode *Single Linkage*. *Dendrogram* ini menggambarkan bagaimana titik data atau kluster digabungkan berdasarkan jarak terdekat (minimum distance). Sumbu Y (Distance) yang tinggi (di atas 5) menunjukkan bahwa penggabungan kluster secara signifikan terjadi pada jarak yang besar, tetapi tidak ada garis horizontal yang panjang yang secara jelas memisahkan sejumlah kecil kluster pada jarak yang lebih rendah. Ini mengindikasikan bahwa data mungkin tidak memiliki struktur kluster yang terpisah dengan baik (well-separated) jika menggunakan *Single Linkage*, atau metode ini mengabaikan kluster yang panjang dan tipis, di mana sebagian besar data tetap tidak terkelompok hingga jarak yang sangat jauh (sekitar 6.0). Oleh karena itu, dendrogram ini tidak memberikan indikasi yang kuat untuk memotong garis horizontal dan mendapatkan jumlah kluster yang optimal, dan metode lain lebih dibutuhkan.

GAMBAR 5

(B)

Analisis visualisasi scatter plot pada gambar 5.9 menunjukkan terjadinya fenomena chaining atau perantaraan yang sangat nyata, di mana hampir seluruh titik data menyatu ke dalam satu kluster besar berwarna hijau. Hal ini disebabkan oleh prinsip kerja algoritma *Single Linkage* yang menggabungkan kluster berdasarkan jarak minimum antar anggota terdekat, sehingga titik-titik data yang memiliki kedekatan jarak pada rentang nilai rata-rata 75 hingga 95 terus bergabung menjadi satu struktur tanpa menciptakan pemisahan kelompok yang seimbang. Akibatnya, metode ini gagal mengidentifikasi segmentasi karakteristik siswa secara mendalam dan hanya mampu memisahkan satu titik biru di pojok kanan bawah, khususnya pada koordinat nilai

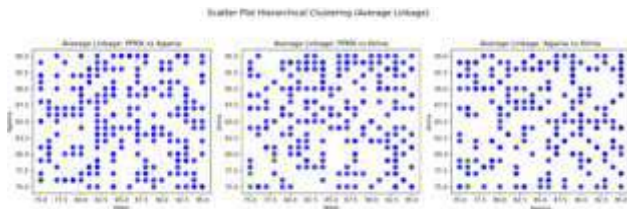


Seni sebesar 95 namun memiliki nilai Bahasa Jawa dan Sejarah yang rendah di angka 75.

GAMBAR 6

(A)

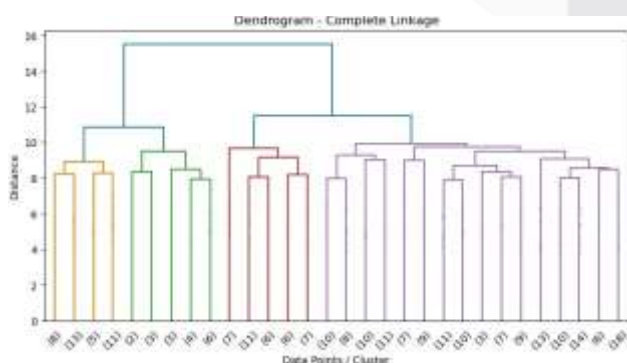
Dendrogram yang dihasilkan menggunakan metode *Average Linkage* menunjukkan struktur hierarkis pengelompokan yang jauh lebih informatif dibandingkan *Single Linkage*. Pada dendrogram ini, kita dapat melihat beberapa garis horizontal yang panjang, menunjukkan adanya pemisahan kluster yang stabil pada jarak yang relatif jauh (misalnya di sekitar jarak 7.5 hingga 8.5). Jika kita memotong garis horizontal di sekitar jarak 7.5, kita dapat mengidentifikasi pembentukan sekitar empat hingga enam kluster utama yang berbeda. Jarak kluster yang lebih pendek (di bawah 7) mengindikasikan bahwa sub-kluster di dalamnya memiliki anggota yang relatif homogen. Secara keseluruhan, metode *Average Linkage* berhasil mengungkapkan adanya beberapa kelompok inti yang terpisah dengan cukup baik dalam data, memberikan panduan yang lebih jelas untuk menentukan jumlah kluster yang optimal, yang mana jumlah tersebut kemungkinan lebih dari dua.



GAMBAR 6

(B)

Analisis terhadap grafik *Average Linkage* menunjukkan pola pengelompokan yang cenderung sangat timpang, di mana hampir seluruh titik data menyatu ke dalam satu kluster besar berwarna biru. Hal ini disebabkan oleh prinsip kerja algoritma *Average Linkage* yang menghitung jarak antar kluster berdasarkan rata-rata jarak seluruh pasangan poin antar kelompok, yang pada dataset ini menghasilkan pengelompokan yang sangat padat pada rentang nilai 75 hingga 95. Metode *Average* pada data ini gagal menciptakan segmentasi yang heterogen dan hanya mampu memisahkan segelintir titik hijau, seperti pada titik PPKN, Agama, dan Kimia di angka 75.



GAMBAR 7

(A)

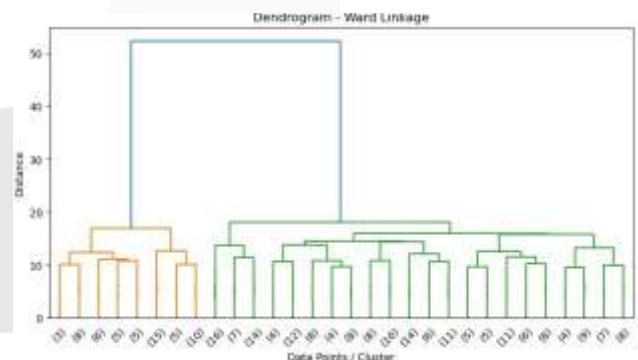
Metode *Complete Linkage* untuk melakukan *Hierarchical Clustering*, yang cenderung menghasilkan kluster yang kompak karena jarak antar kluster diukur dari jarak maksimum antara titik-titik anggotanya. Struktur dendrogram menunjukkan pemisahan yang stabil dan signifikan pada jarak yang tinggi (sumbu Y). Secara khusus, jika garis potong horizontal ditarik pada jarak sekitar 15.5, data akan terbagi menjadi dua kluster utama yang sangat besar. Namun, pemotongan yang lebih rendah, seperti di sekitar jarak 11.5, menghasilkan pemisahan menjadi empat (4) kluster utama yang lebih terisolasi. Kejelasan pemisahan kluster pada metode *Complete Linkage* ini mengindikasikan bahwa data memiliki beberapa kelompok yang kompak dan terpisah dengan jarak yang substansial, memberikan panduan kuat bahwa jumlah kluster optimal (K) kemungkinan adalah 4 atau lebih tinggi.



GAMBAR 7

(B)

Interpretasi terhadap grafik *Hierarchical Clustering* 5.13 memberikan perspektif tambahan mengenai segmentasi siswa berdasarkan nilai akademik rapor menggunakan *Complete Linkage*. Pada *Complete Linkage*, pemisahan antara titik biru dan hijau tampak lebih tersebar dan tumpang tindih (*overlap*), terutama pada perbandingan mata pelajaran Agama vs Kimia dan Kimia vs Sejarah. *Complete Linkage*. Hal ini sejalan dengan tujuan penelitian untuk mengidentifikasi karakteristik siswa secara representatif berdasarkan integrasi data akademik dan non-akademik guna mengurangi risiko salah jurusan yang mencapai angka 87 persen di Indonesia.



GAMBAR 8

(A)

Metode *Ward Linkage*, yang unik karena ia meminimalkan varians internal kluster yang dihasilkan, sehingga cenderung menciptakan kluster yang berukuran relatif sama dan kompak. Struktur hierarkis yang paling jelas terlihat adalah pemisahan menjadi dua kluster utama (Kluster Kiri dan Kluster Kanan) yang bergabung pada jarak yang sangat tinggi, yaitu sekitar 52 pada sumbu Y (*Distance*). Pemisahan yang terjadi pada jarak yang ekstrem ini menunjukkan bahwa kedua kelompok besar ini memiliki

perbedaan varians yang sangat signifikan dan jelas. Jika dilihat lebih detail, Kluster Kiri terbagi menjadi sub-kluster pada jarak ≈ 18 , sedangkan Kluster Kanan menunjukkan beberapa penggabungan utama pada jarak ≈ 15 . Berdasarkan aturan "pemotongan" (cutting) dendrogram untuk mendapatkan kluster optimal, pemotongan di sekitar jarak 18 akan menghasilkan tiga (3) kluster atau pemotongan di sekitar jarak 15 akan menghasilkan lima (5) kluster, yang merupakan kandidat kuat untuk jumlah kluster paling stabil dalam data ini.



GAMBAR 8

(B)

Metode Ward bekerja dengan meminimalkan varians di dalam kelompok (within-cluster variance), sehingga menghasilkan partisi data yang lebih halus dan hierarkis. Visualisasi pada gambar 5.15 Kluster biru secara konsisten mengidentifikasi kelompok siswa dengan performa akademik unggul yang memiliki nilai rapor di kisaran 85 hingga 95, sementara kluster hijau mengelompokkan siswa dengan capaian kompetensi dasar pada rentang nilai 75 hingga 85. Melalui perbandingan multidimensional seperti Sejarah vs Olahraga terlihat bahwa meskipun seorang siswa memiliki nilai tinggi di bidang humaniora, nilai mereka di bidang fisik atau spiritual cenderung lebih bervariasi. Hal ini menunjukkan bahwa metode Ward berhasil menangkap keragaman karakteristik akademik siswa yang tidak selalu linier, memberikan gambaran tentang keseimbangan antara kemampuan kognitif dan keterampilan praktis masing-masing individu.

F. DBSCAN

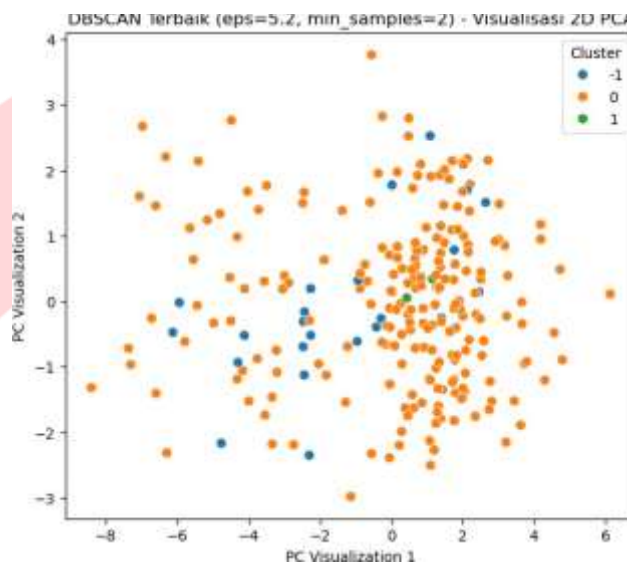
DBSCAN menggunakan *hyperparameter tuning* guna melihat parameter apa yang paling terbaik.

TABEL 3

eps	min_samples	n_clusters	Silhouette	Davies-Bouldin	Time (s)
5.2	2	2	-0.017046	4.224784	0.006002
4.9	3	4	-0.036005	3.443917	0.006997
4.3	4	2	-0.083816	1.899923	0.006016
4.6	5	4	-0.087580	3.369553	0.006996
4.9	6	4	-0.090149	2.879207	0.006009
4.9	2	4	-0.097108	3.086674	0.007025
3.7	4	2	-0.105117	1.648869	0.004997
4.9	2	7	-0.106618	2.644998	0.006001
4.9	4	10	-0.114261	2.388395	0.006526
4.6	6	6	-0.122275	3.090926	0.006309

Hasil Hyperparameter Tuning DBSCAN menunjukkan bahwa algoritma ini gagal menemukan struktur kluster yang kompak dan terpisah dengan baik dalam data yang diuji.

Bukti utama kegagalan ini adalah Silhouette Score yang secara konsisten negatif (nilai terbaik hanya -0.017046), yang mengindikasikan bahwa rata-rata, titik data lebih dekat ke kluster tetangga daripada ke kluster mereka sendiri. Skor yang ideal mendekati +1, sehingga skor negatif menyarankan kluster yang buruk dan tidak terdefinisi. Meskipun kombinasi $\text{eps} = 5.2$ dan $\text{min_samples} = 2$ menghasilkan skor paling tidak negatif, ini hanya menghasilkan dua kluster, yang kemungkinan besar adalah pengelompokan yang dangkal dan tidak mewakili struktur data yang kompleks. Selain itu, nilai Davies-Bouldin Index (DBI) yang tinggi (berkisar antara 1.89 hingga 4.22, di mana nilai ideal mendekati 0) semakin memperkuat bahwa kluster yang terbentuk mengalami tumpang tindih signifikan dan kurang terpisah.



GAMBAR 9

Kegagalan DBSCAN ini menyiratkan bahwa data Anda tidak memiliki wilayah kepadatan tinggi seperti gambar 5.16 yang seragam atau jelas yang menjadi asumsi dasar dari algoritma ini. Data mungkin terlalu tersebar merata di ruang dimensi tinggi, sehingga metode berbasis kepadatan ini tidak cocok.

G. Fitur Terbaik

Pendekatan variance threshold berbasis between-cluster variance digunakan untuk mengidentifikasi fitur-fitur penting pada hasil klusterisasi K-Means dalam penelitian ini.

TABEL 4

(A)

No	Fitur (Variabel)	Nilai Importance
1	Mean Soal 1–8	145.8321
2	Mean Soal 9–16	138.3038
3	Mean Soal 17–23	128.0515

Metode tersebut dipilih karena K-Means merupakan algoritma berbasis centroid, sehingga fitur yang paling berkontribusi adalah fitur yang mampu menghasilkan perbedaan nilai rata-rata (mean) yang besar antar kluster. Secara konseptual, fitur yang memiliki distribusi nilai berbeda secara signifikan pada setiap kelompok akan berperan dominan dalam proses pembentukan kluster. Suatu fitur

dianggap memiliki tingkat kepentingan yang tinggi apabila varians antar klasternya besar, yang menunjukkan adanya perbedaan karakteristik yang kontras antara satu kelompok dengan kelompok lainnya.

Berdasarkan hasil perhitungan between-cluster variance, diperoleh beberapa fitur dengan nilai varians tertinggi yang secara signifikan mendominasi proses pengelompokan. Fitur Mean Soal 1–8 (145.83), Mean Soal 9–16 (138.30), dan Mean Soal 17–23 (128.05) teridentifikasi sebagai variabel paling dominan karena memiliki nilai varians yang sangat besar dibandingkan fitur lainnya. Selain indikator berbasis rata-rata soal tersebut, beberapa nilai akademik seperti Seni (6.64) dan Bahasa Inggris (5.36) juga muncul sebagai fitur penting yang berkontribusi dalam mempertegas struktur klaster yang dihasilkan.

TABEL 4

(B)

No	Fitur (Variabel)	Mean Cluster 1	Mean Cluster 2	Mean Difference
1	Seni	84.6559	95.0000	10.3441
2	Bahasa Jawa	85.0526	75.0000	10.0526
3	Sejarah	84.9271	75.0000	9.9271

Berdasarkan hasil uji signifikansi pada *Hierarchical Clustering*, fitur-fitur yang terbukti paling berpengaruh terhadap pembentukan klaster dapat diidentifikasi dari nilai p-value yang paling kecil. Berdasarkan data nilai rata-rata antar klaster yang dihasilkan, analisis fitur penting pada metode Single Linkage menunjukkan bahwa pembentukan kelompok sangat dipengaruhi oleh variabel yang memiliki rentang nilai kontras. Fitur Seni menjadi variabel paling dominan dengan selisih rata-rata (Mean Difference) sebesar 10.34, di mana Klaster 2 memiliki nilai rata-rata sempurna sebesar 95.0 dibandingkan Klaster 1 yang hanya 84.65. Hal ini mengindikasikan bahwa variabel Seni berfungsi sebagai pembeda utama dalam mengisolasi titik data tertentu dari massa data utama dalam proses penggabungan hierarkis.

TABEL 4

(C)

No	Fitur (Variabel)	Mean Cluster 1	Mean Cluster 2	Mean Difference
1	PPKN	76.00	85.47	9.47
2	Agama	76.00	85.21	9.21
3	Kimia	77.00	86.13	9.13

Berdasarkan data nilai rata-rata antar klaster yang dihasilkan pada tabel 5.6, analisis fitur penting pada metode Average Linkage menunjukkan bahwa pembentukan kelompok sangat didominasi oleh mata pelajaran PPKN, Agama, dan Kimia yang memiliki selisih rata-rata (Mean Difference) di atas 9 poin. Fitur PPKN menjadi variabel pembeda paling kuat dengan selisih rata-rata sebesar 9.47, di mana Klaster 2 memiliki nilai rata-rata sebesar 85.47 sementara Klaster 1 hanya sebesar 76.00. Hal ini mengindikasikan bahwa variabel

kewarganegaraan berfungsi sebagai faktor utama dalam memisahkan kelompok siswa dalam model ini, disusul oleh variabel Agama dan Kimia yang masing-masing memiliki selisih sebesar 9.21 dan 9.14. Munculnya ketiga mata pelajaran ini sebagai top 3 fitur terbaik membuktikan bahwa metode Average Linkage pada dataset ini lebih menyoroti perbedaan kompetensi akademik pada rumpun sosial-spiritual dan sains dibandingkan nilai-nilai praktis atau seni.

TABEL 4

(D)

No	Fitur (Variabel)	Mean Cluster 1	Mean Cluster 2	Mean Difference
1	Agama	83.4488	86.8347	3.3859
2	Kimia	87.5039	84.4793	3.0246
3	Sejarah	86.2835	83.4215	2.8620

Berdasarkan data nilai rata-rata antar klaster yang dihasilkan, analisis fitur penting pada metode Complete Linkage menunjukkan bahwa pembentukan kelompok didorong oleh variabel yang memiliki jarak terjauh antar anggota klaster untuk memastikan kekompakan kelompok. Fitur Agama menjadi variabel pembeda paling signifikan dengan selisih rata-rata (Mean Difference) sebesar 3.3859, di mana Klaster 2 memiliki rata-rata nilai sebesar 86.83 sementara Klaster 1 sebesar 83.45. Hal ini mengindikasikan bahwa variabel nilai spiritualitas berfungsi sebagai faktor penentu utama dalam memisahkan profil siswa pada model ini, disusul oleh variabel Kimia dan Sejarah yang masing-masing memiliki selisih sebesar 3.0246 dan 2.8620. Berbeda dengan metode Single atau Average yang memiliki selisih hingga 9-10 poin, metode Complete Linkage pada dataset ini menunjukkan perbedaan rata-rata yang lebih kecil (kisaran 2-3 poin), menandakan pembagian kelompok yang lebih merata namun dengan batas pemisah yang lebih tipis.

TABEL 4

(E)

No	Fitur (Variabel)	t-stat	p-value
1	Sejarah	10.7785	9.7248e-22
2	Olahraga	7.5202	3.0660e-12
3	Agama	5.8002	3.2153e-08

Kerja algoritma Ward yang meminimalkan varians di dalam klaster, sehingga variabel dengan daya pembeda tinggi seperti Sejarah dan Olahraga (t-stat : 7.52) secara otomatis terpilih sebagai parameter utama dalam mengelompokkan siswa ke dalam tingkatan prestasi yang berbeda. Dalam implementasi sistem rekomendasi jurusan, temuan ini mengonfirmasi bahwa nilai Sejarah, Olahraga, dan Agama (t-stat : 5.80) adalah indikator yang paling andal untuk memetakan profil bakat siswa di sekolah tersebut. Karena fitur-fitur ini memiliki validitas statistik yang kuat dengan p-value jauh di bawah ambang batas 0.05, segmentasi yang dihasilkan dapat digunakan secara objektif untuk memprediksi kecocokan jurusan. Dengan demikian, metode Ward memberikan landasan yang lebih stabil dibandingkan metode linkage lainnya.

karena variabel pembedanya didukung oleh bukti statistik yang signifikan, sehingga mampu mengurangi subjektivitas dalam proses bimbingan karier dan pemilihan program studi bagi mahasiswa.

Karakteristik dasar DBSCAN yang tidak membentuk kluster berdasarkan centroid maupun struktur hierarki, melainkan berdasarkan kepadatan (density) data di ruang fitur. DBSCAN mengelompokkan data berdasarkan dua parameter utama, yaitu ϵ (radius tetangga) dan min_samples (jumlah minimum titik dalam radius ϵ). Kluster terbentuk apabila sejumlah titik berada dalam wilayah yang cukup padat, tanpa mempertimbangkan kontribusi individual masing-masing fitur secara terpisah. Akibatnya, DBSCAN tidak menghasilkan representasi kluster berupa pusat kluster maupun struktur statistik yang memungkinkan perhitungan varians antar kluster atau uji signifikansi fitur.

V. KESIMPULAN

Berdasarkan hasil keseluruhan tahapan analisis yang meliputi eksplorasi data, reduksi dimensi, pemodelan kluster, evaluasi performa algoritma, serta interpretasi karakteristik kluster, penelitian ini berhasil menunjukkan bahwa pendekatan klusterisasi multidimensi mampu mengidentifikasi pola preferensi bidang studi calon mahasiswa secara objektif dan bermakna. Hasil Exploratory Data Analysis (EDA) memperlihatkan bahwa distribusi nilai akademik siswa relatif homogen dengan median berada pada rentang 84–86 pada sebagian besar mata pelajaran, yang menandakan tingkat capaian akademik yang cukup tinggi dan stabil, serta korelasi yang kuat antar mata pelajaran sejenis, khususnya pada kelompok mata pelajaran eksakta dan bahasa. Sementara itu, variabel non-akademik berupa minat, bakat, dan hobi memberikan variasi tambahan yang signifikan sehingga memperkaya struktur data multidimensi, serta hasil deteksi outlier menunjukkan tidak adanya pencilan ekstrem yang berpotensi merusak struktur kluster, sehingga data dinilai layak untuk dilakukannya klusterisasi berbasis jarak. Proses reduksi dimensi menggunakan Principal Component Analysis (PCA) terbukti mampu mempertahankan proporsi variansi yang besar pada beberapa komponen utama, yang mengindikasikan bahwa kombinasi antara variabel akademik dan non-akademik membentuk struktur laten yang relevan dalam menjelaskan perbedaan profil siswa dan mendukung proses pengelompokan secara lebih efisien tanpa kehilangan informasi penting.

Dari sisi performa algoritma, hasil evaluasi menggunakan Silhouette Score dan Davies–Bouldin Index menunjukkan bahwa Hierarchical Clustering memberikan kinerja terbaik dengan Silhouette Score tertinggi sebesar 0,189 dan DBI terendah sebesar 1,874, yang mengindikasikan kluster paling kompak dan terpisah secara relatif lebih baik dibandingkan metode lainnya, diikuti oleh K-Means dengan Silhouette Score sebesar 0,184 dan DBI sebesar 1,915 yang juga secara konsisten mengonfirmasi bahwa struktur data optimal terbentuk pada jumlah kluster $K = 2$. Sebaliknya, DBSCAN menghasilkan Silhouette Score negatif (−0,017) dan DBI sangat tinggi (4,2), yang menunjukkan bahwa data tidak memiliki variasi kepadatan lokal yang cukup kontras sehingga pendekatan berbasis densitas tidak sesuai untuk karakteristik data ini. Temuan ini menguatkan bahwa metode berbasis jarak dan centroid lebih tepat digunakan untuk data preferensi siswa yang cenderung homogen namun tetap memiliki perbedaan struktural global. Dari aspek

interpretasi fitur, pada metode K-Means, pemisahan kluster paling dipengaruhi oleh perbedaan centroid pada mata pelajaran inti seperti Matematika, Fisika, Kimia, dan Biologi, serta indikator minat terhadap aktivitas analitis dan teknologi, sehingga kluster terbentuk terutama berdasarkan kekuatan kompetensi akademik eksakta dan orientasi minat akademik. Sementara itu, pada Hierarchical Clustering, selain mata pelajaran eksakta, variabel Bahasa Indonesia, Bahasa Inggris, serta indikator minat terhadap aktivitas belajar mandiri, eksplorasi pengetahuan, dan perencanaan studi juga menunjukkan perbedaan signifikan antar kluster, sehingga struktur kluster yang dihasilkan lebih komprehensif dalam merepresentasikan kesiapan studi lanjut baik dari aspek kognitif maupun motivasional. Interpretasi lanjutan terhadap pusat kluster terbaik (berdasarkan Hierarchical Clustering) menunjukkan bahwa Cluster 0 merepresentasikan kelompok siswa yang relatif kurang menunjukkan kesiapan dan kompetensi untuk melanjutkan ke studi lanjut, yang dicirikan oleh nilai yang lebih rendah pada mata pelajaran eksakta serta skor minat yang lebih dominan pada aktivitas non-akademik dan praktis, sehingga kelompok ini cenderung membutuhkan pendampingan lebih intensif dalam eksplorasi pilihan studi dan penguatan motivasi akademik. Sebaliknya, Cluster 1 merepresentasikan siswa dengan kesiapan studi lanjut yang lebih kuat, ditandai oleh performa akademik yang lebih tinggi pada mata pelajaran inti, khususnya eksakta dan bahasa, serta indikator minat dan bakat yang konsisten terhadap aktivitas belajar, analitis, dan eksplorasi akademik, sehingga kelompok ini lebih siap diarahkan pada perencanaan pendidikan tinggi secara spesifik dan strategis. Dengan demikian, hasil klusterisasi tidak hanya menggambarkan perbedaan rumpun studi secara umum seperti Saintek dan Soshum, tetapi juga mencerminkan tingkat kesiapan akademik dan orientasi karier siswa, yang sangat relevan untuk mendukung

layanan bimbingan karier berbasis data. Secara keseluruhan, penelitian ini menjawab seluruh rumusan masalah dengan menunjukkan bahwa karakteristik awal data dapat dipahami secara jelas melalui EDA, performa algoritma clustering dapat dibandingkan secara objektif melalui metrik evaluasi kuantitatif, serta variabel multidimensi baik akademik maupun non-akademik terbukti berkontribusi signifikan dalam membentuk pola pengelompokan calon mahasiswa. Hasil ini menegaskan bahwa klusterisasi multidimensi dapat menjadi fondasi yang kuat dalam pengembangan sistem rekomendasi jurusan yang lebih personal, adil, dan berbasis potensi nyata siswa, sehingga berpotensi mengurangi fenomena salah jurusan serta meningkatkan kesesuaian antara profil siswa, pilihan studi, dan arah karier di.

REFERENSI

- [1] Sari, A., Nanere, Y. E. & Ernawati, R. (2023), 'Kematangan karir siswa remaja dalam menghadapi dunia pekerjaan', *Jurnal Suluh Pendidikan (JSP)* 11(1)
- [2] 1Redaksi : "Hasil Survei: 87 Persen Mahasiswa di Indonesia Ternyata Salah Jurusan" <https://jadiasn.id/rumpun-ilmu-sppi-dan-relevansi-akademik-jurusan-harus-linier/> 17 feb 2025
- [3] Oetomo, P. F., Yuwanto, L. & Rahaju, S. (2017), 'Faktor penentu penyesuaian diri pada mahasiswa baru emerging adulthood tahun pertama dan tahun kedua', *Jurnal Ilmiah Psikologi MIND SET* 8(2), 67–77
- [4] Syahril, M., Kusnasari, S., Sobirin, S., Muhazir, A. & Syahputri, A. (2023), 'Implementasi data mining untuk rekomendasi jurusan menggunakan algoritma k-means clustering', *J-SISKO TECH EDISI JANUARI* 6(1).
- [5] Palevi, M. N. et al. (2024), 'Implementasi algoritma k-means clustering dengan pendekatan active learning pada siswa sma untuk menentukan jurusan ke perguruan tinggi', *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informa- tika dan Komputer)* 23(1), 26.
- [6] Tzenios, F. (2020), 'Clustering students for personalized health education ba- sed on learning styles', *Sage Science Review of Educational Technology (SS- RET)* 3(1).
- [7] Fu, N., Ni, W., Hu, H. & Zhang, S. (2023), 'Multidimensional grid-based clustering with local differential privacy', *Information Sciences* 623, 402–420.
- [8] Darmayanti, I., Adhimah, L. F., Sadewo, R., Hidayati, N. & Subarkah, P. (2024), 'Improving k-means clustering accuracy for academic success inves- tigation with extreme gradient boosting algorithm', *Indonesian Journal of Artificial Intelligence and Data Mining* 7(1).
- [9] Kementerian Pendidikan dan Kebudayaan Republik Indonesia, Peraturan Menteri Pen- didikan dan Kebudayaan Republik Indonesia Nomor 154 Tahun 2014 tentang Rumpun Ilmu Pengetahuan dan Teknologi serta Gelar Lulusan Perguruan Tinggi (2014)
- [10] Tisyirin, "Macam-Macam Jurusan Kuliah Berdasarkan Rumpun Keilmuannya," Quipper, 2022
- [11] Kumara, A. R. (2018), *Buku Ajar Penelitian Kualitatif, Program Studi Bim- bingan dan Konseling*, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Ahmad Dahlan.
- [12] Esi Rosita (2021), 'Uji validitas dan reliabilitas kuesioner perilaku prososial', *Fokus Konseling* 4(4), 258–268.
- [13] Sugiyono (2019), *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*, Alfabeta, Bandung.
- [14] Nurhayati, Sigit Sinatrya, N., Wardhani, L. K. & Busman (2018), 'Analysis of k-means and k-medoids's performance using big data technology', *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)* .
- [15] Jain, A. K. (2010), 'Data clustering: 50 years beyond k-means', *Pattern Re- cognition Letters* 31(8), 651–666.
- [16] Han, J., Kamber, M. & Pei, J. (2011), *Data Mining: Concepts and Techniques*, 3 edn, Morgan Kaufmann Publishers, Burlington, MA.
- [17] Kalita, E., Oyelere, S. S., Gaftandzhieva, S. & Kandala, R. N. V. P. S. (2025), 'Educational data mining: A 10-year review', *Discover Computing* .
- [18] Bellaj, M., Bendahmane, A., Boudra, S. & Sefian, M. L. (2024), 'Educational data mining: Employing machine learning techniques and hyperparameter optimization to improve students' academic performance', *International Jo- urnal of Online and Biomedical Engineering (iJOE)* 20(03), 55–74.