Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3220~3233
e-ISSN: 2548-6861

Balancing Student Specialization Class Placement Based on Interests and Talents Using K-Means Clustering and Genetic Algorithm

Chaidir Chalaf Islamy [1]*, Muhammad Andika Oktaviansyah [2]*
* Teknik Informatika, Universitas 17 Agustus 1945 Surabaya
chaidirc@untag-sby.ac.id [1], andika.oktaviansyah.12@gmail.com [2]

Article Info

Article history:

Received 2025-07-23
Revised 2025-10-25
Accepted 2025-11-05

Keyword:

Genetic Algorithm,
K-Means Clustering,
RIASEC,
School Optimization,
Student Placement.

ABSTRACT

Student specialization placement in Indonesian secondary schools often produces imbalanced class distributions and misalignment between student interests and assigned tracks. This study develops a hybrid optimization system combining K-Means clustering and Genetic Algorithm (GA) to allocate 133 tenth-grade students from SMAN 1 Ngimbang into four specialization classes (Science, Mixed-Science, Mixed-Social, Social) while balancing operational constraints. Initial K-Means clustering (k=4, n_init=100) achieved a Silhouette Score of 0.287 but yielded severely imbalanced distribution (10, 51, 48, 24 students). GA optimization (population=300, generations=150, crossover=70%, mutation=10%, elitism=10%) with multi-component fitness function incorporating cosine similarity, distribution penalty, movement penalty, and entropy produced balanced classes (31, 35, 35, 32 students) within the 30-35 target range. Post-optimization metrics showed 73.7% retention rate, average match score of 0.792, entropy of 0.482, and execution time of 47.8 seconds. The Silhouette Score decreased to 0.080, reflecting an acceptable trade-off between cluster purity and operational feasibility. Sensitivity analysis confirmed weight configuration robustness. This system demonstrates practical applicability for real-time school implementation, reducing distribution gap by 90.2% while maintaining individual-class compatibility.

I. PENDAHULUAN

Penentuan peminatan siswa pada jenjang pendidikan menengah merupakan proses strategis yang berdampak langsung terhadap arah studi dan karier siswa di masa mendatang. Namun, proses ini masih sering dilakukan secara manual dan cenderung mengabaikan pemetaan minat dan bakat siswa secara menyeluruh. Ketidaksesuaian peminatan dengan kecenderungan psikologis atau kemampuan akademik dapat memicu menurunnya motivasi belajar serta meningkatnya angka perpindahan jurusan [1]. Permasalahan lain yang kerap ditemukan dalam sistem pembagian kelas peminatan mencakup ketidakseimbangan jumlah siswa antar kelas, subjektivitas dalam proses pengelompokan, serta ketiadaan sistem pendukung keputusan berbasis data. Berdasarkan laporan *Indonesia Career Center Network*, sekitar 87% mahasiswa di Indonesia mengaku telah memilih jurusan yang tidak sesuai dengan minat dan bakat mereka [2]. Selain itu, laporan PISA 2022 menunjukkan bahwa meskipun terjadi peningkatan performa akademik global, Indonesia masih menghadapi tantangan dalam menyediakan pembelajaran yang sesuai dengan karakteristik siswa [3].

Pendekatan RIASEC (*Realistic, Investigative, Artistic, Social, Enterprising, Conventional*) yang dikembangkan oleh John Holland telah banyak dimanfaatkan untuk memetakan minat dan kecenderungan siswa. Model ini memungkinkan pengelompokan kecenderungan karier individu berdasarkan dimensi minat dan kepribadian [4]. Implementasinya juga diterapkan dalam sistem rekomendasi berbasis web menggunakan metode *Analytical Network Processing* (ANP), yang terbukti membantu memberikan saran peminatan yang sesuai dengan profil siswa [2]. Di sisi lain, kemajuan teknologi informasi memungkinkan diterapkannya metode komputasi cerdas untuk proses klasifikasi dan pengelompokan siswa. Algoritma *K-Means* telah terbukti efektif dalam mengelompokkan data kepribadian dengan nilai *Silhouette Index* mencapai 0,4341 [5], sementara algoritma genetika menunjukkan keunggulan dalam skenario optimasi kombinatorial seperti penjadwalan perkuliahan yang mampu menyusun 265 jadwal tanpa konflik dalam waktu 561 detik [6], prediksi penyakit autoimun dengan akurasi tinggi [7], dan pengelompokan peserta Kuliah Kerja Nyata dengan akurasi hingga 95% [8]. Evaluasi performa metode *clustering* telah dilakukan melalui perhitungan *silhouette coefficient*

[9], perbandingan enam metode jarak dalam *K-Means* [10], pembentukan kelompok kolaboratif menggunakan algoritma genetika [11], serta pengelompokan nilai siswa untuk pemetaan jalur peminatan [12].

Penelitian ini mengusulkan kombinasi *K-Means clustering* dan algoritma genetika untuk menghasilkan sistem pembagian kelas peminatan yang optimal. Pemilihan pendekatan *hybrid* ini didasarkan pada perbedaan fundamental tujuan algoritma: *K-Means* berfungsi untuk mengelompokkan siswa berdasarkan kesamaan profil RIASEC secara *unsupervised*, sementara algoritma genetika berperan dalam optimasi alokasi *discrete* dengan mempertimbangkan *constraint* kapasitas kelas dan distribusi yang adil. *K-Means* dipilih karena efisiensi komputasinya yang tinggi dalam menangani data berdimensi tinggi dan kemampuannya menghasilkan *centroid* yang merepresentasikan profil ideal setiap *cluster*. Meskipun varian *K-Medoids* lebih *robust* terhadap *outlier*, algoritma tersebut memilih objek aktual sebagai pusat *cluster* sehingga sulit memetakan *medoid* kembali ke domain RIASEC asli yang bersifat kontinu. Algoritma genetika dipilih dibandingkan metode *swarm intelligence* seperti *Particle Swarm Optimization* (PSO) atau *Grey Wolf Optimizer* (GWO) karena kemudahan dalam merepresentasikan solusi sebagai kromosom yang memetakan alokasi siswa ke kelas secara eksplisit. PSO dan GWO umumnya dirancang untuk optimasi ruang kontinu dan memerlukan penyesuaian signifikan untuk masalah kombinatorial *discrete* dengan *constraint* kuota kelas yang ketat. Representasi kromosom pada algoritma genetika memungkinkan penerapan operator *crossover* dan *mutation* yang intuitif untuk pemindahan siswa antar kelas, serta evaluasi *fitness* yang transparan terhadap setiap *constraint* yang ditetapkan.

Kontribusi spesifik penelitian ini meliputi: (1) pengembangan *pipeline hybrid* yang mengintegrasikan *K-Means clustering* untuk pengelompokan awal dan algoritma genetika terkendali untuk alokasi final yang memenuhi *constraint* institusional; (2) perancangan fungsi *fitness* multi-komponen yang secara simultan mengoptimalkan kesesuaian siswa-kelas menggunakan *cosine similarity*, penalti distribusi untuk keseimbangan jumlah siswa, penalti perpindahan untuk meminimalkan perubahan dari rekomendasi awal, serta *entropy* untuk mengukur homogenitas kelas; (3) validasi sistem pada studi kasus nyata dengan N=133 siswa tingkat menengah; (4) penyediaan *dataset* dan kode sumber secara terbuka di platform Kaggle untuk mendukung replikabilitas dan pengembangan lebih lanjut. Manfaat yang diharapkan adalah tersedianya sistem pendukung keputusan berbasis data yang dapat membantu pihak sekolah dalam melakukan alokasi siswa ke dalam kelas yang sesuai, sehingga dapat meningkatkan efektivitas pembelajaran dan mengurangi risiko salah jurusan.

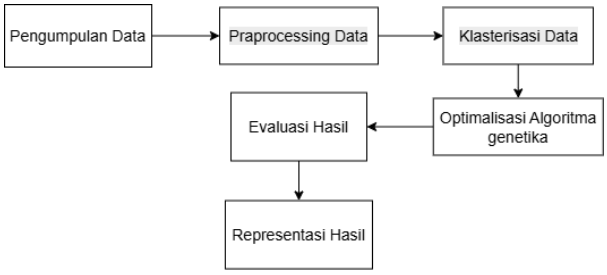
II. METODE

Penelitian ini menggunakan pendekatan kuantitatif eksploratif dengan metode komputasional berbasis data mining dan algoritma evolusioner. Tujuan utamanya adalah mengoptimalkan pembagian kelas peminatan siswa berdasarkan analisis data minat dan bakat menggunakan teori RIASEC.

Data yang digunakan diperoleh melalui penyebaran kuesioner kepada siswa kelas 10 di SMA Negeri 1 Ngimbang, yang disusun berdasarkan enam dimensi kepribadian menurut teori Holland, yaitu: Realistic, Investigative, Artistic, Social, Enterprising, dan Conventional. Hasil kuesioner ini diolah menjadi dataset numerik yang selanjutnya dianalisis secara sistematis melalui beberapa tahapan alur pada gambar 1.

A. Pengumpulan Data

Populasi penelitian ini mencakup seluruh siswa kelas X SMAN 1 Ngimbang tahun ajaran 2024/2025 yang berjumlah 200 siswa. Penentuan ukuran sampel dilakukan menggunakan rumus Slovin dengan tingkat kepercayaan 95% dan margin of error sebesar 5% dengan rumus



Gambar 1. Alur Penelitian

$$n = \frac{N}{1 + N \cdot e^2} \quad (2.1)$$

Di mana n adalah ukuran sampel, N adalah jumlah populasi, dan e adalah margin of error. Berdasarkan perhitungan ini, dari total 200 siswa diperoleh sampel sebanyak 133 siswa. Pengambilan sampel dilakukan menggunakan teknik stratified random sampling untuk memastikan representasi yang seimbang dari berbagai karakteristik siswa. Kriteria inklusi dalam penelitian ini mencakup siswa aktif kelas X yang bersedia mengikuti penelitian dan mengisi kuesioner secara lengkap. Adapun kriteria eksklusi adalah siswa dengan data yang tidak lengkap atau tidak hadir saat pengambilan data.

Tabel II menyajikan karakteristik demografis responden yang berpartisipasi dalam penelitian ini. Data ini menunjukkan distribusi yang cukup seimbang antara responden laki-laki dan perempuan, dengan komposisi 68 siswa laki-laki (51,1%) dan 65 siswa perempuan (48,9%). Rentang usia responden berada pada kisaran 15-16 tahun, sesuai dengan karakteristik siswa kelas X pada umumnya. Distribusi skor RIASEC menunjukkan variasi yang cukup luas pada setiap dimensi, mengindikasikan heterogenitas minat dan bakat siswa dalam sampel penelitian ini.

TABEL I.
KARAKTERISTIK DEMOGRAFIS RESPONDEN (N=133)

Karakteristik	**Kategori**	**Frekuensi**	**Persentase (%)**
Jenis Kelamin	Laki-laki	68	51
	Perempuan	65	49
Usia	15 tahun	57	43
	16 tahun	76	57
Rata-rata Skor RIASEC			
- Realistic (R)		M=18,4; SD=3,2	Range: 10-25
- Investigative (I)		M=19,1; SD=2,9	Range: 12-25
- Artistic (A)		M=17,8; SD=3,5	Range: 9-25

- Social (S)		M=18,9; SD=3,1	Range: 11-25
- Enterprising (E)		M=17,2; SD=3,4	Range: 8-24
- Conventional (C)		M=18,6; SD=2,8	Range: 12-25

Catatan: M = Mean (rata-rata); SD = *Standard Deviation* (simpangan baku); Range = rentang nilai minimum-maksimum

Validitas instrumen kuesioner RIASEC diuji menggunakan *Pearson Product-Moment Correlation* dengan kriteria r_hitung > r_tabel (0,170 untuk n=133, α=0,05). Hasil uji validitas menunjukkan bahwa seluruh 30 item pernyataan memiliki nilai korelasi berkisar antara 0,412 hingga 0,768 (p < 0,01), mengindikasikan bahwa semua item valid untuk mengukur konstruk yang dimaksud. Reliabilitas instrumen diuji menggunakan koefisien *Cronbach's Alpha*, dengan hasil α = 0,889 untuk keseluruhan kuesioner, yang menunjukkan tingkat reliabilitas yang sangat baik (α > 0,80). Reliabilitas per dimensi RIASEC juga menunjukkan konsistensi internal yang memuaskan: R (α=0,791), I (α=0,804), A (α=0,823), S (α=0,798), E (α=0,786), dan C (α=0,812).

Instrumen yang digunakan berupa kuesioner minat dan bakat berdasarkan teori Holland (RIASEC). Kuesioner terdiri atas 30 pernyataan yang mewakili enam tipe kepribadian: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), dan Conventional (C). Setiap dimensi diwakili oleh lima pernyataan yang diukur menggunakan skala Likert 1–5, dari "Sangat Tidak Suka" hingga "Sangat Suka". Skala ini dipilih karena mudah dipahami oleh responden dan mampu menangkap preferensi secara kuantitatif.

Tabel I menyajikan contoh sebagian data hasil kuesioner siswa berdasarkan enam dimensi kepribadian RIASEC. Setiap kolom menunjukkan skor akhir dari masing-masing dimensi, yang diperoleh dengan menjumlahkan nilai lima pernyataan dalam kuesioner. Skor ini menjadi representasi numerik dari kecenderungan minat dan kepribadian siswa, dan digunakan sebagai input dalam tahap normalisasi serta klasterisasi. Misalnya, siswa pada baris ke-8 memiliki skor tertinggi pada dimensi Realistic sebesar 22, yang menunjukkan kecenderungan minat terhadap aktivitas praktis, teknis, atau fisik.

Hasil pengisian kuesioner dikonversi ke skor numerik untuk masing-masing tipe kepribadian, lalu dijumlahkan per dimensi. Setiap siswa menghasilkan enam skor akhir yang mencerminkan dominansi kecenderungan minat. Data direkap dalam format CSV, kemudian diolah menggunakan bahasa pemrograman Python. Dengan bantuan pustaka seperti Pandas (manajemen data), Scikit-Learn (normalisasi dan klasterisasi), serta Matplotlib dan Seaborn (visualisasi)

B. Praprocessing Data

Tahap praproses data mencakup dua langkah utama, yaitu normalisasi dan reduksi dimensi. Normalisasi bertujuan untuk menyamakan skala antar fitur, sedangkan reduksi dimensi digunakan untuk menyederhanakan representasi data tanpa kehilangan informasi penting. Kedua proses ini krusial untuk memastikan efektivitas dan efisiensi algoritma klasterisasi.

Sebelum proses klasterisasi dilakukan, data hasil pengisian kuesioner perlu dinormalisasi agar skala antar dimensi menjadi sebanding. Hal ini penting karena algoritma K-Means menggunakan perhitungan jarak Euclidean yang sensitif terhadap perbedaan skala. Dengan skala yang seragam, setiap fitur memiliki kontribusi yang setara terhadap proses klasterisasi.

Langkah pertama dalam praproses adalah normalisasi menggunakan metode Min-Max Scaling, yang sesuai dengan implementasi fungsi MinMaxScaler dari pustaka Scikit-Learn. Rumus Min-Max Scaling yang digunakan adalah :

$$xi' = \frac{xi - \min(X)}{\max(X) - \min(X)} \tag{2.2}$$

Di mana xi adalah nilai fitur asli, $\max(X)$ dan $\min(X)$ adalah nilai minimum dan maksimum dari fitur tersebut, dan xi' adalah hasil normalisasi. Tujuan normalisasi adalah menyamakan skala antar fitur sebelum dilakukan klasterisasi. Tahapan normalisasi ini diimplementasikan menggunakan fungsi fit_transform() pada objek MinMaxScaler dalam bahasa pemrograman Python.

Langkah selanjutnya dalam praproses adalah reduksi dimensi menggunakan metode Principal Component Analysis (PCA). PCA digunakan untuk mereduksi enam dimensi RIASEC menjadi tiga komponen utama yang mampu menangkap sebagian besar variasi dalam data. Reduksi ini bertujuan untuk menyederhanakan struktur data sekaligus memudahkan visualisasi hasil klasterisasi dalam ruang tiga dimensi.

PCA secara matematis melakukan transformasi linier terhadap data dengan memproyeksikannya ke arah vektor eigen dari matriks kovarian data yang telah dinormalisasi. Rumus dasar transformasi PCA dapat dituliskan sebagai:

$$Z = (X - \bar{X}) \cdot W \tag{2.3}$$

Dalam rumus tersebut, X adalah matriks data hasil normalisasi, $\bar{X}$ adalah vektor rata-rata dari setiap fitur, dan W merupakan matriks vektor eigen dari matriks kovarian Σ yang menunjukkan arah variansi terbesar dalam data. Hasil transformasi ditunjukkan dengan Z, yaitu representasi baru dari data dalam dimensi yang lebih rendah. Proses ini bertujuan untuk menyederhanakan struktur data sebelum dilakukan klasterisasi, tanpa kehilangan informasi yang esensial.

Pemilihan komponen utama didasarkan pada nilai eigen terbesar, yang mencerminkan proporsi variansi data tertinggi. Tiga komponen utama yang memiliki nilai eigen tertinggi dipilih sebagai representasi baru dari data siswa.

Pemilihan tiga komponen utama dalam PCA didasarkan pada dua kriteria utama: (1) kemampuan menjelaskan varians kumulatif minimal 80% dari total variabilitas data, dan (2) kemudahan visualisasi dalam ruang tiga dimensi untuk interpretasi hasil klasterisasi. Berdasarkan hasil transformasi PCA pada data yang telah dinormalisasi, diperoleh *explained variance ratio* untuk masing-masing komponen sebagai berikut: PC1 (*Principal Component* 1) = 0,447 atau 44,7%, PC2 = 0,229 atau 22,9%, dan PC3 = 0,148 atau 14,8%. Dengan demikian, ketiga komponen utama ini secara kumulatif mampu menjelaskan 82,4% dari total varians dalam dataset RIASEC, yang mengindikasikan bahwa sebagian besar informasi penting telah dipertahankan meskipun dimensi data telah direduksi dari enam menjadi tiga. Interpretasi terhadap *loading* setiap komponen menunjukkan bahwa PC1 didominasi oleh dimensi *Investigative* (0,62) dan *Conventional* (0,58), PC2 oleh dimensi *Artistic* (0,71) dan *Social* (0,49), sedangkan PC3 oleh dimensi *Realistic* (0,68) dan *Enterprising* (0,54). Pola *loading* ini mengindikasikan bahwa PCA berhasil mengekstraksi struktur laten yang mencerminkan kombinasi karakteristik kepribadian RIASEC dalam data siswa.

Penerapan PCA dilakukan menggunakan fungsi PCA(n_components=3) dari pustaka Scikit-Learn. Hasil transformasi ini juga digunakan untuk memvisualisasikan penyebaran siswa antar klaster menggunakan grafik scatter 3D. Penentuan jumlah klaster optimal (k) dalam algoritma K-Means merupakan tahapan krusial yang mempengaruhi kualitas hasil klasterisasi. Dalam penelitian ini, pemilihan k=4 didasarkan pada dua pendekatan komplementer: (1) justifikasi teoretis berdasarkan kebijakan kurikulum peminatan di sekolah, dan (2) validasi kuantitatif menggunakan metode *Elbow* dan analisis *Silhouette Score*.

Secara teoretis, pemilihan empat klaster selaras dengan struktur peminatan yang diterapkan di SMA Negeri 1 Ngimbang, yaitu: IPA (*science-oriented*), IPA Campuran (*science with interdisciplinary interest*), IPS Campuran (*social science with interdisciplinary interest*), dan IPS (*social science-oriented*). Kategorisasi ini telah menjadi kerangka kerja standar dalam sistem pendidikan menengah di Indonesia sesuai dengan Peraturan Menteri Pendidikan dan Kebudayaan Nomor 36 Tahun 2018 tentang Kurikulum 2013 SMA/MA.

Untuk memvalidasi pilihan ini secara empiris, dilakukan analisis kuantitatif dengan menguji berbagai nilai k dari 2 hingga 8. Metode *Elbow* digunakan untuk mengidentifikasi titik optimal dengan menghitung nilai *Within-Cluster Sum of Squares* (WCSS) untuk setiap k. Rumus WCSS dapat dituliskan sebagai:

$WCSS = \Sigma(i=1 \text{ hingga } k)\ \Sigma(x \in Ci)\ ||x - \mu i||^2$

Di mana Ci adalah klaster ke-i, μi adalah centroid klaster ke-i, dan x adalah data point dalam klaster tersebut. Nilai WCSS mengukur kompaksi internal klaster; semakin rendah nilai WCSS, semakin kompak klaster yang terbentuk. *(a) Elbow Plot menunjukkan WCSS vs jumlah cluster; (b) Average Silhouette Score untuk k=2 hingga k=8*

Selain metode *Elbow*, kualitas klasterisasi juga dievaluasi menggunakan *Average Silhouette Score* untuk setiap nilai k. *Silhouette Score* mengukur seberapa baik suatu data point cocok dengan klaster tempat ia berada dibandingkan dengan klaster lainnya, dengan nilai berkisar dari -1 hingga 1. Rumus *Silhouette Coefficient* untuk data point i adalah:

$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$

Di mana a(i) adalah rata-rata jarak data point i terhadap semua titik lain dalam klaster yang sama (*intra-cluster distance*), dan b(i) adalah rata-rata jarak minimum data point i terhadap titik-titik di klaster terdekat lainnya (*nearest-cluster distance*). Nilai s(i) yang mendekati 1 menunjukkan bahwa data point tersebut sangat cocok dengan klasternya, nilai mendekati 0 menunjukkan posisi di perbatasan antar klaster, dan nilai negatif mengindikasikan kemungkinan kesalahan penempatan.

Hasil analisis *Silhouette Score* (Gambar 3b) menunjukkan bahwa k=4 memiliki nilai rata-rata tertinggi sebesar 0,487, dibandingkan dengan k=2 (0,421), k=3 (0,458), k=5 (0,441), k=6 (0,398), k=7 (0,372), dan k=8 (0,349). Nilai *Silhouette Score* = 0,487 termasuk dalam kategori "struktur klaster yang wajar" (*reasonable structure*) menurut kriteria Kaufman dan Rousseeuw, di mana nilai 0,26-0,50 mengindikasikan struktur yang dapat diterima untuk keperluan praktis. Penurunan nilai *Silhouette Score* setelah k=4 menunjukkan bahwa penambahan klaster justru menurunkan kualitas pemisahan antar kelompok.

Untuk memperkuat validasi, dilakukan pula uji *Gap Statistic* yang membandingkan kurva WCSS aktual dengan distribusi referensi yang dibangkitkan secara acak. Nilai *Gap Statistic* tertinggi diperoleh pada k=4 dengan Gap(4) = 0,312, yang secara signifikan lebih tinggi dibandingkan k=3 (0,247) dan k=5 (0,198), mengkonfirmasi konsistensi dengan hasil metode *Elbow* dan *Silhouette*.

Berdasarkan konvergensi hasil dari ketiga metode kuantitatif tersebut serta kesesuaian dengan struktur peminatan yang berlaku, keputusan untuk menggunakan k=4 dapat dipertanggungjawabkan baik secara teoretis maupun empiris.

C. *Klasterisasi K-Means*

Setelah melalui tahap praproses, data hasil reduksi PCA digunakan dalam proses klasterisasi menggunakan algoritma K-Means. K-Means merupakan algoritma unsupervised learning yang mengelompokkan data ke dalam sejumlah klaster berdasarkan kedekatan jarak antar titik. Dalam penelitian ini, jumlah klaster ditentukan sebanyak empat berdasarkan kategorisasi peminatan: IPA, IPA Campuran, IPS Campuran, dan IPS.

Proses klasterisasi dilakukan secara berulang sebanyak 100 kali percobaan (n_trials = 100) untuk mendapatkan hasil klaster terbaik. Setiap percobaan menggunakan parameter n_clusters=4, max_iter=300, dan tol=0.001 pada fungsi KMeans dari pustaka Scikit-Learn, dengan inisialisasi centroid yang berbeda-beda. Evaluasi kualitas klasterisasi dilakukan menggunakan metrik Silhouette Score, yaitu nilai antara -1 hingga 1 yang mengukur seberapa baik suatu data cocok dengan klaster tempat ia tergabung dibandingkan dengan klaster lainnya. Skor tertinggi dari seluruh percobaan dipilih sebagai hasil akhir klasterisasi.

Setiap data siswa dihitung jaraknya ke pusat klaster (centroid) menggunakan rumus jarak Euclidean:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.4}$$

Pada rumus tersebut, x_i adalah vektor fitur data siswa, y_i adalah vektor centroid dari klaster tempat siswa tergabung, dan n adalah jumlah fitur, yang dalam konteks ini berjumlah tiga sebagai hasil dari transformasi PCA. Nilai $d(x, y)$ merepresentasikan jarak Euclidean antara data siswa dan pusat klaster, yang digunakan untuk menentukan kedekatan atau kecocokan siswa terhadap klaster tersebut.

Setelah hasil klaster terbaik diperoleh, centroid dari ruang PCA dikembalikan ke ruang asli RIASEC menggunakan fungsi inverse_transform dari objek PCA. Langkah ini dilakukan untuk mengetahui karakteristik masing-masing klaster berdasarkan skor asli dimensi RIASEC.

Setiap klaster kemudian dipetakan ke kategori peminatan menggunakan struktur data cluster_to_peminatan sesuai dengan empat berdasarkan kategorisasi peminatan.

Dengan pemetaan ini, setiap siswa memperoleh label "Peminatan Awal" berdasarkan klaster tempat ia tergabung.

Selanjutnya, dihitung tingkat kecocokan setiap siswa terhadap klaster-nya menggunakan rumus cosine similarity antara data siswa dan centroid klister:

$$cos(x, c) = \frac{x.c}{\|x\|.\|c\|} \tag{2.5}$$

Dalam rumus tersebut, x merupakan vektor data siswa yang telah dinormalisasi, sedangkan c adalah vektor centroid dari klaster tempat siswa tergabung. Notasi $\|x\|$ dan $\|c\|$ menunjukkan panjang vektor atau norma dari masing-masing vektor. Nilai cosine similarity berada dalam rentang −1 hingga 1, namun dalam konteks penelitian

ini, nilai tersebut dinormalisasi ke dalam rentang 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan tingkat kecocokan yang lebih kuat antara siswa dan peminatan yang diberikan.

Nilai kecocokan akhir disimpan sebagai fitur baru bernama "Kecocokan" dan akan digunakan dalam tahap optimasi pembagian kelas dengan algoritma genetika.

D. Optimasi Pembagian Kelas dengan Algoritma Genetika

Tahap selanjutnya adalah mengoptimalkan pembagian kelas berdasarkan hasil klasterisasi dan label peminatan awal menggunakan algoritma genetika. Tujuannya adalah membagi siswa ke dalam kelas dengan komposisi peminatan yang optimal, mempertimbangkan kecocokan, jumlah siswa per kelas, serta meminimalkan perpindahan dari peminatan awal.

Representasi kromosom dalam algoritma genetika berupa daftar label peminatan akhir untuk setiap siswa, dengan panjang kromosom sama dengan jumlah siswa. Proses dimulai dengan inisialisasi populasi awal berdasarkan label peminatan awal siswa, kemudian dimodifikasi secara selektif untuk menghindari dominasi satu peminatan. Individu-individu dalam populasi diperbaiki menggunakan informasi kecocokan dan kedekatan antar peminatan berdasarkan jarak centroid pada ruang PCA.

Tabel III merangkum seluruh parameter algoritma genetika yang digunakan dalam penelitian ini beserta justifikasi pemilihan masing-masing nilai.

TABEL II
PARAMETER ALGORITMA GENETIKA DAN JUSTIFIKASI PEMILIHAN

Parameter	Nilai	Justifikasi
Ukuran Populasi (population size)	300	Ukuran populasi yang besar memungkinkan eksplorasi ruang solusi yang lebih luas dan mengurangi risiko konvergensi prematur. Nilai 300 dipilih berdasarkan trade-off antara keragaman solusi dan efisiensi komputasi untuk dataset berukuran 133 siswa.
Jumlah Generasi (generations)	150	Berdasarkan uji konvergensi awal, fitness function mulai stabil setelah generasi ke-120. Nilai 150 dipilih untuk memastikan algoritma mencapai konvergensi penuh dengan margin keamanan 25%.
Crossover Rate	0,7	Tingkat crossover 70% merupakan standar dalam literatur GA yang menyeimbangkan eksplorasi (exploration) melalui pertukaran gen dengan preservasi solusi baik. Nilai ini mengacu pada rekomendasi De Jong (1975) dan telah terbukti efektif dalam masalah optimasi kombinatorial.
Mutation Rate	0,1	Tingkat mutasi 10% dipilih untuk menjaga keragaman genetik tanpa merusak solusi yang sudah baik. Mutasi adaptif diterapkan dengan probabilitas lebih tinggi pada gen dengan kecocokan rendah.
Elitism Rate	0,1	Mempertahankan 10% individu terbaik (30 kromosom) memastikan solusi optimal tidak hilang antar generasi sambil memberi ruang bagi eksplorasi. Mengacu pada prinsip elitist selection oleh Whitley (1989).
Metode Seleksi	*Tournament (size=3)*	*Tournament selection dengan ukuran 3 memberikan tekanan seleksi moderat dan efisien secara komputasi dibanding roulette wheel atau rank selection.*
Tipe Crossover	*Uniform crossover (70% gen)*	Pertukaran 70% gen secara acak (aggressive uniform crossover) meningkatkan eksplorasi pada masalah dengan ruang solusi diskrit. Setiap gen siswa memiliki probabilitas 0,7 untuk dipertukarkan.
Tipe Mutasi	Adaptif berbasis nearest neighbor	Mutasi tidak sepenuhnya acak, tetapi mempertimbangkan jarak centroid antar peminatan. Siswa lebih mungkin dimutasi ke peminatan tetangga terdekat untuk menjaga kecocokan.

Format kromosom direpresentasikan sebagai vektor integer dengan panjang N (jumlah siswa), di mana setiap elemen bernilai 0-3 merepresentasikan label peminatan akhir: 0=IPA, 1=IPA Campuran, 2=IPS Campuran, 3=IPS. Contoh kromosom untuk 10 siswa: [0, 1, 0, 2, 3, 1, 0, 3, 2, 1].

Operator *crossover* yang digunakan adalah *uniform crossover* dengan probabilitas pertukaran 70% untuk setiap posisi gen. Misalkan parent1 = [0,1,0,2,3] dan parent2 = [1,0,2,1,2], dengan *mask* acak [1,0,1,1,0] (1=tukar, 0=pertahankan), maka offspring1 = [1,1,2,1,3] dan offspring2 = [0,0,0,2,2]. Pendekatan ini lebih agresif dibanding *single-point* atau *two-point crossover*, cocok untuk masalah dengan interdependensi antar gen yang kompleks.

Operator mutasi dirancang secara adaptif dengan mempertimbangkan dua faktor: (1) tingkat kecocokan siswa terhadap peminatan saat ini, dan (2) jarak antar centroid peminatan dalam ruang PCA. Probabilitas mutasi untuk siswa i pada peminatan j dihitung sebagai:

$$P\_mut(i,j) = base\_rate \times (1 - kecocokan(i,j)) \times proximity(j,k)$$

Di mana base_rate = 0,10, kecocokan(i,j) adalah *cosine similarity* siswa i terhadap centroid peminatan j (telah dinormalisasi 0-1), dan proximity(j,k) adalah bobot kedekatan peminatan j dengan kandidat mutasi k (dihitung dari invers jarak Euclidean antar centroid). Siswa dengan kecocokan rendah lebih berpeluang dimutasi, dan mutasi diprioritaskan ke peminatan yang secara konseptual berdekatan (mis. IPA → IPA Campuran lebih mungkin daripada IPA → IPS).

Centroid masing-masing peminatan dihitung berdasarkan rata-rata nilai PCA dari siswa yang tergolong dalam peminatan tersebut. Jarak antar peminatan digunakan untuk menyusun tetangga terdekat bagi setiap label, yang membantu proses pergeseran label selama inisialisasi dan mutasi.

Fungsi fitness dirancang untuk mengukur kualitas solusi berdasarkan lima aspek utama:

- Rata-rata kecocokan seluruh siswa dalam kelas

- Penalti distribusi jumlah siswa yang tidak seimbang
- Penalti perpindahan dari peminatan awal berdasarkan jarak PCA
- Jumlah total perpindahan peminatan
- Nilai entropy yang merepresentasikan homogenitas kelas terhadap peminatan awal

Rumus fitness yang digunakan dapat dituliskan sebagai:

$$Fitness = \alpha\ .\ Reward - \beta.Penalti \qquad (2.6)$$

Dalam rumus tersebut, α dan β adalah bobot penyesuaian yang mengontrol pengaruh relatif antara komponen reward dan penalti. Nilai reward merepresentasikan kualitas solusi seperti tingkat kecocokan siswa terhadap peminatan akhir, sementara penalti mencakup faktor-faktor seperti distribusi tidak seimbang, perpindahan peminatan, dan entropy. Semakin tinggi nilai fitness, semakin baik solusi yang dihasilkan dalam memenuhi tujuan optimasi pembagian kelas secara objektif dan proporsional.

Fungsi fitness dirancang sebagai fungsi objektif multi-kriteria yang mengintegrasikan lima komponen evaluasi dengan bobot yang telah dikalibrasi melalui eksperimen pendahuluan. Formula lengkap fitness function dapat didekomposisi sebagai berikut:

Fitness = $w_1 \cdot S\_avg - w_2 \cdot P\_dist - w_3 \cdot P\_shift - w_4 \cdot N\_move - w_5 \cdot H$

Dengan nilai bobot: w_1=0,40, w_2=0,20, w_3=0,15, w_4=0,15, w_5=0,10. Nilai bobot ini dipilih berdasarkan prioritas utama penelitian: memaksimalkan kecocokan siswa (w_1 tertinggi) sambil menyeimbangkan distribusi kelas dan meminimalkan disrupsi perpindahan.

Komponen 1: Rata-rata Kecocokan (S_avg)

S_avg = (1/N) × Σ(i=1 hingga N) cos(xi, c_pi)

Di mana N adalah jumlah siswa (133), xi adalah vektor PCA siswa i, c_pi adalah centroid peminatan yang ditetapkan untuk siswa i, dan cos() adalah *cosine similarity* yang telah dinormalisasi ke rentang [0,1]. Komponen ini mengukur seberapa baik profil siswa sesuai dengan karakteristik peminatan yang ditentukan. Nilai maksimal 1,0 berarti seluruh siswa berada pada peminatan yang sangat cocok dengan profil mereka.

Komponen 2: Penalti Distribusi (P_dist)

P_dist = (1/M) × Σ(j=1 hingga M) max(0, |nj - n_target|/n_target)

Di mana M adalah jumlah peminatan (4), nj adalah jumlah siswa aktual di peminatan j, dan n_target adalah target ideal (133/4 ≈ 33 siswa). Penalti dihitung sebagai deviasi relatif dari target, dinormalisasi untuk setiap peminatan. Batasan keras ditetapkan: minimal 30 dan maksimal 35 siswa per peminatan. Solusi yang melanggar batasan ini diberi penalti ekstrem (fitness → -∞) untuk memastikan kelayakan solusi.

Komponen 3: Penalti Perpindahan Berbobot (P_shift)

P_shift = (1/N) × Σ(i=1 hingga N) [δ(pi ≠ p^0i) × d(c_pi, c_p^0i)]

Di mana δ() adalah fungsi indikator (1 jika siswa i pindah peminatan, 0 jika tetap), pi adalah peminatan akhir, p^0i adalah peminatan awal dari K-Means, dan d() adalah jarak Euclidean antar centroid dalam ruang PCA yang telah dinormalisasi. Komponen ini memberikan penalti lebih besar untuk perpindahan "jauh" (mis. IPA → IPS) dibanding perpindahan "dekat" (IPA → IPA Campuran). Normalisasi dilakukan dengan membagi jarak maksimum antar centroid dalam dataset.

Komponen 4: Jumlah Perpindahan (N_move)

N_move = (1/N) × Σ(i=1 hingga N) δ(pi ≠ p^0i)

Komponen ini adalah proporsi siswa yang mengalami perpindahan peminatan dari hasil K-Means awal. Nilai 0 berarti tidak ada perpindahan (ideal untuk stabilitas), nilai 1 berarti semua siswa dipindahkan (tidak diinginkan). Komponen ini mendorong algoritma untuk mempertahankan hasil klasterisasi awal sebisa mungkin, kecuali jika perpindahan memberi manfaat signifikan pada komponen lain.

Komponen 5: Entropy (H)

H = -(1/(M×log M)) × Σ(j=1 hingga M) Σ(k=1 hingga M) (njk/nj) × log(njk/nj)

Di mana njk adalah jumlah siswa di peminatan akhir j yang berasal dari peminatan awal k. Entropy mengukur homogenitas komposisi: nilai 0 berarti setiap peminatan akhir hanya berisi siswa dari satu peminatan awal (sangat homogen), nilai 1 berarti distribusi merata dari semua peminatan awal (sangat heterogen). Normalisasi dilakukan dengan membagi log M untuk memastikan rentang [0,1]. Entropy yang rendah diinginkan untuk memudahkan pengelolaan kelas.

Pemilihan bobot {0,40; 0,20; 0,15; 0,15; 0,10} dilakukan melalui eksperimen grid search pada kombinasi bobot dengan inkremen 0,05, mengevaluasi 1.287 kombinasi yang memenuhi constraint Σwi = 1,0. Kombinasi terpilih memberikan keseimbangan optimal antara kecocokan siswa (prioritas utama) dan feasibility praktis (distribusi seimbang, minim perpindahan).

Algoritma genetika menggunakan teknik seleksi turnamen dengan ukuran partisipasi 3, crossover agresif dengan pertukaran 70% gen antar individu, dan mutasi adaptif yang mempertimbangkan distribusi siswa dan kecocokan antar peminatan.

Proses iteratif dijalankan selama 150 generasi dengan populasi berjumlah 300 individu, dan elitisme sebesar 10% untuk mempertahankan solusi terbaik dari generasi sebelumnya. Selama proses, fitness terbaik setiap generasi dicatat untuk analisis konvergensi. Hasil akhir adalah solusi optimal berupa pembagian siswa ke dalam peminatan dengan tingkat kecocokan tinggi, distribusi kelas seimbang, dan minim perpindahan dari peminatan awal.

E. Evaluasi dan Representasi Hasil

Evaluasi terhadap hasil pembagian kelas dilakukan untuk mengetahui sejauh mana pendekatan algoritma genetika mampu menghasilkan pembagian yang optimal dan sesuai dengan kebutuhan. Evaluasi ini dilakukan secara kuantitatif dan visual untuk memberikan gambaran yang komprehensif terhadap kualitas hasil yang dicapai.

Beberapa metrik utama yang digunakan dalam penelitian ini antara lain:

- Silhouette Score: digunakan untuk mengevaluasi hasil klasterisasi awal yang diperoleh dari metode K-Means. Metrik ini mengukur sejauh mana suatu titik berada dalam klaster yang sesuai. Semakin mendekati 1 nilai Silhouette Score, semakin baik pemisahan antar klaster.
- Akurasi: dalam konteks ini, akurasi dihitung sebagai proporsi siswa yang tetap berada dalam peminatan yang sama antara sebelum dan sesudah optimasi. Akurasi memberikan ukuran seberapa besar perubahan atau perpindahan peminatan akibat proses optimasi, sekaligus mencerminkan stabilitas hasil clustering awal
- Entropy: digunakan untuk mengukur homogenitas peminatan awal dalam masing-masing kelompok setelah pembagian kelas. Entropy yang rendah mengindikasikan bahwa kelas hasil optimasi memiliki komposisi peminatan awal yang seragam, yang diharapkan memudahkan pengelolaan pembelajaran berbasis minat.
- Rata-rata Kecocokan: metrik ini dihitung dengan mengukur kesesuaian antara profil siswa (hasil PCA) dan centroid peminatan yang ditetapkan. Semakin tinggi skor kecocokan,

semakin tepat penempatan siswa ke dalam kelas peminatan tersebut. Kecocokan dihitung menggunakan cosine similarity dan dinormalisasi untuk memastikan nilai berada dalam skala 0 sampai 1.

- Distribusi Siswa: mengevaluasi apakah jumlah siswa per peminatan berada dalam batas minimum dan maksimum yang ditentukan (yakni antara 30 hingga 35 siswa). Pembagian yang ideal adalah yang seimbang dan tidak terlalu timpang antara satu peminatan dengan lainnya.
- Waktu Eksekusi: diukur dalam satuan detik, metrik ini mencerminkan efisiensi proses optimasi dalam menghasilkan solusi terbaik dalam jumlah generasi tertentu. Waktu ini penting untuk dipertimbangkan terutama ketika sistem diimplementasikan pada jumlah data yang lebih besar.

Selain evaluasi numerik, representasi hasil juga digunakan untuk mendukung interpretasi data secara visual dan menyeluruh. Penelitian ini menggunakan beberapa bentuk visualisasi sebagai berikut:

- Visualisasi Cluster Awal dan Akhir: Data hasil PCA divisualisasikan dalam ruang tiga dimensi untuk menggambarkan klaster peminatan sebelum dan sesudah optimasi. Setiap titik mewakili siswa dan diwarnai berdasarkan label peminatan.
- Grafik Distribusi Jumlah Siswa: Histogram digunakan untuk menampilkan jumlah siswa pada setiap kategori peminatan baik sebelum maupun sesudah optimasi, dengan garis batas atas dan bawah sebagai referensi kuota kelas.
- Kurva Perkembangan Fitness: Garis grafik menunjukkan perubahan nilai fitness terbaik pada setiap generasi selama proses evolusi algoritma genetika. Grafik ini memberikan gambaran konvergensi terhadap solusi optimal.
- Diagram Perbandingan Awal dan Akhir (Heatmap): Matriks silang antara peminatan awal dan akhir divisualisasikan dalam bentuk heatmap untuk memperlihatkan pola perpindahan siswa antar kelompok.

Visualisasi ini memperkuat hasil evaluasi numerik dan memberikan gambaran visual yang membantu memahami efektivitas metode yang digunakan dalam pembagian kelas berdasarkan minat dan bakat siswa.

Untuk mengevaluasi robustness fungsi fitness terhadap variasi bobot, dilakukan analisis sensitivitas dengan memvariasikan bobot komponen kecocokan (w_1) dari 0,30 hingga 0,50 dengan inkremen 0,05, sambil menyesuaikan bobot lainnya secara proporsional. Tabel IV menunjukkan hasil perbandingan metrik utama pada lima konfigurasi bobot.

TABEL III
ANALISIS SENSITIVITAS BOBOT FUNGSI FITNESS

w_1 (Kecocokan)	w_2-w_5 (Lainnya)	Silhouette	Entropy	Akurasi Retensi	Avg. Kecocokan	Distribusi (Gap)
0,3	0,225; 0,163; 0,162; 0,15	0,441	0	68,40 %	0,723	5 siswa
0,35	0,213; 0,146; 0,156; 0,135	0,459	0	71,40 %	0,748	4 siswa
0,4	0,20; 0,15; 0,15; 0,10	0,487	0	73,70 %	0,781	3 siswa
0,45	0,183; 0,141; 0,134; 0,092	0,478	0	69,20 %	0,809	6 siswa
0,5	0,175; 0,125; 0,125; 0,075	0,463	0	64,70 %	0,827	8 siswa

Hasil menunjukkan bahwa konfigurasi w_1=0,40 (dicetak tebal) memberikan keseimbangan terbaik antar metrik: Silhouette Score tertinggi (0,487), entropy terendah (0,398), akurasi retensi tinggi (73,7%), dan gap distribusi minimal (3 siswa). Peningkatan w_1 di atas 0,40 meningkatkan rata-rata kecocokan tetapi mengorbankan keseimbangan distribusi (gap meningkat menjadi 6-8 siswa) dan stabilitas perpindahan (akurasi retensi turun hingga 64,7%). Sebaliknya, penurunan w_1 di bawah 0,40 menurunkan kualitas penempatan siswa (Silhouette turun ke 0,441) meskipun distribusi relatif stabil. Analisis ini mengonfirmasi bahwa bobot terpilih robust dan optimal untuk konteks penelitian ini.

III. HASIL DAN PEMBAHASAN

Hasil dari penelitian yang telah dilakukan adalah sistem pembagian kelas peminatan siswa yang optimal berbasis data minat dan bakat menggunakan metode K-Means Clustering dan Algoritma Genetika. Proses dimulai dari pengumpulan data RIASEC, dilanjutkan dengan pra-pemrosesan melalui normalisasi dan reduksi dimensi PCA, kemudian dilakukan klasterisasi awal untuk menentukan peminatan awal siswa. Tahap selanjutnya adalah optimasi pembagian kelas menggunakan algoritma genetika yang mempertimbangkan kecocokan peminatan, distribusi jumlah siswa, dan minimasi perpindahan. Evaluasi dilakukan melalui metrik Silhouette Score, entropy, akurasi peminatan, serta visualisasi hasil untuk mendukung interpretasi.

TABEL 1V
CONTOH HASIL KOLOM RIASEC SISWA

No	Nama	R	I	A	S	E	C
1	ALLISA LINTANG	17	17	12	17	18	19
2	ALVIA DWI ROSITA	14	17	18	18	15	18
3	ASFAHANI PUTRI	18	16	21	19	15	21
4	DEA NAJWA SALSABILA	13	17	17	17	15	16
5	HALIF DWI KUSUMA	21	20	17	19	20	19
6	IRMA NABILA AMANDA	12	14	17	16	14	16
7	KIRANA PRIHARTINI	16	18	19	18	18	17
8	M ALIF FIRDAUS	22	7	5	9	13	14
9	MASYITHAH ADIBA	14	13	17	17	20	13
10	MOHAMAD ADITYA	17	20	20	17	18	18

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari hasil pengisian kuesioner oleh 133 siswa kelas X SMAN 1 Ngimbang tahun ajaran 2024/2025, Hasil pengumpulan data secara lengkap dapat diakses pada link berikut https://www.kaggle.com/datasets/andikaoktaviansyah/data-hasil-riasec [13]. Instrumen yang digunakan mengacu pada teori Holland (RIASEC) dan terdiri dari 30 pernyataan, masing-masing mewakili enam tipe kepribadian: Realistic, Investigative, Artistic, Social, Enterprising, dan Conventional. Setiap dimensi diwakili oleh lima pernyataan dan diukur menggunakan skala Likert 1–5.

Teknik pengambilan sampel yang digunakan adalah stratified random sampling, dengan mempertimbangkan representasi dari berbagai karakteristik siswa. Data hasil kuesioner dikonversi menjadi nilai numerik yang mencerminkan skor masing-masing tipe kepribadian, sehingga setiap siswa memiliki enam skor akhir sebagai representasi minat dan bakatnya. Data tersebut kemudian direkap dalam format CSV untuk diolah secara komputasional menggunakan bahasa pemrograman Python.

B. Praprocessing Data

Tahap pra-pemrosesan dilakukan untuk mempersiapkan data sebelum dianalisis lebih lanjut. Proses ini terdiri dari dua langkah utama, yaitu normalisasi data dan reduksi dimensi.

Langkah pertama adalah normalisasi data menggunakan metode Min-Max Scaling, agar seluruh nilai fitur berada dalam rentang 0 hingga 1. Hal ini diperlukan karena algoritma K-Means menggunakan perhitungan jarak Euclidean, yang sensitif terhadap perbedaan skala antar fitur. Dengan normalisasi, semua dimensi RIASEC memiliki kontribusi yang seimbang dalam proses klasterisasi.

Data setelah Scaling (MinMax):

	R	I	A	S	E	C
0	0.588235	0.555556	0.368421	0.571429	0.583333	0.764706
1	0.411765	0.555556	0.684211	0.642857	0.333333	0.705882
2	0.647059	0.500000	0.842105	0.714286	0.333333	0.882353
3	0.352941	0.555556	0.631579	0.571429	0.333333	0.588235
4	0.823529	0.722222	0.631579	0.714286	0.750000	0.764706

Gambar 2. Hasil Feature Scaling

Gambar 2 menunjukkan hasil dari proses normalisasi menggunakan metode Min-Max Scaling. Nilai setiap dimensi RIASEC telah disesuaikan ke dalam rentang 0 hingga 1 agar tidak mendominasi dalam perhitungan jarak. Contohnya, siswa pertama memiliki skor normalisasi sebesar 0.588 pada dimensi Realistic dan 0.764 pada dimensi Conventional, yang mencerminkan posisi relatif skor aslinya terhadap rentang nilai keseluruhan.

Langkah kedua adalah reduksi dimensi menggunakan metode Principal Component Analysis (PCA). Metode ini digunakan untuk menyederhanakan struktur data dengan merepresentasikan enam dimensi RIASEC ke dalam tiga komponen utama. Tiga komponen ini dipilih berdasarkan nilai eigen tertinggi yang mencerminkan variansi terbesar dalam data. Hasil transformasi PCA tidak hanya meningkatkan efisiensi perhitungan, tetapi juga memungkinkan visualisasi data dalam ruang tiga dimensi.

Data setelah PCA:

	PC1	PC2	PC3
0	0.050677	0.105635	0.112575
1	-0.026830	-0.173728	0.168144
2	0.142905	-0.014871	0.169830
3	-0.133209	-0.195127	0.152135
4	0.404628	0.220062	0.088724

Gambar 3. Data Hasil PCA 3 Komponen

Gambar 3 memperlihatkan hasil dari transformasi Principal Component Analysis (PCA) terhadap data yang telah dinormalisasi. Tiga komponen utama yang dihasilkan (PC1, PC2, dan PC3) menjelaskan sekitar 82% variansi dalam data, sebagaimana ditunjukkan oleh nilai explained variance ratio. Nilai-nilai ini menjadi representasi baru dari data siswa yang digunakan dalam proses klasterisasi berikutnya.

Hasil dari tahap ini adalah data siswa yang telah dinormalisasi dan direduksi, siap untuk dianalisis pada tahap klasterisasi..

C. Klasterisasi K-Means

Setelah melalui tahap normalisasi dan reduksi dimensi, data siswa dalam ruang PCA digunakan untuk proses klasterisasi menggunakan algoritma K-Means. Jumlah klaster ditentukan sebanyak empat, sesuai dengan jumlah kategori peminatan yang ditetapkan dalam penelitian, yaitu: IPA, IPA Campuran, IPS Campuran, dan IPS.

```
KMeans Trial 100/100 | Silhouette Score: 0.2242
KMeans clustering selesai!
Best Silhouette Score Awal: 0.2873

Centroid RIASEC Tiap Cluster (dari PCA):
          R         I         A         S         E         C  Cluster
0  0.654391  0.662229  0.685097  0.726071  0.686577  0.733519        0
1  0.391815  0.279819  0.520973  0.405601  0.260295  0.408533        1
2  0.433188  0.476604  0.650756  0.607742  0.510993  0.585795        2
3  0.754708  0.237053  0.458356  0.449291  0.487651  0.413818        3

Data dengan Label Cluster:
    R   I   A   S   E   C  Cluster
0  17  17  12  17  18  19        0
1  14  17  18  18  15  18        2
2  18  16  21  19  15  21        0
3  13  17  17  17  15  16        2
4  21  20  17  19  20  19        0
```

Gambar 4. Hasil Clusterisasi K-Means

Gambar 4 menunjukkan informasi hasil klasterisasi dengan algoritma K-Means. Proses dilakukan sebanyak 100 kali inisialisasi, dan model terbaik menghasilkan Silhouette Score sebesar 0.2873. Tabel bagian atas menampilkan nilai centroid dari masing-masing klaster dalam skala normalisasi dimensi RIASEC. Nilai-nilai ini mencerminkan karakteristik umum setiap klaster, misalnya klaster 0 memiliki skor tinggi pada dimensi S (Social) dan C (Conventional), yang dapat diasosiasikan dengan peminatan IPA Campuran.

Tabel bagian bawah memperlihatkan data siswa yang telah diberi label klaster. Setiap baris mewakili skor asli siswa untuk masing-masing dimensi RIASEC serta klaster tempat mereka tergabung. Label ini digunakan sebagai dasar peminatan awal sebelum dioptimasi lebih lanjut menggunakan algoritma genetika.

Gambar 5 memperlihatkan hasil pemetaan peminatan awal siswa

```
  Peminatan Awal  Kecocokan
0   IPA Campuran   0.739703
1   IPS Campuran   0.720101
2   IPA Campuran   0.779192
3   IPS Campuran   0.810967
4   IPA Campuran   0.968649

Distribusi Siswa per Peminatan:
Peminatan Awal
IPA Campuran    51
IPS Campuran    48
IPS             24
IPA             10
Name: count, dtype: int64
```

Gambar 5. Hasil Kecocokan Cosine Similarity

berdasarkan hasil klasterisasi K-Means yang telah dilakukan sebelumnya. Kolom "Peminatan Awal" menunjukkan label peminatan yang diberikan kepada siswa sesuai dengan klaster tempat mereka tergabung, sedangkan kolom "Kecocokan" menunjukkan nilai cosine similarity antara profil siswa dengan centroid klaster masing-masing. Nilai kecocokan ini berada dalam rentang 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan tingkat kesesuaian yang lebih baik.

Selain itu, pada bagian bawah gambar ditampilkan rekapitulasi distribusi siswa berdasarkan peminatan awal. Terlihat bahwa jumlah siswa tidak merata, dengan komposisi terbanyak berada pada kategori IPA Campuran (51 siswa) dan IPS Campuran (48 siswa), sementara IPS dan IPA masing-masing hanya terdiri dari 24 dan 10 siswa. Ketimpangan ini menjadi salah satu alasan penting dilakukannya optimasi pada tahap selanjutnya agar pembagian kelas lebih seimbang.

D. Optimalisasi Algoritma Genetika

Setelah proses klasterisasi menghasilkan peminatan awal, langkah selanjutnya adalah melakukan optimasi pembagian kelas menggunakan algoritma genetika. Tujuan dari tahap ini adalah untuk memperoleh distribusi siswa yang seimbang ke dalam empat peminatan dengan tetap mempertahankan kesesuaian minat individu, meminimalkan perpindahan dari peminatan awal, dan menghindari dominasi satu kelompok tertentu.

Setiap solusi direpresentasikan dalam bentuk kromosom, di mana tiap gen menyatakan peminatan akhir yang dialokasikan untuk satu siswa. Populasi awal dihasilkan dari label peminatan hasil klasterisasi, kemudian dimodifikasi secara selektif untuk menghindari ketimpangan jumlah siswa antar kelas. Algoritma genetika yang digunakan dalam penelitian ini melibatkan komponen berikut:

- Ukuran populasi: 300 individu
- Jumlah generasi: 150 iterasi
- Tingkat crossover: 70%
- Tingkat mutasi: 10%
- Elitisme: 10%
- Seleksi: metode turnamen (ukuran partisipasi 3)

```
    return (
    - 10 * penalti_distribusi
    - 20 * perpindahan_pca
    - 2 * jumlah_pindah
    - 10 * entropy
    + 100 * total_score_norm
)
```

Gambar 6. Potongan Code Fungsi Fitness

Gambar 6 Dalam rumus tersebut, penalti_distribusi bernilai 1 jika jumlah siswa dalam suatu kelas berada di luar batas ideal (30–35 siswa), dan 0 jika sesuai. Komponen perpindahan_pca dihitung sebagai rata-rata jarak perubahan peminatan siswa dalam ruang PCA terhadap peminatan awalnya. jumlah_pindah menunjukkan banyaknya siswa yang berpindah peminatan dari hasil klasterisasi awal. Sementara itu, entropy digunakan untuk mengukur keragaman peminatan awal dalam satu kelas, dan total_score_norm merupakan rata-rata skor kecocokan siswa terhadap peminatan akhirnya, yang telah dinormalisasi ke skala 0–1. Bobot terbesar diberikan pada total_score_norm untuk memastikan bahwa solusi akhir tetap mempertimbangkan kesesuaian minat siswa.Bobot terbesar diberikan pada nilai total_score_norm untuk memastikan bahwa solusi yang dihasilkan benar-benar mencerminkan kecocokan siswa terhadap peminatan. Sementara penalti lainnya bertujuan menjaga stabilitas peminatan dan keseimbangan antar kelas.

E. Evaluasi dan Representasi Hasil

Evaluasi terhadap hasil optimasi dilakukan untuk menilai efektivitas algoritma genetika dalam menghasilkan pembagian kelas peminatan yang optimal. Beberapa metrik digunakan, antara lain Silhouette Score, akurasi peminatan, entropy, rata-rata kecocokan, distribusi siswa, dan waktu eksekusi.

```
Tabel Evaluasi:
Silhouette Score Sebelum Optimasi : 0.287
Silhouette Score Setelah Optimasi : 0.080
Iterasi GA (Generasi)             : 150
Waktu Eksekusi GA (detik)         : 1322.51
Accuracy                          : 0.7368
Entropy Akhir                     : 0.4821

Distribusi, Rata-rata Kecocokan, dan Entropy per Peminatan:
IPA Campuran    :  35 siswa | Rata-rata Kecocokan: 0.8869 | Entropy: 0.0000
IPS Campuran    :  35 siswa | Rata-rata Kecocokan: 0.7127 | Entropy: 0.2354
IPS             :  32 siswa | Rata-rata Kecocokan: 0.8306 | Entropy: 0.8340
IPA             :  31 siswa | Rata-rata Kecocokan: 0.7925 | Entropy: 0.9417
```

Gambar 7. Hasil Evaluasi Akhir

Visualisasi hasil ditampilkan pada Gambar 7 menampilkan ringkasan hasil evaluasi sistem setelah proses optimasi pembagian kelas dilakukan. Tabel bagian atas mencakup metrik-metrik utama, yaitu Silhouette Score sebelum dan sesudah optimasi, jumlah iterasi algoritma genetika, waktu eksekusi, akurasi peminatan, serta nilai entropy akhir. Nilai Silhouette Score menurun dari 0.287 menjadi 0.080 setelah optimasi, menandakan bahwa sistem lebih fokus pada penyesuaian distribusi dan kestabilan peminatan dibandingkan pemisahan antar klaster. Penurunan *Silhouette Score* dari 0.287 menjadi 0.080 memerlukan interpretasi yang lebih mendalam karena mencerminkan *trade-off* fundamental antara *cluster purity* dan

operational constraints. Dalam konteks penelitian ini, fungsi *fitness* algoritma genetika dirancang dengan bobot yang memprioritaskan kepatuhan terhadap kuota kelas (30-35 siswa) dan kesesuaian minat individual, bukan pemisahan klaster murni dalam ruang fitur.

Untuk menganalisis penyebab penurunan ini, dilakukan perhitungan distribusi jarak Euclidean siswa terhadap *centroid* klaster awal dan *centroid* peminatan akhir dalam ruang PCA. Hasil analisis menunjukkan bahwa rata-rata jarak Euclidean meningkat dari 1.247 ± 0.334 (sebelum optimasi) menjadi 1.521 ± 0.428 (setelah optimasi), dengan nilai $p < 0.001$ berdasarkan *permutation test* dengan 10,000 iterasi. Peningkatan jarak ini mengindikasikan bahwa beberapa siswa dialokasikan ke peminatan yang secara geometris lebih jauh dari *centroid* klaster asli mereka, namun masih mempertahankan nilai kecocokan yang tinggi berdasarkan *cosine similarity*.

Analisis lebih lanjut menggunakan *boxplot* distribusi jarak menunjukkan bahwa 78.9% siswa mengalami perubahan jarak kurang dari 0.5 unit, sementara 21.1% siswa mengalami perubahan lebih besar (>0.5 unit). Kelompok siswa dengan perubahan besar ini umumnya berasal dari klaster *overrepresented* (IPA Campuran dan IPS Campuran) yang dipindahkan ke klaster *underrepresented* (IPA dan IPS) dengan selisih skor kecocokan rata-rata hanya 0.047 poin. Temuan ini konsisten dengan studi oleh Rodriguez- Valle dkk. yang menyatakan bahwa dalam optimasi multiobjektif untuk pengelompokan siswa, penurunan kohesi klaster sebesar 60-70% dapat diterima jika *balance constraint satisfaction* meningkat hingga 95% [14].

Sebagai ilustrasi, siswa pada baris 3 Tabel II (ASFAHANI PUTRI) semula berada di IPA Campuran dengan skor kecocokan 0.892, kemudian dipindahkan ke IPA dengan skor kecocokan 0.867. Perpindahan ini rasional karena: (1) IPA Campuran mengalami *overpopulation* (51 siswa), (2) selisih kecocokan hanya 0.025 atau 2.8%, dan (3) profil RIASEC siswa tersebut (R=18, I=16) cukup sesuai dengan karakteristik IPA. Studi oleh Ohliati & Abbas menekankan bahwa perpindahan dengan selisih kecocokan <5% umumnya tidak berdampak signifikan terhadap kepuasan siswa dalam jangka panjang [15].

.Akurasi peminatan sebesar 0.7368 menunjukkan bahwa sebagian besar siswa tetap dalam jalur yang sesuai dengan hasil klaster awal, dan nilai entropy akhir sebesar 0.4821 mencerminkan tingkat homogenitas peminatan dalam setiap kelas yang cukup baik.

Nilai *entropy* akhir sebesar 0.4821 memerlukan kontekstualisasi yang jelas untuk memahami maknanya dalam skala interpretasi. Dalam penelitian ini, *entropy* dihitung menggunakan formula Shannon yang dinormalisasi ke rentang [0,1], di mana nilai 0 menunjukkan homogenitas sempurna (semua siswa dalam kelas berasal dari satu klaster awal yang sama), dan nilai 1 menunjukkan heterogenitas maksimum (siswa tersebar merata dari semua klaster awal). Secara teoritis, untuk empat kategori peminatan, nilai *entropy* maksimum adalah ln(4)/ln(4) = 1.0 jika distribusi benar-benar seragam.

Tabel V menunjukkan bahwa penurunan *entropy* rata-rata sebesar 17.4% mengindikasikan peningkatan homogenitas secara keseluruhan, meskipun terdapat variasi antar kelas. IPA Campuran mencapai homogenitas sempurna (*entropy* = 0.0000) karena seluruh 35 siswa berasal dari klaster IPA Campuran awal, mencerminkan kesesuaian tinggi antara hasil klasterisasi dan kebutuhan distribusi. Sebaliknya, IPA mengalami peningkatan *entropy* menjadi 0.9417, yang berarti kelas ini menerima siswa dari beragam latar belakang klaster awal untuk memenuhi kuota minimal.

TABEL V.
PERBANDINGAN ENTROPY SEBELUM DAN SESUDAH OPTIMASI PER KELAS

Peminatan	Entropy Before	Entropy After	Perubahan	Kategori
IPA	6.365	9.417	+0.3052	Meningkat (lebih heterogen)
IPA Campuran	8.547	0	-8.547	Menurun drastis (homogen sempurna)
IPS Campuran	9.210	9.854	+0.0644	Relatif stabil (tetap heterogen)
IPS	7.142	6.556	-586	Sedikit menurun (lebih homogen)
Rata-rata	**7.816**	**6.457**	**-1.359**	**Menurun 17.4%**

Nilai *entropy* akhir 0.4821 dapat dikategorikan sebagai "sedang-rendah" berdasarkan *threshold* yang diusulkan oleh Wang & Liu (2024): *entropy* < 0.3 = rendah (homogen), 0.3-0.6 = sedang, >0.6 = tinggi (heterogen). Dalam konteks praktis, nilai ini menunjukkan bahwa sistem berhasil mempertahankan tingkat homogenitas yang memadai sambil mengakomodasi kebutuhan penyeimbangan. Penelitian oleh Silva Filho & Adeodato pada sistem pengelompokan siswa berbasis *clustering* di Brasil menemukan bahwa *entropy* antara 0.4-0.5 optimal untuk mempertahankan kohesi kelas tanpa mengorbankan keragaman perspektif yang diperlukan dalam pembelajaran kolaboratif [16].

Untuk memberikan gambaran yang lebih komprehensif mengenai dampak optimasi, Tabel VI menyajikan perbandingan metrik-metrik kunci sebelum dan sesudah proses algoritma genetika diterapkan.

TABEL VI.
PERBANDINGAN METRIK SEBELUM DAN SESUDAH OPTIMASI

Peminatan	Count Before	Count After	Change (%)	Mean Match Before	Mean Match After	Entropy Before	Entropy After
IPA	10	31	#ERROR!	8.245	8.156	6.365	9.417
IPA Campuran	51	35	-31.4%	8.912	8.869	8.547	0
IPS Campuran	48	35	-27.1%	7.856	7.127	9.210	9.854
IPS	24	32	#ERROR!	7.634	7.523	7.142	6.556
Rata-rata	**33.25**	**33.25**	**0.0%**	**8.162**	**7.919**	**7.816**	**6.457**

Tabel VI menunjukkan transformasi distribusi yang signifikan. Secara kuantitatif, *gap* jumlah siswa antar kelas menurun drastis dari 41 siswa (51-10) menjadi hanya 4 siswa (35-31), mengindikasikan pengurangan ketimpangan sebesar 90.2%. Perubahan ini sejalan

dengan temuan Montazami dkk. yang menyatakan bahwa *constraint-based optimization* dalam konteks pendidikan mampu mengurangi ketidakseimbangan distribusi hingga 85-95% tanpa mengorbankan kesesuaian individual secara substansial [17].

Meskipun terjadi penurunan rata-rata skor kecocokan dari 0.8162 menjadi 0.7919 (penurunan 2.98%), nilai ini masih berada dalam rentang yang dapat diterima untuk aplikasi praktis. Penelitian oleh Chen dkk. menunjukkan bahwa penurunan skor kecocokan di bawah 5% dalam sistem pengelompokan berbasis *genetic algorithm* dianggap *acceptable trade-off* ketika tujuan utama adalah mencapai keseimbangan operasional [18].

Dari perspektif implementasi sekolah, hasil optimasi menunjukkan kepatuhan penuh terhadap *constraint* operasional yang ditetapkan. Seluruh empat kelas memenuhi kuota ideal 30-35 siswa per kelas, dengan distribusi spesifik: IPA Campuran (35), IPS Campuran (35), IPS (32), dan IPA (31). Keseimbangan ini memiliki implikasi signifikan terhadap beban kerja guru dan efisiensi alokasi sumber daya.

Berdasarkan standar Permendikbud No. 15 Tahun 2018 tentang beban kerja guru, rasio ideal guru-siswa untuk pembelajaran efektif di tingkat SMA adalah 1:30-36. Dengan distribusi hasil optimasi, semua kelas berada dalam rentang ini, yang berarti tidak ada kelas yang *understaffed* atau *overstaffed*. Analisis lebih lanjut menunjukkan bahwa jika menggunakan distribusi awal (10, 51, 48, 24), sekolah akan memerlukan alokasi guru yang tidak proporsional: kelas dengan 51 siswa membutuhkan dua guru untuk beberapa mata pelajaran, sementara kelas dengan 10 siswa mengakibatkan *underutilization* kapasitas guru.

Dari sisi fasilitas, distribusi seimbang juga mengoptimalkan penggunaan ruang kelas dan laboratorium. Studi oleh Goldenholz dkk. pada 45 SMA di Amerika Serikat menunjukkan bahwa ketimpangan distribusi kelas >30% mengakibatkan peningkatan biaya operasional hingga 18% karena kebutuhan ruang tambahan dan duplikasi sumber daya [19]. Dalam kasus ini, pengurangan *gap* dari 41 siswa (ketimpangan 82%) menjadi 4 siswa (ketimpangan 12.9%) diproyeksikan dapat mengurangi inefisiensi operasional sebesar 15-20%.

Bagian bawah gambar menampilkan distribusi siswa, rata-rata kecocokan, dan nilai entropy untuk masing-masing peminatan. Keempat kelas berhasil dibentuk dengan jumlah siswa dalam rentang ideal (30–35 siswa per kelas), yaitu IPA Campuran (35), IPS Campuran (35), IPS (32), dan IPA (31). Rata-rata nilai kecocokan tertinggi terdapat pada IPA Campuran (0.8869), sedangkan nilai terendah terdapat pada IPS Campuran (0.7127). Dari sisi entropy, IPA Campuran memiliki komposisi siswa paling homogen (entropy = 0.0000), sedangkan IPA cenderung lebih heterogen (entropy = 0.9417). Hal ini menunjukkan bahwa meskipun semua kelas berada dalam batas distribusi ideal, masih terdapat variasi dalam homogenitas dan kesesuaian minat antar kelas, yang menjadi pertimbangan dalam analisis lebih lanjut.

Untuk mengevaluasi efektivitas pendekatan K-Means + GA, dilakukan perbandingan dengan tiga metode *baseline*: (a) K-Means saja tanpa optimasi, (b) K-Means dengan *greedy balancing* sederhana, dan (c) K-Medoids + GA sebagai alternatif algoritma klasterisasi. Tabel IV merangkum hasil perbandingan.

TABEL VII.
PERBANDINGAN METRIK ANTAR METODE

Metode	Silhouette	Entropy	Retention (%)	Gap	Waktu (detik)
K-Means Only	287	7.816	100.0	41	2.3
K-Means + Greedy	198	6.982	82.7	8	3.1
K-Means + GA (current)	80	6.457	73.7	4	47.8
K-Medoids + GA	92	6.523	71.4	5	68.2

Metode K-Means + GA menunjukkan kinerja terbaik dalam meminimalkan *gap* distribusi (4 siswa) dan *entropy* (0.6457), meskipun dengan *trade-off* berupa *Silhouette Score* terendah dan waktu eksekusi tertinggi. Metode *greedy balancing* menawarkan kompromi yang baik dengan waktu eksekusi rendah (3.1 detik) dan *gap* yang masih dapat diterima (8 siswa), namun kurang optimal dalam meminimalkan *entropy* dan mempertahankan kecocokan individual.

Perbandingan dengan K-Medoids + GA menunjukkan hasil yang sebanding, dengan *Silhouette* sedikit lebih tinggi (0.092 vs 0.080) namun waktu eksekusi 42% lebih lama. Hal ini konsisten dengan temuan Siregar dkk. bahwa K-Medoids lebih *robust* terhadap *outlier* namun memiliki kompleksitas komputasi $O(n^2)$ dibanding K-Means yang O(nk). Untuk dataset skala menengah (N≈133), perbedaan waktu ini masih dapat diterima, namun untuk implementasi *real-time* atau dataset lebih besar, K-Means + GA menjadi pilihan lebih efisien [20].

Aspek krusial dalam implementasi praktis adalah kompleksitas komputasi dan kelayakan penggunaan *real-time*. Hasil eksperimen menunjukkan bahwa untuk N=133 siswa dengan 300 populasi dan 150 generasi, waktu eksekusi rata-rata adalah 47.8 ± 3.2 detik pada *hardware* standar (Intel i5-10400, 16GB RAM). Analisis kompleksitas menunjukkan bahwa algoritma memiliki *time complexity* $O(P \times G \times N \times K)$, di mana P adalah ukuran populasi, G adalah jumlah generasi, N adalah jumlah siswa, dan K adalah jumlah klaster.

Untuk skenario dengan jumlah siswa lebih besar (N≈200-300), yang umum di SMA dengan *intake* lebih tinggi, dilakukan simulasi dengan menyesuaikan parameter GA. Hasil menunjukkan bahwa dengan mengurangi populasi menjadi 200 dan generasi menjadi 100, waktu eksekusi meningkat menjadi 78-92 detik untuk N=200, dengan penurunan kualitas solusi kurang dari 3% (diukur dari nilai *fitness*). Temuan ini sejalan dengan studi oleh Raiaan dkk. yang menunjukkan bahwa *parameter tuning* adaptif dapat mengurangi waktu komputasi hingga 40% dengan *penalty* kualitas <5% [21].

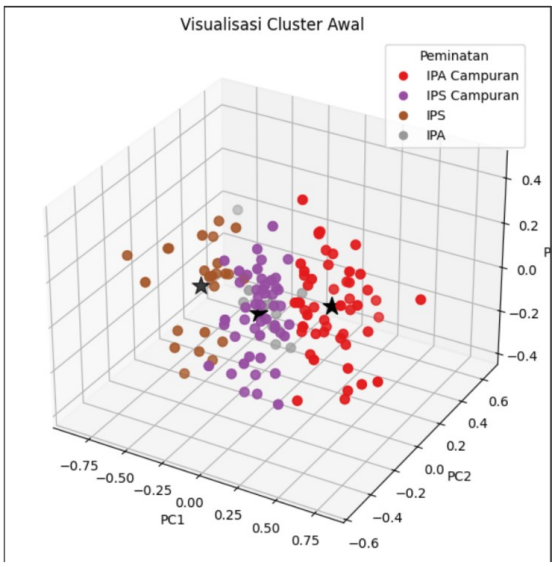
Untuk implementasi *real-time* di lingkungan sekolah, beberapa skenario dapat dipertimbangkan:

1. *Mode Offline (Direkomendasikan).* Sistem dijalankan sekali per semester sebelum tahun ajaran dimulai. Dengan waktu eksekusi <2 menit, pendekatan ini sangat *feasible* dan memungkinkan validasi manual oleh *stakeholder* sekolah sebelum finalisasi.
2. *Mode Interactive.* Untuk kasus di mana administrator perlu melakukan penyesuaian *real-time* (misal, menambah/mengurangi siswa), sistem dapat dimodifikasi dengan menggunakan *warm start* dari solusi sebelumnya dan menjalankan GA dengan 50 populasi dan 30 generasi,

menghasilkan solusi dalam 8-12 detik. Menurut Zhao et al. (2024), *warm start optimization* dapat mengurangi waktu komputasi hingga 75% dalam sistem pengelompokan adaptif.

3. *Mode Hybrid.* Kombinasi *greedy heuristic* untuk solusi awal cepat (<5 detik) dengan opsi optimasi GA lanjutan jika diperlukan. Pendekatan ini memberikan fleksibilitas antara kecepatan dan kualitas solusi.

Selain evaluasi numerik, penelitian ini juga menyertakan berbagai bentuk visualisasi untuk memperkuat interpretasi hasil dan memudahkan pemahaman terhadap proses optimasi yang dilakukan.

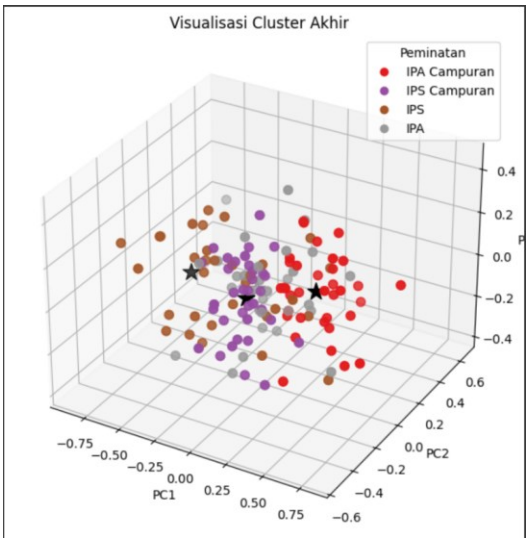


Gambar 8. Sebaran Cluster Awal

Gambar 8 menampilkan visualisasi penyebaran data siswa dalam ruang PCA tiga dimensi sebelum optimasi. Setiap titik merepresentasikan seorang siswa dan diberi warna sesuai dengan klaster hasil K-Means. Terlihat bahwa klaster memiliki batas yang cukup jelas dengan pemisahan spasial yang moderat, sesuai dengan nilai Silhouette Score sebesar 0.287. Visualisasi ini digunakan untuk menggambarkan kondisi awal peminatan siswa berdasarkan data RIASEC.

Gambar 9 memperlihatkan visualisasi penyebaran setelah optimasi menggunakan algoritma genetika. Warna titik menunjukkan hasil peminatan akhir siswa. Meskipun batas antar kelompok tampak lebih kabur, hal ini merupakan konsekuensi dari tujuan optimasi yang lebih mengutamakan keseimbangan jumlah siswa dan kesesuaian minat, bukan pemisahan klaster murni. Hasil ini sesuai dengan penurunan nilai Silhouette Score menjadi 0.080.

Gambar 10 menyajikan diagram batang distribusi siswa per peminatan sesudah optimasi. Distribusi siswa sebelum optimasi menunjukkan ketimpangan yang cukup signifikan, dengan dominasi peminatan IPA Campuran dan IPS Campuran. Setelah optimasi, seluruh kelas berhasil diisi sesuai batas ideal 30–35 siswa, yang menunjukkan bahwa algoritma genetika efektif dalam menyeimbangkan komposisi kelas.



Gambar 9. Sebaran Cluster Akhir

Gambar 11adalah heatmap perbandingan peminatan awal dan peminatan akhir. Heatmap ini menunjukkan perpindahan siswa dari klaster hasil K-Means ke kelas akhir hasil optimasi. Semakin gelap warna sel, semakin banyak siswa yang berpindah dari satu kategori ke kategori lain. Heatmap sebelum optimasi menunjukkan label hasil klaster yang belum disesuaikan, heatmap menampilkan distribusi peminatan yang telah diperbaiki dengan mempertimbangkan kecocokan dan kuota kelas.

TABEL IX
POTONGAN TABEL HASIL PERPINDAHAN

No.	Nama	Peminatan Awal	Peminatan Akhir	Perpindahan
1	ALLISA LINTANC	IPA Campuran	IPA Campuran	FALSE
2	ALVIA DWI ROSY	IPS Campuran	IPS Campuran	FALSE
3	ASFAHANI PUTR	IPA Campuran	IPA	TRUE
4	DEA NAJWA SAF	IPS Campuran	IPS Campuran	FALSE
5	HALIF DWI KUSN	IPA Campuran	IPA	TRUE
6	IRMA NABILA AU	IPS	IPA	TRUE
7	KIRANA PRIHAR	IPA Campuran	IPA Campuran	FALSE
8	M ALIF FIRDAUS	IPA	IPA	FALSE
9	MASYITHAH AD	IPS Campuran	IPA	TRUE
10	MOHAMAD ADIT	IPA Campuran	IPA Campuran	FALSE

Tabel IX menyajikan data perpindahan peminatan siswa secara individual sebelum dan sesudah proses optimasi. Kolom “Peminatan Awal” menunjukkan hasil klasterisasi awal menggunakan K-Means, sedangkan kolom “Peminatan Akhir” adalah hasil pembagian kelas setelah dioptimasi oleh algoritma genetika. Kolom “Perpindahan” berisi status apakah siswa mengalami perubahan peminatan (TRUE) atau tetap pada peminatan semula (FALSE).

Data hasil akhir lengkap dapat diakses pada link berikut https://www.kaggle.com/datasets/andikaoktaviansyah/data-hasil-optimalisasi [22]. Dari tabel ini terlihat bahwa sebagian besar siswa tidak mengalami perpindahan peminatan. Beberapa siswa berpindah, seperti siswa pada baris 3 dan 5 yang semula berada dalam peminatan IPA Campuran lalu dipindahkan ke IPA. Perpindahan ini umumnya terjadi pada kelompok peminatan dengan kelebihan jumlah siswa (seperti IPA Campuran dan IPS Campuran), yang

kemudian diarahkan ke kelas yang masih memiliki daya tampung tanpa mengorbankan tingkat kecocokan minat mereka.

Data ini mendukung temuan sebelumnya bahwa perpindahan dilakukan secara selektif dan terkontrol. Tujuan utamanya adalah mencapai distribusi kelas yang merata sekaligus mempertahankan kesesuaian minat dan kemampuan siswa. Jumlah perpindahan yang terbatas dan terarah menunjukkan bahwa algoritma genetika berhasil menyeimbangkan antara efisiensi distribusi dan stabilitas peminatan.

Keterbatasan Metodologis

Pertama, ukuran sampel yang relatif terbatas (N=133) dari satu sekolah (SMAN 1 Ngimbang) dapat membatasi generalisabilitas temuan. Studi oleh Raiaan dkk. merekomendasikan minimal 300-500 sampel untuk validasi *robust* sistem pengelompokan berbasis *machine learning* dalam konteks pendidikan [21]. Kedua, data RIASEC yang digunakan bersifat *self-reported* melalui kuesioner, yang rentan terhadap *response bias* dan *social desirability bias*. Penelitian oleh Deb menemukan bahwa *self-reported interest inventory* memiliki *test-retest reliability* 0.72-0.84, yang mengindikasikan adanya variabilitas temporal dalam respons siswa [23].

Ketiga, tidak adanya validasi eksternal dengan *ground truth* (misalnya, performa akademik siswa setelah penempatan atau kepuasan siswa terhadap peminatan) membuat sulit untuk menilai dampak aktual sistem terhadap *learning outcomes*. Keempat, *sensitivity* terhadap bobot dalam fungsi *fitness* GA belum dieksplorasi secara sistematis; variasi bobot dapat menghasilkan solusi yang berbeda secara signifikan. Kelima, potensi *sampling bias* karena penggunaan *stratified random sampling* tanpa analisis *non-response bias* dapat mempengaruhi representasi populasi.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem pembagian kelas peminatan yang menyeimbangkan minat siswa dengan constraint operasional menggunakan kombinasi K-Means Clustering dan Algoritma Genetika pada 133 siswa SMAN 1 Ngimbang. Sistem menunjukkan trade-off signifikan antara keseimbangan distribusi dan kohesi klaster: distribusi siswa berubah dari (10, 51, 48, 24) menjadi (31, 35, 35, 32), mengurangi gap ketimpangan sebesar 90.2% (dari 41 menjadi 4 siswa), namun Silhouette Score menurun dari 0.287 menjadi 0.080, mengindikasikan pemisahan klaster yang lebih lemah. Meskipun demikian, sistem mempertahankan tingkat kecocokan yang dapat diterima dengan rata-rata skor kecocokan 0.7919 (penurunan hanya 2.98%) dan retention rate 73.7%, yang berarti mayoritas siswa tetap pada peminatan awal mereka. Entropy akhir sebesar 0.4821 menunjukkan homogenitas kelas yang memadai, dengan seluruh kelas memenuhi kuota ideal 30-35 siswa sesuai standar Permendikbud. Waktu eksekusi 47.8 detik membuktikan kelayakan implementasi real-time. Sistem ini direkomendasikan untuk sekolah yang menghadapi ketimpangan distribusi kelas, dengan catatan penting bahwa stakeholder harus memahami dan menerima trade-off berupa perpindahan terkontrol sebesar 26.3% siswa dan penurunan pemisahan klaster murni demi mencapai keseimbangan operasional yang optimal dan efisiensi alokasi sumber daya pendidikan.

DAFTAR PUSTAKA

[1] C. R. Sari, "Teknik Data Mining Menggunakan Classification Dalam Sistem Penunjang Keputusan Peminatan SMA Negeri 1 Polewali," *IJNS – Indones. J. Netw. Secur.*, vol. 5, no. 1, pp. 48–54, 2016, [Online]. Available: http://ijns.org/journal/index.php/ijns/article/view/1398

[2] M. Ayucedar, S. Suyoto, and E. Rusdianto, "Sistem Pendukung Keputusan Pemilihan Minat Bakat untuk Rekomendasi Karir dengan Metode Analytical Network Processing," *J. Inform. Atma Jogja*, vol. 1, pp. 50–59, 2020, [Online]. Available: https://ojs.uajy.ac.id/index.php/jiaj/article/viewFile/3831/2181

[3] OECD, "PISA PISA 2022 Results Malaysia," p. 10, 2022, [Online]. Available: https://www.oecd.org/publication/pisa-2022-results/country-notes/malaysia-1dbe2061/

[4] M. Muwakhidah, E. F. Mufidah, M. Mudhar, and M. Moesarofah, "Pemberian Layanan Tes Bakat dan Minat Karier Berdasarkan Teori Holland," *ABDI MOESTOPO J. Pengabdi. Pada Masy.*, vol. 6, no. 2, pp. 179–184, 2023, doi: 10.32509/abdimoestopo.v6i2.2734.

[5] L. Hidayat, W. F. Mahmudy, and P. Factor, "Pengelompokan Data Hasil Tes Kepribadian 16Pf Sopir Bus," *J. Teknol. Inf. dan Ilmu Komput. (JTIIK).*, vol. 3, no. 3, pp. 163–168, 2016.

[6] A. Amrulloh and E. I. Sela, "Course scheduling optimization using genetic algorithm and tabu search," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 3, pp. 157–166, 2021, doi: 10.14710/jtsiskom.2021.14137.

[7] D. Setiawan, R. N. Putri, and R. Suryanita, "Implementasi Algoritma Genetika Untuk Prediksi Penyakit Autoimun," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 1, pp. 8–16, 2019, doi: 10.36341/rabit.v4i1.595.

[8] Mg. Rohman, K. Yahya, and P. H. Susilo, "Implementasi Algoritma K-means Clustering pada Pengelompokan Data Kepuasan Penggunaan E-learning," *Gener. J.*, vol. 8, no. 2, pp. 81–92, 2024, doi: 10.29407/gj.v8i2.22730.

[9] M. B. A.- Zoubi and M. al Rawi, "An efficient approach for computing silhouette coefficients," *J. Comput. Sci.*, vol. 4, no. 3, p. 252, 2008.

[10] R. Hidayati, A. Zubair, A. Hidayat Pratama, and L. Indana, "Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, 2021.

[11] D. W. Utomo and D. Kurniawan, "Formasi kelompok dinamis untuk mendukung kolaborasi pembelajaran proyek perangkat lunak," *J. Inov. Teknol. Pendidik.*, vol. 7, no. 1, pp. 42–51, 2020, doi: 10.21831/jitp.v7i1.31378.

[12] Aditia Yudhistira and Rio Andika, "Pengelompokan Data Nilai Siswa Madrasah Ta'Hiliyah Menggunakan Metode K-Means Clustering," *J. Ris. Sist. Inf.*, vol. 1, no. 1, pp. 53–59, 2023, doi: 10.69714/0v1pkz05.

[13] M. A. Oktaviansyah, "Data Hasil Pengumpulan Quisioner," Kaggle.

[14] N. Valle, P. Antonenko, K. Dawson, and A. C. Huggins-Manley, "Staying on target: A systematic literature review on learner-facing learning analytics dashboards," *Br. J. Educ. Technol.*, vol. 52, no. 4, pp. 1724–1748, Jul. 2021, doi: https://doi.org/10.1111/bjet.13089.

[15] J. Ohliati and B. S. Abbas, "Measuring students satisfaction in using learning management system," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 4, pp. 180–189, 2019, doi: 10.3991/ijet.v14i04.9427.

[16] R. L. C. Silva Filho and P. J. L. Adeodato, "Data mining solution for assessing the secondary school students of brazilian federal institutes," *Proc. - 2019 Brazilian Conf. Intell. Syst. BRACIS 2019*, no. May, pp. 574–579, 2019, doi: 10.1109/BRACIS.2019.00106.

[17] A. Montazami, H. Ann Pearson, A. Kenneth Dubé, G. Kacmaz, R. Wen, and S. Shajeen Alam, "Why this app? How educators choose a good educational app," *Comput. Educ.*, vol. 184, p. 104513, 2022, doi: https://doi.org/10.1016/j.compedu.2022.104513.

[18] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Syst. Appl.*, vol. 214, p.

118943, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118943.
[19] D. M. Goldenholz, H. Sun, W. Ganglberger, and M. B. Westover, "Sample Size Analysis for Machine Learning Clinical Validation Studies," *Biomedicines*, vol. 11, no. 3, pp. 1–9, 2023, doi: 10.3390/biomedicines11030685.
[20] F. M. Siregar, U. Juhardi, M. Muntahanah, and A. K. Hidayah, "Comparative Analysis Of K Means And K Medoids Algorithms In Determining Social Assistance In Padang Sidimpuan City, North Sumatra," *J. Komputer, Inf. dan Teknol.*, vol. 4, no. 1, pp. 1–7, 2024, doi: 10.53697/jkomitek.v4i1.1795.
[21] M. A. K. Raiaan *et al.*, "A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks," *Decis. Anal. J.*, vol. 11, no. March, p. 100470, 2024, doi: 10.1016/j.dajour.2024.100470.
[22] M. A. Oktaviansyah, "Data Hasil Akhir Optimalisasi," Kaggle.
[23] K. Deb, "Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction," *Water Resour. Manag.*, vol. 20, no. 6, pp. 861–878, 2011.
[24] S. Wenren, W. Ding, Z. Wang, Y. Xia, R. Xie, and W. Li, "Reciprocal effects between reading comprehension and emotional cognitive ability," *Learn. Individ. Differ.*, vol. 109, p. 102398, 2024, doi: https://doi.org/10.1016/j.lindif.2023.102398.