E-ISSN: 3025-0994
Halaman 5295-5310
Volume 4 Nomor 4 Tahun 2026

Pengelompokan Siswa Berdasarkan Faktor Keberhasilan Belajar Dengan K-Means di SMP Negeri 3 Kota Tambolaka

Wasti Lende[1], Antonius Mbay Ndapamuri[2], Mitra Permata Ayu[3]

[1,2,3] Program Studi Teknik Informatika, Universitas Stella Maris Sumba
e-mail: wastilende182@gmail.com

Abstrak

Keberhasilan belajar siswa berkaitan dengan faktor akademik dan non-akademik, antara lain pendidikan orang tua, kondisi ekonomi keluarga, jarak tempat tinggal ke sekolah, dan capaian akademik. Penelitian ini bertujuan menerapkan algoritma K-Means Clustering untuk mengelompokkan siswa SMP Negeri 3 Kota Tambolaka berdasarkan empat atribut tersebut serta menjelaskan karakteristik kelompok yang terbentuk. Penelitian menggunakan dataset 100 siswa tahun ajaran 2025/2026. Empat atribut numerik distandardisasi menggunakan Z-score sebelum proses K-Means dengan inisialisasi k-means++ (random_state=42, n_init=20). Atribut yang dianalisis meliputi kode pendidikan orang tua, kode penghasilan bulanan orang tua, jarak rumah ke sekolah, dan nilai rata-rata rapor. Hasil dataset menunjukkan tiga kelompok, yaitu cluster tinggi sebanyak 36 siswa (36%), cluster sedang 33 siswa (33%), dan cluster rendah 31 siswa (31%). Rata-rata nilai rapor masing-masing cluster adalah 87,77; 78,44; dan 69,49. Rata-rata jarak rumah ke sekolah pada cluster tinggi, sedang, dan rendah berturut-turut adalah 2,08 km; 4,86 km; dan 7,80 km. Evaluasi jumlah cluster untuk K=2 sampai K=6 menunjukkan bahwa K=3 memiliki SSE/inertia 155,6413 dan Silhouette Coefficient 0,2940, sedangkan nilai silhouette tertinggi terdapat pada K=2 sebesar 0,3948. K=3 digunakan dalam dataset untuk memenuhi kebutuhan pemetaan tiga profil siswa. Hasil ini dapat menjadi informasi awal bagi sekolah dalam merancang pendampingan akademik, dengan tetap memperhatikan bahwa clustering menunjukkan kemiripan data dan bukan hubungan sebab-akibat.

Kata Kunci: K-Means Clustering, Data Mining, Keberhasilan Belajar, Profil Siswa, SMP Negeri 3 Kota Tambolaka.

Abstract

Student learning success is associated with academic and non-academic factors, including parents' education, family income, distance from home to school, and academic achievement. This study applies K-Means Clustering to group 100 students of SMP Negeri 3 Kota Tambolaka in the 2025/2026 academic year based on four attributes: coded parental education, coded monthly parental income, home-to-school distance, and average report-card scores. The dataset reports three groups: 36 students (36%) in the high cluster, 33 (33%) in the medium

cluster, and 31 (31%) in the low cluster. The respective mean report-card scores are 87.77, 78.44, and 69.49, while the mean distances to school are 2.08 km, 4.86 km, and 7.80 km. Evaluation for K values from 2 to 6 reports an SSE/inertia of 155.6413 and a Silhouette Coefficient of 0.2940 for K=3; the highest silhouette value is obtained at K=2 (0.3948). K=3 is retained in the dataset to support a three-profile grouping for school support planning. The findings can inform initial academic-support planning, while clustering should be interpreted as a measure of data similarity rather than evidence of causation.

Keywords : K-Means Clustering, Data Mining, Learning Success, Student Profile, SMP Negeri 3 Kota Tambolaka.

LATAR BELAKANG

Pendidikan merupakan salah satu unsur penting dalam pengembangan sumber daya manusia. Keberhasilan pembelajaran tidak hanya berkaitan dengan kemampuan akademik siswa, tetapi juga dipengaruhi oleh berbagai faktor internal dan eksternal. Faktor eksternal antara lain tingkat pendidikan orang tua, kondisi ekonomi keluarga, lingkungan tempat tinggal, dan aksesibilitas menuju sekolah. Perbedaan karakteristik tersebut dapat memberikan gambaran mengenai kondisi siswa dalam mengikuti proses pembelajaran. Oleh karena itu, identifikasi karakteristik siswa diperlukan sebagai salah satu upaya mendukung peningkatan kualitas pendidikan.

Kondisi sosial-ekonomi keluarga menjadi salah satu aspek yang relevan dalam memahami keberhasilan belajar siswa. Tingkat pendidikan orang tua dapat menggambarkan latar belakang pendidikan keluarga, sedangkan penghasilan orang tua berkaitan dengan kemampuan ekonomi dalam memenuhi kebutuhan pendidikan anak. Selain itu, jarak rumah ke sekolah juga menjadi karakteristik yang dapat menggambarkan aksesibilitas siswa terhadap lingkungan pendidikan. Sentosa (2023) menunjukkan bahwa pendidikan dan penghasilan orang tua memiliki hubungan dengan capaian akademik siswa. Meskipun demikian, faktor-faktor tersebut tidak dapat digunakan secara tunggal untuk menentukan keberhasilan belajar seseorang.

Perkembangan teknologi informasi memberikan peluang untuk memanfaatkan data pendidikan sebagai sumber informasi dalam pengambilan keputusan. Salah satu pendekatan yang dapat digunakan adalah Educational Data Mining, yaitu pemanfaatan teknik data mining untuk menemukan pola dan informasi dari data pendidikan. Chaka (2022) menjelaskan bahwa data mining dapat dimanfaatkan untuk menganalisis dan memprediksi performa akademik siswa. Dengan demikian, data administrasi sekolah dapat dikembangkan menjadi informasi yang lebih bermanfaat melalui proses pengolahan dan analisis data.

Salah satu teknik data mining yang dapat digunakan adalah clustering. Teknik ini mengelompokkan data berdasarkan kemiripan karakteristik tanpa menggunakan label kelas sebelumnya. Algoritma K-Means merupakan salah satu metode clustering yang bekerja dengan menentukan centroid, menghitung jarak setiap data terhadap pusat cluster, dan mengelompokkan data berdasarkan jarak terdekat. Penelitian Alawi et al. (2023) serta

Badhera et al. (2022) menunjukkan bahwa K-Means dapat diterapkan untuk mengelompokkan siswa berdasarkan karakteristik dan performa akademik.

SMP Negeri 3 Kota Tambolaka memiliki siswa dengan latar belakang keluarga dan capaian akademik yang beragam. Namun, data administrasi siswa belum sepenuhnya dimanfaatkan untuk menghasilkan pemetaan karakteristik berdasarkan kombinasi faktor keluarga, aksesibilitas, dan capaian akademik. Penelitian ini menawarkan pendekatan pengelompokan menggunakan empat atribut, yaitu tingkat pendidikan orang tua, penghasilan bulanan orang tua, jarak rumah ke sekolah, dan nilai rata-rata rapor. Kombinasi keempat atribut tersebut diharapkan dapat memberikan gambaran karakteristik siswa secara lebih menyeluruh dibandingkan pengelompokan yang hanya menggunakan nilai akademik.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan menerapkan algoritma K-Means Clustering untuk mengelompokkan siswa SMP Negeri 3 Kota Tambolaka berdasarkan empat atribut yang telah ditentukan, menganalisis karakteristik setiap cluster, serta mengevaluasi kualitas pengelompokan menggunakan Elbow Method dan Silhouette Coefficient. Hasil penelitian diharapkan dapat menjadi informasi awal bagi sekolah dalam merancang program pendampingan akademik dan dukungan pembelajaran sesuai karakteristik kelompok siswa.

Data Mining dan Clustering

Data mining merupakan proses menemukan pola, hubungan, atau informasi yang bermanfaat dari data melalui teknik statistik, matematika, kecerdasan buatan, dan machine learning. Clustering merupakan teknik pengelompokan tanpa label kelas awal. Objek yang memiliki karakteristik relatif serupa ditempatkan pada cluster yang sama, sedangkan objek yang berbeda cenderung ditempatkan pada cluster lain.

Algoritma K-Means

K-Means membagi data ke dalam sejumlah K kelompok. Tahap umumnya meliputi menentukan K, memilih centroid awal, menghitung jarak setiap objek terhadap centroid, menempatkan objek ke centroid terdekat, menghitung centroid baru berdasarkan rata-rata anggota, dan mengulangi proses hingga stabil. Salah satu ukuran jarak yang umum digunakan adalah Euclidean Distance: $d(x,c)=\sqrt{\Sigma(xj-cj)^2}$. Karena K-Means berbasis jarak, skala setiap atribut perlu diperhatikan. Atribut dengan rentang numerik besar dapat mendominasi hasil jika tidak dilakukan standardisasi atau pembobotan yang sesuai.

Faktor yang Digunakan dalam Penelitian

Pendidikan orang tua dan penghasilan bulanan digunakan sebagai indikator latar belakang keluarga. Jarak rumah ke sekolah digunakan untuk menggambarkan aspek aksesibilitas, sedangkan nilai rata-rata rapor digunakan sebagai indikator capaian akademik. Keempat variabel ini merupakan atribut pengelompokan yang tersedia dalam dataset; variabel tersebut tidak mencakup seluruh faktor yang mungkin memengaruhi keberhasilan belajar.

Evaluasi Cluster

Elbow Method mengamati penurunan SSE/inertia ketika jumlah cluster bertambah. Titik perubahan kemiringan kurva dapat dipertimbangkan sebagai jumlah cluster yang memadai. Silhouette Coefficient menilai kesesuaian objek dengan cluster sendiri dibandingkan cluster terdekat lainnya. Nilainya berada pada rentang -1 sampai 1; semakin

tinggi nilainya, secara umum semakin baik kohesi dan pemisahan cluster. Penentuan K perlu mempertimbangkan hasil evaluasi dan kebutuhan interpretasi.

METODE

Jenis dan Sumber Data

Penelitian menggunakan pendekatan data mining dengan teknik clustering K-Means. Data yang dianalisis berasal dari file dataset penelitian yang disertakan, berisi 100 record siswa berkode S001 sampai S100 untuk tahun ajaran 2025/2026. Nama siswa tidak ditampilkan dalam dataset yang dianalisis. Karena file yang tersedia berupa dataset olahan, uraian ini mendeskripsikan data yang diberikan dan tidak mengklaim telah melakukan pengumpulan lapangan secara mandiri.

Tabel 1. Variabel Penelitian

No.	Atribut	Representasi dalam dataset
1	Pendidikan orang tua	SD=1; SMP=2; SMA/SMK=3; Diploma/Sarjana=4.
2	Penghasilan orang tua	< Rp1.000.000=1; Rp1.000.000–2.499.999=2; Rp2.500.000–4.999.999=3; ≥ Rp5.000.000=4.
3	Jarak rumah ke sekolah	Jarak aktual dalam km; kategori ≤2 km=1; 2,01–5 km=2; 5,01–8 km=3; >8 km=4.
4	Nilai rata-rata rapor	Numerik kontinu dalam rentang 0–100.

Tahapan Analisis Data

Tahapan analisis dilakukan secara sistematis melalui preprocessing data, penerapan algoritma K-Means, dan evaluasi kualitas cluster. Preprocessing mencakup seleksi atribut, pembersihan data, transformasi atau pengkodean atribut, serta standardisasi data. Seluruh proses komputasi ulang dilakukan pada 100 record siswa yang memiliki data lengkap.

Preprocessing Data

Seleksi Atribut dan Representasi Data

Empat atribut digunakan sebagai masukan algoritma, yaitu kode pendidikan orang tua (1-4), kode penghasilan orang tua (1-4), jarak rumah ke sekolah dalam kilometer, dan nilai rata-rata rapor. ID siswa hanya berfungsi sebagai identitas record dan tidak dimasukkan ke perhitungan jarak. Kolom hasil cluster yang sudah tersedia juga tidak digunakan sebagai atribut masukan agar tidak terjadi kebocoran label (label leakage). Kode kategori jarak tidak dimasukkan sebagai atribut terpisah karena jarak aktual dalam kilometer telah digunakan sebagai variabel numerik.

Pembersihan Data (Data Cleaning)

Pemeriksaan dilakukan terhadap struktur kolom dan kelengkapan nilai pada empat atribut. Baris ringkasan “Rata-rata / Total” yang bukan merupakan individu siswa dikeluarkan dari proses pemodelan. Setelah pembersihan, data yang dianalisis berjumlah 100 record dan tidak

terdapat nilai kosong pada empat atribut masukan. Data identitas siswa tidak ditampilkan dalam pelaporan hasil.

Transformasi dan Pengkodean Data

Pendidikan dan penghasilan orang tua dikonversi ke kode ordinal sesuai pedoman transformasi pada dataset. Pendidikan orang tua direpresentasikan dengan kode 1–4, sedangkan penghasilan orang tua juga direpresentasikan dengan kode 1–4. Jarak rumah ke sekolah tetap menggunakan nilai aktual dalam kilometer dan nilai rata-rata rapor tetap menggunakan nilai numerik kontinu. Pengkodean ini dilakukan agar atribut kategorikal dapat direpresentasikan dalam bentuk numerik untuk proses analisis.

Normalisasi/Standardisasi Data dengan Z-score

Karena rentang nilai rapor, jarak dalam kilometer, serta kode pendidikan dan penghasilan memiliki skala yang berbeda, keempat atribut numerik distandardisasi menggunakan metode Z-score. Standardisasi dilakukan agar setiap atribut memiliki skala yang sebanding dan tidak ada atribut dengan rentang nilai besar yang mendominasi perhitungan jarak. Rumus yang digunakan adalah $z_{ij} = (x_{ij} - \mu_j) / \sigma_j$, dengan x_{ij} sebagai nilai atribut ke-j pada siswa ke-i, μ_j sebagai rata-rata atribut ke-j, dan σ_j sebagai standar deviasinya. Parameter rata-rata dan standar deviasi dihitung dari 100 record yang dianalisis.

Pengukuran Jarak Euclidean

Kedekatan setiap siswa terhadap centroid dihitung menggunakan Euclidean Distance pada empat atribut yang telah distandardisasi. Untuk siswa ke-i dan centroid ke-k, rumusnya adalah $d(x_i,c_k) = \sqrt{\sum_{j=1}^{4}} (z_{ij} - c_{kj})^2$. Siswa ditempatkan pada cluster dengan jarak terkecil. Dengan empat atribut, perhitungan jarak menjumlahkan kuadrat selisih pada kode pendidikan, kode penghasilan, jarak rumah, dan nilai rapor, kemudian mengambil akar kuadratnya.

Inisialisasi Centroid dan Parameter Algoritma

Pengujian dilakukan dengan implementasi K-Means menggunakan Python dan pustaka scikit-learn. Centroid awal diinisialisasi menggunakan strategi k-means++. Untuk menjaga keterulangan eksperimen, random_state ditetapkan sebesar 42, n_init sebesar 20 (20 inisialisasi), dan batas maksimum iterasi sebesar 300. Pengujian jumlah cluster dilakukan untuk K=2, 3, 4, 5, dan 6. Implementasi dan visualisasi tahapan analisis dilakukan menggunakan Google Colab.

Pembaruan Centroid dan Proses Iterasi

Setelah penempatan awal anggota cluster, centroid baru dihitung sebagai rata-rata setiap atribut dari seluruh anggota cluster. Rumus centroid atribut ke-j pada cluster k adalah $c_{kj} = (1/n_k) \Sigma_{i \in C_k} z_{ij}$, dengan n_k sebagai jumlah anggota cluster k. Proses penghitungan jarak, penempatan ulang anggota, dan pembaruan centroid diulang sampai keanggotaan cluster tidak berubah atau batas iterasi tercapai. Pada eksperimen K=3, proses berhenti setelah 12 iterasi pada eksekusi Google Colab dengan parameter yang dilaporkan.

Centroid akhir K=3 pada ruang Z-score (urutan atribut: kode pendidikan, kode penghasilan, jarak rumah, nilai rapor) adalah: cluster Sedang = (0,1146; −0,0341; 0,0323; −0,0660), cluster Rendah = (−1,0042; −0,9454; 1,0599; −1,0798), dan cluster Tinggi = (0,7597; 0,8454; −0,9423; 0,9903). Nilai centroid terstandar ini memperlihatkan posisi relatif rata-rata anggota tiap cluster terhadap rata-rata keseluruhan data.

Penamaan dan Interpretasi Cluster

Nomor cluster yang dihasilkan algoritma bersifat nominal, sehingga tidak secara otomatis menunjukkan kategori tinggi, sedang, atau rendah. Penamaan dilakukan setelah memeriksa rata-rata nilai rapor dan karakteristik atribut pada masing-masing kelompok. Cluster dengan rata-rata rapor tertinggi diberi label “Tinggi”, berikutnya “Sedang”, dan rata-rata terendah “Rendah”. Label tersebut merupakan ringkasan profil data, bukan diagnosis atau penilaian permanen terhadap individu siswa.

Keterulangan eksperimen: komputasi K-Means dijalankan kembali pada 100 record dengan empat atribut yang distandardisasi menggunakan Z-score. Hasil K=3 mereproduksi pembagian label pada workbook, yaitu 36 siswa pada profil Tinggi, 33 siswa pada profil Sedang, dan 31 siswa pada profil Rendah. Pengaturan yang digunakan adalah k-means++, random_state=42, n_init=20, dan max_iter=300. Dengan demikian, centroid, iterasi, serta metrik evaluasi dalam naskah ini merujuk pada proses komputasi yang terdokumentasi tersebut.

Evaluasi Hasil Clustering

Nilai SSE/inertia dan Silhouette Coefficient dihitung pada data terstandardisasi untuk setiap K dari 2 sampai 6. SSE mengukur jumlah kuadrat jarak setiap data ke centroid cluster-nya; nilai yang lebih kecil umumnya diperoleh ketika K bertambah. Silhouette Coefficient dihitung dari kedekatan rata-rata objek dengan cluster sendiri dan cluster terdekat lainnya. Pada eksperimen ini, K=2 memperoleh silhouette tertinggi (0,3948), sedangkan K=3 memperoleh 0,2940. K=3 tetap digunakan untuk analisis profil tiga kelompok sesuai tujuan pemetaan sekolah, tetapi hasil tersebut harus dipahami sebagai kompromi antara kebutuhan interpretasi dan kualitas pemisahan cluster, bukan sebagai K yang paling unggul menurut Silhouette.

HASIL DAN PEMBAHASAN

Gambaran Umum Dataset

Pengolahan dilakukan terhadap dataset asli yang berisi 100 record siswa. Tampilan Google Colab memperlihatkan delapan kolom pada dataset awal, antara lain ID_Siswa, Pendidikan_Orang_Tua, Kode_Pendidikan, Penghasilan_Orang_Tua, Kode_Penghasilan, jarak rumah ke sekolah, serta atribut lain yang tersedia pada lembar data. ID_Siswa dipertahankan sebagai identitas record dan tidak digunakan dalam perhitungan jarak K-Means.

Jumlah data: 100
Nama kolom: ['ID_Siswa', 'Pendidikan_Orang_Tua', 'Kode_Pendidikan', 'Penghasilan_Orang_Tua', 'Kode_Penghasilan', 'Jarak_Ru

	ID_Siswa	Pendidikan_Orang_Tua	Kode_Pendidikan	Penghasilan_Orang_Tua	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Ra
0	S099	SD	1	< Rp1.000.000	1	6.43	
1	S011	Diploma/Sarjana	4	Rp2.500.000–4.999.999	3	2.68	
2	S005	SMP	2	≥ Rp5.000.000	4	1.56	
3	S019	SMA/SMK	3	Rp2.500.000–4.999.999	3	4.33	
4	S089	SD	1	Rp1.000.000–2.499.999	2	7.57	
...	...	...	...	...	...	...	
95	S078	SMP	2	< Rp1.000.000	1	9.24	
96	S057	SMP	2	Rp1.000.000–2.499.999	2	3.60	
97	S027	SMA/SMK	3	≥ Rp5.000.000	4	0.37	
98	S007	SMA/SMK	3	Rp1.000.000–2.499.999	2	1.22	
99	S002	SMP	2	Rp2.500.000–4.999.999	3	2.96	

100 rows × 8 columns

Gambar 1. Tampilan dataset asli sebanyak 100 record siswa.

Seleksi Atribut Penelitian

Dari dataset awal dipilih empat atribut numerik untuk proses clustering, yaitu Kode_Pendidikan, Kode_Penghasilan, Jarak_Rumah_ke_Sekolah_km, dan Rata_rata_Rapor. Keempat atribut tersebut digunakan karena merepresentasikan kondisi pendidikan orang tua, penghasilan orang tua, jarak tempat tinggal ke sekolah, dan capaian akademik siswa. Kolom identitas dan atribut teks kategorikal tidak dimasukkan langsung ke dalam model; kode numerik yang telah tersedia digunakan sebagai representasi pendidikan dan penghasilan.

Atribut yang digunakan:

	Kode_Pendidikan	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Rata_rata_Rapor
0	1	1	6.43	63.88
1	4	3	2.68	93.23
2	2	4	1.56	92.05
3	3	3	4.33	85.12
4	1	2	7.57	69.98
5	3	4	0.30	90.89
6	4	3	2.38	89.78
7	2	1	5.10	65.52
8	1	1	7.19	59.01
9	4	2	0.77	81.12

Gambar 2. Empat atribut yang dipilih sebagai masukan algoritma K-Means.

Standardisasi Data dengan Z-Score

Sebelum clustering, keempat atribut distandardisasi menggunakan StandardScaler. Hasil standardisasi pada Google Colab menunjukkan 100 baris dan empat kolom.

Standardisasi mengubah skala setiap atribut agar perbedaan rentang nilai tidak menyebabkan atribut dengan skala lebih besar mendominasi pengukuran jarak. Nilai Z-score positif menunjukkan nilai di atas rata-rata atribut, sedangkan nilai negatif menunjukkan nilai di bawah rata-rata.

Hasil Standardisasi Z-Score:

	Kode_Pendidikan	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Rata_rata_Rapor
0	-1.565217	-1.275813	0.580762	-1.715084
1	1.695652	0.586688	-0.731359	1.608064
2	-0.478261	1.517938	-1.123246	1.474458
3	0.608696	0.586688	-0.154026	0.689811
4	-1.565217	-0.344563	0.979647	-1.024412
...	...	...	...	...
95	-0.478261	-1.275813	1.563978	-1.074231
96	-0.478261	-0.344563	-0.409452	1.084965
97	0.608696	1.517938	-1.539625	0.636595
98	0.608696	-0.344563	-1.242211	1.313679
99	-0.478261	0.586688	-0.633387	0.525635

100 rows × 4 columns

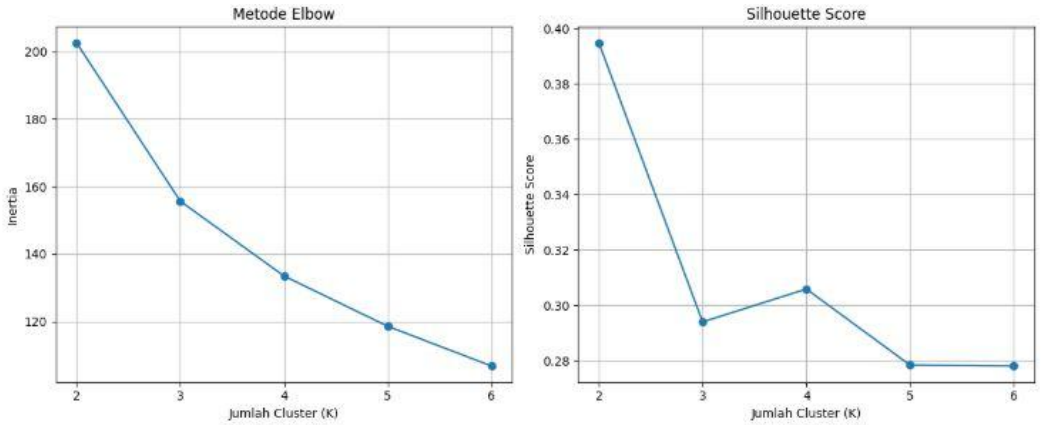
Gambar 3. Hasil standardisasi Z-Score pada 100 data dan empat atribut.

Evaluasi Jumlah Cluster

Evaluasi jumlah cluster dilakukan untuk menentukan jumlah kelompok (*K*) yang sesuai dalam proses pengelompokan data siswa menggunakan algoritma K-Means. Pengujian dilakukan terhadap nilai K=2 sampai dengan K=6 dengan menggunakan dua metode evaluasi, yaitu Elbow Method dan Silhouette Score.

```
Hasil Evaluasi K-Means
K=2 | Inertia=202.5625 | Silhouette=0.3948
K=3 | Inertia=155.6413 | Silhouette=0.2940
K=4 | Inertia=133.4883 | Silhouette=0.3059
K=5 | Inertia=118.5862 | Silhouette=0.2784
K=6 | Inertia=106.8232 | Silhouette=0.2781
```

Gambar 4. Keluaran angka evaluasi Inertia dan Silhouette untuk K=2 sampai K=6.



Gambar 5. Grafik Elbow dan Silhouette hasil evaluasi jumlah cluster.

Nilai inertia menurun ketika jumlah cluster bertambah, dari 202,5625 pada K=2 menjadi 106,8232 pada K=6. Grafik Elbow menunjukkan penurunan yang relatif tajam dari K=2 ke K=3 dan kemudian berlanjut lebih landai. Sementara itu, Silhouette tertinggi diperoleh pada K=2 sebesar 0,3948. Nilai Silhouette untuk K=3 adalah 0,2940, sedikit lebih rendah daripada K=4 sebesar 0,3059. Dengan demikian, K=3 bukan konfigurasi dengan nilai Silhouette tertinggi. Penelitian menggunakan K=3 untuk membentuk tiga profil interpretatif siswa, sehingga pemilihan tersebut didasarkan pada kebutuhan pemetaan tiga kategori dan perlu dipahami dengan mempertimbangkan kualitas pemisahan cluster yang masih terbatas.

Proses Pembentukan Cluster dan Centroid

Model K-Means dijalankan dengan K=3, inisialisasi k-means++, random_state=42, n_init=20, dan max_iter=300. Berdasarkan keluaran Google Colab, jumlah anggota label awal adalah cluster 0 sebanyak 31 siswa, cluster 1 sebanyak 33 siswa, dan cluster 2 sebanyak 36 siswa. Label angka tersebut merupakan penanda kelompok dari algoritma dan belum menunjukkan kategori tinggi, sedang, atau rendah.

```
Jumlah anggota setiap cluster:
Label_Cluster
0    31
1    33
2    36
Name: count, dtype: int64
```

	ID_Siswa	Pendidikan_Orang_Tua	Kode_Pendidikan	Penghasilan_Orang_Tua	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Ra
0	S099	SD	1	< Rp1.000.000	1	6.43	
1	S011	Diploma/Sarjana	4	Rp2.500.000–4.999.999	3	2.68	
2	S005	SMP	2	≥ Rp5.000.000	4	1.56	
3	S019	SMA/SMK	3	Rp2.500.000–4.999.999	3	4.33	
4	S089	SD	1	Rp1.000.000–2.499.999	2	7.57	
...	...	...	...	...	...	...	
95	S078	SMP	2	< Rp1.000.000	1	9.24	
96	S057	SMP	2	Rp1.000.000–2.499.999	2	3.60	
97	S027	SMA/SMK	3	≥ Rp5.000.000	4	0.37	
98	S007	SMA/SMK	3	Rp1.000.000–2.499.999	2	1.22	
99	S002	SMP	2	Rp2.500.000–4.999.999	3	2.96	

100 rows × 9 columns

Gambar 6. Jumlah anggota pada label cluster awal

Centroid akhir untuk ketiga cluster pada ruang Z-score ditampilkan pada Gambar 7. Urutan nilai pada setiap centroid mengikuti empat atribut masukan: kode pendidikan, kode penghasilan, jarak rumah, dan rata-rata rapor.

```
Centroid setiap cluster:
```

	Kode_Pendidikan	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Rata_rata_Rapor
0	-1.004208	-0.945369	1.059898	-1.079783
1	0.114625	-0.034146	0.032269	-0.066016
2	0.759662	0.845368	-0.942270	0.990328

Gambar 7. Centroid akhir K-Means untuk tiga cluster pada skala Z-score.

Penamaan Cluster dan Distribusi Kelompok

Penamaan cluster dilakukan setelah membandingkan rata-rata nilai rapor setiap kelompok. Cluster dengan rata-rata rapor tertinggi diberi nama Tinggi, kelompok dengan rata-rata berikutnya diberi nama Sedang, dan kelompok dengan rata-rata terendah diberi nama Rendah. Berdasarkan hasil tersebut, label 2 menjadi Tinggi, label 1 menjadi Sedang, dan label 0 menjadi Rendah.

Hasil pengelompokan siswa:

	ID_Siswa	Pendidikan_Orang_Tua	Kode_Pendidikan	Penghasilan_Orang_Tua	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Ra
0	S099	SD	1	< Rp1.000.000	1	6.43	
1	S011	Diploma/Sarjana	4	Rp2.500.000–4.999.999	3	2.68	
2	S005	SMP	2	≥ Rp5.000.000	4	1.56	
3	S019	SMA/SMK	3	Rp2.500.000–4.999.999	3	4.33	
4	S089	SD	1	Rp1.000.000–2.499.999	2	7.57	
...	...	...	...	...	...	...	
95	S078	SMP	2	< Rp1.000.000	1	9.24	
96	S057	SMP	2	Rp1.000.000–2.499.999	2	3.60	
97	S027	SMA/SMK	3	≥ Rp5.000.000	4	0.37	
98	S007	SMA/SMK	3	Rp1.000.000–2.499.999	2	1.22	
99	S002	SMP	2	Rp2.500.000–4.999.999	3	2.96	

100 rows × 10 columns

Gambar 8. Dataset setelah penambahan label hasil clustering.

Rata-rata rapor setiap cluster:
Label_Cluster
0 69.490968
1 78.444545
2 87.774167
Name: Rata_rata_Rapor, dtype: float64

	ID_Siswa	Pendidikan_Orang_Tua	Kode_Pendidikan	Penghasilan_Orang_Tua	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Ra
0	S099	SD	1	< Rp1.000.000	1	6.43	
1	S011	Diploma/Sarjana	4	Rp2.500.000–4.999.999	3	2.68	
2	S005	SMP	2	≥ Rp5.000.000	4	1.56	
3	S019	SMA/SMK	3	Rp2.500.000–4.999.999	3	4.33	
4	S089	SD	1	Rp1.000.000–2.499.999	2	7.57	
...	...	...	...	...	...	...	
95	S078	SMP	2	< Rp1.000.000	1	9.24	
96	S057	SMP	2	Rp1.000.000–2.499.999	2	3.60	
97	S027	SMA/SMK	3	≥ Rp5.000.000	4	0.37	
98	S007	SMA/SMK	3	Rp1.000.000–2.499.999	2	1.22	
99	S002	SMP	2	Rp2.500.000–4.999.999	3	2.96	

100 rows × 10 columns

Gambar 9. Rata-rata nilai rapor pada masing-masing label cluster.

Gambar 4.10. Rekapitulasi jumlah siswa berdasarkan kategori cluster.

Jumlah siswa setiap kelompok:

Hasil_Cluster	count
Tinggi	36
Sedang	33
Rendah	31

Gambar 10. Rekapitulasi jumlah siswa berdasarkan kategori cluster.



Gambar 11. Diagram batang distribusi siswa kategori Tinggi, Sedang, dan Rendah.

Profil Karakteristik Setiap Cluster

Profil rata-rata atribut pada Gambar 12. Cluster Tinggi memiliki rata-rata nilai rapor 87,774167 dan jarak rumah 2,077222 km. Cluster Sedang memiliki rata-rata nilai rapor 78,444545 dan jarak rumah 4,862424 km. Cluster Rendah memiliki rata-rata nilai rapor 69,490968 dan jarak rumah 7,799355 km. Rata-rata kode pendidikan dan penghasilan juga lebih tinggi pada kelompok Tinggi dibandingkan kelompok Sedang dan Rendah.

Profil Karakteristik Setiap Cluster:

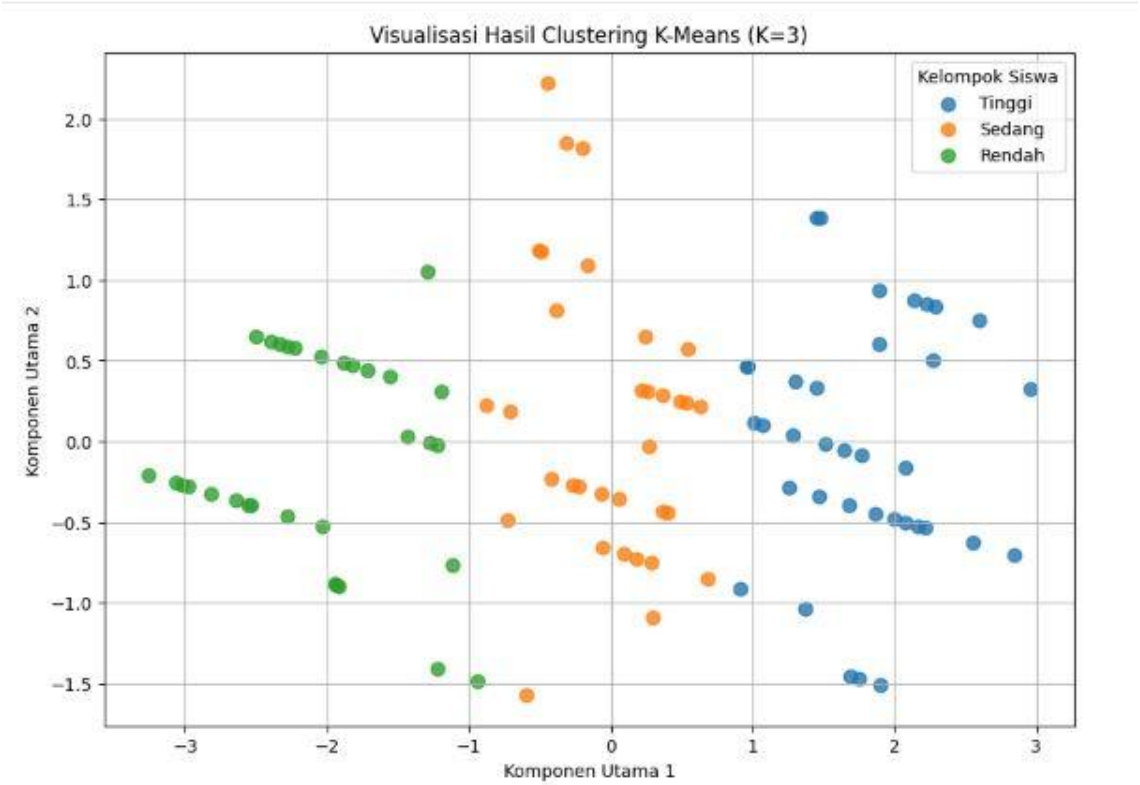
Hasil_Cluster	Kode_Pendidikan	Kode_Penghasilan	Jarak_Rumah_ke_Sekolah_km	Rata_rata_Rapor
Rendah	1.516129	1.354839	7.799355	69.490968
Sedang	2.545455	2.333333	4.862424	78.444545
Tinggi	3.138889	3.277778	2.077222	87.774167

Gambar 12. Profil rata-rata empat atribut pada masing-masing cluster

Secara deskriptif, kelompok Tinggi mempunyai rata-rata rapor tertinggi, kode pendidikan dan penghasilan lebih tinggi, serta jarak rumah rata-rata lebih dekat. Kelompok Rendah menunjukkan pola sebaliknya, sedangkan kelompok Sedang berada di antara keduanya. Kode pendidikan dan penghasilan merupakan kode kategori ordinal, sehingga rata-ratanya ditafsirkan sebagai kecenderungan kode, bukan sebagai ukuran interval yang memiliki jarak makna sama.

Visualisasi Sebaran Cluster dengan PCA

Untuk membantu melihat pola sebaran tiga kelompok, empat atribut terstandardisasi diproyeksikan ke dua komponen utama menggunakan Principal Component Analysis (PCA). Hasilnya ditampilkan pada Gambar 4.13. Setiap titik merepresentasikan satu record siswa dan warna menunjukkan kategori cluster. PCA hanya digunakan untuk visualisasi dua dimensi; proses pembentukan cluster tetap menggunakan empat atribut terstandardisasi.



Gambar 4.13. Visualisasi sebaran hasil clustering K-Means (K=3) menggunakan proyeksi PCA

PEMBAHASAN

Hasil analisis menunjukkan bahwa algoritma K-Means dapat mengelompokkan 100 record siswa ke dalam tiga profil berdasarkan kode pendidikan orang tua, kode penghasilan orang tua, jarak rumah ke sekolah, dan rata-rata rapor. Distribusi anggota relatif berimbang, yaitu 36 siswa pada kelompok Tinggi, 33 siswa pada kelompok Sedang, dan 31 siswa pada kelompok Rendah. Perbedaan rata-rata atribut menggambarkan variasi karakteristik antar kelompok pada dataset yang dianalisis.

Temuan deskriptif memperlihatkan bahwa kelompok dengan rata-rata rapor lebih tinggi juga memiliki rata-rata jarak rumah lebih dekat serta rata-rata kode pendidikan dan penghasilan orang tua lebih tinggi. Namun, hasil clustering tidak membuktikan bahwa faktor-faktor tersebut menyebabkan perbedaan nilai rapor. Analisis ini bersifat eksploratif dan terbatas pada empat atribut yang tersedia.

Evaluasi Silhouette menunjukkan nilai 0,2940 pada K=3, sedangkan K=2 memperoleh nilai 0,3948. Hal ini mengindikasikan bahwa pemisahan tiga cluster belum sekuat pemisahan dua cluster menurut metrik Silhouette. Oleh karena itu, hasil K=3 lebih tepat digunakan sebagai pemetaan awal tiga profil yang mudah ditafsirkan, bukan sebagai pembagian yang secara metrik paling optimal. Sekolah dapat menggunakan informasi ini sebagai bahan telaah awal dan tetap perlu mempertimbangkan faktor lain serta penilaian profesional guru sebelum menentukan bentuk dukungan bagi siswa.

SIMPULAN

Berdasarkan dataset 100 siswa SMP Negeri 3 Kota Tambolaka tahun ajaran 2025/2026, hasil pengelompokan yang tercantum menghasilkan tiga cluster, yaitu cluster tinggi sebanyak 36 siswa (36%), cluster sedang 33 siswa (33%), dan cluster rendah 31 siswa (31%). Rata-rata nilai rapor masing-masing kelompok adalah 87,77; 78,44; dan 69,49. Rata-rata jarak rumah ke sekolah masing-masing kelompok adalah 2,08 km; 4,86 km; dan 7,80 km. Karakteristik cluster tinggi didominasi orang tua berpendidikan SMA/SMK dan Diploma/Sarjana, dengan penghasilan relatif lebih tinggi serta jarak rumah yang cenderung dekat. Cluster sedang didominasi pendidikan SMP dan SMA/SMK dengan jarak rumah umumnya sedang. Cluster rendah didominasi pendidikan SD dan SMP, penghasilan di bawah Rp1.000.000, serta jarak rumah yang cenderung jauh atau sangat jauh. Hasil evaluasi komputasi ulang menunjukkan K=3 memiliki SSE/inertia 155,6413 dan Silhouette Coefficient 0,2940. Silhouette tertinggi diperoleh pada K=2 sebesar 0,3948. Dengan demikian, pemilihan K=3 merupakan keputusan berdasarkan kebutuhan interpretasi tiga profil siswa, bukan karena unggul pada metrik Silhouette. Nilai Silhouette K=3 juga mengindikasikan bahwa pemisahan antarkelompok masih terbatas. Hasil clustering dapat digunakan sebagai informasi awal untuk merancang dukungan belajar yang kemudian diverifikasi melalui asesmen sekolah. Berdasarkan dataset 100 siswa SMP Negeri 3 Kota Tambolaka tahun ajaran 2025/2026, hasil pengelompokan yang tercantum menghasilkan tiga cluster, yaitu cluster tinggi sebanyak 36 siswa (36%), cluster sedang 33 siswa (33%), dan cluster rendah 31 siswa (31%). Rata-rata nilai rapor masing-masing kelompok adalah 87,77; 78,44; dan 69,49. Rata-rata jarak rumah ke sekolah masing-masing kelompok adalah 2,08 km; 4,86 km; dan 7,80 km. Karakteristik cluster tinggi didominasi orang tua berpendidikan SMA/SMK dan Diploma/Sarjana, dengan penghasilan relatif lebih tinggi serta jarak rumah yang cenderung dekat. Cluster sedang didominasi pendidikan SMP dan SMA/SMK dengan jarak rumah umumnya sedang. Cluster rendah didominasi pendidikan SD dan SMP, penghasilan di bawah Rp1.000.000, serta jarak rumah yang cenderung jauh atau sangat jauh. Hasil evaluasi komputasi ulang menunjukkan K=3 memiliki SSE/inertia 155,6413 dan Silhouette Coefficient 0,2940. Silhouette tertinggi diperoleh pada K=2 sebesar 0,3948. Dengan demikian, pemilihan K=3 merupakan keputusan berdasarkan kebutuhan interpretasi tiga profil siswa, bukan karena unggul pada metrik Silhouette. Nilai Silhouette K=3 juga mengindikasikan bahwa pemisahan antarkelompok masih terbatas. Hasil clustering dapat digunakan sebagai informasi awal untuk merancang dukungan belajar yang kemudian diverifikasi melalui asesmen sekolah.

DAFTAR PUSTAKA

Alawi, S. J. S., Mohd Shaharanee, I. N., & Mohd Jamil, J. (2023). Clustering Student Performance Data Using K-Means Algorithms. Journal Of Computational Innovation And Analytics, 2(1), 41–55.

Badhera, U., Verma, A., & Nahar, P. (2022). Applicability Of K-Medoids And K-Means Algorithms For Segmenting Students Based On Their Scholastic Performance. Journal Of Statistics And Management Systems, 25(7), 1621–1632.

Bin Roslan, M. H., & Chen, C. J. (2022). Educational Data Mining For Student Performance Prediction: A Systematic Literature Review (2015–2021). International Journal Of Emerging Technologies In Learning, 17(5), 147–179.

Chaka, C. (2022). Educational Data Mining, Student Academic Performance Prediction, Prediction Methods, Algorithms And Tools: An Overview Of Reviews. Journal Of E-Learning And Knowledge Society, 18(2), 58–69.

Chen, X., Chen, Y., Yu, X., Wei, J., & Yang, X. (2024). The Impact Of Family Socioeconomic Status On Parental Involvement And Student Engagement During COVID-19 In Promoting Academic Achievement: A Longitudinal Study In Chinese Children. Journal Of Experimental Child Psychology, 246, 105992.

Ersozlu, Z., Taheri, S., & Koch, I. (2024). A Review Of Machine Learning Methods Used For Educational Data. Education And Information Technologies, 29, 22125–22145.

Gu, X., Che Hassan, N., & Sulaiman, T. (2024). The Relationship Between Family Factors And Academic Achievement Of Junior High School Students In Rural China: Mediation Effect Of Parental Involvement. Behavioral Sciences, 14(3), 221.

Hilyawan, F. R., Syafrullah, M., & Sudija, I. (2025). A Systematic Literature Review On The Optimization Of K-Means And Agglomerative Clustering For Student Performance Segmentation: A Comparative Analysis Of Elbow And Silhouette Methods. Edcomtech: Jurnal Kajian Teknologi Pendidikan, 10(1), 73–88.

Hutagalung, J., Syahputra, Y. H., & Tanjung, Z. P. (2022). Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi), 9(1), 606–620.

López-Zambrano, J., Lara Torralbo, J. A., & Romero, C. (2021). Early Prediction Of Student Learning Performance Through Data Mining: A Systematic Review. Psicothema, 33(3), 456–465.

Al-Din, M. S. N., & Al Abdulqader, H. A. (2024). Students' Academic Performance Prediction Using Educational Data Mining And Machine Learning: A Systematic Review. International Journal Of Research And Innovation In Social Science.

Muttaqin, S., Chuang, H. H., Et Al. (2022). When Proficiency And Education Matter: The Mediating Role Of English Proficiency And Moderating Effect Of Parents' Education In The SES–Academic Achievement Relationship During EMI. SAGE Open.

Nja, C. O., Neji, H. A., Orim, R. E., Ukwetang, J. O., Ideba, M. A., Cornelius-Ukpepi, B., & Ndifon, R. A. (2022). The Socio-Economic Rank Of Parents And Students' Academic And Cognitive Outcomes: Examining The Physical, Psychological And Social Mediators. Frontiers In Education, 7, 938078.

Qusyairi, M., Hidayatullah, Z., & Sandi, A. (2024). Penerapan K-Means Clustering Dalam Pengelompokan Prestasi Siswa Dengan Optimasi Metode Elbow. Infotek: Jurnal Informatika Dan Teknologi, 7(2), 500–510.

Saputra, E. A., & Nataliani, Y. (2021). Analisis Pengelompokan Data Nilai Siswa Untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means. Journal Of Information Systems And Informatics, 3(3), 424–439.

Sentosa, B. M. (2023). Does Parental Income And Education Improve Academic Achievement? A Case Of Public High School Students At Tangerang Municipality. Jurnal Pendidikan Progresif, 13(2), 387–403.

Tan, C. Y. (2024). Socioeconomic Status And Student Learning: Insights From An Umbrella Review. Educational Psychology Review, 36, 100.

Tsou, T. R. (2024). Family Background, Schooling, And Academic Achievement: Taiwan As An Illustrative Case. SAGE Open.

Tundo, Raihanah, S., Wahyudi, T., & Sugiyono. (2024). K-Means Clustering Of Social Studies Performance At Junior High School. IJID (International Journal On Informatics For Development), 13(2), 460–472.