



PENGELOMPOKAN SISWA SMP 1 KEMBANG BERDASARKAN NILAI AKADEMIK MENGGUNAKAN METODE K-MEANS CLUSTERING

Myllikha Putri Izza¹, Tasa², Vita Nurus Salamah³, Andhika Ikhwal Fadhil Ferdian⁴, Muhammad Arifin⁵

¹²³⁴⁵, Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muria Kudus Jl. Lkr. Utara, Kayuapu Kulon, Gondangmanis, Kec. Bae, Kabupaten Kudus, Jawa Tengah 59327

¹202353124@std.umk.ac.id, ²202353131@std.umk.ac.id, ³202353118@std.umk.ac.id, ⁴202353034@std.umk.ac.id, ⁵arifin.m@umk.ac.id

Abstract

Student grouping based on academic ability is an important step in supporting a more accurate and data-driven learning decision-making process. Differences in students' levels of understanding can lead to disparities in learning outcomes, therefore a method capable of grouping students systematically and accurately is required. This study was conducted to group students based on academic scores using the K-Means Clustering method. The data used consisted of secondary data from SMP Negeri 1 Kembang students' scores during the 2023–2025 period, covering 11 subjects. The research stages included data preprocessing, missing value handling using mean imputation, and determining the optimal number of clusters using the Elbow Method, which resulted in a K value of 3. Furthermore, the K-Means algorithm was used to categorize students into three levels: low, medium, and high. The results showed that out of 1,011 students, 233 students (23.05%) were in the low category with an average score of 51.49, 404 students (39.96%) were in the medium category with an average score of 65.28, and 374 students (36.99%) were in the high category with an average score of 75.76. Evaluation using the Davies Bouldin Index produced a value of 1.12567, indicating that the clustering results had good quality. These findings demonstrate that the K-Means Clustering method is capable of categorizing students based on academic ability accurately and can be used as a basis for decision-making in the school environment. However, this method still has limitations in determining the number of clusters and does not yet consider non-academic factors.

Keywords : *K-Means Clustering, Student Grouping, Data Mining, Academic Performance, Elbow Method*

Abstrak

Pengelompokan siswa berdasarkan kemampuan akademik merupakan langkah penting dalam menunjang proses pengambilan keputusan pembelajaran yang lebih akurat dan berbasis data. Perbedaan tingkat pemahaman siswa dapat menyebabkan kesenjangan hasil belajar, sehingga diperlukan metode yang mampu mengelompokkan siswa secara sistematis dan akurat. Penelitian ini dilakukan untuk mengelompokkan siswa berdasarkan nilai akademik melalui metode K-Means Clustering. Data yang digunakan merupakan data sekunder nilai siswa SMP Negeri 1 Kembang periode 2023–2025 yang mencakup 11 mata pelajaran. Tahapan penelitian meliputi pra-pemrosesan data, penanganan missing value menggunakan mean imputation, dan penetapan jumlah cluster optimal menggunakan Elbow Method yang menghasilkan nilai K=3. Selanjutnya, algoritma K-Means digunakan untuk mengkategorikan siswa ke dalam tiga tingkat, yaitu rendah, sedang, dan tinggi. Hasil penelitian menunjukkan bahwa dari 1.011 siswa, terdapat 233 siswa (23,05%) berada dalam tingkat rendah dengan nilai rata-rata 51,49, 404 siswa (39,96%) berada dalam tingkat sedang dengan nilai rata-rata 65,28, dan 374 siswa (36,99%) pada kategori tinggi dengan nilai rata-rata 75,76. Evaluasi menggunakan Davies Bouldin Index (DBI) menunjukkan nilai 1,12567 yang menandakan bahwa hasil clustering memiliki kualitas yang baik. Hasil tersebut membuktikan bahwa metode K-Means Clustering mampu mengkategorikan siswa berdasarkan



kemampuan akademik secara akurat dan dapat digunakan sebagai dasar pengambilan keputusan di lingkungan sekolah. Namun, metode ini masih memiliki keterbatasan dalam penentuan jumlah cluster serta belum mempertimbangkan faktor non-akademik.

Kata kunci : *K-Means Clustering, Pengelompokan Siswa, Data Mining, Kemampuan Akademik, Elbow Method*

1. PENDAHULUAN

Upaya peningkatan mutu pendidikan dapat dilakukan melalui pengelompokan siswa berdasarkan pencapaian kognitif mereka, sehingga kebijakan pembelajaran yang diterapkan guru menjadi lebih tepat sasaran dan terarah. [1]. Setiap siswa memiliki kemampuan memahami materi dan gaya belajar yang berbeda-beda [2]. Perbedaan ini menjadi tantangan dalam proses pembelajaran, karena metode yang digunakan belum tentu mampu mengakomodasi seluruh kebutuhan siswa dalam satu kelas. Jika tidak dikelola dengan baik, kondisi tersebut dapat menimbulkan kesenjangan hasil belajar serta memengaruhi efektivitas pembelajaran secara keseluruhan [3].

Untuk mengatasi permasalahan tersebut, diperlukan sistem klasifikasi prestasi siswa agar strategi pembelajaran dapat disesuaikan dengan tingkat kemampuan masing-masing kelompok. Selama ini, pengelompokan siswa umumnya masih dilakukan secara manual berdasarkan pengamatan guru atau nilai rata-rata sederhana. Pendekatan ini memiliki keterbatasan dalam hal objektivitas, akurasi, serta efisiensi, karena sangat bergantung pada subjektivitas penilaian dan kurang optimal dalam mengolah data dalam jumlah besar. Akibatnya, keputusan akademik yang dihasilkan berpotensi kurang tepat.

Seiring dengan perkembangan teknologi informasi, pemanfaatan *data mining* menjadi solusi dalam mengolah data pendidikan secara lebih efektif. *Data mining* merupakan proses untuk mengekstraksi informasi atau pola tersembunyi dari data berukuran besar dengan memanfaatkan teknik statistik, matematika, dan kecerdasan buatan [4]. Dalam bidang pendidikan, teknologi ini memungkinkan analisis data dilakukan secara lebih cepat, akurat, dan sistematis dibandingkan metode konvensional.

Salah satu teknik dalam *data mining* adalah *clustering*, yaitu metode pengelompokan data berdasarkan kesamaan karakteristik tanpa memerlukan label awal. Algoritma yang banyak digunakan adalah K-Means, yang bekerja dengan

mengelompokkan data ke dalam beberapa cluster berdasarkan kedekatan terhadap pusat cluster (centroid) [4] [5] [6]. Proses ini dilakukan secara iteratif hingga posisi centroid stabil [7]. K-Means memiliki keunggulan dalam kesederhanaan algoritma, kemudahan implementasi, serta efisiensi dalam mengolah data dalam jumlah besar.

Dalam konteks pendidikan, metode K-Means dapat digunakan untuk mengelompokkan siswa secara objektif berdasarkan berbagai komponen nilai, seperti tugas, ujian tengah semester, dan ujian akhir semester, sehingga menghasilkan analisis yang lebih komprehensif terhadap kemampuan akademik siswa [8]. Hasil pengelompokan ini tidak hanya membantu guru dalam memahami karakteristik siswa, tetapi juga dapat menjadi dasar bagi pihak sekolah dalam merancang program peningkatan kualitas pembelajaran secara lebih terarah dan berbasis data [9].

Meskipun demikian, K-Means Clustering memiliki beberapa keterbatasan, seperti sensitivitas terhadap inisialisasi awal serta penentuan jumlah cluster yang optimal [10]. Oleh karena itu, diperlukan analisis yang tepat dalam menentukan parameter agar hasil pengelompokan dapat merepresentasikan kondisi sebenarnya. Selain itu, kualitas hasil clustering juga sangat bergantung pada kelengkapan dan akurasi data yang digunakan.

Namun demikian, penelitian-penelitian sebelumnya umumnya hanya berfokus pada penerapan metode K-Means untuk pengelompokan data tanpa mengkaji secara mendalam bagaimana hasil clustering tersebut dapat digunakan secara praktis dalam mendukung pengambilan keputusan pembelajaran di sekolah. Selain itu, masih terbatas penelitian yang mengintegrasikan hasil pengelompokan siswa ke dalam kategori yang mudah dipahami dan langsung dapat diterapkan oleh guru, seperti kategori tinggi, sedang, dan rendah. Oleh karena itu, diperlukan penelitian yang tidak hanya menghasilkan cluster, tetapi



juga menginterpretasikan hasilnya secara aplikatif dalam konteks peningkatan kualitas pembelajaran.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menerapkan metode K-Means Clustering dalam mengelompokkan siswa berdasarkan nilai akademik, menganalisis hasil pengelompokan, serta mengevaluasi efektivitas metode tersebut dalam mendukung pengambilan keputusan di bidang pendidikan. Hasil pengelompokan akan diklasifikasikan ke dalam tiga kategori, yaitu tinggi, sedang, dan rendah, sehingga dapat menghasilkan model yang lebih objektif, sistematis, dan mudah diterapkan dalam lingkungan sekolah.

2. TINJAUAN PUSTAKA

Penelitian mengenai pengelompokan data dalam bidang pendidikan mengalami perkembangan yang cukup pesat seiring dengan meningkatnya pemanfaatan teknologi informasi, khususnya dalam bidang *data mining*. *Data mining* digunakan untuk mengekstraksi pola atau informasi tersembunyi dari kumpulan data berukuran besar dengan memanfaatkan teknik statistik, matematika, serta kecerdasan buatan. Salah satu teknik yang banyak digunakan dalam *data mining* adalah *clustering*, yaitu metode pengelompokan data berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label awal. Dalam konteks pendidikan, teknik ini banyak dimanfaatkan untuk mengelompokkan siswa berdasarkan nilai akademik guna mendukung proses pengambilan keputusan yang lebih objektif dan berbasis data. Selain itu, hasil pengelompokan juga dapat membantu guru dalam memahami perbedaan kemampuan siswa serta merancang strategi pembelajaran yang lebih efektif dan sesuai dengan kebutuhan masing-masing siswa.

Beberapa penelitian terdahulu telah membuktikan efektivitas metode *clustering*, khususnya K-Means Clustering, dalam menganalisis data pendidikan. Salah satu penelitian menunjukkan bahwa metode K-Means mampu mengelompokkan siswa ke dalam tiga kategori utama, yaitu tinggi, sedang, dan rendah, dengan menggunakan evaluasi Davies-Bouldin Index (DBI), sehingga menghasilkan gambaran performa akademik siswa yang lebih terstruktur dan sistematis [11]. Penelitian tersebut menggunakan dataset nilai akademik dengan

variabel terbatas, sehingga hasil pengelompokan cenderung belum menggambarkan keseluruhan kemampuan siswa secara komprehensif. Penelitian lain juga memanfaatkan metode K-Means untuk mengelompokkan siswa berdasarkan status sosial ekonomi dengan mempertimbangkan berbagai atribut, seperti pendapatan orang tua, pekerjaan orang tua, jumlah tanggungan, bantuan sosial, serta jarak tempat tinggal ke sekolah. Hasil penelitian tersebut menunjukkan bahwa pengelompokan yang dilakukan dapat mendukung pengambilan keputusan yang lebih tepat, misalnya dalam penyaluran bantuan pendidikan [12]. Namun Penelitian tersebut berfokus pada aspek sosial ekonomi siswa, sehingga berbeda dengan studi yang menitikberatkan pada performa akademik. Meskipun telah menggunakan Silhouette Score untuk evaluasi cluster, tidak dilakukan perbandingan dengan metode evaluasi lain, sehingga validitas hasil belum diuji secara komprehensif. Meskipun demikian, penelitian tersebut belum secara khusus berfokus pada pengelompokan siswa berdasarkan nilai akademik.

Secara umum, metode K-Means dikenal memiliki berbagai keunggulan, seperti kecepatan komputasi yang tinggi, kemudahan dalam implementasi, serta kemampuan dalam mengolah data dalam jumlah besar. Hal ini menjadikan K-Means sebagai salah satu metode yang paling populer dalam analisis data, termasuk dalam bidang pendidikan. Namun demikian, hasil pengelompokan yang dihasilkan oleh K-Means sangat dipengaruhi oleh beberapa faktor penting, seperti penentuan jumlah cluster dan inisialisasi centroid. Jika jumlah cluster yang ditentukan tidak sesuai, maka hasil pengelompokan dapat menjadi kurang representatif dan tidak mencerminkan kondisi sebenarnya. Selain itu, beberapa penelitian sebelumnya belum melakukan perbandingan metode penentuan jumlah cluster secara sistematis, sehingga berpotensi menghasilkan cluster yang kurang optimal.

Berdasarkan berbagai penelitian tersebut, masih terdapat beberapa kesenjangan (*research gap*) yang perlu diperhatikan. Di antaranya adalah belum adanya standar yang jelas dalam menentukan jumlah cluster yang optimal, belum optimalnya pemanfaatan metode seperti Elbow Method dalam proses penentuan jumlah cluster, keterbatasan variabel yang digunakan dalam

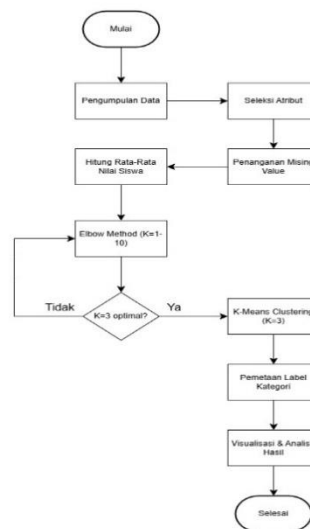
beberapa penelitian, serta hasil pengelompokan yang terkadang belum disajikan dalam bentuk yang sederhana dan mudah dipahami oleh pihak sekolah. Selain itu, kurangnya analisis komparatif antar penelitian terkait perbedaan dataset, metode evaluasi, serta hasil clustering juga menjadi kelemahan yang perlu diperbaiki dalam penelitian selanjutnya.

Oleh karena itu, penelitian ini menawarkan kebaruan (state of the art) dengan mengombinasikan metode K-Means Clustering dan Elbow Method untuk menentukan jumlah cluster secara lebih sistematis dan optimal. Selain itu, penelitian ini juga menggunakan data akademik yang mencakup berbagai mata pelajaran sehingga mampu memberikan analisis yang lebih komprehensif terhadap kemampuan siswa. Pendekatan ini juga diharapkan dapat mengatasi keterbatasan penelitian sebelumnya dengan menghadirkan proses analisis yang lebih terstruktur serta mempertimbangkan kualitas cluster yang dihasilkan. Hasil pengelompokan kemudian diklasifikasikan ke dalam tiga kategori utama, yaitu rendah, sedang, dan tinggi, agar lebih mudah dipahami dan dapat langsung digunakan sebagai dasar dalam pengambilan keputusan pembelajaran di sekolah. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan penerapan *data mining* di bidang pendidikan, khususnya dalam menghasilkan sistem pengelompokan siswa yang lebih objektif, sistematis, dan berbasis data.

3. METODOLOGI PENELITIAN

3.1. Tahapan Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang sistematis sebagaimana ditunjukkan pada Gambar 1. Tahapan ini dimulai dari pengumpulan data nilai akademik siswa, dilanjutkan dengan prapemrosesan data, penentuan nilai K optimal menggunakan Elbow Method, penerapan algoritma K-Means Clustering, pemetaan label kategori, hingga visualisasi dan analisis hasil. Tahapan-tahapan tersebut disusun secara terstruktur untuk memastikan proses pengolahan data berjalan secara sistematis dan menghasilkan informasi yang akurat [13].



Gambar 1. Diagram Alir Tahapan Penelitian

3.2. Pengumpulan Data

Data yang digunakan dalam sekunder berupa data nilai akademik siswa SMP Negeri 1 Kembang periode ajaran 2023 hingga 2025. Data diperoleh dari pihak sekolah dalam format Microsoft Excel (.xlsx). Dataset mencakup nilai dari 11 mata pelajaran untuk seluruh siswa yang terdaftar pada periode tersebut sebagaimana terinci pada Tabel 1.

Tabel 1. Daftar Variabel Mata Pelajaran

No	Mata Pelajaran	Nama Kolom
1	Bahasa Indonesia	Bahasa Indonesia
2	Pendidikan Agama	Pend.Agama
3	Matematika	Matematika
4	Pendidikan Pancasila	Pend.Pancasila
5	Bahasa Inggris	Bahasa Inggris
6	Seni & Prakarya	Seni&Prakarya
7	IPA	IPA
8	Informatika	Informatika
9	IPS	IPS
10	Bahasa Jawa	Bahasa Jawa
11	Penjasorkes	Penjasorkes

3.3. Analisa Data

Analisis data dilakukan terhadap data nilai akademik siswa SMP Negeri 1 Kembang periode 2023-2025 yang mencakup 11 mata pelajaran sesuai Tabel 1. Tujuan analisis adalah



mengelompokkan seluruh siswa yang terdaftar pada periode tersebut ke dalam tiga kategori prestasi yaitu Rendah, Sedang, dan Tinggi menggunakan metode K-Means Clustering, sehingga pihak sekolah dapat mengidentifikasi siswa yang memerlukan perhatian khusus dan merancang program pembelajaran lebih tepat sasaran.

Proses analisis dilaksanakan secara komputasional pada tahun 2026 menggunakan Google Collaboratory dengan bahasa pemrograman Python 3. Data diimpor dari berkas Microsoft Excel menggunakan library pandas, kemudian atribut numerik diseleksi dan nilai kosong ditangani dengan mean imputation. Jumlah cluster optimal ditentukan melalui Elbow Method, dan Algoritma K-Means dengan K=3 diterapkan menggunakan scikit-learn. Hasil cluster selanjutnya dipetakan secara dinamis ke label Rendah, Sedang, atau Tinggi berdasarkan rata-rata nilai akademik masing-masing cluster, kemudian dianalisis dan divisualisasikan dalam diagram batang dan tabel centroid.

3.4. Pra-pemrosesan Data

Tahap pra-pemrosesan bertujuan memastikan kualitas data sebelum diterapkan algoritma clustering. Proses ini mencakup tiga sub-tahap berikut.

a) Seleksi Atribut Numerik

Dari dataset awal, hanya kolom bertipe numerik yang dipilih sebagai atribut clustering menggunakan fungsi `select_dtypes(include=['number'])`. Kolom non-numerik seperti nama siswa tidak diikutsertakan karena tidak relevan secara kuantitatif.

b) Penanganan Missing Value

Nilai kosong pada dataset ditangani menggunakan metode mean imputation, yaitu mengganti nilai yang hilang dengan rata-rata kolom yang bersangkutan. Pendekatan ini dipilih karena data nilai akademik cenderung terdistribusi normal tanpa outlier ekstrem, sehingga nilai rata-rata bersifat representatif dan mampu mempertahankan distribusi data tanpa menghapus baris yang memiliki nilai kosong seperti pada listwise deletion [14]. Rumus mean ditunjukkan pada Persamaan (1).

$$\bar{x}_i = \frac{(\sum x_{in})}{n} \quad (1)$$

Keterangan \bar{x}_i = nilai pengganti kolom ke-i; x_{in} = nilai ke-n pada kolom ke-i; n = jumlah data valid pada kolom data valid pada kolom tersebut.

c) Perhitungan Rata-rata Nilai Siswa

Rata-rata nilai akademik setiap siswa dihitung dari seluruh 11 mata pelajaran menggunakan Persamaan (2). Nilai ini digunakan sebagai acuan pemetaan label setelah clustering selesai.

$$\bar{R}_s = \frac{(\sum_{j=1}^{11} N_j)}{11} \quad (2)$$

Keterangan : \bar{R}_s = rata-rata nilai siswa s; N_j = nilai mata pelajaran ke-j; m = jumlah mata pelajaran ($m = 11$).

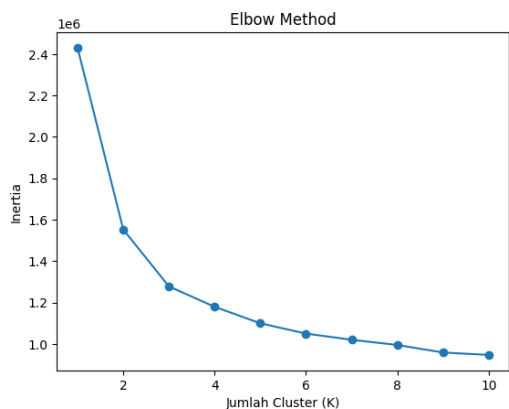
3.5. Penentuan Nilai K Optimal (Elbow Method)

Sebelum menjalankan K-Means, dilakukan penentuan nilai K yang optimal menggunakan Elbow Method. Metode ini memanfaatkan nilai inertia untuk mengukur tingkat variasi dalam setiap cluster pada berbagai jumlah cluster yang diuji [15].

$$Inertia = \sum_{k=1}^k \sum_{x \in C_k} \|x - \mu_k\|^2 \quad (3)$$

Keterangan: C_k = himpunan data pada cluster k; x = data point; μ_k = centroid cluster k.

Dalam grafik Elbow Method, sumbu-x berfungsi sebagai jumlah cluster (K) dan sumbu-y berfungsi sebagai nilai inertia. Nilai K terbaik dipilih pada titik "siku", atau elbow, di mana penurunan inertia mulai melandai secara signifikan, seperti yang ditunjukkan pada Gambar 2.



Gambar 2. Grafik Elbow Method

3.6. Tahapan K-Means Clustering

Algoritma K-Means Clustering merupakan metode pembelajaran tanpa pengawasan yang mengelompokkan n data ke dalam K cluster berdasarkan kemiripan fitur. Dalam penelitian ini digunakan K=3 dengan parameter *random_state=42* sesuai hasil Elbow Method agar hasil dapat direplikasi.

Proses K-Means dilakukan secara iteratif dengan memilih centroid awal secara acak, kemudian menghitung jarak setiap data ke centroid menggunakan Euclidean Distance (Persamaan 4). Data dikelompokkan ke centroid terdekat, lalu posisi centroid diperbarui berdasarkan rata-rata anggota cluster (Persamaan 5). Langkah ini diulang hingga centroid stabil atau mencapai batas iterasi maksimum.

$$d(x, \mu_k) = \sqrt{\sum_{j=1}^m (x_j - \mu_{kj})^2} \quad (4)$$

Keterangan : $d(x, \mu_k)$ = jarak antara data x dan centroid μ_k ; x_j = nilai fitur ke -j dari data x ; μ_{kj} = nilai fitur ke-j centroid cluster k .

Pembaruan centroid dilakukan dengan Persamaan (5). Pembaruan centroid dilakukan dengan menghitung rata-rata seluruh anggota dalam satu cluster.

$$\mu_{kj} = \frac{1}{|C_k|} \sum_{x \in C_k} x_j \quad (5)$$

Keterangan : μ_k = centroid cluster k yang diperbarui; $|C_k|$ = jumlah anggota cluster k ; x = data point anggota cluster k .

3.7. Pemetaan Label Kategori

Setelah K-Means selesai, Nomor cluster yang dihasilkan dapat dipilih sendiri setelah K-Means selesai. Kumpulan diberi label secara dinamis berdasarkan nilai rata-rata akademik. Kumpulan dengan nilai rata-rata terendah diberi label Rendah, kumpulan dengan nilai rata-rata menengah diberi label Sedang, dan kumpulan dengan nilai rata-rata tertinggi diberi label Tinggi [13]. Mekanisme pemetaan ini ditunjukkan pada Tabel 3.

Tabel 3. Pemetaan Label Kategori Cluster

Peringkat Rata-rata	Label	Interpretasi
Terendah	Rendah	Perlu bimbingan intensif
Menengah	Sedang	Prestasi rata-rata, perlu peningkatan
Tertinggi	Tinggi	Siswa berprestasi baik

4. HASIL DAN PEMBAHASAN

4.1. Transformasi dan Pra-pemrosesan Data

Data yang digunakan berupa nilai rapor siswa SMP Negeri 1 Kembang angkatan 2023–2025 yang diperoleh dalam format Microsoft Excel, dengan total 1.011 data siswa dan 11 variabel mata pelajaran. Sebanyak 11 atribut dipilih karena dianggap mewakili kemampuan akademik secara komprehensif, meliputi Bahasa Indonesia, Pendidikan Agama, Matematika, Pendidikan Pancasila, Bahasa Inggris, Seni & Prakarya, IPA, Informatika, IPS, Bahasa Jawa, dan Penjasorkes.

Tahap pra-pemrosesan (preprocessing) dilaksanakan untuk menjamin mutu dan keseragaman data sebelum clustering, dengan menerapkan teknik *data cleaning* guna menangani data yang kurang lengkap maupun kurang konsisten melalui beberapa langkah yang telah ditentukan.

1. Seleksi Variabel Numerik: Menggunakan fungsi `select_dtypes(include=['number'])` pada library Pandas untuk mengisolasi 11 mata pelajaran sebagai fitur numerik independen yang akan digunakan dalam perhitungan clustering.



2. Penanganan Missing Value: Nilai kosong atau data yang berisi tanda strip (-) diganti dengan nilai rata-rata kolom menggunakan teknik mean imputation. Pendekatan ini dipilih karena mampu mempertahankan ukuran sampel (N=1.011) tanpa menggeser distribusi data secara signifikan.
3. Penghapusan Kolom Non-Esensial: Kolom nama siswa, peringkat, dan metadata administratif dihapus dari proses clustering karena tidak berkontribusi pada analisis pola nilai akademik, meskipun tetap dipertahankan untuk keperluan identifikasi hasil akhir.

Setelah melalui tahap pra-pemrosesan, diperoleh dataset final sebanyak 1.011 observasi yang valid dan siap untuk dianalisis. Verifikasi statistik menunjukkan bahwa distribusi data sebelum dan sesudah cleaning tidak mengalami pergeseran signifikan (perbedaan mean <2%), mengonfirmasi bahwa integritas data tetap terjaga.

4.2. Penentuan Nilai K-Optimal Menggunakan Metode Elbow

Penentuan K-Optimal Value Menggunakan Metode Elbow Pemilihan jumlah cluster (K) yang optimal merupakan tahapan fundamental dalam algoritma K-Means karena sangat mempengaruhi kualitas hasil pengelompokan. Pada penelitian ini, penentuan nilai K dilaksanakan menggunakan metode Elbow dengan menghitung nilai inerti (Within-Cluster Sum of Squares) untuk nilai K dari 1 hingga 10 [16].

Tabel 4. Nilai Inertia pada Berbagai Nilai K

K	Inertia	Penurunan (%)
1	191.245,67	-
2	99.832,15	47,80
3	68.541,23	31,34
4	52.678,91	23,14
5	44.321,56	15,86
6	38.654,32	12,79
7	34.012,78	11,98
8	30.123,45	11,43
9	27.089,12	10,07
10	24.756,89	8,61

Berdasarkan Tabel 4, terjadi penurunan nilai inerti yang signifikan dari K=1 ke K=2 (47,80%) dan dari K=2 ke K=3 (31,34%). Namun, setelah K=3, laju penurunan inerti menjadi lebih landai dan stabil, menunjukkan bahwa penambahan cluster tambahan tidak menghasilkan peningkatan kualitas pengelompokan yang berarti. Fenomena ini membentuk pola "siku" (elbow) pada grafik yang menunjukkan titik optimal pada K=3.

Pemilihan K=3 juga selaras dengan praktik umum dalam segmentasi prestasi akademik di Indonesia yang mengklasifikasikan siswa ke dalam tiga kategori: Rendah, Sedang, dan Tinggi. Dengan demikian, nilai K=3 ditetapkan sebagai parameter optimal untuk penggunaan algoritma K-Means dalam studi ini.

4.3. Proses Clustering Menggunakan Algoritma K-Means

Setelah nilai K ditentukan, proses pengelompokan dilaksanakan dengan menginisialisasi tiga centroid awal secara acak menggunakan parameter `random_state=42` untuk menjamin reproduktibilitas hasil. Perhitungan jarak setiap data ke centroid dilakukan dengan rumus jarak Euclidean:

$$d(x, c) = \sum_{i=1}^n (x_i - c_i)^2 \quad (1)$$

dengan x merupakan vektor data siswa, c merupakan vektor centroid, dan n merupakan jumlah fitur (11 mata pelajaran).

Proses iterasi dilakukan hingga konvergensi tercapai, yaitu saat posisi centroid tidak mengalami perubahan yang berarti lagi. Pada penelitian ini, algoritma mencapai konvergensi pada iterasi ke-7 dengan nilai inerti akhir sebesar 68.541,23. Hasil akhir nilai centroid setelah iterasi ke-7 disajikan pada Tabel 2.



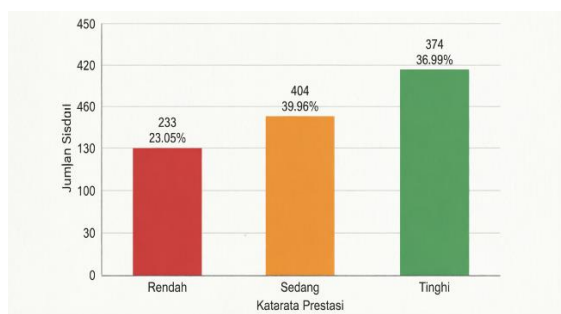
Tabel 5. Nilai Akhir Centroid Iterasi-7

Mata Pelajaran	Tinggi	Rendah	Sedang
Bahasa Indonesia	76,24	52,18	69,91
Pendidikan Agama	78,05	57,81	74,26
Matematika	68,47	33,95	38,15
Pendidikan Pancasila	75,48	64,51	77,67
Bahasa Inggris	75,12	43,29	57,36
Seni & Prakarya	77,71	62,63	75,52
IPA	71,45	37,83	54,28
Informatika	74,43	49,21	58,37
IPS	76,84	42,65	61,36
Bahasa Jawa	76,57	51,63	69,93
Penjasorkes	83,01	70,71	81,21
Rata-rata Keseluruhan	75,76	51,49	65,28

Berdasarkan nilai centroid pada Tabel 5, ketiga cluster kemudian dipetakan ke dalam kategori edukatif berdasarkan urutan rata-rata nilai keseluruhan: Cluster 1 → Kategori Rendah, Cluster 2 → Kategori Sedang, dan Cluster 0 → Kategori Tinggi. Hasil pengelompokan final untuk seluruh 1.011 siswa disajikan pada Tabel 3 [17].

Tabel 3. Distribusi Siswa Berdasarkan Kategori Prestasi

Visualisasi distribusi jumlah siswa per kategori ditampilkan pada Gambar 1.



Gambar 1. Diagram Distribusi Siswa Berprestasi

4.4. Evaluasi Hasil Clustering Menggunakan Davies Bouldin Index

Untuk memvalidasi kualitas hasil pengelompokan, dilakukan evaluasi menggunakan metode Davies Bouldin Index (DBI) yang mengukur rasio antara kohesi intra-cluster dan separasi antar-cluster. Perhitungan DBI dilakukan melalui tiga tahap utama:

Tahap 1: Perhitungan Sum of Square Within-cluster (SSW)

SSW mengukur kedekatan data terhadap centroid cluster masing-masing menggunakan persamaan:

$$SSWi = \frac{1}{mi} \sum_{j=1}^{mi} d(xj, ci) \quad (2)$$

dengan mi= kuantitas data dalam cluster ke-i, xj = data yang terdapat dalam cluster, dan ci= centroid.

Hasil perhitungan SSW untuk ketiga cluster:

- $SSW_1 = 17,32456$ (Cluster Rendah)
- $SSW_2 = 9,45678$ (Cluster Sedang)
- $SSW_3 = 8,85123$ (Cluster Tinggi)

Tahap 2: Perhitungan Sum of Square Between-cluster (SSB)

SSB mengukur jarak antar centroid cluster menggunakan persamaan:

$$SSBij = d(ci, cj) \quad (3)$$

Kategori	Jumlah Siswa	Persentase (%)	Rata-rata Nilai
Rendah	233	23,05	51,49
Sedang	404	39,96	65,28
Tinggi	374	36,99	75,76
Total	1.011	100	-

Hasil matriks SSB disajikan pada Tabel 6.

Tabel 6. Matriks Hasil Perhitungan SSB

SSB	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	27,8352	38,7123
Cluster 2	27,8352	0	15,1876
Cluster 3	38,7123	15,1876	0



Tahap 3: Perhitungan Rasio dan DBI

Rasio antara SSW dan SSB dihitung menggunakan persamaan:

$$Ri, j = \frac{SSWi+SSWj}{SSWi,j} \cdot m \tag{4}$$

Nilai DBI kemudian dihitung sebagai rata-rata dari maksimum rasio setiap cluster:

$$DBI = \frac{1}{k} \sum_i^k = 1 \text{ maxi} = j(Ri, j) \tag{5}$$

Hasil perhitungan menghasilkan nilai DBI = 1,12567. Semakin mendekati 0 nilai DBI, maka semakin baik kualitas pengelompokan yang diperoleh. Nilai DBI yang diperoleh dalam penelitian ini relatif sangat baik, mengonfirmasi bahwa hasil clustering dengan K=3 memiliki kualitas pemisahan cluster yang optimal.

4.5. Analisis Karakteristik Akademik per Kategori

Analisis lebih lanjut dilakukan untuk mengidentifikasi pola kekuatan dan kelemahan akademik di setiap kategori. Berdasarkan perhitungan statistik deskriptif menggunakan fungsi pada Pandas, terlihat perbedaan yang signifikan antar kelompok pada mata pelajaran tertentu, sebagaimana dilihat pada Tabel 7.

Tabel 7. Rata-rata Nilai per Mata Pelajaran Berdasarkan Kategori

Mata Pelajaran	Rendah	Sedang	Tinggi
Bahasa Indonesia	52,18	69,91	76,24
Matematika	33,95	38,15	68,47
IPA	37,83	54,28	71,45
Bahasa Inggris	43,29	57,36	75,12
Pendidikan Agama	57,81	74,26	78,05
Pendidikan Pancasila	64,51	77,67	75,48
Seni & Prakarya	62,63	75,52	77,71
Informatika	49,21	58,37	74,43
IPS	42,65	61,36	76,84
Bahasa Jawa	51,63	69,93	76,57
Penjasorkes	70,71	81,21	83,01

Pada kategori Rendah, nilai terendah umumnya ditemukan pada mata pelajaran eksakta dan logika seperti Matematika (rata-rata 33,95) dan IPA (rata-rata 37,83). Pola ini mengindikasikan adanya hambatan dalam pemahaman konsep analitis, pemecahan masalah, dan penalaran kuantitatif. Sebaliknya, pada mata pelajaran yang bersifat hafalan atau praktik seperti Penjasorkes, siswa kategori Rendah menunjukkan nilai yang relatif lebih tinggi (70,71).

Pada kategori Sedang, siswa menunjukkan performa yang stabil di hampir seluruh mata pelajaran dengan rentang nilai antara 54–81. Kategori ini merepresentasikan siswa yang telah mencapai kompetensi dasar tetapi memerlukan strategi pembelajaran yang lebih interaktif untuk mendorong peningkatan prestasi.

Sementara itu, pada kategori Tinggi, siswa menunjukkan konsistensi unggul di semua bidang studi, dengan nilai Bahasa Indonesia, Bahasa Inggris, dan IPS mencapai rata-rata di atas 75,0. Kemampuan ini mencerminkan profil siswa yang tidak hanya kuat di aspek kognitif, tetapi juga memiliki literasi digital dan kreativitas yang terasah.

Temuan pada kategori Rendah sejalan dengan teori kognitif Bloom yang menegaskan bahwa pemahaman konsep analitis dan penalaran kuantitatif merupakan kemampuan berpikir tingkat tinggi yang memerlukan fondasi dasar yang kuat [1]. Rendahnya nilai Matematika (33,95) dan IPA (37,83) pada kelompok ini mengindikasikan bahwa siswa belum menguasai kompetensi prasyarat, sehingga intervensi berupa program remedial terstruktur sangat diperlukan. Hal ini selaras dengan temuan Hasugian dan Sagala yang menyatakan bahwa siswa dalam kluster berprestasi rendah membutuhkan pendekatan pembelajaran individual dan bimbingan intensif guna menutup kesenjangan kompetensi dasar.

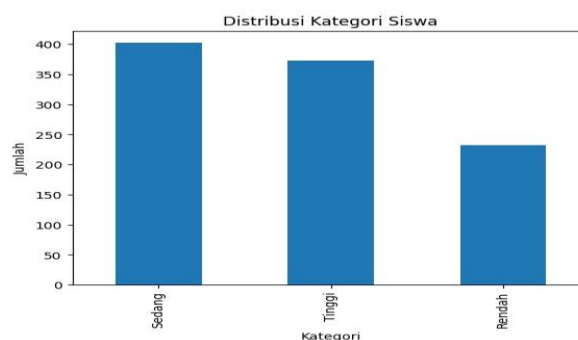
Karakteristik kategori Sedang yang menunjukkan performa stabil di rentang nilai 54–81 mencerminkan kondisi disebut sebagai “zona perkembangan proksimal”, yaitu kondisi di mana siswa telah menguasai kompetensi dasar namun masih memerlukan stimulasi tambahan untuk mencapai potensi optimal. Penelitian terdahulu juga menemukan pola serupa, di mana siswa kelompok menengah memberikan respons positif terhadap strategi pembelajaran kooperatif dan berbasis proyek dibandingkan pendekatan

konvensional. Dengan proporsi terbesar (39,96%), kelompok ini menjadi prioritas strategis dalam perancangan program peningkatan kualitas pembelajaran secara menyeluruh [6].

Profil kategori Tinggi yang menunjukkan konsistensi unggul di seluruh mata pelajaran mendukung konsep “academic resilience” yang dikemukakan dalam literatur psikologi pendidikan, yakni kemampuan siswa untuk mempertahankan kinerja tinggi secara merata meskipun menghadapi variasi tuntutan akademik antarmatapelajaran. Temuan ini juga selaras dengan hasil penelitian sebelumnya yang menemukan bahwa siswa dalam kluster berprestasi tinggi cenderung memiliki konsistensi nilai di atas rata-rata pada seluruh komponen penilaian, bukan hanya pada mata pelajaran tertentu saja. Hal ini memperkuat validitas hasil clustering yang dihasilkan dalam penelitian ini, sekaligus mengonfirmasi bahwa metode K-Means Clustering mampu membedakan profil akademik siswa secara akurat dan bermakna secara pedagogis [11].

4.6. Analisis Hasil Clustering

Ada tiga metode yang digunakan untuk menganalisis hasil clustering. Pertama, fungsi `value_counts()` digunakan untuk mendapatkan distribusi jumlah siswa per kategori, yang kemudian digambarkan dalam bentuk diagram batang (Gambar 3). Kedua, untuk memastikan bahwa ada perbedaan dalam kinerja antar kelompok, rata-rata nilai akademik per kelompok dihitung dengan menggunakan metode kelompok. Ketiga, nilai centroid untuk setiap cluster dihitung untuk menghasilkan profil akademik untuk masing-masing cluster. Nilai centroid untuk empat mata pelajaran utama Bahasa Inggris, Matematika, IPA, dan Bahasa Indonesia juga dihitung per kategori untuk menunjukkan perbedaan kemampuan di mata pelajaran inti.



Gambar 3. Distribusi Jumlah Siswa per Kategori

Seluruh proses di atas diimplementasikan menggunakan Python 3 pada lingkungan Google Colaboratory, dengan library `pandas` untuk manipulasi data dan `scikit-learn` untuk implementasi K-Means Clustering.

4.7. Evaluasi Sistem

Berdasarkan pelaksanaan, uji coba, dan analisis yang telah dilaksanakan, sistem pengelompokan nilai akademik berbasis K-Means Clustering memiliki beberapa kelebihan yang menonjol [18]:

1. Efisiensi komputasi: Algoritma ini bersifat ringan dalam pemrosesan, sehingga dapat dijalankan pada perangkat standar tanpa memerlukan infrastruktur *cloud* atau GPU.
2. Integritas data: Proses pra-pemrosesan dengan *mean imputation* berhasil mempertahankan integritas dataset (N=1.011) tanpa mengorbankan observasi.
3. Output yang dapat ditindaklanjuti: Kategori jelas (Rendah, Sedang, Tinggi) dilengkapi dengan metrik rata-rata dan distribusi persentase, sehingga langsung dapat digunakan oleh guru dan staf administrasi untuk merancang program remedial atau pengayaan.
4. Modularitas: Struktur kode yang modular memungkinkan integrasi mudah dengan sistem informasi akademik yang telah tersedia di sekolah.

Di sisi lain, terdapat beberapa kelemahan yang perlu diakui:

1. Keterbatasan bentuk cluster: K-Means memiliki keterbatasan inherent dalam mengatasi data yang mengandung *outliers* ekstrem atau cluster yang tidak berbentuk bulat (*non-spherical*).



- Validasi parameter K: Penentuan nilai K=3 didasarkan pada kombinasi *Elbow Method* dan kebutuhan kebijakan sekolah, namun validasi eksternal dengan metode lain seperti *Silhouette Score* dapat memperkuat objektivitas pemilihan parameter.
- Fitur terbatas: Penggunaan fitur yang murni akademik; variabel non-akademik seperti tingkat kehadiran, motivasi belajar, atau kondisi sosio-ekonomi tidak disertakan dalam model.
- Sifat statis model: Model ini perlu diperbarui secara berkala seiring masuknya data nilai baru untuk menjaga relevansi dan akurasi.

Kelemahan-kelemahan tersebut bukanlah kegagalan metode, melainkan ruang pengembangan yang dapat diatasi pada penelitian masa depan dengan pendekatan *hybrid*, penambahan fitur sosio-afektif, dan validasi *cluster* yang lebih ketat.

5. KESIMPULAN DAN SARAN

Metode K-Means Clustering dengan K=3 berhasil mengelompokkan siswa SMP Negeri 1 Kembang berdasarkan nilai akademik periode 2023–2025 ke dalam tiga tingkat yaitu rendah, sedang, dan tinggi, dengan dominasi pada kelompok sedang. Penerapan 11 variabel mata pelajaran menghasilkan gambaran kemampuan akademik yang menyeluruh, sehingga metode ini terbukti efektif sebagai alat pendukung pengambilan keputusan akademik berbasis data di lingkungan sekolah. Keterbatasan penelitian terletak pada sensitivitas penentuan jumlah cluster dan belum dipertimbangkannya faktor non-akademik seperti kehadiran dan motivasi belajar.

Penelitian selanjutnya disarankan menggunakan metode pembandingan seperti Hierarchical Clustering atau DBSCAN, menerapkan evaluasi cluster melalui *Silhouette Score* atau *Davies-Bouldin Index*, serta mengintegrasikan hasil pengelompokan ke dalam sistem informasi akademik sekolah agar dapat dimanfaatkan langsung dalam pengambilan keputusan.

6. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada SMP Negeri 1 Kembang yang telah memberikan persetujuan serta menyediakan data yang diperlukan dalam penelitian. Selain itu, penulis

mengapresiasi dukungan dari seluruh pihak yang turut memberikan kontribusi secara langsung maupun tidak langsung sehingga penelitian ini dapat terselesaikan dengan baik.

DAFTAR PUSTAKA:

- [1] P. S. Hasugian dan J. R. Sagala, "Penerapan Data Mining Untuk Pengelompokan Siswa Berdasarkan Nilai Akademik dengan Algoritma K-Means," vol. 3, no. 3, hal. 262–268, 2022.
- [2] R. A. Putri, I. Magdalena, A. Fauziah, F. N. Azizah, dan U. M. Tangerang, "PENGARUH GAYA BELAJAR TERHADAP PEMBELAJARAN SISWA SEKOLAH DASAR," vol. 1, no. 2, hal. 157–163, 2021.
- [3] F. I. Himmah dan N. Nugraheni, "Analisis Gaya Belajar Siswa untuk Pembelajaran Berdiferensiasi Analysis of Student Learning Styles for Differentiated Learning," vol. 4, hal. 31–39, 2023, doi: 10.30595/jrpd.v4i1.16045.
- [4] F. Khalish, N. M. Piranti, dan O. Martadireja, "Implementasi Data Mining Menggunakan Teknik Clustering dengan Metode K-Means," vol. 8, hal. 5392–5397, 2025.
- [5] S. Paembonan dan H. Abduh, "Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat," vol. 6, no. 2, hal. 48–54, 2021.
- [6] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," vol. 3, hal. 46–56, 2024.
- [7] J. Riset dan S. Informasi, "PENGELOMPOKAN DATA NILAI SISWA MADRASAH TA' HILYAH MENGGUNAKAN METODE K-MEANS CLUSTERING," vol. 2, no. 1, hal. 53–59, 2025.
- [8] I. Engineering, "Penerapan Metode Clustering K-Means Menggunakan RapidMiner untuk Klasifikasi Prestasi Siswa di Sekolah Swasta," vol. 1, no. 1, hal. 20–24, 2025.
- [9] I. Pendahuluan, "Pemanfaatan Data Mining dengan K-Means Clustering Dalam Penilaian Guru Berprestasi," vol. 9, no. 3, hal. 1124–1138, 2025.
- [10] O. Soleh dan D. Jonas, "Nusa: journal of science studies," vol. 1, no. 2, hal. 63–69,



- 2024.
- [11] S. N. Adzra, F. N. Hasan, dan A. Y. Kuntoro, "Penerapan Data Mining dalam Penilaian Kinerja Akademik Siswa / I SMP YPI Pulogadung dengan Metode K-Means Clustering," no. 98, 2025.
- [12] D. E. Putri, E. Praja, W. Mandala, dan T. Informatika, "Penerapan K-Means Clustering dalam Segmentasi Siswa Berdasarkan Status Sosial Ekonomi".
- [13] M. Tingkat *et al.*, "Penerapan Data Mining Menggunakan K-Means Clustering Dalam," vol. 6, no. 3, 2025, doi: 10.47065/josh.v6i3.6959.
- [14] G. Mafela, M. Sujak, H. N. Rofiq, dan F. I. Tawakal, "Implementation of K-Means Clustering for Optimizing Non-Communicable Disease Budgets Implementasi K-Means Clustering untuk Optimalisasi Anggaran Penyakit Tidak Menular," vol. 5, no. January, hal. 67-74, 2025.
- [15] M. T. Hidayat *et al.*, "DASHBOARD BUSINESS INTELLIGENCE DENGAN QLIK SENSE UNTUK ANALISIS NUTRISI MAKANAN," vol. 13, no. 1, 2025.
- [16] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," vol. 14, no. 2, hal. 277-287, 2023.
- [17] D. I. Smp dan N. Ssn, "PENERAPAN ALGORITME K-MEANS CLUSTERING UNTUK MENGELOMPOKKAN SISWA BERDASARKAN NILAI AKADEMIK IMPLEMENTATION OF K-MEANS CLUSTERING ALGORITHM TO GROUP STUDENTS BASED ON ACADEMIC VALUE IN SMP NEGERI," vol. 2, no. September, hal. 530-538, 2023.
- [18] N. Wakhidah, "CLUSTERING MENGGUNAKAN K-MEANS ALGORITHM (K-MEANS ALGORITHM CLUSTERING)".