VOL. 03. NO. 5 SEPTEMBER 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

PENERAPAN ALGORITMA K-MEANS UNTUK MENGELOMPOKKAN SISWA SMK NEGERI 1 MAJALAYA BERDASARKAN PRESTASI AKADEMIK

Rifky Rayandi Maulana[1], Nana Suarna[2], Agus bahtiar[3], Faturrohman[4].

Program Studi Teknik Informatika[12]
Program Studi Sistem Informasi[3]
Program Studi Rekayasa Perangkat Lunak[4]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
rifkyrayandi@gmail.com

(*) Corresponding Author : rifkyrayandi@gmail.com
Published : 15 Juli 2026

Abstract—*The manual grouping of students based on academic achievement at SMK Negeri 1 Majalaya has been carried out subjectively and inefficiently, making it difficult for the school to design targeted learning strategies. This study aims to apply the K-Means Clustering algorithm to objectively group students based on academic grades, determine the optimal number of clusters, and utilize the clustering results as a basis for academic decision-making and differentiated learning strategy design. The method used is a quantitative descriptive approach employing the K-Means algorithm implemented in Python on the Google Colab platform. The dataset consists of average grades from 535 students across 12 academic and vocational subjects obtained from SMK Negeri 1 Majalaya. The research stages include data acquisition, preprocessing (data cleaning and normalization using StandardScaler), application of the K-Means algorithm, determination of the optimal number of clusters using the Elbow Method and Silhouette Score, and interpretation of the clustering results. The findings reveal that the optimal number of clusters is two, representing students with sufficient but below-standard academic achievement (322 students) and students with good academic achievement (214 students). A Silhouette Score of 0.5290 and a Calinski-Harabasz Index of 846.14 indicate that the model has a good level of inter-cluster separation. In conclusion, the K-Means algorithm is proven effective in objectively grouping students based on academic achievement and can serve as a foundation for schools to design more targeted enrichment and remedial programs.*

Keywords: *K-Means Clustering, Data Mining, Academic Performance, Clustering, SMK Negeri 1 Majalaya*

Abstrak—Pengelompokan siswa berdasarkan prestasi akademik secara manual di SMK Negeri 1 Majalaya masih dilakukan secara subjektif dan tidak efisien, sehingga menyulitkan pihak sekolah dalam merancang strategi pembelajaran yang tepat sasaran. Penelitian ini bertujuan untuk menerapkan algoritma *K-Means Clustering* guna mengelompokkan siswa berdasarkan nilai akademik secara objektif, menentukan jumlah klaster optimal, serta memanfaatkan hasil klasterisasi sebagai dasar pengambilan keputusan akademik dan perancangan strategi pembelajaran diferensiasi. Metode yang digunakan adalah pendekatan kuantitatif deskriptif dengan menerapkan algoritma *K-Means* menggunakan bahasa pemrograman *Python* di platform *Google Colab*. Data yang digunakan merupakan nilai rata-rata 535 siswa dari 12 mata pelajaran akademik dan kejuruan yang diperoleh dari SMK Negeri 1 Majalaya. Tahapan penelitian meliputi akuisisi data, pra-pemrosesan (pembersihan dan normalisasi menggunakan *StandardScaler*), penerapan algoritma *K-Means*, penentuan jumlah klaster optimal menggunakan metode *Elbow* dan *Silhouette Score*, serta interpretasi hasil klasterisasi. Hasil penelitian menunjukkan bahwa jumlah klaster optimal adalah dua, yaitu kelompok siswa berprestasi cukup/belum optimal (322 siswa) dan berprestasi baik (214 siswa). Nilai *Silhouette Score* sebesar 0,5290 dan *Calinski-Harabasz Index* sebesar 846,14 menunjukkan bahwa model memiliki kemampuan pemisahan antar klaster yang baik. Kesimpulannya, algoritma *K-Means* terbukti efektif digunakan untuk mengelompokkan siswa berdasarkan prestasi akademik secara objektif dan dapat dijadikan dasar bagi sekolah dalam merancang program pengayaan, pembinaan, dan remedial yang lebih terarah.

Kata Kunci: *K-Means Clustering*, Data Mining, Prestasi Akademik, Klasterisasi, SMK Negeri 1 Majalaya

INTRODUCTION

Perkembangan teknologi informasi telah mendorong munculnya berbagai pendekatan analisis data dalam bidang pendidikan, salah satunya melalui *data mining*. Salah satu teknik yang paling banyak digunakan dalam analisis data pendidikan adalah *clustering*, khususnya algoritma *K-Means*. Metode ini memungkinkan pengelompokan siswa ke dalam kelompok yang bermakna seperti berprestasi tinggi, sedang, dan rendah, sehingga dapat memberikan wawasan penting bagi sistem peringatan dini, perancangan intervensi pembelajaran, dan pengambilan kebijakan pendidikan [1].

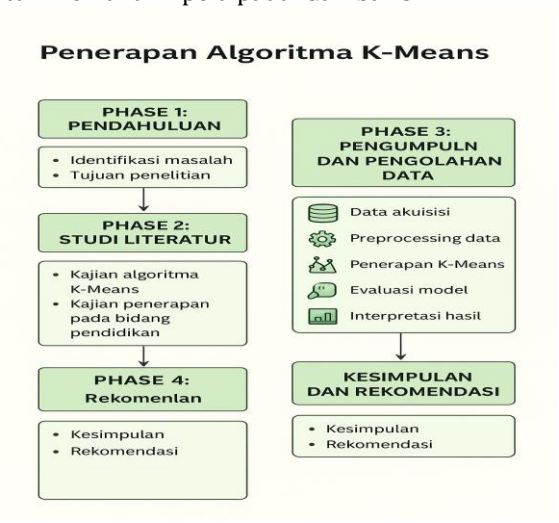
Berbagai penelitian terdahulu menunjukkan bahwa algoritma *K-Means* mampu mengungkap pola dan karakteristik siswa berdasarkan performa akademik mereka. menegaskan bahwa *K-Means* dapat menghasilkan klaster yang mudah diinterpretasikan, sehingga mendukung identifikasi dini terhadap siswa berisiko (*at-risk students*) serta meningkatkan akurasi prediksi ketika label klaster digunakan sebagai variabel tambahan dalam model *supervised learning*. Selain itu, penelitian oleh [2] memperlihatkan bahwa kombinasi *K-Means* dengan metode klasifikasi seperti *Support Vector Machine* dapat membantu mengidentifikasi pola perilaku dan kinerja akademik siswa secara lebih komprehensif. Di sisi lain, [3]menunjukkan pentingnya integrasi *K-Means* dalam proses *feature selection* untuk memaksimalkan efisiensi model prediksi performa akademik.

Kelebihan utama *K-Means* dalam konteks *educational data mining* terletak pada kemampuannya menangani data berukuran besar secara efisien, menghasilkan kelompok yang bermakna, serta memberikan representasi *centroid* yang dapat dijadikan dasar pengambilan keputusan pedagogis [4]. Penelitian lain menunjukkan bahwa penggunaan *K-Means* sebagai tahap prapemrosesan dapat meningkatkan performa model klasifikasi [5]. Selain itu, studi terkini menyoroti skalabilitas dan kompleksitas komputasi yang rendah dari algoritma ini, menjadikannya cocok diterapkan pada konteks sekolah menengah [6] Implementasi *K-Means* juga terbukti efektif dalam mengelompokkan siswa untuk optimalisasi penempatan kelas dan penyesuaian jalur pembelajaran [7], serta memberikan wawasan mengenai pola perilaku akademik siswa di lingkungan sekolah kejuruan [8].

Berdasarkan temuan-temuan tersebut, penerapan algoritma *K-Means* dalam pengelompokan siswa SMK Negeri 1 Majalaya berdasarkan capaian akademik menjadi relevan dan strategis. Melalui pengelompokan ini, sekolah dapat mengenali kelompok siswa dengan karakteristik prestasi yang berbeda, sehingga intervensi pembelajaran dapat dirancang secara lebih tepat sasaran untuk meningkatkan mutu pendidikan dan efektivitas pembinaan akademik.

MATERIALS AND METHODS

Pemilihan desain ini juga didasari oleh karakteristik algoritma *K-Means* yang bekerja secara eksploratif, menyesuaikan pusat klaster (*centroid*) berdasarkan jarak Euclidean untuk meminimalkan variasi dalam kelompok. Hasilnya adalah pembagian siswa ke dalam kelompok-kelompok homogen yang merepresentasikan profil akademik berbeda, sehingga dapat digunakan untuk memahami pola pada Gambar 3.1



Gambar 1 Desain Penelitian

Phase 1: Pendahuluan

Tahap pertama yaitu Pendahuluan merupakan langkah awal yang berfungsi untuk membangun dasar dari penelitian. Pada fase ini dilakukan identifikasi masalah guna memahami isu utama yang menjadi fokus kajian, seperti kesulitan sekolah dalam mengelompokkan siswa berdasarkan prestasi akademik secara objektif dan efisien. Setelah masalah diidentifikasi, peneliti menentukan tujuan penelitian yang pada dasarnya bertujuan untuk menerapkan algoritma *K-Means* dalam mengelompokkan data siswa berdasarkan nilai akademik mereka. Fase ini menjadi landasan penting karena memberikan arah yang jelas terhadap jalannya penelitian.

Phase 2: Studi Literatur

Fase kedua yaitu Studi Literatur, berfungsi untuk memberikan pemahaman teoritis dan mendalam mengenai konsep-konsep yang digunakan dalam penelitian. Pada tahap ini dilakukan kajian terhadap algoritma *K-Means*, meliputi prinsip kerja, tahapan proses, serta kelebihan dan kekurangannya dalam analisis data. Selain itu, dilakukan pula kajian penerapan algoritma pada bidang pendidikan, khususnya pada kasus pengelompokan siswa, analisis prestasi belajar, dan pengambilan keputusan akademik. Melalui studi literatur ini, peneliti memperoleh dasar teori dan referensi ilmiah yang relevan untuk mendukung implementasi dan analisis hasil penelitian.

Phase 3: Pengumpulan dan Pengolahan Data

Fase ketiga merupakan inti dari penelitian, yaitu Pengumpulan dan Pengolahan Data. Tahapan ini terdiri dari beberapa sub-proses penting:

Data Akuisisi Proses pengumpulan data nilai siswa dari sumber yang valid, seperti arsip sekolah atau sistem akademik. Data yang diperoleh berisi informasi numerik berupa nilai mata pelajaran dan identitas siswa.

Preprocessing Data Dilakukan untuk menyiapkan data sebelum dianalisis, termasuk pembersihan data (menghapus nilai kosong dan duplikasi), pemilihan kolom numerik, serta normalisasi agar setiap variabel memiliki skala yang seimbang.

Penerapan *K-Means* ,Algoritma *K-Means* diterapkan untuk membentuk klaster siswa berdasarkan kemiripan nilai akademik. Proses ini melibatkan penentuan jumlah klaster optimal menggunakan metode *Elbow* dan *Silhouette Score*.

Evaluasi Model – Setelah klaster terbentuk, dilakukan pengujian dengan metrik evaluasi seperti *Silhouette Score* dan *Calinski-Harabasz Index* untuk menilai seberapa baik model dalam memisahkan antar klaster.

Interpretasi Hasil – Hasil *clustering* diinterpretasikan untuk memberi makna terhadap setiap klaster, seperti mengategorikan siswa menjadi kelompok berprestasi tinggi, sedang, dan rendah.

Tahap ini menjadi pondasi utama dalam menghasilkan analisis yang akurat dan bermakna bagi pihak sekolah.

Phase 4: Kesimpulan Rekomendasi

Tahapan terakhir dalam penelitian ini merupakan fase Rekomendasi dan Kesimpulan, yang berfungsi untuk menyimpulkan hasil dari seluruh proses penerapan algoritma *K-Means* serta memberikan saran strategis berdasarkan temuan penelitian.

Pada penelitian ini, penerapan algoritma *K-Means Clustering* digunakan untuk mengelompokkan siswa di SMK Negeri 1 Majalaya berdasarkan prestasi akademik yang diperoleh dari nilai rata-rata mata pelajaran. Proses pengolahan data dimulai dari tahap akuisisi data, *preprocessing* (pembersihan dan normalisasi), penerapan algoritma *K-Means* untuk pembentukan klaster, evaluasi model, hingga interpretasi hasil klasterisasi.

RESULTS AND DISCUSSION

Pada tahap ini dilakukan proses penerapan algoritma *K-Means* sebagai inti dari penelitian untuk mengelompokkan siswa berdasarkan kesamaan nilai akademik mereka. Setelah data berhasil dikumpulkan dan melalui tahap *preprocessing*, langkah selanjutnya adalah menerapkan algoritma *K-Means* agar dapat membentuk kelompok siswa dengan karakteristik akademik yang serupa. Penerapan ini bertujuan untuk melihat pola prestasi siswa secara objektif melalui pendekatan berbasis data, sehingga hasil pengelompokan dapat dimanfaatkan untuk mendukung evaluasi dan strategi peningkatan kualitas belajar di SMK Negeri 1 Majalaya.

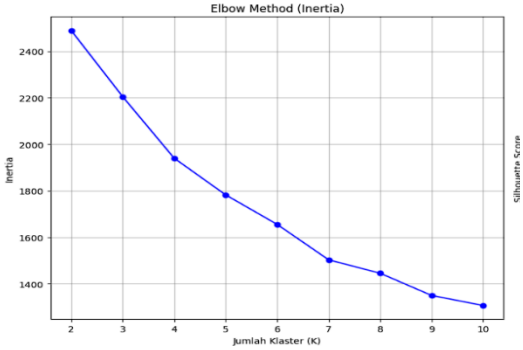
Pada tahap penentuan jumlah klaster optimal, penelitian ini menggunakan pendekatan *Elbow Method* yang dikombinasikan dengan evaluasi menggunakan Silhouette Score dan Calinski-Harabasz Index. Proses dimulai dengan inisialisasi tiga variabel, yaitu *inertia*, *silhouette_scores*, dan *calinski_scores* untuk menyimpan hasil evaluasi dari masing-masing jumlah klaster. Selanjutnya, dilakukan iterasi terhadap jumlah klaster (k) dalam rentang 2 hingga 10. Pada setiap iterasi, algoritma K-Means dijalankan dengan parameter jumlah klaster sesuai nilai k, menggunakan *random_state* untuk memastikan konsistensi hasil dan *n_init* untuk meningkatkan stabilitas model. Setelah model dilatih menggunakan data yang telah distandardisasi (X_scaled), nilai *inertia* dihitung dan disimpan sebagai bagian dari metode Elbow untuk melihat tingkat penurunan variansi dalam klaster. Selain itu, dilakukan perhitungan nilai Silhouette Score untuk mengukur tingkat pemisahan dan kekompakan klaster, serta Calinski-Harabasz Index untuk mengevaluasi kualitas struktur klaster secara keseluruhan. Seluruh nilai evaluasi tersebut kemudian disimpan untuk setiap nilai k, sehingga dapat digunakan sebagai dasar dalam menentukan jumlah klaster yang paling optimal.

Hasil penerapan algoritma *K-Means* pada penelitian ini bertujuan untuk menentukan jumlah klaster optimal dan mengelompokkan siswa berdasarkan kesamaan nilai akademik mereka.

Proses awal dilakukan dengan menggunakan metode Elbow (*Elbow Method*), yaitu teknik visual yang memanfaatkan nilai *inertia* untuk mengevaluasi seberapa baik data terbentuk dalam sejumlah klaster tertentu. Nilai *inertia* menggambarkan total jarak antara setiap titik data terhadap pusat klaster (*centroid*) masing-masing. Semakin kecil nilai *inertia*, semakin baik pengelompokan yang terbentuk.

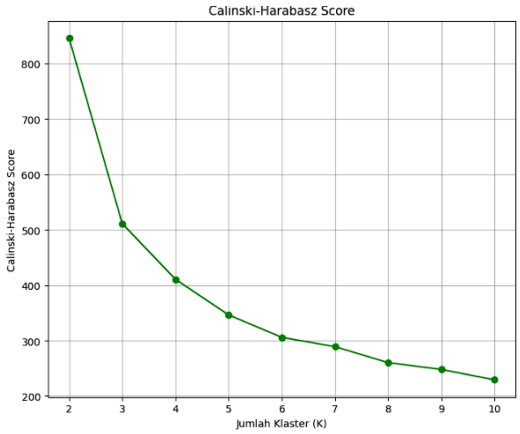
menampilkan grafik *Elbow Method* yang digunakan untuk menentukan jumlah klaster optimal berdasarkan nilai *inertia*. Metode ini merupakan bagian dari pendekatan dalam K-Means Clustering yang bertujuan untuk melihat seberapa besar penurunan variasi dalam klaster seiring dengan penambahan jumlah klaster (k). Pada grafik terlihat bahwa nilai *inertia* mengalami penurunan yang cukup signifikan dari k = 2 hingga k = 4, yang menunjukkan bahwa penambahan jumlah klaster pada rentang tersebut memberikan peningkatan kualitas pengelompokan yang cukup besar. Namun, setelah k = 4, penurunan nilai *inertia* cenderung lebih landai, yang mengindikasikan bahwa penambahan klaster tidak lagi memberikan peningkatan yang signifikan. Titik perubahan drastis (elbow) pada grafik ini berada di sekitar k = 3 atau k = 4, sehingga dapat diinterpretasikan bahwa jumlah klaster optimal berada pada rentang tersebut. Dengan demikian, nilai k = 2 dapat dipertimbangkan sebagai jumlah klaster yang optimal karena mampu memberikan keseimbangan antara kompleksitas model dan kualitas hasil clustering.

Berikut adahal hasil dari Penerapan *K-Means* pada Gambar 4.6



Gambar 2 Hasil Penerapan *K-Means*

gambar 4. 1 *Calinski-Harabasz Score*



Gambar 3 *Calinski-Harabasz Score*

Gambar tersebut menampilkan grafik evaluasi jumlah klaster (K) menggunakan metrik *Calinski-Harabasz Score* dalam konteks analisis clustering. Sumbu horizontal merepresentasikan jumlah klaster yang diuji (K = 2 hingga 10), sedangkan sumbu vertikal menunjukkan nilai skor *Calinski-Harabasz* yang diperoleh untuk masing-masing K. Secara konseptual, metrik ini mengukur rasio antara dispersi antar-klaster (*inter-cluster dispersion*) dan dispersi dalam klaster (*intra-cluster dispersion*), di mana nilai yang lebih tinggi mengindikasikan struktur klaster yang lebih baik dan terpisah secara optimal.

Berdasarkan grafik, terlihat bahwa nilai skor tertinggi terjadi pada K = 2, dengan nilai yang jauh lebih besar dibandingkan jumlah klaster lainnya. Setelah itu, skor mengalami penurunan yang cukup signifikan seiring bertambahnya jumlah klaster. Pola ini menunjukkan bahwa pemisahan data menjadi dua klaster memberikan struktur yang paling optimal menurut kriteria *Calinski-Harabasz*. Penurunan skor yang konsisten pada K yang lebih besar mengindikasikan bahwa penambahan klaster justru menyebabkan pembagian data menjadi kurang kompak dan kurang terpisah dengan baik.

Dengan demikian, secara akademis dapat disimpulkan bahwa jumlah klaster optimal dalam analisis ini adalah K = 2, karena memberikan keseimbangan terbaik antara homogenitas dalam klaster dan heterogenitas antar klaster. Namun demikian, interpretasi akhir tetap perlu mempertimbangkan konteks domain data dan tujuan analisis agar hasil clustering memiliki makna yang relevan secara praktis.

Tabel 1 Perbandingan

K	Inertia	Silhouette	Calinski
2	2488.65	0.5290	846.14
3	2204.35	0.3285	511.11
4	1939.86	0.2255	410.65
5	1782.44	0.2089	346.28
6	1655.49	0.2015	305.84

7	1503.35	0.2050	289.05
8	1445.71	0.1961	260.16
9	1350.24	0.2042	247.93
10	1307.21	0.1898	229.13

K optimal berdasarkan Silhouette: 2 Silhouette Score: 0.5290 Calinski-Harabasz Score: 846.14

Tabel tersebut menyajikan hasil evaluasi performa algoritma clustering (K-Means) dengan variasi jumlah klaster (K = 2 hingga 10) menggunakan tiga metrik utama, yaitu *Inertia*, *Silhouette Score*, dan *Calinski-Harabasz Index*. *Inertia* mengukur total jarak kuadrat antar data terhadap centroid klasternya, sehingga nilai yang lebih kecil menunjukkan klaster yang lebih kompak. *Silhouette Score* mengevaluasi sejauh mana suatu objek lebih dekat dengan klasternya dibandingkan dengan klaster lain, dengan rentang nilai -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan pemisahan klaster yang lebih baik. Sementara itu, *Calinski-Harabasz Index* mengukur rasio antara dispersi antar-klaster dan intra-klaster, dengan nilai yang lebih tinggi menunjukkan struktur klaster yang semakin optimal.

CONCLUSION

Berdasarkan hasil penelitian yang telah dilakukan, dapat ditarik tiga kesimpulan yang menjawab secara langsung rumusan masalah dan tujuan penelitian ini.

Pertama, algoritma *K-Means Clustering* berhasil diterapkan untuk mengelompokkan 535 siswa SMK Negeri 1 Majalaya berdasarkan nilai akademik dari 12 mata pelajaran secara objektif dan terukur. Penerapan dilakukan melalui tahapan sistematis yang meliputi akuisisi data, pra-pemrosesan (pembersihan data dan normalisasi menggunakan *StandardScaler*), serta penerapan algoritma *K-Means*. Hasil pengelompokan menghasilkan tiga *klaster* yang merepresentasikan kelompok siswa berprestasi cukup/belum optimal (Klaster 0, 322 siswa) dan berprestasi baik (Klaster 1, 214 siswa), sehingga menjawab rumusan masalah pertama mengenai bagaimana algoritma ini dapat mengelompokkan siswa berdasarkan nilai akademik dari 12 mata pelajaran secara objektif.

Kedua, jumlah *klaster* optimal ditetapkan sebanyak tiga *klaster* menggunakan metode *Elbow* yang menunjukkan titik siku pada k=2. Kualitas hasil pengelompokan diukur menggunakan dua metrik evaluasi, yaitu *Silhouette Score* sebesar 0,5290 dan *Calinski-Harabasz Index* sebesar 846,14. Nilai *Silhouette Score* tersebut menunjukkan bahwa pemisahan antar *klaster* cukup memadai meskipun terdapat kedekatan karakteristik pada sebagian data, sedangkan nilai *Calinski-Harabasz Index* yang cukup tinggi mengonfirmasi bahwa struktur *klaster* yang terbentuk sudah cukup kompak dan terpisah. Dengan demikian, model *klaster*isasi yang dihasilkan valid dan terukur, sehingga menjawab rumusan masalah kedua mengenai penentuan *klaster* optimal dan evaluasi kualitas pengelompokan.

Ketiga, hasil *klaster*isasi dapat dimanfaatkan secara langsung oleh pihak sekolah sebagai dasar perancangan strategi pembelajaran diferensiasi yang sesuai dengan karakteristik akademik masing-masing kelompok. Klaster 0 (berprestasi cukup/belum optimal) memiliki nilai rata-rata sebesar 72,3 dengan 322 siswa; Klaster 1 (berprestasi baik) memiliki nilai rata-rata sebesar 79,7 dengan 214 siswa. Bagi 322 siswa pada Klaster 0 (nilai rata-rata 72,3), guru disarankan merancang program remedial dan penguatan konsep dasar, terutama pada mata pelajaran Matematika (67,4) yang menjadi kelemahan utama, serta melakukan pendampingan personal secara berkala dan berkoordinasi dengan wali kelas untuk memantau perkembangan akademik siswa secara rutin. Bagi 214 siswa pada Klaster 1 (nilai rata-rata 79,7), guru disarankan memberikan program pengayaan (*enrichment*) berupa proyek akademik lanjutan, kompetisi ilmiah, dan penugasan mandiri bertingkat untuk mempertahankan motivasi serta mengembangkan kemampuan berpikir kritis dan kreatif. Kesimpulan ini menjawab rumusan masalah ketiga sekaligus memenuhi tujuan penelitian ketiga mengenai pemanfaatan hasil *klaster*isasi sebagai landasan strategi pembelajaran diferensiasi yang tepat sasaran.

REFERENCE

[1] Y. Zhang, T. Chen, and L. Wang, “A Comprehensive Review of Student Performance Clustering Techniques in Educational Data Mining,” *IEEE Access*, vol. 9, pp. 109845–109862, 2021, doi: 10.1109/ACCESS.2021.3101234.

[2] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, “Identification of student behavioral patterns in higher education using K-Means clustering and support vector machine,” *Applied Sciences*, vol. 13, no. 5, p. 3267, 2023, doi: 10.3390/app13053267.

[3] A. Al-Zawqari, D. Peumans, and G. Vandersteen, “A flexible feature selection approach for predicting students’ academic performance in online courses,” *Computers & Education: Artificial Intelligence*, vol. 3, p.

100103, 2022, doi: 10.1016/j.caeai.2022.100103.

[4] S. J. Alalawi, I. N. Mohd Shaharanee, and J. Mohd Jamil, "Clustering student performance data using K-Means algorithms," *Journal of Computational Innovation and Analytics (JCIA)*, vol. 2, no. 1, pp. 41–55, 2023, doi: 10.32890/jcia2023.2.1.3.

[5] D. A. N. Wulandari, R. Annisa, L. Yusuf, and T. Prihatin, "Educational data mining for student academic prediction using K-Means clustering and Naïve Bayes classifier," *Jurnal Pilar Nusa Mandiri*, vol. 16, no. 2, pp. 155–160, 2020, doi: 10.33480/pilar.v16i2.1432.

[6] I. J. on I. for Development, "K-means clustering for grouping student performance patterns in Social Studies," *International Journal on Informatics for Development*, 2024, doi: 10.14421/ijid.2024.4632.

[7] R. Nagaswetha, D. Narendar Singh, B. Pavitra, G. Ashwini, and G. Anil Kumar, "Data mining for student performance analysis by clustering K-Means algorithm," *Mathematical Statistician and Engineering Applications*, vol. 71, no. 3, pp. 1271–1287, 2022, doi: 10.17762/msea.v71i3.467.

[8] D. Andi Akram Nur Risal, D. Darma Andayani, M. I. Suherman, and A. Baso Kaswar, "Utilizing the K-Means clustering algorithm for analyzing student achievement assessment at SMK Negeri 1 Gowa," *Journal of Embedded Systems, Security and Intelligent Systems*, vol. 5, no. 1, pp. 60–67, 2024, doi: 10.59562/jessi.v5i1.2178.