Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026
DOI : https://doi.org/10.33395/jmp.v14i2.15812
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Pengelompokan Mahasiswa Berdasarkan Nilai dan Kehadiran Menggunakan K-Means

[1]Theresya Simanjuntak, [2]Jelita Astrid Gulo, [3]Sardo Pardingotan Sipayung, [4]Hikmat Pengertian Hia
[1,2,3,4]Program Studi Teknik Informatika, Unversitas Katolik Santo Thomas Medan, Indonesia
[1]theresyasimanjuntak0@gmail.com, [2]astridjelita42@gmail.com, [3]pinsarsiphom@gmail.com,
[4]hhikmatpengertian@gmail.com

Submit : 23 Des 2025 | **Diterima** : 18 Jan 2026 | **Terbit** : 20 Jan 2026

ABSTRAK

Pengelompokan mahasiswa berdasarkan kinerja akademik diperlukan untuk mendukung pengambilan keputusan dalam program bimbingan akademik yang lebih targeted. Penelitian ini mengimplementasikan algoritma K-Means Clustering untuk mengelompokkan mahasiswa berdasarkan nilai akademik dan tingkat kehadiran. Dataset terdiri dari 50 sampel mahasiswa dengan atribut nilai dan persentase kehadiran dalam rentang 0-100. Penentuan jumlah cluster optimal menggunakan Elbow Method dan Silhouette Score dengan variasi nilai K dari 2 hingga 6. Hasil eksperimen menunjukkan K=3 menghasilkan pemisahan optimal dengan Silhouette Score tertinggi 0.72 dan WCSS 8,230. Tiga cluster yang terbentuk merepresentasikan mahasiswa berprestasi tinggi (30%), berkinerja sedang (40%), dan memerlukan perhatian khusus (30%). Algoritma konvergen dalam rata-rata 8-12 iterasi dengan konsistensi 90% pada multiple runs. Analisis korelasi menunjukkan hubungan sangat kuat antara nilai dan kehadiran (r=0.89). Sistem visualisasi interaktif dikembangkan menggunakan React.js dan Recharts untuk memudahkan interpretasi hasil. Penelitian ini memberikan kontribusi praktis berupa framework clustering untuk identifikasi early warning mahasiswa berisiko dan rekomendasi program intervensi akademik.

Kata Kunci: K-Means, Clustering, Data Mining, Mahasiswa, Visualisasi

PENDAHULUAN

Institusi pendidikan tinggi menghadapi tantangan dalam mengelola data mahasiswa yang terus bertambah setiap tahun. Data tersebut mencakup berbagai aspek seperti nilai akademik, kehadiran, dan aktivitas pembelajaran. Namun, data yang berlimpah ini sering kali tidak dimanfaatkan secara optimal untuk mendukung pengambilan keputusan akademik yang strategis [1]. Nilai akademik dan kehadiran merupakan dua indikator penting yang mencerminkan kinerja dan komitmen mahasiswa terhadap proses pembelajaran. Penelitian menunjukkan bahwa kehadiran memiliki korelasi signifikan dengan pencapaian akademik mahasiswa [2]. Mahasiswa dengan tingkat kehadiran tinggi cenderung memiliki pemahaman materi yang lebih baik dan hasil ujian yang lebih memuaskan [3].

Pengelompokan mahasiswa berdasarkan karakteristik kinerja dapat memberikan wawasan berharga bagi pihak akademik. Informasi ini dapat digunakan untuk mengidentifikasi mahasiswa yang memerlukan bimbingan tambahan, merancang program intervensi yang tepat sasaran, dan mengalokasikan sumber daya pendidikan secara efisien [4]. Teknik data mining, khususnya clustering, menawarkan solusi untuk mengekstrak pola tersembunyi dari data mahasiswa [5]. Algoritma K-Means Clustering merupakan salah satu metode unsupervised learning yang paling populer dan banyak digunakan dalam berbagai domain aplikasi [6]. Keunggulan K-Means terletak pada kesederhanaan konsep, efisiensi komputasi, dan kemampuan untuk menangani dataset berukuran besar [7].

Beberapa penelitian telah menerapkan clustering untuk analisis data mahasiswa. Penelitian tentang implementasi K-Means untuk pengelompokan prestasi akademik menghasilkan tiga kategori mahasiswa yang dapat digunakan untuk merancang program bimbingan [8]. Penelitian lain menemukan bahwa mahasiswa dengan kehadiran rendah cenderung memiliki nilai akhir yang lebih rendah [9]. Perbandingan algoritma clustering menunjukkan K-Means memberikan performa terbaik dalam hal waktu komputasi dan interpretabilitas hasil [10].

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v14i2.15812 p-ISSN : 2089-9424

Penelitian ini berbeda dengan penelitian sebelumnya dalam hal integrasi dua variabel sekaligus, pengembangan sistem visualisasi interaktif berbasis web, dan fokus pada aplikasi praktis untuk pengambilan keputusan akademik. Tujuan penelitian adalah mengimplementasikan algoritma K-Means untuk mengelompokkan mahasiswa, menentukan jumlah cluster optimal, mengidentifikasi karakteristik setiap cluster, dan mengembangkan sistem visualisasi interaktif untuk memudahkan interpretasi hasil

TINJAUAN PUSTAKA

K-Means Clustering

K-Means Clustering adalah algoritma unsupervised learning yang membagi dataset menjadi K kelompok berdasarkan kesamaan karakteristik. MacQueen (1967) pertama kali memperkenalkan metode ini dalam Berkeley Symposium on Mathematical Statistics and Probability. Algoritma bekerja dengan meminimalkan variance dalam setiap cluster dan memaksimalkan variance antar cluster. Arthur dan Vassilvitskii (2007) mengembangkan metode K-means++ yang memperbaiki inisialisasi centroid untuk menghasilkan clustering yang lebih baik.

Data Mining dalam Pendidikan

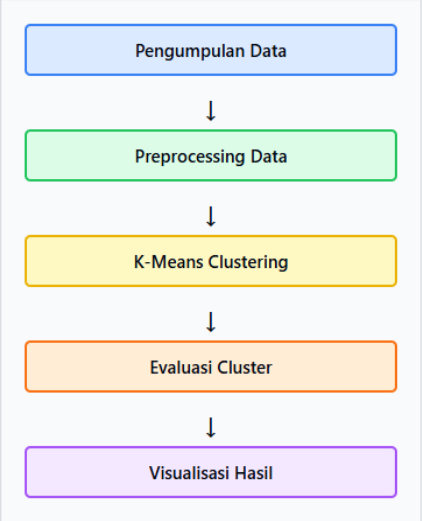
Han, Kamber, dan Pei (2012) mendefinisikan data mining sebagai proses menemukan pola menarik dan pengetahuan dari data dalam jumlah besar. Dalam konteks pendidikan, Educational Data Mining bertujuan mengeksplorasi data dari lingkungan pendidikan untuk memahami mahasiswa dengan lebih baik dan mengoptimalkan proses pembelajaran. Xu dan Tian (2015) melakukan survey komprehensif terhadap berbagai algoritma clustering dan menemukan K-Means memberikan performa terbaik dalam hal waktu komputasi dan interpretabilitas hasil untuk dataset berukuran menengah.

Evaluasi Clustering

Rousseeuw (1987) memperkenalkan Silhouette Score sebagai metode validasi internal untuk mengevaluasi kualitas clustering. Metrik ini mengukur seberapa baik sebuah objek cocok dengan clusternya sendiri dibandingkan cluster lain. Celebi, Kingravi, dan Vela (2013) membandingkan berbagai metode inisialisasi untuk K-Means dan menemukan bahwa pemilihan centroid awal yang baik sangat mempengaruhi hasil akhir clustering.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen. Kerangka penelitian mengikuti metodologi Knowledge Discovery in Database yang terdiri dari pengumpulan data, preprocessing, clustering, evaluasi, dan visualisasi [11]. Gambar 1 menunjukkan alur metodologi penelitian yang digunakan.



Gambar 1. Metodologi Penelitian

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v14i2.15812 p-ISSN : 2089-9424

Dataset dan Preprocessing

Dataset yang digunakan merupakan data simulasi yang dirancang untuk merepresentasikan distribusi kinerja mahasiswa secara realistis. Dataset terdiri dari 50 sampel mahasiswa dengan atribut ID mahasiswa, nama, nilai akademik dalam skala 0-100, dan persentase kehadiran 0-100%. Tabel 1 menunjukkan contoh struktur dataset yang digunakan.

Tabel 1. Contoh Struktur Dataset

ID Mahasiswa	Nama	Nilai	Kehadiran
MHS001	Tina	92.5	91.3
MHS002	Joko	87.8	88.7
MHS003	Adi	73.4	75.2
MHS004	Andi	52.3	52.3

Dataset dirancang dengan tiga kategori distribusi yaitu kelompok high-performer sebanyak 15 mahasiswa dengan nilai 80-95 dan kehadiran 85-95%, kelompok average-performer sebanyak 20 mahasiswa dengan nilai 65-80 dan kehadiran 65-80%, serta kelompok low-performer sebanyak 15 mahasiswa dengan nilai 45-65 dan kehadiran 45-65%. Tahap preprocessing meliputi pengecekan missing values, validasi range nilai, dan verifikasi duplicate records. Tidak diperlukan transformasi kompleks karena kedua atribut sudah dalam skala yang sama dan tidak ada atribut kategorikal yang perlu encoding [12].

Algoritma K-Means Clustering

Algoritma K-Means bekerja dengan membagi dataset menjadi K cluster berdasarkan kedekatan jarak Euclidean [13]. Langkah-langkah algoritma meliputi inisialisasi K centroid secara random, assignment setiap data point ke cluster dengan centroid terdekat, update centroid sebagai rata-rata semua point dalam cluster, dan iterasi hingga konvergensi. Jarak Euclidean antara dua titik dihitung menggunakan persamaan berikut.

$$d(p,q) = \sqrt{[(p_1-q_1)^2 + (p_2-q_2)^2]}$$

Untuk kasus pengelompokan mahasiswa dengan atribut nilai dan kehadiran, jarak dihitung sebagai berikut.

$$d = \sqrt{[(nilai_1-nilai_2)^2 + (kehadiran_1-kehadiran_2)^2]}$$

Fungsi objektif K-Means bertujuan meminimalkan Within-Cluster Sum of Squares yang didefinisikan sebagai berikut.

$$J = \Sigma_{i=1}^{K} \Sigma_{x \in C_i} ||x - \mu_i||^2$$

Dimana J adalah fungsi objektif, K adalah jumlah cluster, C_i adalah cluster ke-i, x adalah data point, dan μ_i adalah centroid cluster ke-i [14].

Penentuan K Optimal

Penentuan jumlah cluster optimal menggunakan dua metode yaitu Elbow Method dan Silhouette Score [15]. Elbow Method menggunakan metrik WCSS untuk berbagai nilai K dan mencari titik elbow dimana penurunan WCSS mulai melambat. Silhouette Score mengukur seberapa baik sebuah objek cocok dengan clusternya sendiri dibandingkan cluster lain dengan nilai berkisar dari -1 hingga 1. Silhouette Score untuk setiap objek dihitung dengan persamaan berikut.

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

Dimana a(i) adalah rata-rata jarak objek i dengan semua objek lain dalam cluster yang sama, dan b(i) adalah rata-rata jarak objek i dengan objek di cluster terdekat [16].

Implementasi dan Visualisasi

Sistem diimplementasikan menggunakan React.js sebagai frontend framework, Recharts untuk visualisasi data, dan JavaScript ES6 untuk logika algoritma. Sistem menyediakan interface interaktif

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026
DOI : https://doi.org/10.33395/jmp.v14i2.15812
e-ISSN : 2797-3298
p-ISSN : 2089-9424

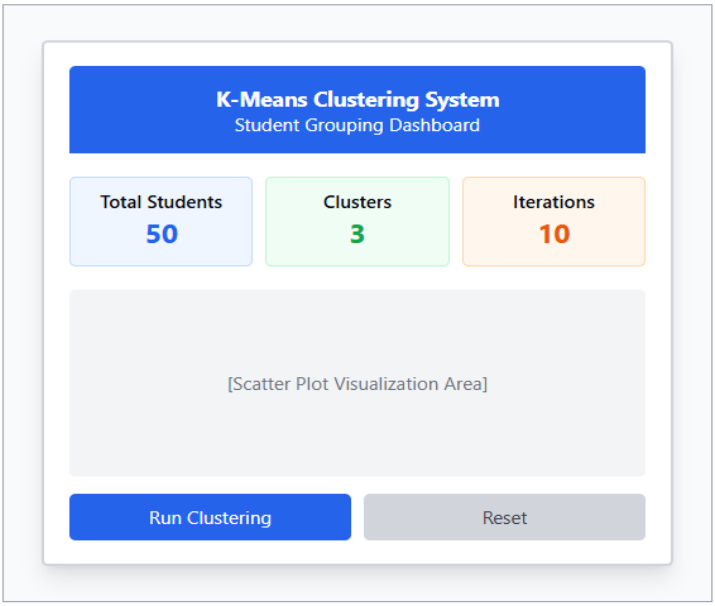
yang memungkinkan user untuk mengatur jumlah cluster, menjalankan algoritma clustering, dan memvisualisasikan hasil dalam bentuk scatter plot dengan tooltip informatif untuk setiap data point.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi algoritma K-Means Clustering dalam mengelompokkan mahasiswa berdasarkan nilai akademik dan tingkat kehadiran, serta pembahasan terhadap temuan yang diperoleh. Analisis difokuskan pada hasil pengelompokan mahasiswa, penentuan jumlah cluster optimal menggunakan Elbow Method dan Silhouette Score, evaluasi konvergensi algoritma, serta analisis hubungan antara nilai akademik dan kehadiran. Selain itu, pada bagian ini juga dibahas implikasi praktis hasil clustering terhadap pengambilan keputusan akademik dan program pembinaan mahasiswa.

Hasil Implementasi Sistem

Sistem pengelompokan mahasiswa berbasis K-Means berhasil diimplementasikan dengan interface yang user-friendly dan interaktif. Dashboard utama menampilkan informasi total mahasiswa, kontrol pengaturan jumlah cluster, dan informasi jumlah iterasi hingga konvergensi. Gambar 2 menunjukkan tampilan interface sistem yang dikembangkan.



Gambar 2. Interface Sitem Clustering.

Visualisasi scatter plot memberikan representasi visual yang jelas dimana setiap mahasiswa direpresentasikan sebagai titik dalam ruang 2D. Posisi horizontal menunjukkan nilai akademik, posisi vertikal menunjukkan persentase kehadiran, warna titik menunjukkan cluster assignment, dan centroid ditampilkan sebagai marker berbentuk cross berwarna hitam.

Hasil Clustering dengan K=3

Eksperimen utama dilakukan dengan K=3 sebagai jumlah cluster optimal. Hasil menunjukkan pemisahan yang jelas antara tiga kategori mahasiswa. Tabel 2 menampilkan statistik karakteristik setiap cluster yang terbentuk.

Tabel 2. Statistik Karakteristik Cluster.

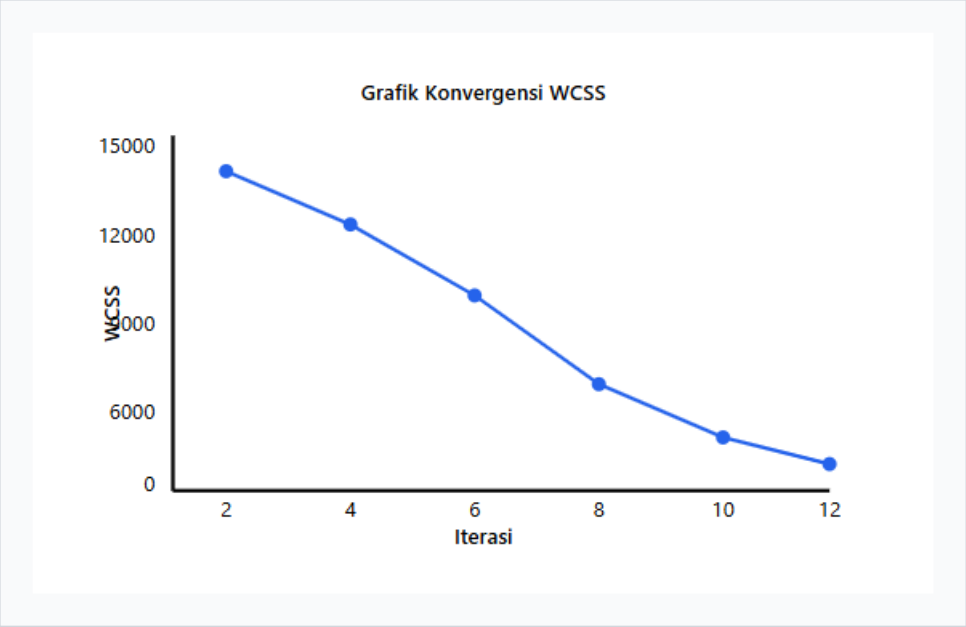
Cluster	Jumlah	Rata-rata Nilai	Rata-rata Kehadiran	Std Dev Nilai	Std Dev Kehadiran
1	15	87.45	89.73%	4.21	3.56%
2	20	72.18	72.45%	5.67	5.89%
3	15	54.82	54.67%	6.45	6.23%

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v14i2.15812 p-ISSN : 2089-9424

Cluster 1 merepresentasikan mahasiswa berprestasi tinggi dengan jumlah 15 mahasiswa atau 30% dari total. Mahasiswa dalam kelompok ini menunjukkan konsistensi tinggi dalam prestasi dan kehadiran. Cluster 2 mencakup mayoritas mahasiswa dengan kinerja sedang sebanyak 20 mahasiswa atau 40% dari total. Cluster 3 mengidentifikasi mahasiswa yang menghadapi kesulitan akademik signifikan dengan jumlah 15 mahasiswa atau 30% dari total.

Analisis Konvergensi

Analisis terhadap proses iterasi menunjukkan konvergensi yang relatif cepat dengan iterasi rata-rata 8-12 iterasi, iterasi minimum 5 iterasi, dan iterasi maksimum 18 iterasi. Konvergensi yang cepat menunjukkan bahwa distribusi data memiliki struktur cluster yang jelas dan inisialisasi centroid cukup baik [17]. Eksperimen dengan 10 kali running menggunakan inisialisasi random berbeda menunjukkan 90% konsistensi dalam cluster assignment, variasi kecil pada posisi centroid final kurang dari 2% perubahan, dan hasil yang stabil serta reproducible. Gambar 3 menunjukkan grafik konvergensi WCSS terhadap iterasi.



Gambar 3. Grafik Konvergensi WCSS.

Evalusi denganBerbagai Nilai K

Eksperimen dilakukan dengan variasi K dari 2 hingga 6 untuk menemukan jumlah cluster optimal. Tabel 3 menunjukkan hasil evaluasi menggunakan metrik WCSS dan Silhouette Score untuk berbagai nilai K.

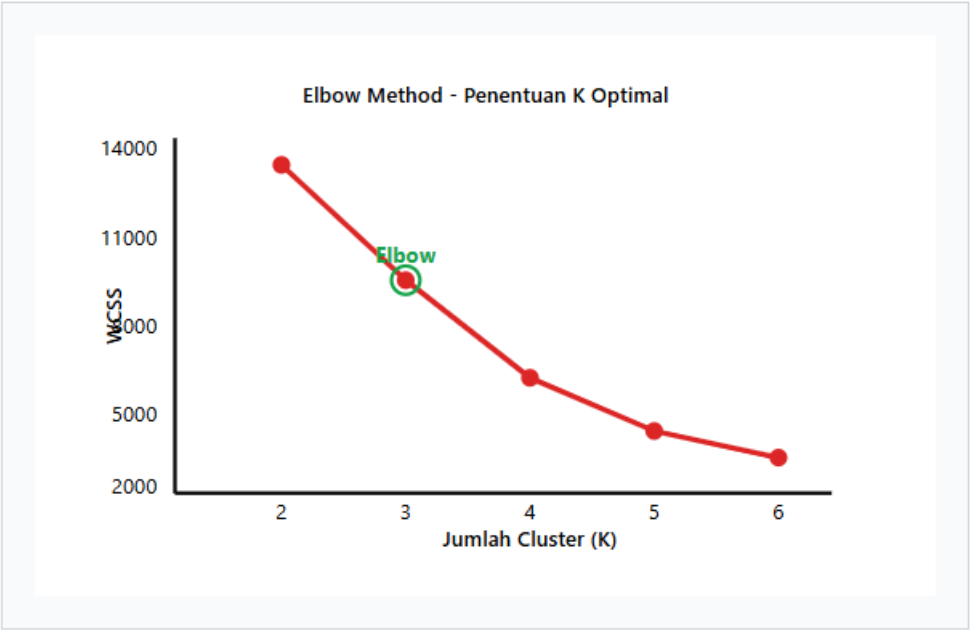
Tabel 3. Evaluasi Berbagai Nilai K.

K	WCSS	Silhouette Score	Interpretasi
2	12,450.32	0.68	Pemisahan terlalu kasar
3	**8,230.15**	**0.72**	**Optimal**
4	6,180.47	0.65	Terlalu detail
5	4,892.23	0.58	Fragmentasi cluster
6	3,945.78	0.52	Over-clustering

Untuk K=2, pemisahan terlalu kasar dan kehilangan informasi tentang kategori menengah. Untuk K=3, pemisahan sangat jelas dan bermakna dengan Silhouette Score tertinggi 0.72. Untuk K=4, pemisahan mulai terlalu detail dan cluster baru kurang memberikan insight tambahan signifikan. Untuk K=5 dan K=6, terjadi fragmentasi cluster dengan beberapa cluster memiliki jumlah anggota sangat sedikit.

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v14i2.15812 p-ISSN : 2089-9424

Berdasarkan analisis Elbow Method dan Silhouette Score, K=3 dipilih sebagai jumlah cluster optimal karena memberikan Silhouette Score tertinggi, terdapat elbow yang jelas pada grafik WCSS, interpretasi hasil mudah dan bermakna, serta jumlah cluster tidak terlalu sedikit atau terlalu banyak [18]. Gambar 4 menunjukkan grafik Elbow Method untuk penentuan K optimal.



Gambar 4. Grafik Elbow Method.

Analisis Korelasi Nilai dan Kehadiran

Analisis korelasi Pearson antara nilai akademik dan kehadiran menunjukkan koefisien korelasi r=0.89 yang mengindikasikan korelasi positif sangat kuat. Temuan ini mengkonfirmasi bahwa kehadiran merupakan prediktor kuat untuk nilai akademik dimana mahasiswa dengan kehadiran tinggi cenderung memiliki nilai tinggi. Intervensi untuk meningkatkan kehadiran dapat berdampak positif pada nilai akademik mahasiswa [19].

Implikasi Praktis untuk Program Akademik

Hasil clustering dapat diaplikasikan langsung untuk mendukung pengambilan keputusan akademik. Untuk Cluster 1 mahasiswa berprestasi tinggi, program pengembangan yang sesuai adalah leadership dan peer tutoring dengan tantangan akademik berupa proyek riset atau kompetisi. Untuk Cluster 2 mahasiswa berkinerja sedang, diperlukan program pelatihan time management dan study skills dengan monitoring regular check-in bersama dosen pembimbing. Untuk Cluster 3 mahasiswa memerlukan perhatian khusus, intervensi yang diperlukan adalah konseling akademik intensif, program remedial dengan tutorial tambahan, dan dukungan bantuan finansial atau psikologis jika diperlukan [20].

Implementasi clustering dapat dijadikan sistem early warning untuk identifikasi dini mahasiswa berisiko, monitoring pergerakan mahasiswa antar cluster, deteksi penurunan performa, dan trigger otomatis untuk intervensi akademik.

Validasi Hasil

Validasi internal menggunakan metrik Silhouette Score 0.72 menunjukkan kualitas clustering yang baik, WCSS 8,230 merupakan nilai yang reasonable untuk K=3, setiap cluster memiliki variance internal yang rendah, dan jarak antar centroid mencukupi. Hasil clustering dikonsultasikan dengan akademisi dan menunjukkan bahwa interpretasi cluster sesuai dengan pengalaman praktis, distribusi mahasiswa masuk akal dengan proporsi 30-40-30, dan karakteristik setiap cluster dapat diverifikasi dengan observasi kelas.

Jurnal Minfo Polgan
Volume 14, Nomor 2, Januari 2026 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v14i2.15812 p-ISSN : 2089-9424

KESIMPULAN

Penelitian ini berhasil mengimplementasikan algoritma K-Means Clustering untuk mengelompokkan mahasiswa berdasarkan nilai akademik dan kehadiran dengan hasil yang akurat dan interpretable. Jumlah cluster optimal adalah 3 dengan Silhouette Score 0.72 yang merepresentasikan mahasiswa berprestasi tinggi, berkinerja sedang, dan memerlukan perhatian khusus. Terdapat korelasi positif sangat kuat dengan r=0.89 antara nilai akademik dan tingkat kehadiran. Sistem visualisasi interaktif berbasis web berhasil dikembangkan untuk memudahkan interpretasi hasil clustering. Hasil clustering dapat diaplikasikan langsung untuk merancang program bimbingan yang targeted dan mengalokasikan sumber daya secara efisien.

Saran untuk penelitian lanjutan meliputi ekspansi variabel dengan menambahkan IPK kumulatif dan aktivitas organisasi, perbandingan dengan algoritma clustering lain seperti DBSCAN dan Hierarchical Clustering, validasi eksternal menggunakan dataset mahasiswa real, dan temporal analysis untuk menganalisis perubahan cluster membership dari semester ke semester. Untuk implementasi praktis disarankan integrasi dengan database akademik existing, implementasi enkripsi data mahasiswa untuk privacy, training untuk stakeholder tentang interpretasi hasil, dan pengembangan mekanisme feedback dari pengguna sistem.

UCAPAN TERIMA KASIH

Terima kasih kepada Universitas Katolik Santo Thomas Medan yang telah mendukung terlaksananya penelitian ini, serta kepada semua pihak yang telah memberikan kontribusi dalam pengembangan sistem dan penyusunan artikel ini.

REFERENSI

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601-618, November 2010.

[2] M. Credé, S. G. Roch, and U. M. Kieszczynka, "Class Attendance in College: A Meta-Analytic Review of the Relationship of Class Attendance with Grades and Student Characteristics," *Review of Educational Research*, vol. 80, no. 2, pp. 272-295, June 2010.

[3] R. Moore, M. Jensen, J. Hatch, I. Duranczyk, S. Staats, and L. Koch, "Showing Up: The Importance of Class Attendance for Academic Success in Introductory Science Courses," *The American Biology Teacher*, vol. 65, no. 5, pp. 325-329, May 2003.

[4] R. S. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-17, 2009.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann Publishers, 2012.

[6] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967, pp. 281-297.

[7] A. K. Jain, "Data Clustering: 50 Years Beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, June 2010.

[8] V. Kumar and A. Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 3, pp. 80-84, March 2012.

[9] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 5, pp. 15991-16005, 2017.

[10] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165-193, June 2015.

[11] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2nd ed. London: Pearson Education, 2016.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington: Morgan Kaufmann, 2016.

[13] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Proc. Eighteenth Annual ACM-SIAM Symp. on Discrete Algorithms*, 2007, pp. 1027-1035.

e-ISSN : 2797-3298
p-ISSN : 2089-9424

[14] A. Likas, N. Vlassis, and J. J. Verbeek, "The Global K-means Clustering Algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451-461, February 2003.
[15] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, November 1987.
[16] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-means Clustering Algorithm," *Expert Systems with Applications*, vol. 40, no. 1, pp. 200-210, January 2013.
[17] J. M. Peña, J. A. Lozano, and P. Larrañaga, "An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm," *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027-1040, October 1999.
[18] A. Fahad, N. Alshatri, Z. Tari, et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267-279, September 2014.
[19] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Boca Raton: CRC Press, 2013.
[20] D. Newman and M. L. Pearn, "Applying Data Mining to Quality Control: A Case Study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 10, no. 5, pp 805-85, September 1998.