E-ISSN: 3025-0994
Halaman 4267-4279
Volume 4 Nomor 4 Tahun 2026

Penerapan Algoritma *K-Means Clustering* Untuk Segmentasi Siswa Berdasarkan Perilaku Belajar Pada *Platform E-Learning* Skillora

Madhuri Lailatul Hamidah[1], Revian Atheylla Liazmi[2], Febti Sofia Loren[3], Risqiyah Hakim[4], Ramadhan Cakra Wibawa[5]

[1,2,3,4,5] Program Studi Pendidikan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Surabaya, Indonesia
e-mail: [1]madhuri.23057@mhs.unesa.ac.id* [2]revian.23054@mhs.unesa.ac.id
[3]febti.23058@mhs.unesa.ac.id [4]rizqiyah.23071@mhs.unesa.ac.id
[5]ramadhanwibawa@unesa.ac.id

Abstrak

Platform e-learning menghasilkan data pendaftaran kursus dan hasil kuis dalam jumlah besar, tetapi pengelola sering memperlakukan seluruh siswa sebagai satu kelompok yang seragam. Penelitian ini menerapkan algoritma K-Means Clustering untuk mengelompokkan siswa platform Skillora berdasarkan perilaku belajarnya. Data yang dipakai meliputi 3.000 siswa, 10.380 pendaftaran kursus, dan 30.875 hasil kuis periode Januari 2022 sampai Maret 2025. Setelah tahap seleksi dan pembersihan, 2.536 siswa yang pernah mendaftar kursus dianalisis dengan empat fitur, yaitu jumlah kursus yang diambil, rata-rata persentase penyelesaian, rata-rata nilai kuis, dan tingkat kelulusan kuis. Fitur dinormalisasi dengan metode min-max, centroid awal dipilih dengan k-means++, dan jarak diukur dengan Euclidean distance. Jumlah klaster dievaluasi dengan metode Elbow dan Silhouette Coefficient untuk k = 2 sampai 8. Dengan k = 4 diperoleh empat klaster yang masing-masing beranggotakan 787, 849, 568, dan 332 siswa. Klaster 1 memiliki rata-rata penyelesaian kursus tertinggi (79,0%), klaster 3 memiliki jumlah kursus per siswa tertinggi (8,2 kursus), dan klaster 4 memiliki tingkat kelulusan kuis terendah (4,4%). Nilai Silhouette sebesar 0,298 menunjukkan struktur klaster yang lemah sampai sedang, sehingga hasil pengelompokan dipakai sebagai alat bantu analisis, bukan pemisahan yang tegas. Uji chi-square menunjukkan keanggotaan klaster tidak berhubungan dengan paket langganan, sumber pendaftaran, maupun kelompok usia (p > 0,05).

Kata Kunci: *K-Means Clustering, Segmentasi Siswa, Perilaku Belajar, E-Learning, Silhouette Coefficient*

Abstract

E-learning platforms generate large volumes of course enrollment and quiz data, yet administrators often treat all students as one homogeneous group. This study applies the K-Means clustering algorithm to group students of the Skillora platform based on their learning

behavior. The dataset contains 3,000 students, 10,380 course enrollments, and 30,875 quiz results recorded between January 2022 and March 2025. After selection and cleaning, 2,536 students who had enrolled in at least one course were analyzed using four features: number of enrolled courses, mean completion percentage, mean quiz score, and quiz pass rate. The features were min-max normalized, initial centroids were chosen with k-means++, and distance was measured with the Euclidean distance. The number of clusters was evaluated with the Elbow method and the Silhouette Coefficient for k = 2 to 8. With k = 4, four clusters were obtained, containing 787, 849, 568, and 332 students. Cluster 1 had the highest mean completion (79.0%), Cluster 3 had the highest number of courses per student (8.2 courses), and Cluster 4 had the lowest quiz pass rate (4.4%). The Silhouette value of 0.298 indicates weak to moderate cluster structure, so the result is best used as an analytical aid rather than a sharp partition. Chi-square tests show that cluster membership is unrelated to subscription plan, acquisition source, or age group (p > 0.05).

Keywords: *K-Means Clustering, Student Segmentation, Learning Behavior, E-Learning, Silhouette Coefficient*

PENDAHULUAN

Pembelajaran daring kini menghasilkan jejak data yang rinci, mulai dari pendaftaran kursus, progres penyelesaian materi, sampai hasil kuis setiap siswa. Data tersebut menjadi bahan utama *educational data mining* dan *learning analytics*, yaitu bidang yang memakai teknik penggalian data untuk memahami dan memperbaiki proses belajar (Romero & Ventura, 2020). Tantangannya, pengelola platform sering membaca data itu sebagai angka agregat, misalnya jumlah pendaftar atau rata-rata nilai, sehingga perbedaan perilaku antarsiswa tidak terlihat.

Angka agregat memang dapat menyesatkan. Jordan (2015) melaporkan bahwa tingkat penyelesaian 221 *Massive Open Online Course* (MOOC) berkisar antara 0,7% dan 52,1% dengan median 12,6%. Kizilcec dkk. (2013) menunjukkan bahwa satu angka penyelesaian menyembunyikan beberapa pola keterlibatan yang berbeda: siswa yang menuntaskan kursus, yang hanya mengaudit, yang berhenti di tengah jalan, dan yang sekadar mencoba. Pola yang berbeda menuntut perlakuan yang berbeda pula, sehingga pengelompokan siswa (segmentasi) menjadi langkah awal yang penting.

Algoritma *K-Means* adalah salah satu metode *clustering* yang paling sering dipakai untuk segmentasi semacam ini karena sederhana dan efisien pada data berukuran besar (Han dkk., 2012; MacQueen, 1967). Moubayed dkk. (2020) memakai *K-Means* pada 12 metrik keterlibatan siswa dan menemukan bahwa model tiga tingkat paling baik mengenali siswa dengan keterlibatan rendah. Di Indonesia, Safitri dkk. (2022) mengelompokkan siswa dari 11.139 catatan log sistem pembelajaran daring menjadi tiga klaster perilaku, sedangkan Evans dan Nataliani (2024) membagi 215 siswa SMP menjadi tiga klaster aktivitas dari log Moodle. Kedua penelitian itu memakai log aktivitas pada satu mata pelajaran atau satu sistem pembelajaran.

Skillora adalah platform belajar data (SQL, Python, Excel, Power BI, *Statistics*, dan *Data Engineering*) yang membuka beberapa kursus sekaligus. Data platform ini mencatat 3.000 siswa, 10.380 pendaftaran pada 45 kursus berstatus terbit, dan 30.875 hasil kuis sepanjang Januari 2022 sampai Maret 2025. Dari seluruh pendaftaran, hanya 3.188 (30,7%) yang sampai selesai, dan 464 siswa terdaftar tidak pernah mengambil satu pun kursus. Angka-angka ini memperlihatkan bahwa siswa Skillora tidak homogen, tetapi pengelola belum punya dasar untuk menentukan siapa yang perlu pendampingan, siapa yang sudah mandiri, dan siapa yang perlu ditawari kursus lanjutan.

Berbeda dari penelitian terdahulu yang bertumpu pada log aktivitas satu kursus (Evans & Nataliani, 2024; Safitri dkk., 2022), penelitian ini memakai data tingkat platform, yaitu gabungan pendaftaran lintas kursus dan hasil kuis, serta menguji apakah klaster yang terbentuk berkaitan dengan karakteristik demografis dan sumber pendaftaran siswa. Penelitian ini difokuskan pada penerapan algoritma *K-Means Clustering*, yaitu penentuan jumlah klaster yang sesuai dan evaluasi kualitas klaster yang terbentuk. Berdasarkan latar belakang tersebut, rumusan masalah penelitian ini adalah sebagai berikut.

1. Bagaimana menerapkan algoritma *K-Means Clustering* untuk mengelompokkan siswa Skillora berdasarkan perilaku belajar?
2. Berapa jumlah klaster yang sesuai, dan seberapa baik kualitas klaster yang terbentuk?
3. Seperti apa karakteristik tiap klaster, dan apakah keanggotaan klaster berkaitan dengan paket langganan, sumber pendaftaran, dan usia siswa?

Tujuan penelitian ini adalah (1) menerapkan *K-Means* untuk mengelompokkan siswa berdasarkan perilaku belajar, (2) mengevaluasi jumlah dan kualitas klaster dengan metode *Elbow* dan *Silhouette Coefficient*, dan (3) menggambarkan karakteristik tiap klaster serta hubungannya dengan karakteristik siswa. Penelitian dibatasi pada siswa yang pernah mendaftar kursus dan pada empat fitur perilaku belajar. Data aktivitas harian seperti durasi menonton atau jumlah *login* tidak tersedia pada *dataset* ini.

KAJIAN PUSTAKA

Data Mining dan Knowledge Discovery in Databases

Data mining adalah proses menemukan pola yang sah, baru, berguna, dan dapat dipahami dari data dalam jumlah besar. Fayyad dkk. (1996) menempatkannya sebagai satu tahap dalam proses *Knowledge Discovery in Databases* (KDD) yang terdiri atas seleksi data, *preprocessing*, transformasi, *data mining*, dan interpretasi. Han dkk. (2012) mengelompokkan teknik *data mining* menjadi pembelajaran terawasi (klasifikasi dan regresi) dan tak terawasi (*clustering*). Pada *clustering*, tidak ada label kelas yang harus ditebak, sehingga kelompok muncul dari kemiripan data itu sendiri.

Algoritma K-Means Clustering

K-Means membagi *n* objek ke dalam *k* klaster dengan meminimalkan jumlah kuadrat jarak setiap objek ke pusat (*centroid*) klasternya. Istilah *k-means* diperkenalkan oleh MacQueen (1967), dan prosedur iteratif yang kini umum dipakai dikenal sebagai algoritma Lloyd (1982). Jain (2010) menilai algoritma ini bertahan lebih dari lima puluh tahun karena mudah diimplementasikan dan cepat, tetapi sensitif terhadap pemilihan *centroid* awal dan skala fitur. Langkah algoritmanya adalah sebagai berikut.

1. Tentukan jumlah klaster *k*.
2. Pilih *k centroid* awal.
3. Hitung jarak setiap objek ke setiap *centroid* dengan jarak *Euclidean*, lalu masukkan objek ke klaster terdekat.
4. Hitung ulang *centroid* sebagai rata-rata anggota klaster.
5. Ulangi langkah 3 dan 4 sampai keanggotaan klaster tidak berubah atau iterasi maksimum tercapai.

Jarak *Euclidean* antara objek *x* dan *centroid c* pada *d* fitur dihitung dengan Persamaan (1), pembaruan *centroid* klaster C_j dengan Persamaan (2), dan fungsi tujuan berupa *Sum of Squared Errors* (SSE) dengan Persamaan (3).

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

$$c_j = \frac{1}{|C_j|}\sum_{x \in C_j} x \qquad (2)$$

$$SSE = \sum_{j=1}^{k}\sum_{x \in C_j} d(x,c_j)^2 \qquad (3)$$

Inisialisasi k-means++ dan Normalisasi Fitur

Pemilihan *centroid* awal secara acak dapat menghasilkan klaster yang buruk. Arthur dan Vassilvitskii (2007) mengusulkan *k-means++*, yaitu memilih *centroid* berikutnya dengan peluang sebanding dengan kuadrat jaraknya ke *centroid* terdekat yang sudah ada, sehingga *centroid* awal tersebar dan hasil akhirnya terjamin kompetitif terhadap solusi optimal.

Karena jarak *Euclidean* dipengaruhi skala, fitur dengan rentang besar akan mendominasi fitur berrentang kecil. Milligan dan Cooper (1988) menunjukkan bahwa cara standardisasi variabel memengaruhi hasil *clustering*. Penelitian ini memakai normalisasi *min-max* pada Persamaan (4) agar semua fitur berada pada rentang 0 sampai 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (4)$$

Evaluasi Klaster

Metode Elbow. Thorndike (1953) mengamati bahwa penambahan klaster akan terus menurunkan SSE, tetapi penurunannya melandai setelah jumlah klaster yang wajar. Titik tekukan pada grafik SSE terhadap k dipakai sebagai kandidat jumlah klaster. Penentuan titik tekuk bersifat subjektif sehingga perlu dikombinasikan dengan ukuran lain.

Silhouette Coefficient. Rousseeuw (1987) mendefinisikan nilai *silhouette* objek *i* pada Persamaan (5), dengan *a(i)* rata-rata jarak *i* ke anggota klasternya sendiri dan *b(i)* rata-rata jarak *i* ke anggota klaster terdekat lain. Nilainya berada pada rentang -1 sampai 1. Kaufman dan Rousseeuw (1990) menafsirkan rata-rata 0,51 sampai 0,70 sebagai struktur

yang masuk akal, 0,26 sampai 0,50 sebagai struktur yang lemah, dan di bawah 0,25 sebagai struktur yang tidak berarti.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (5)$$

Davies-Bouldin Index (DBI). DBI mengukur rata-rata kemiripan setiap klaster dengan klaster yang paling mirip dengannya. Semakin kecil nilainya, semakin terpisah klasternya (Davies & Bouldin, 1979).

Penelitian Terkait

Tabel 1 merangkum penelitian yang memakai *K-Means* pada data perilaku siswa daring. Sebagian besar memakai log aktivitas satu kursus atau satu sistem pembelajaran, dan jumlah klaster yang dipilih berkisar antara tiga dan lima.

Tabel 1. Ringkasan penelitian terkait.

Peneliti	Data	Metode dan evaluasi	Hasil utama
Kizilcec dkk. (2013)	Tiga MOOC Stanford (hingga 46.096 siswa aktif per kursus)	*K-Means* pada label keterlibatan tiap periode, uji *silhouette*	Empat lintasan: *completing*, *auditing*, *disengaging*, *sampling*
Moubayed dkk. (2020)	Log kursus sains blended dengan 12 metrik keterlibatan	*K-Means* dengan model 2, 3, dan 5 tingkat, *silhouette*	Model dua tingkat terpisah terbaik, sedangkan model tiga tingkat lebih baik mendeteksi keterlibatan rendah
Safitri dkk. (2022)	11.139 catatan log SPADA UNS	*K-Means* dilanjutkan *decision tree*, DBI	Tiga klaster (DBI 0,229): keterlibatan rendah, tinggi, dan rajin mengumpulkan tugas
Evans dan Nataliani (2024)	Log Moodle 215 siswa SMP	*K-Means*	Tiga klaster aktivitas: rendah 138, sedang 62, tinggi 15 siswa
Ratnapala dkk. (2014)	Log Moodle 412 mahasiswa teknik, dua mata kuliah	*K-Means*, k dipilih dari SSE untuk k = 2 sampai 6	Lima klaster, dengan sekitar 41% sampai 42% siswa berada pada klaster paling tidak aktif

METODE PENELITIAN

Desain Penelitian

Penelitian ini bersifat kuantitatif dengan pendekatan *data mining* tak terawasi. Prosesnya mengikuti lima tahap KDD (Fayyad dkk., 1996), yaitu seleksi data, *preprocessing*, transformasi, penerapan *K-Means*, dan interpretasi hasil.

Sumber Data

Data berasal dari basis data platform Skillora dalam bentuk lima tabel yang saling terhubung melalui student_id, course_id, dan enrollment_id. Pendaftaran kursus dan hasil kuis tercatat pada periode 14 Januari 2022 sampai 15 Maret 2025.

Tabel 2. Struktur dataset penelitian.

Tabel	Jumlah catatan	Atribut utama yang dipakai
students	3.000	usia, kota, tanggal registrasi, sumber pendaftaran, paket langganan
courses	50 (45 terbit, 5 draf)	kategori, tingkat, harga, durasi
instructors	30	keahlian, kota
enrollments	10.380	tanggal daftar, persentase penyelesaian, sertifikat
quiz_results	30.875	nomor kuis, percobaan, skor, status lulus

Tahapan KDD

Seleksi data. Dari 3.000 siswa terdaftar, hanya 2.536 siswa (84,5%) yang memiliki minimal satu pendaftaran kursus. Sebanyak 464 siswa lainnya tidak memiliki data perilaku belajar sehingga dikeluarkan dari proses *clustering*. Hasil kuis dari 2.457 siswa dipakai untuk menghitung nilai dan tingkat kelulusan.

Preprocessing. Pengecekan menunjukkan tidak ada nilai kosong pada tabel students dan quiz_results, dan semua kunci identitas (student_id, enrollment_id) unik. Kolom completion_date kosong pada 7.192 pendaftaran. Kondisi ini sah karena menandakan kursus belum selesai, bukan data hilang. Sebanyak 79 siswa terdaftar belum pernah mengerjakan kuis, sehingga nilai rata-rata dan tingkat lulusnya diberi nilai 0 agar siswa tersebut tetap dapat dikelompokkan.

Transformasi. Data tingkat pendaftaran dan tingkat kuis diringkas menjadi satu baris per siswa dengan empat fitur perilaku belajar pada Tabel 3. Setiap fitur kemudian dinormalisasi dengan Persamaan (4) memakai nilai minimum dan maksimum dari data.

Tabel 3. Fitur perilaku belajar dan statistik deskriptifnya (n = 2.536 siswa).

Fitur	Definisi	Rata-rata	Simpangan baku	Min	Maks
enroll_count	Jumlah pendaftaran kursus	4,09	2,71	1	11
avg_completion	Rata-rata persentase penyelesaian (%)	54,38	25,29	0	100
avg_score	Rata-rata skor kuis (0 sampai 100)	67,70	15,50	0	97
pass_rate	Persentase kuis yang lulus (%)	55,81	27,21	0	100

Empat fitur ini dipilih karena menggambarkan tiga sisi perilaku siswa, yaitu seberapa banyak ia mengambil kursus (*enroll_count*), seberapa jauh ia menyelesaikannya (*avg_completion*), dan seberapa baik hasil kuisnya (*avg_score* dan *pass_rate*). Keempat fitur inilah yang menjadi masukan algoritma *K-Means*. Algoritma tidak menerima label kelompok apa pun, sehingga pengelompokan sepenuhnya ditentukan oleh kemiripan nilai keempat fitur tersebut.

Penerapan K-Means dan Pemilihan Jumlah Klaster

K-Means dijalankan dengan jarak *Euclidean*, inisialisasi *k-means*++ (Arthur & Vassilvitskii, 2007), satu kali inisialisasi, nilai awal acak (*seed*) 42, dan maksimum 100 iterasi. Algoritma diuji untuk k = 2 sampai 8, dan hasil untuk k = 3, 4, dan 5 dibandingkan secara khusus karena berada pada kisaran jumlah klaster penelitian terdahulu. Untuk tiap k dicatat SSE (Persamaan 3) dan rata-rata *Silhouette Coefficient* (Persamaan 5). Jumlah klaster akhir ditentukan dengan mempertimbangkan titik tekuk *Elbow*, nilai *Silhouette*, dan seberapa mudah karakteristik klaster ditafsirkan.

Interpretasi dan Uji Hubungan

Klaster tidak diberi nama atau label, melainkan hanya dibedakan dengan nomor, yaitu klaster 1 sampai klaster 4. Karakteristik tiap klaster ditafsirkan dari rata-rata asli keempat fitur (bukan nilai yang dinormalisasi). Untuk mengetahui apakah keanggotaan klaster berkaitan dengan karakteristik siswa, dilakukan uji *chi-square* antara klaster dan masing-masing variabel berikut: paket langganan (*free*, *basic*, *pro*), sumber pendaftaran (enam kanal), dan kelompok usia (≤ 22, 23 sampai 30, 31 sampai 40, > 40 tahun). Kekuatan hubungan diukur dengan *Cramér's V*.

Verifikasi Independen

Untuk memeriksa kekokohan hasil implementasi utama, data yang sama dijalankan ulang dengan pustaka *scikit-learn* (Pedregosa dkk., 2011) memakai 10 inisialisasi acak (n_init = 10). Hasilnya dibandingkan lewat SSE, *Silhouette*, DBI, dan *Adjusted Rand Index* (ARI; Hubert & Arabie, 1985). Kestabilan terhadap *seed* juga diuji dengan 20 kali menjalankan *K-Means* satu inisialisasi dengan *seed* berbeda.

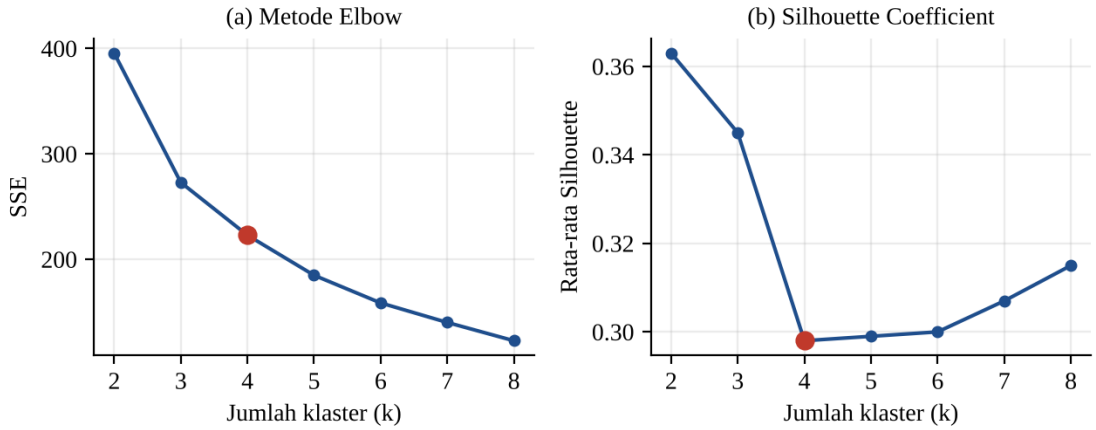
HASIL PENELITIAN DAN PEMBAHASAN

Penentuan Jumlah Klaster

Tidak ada satu nilai k yang unggul pada semua ukuran evaluasi (Tabel 4). Nilai *Silhouette* tertinggi ada pada k = 2 (0,363), dan seluruh nilai k = 2 sampai 8 berada pada rentang sempit 0,298 sampai 0,363 yang menurut skala Kaufman dan Rousseeuw (1990) termasuk struktur lemah. Pada tiga kandidat utama, k = 3 menghasilkan *Silhouette* 0,345 dan DBI 1,027, k = 4 menghasilkan 0,298 dan 1,090, sedangkan k = 5 menghasilkan 0,299 dan 1,064 (DBI dihitung dengan *scikit-learn*). Dengan demikian k = 3 sedikit lebih baik secara statistik daripada k = 4 dan k = 5. Penurunan SSE melandai bertahap, yaitu 31,0% dari k = 2 ke 3, 18,3% dari k = 3 ke 4, lalu 17,0% dan 14,2% pada dua langkah berikutnya, tanpa tekukan yang tajam. Karena itu k = 4 dipilih dengan pertimbangan karakteristik klaster yang dapat dibedakan (Tabel 5), dan penambahan klaster sesudahnya hanya menurunkan SSE dalam jumlah kecil.

Tabel 4. Evaluasi jumlah klaster. SSE dan *Silhouette* implementasi utama berasal dari satu inisialisasi (seed 42), sedangkan kolom *scikit-learn* berasal dari 10 inisialisasi.

k	SSE	Penurunan SSE	*Silhouette* (implementasi utama)	*Silhouette* (*scikit-learn*)	DBI (*scikit-learn*)
2	395,52	-	0,363	0,353	1,098
3	272,89	31,0%	0,345	0,341	1,027
4	**222,86**	**18,3%**	**0,298**	**0,303**	**1,090**
5	184,88	17,0%	0,299	0,300	1,064
6	158,63	14,2%	0,300	0,304	1,071
7	140,11	11,7%	0,307	0,318	0,984
8	122,79	12,4%	0,315	0,317	0,967



Gambar 1. Grafik *Elbow* dan nilai *Silhouette* untuk k = 2 sampai 8 (titik merah menandai k = 4).

Karakteristik Klaster

Pada k = 4, 2.536 siswa terbagi menjadi empat klaster dengan karakteristik pada Tabel 5. Klaster 2 adalah yang terbesar (33,5%), disusul klaster 1 (31,0%), klaster 3 (22,4%), dan klaster 4 (13,1%). Nomor klaster hanya berfungsi sebagai penanda dan tidak menunjukkan tingkat atau peringkat.

Tabel 5. Karakteristik empat klaster siswa (rata-rata asli keempat fitur).

Klaster	Siswa	Persentase	Kursus per siswa	Penyelesaian rata-rata (%)	Skor kuis rata-rata	Lulus kuis (%)
Klaster 1	787	31,0%	2,78	79,00	76,74	76,95
Klaster 2	849	33,5%	3,39	43,49	68,11	52,72
Klaster 3	568	22,4%	8,20	54,04	71,22	61,19
Klaster 4	332	13,1%	2,00	24,46	39,20	4,39

Klaster 1 memiliki rata-rata penyelesaian kursus tertinggi (79,00%) dan kelulusan kuis 76,95%, dengan jumlah kursus yang sedang (2,78 kursus per siswa). Klaster 2 memiliki rata-rata penyelesaian 43,49% dan kelulusan kuis 52,72%, yaitu di kisaran separuh. Klaster 3 memiliki jumlah kursus terbanyak, yaitu 8,20 kursus per siswa atau dua kali lipat rata-rata siswa platform (4,09), dengan penyelesaian 54,04% dan skor kuis 71,22 yang berada pada tingkat menengah. Klaster 4 memiliki penyelesaian terendah (24,46%), skor kuis rata-rata 39,20, dan kelulusan kuis hanya 4,39%. Sebanyak 79 siswa pada klaster 4 tidak pernah mengerjakan kuis, dan secara keseluruhan klaster ini hanya mengerjakan rata-rata 2,6 kuis per siswa (Tabel 6) dengan hasil rendah.

Perilaku Nyata Tiap Klaster

Tabel 6 menghubungkan klaster dengan catatan pendaftaran dan kuis. Jumlah kuis per siswa tidak dipakai sebagai fitur, sehingga angka ini menjadi pemeriksaan tambahan atas karakteristik di atas. Persentase pendaftaran selesai berkaitan erat dengan fitur penyelesaian dan tidak dapat dianggap bukti yang berdiri sendiri.

Tabel 6. Pendaftaran, penyelesaian, dan kuis menurut klaster (total 10.380 pendaftaran, 3.188 selesai, 30.875 hasil kuis).

Klaster	Pendaftaran	Selesai (sertifikat)	Pendaftaran selesai (%)	Kuis per siswa	Lulus per kuis (%)
Klaster 1	2.184	1.246	57,1	11,3	73,8
Klaster 2	2.877	492	17,1	8,3	51,2
Klaster 3	4.656	1.448	31,1	24,8	61,5
Klaster 4	663	2	0,3	2,6	9,2

Klaster 3 hanya 22,4% dari siswa, tetapi menyumbang 44,9% pendaftaran dan 45,6% hasil kuis, serta 1.448 dari 3.188 kursus selesai (45,4%). Jadi klaster ini kecil dari sisi jumlah siswa, tetapi besar dari sisi beban dan kontribusi platform. Sebaliknya klaster 4 hanya menyelesaikan 2 dari 663 pendaftaran.

Hubungan Klaster dengan Karakteristik Siswa

Keanggotaan klaster tidak berhubungan dengan paket langganan, sumber pendaftaran, maupun kelompok usia. Ketiga uji *chi-square* tidak signifikan pada taraf 5%, dan ukuran efek (*Cramér's V*) di sekitar 0,03, yang berarti hubungan nyaris tidak ada.

Tabel 7. Uji chi-square antara klaster dan karakteristik siswa.

Variabel	χ^2	Derajat bebas	p	*Cramér's V*
Paket langganan	6,14	6	0,408	0,035
Sumber pendaftaran	8,26	15	0,913	0,033
Kelompok usia	6,48	9	0,691	0,029

Sebagai contoh, proporsi pengguna paket gratis berkisar 63,4% sampai 68,4% pada keempat klaster, dan proporsi pendaftar dari Instagram berkisar 31,7% sampai 35,8%. Temuan ini penting secara praktis. Kanal akuisisi atau paket langganan tidak dapat dipakai untuk menebak siswa mana yang akan berhenti atau tuntas, sehingga analisis perlu berbasis perilaku belajar yang diamati.

Verifikasi Independen dan Kestabilan

Menjalankan ulang *K-Means* dengan *scikit-learn* dan 10 inisialisasi pada k = 4 menghasilkan SSE 220,87, sedikit lebih rendah (0,9%) daripada hasil implementasi utama

sebesar 222,86. Nilai *Silhouette* seluruh data pada keanggotaan klaster implementasi utama adalah 0,301, dan DBI-nya 1,117. Dua partisi itu memiliki ARI 0,636, yaitu kemiripan sedang. Batas antara klaster yang berada di tengah bergeser, tetapi empat pola karakteristik yang sama tetap muncul, yakni kelompok dengan kelulusan kuis rendah (403 siswa, lulus kuis 8,2%), kelompok dengan penyelesaian tertinggi (721 siswa, penyelesaian 82,0%), kelompok dengan jumlah kursus terbanyak (359 siswa, 9,4 kursus), dan kelompok tengah (1.053 siswa, penyelesaian 45,8%). Pada 20 *seed* berbeda dengan satu inisialisasi, rata-rata ARI antarhasil sebesar 0,794, sehingga hasil cukup stabil tetapi tidak identik. Ini menunjukkan bahwa pola karakteristik klaster kokoh, sedangkan keanggotaan siswa di perbatasan klaster dapat berubah.

PEMBAHASAN

Temuan bahwa sekitar sepertiga siswa (klaster 2) berada di tengah-tengah dan sekitar 13% siswa (klaster 4) hampir tidak berhasil sama sekali sejalan dengan literatur yang menunjukkan bahwa satu angka penyelesaian menyembunyikan beberapa pola keterlibatan (Jordan, 2015; Kizilcec dkk., 2013). Jumlah klaster yang kecil (empat) berada dalam kisaran penelitian terdahulu, yaitu tiga sampai lima klaster (Evans & Nataliani, 2024; Kizilcec dkk., 2013; Moubayed dkk., 2020; Ratnapala dkk., 2014; Safitri dkk., 2022). Data Skillora memberi sudut berbeda, yaitu adanya klaster yang berbeda karena jumlah kursus yang diambil (klaster 3), bukan hanya karena tingkat keaktifan. Hasil ini menunjukkan bahwa algoritma *K-Means* mampu memisahkan siswa berdasarkan keempat fitur perilaku belajar, meskipun batas antarklaster tidak tegas sebagaimana ditunjukkan oleh nilai *Silhouette* yang lemah.

Keterbatasan

Pertama, nilai *Silhouette* yang lemah menunjukkan batas antarklaster tidak tegas, sehingga hasil pengelompokan dipakai sebagai alat bantu analisis. Kedua, implementasi utama menjalankan satu inisialisasi, sehingga hasilnya dapat sedikit lebih buruk daripada optimum yang dicapai dengan banyak inisialisasi (bagian Verifikasi Independen dan Kestabilan). Ketiga, hanya empat fitur yang dipakai, sedangkan data aktivitas harian, durasi belajar, dan interaksi forum tidak tersedia. Keempat, klaster menggambarkan kondisi pada akhir periode data, sehingga siswa yang baru mendaftar dapat tampak rendah penyelesaiannya padahal belum sempat belajar. Kelima, k = 3 memberikan nilai *Silhouette* dan DBI yang sedikit lebih baik daripada k = 4, sehingga pemilihan k = 4 bersifat pertimbangan interpretasi. Keenam, penelitian ini bersifat observasional sehingga tidak menyimpulkan sebab-akibat.

KESIMPULAN

Penelitian ini menerapkan *K-Means Clustering* pada 2.536 siswa Skillora berdasarkan empat fitur perilaku belajar. Dengan normalisasi *min-max*, inisialisasi *k-means++*, dan jarak *Euclidean*, algoritma membentuk empat klaster siswa, yaitu klaster 1 (787 siswa), klaster 2 (849 siswa), klaster 3 (568 siswa), dan klaster 4 (332 siswa). Perbedaan klaster terlihat jelas pada penyelesaian kursus (79,0% sampai 24,5%), kelulusan kuis (77,0% sampai 4,4%), dan jumlah kursus per siswa (2,0 sampai 8,2). Jumlah klaster *k* = 4 dipilih berdasarkan pertimbangan karakteristik klaster yang dapat dibedakan, karena

ukuran statistik tidak menunjuk satu *k* yang tegas. Nilai *Silhouette* sebesar 0,298 menunjukkan struktur lemah, sehingga hasil pengelompokan berfungsi sebagai alat bantu analisis, bukan pemisahan yang tegas. Verifikasi dengan 10 inisialisasi menghasilkan SSE 0,9% lebih rendah dan empat pola klaster yang serupa. Selain itu, keanggotaan klaster tidak berhubungan dengan paket langganan, sumber pendaftaran, atau kelompok usia ($p > 0,05$ dan *Cramér's V* sekitar 0,03), sehingga pengelompokan siswa perlu didasarkan pada perilaku belajar, bukan karakteristik demografis atau kanal akuisisi. Untuk penelitian berikutnya disarankan (1) memakai banyak inisialisasi dan memilih hasil dengan SSE terendah, (2) membandingkan hasil untuk k = 3, 4, dan 5 dengan ukuran evaluasi tambahan, (3) menambahkan fitur dari log aktivitas, seperti durasi belajar, frekuensi *login*, dan interaksi forum, (4) membandingkan *K-Means* dengan metode lain, misalnya *K-Medoids*, *hierarchical clustering*, atau *Gaussian Mixture Model*, dan (5) memperbarui pengelompokan secara berkala karena perilaku siswa berubah seiring waktu.

DAFTAR PUSTAKA

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages Of Careful Seeding. In *Proceedings Of The 18th Annual ACM-SIAM Symposium On Discrete Algorithms (SODA)* (Pp. 1027–1035).

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions On Pattern Analysis And Machine Intelligence, PAMI-1*(2), 224–227.

Evans, J., & Nataliani, Y. (2024). Analisis Perilaku Siswa Pada Pembelajaran Daring Menggunakan K-Means Clustering Dari Log Data Moodle Pada SMPN 6 Salatiga. *Jusitik: Jurnal Sistem Dan Teknologi Informasi Komunikasi, 7*(2), 43–49. Https://Doi.Org/10.32524/Jusitik.V7i2.1108

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining To Knowledge Discovery In Databases. *AI Magazine, 17*(3), 37–54. Https://Doi.Org/10.1609/Aimag.V17i3.1230

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts And Techniques* (3rd Ed.). Morgan Kaufmann.

Hubert, L., & Arabie, P. (1985). Comparing Partitions. *Journal Of Classification, 2*(1), 193–218.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters, 31*(8), 651–666.

Jordan, K. (2015). Massive Open Online Course Completion Rates Revisited: Assessment, Length And Attrition. *International Review Of Research In Open And Distributed Learning, 16*(3), 341–358. Https://Doi.Org/10.19173/Irrodl.V16i3.2112

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups In Data: An Introduction To Cluster Analysis*. Wiley.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations In Massive Open Online Courses. In *Proceedings Of The 3rd International Conference On Learning Analytics And Knowledge (LAK '13)*. Leuven, Belgia.

Lloyd, S. P. (1982). Least Squares Quantization In PCM. *IEEE Transactions On Information Theory, 28*(2), 129–137.

Macqueen, J. (1967). Some Methods For Classification And Analysis Of Multivariate Observations. In *Proceedings Of The 5th Berkeley Symposium On Mathematical Statistics And Probability* (Vol. 1, Pp. 281–297).

Milligan, G. W., & Cooper, M. C. (1988). A Study Of Standardization Of Variables In Cluster Analysis. *Journal Of Classification, 5*(2), 181–204. Https://Doi.Org/10.1007/BF01897163

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student Engagement Level In An E-Learning Environment: Clustering Using K-Means. *American Journal Of Distance Education, 34*(2), 137–156. Https://Doi.Org/10.1080/08923647.2020.1696140

Pedregosa, F., Dkk. (2011). Scikit-Learn: Machine Learning In Python. *Journal Of Machine Learning Research, 12*, 2825–2830.

Ratnapala, I. P., Ragel, R. G., & Deegalla, S. (2014). Students Behavioural Analysis In An Online Learning Environment Using Data Mining. *Arxiv*. Https://Arxiv.Org/Abs/1412.7813

Romero, C., & Ventura, S. (2020). Educational Data Mining And Learning Analytics: An Updated Survey. *Wires Data Mining And Knowledge Discovery, 10*(3), E1355.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid To The Interpretation And Validation Of Cluster Analysis. *Journal Of Computational And Applied Mathematics, 20*, 53–65.

Safitri, S. N., Setiadi, H., & Suryani, E. (2022). Educational Data Mining Using Cluster Analysis Methods And Decision Trees Based On Log Mining. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 6*(3), 448–456. Https://Doi.Org/10.29207/Resti.V6i3.3935

Thorndike, R. L. (1953). Who Belongs In The Family? *Psychometrika, 18*(4), 267–276. Https://Doi.Org/10.1007/BF02289263