



ANALISIS POLA AKTIVITAS BELAJAR MAHASISWA MENGGUNAKAN METODE CLUSTERING PADA OPEN UNIVERSITY LEARNING ANALYTICS DATASET (OULAD)

Ferdi Rifaldo¹, Ahmad Maulana Ani Saputra², Wardianto³

^{1,2,3} Informatics Study Program, Faculty of Science and Technology, Universitas Nurul Huda, Indonesia

Article Info

Article History:

Received: 02-January-2026

Revised: 02-March-2026

Published: 01-June-2026

Keywords:

Learning Analytics,

Clustering,

K-Means,

OULAD

ABSTRACT

Learning analytics has become an important approach for understanding student learning behavior on online learning platforms. This study aims to analyze patterns of student learning activity using the K-Means clustering method on the Open University Learning Analytics Dataset (OULAD). The data used were obtained from studentInfo.csv and studentVle.csv, which were processed through preprocessing stages including removal of irrelevant attributes, missing value handling, duplicate data removal, activity filtering, total click aggregation, dataset merging, and outlier removal, resulting in a clean dataset of 25.559 rows. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, yielding k=3 as the best number of clusters with a Silhouette Score of 0.6543. The clustering results produced three groups of students, namely Cluster 0 with low activity (17.441 students or 68.24%), Cluster 1 with moderate activity (5.991 students), and Cluster 2 with high activity (2.127 students). The dominance of students with low activity indicates a significant engagement challenge on the online learning platform. These findings are expected to serve as a basis for educational institutions in designing early interventions to improve student engagement and academic achievement.

Kata Kunci:

Learning Analytics,

Clustering,

K-Means,

OULAD

ABSTRAK

Learning analytics menjadi salah satu pendekatan penting untuk memahami perilaku belajar mahasiswa pada platform pembelajaran daring. Penelitian ini bertujuan menganalisis pola aktivitas belajar mahasiswa menggunakan metode clustering K-Means pada Open University Learning Analytics Dataset (OULAD). Data yang digunakan berasal dari studentInfo.csv dan studentVle.csv, yang diolah melalui tahapan preprocessing berupa penghapusan atribut tidak relevan, penanganan missing value, penghapusan data duplikat, filtering aktivitas, agregasi total klik, penggabungan dataset, hingga penghapusan outlier, sehingga diperoleh dataset bersih sebanyak 25.559 baris data. Penentuan jumlah cluster optimal dilakukan menggunakan Elbow Method dan Silhouette Score, dan diperoleh k=3 sebagai jumlah cluster terbaik dengan Silhouette Score sebesar 0,6543. Hasil clustering menghasilkan tiga kelompok mahasiswa, yaitu Cluster 0 dengan aktivitas rendah (17.441 mahasiswa atau 68.24%), Cluster 1 dengan aktivitas sedang (5.991 mahasiswa), dan Cluster 2 dengan aktivitas tinggi (2.127 mahasiswa). Dominasi mahasiswa dengan aktivitas rendah mengindikasikan adanya tantangan keterlibatan yang signifikan pada platform pembelajaran daring. Temuan ini diharapkan dapat menjadi acuan bagi institusi pendidikan dalam merancang intervensi dini guna meningkatkan keterlibatan dan capaian akademik mahasiswa.

Corresponding Author:

Ahmad Maulana Ani Saputra,

Study Program of Informatics, Faculty of Science and Technology,

University Nurul Huda, Indonesia

Email: ahmadmaulanaani992@gmail.com

This is an open access article under the [CC BY-SA](#) license



1. PENDAHULUAN

Perkembangan teknologi digital dalam dunia pendidikan telah menghasilkan sejumlah besar data aktivitas pembelajaran yang dapat dimanfaatkan untuk meningkatkan kualitas proses belajar mengajar. Platform pembelajaran daring (e-learning) mampu merekam berbagai bentuk interaksi mahasiswa, seperti akses materi, pengerjaan tugas, diskusi, hingga frekuensi kunjungan ke sistem pembelajaran. Data tersebut dikenal sebagai learning analytics dan menjadi sumber informasi penting dalam memahami perilaku belajar mahasiswa (Ifenthaler & Yau, 2022).

Learning Analytics merupakan bidang penelitian yang berfokus pada pengumpulan, pengukuran, analisis, dan pelaporan data peserta didik untuk memahami serta mengoptimalkan proses pembelajaran dan lingkungan belajar (Viberg et al., 2023). Melalui pendekatan ini, institusi pendidikan dapat mengidentifikasi pola aktivitas belajar, mendeteksi mahasiswa berisiko, serta merancang strategi pembelajaran yang lebih efektif dan personal (Lim et al., 2024).

Salah satu teknik yang banyak digunakan dalam analisis learning analytics adalah data mining. Teknik ini memungkinkan peneliti menemukan pola tersembunyi dari kumpulan data berukuran besar tanpa harus memiliki label kelas sebelumnya. Dalam konteks pendidikan, data mining dapat digunakan untuk mengelompokkan mahasiswa berdasarkan karakteristik perilaku belajar mereka sehingga diperoleh informasi yang lebih mendalam mengenai tingkat keterlibatan mahasiswa selama proses pembelajaran berlangsung (Aljawarneh et al., 2022).

Metode clustering merupakan salah satu teknik unsupervised learning yang bertujuan mengelompokkan objek berdasarkan tingkat kemiripan karakteristik yang dimiliki. Objek yang berada dalam kelompok yang sama memiliki karakteristik yang lebih mirip dibandingkan dengan objek pada kelompok lainnya (Umer et al. 2023). Dalam bidang Educational Data Mining, clustering telah banyak digunakan untuk mengidentifikasi kelompok mahasiswa aktif, mahasiswa dengan aktivitas sedang, maupun mahasiswa yang cenderung pasif selama mengikuti kegiatan pembelajaran daring (Matcha et al. 2024).

Penelitian ini memanfaatkan Open University Learning Analytics Dataset (OULAD) yang merupakan salah satu dataset terbuka paling banyak digunakan dalam penelitian learning analytics. Dataset ini menyediakan informasi profil mahasiswa, aktivitas interaksi pada Virtual Learning Environment (VLE), serta performa akademik mahasiswa yang berasal dari Open University di Inggris (Hlosta et al., 2022). Ketersediaan data aktivitas yang lengkap menjadikan OULAD sangat relevan untuk digunakan dalam analisis pola belajar mahasiswa berbasis data.

Permasalahan yang diangkat dalam penelitian ini adalah bagaimana mengidentifikasi pola aktivitas belajar mahasiswa berdasarkan data interaksi yang tersimpan pada sistem pembelajaran daring. Aktivitas belajar yang berbeda-beda antar mahasiswa sering kali sulit diamati secara manual karena jumlah data yang sangat besar. Oleh karena itu diperlukan metode analisis yang mampu mengelompokkan mahasiswa secara otomatis berdasarkan karakteristik aktivitas belajarnya.

Penelitian ini bertujuan untuk menganalisis pola aktivitas belajar mahasiswa menggunakan metode clustering pada dataset OULAD. Sebelum proses clustering dilakukan, data terlebih dahulu melalui tahapan preprocessing yang meliputi pembersihan data, penanganan nilai kosong, penghapusan data duplikat, penggabungan beberapa sumber data, agregasi aktivitas mahasiswa, serta penghapusan outlier. Tahapan ini dilakukan untuk menghasilkan dataset yang bersih dan siap digunakan dalam proses analisis.

Hasil penelitian diharapkan dapat memberikan gambaran mengenai karakteristik kelompok mahasiswa berdasarkan tingkat aktivitas belajar mereka pada lingkungan pembelajaran daring. Informasi tersebut dapat dimanfaatkan oleh institusi pendidikan sebagai dasar pengambilan keputusan dalam meningkatkan kualitas pembelajaran, memberikan intervensi dini kepada mahasiswa yang kurang aktif, serta mendukung implementasi pembelajaran berbasis data (learning analytics) secara lebih optimal (Gašević et al., 2024).

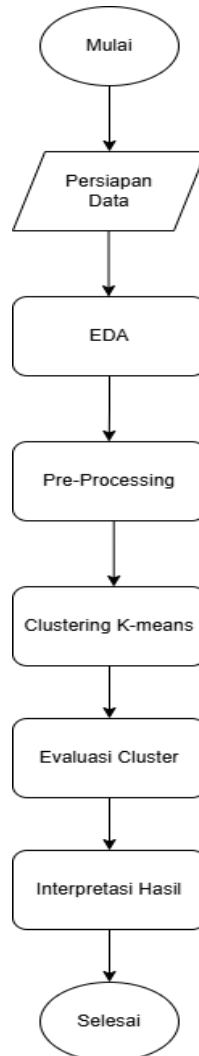
2. METODE PENELITIAN

2.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode Educational Data Mining (EDM) untuk menganalisis pola aktivitas belajar mahasiswa berdasarkan data interaksi pada sistem pembelajaran daring. Teknik yang digunakan adalah clustering, yaitu metode unsupervised learning yang mengelompokkan mahasiswa berdasarkan tingkat kemiripan aktivitas belajar yang tercermin dalam jumlah interaksi pada Virtual Learning Environment (VLE). Tahapan penelitian yang dilakukan terdiri atas:

1. Persiapan lingkungan kerja dan library.
2. Memuat dataset OULAD.
3. Eksplorasi data awal (EDA).

4. Preprocessing data.
5. Pembentukan dataset akhir.
6. Penerapan algoritma clustering.
7. Evaluasi hasil clustering.
8. Interpretasi kelompok mahasiswa



Gambar 1. Flowchart alur penelitian

2.2 Dataset Penelitian

Penelitian menggunakan Open University Learning Analytics Dataset (OULAD) yang tersedia secara terbuka dan banyak digunakan dalam penelitian learning analytics. Dataset ini berisi informasi mengenai aktivitas mahasiswa selama mengikuti proses pembelajaran daring pada Open University.

Data utama yang digunakan meliputi: studentInfo.csv, studentVle.csv, studentAssessment.csv (opsional). Deskripsi masing-masing dataset dijelaskan sebagai berikut:

StudentInfo.csv, ini berisi : Id student, gender, region, highest_education , age_band, disability, dan final_result.

StudentVle.csv, ini berisi data aktivitas mahasiswa pada Virtual Learning Environment berupa jumlah klik (sum_click) yang dilakukan mahasiswa selama proses pembelajaran.

studentAssessment.csv (opsional), digunakan apabila analisis ingin mempertimbangkan performa akademik mahasiswa berdasarkan hasil penilaian.

2.3 Persiapan Environment dan Library

Proses analisis pada penelitian ini dilakukan menggunakan bahasa pemrograman Python, dengan dukungan beberapa pustaka utama untuk pengolahan data dan machine learning.

Library yang digunakan meliputi, yaitu : Pandas untuk manipulasi data, NumPy untuk komputasi numerik, Matplotlib beserta Seaborn untuk visualisasi data, dan Scikit-Learn (sklearn) untuk preprocessing dan clustering.

Tahapan tersebut dipersiapkan lebih dahulu agar seluruh kebutuhan analisis, mulai dari pengolahan data hingga proses clustering, sudah tersedia sebelum data mulai diolah.

2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis dilakukan untuk memahami karakteristik dataset sebelum proses preprocessing (Romero & Ventura 2024).

Tahapan EDA meliputi:

1. Pemeriksaan Struktur Data
Melakukan identifikasi tipe data pada setiap atribut untuk mengetahui apakah data berupa numerik, kategorikal, maupun teks.
2. Pemeriksaan Dimensi Dataset
Mengidentifikasi jumlah baris dan kolom pada dataset guna memastikan data berhasil dimuat secara lengkap.
3. Analisis Distribusi Data, tahap ini melakukan analisis awal terhadap: Distribusi jenis kelamin mahasiswa, distribusi kelompok usia, tingkat pendidikan terakhir, dan sebaran aktivitas belajar mahasiswa.
4. Pemeriksaan Missing Value, ini mengidentifikasi atribut yang memiliki nilai kosong (missing value) sehingga dapat ditentukan metode penanganan yang sesuai.

2.5 Preprocessing Data

Tahapan pre-processing dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses clustering.

a. Penghapusan Kolom Tidak Relevan

Beberapa kolom yang tidak dibutuhkan dalam analisis dihapus agar kompleksitas data berkurang dan komputasi menjadi lebih efisien.

Ketiga kolom yang dihapus adalah: code_module, code_presentation, id_site

Penghapusan atribut ini dilakukan karena ketiganya tidak berkaitan langsung dengan pola aktivitas belajar mahasiswa yang menjadi fokus analisis.

b. Penghapusan Data Duplikat

Data duplikat diidentifikasi lalu dihapus menggunakan fungsi duplicate checking pada Pandas, agar tidak terjadi perhitungan ganda yang dapat memengaruhi hasil analisis.

c. Penanganan Missing Value

Data yang memiliki nilai kosong ditangani berdasarkan tipe datanya.

Untuk atribut kategorikal: Missing Value—"Unknown"

Sedangkan atribut numerik seperti jumlah percobaan atau aktivitas diisi dengan nilai: Missing Value=0

Pendekatan ini dipilih agar jumlah data tetap terjaga, tanpa kehilangan informasi yang masih relevan untuk analisis.

d. Filtering Aktivitas Mahasiswa

Data aktivitas mahasiswa disaring berdasarkan jumlah interaksinya pada VLE.

Kriteria yang digunakan: sum_click > 0

Mahasiswa tanpa aktivitas pada sistem pembelajaran tidak diikutsertakan, sebab datanya tidak mencerminkan perilaku belajar yang dapat diamati.

e. Agregasi Aktivitas

Data aktivitas mahasiswa pada file studentVle.csv dikelompokkan berdasarkan id_student.

Total aktivitas dihitung menggunakan persamaan: $TotalClick_i = \sum_{j=1}^n Click_{ij}$

Dengan keterangan :

Total Click_(i) = total klik mahasiswa ke-i

$Click_{(ij)}$ = jumlah klik pada aktivitas ke-j milik mahasiswa ke-i

n = jumlah aktivitas mahasiswa

Dari proses ini diperoleh satu nilai total aktivitas untuk masing-masing mahasiswa.

f. Penggabungan Data (Merge)

Data profil mahasiswa dan data aktivitas hasil agregasi digabungkan dengan menggunakan atribut: `id_student`

Penggabungan dilakukan menggunakan operasi join pada Pandas sehingga dihasilkan satu dataset yang memuat profil sekaligus aktivitas mahasiswa.

g. Penghapusan Outlier

Data aktivitas dengan nilai ekstrem dihapus menggunakan pendekatan persentil, yaitu dengan membuang: 1% data dengan nilai klik terendah, dan 1% data dengan nilai klik tertinggi.

Langkah ini dilakukan untuk mengurangi pengaruh nilai ekstrem yang berpotensi mendistorsi hasil clustering (Alharbi & Alshehri 2024).

h. Reset Index

Setelah data yang tidak diperlukan dibuang, indeks dataset diatur ulang menggunakan fungsi reset index, agar nomor baris kembali berurutan dan memudahkan proses analisis selanjutnya.

i. Verifikasi Dataset Akhir

Dataset hasil preprocessing diverifikasi dengan:

1. Menampilkan lima baris pertama data.
2. Memeriksa jumlah baris dan kolom akhir.
3. Memastikan tidak terdapat missing value maupun data duplikat.

j. Penyimpanan Dataset

Dataset yang telah bersih dan siap digunakan disimpan ke dalam file: `cleaned_oulad_final.csv`

File ini digunakan sebagai input utama pada tahap clustering dan analisis pola aktivitas belajar mahasiswa (Alharbi & Alshehri 2024).

2.6 Metode Clustering

Setelah preprocessing selesai, dataset akan dianalisis menggunakan algoritma clustering untuk mengelompokkan mahasiswa berdasarkan tingkat aktivitas belajar. Kelompok yang dihasilkan akan dapat merepresentasikan kategori:

1. Mahasiswa dengan aktivitas tinggi.
2. Mahasiswa dengan aktivitas sedang.
3. Mahasiswa dengan aktivitas rendah.

Kualitas cluster nantinya dapat dievaluasi menggunakan metrik Silhouette Score untuk mengukur tingkat pemisahan antar cluster dan kekompakan data dalam setiap cluster (Rousseeuw et al., 2024).

3. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan analisis terhadap dataset Open University Learning Analytics Dataset (OULAD) yang terdiri dari dataset `studentInfo` dan `studentVle`. Dataset `studentInfo` digunakan untuk memperoleh informasi profil mahasiswa, sedangkan dataset `studentVle` digunakan untuk melihat aktivitas mahasiswa pada sistem pembelajaran daring atau Virtual Learning Environment (VLE) dan Kuantitas interaksi (jumlah klik), Keaktifan dan ketidakhadiran mahasiswa, perilaku belajar (misalnya berdasarkan umur, wilayah, atau pengalaman kalau ada), dan stabilitas aktifitas nya.

Tahapan analisis dimulai dengan preprocessing data untuk membersihkan dataset sebelum dilakukan proses clustering menggunakan algoritma K-Means. Tujuan clustering adalah mengelompokkan mahasiswa berdasarkan tingkat aktivitas belajar mereka pada platform pembelajaran daring.

3.1 Hasil Pra-pemrosesan Data

Tahap preprocessing dilakukan untuk meningkatkan kualitas data agar lebih bersih, konsisten, dan siap digunakan pada proses clustering. Dataset awal masih mengandung data duplikat serta beberapa aktivitas yang tidak relevan terhadap analisis pola aktivitas belajar mahasiswa.

Tahapan preprocessing yang dilakukan meliputi:

- Penghapusan atribut yang tidak relevan.
- Penghapusan data duplikat.
- Pengecekan missing value.
- Penghapusan aktivitas dengan nilai sum_click = 0.
- Agregasi aktivitas mahasiswa menjadi fitur total_click.
- Penggabungan dataset studentInfo dan studentVle.
- Penanganan outlier menggunakan persentil 1% dan 99%.

Dataset studentInfo berisi informasi karakteristik mahasiswa seperti jenis kelamin, wilayah, umur, pendidikan terakhir, jumlah SKS, status disabilitas, dan hasil akhir pembelajaran mahasiswa. Sedangkan dataset studentVle berisi data aktivitas mahasiswa pada LMS/VLE seperti jumlah klik aktivitas (sum_click), waktu akses (date), dan materi yang diakses (id_site).

Hasil pengecekan dataset sebelum dan sesudah preprocessing adalah sebagai berikut:

1. Dataset studentInfo

Ukuran data:

Jumlah baris: 32.593

Jumlah kolom: 12

Hasil pengecekan data hilang / rusak Tidak ditemukan data kosong (missing value) pada seluruh kolom.

```
=====
EDA - STUDENTINFO
=====
```

	Dataset	Jumlah Baris	Jumlah Kolom	Jumlah Duplikat	Total Missing Value
0	studentInfo	32593	12	0	0

Missing value studentInfo:

	Kolom	Missing Value
0	code_module	0
1	code_presentation	0
2	id_student	0
3	gender	0
4	region	0
5	highest_education	0
6	imd_band	0
7	age_band	0
8	num_of_prev_attempts	0
9	studied_credits	0
10	disability	0
11	final_result	0

Gambar 2. Hasil pemeriksaan missing value dan informasi dataset studentInfo

final_result , 0

Hasil pengecekan duplikat Jumlah data duplikat: 0

Artinya dataset studentInfo sudah bersih dan tidak memiliki baris ganda.

2. Dataset studentVle

Ukuran data:

Jumlah baris: 1.065.280

Jumlah kolom: 6

Hasil pengecekan data hilang / rusak Tidak ditemukan missing value pada dataset ini

```
=====
EDA - STUDENTVLE
=====
```

	Dataset	Jumlah Baris	Jumlah Kolom	Jumlah Duplikat	Total Missing Value
0	studentVle	10655280	6	787170	0

```
Missing value studentVle:
```

	Kolom	Missing Value
0	code_module	0
1	code_presentation	0
2	id_student	0
3	id_site	0
4	date	0
5	sum_click	0

Gambar 3. Hasil pemeriksaan dataset studentVle

Hasil pengecekan duplikat Jumlah data duplikat: 787.170 baris

Artinya terdapat banyak aktivitas mahasiswa yang tercatat lebih dari satu kali dengan isi data yang sama persis.

Kesimpulan dataset studentVle

Dataset ini:

Tidak memiliki data hilang

Tetapi memiliki 787.170 data duplikat

Sehingga pada tahap preprocessing dilakukan penghapusan data duplikat agar analisis clustering lebih akurat.

3. Dataset hasil preprocessing cleaned_oulad_final.csv

(Ukuran file hasil preprocessing)

Jumlah baris: 25.559

Jumlah kolom: 11

Hasil pengecekan data hilang tidak ditemukan missing value.

Status Jumlah

Missing Value 0

Hasil pengecekan duplikat

Status Jumlah

Data Duplikat 0.

Hasil preprocessing menunjukkan bahwa dataset studentInfo tidak memiliki missing value maupun data duplikat. Pada dataset studentVle ditemukan sebanyak 787.170 data duplikat sehingga dilakukan penghapusan duplicate rows agar data tidak redundan dan hasil analisis menjadi lebih akurat.

Selain itu, aktivitas dengan nilai sum_click = 0 juga dihapus karena dianggap tidak menunjukkan interaksi belajar mahasiswa pada platform VLE. Setelah proses cleaning selesai dilakukan, diperoleh dataset akhir bernama cleaned_oulad_final.csv.

Dataset akhir memiliki:

25.559 baris data

11 atribut/kolom

Dataset tersebut telah bersih dari missing value dan data duplikat sehingga siap digunakan untuk proses clustering.

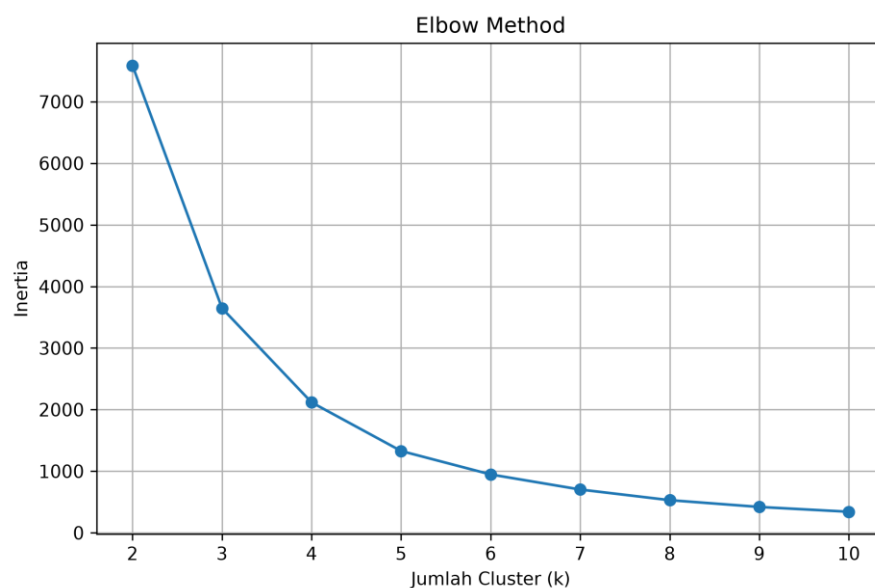
Tahapan	Hasil
0 Duplikat studentVle dihapus	787170
1 Aktivitas sum_click = 0 dihapus	0
2 Agregasi total_click	26074
3 Duplikat profil studentInfo dihapus	3808
4 Merge studentInfo dan studentVle	26074
5 Outlier persentil 1% dan 99% dihapus	515
6 Duplikat akhir dihapus	0
7 Dataset akhir	25559

Dataset akhir:
 (25559, 11)
 Missing akhir: 0
 Duplikat akhir: 0

Gambar 4. Ringkasan hasil preprocessing dataset

3.2 Penentuan Jumlah Cluster Optimal

Sebelum proses clustering dilakukan, dilakukan penentuan jumlah cluster optimal menggunakan dua metode, yaitu Elbow Method dan Silhouette Score. Elbow Method digunakan untuk mengidentifikasi titik belok pada grafik nilai inertia terhadap jumlah cluster k, sedangkan Silhouette Score digunakan untuk mengukur seberapa baik setiap data terwakili oleh clusternya sendiri dibandingkan dengan cluster lain (Rousseeuw et al., 2024).



Gambar 5. Hasil evaluasi model menggunakan metode Elbow Method

Berdasarkan hasil pengujian menggunakan Elbow Method, penurunan nilai inertia yang signifikan terjadi pada k=2 hingga k=3 dan mulai melandai setelah k=3. Hasil evaluasi menggunakan Silhouette Score menunjukkan bahwa nilai tertinggi diperoleh pada k=2 sebesar 0,7119, sedangkan pada k=3 diperoleh nilai sebesar 0,6543. Meskipun nilai tertinggi berada pada k=2, penelitian ini memilih k=3 karena mampu menghasilkan segmentasi mahasiswa yang lebih representatif berdasarkan tingkat aktivitas belajar, yaitu aktivitas rendah, sedang, dan tinggi. Selain itu, nilai Silhouette Score sebesar 0,6543 masih menunjukkan kualitas cluster yang baik dengan tingkat pemisahan antar cluster yang cukup jelas. Oleh karena itu, jumlah cluster sebanyak tiga dipilih untuk proses clustering pada penelitian ini.

3.3 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kualitas cluster yang dihasilkan oleh algoritma K-Means. Pada penelitian ini, evaluasi dilakukan menggunakan Silhouette Score yang mampu mengukur tingkat kemiripan suatu

data terhadap cluster tempat data tersebut berada dibandingkan dengan cluster lainnya. Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan kualitas cluster yang semakin baik (Rousseeuw, 1987). Selain itu, nilai inertia juga digunakan sebagai indikator untuk melihat tingkat kekompakan cluster yang terbentuk.

```

Evaluasi cluster:
  k      Inertia  Silhouette Score
0  2  7588.204062    0.711939
1  3  3646.389495    0.654326
2  4  2117.633702    0.618784
3  5  1328.461752    0.604018
4  6   947.675336    0.585663
5  7   702.758230    0.569962
6  8   529.982057    0.569238
7  9   419.921744    0.570171
8 10   341.965647    0.558213

Silhouette Score k=3: 0.6543

```

Gambar 6. Hasil evaluasi model menggunakan Silhouette Score

Berdasarkan Gambar 6, hasil evaluasi menunjukkan bahwa model K-Means dengan tiga cluster memperoleh nilai Silhouette Score sebesar 0,6543. Nilai tersebut menunjukkan bahwa cluster yang terbentuk memiliki tingkat pemisahan yang baik sehingga setiap kelompok mahasiswa dapat dibedakan secara cukup jelas berdasarkan tingkat aktivitas belajarnya pada platform VLE. Meskipun nilai Silhouette Score tertinggi diperoleh pada k=2 sebesar 0,7119, penelitian ini tetap menggunakan k=3 karena tujuan penelitian adalah mengelompokkan mahasiswa ke dalam tiga kategori aktivitas belajar yang lebih representatif, yaitu aktivitas rendah, sedang, dan tinggi. Dengan demikian, model clustering yang dihasilkan dinilai layak digunakan untuk proses segmentasi mahasiswa berdasarkan tingkat aktivitas belajar pada platform VLE.

3.4 Hasil Clustering K-Means

Algoritma K-Means diterapkan pada dataset `cleaned_oulad_final.csv` dengan menggunakan fitur `total_click` sebagai atribut utama pengelompokan. Sebelum proses clustering, fitur `total_click` dinormalisasi menggunakan `StandardScaler` untuk memastikan skala data yang konsisten dan menghindari bias pada perhitungan jarak antar data point (Pedregosa et al. 2011). Proses clustering dijalankan dengan parameter `n_clusters=3`, `random_state=42`, dan `n_init=10` untuk memastikan stabilitas hasil (Umer et al. 2023).

```

Hasil clustering:
cluster kategori_aktivitas jumlah_mahasiswa rata_rata_total_click minimum_total_click maksimum_total_click persentase
0 0 Aktivitas Rendah 17441 532.22 4 1482 68.24
1 1 Aktivitas Sedang 5991 2432.35 1483 3909 23.44
2 2 Aktivitas Tinggi 2127 5378.46 3911 8765 8.32

=====
SELESAI

```

Gambar 7. Hasil clustering K-Means

Hasil penerapan K-Means dengan k=3 pada dataset OULAD menghasilkan tiga kelompok mahasiswa sebagaimana ditunjukkan pada Tabel 2 berikut.

Tabel 2. Hasil Clustering K-Means (k=3) pada Dataset OULAD

Cluster	Jumlah Mahasiswa	Rata-rata Total Klik	Kategori Aktivitas
0	17.441	532.22	Aktivitas Rendah
1	5.991	2432.35	Aktivitas Sedang
2	2.127	5378.46	Aktivitas Tinggi

Berdasarkan hasil clustering pada Tabel 2, mahasiswa terbagi ke dalam tiga kelompok yang berbeda secara signifikan. Cluster 0 merupakan kelompok terbesar dengan 17.441 mahasiswa dan rata-rata total klik sebesar 532.22, yang dikategorikan sebagai mahasiswa dengan aktivitas rendah. Cluster 1 mencakup 5.991 mahasiswa dengan rata-rata total klik 2432.35, yang dikategorikan sebagai mahasiswa dengan aktivitas sedang. Sementara itu, Cluster 2 adalah kelompok terkecil dengan 2.127 mahasiswa namun memiliki rata-rata total klik tertinggi yaitu 5378.46, yang dikategorikan sebagai mahasiswa dengan aktivitas tinggi.

Nilai Silhouette Score yang diperoleh dari proses clustering adalah 0,6543. Nilai ini mengindikasikan bahwa pemisahan antar cluster cukup baik dan setiap mahasiswa memiliki keanggotaan cluster yang representatif. Skor

di atas 0,50 umumnya dianggap sebagai hasil clustering yang memadai dalam konteks educational data mining (Le Quy et al. 2023).

3.5 Interpretasi dan Pembahasan Hasil Cluster

Berdasarkan hasil clustering K-Means, diperoleh tiga kelompok mahasiswa yang mencerminkan tingkat keterlibatan yang berbeda dalam platform VLE. Interpretasi setiap cluster diuraikan sebagai berikut.

Cluster 0 – Mahasiswa Aktivitas Rendah. Kelompok ini merupakan kelompok terbesar yang mencakup 68.24% dari total mahasiswa. Mahasiswa pada cluster ini memiliki rata-rata total klik yang sangat rendah, yaitu sekitar 532.22 klik selama periode pembelajaran. Rendahnya interaksi pada platform VLE mengindikasikan bahwa mahasiswa dalam kelompok ini cenderung tidak aktif dalam mengakses materi dan sumber belajar daring. Kondisi ini perlu mendapat perhatian dari institusi pendidikan karena berpotensi berdampak negatif terhadap capaian akademik mahasiswa (Aljawarneh et al., 2022).

Cluster 1 – Mahasiswa Aktivitas Sedang. Kelompok ini terdiri atas 5.991 mahasiswa dengan rata-rata total klik sebesar 2432.35. Mahasiswa pada cluster ini menunjukkan tingkat keterlibatan yang cukup dalam menggunakan platform pembelajaran daring. Mereka secara teratur mengakses materi dan menyelesaikan tugas, namun intensitas interaksinya belum seintensif kelompok mahasiswa aktif. Kelompok ini merupakan kelompok menengah yang berpotensi untuk ditingkatkan keterlibatannya melalui intervensi pembelajaran yang tepat (Leitner et al. 2023).

Cluster 2 – Mahasiswa Aktivitas Tinggi. Kelompok ini merupakan kelompok terkecil dengan 2.127 mahasiswa, namun memiliki rata-rata total klik tertinggi yaitu 5378.46 klik. Mahasiswa dalam kelompok ini menunjukkan tingkat keterlibatan yang sangat tinggi dalam menggunakan platform VLE. Mereka secara aktif mengakses berbagai materi, berpartisipasi dalam forum diskusi, dan menyelesaikan tugas dengan frekuensi tinggi. Pola aktivitas ini berkorelasi positif dengan performa akademik yang lebih baik sebagaimana ditunjukkan oleh penelitian sebelumnya terkait hubungan antara keterlibatan daring dan hasil belajar mahasiswa (Viberg et al., 2023).

Hasil penelitian ini sejalan dengan temuan (Matcha et al. 2024) yang menyatakan bahwa metode clustering mampu mengidentifikasi tiga kategori mahasiswa berdasarkan tingkat aktivitas belajar daring, yaitu kelompok mahasiswa aktif,

KESIMPULAN

Penelitian ini berhasil menganalisis pola aktivitas belajar mahasiswa menggunakan metode clustering K-Means pada Open University Learning Analytics Dataset (OULAD). Tahapan preprocessing yang dilakukan meliputi penghapusan atribut tidak relevan, penghapusan 787.170 data duplikat pada dataset studentVle, penanganan missing value, filtering aktivitas dengan $\text{sum_click} = 0$, agregasi total klik per mahasiswa, penggabungan dataset, penghapusan outlier, serta verifikasi dataset akhir. Proses tersebut menghasilkan dataset bersih sebanyak 25.559 baris data yang siap digunakan untuk analisis clustering.

Penentuan jumlah cluster optimal menggunakan Elbow Method dan Silhouette Score menghasilkan nilai $k=3$ sebagai jumlah cluster terbaik dengan Silhouette Score sebesar 0,6543. Penerapan algoritma K-Means menghasilkan tiga kelompok mahasiswa: Cluster 0 (aktivitas rendah) dengan 17.441 mahasiswa dan rata-rata 532.22 klik, Cluster 1 (aktivitas sedang) dengan 5.991 mahasiswa dan rata-rata 2432.35 klik, serta Cluster 2 (aktivitas tinggi) dengan 2.127 mahasiswa dan rata-rata 5378.46 klik.

Dominasi kelompok mahasiswa dengan aktivitas rendah (68.24% dari total mahasiswa) menunjukkan adanya tantangan keterlibatan yang signifikan pada platform pembelajaran daring Open University. Temuan ini dapat menjadi landasan bagi institusi pendidikan dalam merancang intervensi dini berbasis data, seperti sistem peringatan dini, pendampingan akademik, dan pengembangan konten pembelajaran yang lebih interaktif untuk meningkatkan keterlibatan mahasiswa. Untuk penelitian selanjutnya, disarankan untuk mengintegrasikan fitur tambahan seperti data penilaian akademik (studentAssessment) dan demografi mahasiswa guna memperoleh gambaran profil cluster yang lebih komprehensif dan mendukung pengambilan keputusan berbasis data secara lebih optimal.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah mendukung terlaksananya penelitian ini. Apresiasi khusus disampaikan kepada Open University yang telah menyediakan dataset OULAD secara terbuka untuk keperluan penelitian di bidang learning analytics.

REFERENSI

[Referensi akan dilengkapi]

- [1] D. Ifenthaler and J. Y. K. Yau, "Utilising Learning Analytics for Study Success: Reflections on Current Empirical Findings," *Educational Technology Research and Development*, vol. 70, no. 2, pp. 635–649, 2022.
- [2] O. Viberg, A. Khalil, and M. Baars, "Self-Regulated Learning and Learning Analytics in Online Education: A Systematic Review," *Internet and Higher Education*, vol. 57, Art. no. 100891, 2023.
- [3] S. K. Mohamad and N. Tasir, "Learning Analytics Applications in Online Learning: A Systematic Literature Review," *Education Sciences*, vol. 11, no. 8, Art. no. 408, 2021, doi: 10.3390/educsci11080408.
- [4] P. Leitner, M. Khalil, and M. Ebner, "Learning Analytics in Higher Education—A Review of Current Challenges and Future Research Directions," *Computers*, vol. 12, no. 5, Art. no. 102, 2023.
- [4] N. Matcha, D. Gašević and S. Jovanović, "Analytics of Student Engagement in Online Learning Environments," *Journal of Learning Analytics*, vol. 11, no. 1, pp. 1–18, 2024.
- [5] S. Rienties, B. K. Nguyen, and B. Toetenel, "Understanding Student Engagement Through Learning Analytics: Recent Advances and Implications," *Distance Education*, vol. 45, no. 2, pp. 211–229, 2024.
- [6] C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 3, 2024.
- [7] H. Aljohani, "Educational Data Mining and Learning Analytics for Twenty-First Century Higher Education: A Review and Synthesis," *Telematics and Informatics Reports*, vol. 5, Art. no. 100016, 2022.
- [8] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics Dataset," *Scientific Data*, vol. 4, Art. no. 170171, 2017, doi: 10.1038/sdata.2017.171.
- [9] M. Hlosta, Z. Zdrahal, and H. Khalil, "Applications of the Open University Learning Analytics Dataset (OULAD): A Review of Recent Research Trends," *Computers and Education: Artificial Intelligence*, vol. 3, Art. no. 100071, 2022.
- [10] T. Le Quy, G. Friege, and E. Ntoutsis, "A Review of Clustering Models in Educational Data Science Towards Fairness-Aware Learning," *arXiv:2301.03421*, 2023.
- [11] S. Balovskyak, V. Pasichnyk, and Y. Shcherbyna, "Clustering Students According to Their Academic Achievement Using Fuzzy Logic," *arXiv:2312.10047*, 2023.
- [12] M. Umer, M. Sadiq, and A. Mehmood, "Machine Learning and Clustering Approaches for Student Behaviour Analysis in Learning Management Systems," *Applied Sciences*, vol. 13, no. 7, Art. no. 4378, 2023, doi: 10.3390/app13074378.
- [13] D. Gašević S. Dawson, and G. Siemens, "Learning Analytics for Improving Student Success in Digital Education," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 14–32, 2024.
- [14] A. M. Alharbi and A. Alshehri, "Preprocessing Techniques in Educational Data Mining: Impact on Clustering Performance," *IEEE Access*, vol. 12, pp. 24891–24907, 2024.
- [15] J. S. Hwang and Y. Kim, "Data Cleaning and Feature Engineering Strategies for Learning Analytics Research," *Information*, vol. 16, no. 2, Art. no. 87, 2025.
- [16] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
- [17] C. R. Harris et al., "Array Programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.
- [18] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [19] 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, 1987, doi: 10.1016/0377-0427(87)90125-7.