PENERAPAN *DATA MINING* MENGGUNAKAN ALGORITMA *K-MEANS* UNTUK ANALISIS DATA BELANJA *ONLINE* MAHASISWA

(*Application of Data Mining Using the K-Means Algorithm for Analysis of Student Online Shopping Data*)

Deiva Verlyn Marjuki*[1], Mutyara Safitri[2], Harun Al Rosyid[3]
[1] [2] [3]Program Studi Pendidikan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Surabaya
Jl. Ketintang, Gayungan, Surabaya, INDONESIA
Email: [1]deivaverlyn.22009@mhs.unesa.ac.id, [2]mutyarasafitri.22011@mhs.unesa.ac.id, [3]harunrosyid@unesa.ac.id

Abstract

This study was conducted to analyze the online shopping behavior of college students using the K-Means algorithm as a clustering technique in data mining. This study was motivated by the lack of systematic segmentation of student shopping behavior, which limits the understanding of purchasing characteristics within this consumer group. Unlike previous studies that mostly examine general retail customers or broad e-commerce users, this study specifically focuses on university students by integrating demographic and behavioral attributes. The originality of this study is reflected in the simultaneous use of six variables, namely gender, shopping time, product type, expenditure level, payment method, and purchase decision factors. Data were collected through an online survey involving 200 active college students. The research stages consisted of data cleaning, data category transformation using One-Hot Encoding, clustering model construction using the K-Means algorithm, and cluster evaluation using the Silhouette method. The evaluation results showed that the optimal number of clusters was k = 3, achieving the highest Silhouette score of 0.0913. Three distinct segments of college students' online shopping behavior were identified, providing insights that can support more targeted marketing strategies and student-oriented e-commerce services.

Keywords: *Data Mining, K-Means, Clustering, Online Shopping, Silhouette*

**Corresponding Author*

1. PENDAHULUAN

Perilaku belanja mahasiswa telah berubah secara signifikan sebagai akibat dari perkembangan e-commerce. Konsumen mahasiswa memiliki karakteristik khusus, seperti keterbatasan anggaran, literasi digital yang tinggi, dan kecenderungan untuk memilih barang berdasarkan harga, promosi, dan kemudahan transaksi. Belanja online mahasiswa tidak hanya dipengaruhi oleh kebutuhan fungsional mahasiswa ada juga faktor gaya hidup, waktu yang dihabiskan, dan preferensi pembayaran digital. Oleh karena itu, sangat penting untuk memahami pola perilaku belanja mahasiswa saat membuat strategi e-commerce, instruksi keuangan, dan kebijakan pemasaran yang lebih tepat sasaran [1].

Penelitian terdahulu menunjukkan bahwa *Clustering* adalah teknik yang mampu mengidentifikasi pola dan fitur data yang tidak tampak secara langsung [2]. Penelitian telah menunjukkan bahwa *Clustering* dapat memberikan gambaran segmentasi yang dapat digunakan sebagai dasar strategi kebijakan dan bisnis [3]. Ini ditunjukkan oleh pengelompokan data tentang konsumsi produk kosmetik, data COVID-19 berdasarkan wilayah, dan data transaksi pembayaran. Banyak orang menggunakan model *Clustering*, terutama algoritma *K-Means*, untuk analisis produk yang laku di pasaran, pengelompokan potensi perikanan berdasarkan provinsi, dan analisis sentra industri. Menurut beberapa penelitian, metode clustering, terutama algoritma K-Means, dapat digunakan untuk menemukan pola perilaku konsumsi pada kelompok pengguna tertentu, seperti mahasiswa, yang merupakan segmen konsumen digital.

Tahap evaluasi diperlukan dalam proses pembuatan model *Clustering* untuk menentukan jumlah *Cluster* yang paling ideal [4]. *Silhouette Coefficient*, *Davies-Bouldin Index*, dan *Elbow Method* adalah beberapa metode evaluasi yang sering digunakan [5]. Metode *Silhouette* menunjukkan tingkat keseragaman objek dengan anggota kelompoknya dibandingkan dengan *Cluster* lain, menunjukkan kualitas pengelompokan yang terbentuk

[6]. Penelitian sebelumnya yang menggunakan metode *Silhouette* untuk menilai pengelompokan data obat, data kesiapan ujian, dan fitur kanal YouTube menunjukkan bahwa metode ini dapat memvalidasi hasil pengelompokan dengan lebih objektif [7]. Oleh karena itu, dalam penelitian ini, penggunaan Silhouette Coefficient sangat penting untuk memastikan bahwa jumlah klaster yang terbentuk benar-benar mewakili pola perilaku belanja mahasiswa secara optimal.

Berbagai fakta menyatakan bahwa banyak penelitian telah dilakukan mengenai perilaku belanja online, sebagian besar penelitian masih berfokus pada pelanggan e-commerce secara keseluruhan atau pelanggan retail skala besar. Penelitian ini masih berfokus pada faktor-faktor seperti demografi mahasiswa, waktu yang mahasiswa habiskan untuk berbelanja, jenis produk, metode pembayaran, dan faktor-faktor lainnya yang memengaruhi keputusan mahasiswa untuk berbelanja online. Selain itu, penelitian sebelumnya biasanya hanya menggunakan satu atau dua variabel utama, sehingga belum mampu menjelaskan secara menyeluruh segmentasi perilaku belanja mahasiswa. Kondisi ini menunjukkan bahwa kurangnya penelitian mengenai kebutuhan akan segmentasi perilaku belanja mahasiswa yang lebih sistematis dan berbasis data.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk memanfaatkan model *data mining* menggunakan algoritma *K-Means* untuk menganalisis data belanja *online* mahasiswa [8]. Data belanja *online* semakin meningkat, terutama di kalangan mahasiswa, menghasilkan pola perilaku konsumsi yang menarik untuk dipelajari [9]. Diharapkan bahwa penggunaan *K- Means* akan menghasilkan berbagai kategori belanja, seperti kategori kebutuhan primer, sekunder, dan gaya hidup. Metode *Silhouette* digunakan untuk mengevaluasi hasil pengelompokan untuk menentukan jumlah *Cluster* terbaik [10].

Berdasarkan uraian tersebut, pertanyaan penelitian dirumuskan sebagai berikut:

1. Bagaimana algoritma K-Means digunakan untuk membagi perilaku belanja online siswa berdasarkan atribut perilaku?
2. Berapa banyak cluster optimal yang dihasilkan menggunakan evaluasi Silhouette Coefficient?
3. Bagaimana karakteristik masing-masing cluster perilaku belanja online siswa dibentuk.

Diharapkan bahwa penelitian ini akan meningkatkan pemahaman kita tentang perilaku belanja mahasiswa. Ini juga akan menjadi referensi untuk strategi layanan *e-commerce*.

2. TINJAUAN PUSTAKA DAN TEORI PENUNJANG

2.1 Tinjauan Pustaka

Berikut adalah hasil dari tinjauan literatur yang telah dilakukan pada berbagai jurnal penelitian yang berkaitan dengan topik metode *Clustering K-Means*:

Penelitian [11] menunjukkan bahwa data mining, khususnya metode clustering, dapat dengan mudah menemukan pola tersembunyi dalam data perilaku pengguna. Dengan mengelompokkan data berdasarkan kemiripan fitur, teknik ini dapat menemukan kecenderungan waktu aktivitas, preferensi, dan pola interaksi pengguna. Hasilnya menunjukkan bahwa clustering adalah metode yang baik untuk memahami perilaku pengguna dalam berbagai konteks, seperti media sosial, pendidikan, dan bisnis.

Seperti yang ditunjukkan oleh sejumlah penelitian [12] yang menggunakan algoritma K-Means, algoritma K-Means memiliki kemampuan untuk menghasilkan segmentasi yang jelas dan mudah dipahami. Dalam dunia pendidikan, K-Means digunakan untuk mengelompokkan mahasiswa berdasarkan nilai akademik dan pencapaian pembelajaran mahasiswa. Ini juga menghasilkan nilai Silhouette Coefficient yang tinggi, yang menunjukkan kualitas cluster yang baik, menunjukkan bahwa K-Means efektif dalam menangani data tanpa label dan mampu memberikan informasi yang bermanfaat untuk pengambilan keputusan berbasis data. Namun, aspek perilaku konsumsi mahasiswa tidak dipertimbangkan dalam penelitian ini karena fokusnya lebih pada evaluasi hasil belajar.

K-Means telah digunakan dalam bidang bisnis dan manajemen persediaan untuk mengelompokkan produk berdasarkan tingkat permintaan dan potensi penjualan. Hasil clustering digunakan untuk mengoptimalkan strategi pemasaran dan pengelolaan stok. Meskipun demikian, penelitian [13] ini biasanya menggunakan data transaksi produk secara keseluruhan dan belum mengaitkannya dengan karakteristik konsumen individu khususnya, karena mahasiswa adalah kelompok pengguna e-commerce yang berbeda dari konsumen umum dalam hal pola konsumsi mereka.

Penelitian terbaru menunjukkan bahwa faktor-faktor seperti harga, promosi, kemudahan transaksi, waktu belanja, dan preferensi pembayaran digital memengaruhi perilaku belanja online mahasiswa. Dibandingkan dengan demografi konsumen lainnya,

mahasiswa lebih sensitif terhadap uang. Sebagian besar penelitian, bagaimanapun, masih bersifat deskriptif.

Untuk evaluasi model clustering, Silhouette Coefficient banyak digunakan karena mampu mengukur tingkat kohesi dan separasi antar cluster secara bersamaan. Penelitian sebelumnya menunjukkan bahwa Silhouette dapat memberikan dasar objektif untuk menentukan jumlah cluster yang ideal. Namun, metode ini masih sangat terbatas untuk menilai perilaku belanja online mahasiswa [3].

Penelitian ini menunjukkan bahwa algoritma K-Means telah banyak digunakan dalam berbagai bidang, tetapi belum banyak yang menggunakannya untuk menganalisis perilaku belanja online siswa dengan menggabungkan berbagai atribut perilaku dan evaluasi Silhouette. Oleh karena itu, penelitian ini membantu menerapkan K-Means untuk mengelompokkan perilaku belanja online siswa secara menyeluruh, sehingga hasil clustering dapat digunakan untuk membuat rencana untuk layanan e-commerce, promosi, dan pendidikan keuangan yang lebih tepat sasaran untuk mahasiswa .

2.2 Teori Penunjang

2.2.1 Data mining

Metode data mining digunakan untuk mengidentifikasi pola perilaku belanja mahasiswa dari data survei yang kompleks dan tidak berlabel. *Data mining* dapat dipahami sebagai proses menggali atau menemukan informasi baru melalui identifikasi pola dan aturan tertentu dari data dalam jumlah besar. Secara umum, *data mining* merupakan serangkaian teknik yang digunakan untuk memperoleh nilai dan pengetahuan tambahan yang sebelumnya belum diketahui secara manual dari suatu himpunan data. Proses ini juga sering disebut sebagai *knowledge discovery in databases* (KDD) [14].

2.2.2 Clustering

Clustering adalah proses mengelompokkan sejumlah besar data yang memiliki banyak kesamaan satu sama lain ke dalam kelompok atau klaster. Berdasarkan kemiripan karakteristik dalam perilaku belanja online, teori clustering membantu proses pengelompokan siswa. Tujuan *Clustering* adalah untuk memastikan bahwa data dalam satu klaster memiliki tingkat kemiripan yang paling tinggi dan data di antaranya memiliki tingkat kemiripan yang paling rendah. *Clustering* juga dapat dianggap sebagai teknik segmentasi data yang digunakan dalam sejumlah industry [15].

2.2.3 Algoritma *K-Means*

Algoritma K-Means dipilih karena kesederhanaannya dan kemampuan untuk membuat segmentasi data perilaku mahasiswa yang mudah dipahami. Algoritma K-Means merupakan metode pengelompokan non-hierarki yang diperkenalkan oleh Stuart Lloyd pada tahun 1984. Teknik ini digunakan untuk membagi sekumpulan data ke dalam dua atau lebih kelompok. *K-means* menghitung setiap *record* dari *Cluster* awal secara iteratif menggunakan *Euclidean distance*.

Untuk memastikan bahwa setiap fitur memiliki skala yang sebanding, tahap normalisasi data sangat penting sebelum proses clustering. Karena algoritma K-Means sensitif terhadap perbedaan skala data, normalisasi diperlukan. Akibatnya, atribut dengan rentang nilai yang lebih besar dapat mendominasi perhitungan jarak dan berdampak pada hasil pengelompokan. Oleh karena itu, semua variabel angka dinormalisasi terlebih dahulu menggunakan metode normalisasi Min–Max untuk memastikan bahwa mereka berada dalam rentang nilai yang sama, seperti persamaan (1).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Keterangan :
X' = nilai data hasil normalisasi
X = nilai data asli
$X_{\min}$ = nilai minimum dari suatu atribut
$X_{\max}$ = nilai maksimum dari suatu atribut

Proses dimulai dengan memilih k daftar awal sebagai pusat *Cluster* (atau *seed* awal) [15]. Adapun rumus yang digunakan, seperti persamaan (2).

$$\mathrm{d}(x_{i,}c_j) = \sqrt{\sum_{m=1}^{M}(X_{im} - C_{jm})^2} \quad (2)$$

Keterangan :
$\mathrm{d}(x_{i,}c_j)$ = Jarak antara titik data x_i ke c_j
M = Jumlah fitur
X_{im} = Nilai fitur ke $-m$ dari titik data x_i
X_{jm} = Nilai fitur ke $-m$ dari titik data c_j

2.2.4 Analisis data

Analisis data merupakan tahapan penting dalam penelitian yang bertujuan mengolah data mentah hasil pengumpulan menjadi informasi yang bermakna. Singkatnya, analisis data penelitian adalah proses yang menggunakan metode statistik atau kualitatif untuk mengubah data yang belum diolah menjadi informasi yang relevan dan bermakna. Tujuan analisis data adalah untuk menemukan tren, pola, atau hubungan

dalam data yang dapat digunakan untuk menguji teori atau menyelesaikan masalah penelitian. Penelitian berhasil dan validitas hasilnya bergantung pada pemilihan metode analisis yang tepat dan interpretasi data yang akurat [16].

3. METODE PENELITIAN

3.1 Uji Validitas dan Reliabilitas

Instrumen penelitian berupa kuesioner terlebih dahulu diuji validitas dan reliabilitasnya sebelum digunakan dalam proses pengumpulan data. Uji validitas dilakukan untuk memastikan bahwa setiap pertanyaan memiliki kemampuan untuk mengukur variabel perilaku belanja online siswa secara akurat. Untuk menguji validitas item pernyataan, uji korelasi Pearson Product Moment digunakan. Nilai r hitung dan r tabel dibandingkan pada taraf signifikansi 0,05. Apabila nilai r hitung lebih besar dari nilai r tabel, item pernyataan dianggap valid, seperti Tabel I. hasil uiji validitas intrumen kuisioner.

Tabel I. Hasil Uji Validitas Instrumen Kuesioner

Indikator Pernyataan	r hitung	r tabel	Keterangan
Frekuensi belanja online mahasiswa	0,612	0,138	Valid
Jenis produk yang sering dibeli	0,584	0,138	Valid
Waktu belanja online	0,701	0,138	Valid
Metode pembayaran yang digunakan	0,667	0,138	Valid
Faktor harga dalam keputusan belanja	0,724	0,138	Valid
Pengaruh promosi dan diskon	0,689	0,138	Valid

Uji reliabilitas juga dilakukan untuk mengetahui seberapa konsisten instrumen mengukur variabel penelitian. Metode Cronbach's Alpha digunakan untuk menguji reliabilitas instrument, instrumen dianggap reliabel jika nilai Cronbach's Alphanya lebih besar dari 0,70, yang menunjukkan bahwa kuesioner memiliki tingkat keandalan yang tinggi. seperti Tabel II. hasil uiji reliabilitas intrumen kuisioner.

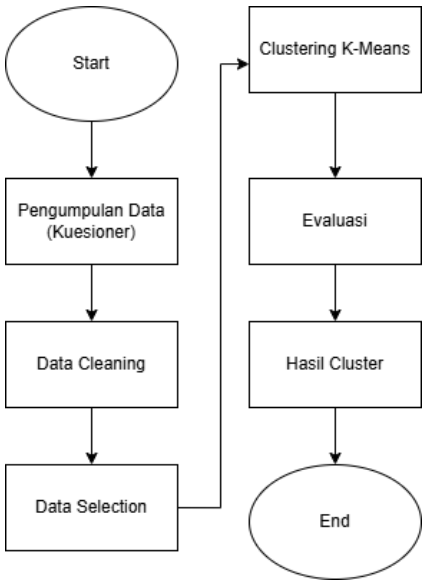
Tabel II. Hasil Uji Reliabilitas Instrumen Kuesioner

Variabel	Jumlah Item	Cronbach's Alpha	Keterangan
Perilaku Belanjan Online Mahasiswa	6	0,812	Reliabel

Berdasarkan hasil uji validitas, setiap item pernyataan memiliki nilai r hitung yang lebih besar daripada r tabel, yang menunjukkan bahwa semua item pertanyaan valid. Selain itu, hasil uji reliabilitas menunjukkan instrumen kuesioner memiliki nilai alfa Cronbach sebesar 0,812, yang menunjukkan bahwa itu layak untuk digunakan dalam penelitian.

3.2 Proses Data Mining

Penelitian ini dimulai dari pengumpulan data melalui kuesioner *online*, setelah itu dilakukan data *cleaning*, data *selection*, pemodelan *Cluster*ing, evaluasi model, dan terakhir hasil *Cluster*. Berikut merupakan *flowchart* dari metode penelitian pada Gambar 1.



Gambar 1. *Flowchart* Metode Penelitian

3.3 Peralatan Penelitian

Pembuatan model Clustering menggunakan Python. Bahasa pemrograman ini memberikan dukungan yang kuat dalam penerapan proses data mining dan menggunakan library yang sudah tersedia untuk menghubungkan model Clustering, seperti library untuk mengevaluasi kinerja dan sklearn.Cluster. KMeans [17].

3.4 *Datasheet*

Datasheet dalam penelitian ini diperoleh dari hasil pengisian kuesioner online yang disebarkan kepada 200 mahasiswa yang masih aktif. *Datasheet* berisi informasi mengenai karakteristik perilaku belanja *online* mahasiswa dan terdiri dari 200 *record* dengan 6 atribut utama. Untuk mewakili aspek utama perilaku belanja mahasiswa, 6 fitur dipilih. Ini juga dilakukan untuk mengimbangi antara kelengkapan informasi dan interpretabilitas hasil clustering. Data berisi jawaban responden berdasarkan pengalaman dalam melakukan transaksi belanja *online*.

Atribut pada dataset meliputi:

a. Jenis kelamin, terdiri dari kategori laki-laki dan perempuan.
b. Waktu terakhir berbelanja *online*, berupa pilihan waktu seperti pagi, siang, sore, atau malam.
c. Jenis produk terakhir yang dibeli, misalnya fashion, elektronik, kecantikan, kebutuhan rumah, dan lainnya.
d. Rata-rata pengeluaran terakhir berbelanja *online*, berupa sangat rendah, rendah, sedang, tinggi, dan sangat tinggi.
e. Metode pembayaran, meliputi pilihan pembayaran seperti *e-Wallet* , Transfer Bank, COD (*Cash on Delivery*), Kartu Kredit, dan lainnya.
f. Faktor yang memengaruhi keputusan belanja, seperti harga, promo/diskon, ulasan pengguna, dan lainnya.

3.5 Model *Clustering*

Dalam membangun sebuah model *clustering*, dapat diterapkan berbagai algoritma seperti *K-Means*, *K-Medoids*, dan *DBSCAN* dapat digunakan [18]. Model *Clustering* merupakan salah satu teknik dalam data mining yang berfungsi untuk mengelompokkan data ke dalam beberapa grup berdasarkan kemiripan karakteristik. Salah satu metode yang umum dipakai untuk membentuk kelompok tersebut adalah algoritma K-Means.

3.6 Evaluasi *Clustering*

Metode *Silhouette* untuk evaluasi model digunakan untuk melakukan proses penentuan banyak kelompok ideal.

3.7 Metode *Silhouette*

Metode silhouette coefficient menggabungkan dua metode yaitu metode cohesion, yang menentukan seberapa dekat satu objek dengan objek lain dalam suatu Cluster, dan metode separation yang menentukan seberapa jauh suatu Cluster terpisah dari Cluster lain [19].

4. HASIL DAN PEMBAHASAN

4.1 Data Cleaning

Dalam tahap pembersihan data, diterapkan dua langkah pokok untuk menjamin mutu dataset sebelum diterapkan pada pemodelan. Langkah awal melibatkan penghapusan data duplikat *(drop duplicates*) dengan tujuan membuang baris berulang yang berpotensi menimbulkan bias serta distorsi pada hasil analisis. Setelah langkah ini, ukuran data berkurang menjadi 154 baris dan 6 atribut, yang menandakan bahwa entri duplikat telah berhasil dihapus**.** Hasil penghapusan data duplikat ditampilkan pada Gambar 2.

```
df = df.drop_duplicates()
print(f"Jumlah data setelah menghapus duplikasi: {df.shape}")
```

```
Jumlah data setelah menghapus duplikasi: (154, 6)
```

Gambar 2. Penghapusan Data Duplikat

Langkah selanjutnya adalah penghapusan baris dengan missing value (*drop*na) untuk memastikan setiap atribut pada setiap rekaman terisi secara penuh. Akibatnya, jumlah data tetap 154 baris, yang mengindikasikan bahwa dataset tidak lagi mengandung nilai kosong setelah duplikasi dikeluarkan. Hasil penghapusan baris dengan missing value ditampilkan pada Gambar 3.

```
df = df.dropna()
print(f"Jumlah data setelah menghapus missing value: {df.shape}")
```

```
Jumlah data setelah menghapus missing value: (154, 6)
```

Gambar 3. Penghapusan *Missing Value*

Proses pembersihan ini menekankan pentingnya akurasi, konsistensi, dan kelengkapan data sebelum melanjutkan ke tahap preprocessing serta analisis lanjutan menggunakan algoritma *K-Means* [20].

4.2 Data Selection

Proses pemilihan data bertujuan untuk menetapkan atribut maupun cakupan data yang akan dipakai pada proses *data mining* [21]. Proses ini meliputi pengecekan type atribut, menghapus atribut yang tidak digunakan, dan merubah type atribut.

Untuk mengidentifikasi jenis data pada atribut, metode pengecekan type attribute digunakan. Dalam proses pengecekan ini, jenis kelamin, waktu, produk, rata rata, pembayaran, dan faktor diidentifikasi.

Attribute jenis kelamin, waktu, produk, rata rata, pembayaran, dan faktor ini diubah menjadi data numerik menggunakan metode *One-Hot Encoding*.

Menggunakan perintah encoded_df untuk mengubah atribut kategorikal menjadi numerik dapat mempercepat proses *data mining*. Jenis kelamin, waktu, produk, rata rata, pembayaran, dan faktor diubah, dan *One-Hot Encoding* mengubah setiap kategori menjadi kolom biner (0/1) dan dapat melanjutkan proses *Clustering*.

4.3 Data Mining

Tahap *data mining* dalam penelitian ini menekankan pada proses pemodelan klasterisasi dengan menggunakan algoritma *K-Means*. Sebelum pemodelan tersebut dilaksanakan, semua atribut kategorikal dalam dataset dikonversi ke format numerik melalui metode *One-Hot Encoding*, Konversi ini bertujuan untuk menggambarkan setiap kategori dalam ruang vektor, sehingga algoritma *K-Means* dapat menghitung jarak antara data. Gambar 4 menunjukkan hasil konversi melalui metode *One-Hot Encoding*.

```
encoder = OneHotEncoder(sparse_output=False)
encoded_array = encoder.fit_transform(selected_columns)

encoded_df = pd.DataFrame(encoded_array, columns=encoder.get_feature_names_out())
print("Dimensi fitur setelah encoding:", encoded_df.shape)
print("Contoh nama fitur:", encoded_df.columns[:15])

Dimensi fitur setelah encoding: (154, 51)
Contoh nama fitur: Index(['JENIS KELAMIN_Laki - Laki', 'JENIS KELAMIN_Perempuan',
       'Waktu_Malam (19.00-23.00)', 'Waktu_Pagi (06.00-10.00)',
       'Waktu_Siang (11.00-15.00)', 'Waktu_Sore (16.00-18.00)',
       'Produk_Aksesoris HP', 'Produk_Buku', 'Produk_Casing Handphone',
       'Produk_Diecast mini gt', 'Produk_Elektronik / Gadget',
       'Produk_Fashion (pakaian, sepatu, aksesoris)', 'Produk_Jajan',
       'Produk_Kebutuhan rumah / kos', 'Produk_Kecantikan / Perawatan diri'],
      dtype='object')
```

Gambar 4. Hasil Konversi

Setelah seluruh data telah ditransformasikan dengan sukses, langkah berikutnya melibatkan penerapan algoritma *K-Means* untuk membentuk kelompok. Dalam tahap ini, algoritma dimulai dengan menginisialisasi *centroid*, kemudian menghitung jarak antara titik-titik data, dan secara berulang memperbarui posisi *centroid* hingga model mencapai kondisi konvergensi. Proses tersebut menghasilkan label kelompok awal untuk setiap entri dalam dataset. Gambar 5 dan 6 menunjukkan proses pemodelan *K-Means*.

```
=== Hasil Pengelompokan (K=2) ===
   Cluster
0        0
1        1
2        0
3        1
4        1

Jumlah anggota setiap cluster:
Cluster
0    89
1    65
```

Gambar 5. Permodelan K=2

```
k = 3
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(encoded_df)
df['Cluster'] = kmeans.labels_

print("\nHASIL CLUSTER (K=3):")
display(df[['Cluster']].head())
```

HASIL CLUSTER (K=3):

	Cluster
0	2
1	0
2	2
3	1
4	0

Gambar 6. Permodelan K=3

4.4 Evaluasi Model

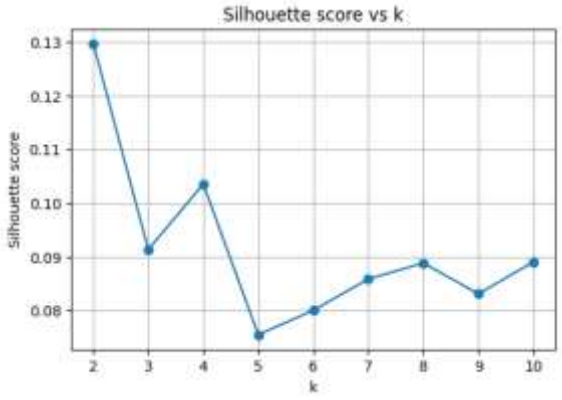
Untuk menentukan pengelompokan data yang ideal, evaluasi model *Clustering* digunakan. Pada penelitian ini, evaluasi *Clustering silhouette* digunakan. Evaluasi ini digunakan untuk menentukan akurasi yang optimal dari proses clustering dan memberikan rekomendasi K untuk hasilnya [12].

4.5 Evaluasi Metode *Silhoutte*

Evaluasi dilakukan dengan menerapkan metode *Silhouette* melalui pengujian berbagai jumlah *Cluster* dan perbandingan skor *Silhouette* untuk setiap nilai k. Metode ini dimaksudkan untuk mengidentifikasi jumlah *Cluster* yang paling ideal dengan memilih skor *Silhouette* yang paling tinggi, sebab skor yang lebih tinggi menunjukkan pemisahan antar *Cluster* yang lebih baik. Hasil evaluasi tersebut disajikan dalam Gambar 7 dan Gambar 8, yang masing-masing menunjukkan pengujian skor *Silhouette* untuk model dengan jumlah *Cluster* yang bervariasi.

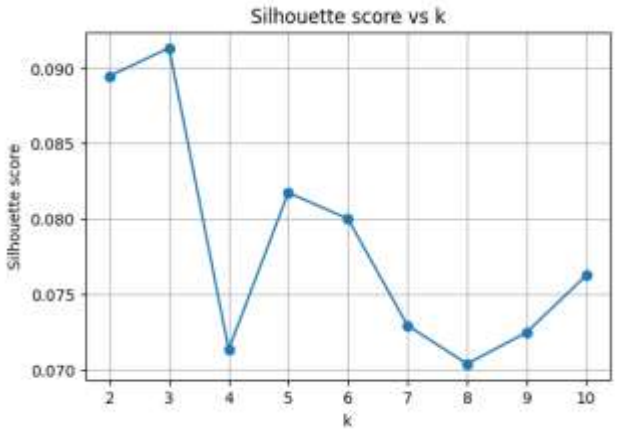
Dalam pengujian awal, analisis *Silhouette* dilakukan menggunakan dua *Cluster* (k = 2), yang hasilnya divisualisasikan pada Gambar 7. Perhitungan menunjukkan bahwa skor *Silhouette* untuk pengaturan dua *Cluster* mencapai 0.0894, yang termasuk rendah. Nilai tersebut menunjukkan bahwa, meskipun algoritma K-Means berhasil membentuk dua kelompok data, tingkat pemisahan antar cluster belum optimal dan masih ada kedekatan yang signifikan di antara data dari kelompok yang berbeda. Hal ini mencerminkan sifat data belanja online mahasiswa yang cukup homogen, sehingga batasan antar cluster tidak terbentuk dengan jelas. Skor *Silhouette* yang rendah ini menjadi salah satu batasan dalam penelitian, karena menggambarkan kualitas pemisahan cluster yang belum maksimal.

Namun, hasil pengelompokan tersebut tetap dianggap sah secara metodologis, sebab prosesnya berhasil mengenali pola dasar dan segmentasi awal berdasarkan kesamaan atribut data. Nilai *Silhouette* yang rendah menunjukkan keterbatasan dalam kualitas pemisahan cluster, bukan kegagalan pada metode clustering yang diterapkan. Oleh karena itu, hasil ini masih berguna untuk analisis eksploratif dan dapat dijadikan landasan untuk melanjutkan pengujian dengan jumlah cluster yang berbeda guna mencapai konfigurasi pengelompokan yang lebih baik.



Gambar 7. Grafik *Silhouette* K=2

Pengujian selanjutnya dilakukan dengan menggunakan tiga *Cluster* (k = 3), sebagaimana yang diperlihatkan dalam Gambar 8. Evaluasi hasil menunjukkan bahwa skor *Silhouette* untuk pengaturan tiga *Cluster* mencapai 0.0913, angka yang lebih tinggi daripada skor untuk k = 2. Kenaikan skor ini mengindikasikan bahwa model dengan tiga *Cluster* menghasilkan struktur kelompok yang lebih efektif serta pemisahan antar kelompok yang lebih tajam. Representasi visual pada grafik menunjukkan perbaikan dalam kualitas pengelompokan, sehingga pengaturan tiga *Cluster* dinilai lebih sesuai untuk merepresentasikan pola keseluruhan data.



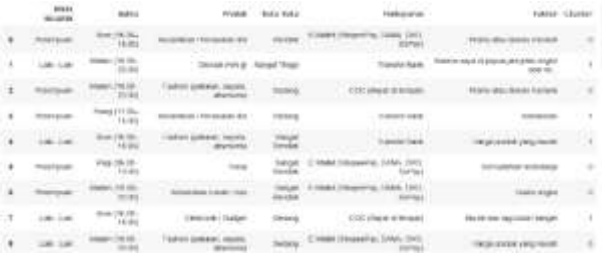
Gambar 8. Grafik *Silhouette* K=3

4.6 Hasil Pengelompokkan

Hasil evaluasi metode *silhouette* menunjukkan bahwa pengelompokan K=2 dan K=3 adalah yang terbaik. Hasil ini dapat digunakan sebagai dasar untuk mengorganisir datasheet yang digunakan. Nama kategori dapat diklasifikasikan dengan K=2. Hasil dari *Cluster* 0 menunjukkan bahwa laki-laki biasanya berbelanja di siang atau sore hari, mencari harga murah, dan lebih sering menggunakan transfer bank sebagai metode pembayaran dan hasil dari *Cluster* 1 menunjukkan bahwa perempuan memiliki pengeluaran sedang, sering membeli barang fashion, berbelanja di sore atau malam hari, dan dipengaruhi oleh promosi dan menggunakan dompet digital sebagai metode pembayaran utama. Sedangkan untuk K=3, hasil dari *Cluster* 0 laki-laki mencari harga murah, belanja malam, dan menggunakan transfer bank. *Cluster* 1 wanita mengeluarkan uang sedang, membeli pakaian di sore hari, dan terpengaruh oleh promo. *Cluster* 2 wanita belanja malam, mengeluarkan uang rendah, suka promo, dan menggunakan dompet digital sebagai cara pembayaran utama.

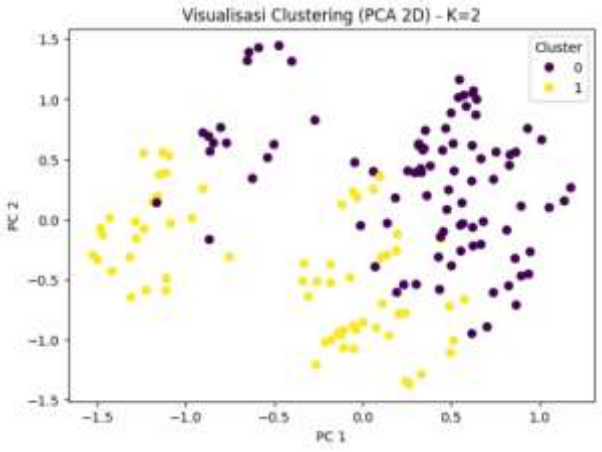
Interpretasi dari hasil pengelompokan ini sesuai dengan konsep segmentasi pasar yang didasarkan pada faktor demografis dan perilaku (behavioral segmentation), yaitu pendekatan di mana konsumen diklasifikasikan menurut atribut pribadi, pola konsumsi, serta reaksi mereka terhadap respons pemasaran. Temuan tersebut diperkuat oleh penelitian yang menjelaskan bahwa algoritma K-Means efektif dalam menemukan kelompok konsumen dengan pola perilaku yang cukup seragam, walaupun pemisahan antarsegmen tidak selalu bersifat mutlak

[22]. Oleh karena itu, hasil clustering dalam penelitian ini mengindikasikan bahwa pola belanja online mahasiswa sering kali tumpang tindih dan merefleksikan segmentasi umum pada konsumen digital. Hal ini menunjukkan bahwa perbedaan perilaku belanja mahasiswa lebih dipengaruhi oleh faktor situasional dan preferensi pribadi daripada perbedaan demografis yang signifikan. Gambar 9 dan 10 menunjukkan hasil pengelompokan dalam bentuk tabel, dan gambar 11 dan 12 menunjukkan hasil pengelompokan dalam bentuk visual

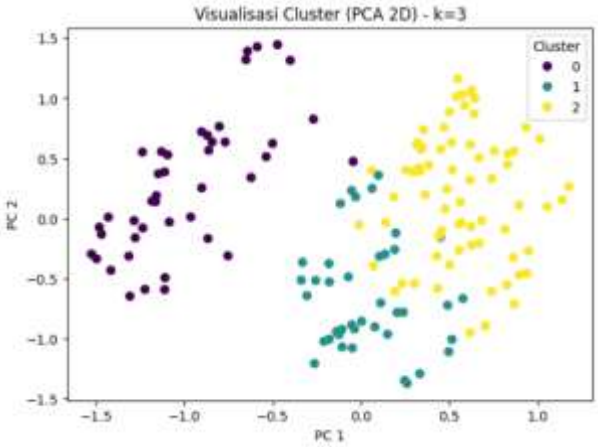


Gambar 9. Data K=2

Gambar 10. Data K=3



Gambar 11. Visualisasi *Clustering* K=2



Gambar 12. Visualisasi *Clustering* K=3

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian diatas maka dapat disimpulkan *datasheet* yang berfokus pada kategori tertentu dapat dipetakan dengan menggunakan metode pengelompokan. Hasil penelitian yang dilakukan untuk mengelompokkan *datasheet* menunjukkan banyak pengelompokan K=2 dan K=3. Proses menentukan pengelompokan K=2 dan K=3 dengan karakteristik *Cluster* 0 menunjukkan bahwa laki-laki biasanya berbelanja di siang atau sore hari, mencari harga murah, dan lebih sering menggunakan transfer bank sebagai metode pembayaran dan hasil dari *Cluster* 1 menunjukkan bahwa perempuan memiliki pengeluaran sedang, sering membeli barang fashion, berbelanja di sore atau malam hari, dan dipengaruhi oleh promosi dan menggunakan dompet digital sebagai metode pembayaran utama. Sedangkan untuk K=3, hasil dari *Cluster* 0 laki-laki mencari harga murah, belanja malam, dan menggunakan transfer bank. *Cluster* 1 wanita mengeluarkan uang sedang, membeli pakaian di sore hari, dan terpengaruh oleh promo. *Cluster* 2 wanita belanja malam, mengeluarkan uang rendah, suka promo, dan menggunakan dompet digital sebagai cara pembayaran utama. Pengujian evaluasi kinerja dengan menggunakan metode *silhouette* menghasilkan pengelompokan terbaik.

Kontribusi ilmiah penelitian ini terletak pada penerapan metode data mining untuk mengidentifikasi segmentasi perilaku pembelian daring mahasiswa yang bersifat laten dan sulit diamati secara langsung, sehingga memperdalam pemahaman analisis perilaku konsumen pada generasi muda yang terbiasa dengan teknologi digital. Meskipun nilai *Silhouette* yang diperoleh tergolong rendah, hasil pengelompokan tetap dinilai valid secara metodologis karena mampu mengungkap pola segmentasi dasar yang bermakna

dan berfungsi sebagai kajian awal terhadap karakteristik data yang cenderung homogen. Adapun keterbatasan penelitian ini meliputi jumlah responden yang masih terbatas, penggunaan variabel yang relatif sederhana, serta rendahnya nilai *Silhouette* yang menunjukkan keterbatasan kualitas pemisahan kelompok, sehingga hasil pengelompokan yang diperoleh lebih bersifat eksploratif dan ditujukan untuk memberikan gambaran awal mengenai segmentasi perilaku pembelian online mahasiswa.

5.2 Saran

Berdasarkan hasil penelitian, penulis membuat saran sebagai berikut.

a. Penelitian selanjutnya harus melibatkan lebih banyak responden dan variabel tambahan seperti frekuensi belanja, jenis platform *e-commerce*, dan faktor sosial.
b. Penelitian selanjutnya dapat membandingkan *K-Means* dengan algoritma lain seperti *K-Medoids*, *DBSCAN*, atau Pengelompokan Hierarki untuk mendapatkan hasil pengelompokan yang lebih baik.
c. Penelitian selanjutnya harus melakukan standarisasi dan normalisasi agar perbedaan skala tidak mempengaruhi proses pembentukan *Cluster*.
d. Penelitian selanjutnya dapat menggunakan metode evaluasi seperti Davies-Bouldin Index atau Calinski-Harabasz Index selain menggunakan *Silhouette* Coefficient.

DAFTAR PUSTAKA

[1] N. L. Hasibuan and A. Al Fauzan, "Perilaku Konsumtif Mahasiswa Di Era Digital Pada Penggunaan E- Wallet Dan E-Commerce," vol. 3, no. 6, pp. 301–311, 2025.
[2] M. Fajar, S. Adam, B. Putra, S. I. Puteri, A. Fajrissiddiq, and L. Sani, "Eksplorasi dan Analisis Data Mining untuk Prediksi Pola Konsumen Menggunakan Teknik Klasifikasi dan Clustering," 2025.
[3] J. Teknologi, E. Febrianty, L. Awalina, and W. I. Rahayu, "Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail Optimizing Marketing Strategies with Customer Segmentation Using K-Means Clustering on Online Retail Transactions," vol. 13, no. September, pp. 122–137, 2023.
[4] M. Sholeh *et al.*, "PERBANDINGAN EVALUASI METODE DAVIES BOULDIN , ELBOW DAN SILHOUETTE PADA MODEL CLUSTERING DENGAN," vol. 8, no. 1, 2023.
[5] N. A. Yolandari, L. E. Butarbutar, G. Citra, and H. Rajagukguk, "ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX," vol. 13, no. 3, 2025.
[6] T. Abdulpatah and B. N. Sari, "ALGORITMA K-MEANS DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING UNTUK PENGELOMPOKAN DAERAH PENGHASIL PADI," vol. 13, no. 3, 2025.
[7] D. Fitriyani and M. Jajuli, "IMPLEMENTASI ALGORITMA K-MEANS UNTUK KLASTERISASI DALAM PENGELOLAAN PERSEDIAAN OBAT (STUDI KASUS : APOTEK NAZA)," vol. 12, no. 3, pp. 2841–2848, 2024.
[8] T. T. Alifa *et al.*, "IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITMA K- MEANS," vol. 8, no. 1, pp. 602–607, 2024.
[9] D. Sartika *et al.*, "Fenomena Penggunaan E-Commerce terhadap Perilaku Konsumsi Mahasiswa," no. 3, 2024.
[10] M. R. Syahkur and D. Hartama, "Evaluasi Jumlah Cluster pada Algoritma K-Means ++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita," vol. 13, no. 3, pp. 487–496, 2024.
[11] J. M. Polgan *et al.*, "Penggunaan Teknik Data Mining untuk Analisis Perilaku Pengguna pada Media Sosial," vol. 13, pp. 1074–1078, 2024.
[12] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," vol. 3, pp. 46–56, 2024.
[13] S. Pujiono, R. Astuti, and F. M. Basysyar, "IMPLEMENTASI DATA MINING UNTUK MENENTUKAN POLA PENJUALAN PRODUK MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING," vol. 8, no. 1, 2024.
[14] R. D. Hadisaputro and A. Zubaidi, "FREQUENT ITEMSET MINING PADA ARTIKEL COVID-19 MENGGUNAKAN WEB CRAWLING DAN ALGORITMA FP- GROWTH (Frequent Itemset Mining On Covid-19 Articles Using Web Crawling And Fp-Growth Algorithm)," vol. 4, no. 2, pp. 242–252, 2022.
[15] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018 / 2019," vol. 4, pp. 51–58, 2020.
[16] P. C. Susanto, D. U. Arini, L. Yuntina, and J. Panatap, "Konsep Penelitian Kuantitatif : Populasi

, Sampel , dan Analisis Data (Sebuah Tinjauan Pustaka),” vol. 3, no. 1, pp. 1–12, 2024.

[17] E. Retnoningsih and R. Pramudita, “Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python,” vol. 7, no. 2, pp. 156–165, 2020.

[18] R. Adha, N. Nurhaliza, U. Soleha, P. Studi, S. Informasi, and F. Sains, “Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia,” vol. 18, no. 2, pp. 206–211, 2021.

[19] K. P. Simanjuntak and U. Khaira, “Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering,” vol. 1, no. April, pp. 7–16, 2021.

[20] B. G. Sudarsono and M. I. Leo, “ANALISIS DATA MINING DATA NETFLIX MENGGUNAKAN APLIKASI RAPID MINER ANALYSIS DATA MINING NETFLIX DATA USING THE RAPID MINER,” vol. 4, no. 1, pp. 13–21, 2021.

[21] L. Canchen, “Preprocessing Methods and Pipelines of Data Mining : An Overview,” no. June, pp. 1–7, 2019.

[22] S. Konsumen, B. Online, and B. Karakteristik, “Indonesia Economic Journal,” vol. 1, no. 2, pp. 1926–1933, 2025.