ANALISIS KOMPARASI ALGORITMA K-MEANS DAN HIERARCHICAL CLUSTERING UNTUK SEGMENTASI PELANGGAN DALAM MENGGUNAKAN METODE PEMBAYARAN

Rifqi Muhammad Rifat, M Andre Yunizar, Hori Fahmi, Winda Kurnia Sari, Ken Ditha Tania, Fathoni
Sistem Informasi, Universitas Sriwijaya
Jl. Raya Palembang - Prabumulih No. KM. 32, Indralaya Indah, Kabupaten Ogan Ilir, Sumatera Selatan
09031282328062@student.unsri.ac.id

ABSTRAK

Transformasi digital dalam industri ritel menuntut pemanfaatan data transaksi melalui *machine learning* untuk memahami perilaku belanja pelanggan secara tepat. Namun, keterbatasan dalam mengenali profil pelanggan berdasarkan pola pembayaran digital sering kali menyebabkan strategi pemasaran tidak tepat sasaran dan kurang personal. Penelitian ini bertujuan melakukan analisis komparatif antara algoritma *K-Means* dan *Hierarchical Clustering* untuk menentukan metode segmentasi pelanggan yang paling akurat berdasarkan metode pembayaran. Penelitian ini menggunakan pendekatan kuantitatif pada 3900 data sekunder, kualitas klaster dievaluasi melalui metrik *Silhouette Score*, *Calinski-Harabasz*, dan *Davies-Bouldin Index*. Hasil penelitian menunjukkan bahwa kedua algoritma mampu membentuk lima klaster pelanggan dengan kualitas pemisahan yang sangat baik. Secara spesifik, *K-Means* menunjukkan kinerja sedikit lebih unggul dengan nilai *Silhouette Score* 0,8606 dibandingkan *Hierarchical Clustering* sebesar 0,8595. Temuan ini berhasil memetakan profil pelanggan berdasarkan metode pembayaran dominan seperti *Debit Card*, *PayPal, dan Cash*, yang dapat menjadi landasan strategis bagi perusahaan dalam merancang strategi pemasaran yang lebih personal serta tepat sasaran.

Kata kunci : *K-means, Hierarchical, Komparasi, Segmentasi, Shopping Behavior, Paymment Method*

1. PENDAHULUAN

Fenomena transformasi digital telah mengubah wajah industri ritel secara global, di mana volume data transaksi tumbuh secara eksponensial seiring dengan adopsi teknologi digital yang masif. Dalam pasar yang semakin kompetitif dan terintegrasi secara global, data bukan lagi sekadar catatan administratif, melainkan aset strategis yang menentukan daya saing perusahaan. Di era pesatnya persaingan industri ritel saat ini, keberhasilan perusahaan dalam menghadapi tantangan bisnis sangat dipengaruhi oleh kemampuan mereka dalam mengekstraksi nilai dari data transaksi yang dihasilkan melalui sistem *Point of Sales* (POS). Teknologi ini kini berfungsi lebih dari sekadar alat pencatat transaksi, POS telah bertransformasi menjadi sumber informasi primer untuk memahami perilaku belanja pelanggan dalam skala besar. Sejalan dengan temuan dalam literatur, pemanfaatan *machine learning* memegang peran vital dalam membedah pola konsumsi pelanggan yang sering kali kompleks dan sulit diprediksi secara konvensional [1].

Namun, tantangan nyata muncul ketika perusahaan belum sepenuhnya menyadari pola perilaku pelanggan secara mendalam, termasuk dalam hal preferensi metode pembayaran, yang sering kali menyebabkan strategi pemasaran menjadi tidak tepat sasaran dan kurang personal. Melalui pendekatan berbasis data yang presisi, perusahaan dapat memiliki fondasi yang kuat untuk mengidentifikasi karakteristik pelanggan secara lebih akurat. Hal ini memungkinkan perancangan program pemasaran yang tepat sasaran, yang pada akhirnya berfungsi untuk memperkuat loyalitas pelanggan untuk jangka waktu panjang di tengah dinamika pasar yang terus berubah.

Meskipun transformasi digital menyediakan data yang melimpah, tantangan utama muncul dalam mengekstraksi wawasan yang relevan dari kompleksitas data tersebut. Kehadiran berbagai metode pembayaran digital saat ini memang mempermudah proses transaksi, namun di sisi lain, perusahaan sering kali terjebak pada penggunaan strategi pemasaran yang bersifat generalis. Strategi pemasaran sering kali tidak tepat sasaran karena perusahaan belum mampu membedah secara mendalam bagaimana pola perilaku pelanggan dalam membayar sebenarnya sangat dipengaruhi oleh variabel demografis, khususnya kelompok usia dan gender pelanggan. Permasalahan ini sejalan dengan apa yang ditemukan dalam riset sebelumnya, di mana keterbatasan dalam memahami pola transaksi mendalam menjadi hambatan nyata dalam menciptakan profil pelanggan yang akurat [2].

Tanpa identifikasi pola yang presisi, perusahaan akan sulit memberikan layanan yang benar-benar personal dan gagal menangkap preferensi unik dari setiap segmen konsumen. Kegagalan dalam melakukan segmentasi yang tajam berdasarkan metode pembayaran ini tidak hanya berdampak pada efisiensi biaya pemasaran, tetapi juga pada hilangnya peluang untuk membangun interaksi yang lebih intim dengan pelanggan. Hal inilah yang mendasari perlunya evaluasi terhadap metode pengelompokan data yang sudah ada guna menemukan pendekatan yang lebih efektif dalam memetakan karakteristik subjek penelitian secara spesifik.

Peneliti terdahulu telah berupaya mengenali profil pelanggan melalui metode segmentasi. Penerapan model RFM (*Recency, Frequency, Money*) adalah salah satu contoh sukses yang dikombinasikan dengan algoritma *K-Means* untuk memahami perilaku para pengguna transportasi umum [3].

Selain itu, penelitian lainnya menunjukkan bahwa data transaksi yang lebih detail, seperti contoh uang muka dan angsuran cicilan, mampu memberikan gambaran yang sangat jelas dalam melihat profil risiko pelanggan [4].

Studi-studi ini membuktikan bahwa pendekatan *data mining* efektif untuk membedah karakteristik subjek penelitian secara spesifik.

Meskipun berbagai penelitian telah memanfaatkan metode segmentasi pelanggan, sebagian besar studi lebih berfokus pada analisis frekuensi pembelian dan nilai transaksi. Penelitian yang secara khusus membandingkan performa algoritma *clustering* dalam mengelompokkan perilaku pelanggan berdasarkan metode pembayaran masih relatif terbatas. Oleh karena itu, diperlukan kajian komparatif untuk mengetahui algoritma yang lebih efektif dalam mengidentifikasi pola transaksi pelanggan.

Penelitian ini dilakukan sebagai upaya menutup celah tersebut melalui perbandingan antara algoritma *K-means* dan *Hierarchical Clustering* dalam mengelompokkan perilaku pelanggan dalam bertransaksi. Fokus utama adalah mencari algoritma mana yang lebih efektif dalam mengelompokkan pola perilaku pelanggan berdasarkan cara mereka bertransaksi. Penelitian ini juga menekankan pentingnya pendekatan yang lebih mendalam untuk mengidentifikasi pola perilaku pelanggan yang tersembunyi. [5].

Tujuan utama penelitian ini adalah untuk melakukan analisis komparatif antara algoritma *K-Means* dan *Hierarchical Clustering* guna menentukan metode yang paling akurat dalam mengelompokkan perilaku transaksi pelanggan berdasarkan metode pembayaran. Melalui evaluasi segmentasi menggunakan metrik *Silhouette Score* dan *Calinski-Harabasz Index*, penelitian ini diharapkan dapat memberikan kontribusi ilmiah berupa model pengelompokan pola inklusi digital yang lebih tajam dan teruji secara saintifik. Hasil penelitian ini diharapkan menjadi landasan strategis bagi industri ritel dalam memahami interaksi transaksi pelanggan secara mendalam, sehingga perusahaan dapat merancang program pemasaran yang lebih personal, akurat, dan tepat sasaran.

2. TINJAUAN PUSTAKA

Clustering adalah metode yang digunakan untuk memisahkan kumpulan data menjadi beberapa kategori sesuai dengan kemiripan antar data. Berdasarkan riset yang dilakukan P. Aselnino and A. W. Wijayanto, tujuan utama *clustering* adalah untuk mengelompokkan informasi berdasarkan kesamaan sifat-sifatnya dan tidak memerlukan penamaan atau hasil yang ditargetkan sehingga melakukan proses validasi dapat mengukur seberapa baik metode pengelompokkan [6].

Melalui proses ini, data yang sebelumnya tidak teratur dapat disusun ke dalam kelompok-kelompok tertentu sehingga pola atau struktur yang terpendam dalam data menjadi lebih mudah dipahami dan dianalisis.

Algoritma *K-Means* merupakan salah satu metode *clustering* yang banyak digunakan dalam teknik *unsupervised learning* karena kemampuannya dalam mengelompokkan data secara efisien. Prinsip kerja algoritma ini cukup sederhana namun efektif, yakni membagi data dalam beberapa kelompok (k) berdasarkan kedekatannya dengan dengan titik pusat klaster (*centroid*) [7].

Melalui proses ini, *k-means* mampu memastikan data yang berada di satu kelompok secara rapi berdasarkan kesamaan karakteristiknya. *K-Means* memiliki kelebihan dalam menghasilkan pengelompokan yang presisi di mana algoritma ini terdefinisi dengan baik dalam memetakan data layanan publik yang jelas antar-kelompok [3].

Hierarchical Clustering adalah teknik pengelompokan data yang menciptakan struktur kluster dalam bentuk bertingkat atau hierarkis. Teknik ini beroperasi dengan cara menggabungkan atau memisahkan data secara bertahap sesuai dengan derajat kesamaan antara objek-objek. Proses ini menghasilkan sebuah struktur yang menyerupai pohon yang dinamakan dendrogram, yang berfungsi untuk menampilkan hubungan antar data serta mengidentifikasi jumlah kluster yang ada. Metode ini memberi kesempatan kepada peneliti untuk memahami lebih dalam mengenai struktur data karena pengelompokan dilakukan secara bertahap hingga membentuk kelompok yang lebih khusus. Walaupun *K-Means* unggul dari sisi efisiensi komputasi, namun pendekatan *Hierarchical* memberikan kelebihan dalam pemahaman struktur data yang lebih mendalam. Penelitian yang dilakukan S. B. W. Rizky menunjukkan *Hierarchical Clustering* menghasilkan pengelompokan yang lebih stabil dibandingkan *K-Means* [8].

Penelitian terdahulu implementasi algoritma *clustering* dalam riset sebelumnya telah membuktikan efektivitas pendekatan data mining dalam membedah karakteristik subjek secara spesifik di berbagai sektor. Penerapan model RFM (*Recency, Frequency, Monetary*) yang dikombinasikan dengan K-Means telah sukses memetakan perilaku pengguna transportasi umum serta mendeteksi risiko churn pelanggan [3].

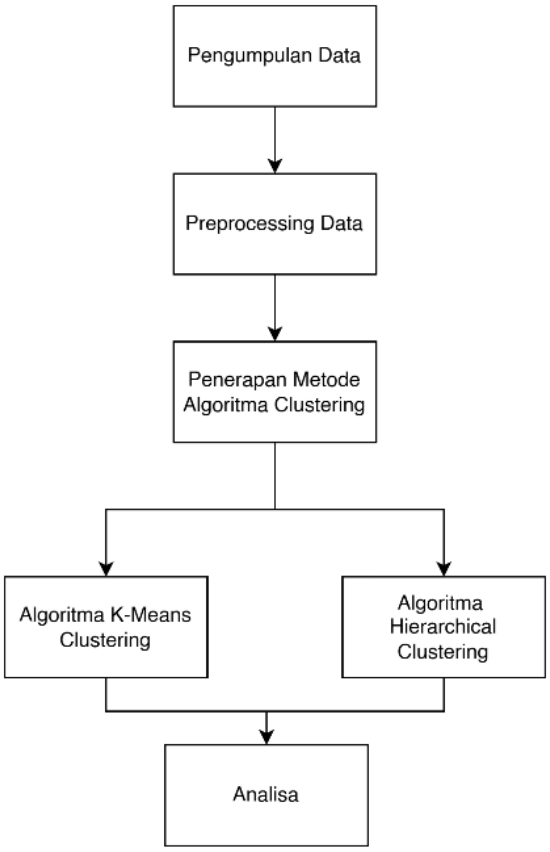
Evolusi model ini terus berkembang melalui kerangka LRFMV dan DCV, yang menunjukkan bahwa segmentasi yang baik harus selalu adaptif terhadap dinamika perilaku pasar yang dinamis [5], [10].

Di sektor ritel, penggunaan *machine learning* terbukti mampu meningkatkan strategi pemasaran PT Maspion menjadi lebih personal, sementara analisis transaksi yang mendalam pada nasabah kartu kredit berhasil menciptakan profil risiko yang lebih akurat dan transparan [1].

Berbagai studi ini memberikan landasan bagi penelitian saat ini untuk menutup celah (*research gap*) terkait terbatasnya kajian komparatif yang secara khusus memfokuskan variabel pada metode pembayaran guna menggambarkan tingkat inklusi digital pelanggan secara lebih mendalam.

Beberapa teknik yang sering digunakan dalam proses penilaian pengelompokan meliputi metode *Elbow* yang berguna untuk menentukan jumlah klaster yang paling optimal, serta metrik penilaian seperti *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index* yang digunakan untuk menilai tingkat pemisahan dan konsistensi kelompok yang dihasilkan [9]. Dengan menerapkan metode evaluasi ini, para peneliti dapat memastikan bahwa hasil segmentasi data yang diperoleh memiliki kualitas pengelompokan yang tinggi dan dapat dimanfaatkan untuk analisis selanjutnya.

3. METODE PENELITIAN



Gambar 1. Diagram tahapan penelitian

Penelitian ini adalah studi kuantitatif dengan menggunakan pendekatan komparatif yang bertujuan untuk menganalisis kinerja algoritma *K-Means Clustering* dan *Hierarchical Clustering* dalam segementasi pelanggan berdasarkan karakteristik perilaku belanja serta metode pembayaran yang dipakai. Segmentasi pelanggan ini ditujukan untuk menemukan pola perilaku dari konsumen agar dapat memberikan wawasan mengenai kecenderungan pelanggan saat melakukan pembelian.

Penelitian ini dilakukan secara eksploratif dengan menggunakan data sekunder yang diambil dari platform *Kaggle* yang menyimpan informasi tentang perilaku konsumen di sektor ritel. Dataset ini mencakup berbagai detail mengenai karakteristik pelanggan dan rekam jejak transaksi pembelian.

Tahapan dalam penelitian ini meliputi pengumpulan data, pra-pemrosesan data, penerapan algoritma *clustering* metode *K-Means* dan *Hierarchical Clustering*, serta analisis hasil segmentasi pelanggan.

Pada tahap pengumpulan data, penelitian ini menggunakan data sekunder dataset *Customer Shopping Behaviour Analysis* dengan 3900 data dan 16 variabel. Dataset ini menawarkan gambaran mendalam tentang kebiasaan belanja di sektor ritel. Dengan memadukan profil pengguna dan riwayat pembelian, dapat diidentifikasi bahwa kecenderungan pelanggan dalam mengambil keputusan serta bagaimana mereka membelanjakan uangnya secara lebih sistematis.

Variabel yang terdapat dalam kumpulan data mencakup umur dan gender pelanggan, produk yang dibeli, total belanja, cara pembayaran, pemanfaatan diskon, dan seberapa sering mereka berbelanja. Dari Dataset ini peneliti hanya memakai lima variabel saja yang terdiri dari metode pembayaran, umur, jumlah pembelian, pembelian sebelumnya, dan tingkat kepuasan.

Dengan menganalisis variabel-variabel ini, penelitian ini dapat mengidentifikasi pola preferensi pelanggan, tren pengeluaran, serta dampak promosi terhadap tindakan pembelian.

3.1. Data Preprocessing

Pra-pemrosesan data bertujuan untuk mengonversi data mentah menjadi bentuk yang dapat digunakan dalam algoritma [9].

Proses ini memiliki peranan yang sangat penting karena data yang belum diproses sering kali tidak teratur. Selain itu, algoritma seperti *K-Means* dan *Hierarchical* tidak bisa langsung diterapkan pada data yang belum terstruktur. Oleh karena itu, tahap pemrosesan awal diperlukan untuk mempermudah analisis data menggunakan algoritma. Langkah-langkah dalam pra-pemrosesan dimulai dari pemilihan variabel yang dapat merepresentasikan karakteristik pelanggan dalam menggunakan metode pembayaran, penghapusan variabel yang tidak relevan, dan pemerikasaan terhadap kemungkinan adanya nilai yang hilang di dalam data.

Setelah menentukan lima variabel yang berhubungan dengan cara pelanggan menggunakan metode pembayaran, langkah selanjutnya yaitu

encoding. K-Mean dan *Hierarchical clustering* hanya dapat memproses data dalam bentuk numerik. Variabel metode pembayaran merupakan data kategorikal yang berisi beberapa jenis metode pembayaran dalam penelitian ini, sehingga data kategorikal pada variabel tersebut harus diubah menjadi data numerik menggunakan teknik *one-hot-encoding.*

Teknik *one-hot-encoding* akan mengubah satu kolom kategori menjadi kolom biner dimana setiap kolom merepresentasikan satu jenis metode pembayaran. Melalui teknik transformasi ini, data yang bersifat kategorikal bisa dialihkan ke format numerik agar dapat diolah oleh algoritma clustering.

Pada tahap *scalling*, data dinormalisasi agar tidak ada variabel yang berpengaruh lebih besar karena memiliki nilai numerik yang lebih tinggi. Melalui proses *scalling*, variabel memiliki skala yang sebanding sehingga algoritma klastering lebih optimal dalam mengidentifikasi pola pada data.

3.2. Penerapan Algoritma K-Means Clustering

K-Means Clustering adalah teknik *unsupervised learning* yang umum digunakan untuk mengelola data berdasarkan kesamaan sifat antar objek. Metode ini berfungsi dengan memisahkan data menjadi beberapa klaster, sehingga objek dalam satu klaster memiliki kesamaan yang lebih besar dibandingkan objek dari klaster lain [10].

Langkah pertama dalam K-Means adalah menetapkan jumlah *cluster* (k) yang akan diterapkan. Setelah itu, algoritma akan memilih posisi awal titik pusat *cluster* (*centroid*), lalu menghitung jarak antara masing-masing data dan *centroid* dengan menggunakan metode pengukuran jarak tertentu, seperti jarak *Euclidean*. Data selanjutnya akan dikelompokkan ke dalam *cluster* berdasarkan jarak terpendek ke *centroid*. Setelah pengelompokan awal selesai, posisi *centroid* diperbaharui berdasarkan rata-rata nilai semua data dalam cluster itu. Proses ini akan berlanjut sampai posisi *centroid* tidak mengalami perubahan yang signifikan atau hingga kondisi konvergen tercapai.

Dalam penelitian ini, algoritma *K-Means Clustering* diterapkan untuk mengelompokkan pelanggan menurut karakteristik perilaku belanja dan metode pembayaran yang mereka gunakan. Penentuan jumlah klaster yang paling sesuai dilakukan dengan menggunakan metode *Elbow* dan didukung oleh *Silhouette Score* untuk menilai kualitas dari hasil pengelompokan.

3.3. Penerapan Algoritma Hierarchical Clustering

Hierarchical Clustering merupakan metode pengelompokan data kedalam struktur hierarki yang menyerupai cabang pohon (*dendrogram*), di mana objekobjek yang memiliki tingkat kemiripan tertinggi dikelompokkan terlebih dahulu.

Pendekatan yang digunakan dalam penelitian ini adalah *Agglomerative Hierarchical Clustering*, yaitu proses penggabungan yang dimulai dengan setiap objek sebagai satu *cluster* terpisah, kemudian secara bertahap menggabungkan dua *cluster* yang memiliki kemiripan tertinggi hingga terbentuk satu *cluster* besar [11].

Dalam penelitian ini, *Hierarchical Clustering* digunakan sebagai pembanding terhadap *K-Means.* Untuk segmentasi pelanggan dalam menggunakan metode pembayaran.

3.4. Analisis Data

Teknik analisis data dalam penelitian ini dilakukan dengan cara membandingkan hasil pengelompokan data melalui dua algoritma clustering, yaitu *K-Means Clustering* dan *Hierarchical Clustering*. Tujuan dari analisis ini adalah untuk menemukan metode *clustering* yang dapat memberikan segmentasi pelanggan yang lebih baik, dilihat dari karakteristik perilaku belanja dan jenis pembayaran yang digunakan.

Untuk menentukan jumlah klaster yang paling sesuai, penelitian ini mengaplikasikan metode *Elbow*. Proses ini dilakukan dengan menilai perubahan nilai *Within Cluster Sum of Squares* (WCSS) sehubungan dengan jumlah klaster yang dihasilkan. Titik siku pada grafik menunjukkan jumlah cluster yang dianggap paling cocok untuk digunakan selama proses *clustering.*

Kualitas hasil *clustering* juga dinilai dengan menggunakan *Silhouette Score*. Metode ini bertujuan untuk mengevaluasi seberapa mirip suatu data dengan *cluster* yang diikuti, dibandingkan dengan cluster lainnya. Nilai *silhouette* memiliki rentang antara -1 dan 1. Nilai yang mendekati 1 mengindikasikan bahwa data memiliki kesesuaian yang baik dengan *cluster* yang diikutinya serta adanya pemisahan yang jelas dari *cluster* yang lain.

Hasil dari penilaian kedua metode *clustering* tadi selanjutnya dianalisis untuk mengidentifikasi algoritma mana yang memberikan hasil segmentasi pelanggan yang lebih optimal.

4. HASIL DAN PEMBAHASAN

4.1. Data Preprocessing

```
2. DATA PREPROCESSING
----------------------------------------
Missing values:
Age                       0
Gender                    0
Item Purchased            0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Subscription Status       0
Discount Applied          0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
```

Gambar 2. Hasil Pre-Processing Data

Dalam fase pra-pemrosesan, dilakukan pembersihan data untuk menjamin integritas dataset dengan mengeliminasi entri duplikat serta menangani nilai yang hilang (missing values). Setelah dibersihkan, data tersebut siap diproses lebih lanjut untuk keperluan analisis clustering.

Gambar 2 memperlihatkan hasil dari proses pembersihan data yang telah dilakukan. Semua bagian dalam dataset ini tidak memiliki nilai yang kosong, jadi data ini bisa segera digunakan untuk analisa selanjutnya.

```
[4]
✓ 0d
    data = df[['Age',
               'Purchase Amount (USD)',
               'Payment Method',
               'Previous Purchases',
               'Review Rating']]

[5]
✓ 0d
    features_for_clustering = ['Payment Method']
```

Gambar 3. Variabel yang digunakan dalam analisis

Gambar 3 memperlihatkan tahapan *feature selection* yaitu memilih variabel yang relevan dengan tujuan penelitian, variabel ini dimaksudkan untuk mencerminkan sifat dan tindakan pelanggan saat berbelanja serta cara pembayaran yang mereka pilih.

Dengan melakukan pemilihan variabel ini, dataset menjadi lebih terarah sehingga dapat meningkatkan kualitas hasil *clustering.*

Variabel *features_for_clustering* memuat daftar karakteristik yang digunakan secara langsung dalam proses *clustering*. Dalam studi ini, karakteristik yang diterapkan adalah Metode Pembayaran, karena penelitian ini berfokus pada analisis segmentasi pelanggan berdasarkan cara pembayaran yang dipilih.

```
Numerical features (0): []
Categorical features (1): ['Payment Method']
```

Gambar 4. Mengidentifikasi Fitur Kategorikal

Gambar 4 memperlihatkan berdasarkan analisis fitur, data yang digunakan dalam proses *clustering* hanya memiliki satu fitur kategorikal, yaitu Metode Pembayaran, dan tidak ada fitur angka dalam subset data yang diterapkan. Oleh karena itu, tahap *preprocessing* selanjutnya difokuskan pada proses *encoding* terhadap fitur kategorikal tersebut agar dapat diproses oleh algoritma clustering.

```
Encoding categorical features...
Final feature matrix shape: (3900, 5)
```

Gambar 5. *Encoding* Variabel Kategorikal

Gambar 5 memperlihatkan bahwa Tahap encoding dilakukan untuk mentransformasikan variabel kategorikal Payment Method menjadi beberapa variabel numerik menggunakan teknik *One-Hot Encoding*. Hasil dari proses ini adalah matriks fitur yang seluruhnya berbentuk numerik sehingga dapat digunakan dalam proses *clustering* menggunakan algoritma *K-Means* dan *Hierarchical Clustering*. Dengan 3900 baris data dan 5 fitur yang digunakan dalam proses *clustering.*

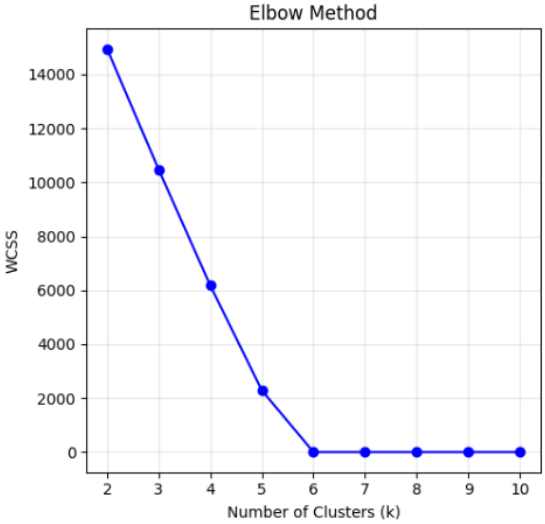
```
# Scale numerical features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(df_features)
df_scaled = pd.DataFrame(features_scaled, columns=df_features.columns)

4. FEATURE SCALING
----------------------------------------
Feature scaling completed.
Scaled data shape: (3900, 5)
```

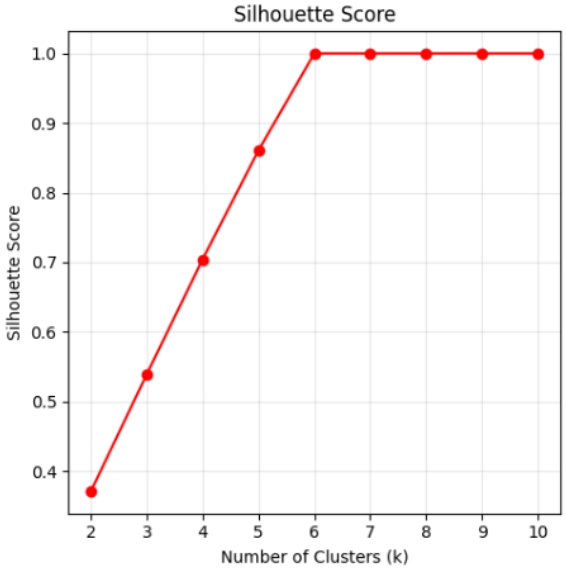
Gambar 6. Proses *Scaling*

Gambar 6 memperlihatkan tahap *feature scaling* dilakukan dengan StandardScaler guna mencegah adanya fitur yang mendominasi perhitungan jarak. Hasil standarisasi ini menjadi input utama dalam menjalankan algoritma *clustering* berbasis jarak seperti *K-Means* dan *Hierarchical.*

4.2. Penerapan Algoritma K-Means Clustering



Gambar 7. Elbow Method



Gambar 8. *Silhouette Score*

Hasil evaluasi menggunakan Metode Elbow menunjukkan penurunan *Within-Cluster Sum of Squares* (WCSS) yang signifikan pada rentang k=2 hingga k=6. Penurunan grafik tersebut mulai melandai dan cenderung stabil setelah titik k=6, yang mengindikasikan bahwa penambahan klaster lebih lanjut tidak memberikan peningkatan kualitas pengelompokan yang substansial. Oleh karena itu, jumlah klaster yang ditetapkan sebagai nilai optimal untuk model ini adalah 6 klaster.

Pengujian menggunakan metrik Silhouette Score menunjukkan tren peningkatan yang mencapai titik puncaknya pada k=6. Nilai metrik yang mendekati angka 1 ini mengindikasikan tingginya tingkat kohesi dan separasi. Mengingat pemisahan klaster tervalidasi dengan sangat baik pada titik ini, evaluasi Silhouette Score mengonfirmasi bahwa 6 merupakan jumlah klaster paling optimal untuk dataset tersebut.

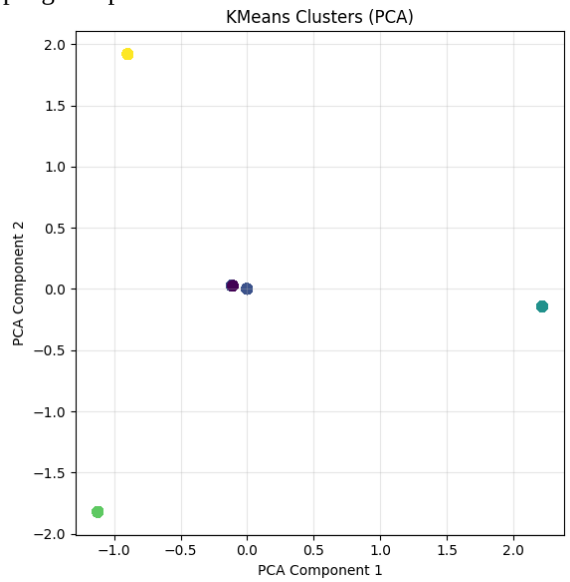
```
6. CLUSTERING WITH DIFFERENT ALGORITHMS
----------------------------------------

1. KMeans Clustering...
2. Hierarchical Clustering...

Clustering completed for all algorithms.
```

Gambar 9. *Cluster* Dengan Dua Alogritma Berbeda

Tahap ini difokuskan pada pengelompokan pelanggan menggunakan algoritma *K-Means* dan *Hierarchical Clustering* dengan konfigurasi 5 klaster. Output dari kedua metode tersebut kemudian diintegrasikan ke dalam dataset untuk dievaluasi secara komparatif. Evaluasi kualitas segmentasi dilakukan menggunakan metrik *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index* guna menentukan algoritma dengan performa pengelompokan terbaik.



Gambar 10. *K-means Clusters (PCA)*

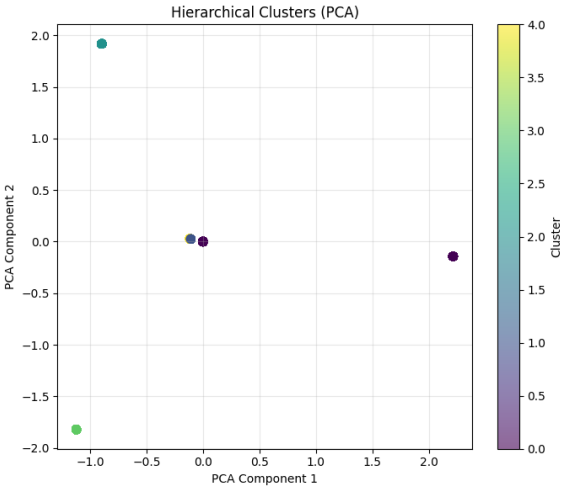
Pada grafik terlihat bahwa data pelanggan terbagi ke dalam lima *cluster* yang berbeda. Setiap *cluster* ditandai dengan warna yang berbeda dan memiliki posisi yang relatif terpisah pada ruang dua dimensi PCA. Hal ini menunjukkan bahwa algoritma *K-Means* mampu mengelompokkan data pelanggan dengan cukup jelas berdasarkan pola yang terdapat pada fitur yang digunakan.

4.3. Penerapan Algoritma Hierarchical Clustering

```
# Algorithm 2: Hierarchical Clustering
print("2. Hierarchical Clustering...")
agg_clustering = AgglomerativeClustering(n_clusters=n_clusters)
agg_labels = agg_clustering.fit_predict(df_scaled)
df['Hierarchical_Cluster'] = agg_labels
```

Gambar 11. Metode *Agglomerative Clustering Pada Hierarchical Clustering*

Berdasarkan proses yang dilakukan, metode *Hierarchical Clustering* dengan pendekatan *Agglomerative* bekerja dengan cara menggabungkan data secara bertahap berdasarkan kedekatan jarak antar data. Proses dimulai dengan menganggap setiap data sebagai satu *cluster* tersendiri, kemudian *cluster* yang memiliki jarak paling dekat akan digabungkan secara bertahap hingga mencapai jumlah *cluster* yang telah ditentukan. Metode ini memungkinkan terbentuknya struktur hierarki pada data sehingga hubungan antar *cluster* dapat dianalisis secara lebih sistematis



Gambar 12. *Hierarchical Clusters (PCA)*

Pada grafik kedua terlihat bahwa hasil clustering menggunakan *Hierarchical Clustering* juga membentuk lima *cluster* yang berbeda. Pola distribusi cluster yang terbentuk relatif mirip dengan hasil yang diperoleh dari algoritma K-Means, meskipun terdapat sedikit perbedaan pada posisi dan distribusi beberapa cluster.

4.4. Analisis Data

Penelitian ini melakukan studi komparatif untuk mengevaluasi efektivitas algoritma *K-Means* dan *Hierarchical Clustering* dalam melakukan segmentasi pelanggan berdasarkan variabel metode pembayaran. Kualitas pengelompokan dari kedua metode tersebut diukur secara objektif melalui tiga instrumen validasi

internal, yaitu *Silhouette Score, Calinski-Harabasz (CH) Index, dan Davies-Bouldin Index.* Perbandingan performa dari masing-masing algoritma disajikan secara detail dalam tabel berikut.

Tabel 1. Hasil Evaluasi Metrik Clusterting

Metrik Evaluasi	K- Means	Hierarchical
Silhouette Score	0.8606	0.8595
Calinski-Harabasz	7335.58	7500.76
Davies-Bouldin	0.45	0.43

Tabel 1 nilai *Silhouette Score*, algoritma *K-Means* memperoleh nilai sebesar 0.8606, sedangkan *Hierarchical Clustering* memperoleh nilai 0.8595. Kedua nilai tersebut tergolong tinggi dan mendekati angka 1, yang menunjukkan bahwa struktur cluster yang terbentuk memiliki pemisahan yang baik antar kelompok data. Namun demikian, nilai yang diperoleh *K-Means* sedikit lebih tinggi dibandingkan *Hierarchical*, sehingga menunjukkan bahwa *K-Means* mampu menghasilkan pemisahan cluster yang sedikit lebih optimal.

Selanjutnya, berdasarkan *Calinski–Harabasz Index*, algoritma *Hierarchical Clustering* memperoleh nilai yang lebih tinggi yaitu 7500.76, dibandingkan dengan *K-Means* yang memperoleh nilai 7335.58. Nilai *Calinski–Harabasz* yang lebih besar menunjukkan bahwa jarak antar *cluster* semakin jelas dan variasi antar cluster lebih besar dibandingkan variasi di dalam *cluster*. Dengan demikian, berdasarkan metrik ini, *Hierarchical Clustering* menunjukkan performa yang lebih baik dalam memisahkan kelompok data.

Sementara itu, jika dilihat dari nilai *Davies–Bouldin Index*, algoritma *Hierarchical Clustering* menghasilkan nilai 0.43, yang lebih rendah dibandingkan *K-Means* dengan nilai 0.45. Dalam metrik ini, nilai yang lebih kecil menunjukkan kualitas *cluster* yang lebih baik karena jarak antar *cluster* lebih besar dibandingkan dengan penyebaran data di dalam *cluster*.

KMeans_Cluster	
1	1248
2	677
3	671
4	670
0	634

Hierarchical_Cluster	
0	1289
3	671
2	670
4	636
1	634

Gambar 13. Tabel Clustering Berdasarkan Metode Pembayaran

Pada Gambar 13 Berdasarkan hasil proses *clustering* menggunakan algoritma *K-Means* dan *Hierarchical Clustering*, dataset pelanggan berhasil dikelompokkan ke dalam lima *cluster* yang berbeda. Pemilihan jumlah 5 *cluster* dilakukan berdasarkan hasil evaluasi menggunakan metrik *Silhouette Score, Calinski–Harabasz Index, dan Davies–Bouldin Index.*

Dari hasil distribusi *clustering* menggunakan algoritma *K-means*, terlihat bahwa *Cluster 1* memiliki jumlah data paling besar, yaitu sebanyak 1248 pelanggan. Hal ini menunjukkan bahwa sebagian besar pelanggan memiliki karakteristik yang relatif mirip sehingga tergabung dalam *cluster* tersebut.

Sementara itu, kelompok lainnya memiliki jumlah anggota yang relatif seimbang, yakni berkisar di angka 634 sampai 677 pelanggan. Distribusi ini membuktikan kemampuan algoritma *K-Means* dalam membagi populasi pelanggan ke dalam segmen-segmen yang ukurannya cukup seragam.

Sedangkan hasil distribusi *clustering*.menggunakan algoritma *Hierarcical Clustering* menunjukkan bahwa *Cluster 0* memiliki jumlah anggota paling banyak, yaitu sebanyak 1289 pelanggan. Hal ini menunjukkan adanya kelompok pelanggan dengan karakteristik dominan yang lebih besar dibandingkan kelompok lainnya.

Sementara itu, *cluster* lainnya memiliki jumlah anggota yang relatif serupa, yaitu berada pada kisaran 634 hingga 671 pelanggan, yang menunjukkan distribusi data yang cukup seimbang di antara *cluster* yang terbentuk.

	Cluster	Size	Avg Age	Avg Purchase	Avg Previous Purchases	Dominant Payment
0	0	634	43.69	58.95	25.65	Venmo
1	1	1248	43.97	60.33	25.04	Debit Card
2	2	677	44.03	59.25	25.51	PayPal
3	3	671	44.63	60.07	25.59	Credit Card
4	4	670	44.08	59.70	25.25	Cash

Gambar 14. Tabel Karakteristik *Cluster* Pelanggan *K-means*

a. Klaster 0 ini terdiri dari 634 pelanggan dengan rata-rata usia 43 tahun. Rata-rata nilai pembelian pelanggan pada *cluster* ini adalah 58.95 USD dengan rata-rata jumlah pembelian sebelumnya sebesar 25.65 kali. Metode pembayaran yang paling dominan pada Klaster ini adalah Venmo.
b. Klaster 1 merupakan klaster dengan jumlah anggota terbesar yaitu 1248 pelanggan. Pelanggan dalam klaster ini memiliki rata-rata usia 43 tahun dengan rata-rata nilai pembelian sebesar 60.33 USD dan rata-rata pembelian sebelumnya sebanyak 25.04 kali. Metode pembayaran yang paling banyak digunakan pada cluster ini adalah Debit Card.
c. Klaster 2 ini terdiri dari 677 pelanggan dengan rata-rata usia 44 tahun. Nilai rata-rata pembelian pelanggan pada cluster ini sebesar 59.25 USD dengan rata-rata pembelian sebelumnya 25.51 kali. Metode pembayaran yang paling dominan pada klaster ini adalah PayPal.

d. Klaster 3 ini memiliki 671 pelanggan dengan rata-rata usia 44 tahun, yang merupakan usia rata-rata tertinggi dibanding cluster lainnya. Rata-rata nilai pembelian pada klaster ini adalah 60.07 USD dengan rata-rata pembelian sebelumnya sebesar 25.59 kali. Metode pembayaran yang paling dominan pada klaster ini adalah Credit Card.
e. Klaster ini terdiri dari 670 pelanggan dengan rata-rata usia 44 tahun. Nilai rata-rata pembelian pada cluster ini sebesar 59.70 USD dengan rata-rata pembelian sebelumnya sebesar 25.25 kali. Metode pembayaran yang paling banyak digunakan pada Klaster ini adalah Cash.

	Cluster	Size	Avg Age	Avg Purchase	Avg Previous Purchases	Dominant Payment
0	0	1289	43.95	59.47	25.03	PayPal
1	1	634	43.69	58.95	25.65	Venmo
2	2	670	44.08	59.70	25.25	Cash
3	3	671	44.63	60.07	25.59	Credit Card
4	4	636	44.07	60.92	25.56	Debit Card

Gambar 15. Tabel Karakteristik Cluster Pelanggan *Hierarchical Clustering*

a. Klaster 0 ini memiliki jumlah anggota terbesar yaitu 1289 pelanggan dengan rata-rata usia 43 tahun. Rata-rata nilai pembelian pada cluster ini sebesar 59.47 USD dengan rata-rata pembelian sebelumnya sebesar 25.03 kali. Metode pembayaran yang paling dominan pada Klaster ini adalah PayPal.
b. Klaster 1 ini terdiri dari 634 pelanggan dengan rata-rata usia 43 tahun. Nilai rata-rata pembelian pada cluster ini sebesar 58.95 USD dengan rata-rata pembelian sebelumnya sebesar 25.65 kali. Metode pembayaran yang paling dominan pada klaster ini adalah Venmo.
c. Klaster 2 ini terdiri dari 670 pelanggan dengan rata-rata usia 44 tahun. Rata-rata nilai pembelian pelanggan pada cluster ini sebesar 59.70 USD dengan rata-rata pembelian sebelumnya sebesar 25.25 kali. Metode pembayaran yang paling dominan pada klaster ini adalah Cash.
d. Klaster 3 ini memiliki 671 pelanggan dengan rata-rata usia 44 tahun. Nilai rata-rata pembelian pelanggan pada cluster ini adalah 60.07 USD dengan rata-rata pembelian sebelumnya sebesar 25.59 kali. Metode pembayaran yang paling dominan pada klaster ini adalah Credit Card.
e. Klaster 4 ini terdiri dari 636 pelanggan dengan rata-rata usia 44 tahun. Rata-rata nilai pembelian pada cluster ini sebesar 60.92 USD, yang merupakan nilai pembelian tertinggi di antara cluster lainnya. Metode pembayaran yang paling dominan pada klaster ini adalah Debit Card.

4.5. Analisis Perbandingan 2 Metode *Clustering*

Berdasarkan hasil proses *clustering* yang dilakukan menggunakan algoritma *K-Means* dan *Hierarchical Clustering*, dilakukan evaluasi terhadap kualitas *cluster* menggunakan metrik *Silhouette Coefficient*. Nilai *Silhouette Coefficient* digunakan untuk mengukur tingkat kedekatan antar data dalam satu *cluster* serta jarak pemisahan antar *cluster* yang berbeda. Hasil evaluasi nilai *Silhouette Coefficient* ditampilkan pada Gambar 16.

	Metode Clustering	k=2	k=3	k=4	k=5
0	K-Means	0.370663	0.538902	0.704112	0.860653
1	Hierarchical	0.371133	0.539050	0.700853	0.859587

Gambar 16. Perbandingan Hasil Rata-Rata *Silhouette Coefficient*

Silhouette Coefficient memiliki rentang nilai antara -1 hingga 1, di mana nilai yang semakin mendekati 1 menunjukkan bahwa struktur *cluster* yang terbentuk semakin baik karena data dalam *cluster* memiliki tingkat kemiripan yang tinggi dan memiliki jarak yang cukup jelas dengan *cluster* lainnya. Sebaliknya, nilai yang mendekati -1 menunjukkan bahwa kualitas *clustering* kurang baik karena data cenderung berada pada *cluster* yang tidak tepat.

Berdasarkan hasil evaluasi yang diperoleh, algoritma *K-Means* menghasilkan nilai *Silhouette Score* sebesar 0.8606, sedangkan algoritma *Hierarchical Clustering* memperoleh nilai 0.8595. Kedua nilai tersebut menunjukkan bahwa kualitas *cluster* yang dihasilkan oleh kedua algoritma tergolong sangat baik karena mendekati nilai maksimum dari *Silhouette Coefficient.*

Jika dibandingkan secara langsung, algoritma *K-Means* memiliki nilai *Silhouette Score* yang sedikit lebih tinggi dibandingkan dengan *Hierarchical Clustering*. Hal ini menunjukkan bahwa algoritma *K-Means* mampu menghasilkan struktur *cluster* yang sedikit lebih optimal dalam memisahkan data pelanggan berdasarkan karakteristik yang digunakan dalam penelitian.

Dengan demikian, berdasarkan hasil evaluasi menggunakan *Silhouette Coefficient,* dapat disimpulkan bahwa kedua algoritma memiliki performa *clustering* yang baik, namun algoritma *K-Means* menunjukkan kinerja yang sedikit lebih unggul dibandingkan dengan *Hierarchical Clustering* dalam proses segmentasi data pelanggan pada penelitian ini.

5. KESIMPULAN DAN SARAN

Kesimpulan Penelitian ini menyimpulkan bahwa algoritma *K-Means* dan *Hierarchical Clustering* memiliki performa yang sangat baik dalam melakukan segmentasi pelanggan berdasarkan metode pembayaran, dengan nilai *Silhouette Score* yang tinggi (di atas 0,85) yang menunjukkan pemisahan klaster yang jelas. Secara komparatif, algoritma *K-Means* menunjukkan kinerja yang sedikit lebih unggul dibandingkan *Hierarchical Clustering* berdasarkan metrik *Silhouette Score* (0,8606 vs 0,8595), yang berarti *K-Means* lebih optimal dalam menghasilkan struktur kelompok yang kohesif. Melalui evaluasi gabungan berbagai metrik, penelitian ini berhasil memetakan pelanggan ke dalam lima klaster utama yang didominasi oleh preferensi pembayaran yang

berbeda, yaitu *Debit Card, Credit Card, PayPal, Venmo,* dan *Cash*. Saran bagi pelaku industri ritel, hasil segmentasi ini sebaiknya digunakan sebagai landasan strategis dalam merancang program pemasaran yang lebih personal dan tepat sasaran guna memperkuat loyalitas pelanggan jangka panjang. Untuk penelitian selanjutnya, disarankan untuk melakukan eksplorasi dengan menambahkan variabel pendukung lainnya atau menggunakan dataset yang lebih kompleks untuk melihat stabilitas algoritma dalam cakupan yang lebih luas. Selain itu, pengembangan penelitian dapat diarahkan pada penggunaan metode unsupervised learning lainnya guna memperkaya literatur mengenai efektivitas algoritma pengelompokan dalam konteks perilaku konsumen digital.

DAFTAR PUSTAKA

[1] B. I. Pratama et al., "PENERAPAN MACHINE LEARNING DALAM PENGELOMPOKAN PELANGGAN MENGGUNAKAN K-MEANS CLUSTERING UNTUK MENINGKATKAN STRATEGI PEMASARAN PT MASPION," vol. 10, no. 4, pp. 3573–3585, 2025.

[2] I. Budiyanto and A. Hermawan, "SEGMENTASI NASABAH KARTU KREDIT MENGGUNAKAN ALGORITMA K-MEANS BERDASARKAN POLA TRANSAKSI UNTUK PENENTUAN PROFIL NASABAH," vol. 9, no. 3, pp. 560–571, 2025, doi: 10.26798/jiko.v9i3.1669.

[3] T. W. Ningsih et al., "METODE CLUSTERING DENGAN MODEL RFM DAN ALGORITMA K-MEANS UNTUK SEGMENTASI PELANGGAN TRANSJAKARTA," vol. 9, no. 5, pp. 7898–7905, 2025.

[4] A. Syaharani et al., "ANALISIS CLUSTERING SEGMENTASI PELANGGAN PADA CV PRIMA JAYA UTAMA MENGGUNAKAN METODE K-MEANS," vol. 9, no. 5, pp. 8847–8854, 2025.

[5] A. N. S. Wawagalang, S. Syahrullah, R. Ardiyansyah, D. S. Angreni, S. A. Pratama, and D. W. Nugraha, "Segmentasi Pelanggan Menggunakan Kerangka LRFMV dan Algoritma K-Means untuk Optimalisasi Strategi Pemasaran", Edumatic, vol. 9, no. 2, pp. 639–648, Aug. 2025.

[6] P. Aselnino and A. W. Wijayanto, "Analisis Perbandingan Metode Hierarchical dan Non-Hierarchical dalam Pembentukan Cluster Provinsi di Indonesia Berdasarkan Indikator Women Empowerment," vol. 6, no. 1, pp. 57–68, 2023.

[7] A. Rahmawati et al., "Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Untuk Optimalisasi Penjualan Sembako," vol. 8, no. 2, pp. 233–243, 2025.

[8] S. B. W. Rizky, "PERBANDINGAN ALGORITMA K-MEANS DAN HIERARCHICAL CLUSTERING DALAM SEGMENTASI KABUPATEN / KOTA DI JAWA TIMUR BERDASARKAN DATA PERJALANAN DAN PERGERAKAN WISATAWAN," vol. 9, no. 5, pp. 8499–8506, 2025.

[9] N. Siswadi et al., "PERBANDINGAN ALGORITMA K-MEANS DAN HIERARCHICAL CLUSTERING UNTUK PENGELOMPOKKAN WILAYAH BERDASARKAN PRODUKSI DAN HARGA CABAI DI INDONESIA," vol. 9, no. 5, pp. 7713–7720, 2025.

[10] B. Satria and M. Al Hafiz, "Segmentasi pelanggan layanan air bersih di kawasan pedesaan menggunakan k-means clustering berdasarkan metrik durasi, konsumsi rata-rata, dan total volume (dcv)," vol. 10, no. 4, pp. 3896–3906, 2025. PL

[11] T. Ardyanti, M. Furqan, F. Science, U. Islam, and N. Sumatera, "Implementation of The Agglomerative Hierarchical Clustering Method In Ordering Hijab Products," vol. 8, no. October, pp. 2479–2489, 2024.