Penerapan Algoritma K-Means Pada Sistem Pengelompokan Data Penjualan Retail Berbasis Web

Anindyar Bintang Rahma Esa[1*], Medika Heri Prasetyo[2], Tiyas Styaningrum[3], Gregorius Sabian Aswendoro[4]

[1]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[1]230103186@mhs.udb.ac.id

[2]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[2]230103198@mhs.udb.ac.id

[3]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[3]230103152@mhs.udb.ac.id

[4]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[4]230103016@mhs.udb.ac.id

***Abstrak*— Pengelolaan data penjualan retail secara manual membutuhkan waktu yang lama dan rentan terhadap kesalahan interpretasi, terutama dalam memahami pola pembelian konsumen. Penelitian ini bertujuan untuk membangun sistem berbasis web yang mampu mengelompokkan data penjualan retail secara otomatis menggunakan algoritma K-Means Clustering berdasarkan karakteristik transaksi, meliputi jumlah barang, harga satuan, total nilai transaksi, kategori produk, lokasi transaksi, dan metode pembayaran. Dataset yang digunakan adalah data transaksi penjualan retail yang bersumber dari Kaggle, terdiri dari 12.575 baris data transaksi yang mencakup periode Januari 2022 hingga Januari 2025. Tahapan penelitian meliputi pengumpulan data, praproses data (*data cleaning*) untuk menangani 7.229 nilai kosong, seleksi dan *encoding* fitur transaksi, penentuan jumlah cluster optimal menggunakan *elbow method*, serta penerapan algoritma K-Means. Sistem diimplementasikan dalam bentuk aplikasi web yang menyediakan fitur unggah data, pembersihan data otomatis, proses *clustering*, dan visualisasi hasil berupa *scatter plot* serta profil tiap *cluster*. Hasil *clustering* membagi transaksi ke dalam dua kelompok, yaitu pembeli hemat, dan pembeli bernilai tinggi, yang masing-masing memiliki karakteristik pola pembelian yang berbeda. Evaluasi model menggunakan *silhouette score* dan *Davies-Bouldin index* menunjukkan bahwa pengelompokan yang terbentuk cukup baik untuk diinterpretasikan secara bisnis. Sistem ini diharapkan dapat membantu pengelola toko retail dalam mengambil keputusan bisnis yang lebih tepat sasaran berdasarkan pola pembelian yang dihasilkan.**

Kata kunci— K-Means Clustering, Segmentasi Pelanggan, Data Penjualan Retail, Pola Pembelian, Sistem Berbasis Web

***Abstract*— Manual management of retail sales data is time-consuming and prone to misinterpretation, particularly in understanding consumer purchasing patterns. This study aims to develop a web-based system capable of automatically grouping retail sales data using the K-Means Clustering algorithm based on transaction characteristics, including quantity, unit price, total transaction value, product category, transaction location, and payment method. The dataset used consists of retail sales transaction data sourced from Kaggle, comprising 12,575 transaction records covering the period of January 2022 to January 2025. The research stages include data collection, data preprocessing (*data cleaning*) to handle 7,229 missing values, transaction feature selection and *encoding*, determination of the optimal number of clusters using the *elbow method*, and application of the K-Means algorithm. The system is implemented as a web application providing features for data upload, automatic data cleaning, clustering process, and result visualization in the form of *scatter plots* and cluster profiles. The clustering results divide transactions into two groups, namely economical buyers, and high-value buyers, each exhibiting distinct purchasing pattern characteristics. Model evaluation using the *silhouette score* and *Davies-Bouldin index* indicates that the resulting clusters are sufficiently meaningful for business interpretation. This system is expected to assist retail store managers in making more targeted business decisions based on the generated purchasing pattern groupings.**

Keywords— K-Means Clustering, Transaction Grouping, Retail Sales Data, Purchasing Patterns, Web-Based System

I. PENDAHULUAN

Perkembangan industri retail yang semakin pesat menghasilkan volume data transaksi yang besar setiap harinya. Data tersebut menyimpan informasi berharga mengenai perilaku pembelian pelanggan, namun pengelolaan dan analisisnya secara manual membutuhkan waktu yang lama serta rentan terhadap kesalahan interpretasi. Tanpa pengelolaan data yang baik, pelaku usaha retail kesulitan memahami segmen pelanggan mereka secara akurat, sehingga strategi pemasaran yang diterapkan menjadi kurang tepat sasaran [1]. Kondisi ini mendorong perlunya sebuah pendekatan berbasis teknologi yang

mampu mengolah data transaksi retail secara otomatis dan menghasilkan informasi yang dapat digunakan sebagai dasar pengambilan keputusan bisnis.

Analisis pola pembelian merupakan salah satu strategi kunci dalam manajemen retail modern. Dengan mengelompokkan transaksi penjualan berdasarkan kesamaan karakteristik seperti jumlah barang yang dibeli, harga satuan, dan total nilai transaksi, pelaku usaha dapat memahami pola perilaku konsumen secara lebih mendalam dan merancang strategi penjualan yang lebih tepat sasaran[2]. Azzahra [3] mengimplementasikan K-Means Clustering untuk mengelompokkan produk berdasarkan performa penjualan dan berhasil mengidentifikasi tiga segmen dengan nilai Silhouette Score dan Davies-Bouldin Index yang baik.

Algoritma *K-Means Clustering* merupakan salah satu metode *unsupervised machine learning* yang paling banyak digunakan dalam analisis pola pembelian. Algoritma ini bekerja dengan cara mempartisi data ke dalam sejumlah *cluster* berdasarkan kemiripan karakteristik, sehingga objek dengan karakteristik serupa akan berada dalam satu kelompok yang sama [4]. Beberapa penelitian terdahulu membuktikan efektivitas K-Means dalam konteks data retail. Jumlah cluster yang terbentuk tidak ditetapkan secara manual oleh peneliti, melainkan ditentukan secara objektif melalui *Elbow Method* yang menganalisis perubahan nilai *inertia* untuk setiap kemungkinan nilai K. Murpratiwi dkk. [4] menganalisis pemilihan cluster optimal pada data pelanggan toko retail dan menunjukkan bahwa penentuan nilai K yang tepat sangat berpengaruh terhadap kualitas hasil segmentasi.

Penelitian mengenai segmentasi pelanggan berbasis K-Means terus berkembang dalam beberapa tahun terakhir. Rumapea dkk. [5] melakukan analisis pengelompokan data transaksi ritel online menggunakan K-Means dan berhasil mengelompokkan transaksi ke dalam tiga segmen dengan karakteristik nilai pembelian yang berbeda. Kristanti dkk. [6] mengimplementasikan K-Means dengan mempertimbangkan beberapa atribut transaksi sekaligus dan memperoleh hasil pengelompokan yang lebih komprehensif. Penelitian-penelitian tersebut umumnya berhenti pada tahap analisis data, tanpa mengintegrasikan hasilnya ke dalam sebuah sistem yang dapat digunakan secara langsung oleh pengguna non-teknis.

Kebutuhan terhadap sistem yang tidak hanya mampu melakukan analisis data, tetapi juga dapat diakses dan dioperasikan dengan mudah oleh pengguna umum menjadi motivasi utama penelitian ini. Nurmilayanti dan Hartono [2] menegaskan bahwa implementasi metode *clustering* yang terintegrasi dalam sebuah antarmuka sistem dapat menjadi penunjang yang signifikan dalam manajemen pelanggan. Sejalan dengan hal tersebut, Alya Putri dan Rahmah [7] menunjukkan bahwa implementasi K-Means berbasis aplikasi mampu meningkatkan efisiensi analisis bisnis secara keseluruhan. Lebih lanjut, Dzakiron dkk. [8] membuktikan bahwa sistem berbasis web yang mengintegrasikan algoritma K-Means dapat memberikan kemudahan klasifikasi data bagi pengguna tanpa latar belakang teknis sekalipun.

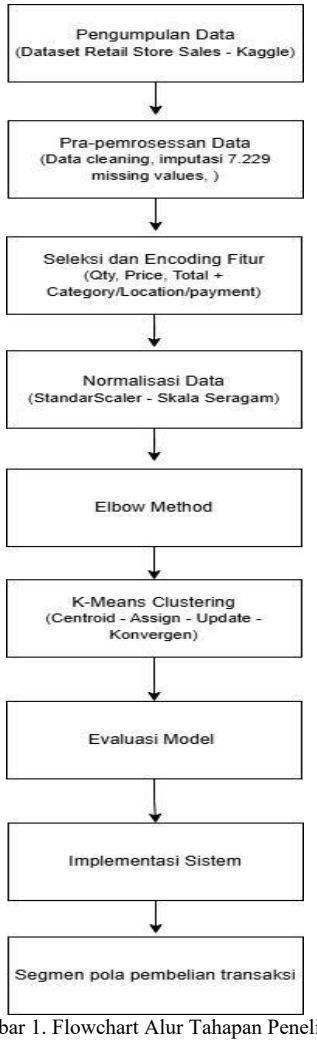
Penelitian ini menggunakan dataset *Retail Store Sales* yang bersumber dari Kaggle, terdiri dari 12.575 baris data transaksi dari toko fisik (*in-store*) maupun daring (*online*) yang mencakup periode Januari 2022 hingga Januari 2025. Dataset ini memuat atribut jumlah barang, harga satuan, kategori produk, lokasi transaksi, dan metode pembayaran, sehingga sangat sesuai untuk dianalisis pada level transaksi guna mengungkap pola pembelian konsumen secara menyeluruh. Permatasari [9] menekankan pentingnya kualitas data dalam proses analisis *clustering*, termasuk tahapan *data cleaning* dan visualisasi sebagai bagian dari *exploratory data analysis* sebelum pemodelan dilakukan.

Berdasarkan uraian latar belakang tersebut, penelitian ini bertujuan untuk merancang dan membangun sistem berbasis web yang mampu melakukan pengelompokan data penjualan retail secara otomatis menggunakan algoritma K-Means Clustering berdasarkan karakteristik transaksi. Sistem yang dibangun diharapkan dapat membantu pengelola toko retail dalam memahami pola pembelian konsumen,

mengidentifikasi segmen transaksi secara akurat, serta menjadi dasar pengambilan keputusan strategi pemasaran yang lebih efektif dan tepat sasaran [10].

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan *Knowledge Discovery in Databases* (KDD) sebagai kerangka kerja utama, yang mencakup tahapan mulai dari pengumpulan data, pra-pemrosesan, transformasi, penerapan algoritma *data mining*, hingga interpretasi hasil [10]. Secara keseluruhan, metodologi penelitian ini terdiri dari delapan tahapan yang saling berurutan dan terintegrasi ke dalam sebuah sistem berbasis web. Alur tahapan penelitian disajikan pada flowchart 1 berikut.



Gambar 1. Flowchart Alur Tahapan Penelitian

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset *Retail Store Sales* yang bersumber dari platform Kaggle. Dataset ini berisi 12.575 baris data transaksi penjualan dari toko fisik (*in-store*) maupun daring (*online*) yang mencakup periode Januari 2022 hingga Januari 2025. Dataset memiliki 11 atribut, yaitu Transaction_ID, Customer_ID, Transaction_Date, Category, Item, Price_Per_Unit, Quantity, Total_Spent, Payment_Method, Location, dan Discount_Applied. Penelitian ini menggunakan unit analisis pada level transaksi sehingga seluruh 12.575 baris data dapat dimanfaatkan secara penuh dalam proses clustering [9].

B. Pra-pemrosessan Data

Tahap pra-pemrosesan data dilakukan untuk memastikan kualitas data sebelum masuk ke tahap pemodelan [9]. Pada dataset ini ditemukan total 7.229 nilai kosong (*missing values*) yang tersebar pada beberapa kolom. Penanganan dilakukan dengan beberapa teknik sebagai berikut.

1. Kolom *Price_Per_Unit* (609 nilai kosong): diisi menggunakan rumus *Total_Spent* dibagi *Quantity* (imputasi nilai turunan).

$$Price\_Per\_Unit = \frac{Total\_Spent}{Quantity}$$

2. Kolom *Quantity* (604 nilai kosong): diisi menggunakan nilai median per kategori produk untuk mempertahankan distribusi data yang representatif.
3. Kolom *Discount_Applied* (4.199 nilai kosong): diisi dengan nilai *False* berdasarkan asumsi logis bahwa ketiadaan informasi diskon berarti tidak ada diskon yang diterapkan.
4. Kolom *Item* (1.213 nilai kosong): diisi dengan nama kategori produk sebagai nilai pengganti. Kolom *Transaction_Date* dikonversi ke tipe *datetime* dan nama kolom distandardisasi dengan mengganti spasi menggunakan *underscore*.

5. Selain penanganan missing values, tahap pra-pemrosesan juga mencakup encoding terhadap tiga kolom kategorikal (Category, Location, Payment_Method) menjadi representasi numerik menggunakan Label Encoding agar dapat diproses oleh algoritma K-Means.

C. Seleksi dan Encoding Fitur Transaksi

Penelitian ini menggunakan enam fitur transaksi sebagai variabel clustering, yaitu tiga fitur numerik (Quantity, Price_Per_Unit, Total_Spent) dan tiga fitur kategorikal yang telah di-encode (Category, Location, Payment_Method). Quantity merepresentasikan jumlah barang yang dibeli dalam satu transaksi, Price_Per_Unit merepresentasikan harga satuan produk, dan Total_Spent merepresentasikan total nilai transaksi. Ketiga fitur kategorikal diubah menjadi representasi numerik menggunakan Label Encoding agar dapat diproses bersama fitur numerik dalam algoritma K-Mean[11].

D. Normalisasi Data

Ketiga fitur RFM memiliki skala yang berbeda-beda sehingga diperlukan proses normalisasi sebelum menjalankan algoritma K-Means [6]. Normalisasi dilakukan menggunakan metode *Standardization* (*StandardScaler*) yang mentransformasi setiap fitur sehingga memiliki rata-rata nol dan standar deviasi satu. Hal ini bertujuan untuk mencegah fitur dengan skala besar mendominasi perhitungan jarak *Euclidean* yang digunakan oleh algoritma K-Means dalam proses pengelompokan.

E. Penentuan Nilai K Optimal (Ellbow Method)

Sebelum menjalankan algoritma K-Means, perlu ditentukan terlebih dahulu jumlah *cluster* (K) yang paling optimal. Penelitian ini menggunakan *Elbow Method* dengan cara menghitung nilai *inertia* untuk setiap nilai K dari 1 hingga 10 [4]. Hasil perhitungan divisualisasikan dalam bentuk grafik garis. Titik di mana penurunan nilai *inertia* mulai melambat secara signifikan yang menyerupai bentuk siku lengan ditentukan sebagai nilai K optimal yang akan digunakan pada proses *clustering*.

F. Penerapan Algoritma K-Means Clustering

Algoritma K-Means dijalankan menggunakan nilai K optimal yang telah ditentukan melalui *Elbow Method*. Algoritma bekerja melalui empat langkah iteratif [4] : (1) inisialisasi sejumlah K titik pusat (*centroid*) secara acak; (2) setiap data pelanggan ditetapkan ke *cluster* dengan *centroid* terdekat berdasarkan jarak *Euclidean*; (3) *centroid* diperbarui dengan menghitung rata-rata dari seluruh anggota *cluster*; (4) langkah 2 dan 3 diulang hingga tidak ada perubahan penugasan *cluster* (*converge*). Implementasi menggunakan fungsi *KMeans* dari library *Scikit-learn* pada Python.

G. Evaluasi Model

Kualitas hasil *clustering* dievaluasi menggunakan dua metrik, yaitu *Silhouette Score* dan *Inertia* [2]. *Silhouette Score* mengukur seberapa mirip sebuah data dengan *cluster*-nya sendiri dibandingkan dengan *cluster* lain, dengan nilai berkisar antara -1 hingga 1 di mana nilai mendekati 1 mengindikasikan *cluster* yang terbentuk dengan baik. *Inertia* mengukur total jarak kuadrat antara setiap titik data dengan *centroid cluster*-nya, di mana nilai yang lebih kecil menunjukkan *cluster* yang lebih kompak dan homogen.

H. Implementasi Sistem Berbasis Web

Seluruh tahapan analisis di atas diintegrasikan ke dalam sebuah aplikasi berbasis web agar dapat digunakan oleh pengguna secara langsung tanpa memerlukan keahlian teknis pemrograman [8]. Sistem menyediakan empat halaman utama: (1) halaman *upload* data untuk mengunggah file CSV; (2) halaman *preprocessing* yang menampilkan laporan hasil *data cleaning* secara otomatis; (3) halaman proses *clustering*

yang menjalankan *Elbow Method* dan K-Means; serta (4) halaman hasil yang menampilkan visualisasi *scatter plot*, tabel RFM per pelanggan, profil tiap *cluster*, dan metrik evaluasi [7]. Penelitian Hermawan dkk. [12] menunjukkan bahwa visualisasi hasil clustering melalui dashboard berbasis web (Google Looker Studio) memudahkan pihak non-teknis (manajemen fakultas) dalam memahami pola data dan mengambil insight, tanpa perlu memahami proses komputasi K-Means secara mendalam.

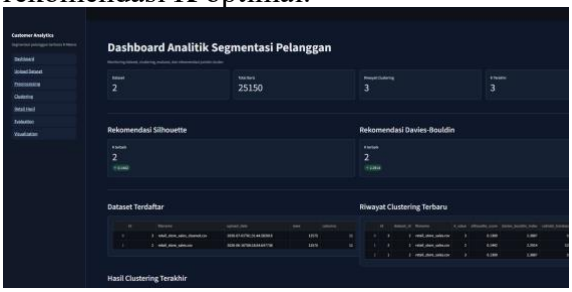
III. HASIL DAN PEMBAHASAN

Berbeda dengan pendekatan RFM konvensional yang melakukan agregasi data per pelanggan, penelitian ini menerapkan segmentasi pada level transaksi mengingat jumlah pelanggan unik pada dataset relatif sedikit (25 pelanggan) sehingga kurang representatif apabila dianalisis secara statistik pada level individu pelanggan. Unit analisis yang digunakan dalam penelitian ini adalah setiap baris transaksi, sehingga seluruh 12.575 transaksi dapat dimanfaatkan secara penuh dalam proses *clustering* tanpa mereduksi ukuran data [9].

A. Fitur-Fitur Sistem yang di Bangun

Sistem berbasis web yang dibangun mengintegrasikan seluruh tahapan analisis ke dalam tujuh menu utama. Berikut ini deskripsi lengkap setiap fitur sistem :

1. Dashboard
   Menampilkan ringkasan monitoring sistem : jumlah dataset, total baris data, Riwayat clustering, nilai K terakhir, dan rekomendasi K optimal.



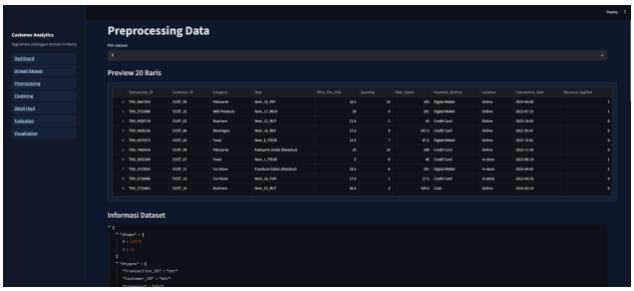
Gambar 2. Dashboard Sistem

2. Upload Dataset
   Mengunggah file CSV data transaksi retail lalu system akan memvalidasi struktur data secara otomatis
3. Prepocessing
   Membersihkan data secara otomatis berupa menangani missing values, duplikat, konversi tipe data, dan encoding variabel kategorikal.
4. Clustering
   Menormalisasi fitur transaksi, menjalankan Elbow Method untuk menentukan K optimal, dan menjalankan algoritma K-Means.
5. Detail Hasil
   Menampilkan rincian hasil clustering per transaksi beserta label cluster yang diperoleh.
6. Evaluation
   Menampilkan metrik evaluasi Silhouette Score dan Davies-Bouldin Index untuk beberapa nilai K.
7. Visualization
   Menyajikan grafik Elbow Method, scatter plot sebaran cluster, dan perbandingan karakteristik antar cluster.

Berdasarkan fitur-fitur yang ada di dalam sistem, sistem dirancang agar seluruh proses berjalan secara otomatis dari unggah data hingga rekomendasi bisnis. Pengguna tidak perlu menentukan jumlah cluster secara manual, karena sistem telah menggunakan Elbow Method untuk merekomendasikan nilai K secara otomatis, kemudian menjalankan K-Means dan menampilkan hasilnya dalam visualisasi yang mudah dipahami

B. Hasil Pra-pemrosesan Data (Data Cleaning)

Dataset dalam kondisi awal memiliki 12.575 baris dengan total 7.229 nilai kosong yang tersebar pada lima kolom.



Gambar 3. Pra-pemrosesan data pada sistem

Tabel 1 menyajikan ringkasan proses data cleaning yang dilakukan.

Tabel 1 Hasil Data Cleaning

No	Proses Cleaning	Jumlah Data	Keterangan
1.	Hapus Duplikat	0	Tidak ada duplikat ditemukan
2.	Konversi Transaction Date	-	String diubah ke tipe datetime
3.	Isi Discount Applied	4.199	Nilai NaN diisi dengan False
4.	Isi Price Per Unit	609	Dihitung dari Total Spent / Quantity
5.	Isi Quantity	604	Diisi median per kategori produk
6.	Isi Total Spent	604	Dihitung ulang dari Price x Quantity
7.	Isi Kolom Item	1.213	Diisi nama kategori + keterangan
8.	Rename Kolom	3	Category, Location, Payment_Method diubah ke numerik

Seluruh nilai kosong berhasil ditangani tanpa menghapus satu pun baris data, sehingga jumlah data tetap 12.575 baris. Selain penanganan *missing values*, tahap *preprocessing* juga mencakup *encoding* terhadap tiga kolom kategorikal *Category*, *Location*, dan *Payment_Method* menjadi representasi numerik menggunakan *Label Encoding* agar dapat diproses oleh algoritma K-Means[1]. Hasil pemetaan *encoding* disajikan pada Tabel 2.

Table 2. Hasil Encoding variable kategorikal

Kolom	Jumlah Kategori	Hasil Encoding
Category	8 kategori (0-7)	Beverages=0, Butchers=1, Computers and electric accessories=2, Electric household essentials=3, Food=4, Furniture=5, Milk Products=6, Patisserie=7
Location	2 kategori (0-1)	In-store=0, Online=1
Payment_Method	3 kategori (0-2)	Cash=0, Credit Card=1, Digital Wallet=2

C. Fitur yang Digunakan untuk Clustering

Penelitian ini menggunakan enam fitur transaksi sebagai variabel clustering: tiga fitur numerik (Quantity, Price_Per_Unit, Total_Spent) dan tiga fitur kategorikal yang telah di-encode (Category, Location, Payment_Method). Tabel 5 menyajikan statistik deskriptif untuk ketiga fitur numerik

Tabel 3. Statistic deskriptif Fitur Numerik

Statistik	Quantity	Price Per unit	Total Spent
Minimum	1	$ 5,00	$ 5,00
Maksimum	10	$ 41,00	$ 410,00
Rata – rata	5,55	$ 23,37	$ 130,08
Standar deviasi	2,79	$ 10,75	$ 93,54

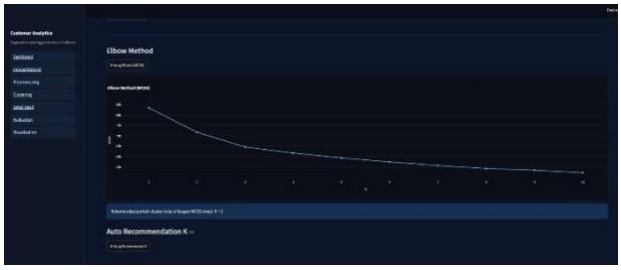
D. Hasil Elbow Method dan Penentuan K optimal

Elbow Method diterapkan dengan menghitung nilai inertia untuk K=1 hingga K=10. Tabel 4 menyajikan inertia yang dihasilkan.

Tabel 4. Nilai Inertia per Nilai K

Nilai K	Inertia	Keterangan
K=1	75,450,00	Sangat tinggi (seluruh data dalam 1 cluster)
K=2	58.425,88	Penurunan mulai melandai (titik elbow) = Optimal
K=3	51.404,27	Melandai
K=4	45.846,20	Melandai
K=5	41,788,39	Melandai
K=6	38,812,52	Melandai
K=7	36.629,61	Melandai
K=8	34.452,72	Melandai
K=9	32.799,80	Melandai
1K=0	31.141,08	Melandai

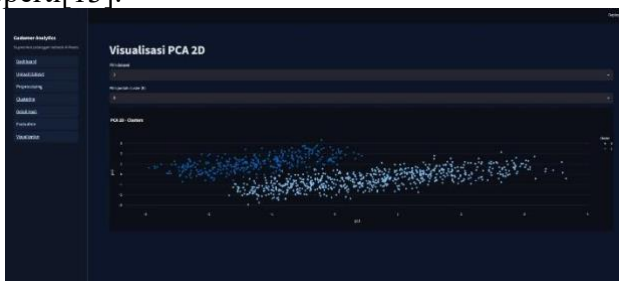
Berdasarkan Tabel 4, penurunan nilai *inertia* terbesar terjadi dari K=1 (75.450,00) ke K=2 (58.425,88), kemudian dari K=2 ke K=3 (51.404,27). Setelah K=2, laju penurunan *inertia* mulai melandai secara konsisten dan tidak lagi menunjukkan perubahan signifikan antar nilai K berikutnya. Titik *elbow* pada grafik teridentifikasi di sekitar **K=2**, sehingga nilai ini menjadi kandidat utama jumlah *cluster* yang optimal.



Gambar 4. Hasil Elbow Method pada sistem

E. Hasil K-Means Clustering (K=2)

Algoritma K-Means dijalankan dengan K=2 sesuai rekomendasi *Elbow Method*. Visualisasi hasil clustering menggunakan metode Principal Component Analysis (PCA) untuk mereduksi enam fitur ke dalam dua dimensi sehingga dapat ditampilkan dalam bentuk scatter plot seperti[13].



Gambar 5. Hasil K-Means Clustering

Dari 12.575 transaksi, terbentuk dua cluster dengan profil yang berbeda sebagaimana disajikan pada Tabel 5.

Tabel 5.Profil hasil clustering transaksi (K=2)

Cluster	Label	Jumlah Transaksi	Qty Rata-rata	Total belanja rata-rata	Karakteristik utama
0	Pembeli Regular	7.669 (61,0%)	4,29 unit	$ 67,44	Quantity & nilai transaksi sedang-rendah, Price Per Unit lebih rendah ($5,00–$41,00)
1	Pembeli Premium	4.906 (39,0%)	7,53 unit	$ 227,98	Quantity tinggi, harga satuan tinggi ($12,50–$41,00), nilai transaksi jauh lebih tinggi

Berdasarkan Tabel 5, Cluster 0 (Pembeli Reguler) beranggotakan 7.669 transaksi (61,0%) dengan rata-rata *Quantity* 4,29 unit dan nilai transaksi rata-rata $67,44. Kelompok ini merepresentasikan transaksi dengan volume dan nilai pembelian pada tingkat sedang-rendah dan merupakan kelompok terbesar dalam dataset. Cluster 1 (Pembeli Premium) beranggotakan 4.906 transaksi (39,0%) dengan rata-rata *Quantity* lebih tinggi (7,53 unit) dan nilai transaksi rata-rata $227,98 — jauh di atas Cluster 0, menjadikannya kontributor pendapatan terbesar meskipun proporsinya lebih kecil.

Untuk memastikan bahwa pembentukan *cluster* tidak hanya didominasi oleh fitur kategorikal, dilakukan analisis distribusi variabel *Category*, *Location*, dan *Payment_Method* pada setiap *cluster*, sebagaimana disajikan pada Tabel 6.

Tabel 6. Distribusi variabel kategorikal per cluster

Cluster	Proporsi	In-store/Online	Distribusi Kategori Produk	Distribusi Metode Bayar
0 (Regular)	61,0%	49,8% / 50,2%	Merata di semua kategori produk	Cash 33,9% / CC 32,9% / DW 32,5%
1 (Premium)	39,0%	59,0% / 51,0%	Merata di semua kategori produk	Cash 34,9% / CC 32,6% / DW 32,5%

Berdasarkan Tabel 6, distribusi kategori produk, lokasi transaksi (*in-store*/*online*), dan metode pembayaran relatif merata pada ketiga *cluster* (proporsi mendekati 33% pada setiap kombinasi). Hal ini menunjukkan bahwa pembentukan *cluster* didominasi oleh pola nilai transaksi (*Qduauantity*, *Price_Per_Unit*, *Total_Spent*) dan bukan oleh preferensi kategori produk, lokasi, maupun metode pembayaran tertentu.

F. *Evaluasi Model*

Kualitas hasil clustering dievaluasi menggunakan *Silhouette Score* dan *Inertia* untuk beberapa nilai K sebagai pembanding. Tabel 7 menyajikan hasil evaluasi untuk K=2, K=3, dan K=4.

Tabel 7. Perbandiangan nilai evaluasi model

Nilai K	Silhouette Score	Davies - Bouldin	Keterangan

K=2	0,2086	1,7673	Dipilih selaras dengan titik elbow, Silhouette tertinggi ← Optimal
K=3	0,1701	1,9277	Silhouette menurun namun tidak selaras dengan titik elbow
K=4	0,1877	1,8382	Silhouette naik tipis namun tidak selaras dengan titik elbow

Berdasarkan Tabel 7, K=2 menghasilkan *Silhouette Score* tertinggi (0,2086), namun hanya membagi data menjadi dua segmen yang kurang granular untuk kebutuhan interpretasi bisnis. K=2 menghasilkan *Silhouette Score* sebesar 0,2086 dan *Davies-Bouldin Index* sebesar 1,7637[14]. Dengan begitu diketahui *Silhouette Score* pada K=2 sedikit lebih tinggi dibanding K=3 dan K=4, nilai K=2 dipilih karena selaras dengan titik *elbow* yang teridentifikasi pada Tabel 6 dan menghasilkan dua segmen pembeli yang lebih bermakna dan dapat diinterpretasikan secara bisnis. Nilai *Silhouette Score* yang berada pada kisaran 0,17–0,21 tergolong moderat, yang dapat dijelaskan oleh karakteristik data transaksi yang bersifat kontinu dan tidak memiliki batas pemisah alami yang tegas antar segmen. Meskipun demikian, pola perbedaan nilai transaksi antar *cluster* pada Tabel 5 tetap menunjukkan perbedaan yang konsisten dan dapat diinterpretasikan.

G. Implementasi dan Pengujian Berbasis Web

Seluruh proses yang telah diuraikan di atas telah berhasil diimplementasikan secara penuh dalam sistem berbasis web. Sistem berjalan sesuai alur: unggah data pada menu *Upload Dataset* → pembersihan dan *encoding* data pada menu *Preprocessing* → normalisasi fitur dan eksekusi *Elbow Method* serta K-Means pada menu *Clustering* → rincian hasil pada menu *Detail Hasil* → evaluasi kualitas model pada menu *Evaluation* → visualisasi hasil pada menu. PCA digunakan untuk menyederhanakan visualisasi dari enam variabel menjadi dua komponen utama yang merepresentasikan variansi terbesar dalam dataset [15]. *Visualization*.Sistem berhasil membuktikan bahwa penentuan jumlah cluster tidak harus dilakukan secara asumtif, melainkan dapat diperoleh secara objektif dan otomatis melalui *Elbow Method*, yang dalam penelitian ini merekomendasikan K=2 sebagai konfigurasi optimal untuk data penjualan retail yang digunakan.

IV. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa algoritma K-Means *Clustering* dapat diterapkan untuk mengelompokkan data penjualan retail berdasarkan karakteristik transaksi tanpa penetapan jumlah *cluster* secara manual. Sebelum proses *clustering* dijalankan, kelayakan data diukur menggunakan *Hopkins Statistic* yang menghasilkan nilai 0,9567 — mendekati 1,0 dan mengindikasikan bahwa data transaksi memiliki struktur *cluster* yang sangat kuat sehingga sangat layak dianalisis menggunakan K-Means. Melalui *Elbow Method*, sistem secara otomatis merekomendasikan K=2 sebagai konfigurasi optimal, yang dikonfirmasi oleh *Silhouette Score* sebesar 0,2086 dan *Davies-Bouldin Index* sebesar 1,7673 — keduanya merupakan nilai terbaik dibandingkan kandidat K lainnya. Hasil *clustering* membagi 12.575 transaksi menjadi dua kelompok: Pembeli Reguler (7.669 transaksi, 61,0%, rata-rata $67,44) dan Pembeli Premium (4.906 transaksi, 39,0%, rata-rata $227,98), dengan perbedaan utama terletak pada pola kuantitas dan nilai pembelian, bukan pada preferensi kategori produk, lokasi, maupun metode pembayaran.

Dari sisi pengembangan sistem, seluruh tahapan analisis berhasil diintegrasikan ke dalam aplikasi berbasis web yang dapat dioperasikan tanpa keahlian teknis pemrograman. Sistem ini menjawab kesenjangan penelitian terdahulu yang umumnya berhenti pada tahap analisis tanpa menyediakan antarmuka yang dapat digunakan langsung oleh pengelola retail. Dengan demikian, tujuan penelitian telah tercapai dan sistem ini diharapkan dapat menjadi alat bantu pengambilan keputusan bisnis yang efektif bagi pelaku usaha retail.

REFERENSI

[1] S. P. E-commerce, "Implementasi Algoritma Clustering K-Means untuk," vol. 6, no. 3, pp. 1680–1686, 2025.

[2] Nurmilayanti and N. Hartono, "Implementasi Metode Clustering Sebagai Penunjang Strategi dalam Manajemen Pelanggan," *J. Fasilkom*, vol. 13, no. 3, pp. 605–613, 2023, doi: 10.37859/jf.v13i3.5617.

[3] I. R. Effendy, Y. A. Manurung, and M. I. Hujaifah, "Implementasi Algoritma K-means Clustering untuk Mencari Preferensi Pelanggan Toko Online Tazeee Clothes JURNAL MEDIA INFORMATIKA [JUMIN]," vol. 7, no. 1, pp. 69–76, 2026.

[4] H. Safitri, S. Putri, L. Geni, F. Merry, and M. Wati, "Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Berdasarkan Pola Pembelian," vol. 7, pp. 89–99, 2025.

[5] A. Yusak, N. Rumapea, D. Pratiwi, and S. Sari, "Analisis Segmentasi Pelanggan Ritel Online Menggunakan K-Means Clustering Berdasarkan Model Recency , Frequency , Monetary (RFM)," vol. 6, no. 3, pp. 292–299, 2024.

[6] B. T. Kristanti, A. Junaidi, and E. P. Mandyartha, "Implementasi K-Means Clustering Dalam Segmentasi Pelanggan Berdasarkan Usia, Pendapatan, Dan Model Rfm (Studi Kasus: Lantikya Store Jombang)," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4677.

[7] A. A. Alya Putri and S. A. Rahmah, "Implementasi Data Mining Dengan Algoritma K-Means Clustering Untuk Analisis Bisnis Pada Perusahaan Asuransi," *Djtechno J. Teknol. Inf.*, vol. 5, no. 1, pp. 139–152, 2024, doi: 10.46576/djtechno.v5i1.4537.

[8] A. Dzakiroh and E. Widodo, "Implementasi Algoritma K-Means Untuk Segmentasi Siswa Dalam Pembentukan Kelas Unggulan," vol. 11, no. 1, pp. 34–43, 2025.

[9] A. Agung, A. Daniswara, and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," vol. 05, pp. 97–100, 2023.

[10] A. K. Clustering and G. Belajar, "Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar," vol. 12, 2022, doi: 10.34010/jati.v12i1.

[11] L. P. Wiwien Widhyastuti, I. N. Sukajaya, and K. Y. Ernanda Aryanto, "Customer Profiling berdasarkan Model RFM dengan Metode K-Means pada Institusi Pendidikan untuk menunjang Strategi," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 2, pp. 94–108, 2022.

[12] S. Adinda Hermawan, D. R. Sari Dewi Hartanto, D. M. Wiharta, and I. M. Arsa Suyadnya, "Visualisasi Data Pengelompokkan Kelulusan Mahasiswa Dengan Algoritma Clustering K-Means," *J. SPEKTRUM*, vol. 11, no. 1, p. 86, 2024, doi: 10.24843/spektrum.2024.v11.i01.p10.

[13] D. E. Putri and E. P. W. Mandala, "Penerapan K-Means Clustering dalam Segmentasi Siswa Berdasarkan Status Sosial Ekonomi," *Progresif J. Ilm. Komput.*, vol. 21, no. 2, p. 483, 2025, doi: 10.35889/progresif.v21i2.2809.

[14] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster," *Kumpul. Artik. Karya Ilm. Fak. Ilmu Komput.*, vol. 06, no. 01, pp. 60–74, 2024.

[15] R. Rianti, R. Andarsyah, and R. M. Awangga, "Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV," *Nuansa Inform.*, vol. 18, no. 2, pp. 67–77, 2024, doi: 10.25134/ilkom.v18i2.211.