Implementasi K-Means Clustering Untuk Segmentasi Pelanggan Menggunakan Customer Shopping Trend Dataset

Ethic Nova Herlina [1*], Septian Damar Pamudya[2], Erika Jonet Febrianto [3], Victorio Mikhael David Pratama[4]

[1]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[1*]230103267@mhs.udb.ac.id (penulis korespondensi

[2] Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[2] 230103208@mhs.udb.ac.id

[3] Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[3] 202030264@mhs.udb.ac.id

[4]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[4] 230103210@mhs.udb.ac.id

Abstrak— **Perkembangan perdagangan digital menuntut perusahaan ritel memahami perilaku pelanggan secara mendalam untuk menyusun strategi pemasaran yang efektif. Penelitian ini bertujuan mengelompokkan pelanggan berdasarkan kemiripan pola perilaku belanja konsumen menggunakan algoritma *K-Means clustering* melalui pendekatan *Knowledge Discovery in Databases* (KDD). Eksperimen dilakukan terhadap 1.000 *records* data dari *Customer Shopping Trends Dataset* dengan memfokuskan analisis pada empat atribut numerik utama, yaitu *Age*, *Purchase Amount (USD)*, *Previous Purchases*, dan *Review Rating*. Tahap pra-pemrosesan data diterapkan secara konsisten menggunakan metode *StandardScaler* untuk menyetarakan skala dimensi antar-fitur guna menghindari bias dimensi. Berdasarkan hasil pengujian menggunakan *Elbow Method*, jumlah kelompok paling optimal yang terbentuk adalah sebanyak tiga klaster ($K = 3$). Proses *profiling* berbasis nilai rata-rata berhasil mengidentifikasi Klaster 0 sebagai segmen *"Loyal High-Spenders (Mature)"*, Klaster 1 sebagai segmen *"Satisfied New or Occasional Shoppers"*, dan Klaster 2 sebagai segmen *"Dissatisfied Moderate Shoppers"* yang memiliki risiko tinggi melakukan migrasi ke kompetitor (*customer churn risk*). Pengujian kualitas menggunakan *Silhouette Score* menghasilkan nilai akhir sebesar 0,187, yang mengindikasikan bahwa hasil klasterisasi berada pada kategori struktur yang lemah (*weak structure*). Karakteristik batas pemisah yang cair dan saling tumpang tindih (*overlapping*) ini secara objektif merepresentasikan sifat asli distribusi data perilaku belanja manusia di dunia nyata yang bersifat kontinu dan seragam. Hasil segmentasi berbasis data komputasi ini memberikan implikasi manajerial yang berharga bagi manajemen perusahaan ritel dalam merumuskan strategi pemasaran dan retensi pelanggan yang lebih personal serta tepat sasaran.**
Kata kunci— K-Means Clustering, Segmentasi Pelanggan, KDD, StandardScaler, Silhouette Score.

Abstract— ***The rapid growth of digital commerce demands retail companies to understand customer behavior deeply to formulate effective marketing strategies. This study aims to group customers based on similarities in consumer shopping patterns using the K-Means clustering algorithm within the Knowledge Discovery in Databases (KDD) framework. Experiments were conducted on 1,000 records from the Customer Shopping Trends Dataset, focusing on four primary numerical attributes: Age, Purchase Amount (USD), Previous Purchases, and Review Rating. The data preprocessing phase consistently applied the StandardScaler method to balance the scale across feature dimensions and prevent dimensional bias. Based on the Elbow Method evaluation, the most optimal number of clusters formed was three (K=3). Data-driven profiling identified Cluster 0 as the "Loyal High-Spenders (Mature)" segment, Cluster 1 as the "Satisfied New or Occasional Shoppers" segment, and Cluster 2 as the "Dissatisfied Moderate Shoppers" segment, which presents a high customer churn risk. Model quality validation using the Silhouette Score yielded a final score of 0.187, indicating a weak cluster structure. This fluid and overlapping boundary objectively represents the intrinsic nature of real-world human shopping data, which is continuous and uniformly distributed. The findings of this computational segmentation offer valuable managerial implications for retail management in designing personalized and targeted marketing and customer retention strategies.***
Keywords— K-Means Clustering, Customer Segmentation, KDD, StandardScaler, Silhouette Score.

V. PENDAHULUAN

Perkembangan perdagangan digital yang masif dan meningkatnya intensitas persaingan di sektor ritel modern mendorong perusahaan untuk memahami perilaku pelanggan secara lebih mendalam. Di tengah dinamika pasar yang kompetitif, perusahaan ritel tidak lagi dapat mengandalkan strategi pemasaran massal yang

seragam untuk semua konsumen. Segmentasi pelanggan menjadi pendekatan yang sangat krusial karena memungkinkan perusahaan mengelompokkan basis data pelanggan ke dalam kelompok-kelompok homogen berdasarkan kemiripan karakteristik atau pola pembelian tertentu. Melalui segmentasi yang tepat, strategi promosi, alokasi sumber daya, program retensi, dan kualitas pelayanan dapat disesuaikan secara personal dengan kebutuhan spesifik masing-masing segmen. Dalam era berbasis data saat ini, proses segmentasi tidak lagi boleh hanya mengandalkan intuisi manajemen, melainkan harus memanfaatkan teknik data mining untuk memperoleh pola perilaku yang lebih objektif, terukur, dan berbasis data ilmiah [1].

Salah satu metode pembelajaran tanpa pengawasan yang paling banyak digunakan dalam segmentasi pelanggan adalah algoritma *K-Means clustering*. Algoritma ini bekerja secara iteratif untuk mengelompokkan data ke dalam beberapa klaster berdasarkan tingkat kemiripan terkecil antar-atribut menggunakan perhitungan jarak spasial. Hasil akhir dari algoritma ini adalah terbentuknya kelompok pelanggan yang memiliki homogenitas internal yang tinggi di dalam klaster, namun memiliki heterogenitas eksternal yang kontras antar-klaster. Penggunaan *K-Means* dinilai sangat efisien dari segi komputasi dan memiliki kemampuan adaptasi yang baik ketika diimplementasikan pada dataset transaksi ritel berskala besar [2].

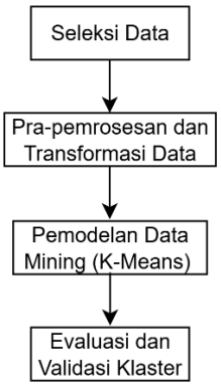
Meskipun penerapan *K-Means* untuk segmentasi pelanggan telah banyak diteliti sebelumnya, terdapat *research gap* yang signifikan dalam literatur yang ada. Sebagian besar penelitian terdahulu cenderung melakukan simplifikasi dengan hanya berfokus pada analisis segmentasi statis yang menitikberatkan pada aspek demografi atau volume transaksi tunggal. Pendekatan tradisional tersebut kerap kali mengabaikan interaksi multidimensi antara rekam jejak finansial dengan indikator kepuasan pelanggan secara simultan. Kekurangan utama dari model yang terlalu kaku ini adalah ketidakmampuannya dalam menangkap anomali perilaku pasar di dunia nyata—misalnya, mendeteksi kelompok pelanggan yang memiliki volume belanja tinggi namun secara emosional memiliki tingkat kepuasan yang rendah. Kegagalan dalam mengintegrasikan variabel kepuasan ini menyebabkan strategi pemasaran yang dihasilkan menjadi kurang personal, tidak akurat, serta gagal mengantisipasi risiko migrasi pelanggan ke kompetitor sejak dini [3].

Oleh karena itu, penelitian ini menerapkan kerangka kerja *Knowledge Discovery in Databases* (KDD) untuk mengeksplorasi interaksi multidimensi dari empat atribut utama: *Age*, *Purchase Amount*, *Previous Purchases*, dan *Review Rating*. Kebaruan yang ditawarkan dalam penelitian ini terletak pada integrasi multidimensi antara rekam jejak perilaku transaksi ritel komersial dengan metrik penilaian subjektif kepuasan pelanggan (Review Rating) di bawah kerangka kerja KDD, guna mengidentifikasi pergeseran segmen pasar secara lebih realistis.

Melalui penerapan algoritma *K-Means* berbasis kerangka kerja KDD ini, penelitian ini bertujuan untuk mengelompokkan pelanggan ke dalam segmen-segmen strategis yang lebih realistis berdasarkan kemiripan pola belanja, serta mengidentifikasi profil karakteristik utama dari setiap segmen. Hasil dari proses klasterisasi berbasis data komputasi ini diharapkan dapat memberikan kontribusi praktis yang bernilai bagi manajemen perusahaan ritel dalam merumuskan peta jalan operasional, memitigasi risiko *customer churn*, serta merancang strategi pemasaran serta retensi yang lebih personal, efisien, dan tepat sasaran.

VI. METODOLOGI PENELITIAN

Bagian ini menjelaskan secara rinci tentang tahapan penelitian yang dilakukan dalam mengimplementasikan algoritma K-Means *clustering* untuk segmentasi pelanggan. Alur penelitian yang digunakan dapat dilihat di Gambar 1.



Gambar 1. Diagram alir

Diagram Alir menggambarkan empat tahapan KDD dalam melakukan segmentasi pelanggan. Proses dimulai dengan Seleksi Data untuk memilih empat atribut perilaku belanja utama dari *dataset*. Selanjutnya, dilakukan Pra-pemrosesan dan Transformasi Data melalui pembersihan serta standardisasi skala nilai menggunakan *StandardScaler*. Data yang telah seragam kemudian diproses pada tahap Pemodelan Data Mining (K-Means) untuk dikelompokkan secara iteratif berdasarkan jarak *Euclidean*. Rangkaian metodologi ini diakhiri dengan Evaluasi dan Validasi Klaster menggunakan metode *Elbow* dan *Silhouette Score* guna menjamin kualitas serta optimalitas jumlah segmen yang terbentuk.

A. Pengumpulan data

Data yang digunakan dalam penelitian ini adalah *Customer Shopping Trend Dataset* yang memuat informasi mengenai karakteristik demografis dan perilaku transaksi pelanggan ritel. Atribut di dalam dataset ini terdiri dari kombinasi data numerik (seperti *Age, Purchase Amount*, dan *Review Rating*) serta data kategorikal (seperti *Subscription Status, Frequency of Purchases*, dan *Payment Method*). Informasi umum mengenai variabel utama yang digunakan dalam proses klasterisasi ditampilkan pada Tabel I. Atribut-atribut ini dipilih secara selektif dari repositori publik untuk merepresentasikan preferensi serta rekam jejak loyalitas konsumen secara berkala [4]

Tabel 1. Karakteristik Variabel Dataset

Nama Variabel	Jenis Data	Keterangan
Age	Numerik	Usia dari pelanggan retail
Purchase Amount (USD)	Numerik	Nilai transasi dalam satuan USD
Review Rating	Numerik	Skala kepuasan pelanggan terhadap produk
Previous Purchases	Numerik	Jumlah total transaksi terdahulu

B. Pra Pemrosesan dan Tranformasi Data

Tahap pra-pemrosesan dan transformasi data merupakan langkah kritis yang menentukan keberhasilan dan objektivitas pembentukan klaster. Di dalam draf data mentah, setiap atribut sering kali memiliki unit satuan, skala, dan rentang nilai yang sangat bervariasi. Sebagai contoh pada penelitian ini, atribut *Age* memiliki rentang nilai puluhan tahun dan *Purchase Amount* mencapai puluhan USD, sedangkan *Review Rating* hanya memiliki rentang sempit antara angka 1 hingga 5.

Algoritma K-Means secara inheren mengandalkan perhitungan jarak geometris (seperti *Euclidean Distance*) yang memperlakukan semua dimensi secara setara. Jika data dengan ketimpangan skala ini langsung diproses tanpa transformasi, atribut dengan rentang nilai yang besar akan mendominasi fungsi objektif dalam perhitungan jarak spasial. Sebaliknya, kontribusi dari atribut berrentang kecil seperti *Review Rating* akan terabaikan secara signifikan, sehingga menghasilkan pengelompokan yang bias dan tidak akurat [5].

Untuk mengatasi kendala tersebut dan memastikan seluruh fitur memiliki bobot yang setara (*equally weighted*), dilakukan proses standardisasi data menggunakan metode Z-score Standardization (StandardScaler). Metode ini mentransformasikan distribusi setiap atribut numerik agar berpusat pada nilai rata-rata (μ) sama dengan 0 dan memiliki varians atau standar deviasi (σ) sama dengan 1. Melalui transformasi ini, data diubah menjadi unit tanpa dimensi yang mencerminkan seberapa jauh suatu titik data menyimpang dari rata-ratanya [6]. Proses ini sangat ideal untuk K-Means karena mampu menyetarakan skala tanpa mengubah bentuk distribusi asli data maupun menghilangkan informasi *outlier* yang mungkin penting. Formula matematis yang digunakan untuk menghitung standardisasi Z-score adalah sebagai berikut:

$$Z = \frac{x - \mu}{\sigma}$$

1. Z: Nilai fitur hasil standardisasi (*standardized value*)
2. x: Nilai fitur asli sebelum transformasi (*raw value*)
3. μ: Nilai rata-rata (*mean*) dari keseluruhan sampel pada atribut tersebut
4. σ: Nilai standar deviasi (*standard deviation*) dari keseluruhan sampel pada atribut tersebut.

C. Pemodelan Data Mining (K-Means Clustering)

Setelah tahapan transformasi data selesai dan seluruh fitur berada pada skala metrik yang

seragam, langkah berikutnya adalah melakukan pemodelan menggunakan algoritma K-Means Clustering. Pemilihan algoritma K-Means didasarkan pada efisiensi komputasinya yang tinggi dalam menangani dataset numerik berskala besar serta kemampuannya yang adaptif dalam mempartisi objek ke dalam sejumlah K kelompok yang saling terpisah (*disjoint clusters*) berdasarkan kesamaan karakteristik perilaku belanja. Secara prinsip dasar, K-Means beroperasi sebagai algoritma pembelajaran tanpa pengawasan (*unsupervised learning*) yang bertujuan untuk meminimalkan varians di dalam klaster (*intra-cluster variance*) sekaligus memaksimalkan jarak antar-klaster (*inter-cluster distance*).

Rangkaian pemodelan ini diawali dengan penentuan jumlah klaster (K) yang merepresentasikan target jumlah kelompok pelanggan yang ingin dibentuk. Pada penelitian ini, nilai konstanta K tidak ditentukan secara subjektif, melainkan dikomparasikan secara empiris melalui *Elbow Method* untuk menjamin efisiensi kelompok yang dihasilkan. Setelah nilai K ditetapkan, sistem melakukan inisialisasi pusat klaster (*initial centroids*) sebanyak K secara acak di dalam ruang dimensi data sebagai titik jangkar geometris sementara. Langkah berikutnya adalah mengalkulasi jarak spasial antara setiap titik data pelanggan dengan seluruh *centroid* yang tersedia menggunakan metode *Euclidean Distance* yang dirumuskan sebagai berikut:

$$d(x,y) = \sqrt{\sum_{j=1}^{m} (x_j - y_j)^2}$$

Dalam persamaan tersebut, $d(x,y)$ menunjukkan nilai jarak Euclidean antara objek data pelanggan x dan objek pusat klaster y. Simbol m merepresentasikan jumlah dimensi atau atribut fitur yang dianalisis, yang dalam eksperimen ini bernilai 4 yaitu meliputi atribut *Age*, *Purchase Amount*, *Previous Purchases*, dan *Review Rating*. Sementara itu, variabel x_j merupakan nilai koordinat atribut ke-j pada objek data pelanggan x, dan y_j adalah nilai koordinat atribut ke-j pada objek pusat klaster (*centroid*) y.

Setelah seluruh jarak spasial berhasil dikalkulasi, proses dilanjutkan dengan melakukan alokasi objek ke klaster terdekat berdasarkan prinsip jarak minimum. Objek data pelanggan secara otomatis akan dipetakan ke dalam suatu kelompok jika hasil perhitungan jarak Euclideannya menunjukkan nilai paling kecil dibandingkan dengan jarak ke pusat klaster lainnya. Segera setelah seluruh data menempati kelompok masing-masing, sistem akan memperbarui posisi pusat klaster dengan menghitung kembali nilai rata-rata (*mean*) aritmetika dari keseluruhan nilai atribut objek data pelanggan yang telah teralokasi di dalam klaster tersebut. Seluruh rangkaian kalkulasi yang meliputi perhitungan jarak, alokasi objek, dan pembaruan *centroid* ini diulangi secara iteratif hingga sistem mencapai kondisi konvergen, yaitu ketika posisi pusat klaster tidak lagi mengalami pergeseran atau tidak ada lagi data pelanggan yang berpindah orientasi klasternya [7].

D. Evaluasi dan Validasi Klaster

Untuk memastikan bahwa hasil segmentasi pelanggan bersifat optimal, objektif, dan memiliki kualitas pengelompokan yang baik, dilakukan tahap evaluasi dan validasi kuantitatif terhadap model klaster yang terbentuk. Langkah validasi pertama diterapkan menggunakan Metode Elbow yang mengandalkan kalkulasi nilai *Within-Cluster Sum of Squares* (WCSS) untuk mengidentifikasi jumlah kelompok (K) terbaik secara empiris. Secara matematis, nilai WCSS dihitung dengan mengakumulasikan kuadrat jarak antara setiap titik data dengan pusat klaster atau *centroid* kelompoknya sendiri melalui persamaan sebagai berikut:

$$WCSS = \sum_{k=1}^{K} \sum_{x \in C_k} d\,(x, \mu_k)^2$$

Dalam formula tersebut, variabel K menunjukkan jumlah keseluruhan klaster yang sedang diuji, sementara C_k merepresentasikan objek klaster ke-k. Simbol x mengacu pada titik data pelanggan yang menjadi anggota di dalam klaster C_k, sedangkan μ_k menyatakan titik pusat klaster dari kelompok tersebut. Nilai fungsi jarak spasial antara titik data dengan pusat klasternya disimbolkan oleh $d(x, \mu_k)$ yang dihitung

berdasarkan metode jarak Euclidean. Nilai WCSS ini secara inheren akan mengalami penurunan seiring dengan bertambahnya jumlah kelompok yang diuji, sehingga penentuan jumlah klaster paling optimal dilakukan dengan menganalisis kurva grafik guna menemukan area "titik siku" (*elbow point*). Kondisi optimal ini ditunjukkan oleh titik spesifik di mana terjadi penurunan nilai WCSS yang sangat drastis sebelum titik tersebut dan mulai melandai secara signifikan setelah melewati titik tersebut [8].

Setelah estimasi jumlah klaster optimal diprediksi melalui pendekatan kurva Elbow, kekuatan dan kualitas dari struktur internal klaster yang terbentuk divalidasi lebih lanjut menggunakan pengujian *Silhouette Coefficient*. Metode validasi ini bekerja dengan memberikan penilaian numerik individual terhadap kelayakan penempatan suatu titik data ke dalam klasternya berdasarkan komparasi dua aspek utama, yaitu aspek kohesi yang mengukur kedekatan internal kelompok dan aspek separasi yang menilai tingkat keterpisahan eksternal antar-kelompok. Skor *Silhouette Coefficient* untuk sebuah objek data pelanggan i dirumuskan melalui persamaan matematika sebagai berikut:

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

Pada persamaan ini, simbol $s(i)$ melambangkan hasil akhir nilai koefisien *Silhouette* untuk titik data ke-i. Komponen $a(i)$ merepresentasikan rata-rata jarak spasial antara objek data i dengan seluruh objek data lain yang berada di dalam satu kelompok atau klaster yang sama, yang berfungsi sebagai indikator kekuatan kohesi internal. Sementara itu, komponen $b(i)$ mengukur rata-rata jarak terdekat antara objek data i terhadap seluruh titik data yang berada pada kelompok tetangga terdekatnya sebagai representasi dari aspek separasi (*nearest-cluster distance*). Nilai akhir dari pengujian ini diperoleh dengan merata-ratakan seluruh nilai koefisien objek ke dalam rentang skor antara -1 hingga +1. Skor yang bergerak mendekati nilai +1 mengindikasikan bahwa sistem berhasil memisahkan kelompok dengan sangat tegas karena jarak antar-klaster berjauhan dan struktur internalnya padat. Sebaliknya, skor yang mendekati nilai 0 mencerminkan adanya fenomena tumpang tindih (*overlapping*) di mana batas pemisah antar-kelompok pelanggan menjadi cair dan tidak tegas, serta nilai mendekati -1 menandakan tingginya probabilitas kesalahan alokasi penempatan objek data [9].

VII. HASIL DAN PEMBAHASAN

Pada penelitian ini mengimplementasikan sebuah sistem sederhana yang menerapkan algoritma K-Means Clustering untuk melakukan segmentasi pelanggan berdasarkan karakteristik perilaku belanja. Sistem dikembangkan menggunakan bahasa pemrograman Python dengan bantuan library Pandas, Scikit-Learn, dan Matplotlib. Dataset yang digunakan merupakan Customer Shopping Trends Dataset yang diperoleh dari Kaggle dengan jumlah data sebanyak 1000 pelanggan. Sebelum proses clustering dilakukan, data terlebih dahulu melalui tahap preprocessing. Tahap ini meliputi pemilihan atribut yang digunakan sebagai dasar pengelompokan, yaitu Age, Purchase Amount (USD), Previous Purchases, dan Review Rating. Keempat atribut tersebut kemudian dinormalisasi menggunakan metode StandardScaler agar seluruh atribut memiliki skala yang seimbang sehingga proses pengelompokan menjadi lebih optimal [10].

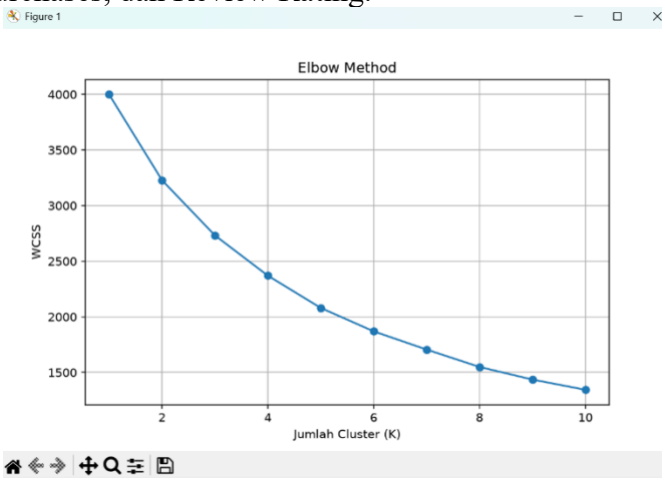
Setelah proses normalisasi selesai, algoritma K-Means diterapkan dengan jumlah cluster (K) sebanyak 3. Hasil dari proses clustering berupa label cluster yang diberikan kepada setiap pelanggan berdasarkan tingkat kemiripan karakteristiknya[11]. Label tersebut disimpan pada kolom Cluster dan selanjutnya diekspor ke dalam file hasil clustering.csv.


shopping_trends_updated.csv

```
Customer ID,Age,Gender,Item Purchased,Category,Purchase Amount (USD),Location,Size,Color,Season,Revi
1,55,Male,Blouse,Clothing,53,Kentucky,L,Gray,Winter,3.1,Yes,Express,Yes,Yes,14,Venmo,Fortnightly
2,19,Male,Sweater,Clothing,64,Maine,L,Maroon,Winter,3.1,Yes,Express,Yes,Yes,2,Cash,Fortnightly
3,50,Male,Jeans,Clothing,73,Massachusetts,S,Maroon,Spring,3.1,Yes,Free Shipping,Yes,Yes,23,Credit Ca
4,21,Male,Sandals,Footwear,90,Rhode Island,M,Maroon,Spring,3.5,Yes,Next Day Air,Yes,Yes,49,PayPal,We
5,45,Male,Blouse,Clothing,49,Oregon,M,Turquoise,Spring,2.7,Yes,Free Shipping,Yes,Yes,31,PayPal,Annua
6,46,Male,Sneakers,Footwear,20,Wyoming,M,White,Summer,2.9,Yes,Standard,Yes,Yes,14,Venmo,Weekly
7,63,Male,Shirt,Clothing,85,Montana,M,Gray,Fall,3.2,Yes,Free Shipping,Yes,Yes,49,Cash,Quarterly
8,27,Male,Shorts,Clothing,34,Louisiana,L,Charcoal,Winter,3.2,Yes,Free Shipping,Yes,Yes,19,Credit Car
9,26,Male,Coat,Outerwear,97,West Virginia,L,Silver,Summer,2.6,Yes,Express,Yes,Yes,8,Venmo,Annually
10,57,Male,Handbag,Accessories,31,Missouri,M,Pink,Spring,4.8,Yes,2-Day Shipping,Yes,Yes,4,Cash,Quart
11,53,Male,Shoes,Footwear,34,Arkansas,L,Purple,Fall,4.1,Yes,Store Pickup,Yes,Yes,26,Bank Transfer,Bi
12,30,Male,Shorts,Clothing,68,Hawaii,S,Olive,Winter,4.9,Yes,Store Pickup,Yes,Yes,10,Bank Transfer,Fo
13,61,Male,Coat,Outerwear,72,Delaware,M,Gold,Winter,4.5,Yes,Express,Yes,Yes,37,Venmo,Fortnightly
14,65,Male,Dress,Clothing,51,New Hampshire,M,Violet,Spring,4.7,Yes,Express,Yes,Yes,31,PayPal,Weekly
15,64,Male,Coat,Outerwear,53,New York,L,Teal,Winter,4.7,Yes,Free Shipping,Yes,Yes,34,Debit Card,Week
16,66,Male,Skirt,Clothing,85,Rhode Island,M,Teal,Winter,3.8,Yes,Store Pickup,Yes,Yes,8,PayPal,Monthl
17,25,Male,Sunglasses,Accessories,36,Alabama,S,Gray,Spring,4.1,Yes,Next Day Air,Yes,Yes,44,Debit Car
18,53,Male,Dress,Clothing,38,Mississippi,XL,Lavender,Winter,4.7,Yes,2-Day Shipping,Yes,Yes,36,Venmo,
19,52,Male,Sweater,Clothing,48,Montana,S,Black,Summer,4.6,Yes,Free Shipping,Yes,Yes,17,Cash,Weekly
20,66,Male,Pants,Clothing,90,Rhode Island,M,Green,Summer,3.3,Yes,Standard,Yes,Yes,46,Debit Card,Bi-w
21,21,Male,Pants,Clothing,51,Louisiana,M,Black,Winter,2.8,Yes,Express,Yes,Yes,50,Cash,Every 3 Months
22,31,Male,Pants,Clothing,62,North Carolina,M,Charcoal,Winter,4.1,Yes,Store Pickup,Yes,Yes,22,Debit
23,56,Male,Pants,Clothing,37,California,M,Peach,Summer,3.2,Yes,Store Pickup,Yes,Yes,32,Debit Card,An
24,31,Male,Pants,Clothing,88,Oklahoma,XL,White,Winter,4.4,Yes,Express,Yes,Yes,40,Credit Card,Weekly
25,18,Male,Jacket,Outerwear,22,Florida,M,Green,Fall,2.9,Yes,Store Pickup,Yes,Yes,16,Debit Card,Weekl
26,18,Male,Hoodie,Clothing,25,Texas,M,Silver,Summer,3.6,Yes,Express,Yes,Yes,14,PayPal,Annually
27,38,Male,Jewelry,Accessories,20,Nevada,M,Red,Spring,3.6,Yes,Next Day Air,Yes,Yes,13,Credit Card,An
28,56,Male,Shorts,Clothing,56,Kentucky,L,Cyan,Summer,5,Yes,Next Day Air,Yes,Yes,7,Bank Transfer,Ever
29,54,Male,Handbag,Accessories,94,North Carolina,M,Gray,Fall,4.4,Yes,Free Shipping,Yes,Yes,41,PayPal
30,31,Male,Dress,Clothing,45,Wyoming,S,Black,Fall,4.1,Yes,Store Pickup,Yes,Yes,14,Credit Card,Weekly
31,57,Male,Jewelry,Accessories,31,North Carolina,L,Black,Winter,4.7,Yes,Standard,Yes,Yes,16,Credit C
```

Gambar 2. Dataset Customer Shopping Trends

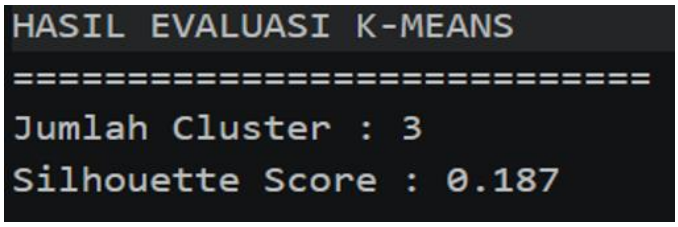
Dataset yang digunakan terdiri atas berbagai atribut pelanggan. Namun pada penelitian ini hanya digunakan empat atribut utama yang dianggap mampu merepresentasikan karakteristik pelanggan, yaitu Age, Purchase Amount (USD), Previous Purchases, dan Review Rating.



Gambar 3. Grafik Elbow Method

Sebelum proses clustering dilakukan, penelitian ini menentukan jumlah cluster menggunakan metode Elbow. Metode ini dilakukan dengan menghitung nilai Within Cluster Sum of Squares (WCSS) untuk beberapa nilai K. Nilai WCSS menunjukkan jumlah variasi data di dalam setiap cluster. Semakin kecil nilai WCSS, maka data dalam cluster semakin homogen [12].

Berdasarkan Gambar 3. terlihat bahwa terjadi penurunan WCSS yang cukup signifikan hingga K = 3, kemudian setelah itu penurunan nilai WCSS mulai melandai. Oleh karena itu, penelitian ini menggunakan 3 cluster sebagai jumlah cluster yang paling optimal dalam proses K-Means Clustering.



Gambar 4. Evaluasi Hasil Clustering Menggunakan Silhouette Score

Penelitian ini menunjukkan hasil pengujian kualitas klaster dengan *Silhouette Score* sebesar 0,187, yang menempatkan model ini dalam kategori struktur yang lemah (*weak structure*). Nilai yang rendah ini secara objektif merepresentasikan adanya tumpang tindih (*overlapping*) yang signifikan antar-segmen pelanggan, sehingga batas kelompok menjadi sangat cair dan tidak terpolarisasi secara tegas.

Hasil ini menyingkap keterbatasan teoretis algoritma K-Means yang berasumsi bahwa klaster ideal harus berbentuk bulat (*spherical*) dan terpisah jelas. Pada dataset perilaku belanja ini, sebaran data numerik bersifat kontinu dan merata (*uniform distribution*). Akibatnya, K-Means terpaksa menarik garis centroid secara artifisial di tengah kepadatan data, yang berujung pada rendahnya kohesi internal kelompok dan rapatnya jarak antar-klaster.

Dikaitkan dengan tujuan penelitian untuk segmentasi pasar, struktur yang lemah ini mengindikasikan bahwa perusahaan retail sebaiknya menerapkan strategi pemasaran yang fleksibel (*fluid marketing strategy*) daripada target promosi yang kaku, terutama pada pelanggan di area perbatasan klaster. Temuan ini juga sejalan dengan penelitian terdahulu yang menyatakan bahwa pelibatan unsur demografi kontinu (seperti usia) rentan menimbulkan *noise* dan menurunkan *Silhouette Score*. Sebaliknya, studi terdahulu yang murni menggunakan variabel transaksional seperti kerangka *Recency, Frequency, and Monetary* (RFM) terbukti mampu menghasilkan pemisahan klaster yang jauh lebih tegas [13].

Guna mengatasi keterbatasan ini, beberapa alternatif perbaikan untuk penelitian selanjutnya meliputi:

A. Optimasi Fitur (RFM): Menerapkan variabel perilaku transaksional murni (RFM) untuk mengurangi *noise* dari atribut demografi atau rating subjektif.
B. Algoritma Clustering Alternatif: Menggunakan metode *soft clustering* seperti *Fuzzy C-Means* (FCM) atau *Gaussian Mixture Models* (GMM) yang lebih adaptif pada area *overlapping*, atau algoritma berbasis kerapatan seperti DBSCAN.
C. Reduksi Dimensi (PCA): Mengintegrasikan metode seperti PCA atau t-SNE sebelum proses clustering untuk mengeliminasi multikolinearitas dan menekan hilangnya informasi visual (*projection loss*).

=== HASIL CLUSTERING ===

	Age	Purchase Amount (USD)	Previous Purchases	Review Rating	Cluster
0	55	53	14	3.1	2
1	19	64	2	3.1	1
2	50	73	23	3.1	2
3	21	90	49	3.5	0
4	45	49	31	2.7	2
5	46	20	14	2.9	2
6	63	85	49	3.2	0
7	27	34	19	3.2	2
8	26	97	8	2.6	2
9	57	31	4	4.8	1
10	53	34	26	4.1	1
11	30	68	10	4.9	1
12	61	72	37	4.5	0
13	65	51	31	4.7	0
14	64	53	34	4.7	0
15	64	81	8	2.8	2
16	25	36	44	4.1	0
17	53	38	36	4.7	0
18	52	48	17	4.6	1
19	66	90	46	3.3	0

Gambar 5. Hasil Clustering

Gambar tersebut menunjukkan bahwa sistem berhasil mengelompokkan setiap pelanggan ke dalam salah satu cluster berdasarkan karakteristik yang dimiliki. Setiap pelanggan memperoleh label cluster secara otomatis setelah proses perhitungan selesai dilakukan.

hasil_clustering.csv

```
Age,Purchase Amount (USD),Previous Purchases,Review Rating,Cluster
55,53,14,3.1,2
19,64,2,3.1,1
50,73,23,3.1,2
21,90,49,3.5,0
45,49,31,2.7,2
46,20,14,2.9,2
63,85,49,3.2,0
27,34,19,3.2,2
26,97,8,2.6,2
57,31,4,4.8,1
53,34,26,4.1,1
30,68,10,4.9,1
61,72,37,4.5,0
65,51,31,4.7,0
64,53,34,4.7,0
64,81,8,2.8,2
25,36,44,4.1,0
53,38,36,4.7,0
52,48,17,4.6,1
66,90,46,3.3,0
21,51,50,2.8,2
31,62,22,4.1,1
56,37,32,3.2,2
31,88,40,4.4,0
18,22,16,2.9,2
18,25,14,3.6,1
38,20,13,3.6,1
56,56,7,5.0,1
54,94,41,4.4,0
31,48,14,4.1,1
57,31,16,4.7,1
```

Gambar 6. Hasil Clustering dalam Format CSV

Hasil clustering kemudian disimpan dalam file CSV agar dapat digunakan untuk proses analisis lanjutan. File tersebut hanya memuat atribut yang digunakan dalam penelitian beserta hasil cluster masing-masing pelanggan.

Untuk mengetahui karakteristik masing-masing kelompok pelanggan, dilakukan perhitungan nilai rata-rata pada setiap atribut berdasarkan cluster yang dihasilkan.

=== RATA-RATA TIAP CLUSTER ===

Cluster	Age	Purchase Amount (USD)	Previous Purchases	Review Rating
0	48.161194	67.516418	39.017910	4.062687
1	38.408497	55.741830	15.284314	4.254575
2	45.509749	55.649025	23.540390	3.004178

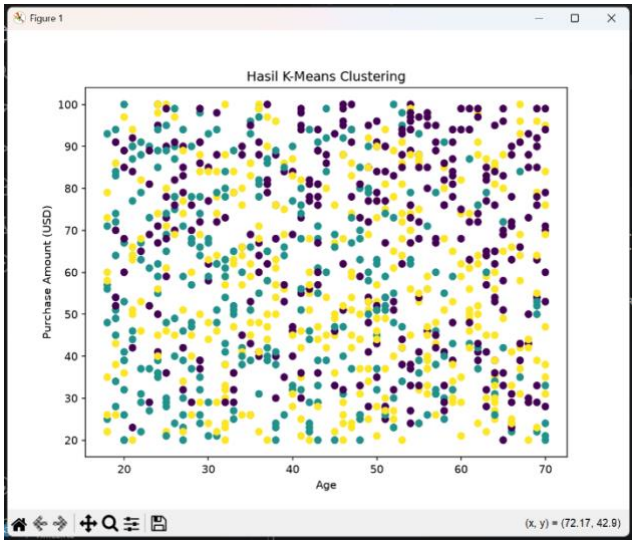
Gambar 7. Ringkasan Karakteristik Tiap Cluster

Setelah proses iterasi K-Means mencapai tahap konvergen, analisis deskriptif dilakukan dengan menghitung nilai rata-rata (*mean*) dari setiap atribut untuk memetakan karakteristik unik dari ketiga segmen pelanggan yang terbentuk.

Klaster 0 diidentifikasi sebagai segmen "Loyal High-Spenders (Mature)" yang didominasi konsumen paruh baya dengan rata-rata usia tertua (48,16 tahun). Segmen paling menguntungkan ini mencatatkan rata-rata nilai transaksi (*Purchase Amount*) tertinggi sebesar 67,51 USD dan akumulasi riwayat kunjungan (*Previous Purchases*) paling masif sebanyak 39,01 kali, yang berjalan selaras dengan tingkat loyalitas yang stabil melalui skor *Review Rating* sebesar 4,06.

Klaster 1 dikategorikan sebagai segmen "Satisfied New or Occasional Shoppers" yang merepresentasikan kelompok konsumen dengan rata-rata usia termuda (38,40 tahun). Kelompok ini memiliki nilai pengeluaran moderat sebesar 55,74 USD dengan rekam jejak transaksi terdahulu paling rendah, yaitu hanya 15,28 kali, namun menunjukkan potensi keterikatan emosional yang besar berkat perolehan skor kepuasan *Review Rating* tertinggi mencapai 4,25.

Sebagai pembanding, Klaster 2 mempresentasikan segmen "Dissatisfied Moderate Shoppers" dengan rata-rata usia pelanggan 45,50 tahun, nilai pengeluaran 55,64 USD, dan riwayat transaksi menengah sebanyak 23,54 kali. Fokus kritis pada kelompok ini terletak pada nilai *Review Rating* yang menyentuh angka terendah, yaitu 3,00, sehingga segmen ini dikategorikan memiliki risiko tinggi untuk bermigrasi ke kompetitor ritel lain.



Gambar 8. Visualisasi Hasil K-Means Clustering

menampilkan visualisasi sebaran spasial dari hasil segmentasi pelanggan menggunakan grafik *scatter plot* dua dimensi dengan memetakan variabel Age pada sumbu X dan Purchase Amount (USD) pada sumbu Y. Secara visual, representasi warna dari ketiga klaster pelanggan tampak saling membaur secara masif di sepanjang koordinat plot tanpa adanya garis pemisah (*clear decision boundary*) yang tegas. Fenomena tumpang tindih (*overlapping*) yang ekstrem ini secara objektif mengonfirmasi adanya keterbatasan teoretis dan metodologis dalam pemodelan yang dilakukan. Secara teknis komputasi, distorsi visual ini terjadi akibat adanya fonomena *projection loss* atau kehilangan informasi spasial akibat proses reduksi dimensi dari ruang multidimensi ke ruang dua dimensi. Algoritma K-Means beroperasi mengevaluasi jarak geometris pada ruang empat dimensi (4D) secara bersamaan, yaitu meliputi atribut Age, Purchase Amount, Previous Purchases, dan Review Rating. Ketika hasil partisi ruang empat dimensi tersebut diproyeksikan secara paksa ke dalam grafik dua dimensi (2D) yang hanya mengevaluasi dua variabel, kontribusi kedekatan jarak yang disumbangkan oleh atribut Previous Purchases dan Review Rating menjadi terkompresi sehingga tidak tervisualisasikan dengan akurat. Akibatnya, dua titik data pelanggan yang tampak bertumpukan rapat pada koordinat grafik Gambar 8 sebenarnya memiliki jarak geometris yang berjauhan pada ruang metrik riilnya, yang dipisahkan oleh perbedaan signifikan pada akumulasi transaksi terdahulu atau penilaian rating kepuasan mereka. [14].

Di luar faktor *projection loss*, sebaran visual yang sangat cair ini secara substantif menyingkap keterbatasan metode K-Means ketika dihadapkan pada karakteristik *Customer Shopping Trends Dataset* dunia nyata. Ditinjau dari tujuan penelitian untuk membentuk segmen pelanggan yang homogen, visualisasi ini membuktikan bahwa pencampuran data demografi kontinu dan transaksi finansial kurang efektif dalam membentuk polarisasi kelompok alami (*natural clusters*). Karakteristik data perilaku belanja pada umumnya memiliki distribusi yang cenderung seragam (*uniform distribution*) dan kontinu, sehingga pemodelan berbasis jarak seperti K-Means akan menghasilkan batas klaster artifisial yang rapuh, kabur, dan saling bersinggungan satu sama lain. Sifat pemaksaan batas artifisial inilah yang menyebabkan batas antar-segmen pelanggan pada grafik visual menjadi sangat cair, yang secara teoretis sejalan dengan beberapa penelitian terdahulu yang menunjukkan bahwa pencampuran variabel non-linear rentan menurunkan ketegasan visualisasi kelompok [15].

Secara kuantitatif, kekaburan visual pada Gambar 8 ini mengonfirmasi dan memvalidasi secara selaras mengapa nilai *Silhouette Coefficient* pada penelitian ini berada pada angka yang rendah, yaitu sebesar 0,187 yang menandakan kategori struktur lemah (*weak structure*). Visualisasi sebaran plot yang saling membaur ini memberikan bukti empiris bahwa nilai kohesi internal di dalam klaster tidak cukup padat dan nilai separasi antar-klaster terlalu rapat. Implikasi manajerial dari fenomena visual ini menegaskan bahwa pihak manajemen perusahaan ritel tidak dapat menerapkan strategi pemasaran yang kaku (*rigid target marketing*) berbasis pengelompokan klaster mutlak. Sebaliknya, perusahaan harus mengadopsi pendekatan kampanye promosi yang lebih dinamis dan fleksibel (*fluid marketing strategy*) untuk mengantisipasi perilaku konsumen, terutama bagi kelompok pelanggan yang secara spasial terdeteksi berada di area tumpang tindih (*overlapping*) antar-klaster [16].

VIII. KESIMPULAN

Penerapan algoritma K-Means dengan kerangka kerja *Knowledge Discovery in Databases* (KDD) telah berhasil mencapai tujuan penelitian dengan mempartisi 1.000 data pelanggan secara objektif ke dalam tiga segmen perilaku belanja strategis ($K = 3$) berdasarkan optimasi *Elbow Method*. Hasil penelitian berhasil memetakan karakteristik unik masing-masing kelompok, yaitu Klaster 0 sebagai segmen "*Loyal High-Spenders (Mature)*" dengan kontribusi finansial tertinggi, Klaster 1 sebagai segmen "*Satisfied New or Occasional Shoppers*" yang didominasi konsumen muda potensial, dan Klaster 2 sebagai segmen "*Dissatisfied Moderate Shoppers*" yang memiliki risiko tinggi melakukan migrasi (*customer churn risk*). Implikasi praktis dari temuan ini memberikan peta jalan bagi manajemen ritel untuk menerapkan strategi promosi dan retensi pelanggan secara personal serta dinamis (*fluid*

marketing strategy) guna memitigasi risiko kehilangan pasar pada segmen yang kritis.

Meskipun tujuan segmentasi tercapai, penelitian ini memiliki keterbatasan berupa struktur klaster yang lemah dengan nilai *Silhouette Score* sebesar 0,187 akibat penggunaan kombinasi variabel demografi-transaksi pada karakteristik data dunia nyata yang bersifat kontinu dan tersebar merata. Sebagai usulan pengembangan selanjutnya, disarankan untuk mengoptimalkan pemilihan atribut menggunakan pendekatan perilaku transaksional murni (*Recency, Frequency, and Monetary*/RFM), menerapkan metode *soft clustering* (*Fuzzy C-Means*/GMM) atau berbasis kerapatan (*DBSCAN*), serta mengintegrasikan teknik reduksi dimensi (*Principal Component Analysis*/PCA) sebelum proses klasterisasi dilakukan.

REFERENSI

[1] P. Mikael Kia Mado, "Implementasi Algoritma Clustering K-Means untuk Segmentasi Pelanggan di E-Commerce," 2025. [Online]. Available: https://journal.stmiki.ac.id

[2] R. B. Ardi, F. Ely Nastiti, and S. Sumarlinda, "ALGORITMA K-MEANS CLUSTERING UNTUK SEGMENTASI PELANGGAN (STUDI KASUS : FASHION VIRAL SOLO)," *INFOTECH journal*, vol. 9, no. 1, pp. 124–131, May 2023, doi: 10.31949/infotech.v9i1.5214.

[3] V. Wulandari, Y. Syarif, Z. Alfian, M. A. Althof, and M. Mufidah, "Comparison of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-Means and X-Means Algorithms on Shopping Trends Data," *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, vol. 1, no. 1, pp. 1–8, Jan. 2024, doi: 10.57152/ijatis.v1i1.1135.

[4] I. Iqbal, N. Hidayat, D. P. Gevano, and A. P. R. Ilahi, "Segmentasi Pelanggan Menggunakan K-Means Clustering Berdasarkan Data Kepribadian dan Pola Konsumsi," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 5, pp. 3914–3924, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5140.

[5] N. Oktaviany, N. Suarna, and W. Prihartono, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING DALAM MENGELOMPOKKAN KEPADATAN PENDUDUK DI PROVINSI DKI JAKARTA," 2024.

[6] N Tahta Phudjashakty, Hasbi Firmansyah, Wahyu Asriyani, and Ali Sofyan, "Segmentasi Pelanggan Grosir Menggunakan K-Means: Analisis Outlier dan Ketidakseimbangan Data," *Jurnal Pengabdian Masyarakat dan Riset Pendidikan*, vol. 4, no. 3, pp. 15993–16002, Jan. 2026, doi: 10.31004/jerkin.v4i3.4771.

[7] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *Jurnal SIMETRIS*, vol. 14, 2023.

[8] R. Mukhlis Juliyanto, F. Patria, H. Danggo Pamungkas, and F. Surya Nugraha, "SEMINAR NASIONAL AMIKOM SURAKARTA (SEMNASA) 2024 PENERAPAN ALGORITMA CLUSTERING UNTUK SEGMENTASI PELANGGAN BERDASARKAN PERILAKU PEMBELIAN".

[9] L. C. Fahsya, C. Wijaya, F. M. Bintang, J. J. Mulyono, F. Ramadhan, and F. Amsury, "PENERAPAN CLUSTERING K-MEANS UNTUK SEGMENTASI PELANGGAN PADA BISNIS RETAIL", doi: 10.52972/hoaq.vol17no1.

[10] S. L. Achmad, A. Fauzi, R. Rahmat, and J. Indra, "SEGMENTASI PELANGGAN MENGGUNAKAN K-MEANS CLUSTERING DI TOKO RETAIL," *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 7, no. 2, p. 736, Dec. 2024, doi: 10.37600/tekinkom.v7i2.1226.

[11] E. Tambunan, Y. B. Limbeng, and S. Sipayung, "Implementasi Algoritma K-Means Dengan Normalisasi Min-Max Pada Analisis Data Ketidakbersekolahan Anak," *JDMIS: Journal of Data Mining and Information Systems*, vol. 4, no. 1, pp. 26–32, 2026, doi: 10.54259/jdmis.v4i1.7064.

[12] N. A. Maori, "METODE ELBOW DALAM OPTIMASI JUMLAH CLUSTER PADA K-MEANS CLUSTERING," *Jurnal SIMETRIS*, vol. 14, 2023.

[13] Muhammad Raqib Syahkur, D. Hartama, and S. Solikhun, "Evaluasi Jumlah Cluster pada Algoritma K-Means++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita," *JST (Jurnal Sains dan Teknologi)*, vol. 13, no. 3, pp. 487–496, Oct. 2024, doi: 10.23887/jstundiksha.v13i3.86419.

[14] B. Hartono and V. Lusiana, "Analisis Metode Elbow SSE, Silhouette Score, dan Jaccard Stability dalam Pemilihan Jumlah Klaster Data yang Optimal," vol. 6, no. 8, pp. 1521–1532, 2026, doi: 10.47065/tin.v6i8.9271.

[15] W. Apriani1, Y. Perwira, R. Herissandi, and S. J. Purba, "Penerapan Metode K-Means Clustering Dalam Analisis Profil Konsumen Untuk Strategi Manajemen Pemasaran," *JURNAL SISTEM INFORMASI TGD*, vol. 5, no. 1, pp. 354–360, 2026, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi

[16] A. Purwanto, I. Ibrahim, H. Gunawan, P. Studi Teknik Informatika, and P. Studi Manajemen Informatika, "INFORMASI (Jurnal Informatika dan Sistem Informasi) Penerapan Algoritma K-Means Clustering pada Mall XYZ Menggunakan Teknik Data Mining."