The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 2, April 2025
https://doi.org/10.33022/ijcs.v14i2.4713

# Komparasi K-Means Clustering dengan *Euclidean* dan *Cosine Similarity* untuk Segmentasi dan Rekomendasi Produk pada Data *E-Commerce*

**Annisa Khoirala Ermy Pily[1], Susanti[2], Unang Rio[3], Tashid[4]**
annkhoiermy210@gmail.com[1], susanti@usti.ac.id[2], unangrio@usti.ac.id[3], tashid@sar.ac.id[4]
[1,2,3] Universitas Sains dan Teknologi Indonesia

**Informasi Artikel**

Diterima : 7 Feb 2025
Direview : 23 Feb 2025
Disetujui : 5 Mar 2025

**Kata Kunci**

*Cosine Similarity*, *Euclidean Distance*, *Hybrid Recommendation*, K-Means Clustering, Segmentasi Pelanggan.

**Abstrak**

Segmentasi pelanggan dan sistem rekomendasi berperan penting dalam meningkatkan pengalaman pengguna dan efektivitas analisis perilaku pelanggan. Penelitian ini membandingkan K-Means Clustering dengan Euclidean Distance dan Cosine Similarity untuk segmentasi pelanggan. Hasil evaluasi menunjukkan K-Means dengan Cosine Similarity lebih unggul, dengan Silhouette Score 0.6918, Calinski-Harabasz Score 3016.781, dan Davies-Bouldin Score 0.951, dibandingkan dengan 0.2363, 1257.178, dan 1.368 pada Euclidean. Clustering diterapkan pada sistem rekomendasi hybrid yang menggabungkan CBF, CF, dan CBPR. Hasil eksperimen menunjukkan bahwa hybrid recommendation dengan bobot CF 0.3, CBF 0.5, dan CBPR 0.2 menghasilkan Precision 0.8266, Recall 0.6198, NDCG 0.9964, dan Hit Ratio 0.9964. Penelitian ini menyimpulkan bahwa K-Means dengan Cosine Similarity lebih efektif dalam membentuk klaster pelanggan, sementara hybrid recommendation dengan bobot CF 0.3, CBF 0.5, dan CBPR 0.2 meningkatkan kualitas rekomendasi secara lebih relevan dan terstruktur.

***Keywords***

*Cosine Similarity, Customer Segmentation, Euclidean Distance, Hybrid Recommendation, K-Means Clustering.*

**Abstract**

*Customer segmentation and recommendation systems play a crucial role in enhancing user experience and the effectiveness of customer behavior analysis. This study compares K-Means Clustering with Euclidean Distance and Cosine Similarity for customer segmentation. The evaluation results show that K-Means with Cosine Similarity outperforms, with a Silhouette Score of 0.6918, Calinski-Harabasz Score of 3016.781, and Davies-Bouldin Score of 0.951, compared to 0.2363, 1257.178, and 1.368 for Euclidean. Clustering is applied to a hybrid recommendation system combining CBF, CF, and CBPR. The experimental results show that the hybrid recommendation with CF 0.3, CBF 0.5, and CBPR 0.2 yields Precision of 0.8266, Recall of 0.6198, NDCG of 0.9964, and Hit Ratio of 0.9964. This study concludes that K-Means with Cosine Similarity is more effective in forming customer clusters, while the hybrid recommendation with CF 0.3, CBF 0.5, and CBPR 0.2 enhances the relevance and structure of recommendations..*

Attribution-ShareAlike 4.0 International License

## A. Pendahuluan

Perkembangan teknologi informasi mendorong perubahan strategi bisnis di berbagai sektor digital, termasuk pemanfaatan data transaksi yang terus meningkat [1]. Agar strategi bisnis lebih efektif, analisis data menjadi kunci untuk memahami pola transaksi dan perilaku pelanggan. Melalui analisis ini, pelanggan dapat dikelompokkan berdasarkan karakteristik tertentu, memungkinkan penerapan strategi yang lebih personal dan relevan [2]. Dalam konteks ini, segmentasi pelanggan berperan penting dalam mengelompokkan pelanggan berdasarkan kesamaan, seperti preferensi belanja atau pola transaksi. Dengan segmentasi yang tepat, bisnis dapat meningkatkan kepuasan pengguna dan merancang strategi pemasaran yang lebih efektif [2].

Salah satu teknik yang umum digunakan untuk segmentasi pelanggan adalah *clustering*, yang bertujuan untuk menemukan pola tersembunyi dalam data dan mengelompokkan entitas dengan karakteristik serupa. K-Means merupakan salah satu algoritma clustering yang populer karena kemampuannya dalam mengelompokkan data dalam skala besar dengan efisien [3]. Namun, dalam penerapannya, pemilihan metrik jarak sangat memengaruhi hasil *clustering*. Oleh karena itu, penting untuk mengevaluasi efektivitas kedua metrik tersebut dalam membentuk klaster yang optimal [4].

Selain segmentasi pelanggan, sistem rekomendasi berperan penting dalam meningkatkan pengalaman pengguna dengan menyajikan produk yang lebih relevan dan sesuai dengan preferensi mereka. Dengan memanfaatkan teknik *hybrid recommendation* sistem rekomendasi dapat mengintegrasikan berbagai metode untuk menghasilkan rekomendasi yang lebih presisi dan sesuai kebutuhan pelanggan lain [4].

Segmentasi pelanggan memiliki keterkaitan dengan sistem rekomendasi, di mana hasil clustering dapat dimanfaatkan untuk meningkatkan relevansi dan akurasi rekomendasi. Dengan adanya klaster pelanggan yang lebih terstruktur, sistem rekomendasi dapat menyusun strategi rekomendasi yang lebih personal berdasarkan kelompok pelanggan dengan pola transaksi serupa[5]. Penelitian ini menggunakan pendekatan hybrid recommendation yang menggabungkan berbagai metode dapat lebih optimal jika didukung dengan segmentasi pelanggan yang akurat, karena rekomendasi yang diberikan tidak hanya mempertimbangkan interaksi individu tetapi juga pola kelompok pelanggan dengan karakteristik serupa.

Pada penelitian [4], penerapan K-Means Clustering untuk mengelompokkan bibit padi varietas unggul berdasarkan karakteristiknya, menggunakan *Euclidean Distance, Manhattan Distance, dan Cosine Similarity* sebagai metode pengukuran jarak. Evaluasi dengan Davies-Bouldin Index menunjukkan bahwa Euclidean Distance dan Cosine Similarity memperoleh nilai 0.307, lebih baik dibandingkan Manhattan Distance yang menghasilkan nilai 0.318, menyimpulkan bahwa keduanya lebih optimal untuk pengelompokkan bibit padi.

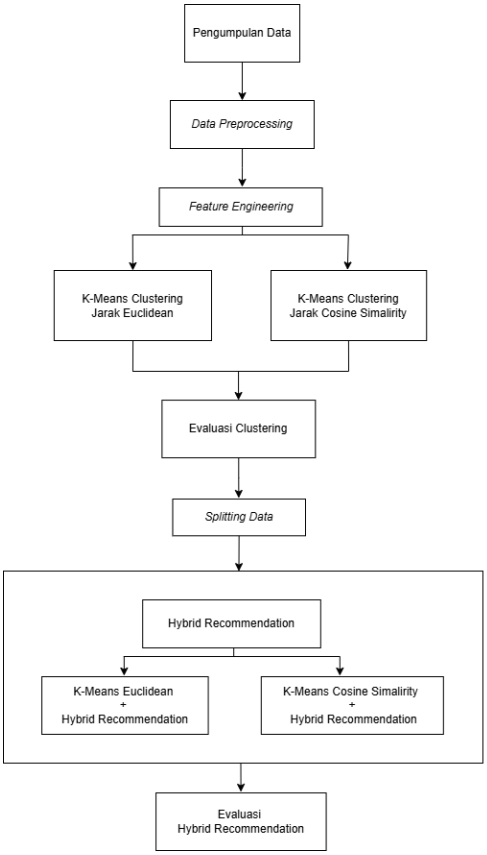
Lalu pada penelitian [5], mengembangkan sistem rekomendasi mata kuliah dengan K-Means Clustering untuk mengoptimalkan rekomendasi berdasarkan pola pemilihan mata kuliah mahasiswa. Metode *Collaborative Filtering* diterapkan menggunakan *Jaccard, Euclidean Distance, Cosine Similarity*, dan *Pearson Correlation*. Hasil evaluasi menunjukkan *Cosine Similarity* memiliki akurasi tertinggi

sebesar 64.76%, diikuti oleh *Euclidean Distance* sebesar 62.04%, *Jaccard* sebesar 53.66%, dan *Pearson Correlation* sebesar 43.71% [5].

Penelitian ini berfokus pada komparasi K-Means dengan *Euclidean Distance* dan *Cosine Similarity* dalam *clustering* data berbasis *e-commerce*, serta penerapannya dalam sistem rekomendasi produk. Evaluasi dilakukan untuk menentukan pendekatan yang lebih unggul dalam membentuk klaster yang optimal dan meningkatkan performa sistem rekomendasi. Hasil dari penelitian ini diharapkan dapat memberikan wawasan bagi bisnis yang ingin mengoptimalkan strategi rekomendasi produk berdasarkan pola transaksi pelanggan. Dengan metode clustering yang lebih akurat, bisnis dapat meningkatkan kepuasan pelanggan, memperkuat loyalitas, dan mendorong pertumbuhan melalui sistem rekomendasi yang lebih personal dan relevan.

## B. Metode Penelitian

Metode penelitian merupakan gambaran umum terkait tahapan proses dari penelitian yang dilakukan dari awal hingga akhir untuk mencapai tujuan penelitian. Tahapan penelitian disajikan sesuai pada Gambar 1.




**Gambar 1.** Metode Penelitian

### 1. Pengumpulan Data

Pengumpulan data merupakan tahap awal dalam penelitian ini. Data dapat diperoleh dari berbagai sumber, salah satunya melalui internet. Pada penelitian ini, peneliti menggunakan data yang diambil dari *website* Kaggle, dimana dataset ini bersifat data publik. Data yang dikumpulkan merupakan data transaksional yang

mencatat semua transaksi aktual selama kurang lebih satu tahun untuk ritel daring yang terdaftar dan berbasis di Inggris. Data tersebut mencakup berbagai jenis, yaitu numerik, kategorikal, temporal, dan teks. Peneliti mengunduh dataset ini dari Kaggle menggunakan kata kunci "*E-commerce Data*". Dataset tersebut berisi 541.909 baris data dengan 8 fitur, yaitu *Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, dan Country*.

## 2. *Data Preprocessing*

*Data preprocessing* merupakan tahap untuk mengolah data mentah menjadi data yang siap untuk diolah sehingga dapat masuk ketahap selanjutnya [6]. Tahap ini dapat membantu meningkatkan kualitas data yang digunakan dalam analisis dan memastikan hasil analisis yang lebih akurat [7].

Terdapat beberapa tahapan dalam data *preprocessing* yang digunakan pada penelitian ini, yaitu sebagai berikut.

a. *Data Cleaning*

*Data cleaning* bertujuan untuk membersihkan data dari *noise* atau data yang tidak relevan seperti *missing values* (terdapat nilai yang hilang), *outliers*, data duplikat, dan data yang tidak konsisten, kemudian mempersiapkan data agar dapat diolah dan dianalisis dengan baik. *Data cleaning* dilakukan untuk meningkatkan kualitas data, menghilangkan kesalahan atau ketidaksesuaian, dan membuat data lebih siap untuk digunakan dalam analisis atau pemodelan [8] .

b. Transformasi Data

Transformasi data adalah proses mengubah nilai-nilai dalam dataset untuk memperbaiki karakteristik atau distribusinya agar lebih sesuai dengan kebutuhan analisis. Tujuan utamanya adalah untuk meningkatkan kualitas data, seperti mengurangi *skewness*, meningkatkan kinerja model, dan memenuhi asumsi statistik yang diperlukan. Beberapa teknik transformasi yang umum digunakan antara lain normalisasi, standarisasi, log transformasi, dan *encoding* untuk data kategorikal. Proses ini membantu memastikan data siap untuk analisis lebih lanjut atau diterapkan dalam algoritma *machine learning* secara efektif [9].

c. *Normalization*

*Normalization* bertujuan untuk memastikan bahwa variabel-variabel yang memiliki skala atau rentang nilai yang berbeda-beda dapat diukur atau dibandingkan secara adil. Normalisasi membantu menghindari dominasi variabel dengan skala besar terhadap variabel dengan skala kecil, sehingga mencegah kesalahan dalam analisis dan *modelling*. Hal ini memungkinkan komputer untuk memproses data numerik dengan lebih efisien dan membuat proses analisis menjadi lebih akurat [9].

## 3. *Feature Engineering*

Data yang sudah melalui tahap *preprocessing* akan dilanjutkan dengan tahap *Feature Engineering* yang merupakan proses untuk menciptakan atau mengubah fitur dari data mentah agar dapat digunakan lebih optimal dalam analisis. Pada penelitian ini, metode RFM (*Recency*, *Frequency*, *Monetary*) digunakan untuk menganalisis nilai pelanggan dan melakukan segmentasi basis pelanggan, sehingga

menghasilkan fitur yang relevan untuk analisis segmentasi pelanggan. Tahapan dalam analisis RFM adalah sebagai berikut:

1. *Recency* (R)

Tahap ini menghitung waktu terakhir pelanggan melakukan transaksi dalam rentang waktu tertentu dengan cara menghitung selisih antara tanggal transaksi terakhir setiap pelanggan dengan tanggal akhir periode data. Nilai *recency* yang lebih rendah menunjukkan pelanggan yang lebih aktif dalam bertransaksi [10].

2. *Frequency* (F)

Tahap ini menghitung jumlah transaksi yang dilakukan oleh pelanggan dalam rentang waktu tertentu dengan menghitung frekuensi transaksi berdasarkan jumlah faktur unik yang tercatat. Semakin tinggi nilai *frequency*, semakin menunjukkan tingkat loyalitas pelanggan yang lebih baik [10].

3. *Monetary* (M)

Tahap ini menghitung total nilai uang yang dihasilkan oleh pelanggan dari transaksi mereka dalam rentang waktu tertentu dengan menjumlahkan total pembelian setiap pelanggan berdasarkan nilai transaksi. Nilai *monetary* yang lebih tinggi mencerminkan kontribusi finansial pelanggan yang lebih besar terhadap bisnis [11].

Setelah seluruh tahap perhitungan selesai, hasil dari setiap metrik R, F, dan M akan digunakan sebagai fitur utama dalam proses segmentasi pelanggan menggunakan algoritma K-Means clustering. Proses ini bertujuan untuk mengidentifikasi perilaku pelanggan berdasarkan nilai, frekuensi, dan keterlibatan mereka dalam transaksi bisnis, sehingga mendukung pengelompokan pelanggan yang lebih akurat [11].

## 4. K-Means Clustering

Setelah tahap *feature engineering*, proses selanjutnya adalah clustering data menggunakan algoritma K-Means. Dalam penelitian ini, nilai R, F, dan M yang dihasilkan dari *feature engineering* digunakan sebagai input untuk proses *clustering*. K-Means adalah algoritma unsupervised learning yang membagi dataset ke dalam K kelompok berdasarkan kemiripan data. Setiap data point akan dimasukkan ke dalam cluster dengan centroid terdekat menggunakan ukuran kesamaan tertentu. Algoritma ini bekerja secara iteratif hingga posisi centroid stabil dan tidak berubah lagi [3].

Dalam penelitian ini, K-Means diuji menggunakan dua pendekatan berbeda dalam mengukur jarak atau kesamaan antar data, yaitu:

### 4.1 K-Means dengan Euclidean Distance

Euclidean Distance adalah ukuran jarak antara dua titik dalam ruang multidimensi, yang dihitung berdasarkan akar kuadrat dari jumlah selisih kuadrat antara koordinat dua titik tersebut [4]. Secara matematis, Euclidean Distance dirumuskan sebagai berikut:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

dimana :

$d(x, y)$ : jarak antara data $x_i$ dan *centroid* $y_i$,

$x_i$ : data *i* dengan *n* fitur,
$y_i$ : koordinat *centroid k,*
$n$ : jumlah fitur.

Euclidean Distance mengukur jarak absolut antar titik data dalam ruang fitur, sehingga lebih cocok untuk data yang memiliki skala yang seragam [5].

### 4.2 K-Means dengan Cosine Similarity

Cosine Similarity mengukur kemiripan antara dua vektor berdasarkan sudut yang terbentuk di antara mereka dalam ruang multidimensi. Berbeda dengan Euclidean Distance yang menghitung jarak absolut, Cosine Similarity hanya mempertimbangkan arah vektor, bukan panjangnya [12]. Rumus Cosine Similarity:

$$\cos(\theta) = \frac{x.y}{||x||.||y||}$$

dimana:

$x.y$ : perkalian dot product dari dua vektor.
$||x||$ dan $||y||$ : panjang (magnitude) dari masing-masing vektor.

Cosine Similarity lebih efektif dalam menangani data dengan skala yang berbeda dan lebih cocok untuk data berbasis perilaku, seperti pola transaksi pelanggan dalam data e-commerce [12].

Tahapan dalam algoritma K-Means meliputi :

1. Inisialisasi Centroid
   Tentukan jumlah cluster KKK dan pilih KKK titik awal sebagai centroid awal.
2. Penentuan Cluster
   a. Hitung jarak antara setiap data point dengan semua centroid menggunakan Euclidean Distance dan Cosine Similarity.
   b. Setiap data point dimasukkan ke dalam cluster dengan centroid terdekat atau dengan kemiripan tertinggi.
3. Pembaruan Centroid
   Hitung ulang centroid setiap cluster dengan mengambil rata-rata dari semua anggota cluster.

$$R_k = \frac{1}{N_k}(\sum x_i)$$

   dimana :
   $R_k$ : *centroid* baru untuk *cluster k*,
   $N_k$ : jumlah data dalam *cluster k,*
   $x_i$ : titik data ke *i* dalam *cluster k*.
4. Iterasi Hingga Konvergensi
   Ulangi langkah 2 dan 3 hingga tidak ada perubahan signifikan pada posisi centroid atau jumlah iterasi maksimum tercapai [13].

Menentukan jumlah cluster yang optimal sangat penting dalam K-Means agar segmentasi pelanggan lebih akurat [13]. Dalam penelitian ini, pemilihan jumlah cluster dilakukan menggunakan *Elbow Method.*

Elbow Method bekerja dengan menghitung Within-Cluster Sum of Squares (WCSS) untuk berbagai nilai [14]. WCSS mengukur seberapa dekat data dalam satu cluster terhadap centroid-nya, dengan rumus:

$$WCSS = \sum_{i=1}^{K} \sum x \in C_i \ ||x - \mu_i||^2$$

dimana:
K adalah jumlah cluster,
x adalah data dalam cluster $C_i$,
$\mu_i$ adalah centroid dari cluster $C_i$.

Plot WCSS terhadap jumlah cluster akan menghasilkan pola yang menyerupai siku (elbow). Titik di mana kurva mulai melandai menunjukkan jumlah cluster optimal, karena setelah titik ini, penambahan jumlah cluster tidak secara signifikan mengurangi WCSS [15]. Berdasarkan hasil analisis, jumlah cluster optimal dalam penelitian ini adalah, yang digunakan dalam proses clustering dengan K-Means.

Hasil dari clustering ini akan membagi pelanggan ke dalam beberapa kelompok berdasarkan pola transaksi mereka. Cluster yang terbentuk kemudian digunakan dalam hybrid recommendation untuk meningkatkan personalisasi rekomendasi produk serta menyusun strategi pemasaran yang lebih efektif.

## 5. Evaluasi K-Means Clustering

Evaluasi model K-Means clustering dilakukan setelah proses clustering selesai untuk mengukur kualitas hasil yang dihasilkan oleh algoritma. Evaluasi ini bertujuan memastikan bahwa model dapat mengelompokkan data dengan baik sesuai dengan pola yang ada serta menilai tingkat kualitas cluster yang terbentuk [16]. Dalam penelitian ini, tiga metode evaluasi yang digunakan adalah *Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index.*

1. Silhouette Score

   Silhouette Score mengukur seberapa mirip suatu data dengan cluster tempatnya berada dibandingkan dengan cluster lain [16]. Rumus Silhouette score untuk suatu data individu dapat dituliskan sebagai berikut:

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

   dimana:
   $S(i)$ : nilai silhouette,
   $b(i)$ : rata-rata jarak dari objek i dengan seluruh objek dalam cluster yang sama,
   $a(i)$ : rata-rata jarak dari objek i dengan objek berada di cluster yang berbeda.
   Nilai Silhouette Score berkisar antara -1 hingga 1, di mana:

   a. Nilai mendekati 1 menunjukkan bahwa data berada dalam cluster yang tepat.
   b. Nilai mendekati 0 menunjukkan bahwa data berada di batas antara dua cluster.
   c. Nilai negatif menunjukkan bahwa data mungkin salah dimasukkan ke dalam cluster [3].

   Evaluasi menggunakan Silhouette Score memberikan gambaran yang jelas tentang efektivitas algoritma K-Means dalam membentuk cluster yang terpisah dan kohesif. Nilai rata-rata dari Silhouette Score untuk seluruh data digunakan sebagai indikator kualitas clustering [3].

2. Calinski-Harabasz Index

Calinski-Harabasz Index atau Variance Ratio Criterion mengukur rasio antara dispersi antar cluster dengan dispersi dalam cluster **eva**. Rumusnya:

$$CH = \frac{Tr\ (B_k)}{Tr\ (W_k)} \text{ x } \frac{N-K}{K-1}$$

dimana:

$Tr\ (B_k)$ : jumlah total varians antar cluster.
$Tr\ (W_k)$ : jumlah total varians dalam cluster.
N : jumlah data.
K : jumlah cluster.

Nilai Calinski-Harabasz yang lebih tinggi menunjukkan clustering yang lebih baik karena menandakan bahwa data dalam cluster lebih padat dan cluster satu dengan yang lain lebih terpisah [17].

3. *Davies-Bouldin Index*

Davies-Bouldin Index mengukur rata-rata kesamaan antara setiap cluster dengan cluster lain. Nilainya dihitung berdasarkan rasio antara jarak dalam cluster dan jarak antar cluster eva. Rumusnya:

$$DB = \frac{1}{K}\sum_{i=1}^{K} \underset{j\neq 1}{max} \left(\frac{s_i + s_j}{d_{i,j}}\right)$$

dimana:

$s_i$ : dispersi dalam cluster i.
$s_j$ : dispersi dalam cluster j.
$d_{i,j}$ : jarak antara centroid cluster iii dan centroid cluster j.

Nilai Davies-Bouldin yang lebih rendah menunjukkan clustering yang lebih baik, karena menandakan bahwa cluster lebih terpisah satu sama lain.

Hasil dari evaluasi ini digunakan untuk memastikan bahwa jumlah cluster yang dipilih optimal dan model clustering bekerja dengan baik dalam mengelompokkan data secara efekti [17].

## 6. *Splitting* Data

Setelah proses clustering menggunakan metode K-Means dan evaluasi hasilnya, tahap berikutnya adalah splitting data. Splitting data dilakukan untuk membagi dataset menjadi dua bagian, yaitu data latih dan data uji, yang digunakan dalam pengembangan dan evaluasi sistem rekomendasi. Langkah ini memastikan bahwa model rekomendasi dapat digeneralisasi dengan baik pada data yang belum pernah dilihat sebelumnya [18]. Pada penelitian ini, dilakukan 3 splitting data dengan proporsi 70:30, 80:20, dan 90:10. Splitting data dilakukan setelah proses dan evaluasi K-Means clustering, untuk memastikan bahwa data yang dibagi mencakup segmen pelanggan yang telah diidentifikasi sebelumnya.

## 7. Hybrid Recommendation

Setelah data dibagi menjadi data latih dan data uji, langkah selanjutnya adalah penerapan sistem rekomendasi. Dalam penelitian ini, sistem rekomendasi dirancang menggunakan pendekatan *hybrid recommendation* yang menggabungkan

metode *content-based filtering*, *collaborative filtering,* dan *Cluster-Based Popularity Ranking*. Pendekatan ini bertujuan untuk menghasilkan rekomendasi produk yang lebih relevan dan personal bagi pelanggan. Tahapan dalam pembangunan sistem rekomendasi adalah sebagai berikut:

### 7.1 Content-Based Filtering (CBF)

Pada tahap ini, sistem akan merekomendasikan produk berdasarkan kesamaan fitur antar produk. Fitur produk, seperti deskripsi, kategori, atau atribut lainnya, digunakan untuk membandingkan produk yang pernah dibeli atau memiliki interaksi positif dengan pelanggan. Sistem ini berfokus pada karakteristik produk yang relevan bagi pelanggan. Salah satu metode umum dalam Content-Based Filtering adalah Cosine Similarity, yang digunakan untuk menghitung kemiripan antara produk berdasarkan vektorisasi fitur. Sistem akan membuat matriks kesamaan antar produk berdasarkan Cosine Similarity. Produk yang memiliki nilai kesamaan tertinggi dengan produk yang pernah dibeli oleh pelanggan akan direkomendasikan [19]

### 7.2 Collaborative Filtering (CF)

Selanjutnya, collaborative filtering diterapkan dengan memanfaatkan pola interaksi antar pelanggan dan produk untuk memberikan rekomendasi. Sistem ini menganalisis data historis pembelian pelanggan, sehingga dapat menemukan hubungan yang relevan antar produk. Dengan demikian, produk yang sesuai dengan preferensi pelanggan dapat direkomendasikan [20]. Pendekatan yang umum digunakan dalam Collaborative Filtering adalah Pearson Correlation Similarity, yang mengukur kesamaan antar pelanggan berdasarkan pola rating atau transaksi mereka [20]. Rumus Pearson Correlation Similarity :

$$sim(A,B) = \frac{\sum i \in P \left(r_{A,i} - \bar{r}_A\right)\left(r_{B,i} - \bar{r}_B\right)}{\sqrt{\sum i \in P \left(r_{A,i} - \bar{r}_A\right)^2}\sqrt{\sum i \in P \left(r_{B,i} - \bar{r}_B\right)^2}}$$

dimana:

$sim(A,B)$ : tingkat kemiripan antara pelanggan A dan pelanggan B.
$r_{A,i}$ dan $r_{B,i}$ : rating atau interaksi pelanggan A dan B terhadap produk i.
$\bar{r}_A$ dan $\bar{r}_B$ : rata-rata rating pelanggan A dan B.
$P$ : himpunan produk yang telah diberi rating oleh kedua pelanggan.

### 7.3 Cluster-Based Popularity Ranking (CBPR)

Langkah berikutnya adalah menyusun rekomendasi berdasarkan popularitas dalam setiap cluster. Pendekatan ini bertujuan untuk memberikan rekomendasi yang lebih relevan dengan mempertimbangkan preferensi pelanggan dalam cluster yang sama [21].

Tahapan dalam Cluster-Based Popularity Ranking :

1. Identifikasi Produk Populer dalam Setiap Cluster

   Produk yang paling sering dibeli dalam suatu cluster ditandai sebagai produk populer dalam kelompok tersebut.

2. Perhitungan Skor Popularitas
   a. Skor popularitas suatu produk dalam cluster dihitung berdasarkan frekuensi pembelian dan jumlah pelanggan yang berinteraksi dengan produk tersebut.
   b. Rumus dasar yang dapat digunakan untuk menghitung skor popularitas dalam cluster:

$$Score(p,c) = \frac{Freq(p,c)}{Total_Users(c)}$$

   dimana:
   $Score(p,c)$ :Skor popularitas produk ppp dalam cluster ccc,
   $Freq(p,c)$ :Frekuensi produk p muncul dalam transaksi pelanggan dalam cluster c,
   $Total_Users(c)$ :Jumlah total pelanggan dalam cluster c.
3. Penyusunan Daftar Rekomendasi
   Produk dengan skor tertinggi dalam suatu cluster akan memiliki prioritas lebih tinggi dalam daftar rekomendasi pelanggan di cluster tersebut [21].

Pendekatan ini memanfaatkan keunggulan masing-masing metode untuk menghasilkan rekomendasi yang lebih akurat, relevan, dan bervariasi. Rekomendasi yang dihasilkan lebih sesuai dengan karakteristik pelanggan dalam kelompoknya dibandingkan dengan sekadar mengandalkan popularitas global di seluruh dataset [22]

## 7.4 Penggabungan dengan Weighted Hybrid

Setelah hasil dari CBF, CF, dan CPR diperoleh, ketiga metode tersebut digabungkan menggunakan pendekatan weighted hybrid. Pendekatan ini memungkinkan sistem rekomendasi untuk mengombinasikan keunggulan dari ketiga metode, menghasilkan rekomendasi yang lebih akurat dan relevan bagi pelanggan. Dalam metode ini, setiap sistem rekomendasi diberi bobot tertentu berdasarkan kontribusinya dalam menghasilkan rekomendasi [23]. Perhitungan skor akhir produk dilakukan dengan menggabungkan skor dari CBF, CF, dan CPR menggunakan persamaan berikut :

$$S_{final} = w_1 S_{CBF} + w_2 S_{CF} + w_3 S_{CBPR}$$

Di mana:
$S_{final}$ : skor akhir produk yang akan direkomendasikan.
$S_{CBF}$ : skor rekomendasi dari Content-Based Filtering.
$S_{CF}$ : skor rekomendasi dari Collaborative Filtering.
$S_{CBPR}$ : skor berdasarkan popularitas dalam cluster menggunakan Cluster-Based Popularity Ranking.
$w_1, w_2, w_3$ :bobot yang diberikan untuk masing-masing metode rekomendasi, dengan $w_1 + w_2 + w_3 = 1$

Pendekatan ini memastikan bahwa rekomendasi yang diberikan tidak hanya mempertimbangkan kesamaan fitur antar produk (CBF) dan pola interaksi pelanggan (CF), tetapi juga mempertimbangkan popularitas produk di dalam cluster pelanggan yang memiliki karakteristik serupa (CBPR). Dengan demikian, sistem rekomendasi dapat memberikan hasil yang lebih

relevan, personal, dan sesuai dengan preferensi pelanggan serta tren dalam kelompok pengguna tertentu [24].

Hybrid recommendation menggabungkan CBF dan CF untuk mengatasi cold-start problem serta meningkatkan akurasi rekomendasi yang lebih relevan dan personal. Metode ini cocok untuk dataset dengan kepadatan rendah, mengurangi masalah sparsity, dan mampu memberikan rekomendasi yang lebih beragam. Selain itu, pendekatan ini lebih personal dan kontekstual, menyesuaikan rekomendasi dengan preferensi pelanggan [24]. Dengan Cluster-Based Popularity Ranking (CBPR), rekomendasi juga dapat disesuaikan dengan tren dalam segmen pelanggan tertentu.Pendekatan Hybrid Recommendation memungkinkan sistem rekomendasi yang lebih komprehensif, akurat, dan fleksibel, sehingga dapat meningkatkan pengalaman pengguna serta efektivitas strategi pemasaran.

## 8. Evaluasi Sistem Rekomendasi

Evaluasi sistem rekomendasi dilakukan setelah proses pengujian selesai untuk mengukur sejauh mana model dapat memberikan rekomendasi yang relevan dan sesuai dengan kebutuhan pengguna. Tahapan ini bertujuan untuk memastikan bahwa sistem rekomendasi hybrid yang dikembangkan mampu memberikan rekomendasi yang akurat dan optimal [25].

Dalam penelitian ini, evaluasi sistem rekomendasi dilakukan menggunakan beberapa metrik utama:

1. Precision

   Mengukur proporsi item yang direkomendasikan dan relevan terhadap total item yang direkomendasikan. Semakin tinggi nilai precision, semakin baik sistem dalam memberikan rekomendasi yang tepat [26].

$$Precision = \frac{|\ Item\ relevan \cap Item\ direkomendasikan\ |}{|\ Item\ direkomendasikan\ |}$$

2. Recall

   Mengevaluasi kemampuan sistem dalam menemukan item relevan dari keseluruhan item yang tersedia. Metrik ini menunjukkan sejauh mana sistem dapat menangkap preferensi pengguna secara menyeluruh [26].

$$Precision = \frac{|\ Item\ relevan \cap Item\ direkomendasikan\ |}{|\ Total\ item\ relevan\ |}$$

3. Normalized Discounted Cumulative Gain (NDCG)

   Mengukur kualitas urutan rekomendasi dengan mempertimbangkan posisi item relevan dalam daftar rekomendasi. Item yang lebih relevan dan muncul lebih awal mendapatkan bobot lebih tinggi [27].

$$NDCG = \frac{DCGk}{IDCGk}, \qquad DCGk = \sum_{i=1}^{K} \frac{relevansi_i}{log_2(i+1)}$$

4. Hit Ratio

   Mengukur persentase pengguna yang menerima setidaknya satu item relevan dalam daftar rekomendasi. Metrik ini sangat berguna dalam mengevaluasi efektivitas sistem rekomendasi dalam skenario nyata [24].

$$HitRatio = \frac{Jumlah\ pengguna\ dengan\ minimal\ satu\ item\ relevan}{Total\ Pengguna}$$

Dengan mengukur performa sistem menggunakan metrik-metrik ini, dapat diperoleh gambaran yang lebih komprehensif mengenai efektivitas sistem rekomendasi. Evaluasi ini juga membantu dalam mengidentifikasi area yang perlu ditingkatkan agar rekomendasi yang diberikan semakin relevan dan sesuai dengan preferensi pengguna [25].

## C. Hasil dan Pembahasan

### 1. Pengumpulan Data

Pada tahap ini data yang digunakan diperoleh dari situs Kaggle yang berjumlah 541.909 baris data dengan 8 fitur. Dataset yang digunakan disajikan pada Tabel 1.

**Tabel 1.** Dataset

| No | *Invoice No* | *Stock Code* | *Description* | *Qty* | *Invoice Date* | *Unit Price* | *Customer ID* | *Country* |
|---|---|---|---|---|---|---|---|---|
| 1 | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 2.55 | 17850.0 | United Kingdom |
| 2 | 536365 | 71053 | WHITE METAL LANTERN | 6 | 12/1/2010 8:26 | 3.39 | 17850.0 | United Kingdom |
| 3 | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 12/1/2010 8:26 | 2.75 | 17850.0 | United Kingdom |
| 4 | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 12/1/2010 8:26 | 3.39 | 17850.0 | United Kingdom |
| 5 | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 12/1/2010 8:26 | 3.39 | 17850.0 | United Kingdom |
| … | … | … | … | … | … | … | … | … |
| 541 909 | 581587 | 22138 | BAKING SET 9 PIECE RETROSPOT | 3 | 12/9/2011 12:50 | 4.95 | 12680.0 | France |

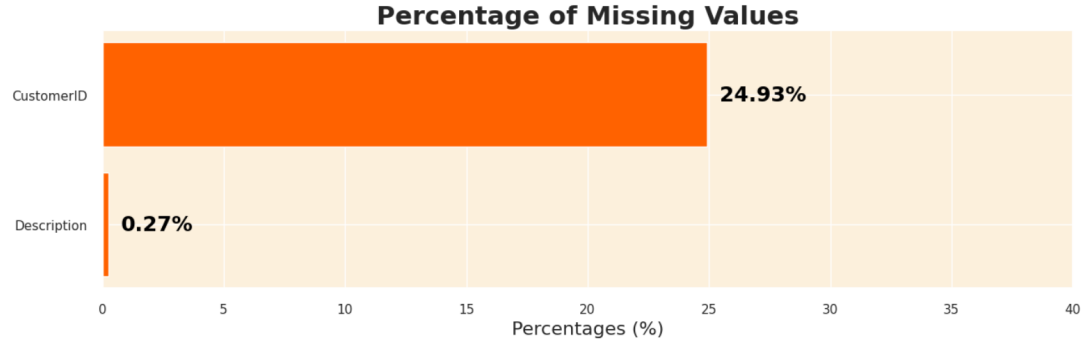
Pada Tabel 1 merupakan dataset *e-commerce* yang terdiri dari fitur, yaitu yaitu *Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, dan Country*. Berikut penjelasan masing-masing fitur :

a. *InvoiceNo* adalah nomor unik untuk setiap transaksi atau pesanan.
b. *StockCode* adalah kode unik yang mewakili setiap produk.
c. *Description* adalah deskripsi produk yang dibeli.
d. *Quantity* adalah jumlah unit produk yang dibeli dalam satu transaksi.
e. *InvoiceDate* adalah tanggal dan waktu transaksi dilakukan.
f. *UnitPrice* adalah harga satuan produk dalam transaksi.
g. *CustomerID* adalah id unik pelanggan yang melakukan transaksi.
h. *Country* adalah negara tempat pelanggan berasal atau transaksi dilakukan.

### 2. *Data Preprocessing*

Pada data preprocessing, langkah-langkah yang diambil meliputi cleaning, transformasi, dan normalisasi untuk memastikan data bersih, terstruktur, dan siap untuk analisis. Ditemukan missing values pada kolom *CustomerID* dan *Description*.

Sebagai penanganan, missing values tersebut dihapus. Gambar 2 menunjukkan distribusi dataset sebelum dilakukan penanganan terhadap missing values, memberikan gambaran bahwa data masih mengandung ketidakteraturan dan belum sepenuhnya siap untuk analisis lebih lanjut. Setelah penanganan, distribusi dataset menjadi lebih terkontrol dan siap digunakan untuk proses berikutnya.




**Gambar 2.** Missing Values

### 3. *Feature Engineering*

Setelah melalui tahap *preprocessing*, data kemudian dianalisis menggunakan metode RFM (*Recency, Frequency, Monetary*) untuk menganalisis dan melakukan segmentasi pelanggan. Hasil analisis menunjukkan bahwa pelanggan dengan nilai *recency* yang rendah cenderung lebih aktif bertransaksi, pelanggan dengan *frequency* tinggi menunjukkan loyalitas yang lebih baik, dan pelanggan dengan nilai *monetary* tinggi memberikan kontribusi finansial yang signifikan.

Setelah proses RFM, dihasilkan beberapa fitur baru, seperti *Days_Since_Last_Purchase*, *Total_Transactions*, *Total_Spend*, *Spending_Trend*, dan lain-lain, yang menggambarkan karakteristik transaksi pelanggan. Fitur-fitur baru ini memberikan gambaran yang lebih mendalam tentang perilaku pelanggan yang dapat digunakan untuk segmentasi lebih lanjut. Data tersebut kemudian kembali dinormalisasi untuk memastikan bahwa setiap fitur berada dalam skala yang seragam, yang memungkinkan analisis lebih akurat dalam pengelompokan pelanggan.

Setelah proses RFM, data juga disusun berdasarkan *CustomerID*, sehingga menghasilkan dataset akhir dengan 4.067 baris dan 16 fitur. Hasil normalisasi ini dapat dilihat pada Tabel 2 dan 3 berikut, di mana setiap nilai untuk variabel seperti *Days_Since_Last_Purchase*, *Total_Transactions*, dan *Total_Spend* telah disesuaikan dengan skala yang lebih rendah dan seragam.

**Tabel 2.** Hasil RFM dan Normalisasi (Bagian 1)

| No | Customer ID | Days_Since_ Last_Purchase | Total_ Transactions | Total_Products_ Purchased | Total_ Spend | Average_ Transaction_ Value |
|---|---|---|---|---|---|---|
| 1 | 12346.0 | 2.345802 | -0.477589 | -0.754491 | -0.813464 | -1.317106 |
| 2 | 12347.0 | -0.905575 | 0.707930 | 2.005048 | 2.366920 | 1.528132 |
| 3 | 12348.0 | -0.170744 | -0.003381 | 1.863591 | 0.247087 | 0.343279 |

**Tabel 3.** Hasil RFM dan Normalisasi (Bagian 2)

| No | Customer ID | Unique_ Products_ Purchased | Average_ Days_ Between_ Purchases | Day_ Of_ Week | Hour | Is_UK |
|---|---|---|---|---|---|---|
| 1 | 12346.0 | -0.908471 | -0.310564 | 1 | -1.086929 | 1 |
| 2 | 12347.0 | 0.815119 | -0.128438 | 1 | 0.647126 | 0 |
| 3 | 12348.0 | -0.570512 | 0.672476 | 3 | 2.814696 | 0 |

**Tabel 4.** Hasil RFM dan Normalisasi (Bagian 3)

| No | Customer ID | Cancellation_ Frequency | Cancellation_ Rate | Monthly Spending_ Mean | Spending_Trend |
|---|---|---|---|---|---|
| 1 | 12346.0 | 0.420541 | 0.417623 | -1.329018 | -0.713318 |
| 2 | 12347.0 | -0.545753 | -0.432111 | 0.989511 | 1.259961 |
| 3 | 12348.0 | -0.545753 | -0.432111 | 0.023997 | 0.466213 |

## 4. K-Means Clustering

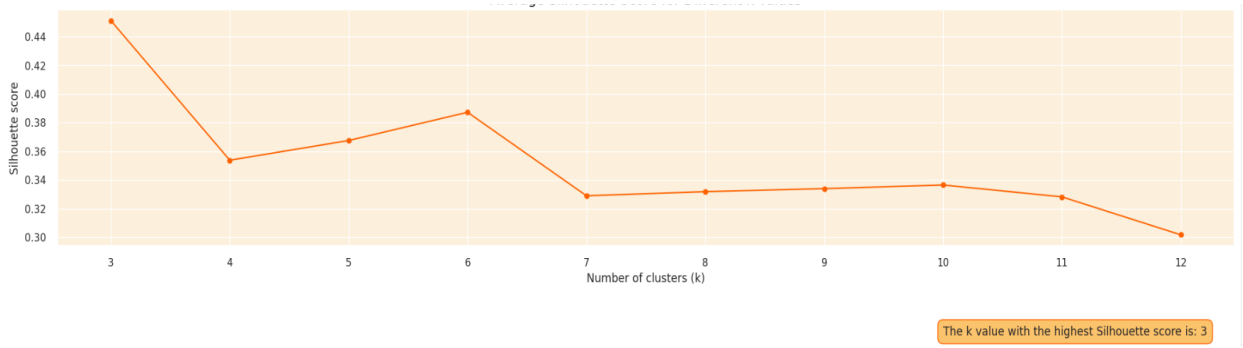
Setelah tahap *feature engineering* menggunakan metode RFM, segmentasi pelanggan dilakukan dengan menerapkan algoritma *K-Means Clustering*. Untuk menentukan jumlah klaster yang optimal, dilakukan analisis menggunakan *Elbow Method* guna mengidentifikasi titik optimal dalam pengelompokan data.

### 4.1 Hasil *Elbow Method*

Hasil dari *Elbow Method* menunjukkan bahwa nilai optimal terdapat pada k = 3, di mana terjadi perubahan signifikan dalam nilai inertia sebelum grafik mulai melandai. Untuk memvalidasi kualitas klasterisasi, dilakukan pengukuran *Silhouette Score* menggunakan dua metode pengukuran jarak, yaitu:

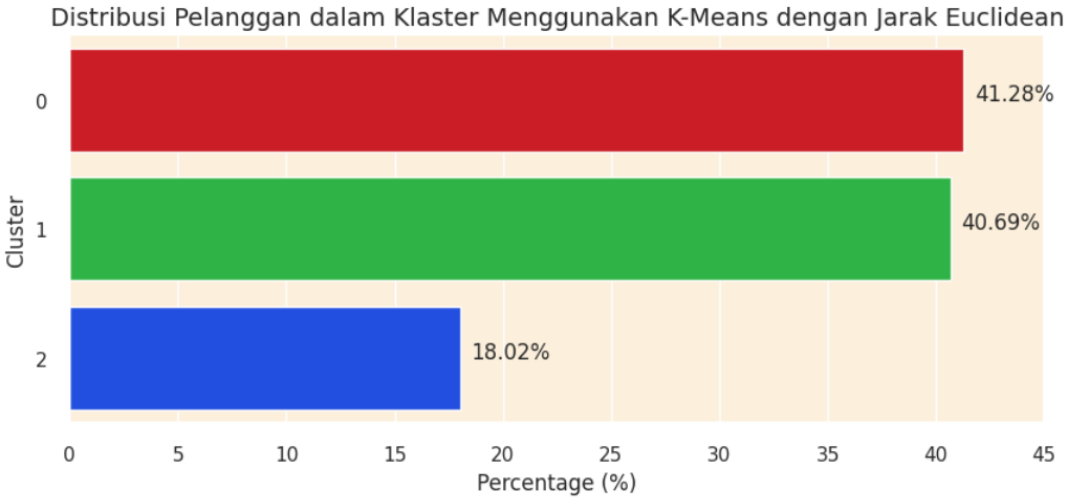
a. Cosine Similarity, yang menghasilkan Silhouette Score tertinggi sebesar 0.45 pada k=3.
b. Euclidean Distance, yang menghasilkan Silhouette Score sebesar 0.24 pada k =3.

Berdasarkan hasil tersebut, proses K-Means Clustering dilanjutkan dengan k = 3 untuk membentuk tiga segmen pelanggan berdasarkan fitur *Recency, Frequency, dan Monetary*. Visualisasi hasil klasterisasi dapat dilihat pada Gambar 3.




**Gambar 3.** Visualisasi *Missing Values*

## 4.2 K-Means Clustering dengan Jarak *Euclidean*



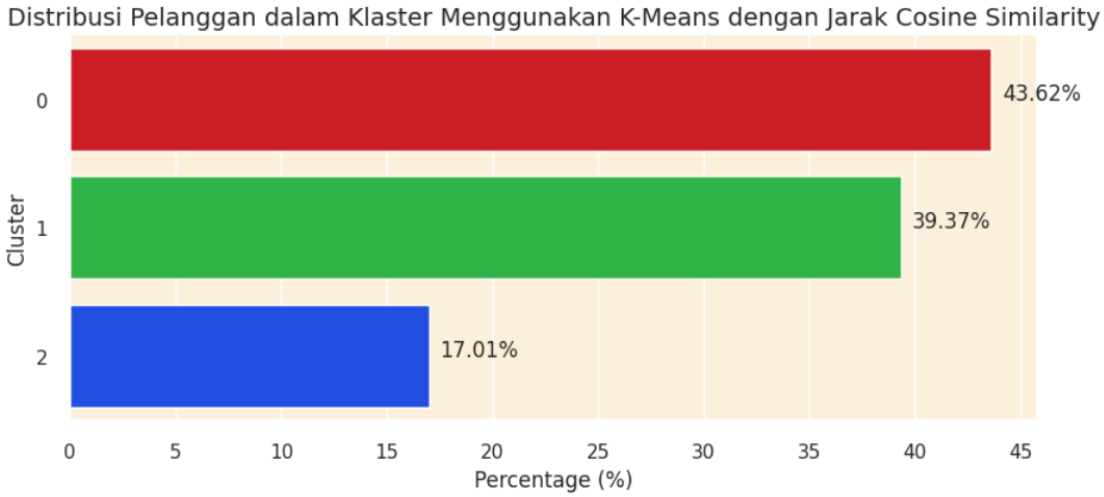

**Gambar 4.** Distribusi Klaster K-Means *Euclidean*

Hasil pengelompokan pelanggan menggunakan metode K-Means dengan metrik jarak Euclidean, sebagaimana ditampilkan pada Gambar 4, menunjukkan bahwa pelanggan terbagi ke dalam tiga klaster dengan proporsi sebagai berikut:

a. Klaster 0 mencakup 41,28% dari total pelanggan.
b. Klaster 1 mencakup 40,69% dari total pelanggan.
c. Klaster 2 mencakup 18,2% dari total pelanggan.

Distribusi ini menunjukkan bahwa mayoritas pelanggan berada di Klaster 0 dan Klaster 1, sementara Klaster 2 memiliki proporsi yang lebih kecil. Proporsi yang lebih besar pada Klaster 0 dan 1 mengindikasikan bahwa sebagian besar pelanggan memiliki karakteristik yang lebih mirip satu sama lain dalam hal *Recency*, *Frequency*, dan *Monetary*. Sementara itu, Klaster 2 yang lebih kecil dapat merepresentasikan kelompok pelanggan dengan pola pembelian yang berbeda, seperti pelanggan dengan transaksi lebih jarang atau lebih spesifik.

### 4.3 K-Means Clustering dengan Jarak *Cosine Similarity*




**Gambar 5.** Distribusi Klaster K-Means Cosine

Sementara itu, pengelompokan menggunakan K-Means dengan metrik jarak *Cosine*, sebagaimana ditampilkan pada Gambar 5, dibandingkan dengan metode *Euclidean*, yaitu:

a. Klaster 0 mencakup 43,62% dari total pelanggan.
b. Klaster 1 mencakup 39,37% dari total pelanggan.
c. Klaster 2 mencakup 17,01% dari total pelanggan.

Meskipun terdapat perbedaan distribusi antar klaster dibandingkan dengan metode *Euclidean*, pola segmentasi pelanggan secara keseluruhan tetap menunjukkan bahwa klaster 0 dan 1 memiliki proporsi yang dominan dibandingkan klaster 2. Meskipun terdapat perbedaan dalam proporsi antar klaster dibandingkan dengan metode *Euclidean*, pola segmentasi pelanggan secara keseluruhan tetap menunjukkan bahwa Klaster 0 dan Klaster 1 memiliki proporsi yang dominan dibandingkan Klaster 2.

Perbedaan ini menunjukkan bahwa metrik jarak yang digunakan dalam K-Means dapat mempengaruhi hasil segmentasi pelanggan. Dengan *Cosine Similarity*, pengelompokan lebih mempertimbangkan kemiripan pola belanja relatif terhadap keseluruhan data, dibandingkan dengan Euclidean Distance yang lebih memperhatikan jarak absolut antar pelanggan dalam ruang fitur.

### 4.4 Karakteristik Klaster Pelanggan

Berdasarkan hasil metode *Elbow Method*, diperoleh jumlah klaster optimal sebanyak tiga klaster. Setiap klaster menunjukkan karakteristik pelanggan yang berbeda berdasarkan pola belanja, nilai transaksi, dan tingkat pembatalan pesanan. Berikut adalah deskripsi masing-masing klaster :

1. Klaster 0: Pembeli Sesekali dengan Preferensi Belanja Akhir Pekan
   a. Pelanggan dalam klaster ini memiliki aktivitas belanja yang jarang dengan total transaksi dan pengeluaran yang rendah.
   b. Mereka cenderung lebih sering berbelanja di akhir pekan.

c. Tren pengeluaran stabil tetapi tetap rendah, dengan variasi pengeluaran bulanan yang kecil.
d. Jarang melakukan pembatalan pesanan dan nilai transaksi per pembelian tergolong kecil.

2. Klaster 1: Pembelanja Jarang tetapi Mengeluarkan Uang dalam Jumlah Besar
   a. Pelanggan dalam klaster ini memiliki frekuensi belanja yang rendah, namun jumlah uang yang mereka keluarkan per transaksi tergolong tinggi.
   b. Tren pengeluaran mereka meningkat seiring waktu.
   c. Mayoritas pelanggan dalam klaster ini berasal dari Inggris dan lebih sering berbelanja pada malam hari.
   d. Tingkat pembatalan pesanan berada di tingkat sedang.
3. Klaster 2: Pembeli Aktif dengan Pengeluaran Besar dan Tingkat Pembatalan Tinggi
   a. Pelanggan dalam klaster ini sering melakukan transaksi dan memiliki total pengeluaran yang sangat tinggi.
   b. Mereka membeli banyak produk unik tetapi juga memiliki tingkat pembatalan pesanan yang tinggi.
   c. Lebih sering berbelanja di pagi hari dan menunjukkan pola pengeluaran yang tidak konsisten.
   d. Meskipun pengeluaran mereka besar, tren belanja mereka menunjukkan penurunan dalam beberapa waktu terakhir.

Pemetaan karakteristik pelanggan berdasarkan klaster ini dapat membantu dalam pengambilan keputusan strategis, terutama dalam menyusun strategi pemasaran yang lebih efektif sesuai dengan pola belanja pelanggan di masing-masing segmen.

## 5. Evaluasi K-Means Clustering

Setelah clustering selesai, evaluasi hasil clustering dilakukan menggunakan tiga metrik utama, yaitu *Silhouette Score, Calinski-Harabasz Score*, dan *Davies-Bouldin Score*. Tabel 5 menyajikan hasil evaluasi untuk dua metode jarak yang digunakan, yaitu *Euclidean* dan *Cosine Similarity* dalam algoritma K-Means.

**Tabel 5.** Hasil Perbandingan *K-Means Euclidean* dan *Cosine Similarity*

| Algoritma | *Silhouette Score* | *Calinski-Harabasz Score* | *Davies-Bouldin Score* |
|---|---|---|---|
| K-Means *Euclidean* | 0.2364 | 1257.1783 | 1.3682 |
| K-Means *Cosine Similarity* | 0.6918 | 3016.6532 | 0.9508 |

Berdasarkan *Silhouette Score*, K-Means dengan *Cosine Similarity* menunjukkan nilai 0.6918, yang jauh lebih tinggi dibandingkan dengan K-Means *Euclidean* yang hanya 0.2364. Nilai yang lebih tinggi menunjukkan bahwa klaster yang terbentuk memiliki pemisahan yang lebih baik dan lebih terstruktur.

Dari *Calinski-Harabasz Score*, K-Means dengan *Cosine Similarity* juga memiliki skor yang lebih tinggi sebesar 3016.6532 dibandingkan dengan K-Means *Euclidean* sebesar 1257.1783. Skor ini mengindikasikan bahwa metode *Cosine Similarity* lebih optimal dalam menghasilkan klaster yang terpisah secara signifikan dengan variansi intra-klaster yang lebih kecil.

Sementara itu, *Davies-Bouldin Score* menunjukkan bahwa K-Means dengan *Cosine Similarity* memiliki nilai 0.9508, yang lebih rendah dibandingkan dengan K-Means *Euclidean* sebesar 1.3682. Nilai yang lebih rendah pada metrik ini menunjukkan bahwa klaster yang terbentuk memiliki tingkat pemisahan yang lebih baik dan tidak terlalu tumpang tindih.

Secara keseluruhan, berdasarkan ketiga metrik evaluasi ini, metode K-Means dengan *Cosine Similarity* memberikan hasil clustering yang lebih baik dibandingkan dengan K-Means *Euclidean*, karena menghasilkan klaster yang lebih kompak, terpisah dengan baik, dan memiliki struktur yang lebih optimal.

### 6. *Splitting* Data

Penelitian ini melakukan 3 kali *splitting* data dengan proporsi 70:30, 80:20, dan 90:10. Pada *splitting* data 70:30 berarti 70% dari total data digunakan untuk data latih, dan 30% digunakan untuk data uji. Lalu, pada *splitting* data 80:20 berarti 80% dari total data digunakan untuk data latih, dan 20% digunakan untuk data uji. Dan, pada *splitting* data 90:10 berarti 90% dari total data digunakan untuk data latih, dan 10% digunakan untuk data uji. *Splitting* data dilakukan setelah proses dan evaluasi K-Means clustering, untuk memastikan bahwa data yang dibagi telah mencakup segmen pelanggan yang telah teridentifikasi. Hal ini memungkinkan model rekomendasi mempertimbangkan pola perilaku pelanggan dari setiap klaster saat memberikan rekomendasi.

### *7. Hybrid Recommendation*

Setelah membagi data menjadi data latih dan data uji, sistem rekomendasi diterapkan menggunakan pendekatan *Hybrid Recommendation*, yang mengombinasikan *Content-Based Filtering* (CBF*), Collaborative Filtering* (CF), dan *Cluster-Based Popularity Ranking* (CBPR). Hasil rekomendasi dari masing-masing metode dihitung dan digabungkan menggunakan pendekatan *Weighted Hybrid* untuk menghasilkan rekomendasi yang lebih akurat dan relevan.

Pada penelitian ini, dilakukan perbandingan antara dua pendekatan *clustering*, yaitu *Hybrid Recommendation* dengan K-Means *Euclidean* dan *Hybrid Recommendation* dengan K-Means *Cosine Similarity*. Setiap metode diuji dengan tiga kombinasi bobot yang berbeda dalam pendekatan *Weighted Hybrid*, yaitu:

a. CF = 0.4, CBF = 0.4, CBPR = 0.2
b. CF = 0.3, CBF = 0.5, CBPR = 0.2
c. CF = 0.5, CBF = 0.3, CBPR = 0.2

Selanjutnya, diambil 3 rekomendasi teratas sesuai dengan bobot yang telah ditentukan. Hasil akhir dari setiap skenario kemudian dibandingkan untuk

menentukan kombinasi terbaik yang memberikan rekomendasi paling optimal. Contoh hasil rekomendasi yang dihasilkan dapat dilihat pada Tabel 6 dan Tabel 7.

**Tabel 6**. *Hybrid Recommendation* dan *K-Means Euclidean*

| Customer ID | Rec1 Stock Code | Rec2 Stock Code | Rec3 StockCode | Rec1 Description | Rec2 Description | Rec3 Description |
|---|---|---|---|---|---|---|
| 12619.0 | 22952 | 22078 | 22776 | 72 CAKE CASES VINTAGE CHRISTMAS | RIBBON REEL LACE DESIGN | SWEETHEART 3 TIER CAKE STAND |
| 14023.0 | 22537 | 22535 | 23078 | MAGIC DRAWING SLATE DINOSAUR | MAGIC DRAWING SLATE BUNNIES | ICE CREAM PEN LIP GLOSS |
| 17046.0 | 21121 | 22952 | 22417 | SET/10 RED POLKADOT PARTY CANDLES | 72 CAKE CASES VINTAGE CHRISTMAS | PACK OF 60 SPACEBOY CAKE CASES |

**Tabel 7.** *Hybrid Recommendation* dan K-Means *Cosine Similarity*

| Customer ID | Rec1 Stock Code | Rec2 Stock Code | Rec3 StockCode | Rec1 Description | Rec2 Description | Rec3 Description |
|---|---|---|---|---|---|---|
| **12619.0** | 22952 | 22078 | 22776 | 72 CAKE CASES VINTAGE CHRISTMAS | RIBBON REEL LACE DESIGN | SWEETHEART 3 TIER CAKE STAND |
| **14023.0** | 22537 | 22535 | 23078 | MAGIC DRAWING SLATE DINOSAUR | MAGIC DRAWING SLATE BUNNIES | ICE CREAM PEN LIP GLOSS |
| **17046.0** | 21121 | 22952 | 22417 | SET/10 RED POLKADOT PARTY CANDLES | 72 CAKE CASES VINTAGE CHRISTMAS | PACK OF 60 SPACEBOY CAKE CASES |

Kedua tabel di atas menampilkan hasil rekomendasi dari *Hybrid Recommendation* dengan K-Means *Euclidean* dan *Cosine Similarity* menggunakan bobot CF 0.4, CBF 0.4, dan CBPR 0.2. Hasilnya menunjukkan kemiripan dalam pemilihan item, dengan sedikit perbedaan dalam urutan rekomendasi. Namun, karena hanya tiga rekomendasi teratas yang ditampilkan, perbedaan antara kedua metode mungkin tidak terlihat jelas. Jika lebih banyak rekomendasi ditampilkan, variasi hasil bisa lebih terlihat.

Secara keseluruhan, rekomendasi yang dihasilkan tetap relevan dengan preferensi pelanggan, dan metode hybrid yang digunakan mampu memberikan rekomendasi yang seimbang berdasarkan kesamaan perilaku pelanggan serta karakteristik produk.

## 8. Evaluasi Hybrid Recommendation

Selanjutnya, setelah model hybrid dibangun, dilakukan evaluasi untuk mengukur kinerjanya. Evaluasi dilakukan menggunakan metrik precision, recall, nDCG, dan hit ratio. Hasil perbandingan evaluasi antara masing-masing metode, baik dari segi splitting maupun bobot, dapat dilihat pada Tabel 7.

| Hybrid Filtering | Splitting | CF | CBF | CBPR | Precision | Recall | NDCG | Hit Ratio |
|---|---|---|---|---|---|---|---|---|
| K-Means Euclidean | 70:30 | 0.4 | 0.4 | 0.2 | 0.9225 | 0.366 | 0.9937 | 0.9937 |
| | | 0.5 | 0.3 | 0.2 | 0.9244 | 0.3651 | 0.9952 | 0.9952 |
| | | 0.3 | 0.5 | 0.2 | 0.9202 | 0.3659 | 0.9952 | 0.9952 |
| | 80:20 | 0.4 | 0.4 | 0.2 | 0.8967 | 0.4556 | 0.9948 | 0.9948 |
| | | 0.5 | 0.3 | 0.2 | 0.8925 | 0.4571 | 0.9956 | 0.9956 |
| | | 0.3 | 0.5 | 0.2 | 0.8959 | 0.4559 | 0.9964 | 0.9964 |
| | 90:10 | 0.4 | 0.4 | 0.2 | 0.8253 | 0.6176 | 0.9975 | 0.9974 |
| | | 0.5 | 0.3 | 0.2 | 0.8332 | 0.6184 | 0.9969 | 0.9969 |
| | | 0.3 | 0.5 | 0.2 | 0.8379 | 0.6151 | 0.9969 | 0.9969 |
| K-Means Cosine Similarity | 70:30 | 0.4 | 0.4 | 0.2 | 0.9229 | 0.3661 | 0.995 | 0.995 |
| | | 0.5 | 0.3 | 0.2 | 0.9201 | 0.3668 | 0.9945 | 0.9945 |
| | | 0.3 | 0.5 | 0.2 | 0.9170 | 0.3691 | 0.9947 | 0.9947 |
| | 80:20 | 0.4 | 0.4 | 0.2 | 0.8952 | 0.4594 | 0.9967 | 0.9967 |
| | | 0.5 | 0.3 | 0.2 | 0.8944 | 0.4577 | 0.9954 | 0.9954 |
| | | 0.3 | 0.5 | 0.2 | 0.8944 | 0.4575 | 0.9938 | 0.9938 |
| | 90:10 | 0.4 | 0.4 | 0.2 | 0.8298 | 0.6151 | 0.9956 | 0.9956 |
| | | 0.5 | 0.3 | 0.2 | 0.8341 | 0.6129 | 0.9956 | 0.9955 |
| | | 0.3 | 0.5 | 0.2 | 0.8266 | 0.6198 | 0.9964 | 0.9964 |

Berdasarkan hasil evaluasi hybrid recommendation dengan K-Means menggunakan *Euclidean* dan *Cosine Similarity*, terlihat bahwa performa sistem bervariasi bergantung pada metode *splitting* data serta kombinasi bobot *Collaborative Filtering* (CF), *Content-Based Filtering* (CBF), dan *Collaborative-Based Personalized Ranking* (CBPR). Analisis lebih lanjut dilakukan untuk memahami pengaruh masing-masing faktor terhadap kualitas rekomendasi :

1. Pengaruh *Splitting* Data
   a. Pada *splitting* 70:30 *precision* relatif tinggi, berada di atas 0.92 untuk kedua metode, namun *recall* masih rendah sekitar 0.36. Hal ini menunjukkan bahwa model mampu memberikan rekomendasi yang sangat akurat, tetapi belum cukup luas dalam menangkap item-item relevan.
   b. Pada *splitting* 80:20 *precision* sedikit menurun dibandingkan dengan 70:30, tetapi terjadi peningkatan *recall* menjadi sekitar 0.45. Ini menunjukkan bahwa model lebih mampu mencakup rekomendasi yang lebih luas, meskipun akurasinya sedikit berkurang.
   c. Pada splitting 90:10 *precision* mengalami penurunan lebih lanjut ke kisaran 0.82–0.83, tetapi *recall* meningkat secara signifikan hingga mencapai 0.61–0.62. Hasil ini mengindikasikan bahwa dengan jumlah data latih yang lebih sedikit, model lebih fokus dalam mencakup lebih banyak item yang relevan, meskipun dengan konsekuensi penurunan akurasi rekomendasi.

2. Perbandingan K-Means *Euclidean* dan *Cosine Similarity*
   a. K-Means *Euclidean* cenderung menghasilkan *precision* yang sedikit lebih tinggi dalam beberapa skenario, tetapi *recall* lebih rendah dibandingkan dengan *Cosine Similarity.*
   b. K-Means *Cosine Similarity* menunjukkan stabilitas yang lebih baik dalam *recall*, terutama pada split 80:20 dan 90:10, yang menunjukkan bahwa metode ini lebih efektif dalam mencakup lebih banyak rekomendasi tanpa mengorbankan kualitas klaster.
3. Evaluasi Metrik NDCG dan *Hit Ratio*
   a. NDCG dan *hit ratio* secara konsisten mendekati nilai 1, yang mengindikasikan bahwa sistem mampu memberikan rekomendasi yang sangat relevan serta mempertahankan urutan peringkat yang baik.
   b. Meskipun terdapat perbedaan kecil dalam *precision* dan *recall*, nilai NDCG dan *hit ratio* yang tinggi menunjukkan bahwa kualitas rekomendasi tetap optimal dalam menyesuaikan preferensi pelanggan.
4. Kombinasi *Hybrid Recommendation* Terbaik
   a. Pada K-Means *Euclidean*, kombinasi *Hybrid Recommendation* dengan bobot CF 0.4, CBF 0.4, dan CBPR 0.2 pada *splitting* 90:10 menunjukkan performa terbaik berdasarkan metrik NDCG dan *hit ratio*, menandakan model yang lebih stabil dan relevan.
   b. Pada K-Means *Cosine Similarity*, kombinasi *Hybrid Recommendation* dengan bobot CF 0.3, CBF 0.5, dan CBPR 0.2 pada *splitting* 90:10 memberikan keseimbangan optimal antara *precision* dan *recall*, memungkinkan cakupan rekomendasi yang lebih luas tanpa mengorbankan akurasi.

Secara keseluruhan, kombinasi *Hybrid Recommendation* dengan bobot CF 0.3, CBF 0.5, dan CBPR 0.2 menggunakan K-Means *Cosine Similarity* pada splitting 90:10 merupakan konfigurasi terbaik. Model ini menunjukkan keseimbangan optimal di semua metrik evaluasi, dengan *Precision* sebesar 0.8266, *Recall* 0.6198, serta NDCG dan *Hit Ratio* yang sama-sama mencapai 0.9964. Selain itu, model ini juga menghasilkan *Silhouette Score* sebesar 0.6918, yang menandakan kualitas klaster yang lebih baik. Dengan demikian, sistem rekomendasi yang dihasilkan lebih menyeluruh, relevan, dan terstruktur dalam memberikan rekomendasi produk kepada pelanggan.

## D. Simpulan

Berdasarkan hasil penelitian, berikut adalah kesimpulan yang dapat diambil:

1. K-Means dengan *Cosine Similarity* terbukti lebih unggul dibandingkan dengan K-Means *Euclidean* dalam kualitas klasterisasi. Hal ini ditunjukkan oleh *Silhouette Score* sebesar 0.6918, *Calinski-Harabasz Score* 3016.6532, dan *Davies-Bouldin Score* 0.9508, yang mengindikasikan klaster yang lebih terpisah, kompak, dan stabil.

2. Dalam sistem rekomendasi, *Hybrid Recommendation* dengan bobot (3,5,2) menggunakan K-Means *Cosine Similarity* pada *splitting* 90:10 memberikan keseimbangan optimal di semua metrik evaluasi. Kombinasi ini menghasilkan rekomendasi yang lebih relevan dan terstruktur, didukung oleh kualitas klasterisasi yang lebih baik.
3. Hasil Klasterisasi
   Analisis klaster menghasilkan tiga segmen utama pelanggan:
   a. Klaster 0: Pelanggan dengan pola belanja tidak terlalu aktif, lebih banyak berbelanja di akhir pekan dengan jumlah transaksi kecil.
   b. Klaster 1: Pelanggan yang jarang berbelanja tetapi melakukan transaksi dengan nominal besar, dengan tren pengeluaran yang meningkat.
   c. Klaster 2: Pelanggan yang sering berbelanja dan memiliki total pengeluaran tinggi, tetapi tingkat pembatalan pesanan mereka juga tinggi.
4. Implikasi Bisnis
   Berdasarkan hasil klasterisasi, strategi pemasaran yang disarankan adalah sebagai berikut:
   a. Klaster 0: Menargetkan pelanggan dengan promosi akhir pekan untuk meningkatkan frekuensi belanja.
   b. Klaster 1: Memberikan program loyalitas atau diskon untuk mempertahankan tren peningkatan pengeluaran mereka.
   c. Klaster 2: Mengembangkan strategi untuk mengurangi tingkat pembatalan pesanan dan menjaga stabilitas pengeluaran pelanggan.

Secara keseluruhan, metode yang digunakan dalam penelitian ini menunjukkan efektivitas dalam segmentasi pelanggan dan optimalisasi sistem rekomendasi, yang dapat membantu meningkatkan strategi pemasaran dalam e-commerce.

## E. Referensi

[1] S. A. Perdana, S. F. Florentin, and A. Santoso, "Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Aplikasi Alfagift," *Sebatik*, vol. 26, no. 2, pp. 420–427, Dec. 2022, doi: 10.46984/sebatik.v26i2.2134.

[2] A. Lomaksan, M. Tampubolon, T. Marta, E. Yuridis Butar Butar, and S. Rochimah, "Segmentasi Pelanggan Majalah Pada Situs Web E-Commerce Dengan K-Means++ Dan Metode Rfm," vol. 11, no. 6, pp. 1243–1252, Dec. 2024, doi: 10.25126/jtiik.2024118208.

[3] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, vol. 3, no. 1, pp. 46–56, Mar. 2024, doi: 10.58602/jima-ilkom.v3i1.26.

[4] E. Fernando, P. Mudjiraharjo, and M. Aswin, "Implementasi Pendekatan Collaborative Filtering Dan K-Means Clustering Pada Sistem Rekomendasi Mata Kuliah," *Jurnal Informatika dan Komputer) Akreditasi KEMENRISTEKDIKTI*, vol. 5, no. 2, 2022, doi: 10.33387/jiko.

[5] M. S. Pangestu and M. A. Fitriani, "Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan

Data Bibit Padi Menggunakan Algoritma K-Means," *Sainteks*, vol. 19, no. 2, p. 141, Oct. 2022, doi: 10.30595/sainteks.v19i2.14495.

[6] R. Rahmaddeni and F. Akbar, "Comparison of Naïve Bayes Algorithm, Support Vector Machine and Decision Tree in Analyzing Public Opinion on COVID-19 Vaccination in Indonesia," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 6, no. 1, p. 8, Apr. 2023, doi: 10.24014/ijaidm.v6i1.19966.

[7] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022, doi: 10.30812/matrik.v21i3.1851.

[8] F. Akbar, H. Wira Saputra, A. Karel Maulaya, and M. Fikri Hidayat, "Implementation of Decision Tree Algorithm C4.5 and Support Vector Regression for Stroke Disease Prediction Implementasi Algoritma Decision Tree C4.5 dan Support Vector Regression untuk Prediksi Penyakit Stroke," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, pp. 61–67, 2022.

[9] M. Z. Al-Taie, S. Kadry, and J. P. Lucas, "Online data preprocessing: A case study approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2620–2626, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2620-2626.

[10] R. Y. Firmansah, J. Dedy Irawan, and N. Vendyansyah, "Analisis Rfm (Recency, Frequency And Monetary) Produk Menggunakan Metode K-Means," Mar. 2021. Accessed: Feb. 07, 2025. [Online]. Available: https://ejournal.itn.ac.id/index.php/jati/article/view/3282

[11] M. Fithriyah, M. A. Yaqin, and S. Zaman, "K-Means Clustering Untuk Segmentasi Produk Berdasarkan Analisis Recency, Frequency, Monetary (RFM) Pada Data Transaksi Penjualan," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 3, no. 2, pp. 151–164, Aug. 2021, doi: 10.28926/ilkomnika.v3i2.284.

[12] D. Fenny, "Analisis Perbandingan Cosine Normalization Dan Min-Max Normalization Pada Pengelompokan Terjemahan Ayat Al Quran Menggunakan Algoritma K-Means Clustering," *Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta*, pp. 1–115.

[13] L. Qadrini, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," 2020. [Online]. Available: www.unipasby.ac.id

[14] Y. Florenza, "Tinjauan Naratif Tentang Mengeksplorasi Menyeluruh Kehebatan K-Means Algoritma Klasterisasi," 2023. Accessed: Dec. 31, 2024. [Online]. Available: http://ilmudata.org/index.php/ilmudata/article/view/317/303

[15] N. A. Maori, "Metode Elbow Dalam Optimasi Jumlah Cluster Pada K-Means Clustering," *Jurnal SIMETRIS*, vol. 14, 2023.

[16] M. A. Satriawan, R. Andreswari, and O. N. Pratiwi, "Segmentasi Pelanggan Telkomsel Menggunakan Metode Clustering Dengan Rfm Model Dan Algoritma K-Means," Apr. 2021. Accessed: Dec. 27, 2024. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14687/0

[17] I. Firman Ashari, E. Dwi Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," 2023. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[18] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/4832864.

[19] F. B. A. Larasati and H. Februariyanti, "Sistem Rekomendasi Product Emina Cosmetics Dengan Menggunakan Metode Content - Based Filtering," *Jurnal Manajemen informatika & Sistem Informasi*, vol. 4, pp. 45–54, Jan. 2021, Accessed: Dec. 27, 2024. [Online]. Available: https://e-journal.stmiklombok.ac.id/index.php/misi

[20] H. Februariyanti, A. D. Laksono, J. S. Wibowo, and M. S. Utomo, "Implementasi Metode Collaborative Filtering Untuk Sistem Rekomendasi Penjualan Pada Toko Mebel," *Jurnal Khatulistiwa Informatika*, vol. IX, pp. 43–50, Jun. 2021, [Online]. Available: www.unisbank.ac.id

[21] M. M. Sulthan, "Analisis Segmentasi, Targeting, dan Positioning Bahasa Indonesia Pasar Mobil dalam Menentukan Pasar yang Tepat untuk Wuling Motors."

[22] M. D. D. Maristha, I. Albertus Joko Santoso, and F. K. S. Dewi, "Sistem Rekomendasi Pembelian Produk Kesehatan pada E-Commerce ABC berbasis Graph Database Amazon Neptune menggunakan Metode Hybrid Content-Collaborative Filtering," 2021. [Online]. Available: https://ojs.uajy.ac.id/index.php/jbi/about/editorialTeam

[23] Kanafi and M. Fitriana, "Sistemasi: Jurnal Sistem Informasi Model Sistem Cerdas Rekomendasi Wisata Terbaik: Metode Weighted Product dan K-Means The Best Tourism Recommendation Intelligent System Model: Weighted Product and K-Means Methods," 2024. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[24] H. H. Arfisko and A. T. Wibowo, "Sistem Rekomendasi Film Menggunakan Metode Hybrid Collaborative Filtering Dan Content-Based Filtering," Jun. 2022. Accessed: Dec. 31, 2024. [Online]. Available: file:///D:/Downloads/18066-Article%20Text-35704-1-10-20220630.pdf

[25] L. Tommy, D. Novianto, and Y. Setiawan Japriadi, "Sistem Rekomendasi Hybrid untuk Pemesanan Hidangan Berdasarkan Karakteristik dan Rating Hidangan," Dec. 2020. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[26] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," 2020. [Online]. Available: www.ijacsa.thesai.org

[27] F. Fitriyeh, N. Haqqi, L. M. Choiriyah, and N. Ifada, "Dampak pembobotan pada metode hybrid user-based dan item-based untuk sistem rekomendasi film," *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 5, no. 3, pp. 516–525, 2024, doi: 10.37859/coscitech.v5i3.7637.