Rekomendasi Produk Berdasarkan Kategorisasi Pelanggan dengan Metode K-Means Clustering

Susy Susanty
Sains Data
Universitas Negeri Surabaya
susy.22012@mhs.unesa.ac.id

Nahdliyah Zahrah
Sains Data
Universitas Negeri Surabaya
nahdliyah.22024@mhs.unesa.ac.id

Aldora Novrizal N
Sains Data
Universitas Negeri Surabaya
aldora.22033@mhs.unesa.ac.id

Abstract—Sistem rekomendasi adalah sistem yang memprediksi preferensi pengguna terhadap suatu item berdasarkan profil pengguna. Sistem rekomendasi bermanfaat untuk mengatasi masalah kelebihan informasi dan meningkatkan kualitas pengambilan keputusan. Sistem rekomendasi dapat diklasifikasikan menjadi berbasis kolaboratif, berbasis konten, dan hibrida. Salah satu metode yang digunakan dalam sistem rekomendasi adalah K-Means Clustering, yaitu algoritma yang mengelompokkan data ke dalam sejumlah cluster berdasarkan kesamaan karakteristik. Algoritma ini bekerja berdasarkan jarak antar data dan hanya bekerja dengan data numerik. Algoritma ini telah diterapkan dalam berbagai bidang, seperti gizi, penjualan, dan bahan bangunan. Dalam konteks bisnis, algoritma ini dapat digunakan untuk memberikan rekomendasi produk berdasarkan kategorisasi pelanggan. Namun, algoritma ini memiliki beberapa kelemahan, seperti pemilihan jumlah cluster secara manual dan kurang cocok untuk data dengan variabel kategori. Algoritma ini juga relatif sederhana dan mampu mengolah data besar dengan cepat. Tujuan dari penelitian ini adalah untuk mengkaji penerapan metode K-Means Clustering dalam sistem rekomendasi dan mengevaluasi kinerja dan akurasi algoritma ini. Penelitian ini menggunakan data transaksi penjualan online dari sebuah perusahaan e-commerce. Data tersebut diolah dengan menggunakan algoritma K-Means Clustering untuk mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan perilaku pembelian mereka. Hasil penelitian menunjukkan bahwa algoritma K-Means Clustering dapat menghasilkan cluster yang homogen dan heterogen. Cluster yang homogen memiliki karakteristik pelanggan yang serupa, sedangkan cluster yang heterogen memiliki karakteristik pelanggan yang berbeda. Dengan demikian, algoritma ini dapat digunakan untuk memberikan rekomendasi produk yang sesuai dengan preferensi masing-masing segmen pelanggan. Rekomendasi produk ini diharapkan dapat meningkatkan loyalitas pelanggan dan pendapatan perusahaan.

Kata kunci: sistem rekomendasi, K-Means Clustering, data numerik, kategorisasi pelanggan, rekomendasi produk.

I. Pendahuluan

Sistem rekomendasi atau sistem rekomendasi adalah subkelas dari sistem penyaringan informasi yang berusaha memprediksi "rating" atau "preferensi" yang akan diberikan pengguna kepada suatu item. Sistem rekomendasi adalah sistem penyaringan informasi yang menangani masalah kelebihan informasi dengan menyaring fragmen informasi penting dari sejumlah besar informasi yang dihasilkan secara dinamis sesuai dengan preferensi, minat (atau) perilaku pengguna tentang item. Sistem rekomendasi memiliki kemampuan untuk memprediksi apakah pengguna tertentu akan menyukai suatu item atau tidak berdasarkan profil pengguna. Sistem rekomendasi bermanfaat bagi penyedia layanan dan pengguna. Mereka mengurangi biaya transaksi mencari dan memilih item di lingkungan belanja online. Sistem rekomendasi semakin populer dalam beberapa tahun terakhir, dan digunakan di berbagai bidang termasuk film, musik, berita, buku, artikel penelitian, kueri pencarian, tag sosial, dan produk secara umum. Sistem rekomendasi juga terbukti meningkatkan proses dan kualitas pengambilan keputusan. Dalam pengaturan e-commerce, sistem rekomendasi meningkatkan pendapatan, karena mereka adalah sarana yang efektif untuk menjual lebih banyak produk. Di perpustakaan ilmiah, sistem rekomendasi mendukung pengguna dengan memungkinkan mereka untuk melampaui pencarian katalog. Oleh karena itu, perlunya menggunakan teknik rekomendasi yang efisien dan akurat dalam

sistem yang akan memberikan rekomendasi yang relevan dan dapat diandalkan tidak dapat dilebih-lebihkan. Secara umum, sistem rekomendasi diklasifikasikan sebagai sistem rekomendasi berbasis kolaboratif (CF), berbasis konten, dan hibrida. CF banyak digunakan dalam RS, dan rekomendasi ini dapat dibagi menjadi berbasis pengguna dan berbasis item.

Salah satu metode yang digunakan dalam sistem rekomendasi adalah K-Means Clustering. K-Means Clustering adalah algoritma yang membagi data ke dalam kelompok berdasarkan kesamaan karakteristiknya. Halo, ini adalah Bing. Saya dapat membantu Anda menyatukan dua penjelasan yang Anda berikan menjadi satu. Berikut adalah hasilnya:

Metode K-Means Clustering adalah salah satu algoritma yang digunakan untuk mengelompokkan data ke dalam sejumlah cluster berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Algoritma ini bekerja berdasarkan jarak antar data dan hanya bekerja dengan data numerik. Langkah-langkah umum dalam algoritma K-Means meliputi menentukan jumlah kelompok, alokasi data ke dalam kelompok secara acak, dan menghitung pusat kelompok. Metode ini telah diterapkan dalam berbagai studi, seperti untuk menentukan status gizi balita, pola penjualan, dan pengelompokan bahan bangunan. Dalam konteks bisnis, K-Means Clustering dapat digunakan untuk memberikan rekomendasi produk berdasarkan kategorisasi pelanggan. Algoritma ini dapat membantu dalam pengelompokan data tanpa label kategori, seperti dalam segmentasi pelanggan, riset pasar, dan analisis pola penjualan3. Namun, algoritma ini kurang cocok jika diterapkan pada data dengan variabel kategori. Proses K-Means Clustering melibatkan pemilihan jumlah kluster secara manual, yang dapat menjadi subjektif. Namun, algoritma ini relatif sederhana dan mampu mengolah kumpulan data besar tanpa menghabiskan terlalu banyak waktu. Dengan demikian, dalam konteks bisnis, K-Means Clustering dapat menjadi alat yang berguna untuk menganalisis data pelanggan dan memberikan rekomendasi produk berdasarkan pola pembelian atau perilaku konsumen.

II. Tinjauan Pustaka

2.1 RFM

Analisis Recency, Frequency, and Monetary (RFM) merupakan metode yang sudah lama popular untuk mengukur hubungan dengan pelanggan, metode ini berbasis perilaku yang digunakan untuk menganalisis perilaku pelanggan dan kemudian membuat prediksi berdasarkan database (Hardiani dkk, 2015). Model yang melibatkan 3 variabel ini pertama kali diperkenalkan oleh Hughes, dengan penjelasan sebagai berikut (Saputra & Riksaomara, 2018):

1. Recency merupakan jarak dari waktu transaksi terrakhir kali dilakukan dengan waktu saat ini. Semakin kecil nilai karak waktu menandakan semakin besarnya nilai R.
2. Frequency merypakan tital jumlah transaksi yang dilakukan selama periode tertentu. Semakin besar jumlah transaksi, maka semakin besar nilai F.
3. Monetary merupakan tital nilai produk dalam bentuk uang dalam periode tertentu. Dengan semakin besarnya nilai produk, menandakan semakin besar pula nilai M.

2.2 Clustering

Menurut Rahmah dan Antares (2021), bahwa clustering dapat diartikan sebagai identifikasi kelas objek yang memiliki kemiripan dengan data yang lain. Teknik clustering ini dapat mengidentifikasi kepadatan dan jarak daerah dalam objek ruang dan dapat menemukan secara keseluruhan pola distribusi dan korelasi antara atribut. Terdapat dua jenis data clustering yang sering dugunakan dalam proses pengelompokkan yaitu hierarchial (hirarki) dan non-hierarchial (non hirarki) (Muningsih & Kiswati, 2018).

2.3 Algoritma K-Means

Algoritma K-Means merupakan algoritma yang paling popular dan menjadi salah satu algoritma yang paling penting dalam bidang data mining. Hal itu karena K-Means memiliki kelebihan sebagai algoritma yang mudah untuk diimplementasikan, waktu komputasi yang cepat, dan telah digunakan untuk menyelesaikan berbagai persoalan komputasi (Triyyansah & Fitrianah, 2018).

Algoritma K-Means merupakan algoritma pada unsupervised learning pada proses clustering yang mengelompokkan data berdasarkan kemiripan atau kesamman. Sehingga data yang memiliki karakter sama akan berada pada satu klaster dan data yang memiliki karakter yang berbeda akan berada pada klaster yang berbeda (Rahmah &Antares, 2021). Menurut Nuryani dan Darwis (2021), tujuan pengelompokkan untuk meminimalkan fungsi objektif yang ditetapkan selama proses pengelompokkan, biasanya proses pengelompokkan akan meminimalkan perbedaan dalam kelompok dan memaksimalkn perbedaan anatar kelompok.

Menurut Muningsih dan Kiswati (2019), algoritma K-Means merupakan sebuah metode sederhana untuk membagi suatu data sebanyak k jumlah cluster yang merupakan angka spesifik.

Langkah-langkah melakukan clustering dengan metode K-Means menurut Sulistiyawati dan Supriyanto (2023), yaitu:

1. Menentukan nilai k sebagai jumlah cluster yang ingin dibentuk.
2. Inisialisasi k pusat cluster atau centroid Langkah ini ketika diawal biasanya dilakukan secara random
3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak Euclidean (Euclidean Distance) hingga menemukan jarak paling dekat dari setiap data dengan centroid. Berikut persamaan Euclidean Distance:

$$\mathrm{De} = \sqrt{(\mathrm{x_i} - \mathrm{s_i})^2 (\mathrm{y_{i-t_i}})^2}$$

dimana:

De adalah *Euclidean Distance*, i merupakan banyaknya objek, (x,y) merupakan koordinat objek dan, (s,t) merupakan koordinat centroid.

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).
5. Memperbarui perbaharui nilai centroid, yang mana nilai centroid yang baru didapat dari rata-rata cluster yang bersangkutan menggunakan rumus berikut:

$$\mathrm{v_{ij}} = \frac{1}{\mathrm{N_i}} \sum_{\mathrm{k=0}}^{\mathrm{N_i}} \mathrm{X_{kj}}$$

dimana:

vij merupakan centroid atau rata rata cluster ke-i untuk variable ke-j, lalu Ni merupakan jumlah data yang menjadi anggota cluster ke-i, (i,k) adalah indeks cluster, j adalah indeks variable, dan Xkj adalah nilai data ke-k yang ada di dalam cluster tersebut untuk variable ke-j.

6. Melakukan perulangan dari Langkah 2 hingga 5, sampai anggota setiap cluster tidak ada yang berubah.

2.4 Elbow Method

Menurut Izzadin (2020), metode Elbow merupakan metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat presentasi hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik. Pada metode ini, jumlah kluster diperoleh dengan mencari titik siku pada grafik yang menunjukkan penurunan jumlah kesalahan yang signifikan. Oleh karena itu, nilai k yang dipilih adalah nilai dimana penurunan kesalahan menjadi lebih lambat dan grafik membentuk sudut tajam, yang disebut elbow (Syakur dkk, 2018).

Cara untuk mendapatkan nilai perbandingannya adalah dengan menghitung SSE (Sum of Square Error) pada setiap cluster, ketika nilai K semakin besar maka nilai SSE akan semakin kecil, berikut rumus SSE pada K-Means

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \|X_i - C_k\|_2^2$$

dimana:

k = banyak cluster yang terbentuk, Ci adaalah cluster ke-i, dan x merupakan data yang ada di setiap cluster.

Dalam penjelasan Marliana dkk, dijelaskan bahwa setelah melakukan perhitungan SSE dengan k yang sudah inisalisasi dan menentukan dititik mana penurunan SSE yang signifikan atau titik dimana grafik berbentuk siku maka pada nilai tersebutlah nilai k optimal berada dari perhitungan K-Means yang dilakukan.

2.5 Silhouette Coefficient

Silhouette Coefficient merupakan bentuk gabungan dari metode cohesion dan separation. Yang mana tujuan dari cohesion yaitu mengukur seberapa dekat hubungan antara objek salah satu kelompok cluster, dan tujuan dari separation ialah mengukur jarak antar cluster dengan cluster lainnya. Jadi silhouette coefficient merupakan suatu cara yang dugunakan untuk mengukur seberapa baik sebuah objek diklasifasikan saat menggunakan algoritma clustering. Nilai cluster terbaik menurut perhitungan Silhouttee coefficient ini adalah nilai dengan rata-rata mendekati satu (Simanjuntak dan Khaira, 2021). Berikut rumus Silhouette Coefficient dalam tulisan Rochman dkk (2022):

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

dimana: a(i) adalah jarak anatara objek data i dengan cluster tempat objek i berada yang berbeda, sedangkan b(i) adalah jarak antara objek data i dengan cluster tetangga. Nilai s(i) juga bisa didapat dengan beberapa ketentuan seperti berikut Rousseeuw, P. J. (1987):

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ b(i)/a(i) - 1, & if\ a(i) > b(i) \end{cases}$$

Ketika claster A hanya berisi satu objek dan membuat definisi nilai a(i) tidak jelas. Oleh karena itu maka ditetapkan untuk s(i) sama dengan nol. Pilihan ini dapat sewenang-wenang namun nilai nol tampaknya nilai paling netral dalam konteks ini. Dengan kata lain, ketika suatu cluster hanya terdiri dari satu objek, maka susah untuk menentukan seberapa baik atau buruknya objek tersebut. Oleh karena dipilihlah nilai yang dianggap paling netral untuk nilai s(i) yaitu nol.

Kategori dari nilai silhouette coefficient terbagi kedalam 4 rentang, seperti pada tabel 1 berikut (Rochman dkk, 2022):

Rentang nilai	Kategori Structure
0.7<SC<=1	Strong Structure
0.5<SC<=0.7	Medium Structure
0.25<SC<=0.5	Weak Structure
SC<=0.25	No Structure

Tabel 1. Kategori hasil Silhouette coefficient

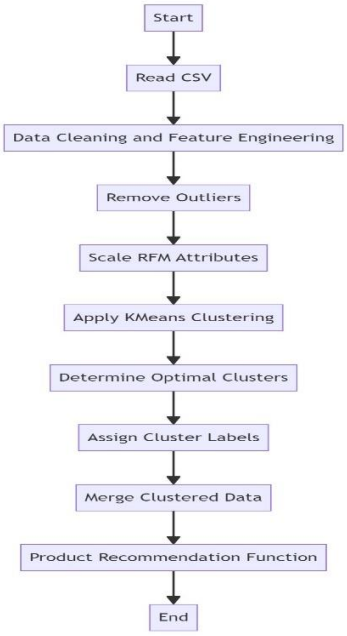
2.6 Cosine Similarity

Metode Cosine Similarity adalah sebuah metode yang digunakan untuk mengukur kemiripan antara dua dokumen atau teks dengan menggunakan konsep geometri vektor. Metode ini menganggap dokumen atau teks sebagai vektor dalam ruang multidimensi, di mana setiap dimensi merepresentasikan sebuah kata atau istilah yang muncul dalam dokumen atau teks. Kemiripan antara dua dokumen atau teks dihitung dengan menggunakan rumus cosinus dari sudut antara kedua vektor tersebut. Nilai cosinus dari sudut antara dua vektor berkisar antara -1 sampai 1, di mana nilai 1 menunjukkan bahwa kedua vektor memiliki arah yang sama (sudut 0 derajat), nilai 0 menunjukkan bahwa kedua vektor saling tegak lurus (sudut 90 derajat), dan nilai -1 menunjukkan bahwa kedua vektor memiliki arah yang berlawanan (sudut 180 derajat). Dalam konteks kemiripan dokumen atau teks, nilai cosinus yang mendekati 1 menunjukkan bahwa kedua dokumen atau teks memiliki kemiripan yang tinggi, sedangkan nilai cosinus yang mendekati -1 menunjukkan bahwa kedua dokumen atau teks memiliki kemiripan yang rendah (Samuel et al., 2018). Rekomendasi sistem adalah sistem yang memberikan saran atau rekomendasi kepada pengguna tentang produk atau layanan yang sesuai dengan preferensi atau kebutuhan mereka. Metode Cosine Similarity dapat digunakan untuk menghitung kemiripan antara profil pengguna dengan profil produk atau layanan, dan merekomendasikan produk atau layanan yang memiliki nilai kemiripan tertinggi kepada pengguna (Hung et al., 2019). Rumus perhitungan Cosine Similarity sebagai berikut (Sujasman dkk, 2020):

$$Sim = \frac{\sum_{k-1}^{l} weight_{lk} * weight_{qk}}{\sqrt{\sum_{k-1}^{l} (weight_{lk}^2 * weight_{qk}^2)}}$$

Dimana: $weight_{lk}$ = bobot setiap dokumen
$weight_{qk}$ = invest frekuensi dokumen log10 (n/df)

III. Metode



Gambar 1. Flowchart Metode

Penelitian ini mengintegrasikan beberapa metode secara sistematis yang akan diterapkan pada data yang telah dikumpulkan, untuk mencapai tujuan dalam memberikan rekomendasi produk berdasarkan perilaku pembelian serupa sebagai berikut:

3.1 Pengumpulan Data

Pengumpulan dataset yang akan digunakan dalam projek ini merupakan dataset yang berisi tentang riwayat pembelian barang, dataset didapat dari kaggle dengan jumlah baris sebanyak 541909 dengan 8 kolom yang terdiri dari kolom InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, Customer, dan Country (Pedersen, 2022).

Setiap kolom menggambarkan informasi spesifik terkait dengan transaksi penjualan produk secara online sebagai berikut:

- InvoiceNo: Nomor faktur. Merupakan angka integral 6 digit yang secara unik diberikan untuk setiap transaksi. Jika kode ini diawali dengan huruf 'c', menunjukkan bahwa transaksi tersebut merupakan pembatalan.
- StockCode: Kode produk (barang). Merupakan angka integral 5 digit yang secara unik diberikan untuk setiap produk yang berbeda.
- Description: Nama produk (barang). Merupakan data nominal yang menjelaskan nama atau deskripsi produk.
- Quantity: Jumlah produk (barang) dalam setiap transaksi. Merupakan data numerik yang menunjukkan jumlah produk yang dibeli atau dijual.

- InvoiceDate: Tanggal dan waktu faktur. Merupakan data numerik yang mencatat tanggal dan waktu ketika setiap transaksi dihasilkan.
- UnitPrice: Harga per unit. Merupakan data numerik yang menunjukkan harga per unit produk dalam mata uang sterling.
- CustomerID: Nomor pelanggan. Merupakan angka integral 5 digit yang secara unik diberikan untuk setiap pelanggan.
- Country: Nama negara. Merupakan data nominal yang mencatat nama negara tempat tinggal setiap pelanggan.

3.2 Preprocessing Data

Preproses data merupakan tahap awal dalam analisis data yang bertujuan untuk memastikan bahwa data yang akan dianalisis merupakan data yang sudah siap untuk diolah. Praprosess data ini mencakup memahami tipe data, kemudian handling missing values untuk mengatasi nilai-nilai yang hilang dalam data set, pembersihan data untuk membuang data yang mungkin tidak digunakan, handling categorical data untuk mengubah variable kategori menjadi bentuk numerik sehingga dapat diolah oleh algoritma yang akan digunakan, dan analisis statistik untuk mengetahui karakteristik data. Tujuan utama praprroses data ini adalah untuk meningkatkan kualitas data dan mempermudah untuk interpertasi. Langkah-langkah preprocessing data sebgai berikut:

3.2.1 Data Preparation

Pembersihan data yang tidak valid atau missing value. Pada data yang digunakan terdapat missing value pada kolom Description dan CustomerID, lalu dilakukan penghapusan kolom baris data pada data yang mengalami missing value tersebut.

3.2.2 Memahami Data

Memahami data merupakan tahap awal dalam proses analisis data. Pada tahap ini, dilakukan pemeriksaan dan pemahaman terhadap karakteristik data yang telah didapat, dengan tujuan untuk menghindari eror yang disebabkan oleh tidak tepatnya tipe data ketika diproses sebelum melangkah ke tahap pengolahan lebih lanjut.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Gambar 2. Info dataset

Pada Gambar 2, ditunjukkan bahwa tipe setiap data berbeda yang nantinya akan diubah sesuai tipe data yang dapat diolah pada model.

3.2.3 Menyiapkan Data RFM

Selanjutnya menyiapkan data RFM yang akan diproses lebih lanjut, pada data mentah yang didapat belum ada kolom yang merepresentasikan nilai RFM, maka dari itu langkah selanjutnya menyiapkan tabel yang merepresentasikan nilai RFM dari dataset yang di dapat sebagai berikut:
- Recency (kapan terakhir melakukan transaksi), Frequency (berapa banyak pelanggan yang melakukan transaksi), dan Monetary (jumlah uang yang dibelanjakan). Dalam perhitungan Recency yang dilakukan menggunakan dataset kolom InvoiceDate dengan mengekstraksinya menggunakan fungsi Diff dan kemudian mengambil bagian harinya saja.
- Frequency (jumlah transaksi) perhitunagn ini menggunakan kolom CustomerID dan InvoiceNo dengan menghitung InvoiceNo berdasarkan CustomerID.
- Monetary (total transaksi) perhitungan dilakukan dengan mengkalikan nilai Quantity dengan UnitPrice yang nantinya hasil kali tersebut dijumlahkan berdasarkan dengan kolom CustomerID.

3.2.4 Penghapusan Outlier

Algoritma K-Means perlu diperhatikan lebih dalam karena algoritma K-Means sangat sensitive terhadap outlier, dan outlier tersebut mungkin memiliki dampak yang tidak proporsional pada konfigurasi cluster akhir. Hal ini dapat mengakibatkan banyak kesalahan negatif yaitu, titik data yang seharusnya dinyatakan outlier ditutupi oleh pengelompokkan dan juga kesalahan positif (Chawla & Gionis (2013).

Jarak interkuartil atau IQR dapat digunakan secara obyektif untuk menentukan keabsahan suatu kasus outlier dengan menilai kasus tersebut dengan jarak dari nilai pusat atau tengah. Sebagai aturan, suatu kasus dianggap outlier univariat jika nilainya terletak setidaknya 1,5 kali panjang kotak IQR di luar salah satu kedua sisi tepi kotak. Artinya, jika nilai kasus tersebut terletak diluar rentang 1,5 kali IQR di atas atau bawah tepi kotak, maka kasus tersebut dapat dianggap outlier (Mowbrey dkk, (2019).

Penanganan outlier dengan menghapusnya adalah pendekatan yang paling konservatif dan mungkin merupakan pendekatan yang paling aman. Hal ini karena, menurut definisi, suatu kasus outlier tidak termasuk dalam populasi yang diminati, atau dalam scenario terbaik outlier adalah kasus ekstrem dari populasi tersebut. Ada dua metode penghapusan outlier, yitu penghapusan kasus (case deletation) dan

penghapusan variable (variable deletation) (Mowbrey dkk, (2019).

3.2.5 Normalisasi

Tahap ini merupakan tahap menormalisasikan data yang akan diolah dengan mentransformasi data dengan penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu (Nasution dkk (2019).

Normalisasi yang digunakan dalam projek ini adalah Z-Score. Z-score normalization merupakan metode normalisasi berdasarkan mean (nilai rata-rata) dan standard deviation (devisi standar) dari data. Rumus yang digunakan sebagai berikut (Nasution dkk, (2019):

$$nilai\ baru = \frac{nilailama - mean}{stdev}$$

3.3 Algoritma K-Means

K-Means adalah metode clustering berbasis jarak yang membagi data kedalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numerik. Algoritma K-Means merupakan metode non-heararki yang awalnya mengambil sebagian banyaknya komponen populasi untuk dijadikan pusat cluster awal (Metisen dan Sari, 2015).

Langkah-langkah melakukan clustering dengan metode K-Means menurut Sulistiyawati dan Supriyanto (2023), yaitu:

1. Menentukan nilai k sebagai jumlah cluster yang ingin dibentuk.

2. Inisialisasi k pusat cluster atau centroid Langkah ini ketika diawal biasanya dilakukan secara random

3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak Euclidean (Euclidean Distance) hingga menemukan jarak paling dekat dari setiap data dengan centroid. Berikut persamaan Euclidean Distance:

$$\mathrm{De} = \sqrt{(\mathrm{x_i} - \mathrm{s_i})^2 (\mathrm{y_{i-t_i}})^2}$$

dimana:

De adalah Euclidean Distance, i merupakan banyaknya objek, (x,y) merupakan koordinat objek dan, (s,t) merupakan koordinat centroid.

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).

5. Memperbarui perbaharui nilai centroid, yang mana nilai centroid yang baru didapat dari rata-rata cluster yang bersangkutan menggunakan rumus berikut:

$$\mathrm{v_{ij}} = \frac{1}{\mathrm{N_i}} \sum_{\mathrm{k=0}}^{\mathrm{N_i}} \mathrm{X_{kj}}$$

dimana:

vij merupakan centroid atau rata rata cluster ke-i untuk variable ke-j, lalu Ni merupakan jumlah data yang menjadi anggota cluster ke-i, (i,k) adalah indeks cluster, j adalah indeks variable, dan Xkj adalah nilai data ke-k yang ada di dalam cluster tersebut untuk variable ke-j.

6. Melakukan perulangan dari Langkah 2 hingga 5, sampai anggota setiap cluster tidak ada yang berubah.

3.4 Evaluasi Measure

Algoritma K-means yang telah dilakukan akan di evaluasi matriks menggunakan Silhouette coefficient untuk melihat kualitas klaster berdasarkan jarak kerapatan antar objek.

3.5 Rekomendasi produk

Rekomendasi produk ini menggunakan pendekatan cosine similarity terhadap data hasil cluster dengan mencari objek yang paling mirip terhadap data yang nantinya akan diberikan sebagai acuan pencarian kesamaan. Cosine similarity dalam projek ini mencari dengan batas cluster sehingga luas pencarian dipersempit, hal ini diharapkan dapat memberikan rekomendasi terbaik dengan nilai similarity yang tinggi.

IV. HASIL DAN PEMBAHASAN

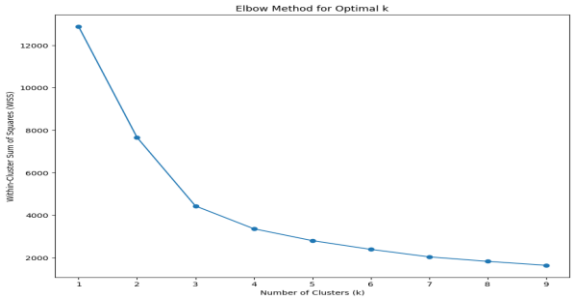
Pembahasan pada hasil proses projek rekomendasi produk berdasarkan kategorisasi pelanggan ini meliputi poin-poin sebagai berikut:

4.1 Data Preparation

Pada tahap ini persiapan data yang dilakukan tidak lagi pada data mentah, namun pada data RFM yang akan diproses lebih lanjut. Hasil pemeriksaan missing value pada data RFM ialah tidak ada data yang mengalami missing value, sehingga data dinilai sudah bersih. Kemudian penghapusan data outliers dan menormalisasikan data RFM yang akan di clustering, dengan begitu data siap untuk diolah.

4.2 Menentukan nilai K optimal

Tahap ini merupakan tahap menentukan nilai k optimal menggunakan elbow method. Metode ini mencari penurunan jumlah kesalahan nilai k dengan siku sebagai tandanya. Rentang iterasi yang diterapkan dalam perhitungan ini, mulai rentang 2 sampai 9. Hasil plot elbow terdapat pada gambar 2.

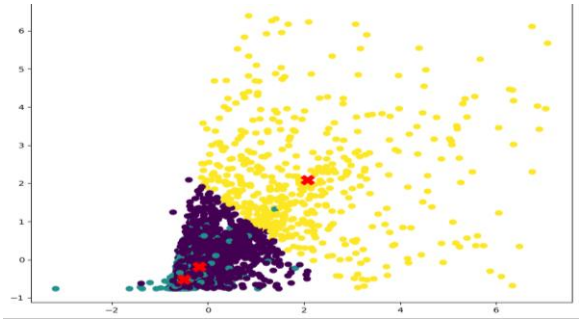


Gambar 3. Plot hasil Elbow Method

Pada Gambar 3. Didapatkan hasil bahwa nilai k optimal untuk klaster pada data RFM adalah 3, sehingga perhitungan K-means selanjutnya akan menggunakan k sama dengan 3.

4.3 Modelling K-Means

Pada proses pengelompokkan menggunakan K-means dengan k sama dengan 3. K-Means adalah salah satu algoritma pengelompokkan yang umum digunakan untuk memisahkan data menjadi beberapa kluster berdasarkan kemiripan karakteristik. Hasil pengelompokkan K-Means terdapat pada gambar 4.



Gambar 4. Merupakan plot hasil algoritma k-means

Dengan nilai rata-rata cluster sebagai berikut, pada tabel 2.

Cluster	Monetary	Frequency	Recency
1	-0.182000	-0.179593	-0.475793
2	-0.502376	-0.518466	1.542128
3	2.064973	2.086226	-0.700911

Tabel 2. Nilai rata-rata cluster

Berdasarkan hasil rata-rata cluster yang terdapat pada tabel 2 menunjukkan karakteristik setiap cluster sebagai berikut:

- Cluster 3 dapat dianggap sebagai kelompok pelanggan yang berharga tinggi, karena mereka memiliki total pembelian yang tinggi dan sering berbelanja dengan waktu sejak pembelian yang relatif baru.
- Cluster 1 mungkin adalah kelompok pelanggan yang berinteraksi sedang dengan bisnis, dengan total pembelian dan frekuensi belanja yang sedang, serta waktu sejak pembelian yang tidak terlalu baru.
- Cluster 2 mungkin mencakup pelanggan yang berinteraksi kurang dengan bisnis, dengan total pembelian dan frekuensi belanja yang lebih rendah, serta waktu sejak pembelian yang lebih lama.

4.4 Evaluasi Matriks

Pada tahap ini, dilakukan evaluasi hasil pengelompokkan menggunakan metode Silhouette coefficient. Silhouette coefficient adalah metrik avaluasi yang membantu mengukur seberapa baik objek-objek dalam suatu kluster dikelompokkan dan seberapa terpisah kluster tersebut dari kluster lainnya. Metrik ini memberikan skor antara -1 hingga1, dimana skor yang terbaik adalah skor yang paling tinggi (Simanjuntak dan Khaira, 2021). Hasil silhouette dengan nilai k sama dengan 3 adalah 0.51 yang menandakan bahwa kluster termasuk kategori medium structure.

4.5 Rekomendasi Produk

Hasil pengelompokkan K-Means yang telah dilakukan pada tabel RFM menjadi patokan dalam memberikan rekomendasi produk. Rekomendasi produk ini berdasarkan pendekatan personal pada perilaku pelanggan, Langkah selanjutnya adalah menggunakan cosine similarity untuk memberikan rekomendasi produk yang lebih spesifik.

Cosine similaritya dalah perhitungan kesamaan antara dua vektor n dimensi dengan mencari kosinus dari sudut diantara keduanya dan sering digunakan untuk membandingkan dokumen dalam text mining (Sujasman dkk, 2020). Dalam konteks ini, cosine similarity akan mempersempit pencarian sesuai dengan hasil cluster yang telah di dapat dan juga nilai similarity ini memiliki nilai yang tinggi. Gambar 4, merupakan hasil uji coba rekomendasi produk menggunakan Cosine similarity dengan menampilkan nama produk yang serupa dari yang diminta dan nilai tingkat similarity.

```
Rekomendasi untuk produk 'ALARM CLOCK BAKELIKE GREEN' (Cluster_Id 2.0):
- ALARM CLOCK BAKELIKE IVORY (ID: 22730), Similarity: 1.00
- BOX OF 6 CHRISTMAS CAKE DECORATIONS (ID: 23382), Similarity: 1.00
- WOODLAND CHARLOTTE BAG (ID: 20719), Similarity: 1.00
- WOODLAND CHARLOTTE BAG (ID: 20719), Similarity: 1.00
- WOODLAND CHARLOTTE BAG (ID: 20719), Similarity: 1.00
```

Gambar 4. Hasil rekomendasi produk cosine similarity.

Pada hasil tersebut semua barang rekomendasi produk memiliki niali similarity yang tinggi, hal tersebut mengindikasikan bahwa produk-produk tersebut memiliki kemiripan yang signifikan dengan produk yang sedang ditinjau, sehingga dianggap pilihan rekomendasi relevan.

Keberhasilan rekomendasi produk dengan nilai similarity tinggi dapat diartikan bahwa metode cosine similarity yang diimplementasikan telah efektif. Dengan demikian, hubungan antara sistem rekomendasin dan RFM telah bekerja sama untuk

memberikan rekomendasi produk sesuai dengan preferensi personal setiap pembeli.

V. KESIMPULAN

Sistem rekomendasi memiliki peran penting dalam menyaring informasi dan memberikan prediksi tentang preferensi pengguna terhadap suatu item. Metode seperti Analisis Recency, Frequency, and Monetary (RFM), Clustering, dan Algoritma K-Means telah digunakan secara luas dalam sistem rekomendasi untuk mengelompokkan data berdasarkan kesamaan karakteristik.

Analisis RFM penting untuk memahami perilaku pelanggan. Dengan memahami kapan transaksi terakhir dilakukan (Recency), berapa kali transaksi dilakukan (Frequency), dan total nilai transaksi (Monetary), kita dapat mendapatkan gambaran awal tentang preferensi pelanggan.
Setelah itu,

Clustering dan **Algoritma K-Means** dapat digunakan untuk mengelompokkan pelanggan berdasarkan perilaku mereka. Misalnya, pelanggan yang sering melakukan transaksi (frekuensi tinggi) dan baru saja melakukan transaksi (recency rendah) dapat dikelompokkan menjadi satu kluster.

Metode Elbow dan **Silhouette Coefficient** kemudian digunakan untuk menentukan jumlah kluster optimal. Dengan mengetahui jumlah kluster yang tepat, kita dapat memastikan bahwa setiap kluster memiliki karakteristik yang benar-benar unik dan berbeda dari kluster lainnya.

Terakhir, **Cosine Similarity** digunakan untuk mengukur sejauh mana profil pelanggan (berdasarkan RFM dan kluster mereka) mirip dengan profil produk atau layanan. Dengan demikian, sistem rekomendasi dapat memberikan rekomendasi yang paling sesuai dengan preferensi pelanggan.
Dengan cara ini, metode-metode tersebut saling melengkapi satu sama lain dalam menciptakan sistem rekomendasi yang efektif dan akurat. Dari analisis awal dengan RFM, pengelompokan dengan Clustering dan K-Means, penentuan jumlah kluster dengan Elbow dan Silhouette Coefficient, hingga pencocokan profil pelanggan dan produk dengan Cosine Similarity, setiap metode memiliki peran penting dalam proses tersebut.

Referensi

Hardiani, T., Sulistyo, S., & Hartanto, R. (2015). Segmentasi Nasabah Tabungan Menggunakan Model RFM (Recency, Frequency, Monetary) dan K-Means Pada Lembaga Keuangan Mikro. Dalam *Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (SEMANTIK) 2015*. Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada.

Sjhgfaputra, D. B., & Riksakomara, E. (2018). Implementasi Fuzzy C-Means dan Model RFM untuk Segmentasi Pelanggan (Studi Kasus: PT. XYZ). *JURNAL TEKNIK ITS*, 7(1), 2337-3520. Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS). (ISSN: 2301-928X Print).

Triyansyah, D., & Fitrianah, D. (2018). Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing: Studi Kasus Hoyweapstore. *IncomTech, Jurnal Telekomunikasi dan Komputer*, 8(3), SSN 2085-4811, eISSN: 2579-6089. DOI: 10.22441/incomtech.v8i2.4174. Teknik Informatika, Universitas Mercu Buana, Jakarta. (Accepted for Publication September 2018).

Rahmah, S. A., & Antares, J. (2021). Klasterisasi Seleksi Mahasiswa Calon Penerima Beasiswa Yayasan Menggunakan K-Means Clustering. *Jurnal Informatika, Manajemen dan Komputer*, 13(2), 25. eISSN: 2580-3042, pISSN: 1979-0694. Universitas Dharmawangsa, Jl. K.L.Yos Sudarso No. 224, Medan, 20115.

Nuryani, I., & Darwis, D. (2021). Analisis Clustering Pada Pengguna Brand Hp Menggunakan Metode K-Means. *Seminar Nasional Ilmu Komputer (SNASIKOM)*, *1*(1), 190–211. Retrieved from https://proceeding.unived.ac.id/index.php/snasikom/article/view/62

Sulistiyawati, A., & Supriyanto, E. (2023). Implementasi Algoritma K-means Clustering dalam Penetuan Siswa Kelas Unggulan. *Jurnal TEKNO KOMPAK*, 15(2), 25-36. P-ISSN: 1412-9663, E-ISSN: 2656-3525. Fakultas Teknik dan Ilmu Komputer, Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia.

Muningsih, E., & Kiswati, S. (2018). SISTEM APLIKASI BERBASIS OPTIMASI METODE ELBOW UNTUK PENENTUAN CLUSTERING PELANGGAN. *Joutica: Journal of Informatic Unisla*, *3*(1), 117–124. https://doi.org/10.30736/jti.v3i1.196

Izzadin, F. M. (2020). Optimasi Jumlah Cluster K-Means Dengan Metode Elbow Dan Silhouette Pada Produktivitas Tanaman Pangan Di Provinsi Jawa Tengah Tahun 2018.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of The Best Customer Profile Cluster. IOP Conf. Series: Materials Science and Engineering, 336(1), 012017. doi:10.1088/1757-899X/336/1/012017.

Merliana, N. P. E., Ernawati, & Santoso, A. J. Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means Clustering. Dalam Prosiding Seminar Nasional Multi Disiplin Ilmu. ISBN: 978-979-3649-81-8.

Rochman, E. M. S., Khozaimi, A., Suzanti, I. O., Husni, Rachmad, A., Jannah, R., & Khotimah, B. K. (2022). A Combination of Algorithm Agglomerative Hierarchical Cluster (AHC) and K-Means for Clustering Tourism in Madura, Indonesia. *Journal of Mathematics, Computational Science*, 12(62), https://doi.org/10.28919/jmcs/7086. ISSN: 1927-5307. Tersedia online di http://scik.org.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. doi:10.1016/0377-0427(87)90125-7.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *CESS (Journal of Computer Engineering System and Science)*, Vol. 4, No. 1, Januari 2019, hlm. 78. ISSN: 2502-7131 (p-ISSN), e-ISSN: 2502-714x.

Chawla, S., & Gionis, A. (2013). k-means–: A unified approach to clustering and outlier detection. Dalam *Proceedings of the 2013 SIAM International Conference on Data Mining* (hal. 189-197). DOI: https://doi.org/10.1137/1.9781611972832.21

Mowbray, F. I., Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2019). Univariate Outliers: A Conceptual Overview for the Nurse Researcher. *Canadian Journal of Nursing Research*, 51(1), 31–37. https://doi.org/10.1177/0844562118786647

Simanjuntak, K. P., & Khaira, U. (2021). Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 1(1), 7-16. Retrieved from https://journal.irpi.or.id/index.php/malcom/article/view/6. P-ISSN: 2797-2313, E-ISSN: 2775-85757.

Metisen, B. M., & Sari, H. L. (2015). Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan Produk pada Swalayan Fadhila. Jurnal Media Infotama, 11(2), 110. ISSN 1858-2680. Fakultas Ilmu Komputer, Universitas Dehasen Bengkulu.

Hung, P. D., Nguyen, T. T., Nguyen, T. H., & Nguyen, T. T. (2019). Customer Segmentation Using RFM Model and K-Means Clustering. International Journal of Advanced Computer Science and Applications, 10(11), 1-7.

Sujasman, M. B., Diana, & Syazili, A. (2020). Implementasi Metode Cosine Similarity untuk Rekomendasi Produk pada Aplikasi Penjualan Berbasis Mobile. Dalam Proceedings of the Bina Darma Conference on Computer Science (e-ISSN: 2685-2683, p-ISSN: 2685-2675), halaman 162.

Pedersen, U. T. (2022). Online Retail Dataset. Retrieved from Kaggle: https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset