**Modem: Jurnal Informatika dan Sains Teknologi**
**Volume. 3, Nomor. 3 Juli 2025**
e-ISSN: 3046-7209; p-ISSN: 3046-7217, Hal. 47-65
DOI: https://doi.org/10.62951/modem.v3i3.533
*Available Online at:* https://journal.aptii.or.id/index.php/Modem

# Penerapan Klasifikasi Pelanggan Berdasarkan Segmentasi Pelanggan pada UMKM *Monex Toys* Bekasi

**Eugenea Chiquita Zahrani Assyarif[1*], I Kadek Dwi Nuryana[2]**
[1-2]Universitas Negeri Surabaya, Surabaya, Indonesia

Alamat: Jl. Raya Kampus Unesa, Lidah Wetan, Kec. Lakarsantri, Surabaya
*Korespondensi penulis: eugenea.21019@mhs.unesa.ac.id*

***Abstract.*** *This study aims to conduct customer segmentation and develop a classification model to predict the clusters of new customers at Monex Toys Abadi Bekasi, a micro, small, and medium enterprise (MSME). Segmentation was performed using the K-Means Clustering algorithm, incorporating parameters such as Recency, Frequency, Monetary (RFM), purchased products, payment methods, shipping cost discounts, and the total number of products purchased by customers. The segmentation results revealed two clusters: (1) Discount Hunters and (2) Loyal Customers. Subsequently, a classification process was conducted to predict customer clusters using the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms. Evaluation results indicated that all models achieved high accuracy exceeding 98%. The best-performing model was obtained with SVM using a 70:30 data split, achieving an accuracy of 98.81%. This classification model was then implemented into a Streamlit-based cluster prediction application, enabling users to identify customer segments in real-time. The findings of this research are expected to assist MSMEs in understanding customer behavior, enhancing service quality, and supporting more effective marketing strategies.*

***Keywords****: Classification, Customer Segmentation, K-Means Clustering, SVM, UMKM.*

**Abstrak**. Penelitian ini bertujuan untuk melakukan segmentasi pelanggan serta membangun model klasifikasi guna memprediksi klaster pelanggan baru pada UMKM *Monex Toys* Abadi Bekasi. Segmentasi dilakukan dengan algoritma *K-Means Clustering* menggunakan *parameter Recency, Frequency, Monetary* (RFM), produk yang dibeli, metode pembayaran, potongan ongkos kirim, dan jumlah total produk yang dibeli pelanggan. Hasil segmentasi menunjukkan dua klaster, yaitu: (1) Pelanggan Pemburu Diskon (2) Pelanggan Setia. Kemudian dilakukan proses klasifikasi untuk memprediksi klaster pelanggan menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Support Vector Machine* (SVM). Hasil evaluasi menunjukkan bahwa seluruh model memiliki akurasi tinggi di atas 98%. Model terbaik diperoleh pada SVM dengan pembagian data 70:30, mencapai akurasi 98,81%. Model klasifikasi ini kemudian diimplementasikan ke dalam aplikasi prediksi klaster berbasis *Streamlit*, yang memungkinkan pengguna untuk mengidentifikasi segmen pelanggan secara *real-time*. Hasil penelitian ini diharapkan dapat membantu UMKM dalam memahami perilaku pelanggan, meningkatkan pelayanan, serta mendukung strategi pemasaran yang lebih efektif.

**Kata kunci**: K-Means Clustering, Klasifikasi, Segmentasi Pelanggan, SVM, UMKM.

## 1. LATAR BELAKANG

Perkembangan teknologi informasi (TI) terus mengalami kemajuan pesat dan memberikan dampak signifikan pada berbagai sektor, seperti bisnis, kesehatan, pendidikan, dan pemerintahan. Seluruh pelaku usaha, termasuk usaha mikro, kecil, dan menengah (UMKM), dituntut untuk mampu memanfaatkan teknologi agar tetap kompetitif. Salah satu strategi penting dalam mempertahankan pelanggan adalah dengan menerapkan manajemen hubungan pelanggan atau *Customer Relationship Management* (CRM). Segmentasi pelanggan merupakan aspek utama dalam CRM karena memungkinkan perusahaan menyesuaikan pelayanan sesuai karakteristik kelompok pelanggan (Afthoni et al., 2021).

*Received: Juni 02, 2025; Revised: Juni 16, 2025; Accepted: Juni 30 2025; Published: Juli 02, 2025*

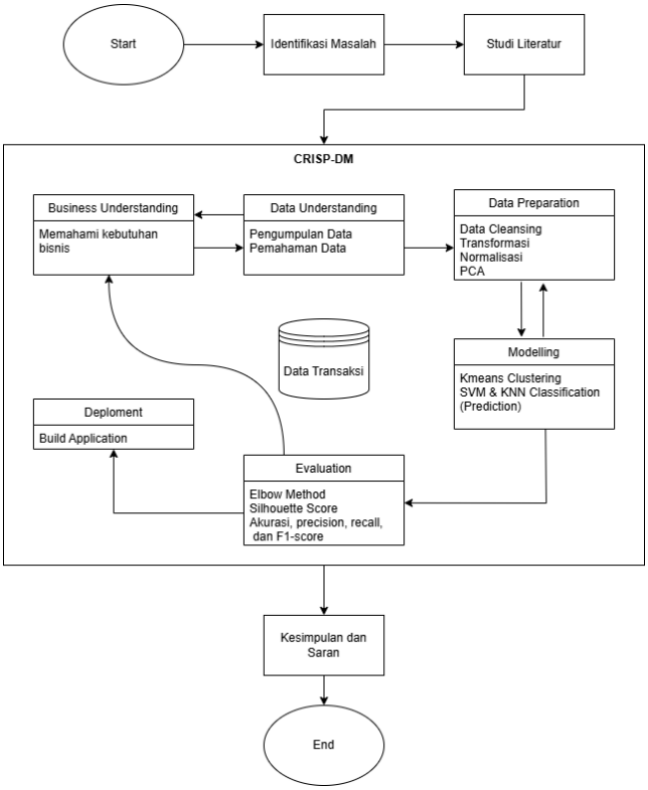
Monex Toys Abadi merupakan UMKM yang bergerak di bidang produksi dan penjualan boneka yang berlokasi di Bantar Gebang, Kota Bekasi. Usaha ini didirikan pada tahun 2013 dan mulai fokus pada penjualan online melalui Toko Mutiara Boneka sejak tahun 2022. Meski memiliki minat tinggi dalam penjualan *online*, *Monex Toys* Abadi belum menerapkan pengelompokan pelanggan dalam strategi pemasarannya. Selama ini, perlakuan yang diberikan kepada pelanggan masih seragam, seperti dalam pemberian diskon dan pelayanan. Hal ini menyebabkan sebagian pelanggan sulit dipertahankan, dan pelaku usaha kesulitan memahami perilaku konsumen, yang pada akhirnya menghambat pengembangan strategi pemasaran yang efektif. Untuk mengatasi permasalahan tersebut, diperlukan analisis perilaku pelanggan melalui segmentasi berdasarkan data transaksi. Segmentasi dilakukan dengan mengelompokkan pelanggan ke dalam beberapa segmen sesuai dengan perilaku pembeliannya. Salah satu metode yang umum digunakan adalah analisis *Recency, Frequency, dan Monetary* (RFM). Model RFM dipilih karena telah terbukti efektif dalam menggambarkan kebiasaan pembelian pelanggan.

Namun, penggunaan tiga variabel RFM saja dinilai kurang memberikan gambaran yang menyeluruh. Oleh karena itu, penelitian ini menambahkan beberapa indikator tambahan, yaitu jumlah jenis produk yang dibeli, jumlah metode pembayaran yang digunakan, ongkos kirim, dan jumlah produk yang dipesan. Penambahan indikator ini merujuk pada penelitian sebelumnya yang menyatakan bahwa produk, metode pembayaran, dan promosi seperti gratis ongkir memiliki pengaruh signifikan terhadap keputusan pembelian konsumen (Komariyah & Subiyantoro, 2023)(Istikomah & Hartono, 2022).

Dalam segmentasi ini digunakan algoritma *clustering K-Means* karena bersifat interaktif, mudah diterapkan, dan fleksibel dalam menangani data yang tersebar. Selain itu, hasil segmentasi akan dianalisis lebih lanjut dengan teknik klasifikasi untuk memprediksi cluster pelanggan pada data baru. Hal ini bertujuan untuk membangun sistem pendukung keputusan yang dapat digunakan secara berkelanjutan. Melalui penelitian ini, diharapkan *Monex Toys* Abadi dapat lebih mengenali karakteristik pelanggannya, meningkatkan kualitas pelayanan, serta memperkuat loyalitas dan kepuasan pelanggan. Dengan demikian, UMKM ini dapat meningkatkan pendapatan dan mempertahankan daya saingnya dalam jangka panjang.

## 2. METODE PENELITIAN

Penelitian ini mengikuti tahapan CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang terdiri dari enam fase utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* (Schröer et al., 2021). Alur lengkap proses penelitian digambarkan pada Gambar 1.




**Gambar 1.** Alur Penelitian

### *Business Understanding*

Masalah utama yang dihadapi oleh *Monex Toys* Bekasi adalah belumnya diterapkan Segmentasi Pelanggan. Hal ini menyebabkan pemilik usaha mengalami kesulitan dalam memahami perilaku konsumen serta menentukan cara yang tepat dalam melakukan pengelompokan pelanggan sebagai dasar dalam penyusunan strategi pemasaran yang tepat dan efektif dalam meningkatkan penjualan dan loyalitas pelanggan.

Penelitian ini bertujuan untuk melakukan pengelompokan pelanggan berdasarkan dengan perilaku pembelian mereka. Dengan pendekatan ini, *Monex Toys* Bekasi diharapkan dapat mengidentifikasi segmen pelanggan dengan potensi transaksi tinggi seperti pelanggan yang baru, pelanggan yang sering berbelanja, dan memiliki nilai belanja yang besar. Selain itu, penelitian ini juga akan menggambarkan preferensi pelanggan terhadap jenis produk, penggunaan diskon ongkos kirim, serta metode pembayaran yang

digunakan. Hal ini akan dijadikan sebagai acuan dalam merancang strategi pemasaran yang lebih personal dan tepat sasaran sesuai dengan karakteristik masing-masing kelompok pelanggan.

***Data Understanding***

Dataset terdiri dari 3094 baris data dengan 8 atribut utama. Data disajikan pada tabel berikut.

**Tabel 1.** Data Awal

| No | Nama Kolom | Tipe Data | Deskripsi |
|---|---|---|---|
| 1 | Waktu Pembayaran Dilakukan | *Datetime[ns]* | Tanggal dan waktu pembayaran dilakukan |
| 2 | Username (Pembeli) | *Object* | Identitas unik pelanggan |
| 3 | Total pembayaran | *Float64* | Total nominal pembayaran |
| 4 | Metode Pembayaran | *Object* | Jenis metode pembayaran yang digunakan |
| 5 | Nama Produk | *Object* | Nama produk yang dibeli |
| 6 | Perkiraan Ongkos Kirim | *Float64* | Estimasi ongkir sebelum potongan |
| 7 | Estimasi Potongan Biaya Pengiriman | *Float64* | Potongan biaya pengiriman dalam rupiah |
| 8 | Jumlah Produk di Pesan | *Int64* | Total kuantitas produk yang dipesan dalam satu transaksi |

***Data Preparation***

Langkah awal yang penting dilakukan dalam proses analisis data adalah melakukan data preparation untuk memahami struktur dan kualitas data yang akan digunakan (Wororomi et al., 2024).

**Mengatasi Data Hilang dan Duplikat**

Pada tahap pra-pemrosesan data, dilakukan pemeriksaan nilai hilang (*missing value*) menggunakan fungsi *isnull().sum()*. Ditemukan 7 nilai kosong pada kolom *Username* (pembeli), yang kemudian dihapus menggunakan *dropna()* karena kolom ini merupakan identitas utama pelanggan (Wororomi et al., 2024). Setelah itu, dilakukan pengecekan ulang untuk memastikan tidak ada nilai kosong tersisa. Selanjutnya, pemeriksaan data duplikat dilakukan menggunakan *df.duplicated().sum()* dan ditemukan 137 baris duplikat. Untuk menghindari bias dalam analisis, baris duplikat tersebut dihapus menggunakan fungsi *drop_duplicates()* (Handijono & Suhatman, 2024).

**Membuat Parameter RFM, Persen Ongkos Kirim Per Pelanggan, Jumlah Jenis Produk dan Metode Pembayaran, dan Jumlah Produk yang Dipesan**

a. Menentukan Tanggal Referensi Terakhir

Proses pertama yang harus dilakukan untuk analisis berbasis waktu, terutama dalam dalam perhitungan RFM nantinya diperlukan penentuan tanggal referensi terakhir yang digunakan sebagai acuan. Tanggal diperoleh dari nilai maksimum pada kolom Waktu Pembayaran Terakhir Dilakukan. Diketahui bahwa tanggal refernsi yang akan digunakan adalah 31 Desember 2023.

*b. Recency*

*Recency* adalah metrik yang mengukur selisih waktu antara transaksi terakhir pelanggan dengan tanggal referensi tertentu. Metrik ini menunjukkan seberapa baru atau lama pelanggan terakhir kali melakukan transaksi, dan digunakan untuk menganalisis perilaku pelanggan (Ariesty, 2015).

**Tabel 2.** *Recency 1*

| Username (Pembeli) | Recency |
|---|---|
| wedct9b3ca | 12/31/2023 |
| ryvee_ | 12/31/2023 |
| kylhvyki4x | 12/31/2023 |
| hnvnix_1pq | 12/31/2023 |
| hendang06 | 12/31/2023 |

**Tabel 3.** *Recency 2*

| Username (Pembeli) | Recency |
|---|---|
| wedct9b3ca | 0 |
| ryvee_ | 0 |
| kylhvyki4x | 0 |
| hnvnix_1pq | 0 |
| hendang06 | 0 |

*c. Frequency*

Langkah berikutnya dalam analisis RFM adalah menghitung *Frequency*, yaitu jumlah transaksi yang dilakukan masing-masing pelanggan selama periode pengamatan. Metrik ini digunakan untuk mengukur seberapa aktif pelanggan dalam bertransaksi. Perhitungan dilakukan dengan mengelompokkan data berdasarkan kolom *Username* (Pembeli) dan menghitung jumlah transaksi (Ariesty, 2015).

**Tabel 4.** *Frequency*

| Username (Pembeli) | Frequency |
|---|---|
| 01_dalawiguna.1008 | 1 |
| 020485erna | 1 |
| 02gedeprasetiawiguna | 1 |
| 02nxxlgh6v | 1 |
| 03c522sjiq | 1 |

*d. Monetary*

Komponen terakhir dalam analisis RFM adalah *Monetary*, yaitu total nilai transaksi yang dilakukan oleh masing-masing pelanggan selama periode pengamatan. Semakin tinggi nilai *monetary*, semakin besar pengeluaran pelanggan. Perhitungannya dilakukan dengan mengelompokkan data berdasarkan kolom *Username* (Pembeli) dan menjumlahkan nilai pada kolom Total Pembayaran (Ariesty, 2015).

**Tabel 5.** *Monetary*

| Username (Pembeli) | Monetary |
|---|---|
| 01_dalawiguna.1008 | 23776 |
| 020485erna | 40500 |
| 02gedeprasetiawiguna | 58200 |
| 02nxxlgh6v | 45000 |
| 03c522sjiq | 25000 |

e. Persen Ongkos Kirim

Selain analisis RFM, penelitian ini juga mengevaluasi efisiensi biaya pengiriman melalui indikator persentase potongan ongkos kirim. Menurut (Istikomah & Hartono, 2022), penawaran gratis ongkos kirim berpengaruh positif terhadap keputusan pembelian konsumen di platform seperti Shopee. Promosi gratis ongkir memengaruhi keputusan pembelian dengan cara menekan pengeluaran dan meningkatkan minat bertransaksi. Oleh karena itu persentase ongkos kirim fkigunakan. Nilai ini dihitung dari perbandingan estimasi potongan dengan perkiraan ongkir, lalu dikalikan 100. Data dikelompokkan berdasarkan *Username* (Pembeli) dan dihitung rata-ratanya, lalu disimpan dalam *DataFrame* Rata-rata potongan ongkos kirim. Indikator ini membantu menilai pengaruh promo ongkir terhadap perilaku pelanggan.

**Tabel 6.** Persentase Ongkos Kirim

| Username (Pembeli) | Potongan Ongkir (Persen) |
|---|---|
| 01_dalawiguna.1008 | 100 |
| 020485erna | 100 |
| 02gedeprasetiawiguna | 86.956522 |
| 02nxxlgh6v | 93.75 |

| | |
|---|---|
| 03c522sjiq | 100 |

f. Jumlah Jenis Produk Dibeli

Sebagai pelengkap analisis RFM, penelitian ini juga menambahkan indikator jumlah jenis produk yang dibeli untuk mengukur keragaman pembelian pelanggan. Data dikelompokkan berdasarkan *Username* (Pembeli), lalu dihitung jumlah unik dari kolom Nama Produk.

**Tabel 7.** Jumlah Jenis Produk Dibeli

| **Username (Pembeli)** | **Jumlah Jenis Produk Dibeli** |
|---|---|
| 01_dalawiguna.1008 | 1 |
| 020485erna | 1 |
| 02gedeprasetiawiguna | 1 |
| 02nxxlgh6v | 1 |
| 03c522sjiq | 1 |

g. Jumlah Jenis Metode Pembayaran

Penelitian ini juga menganalisis keragaman metode pembayaran yang digunakan pelanggan. Indikatornya adalah jumlah metode pembayaran berbeda yang tercatat dalam riwayat transaksi. Data dikelompokkan berdasarkan *Username* (Pembeli), lalu dihitung jumlah unik dari kolom Metode Pembayaran.

**Tabel 8.** Jumlah Jenis Metode Pembayaran

| **Username (Pembeli)** | **Jumlah Jenis Metode Pembayaran** |
|---|---|
| 01_dalawiguna.1008 | 1 |
| 020485erna | 1 |
| 02gedeprasetiawiguna | 1 |
| 02nxxlgh6v | 1 |
| 03c522sjiq | 1 |

h. Jumlah Produk yang Dipesan

Penelitian ini juga menambahkan indikator Total Produk Dipesan untuk menggambarkan intensitas pembelian pelanggan. Nilai ini diperoleh dengan mengelompokkan data berdasarkan *Username* (Pembeli) dan menjumlahkan isi kolom Jumlah Produk di Pesan. Hasilnya disimpan dalam *DataFrame* Total Produk Dipesan.

**Tabel 9.** Jumlah Produk Dipesan

| Username (Pembeli) | Total Produk Dipesan |
|---|---|
| 01_dalawiguna.1008 | 1 |
| 020485erna | 1 |
| 02gedeprasetiawiguna | 2 |
| 02nxxlgh6v | 2 |
| 03c522sjiq | 1 |

**Normalisasi Data**

Tahap ini dilakukan normalisasi data menggunakan metode *Min-Max Scaling* untuk menyetarakan skala setiap variabel dalam rentang 0 hingga 1 (Dorojatun, 2023). Normalisasi penting agar semua variabel memiliki kontribusi yang seimbang dalam proses klasterisasi (Wardhana & Winarno, 2024). Proses ini menggunakan fungsi *MinMaxScaler()* dari pustaka *scikit-learn*, dan hasilnya disimpan dalam *DataFrame* bernama *data_scaled.*

**PCA**

Reduksi dimensi dilakukan menggunakan metode *Principal Component Analysis* (PCA) untuk menyederhanakan jumlah fitur tanpa menghilangkan informasi penting secara signifikan (Sari, 2023). PCA mentransformasikan data ke dalam dua komponen utama yang merupakan kombinasi linier dari fitur-fitur asli. Proses ini menggunakan *sklearn.decomposition* dengan parameter *n_components=2*.

***Modeling***

Segmentasi dilakukan menggunakan algoritma *K-Means Clustering* dengan fitur-fitur hasil RFM, persentase diskon ongkos kirim, jumlah jenis produk yang dibeli, jumlah jenis metode pembayaran yang digunakan, jumlah jenis metode pembayaran yang digunakan, dan jumlah produk yang dibeli. Klasifikasi dilakukan menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan k = 1 hingga 5 dan *Support Vector Machine* (SVM), Split Data dibagi dalam dua skenario, yaitu 70:30 dan 80:20.

***Evaluation***

Pada tahap segmentasi pelanggan, algoritma *K-Means* digunakan untuk mengelompokkan pelanggan berdasarkan karakteristik tertentu. Jumlah cluster optimal ditentukan menggunakan metode *Elbow* dan *Silhouette Score*. Setelah segmentasi,

dilakukan klasifikasi untuk memprediksi klaster pelanggan baru, yang dievaluasi menggunakan *confusion matrix* dengan metrik akurasi, *precision*, *recall*, dan *F1 score*.

### *Deployment*

Pembuatan mini aplikasi menggunakan *Streamlit* memungkinkan pengguna berinteraksi langsung dengan model melalui antarmuka *web* sederhana. Aplikasi ini menerima input data pelanggan, memprediksi klaster menggunakan model klasifikasi, dan menampilkan hasilnya secara instan.

## 3. HASIL DAN PEMBAHASAN

### *Clustering*

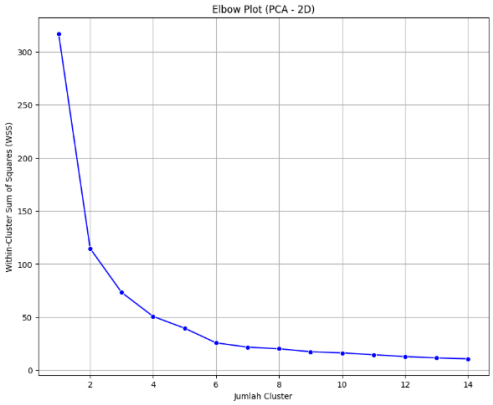
*Clustering* adalah teknik analisis data yang digunakan dalam pengelompokkan objek-objek yang memiliki kemiripan dalam satu kelompok atau klaster, sehingga objek dalam satu kelompok lebih mirip satu sama lain dibandingkan dengan objek pada kelompok lainnya (Nisa, 2019).

### *Elbow Method*

Berdasarkan grafik *Elbow Plot* dan hasil perhitungan *Within-Cluster Sum of Square* (WCSS) untuk nilai k dari 1 hingga 14, diperoleh penurunan WCSS yang paling signifikan terjadi dari k = 1 ke k = 2, yaitu dari 317.03 menjadi 114.66. Setelah itu, penurunan WCSS terus terjadi namun dengan laju yang semakin melandai. Hal ini mengindikasikan bahwa k = 2 merupakan titik siku (*elbow*) yang paling jelas, yang dapat dijadikan sebagai kandidat kuat untuk jumlah klaster optimal. Oleh karena itu, jumlah klaster yang digunakan dalam penelitian ini ditetapkan sebanyak dua klaster.

```
[317.02897422756047,
 114.66015062007277,
 73.2016202968734,
 50.590681492965906,
 39.49994215026881,
 25.668873440713078,
 21.647137802941625,
 20.085412141696377,
 17.32919490736908,
 16.22906945499342,
 14.500161757096473,
 12.761798359362944,
 11.492719018176025,
 10.643877156625008]
```

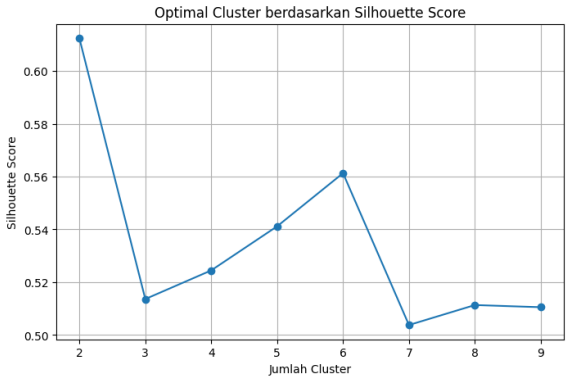
**Gambar 2.** Nilai WCSS untuk Berbagai Jumlah Klaster (k)




**Gambar 3.** Grafik *Elbow Method*

***Silhouette Score***

*Silhouette Score* digunakan sebagai validasi tambahan setelah metode *Elbow* untuk menilai kualitas klaster, dengan skor antara -1 hingga 1 (Mulyani et al., 2023). Semakin tinggi nilainya, semakin baik pemisahan antar klaster (Putra & Fadhillah, 2025). Berdasarkan pengujian dengan jumlah klaster 2 hingga 9, skor tertinggi diperoleh saat jumlah klaster 2, mendukung hasil dari metode *Elbow*.



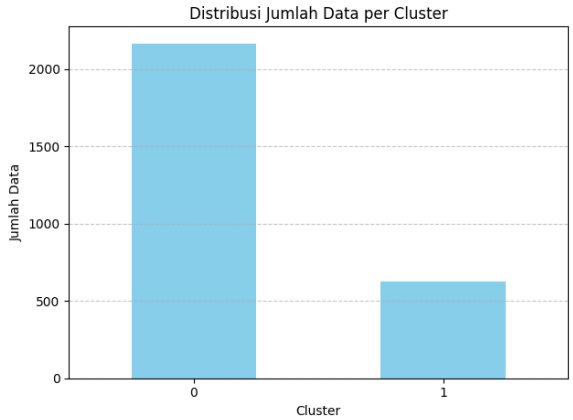

**Gambar 4.** Grafik *Silhouette Score*

**Hasil Segmentasi Pelanggan**

Setelah menentukan bahwa jumlah klaster yang optimal adalah dua, proses klasterisasi dilakukan dengan menggunakan algoritma *K-Means*. Algoritma ini bekerja dengan mengelompokkan data berdasarkan kedekatan terhadap pusat klaster (*centroid*) (Kavlakoglu & Winland, 2024). Setelah proses klasterisasi selesai, setiap data pada dataset *rfm_all* diberi label klaster sesuai hasil prediksi model dan disimpan dalam kolom baru bernama *Cluster*.

**Tabel 10.** Karakteristik Kluster

| Cluster | Recency | Frequency | Monetary | Potongan Ongkir (%) | Jumlah Jenis Metode Pembayaran | Jumlah Jenis Produk Dibeli | Total Produk Dipesan |
|---|---|---|---|---|---|---|---|
| 0 | 149.1976 | 1.05494 | 41692.92 | 89.14853 | 1.001847 | 1.041551 | 1.216528 |
| 1 | 200.451 | 1.067416 | 58459.01 | 11.25233 | 1 | 1.040128 | 1.340289 |




**Gambar 5.** Distribusi *Cluster*

Adapun interpretasi dari masing-masing klaster adalah sebagai berikut:

a) Klaster 0 terdiri dari pelanggan yang baru bertransaksi (*recency* rendah), memiliki pengeluaran kecil (*monetary* rendah), dan sangat bergantung pada promo (diskon rata-rata 89%). Mereka umumnya hanya membeli sekali, satu produk, dan menggunakan satu metode pembayaran, menunjukkan karakteristik sebagai pemburu diskon.

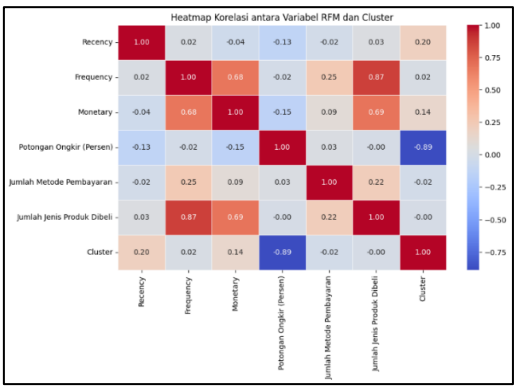
b) Klaster 1 berisi pelanggan yang sudah lama tidak membeli (*recency* tinggi), namun dengan nilai belanja lebih besar dan potongan ongkir rendah (rata-rata 11%). Mereka cenderung tidak menunggu promo dan memiliki daya beli lebih tinggi.

Perbedaan utama kedua klaster terletak pada *recency*, *monetary*, dan diskon ongkir, sementara atribut lain seperti frekuensi, metode pembayaran, jenis, dan jumlah produk menunjukkan pola serupa.

**Korelasi Antar Variabel**

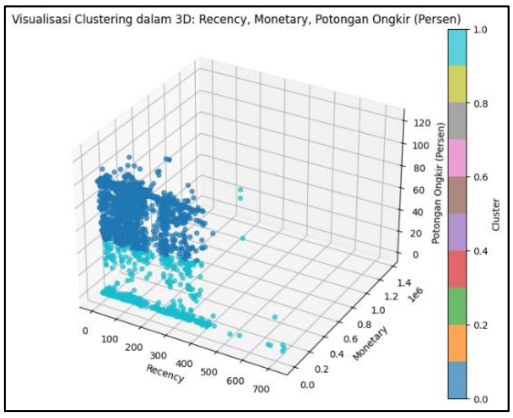
Berdasarkan *heatmap* korelasi, ditemukan bahwa hanya beberapa variabel yang berpengaruh signifikan terhadap pembentukan klaster. Potongan Ongkir (Persen) memiliki korelasi negatif tinggi dengan klaster (-0,89), yang berarti semakin besar potongan ongkir

yang diterima, semakin kecil nomor klaster pelanggan tersebut. Sementara itu, *Recency* (0,20) dan *Monetary* (0,14) menunjukkan korelasi positif, artinya semakin tinggi nilai keduanya, semakin besar pula nomor klaster pelanggan. Hal ini menunjukkan bahwa potongan ongkir, *recency*, dan *monetary* adalah variabel paling berpengaruh dalam segmentasi pelanggan.

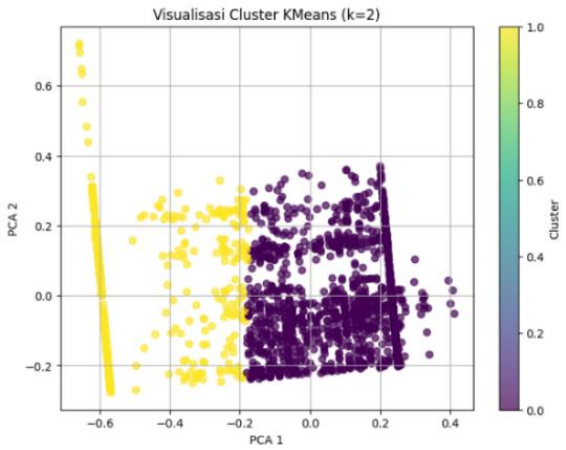



**Gambar 6.** Korelasi Antar Variabel

**Visualisasi Hasil *Cluster***




**Gambar 7.** Visualisasi




**Gambar 8.** Visualisasi PCA

Gambar di atas menampilkan visualisasi hasil clustering 3D berdasarkan tiga variabel paling berpengaruh, yaitu Potongan Ongkir (Persen), *Recency*, dan *Monetary*. Tiap titik merepresentasikan satu pelanggan, dengan warna berbeda menunjukkan klaster yang terbentuk. Terlihat dua kelompok yang cukup jelas terpisah, menandakan bahwa segmentasi berhasil mengelompokkan pelanggan berdasarkan karakteristik transaksinya.

**Penamaan Label *Cluster* Berdasarkan Karakteristik**

**Tabel 11.** Penamaan Label *Cluster*

| | |
|---|---|
| *Cluster* 0: Pelanggan Pemburu Diskon | *Recency* ter-rendah (149 hari) → Paling Aktif<br>*Monetary* ter-rendah → Paling sedikit uang dikeluarkan<br>Potongan Ongkir tertinggi (89,15%) → Memanfaatkan diskon ongkir |
| *Cluster* 1: Pelanggan Setia | *Recency* tertinggi (200 hari) → Tidak Aktif<br>*Monetary* tertinggi → uang yang dikeluarkan banyak<br>Potongan ongkir sedang (11,25%) → kurang memanfaatkan diskon |

**Prediksi (Klasifikasi)**

Tahap ini, dilakukan evaluasi pada model klasifikasi yang bertujuan untuk menentukan metode prediksi yang terbaik dalam melakukan pengelompokan pelanggan berdasarkan fitur-fitur RFM (*Recency, frequency, Monetary*), Potongan Ongkir (Persen), Jumlah Jenis Metode Pembayaran, Jumlah Jenis Produk yang Dibeli, dan Jumlah Produk yang Dibeli. Tujuh fitur tersebut kemudian digunakan sebagai input (variabel independen) yang disimpan dalam variabel X. Sedangkan label atau hasil klasterisasi yang sebelumnya telah diperoleh dari algoritma *K-Means* disimpan dalam variabel y sebagai target (variabel dependen). Evaluasi dilakukan dengan menggunakan algoritma *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM). Setiap skenario diuji dengan kombinasi skema data, Data penuh, Split data 80:20, dan 70:30.

***KNN***

K-Nearest Neighbor (KNN) merupakan salah satu metode supervised learning yang digunakan dalam tugas klasifikasi dan prediksi. Algoritma ini bekerja dengan mencari sejumlah tetangga terdekat dari data uji berdasarkan jarak tertentu, lalu menentukan kelas berdasarkan mayoritas label dari tetangga tersebut (Hasran, 2020). Evaluasi algoritma *K-Nearest Neighbors* dilakukan dengan data yang telah distandardisasi.

Dua skenario diuji: pertama tanpa pembagian data (menggunakan seluruh data), dan kedua dengan pembagian data menggunakan rasio 80:20 dan 70:30. Dilakukan pengujian nilai k dari 2 hingga 5 untuk mengevaluasi performa KNN terhadap data uji.

**Tabel 12.** Evaluasi *KNN*

| Data | K= | Akurasi |
|---|---|---|
| Full Data | 2 | 99.35% |
| | 3 | 99.35% |
| | 4 | 99.03% |
| | 5 | 99.07% |
| 70:30 | 2 | 98.45% |
| | 3 | 98.69% |
| | 4 | 98.45% |
| | 5 | 98.45% |
| 80:20 | 2 | 98.39% |
| | 3 | 98.75% |
| | 4 | 98.57% |
| | 5 | 98.57% |

***SVM***

Support Vector Machine (SVM) merupakan algoritma machine learning yang cukup baik dalam klasifikasi. Algoritma ini menggunakan hyperplane dan support vector untuk memisahkan kelas-kelas dalam ruang dimensi secara optimal (Nugroho et al., 2021). Evaluasi model *Support Vector Machine* (SVM) dilakukan pada data yang telah diskalakan melalui dua skenario: tanpa pembagian data dan dengan pembagian data (80:20 dan 70:30). Pada skenario awal, model dilatih dan diuji pada seluruh data. Selanjutnya, evaluasi dilakukan menggunakan metrik akurasi, *classification report (precision, recall, f1-score),* dan *confusion matrix*.

**Tabel 13.** Evaluasi *SVM*

| Data | Akurasi |
|---|---|
| Full Data | 98.85% |
| **70:30** | **98.81%** |
| 80:20 | 98.75% |

Meskipun akurasi pada data penuh tinggi, skenario pembagian data 70:30 memberikan hasil terbaik dengan akurasi 98,81%. Ini menunjukkan bahwa pembagian 70:30 memberikan keseimbangan optimal antara data latih dan uji serta kemampuan generalisasi yang baik.

**Implementasi SVM (70:30)**

Dari hasil perbandingan antara algoritma *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM) dengan berbagai skenario pembagian data, diketahui bahwa evaluasi pada seluruh data memang menghasilkan akurasi tinggi. Namun, hal tersebut tidak mencerminkan performa sesungguhnya karena berisiko mengalami *overfitting*. Oleh karena itu, dilakukan pemisahan data untuk menguji kemampuan generalisasi model.

Setelah dilakukan pembagian data dengan skenario 80:20 dan 70:30, diperoleh bahwa implementasi SVM dengan skenario 70:30 memberikan hasil paling optimal. Model ini menunjukkan akurasi tertinggi sebesar 98,81%, serta kesalahan klasifikasi yang sangat minim berdasarkan confusion matrix. Oleh karena itu, konfigurasi tersebut dipilih sebagai model akhir dalam penelitian karena mampu memberikan performa klasifikasi yang paling baik.

**Evaluasi**

Berdasarkan hasil evaluasi model *Support Vector Machine* (SVM) dengan pembagian data 70:30 memberikan hasil yang sangat baik dengan akurasi 98,81% Dengan *precision, recall,* dan *F1-score* yang tinggi, model ini menunjukkan kemampuan yang sangat baik dalam mengklasifikasikan data ke dalam enam kelas yang berbeda.

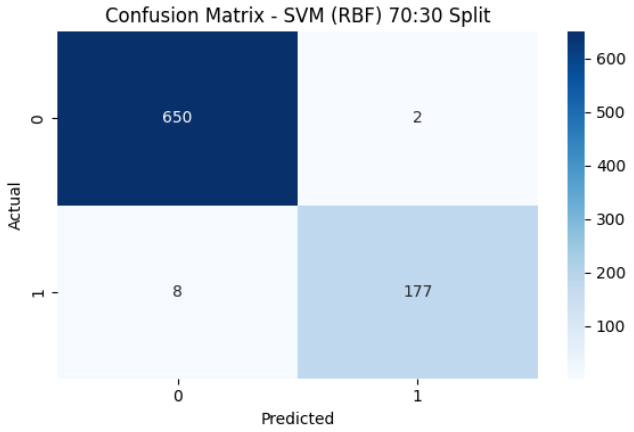
Meski ada sedikit kesalahan dalam klasifikasi, kesalahan tersebut sangat kecil dan tidak mempengaruhi kinerja keseluruhan model. Secara keseluruhan, model ini sangat layak digunakan untuk tugas klasifikasi dengan dataset ini. Hasil evaluasi disajikan pada gambar berikut.

```
Akurasi: 0.9881
Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       652
           1       0.99      0.96      0.97       185

    accuracy                           0.99       837
   macro avg       0.99      0.98      0.98       837
weighted avg       0.99      0.99      0.99       837
```

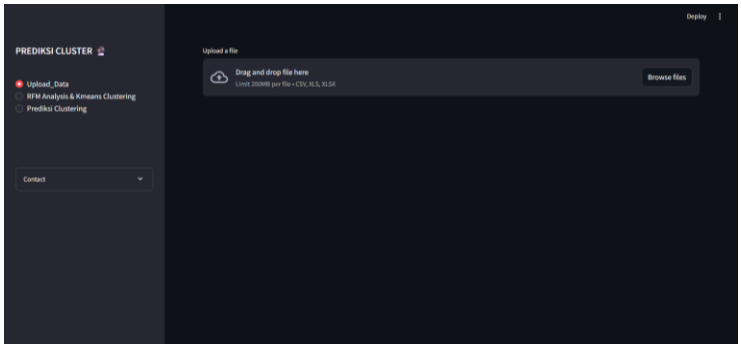
**Gambar 9.** *Classification Report*




**Gambar 10.** *Confusion Matrix*

***Deployment***

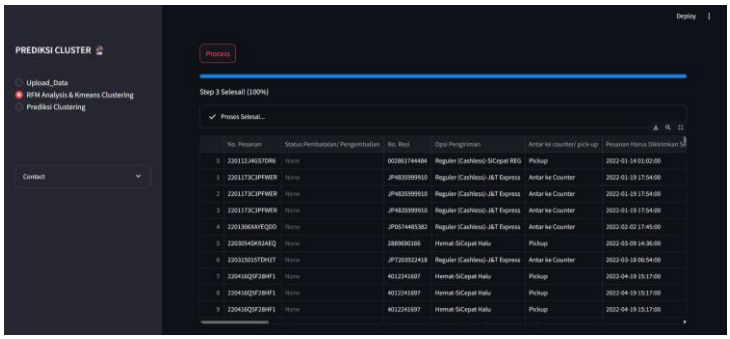
1) Upload File




**Gambar 11.** *Streamlit*

Halaman *Upload File* berfungsi sebagai awal bagi pengguna untuk mengunggah *file* dataset yang akan digunakan dalam proses *clustering* dan prediksi. Pada halaman ini, pengguna diminta untuk mengunggah *file* dalam format *.csv*, *.xls*, atau *.xlsx*, yang berisikan data transaksi yang sesuai.

2) Model




**Gambar 12.** *Streamlit 2*

Setelah mengunggah, pada halaman ini akan dilakukan proses lebih lanjut oleh sistem yaitu mulai pembersihan data hingga *clustering* dengan *K-Means*.

3) Prediksi Data Baru


**Gambar 13.** *Streamlit 3*

Pada halaman prediksi, pengguna diminta menginputkan data pelanggan baru untuk selanjutnya akan dilakukan prediksi pelanggan dengan tersebut termasuk dalam klaster apa.

## 4. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk melakukan segmentasi pelanggan dan membangun model klasifikasi untuk memprediksi klaster pelanggan baru pada *Monex Toys* Bekasi. Hasil segmentasi menggunakan algoritma *K-Means* menunjukkan terbentuknya dua klaster pelanggan, yaitu *Cluster* 0 (Pelanggan Pemburu Diskon) yang cenderung hanya bertransaksi saat ada promo besar dan memiliki nilai *recency* serta monetary yang rendah, serta *Cluster* 1 (Pelanggan Setia) dengan nilai monetary lebih tinggi, tidak terlalu bergantung pada diskon, meskipun sudah lama tidak bertransaksi (*recency* tinggi). Perbedaan utama terletak pada nilai *recency*, *monetary*, dan potongan ongkir, sementara variabel lain seperti frekuensi pembelian, metode pembayaran, dan jenis produk relatif seragam di kedua klaster. Setelah segmentasi, dilakukan klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN) dan *Support Vector Machine* (SVM) untuk memprediksi klaster pelanggan baru. Evaluasi dilakukan pada beberapa skenario, termasuk full data, split 80:20, dan split 70:30. Hasil terbaik diperoleh dari model SVM dengan pembagian data 70:30, dengan akurasi sebesar 98,81%. Model terbaik ini kemudian diterapkan dalam aplikasi berbasis *Streamlit*, yang memungkinkan pengguna memprediksi klaster pelanggan baru secara langsung, sehingga mendukung pengambilan keputusan bisnis yang lebih cepat dan efisien.

**UCAPAN TERIMA KASIH**

Saya ucapakn terima kasih yang setulus-tulusnya kepada Bapak I Kadek Dwi Nuryana, S.Kom., M.Kom. selaku Dosen Pembimbing saya, atas arahan, bimbingan, dan masukan yang sangat berarti selama penyusunan skripsi ini. Ucapan terima kasih juga saya sampaikan kepada seluruh dosen dan staf Departemen Sistem Informasi, Universitas Negeri Surabaya, yang telah memberikan ilmu dan pengalaman berharga selama masa studi.

**DAFTAR REFERENSI**

Afthoni, R., Hamdhani, M., Ardianto, Karimah, A. F., & Patria, H. (2021). Pemanfaatan algoritma machine learning untuk segmentasi pelanggan berbasis data konsumsi listrik di PT PLN XYZ. *Seminar Nasional Teknik dan Manajemen Industri*, *1*(1). https://doi.org/10.28932/sentekmi2021.v1i1.85

Ariesty, R. (2015). *Segmentasi pelanggan menggunakan SOM, algoritma K-Means dan analisis LRFM untuk penyusunan rekomendasi strategi pemasaran pada Klinik Kecantikan Nanisa, Sidoarjo* [Skripsi, Institut Teknologi Sepuluh Nopember]. https://repository.its.ac.id/70832/

Dorojatun, H. (2023). Normalisasi dan standardisasi dalam data mining. *Artikel DJKN KPKNL Mamuju*. https://www.djkn.kemenkeu.go.id/artikel/baca/15943/Seri-Artikel-DDDM-KPKNL-Mamuju-Normalisasi-dan-Standardisasi-dalam-Data-Mining.html

Handijono, A., & Suhatman, Z. (2024). Meningkatkan deduplikasi data melalui kesamaan teks dalam pembelajaran mesin: Pendekatan komprehensif. *AKADEMIK: Jurnal Mahasiswa Humanis*, *4*(2), 602–615. https://doi.org/10.37481/jmh.v4i2.955

Hasran. (2020). Klasifikasi penyakit jantung menggunakan metode K-Nearest Neighbor. *Indonesian Journal of Data and Science*, *1*(1). https://doi.org/10.33096/ijodas.v1i1.3

Istikomah, N., & Hartono, B. (2022). Analisis persepsi promosi gratis ongkos kirim (ongkir) Shopee terhadap keputusan pembelian. *Jurnal Bisnis Kompetitif*, *1*(2), 49–57. https://doi.org/10.35446/bisniskompetif.v1i2.1011

Kavlakoglu, E., & Winland, V. (2024). Apa itu K-Means clustering? *IBM*. https://www.ibm.com/id-id/topics/k-means-clustering

Komariyah, S., & Subiyantoro, H. (2023). Pengaruh kualitas produk, promosi, dan metode pembayaran terhadap keputusan pembelian konsumen pada marketplace Shopee (Studi kasus pada mahasiswa Universitas Bhinneka PGRI Tulungagung angkatan tahun 2019–2020). *JURNAL ECONOMINA*, *2*(9), 2644–2659. https://doi.org/10.55681/economina.v2i9.839

Mulyani, H., Setiawan, R. A., & Fathi, H. (2023). Optimization of K value in clustering using silhouette score (Case study: Mall customers data). *Journal of Information Technology and Its Utilization*, *6*(2), 45–50. https://doi.org/10.56873/jitu.6.2.5243

Nisa, K. (2019). *Analisis cluster dengan menggunakan metode hierarki untuk pengelompokan kecamatan di Kabupaten Langkat berdasarkan indikator kesehatan* [Skripsi, Universitas Islam Negeri Sumatera Barat]. http://repository.uinsu.ac.id/11406/

Nugroho, B. A., Pradana, A. K. A., & Nurfarida, E. (2021). Prediksi waktu kedatangan pelanggan servis kendaraan bermotor berdasarkan data historis menggunakan Support Vector Machine. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, *7*(1), 25. https://doi.org/10.26418/jp.v7i1.42964

Putra, I. P., & Fadhillah, A. (2025). Perbandingan metode K-Means dan Hierarchical Clustering dalam pengelompokan data penduduk miskin di Kabupaten Cianjur. *LANCAH: Jurnal Inovasi dan Tren*, *3*(1). https://doi.org/10.35870/ljit.v3i1.4028

Sari, D. R. P. (2023). Metode Principal Component Analysis (PCA) sebagai penanganan asumsi multikolinearitas. *PARAMETER: Jurnal Matematika, Statistika dan Terapannya*, *2*(02), 115–124. https://doi.org/10.30598/parameterv2i02pp115-124

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Wardhana, F. P., & Winarno, S. (2024). Analisis pemain terbaik sepak bola dengan menggunakan algoritma K-Means. *Edumatic: Jurnal Pendidikan Informatika*, *8*(2), 409–417. https://doi.org/10.29408/edumatic.v8i2.27105

Wororomi, J. K., Reba, F., Mandowen, S. A., Sroyer, A. M., Manurung, H. E., Koibur, M. E., Pujiarini, E. H., Edy, M. R., Yuliana, H., Yusuf, M., & Dinata, R. M. (2024). *Data mining (memahami pola di balik angka)* (N. R. Jayanti, Ed.). Eureka Media Aksara.