PENERAPAN METODE K-MEANS CLUSTERING DAN SUPPORT VECTOR MACHINE (SVM) BERBASIS MODEL RFM UNTUK KLASIFIKASI TIER PELANGGAN

Tariq[1*], Nurmalitasari[2], Faulinda Ely Nastiti[3]

[1,2,3] Fakultas Ilmu Komputer Universitas Duta Bangsa Surakarta

[1*]220101037@mhs.udb.ac.id,[2]nurmalitasari@udb.ac.id,[3]faulinda_ely@udb.ac.id

DOI: https://doi.org/10.46880/mtk.v12i2.5958

ABSTRACT

Suboptimal management of large-scale transaction data can lead to marketing inefficiencies, particularly in determining promotional strategies that do not align with customer characteristics. This study aims to map the customer loyalty of CV Ekasa's client partners, by segmenting its customers using an integrated Recency, Frequency, Monetary (RFM) model, K-Means Clustering, and Support Vector Machine (SVM) classification. The dataset comprises 287,512 raw point-of-sale transaction records collected between October 2022 and September 2025, which after preprocessing yielded 341 valid customers for RFM modeling. Following the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, RFM features were log-transformed and standardized before clustering. Silhouette Score evaluation across k = 1–10 identified two customer segments (k = 2, Silhouette Score = 0.482) as optimal, labeled Passive Tier and Active Tier. These cluster labels were then used as classification targets for a linear-kernel SVM, evaluated under two data-splitting scenarios (80:20 and 70:30). The model achieved 97.10% accuracy with the 80:20 split and 98.06% with the 70:30 split, with precision, recall, and F1-scores above 0.97 for both tiers in both scenarios. These findings indicate that the integrated RFM–K-Means–SVM pipeline classifies customer loyalty tiers reliably and stably. The resulting model was deployed as an interactive Streamlit dashboard, giving CV Ekasa's client partner a practical, data-driven basis for designing more targeted and efficient marketing and retention strategies.

Keyword : RFM, K-Means Clustering, Support Vector Machine, Customer Segmentation, CRISP-DM

I. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pertumbuhan volume data transaksi pada sektor retail secara signifikan. Setiap transaksi menyimpan informasi mengenai perilaku pembelian konsumen yang bila dianalisis dapat menjadi dasar pengambilan keputusan bisnis, khususnya dalam mempertahankan pelanggan lama. Permasalahan ini secara nyata dialami oleh Klien Mitra CV Ekasa di Malang: selama periode observasi tiga tahun, yaitu 1 Oktober 2022 hingga 17 September 2025, sistem kasir (*Point of Sale*) mitra telah mencatat sebanyak 287.512 baris data transaksi mentah, namun setelah proses seleksi hanya tersisa 341 pelanggan dengan riwayat transaksi valid yang dapat dianalisis secara individual. Kondisi ini diperparah oleh tidak adanya pembeda antara segmen pelanggan, yang mengakibatkan penawaran promosi dipukul rata dan pada akhirnya berdampak pada pemborosan anggaran pemasaran [1].

Permasalahan tersebut menunjukkan perlunya suatu pendekatan analisis data yang dapat mengidentifikasi segmen pelanggan secara objektif berdasarkan riwayat transaksi yang dimiliki. *Data mining* merupakan pendekatan yang digunakan untuk menggali informasi dari kumpulan data berukuran besar. Salah satu teknik analisis yang relevan untuk mengevaluasi kualitas dan loyalitas pelanggan adalah metode *Recency, Frequency, Monetary* (RFM), yaitu metode yang bertujuan menilai perilaku belanja berdasarkan waktu transaksi terakhir, frekuensi pembelian, dan total nominal yang dibelanjakan. Hasil dari segmentasi berbasis model RFM ini dapat dimanfaatkan untuk menyusun strategi promosi yang tepat sasaran, meningkatkan retensi pelanggan, serta mengoptimalkan alokasi anggaran pemasaran [2].

Untuk dapat menerapkan model RFM tersebut secara objektif pada data pelanggan berskala besar, diperlukan suatu teknik pengelompokan yang mampu memisahkan pelanggan secara otomatis berdasarkan kemiripan karakteristik RFM yang dimiliki. Algoritma *K-Means Clustering* merupakan metode pengelompokan yang banyak diterapkan untuk membagi pelanggan berdasarkan variabel RFM karena kemampuannya dalam memisahkan sekumpulan data skala besar dengan efektif. Beberapa penelitian sebelumnya telah membuktikan efektivitas kombinasi metode ini dalam berbagai konteks retail. Syahra et al. menerapkan *K-Means* dan RFM pada industri ritel dan berhasil membentuk kelompok pelanggan untuk mendukung manajemen hubungan pelanggan. Penelitian lain oleh Sharyanto dan Lestari [3] pada sektor perdagangan elektronik menunjukkan bahwa algoritma *K-Means* mampu mengidentifikasi kelompok pelanggan potensial secara objektif. Akande et al. juga berhasil membuktikan bahwa pengelompokan yang didasarkan pada data riwayat transaksi mampu mengatasi penilaian subjektif yang sering terjadi pada pemilahan tingkatan pelanggan secara manual. Penelitian Ardi et al. juga menerapkan kombinasi model RFM dan algoritma K-Means pada segmentasi pelanggan sebuah distributor fashion, yang menghasilkan 2 klaster optimal berdasarkan pengujian Silhouette Score dengan skor sebesar 0,915 [4].

Meskipun penelitian-penelitian tersebut menunjukkan hasil yang baik, sebagian besar

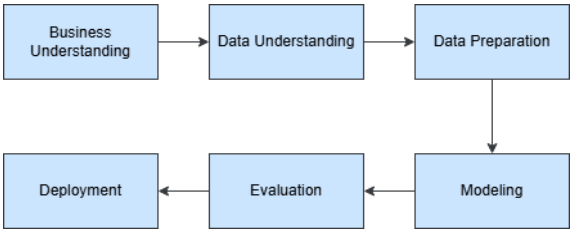
This work is licensed under a Creative Commons Attribution 4.0 International License.

penerapannya hanya berhenti pada pengelompokan data historis [5]. Meskipun algoritma *K-Means* mampu mengelompokkan data dengan baik, validasi dan pemodelan lanjutan diperlukan untuk memastikan pola RFM yang terbentuk memiliki margin yang terukur secara matematis. Oleh karena itu, algoritma *Support Vector Machine* (SVM) diterapkan untuk mempelajari pola hasil klasterisasi tersebut menjadi pemodelan segmentatif yang solid [6].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menerapkan pemodelan RFM dan algoritma *K-Means Clustering* yang digabungkan dengan algoritma *Support Vector Machine* (SVM) pada data transaksi Klien Mitra CV Ekasa. Penelitian dilakukan dengan membagi data transaksi pelanggan menjadi beberapa kelompok segmen secara optimal menggunakan *K-Means*, yang hasilnya kemudian diolah sebagai data latih bagi model SVM. Hasil penelitian ini diimplementasikan ke dalam *dashboard web* statis sebagai basis pendukung keputusan. *Dashboard* ini mempermudah pihak manajemen dalam mengambil keputusan strategis, menargetkan promosi sesuai segmen pelanggan, dan meningkatkan efisiensi operasional tanpa perlu menganalisis data mentah secara manual.

II. METODE

Penelitian ini menggunakan metode *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai metodologi utama untuk menganalisis klasifikasi kelompok pelanggan berdasarkan model perilaku RFM (*Recency*, *Frequency*, *Monetary*) menggunakan algoritma *K-Means Clustering* dan *Support Vector Machine* (SVM) [7]. CRISP-DM dipilih karena menyediakan alur kerja yang sistematis, terstruktur, dan terukur di setiap tahapan proses *data mining* [8]. Dalam penerapannya, penelitian ini mengikuti enam fase standar CRISP-DM secara runtut yang mencakup *business understanding*, *data understanding*, *data preparation* (yang di dalamnya mencakup prapemrosesan dan normalisasi data), *modeling* (melalui implementasi *K-Means clustering* dan *SVM classification*), *evaluation* (menggunakan pengujian *Silhouette Score* dan matriks evaluasi SVM), serta diakhiri dengan fase *deployment* menggunakan *Streamlit*. Fase *deployment* diikutsertakan karena penelitian ini tidak hanya berfokus pada analisis teoritis, melainkan juga menerapkan hasil pemodelan ke dalam antarmuka web interaktif yang siap digunakan sebagai prototipe sistem pendukung keputusan operasional perusahaan. Alur keenam tahapan tersebut divisualisasikan pada Gambar 1.



Gambar 1. Alur Tahapan Metodologi Penelitian yang Mengacu pada Kerangka Kerja CRISP-DM, Dimulai dari Business Understanding hingga Deployment Sistem

Business Understanding

Fokus utama pada tahap ini adalah mendefinisikan tujuan dari sudut pandang manajerial perusahaan. Permasalahan utama yang diangkat adalah tidak ada sistem segmentasi loyalitas pelanggan yang terukur, sehingga mengaikbatkan inefisiensi pada strategi promosi. Pemetaan profil pelanggan ke dalam segmentasi yang terukur secara objektif sangat diperlukan untuk mengoptimalkan target pemasaran perusahaan [9]. Hasil klasifikasi ini akan menjadi basis utama dalam model pendukung keputusan.

Data Understanding

Fase ini bertujuan untuk mengidentifikasi dan memahami karakteristik awal dari data transaksi ritel. Variabel yang dieksplorasi mencakup tanggal transaksi, identitas pelanggan, dan nilai belanja. Melalui analisis eksplorasi data, diidentifikasi adanya nilai transaksi yang fluktuatif serta kelompok pengguna dengan transaksi yang sangat ekstrem (*outlier*) [10]. Karakteristik anomali pada fase awal ini menjadi indikator penting bahwa tahapan *data preparation* sangat diperlukan sebelum algoritma dilatih.

Data Preparation

Pada tahap prapemrosesan, data transaksi dibersihkan dari nilai anomali dan baris transaksi yang dibatalkan. Data pencilan (*outlier*) ekstrim seperti akun transaksi kasir internal (ID-211007-104810-0001) juga dihapus agar tidak merusak titik pusat klaster. Fitur RFM (*Recency*, *Frequency*, *Monetary*) kemudian dibentuk berdasarkan agregasi data pelanggan. Karena ketiga variabel ini memiliki rentang nilai yang sangat jauh berbeda dimana *Monetary* berada dalam skala jutaan, sedangkan *Frequency* dalam hitungan satuan atau puluhan maka dilakukan transformasi logaritma (*Log Transform*) untuk menormalkan kemiringan distribusi data (*skewness*). Setelah itu, data diskalakan secara menyeluruh menggunakan metode *StandardScaler* sehingga seluruh variabel terpusat pada skala yang sama. Langkah ini krusial agar perhitungan pada *K-Means* nantinya tidak didominasi oleh variabel *Monetary* [11].

Sebagai gambaran awal, tahap pra-pemrosesan ini menangani dataset transaksi mentah berukuran 287.512 baris dan 16 kolom, dengan rincian atribut sebagaimana dirangkum pada Tabel 1. Pengecekan nilai kosong (*missing value*) menunjukkan bahwa keempat variabel pembentuk fitur RFM pada Tabel 2 seluruhnya terisi lengkap sehingga tidak memerlukan proses imputasi. Sebaliknya, kolom cust_name kosong di seluruh baris data dan kolom sal_no memiliki 4.835 nilai hilang; namun karena keduanya tidak dilibatkan dalam pembentukan fitur RFM, kekosongan tersebut tidak berdampak pada hasil pemodelan.

Tabel 1. Variabel Dataset Transaksi Mentah

No	Nama Variabel	Kategori	Keterangan
1	jual_no	Identitas Transaksi	Nomor identitas transaksi penjualan
2	sal_no	Identitas Transaksi	Nomor identitas kasir/sales
3	let_no	Identitas Transaksi	Nomor identitas pencatatan transaksi
4	person_no	Identitas Pelanggan	Kode unik identitas pelanggan

This work is licensed under a Creative Commons Attribution 4.0 International License.

5	cust_name	Identitas Pelanggan	Nama pelanggan
6	jual_sub_total	Nilai Pembelian	Subtotal pembelian sebelum diskon dan pajak
7	jual_disc0_rp	Nilai Pembelian	Nilai diskon tahap pertama
8	jual_disc1_rp	Nilai Pembelian	Nilai diskon tahap kedua
9	jual_ppn_rp	Nilai Pembelian	Nilai pajak pertambahan nilai (PPN)
10	jual_total	Nilai Pembelian	Total nilai transaksi setelah diskon dan pajak
11	pay_cash	Metode Pembayaran	Nilai pembayaran tunai
12	pay_card	Metode Pembayaran	Nilai pembayaran kartu debit/kredit
13	pay_credit	Metode Pembayaran	Nilai pembayaran kredit toko
14	is_delete	Status Transaksi	Penanda status transaksi (1 = dibatalkan, 0 = valid)
15	jual_date	Waktu Transaksi	Tanggal dan waktu transaksi dilakukan

Tabel 2. Variabel yang Digunakan untuk Pembentukan Fitur RFM

No	Nama Variabel	Fitur RFM	Keterangan
1	jual_date	Recency	Digunakan untuk menghitung selisih waktu sejak transaksi terakhir pelanggan
2	jual_no	Frequency	Digunakan untuk menghitung jumlah transaksi yang dilakukan pelanggan
3	person_no	Kunci Agregasi	Digunakan sebagai kunci pengelompokan data transaksi per pelanggan
4	jual_total	Monetary	Digunakan untuk menghitung total nominal pembelian pelanggan

Proses eksklusi data dilakukan secara bertahap. Baris dengan status is_delete = 1 dibuang terlebih dahulu agar hanya transaksi valid yang diproses. Akun kasir internal yang telah disebutkan sebelumnya juga turut disingkirkan pada tahap ini, mengingat kontribusinya mencapai 251.446 baris atau sekitar 87,5% dari total data mentah—apabila dibiarkan, volume sebesar itu akan sangat mendominasi dan menggeser posisi titik pusat klaster. Selanjutnya, baris hasil agregasi dengan nilai *Monetary* ≤ 0 turut dikeluarkan karena tidak mencerminkan transaksi pembelian yang sah. Rangkaian penyaringan ini menyisakan 36.066 baris transaksi valid dari 342 pelanggan unik, yang setelah agregasi RFM dan penyaringan akhir menghasilkan 341 pelanggan sebagai unit analisis pada tahap pemodelan berikutnya.

Modeling

Pada tahap ini dijabarkan model yang dipakai dalam penelitian ini yaitu *K-Means Clustering* dan *Support Vector Machine* (SVM).

Algoritma *K-Means*

K-Means Clustering adalah algoritma *unsupervised learning* yang memecah kumpulan data pelanggan ke dalam *k* kelompok (*cluster*) berdasarkan tingkat kemiripan atribut RFM. Pada penelitian ini, penentuan jumlah klaster optimal (*k*) tidak ditentukan secara subjektif, melainkan dievaluasi murni secara objektif dan matematis menggunakan metrik *Silhouette Score* tertinggi. Hasil pelabelan klaster dari tahapan ini kemudian diekstraksi menjadi hierarki kelas pelanggan yang akan dipelajari oleh algoritma klasifikasi. Fungsi objektif dari algoritma *K-Means* dirumuskan sebagai berikut [12]:

$$J = \sum_{j=1}^{k} \quad \sum_{i=1}^{n} \quad ||x_i^{(j)} - c_j||^2 \quad (1)$$

Diketahui:
J : Fungsi objektif *K-Means* (total jarak kuadrat)
k : Jumlah klaster
n : Jumlah titik data dalam klaster
$x_i^{(j)}$: Titik data ke-i pada klaster ke - j
c_j : Titik pusat (*centroid*) dari klaster ke - j

Secara sederhana, model *K-Means* pada penelitian ini dibangun menggunakan *library scikit-learn*. Agar hasil pengelompokan tidak asal-asalan, algoritma ini dijalankan berulang sebanyak 10 kali (*n_init* = 10) dengan titik awal (*centroid*) yang berbeda-beda secara acak pada setiap percobaan, sehingga hasilnya tidak terjebak pada titik pengelompokan yang kurang optimal (*local minimum*). Batas maksimal proses perhitungan (*max_iter*) mengikuti pengaturan bawaan *library*, yaitu 300 kali putaran, sedangkan *random_state* dikunci pada angka 42 agar hasilnya selalu sama setiap dijalankan ulang. Untuk menentukan jumlah kelompok (*k*) yang paling sesuai, penelitian ini mencoba beberapa pilihan mulai dari *k* = 1 sampai *k* = 10, lalu memilih nilai *k* yang memberikan skor *Silhouette Score* paling tinggi sebagai jumlah klaster yang paling optimal.

***Support Vector Machine* (SVM)**

Pemodelan menggunakan algoritma *Support Vector Machine* (SVM) bertujuan agar sistem bisa belajar dan menebak kelompok pelanggan secara otomatis. Pada penelitian ini, algoritma SVM bekerja dengan cara membentuk sebuah batas pemisah untuk mengelompokkan data pelanggan ke dalam segmen yang tepat [13]. Cara ini dipilih karena data perilaku pelanggan sebelumnya sudah dirapikan (*StandardScaler*), sehingga jarak antar kelompoknya sudah terlihat sangat jelas. Untuk mengukur seberapa pintar sistem ini dalam menebak, dilakukan pengujian (evaluasi) menggunakan perhitungan akurasi dengan membagi porsi data uji sebesar 20% dan 30%. [14]
Secara matematis, fungsi batas pemisah pada SVM dirumuskan sebagai berikut:

$$f(x) = w \cdot x + b = 0 \quad (2)$$

Keterangan:
$f(x)$: Fungsi batas pemisah kelompok pelanggan.
w : Bobot yang menentukan kemiringan batas pemisah.
x : Titik data (*Recency*, *Frequency*, *Monetary*).

This work is licensed under a Creative Commons Attribution 4.0 International License.

b : Nilai *bias* agar batas pemisah berada tepat di tengah antar-kelompok.

Model SVM pada penelitian ini dibangun dengan jenis *kernel linear*, yaitu batas pemisah antar kelompok pelanggan berbentuk garis lurus, dengan parameter *C* sebesar 1,0 mengikuti pengaturan bawaan *library scikit-learn*, dan *random_state* = 42. Pengaturan ini digunakan karena data pelanggan sudah dirapikan lebih dulu sehingga kelompok *Tier* Pasif dan *Tier* Aktif sudah cukup mudah dipisahkan tanpa perlu penyesuaian tambahan. Oleh karena itu, penelitian ini tidak melakukan uji coba parameter lain, seperti pencarian nilai *C* terbaik (*grid search*) maupun mencoba *kernel* berbentuk *non-linear*. Keterbatasan ini akan dibahas lebih lanjut pada bagian Hasil dan Pembahasan.

Evaluation

Tahap evaluasi pada penelitian ini menggunakan dua jenis pengukuran, yaitu evaluasi kualitas klasterisasi dan evaluasi performa klasifikasi.

Silhouette Score

Silhouette Score digunakan untuk mengevaluasi kualitas klaster *K-Means* dengan mengukur seberapa dekat suatu titik data dengan klasternya sendiri dibandingkan dengan klaster lain [15].

$$s(i) = \frac{b(i)-a(i)}{(a(i),b(i))} \quad (3)$$

Diketahui:

$s(i)$: Nilai *Silhouette* untuk titik data ke-i

$a(i)$: Rata-rata jarak titik data ke-i terhadap titik lain di klaster yang sama

$b(i)$: Rata-rata jarak terdekat titik data ke-i terhadap klaster tetangga

Confussion Matrix

Kinerja model prediksi SVM diukur menggunakan *Confusion Matrix* yang menghasilkan empat metrik evaluasi utama, yaitu *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. *Accuracy* mengukur persentase total tebakan yang benar secara keseluruhan. *Precision* mengukur ketepatan prediksi kelas positif, *Recall* mengukur sensitivitas model mengenali kelas aktual, dan *F1-Score* memberikan rata-rata harmonik untuk memastikan stabilitas prediksi pada data yang tidak seimbang.

Deployment

Pada tahap ini, *output* dari model SVM diekspor ke dalam format CSV, lalu divisualisasikan menjadi *Dashboard* interaktif berbasis *Streamlit*. Hasil visualisasi ini akan menjadi acuan bagi Klien Mitra CV Ekasa untuk menganalisis profil pelanggan dan menentukan target promosi yang sesuai untuk masing-masing Klaster.

III. HASIL DAN PEMBAHASAN

Analisis dan Prapemrosesan Data RFM

Penelitian ini menggunakan dataset riwayat transaksi penjualan dari Klien Mitra CV Ekasa. Tahap pertama pemodelan adalah mengekstraksi data transaksi mentah menjadi tiga metrik perilaku pelanggan, yakni *Recency* (interval hari sejak transaksi terakhir pelanggan hingga tanggal analisis), *Frequency* (total kemunculan atau jumlah transaksi yang dilakukan), dan *Monetary* (total nominal pembelanjaan dalam rupiah). Transformasi data ini menghasilkan variabel RFM yang menjadi fondasi utama dalam mengukur tingkat profitabilitas setiap pelanggan. Contoh hasil ekstraksi RFM untuk lima pelanggan disajikan pada Tabel 3.

Tabel 3. Sampel Data RFM Awal

Person_no	Recency (hari)	Frequency (kali)	Monetary (Rp)
ID-120131-194130-0001	37	341	1.254214
ID-120405-172133-0001	5	337	8.188988
ID-120411-084514-0001	56	1373	5.792668
ID-150622-082111-0001	353	21	1.043660
ID-160427-091223-0001	29	48	3.365004

Sebagaimana terlihat pada Tabel 3, rentang nilai ketiga variabel RFM berbeda cukup jauh antar pelanggan, dengan Monetary bernilai jutaan rupiah sementara Frequency hanya berkisar puluhan hingga ratusan kali transaksi. Mengingat ketiga variabel RFM memiliki rentang nilai yang berbeda dan cenderung memiliki distribusi yang tidak normal, dilakukan *log transformation* pada seluruh variabel untuk mengurangi tingkat kemiringan data (*skewness*). Setelah itu dilakukan proses standarisasi menggunakan metode *StandardScaler* sehingga setiap variabel memiliki rata-rata mendekati nol dan deviasi standar sebesar satu. Proses ini bertujuan agar perhitungan jarak *Euclidean* pada algoritma *K-Means* tidak didominasi oleh variabel *Monetary* yang memiliki rentang nilai jauh lebih besar dibandingkan variabel lainnya. Hasil transformasi untuk lima sampel pelanggan yang sama ditunjukkan pada Tabel 4.

Tabel 4. Sampel Hasil Transformasi Logaritmik dan Standardisasi Data RFM

person_no	Recency (Scaled)	Frequency (Scaled)	Monetary (Scaled)
ID-120131-194130-0001	-0.368	1.454	1.309
ID-120405-172133-0001	-1.346	1.447	1.076
ID-120411-084514-0001	-0.154	2.334	2.144
ID-150622-082111-0001	0.814	-0.282	-0.048
ID-160427-091223-0001	-0.493	0.225	0.591

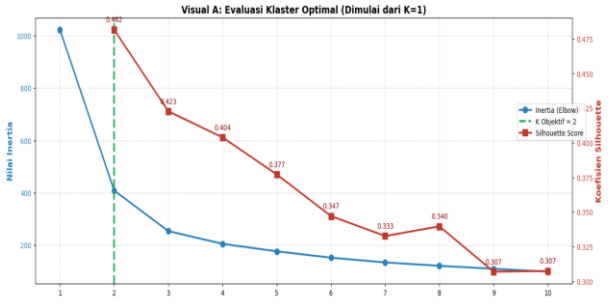
Tabel 4 memperlihatkan bahwa setelah proses transformasi logaritmik dan standardisasi, nilai ketiga variabel RFM berada pada skala yang seimbang dengan rata-rata mendekati nol, sehingga siap digunakan sebagai input algoritma K-Means tanpa didominasi oleh variabel bernilai besar seperti Monetary.

Hasil Klasterisasi *K-Means* dan Pembentukan *Klaster*

Proses segmentasi pelanggan dilakukan menggunakan algoritma *K-Means Clustering* terhadap data RFM yang telah melalui tahapan transformasi logaritmik dan

This work is licensed under a Creative Commons Attribution 4.0 International License.

standarisasi. Untuk menentukan jumlah klaster yang paling optimal, dilakukan pengujian menggunakan metode *Silhouette Score* pada beberapa variasi nilai k, sebagaimana ditampilkan pada Gambar 2.



Gambar 2. Grafik Silhouette Score untuk Setiap Nilai k (Sumbu-x: Jumlah Klaster k = 1 sampai 10; Sumbu-y: Nilai Silhouette Score), dengan Skor Tertinggi pada k = 2

Berdasarkan hasil evaluasi, nilai *Silhouette Score* tertinggi diperoleh pada jumlah klaster sebanyak dua kelompok (k = 2). Hasil tersebut menunjukkan bahwa pembagian pelanggan ke dalam dua kelompok memberikan tingkat pemisahan antar klaster yang paling baik dibandingkan jumlah klaster lainnya. Oleh karena itu, penelitian ini menggunakan dua klaster sebagai hasil segmentasi akhir pelanggan.

Tabel 5. Profil Rata-Rata Klaster

Klaster	Recency (hari)	Frequency (kali)	Monetary (Rp)
Pasif	412.81	12.25	5.127.666
Aktif	52.25	200.91	111.534.800

Tabel 6. Sampel Pelabelan Cluster

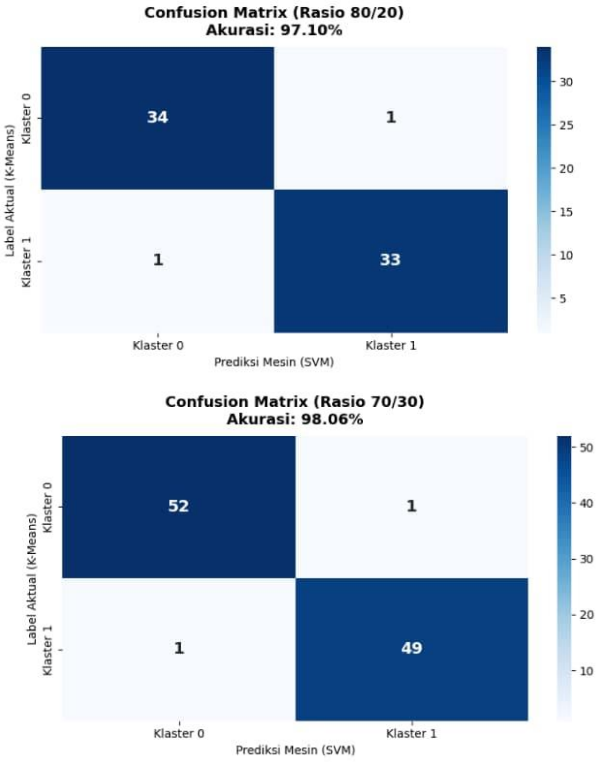
Person_no	Recency (hari)	Frequency (kali)	Monetary (Rp)	*Tier*
ID-120131-194130-0001	37	341	1.254214	*Tier* Aktif
ID-120405-172133-0001	5	337	8.188988	*Tier* Aktif
ID-120411-084514-0001	56	1373	5.792668	*Tier* Aktif
ID-150622-082111-0001	353	21	1.043660	*Tier* Pasif
ID-160427-091223-0001	29	48	3.365004	*Tier* Aktif

Profil rata-rata tiap klaster ditunjukkan pada Tabel 5, sedangkan Tabel 6 menyajikan contoh hasil pelabelan Tier untuk lima pelanggan berdasarkan nilai RFM masing-masing. Berdasarkan hasil klasterisasi pada Tabel 5, Klaster 0 dikategorikan sebagai *Tier* Pasif dan Klaster 1 dikategorikan sebagai *Tier* Aktif. Berdasarkan karakteristik rata-rata masing-masing klaster, *Tier* Pasif memiliki nilai *Recency* yang relatif tinggi, menunjukkan bahwa sebagian besar pelanggan pada kelompok ini sudah lama tidak melakukan transaksi. Selain itu, nilai *Frequency* dan *Monetary* yang rendah mengindikasikan bahwa pelanggan dalam kelompok ini memiliki aktivitas transaksi yang relatif sedikit serta kontribusi pendapatan yang rendah terhadap perusahaan.

Sebaliknya, *Tier* Aktif menunjukkan nilai *Recency* yang jauh lebih rendah, yang berarti pelanggan masih aktif melakukan transaksi dalam periode yang relatif dekat dengan tanggal analisis. Nilai *Frequency* dan *Monetary* yang tinggi menunjukkan bahwa kelompok ini merupakan pelanggan aktif dengan kontribusi transaksi yang lebih besar. Dengan demikian, *Tier* Aktif dapat dikategorikan sebagai kelompok pelanggan bernilai tinggi yang berpotensi menjadi fokus utama dalam strategi retensi pelanggan.

Evaluasi Kinerja Klasifikasi *Support Vector Machine* (SVM)

Pelabelan klaster dari algoritma *K-Means* ($k = 2$) diubah menjadi label target aktual untuk melatih model klasifikasi *Support Vector Machine* (SVM). Pada penelitian ini, algoritma SVM dilatih menggunakan fungsi *kernel linear* berdasarkan persamaan matematis $f(x) = w \cdot x + b = 0$. Penggunaan batas pemisah garis lurus (*linear*) dipilih karena data fitur RFM sebelumnya telah dinormalisasi menggunakan *StandardScaler*, sehingga distribusi data sudah terpusat dan batas pemisah antar kelas pelanggan dapat ditarik secara optimal. Untuk menguji keandalan model, dataset dibagi menjadi data latih (*training set*) dan data uji (*testing set*) melalui dua skenario data splitting, yakni rasio 80:20 dan 70:30. Hasil confusion matrix dari kedua skenario tersebut ditampilkan pada Gambar 3.



Gambar 3. Confusion Matrix Hasil Klasifikasi SVM untuk Tier Pasif dan Tier Aktif: (a) Skenario Data Splitting 80:20; (b) Skenario Data Splitting 70:30

Berdasarkan hasil pengujian, performa model klasifikasi SVM menunjukkan tingkat kemampuan prediksi yang

This work is licensed under a Creative Commons Attribution 4.0 International License.

sangat presisi. Pada skenario pengujian pertama dengan porsi data uji 20% (rasio 80:20), model berhasil memprediksi kelompok loyalitas pelanggan dengan tingkat akurasi global mencapai 97,10%. Lebih lanjut, ketika skenario porsi data uji diperbesar menjadi 30% (rasio 70:30) guna menguji ketangguhan algoritma terhadap data yang belum pernah dilihat sebelumnya, tingkat akurasi model tetap stabil dan justru mengalami sedikit peningkatan di angka 98,06%..

Tabel 7. Classification Report Hasil Klasifikasi SVM

Tier (Skenario 80:20)	**Precision**	**Recall**	**F1-Score**	**Support**
Tier Pasif	0.97	0.97	0.97	35
Tier Aktif	0.97	0.97	0.97	34
Accuracy			0.97	69
Macro avg	0.97	0.97	0.97	69
Weighted avg	0.97	0.97	0.97	69
Tier (Skenario 70:30)	**Precision**	**Recall**	**F1-Score**	**Support**
Tier Pasif	0.98	0.98	0.98	53
Tier Aktif	0.98	0.98	0.98	50
Accuracy			0.98	103
Macro avg	0.98	0.98	0.98	103
Weighted avg	0.98	0.98	0.98	103

Evaluasi melalui *Classification Report* (Tabel 7) memperlihatkan bahwa nilai *Precision*, *Recall*, dan *F1-Score* rata-rata untuk setiap klaster selalu berada di atas angka 0,97. Tingginya skor metrik evaluasi pada kedua skenario pengujian tersebut menegaskan bahwa pemodelan klasifikasi ini sangat efektif. Sistem mampu mempelajari pola pemisahan karakteristik antara pelanggan di *Tier* Pasif dan *Tier* Aktif secara akurat tanpa mengalami indikasi *overfitting*. Hasil prediksi inilah yang selanjutnya diimplementasikan sebagai dasar sistem *dashboard* interaktif.

Perbedaan akurasi antara kedua skenario pengujian tergolong tipis, hanya sekitar 0,96 poin persentase, sementara *precision*, *recall*, dan *F1-Score* tetap konsisten di atas 0,97 untuk kedua *Tier*—indikasi bahwa model tidak mengalami *overfitting*, meskipun pengujian baru dilakukan satu kali untuk masing-masing rasio pembagian data. Dibandingkan dengan penelitian sejenis pada segmentasi pelanggan distributor fashion yang mencatat *Silhouette Score* hingga 0,915, angka 0,482 yang diperoleh pada penelitian ini tergolong lebih rendah, kemungkinan dipengaruhi oleh karakteristik pelanggan Klien Mitra CV Ekasa yang lebih beragam serta nilai *Frequency* dan *Monetary* yang masih menunjukkan sebaran ekstrem walau telah dinormalisasi. Meski demikian, kondisi tersebut tidak menurunkan performa klasifikasi SVM, yang tetap stabil di atas 97% pada kedua skenario.

Perlu dicatat bahwa penelitian ini belum melibatkan *cross-validation* berlapis ataupun penyetelan *hyperparameter* SVM, sehingga angka performa yang dilaporkan mencerminkan satu konfigurasi model saja. Dari sisi penerapan, hasil segmentasi ini tetap dapat menjadi acuan bagi Klien Mitra CV Ekasa dalam menyusun kampanye retensi untuk *Tier* Aktif maupun kampanye reaktivasi untuk *Tier* Pasif, dengan tetap memperhatikan kemungkinan kesalahan klasifikasi pada pelanggan yang berada di sekitar garis batas keputusan antar *Tier*.

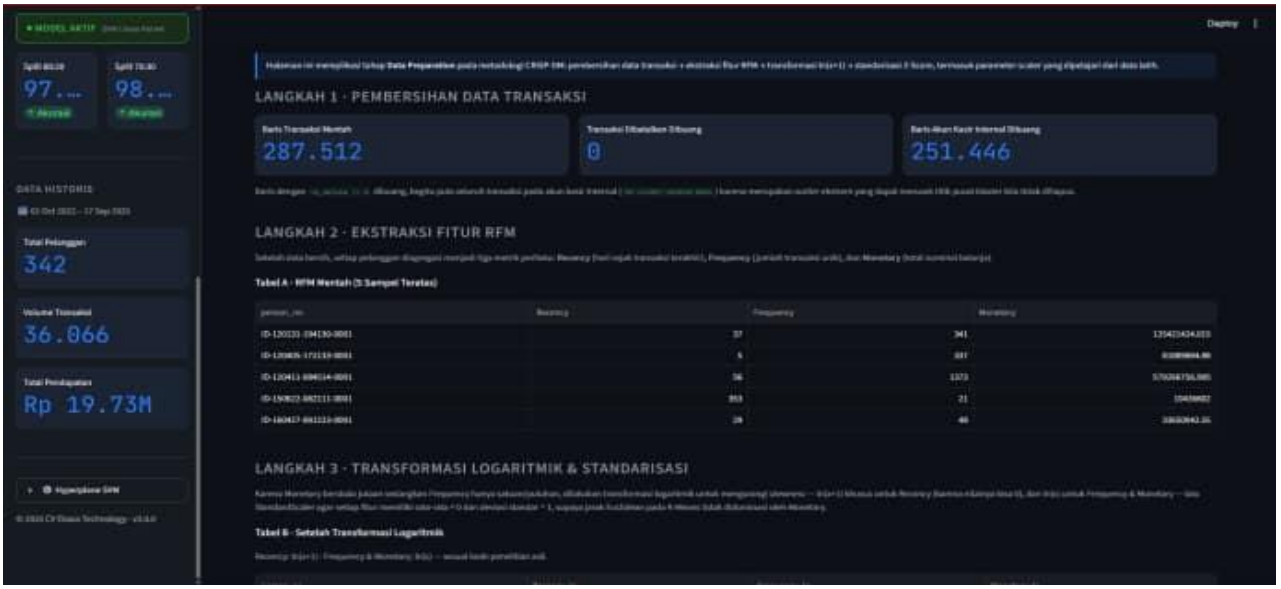
Implementasi *Dashboard* Interaktif *Streamlit*

Implementasi akhir penelitian ini berupa visualisasi *Dashboard* menggunakan *Streamlit*. Untuk memastikan kinerja web tetap efisien, data prediksi SVM diekspor sebagai file statis (CSV). Dashboard web ini merangkum distribusi pelanggan per klaster secara visual. *Dashboard* ini sebagai panduan bagi Klien Mitra CV Ekasa untuk mengatur strategi pemasaran berbasis data. Tampilan setiap halaman dashboard secara berurutan ditunjukkan pada Gambar 4 sampai dengan Gambar 9.

Gambar 4. Halaman Beranda Dashboard Streamlit yang Menyajikan Ringkasan Statistik Pelanggan beserta Diagram Donat Komposisi Tier Pasif dan Aktif

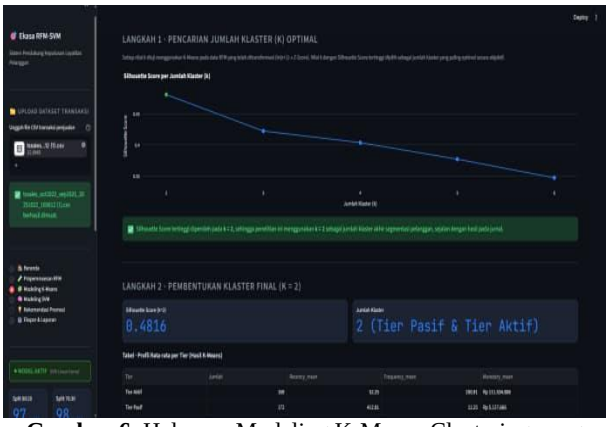
This work is licensed under a Creative Commons Attribution 4.0 International License.

Halaman Beranda menampilkan form *input* dataset yang apabila diinput akan menampilkan ringkasan statistik utama dataset (total pelanggan, volume transaksi, dan pendapatan kotor) beserta komposisi proporsi pelanggan pada *Tier* Pasif dan *Tier* Aktif dalam bentuk diagram donat.



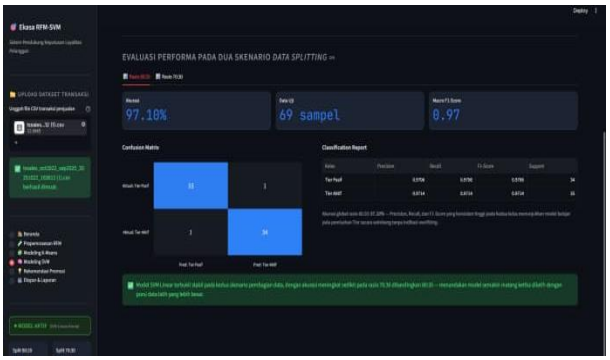
Gambar 5. Halaman Pra-pemrosesan RFM yang Memperlihatkan Tahapan Pembersihan Data, Ekstraksi Fitur RFM, hingga Transformasi Logaritmik dan Standardisasi

Halaman pra-pemrosesan RFM berisi 3 langkah yaitu pembersihan data transaksi, ekstraksi fitur *Recency, Frequency, dan Monetary* lalu transformasi logaritmik dan standarisasi sebelum di proses ke langkah selanjutnya untuk memprediksi klaster loyalitas pelanggan secara *real-time*, lengkap dengan *K-Means Clustering*.



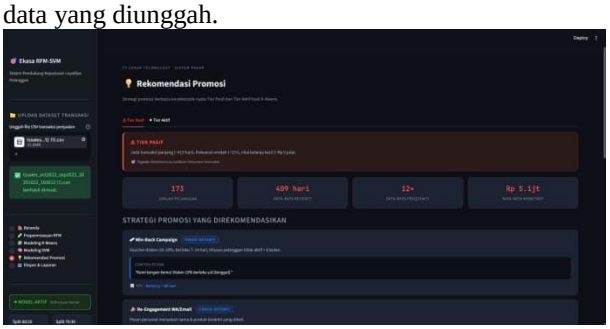
Gambar 6. Halaman Modeling K-Means Clustering yang Menyajikan Grafik Silhouette Score, Profil Rata-Rata Setiap Tier, dan Visualisasi Sebaran Pelanggan Berdasarkan Klaster

Halaman *Modeling K-Means Clustering* menampilkan pencarian jumlah klaster optimal melalui evaluasi *Silhouette Score* pada beberapa variasi *k*, profil rata-rata setiap *Tier*, sampel pelabelan klaster, serta visualisasi sebaran pelanggan berdasarkan hasil *K-Means*. Label *Tier* yang dihasilkan selanjutnya menjadi target aktual untuk pelatihan model SVM.



Gambar 7. Halaman Modeling SVM yang Menampilkan Formula Hyperplane beserta Confusion Matrix dan Classification Report pada Kedua Skenario Pembagian Data

Halaman *Modeling SVM* menampilkan Klasifikasi SVM *Linear* menggunakan label hasil *K-Means* sebagai target yang mencakup tampilan Rumus perhitungan *HyperPlane*, dan Evaluasi Performa pada dua data *Splitting* yang meliputi akurasi, *Confussion Matrix* serta *Classification Report* yang dihitung secara real-time dari data yang diunggah.

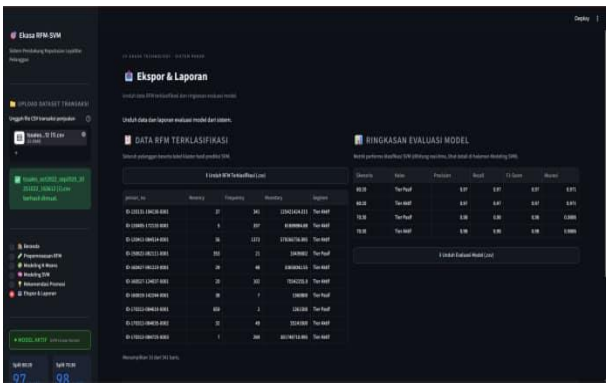


Gambar 8. Halaman Rekomendasi Promosi yang Menyajikan Strategi Pemasaran, Contoh Pesan, Indikator Keberhasilan (KPI), dan Prioritas Eksekusi untuk Setiap Tier

Halaman Rekomendasi Promosi menampilkan strategi promosi yang disesuaikan dengan karakteristik nyata

This work is licensed under a Creative Commons Attribution 4.0 International License.

masing-masing *tier*, lengkap dengan contoh pesan pemasaran, indikator keberhasilan (KPI), dan prioritas eksekusi untuk setiap strategi yang diusulkan



Gambar 9. Halaman Ekspor dan Laporan yang Menyediakan Fitur Unduh Data RFM Terklasifikasi beserta Ringkasan Metrik Evaluasi SVM per Tier

Halaman Ekspor & Laporan memungkinkan pengguna mengunduh data RFM terklasifikasi seluruh pelanggan beserta label Tier dalam format CSV, serta ringkasan metrik evaluasi model SVM (*Precision*, *Recall*, *F1-Score*, dan Akurasi) untuk kedua skenario pembagian data, serta statistik ringkas per *Tier*.

IV. KESIMPULAN

Penelitian ini berhasil mengintegrasikan pendekatan RFM, *K-Means Clustering*, dan *SVM* untuk mengklasifikasikan *Tier* loyalitas pelanggan berdasarkan data transaksi Klien Mitra CV Ekasa. Berdasarkan evaluasi *Silhouette Score*, diperoleh dua segmen paling optimal (k = 2, skor 0,482) yang terdiri atas *Tier* Pasif dan *Tier* Aktif, dan kedua label ini kemudian digunakan sebagai target dalam pelatihan model SVM berkernel *linear*. Model klasifikasi yang dihasilkan mampu mencapai akurasi sebesar 97,10% pada skema pembagian data 80:20 serta 98,06% pada skema 70:30, dengan nilai *precision*, *recall*, dan *F1-Score* yang tetap berada di atas 0,97 untuk kedua *Tier* pada kedua skenario tersebut, sehingga model dapat dikatakan bekerja secara stabil dan dapat diandalkan. Alur kerja terintegrasi ini turut tervalidasi pada data transaksi ritel yang nyata, dan hasilnya diimplementasikan dalam bentuk *dashboard* interaktif berbasis *Streamlit* yang dapat langsung digunakan oleh Klien Mitra CV Ekasa untuk menyusun strategi retensi bagi *Tier* Aktif maupun reaktivasi bagi *Tier* Pasif, tanpa perlu mengolah data mentah secara manual.

Penelitian ini masih memiliki keterbatasan, yaitu belum menerapkan *cross-validation* berlapis dan belum melakukan penyetelan *hyperparameter* pada model SVM. Akibatnya, performa yang dilaporkan hanya berasal dari satu konfigurasi model dan belum tentu menunjukkan performa terbaik yang bisa dicapai. Beberapa saran untuk penelitian selanjutnya adalah sebagai berikut. Pertama, model perlu diuji pada periode transaksi yang berbeda untuk melihat apakah hasil segmentasi tetap konsisten dari waktu ke waktu. Kedua, fitur demografis pelanggan, seperti usia, lokasi, atau kategori produk, dapat ditambahkan agar profil pelanggan lebih lengkap. Ketiga, *cross-validation* berlapis dan pencarian *hyperparameter* SVM perlu diterapkan untuk menguji apakah performa model dapat ditingkatkan.

V. REFERENSI

[1] H. Hairani, R. Wahyudi, G. Y. Pratama, and M. K. Ihsan, "Application of RFM Model and K-Means Algorithm in Customer Loyalty Segmentation," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 5, no. 1, pp. 1–8, Mar. 2026, doi: 10.30812/ijecsa.v5i1.6087.

[2] Nugraha, Thoib, Sururi, Bayu F, and Candra, "Segmentasi Pelanggan Berbasis Analisis RFM Menggunakan Algoritma K-Means Clustering," 2025.

[3] S. Sharyanto and D. Lestari, "Penerapan data mining untuk menentukan segmentasi pelanggan dengan menggunakan algoritma K-Means dan model RFM pada e-commerce," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, pp. 866–871, 2022, doi: 10.30865/jurikom.v9i4.4525.

[4] R. B. Ardi, F. E. Nastiti, and S. Sumarlinda, "Algoritma K-Means Clustering Untuk Segmentasi Pelanggan (Studi Kasus: Fashion Viral Solo)," *INFOTECH journal*, vol. 9, no. 1, pp. 124–131, 2023, doi: 10.31949/infotech.v9i1.5214.

[5] Syahra, Fadlil, and Yuliansyah, "Customer Segmentation Using RFM and K-Means Clustering to Support CRM in Retail Industry," 2025.

[6] M. D. Andini *et al.*, "Analisis Segmentasi Pelanggan Menggunakan RFM dan K-Means Clustering sebagai Dasar Penyusunan Aturan Pendukung Keputusan," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 4, pp. 2752–2760, Mar. 2026, doi: 10.47065/bits.v7i4.9511.

[7] A. M. M. Fattah, A. Voutama, N. Heryana, and N. Sulistiyowati, "Pengembangan Model Machine Learning Regresi sebagai Web Service untuk Prediksi Harga Pembelian Mobil dengan Metode CRISP-DM," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 5, pp. 1669–1678, Oct. 2022, doi: 10.30865/jurikom.v9i5.5021.

[8] F. Ranja, A. H. Ginting, I. S. Putra, and R. Bukit, "Customer Segmentation of E-Wallet Top-Up Users Based on RFM and K-Means Clustering," *JUKTISI*, vol. 4, no. 3, Aug. 2025, doi: 10.62712/juktisi.v4i3.887.

[9] F. Nazihah, A. Danniswara, and A. Wibowo, "Analisis Segmentasi Pelanggan dengan Algoritma K-Means pada Data Penjualan," *Jurnal Algoritma*, vol. 22, no. 2, Nov. 2025, doi: 10.33364/algoritma/v.22-2.2489.

[10] P. R. Sihombing, S. Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," *Jurnal Ekonomi dan Statistik Indonesia*, vol. 2, no. 3, pp. 307–316, 2022, doi: 10.11594/jesi.02.03.07.

[11] M. A. Irwansyah, A. Meiriza, and D. Lestarini, "Analisis Klasterisasi Kualitas Internet Seluler Menggunakan Metode K-Means dan Gaussian Mixture Model," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 3, p. 1694, Dec. 2025, doi: 10.47065/bits.v7i3.8615.

[12] B. T. Kristanti, A. Junaidi, and E. P. Mandyartha, "Implementasi K-Means Clustering Dalam Segmentasi Pelanggan Berdasarkan Usia, Pendapatan, dan Model RFM (Studi Kasus: Lantikya Store Jombang)," *Jurnal Informatika dan Teknik Elektro*

This work is licensed under a **Creative Commons Attribution 4.0 International License.**

Terapan, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4677.

[13] M. Y. Smaili and H. Hachimi, "New RFM-D Classification Model for Improving Customer Analysis and Response Prediction," *Ain Shams Engineering Journal*, vol. 14, no. 12, p. 102254, 2023, doi: 10.1016/j.asej.2023.102254.

[14] F. R. Lumbanraja, E. C. L. Gaol, D. A. Shofiana, and A. Junaidi, "Implementasi SMOTE dan Support Vector Machine Pada Klasifikasi Data Tidak Seimbang Metilasi Arginin," *Jurnal Pepadun*, vol. 5, no. 1, pp. 27–37, Apr. 2024, doi: 10.23960/pepadun.v5i1.209.

[15] B. N. Yulisasih, H. Herman, S. Sunardi, and H. Yuliansyah, "Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation," *Ilk. J. Ilm.*, vol. 16, no. 3, pp. 330–342, Dec. 2024, doi: 10.33096/ilkom.v16i3.2325.330-342.

This work is licensed under a **Creative Commons Attribution 4.0 International License.**