JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

Analisis Segmentasi dan Prediksi Pola Pembelian IC Label Gamis Menggunakan Hybrid K-Means Random Forest

Chandra Ayu Fatikasari[1*], Alifa Marsha Rahmania[2], Lu'lu'il Laili[3], Regizka Ayu Mega Saputri [4], Muhammad Arifin[5]

[1,2,3,4,5,] Fakultas Teknik, Sistem Informasi, Universitas Muria Kudus, Kudus, Indonesia

Email: [1*]202353080@std.umk.ac.id, [2]202353056@std.umk.ac.id, [3]202353057@std.umk.ac.id, [4]202353087@std.umk.ac.id, [5]arifin.m@umk.ac.id
(*Email Corresponding Author: 202353080@std.umk.ac.id)
Received: April 30, 2026 | Revision: May 2, 2026 | Accepted: May 8, 2026

Abstrak

Industri *fashion* muslim, khususnya busana gamis, menghadapi tantangan dalam memahami pola pembelian konsumen yang bersifat fluktuatif, sehingga kerap menimbulkan permasalahan *overstock* dan *stockout*. Penelitian ini bertujuan untuk menganalisis segmentasi pelanggan dan memprediksi pola pembelian gamis pada toko IC Label melalui pendekatan hibrida yang mengintegrasikan algoritma *K-Means Clustering* dan *Random Forest Classification*. Data yang digunakan merupakan 1.000 transaksi penjualan dari platform *e-commerce* Shopee pada periode Januari 2025 hingga Maret 2026. Metodologi penelitian mengacu pada kerangka kerja CRISP-DM, yang meliputi eksplorasi data, *preprocessing*, ekstraksi fitur RFM, pemodelan, serta evaluasi. Hasil penerapan *K-Means* dengan K=4 menghasilkan empat segmen pelanggan yang terdistribusi secara merata, yaitu: Pembeli Premium Wilayah Jawa (26,1%), Pembeli Sumatera Sensitif Diskon (24,9%), Pembeli Digital Wilayah Kepulauan (23,3%), dan Pembeli Mitra Wilayah Sumatera Bagian Selatan (25,7%). Label klaster hasil segmentasi selanjutnya digunakan sebagai variabel target pada model *Random Forest*, yang menghasilkan akurasi 96,38% pada data latih dan 68,00% pada data uji dengan konsistensi *cross-validation* 5-fold sebesar 68,70% (±3,50%). Analisis *feature importance* mengidentifikasi variabel provinsi (40,5%) dan jasa pengiriman (26,8%) sebagai faktor paling dominan dalam menentukan segmen pembeli. Simulasi prediksi pada pembeli baru membuktikan kemampuan model dalam mengklasifikasikan segmen secara *real-time*. Penelitian ini berkontribusi secara praktis dalam mendukung strategi pemasaran tersegmentasi, optimasi manajemen inventori, dan penguatan loyalitas pelanggan berbasis data pada industri *fashion* muslim.

Kata Kunci: K-Means Clustering, Random Forest, Segmentasi Pelanggan, Prediksi Pola Pembelian, Fashion Muslim

Abstract

The Muslim fashion industry, particularly the gamis (Islamic dress) segment, faces challenges in understanding fluctuating consumer purchasing patterns, which often leads to overstock and stockout issues. This study aims to analyze customer segmentation and predict gamis purchasing patterns at IC Label store through a hybrid approach integrating K-Means Clustering and Random Forest Classification algorithms. The dataset comprises 1,000 sales transactions from the Shopee e-commerce platform spanning January 2025 to March 2026. The research methodology follows the CRISP-DM framework, encompassing data exploration, preprocessing, RFM feature extraction, modeling, and evaluation. Application of K-Means with K=4 yielded four evenly distributed customer segments: Premium Buyers in Java Region (26.1%), Discount-Sensitive Buyers in Sumatra (24.9%), Digital Buyers in Island Regions (23.3%), and Partner-Transaction Buyers in Southern Sumatra (25.7%). Cluster labels from segmentation were subsequently used as target variables in the Random Forest model, achieving an accuracy of 96.38% on training data and 68.00% on test data, with 5-fold cross-validation consistency of 68.70% (±3.50%). Feature importance analysis identified province (40.5%) and shipping service (26.8%) as the most dominant factors in determining buyer segments. Prediction simulations on new buyers demonstrated the model's capability to classify segments in real-time. This research practically contributes to supporting segmented marketing strategies, inventory management optimization, and data-driven customer loyalty enhancement in the Muslim fashion industry.

Keywords; *K-Means Clustering, Random Forest, Customer Segmentation, Purchase Pattern Prediction, Muslim Fashion*

1. PENDAHULUAN

Sektor busana muslim, terutama kategori gamis, menunjukkan peningkatan tren yang sangat pesat belakangan ini. Dinamika ini didorong oleh penguatan kesadaran konsumen terhadap gaya busana syar'i yang modern, yang pada akhirnya menciptakan ekosistem persaingan bisnis yang semakin ketat. Kondisi pasar yang kompetitif tersebut menuntut para pelaku industri untuk lebih adaptif dalam merespons selera pasar yang terus berubah. Dalam ekosistem bisnis yang ketat, pemahaman terhadap perilaku konsumen menjadi determinan utama keberhasilan operasional [1]. Namun, entitas bisnis seperti IC Label kerap menghadapi kendala dalam mengidentifikasi pola konsumsi pelanggan yang bersifat fluktuatif [2]. Ketidakmampuan dalam memproyeksikan permintaan pasar secara presisi berdampak pada permasalahan manajerial klasik, yakni anomali stok berupa kelebihan persediaan (overstock) pada item tertentu dan kelangkaan stok (stockout) pada produk unggulan [3]. Oleh karena itu, diperlukan implementasi teknologi informasi yang mampu mengekstraksi data transaksi historis menjadi wawasan strategis melalui segmentasi pelanggan dan estimasi pembelian masa depan.

Data mining merupakan instrumen yang terbukti efektif dalam mengeksplorasi pola tersembunyi pada basis data pelanggan . Dalam domain pemasaran, algoritma *K-Means Clustering* sering diadopsi untuk mengelompokkan pelanggan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

berdasarkan kemiripan perilaku. Studi relevan pada unit bisnis Hijab Miulan mengonfirmasi bahwa *K-Means* mampu melakukan segmentasi pelanggan melalui parameter *Recency, Frequency,* dan *Monetary* (RFM) sebagai landasan strategi *Customer Relationship Management* (CRM) yang personal [4]. Melalui pendekatan ini, pelaku bisnis dapat mengidentifikasi kategori pelanggan potensial seperti "*superstar customer*" maupun kelompok yang berisiko meninggalkan layanan (*dormant customer*) [5]. Kendati demikian, segmentasi saja belum memadai tanpa dukungan model prediktif yang mampu mengestimasi perilaku konsumsi di masa mendatang.

Sejumlah penelitian terkini telah membuktikan relevansi pendekatan berbasis *data mining* dalam konteks ritel. Studi mengenai tren perilaku pelanggan [6] mengonfirmasi bahwa penerapan teknik *data mining* memiliki kemampuan untuk memetakan pola belanja sekaligus memproyeksikan dinamika pasar di platform *e-commerce* dengan tingkat presisi yang tinggi. Penggunaan metode ini memberikan dukungan signifikan bagi manajemen dalam merumuskan strategi promosi serta mengoptimalkan kontrol persediaan barang secara lebih efisien. Sementara itu, terdapat pembuktian bahwa integrasi metode klasifikasi dan *clustering* memberikan wawasan strategis yang komprehensif untuk pengambilan keputusan bisnis berbasis data [7]. Di sisi segmentasi pelanggan yang dilakukan oleh Pratama, berhasil menerapkan *K-Means Clustering* berbasis model RFM pada data transaksi UMKM di Manggarai Barat dan menghasilkan dua segmen pelanggan yang dapat dimanfaatkan sebagai dasar strategi retensi [8]. Implementasi algoritma *K-Means Clustering* yang mengintegrasikan data karakteristik personal dan kebiasaan belanja terbukti mampu menciptakan empat segmentasi konsumen yang unik. [9]. Namun, penelitian-penelitian tersebut umumnya masih menerapkan segmentasi dan prediksi secara terpisah, sehingga belum sepenuhnya mengoptimalkan sinergi antara kedua pendekatan dalam satu kerangka analisis yang terpadu.

Berdasarkan kesenjangan tersebut, penelitian ini mengusulkan integrasi metode klasifikasi *Random Forest* yang memiliki keunggulan dalam hal akurasi serta resistansi terhadap *overfitting* [10]. Literatur mengenai prediksi penjualan busana muslim menunjukkan bahwa *Random Forest* sebagai metode *ensemble* berbasis pohon keputusan mampu mencapai tingkat akurasi yang optimal, bahkan menyentuh angka 100% pada kategori produk tertentu [11]. Melalui penggabungan kekuatan segmentasi *K-Means* dan kapabilitas klasifikasi *Random Forest*, diharapkan model hibrida yang dibangun dapat memberikan proyeksi pola pembelian yang lebih presisi dan terukur.

Studi oleh Anam dkk, memberikan bukti empiris bahwa algoritma *Random Forest* pada konteks prediksi *churn* pelanggan ritel online menunjukkan bahwa analisis *feature importance* pada model tersebut mampu mengidentifikasi variabel transaksi dan kepuasan pelanggan sebagai faktor dominan, sekaligus menegaskan pentingnya kualitas dan representativitas data dalam membangun model prediksi yang andal [12]. Temuan ini relevan dengan pendekatan yang digunakan dalam penelitian ini, di mana fitur-fitur transaksi seperti nilai belanja, frekuensi pembelian, dan lokasi geografis menjadi prediktor utama dalam pembentukan segmen dan prediksi pola pembelian

Meskipun penelitian terpisah tentang clustering dan klasifikasi telah banyak dilakukan, masih terdapat kesenjangan penelitian terkait efektivitas model hibrida yang mengintegrasikan hasil segmentasi *K-Means* sebagai fitur masukan (*input features*) pada model *Random Forest* dalam konteks produk gamis. Mayoritas studi terdahulu cenderung menerapkan metode tersebut secara terpisah atau hanya berfokus pada salah satunya [13][14]. Padahal, sinergi kedua metode ini secara empiris mampu meningkatkan performa prediksi konsumsi pengguna hingga mencapai akurasi di atas 94% dalam jangka panjang [15]. Dengan demikian, penelitian ini bertujuan untuk menjembatani celah tersebut melalui implementasi *Hybrid K-Means Clustering* dan *Random Forest Classification* guna menganalisis segmentasi sekaligus memprediksi pola penjualan.

Riset ini menerapkan alur CRISP-DM dengan mengintegrasikan fitur RFM ke dalam model hibrida *K-Means* dan *Random Forest* untuk memetakan profil sekaligus memprediksi perilaku belanja pelanggan. Secara teoretis, penelitian ini memperkaya literatur pemodelan data pada ritel fesyen, sementara secara praktis memberikan rujukan bagi pelaku usaha dalam mengoptimalkan stok dan strategi promosi regional. Hasil analisis ini diharapkan mampu memperkuat loyalitas konsumen melalui kebijakan manajemen inventori dan pemasaran yang berbasis pada data transaksional yang terukur.

2. METODOLOGI PENELITIAN

2.1 Data Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan data sekunder yang bersifat privat dan terstruktur, yang bersumber dari catatan transaksi internal toko gamis pada platform e-commerce. Data yang digunakan berupa dataset transaksi penjualan yang mencakup periode Januari 2025 hingga Maret 2026 dengan total sebanyak 1.000 data transaksi. Berdasarkan dataset yang digunakan dalam penelitian ini, terdapat beberapa kelompok atribut yang digunakan dalam proses analisis. Kelompok atribut tersebut disajikan pada Tabel 1. Kelompok Atribut Dataset Penelitian sebagai berikut:

Tabel 1. Kelompok Atribut Dataset Penelitian

Kelompok Atribut	Variabel	Deskripsi
Identitas Transaksi	No Pesanan	Nomor unik yang digunakan untuk mengidentifikasi setiap transaksi

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

Waktu	Waktu Pesanan	Tanggal atau waktu terjadinya transaksi untuk analisis pola temporal
Produk	Nama Produk	Nama produk gamis yang dibeli
	SKU	Kode unik produk
	Variasi	Jenis atau model gamis (misalnya abaya, crinkle, marbella)
	Total Harga	Total nilai pembelian dalam satu transaksi
	Diskon	Besaran potongan harga yang diberikan
Logistik & Perilaku Konsumen	Provinsi	Lokasi asal pembeli
	Metode Bayar	Jenis metode pembayaran yang digunakan
	Pengiriman	Jenis layanan pengiriman yang dipilih

Data yang digunakan dalam penelitian ini bersifat terstruktur dalam bentuk tabel, sehingga memudahkan dalam proses pengolahan menggunakan metode data mining. Selain itu, data telah melalui proses seleksi awal untuk memastikan kelengkapan dan konsistensi, sehingga layak digunakan untuk tahap analisis lebih lanjut. Penggunaan data sekunder dalam penelitian ini bertujuan untuk memperoleh gambaran perilaku konsumen secara nyata berdasarkan aktivitas transaksi yang telah terjadi.

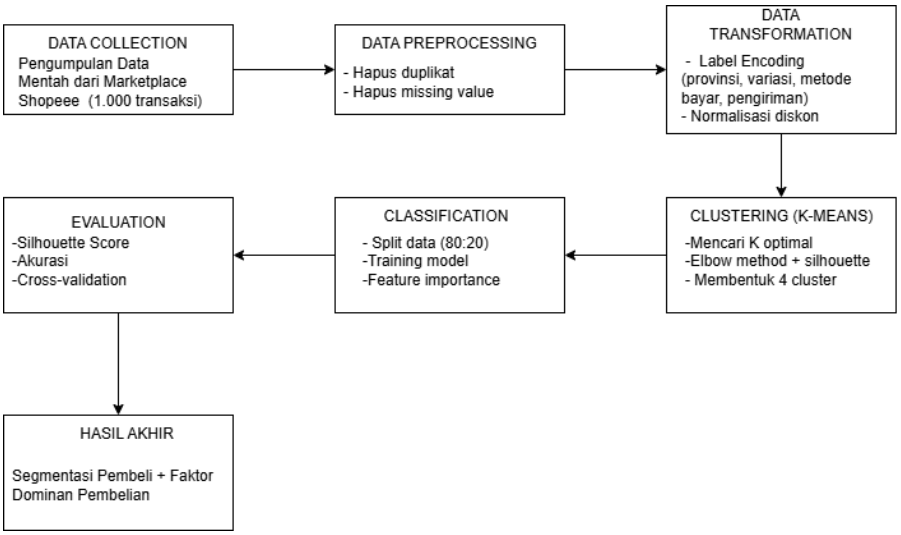
2.2 Teknik Pengumpulan Data

Data dikumpulkan melalui dua metode utama:

1. Wawancara & Akses Langsung
   Data diperoleh melalui wawancara dengan admin toko Shopee yang memiliki akses terhadap data transaksi. Dari proses ini diperoleh data mentah (*raw data*) berupa catatan transaksi penjualan yang bersifat privat, meliputi informasi waktu pemesanan, produk, variasi, metode pembayaran, pengiriman, dan lokasi pembeli. Metode ini digunakan untuk memastikan keakuratan dan keaslian data.
2. Studi Literatur
   Studi literatur dilakukan dengan mengkaji jurnal dan referensi ilmiah yang relevan dengan data mining, khususnya metode K-Means dan Random Forest. Literatur ini digunakan sebagai dasar dalam pemilihan metode dan analisis penelitian.

2.3 Modul Penelitian

Model penelitian pada studi ini dirancang untuk mengolah data transaksi penjualan gamis secara sistematis menggunakan pendekatan data mining. Proses penelitian dimulai dari tahap pengumpulan data hingga menghasilkan informasi berupa segmentasi pembeli dan faktor dominan yang mempengaruhi pola pembelian. Alur penelitian secara keseluruhan dapat dilihat pada Gambar 1. Alur Olah Data Penelitian:



Gambar 1. Alur Olah Data Penelitian

Penjelasan setiap tahap:

1. *Data Collection*
   Pengumpulan data transaksi penjualan gamis sebanyak 1.000 data dari Shopee yang mencakup informasi transaksi, produk, dan pembeli.
2. *Data Preprocessing*

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

Pembersihan data dengan menghapus duplikat dan menangani *missing value* untuk memastikan kualitas data.

3. *Data Transformation*
   Transformasi data melalui encoding variabel kategorikal dan normalisasi data numerik.
4. *Clustering (K-Means)*
   Pengelompokan data menggunakan K-Means untuk membentuk 4 cluster berdasarkan kemiripan karakteristik.
5. *Classification (Random Forest)*
   Klasifikasi dilakukan untuk menganalisis faktor yang mempengaruhi segmentasi dengan pembagian data latih dan uji.
6. *Evaluation*
   Evaluasi model menggunakan Silhouette Score untuk *clustering* serta akurasi dan *cross-validation* untuk klasifikasi.
7. Hasil Akhir
   Menghasilkan segmentasi pembeli dan faktor dominan yang mempengaruhi pola pembelian.

2.4 Metode Analisis Data

a. *Data Processing*

Tahap awal meliputi *data cleaning* dengan menghapus data duplikat dan menangani *missing value* pada atribut utama. Selanjutnya dilakukan *Data Transformation* berupa *Label Encoding* untuk mengubah data kategorikal (provinsi, variasi, metode bayar, pengiriman) menjadi numerik. Kolom numerik seperti diskon dan total harga dilakukan normalisasi skala menggunakan *StandardScaler* guna menyamakan rentang nilai agar tidak terjadi dominasi fitur pada perhitungan jarak algoritma.

b. *K-Means Clustering*

Algoritma K-Means Clustering diterapkan untuk mengelompokkan transaksi penjualan gamis melalui identifikasi kesamaan karakteristik antar data. Teknik ini bekerja dengan cara mereduksi variansi di dalam kelompok (*cluster*) sekaligus memperlebar perbedaan antar kelompok tersebut. Penentuan posisi data dalam klaster didasarkan pada perhitungan kedekatan matematis menggunakan Euclidean Distance. Rumus yang digunakan untuk mengukur jarak tersebut adalah sebagai berikut:

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

Jumlah cluster ditetapkan sebanyak 4 kelompok (k=4) untuk hasil segmentasi yang optimal. Kualitas cluster dievaluasi menggunakan Silhouette Score yang mengukur kekompakan data dalam satu cluster dan jaraknya terhadap cluster lain.

c. *RF Classification*

Hasil pelabelan klaster dari tahap K-Means kemudian dijadikan sebagai variabel target dalam pemodelan *Random Forest Classification*. Algoritma ini membangun serangkaian pohon keputusan secara acak untuk mengklasifikasikan segmen pembeli baru. Alur ini memungkinkan sistem untuk mengidentifikasi *Feature Importance*, yaitu variabel yang paling berpengaruh terhadap pembentukan segmen pelanggan.

3. HASIL DAN PEMBAHASAN

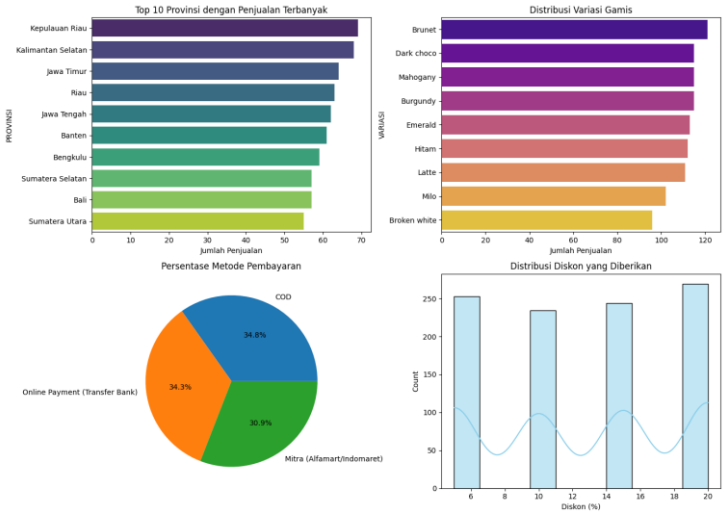
3.1 Preprocessing dan Eksplorasi data

Dataset yang digunakan dalam penelitian ini merupakan data transaksi penjualan gamis dari toko IC Label pada platform *e-commerce* Shopee yang diperoleh secara langsung melalui akses admin toko. Dataset bersifat privat dan terstruktur, terdiri dari 1.000 baris transaksi pada periode 1 Januari 2025 – 18 Maret 2026. Atribut dataset meliputi profil pembeli (*username*), detail transaksi (waktu pesan, total harga, diskon), spesifikasi produk (nama SKU, variasi warna), serta informasi logistik yang mencakup provinsi/kota tujuan, metode pembayaran, dan jasa pengiriman.

3.1.1 Eksplorasi Data Awal (EDA)

Tahap eksplorasi data awal dilakukan untuk memahami distribusi dan karakteristik dataset sebelum masuk ke tahap pemodelan. Hasil eksplorasi divisualisasikan pada Gambar 2 yang mencakup empat aspek utama.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001



Gambar 2. Karakter Data Set

Berdasarkan Gambar 2, karakteristik utama dataset dapat diidentifikasi sebagai berikut.

a. Distribusi Geografis**:** Kepulauan Riau (69 transaksi), Kalimantan Selatan (68 transaksi), dan Jawa Timur (64 transaksi) mendominasi penjualan dari sisi provinsi. Dataset mencakup 18 provinsi yang tersebar di seluruh Indonesia.
b. Variasi Produk: Terdapat 9 variasi warna gamis yang dipasarkan. Variasi Brunet merupakan warna terlaris dengan 121 transaksi (12,1%), disusul Mahogany dan Burgundy masing-masing 115 transaksi (11,5%).
c. Metode Pembayaran: Ketiga metode pembayaran terdistribusi merata, yaitu COD (34,8%), *Online Payment* Transfer Bank (34,3%), dan Mitra Alfamart/Indomaret (30,9%).
d. Diskon: Diskon berkisar antara 5% hingga 20% dengan rata-rata sebesar 12,6% dan median 15%.

3.1.2 Tahapan *Preprocessing*

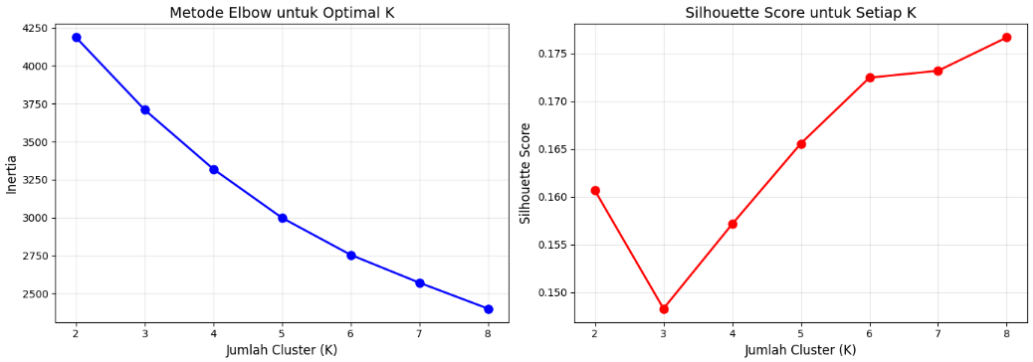
Proses *preprocessing* dilakukan untuk mengubah data mentah menjadi dataset yang siap dimodelkan. Tahapan yang dilaksanakan adalah sebagai berikut.

a. Data Cleaning: Menghapus baris duplikat dan transaksi dengan nilai kosong (*missing value*) pada kolom Nama Produk dan Variasi.
b. Ekstraksi Fitur Waktu: Mengekstrak informasi bulan, hari, dan jam dari kolom tanggal pemesanan menggunakan pustaka pandas, serta memisahkan nama provinsi dari kolom gabungan PROVINSI/KOTA.
c. Encoding Kategorikal: *Encoding* fitur kategorikal menggunakan *Label Encoder* dari Scikit-Learn, menghasilkan empat fitur numerik baru: PROVINSI_EN (18 kategori, 0: Aceh – 17: Sumatera Utara), VARIASI_EN (9 kategori), METODE_EN, dan PENGIRIMAN_EN.
d. Normalisasi: Standardisasi fitur dilakukan menggunakan *StandardScaler* untuk menyetarakan skala seluruh variabel agar perhitungan jarak dalam algoritma *clustering* tidak didominasi oleh satu fitur tertentu.

3.2 Hasil K-Means Clustering

Penentuan jumlah klaster (K) yang optimal merupakan tahap krusial dalam algoritma K-Means. Fitur yang digunakan sebagai masukan meliputi PROVINSI_EN, VARIASI_EN, METODE_EN, PENGIRIMAN_EN, dan DISKON_NUM. Dua metode digunakan secara bersamaan untuk menentukan nilai K terbaik, yaitu *Elbow Method* yang mengukur nilai *Within-Cluster Sum of Squares* (WCSS/*Inertia*), dan *Silhouette Score* yang mengukur kualitas keterpisahan antar klaster. Pengujian dilakukan pada rentang K=2 hingga K=8 dengan *random_state*=42 dan *n_init*=10.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001



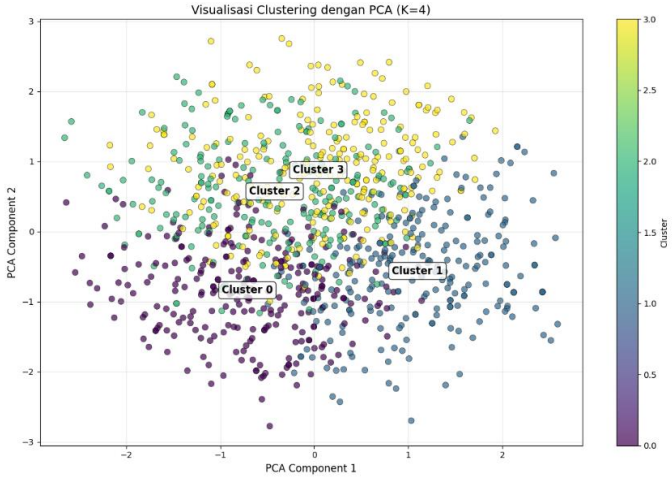
Gambar 3. Nilai Silhoutte

Berdasarkan gambar 3, nilai Silhouette Score tertinggi secara teknis dicapai pada K=8 (0,1766). Namun, nilai silhouette yang rendah pada keseluruhan rentang K (di bawah 0,2) merupakan kondisi yang lazim pada data transaksional ritel karena pola pembelian antar konsumen cenderung tumpang tindih. Grafik Elbow menunjukkan penurunan inertia yang mulai melandai setelah K=4, membentuk titik siku yang mengindikasikan K=4 sebagai pilihan efisien. Dengan mempertimbangkan keseimbangan distribusi klaster dan interpretabilitas bisnis, penelitian ini menetapkan K=4 (Silhouette Score: 0,1571).

Selanjutnya, algoritma K-Means dijalankan dengan K=4, random_state=42, dan n_init=10. Kode implementasi ditampilkan pada Gambar 4. Hasil clustering menghasilkan empat klaster dengan distribusi yang relatif seimbang. Visualisasi sebaran klaster menggunakan reduksi dimensi Principal Component Analysis (PCA) ditampilkan pada Gambar 5.

```
# ============================================
# CELL 6: Implementasi K-Means dengan K Optimal
# ============================================
# Gunakan K optimal (misal 4)
k_final = 4  # Bisa disesuaikan berdasarkan hasil CELL 5
kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
df_clean['CLUSTER'] = kmeans_final.fit_predict(X_scaled)

# Evaluasi final
final_silhouette = silhouette_score(X_scaled, df_clean['CLUSTER'])
print("="*50)
print("✅ HASIL CLUSTERING FINAL")
print("="*50)
print(f"Jumlah cluster: {k_final}")
print(f"Silhouette Score final: {final_silhouette:.4f}")

# Profil setiap cluster
print("\n📋 PROFIL SETIAP CLUSTER:")
for cluster in range(k_final):
    print(f"\n{'='*40}")
    print(f"🔵 CLUSTER {cluster}")
    print(f"{'='*40}")
    cluster_data = df_clean[df_clean['CLUSTER'] == cluster]
    print(f"Jumlah anggota: {len(cluster_data)} ({len(cluster_data)/len(df_clean)*100:.1f}%)")
    print(f"Provinsi terbanyak: {cluster_data['PROVINSI'].mode()[0] if len(cluster_data['PROVINSI'].mode())>0 else 'N/A'}")
    print(f"Variasi terbanyak: {cluster_data['VARIASI'].mode()[0] if len(cluster_data['VARIASI'].mode())>0 else 'N/A'}")
    print(f"Metode bayar terbanyak: {cluster_data['METODE BAYAR'].mode()[0] if len(cluster_data['METODE BAYAR'].mode())>0 else 'N/A'}")
    print(f"Rata-rata diskon: {cluster_data['DISKON_NUM'].mean():.1f}%")
    print(f"Rata-rata harga: Rp{cluster_data['TOTAL HARGA'].mean():,.0f}")
```

Gambar 4. Kode Implementasi



Gambar 5. Grafik Cluster

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

Berdasarkan Gambar 4 dan 5, keempat klaster diinterpretasikan sebagai berikut. Klaster 0 dengan Pembeli Premium Wilayah Jawa (261 transaksi, 26,1%), Klaster terbesar dengan rata-rata harga tertinggi (Rp138.584) dan diskon terendah (10,6%). Didominasi wilayah Jawa Tengah, memilih variasi Brunet dengan metode pembayaran COD dan layanan Anteraja. Segmen ini merepresentasikan konsumen yang bersedia membayar harga lebih tinggi tanpa bergantung pada diskon besar. Klaster 1 dengan Pembeli Sumatera Sensitif Diskon (249 transaksi, 24,9%): Klaster dengan rata-rata diskon tertinggi (13,9%) dan terkonsentrasi di Sumatera Utara dan sekitarnya. Variasi Burgundy paling diminati dengan layanan Shopee Standard sebagai pilihan pengiriman dominan. Segmen ini merespons kuat terhadap program promosi dan diskon. Klaster 2 dengan Pembeli Digital Wilayah Kepulauan (233 transaksi, 23,3%): Dicirikan oleh dominasi metode *Online Payment* (Transfer Bank) dengan konsentrasi di Kepulauan Riau. Variasi Mahogany paling diminati. Segmen ini merepresentasikan konsumen yang melek teknologi finansial dan terbiasa bertransaksi secara digital. Sementara itu, Klaster 3 dengan Pembeli Mitra Wilayah Sumatera Bagian Selatan (257 transaksi, 25,7%): Didominasi metode pembayaran melalui Mitra Alfamart/Indomaret (91 transaksi) dan terkonsentrasi di Bengkulu. Variasi Mahogany dan Shopee Standard menjadi pilihan utama. Segmen ini lebih nyaman dengan transaksi fisik di minimarket terdekat.

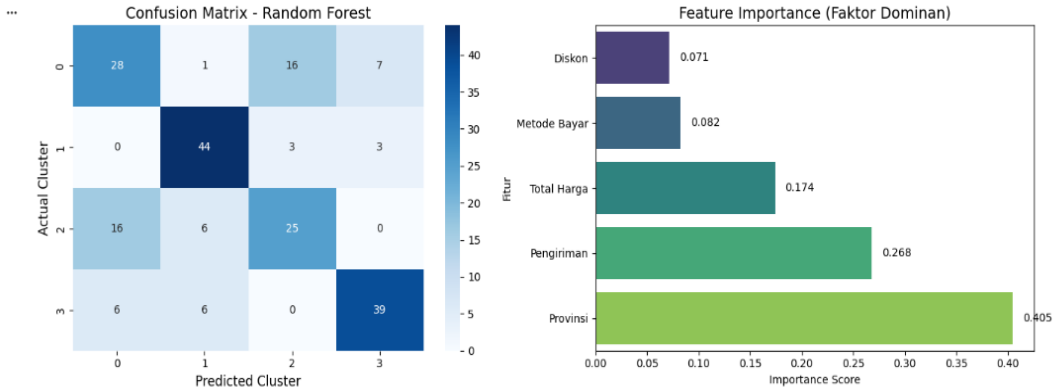
3.2 Evaluasi Model Random Forest Classification

Tahap evaluasi model dilakukan untuk menilai efektivitas kinerja Random Forest dalam memprediksi klaster pembeli baru. Label klaster hasil K-Means digunakan sebagai variabel target (y), sedangkan fitur masukan (X) terdiri dari PROVINSI_EN, METODE_EN, PENGIRIMAN_EN, DISKON_NUM, dan TOTAL HARGA. Penambahan variabel TOTAL HARGA pada tahap klasifikasi bertujuan memperkaya informasi pembelian agar model dapat memproyeksikan kecenderungan klaster secara lebih komprehensif.

Konfigurasi model Random Forest menggunakan *n_estimators=100* dan *max_depth=10* dengan pembagian data 80:20 (800 latih, 200 uji) menggunakan teknik *stratified sampling*. Performa algoritma divalidasi melalui *confusion matrix*, laporan klasifikasi, serta pengujian stabilitas model lewat *5-fold cross-validation*. *Output* model ditampilkan pada Gambar 6 dan 7.

```
==================================================
🌲 HASIL RANDOM FOREST CLASSIFICATION
==================================================
Accuracy Training: 0.9637 (96.38%)
Accuracy Testing:  0.6800 (68.00%)
Gap (overfitting): 0.2837

Cross-validation (5-fold):
  Rata-rata: 0.6870 (+/- 0.0350)
  Detail: [0.695 0.675 0.695 0.71  0.66 ]

📋 CLASSIFICATION REPORT (Testing):
              precision    recall  f1-score   support

           0       0.56      0.54      0.55        52
           1       0.77      0.88      0.82        50
           2       0.57      0.53      0.55        47
           3       0.80      0.76      0.78        51

    accuracy                           0.68       200
   macro avg       0.67      0.68      0.68       200
weighted avg       0.68      0.68      0.68       200
```

Gambar 6. Hasil Random Forest



Gambar 7. Hasil Matriks Random Forest

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

Model Random Forest menghasilkan akurasi 96,38% pada data latih dan 68,00% pada data uji, dengan konsistensi *cross-validation* 5-fold sebesar 68,70% (±3,50%). Kesenjangan akurasi sebesar 28,38% mengindikasikan adanya *overfitting*, sebuah kondisi lazim saat menggunakan label klaster *probabilistik* sebagai target klasifikasi. Secara spesifik, model berkinerja optimal pada Klaster 1 (F1-score 0,82) dan Klaster 3 (F1-score 0,78) karena memiliki karakteristik geografis yang kontras. Sebaliknya, Klaster 0 dan 2 memiliki performa lebih rendah (F1-score 0,55) akibat kemiripan profil rata-rata harga dan diskon yang menyebabkan batas keputusan klasifikasi menjadi kurang tegas.

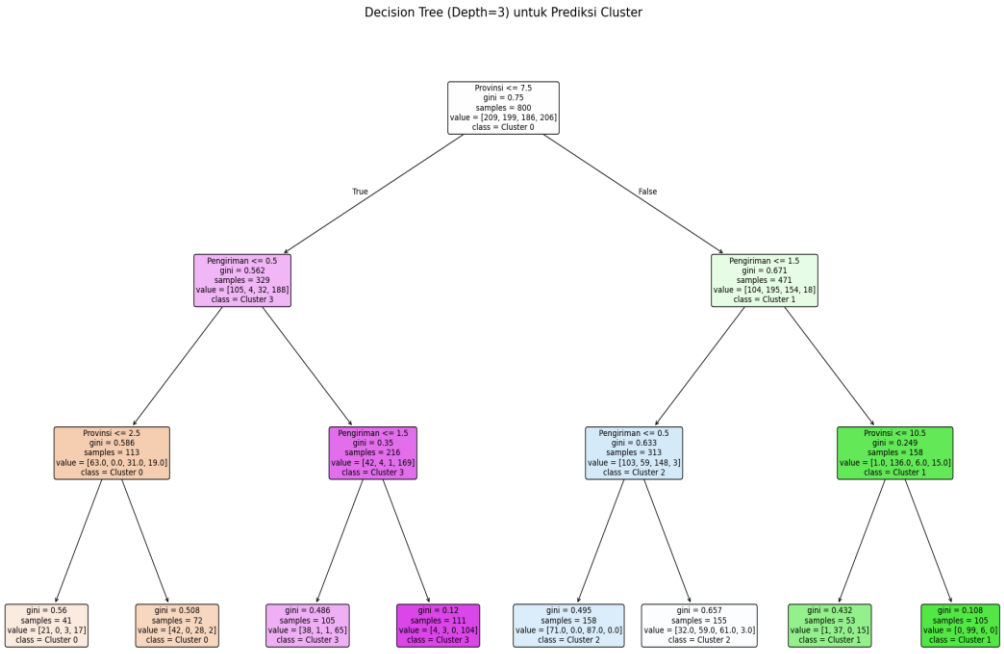
3.3 Analisis Feature Importance

Analisis *feature importance* dalam model *Random Forest* digunakan untuk mengukur kontribusi masing-masing variabel prediktor selama proses klasifikasi. Skor ini ditentukan berdasarkan frekuensi dan signifikansi sebuah fitur saat dijadikan parameter pemisahan (*split*) pada seluruh kumpulan pohon keputusan dalam model tersebut. Berdasarkan Gambar 7, rangkuman nilai kepentingan setiap fitur disajikan pada Tabel 1.

Tabel 2. Analisis Nilai Prediksi

Rank	Fitur	Importance Score	Kontribusi (%)
1	Provinsi	0,4048	40,5%
2	Pengiriman	0,2680	26,8%
3	Total Harga	0,1740	17,4%
4	Metode Bayar	0,0819	8,2%
5	Diskon	0,0713	7,1%

Berdasarkan Tabel 2, variabel Provinsi (40,5%) dan Pengiriman (26,8%) merupakan faktor paling dominan dalam menentukan pola pembelian gamis IC Label. Temuan ini mengonfirmasi bahwa preferensi konsumen sangat dipengaruhi oleh letak geografis dan ketersediaan infrastruktur logistik di tiap wilayah. Variabel Total Harga (17,4%) turut memperkuat adanya segmentasi berbasis daya beli, sementara Metode Bayar dan Diskon memberikan kontribusi yang lebih kecil namun tetap relevan. Hasil ini dipertegas oleh visualisasi *Decision Tree* (Gambar 8), di mana variabel Provinsi dan Pengiriman menempati posisi percabangan teratas (*top-level split*), membuktikan keduanya sebagai prediktor utama dalam mengklasifikasikan segmen pembeli.



Gambar 8. Alur Decision Tree

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

3.4 Implementasi

3.4.1 Simulasi Prediksi Klaster Pembeli Baru

Simulasi prediksi dilakukan menggunakan skenario pembeli baru dengan atribut: Provinsi Jawa Barat, metode pembayaran COD, jasa pengiriman Shopee Express, diskon 15%, dan total harga Rp150.000. Proses prediksi dimulai dengan *encoding* nilai kategorikal menggunakan objek *LabelEncoder* yang telah dilatih melalui fungsi *safe_transform()* untuk menangani nilai yang tidak dikenali (*unseen labels*).

Tabel 3. Prediksi Pembeli Baru

Profil Pembeli Baru	Klaster 0	Klaster 1	Klaster 2	Klaster 3
Jawa Barat | COD | Shopee Express | 15% | Rp150.000	52,01% ✓	17,07%	0,00%	30,92%

Berdasarkan Tabel 3, pembeli baru tersebut diprediksi masuk ke dalam Klaster 0 dengan probabilitas tertinggi sebesar 52,01%, yaitu segmen Pembeli Premium Wilayah Jawa. Alternatif terdekat adalah Klaster 3 (30,92%), yang menandakan preferensi pembeli juga condong ke segmen Mitra. Probabilitas Klaster 2 sebesar 0,00% menunjukkan profil transaksi ini sangat tidak cocok dengan karakteristik klaster digital kepulauan.

Informasi hasil prediksi ini dapat langsung dimanfaatkan IC Label untuk merancang strategi pemasaran yang tersegmentasi. Untuk memperjelas dasar matematis prediksi tersebut, berkut disajikan perhitungan *Euclidean Distance* dari profil pembeli baru [PROVINSI=4, VARIASI=0, METODE=0, PENGIRIMAN=1, DISKON=15] ke centroid setiap klaster, sebagai landasan keputusan klasifikasi:

$$d(k0) = \sqrt{[(4-6)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (15-10{,}6)^2]}$$
$$= \sqrt{[4+0+0+1+19{,}36]} = \sqrt{24{,}36} = 4{,}936$$
$$d(k1) = \sqrt{[(4-14)^2 + (0-1)^2 + (0-0)^2 + (1-2)^2 + (15-13{,}9)^2]}$$
$$= \sqrt{[100+1+0+1+1{,}21]} = \sqrt{103{,}21} = 10{,}159$$
$$d(k2) = \sqrt{[(4-8)^2 + (0-4)^2 + (0-2)^2 + (1-1)^2 + (15-12{,}6)^2]}$$
$$= \sqrt{[16+16+4+0+5{,}76]} = \sqrt{41{,}76} = 6{,}462$$
$$d(k3) = \sqrt{[(4-2)^2 + (0-4)^2 + (0-1)^2 + (1-2)^2 + (15-12{,}6)^2]}$$
$$= \sqrt{[4+16+1+1+5{,}76]} = \sqrt{27{,}76} = 5{,}269$$

Berdasarkan perhitungan di atas, jarak terkecil adalah d(K0) = 4,936, sehingga secara matematis pembeli baru tersebut paling dekat dengan centroid Klaster 0. Hasil ini selaras dengan probabilitas prediksi Random Forest sebesar 52,01% untuk Klaster 0 pada Tabel 2. Alternatif terdekat adalah Klaster 3 dengan d = 5,269, yang juga tercermin pada probabilitas kedua tertinggi sebesar 30,92%. Konsistensi antara perhitungan jarak *Euclidean* dan probabilitas output Random Forest membuktikan bahwa model klasifikasi bekerja secara koheren dengan struktur klaster yang dihasilkan oleh K-Means.

a. Merekomendasikan produk dengan harga menengah-atas seperti Fatimah Abaya atau Renata Marbella dengan variasi Brunet, sesuai preferensi dominan Klaster 0.
b. Memprioritaskan layanan COD dengan mitra pengiriman Anteraja sebagai opsi utama untuk wilayah Jawa Barat.
c. Menetapkan diskon pada kisaran 10–11% sesuai rata-rata diskon profil Klaster 0, tanpa perlu memberikan promosi diskon besar yang tidak efisien untuk segmen premium.

Hasil akhir seluruh transaksi beserta label klaster juga diekspor ke dalam format CSV (*hasil_clustering_gamis.csv*) yang dapat diintegrasikan dengan sistem manajemen toko IC Label untuk keperluan analisis lanjutan dan pembaruan model secara berkala.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 355-364
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1001

4. KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan model hibrida yang mengintegrasikan algoritma *K-Means Clustering* dan *Random Forest Classification* di bawah kerangka kerja CRISP-DM untuk mengatasi anomali stok pada toko IC Label di platform Shopee. Pendekatan berbasis data ini terbukti efektif dalam memproyeksikan permintaan pasar secara sistematis guna meminimalisir risiko *overstock* dan *stockout*. Melalui implementasi *K-Means* dengan nilai optimal $K=4$, penelitian ini berhasil memetakan empat segmen pelanggan yang memiliki karakteristik bisnis yang unik: Pembeli Premium Wilayah Jawa (26,1%), Pembeli Sumatera Sensitif Diskon (24,9%), Pembeli Digital Wilayah Kepulauan (23,3%), dan Pembeli Mitra Wilayah Sumatera Bagian Selatan (25,7%). Meskipun nilai *Silhouette Score* yang dihasilkan relatif rendah sebesar 0,1571 akibat adanya pola pembelian yang tumpang tindih secara transaksional, distribusi klaster yang seimbang memberikan fondasi yang kuat bagi interpretasi strategi bisnis yang relevan. Pada tahap klasifikasi, integrasi label klaster sebagai variabel target dalam model *Random Forest* menunjukkan performa yang stabil dengan akurasi uji sebesar 68,00% dan konsistensi *cross-validation* 5-fold pada angka 68,70%. Analisis *feature importance* mengungkapkan bahwa faktor geografis yang diwakili oleh variabel Provinsi (40,5%) dan preferensi logistik melalui variabel Pengiriman (26,8%) menjadi determinan utama dalam pembentukan pola pembelian pelanggan. Hal ini mengonfirmasi bahwa strategi pemasaran regional akan jauh lebih efektif bagi IC Label dibandingkan pendekatan yang bersifat seragam. Secara praktis, kemampuan model hibrida ini dalam memprediksi segmen pembeli baru secara *real-time* memberikan kontribusi signifikan terhadap optimasi inventori dan manajemen hubungan pelanggan berbasis data. Untuk pengembangan riset selanjutnya, disarankan untuk memperluas rentang waktu pengambilan data serta mengeksplorasi teknik *oversampling* atau algoritma *clustering* alternatif seperti DBSCAN guna mengatasi kendala *overfitting* dan meningkatkan kualitas keterpisahan antar segmen.

REFERENCES

[1] M. Dora, R. Khairul, and W. M. Sari, "Analisa Transaksi Penjualan Dalam peningkatan Promosi Penjualan Berbasis Sistem Informasi How to Cite :," *J. Ekombis Rev. – J. Ilm. Ekon. dan Bisnis*, vol. 11, no. 1, pp. 357–368, 2023.
[2] A. Gustiana, "Comparative Analysis of Muslim Clothing Sales Predictions Using the C4 . 5 Method and Linear Regression," *J. Syst. Eng. Inf. Technol.*, vol. 03, no. 01, pp. 30–36, 2024, doi: 10.29207/joseit.v3i1.5678.
[3] D. A. Kurnia, K. Anam, C. L. Rohma, and Fathurrohman, "Analisis Data Penjualan Toko Menggunakan Power Bi Untuk Meningkatkan Strategi Bisnis Studi Kasus : Toko XYZ," *J. Comput. Sci. Artif. Intell.*, vol. 06, no. 02, pp. 1–8, 2025, doi: 10.32485/jcsai.JOURNAL.
[4] S. Siregar and A. Rahman, "Pengelompokkan Jenis Surat Masuk di Dinas Komunikasi dan Informatika Menggunakan Metode K-Means Clustering," *J. Komput. Teknol. Sist. Inf.*, vol. 5, no. 1, pp. 200–211, 2026.
[5] A. K. Ramadhan, A. Sakti, and B. O. Rosdiyanti, "Analisis Dampak Digitalisasi Penjualan terhadap Kinerja UMKM Retail Menggunakan Algoritma C4 . 5," *Siwah Multidiscip. Sci. J.*, vol. 2, pp. 63–76, 2026.
[6] R. Ulya, "Pemanfaatan Data Mining untuk Prediksi Tren Konsumen pada Platform E - Retail," *J. Ilmu Komput.*, vol. 1, no. 1, pp. 27–32, 2026.
[7] M. Fajar, S. Adam, B. Putra, S. I. Puteri, A. Fajrissiddiq, and L. S. Parwati, "Eksplorasi dan Analisis Data Mining untuk Prediksi Pola Konsumen Menggunakan Teknik Klasifikasi dan Clustering," *Sentim. (Seminar Nas. Teknol. Informasi, Mekatronika dan Ilmu Komputer)*, vol. 4, 2025.
[8] I. W. P. Pratama, W. S. Jangku, S. F. Djun, and I. P. Eka, "Segmentasi Pelanggan UMKM Berdasarkan Pola Distribusi Produk di Manggarai Barat Menggunakan K-Means Clustering," in *National Conference on Tourism (NCT) 2025 Universitas Triatma Mulya*, 2025, pp. 70–85.
[9] N. Hidayat, D. P. Gevano, and A. P. R. Ilahi, "Segmentasi Pelanggan Menggunakan K-Means Clustering Berdasarkan Data Kepribadian dan Pola Konsumsi," *J. Tek. Inform.*, vol. 6, no. 5, pp. 3914–3924, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.5.5140.
[10] L. Alfiyani, "Penerapan Data Mining Dengan Algoritma Kmeans Terhadap Data Transaksi Penjualan (Studi Kasus : CV . Sogan Batik Rejodani)," Universitas Islam Indonesia, 2023.
[11] G. T. Atmaja, "Enhancing Fashion Product Sales Segmentation Using Random Forest with SMOTE and Hyperparameter Optimization," *J. Ilmu Komput. dan Inform.*, vol. 5, no. 1, pp. 45–56, 2025.
[12] M. Haris, K. Anam, D. Kurnianingtyas, and A. A. Soebroto, "Implementasi Algoritma Random Forest Untuk Prediksi Churn Pada Pelanggan Retail Online," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4, pp. 1–10, 2026.
[13] D. Ramdhan, G. Dwilestari, R. D. Dana, A. Ajiz, and Kaslani, "Clustering Data Persediaan Barang dengan Menggunakan Metode," *MEANS (Media Inf. Anal. dan Sist.*, vol. 7, no. 1, pp. 1–9, 2022.
[14] S. I. Puro, J. Hariyan, R. Rafliansyah, R. Aziz, and P. V. Rajagukguk, "Analisa Data Shopping Trends Menggunakan Algoritma Klasifikasi Dengan Metode Naive Bayes," *Repeater Publ. Tek. Inform. dan Jar.*, vol. 2, no. 3, pp. 199–134, 2024.
[15] A. Luthfi, D. Alfareza, J. P. Natarendra, and M. A. Pratama, "Analisis Klaster Pembiayaan UMKM dan Sektor Ekonomi Menggunakan Metode K-Means di Jawa Tengah," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 5469–5473, 2025.