

Pengaruh Feature Engineering terhadap Kualitas Clustering pada Segmentasi Pelanggan Ritel Daring

Amril Mutoi Siregar^{1*}, Bella Adisty², Surjandy³, Ely Aprilia⁴

^{1,2,3,4}Department Informatics, Faculty of digital design and business, Institut Teknologi Sains Bandung, Indonesia
amrilsiregar99@gmail.com, surjandy@iitb.ac.id, aprilia.ely@gmail.com

Info Artikel

Riwayat Artikel:

Received 2026-02-02

Revised 2026-08-14

Accepted 2026-08-23

Abstract – Customer segmentation based on transaction data is a significant challenge in the realm of digital marketing, as the effectiveness of clustering results is highly dependent on the features used. Reliance on imperfect transaction features often results in improved numerical evaluation metrics; however, this does not sufficiently capture the nuances of actual customer behaviour. This research effort seeks to investigate the impact of feature engineering on the quality of segmentation related to online retail customers. The methodology used is an experimental quantitative framework that compares two feature models, specifically Model 1 based on fundamental features and Model 2 anchored in customer behaviour engineering features, including RFM, Average Order Value, and product diversity. For this purpose, three unsupervised clustering algorithms were used, namely K-Means, Gaussian Mixture Model, and DBSCAN, with evaluation conducted using Silhouette Score and Davies-Bouldin Index. Findings indicate that while Model 1 produces superior metric values, it shows reduced relevance to the business context. Conversely, Model 2 produced clusters characterized by increased stability, consistency, and interpretability for marketing applications. This study ultimately concludes that the success of customer segmentation is largely influenced by the caliber of feature engineering rather than simply increased values.

Keywords: Customer Segmentation; Clustering; Feature Engineering; Online Retail; Machine Learning.

Corresponding Author:

Amril Mutoi Siregar

Email: amrilsiregar99@gmail.com



This is an open access article under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.

Abstrak – Segmentasi pelanggan yang didasarkan pada data transaksi merupakan tantangan signifikan dalam ranah pemasaran digital, karena kemandirian hasil pengelompokan sangat bergantung pada fitur yang digunakan. Ketergantungan pada fitur transaksi yang belum sempurna sering menghasilkan metrik evaluasi numerik yang meningkat; Namun, itu tidak cukup menangkap seluk-beluk perilaku pelanggan yang sebenarnya. Upaya penelitian ini berusaha untuk menyelidiki dampak rekayasa fitur pada kualitas segmentasi yang berkaitan dengan pelanggan ritel online. Metodologi yang digunakan adalah kerangka kerja kuantitatif eksperimental yang menyandingkan dua model fitur, khususnya Model 1 yang didasarkan pada fitur fundamental dan Model 2 yang berlabuh dalam fitur rekayasa perilaku pelanggan, mencakup RFM, Nilai Pesanan Rata-Rata, dan keragaman produk. Untuk tujuan ini, tiga algoritma pengelompokan tanpa pengawasan digunakan, yaitu K-Means, Model Campuran Gaussian, dan DBSCAN, dengan evaluasi dilakukan menggunakan Skor Silhouette dan Indeks Davies-Bouldin. Temuan menunjukkan bahwa sementara Model 1 menghasilkan nilai metrik yang unggul, ia menunjukkan relevansi yang berkurang dengan konteks bisnis. Sebaliknya, Model 2 menghasilkan cluster yang ditandai dengan peningkatan stabilitas, konsistensi, dan interpretabilitas untuk aplikasi pemasaran. Studi ini akhirnya menyimpulkan bahwa keberhasilan segmentasi pelanggan sebagian besar dipengaruhi oleh kaliber rekayasa fitur daripada hanya nilai yang meningkat dari metrik evaluasi internal.

Kata Kunci: Customer Segmentation; Clustering; Feature Engineering; Ritel Daring; Machine Learning

I. PENDAHULUAN

Dalam lingkungan digital modern, data pelanggan dianggap sebagai salah satu aset strategis paling penting bagi organisasi yang bercita-cita mempertahankan keunggulan kompetitif dan meningkatkan profitabilitas. Volume data transaksional yang terus berkembang yang berasal dari platform ritel online menciptakan peluang untuk penerapan metodologi pembelajaran mesin yang memiliki kapasitas untuk mengungkap pola perilaku pelanggan yang tidak mudah diidentifikasi melalui metode konvensional. Aplikasi pembelajaran mesin yang relevan dalam kerangka kerja ini adalah segmentasi pelanggan berbasis pengelompokan, yang berusaha untuk mengkategorikan pelanggan ke dalam kelompok homogen berdasarkan kesesuaian atribut perilaku mereka seperti frekuensi pembelian dan nilai transaksi kumulatif tanpa memerlukan label kelas yang sudah ada sebelumnya [1].

Algoritma pengelompokan tanpa pengawasan seperti K-Means, Gaussian Mixture Model (GMM), dan DBSCAN telah banyak digunakan dalam penelitian segmentasi pelanggan karena kemahirannya dalam mendeteksi struktur data tersembunyi dalam kumpulan data yang tidak berlabel. Pendekatan metodologis ini membantu dalam penggambaran kelompok pelanggan yang menunjukkan karakteristik perilaku analog, sehingga memperkuat inisiatif

organisasi dalam merancang strategi pemasaran yang lebih personal dan efektif [1], [2]. Namun demikian, efektivitas segmentasi secara signifikan bergantung pada kualitas dan relevansi fitur yang digunakan sebagai input ke model.

Dalam bidang pembelajaran mesin, rekayasa fitur merupakan proses penting yang mengubah data mentah menjadi fitur yang lebih informatif untuk model. Proses ini mencakup pemilihan fitur, transformasi, dan pembuatan variabel baru yang mampu mencerminkan pola perilaku pelanggan secara lebih akurat, seperti skor RFM (Terbaru, Frekuensi, Moneter), intensitas pembelian, dan fluktuasi pola pembelian [3][4]. Literatur yang ada menunjukkan bahwa rekayasa fitur sering memberikan pengaruh dominan pada kinerja model, seringkali melampaui pentingnya pemilihan algoritma itu sendiri, terutama ketika menangani data pelanggan yang dicirikan oleh dimensi tinggi dan variasi perilaku yang rumit.

Meskipun demikian, tantangan yang cukup besar dalam praktik segmentasi pelanggan adalah ketergantungan berlebihan pada fitur mendasar seperti RFM, tanpa penyelidikan lebih lanjut ke dalam fitur perilaku atau domain spesifik yang dapat meningkatkan stabilitas dan interpretasi hasil cluster dalam kerangka aplikasi bisnis nyata. Beberapa penelitian konvensional terus mengandalkan algoritma tunggal seperti K-Means yang diterapkan pada fitur transaksional dasar dan menilai kualitas cluster semata-mata melalui metrik internal seperti Skor Silhouette dan Indeks Davies-Bouldin, tanpa menghubungkannya dengan nilai strategis untuk pemasaran praktis [2]. Akibatnya, hasil segmentasi yang valid secara numerik tidak selalu menunjukkan relevansi substansif untuk pengambilan keputusan bisnis.

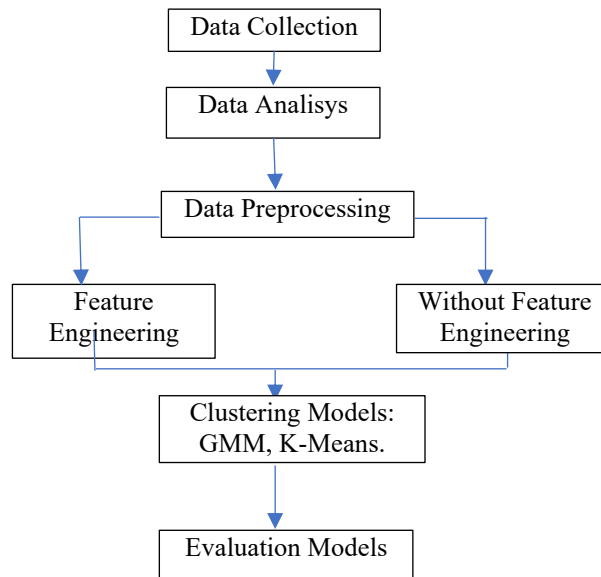
Penelitian sebelumnya telah menunjukkan bahwa penerapan model RFM dalam segmentasi pelanggan dapat menghasilkan cluster yang mengartikulasikan variasi dalam perilaku pelanggan [4], namun sebagian besar studi tersebut belum secara komprehensif menyelidiki sejauh mana rekayasa fitur berdampak pada kualitas segmentasi dan kontribusinya terhadap keputusan pemasaran strategis. Selanjutnya, studi internasional mengenai metodologi segmentasi pelanggan menggarisbawahi perlunya pendekatan algoritmik yang lebih ekspansif, yang mencakup pemanfaatan beragam teknik pengelompokan dan integrasi fitur yang lebih relevan untuk mencapai segmentasi yang bermakna bisnis [2]. Di antara kekurangan penelitian sebelumnya yang perlu diperhatikan dalam penelitian ini adalah: 1). Ketergantungan pada fitur yang berasal dari data transaksi fundamental, ditambah dengan kurangnya eksplorasi ke dalam fitur perilaku canggih yang dapat meningkatkan interpretabilitas cluster. 2) Tidak adanya pemeriksaan sistematis pengaruh rekayasa fitur pada stabilitas dan kualitas segmentasi dari perspektif bisnis numerik dan praktis. 3). Keterbatasan ada dalam penilaian metodologi segmentasi dengan menghubungkan hasil pengelompokan kuantitatif dengan indikator kemandirian pemasaran, seperti loyalitas pelanggan dan tingkat retensi.

Untuk mengatasi kekurangan ini, penelitian memperkenalkan kerangka segmentasi pelanggan yang menggabungkan atribut fundamental (misalnya, RFM) dengan karakteristik rekayasa inovatif yang berasal dari perilaku pelanggan, yang mencakup metrik seperti skor intensitas pembelian, fluktuasi temporal dalam pembelian, dan volume pembelian. Kerangka kerja ini menjalani validasi melalui implementasi tiga algoritma pengelompokan bebas label—K-Means, Gaussian Mixture Model (GMM), dan DBScan— untuk memastikan sejauh mana rekayasa fitur dapat meningkatkan kualitas cluster, dievaluasi melalui metrik numerik dan kapasitas interpretatif dalam kerangka paradigma pemasaran berbasis data [5].

Solusi yang dikemukakan dalam penelitian ini meliputi: 1). Pendekatan sistematis untuk rekayasa fitur yang bertujuan menghasilkan atribut baru dari data transaksional mentah, yang diharapkan memiliki signifikansi informasi yang cukup besar. 2). Kerangka evaluasi untuk segmentasi yang menggabungkan metrik kuantitatif (Skor Silhouette, Indeks Davies-Bouldin) di samping penilaian kualitas cluster dari sudut pandang komersial. 3). Analisis komparatif yang meneliti kinerja berbagai algoritma pengelompokan untuk menunjukkan dengan tepat algoritma yang menunjukkan sensitivitas tinggi terhadap perubahan dalam fitur yang direkayasa. Melalui fokus ini, penelitian ini memberikan wawasan teoritis dan praktis yang substansif untuk menjelaskan pengaruh rekayasa fitur pada segmentasi pelanggan, sambil menyediakan kerangka kerja serbaguna yang berlaku untuk beragam perusahaan ritel online dan offline di Indonesia dan pasar global.

II. METODE

Investigasi ini menggunakan metodologi kuantitatif eksperimental komparatif untuk menilai dampak rekayasa fitur pada kemandirian segmentasi terkait pelanggan ritel online. Metodologi ini dipilih karena kapasitasnya untuk penilaian obyektif kinerja berbagai algoritma pengelompokan yang didasarkan pada metrik evaluasi internal, menghilangkan kebutuhan untuk data berlabel. Alur kerja penelitian diatur secara sistematis, mencakup tahapan pengumpulan data, analisis data awal, pra-pemrosesan, konstruksi fitur, pemodelan pengelompokan, dan evaluasi hasil. Tahapan metodologi penelitian sebagai unit kohesif diilustrasikan pada Gambar 1.



Gambar 1. Aliran Penelitian

A. Sumber dan Karakteristik Data

Dataset yang digunakan dalam penelitian ini adalah Dataset Ritel Online yang diperoleh dari Repositori Pembelajaran Mesin UCI. Dataset ini merangkum informasi yang berkaitan dengan transaksi penjualan perusahaan ritel online yang terletak di Inggris selama jangka waktu dari Desember 2010 hingga Desember 2011. Dataset terdiri dari 541.909 entri transaksi dengan delapan atribut utama, khususnya InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, dan Country.

Dataset ini banyak digunakan dalam penyelidikan segmentasi pelanggan karena representasinya dari perilaku transaksi aktual dalam lingkungan ritel digital dan atributnya yang kaya yang dapat dianalisis melalui teknik pembelajaran mesin [6], [7]. Atribut transaksi yang tersedia memfasilitasi pembangunan beragam fitur perilaku pelanggan yang berkaitan dengan upaya pengelompokan.

B. Analisis Data dan Pra-Pemrosesan

Tahap awal penelitian dilakukan melalui analisis data eksplorasi (EDA) untuk memahami struktur data, memeriksa distribusi variabel, dan mengidentifikasi contoh nilai nol dan anomali. Beberapa prosedur pra-pemrosesan yang diterapkan meliputi: 1). Penghapusan transaksi yang menunjukkan nilai negatif untuk Quantity dan UnitPrice karena dianggap tidak valid. 2). Penghapusan entri data yang tidak memiliki atribut CustomerId. 3). Transformasi atribut InvoiceDate ke dalam format datetime untuk memfasilitasi perhitungan fitur berbasis waktu seperti kebaruan. 4). Pengecualian entri duplikat untuk menegaskan integritas dataset.

Selain itu, semua fitur numerik menjalani normalisasi menggunakan metode StandardScaler untuk memastikan bahwa setiap fitur mempertahankan skala yang seimbang, sehingga mencegah fitur tunggal mendominasi proses pengelompokan [8]. Normalisasi ini sangat penting untuk meningkatkan stabilitas algoritma pengelompokan berbasis jarak seperti K-Means.

C. Pembentukan Model dan Rekayasa Fitur

Untuk menyelidiki pengaruh rekayasa fitur pada kualitas segmentasi, penelitian ini menggunakan dua pendekatan pemodelan yang berbeda, yaitu:

1) Model 1 - Tanpa Rekayasa Fitur

Model awal menggunakan fitur transaksi fundamental tanpa proses rekayasa fitur canggih. Atribut yang digunakan dalam model ini hanya mencakup: Kuantitas dan Harga satuan. Model ini berfungsi sebagai titik referensi untuk mengevaluasi sejauh mana fitur dasar dapat memfasilitasi segmentasi pelanggan yang efektif tanpa transformasi atau pembentukan fitur tambahan.

2) Model 2 - Dengan Rekayasa Fitur

Model berikutnya menggabungkan proses rekayasa fitur untuk membangun penggambaran perilaku pelanggan yang lebih informatif. Fitur yang dikembangkan meliputi: 1). Recency — contoh terbaru dari transaksi pelanggan. 2). Frekuensi — jumlah transaksi yang dilakukan oleh pelanggan. 3). Moneter — nilai kumulatif dari

transaksi pelanggan. 3). Rata-rata Nilai Pesanan (AOV). 4). Keragaman Produk — bermacam-macam produk yang dibeli oleh pelanggan.

Metodologi ini mengadopsi kerangka kerja RFM (Recency, Frequency, Monetary), yang telah banyak digunakan dalam praktik segmentasi pelanggan kontemporer [9], [10]. Fitur yang menunjukkan distribusi miring mengalami transformasi logaritmik untuk mengurangi efek kemiringan sebelum prosedur normalisasi. Rekayasa fitur diantisipasi untuk merangkum perilaku pelanggan lebih komprehensif daripada fitur transaksi yang tidak diproses.

3) *Algoritma Clustering*

Penelitian menerapkan tiga algoritma pengelompokan tanpa pengawasan yang berbeda untuk menilai sensitivitas algoritma ini dalam kaitannya dengan variasi fitur, khususnya [11]:

- K-Means Pengelompokan Algoritma K-Means digunakan karena kesederhanaan, efisiensi, dan aplikasi yang luas dalam segmentasi pelanggan. Teknik ini mengkategorikan data menurut kedekatannya dengan titik pusat (centroid) cluster [12].
- Model Campuran Gaussian (GMM) GMM digunakan untuk mewakili data sebagai campuran distribusi Gaussian, sehingga menawarkan fleksibilitas yang ditingkatkan dalam mengakomodasi cluster yang menunjukkan geometri non-bola [13].
- DBSCAN (Pengelompokan Spasial Berbasis Kepadatan Aplikasi dengan Kebisingan) DBSCAN dipilih karena kemahirannya dalam mengidentifikasi cluster berbasis kepadatan sambil secara bersamaan mengenali data outlier tanpa memerlukan penentuan sebelumnya dari jumlah cluster [14].

Ketiga algoritma dieksekusi pada Model 1 dan Model 2 untuk memastikan dampak variasi fitur pada kemanjuran setiap algoritma.

4) *Penilaian Hasil Pengelompokan*

Penilaian kualitas cluster dilakukan dengan menggunakan dua metrik internal yang biasa digunakan dalam domain penelitian clustering, khususnya:

- Skor Siluet Metrik ini mengukur tingkat kohesi intra-cluster dan pemisahan antar-cluster. Nilai mendekati 1 menandakan kualitas pengelompokan yang tinggi.
- Indeks Davies-Bouldin (DBI) DBI mengevaluasi proporsi jarak intra-cluster ke jarak antar-cluster. Nilai DBI yang lebih rendah menunjukkan kualitas unggul dalam pengelompokan yang dihasilkan [15].

Dalam hubungannya dengan penilaian kuantitatif, analisis kualitatif juga dilakukan melalui interpretasi karakteristik cluster yang didirikan, untuk memastikan apakah segmentasi yang dihasilkan memiliki implikasi bisnis yang signifikan, seperti identifikasi pelanggan bernilai tinggi, pelanggan aktif, dan pelanggan pasif. Metodologi evaluasi ganda ini sejalan dengan rekomendasi penelitian kontemporer yang menggarisbawahi perlunya menjaga keseimbangan antara validitas numerik dan interpretabilitas bisnis [16].

III. HASIL DAN PEMBAHASAN

Bagian ini menggambarkan hasil eksperimen segmentasi pelanggan yang dilakukan menggunakan dua metodologi fitur yang berbeda, khususnya tanpa rekayasa fitur (Model 1) dan menggabungkan rekayasa fitur berbasis perilaku (Model 2). Analisis dilakukan dari perspektif ganda, mencakup penilaian kuantitatif melalui metrik evaluasi internal, di samping interpretasi kualitatif yang berasal dari profil cluster. Tujuan evaluasi adalah untuk memastikan sejauh mana proses rekayasa fitur mempengaruhi kualitas cluster yang dihasilkan oleh algoritma K-Means, Gaussian Mixture Model (GMM), dan DBSCAN yang diterapkan pada data pelanggan ritel online.

CustomerID	Quantity	UnitPrice
12346	74215	1.04
12347	12	2.10
12348	72	0.55
12349	2	7.50
12350	12	2.10

CustomerID	Quantity	UnitPrice
12346	65.364946	-0.114401
12347	-0.025902	-0.083531
12348	0.026972	-0.128671
12349	-0.034715	0.073730
12350	-0.025902	-0.083531

Gambar 2. Fitur Rekayasa Proses

A. *Hasil Eksperimen Segmentasi*

Eksperimen dilakukan dengan menerapkan tiga algoritma pengelompokan di dua model fitur yang berbeda. Model 1 menggunakan fitur transaksi fundamental (Quantity dan UnitPrice) tanpa penerapan rekayasa fitur canggih, sedangkan Model 2 memanfaatkan fitur rekayasa, termasuk RFM, Nilai Pesanan Rata-rata (AOV), dan keragaman produk. Kualitas cluster dinilai menggunakan dua metrik internal, Skor Silhouette dan Davies-Bouldin Index (DBI),

seperti yang dianjurkan dalam literatur yang berkaitan dengan studi segmentasi tanpa pengawasan. Tabel 1 merupakan hasil kinerja model 1 dan model 2.

TABEL 1
PERBANDINGAN METRIK EVALUASI KUANTITATIF ANTARA MODEL 1 DAN MODEL 2

Algorithm Cluster	Silhouette Score (Model 1)	Silhouette Score (Model 2)	Davies-Bouldin Index (Model 1)	Davies-Bouldin Index (Model 2)
K-means	0.981	0.337	0.171	1.010
GMM	0.027	0.330	0.827	0.992
DBSCAN	0.981	0.300	0.093	1.046

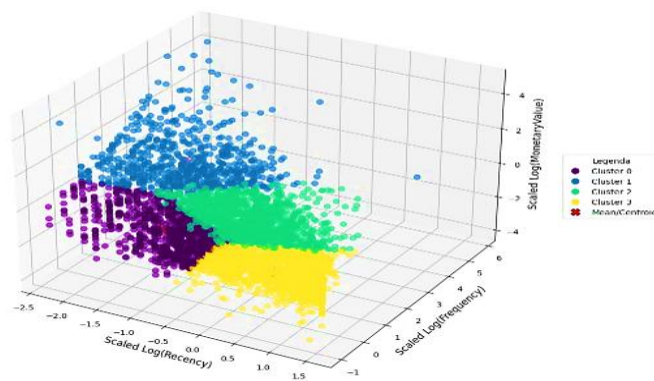
Hasil evaluasi menunjukkan perbedaan kinerja yang nyata antara dua model fitur. Dalam Model 1, algoritma K-Means dan DBSCAN mencapai nilai Skor Silhouette yang sangat tinggi (0,981) disertai dengan nilai DBI yang rendah. Secara numerik, ini menunjukkan pembentukan cluster yang kompak dan terpisah dengan jelas. Sebaliknya, kinerja algoritma GMM sangat rendah, menghasilkan nilai Silhouette hanya 0,027, yang menyiratkan bahwa asumsi distribusi Gaussian tidak cocok untuk fitur transaksi dasar.

Sebaliknya, dalam Model 2, nilai Skor Silhouette terletak dalam kisaran yang lebih moderat (0,300-0,337) di ketiga algoritma. Meskipun hasil ini secara numerik lebih rendah daripada yang diamati pada Model 1, hasil dalam Model 2 menunjukkan homogenitas dan konsistensi yang lebih besar. Nilai DBI mendekati 1 menandakan bahwa cluster yang terbentuk lebih rumit, namun juga lebih seimbang dan realistis dalam representasi perilaku pelanggan mereka.

B. Analisis Kuantitatif Hasil Pengelompokan

Dari sudut pandang kuantitatif, temuan ini membuktikan bahwa peningkatan nilai metrik evaluasi internal tidak selalu sesuai dengan segmentasi yang memiliki relevansi bisnis. Dalam Model 1, Skor Silhouette yang meningkat sebagian besar disebabkan oleh dimensi rendah dari fitur yang digunakan. Dengan hanya dua fitur mendasar, jarak antar-cluster menjadi terlalu besar, menghasilkan metrik internal yang cenderung miring dan tidak secara akurat mencerminkan pola perilaku pelanggan asli.

Fenomena ini menggambarkan kendala evaluasi yang bergantung secara eksklusif pada metrik numerik, terutama dengan kumpulan data dimensi rendah. Temuan yang sebanding telah diartikulasikan dalam studi sebelumnya, yang menegaskan bahwa metrik internal seringkali tidak cukup representatif ketika diterapkan pada fitur transaksi mentah.



Gambar 3. Visualisasi 3D Hasil Pengelompokan K-Means dalam Model Berbasis RFM 2

Dalam Model 2, meskipun nilai metrik mencerminkan rentang moderat, kinerja konsisten yang ditunjukkan di seluruh algoritma yang beragam menunjukkan kemanjuran rekayasa fitur dalam pengembangan representasi data yang lebih otentik. Fitur perilaku yang didasarkan pada analisis Terbaru, Frekuensi, dan Moneter (RFM), serta Nilai Pesanan Rata-rata (AOV), memberdayakan algoritme untuk menangkap pola perilaku pelanggan secara menyeluruh. Pengamatan ini menguatkan penelitian ilmiah kontemporer yang menegaskan bahwa rekayasa fitur berperan penting dalam meningkatkan stabilitas dan interpretabilitas hasil pengelompokan.

C. Analisis Kualitatif dan Interpretasi Cluster

Di luar penilaian kuantitatif, analisis kualitatif dilakukan melalui interpretasi profil cluster yang berasal dari Model 2, khususnya temuan dari algoritma K-Means. Implementasi visualisasi tiga dimensi yang menggunakan fitur Recency, Frekuensi, dan Moneter menjelaskan pemisahan cluster yang berbeda, yang kemudian dapat dikontekstualisasikan dalam kerangka bisnis. Dari analisis interpretatif cluster, beberapa segmen pelanggan yang signifikan digambarkan, termasuk:

- a. Pelanggan bernilai tinggi: ditandai dengan kontribusi moneter yang besar, frekuensi transaksi yang tinggi, dan pembaruan minimal.
- b. Pelanggan aktif reguler: ditentukan oleh frekuensi transaksi moderat ditambah dengan nilai moneter yang konsisten.
- c. Pelanggan pasif: diidentifikasi sebagai transaktor yang jarang dengan tingkat pengeluaran rendah.

Strategi segmentasi ini terbukti jauh lebih relevan dengan kebutuhan strategi pemasaran dibandingkan dengan hasil Model 1, yang hanya mengkategorikan pelanggan berdasarkan variasi ekstrem dalam atribut transaksional mendasar. Pengamatan ini mengindikasikan bahwa efektivitas segmentasi tidak hanya bergantung pada metrik internal tetapi juga pada kapasitas cluster untuk secara otentik mewakili perilaku pelanggan dalam istilah praktis.

D. Perbandingan Kinerja Antar Algoritma

Hasil empiris menunjukkan bahwa algoritma menunjukkan respons yang bervariasi terhadap intervensi rekayasa fitur. Dalam Model 1, baik K-Means dan DBSCAN menunjukkan keunggulan numerik; Namun, kinerja mereka menunjukkan penurunan pada Model 2. Sebaliknya, algoritma Gaussian Mixture Model (GMM) menunjukkan peningkatan kinerja ketika diterapkan pada Model 2, meskipun tetap relatif terbatas dalam istilah numerik.

Hasil ini mendukung pernyataan bahwa tidak ada algoritma pengelompokan yang memiliki keunggulan universal. Kemanjuran suatu algoritma terutama dipengaruhi oleh atribut fitur dan distribusi data yang digunakan. Akibatnya, pemilihan algoritma harus memperhitungkan keselarasan antara tujuan analitis, karakteristik struktural data, dan sifat fitur yang diterapkan.

E. Diskusi dan Implikasi Penelitian

Temuan penyelidikan ini menjelaskan defisit yang signifikan dalam metodologi segmentasi pelanggan, khususnya keunggulan evaluasi berdasarkan metrik numerik yang mengabaikan relevansi bisnis. Studi ini membuktikan bahwa:

- a. Rekayasa fitur tidak selalu menambah nilai metrik evaluatif internal secara langsung.
- b. Rekayasa fitur secara nyata meningkatkan kualitas interpretatif dan penerapan praktis dari hasil segmentasi.
- c. Evaluasi segmentasi memerlukan pendekatan integratif yang menggabungkan perspektif kuantitatif dan kualitatif untuk meningkatkan relevansinya untuk pengambilan keputusan bisnis.

Konsekuensi praktis dari penelitian ini menunjukkan bahwa perusahaan ritel harus menghindari ketergantungan pada atribut transaksional yang belum sempurna untuk segmentasi pelanggan. Penggabungan fitur perilaku yang lebih bernuansa memfasilitasi identifikasi yang tepat dari segmen pelanggan dan mendukung penerapan strategi pemasaran yang lebih personal dan bertarget. Metodologi ini sejalan dengan rekomendasi ilmiah terbaru mengenai keharusan sinergi antara evaluasi algoritmik dan nilai bisnis dalam analisis segmentasi. Dari sudut pandang akademis, penelitian ini berkontribusi melalui:

- a. Analisis komparatif yang menggambarkan dampak rekayasa fitur pada tiga algoritma pengelompokan yang lazim.
- b. Validasi empiris bahwa stabilitas dan interpretabilitas cluster memiliki signifikansi yang lebih besar daripada nilai-nilai metrik internal belaka.
- c. Kerangka metodologis yang dapat direplikasi yang kondusif untuk penelitian segmentasi pelanggan berbasis data transaksional.

IV. SIMPULAN

Investigasi ini menjelaskan bahwa implementasi rekayasa fitur memberikan pengaruh yang cukup besar pada kemanjuran segmentasi pelanggan yang berasal dari metodologi pengelompokan. Temuan empiris menunjukkan bahwa penggabungan fitur transaksi mendasar menghasilkan nilai yang meningkat untuk metrik evaluasi internal; Namun, fitur-fitur ini kurang mahir dalam merangkum pola perilaku pelanggan secara akurat. Sebaliknya, model yang didasarkan pada fitur rekayasa, seperti metrik Terbaru, Frekuensi, dan Moneter (RFM), Nilai Pesanan Rata-rata (AOV), dan keragaman produk, memuncak dalam cluster yang menunjukkan peningkatan stabilitas, konsistensi, dan interpretabilitas untuk aplikasi bisnis, meskipun nilai metrik kuantitatif mereka relatif moderat. Hasil ini mendukung gagasan bahwa kemanjuran segmentasi pelanggan tidak hanya bergantung pada metrik kinerja numerik tetapi juga pada relevansi fitur dalam kerangka proses pengambilan keputusan pemasaran. Untuk pertanyaan selanjutnya, disarankan untuk memasukkan atribut kontekstual, termasuk pengaruh musiman dan preferensi kategori produk, serta

untuk menyelidiki teknik pengelompokan lanjutan seperti pengelompokan ensemble atau metodologi pembelajaran mendalam. Selain itu, evaluasi eksternal yang didasarkan pada indikator bisnis otentik harus diintegrasikan untuk menambah validitas praktis dari hasil segmentasi.

DAFTAR PUSTAKA

- [1] R. M. Fauzan and G. Alfian, "Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 9, no. 3, pp. 170–177, Sep. 2024, doi: 10.14421/jiska.2024.9.3.170-177.
- [2] J. Salminen, M. Mustak, M. Sufyan, and B. J. Jansen, "How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation," *J. Mark. Anal.*, vol. 11, no. 4, pp. 677–692, Dec. 2023, doi: 10.1057/s41270-023-00235-5.
- [3] N. Kumar, "Intelligent customer segmentation: unveiling consumer patterns with machine learning," *J. Umm Al-Qura Univ. Eng. Archit.*, vol. 16, no. 3, pp. 774–783, Sep. 2025, doi: 10.1007/s43995-025-00180-7.
- [4] Z. Wang, "Customer Segmentation Based on Machine Learning Methods," *Highlights Sci. Eng. Technol.*, vol. 92, pp. 126–132, Apr. 2024, doi: 10.54097/g70xqb16.
- [5] A. M. Haddadi and H. Hamidi, "A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm," *Comput. Hum. Behav. Reports*, vol. 18, p. 100616, May 2025, doi: 10.1016/j.chbr.2025.100616.
- [6] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.
- [7] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Inf. Syst. E-bus. Manag.*, vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.
- [8] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [9] Z.-J. Lee, C.-Y. Lee, L.-Y. Chang, and N. Sano, "Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation," *Symmetry (Basel)*, vol. 13, no. 9, p. 1557, Aug. 2021, doi: 10.3390/sym13091557.
- [10] S. Monalisa, Y. Juniarti, E. Saputra, F. Muttakin, and T. K. Ahsyar, "Customer segmentation with RFM models and demographic variable using DBSCAN algorithm," *TELKOMNIKA (Telecommunication Comput. Electron. Control)*, vol. 21, no. 4, p. 742, Aug. 2023, doi: 10.12928/telkomnika.v21i4.22759.
- [11] H. Ramadhan, M. R. Abdan Kamaludin, M. A. Nasrullah, and D. Rolliawati, "Comparison of Hierarchical, K-Means and DBSCAN Clustering Methods for Credit Card Customer Segmentation Analysis Based on Expenditure Level," *J. Appl. Informatics Comput.*, vol. 7, no. 2, pp. 246–251, Dec. 2023, doi: 10.30871/jaic.v7i2.5790.
- [12] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.
- [13] Y. Zhang *et al.*, "Gaussian Mixture Model Clustering with Incomplete Data," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, no. 1s, pp. 1–14, Jan. 2021, doi: 10.1145/3408318.
- [14] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Sep. 2020, pp. 949–953, doi: 10.1109/IFEEA51475.2020.00199.
- [15] - William and M. E. Johan, "Implementation of Customer Segmentation Model using K-Means and DBSCAN for Fashion Industry Product Transaction," *JOIV Int. J. Informatics Vis.*, vol. 9, no. 6, p. 2559, Nov. 2025, doi: 10.62527/joiv.9.6.2978.
- [16] A. S. Paramita and T. Hariguna, "Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 43–62, Jun. 2024, doi: 10.47738/jdmcc.v1i1.3.