PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 918-926
E-ISSN: 3109-8886

Segmentasi Pelanggan dan Prediksi Churn E-Commerce Menggunakan K-Means Clustering dan Random Forest: Studi Kasus Olist Brazil

[1]Muhamad Yumni, [2]Devira Nazra Suhendra, [3]Najwa Rena Amanda

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Banten

[1]muhammadmumin082@gmail.com, [2]deviranazrasuhendra@gmail.com, [3]najwrenaamanda@gmail.com

Abstract

Among 93,357 customers analyzed from the Olist Brazil e-commerce platform, nearly four in ten were found to be in a state of permanent churn a condition invisible to conventional transaction reporting without data-driven segmentation. This study proposes a two-stage analytical pipeline integrating RFM-based (Recency, Frequency, Monetary) K-Means Clustering with a Random Forest Classifier for churn prediction, structured within the CRISP-DM framework. Data were drawn from the Olist Brazilian E-Commerce Public Dataset covering 115,653 orders between 2016 and 2018. Churn was operationalized as customers with recency exceeding 180 days and a transaction frequency of one, yielding a churn proportion of 56.4% across the sample. Clustering at K=4 (Silhouette Score=0.526) partitioned customers into four behaviorally distinct segments: Active (53%, churn rate 29%), Lost (39%, churn rate 100%), Big Spender (4%, churn rate 60%), and Loyal (3%, churn rate 0%). Cluster labels were subsequently incorporated as input features into the Random Forest model a design decision that proved consequential, as the cluster variable emerged as the single strongest predictor with a feature importance score of 0.826, outweighing all individual behavioral features combined. The model achieved an ROC-AUC of 0.897, accuracy of 82.9%, precision of 97.7%, recall of 71.5%, and F1-Score of 82.6%. These results demonstrate that customer segmentation, when embedded within a predictive pipeline rather than used in isolation, yields substantial gains in churn detection capability.

Keywords: *Customer Segmentation, Churn Prediction, K-Means Clustering, Random Forest, RFM, E-Commerce, CRISP-DM*

Abstrak

Dari 93.357 pelanggan yang dianalisis pada platform e-commerce Olist Brasil, hampir empat dari sepuluh berada dalam kondisi churn permanen sebuah kondisi yang tidak terdeteksi melalui pelaporan transaksi konvensional tanpa pendekatan segmentasi berbasis data. Penelitian ini mengusulkan pipeline analitik dua tahap yang mengintegrasikan K-Means Clustering berbasis RFM (Recency, Frequency, Monetary) dengan Random Forest Classifier untuk prediksi churn, mengikuti kerangka kerja CRISP-DM. Data bersumber dari Olist Brazilian E-Commerce Public Dataset yang mencakup 115.653 pesanan periode 2016–2018. Churn didefinisikan sebagai kondisi pelanggan dengan recency melebihi 180 hari dan frekuensi transaksi satu kali, menghasilkan proporsi churn sebesar 56,4% dari total sampel. Proses clustering dengan K=4 (Silhouette Score=0,526) berhasil memilah pelanggan ke dalam empat segmen yang terbedakan secara perilaku: Active (53%, churn rate 29%), Lost (39%, churn rate 100%), Big Spender (4%, churn rate 60%), dan Loyal (3%, churn rate 0%). Label segmen kemudian diintegrasikan sebagai fitur masukan dalam model Random Forest, dan hasilnya signifikan variabel cluster menjadi prediktor tunggal terkuat dengan feature importance 0,826, mengungguli seluruh fitur perilaku individual. Model mencapai ROC-AUC 0,897, akurasi 82,9%, precision 97,7%, recall 71,5%, dan F1-Score 82,6%. Temuan ini mengonfirmasi bahwa segmentasi pelanggan bukan sekadar alat deskriptif, melainkan kontributor prediktif yang substansial ketika diintegrasikan dalam pipeline klasifikasi.

Kata Kunci: Segmentasi Pelanggan, Prediksi Churn, K-Means Clustering, Random Forest, RFM, E-Commerce, CRISP-DM

A. PENDAHULUAN

Brasil mencatat volume transaksi e-commerce senilai USD 49,2 miliar pada 2021, menjadikannya pasar digital terbesar di Amerika Latin sekaligus salah satu dari sepuluh pasar e-commerce terbesar di dunia [1]. Angka ini tumbuh dari basis yang jauh lebih kecil pada 2016, bertepatan dengan periode data yang digunakan dalam penelitian ini, sehingga dataset Olist Brazilian E-Commerce secara tepat merepresentasikan fase ekspansi yang dinamis dan penuh tekanan kompetitif. Di tengah pertumbuhan tersebut, salah satu persoalan struktural yang dihadapi platform marketplace adalah tingginya tingkat pelanggan yang tidak pernah kembali bertransaksi setelah pembelian pertama. Fenomena ini bukan sekadar statistik pasif, melainkan sinyal dini dari erosi basis pelanggan yang, jika dibiarkan

Diterima Redaksi: 10-06-2026 | Selesai Revisi: 22-06-2026 | Diterbitkan Online: 23-06-2026

tanpa intervensi, akan menekan margin dan meningkatkan beban akuisisi secara bersamaan.

Customer churn dalam konteks e-commerce memiliki karakter yang berbeda dari industri berlangganan seperti telekomunikasi atau perbankan. Tidak ada kontrak yang putus, tidak ada notifikasi pembatalan pelanggan cukup berhenti memesan, dan platform sering kali tidak menyadari kehilangan tersebut hingga jendela waktu untuk intervensi sudah terlewat. Penelitian pada platform marketplace multiseller menunjukkan bahwa pelanggan yang berpotensi churn umumnya menunjukkan pola penurunan frekuensi transaksi yang dapat dideteksi jauh sebelum mereka benar-benar berhenti, asalkan data perilaku historis dianalisis secara sistematis [3]. Fakta ini menegaskan bahwa deteksi dini bukan sekadar keunggulan operasional, melainkan prasyarat bagi strategi retensi yang efektif secara biaya.

Pendekatan data mining berbasis perilaku historis telah terbukti mampu mengungkap pola tersembunyi yang tidak terlihat melalui analisis konvensional. Karimah dan Marwati [4] mendemonstrasikan bahwa model data mining yang dibangun dari rekam jejak aktivitas historis mampu menghasilkan prediksi yang akurat dan dapat ditindaklanjuti secara praktis, bahkan pada domain di luar e-commerce. Prinsip yang sama berlaku pada konteks pelanggan: perilaku transaksi masa lalu seberapa baru, seberapa sering, dan seberapa besar nilainya merupakan prediktor yang konsisten untuk perilaku di masa mendatang. Kerangka RFM (Recency, Frequency, Monetary) mengoperasionalkan ketiga dimensi ini menjadi fitur kuantitatif yang dapat diproses oleh algoritma machine learning [5].

Namun, RFM sebagai metode deskriptif semata memiliki batas yang jelas: ia dapat menggambarkan posisi pelanggan saat ini, tetapi tidak menjawab siapa di antara mereka yang akan pergi besok. Di sinilah integrasi antara clustering berbasis RFM dan model prediksi berbasis ensemble menjadi relevan. K-Means Clustering mempartisi pelanggan ke dalam kelompok-kelompok yang homogen secara internal berdasarkan jarak Euclidean pada ruang fitur ternormalisasi, sementara Random Forest mampu menangkap hubungan non-linear antara fitur perilaku dan probabilitas churn dengan tingkat akurasi yang kompetitif dibanding algoritma lain [6]. Kombinasi keduanya memungkinkan label segmen hasil clustering digunakan sebagai fitur tambahan dalam model prediktif, mengintegrasikan informasi segmental ke dalam keputusan pada level individu pelanggan sebuah pendekatan yang masih jarang dieksplorasi secara eksplisit dalam literatur berbahasa Indonesia.

Dataset yang digunakan adalah Olist Brazilian E-Commerce Public Dataset, mencakup 115.653 pesanan dari September 2016 hingga Oktober 2018 dalam lima tabel relasional yang merepresentasikan pesanan, pelanggan, item produk, pembayaran, dan ulasan [7]. Setelah filtering pada pesanan berstatus *delivered* dan pembersihan data secara menyeluruh, sebanyak 93.357 pelanggan unik dianalisis. Olist beroperasi sebagai penghubung antara penjual kecil dan konsumen akhir melalui kanal marketplace utama Brasil, sehingga dinamika datanya mencerminkan pola belanja ritel nyata yang kompleks, melibatkan ribuan kategori produk dan ratusan kota.

Penelitian ini dirancang dengan tiga tujuan yang saling terhubung. Pertama, membentuk segmen pelanggan yang bermakna secara bisnis menggunakan K-Means Clustering pada fitur RFM, dengan menentukan jumlah cluster optimal melalui Elbow Method dan Silhouette Score. Kedua, membangun model prediksi churn menggunakan Random Forest Classifier yang memanfaatkan label segmen sebagai salah satu fitur masukan, dievaluasi menggunakan ROC-AUC, precision, recall, dan F1-score. Ketiga, menerjemahkan temuan kuantitatif tersebut ke dalam rekomendasi strategi retensi yang spesifik per segmen, sehingga hasil penelitian memiliki nilai aplikatif langsung bagi pengelola platform. Seluruh alur kerja penelitian ini diorganisasikan menggunakan CRISP-DM sebagai panduan iteratif, yang dalam implementasinya pada dataset Olist mencakup tahap pemahaman karakteristik marketplace Brasil, eksplorasi dan pembersihan data transaksi, rekayasa fitur RFM, pembangunan dan evaluasi model, hingga penerjemahan hasil ke dalam rekomendasi yang dapat dieksekusi [8].

B. METODE

Penelitian ini mengikuti kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang terdiri dari enam fase iteratif: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penyebaran [8]. Pemilihan kerangka ini didasarkan pada sifatnya yang agnostik terhadap domain dan kemampuannya mengakomodasi siklus analisis yang berulang, di mana temuan pada fase lebih lanjut dapat memicu penyesuaian pada fase sebelumnya. Schröer, Kruse, dan Gómez [8] menegaskan bahwa CRISP-DM tetap menjadi kerangka proses data mining yang paling banyak diadopsi dalam penelitian akademik maupun industri hingga saat ini.

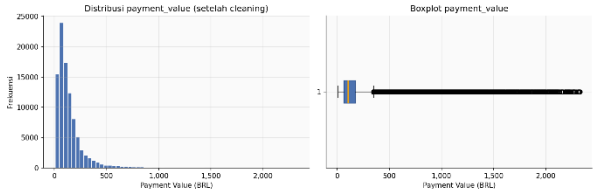
1 Sumber Data

Data yang digunakan bersumber dari Olist Brazilian E-Commerce Public Dataset yang tersedia secara publik di platform Kaggle [7]. Dataset ini terdiri dari lima tabel relasional yang saling terhubung melalui kunci primer dan kunci asing: tabel *orders* yang memuat informasi status dan waktu pesanan, tabel *customers* yang berisi identitas dan lokasi pelanggan, tabel *order_items* yang merinci produk dan harga per item, tabel *order_payments* yang mencatat metode dan nilai pembayaran, serta tabel *order_reviews* yang menyimpan skor ulasan dan komentar dari pelanggan. Cakupan data meliputi periode September 2016 hingga Oktober 2018 dengan total 115.653 entri pesanan sebelum proses pembersihan.

2 Persiapan Data

Proses persiapan data dimulai dengan penggabungan (*join*) kelima tabel menggunakan *order_id* dan *customer_id* sebagai kunci relasi, menghasilkan satu tabel analitik terpadu. Selanjutnya, filtering diterapkan hanya pada pesanan dengan status *delivered*, karena hanya transaksi yang benar-benar selesai yang relevan untuk analisis perilaku pelanggan. Baris dengan nilai hilang pada kolom kritis seperti *order_delivered_customer_date* dan *review_score* dihapus sepenuhnya, sementara nilai hilang pada kolom non-kritis diimputasi menggunakan median kolom bersangkutan. Duplikasi pada *order_id* juga dieliminasi untuk memastikan setiap transaksi dihitung tepat satu kali.

Penanganan outlier diterapkan pada variabel *payment_value*: nilai yang sama dengan atau lebih kecil dari nol dihapus karena tidak memiliki makna bisnis, sementara nilai di atas persentil ke-99,9 (P99.9) turut dieliminasi untuk mencegah distorsi pada proses normalisasi dan clustering. Keputusan menggunakan P99.9 bukan P99 atau P95 diambil secara sadar untuk mempertahankan sebanyak mungkin pelanggan bernilai tinggi yang memang memiliki pengeluaran besar secara legitim, sambil tetap membuang nilai ekstrem yang kemungkinan besar merupakan anomali data.



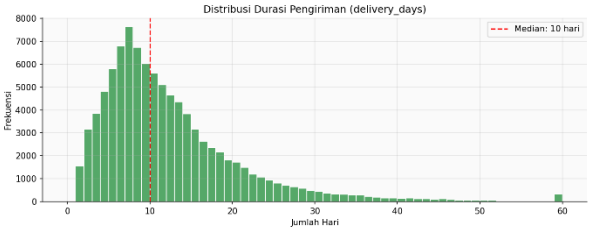
Gambar 1. Distribusi dan Boxplot Payment Value Setelah Penanganan Outlier

Setelah seluruh tahap pembersihan selesai, diperoleh 93.357 pelanggan unik yang menjadi unit analisis dalam penelitian ini.

3 Feature Engineering

Tiga fitur utama dibangun dari data transaksi bersih menggunakan kerangka RFM. *Recency* didefinisikan sebagai selisih dalam satuan hari antara tanggal referensi ditetapkan sebagai tanggal maksimum dalam dataset dengan tanggal pembelian terakhir pelanggan; nilai recency yang lebih kecil menandakan pelanggan yang lebih baru bertransaksi. *Frequency* dihitung sebagai total jumlah pesanan unik yang pernah dilakukan pelanggan selama periode observasi. *Monetary* dihitung sebagai rata-rata nilai pembayaran per transaksi, bukan total kumulatif, untuk menghindari bias terhadap pelanggan lama yang secara alami memiliki lebih banyak kesempatan bertransaksi.

Selain fitur RFM, dua fitur tambahan direkayasa untuk memperkaya representasi perilaku pelanggan. *avg_review_score* merupakan rata-rata skor ulasan yang diberikan pelanggan pada seluruh pesanannya, mencerminkan dimensi kepuasan. *avg_delivery_delay* dihitung sebagai rata-rata selisih hari antara tanggal pengiriman aktual dan tanggal estimasi pengiriman yang dijanjikan; nilai positif menandakan keterlambatan, nilai negatif menandakan pengiriman lebih awal dari estimasi.



Gambar 2. Distribusi Durasi Pengiriman Pesanan

Label churn didefinisikan secara biner: pelanggan dikategorikan sebagai churn (1) apabila nilai recency-nya melebihi 180 hari dan frekuensi transaksinya hanya satu kali, dan sebagai non-churn (0) apabila sebaliknya. Ambang 180 hari dipilih berdasarkan pertimbangan bahwa jendela waktu enam bulan merupakan periode yang lazim digunakan sebagai batas inaktivitas dalam studi retensi e-commerce [2], sekaligus relevan secara operasional mengingat mayoritas produk pada platform Olist bersifat konsumsi berulang dalam siklus bulanan hingga kuartalan.

4 Normalisasi dan Persiapan untuk Clustering

K-Means Clustering sensitif terhadap perbedaan skala antar fitur karena algoritma ini menggunakan jarak Euclidean sebagai ukuran kemiripan [5]. Fitur dengan rentang nilai yang lebih besar akan mendominasi perhitungan jarak dan mengaburkan kontribusi fitur lain. Oleh karena itu, ketiga fitur RFM dinormalisasi menggunakan MinMaxScaler yang mentransformasikan setiap nilai ke dalam rentang [0, 1] berdasarkan persamaan berikut:

x′=x−xminxmax−xminx' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}x′=xmax−xminx−xmin

Normalisasi hanya diterapkan pada fitur RFM yang digunakan dalam proses clustering, sementara fitur tambahan (*avg_review_score* dan *avg_delivery_delay*) dipertahankan dalam skala aslinya untuk tahap klasifikasi.

5 K-Means Clustering dan Pemilihan K Optimal

Algoritma K-Means bekerja dengan cara menginisialisasi sejumlah K centroid secara acak, kemudian secara iteratif mengalokasikan setiap titik data ke centroid terdekat dan memperbarui posisi centroid berdasarkan rata-rata anggota klusternya, hingga konvergensi tercapai [5]. Inisialisasi centroid menggunakan metode *k-means++* untuk mempercepat konvergensi dan mengurangi sensitivitas terhadap inisialisasi acak.

Pemilihan jumlah cluster optimal dilakukan menggunakan dua metrik secara bersamaan. Pertama, Elbow Method yang memplot nilai inersia (jumlah kuadrat jarak dalam cluster) terhadap berbagai nilai K dari 2 hingga 9; titik "siku" pada kurva mengindikasikan titik di mana penambahan cluster tidak lagi memberikan penurunan inersia yang signifikan. Kedua, Silhouette Score yang mengukur seberapa mirip suatu titik dengan clusternya

sendiri dibandingkan dengan cluster lain, dengan rentang nilai dari -1 (salah penempatan) hingga +1 (penempatan sempurna) [9].

Eksperimen dijalankan dengan mengevaluasi nilai K dari 2 hingga 9 secara berurutan. Untuk setiap nilai K, dihitung nilai inersia sebagai ukuran kompaktness cluster, serta Silhouette Score sebagai ukuran kualitas pemisahan antar cluster. Kedua metrik ini kemudian dibandingkan secara bersamaan untuk menghindari kelemahan masing-masing metode apabila digunakan secara terpisah Elbow Method cenderung subjektif dalam menentukan titik siku, sementara Silhouette Score dapat memberikan nilai tinggi pada K yang terlalu kecil jika distribusi data tidak seimbang. Hasil perbandingan kedua metrik ini disajikan dan dibahas lebih lanjut pada bagian Hasil dan Pembahasan.

6 Definisi Label dan Profil Cluster

Setelah K optimal ditentukan dan clustering dijalankan, setiap cluster dianalisis berdasarkan nilai rata-rata ketiga fitur RFM serta churn rate-nya. Dari karakteristik tersebut, masing-masing cluster diberi label bisnis yang deskriptif untuk memudahkan interpretasi dan komunikasi hasil kepada pemangku kepentingan. Label cluster kemudian diencode sebagai fitur numerik ordinal dan dimasukkan ke dalam tahap pemodelan klasifikasi.

7 Random Forest Classifier

Pada penelitian ini, prediksi churn dimodelkan menggunakan Random Forest pendekatan yang bekerja layaknya pemungutan suara dari ratusan model keputusan yang masing-masing dilatih pada subset data berbeda, sehingga keputusan akhir lebih stabil dan tidak bergantung pada satu pola tunggal, di mana setiap pohon dilatih pada subsampel data yang berbeda dan prediksi akhir ditentukan melalui voting mayoritas [6]. Keunggulan utamanya terletak pada ketahanannya terhadap overfitting, kemampuan menangani fitur dengan skala berbeda, serta kapasitasnya menghasilkan peringkat kepentingan fitur (*feature importance*) yang dapat diinterpretasikan.

Fitur yang digunakan sebagai masukan model klasifikasi terdiri dari lima variabel: *frequency*, *monetary*, *avg_review_score*, *avg_delivery_delay*, dan *cluster*. Variabel *recency* secara eksplisit dikecualikan dari model karena berpotensi menyebabkan *data leakage* mengingat definisi label churn itu sendiri berbasis nilai recency, menyertakan recency sebagai fitur masukan akan membuat model belajar dari informasi yang sudah tertanam langsung dalam label, sehingga performa evaluasi menjadi tidak valid.

Ketidakseimbangan kelas (*class imbalance*) ditangani menggunakan parameter *class_weight='balanced'* pada implementasi Random Forest, yang secara otomatis menyesuaikan bobot setiap kelas berbanding terbalik dengan frekuensinya dalam data latih. Pendekatan ini dipilih dibandingkan teknik oversampling seperti SMOTE karena lebih sederhana secara komputasional dan tidak memperkenalkan data sintetis yang dapat mengubah distribusi asli [10]. Data dibagi menjadi set latih dan set uji dengan rasio 80:20 menggunakan stratified splitting untuk memastikan proporsi kelas terjaga di kedua subset.

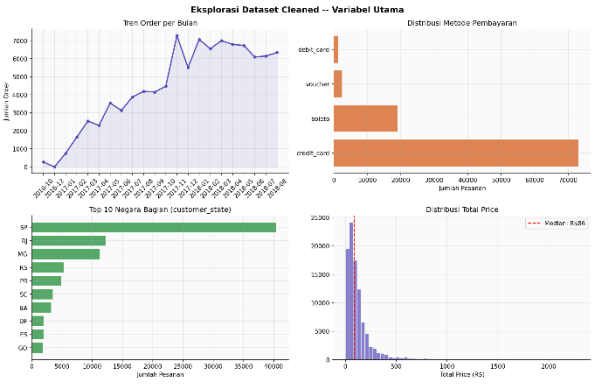
8 Evaluasi Model

Performa model dievaluasi menggunakan empat metrik utama yang saling melengkapi. *Accuracy* mengukur proporsi prediksi yang benar secara keseluruhan. *Precision* mengukur proporsi prediksi churn yang benar-benar merupakan churn, relevan untuk menghindari pemborosan sumber daya retensi pada pelanggan yang sebenarnya tidak akan pergi. *Recall* mengukur proporsi pelanggan churn sesungguhnya yang berhasil terdeteksi oleh model, penting untuk meminimalkan pelanggan yang terlewat. *F1-Score* menyajikan rata-rata harmonik antara precision dan recall sebagai metrik tunggal yang menyeimbangkan keduanya. Selain itu, kurva ROC dan nilai AUC (*Area Under Curve*) digunakan untuk mengukur kemampuan diskriminatif model secara keseluruhan pada berbagai ambang klasifikasi [9].

C. HASIL DAN PEMBAHASAN

1 Eksplorasi Data Awal

Setelah proses penggabungan kelima tabel dan filtering pada status pesanan *delivered*, diperoleh dataset terpadu yang siap untuk dianalisis. Eksplorasi awal menunjukkan bahwa distribusi pesanan tidak merata secara temporal volume transaksi meningkat signifikan mulai pertengahan 2017 dan mencapai puncaknya pada kuartal keempat 2017, yang bertepatan dengan periode belanja akhir tahun. Secara geografis, konsentrasi pelanggan terpusat di negara bagian São Paulo dan Rio de Janeiro, mencerminkan distribusi populasi dan infrastruktur digital Brasil yang tidak merata antar wilayah.



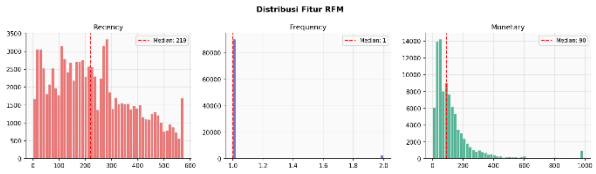
Gambar 3. Hasil Eksplorasi Data Awal Dataset Olist

Dari total 115.653 entri pesanan awal, proses pembersihan data menghasilkan 93.357 pelanggan unik yang layak dianalisis. Sebagian besar pelanggan sekitar 97% hanya melakukan satu kali transaksi selama periode observasi, sebuah karakteristik yang umum ditemukan pada platform e-commerce ritel generalis dan sekaligus menjadi motivasi utama penelitian ini: memahami siapa di antara pelanggan

satu kali tersebut yang memiliki potensi untuk kembali, dan siapa yang sudah pergi secara permanen.

2 Distribusi Fitur RFM

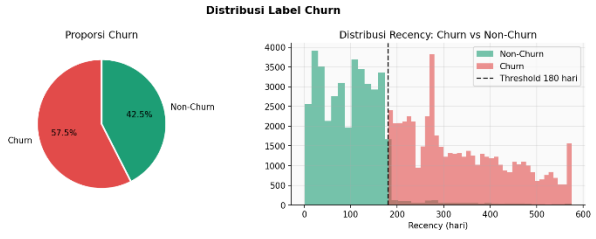
Distribusi ketiga fitur RFM memperlihatkan pola yang berbeda satu sama lain. Recency menunjukkan distribusi yang relatif tersebar luas dengan rentang dari beberapa hari hingga lebih dari 700 hari, mengindikasikan keragaman tingkat keaktifan pelanggan yang tinggi. Frequency sangat terpusat pada nilai satu, konsisten dengan karakteristik one-time buyer yang mendominasi dataset. Monetary memiliki distribusi yang miring ke kanan (*right-skewed*) dengan sebagian besar transaksi bernilai di bawah R$300, sementara sebagian kecil pelanggan mencatat nilai pembelian yang jauh lebih tinggi segmen inilah yang kemudian membentuk cluster Big Spender.



Gambar 4. Distribusi Fitur RFM Pelanggan Olist

3 Distribusi Label Churn

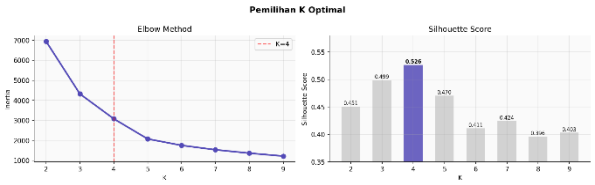
Berdasarkan definisi yang ditetapkan recency lebih dari 180 hari dan frequency sama dengan satu sebanyak 52.671 pelanggan (56,4%) dikategorikan sebagai churn, sementara 40.686 pelanggan (43,6%) dikategorikan sebagai non-churn. Proporsi ini mencerminkan kondisi riil platform e-commerce ritel di mana mayoritas pelanggan memang tidak melakukan pembelian ulang dalam jangka waktu enam bulan. Meskipun demikian, distribusi yang relatif mendekati seimbang ini cukup menguntungkan dari perspektif pemodelan, karena mengurangi tingkat ketidakseimbangan kelas yang harus ditangani.



Gambar 5. Distribusi Label Churn dan Sebaran Recency Berdasarkan Threshold 180 Hari

4 Pemilihan Jumlah Cluster Optimal

Eksperimen clustering dijalankan untuk nilai K dari 2 hingga 9, dengan kedua metrik evaluasi inersia dan Silhouette Score dihitung untuk setiap nilai K. Seperti terlihat pada Gambar 6, kurva Elbow Method menunjukkan penurunan inersia yang tajam dari K=2 ke K=4, setelah itu laju penurunan melambat secara signifikan dan kurva mulai mendatar. Sementara itu, Silhouette Score mencapai nilai tertinggi pada K=4 dengan skor 0,526, yang mengindikasikan bahwa pada konfigurasi empat cluster, pemisahan antar kelompok dan kekompakan dalam kelompok berada pada kondisi yang paling optimal.



Gambar 6. Elbow Method dan Silhouette Score untuk Pemilihan Jumlah Cluster Optimal

Nilai Silhouette Score sebesar 0,526 tergolong moderat-baik berada di atas ambang 0,5 yang secara konvensional diinterpretasikan sebagai struktur cluster yang *reasonable* dan dapat diinterpretasikan [9]. Dengan demikian, K=4 ditetapkan sebagai jumlah cluster optimal untuk seluruh analisis selanjutnya.

5 Profil Cluster dan Interpretasi Bisnis

Hasil K-Means Clustering dengan K=4 menghasilkan empat segmen pelanggan dengan karakteristik yang terbedakan secara jelas. Tabel 1 merangkum profil masing-masing cluster berdasarkan rata-rata fitur RFM, churn rate, dan ukuran segmen.

Tabel 1. Profil Cluster Hasil K-Means Clustering (K=4)

Cluster	Label	n	%	Churn Rate	Avg Recency (hari)	Avg Monetary (R$)	Avg Review Score
C0	Active	49.473	53 %	29%	127,68	103,25	4,16
C1	Lost	36.824	39 %	100 %	387,21	104,87	4,18
C2	Big Spender	4.158	4%	60%	231,30	690,75	4,00
C3	Loyal	2.799	3%	0%	219,90	246,36	4,21

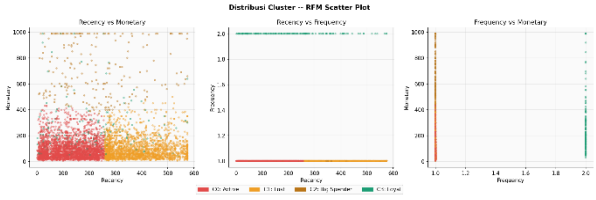
Cluster C0 (*Active*) merupakan segmen terbesar dengan 49.473 pelanggan atau 53% dari total. Rata-rata recency sebesar 127,68 hari menunjukkan bahwa pelanggan di segmen ini relatif baru bertransaksi dibandingkan cluster lain. Churn rate sebesar 29% mengindikasikan bahwa sebagian dari mereka sudah mulai menunjukkan tanda-tanda inaktivitas, meskipun mayoritas masih dapat diklasifikasikan sebagai pelanggan aktif. Nilai monetary yang moderat (R$103,25) dan skor ulasan yang baik (4,16) mencerminkan segmen pelanggan tipikal yang puas namun belum menunjukkan komitmen loyalitas yang kuat.

Cluster C1 (*Lost*) mencakup 36.824 pelanggan dengan churn rate 100% seluruh anggota segmen ini memenuhi definisi churn yang ditetapkan. Rata-rata recency 387,21 hari jauh melampaui ambang 180 hari, mengkonfirmasi bahwa pelanggan di segmen ini sudah lama tidak aktif.

Yang menarik, nilai monetary dan skor ulasan segmen ini (R$104,87 dan 4,18) tidak berbeda signifikan dari C0, mengisyaratkan bahwa churn yang terjadi bukan disebabkan oleh pengalaman belanja yang buruk, melainkan kemungkinan besar oleh faktor-faktor di luar pengalaman transaksi itu sendiri, seperti pergeseran kebutuhan atau kompetisi platform.

Cluster C2 (*Big Spender*) adalah segmen terkecil kedua dengan hanya 4.158 pelanggan (4%), namun memiliki nilai monetary rata-rata tertinggi secara signifikan, yakni R$690,75 hampir tujuh kali lipat dibandingkan C0 dan C1. Churn rate sebesar 60% pada segmen bernilai tinggi ini merupakan temuan yang paling kritis dari perspektif bisnis: platform berpotensi kehilangan pelanggan dengan kontribusi pendapatan terbesar. Skor ulasan yang sedikit lebih rendah (4,00) dibanding cluster lain memberi petunjuk bahwa ekspektasi pelanggan bernilai tinggi mungkin tidak sepenuhnya terpenuhi, khususnya pada dimensi pengiriman dan layanan purna jual.

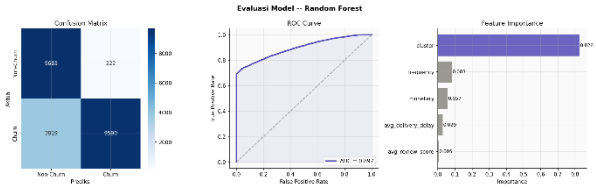
Cluster C3 (*Loyal*) meskipun paling kecil secara ukuran (2.799 pelanggan, 3%), merupakan segmen paling berharga dari perspektif retensi jangka panjang. Churn rate nol persen berarti tidak satu pun anggota segmen ini memenuhi definisi churn, sementara nilai monetary rata-rata R$246,36 menunjukkan kontribusi finansial yang substansial. Skor ulasan tertinggi di antara semua cluster (4,21) menegaskan bahwa kepuasan pelanggan dan loyalitas berjalan beriringan dan bahwa segmen ini merepresentasikan kondisi ideal yang ingin dicapai platform untuk pelanggan-pelanggan di segmen lain.



Gambar 7. Visualisasi Sebaran Cluster pada Ruang Fitur RFM

6 Distribusi Pelanggan dan Churn Rate per Cluster

Gambaran keseluruhan distribusi segmen dan churn rate masing-masing cluster disajikan pada Gambar 8. Terlihat jelas bahwa dominasi C0 dan C1 mencerminkan pola umum platform e-commerce ritel: sebagian besar pelanggan berada di zona abu-abu antara aktif dan hilang, sementara segmen loyal dan bernilai tinggi hanya membentuk ekor kecil dari distribusi.



Gambar 8. Distribusi Jumlah Pelanggan dan Churn Rate per Cluster

Kontras yang paling mencolok adalah antara C1 dengan churn rate 100% dan C3 dengan churn rate 0%, yang secara visual menegaskan bahwa algoritma K-Means berhasil memisahkan pelanggan yang sudah pasti pergi dari yang tetap setia sebuah pemisahan yang tidak dapat dicapai oleh analisis deskriptif sederhana tanpa bantuan clustering.

7 Hasil Pemodelan Random Forest

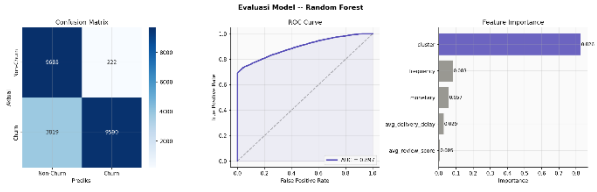
Model Random Forest dilatih menggunakan lima fitur masukan: *frequency*, *monetary*, *avg_review_score*, *avg_delivery_delay*, dan *cluster*. Pembagian data menggunakan rasio 80:20 dengan stratified splitting menghasilkan set latih sebanyak 74.685 sampel dan set uji sebanyak 18.672 sampel. Tabel 2 merangkum performa model pada set uji.

Tabel 2. Hasil Evaluasi Model Random Forest pada Set Uji

Metrik	Nilai
Accuracy	82,9%
Precision	97,7%
Recall	71,5%
F1-Score	82,6%
ROC-AUC	0,897

Model mencapai akurasi keseluruhan sebesar 82,9% dengan ROC-AUC 0,897, mengindikasikan kemampuan diskriminatif yang kuat dalam membedakan pelanggan churn dari yang tidak. Precision sebesar 97,7% sangat tinggi, artinya ketika model memprediksi seorang pelanggan akan churn, prediksi tersebut hampir selalu benar hanya 2,3% prediksi churn yang ternyata merupakan false positive. Ini memiliki implikasi praktis yang penting: sumber daya retensi dapat dialokasikan dengan keyakinan tinggi bahwa target yang dipilih model memang benar-benar berisiko.

Di sisi lain, recall sebesar 71,5% berarti model melewatkan sekitar 28,5% pelanggan yang sesungguhnya akan churn. Dari confusion matrix TN=9.683, FP=222, FN=3.819, TP=9.590 terlihat bahwa false negative (3.819 kasus) jauh lebih banyak dibandingkan false positive (222 kasus). Ketidaksimetrisan ini merupakan konsekuensi langsung dari pilihan desain yang memprioritaskan precision: dengan *class_weight='balanced'* dan tanpa penyesuaian threshold, model cenderung konservatif dalam melabeli churn. Dalam konteks bisnis, *trade-off* ini dapat diterima apabila biaya intervensi retensi per pelanggan cukup tinggi, karena false positive yang rendah berarti minimnya pemborosan sumber daya pada pelanggan yang sebenarnya tidak membutuhkan intervensi.



Gambar 9. Confusion Matrix, ROC Curve, dan Feature Importance Model Random Forest

8 Analisis Feature Importance

Temuan paling menonjol dari analisis feature importance adalah dominasi variabel *cluster* dengan skor 0,826 jauh melampaui keempat fitur lainnya secara gabungan. Ini mengkonfirmasi bahwa informasi segmental yang dihasilkan oleh K-Means Clustering membawa kandungan prediktif yang sangat tinggi terhadap probabilitas churn, dan sekaligus memvalidasi pendekatan pipeline terintegrasi yang digunakan dalam penelitian ini: clustering bukan hanya alat deskriptif, melainkan kontributor aktif dalam pemodelan prediktif.

Fitur berikutnya dalam urutan kepentingan adalah *frequency* (0,083) dan *monetary* (0,057), yang secara bersama-sama mencerminkan dimensi perilaku transaksi yang paling fundamental. *avg_delivery_delay* menempati posisi keempat dengan skor 0,029, menunjukkan bahwa pengalaman logistik memiliki pengaruh terhadap kecenderungan churn meskipun bukan faktor dominan. *avg_review_score* berada di posisi terakhir dengan kontribusi 0,005, mengisyaratkan bahwa kepuasan yang diekspresikan melalui ulasan tidak secara langsung memprediksi retensi pelanggan yang memberikan ulasan positif pun tetap dapat mengalami churn karena alasan di luar pengalaman transaksi itu sendiri.

9 Rekomendasi Strategi Retensi per Segmen

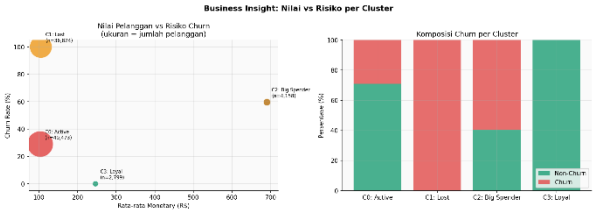
Temuan kuantitatif di atas diterjemahkan ke dalam rekomendasi strategi retensi yang berbeda untuk masing-masing segmen, disesuaikan dengan profil risiko dan nilai kontribusi tiap cluster. Pendekatan segmentasi berbasis data sebagai landasan strategi retensi telah terbukti lebih efektif dibandingkan pendekatan generik karena memungkinkan alokasi sumber daya yang proporsional terhadap potensi nilai pelanggan [11].

Untuk segmen *Lost* (C1), mengingat seluruh anggotanya sudah melewati batas churn dengan recency rata-rata hampir 13 bulan, intervensi yang direkomendasikan bersifat *win-back* dengan intensitas rendah: satu email reaktivasi per kuartal dengan insentif yang dipersonalisasi berdasarkan kategori produk terakhir yang dibeli. Penelitian menunjukkan bahwa kampanye win-back yang dipersonalisasi memiliki tingkat keberhasilan dua hingga tiga kali lebih tinggi dibandingkan pesan generik, meskipun tetap efektif hanya pada segmen pelanggan yang sebelumnya memiliki pengalaman positif [12]. Jika tidak ada respons setelah dua kali upaya, pelanggan sebaiknya dihapus dari daftar aktif untuk menghindari pemborosan anggaran pemasaran dan menjaga reputasi pengirim email.

Segmen *Active* (C0) membutuhkan pendekatan proaktif sebelum pelanggan bergeser ke zona churn. Program *follow-up* pada hari ke-14 pasca pembelian, pemberian poin loyalitas pada transaksi berikutnya, dan aktivasi otomatis kampanye retensi apabila pelanggan tidak bertransaksi selama 45 hari merupakan langkah-langkah yang dapat diimplementasikan melalui sistem CRM (*Customer Relationship Management*) yang sudah ada. Intervensi berbasis trigger waktu terbukti meningkatkan kemungkinan pembelian ulang secara signifikan dibandingkan kampanye massal yang tidak mempertimbangkan tahap siklus hidup pelanggan [11].

Segmen *Big Spender* (C2) memerlukan perhatian khusus mengingat nilai monetarynya yang tinggi namun churn rate-nya yang mengkhawatirkan sebesar 60%. Survei kepuasan tujuh hari pasca pengiriman, penugasan *customer success agent* untuk akun-akun dengan nilai di atas ambang tertentu, serta jaminan kepuasan yang diperkuat merupakan investasi yang dapat dibenarkan secara finansial. Dalam konteks platform marketplace, pelanggan bernilai tinggi umumnya memiliki ekspektasi layanan yang lebih tinggi pula, sehingga ketidaksesuaian antara ekspektasi dan pengalaman aktual menjadi pemicu churn yang dominan pada segmen ini [12].

Segmen *Loyal* (C3), sebagai aset paling berharga platform, layak mendapatkan perlakuan VIP yang eksplisit: program keanggotaan premium dengan keuntungan nyata seperti gratis ongkir tanpa minimum pembelian, akses prioritas ke produk baru, dan program referral dengan insentif menarik. Tujuannya bukan hanya mempertahankan loyalitas yang sudah ada, tetapi menjadikan pelanggan C3 sebagai kanal akuisisi organik melalui rekomendasi dari mulut ke mulut sebuah mekanisme yang biaya akuisisinya jauh lebih rendah dibandingkan iklan berbayar [11].



Gambar 10. Pemetaan Nilai Pelanggan terhadap Risiko Churn dan Komposisi Churn per Cluster

D. PENUTUP

Kesimpulan

Penelitian ini berhasil membangun pipeline analitik terintegrasi yang menggabungkan segmentasi pelanggan berbasis RFM menggunakan K-Means Clustering dengan prediksi churn menggunakan Random Forest Classifier pada dataset Olist Brazilian E-Commerce. Dari 93.357 pelanggan unik yang dianalisis, proses clustering dengan K=4 dipilih berdasarkan konvergensi antara Elbow Method dan Silhouette Score tertinggi sebesar 0,526 menghasilkan empat segmen yang terbedakan secara jelas baik dari karakteristik perilaku maupun profil risiko churnnya.

Keempat segmen tersebut mencerminkan spektrum loyalitas pelanggan yang nyata: C0 *Active* (53%, churn rate 29%) merepresentasikan mayoritas pelanggan yang masih dalam jangkauan retensi; C1 *Lost* (39%, churn rate 100%) mengkonfirmasi bahwa hampir dua dari lima pelanggan sudah melewati titik inaktivitas yang kritis; C2 *Big Spender* (4%, churn rate 60%) menjadi temuan paling

mengkhawatirkan secara bisnis karena platform berpotensi kehilangan segmen dengan kontribusi pendapatan tertinggi; sementara C3 *Loyal* (3%, churn rate 0%) membuktikan bahwa kelompok pelanggan yang benar-benar setia meski kecil jumlahnya memang ada dan dapat diidentifikasi secara algoritmik.

Model Random Forest yang dibangun di atas lima fitur *frequency*, *monetary*, *avg_review_score*, *avg_delivery_delay*, dan *cluster* mencapai performa yang kuat dengan ROC-AUC 0,897, akurasi 82,9%, precision 97,7%, recall 71,5%, dan F1-Score 82,6%. Temuan terpenting dari analisis feature importance adalah dominasi variabel *cluster* dengan skor 0,826, yang secara empiris membuktikan bahwa informasi segmental hasil clustering membawa kandungan prediktif yang jauh lebih besar dibandingkan fitur-fitur perilaku individual lainnya. Hasil ini mengonfirmasi nilai tambah dari pendekatan pipeline terintegrasi: clustering tidak hanya berfungsi sebagai alat deskriptif, melainkan sebagai penghasil fitur prediktif yang krusial.

Secara metodologis, keputusan untuk mengecualikan variabel *recency* dari model klasifikasi terbukti tepat selain mencegah *data leakage*, hal ini juga memaksa model untuk belajar dari pola perilaku yang lebih kaya dan beragam. Penanganan ketidakseimbangan kelas melalui *class_weight='balanced'* menghasilkan komposisi error yang asimetris namun dapat dipertanggungjawabkan secara bisnis: false positive yang sangat rendah (222 kasus) memastikan bahwa intervensi retensi yang dihasilkan model bersifat presisi tinggi dan hemat sumber daya.

Saran

Meskipun hasil penelitian ini menjanjikan, terdapat beberapa arah pengembangan yang dapat memperkuat temuan dan memperluas aplikabilitasnya. Pertama, eksplorasi algoritma clustering alternatif seperti DBSCAN atau Gaussian Mixture Model perlu dilakukan untuk membandingkan kualitas segmentasi, mengingat K-Means memiliki asumsi bahwa cluster berbentuk sferis dan berukuran seimbang asumsi yang belum tentu sepenuhnya terpenuhi pada data pelanggan yang kompleks. Kedua, threshold churn 180 hari yang digunakan dalam penelitian ini bersifat heuristik; penelitian lanjutan dapat mengeksplorasi pendekatan berbasis *survival analysis* untuk menentukan threshold yang lebih adaptif terhadap karakteristik tiap segmen secara individual.

Ketiga, recall sebesar 71,5% mengindikasikan masih adanya ruang perbaikan dalam mendeteksi pelanggan churn yang terlewat. Teknik seperti penyesuaian *classification threshold*, ensemble stacking dengan algoritma lain, atau penambahan fitur berbasis data interaksi pelanggan seperti frekuensi kunjungan halaman atau waktu respons terhadap email promosi berpotensi meningkatkan recall tanpa mengorbankan precision secara signifikan. Keempat, penelitian ini menggunakan data historis statis; implementasi dalam lingkungan produksi yang sesungguhnya membutuhkan pipeline yang mampu memperbarui model secara berkala seiring masuknya data transaksi baru, serta mekanisme monitoring untuk mendeteksi *concept drift* pada perilaku pelanggan dari waktu ke waktu.

Dari sisi praktis, rekomendasi strategi retensi yang dihasilkan penelitian ini perlu divalidasi melalui eksperimen terkontrol misalnya A/B testing pada kampanye retensi berbasis segmen untuk mengukur efektivitas aktual di lapangan sebelum diterapkan secara penuh pada skala platform.

E. DAFTAR PUSTAKA

[1] Statista. (2023). *E-commerce revenue in Brazil from 2017 to 2027*. Statista Digital Market Outlook. https://www.statista.com/forecasts/1151564/e-commerce-revenue-brazil

[2] Dursun, A., & Caber, M. (2022). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, 41, 100922. https://doi.org/10.1016/j.tmp.2021.100922

[3] Kasiran, M. F., Nordin, N. A., & Othman, M. F. (2022). Customer churn prediction using machine learning: A systematic review. *Journal of Physics: Conference Series*, 2319(1), 012006. https://doi.org/10.1088/1742-6596/2319/1/012006

[4] Karimah, M., & Marwati, F. (2024). Sustainability of quality management by implementing data mining to predict academic achievement. *Journal of Social Science and Business Studies*, 2(3), 240–250. https://doi.org/10.61487/jssbs.v2i3.90

[5] Prayitno, A., & Riastuti, R. (2023). Customer segmentation using RFM analysis and K-Means clustering on e-commerce transaction data. *Jurnal Ilmu Komputer dan Informasi*, 16(1), 45–54. https://doi.org/10.21609/jiki.v16i1.1124

[6] Arifin, T., Wahyudi, M., & Sulistyo, S. (2023). Comparative analysis of machine learning algorithms for customer churn prediction in e-commerce. *Journal of Theoretical and Applied Information Technology*, 101(8), 3012–3024. http://www.jatit.org/volumes/Vol101No8/1Vol101No8.pdf

[7] Olist. (2018). *Brazilian E-Commerce Public Dataset by Olist* [Data set]. Kaggle. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

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

[9] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2022). Integration K-Means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 336(1), 012017. https://doi.org/10.1088/1757-899X/336/1/012017

[10] Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10, 47643–47660. https://doi.org/10.1109/ACCESS.2022.3169512

[11] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2022). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 76, 226–242. https://doi.org/10.1016/j.simpat.2021.102517

[12] Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using

deep learning and PCA. *Journal of Big Data*, 7(1), 9. https://doi.org/10.1186/s40537-020-00290-0