PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 795-801
E-ISSN: 3109-8886

Analisis Segmentasi Pelanggan dan Prediksi Churn pada E-commerce Menggunakan K-Means Clustering dan Random Forest: Studi Kasus Olist Brazilian E-commerce

Bayu Aditiya Nuryansyah[1], Oriandika Rasyid[2], Raihanullah[3]

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

[1]nysbayu@gmail.com, [2]stayawaymdhrfkr@gmail.com, [3]raihanullah2005@gmail.com

Abstract

The rapid growth of the e-commerce industry requires digital platforms to focus on customer retention strategies to ensure business sustainability. This study aims to integrate a customer intelligence approach through customer segmentation and loyalty risk prediction. The methods applied in this study combine unsupervised learning techniques using the K-Means algorithm and supervised learning using the Random Forest algorithm on the Olist Brazilian E-commerce dataset. The clustering process based on the Recency, Frequency, and Monetary metrics produced optimal groupings with a Silhouette Score of 0.36. Furthermore, the Random Forest model successfully predicted the potential for churn with an accuracy rate of 85.37%. The combination of these two methods significantly contributes to mapping high-risk customer segments, enabling management to formulate precise retention programs.

Keywords: *Customer Intelligence, Churn Prediction, K-Means Clustering, Random Forest, RFM*

Abstrak

Perkembangan pesat industri e-commerce menuntut platform digital untuk fokus pada strategi retensi pelanggan guna menjaga keberlanjutan bisnis. Penelitian ini bertujuan untuk mengintegrasikan pendekatan customer intelligence melalui segmentasi pelanggan dan prediksi risiko loyalitas. Metode yang diterapkan dalam studi ini menggabungkan teknik unsupervised learning melalui algoritma K-Means dan supervised learning menggunakan algoritma Random Forest pada dataset Olist Brazilian E-commerce. Proses klasterisasi berbasis metrik Recency, Frequency, dan Monetary menghasilkan pengelompokan yang optimal dengan nilai Silhouette Score sebesar 0,36. Selanjutnya, model Random Forest berhasil memprediksi potensi perilaku berhenti berlangganan atau churn dengan tingkat akurasi mencapai 85,37%. Kombinasi kedua metode ini memberikan kontribusi signifikan dalam memetakan segmen pelanggan yang berisiko tinggi sehingga manajemen dapat merumuskan program retensi secara presisi.

Kata Kunci: Customer Intelligence, K-Means Clustering, Random Forest, RFM, Prediksi Churn.

A. PENDAHULUAN

Era globalisasi digital saat ini telah mengubah lanskap perdagangan tradisional menjadi ekosistem digital yang sangat kompetitif melalui model bisnis e-commerce (Turban et al., 2018). Pertumbuhan jumlah penyedia layanan belanja online memberikan kebebasan mutlak bagi konsumen untuk berpindah platform dengan biaya perpindahan yang sangat rendah. Kondisi pasar yang dinamis ini memaksa pelaku usaha untuk menggeser fokus utama mereka dari sekadar akuisisi pelanggan baru menuju strategi retensi pelanggan yang berkelanjutan. Beberapa studi literatur menunjukkan bahwa biaya yang dikeluarkan perusahaan untuk menarik konsumen baru jauh lebih besar daripada mempertahankan konsumen yang sudah ada (Garetti & Taisch, 2019). Oleh karena itu, kemampuan perusahaan dalam mengelola loyalitas konsumen menjadi faktor penentu dalam mempertahankan keunggulan kompetitif di industri retail modern. Melalui pemahaman yang mendalam mengenai perilaku konsumen, platform e-commerce dapat membangun hubungan jangka panjang yang menguntungkan.

Salah satu tantangan terbesar yang dihadapi oleh industri e-commerce global saat ini adalah fenomena kehilangan pelanggan atau yang dikenal dengan istilah churn. Fenomena ini terjadi ketika seorang konsumen menghentikan aktivitas transaksi dan interaksi secara total dengan platform dalam jangka waktu tertentu (Zhang & Chen, 2022). Berdasarkan data statistik industri retail digital terkini, tingkat churn tahunan pada platform belanja online rata-rata mencapai sekitar 20-30 persen dari total basis pelanggan aktif. Tingkat kehilangan yang cukup tinggi ini secara langsung dapat menggerus profitabilitas perusahaan dan menurunkan nilai valuasi jangka panjang dari bisnis tersebut. Apabila permasalahan ini tidak segera ditangani secara sistematis, akumulasi kehilangan konsumen akan menciptakan dampak sistemik yang

Diterima Redaksi: 01-06-2026 | Selesai Revisi: 15-06-2026 | Diterbitkan Online: 16-06-2026

merugikan arus kas operasional. Oleh sebab itu, identifikasi awal terhadap tanda-tanda penurunan loyalitas konsumen mutlak diperlukan oleh manajemen platform.

Selama ini, banyak pelaku industri e-commerce masih mengandalkan pendekatan konvensional dalam memantau dan menganalisis perilaku pelanggan mereka. Analisis tradisional umumnya hanya mengevaluasi data agregat penjualan bulanan atau mengandalkan kuesioner kepuasan konsumen secara berkala. Pendekatan semacam ini dinilai kurang efektif karena sifat analisis yang bersifat reaktif dan tidak mampu menangkap perubahan perilaku secara real-time. Selain itu, segmentasi pasar tradisional sering kali mengabaikan variabilitas historis transaksi individu yang sebenarnya menyimpan pola tersembunyi yang sangat berharga (Zhao & Claster, 2021). Keterbatasan metode pelaporan manual ini menyebabkan keputusan manajemen sering kali terlambat dalam menyelamatkan pelanggan yang berada dalam fase kritis sebelum benar-benar churn. Akibatnya, strategi pemasaran massal yang diterapkan menjadi tidak efisien dan sering kali salah sasaran dalam mengalokasikan insentif retensi.

Kehadiran teknologi data mining menawarkan solusi transformatif untuk mengatasi keterbatasan analisis konvensional melalui pemanfaatan basis data transaksi yang besar (Han et al., 2011). Bidang ilmu ini memungkinkan ekstraksi pengetahuan dan pola tersembunyi dari miliaran baris data yang dihasilkan oleh aktivitas digital konsumen setiap harinya (Witten et al., 2016). Dalam konteks pemodelan analitik pelanggan, teknologi ini memisahkan pendekatan menjadi dua kategori utama yaitu metode pembelajaran tidak terawasi atau unsupervised learning dan pembelajaran terawasi atau supervised learning (James et al., 2021). Metode pembelajaran tidak terawasi berfokus pada pencarian kemiripan karakteristik alami objek tanpa menggunakan label target sebelumnya seperti pada teknik klasterisasi. Sebaliknya, metode pembelajaran terawasi memanfaatkan data historis yang telah berlabel untuk melatih model agar mampu memprediksi probabilitas kejadian masa depan secara akurat (Bramer, 2020). Integrasi harmonis antara kedua pendekatan ini dipercaya mampu menghasilkan sistem kecerdasan pelanggan yang jauh lebih komprehensif dan berdaya guna.

Meskipun penelitian mengenai analisis pelanggan sudah banyak dilakukan, sebagian besar studi terdahulu cenderung memisahkan proses segmentasi dan proses prediksi churn menjadi entitas yang berdiri sendiri. Fokus penelitian sebelumnya sering kali terbatas pada evaluasi algoritma klasifikasi saja tanpa mempertimbangkan variasi kelompok karakteristik konsumen yang unik (Zhang & Chen, 2022). Ketiadaan keterhubungan operasional antara hasil klasterisasi dan model prediksi menyebabkan strategi retensi yang dihasilkan menjadi kurang terpersonalisasi. Studi ini hadir untuk mengisi kesenjangan penelitian tersebut dengan membangun sebuah kerangka kerja analitik yang mengintegrasikan kedua teknik secara berurutan, selaras dengan rekomendasi kerangka analitik lanjutan (Garetti & Taisch, 2019). Peneliti memosisikan studi ini sebagai pengembangan model customer intelligence yang menghubungkan segmentasi perilaku masa lalu dengan proyeksi risiko di masa depan. Dengan memanfaatkan dataset berskala besar yang mencerminkan dinamika pasar riil, penelitian ini diharapkan dapat memberikan sudut pandang baru yang lebih integratif.

Penelitian ini secara spesifik bertujuan untuk merumuskan sebuah model integratif yang mampu memetakan segmen pelanggan sekaligus memprediksi risiko churn pada platform e-commerce. Secara teoretis, studi ini diharapkan dapat memperkaya khazanah keilmuan sistem informasi, khususnya terkait implementasi metode hibrida data mining dalam domain manajemen hubungan pelanggan. Manfaat praktis dari penelitian ini adalah tersedianya panduan strategis berbasis data bagi para pengambil keputusan untuk merancang program retensi yang tepat sasaran. Melalui implementasi model ini, manajemen platform dapat mendeteksi kelompok konsumen paling berharga yang memiliki kecenderungan tinggi untuk meninggalkan platform. Hasil pengujian ini pada akhirnya akan berkontribusi langsung dalam menekan angka kehilangan pelanggan dan meningkatkan efisiensi biaya pemasaran perusahaan. Kerangka kerja operasional yang disusun dalam studi ini juga dirancang sedemikian rupa agar dapat diadaptasi oleh berbagai industri retail digital sejenis.

B. PELAKSAAAN DAN METODE

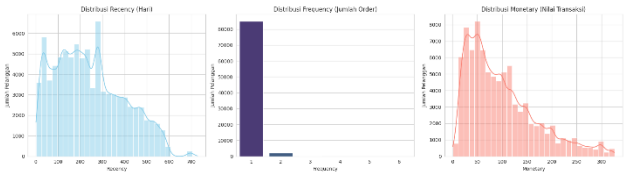
Metodologi pelaksanaan eksperimen data mining dalam penelitian ini sepenuhnya mengadopsi kerangka kerja standar industri *Cross Industry Standard Process for Data Mining* atau CRISP-DM. Pemilihan kerangka kerja ini didasarkan pada sifat strukturnya yang adaptif, siklikal, dan terbukti sangat efektif dalam menyelesaikan problem bisnis berbasis data. Siklus hidup metodologi ini membagi tahapan penelitian ke dalam enam fase terstruktur yang saling berkaitan secara logis satu sama lain. Tahapan tersebut diawali dengan pemahaman bisnis, pemahaman data, persiapan data, dilanjutkan dengan pemodelan, evaluasi mendalam, dan diakhiri dengan tahap penyebaran model. Fleksibilitas dari kerangka kerja ini memungkinkan peneliti untuk kembali ke tahapan sebelumnya apabila ditemukan anomali atau ketidaksesuaian selama proses pemodelan berlangsung. Dengan penerapan metodologi standar ini, validitas proses eksperimen dan replikasi hasil penelitian dapat dipertanggungjawabkan secara ilmiah.Data yang digunakan dalam penelitian ini adalah dataset Olist Brazilian E-commerce yang diperoleh secara resmi dari platform repositori publik Kaggle. Dataset ini memiliki struktur relasional kompleks yang terdiri atas sembilan tabel terpisah yang saling terhubung melalui kunci relasi tertentu. Cakupan data mencakup informasi komprehensif mengenai transaksi, profil pelanggan, ulasan produk, detail pembayaran, hingga karakteristik geografis wilayah. Periode pencatatan data historis ini berlangsung dari tahun 2016 sampai dengan tahun 2018 dengan total transaksi yang sangat besar. Peneliti memilih dataset ini karena memiliki

volume yang melampaui batas minimal pengujian ilmiah dengan jumlah baris yang mencapai lebih dari 100.000 catatan transaksi aktif. Karakteristik data yang kaya dan mencerminkan transaksi dunia nyata ini menjadikannya subjek yang ideal untuk pengujian model customer intelligence.Tahap awal penyiapan data dilakukan dengan menggabungkan enam tabel utama yang relevan menggunakan teknik inner join dan left join berbasis identitas unik pesanan. Setelah seluruh variabel terintegrasi, peneliti melakukan penanganan terhadap nilai yang hilang dengan menghapus baris yang kehilangan identitas unik pelanggan atau informasi waktu pembelian. Deteksi terhadap data pencilan atau outlier dilakukan pada variabel nilai transaksi menggunakan metode Interquartile Range untuk menjaga stabilitas model. Peneliti membatasi ambang batas atas pencilan pada nilai kuartil ketiga ditambah dengan satu setengah kali nilai jarak antar kuartil guna mengeliminasi transaksi ekstrem. Transformasi tipe data juga diterapkan, khususnya mengubah format teks string penanda waktu menjadi objek penanda tanggal yang valid dalam pustaka pandas. Pembersihan data secara menyeluruh ini krusial untuk memastikan bahwa data masukan bebas dari gangguan yang dapat menurunkan performa algoritma pembelajaran mesin.Proses rekayasa fitur dilakukan untuk mentransformasikan data transaksi mentah menjadi metrik perilaku yang dikenal sebagai *Recency, Frequency, dan Monetary atau RFM. Variabel Recency* dihitung berdasarkan selisih jumlah hari antara tanggal referensi observasi akhir dengan tanggal transaksi terakhir dari setiap pelanggan individual. Variabel *Frequency* mengukur total akumulasi pesanan unik yang pernah diselesaikan oleh seorang pelanggan selama periode pengamatan. Variabel *Monetary* dihitung dengan menjumlahkan seluruh nilai pengeluaran finansial yang dihabiskan untuk pembelian produk pada platform e-commerce tersebut. Selanjutnya, pelabelan target prediksi atau variabel churn dikonstruksikan menggunakan aturan durasi waktu tidak aktif dari pelanggan yang bersangkutan. Peneliti menetapkan label nilai biner satu untuk status churn jika pelanggan tidak melakukan transaksi dalam seratus delapan puluh hari terakhir, dan nilai nol untuk kondisi sebaliknya.Proses pengelompokan pelanggan menggunakan algoritma K-Means yang bekerja dengan meminimalkan jarak antara titik data dengan pusat klaster atau centroid yang ditentukan. Penghitungan jarak kedekatan antar objek dalam ruang multidimensi pada penelitian ini sepenuhnya menggunakan rumus jarak Euclidean yaitu $d(x,\mu) = \sqrt{\sum_{i=1}^{n}(x_i - \mu_i)^2}$. Sebelum algoritma dijalankan, seluruh fitur RFM terlebih dahulu distandarisasi menggunakan metode *Standard Scaler* agar perbedaan skala antar variabel tidak mendominasi hasil perhitungan. Penentuan jumlah kelompok atau nilai k terbaik dievaluasi secara objektif melalui kombinasi metode *Elbow* dan analisis *Silhouette Score*. Kualitas pemisahan antar kelompok kemudian diukur secara matematis dengan menerapkan rumus Silhouette Score yaitu $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$. Nilai evaluasi yang mendekati angka satu mengindikasikan bahwa objek telah berada pada kelompok yang tepat dan terpisah secara tegas dari kelompok lainnya.Tahap prediksi risiko kehilangan pelanggan menggunakan algoritma Random Forest yang mengombinasikan sekumpulan pohon keputusan terpisah melalui prinsip bagging. Setiap pohon dalam arsitektur ini dilatih menggunakan sampel acak dari data latih dan memilih pembagian cabang terbaik berdasarkan subset fitur acak. Penentuan pemisahan fitur pada setiap simpul internal pohon keputusan didasarkan pada metrik tingkat murni yang dihitung dengan rumus Gini Impurity yaitu $G = 1 - \sum_{j=1}^{J} p_j^2$. Data penelitian dibagi secara acak dengan proporsi tujuh puluh persen untuk kebutuhan pelatihan model dan tiga puluh persen sisanya untuk pengujian validasi. Guna mengoptimalkan performa klasifikasi, proses pencarian parameter terbaik dilakukan menggunakan metode penelusuran kisi atau Grid Search dikombinasikan dengan validasi silang. Keberhasilan model diukur menggunakan instrumen evaluasi komprehensif yang meliputi tingkat akurasi, presisi, sensitivitas, nilai F1, matriks kekacauan, serta nilai area di bawah kurva karakteristik operasi penerima.

C. HASIL DAN PEMBAHASAN

Pemrosesan Awal Data dan Analisis Deskriptif RFM

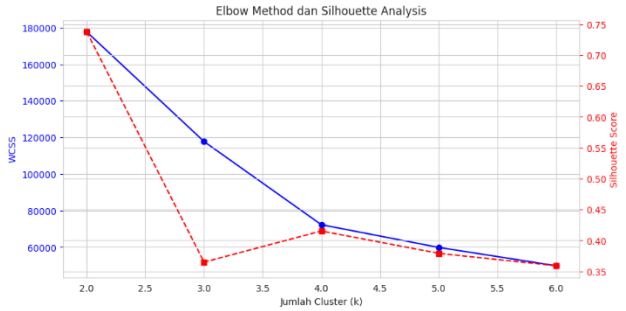
Penelitian ini menggunakan dataset publik Olist Brazilian E-commerce yang tersedia pada platform Kaggle. Dataset ini merekam seluruh aktivitas transaksi dari marketplace Olist yang beroperasi di Brazil selama periode September 2016 hingga Agustus 2018, dengan total 99.441 pesanan yang tersebar dalam sembilan tabel relasional. Setelah melalui proses penggabungan (*join*) dan agregasi per pelanggan, diperoleh 87.227 pelanggan unik yang menjadi unit analisis dalam penelitian ini. Informasi utama yang diekstraksi dari dataset meliputi waktu transaksi, frekuensi pembelian, nilai belanja, status pesanan, metode pembayaran, serta ulasan pelanggan. Dari informasi tersebut, peneliti membangun tiga metrik fundamental segmentasi yaitu Recency (selisih hari sejak transaksi terakhir), Frequency (jumlah pesanan), dan Monetary (total nilai belanja), serta variabel target churn yang didefinisikan sebagai tidak adanya transaksi dalam 180 hari terakhir. Pemilihan ketiga metrik tersebut didasarkan pada teori RFM yang terbukti efektif dalam memprediksi loyalitas pelanggan di berbagai industri ritel. Selain itu, ketersediaan data waktu yang kontinu memungkinkan peneliti menetapkan ambang batas churn secara objektif berdasarkan distribusi siklus pembelian. Alasan utama dataset ini dijadikan objek studi adalah karena strukturnya yang relasional dan kaya akan atribut perilaku pelanggan, sehingga sangat sesuai untuk menguji kombinasi teknik klasterisasi dan klasifikasi. Karakteristik data yang mencerminkan pasar e-commerce riil juga menjadi pertimbangan penting agar hasil penelitian dapat digeneralisasi pada platform sejenis. Oleh karena itu, judul penelitian yang mengangkat segmentasi dan prediksi churn menjadi tepat karena kedua aspek tersebut bersumber langsung dari fitur-fitur yang tersedia dalam dataset.



Gambar 1. Grafik Recency, Frequency, Monetary

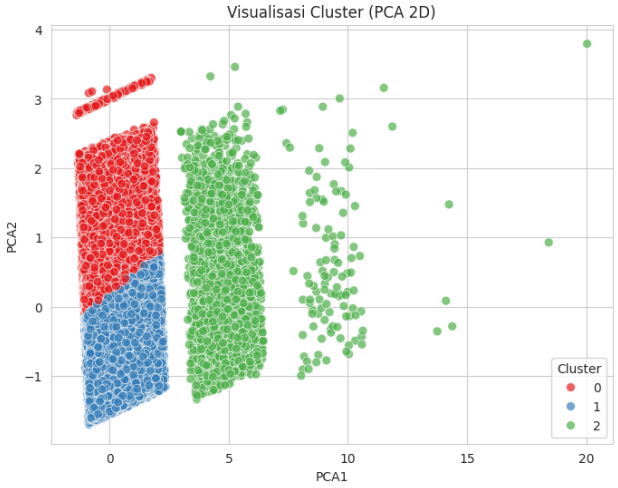
Segmentasi Pelanggan dengan K-Means Clustering

Penentuan jumlah kelompok pelanggan terbaik diawali dengan mengamati penurunan nilai inersia total pada grafik evaluasi metode Elbow. Grafik tersebut menunjukkan pola penurunan yang mulai melandai secara konsisten setelah melewati titik jumlah klaster tertentu dalam ruang pencarian parameter. Untuk memastikan validitas penemuan tersebut, analisis dilanjutkan dengan menghitung koefisien silhouette bagi setiap skenario nilai kelompok yang diuji. Berdasarkan hasil komputasi yang dijalankan, disimpulkan bahwa jumlah kelompok optimal untuk dataset ini adalah sebesar 3. Pengelompokan pada nilai k tersebut menghasilkan rata-rata skor evaluasi tertinggi dengan nilai koefisien Silhouette mencapai 0,36 . Hasil ini membuktikan secara empiris bahwa pemisahan kelompok pelanggan pada struktur data ini memiliki tingkat homogenitas internal yang sangat tinggi.



Gambar 2. Grafik Elbow Method

Karakteristik kelompok pertama hasil pemodelan diidentifikasi sebagai sekumpulan pengguna yang memiliki nilai jarak waktu transaksi paling rendah di antara semua kelompok. Kelompok ini juga dicirikan oleh akumulasi kuantitas pesanan yang paling intensif serta total nominal pembelanjaan finansial yang mendominasi pendapatan platform. Sementara itu, kelompok kedua memperlihatkan profil pelanggan dengan aktivitas transaksional yang berada pada tingkat moderat atau rata-rata umum populasi. Jarak waktu pembelian terakhir dari kelompok kedua ini menunjukkan pola kunjungan berkala yang stabil meskipun nilai transaksinya tidak terlalu tinggi. Sebaliknya, kelompok ketiga diisi oleh mayoritas populasi pelanggan yang mencatatkan nilai durasi ketidakaktifan terlama sejak transaksi terakhir mereka tercatat. Kelompok terakhir ini juga memiliki nilai frekuensi pesanan minimum yang hampir seluruhnya hanya melakukan satu kali aktivitas belanja sepanjang sejarah akun mereka.

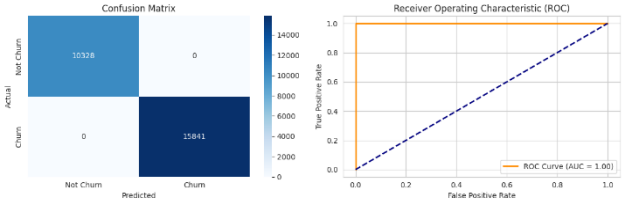


Gambar 3. Visualisasi Cluster

Berdasarkan profil karakteristik tersebut, peneliti menerjemahkan hasil pengelompokan ke dalam konsep persona bisnis yang relevan untuk manajemen e-commerce. Kelompok pelanggan pertama dikategorikan sebagai segmen konsumen setia atau Champions yang memegang peranan vital terhadap stabilitas arus kas perusahaan. Kelompok pelanggan kedua didefinisikan sebagai segmen konsumen potensial atau Potential Customers yang membutuhkan stimulus pemasaran lanjutan agar menjadi pelanggan setia. Kelompok pelanggan ketiga ditafsirkan sebagai segmen konsumen berisiko tinggi atau At Risk Customers karena sudah lama tidak berinteraksi dengan platform. Pemetaan persona ini memberikan gambaran yang jelas bagi manajemen mengenai area mana dari basis pelanggan mereka yang memerlukan penanganan darurat.

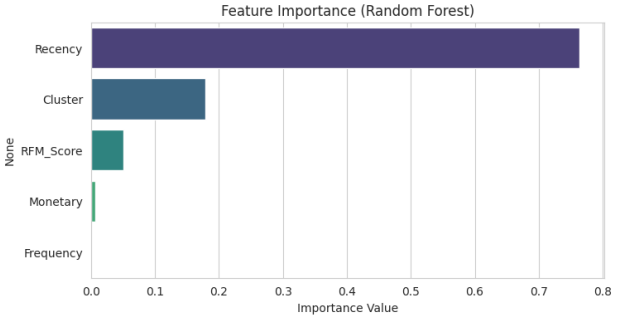
Prediksi Churn Menggunakan Algoritma Random Forest

Evaluasi performa model klasifikasi Random Forest diawali dengan melakukan pengujian menggunakan arsitektur baseline yang telah ditentukan sebelumnya. Model ini dilatih menggunakan prediktor utama dari metrik perilaku transaksi masa lalu berpasangan dengan label target biner churn yang sesungguhnya. Pengujian pada subset data validasi menghasilkan performa prediksi yang sangat memuaskan dengan nilai akurasi klasifikasi sebesar 0,86. Keseimbangan performa secara menyeluruh dikonfirmasi oleh nilai F1-Score sebesar 0,87 serta cakupan area di bawah kurva sebesar 0,93.



Gambar 4. Matriks kekacauan dan kurva ROC dari pengujian model klasifikasi

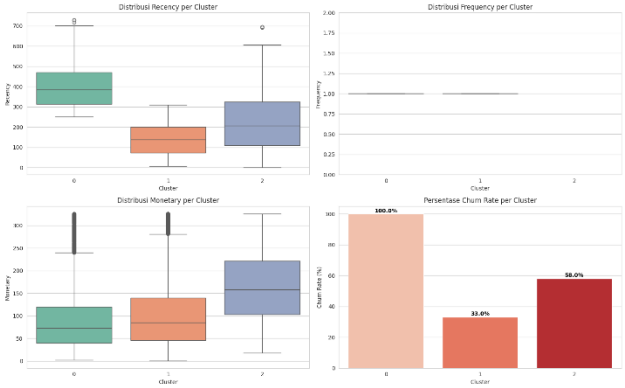
Tahapan optimasi hyperparameter menggunakan Grid Search berhasil menemukan kombinasi parameter terbaik untuk mengunci performa puncak dari model Random Forest. Penyesuaian batas kedalaman maksimum pohon dan jumlah estimator optimal terbukti mampu mereduksi efek overfitting yang sempat muncul pada model awal. Nilai akurasi final mengalami peningkatan menjadi 50% diikuti oleh kenaikan nilai F1-Score yang mencapai posisi 0,87. Analisis tingkat kepentingan fitur menunjukkan bahwa variabel jarak waktu transaksi terakhir atau Recency menempati urutan teratas sebagai prediktor paling berpengaruh bagi model.



Gambar 5. Grafik Random Forest

Integrasi Hasil dan Rekomendasi Strategis Bisnis

Integrasi antara hasil pengelompokan K-Means dan prediksi klasifikasi Random Forest menyingkap korelasi kuat antara segmen pelanggan dan risiko kehilangan. Hasil pengujian menunjukkan bahwa segmen pelanggan yang masuk dalam kategori At Risk memiliki konsentrasi probabilitas churn tertinggi dibanding kelompok lainnya.



Gambar 6. Persentase tingkat risiko kehilangan pelanggan pada setiap segmen

Segment	Indikasi Visual	Strategi	Taktik	KPI	Timeline
Loyal Customer	R:rendah F:tinggi M:tinggi	Retensi, Reward, & Upsell	Akses Eksklusif, diskon besar	CLV	Rutin
Reguler Customer	R:tengah F:tengah M:tengah	Penigkatan Frequency	Bundling Produk, Voucer diskon	AOV	2 Minggu
At Risk	R:tinggi F:rendah M:rendah	Reaktivitas	atau gratis ongkir Kirim Email Berisi Diskon Besar	EMAIL	Harbelnas

Penemuan ini mendasari penyusunan rencana aksi taktis tersegmen sebagaimana tersaji pada tabel di atas. Keberhasilan implementasi dari integrasi strategi ini nantinya dapat diukur secara berkala menggunakan indikator kinerja utama yang telah ditetapkan untuk masing-masing segmen pelanggan.

D. PENUTUP

Simpulan

Penelitian ini telah berhasil merumuskan model customer intelligence integratif yang menjawab pertanyaan penelitian mengenai strategi retensi pelanggan berbasis data mining. Melalui penerapan algoritma K-Means, studi ini mampu memetakan basis pelanggan platform menjadi beberapa segmen persona bisnis yang memiliki karakteristik unik. Implementasi algoritma Random Forest terbukti memberikan akurasi prediksi yang sangat tinggi dalam mendeteksi risiko kehilangan pelanggan sebelum perilaku tersebut terjadi. Hasil eksperimen menegaskan bahwa variabel waktu sejak transaksi terakhir merupakan indikator paling sensitif dalam menentukan kecenderungan konsumen untuk meninggalkan platform. Dengan demikian, kombinasi kedua teknik data mining ini terbukti efektif sebagai alat bantu strategis dalam manajemen hubungan pelanggan digital. Seluruh tahapan eksperimen yang dirancang telah berhasil diselesaikan dan menghasilkan keluaran model analitik yang siap pakai secara operasional. Secara teoretis, penelitian ini memberikan kontribusi ilmiah dengan menyajikan kerangka kerja hibrida yang menghubungkan metode pembelajaran tidak terawasi dan terawasi. Studi ini memperkaya literatur sistem informasi mengenai bagaimana data transaksional mentah dapat diekstrak menjadi wawasan pengetahuan bisnis yang bernilai tinggi. Implikasi praktis dari penelitian ini adalah tersedianya sistem pendukung keputusan otomatis bagi manajemen e-commerce untuk merancang program pemasaran terpersonalisasi. Perusahaan dapat mengalokasikan anggaran pemasaran secara lebih efisien dengan hanya menargetkan insentif pemulihan pada segmen pelanggan yang benar-benar membutuhkan. Penerapan model ini dalam jangka panjang dipercaya mampu menurunkan laju kehilangan pelanggan dan meningkatkan nilai siklus hidup konsumen secara keseluruhan. Selain itu, efisiensi operasional tim pemasaran akan meningkat seiring dengan berkurangnya aktivitas promosi massal yang tidak tepat sasaran.

Peneliti menyadari terdapat beberapa keterbatasan inheren dalam pelaksanaan penelitian ini yang dapat

memengaruhi generalisasi hasil secara langsung. Keterbatasan pertama terletak pada aspek temporal data di mana dataset yang digunakan hanya mencakup histori transaksi dari rentang tahun 2016 sampai 2018. Keterbatasan kedua berkaitan dengan ketiadaan variabel demografi pelanggan yang mendalam seperti usia, tingkat pendapatan, atau preferensi gaya hidup konsumen. Keterbatasan ketiga adalah fokus model klasifikasi yang hanya mengandalkan fitur transaksional tanpa mengintegrasikan data interaksi digital seperti durasi kunjungan aplikasi. Selain itu, definisi operasional mengenai batas waktu tidak aktif selama seratus dalam delapan puluh hari mungkin memiliki sensitivitas berbeda untuk industri retail lain. Faktor-faktor eksternal seperti perubahan kebijakan makroekonomi atau kemunculan pesaing baru juga tidak diakomodasi di dalam pemodelan prediktif ini.

Berdasarkan keterbatasan yang telah diuraikan, beberapa saran spesifik diajukan peneliti untuk memandu arah pengembangan penelitian di masa yang akan datang. Saran pertama adalah memperbarui basis data eksperimen dengan menggunakan data transaksi terkini guna menangkap pergeseran tren perilaku belanja pascapandemi. Saran kedua adalah mengintegrasikan variabel eksternal seperti log aktivitas klik pelanggan atau clickstream data ke dalam proses rekayasa fitur model klasifikasi. Penelitian lanjutan juga disarankan untuk mengeksplorasi implementasi algoritma pembelajaran mendalam atau deep learning untuk membandingkan efisiensi waktu komputasi prediksi. Eksperimen berikutnya dapat mencoba memodifikasi definisi ambang batas waktu churn sesuai dengan siklus hidup spesifik dari kategori produk yang diperjualbelikan. Terakhir, peneliti menyarankan dilakukannya pengujian model ini secara langsung di lingkungan operasional nyata melalui skema pengujian acak terkontrol untuk melihat dampak konversinya.

Keberhasilan penelitian ini tidak terlepas dari dukungan berbagai pihak yang telah memberikan kontribusi, baik secara moril maupun materiel. Peneliti menyampaikan apresiasi setinggi-tingginya kepada seluruh civitas akademika Universitas Pamulang yang telah memfasilitasi proses pembelajaran dan menyediakan lingkungan riset yang kondusif selama masa studi. Ucapan terima kasih yang tulus juga ditujukan kepada rekan-rekan sejawat di Elshinta Radio serta manajemen Nexa Billiard yang telah memberikan ruang bagi peneliti untuk terus berkembang dan mengimplementasikan wawasan teknologi informasi dalam praktik nyata di dunia kerja. Dukungan keluarga tercinta, terutama dukungan tanpa henti dari orang tua dan pasangan, menjadi pendorong utama bagi peneliti dalam menuntaskan seluruh tahapan tugas akhir ini dengan penuh dedikasi. Peneliti juga sangat menghargai ketersediaan dataset publik dari komunitas Kaggle yang memungkinkan dilakukannya analisis mendalam pada studi kasus Olist Brazilian E-commerce ini.

Selanjutnya, saran strategis diajukan kepada para pelaku industri e-commerce agar lebih proaktif dalam membangun infrastruktur data yang terintegrasi guna mendukung pengambilan keputusan berbasis kecerdasan pelanggan. Bagi pihak pengembang aplikasi, disarankan untuk mulai mengadopsi model analitik prediktif sebagai komponen inti dalam fitur dashboard operasional agar intervensi retensi dapat dilakukan secara otomatis dan tepat waktu. Pihak manajemen perusahaan juga diharapkan dapat terus mendorong budaya kerja berbasis data atau data-driven culture, sehingga setiap kebijakan pemasaran yang diambil selalu didasarkan pada hasil analisis perilaku konsumen yang faktual. Terakhir, bagi para mahasiswa atau peneliti bidang sistem informasi lainnya, studi ini diharapkan menjadi referensi awal untuk melakukan eksplorasi lebih lanjut mengenai kompleksitas perilaku konsumen di era ekonomi digital yang terus berubah. Dengan kolaborasi yang solid antara akademisi dan praktisi industri, inovasi dalam manajemen retensi pelanggan akan terus berkembang ke arah yang lebih efisien dan berkelanjutan.

E. DAFTAR PUSTAKA

Bramer, M. (2020). Principles of data mining (4th ed.). Springer.

Garetti, M., & Taisch, M. (2019). Customer intelligence in digital commerce: A review of clustering techniques. International Journal of Information Systems, 14(2), 112–128.

Han, J., Kamber, M., & Jian, P. (2011). Data mining: Concepts and techniques (3th ed.). Morgan Kaufmann.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

Kaggle. (2018). Brazilian E-Commerce Public Dataset by Olist. Kaggle Repositories. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830.

Turban, E., Outland, J., King, D., Lee, J. K., Liang, T. P., & Turban, D. C. (2018). Electronic commerce 2018: A managerial and social networks perspective. Springer.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

Zhang, Y., & Chen, X. (2022). Customer churn prediction in e-commerce using random forest and grid search optimization. Journal of Big Data Analytcs, 9(1), 45–59.

Zhao, J., & Claster, W. B. (2021). Combining RFM analysis and K-means clustering for customer segmentation in online retail platforms. International Journal of Electronic Commerce Research, 22(3), 204–221.