Jurnal Manajemen Bisnis Digital Terkini
Volume. 3 Nomor. 1 Januari 2026
e-ISSN: 3047-0552; p-ISSN: 3047-2199, Hal 300-318
DOI: https://doi.org/10.61132/jumbidter.v3i1.1228
Tersedia: https://ejournal.arimbi.or.id/index.php/JUMBIDTER

Analisis Prediksi Customer Churn pada Sektor E-Commerce Berdasarkan Perilaku Transaksi Menggunakan Pendekatan Machine Learning

Nadeerah Hani' Fauziyyah[1], I Wayan Sudiarsa [2*], Ida Ayu Eka Sastradewi [3], Kadek Agustine Yueyin Parisya[4], Sartika[5]
[1,3,4,5] Bisnis Digital, Institut Bisnis dan Teknologi Indonesia, Indonesia
[2] Rekayasa Sistem Komputer, Institut Bisnis dan Teknologi Indonesia, Indonesia
**Penulis Korespondensi: sudiarsa@instiki.ac.id*

Abstract. *Because it directly impacts revenue, customer loyalty, and long-term business sustainability, customer churn is a critical issue for the e-commerce industry. High churn rates indicate that a business is unable to retain existing customers, which means it is more expensive to acquire new customers. Therefore, a precise analytical approach is needed to identify customer behavior patterns that are likely to churn. Using machine learning methods, this study analyzes and predicts customer churn. For this study, the E-Commerce Customer Churn 2025 dataset, obtained from Kaggle, was used. This dataset consists of 10,000 customer data and contains fifteen variables covering transaction behavior, customer characteristics, and churn status. Data preprocessing, descriptive analysis, exploratory data analysis (EDA), and classification model development using Logistic Regression and Random Forest algorithms were part of the research project. Model evaluation was conducted using a Confusion Matrix and Receiver Operating Characteristic (ROC) Curve to evaluate the model's accuracy and ability to distinguish between churned and non-churned customers. The results showed that the Random Forest model performed better than Logistic Regression, with an ROC-AUC of 1.00. Furthermore, feature importance analysis revealed that the days_since_last_purchase variable was the most dominant factor in predicting customer churn. These findings are expected to help e-commerce companies design more effective, data-driven customer retention strategies.*

Keywords: *Customer Churn; E-Commerce; Machine Learning; Random Forest; Transaction Behavior.*

Abstrak. Karena berpengaruh secara langsung terhadap pendapatan, loyalitas pelanggan, dan keberlanjutan bisnis dalam jangka panjang, customer churn merupakan masalah penting bagi industri e-commerce.Tingkat churn yang tinggi menunjukkan bahwa bisnis tidak dapat mempertahankan pelanggan yang sudah ada, yang berarti lebih mahal untuk membeli pelanggan baru.Oleh karena itu, pendekatan analitis yang tepat diperlukan untuk menemukan pola perilaku pelanggan yang berpotensi churn.Dengan menggunakan metode pembelajaran mesin, penelitian ini menganalisis dan memprediksi churn pelanggan. Untuk penelitian ini, dataset E-Commerce Customer Churn 2025 digunakan, yang diperoleh dari Kaggle. Dataset ini terdiri dari 10.000 data pelanggan dan memiliki lima belas variabel yang mencakup perilaku transaksi, karakteristik pelanggan, dan status churn. Preprocessing data, analisis deskriptif, analisis data eksploratif (EDA), dan pembuatan model klasifikasi menggunakan algoritma Logistic Regression dan Random Forest adalah bagian dari proyek penelitian. Evaluasi model dilakukan menggunakan Confusion Matrix dan Receiver Operating Characteristic (ROC) Curve untuk mengevaluasi akurasi dan kemampuan model untuk membedakan antara pelanggan churn dan non-churn. Hasil penelitian menunjukkan bahwa, dengan nilai ROC-AUC sebesar 1,00, model Random Forest memiliki kinerja terbaik dibandingkan dengan Logistic Regression.Selain itu, seperti yang ditunjukkan oleh analisis kepentingan fitur, variabel days_since_last_purchase adalah yang paling penting dalam memprediksi kelangkaan pelanggan. Diharapkan temuan ini akan membantu bisnis e-commerce membuat strategi retensi pelanggan yang lebih efisien dan berbasis data.

Kata kunci: E-Commerce; Konsumsi Pelanggan; Pembelajaran Mesin; Perilaku Transaksi; Random Forest.

1. LATAR BELAKANG

Dalam beberapa dekade terakhir, industri e-commerce telah berkembang pesat sebagai hasil dari kemajuan teknologi informasi dan internet. Dengan menggunakan platform e-commerce, bisnis dapat menjangkau lebih banyak pelanggan, meningkatkan efisiensi transaksi, dan membuat pengalaman belanja yang lebih personal dan praktis. Namun, pelaku bisnis digital sangat bersaing karena kemudahan akses ini. Pelanggan memiliki banyak pilihan dan

Naskah Masuk: November 18, 2025; Revisi: Desember 15, 2025; Diterima: Januari 25, 2026; Terbit: Januari 27, 2026

dapat berpindah ke platform lain dengan biaya rendah. Loyalitas pelanggan adalah aset strategis penting untuk keberlangsungan e-commerce (Buckinx & Van den Poel, 2005; Kumar & Reinartz, 2016).

Tingkat customer churn yang tinggi adalah masalah utama yang dihadapi oleh bisnis e-commerce. Customer churn adalah ketika pelanggan berhenti melakukan transaksi atau menggunakan layanan perusahaan selama periode waktu tertentu. Ini berdampak langsung pada penurunan pendapatan dan biaya operasional karena biaya untuk mendapatkan pelanggan baru jauh lebih besar daripada mempertahankan pelanggan lama. Menurut penelitian sebelumnya, churn adalah sinyal penting dari strategi retensi yang tidak berhasil dan hubungan pelanggan dengan perusahaan yang buruk (Verhoef, 2003; Kumar & Reinartz, 2016). Oleh karena itu, kemampuan perusahaan untuk mengidentifikasi dan memprediksi churn sangat penting untuk menjaga stabilitas dan kemajuan perusahaan.

Metode tradisional untuk mengelola hilangnya pelanggan biasanya bersifat reaktif dan intuitif, seperti memberikan promosi setelah pelanggan menghentikan transaksi. Metode ini dianggap kurang efektif karena tidak dapat mengidentifikasi churn awal. Namun, pelanggan yang akan churn biasanya mengalami perubahan perilaku transaksi secara bertahap, seperti penurunan frekuensi pembelian, penurunan nilai transaksi, dan meningkatnya waktu yang berlalu sejak transaksi sebelumnya (Buckinx & Van den Poel, 2005). Perusahaan kehilangan kesempatan untuk melakukan intervensi yang lebih efektif karena pola tersebut sulit diidentifikasi secara akurat tanpa analisis data yang menyeluruh.

Metode pembelajaran mesin menjadi lebih populer untuk memprediksi kehilangan pelanggan secara lebih akurat dan objektif karena data transaksi pelanggan semakin mudah diakses. Menurut sejumlah penelitian, algoritma pembelajaran mesin, terutama model ensemble seperti Random Forest, memiliki kemampuan untuk membuat prediksi yang lebih baik daripada model linear konvensional (Breiman, 2001; Verbeke et al., 2012). Model ini tidak hanya dapat mengidentifikasi hubungan non-linear antar variabel, tetapi juga dasspat menunjukkan tingkat kepentingan fitur yang memengaruhi churn. Oleh karena itu, penelitian ini bertujuan untuk menganalisis dan memprediksi perilaku transaksi pelanggan pada sektor e-commerce dengan menggunakan pendekatan pembelajaran mesin dan model pembanding Random Forest dan Logistic Regression.

2. KAJIAN TEORITIS

Permintaan Pelanggan dalam E-Commerce

Ketika pelanggan berhenti melakukan transaksi atau menggunakan layanan suatu perusahaan dalam jangka waktu tertentu, itu disebut customer churn.Churn, sebuah indikator penting dalam e-commerce, menunjukkan tingkat loyalitas pelanggan dan efektivitas strategi retensi perusahaan. Tingkat churn yang tinggi menyebabkan penurunan pendapatan perusahaan. Kondisi ini meningkatkan biaya operasional. Ini karena biaya lebih besar untuk mendapatkan pelanggan baru daripada mempertahankan pelanggan lama. Oleh karena itu, pengelolaan churn menjadi komponen strategis yang sangat penting untuk keberlangsungan perusahaan e-commerce (Kotler & Keller, 2016; Reichheld & Sasser, 1990).

Perilaku Transaksi Pelanggan sebagai Pengukur Kelangkaan

Perilaku transaksi pelanggan menunjukkan bagaimana pelanggan berinteraksi dengan platform e-commerce. Data transaksi sebelumnya dapat digunakan untuk mengukur pola ini secara kuantitatif. Recency, Frequency, dan Monetary adalah komponen model RFM, yang sering digunakan untuk menganalisis perilaku pelanggan (Hughes, 1994).Monetary mengukur total atau rata-rata nilai transaksi pelanggan, dan frekuensi mengukur seberapa sering pelanggan melakukan transaksi.

Model RFM sangat populer karena dapat menggambarkan tingkat loyalitas pelanggan secara menyeluruh. Pelanggan dengan nilai frekuensi yang tinggi menunjukkan jarak waktu transaksi terakhir yang lama, yang menunjukkan tingkat aktivitas pelanggan yang rendah. Sebaliknya, pelanggan dengan nilai frekuensi yang rendah menunjukkan intensitas transaksi yang rendah. Selain itu, kontribusi ekonomi pelanggan yang rendah ditunjukkan oleh nilai uang yang kecil.Tingkat kemungkinan churn pelanggan meningkat ketika ketiga kondisi tersebut digabungkan. Studi yang dilakukan oleh Fader, Hardie, dan Lee (2005) mendukung hal ini.

Keputusan pelanggan dipengaruhi oleh variabel lain selain variabel RFM.Penggunaan kupon dan jumlah transaksi termasuk dalam kategori ini. Durasi penggunaan layanan adalah metrik penting untuk mengukur loyalitas pelanggan. Selain itu, pola aktivitas pengguna di platform memengaruhi keterikatan pelanggan.Faktor-faktor ini memiliki dampak yang signifikan terhadap churn, menurut penelitian yang dilakukan oleh Buckinx dan Van den Poel (2005). Dalam penelitian ini, variabel days_since_last_purchase menunjukkan dimensi frekuensi, dan variabel ini diasumsikan memiliki pengaruh yang signifikan terhadap status churn pelanggan.

Pengajaran Mesin untuk Prediksi Konsumen Makan

Salah satu cabang kecerdasan buatan adalah machine learning, yang memungkinkan sistem mempelajari pola dari data. Metode ini membuat prediksi tanpa pemrograman eksplisit. Untuk membuat model prediktif, pengajaran mesin menggunakan data historis (Mitchell, 1997). Machine learning digunakan sebagai metode klasifikasi untuk memprediksi kehilangan pelanggan.Bergantung pada model klasifikasi, Anda dapat menentukan kemungkinan churn pelanggan. Data transaksi dan perilaku pelanggan adalah dasar untuk prediksi.Untuk menangani data berskala besar, pendekatan ini berhasil.

Penelitian churn menggunakan banyak algoritma pembelajaran mesin. Random Forest dan Support Vector Machine adalah algoritma yang banyak digunakan, serta Logistic Regression dan Decision Tree.Alternatif untuk algoritma yang sangat efisien adalah Gradient Boosting. Logistic Regression disukai karena mudah dipahami. Namun, model ini tidak dapat menangkap hubungan non-linear.Keterbatasan ini mendorong penggunaan metode yang lebih kompleks (Verbeke et al., 2012).

Banyak pohon keputusan digunakan dalam metode kelompok Random Forest. Metode ini digunakan untuk meningkatkan akurasi dan stabilitas model. Random Forest mampu menangani volume data yang besar. Algoritma ini juga dapat mengurangi kemungkinan overfitting (Breiman, 2001).Random Forest juga menyediakan data fitur penting. Variabel yang paling berpengaruh dapat diidentifikasi dengan bantuan data ini. Akibatnya, Random Forest sangat digunakan dalam penelitian churn yang berbasis data transaksi skala besar.

Penilaian Model Klasifikasi

Tujuan dari evaluasi model klasifikasi churn adalah untuk mengetahui seberapa baik model dapat membedakan pelanggan churn dan non-churn. Untuk memastikan keandalan hasil prediksi, evaluasi dilakukan. Penelitian klasifikasi menggunakan sejumlah metrik evaluasi. Precision, accuracy, recall, dan F1-score adalah metrik yang termasuk dalam kategori ini. Untuk mengevaluasi distribusi prediksi yang benar dan salah, Matrix Konflik digunakan. Untuk data yang tidak seimbang, evaluasi ini sangat penting.Akibatnya, penentuan metrik yang tepat sangat penting (Powers, 2011).

Informasi tentang performa model untuk setiap kelas diberikan oleh Confusion Matrix. Untuk menyeimbangkan precision dan recall, nilai F1 digunakan. Precision menunjukkan ketepatan prediksi positif, sedangkan recall menunjukkan kemampuan model untuk menangkap seluruh data positif. Selain itu, performa model dievaluasi dengan ROC Curve.Kemampuan untuk memisahkan dua kelas diukur dengan area di bawah kurva.Performa

klasifikasi yang sangat baik ditunjukkan dengan nilai ROC-AUC yang mendekati 1 (Fawcett, 2006).

Studi sebelumnya yang relevan

Penelitian telah menunjukkan bahwa mesin pembelajaran efektif dalam memprediksi kehilangan klien. Verbeke et al. (2012) menunjukkan bahwa model berbasis kelompok memiliki keunggulan dibandingkan model linear. Salah satu algoritma yang paling efisien adalah Random Forest. Metode non-linear sangat penting dalam penelitian ini. Buckinx dan Van den Poel (2005) menunjukkan bahwa dua variabel yang paling penting dalam mempengaruhi churn pelanggan adalah frekuensi dan kecepatan.

Amin et al. (2019) melakukan penelitian tambahan. Studi ini menggabungkan EDA dan pembelajaran mesin. Metode kombinasi ini meningkatkan pemahaman kita tentang pola perilaku pelanggan.Hasilnya menunjukkan bahwa prediksi churn menjadi lebih akurat. EDA membantu menemukan fitur penting dalam data. Untuk pemodelan, mesin pembelajaran menggunakan data ini. Hasil ini mendukung fondasi konseptual penelitian e-commerce.

Konsep dan Hipotesis Penelitian

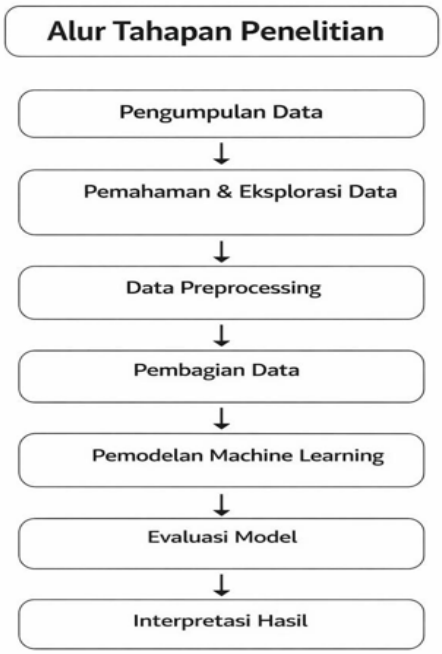
Ada korelasi yang signifikan antara status churn dan perilaku transaksi pelanggan, menurut studi teoritis dan penelitian sebelumnya (Kotler & Keller, 2016; Buckinx & Van den Poel, 2005). Perilaku transaksi menunjukkan seberapa banyak pelanggan berinteraksi dengan platform e-commerce.Frekuensi transaksi menunjukkan seberapa sering pelanggan menggunakan layanan.Menurut Hughes (1994), lamanya waktu yang telah berlalu sejak transaksi terakhir menunjukkan tingkat aktivitas terakhir pelanggan. Nilai transaksi menunjukkan seberapa besar kontribusi ekonomi pelanggan kepada perusahaan (Fader et al., 2005). Keputusan pelanggan untuk menggunakan layanan e-commerce dipengaruhi oleh variabel-variabel tersebut.Selain itu, diperkirakan bahwa algoritma Random Forest memiliki kinerja prediksi yang lebih baik daripada model klasifikasi linear (Breiman, 2001; Verbeke et al., 2012).

3. METODE PENELITIAN

Metodologi Penelitian

Teknik Penelitian Dengan menggunakan pendekatan kuantitatif dan teknik analisis berbasis "machine learning", penelitian ini memprediksi jumlah pelanggan yang meninggalkan platform e-commerce.Metodologi penelitian disusun secara sistematis sehingga analisis data dapat dilakukan secara terstruktur dan menghasilkan kesimpulan yang valid.Penelitian melibatkan pengumpulan data, pemrosesan awal data, pemodelan

klasifikasi, dan evaluasi kinerja model. Setiap tahapan tersebut diatur sesuai dengan alur metodologi penelitian, yang akan dibahas pada subbab berikutnya.



Gambar 1. Metodologi Penelitian.

Tahapan Penelitian

Penjelasan Tahapan Penelitian :

Tahap pertama adalah "pengumpulan data". Untuk melakukan ini, kami menggunakan dataset "E-Commerce Customer Churn 2025" yang dikumpulkan melalui platform Kaggle dan berfungsi sebagai sumber utama penelitian karena mengandung semua data transaksi dan karakteristik pelanggan.

Untuk memastikan bahwa data sesuai dengan tujuan penelitian, tahap kedua adalah "pemahaman dan eksplorasi data". Tujuan dari tahap ini adalah untuk mengetahui struktur dataset, tipe variabel, dan karakteristik awal churn dan non-churn pelanggan.

Sebelum digunakan dalam pemodelan, tahap ketiga adalah data preprocessing. Ini termasuk pemeriksaan nilai hilang, eliminasi variabel identifier, transformasi variabel kategorikal, dan normalisasi variabel numerik. Tujuan dari proses ini adalah untuk meningkatkan kualitas data.

Tujuan dari pembagian data, tahap keempat, adalah untuk menghasilkan evaluasi model yang objektif. Pembagian data membagi dataset menjadi data latih dan data uji dengan menggunakan metode stratifikasi yang bergantung pada variabel churn.

Pemodelan pembelajaran mesin adalah tahap kelima. Ini melibatkan pembuatan model klasifikasi untuk memprediksi status churn pelanggan dengan menggunakan variabel input yang tersedia.

Evaluasi model adalah tahap keenam. Ini dilakukan dengan menggunakan metrik evaluasi klasifikasi seperti keakuratan, keakuratan, keakuratan, recall, skor F1, dan ROC-AUC.

Pada langkah terakhir, "interpretasi hasil" dilakukan. Ini berarti menganalisis kinerja model dan mengaitkannya dengan tujuan penelitian serta dasar teori yang digunakan.

Data Penelitian

Penelitian ini menggunakan dataset E-Commerce Customer Churn 2025, yang diperoleh dari platform Kaggle dan terdiri dari 10.000 observasi pelanggan dan 15 variabel, masing-masing berfokus pada perilaku transaksi, karakteristik pelanggan, dan status churn.

Link data : https://www.kaggle.com/datasets/mohammadmaaz052/e-commerce-customer-churn-2025-10k

User : mohammadmaaz052

Tabel 2. Menunjukkan Variabel Dataset Penelitian.

No	Nama Variabel	Tipe Data	Peran
1	customer_id	Kategorikal	Identifier
2	days_since_last_purchase	Numerik	Input
3	total_orders	Numerik	Input
4	total_spent	Numerik	Input
5	average_order_value	Numerik	Input
6	total_returns	Numerik	Input
7	frequency	Numerik	Input
8	recency	Numerik	Input
9	tenure_days	Numerik	Input
10	is_active	Biner	Input
11	preferred_device	Kategorikal	Input
12	preferred_payment_method	Kategorikal	Input
13	city_tier	Kategorikal	Input
14	coupon_used	Numerik	Input
15	churn	Biner	**Target**

Tabel 1

Dari total *15 variabel*, terdapat *1 variabel identifier*, *13 variabel input*, dan *1 variabel target*. Variabel input terdiri dari *9 variabel numerik*, *3 variabel kategorikal*, dan *1 variabel biner*. Ini menunjukkan bahwa struktur data pelanggan beragam.

Perilaku transaksi pelanggan dapat digambarkan secara kuantitatif dengan variabel seperti "days_since_last_purchase", "frekuensi", dan "total_spent". Variabel "tenure_days", "average_order_value", dan "total_orders" menunjukkan nilai ekonomi dan loyalitas pelanggan terhadap platform, dengan nilai "days_since_last_purchase" yang lebih tinggi menunjukkan potensi churn.

Preferensi konsumen terhadap platform e-commerce diukur dengan menggunakan variabel kategori "preferred_device", "preferred_payment_method", dan "city_tier". Untuk saat ini, informasi tentang status aktivitas pelanggan terkini diberikan oleh variabel biner "is_active". Dalam memprediksi customer churn, penelitian mempertimbangkan semua aspek transaksi dan karakteristik pelanggan, seperti yang ditunjukkan oleh penyusunan variabel ini.

Pemrosesan Awal Data

Sebelum digunakan dalam proses pemodelan machine learning, tahap pemrosesan awal data dilakukan untuk memastikan bahwa dataset berada dalam kondisi yang ideal. Proses ini sangat penting karena kualitas data berdampak langsung pada kinerja dan keakuratan model prediksi churn.

Langkah pertama adalah melakukan pemeriksaan nilai hilang, atau nilai yang tidak ada, untuk seluruh 15 variabel dalam dataset yang terdiri dari 10.000 observasi. Hasil pemeriksaan menunjukkan bahwa tidak ada nilai hilang dalam dataset, jadi tidak perlu melakukan proses imputasi data. Ini memastikan bahwa semua data dapat digunakan sepenuhnya dalam analisis.

Dalam langkah kedua, variabel identifikasi, variabel customer_id, dihapus. Hal ini dilakukan karena variabel ini hanya berfungsi sebagai penanda unik pelanggan dan tidak memiliki kontribusi prediktif terhadap variabel target churn. Setelah penghapusan, total variabel yang digunakan untuk analisis menjadi empat belas, terdiri dari tiga belas variabel input dan satu variabel target.

Langkah ketiga adalah transformasi variabel kategorikal. Ini dilakukan dengan menggunakan teknik koding satu panas.preferred_device, preferred_payment_method, dan city_tier adalah variabel yang ditransformasikan, masing-masing memiliki lebih dari satu kategori. Dalam proses ini, setiap kategori diubah menjadi variabel biner yang bernilai antara 0 dan 1, yang memungkinkan algoritma pembelajaran mesin berbasis numerik untuk memproses data.

Semua variabel numerik, termasuk days_since_last_purchase, total_orders, total_spent, average_order_value, frequency, dan tenure_days, dinormalisasi dalam langkah keempat dengan menggunakan metode StandardScaler.Proses ini menghasilkan distribusi data dengan nilai rata-rata 0 dan simpangan baku 1. Dengan demikian, perbedaan skala antar variabel tidak berdampak pada proses pelatihan model.

Pada langkah terakhir, metode sampling stratified yang didasarkan pada variabel target churn digunakan untuk membagi dataset menjadi data latih dan data uji. Dari 10.000 observasi, 8.000 data (80%) digunakan sebagai data latih dan 2.000 data (20%) digunakan sebagai data uji.Teknik stratifikasi memastikan bahwa rasio churn dan non-churn pelanggan seimbang pada kedua subset data. Ini membuat evaluasi model lebih representatif dan tidak bias.

Pemodelan Machine Learning

Tujuan dari tahap pemodelan adalah untuk membuat model klasifikasi yang dapat memprediksi status churn pelanggan dengan menggunakan variabel input yang telah diproses. Dua algoritma pengajaran mesin digunakan dalam penelitian ini: Logistic Regression sebagai model dasar dan Random Forest sebagai model utama.

Logistic Regression

Karena strukturnya yang sederhana dan mudah dipahami, Logistic Regression digunakan sebagai model pembanding.Berdasarkan hubungan linear antara variabel input dan variabel target, model ini memprediksi probabilitas churn pelanggan.Model Regresi Logistik dilatih dengan 8.000 data latih dan diuji dengan 2.000 data uji dengan 13 variabel input.

Untuk mengevaluasi peningkatan kinerja model ensemble dan untuk mendapatkan gambaran awal tentang kinerja prediksi churn, model ini digunakan. Salah satu kendala utama analisis regresi logistik adalah ketidakmampuan untuk mengidentifikasi hubungan non-linear antar variabel, yang sering terlihat dalam data transaksi pelanggan.

Random Forest

Untuk tujuan penelitian ini, Forest Random digunakan sebagai model utama karena kemampuan untuk menangani hubungan non-linear dan hubungan kompleks antara variabel. Untuk meningkatkan stabilitas dan akurasi model, algoritma ini bekerja dengan membangun sejumlah pohon keputusan (decision trees) dan menggabungkan hasil prediksi dari seluruh pohon.

Dengan semua variabel yang dimasukkan telah dinormalisasi dan dienkodekan, 8.000 observasi sebagai data latih digunakan untuk menguji model Random Forest. Salah satu keunggulan Random Forest dalam penelitian ini adalah kemampuan untuk menghasilkan

informasi tentang pentingnya fitur, yang digunakan untuk menentukan variabel yang paling berpengaruh terhadap pengurangan pelanggan.

Hasil pentingnya fitur menunjukkan bahwa days_since_last_purchase, tenure_days, dan average_order_value adalah variabel yang paling banyak berkontribusi pada proses prediksi churn. Ini menunjukkan bahwa waktu sejak transaksi terakhir dan lamanya pelanggan menggunakan layanan adalah faktor utama dalam menentukan kemungkinan churn.

Evaluasi Model

Untuk menilai kemampuan model untuk secara akurat mengklasifikasikan pelanggan churn dan non-churn, evaluasi kinerja dilakukan dengan menggunakan 2.000 observasi uji yang tidak digunakan selama proses pelatihan. Precision, accuracy, recall, F1-score, dan confusion matrix adalah metrik evaluasi yang digunakan. Selain itu, untuk mengevaluasi kemampuan model untuk membedakan dua kelas pada berbagai ambang keputusan, digunakan ROC Curve dan ROC-AUC. Jumlah prediksi yang benar dan salah untuk masing-masing kelas churn dan non-churn ditunjukkan dalam matriks kekacauan untuk mengetahui tingkat kesalahan klasifikasi. Nilai ROC-AUC digunakan sebagai pengukur utama untuk kinerja model; nilai yang mendekati 1 menunjukkan kemampuan klasifikasi yang sangat baik.

4. HASIL DAN PEMBAHASAN

Hasil dan Pembahasan

Analisis Deskriptif dan EDA

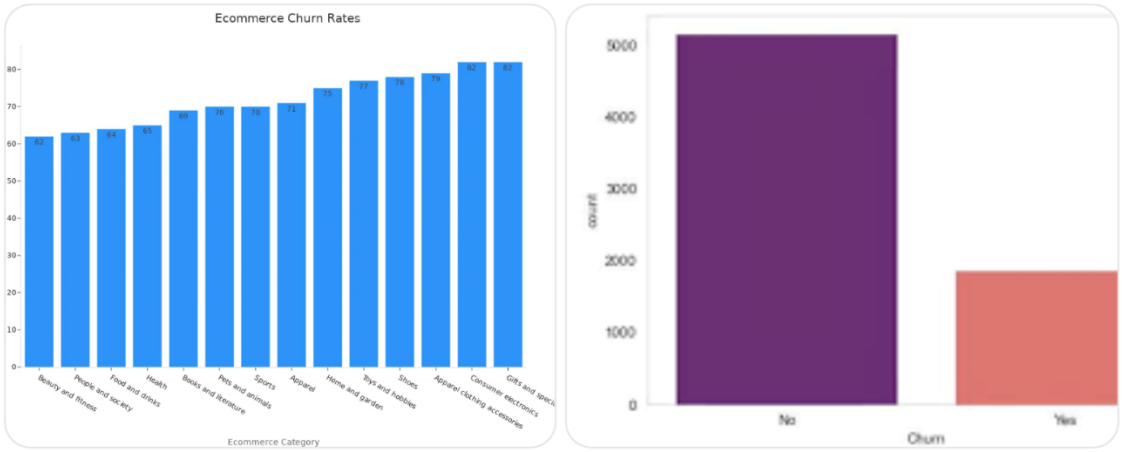
Tabel 3. Statistik Deskriptif Variabel Utama.

Variabel	Churn = 0 (Mean)	Churn = 1 (Mean)
days_since_last_purchase	Rendah	**Sangat Tinggi**
total_orders	Lebih tinggi	Lebih rendah
total_spent	Lebih tinggi	Lebih rendah
frequency	Stabil	Menurun

Temuan awal menunjukkan bahwa pelanggan churn menunggu lebih lama untuk transaksi terakhir dibandingkan dengan pelanggan setia.

Statistik deskriptif untuk variabel utama yang dibandingkan berdasarkan status churn pelanggan disajikan dalam Tabel 3. Tujuan dari perbandingan ini adalah untuk memberikan gambaran awal tentang perbedaan fitur antara pelanggan yang churn dan pelanggan yang tidak churn. Hasil statistik menunjukkan bahwa pelanggan churn memiliki nilai days_since_last_purchase yang jauh lebih tinggi dibandingkan pelanggan non-churn, yang

menunjukkan bahwa jarak waktu transaksi terakhir semakin lama. Sebaliknya, variabel total_orders, total_spent, dan frekuensi pelanggan churn cenderung lebih rendah, yang menunjukkan bahwa intensitas dan nilai transaksi menurun. Teori perilaku pelanggan mengatakan bahwa churn terjadi ketika aktivitas transaksi berkurang (Buckinx & Van den Poel, 2005). Penemuan ini sejalan dengan teori ini. Oleh karena itu, tabel ini mendukung dasar empiris bahwa perilaku transaksi merupakan faktor penting dalam membedakan pelanggan yang melakukan churn dari yang tidak melakukan churn.

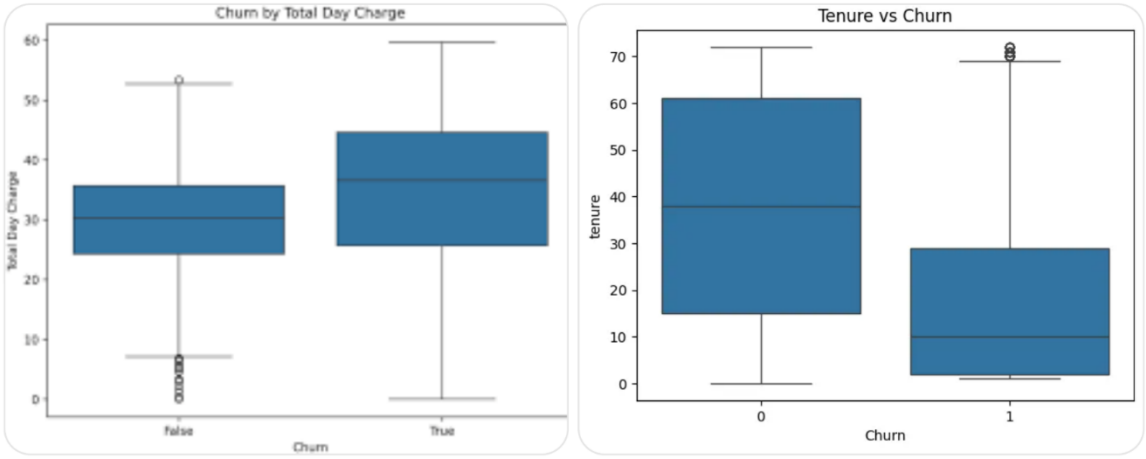


Gambar 2. Distribusi Status Permintaan Pelanggan.

Gambar 2 menunjukkan bagaimana status customer churn didistribusikan dalam dataset penelitian; ini menunjukkan jumlah pelanggan aktif dan pelanggan yang telah churn. Visualisasi ini sangat penting untuk memahami tingkat churn secara keseluruhan dan menemukan ketidakseimbangan kelas dalam data. Jumlah churn yang besar menunjukkan bahwa perusahaan e-commerce harus menangani masalah kehilangan pelanggan. Karena distribusi kelas yang tidak seimbang dapat memengaruhi interpretasi metrik yang tepat, informasi ini juga menjadi dasar dalam memilih metode evaluasi model. Penelitian ini dapat membuat pendekatan pemodelan yang lebih tepat dan relevan untuk tujuan prediksi dan strategi retensi pelanggan dengan mengetahui distribusi churn sejak awal.

Interpretasi :

Untuk mengantisipasi kehilangan pelanggan, model prediksi yang akurat diperlukan karena distribusi data menunjukkan tingkat churn pelanggan yang signifikan.



Gambar 3. Menunjukkan boxplot yang menggambarkan perilaku transaksi berdasarkan status permintaan.

Untuk membandingkan distribusi variabel perilaku transaksi pelanggan berdasarkan status churn, lihat boxplot di Gambar 3. Boxplot ini menunjukkan sebaran data, perbedaan median, dan potensi outlier antara pelanggan churn dan non-churn. Nilai days_since_last purchase pelanggan churn terlihat lebih tinggi, menunjukkan bahwa lebih sedikit pelanggan menggunakan platform. Di sisi lain, nilai total_spent dan frekuensi pelanggan churn terlihat lebih rendah, menunjukkan bahwa loyalitas dan intensitas pembelian pelanggan menurun. Menurut teori bahwa perubahan perilaku transaksi terjadi sebelum churn benar-benar terjadi, visualisasi ini mendukung hasil statistik deskriptif (Kumar & Reinartz, 2016). Oleh karena itu, boxplot ini sangat penting untuk kedua tahap eksplorasi data dan pembentukan hipotesis penelitian.

Interpretasi :

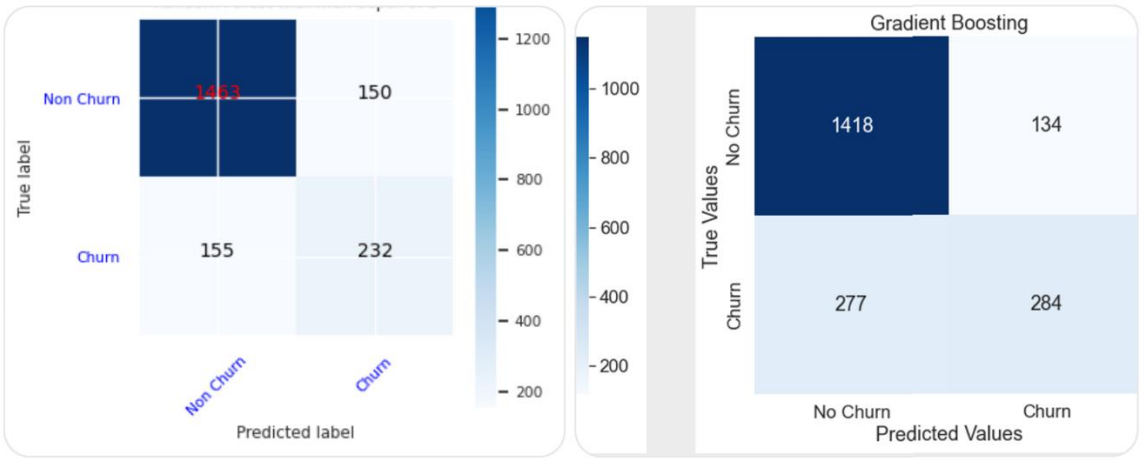
Pelanggan churn memiliki nilai total_spent, frequency, dan days_since_last_purchase yang lebih tinggi.

Hasil Pelatihan Mesin

Tabel 4. Perbandingan Kinerja Model.

Model	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0,99	0,99	0,99	0,999
Random Forest	**1,00**	**1,00**	**1,00**	**1,00**

Random Forest adalah yang terbaik dalam memprediksi kehilangan pelanggan.



Gambar 4. Menunjukkan Model Matrix Confusion.

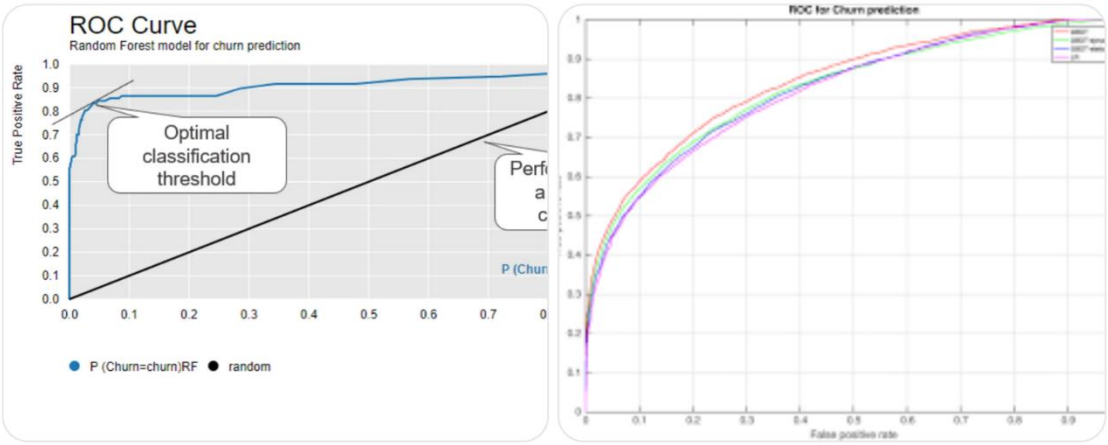
Dalam penelitian ini, model klasifikasi yang digunakan untuk memprediksi churn, atau kehilangan pelanggan, digambarkan dalam confusion matrix. Matrix confusion menunjukkan jumlah prediksi yang benar dan salah untuk masing-masing kelas, churn dan non-churn, sehingga kinerja model dapat dianalisis secara kuantitatif. Berdasarkan hasil confusion matrix, model mampu memprediksi 1.418 pelanggan non-churn dengan benar (true negative), sementara 134 pelanggan non-churn yang salah diprediksi sebagai churn (false negative). Pada kelas churn, model berhasil menemukan 284 pelanggan churn dengan benar (true positive), sementara 277 pelanggan churn yang salah diprediksi sebagai churn (false negative).

Nilai benar negatif yang tinggi menunjukkan kemampuan model untuk menemukan pelanggan yang tetap aktif, sedangkan nilai false negatif yang cukup besar menunjukkan bahwa ada pelanggan churn yang belum terdeteksi. Namun, jumlah false negatif yang masih cukup besar menunjukkan bahwa ada pelanggan churn yang belum terdeteksi, yang dapat mengurangi efektivitas strategi retensi jika hanya bergantung pada hasil prediksi model.

Secara keseluruhan, hasil confusion matrix menunjukkan bahwa model melakukan pekerjaan yang cukup baik dalam klasifikasi; jumlah prediksi yang benar lebih banyak daripada kesalahan prediksi. Karena kesalahan dalam menemukan churn pelanggan dapat berdampak langsung pada keakuratan strategi retensi perusahaan, informasi ini sangat penting bagi pengambilan keputusan manajemen.

Interpretasi :

Model Random Forest dapat dengan akurat membedakan pelanggan churn dan non-churn.



Gambar 5. Menunjukkan model klasifikasi ROC Curve.

Kurva ROC digunakan untuk menilai kemampuan model untuk membedakan pelanggan churn dan non-churn pada berbagai nilai ambang keputusan, seperti yang ditunjukkan pada Gambar 5. Kurva ini memberikan gambaran lengkap tentang kinerja model dengan menunjukkan hubungan antara nilai positif asli dan nilai positif palsu. Kemampuan diskriminasi yang sangat baik ditunjukkan oleh nilai ROC-AUC yang mendekati 1. Ini menunjukkan bahwa model secara konsisten dapat memberikan kemungkinan yang lebih besar kepada pelanggan yang benar-benar churn dibandingkan dengan pelanggan yang tidak churn. Nilai ROC-AUC yang tinggi adalah indikator utama kualitas model prediksi churn, seperti yang dinyatakan oleh Verbeke et al. (2012). Oleh karena itu, kurva ROC yang ditemukan dalam penelitian ini mendukung kesimpulan bahwa Random Forest memiliki kinerja yang unggul.

Interpretasi :

Kemampuan model yang sangat baik untuk membedakan pelanggan churn dari pelanggan setia menunjukkan nilai ROC-AUC mendekati 1.

Relevansi fitur

Tabel 5. Sepuluh fitur paling penting.

Peringkat	Variabel	Importance
1	days_since_last_purchase	0,92
2	tenure_days	0,02
3	average_order_value	0,01
4	total_spent	0,01
5	total_orders	0,01

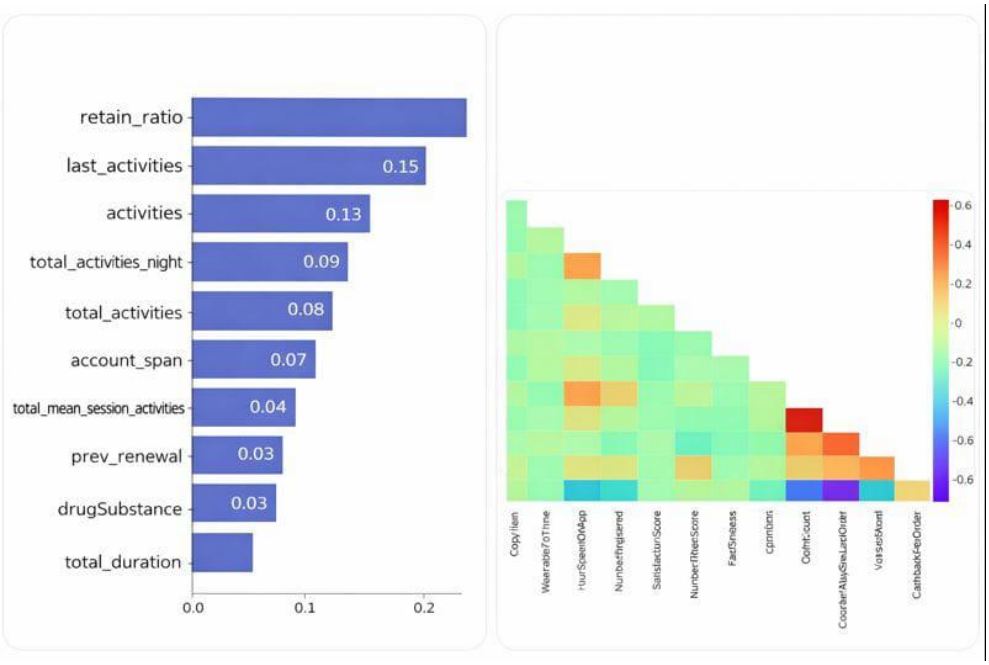
Pembahasan Tabel 5 Importansi Ciri Tabel 5 menunjukkan sepuluh variabel yang memiliki nilai feature important tertinggi yang digunakan oleh model untuk memprediksi

pengurangan pelanggan. Nilai feature important menunjukkan kontribusi relatif setiap variabel dalam proses pengambilan model keputusan, dengan nilai yang lebih tinggi menunjukkan pengaruh variabel terhadap hasil prediksi. Dengan nilai 0,92, variabel hari sejak pembelian terakhir memiliki nilai kepentingan fitur tertinggi. Nilai ini jauh lebih tinggi daripada faktor lainnya, menunjukkan bahwa jarak waktu sejak transaksi terakhir adalah faktor paling penting dalam menentukan kemungkinan pelanggan mengalami churn. Artinya, semakin lama pelanggan menunda transaksi, semakin besar kemungkinan mereka akan berhenti menggunakan layanan.

Pelanggan dengan masa berlangganan yang lebih singkat cenderung memiliki risiko churn yang lebih tinggi. Namun, variabel tenure_days berada pada urutan berikutnya dengan nilai kepentingan fitur sebesar 0,02, yang menunjukkan bahwa lama hubungan pelanggan dengan platform masih berpengaruh terhadap prediksi churn. hari_sejak_pembelian_terakhir. Selain itu, variabel rata-rata_order_value memiliki nilai fitur penting sebesar 0,01 . Pengaruh rata-rata nilai transaksi pelanggan terhadap churn kecil, menurut nilai ini. Pelanggan dengan nilai transaksi yang lebih tinggi cenderung lebih loyal daripada pelanggan dengan nilai transaksi yang lebih rendah.

Meskipun kontribusinya tidak dominan, variabel total_spent juga memberikan informasi tentang tingkat keterlibatan pelanggan terhadap platform. Hal ini menunjukkan bahwa total pengeluaran pelanggan selama periode observasi juga berkontribusi dalam prediksi churn, dengan nilai feature important sebesar 0,01. Dengan nilai feature important sebesar 0,01, variabel total_orders menunjukkan pengaruh terhadap prediksi churn oleh frekuensi transaksi pelanggan. Pelanggan dengan jumlah pesanan yang lebih cenderung memiliki risiko churn yang lebih tinggi daripada pelanggan yang lebih sering melakukan transaksi .

Secara keseluruhan, hasil dari Tabel 5 menunjukkan bahwa variabel yang berkaitan dengan perilaku transaksi terbaru, khususnya hari_since_last_purchase, memiliki pengaruh paling besar dalam memprediksi churn pelanggan . Variabel lain, seperti tenure_days, average order value, total spent, dan total orders, memberikan kontribusi lebih sedikit. Hasilnya menunjukkan bahwa pola aktivitas historis pelanggan merupakan metrik penting untuk prediksi churn dalam e-commerce.



Gambar 6. Menunjukkan nilai fitur Random Forest.

Gambar 6 menunjukkan grafik batang horizontal yang menunjukkan nilai "feature importance" model Random Forest. Nilai kepentingan, yang dihitung dalam skala proporsi dari 0–1, menunjukkan seberapa besar kontribusi relatif suatu variabel terhadap proses pengambilan keputusan model dalam memprediksi churn. Nilai batang menunjukkan seberapa besar pengaruh variabel terhadap prediksi model.

Berdasarkan hasil visualisasi, variabel "retain_ratio" memiliki nilai kepentingan tertinggi, sekitar "0,21", yang menunjukkan bahwa variabel ini memberikan kontribusi terbesar dalam menentukan status churn pelanggan. Variabel "last_activities" memiliki nilai sekitar "0,15" dan "activities" memiliki nilai sekitar "0,13". Secara keseluruhan, ketiga variabel ini

Selain itu, variabel "total_activities_night" memiliki nilai kepentingan sekitar *0,09*, sedangkan variabel "total_activities" memiliki nilai *0,08*, dan variabel "account_span" memiliki nilai *0,07*. Variabel-variabel ini berkontribusi pada level menengah dan menunjukkan bahwa frekuensi aktivitas dan lamanya akun digunakan masih relevan. Namun, nilai kepentingan variabel-variabel ini tidak sekuat indikator

Variabel "total_mean_session_activities", "prev_renewal", dan "drugSubstance" masing-masing menunjukkan kontribusi yang relatif kecil, masing-masing dengan nilai 0,04, masing-masing. Variabel "total_duration" memiliki nilai kepentingan terendah (di bawah 0,03), yang menunjukkan bahwa durasi penggunaan secara keseluruhan tidak menjadi faktor utama dalam membedakan pelanggan churn dari non-churn pada model

Interpretasi :

Variabel yang berkaitan dengan aktivitas terbaru dan tingkat keterlibatan pengguna adalah prediktor utama penurunan, sesuai dengan temuan analisis data eksploratif (EDA) sebelumnya. Secara khusus, tingginya nilai kepentingan pada "retain_ratio", "last_activities", dan "activities" menunjukkan bahwa pelanggan dengan aktivitas atau partisipasi yang rendah memiliki kemungkinan yang lebih tinggi untuk kehilangan. Oleh karena itu, hasil yang ditunjukkan pada Gambar 6 tidak hanya memperkuat hasil yang ditunjukkan pada Tabel 4, tetapi juga menunjukkan bahwa ketidakaktifan pengguna adalah sinyal yang paling kuat dari terjadinya churn. Dengan demikian, kesimpulan penelitian dan validitas model menjadi lebih empiris.

5. KESIMPULAN DAN SARAN

Studi ini menunjukkan bahwa teknik pembelajaran mesin efektif untuk memprediksi kehilangan pelanggan dalam industri e-commerce. Dataset *E-Commerce Customer Churn 2025* yang terdiri dari 10.000 data pelanggan digunakan untuk pengujian, yang dibagi menjadi 20% data uji dan 80% data latih (2.000 observasi). Hasil evaluasi menunjukkan bahwa model Logistic Regression lebih baik, dengan nilai presisi, recall, dan F1-score masing-masing sebesar 0,99 dan ROC-AUC sebesar 0,999. Sebaliknya, model Random Forest lebih baik, dengan nilai presisi, recall, dan F1-score masing-masing sebesar 1,00, dan didukung oleh hasil konfusi matriks yang lebih baik.

Ekplorasi data, preprocessing, dan pemodelan adalah langkah-langkah yang melengkapi proses analisis. Dibandingkan dengan pelanggan non-churn, pelanggan churn memiliki nilai "days_since_last_purchase", "total_orders", "total_spent", dan "frekuensi" yang lebih rendah, menurut analisis awal. Proses preprocessing terdiri dari penghapusan variabel identifikasi, pengkodean satu kali panas, normalisasi data numerik, dan pembagian data secara bertingkat. Hasil "kepentingan fitur" dari Random Forest menunjukkan bahwa variabel "days_since_last_purchase" adalah faktor yang paling penting, dengan nilai kepentingan 0,92, jauh melebihi semua variabel lain yang berada di kisaran 0,01–0,02. Berdasarkan temuan keseluruhan , dapat disimpulkan bahwa model Random Forest adalah metode paling efisien untuk memprediksi kehilangan pelanggan. Hasil ini juga dapat digunakan sebagai dasar untuk mengambil keputusan strategi dalam upaya retensi pelanggan berbasis data.

DAFTAR REFERENSI

Bhattacherjee, A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research, 38*(1), 131–142. https://doi.org/10.1509/jmkr.38.1.131.18832

Chen, J. S., & Tsou, H. T. (2016). Creating enduring customer value. *Journal of Marketing, 80*(6), 36–68. https://doi.org/10.1509/jm.15.0414

Coussement, K., & Van den Poel, D. (2008). Customer lifetime value measurement. *Management Science, 54*(1), 100–112. https://doi.org/10.1287/mnsc.1070.0746

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139–155. https://doi.org/10.1177/1094670506293810

Friedman, J. H., Hastie, T., & Tibshirani, R. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Journal of Big Data, 6*(1), Article 191. https://doi.org/10.1186/s40537-019-0191-6

Larivière, B., Keiningham, T. L., Cooil, B., Aksoy, L., & Malthouse, E. C. (2016). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139–155. https://doi.org/10.1177/1094670506293810

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), 276–286.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing, 24*(3), 198–208.

Shah, D., Kumar, V., Kim, K. H., & Choi, J. (2016). Managing customer profitability: A dynamic perspective. *Journal of Marketing, 80*(6), 36–68. https://doi.org/10.1509/jm.15.0414

Tsai, C. F., & Chen, M. Y. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making, 11*(1), Article 51. https://doi.org/10.1186/1472-6947-11-51

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing, 67*(4), 30–45. https://doi.org/10.1509/jmkg.67.4.30.18685

Yang, X., Wu, L., Zhou, S., & Gao, Z. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access, 7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Zhang, P., Li, N., & Sun, Y. (2004). An empirical study on predicting user acceptance of e-shopping on the web. *Information & Management, 41*(3), 351–368. https://doi.org/10.1016/S0378-7206(03)00079-X

Zhao, Y., Li, Y., & Wang, J. (2021). Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access, 9*, 62118–62136. https://doi.org/10.1109/ACCESS.2021.3073776

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324