

Peningkatan Akurasi Prediksi Churn Pelanggan Menggunakan SMOTE dan RFE pada Algoritma Random Forest: Studi Komparasi Sektor E-Commerce dan Telekomunikasi

M. Rizki Al Akbar^{*1}, Apriansyah Putra², Ken Ditha Tania³, Ahmad Rifai

^{1,2,3,4}Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya, Indonesia
Email: ¹rizkialakbar0@gmail.com, ²apriansyah@unsri.ac.id, ³kenya.tania@gmail.com,
⁴ahmadrifai@ilkom.unsri.ac.id

Abstrak

Tingkat *churn* pelanggan menjadi tantangan krusial bagi keberlangsungan bisnis di sektor e-commerce dan telekomunikasi. Pemodelan prediktif untuk masalah ini kerap terkendala oleh tingginya dimensi fitur data dan ketidakseimbangan kelas (*class imbalance*). Penelitian ini bertujuan untuk mengukur signifikansi peningkatan performa algoritma *Random Forest* yang dioptimasi menggunakan integrasi *Recursive Feature Elimination* (RFE) dan *Synthetic Minority Over-sampling Technique* (SMOTE) melalui studi komparasi lintas sektor. Mengadopsi kerangka kerja *Knowledge Discovery in Databases* (KDD), eksperimen dieksekusi dalam empat skenario komparatif pada dataset publik. Hasil pengujian menunjukkan bahwa efektivitas optimasi bersifat sangat kontekstual terhadap karakteristik kelas data. Pada sektor E-Commerce dengan rasio ketidakseimbangan rendah ($\pm 16,8\%$), skenario optimasi RFE terbukti paling dominan; metode ini mampu mereduksi 41,4% fitur (dari 29 menjadi 17) sekaligus menaikkan *Recall* sebesar 4,73 poin persentase dan *F1-Score* sebesar 1,81 poin persentase dibanding *baseline*. Sebaliknya, pada sektor Telekomunikasi dengan tingkat ketidakseimbangan lebih tinggi ($\pm 26,5\%$), penerapan SMOTE menjadi komponen determinan yang secara krusial mendongkrak *Recall* sebesar 11,50 poin persentase untuk mendeteksi pelanggan berisiko. Disimpulkan bahwa strategi optimasi algoritma bergantung pada tingkat ketidakseimbangan data; RFE direkomendasikan sebagai intervensi utama untuk *imbalance* di bawah 20%, sedangkan SMOTE wajib diprioritaskan untuk kondisi di atas ambang batas tersebut.

Kata kunci: *Churn Pelanggan, Random Forest, SMOTE, Recursive Feature Elimination, Klasifikasi.*

Improving Customer Churn Prediction Accuracy Using SMOTE and RFE on Random Forest Algorithm: A Comparative Study of E-Commerce and Telecommunication Sectors

Abstract

Customer churn remains a critical business challenge in the e-commerce and telecommunication sectors, where predictive modeling is frequently hindered by high data dimensionality and class imbalance. This study aims to quantify the performance improvement of the Random Forest algorithm optimized through the integration of Recursive Feature Elimination (RFE) and the Synthetic Minority Over-sampling Technique (SMOTE) via a cross-sector comparative study. Adopting the Knowledge Discovery in Databases (KDD) framework, experiments were executed across four comparative scenarios using public datasets. The results reveal that optimization effectiveness is highly contextual and contingent upon class characteristics. In the E-commerce sector, characterized by a lower imbalance ratio ($\pm 16.8\%$), RFE emerged as the dominant component; it successfully reduced features by 41.4% (from 29 to 17) while improving Recall by 4.73 percentage points and F1-Score by 1.81 percentage points over the baseline. Conversely, in the Telecommunication sector, which exhibits a higher imbalance ($\pm 26.5\%$), SMOTE acted as the determining factor, crucially boosting Recall by 11.50 percentage points to detect high-risk customers. It is concluded that optimization strategies strictly depend on the class imbalance level; RFE is recommended as the primary intervention for imbalances below 20%, whereas SMOTE must be prioritized for conditions exceeding this threshold.

Keywords: *Customer Churn, Random Forest, SMOTE, Recursive Feature Elimination, Classification.*

1. PENDAHULUAN

Pertumbuhan industri digital dan e-commerce saat ini menciptakan persaingan pasar yang sangat ketat, di mana pelanggan memiliki kemudahan untuk berpindah platform layanan. Tingkat churn tahunan di industri

telekomunikasi diketahui melampaui 30%, menjadikannya salah satu sektor dengan laju kehilangan pelanggan tertinggi [1]. Fenomena perpindahan pelanggan atau churn menjadi tantangan utama bagi keberlangsungan bisnis perusahaan. Mempertahankan pelanggan lama terbukti jauh lebih krusial, mengingat biaya untuk mengakuisisi pelanggan baru bisa berkali-kali lipat lebih besar [2]. Oleh karena itu, memprediksi potensi pelanggan yang akan churn menjadi fokus utama dalam strategi Customer Relationship Management (CRM). Penelitian terkini telah sukses mengimplementasikan algoritma machine learning untuk memprediksi customer churn dengan tingkat akurasi yang tinggi pada industri telekomunikasi dan e-commerce dengan menganalisis data historis transaksi pelanggan [1]. Model komputasi terbukti mampu menggali pola perilaku laten untuk menentukan pelanggan mana yang berisiko tinggi meninggalkan layanan [2]. Dalam praktiknya, analisis ini memungkinkan perusahaan untuk mengambil tindakan preventif dan merancang strategi retensi yang tepat sasaran [3].

Penerapan model prediksi churn pada data pelanggan nyata menghadapi dua tantangan teknis utama, yaitu tingginya dimensi fitur data dan ketidakseimbangan kelas (class imbalance). Salah satu algoritma machine learning yang paling direkomendasikan dalam menangani klasifikasi adalah Random Forest. Algoritma ensemble ini bekerja dengan membangun banyak pohon keputusan untuk mencapai prediksi yang stabil. Literatur terbaru menunjukkan bahwa Random Forest dan model tree-based lainnya memiliki kinerja komparatif yang superior untuk memprediksi probabilitas churn [4]. Keunggulan utama dari metode ini terletak pada kemampuannya menangani dataset berukuran besar dengan fitur yang kompleks [5] serta mencegah overfitting yang sering dialami algoritma linier tunggal [6].

Meskipun Random Forest sangat kuat, penerapannya pada dataset churn di dunia nyata seringkali dihadapkan pada kendala tingginya dimensi data. Dataset pelanggan sering kali memiliki puluhan atribut, di mana fitur yang tidak relevan (noise) justru dapat menurunkan performa model. Oleh karena itu, teknik seleksi fitur menjadi esensial. Metode Recursive Feature Elimination (RFE) diakui sebagai teknik seleksi fitur yang efektif karena mengeliminasi atribut berbobot rendah secara berulang [7]. Integrasi RFE dengan Random Forest terbukti sukses mereduksi variabel tanpa menurunkan akurasi, sebagaimana dibuktikan pada penelitian klasifikasi kompleks seperti deteksi dini stunting [8] dan prediksi keberhasilan operasi toraks [9].

Selain masalah fitur, dataset churn selalu identik dengan masalah ketidakseimbangan kelas, di mana jumlah pelanggan yang churn jauh lebih sedikit daripada yang bertahan. Jika dibiarkan, model akan bias memprediksi kelas mayoritas sehingga nilai Recall menjadi rendah padahal metrik inilah yang paling kritis dalam konteks bisnis retensi pelanggan. Penggunaan teknik Synthetic Minority Over-sampling Technique (SMOTE) telah terbukti secara empiris mampu mengatasi ketidakseimbangan ini dan meningkatkan sensitivitas algoritma dalam mendeteksi kelas minoritas [10]. Penelitian terbaru mengonfirmasi bahwa kombinasi SMOTE dengan metode ensemble secara konsisten meningkatkan performa deteksi pelanggan yang berisiko churn pada berbagai tingkat ketidakseimbangan data [11]. Bahkan pada imbalanced data ekstrem, kombinasi SMOTE dan RFE pada Random Forest mampu mendorong skor evaluasi secara signifikan [12].

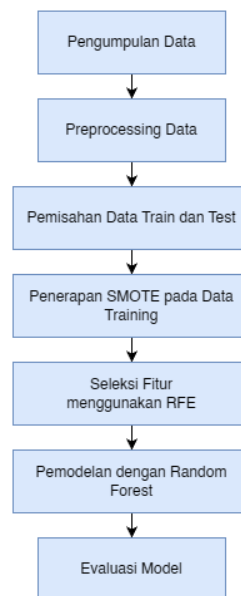
Meskipun efektivitas masing-masing komponen telah dibuktikan secara individual, belum terdapat studi yang secara eksplisit mengukur kontribusi kuantitatif integrasi simultan RFE dan SMOTE pada *Random Forest* untuk prediksi *churn* lintas sektor. Penelitian yang ada umumnya hanya menguji satu atau dua komponen optimasi [1], [5] atau mengaplikasikan kombinasi lengkap pada satu domain tunggal tanpa validasi komparatif [12], [13]. Kesenjangan ini penting untuk diisi mengingat karakteristik distribusi data di sektor e-commerce dan telekomunikasi memiliki perbedaan mencolok dari sisi struktur fitur dan tingkat ketidakseimbangan kelas.

Berdasarkan celah literatur tersebut, penelitian ini bertujuan untuk mengukur signifikansi peningkatan performa algoritma *Random Forest* setelah dioptimasi menggunakan kombinasi RFE dan SMOTE, sekaligus menguji *robustness* metode yang diusulkan melalui studi komparasi pada sektor E-Commerce dan Telekomunikasi. *Novelty* penelitian ini terletak pada pengukuran dampak setiap tahapan optimasi secara sistematis *baseline Random Forest*, *Random Forest+RFE*, *Random Forest+SMOTE*, dan *Random Forest+RFE+SMOTE* pada dua sektor dengan karakteristik data yang berbeda dalam satu kerangka eksperimen yang seragam.

2. METODE PENELITIAN

2.1. Desain Penelitian

Penelitian ini mengadopsi kerangka kerja Knowledge Discovery in Databases (KDD) yang disesuaikan untuk kebutuhan optimasi klasifikasi machine learning. Alur penelitian dirancang secara terstruktur mulai dari akuisisi data mentah hingga evaluasi performa algoritma. Pendekatan penelitian yang digunakan bersifat kuantitatif eksperimental, di mana kinerja model diukur secara objektif menggunakan metrik numerik yang terstandar. Ilustrasi tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Secara keseluruhan, penelitian ini menggunakan pendekatan kuantitatif-komputasional dengan paradigma eksperimental berbasis data sekunder. Studi dirancang sebagai penelitian komparasi dua domain (*cross-domain comparative study*) yang bertujuan mengevaluasi konsistensi dan generalisabilitas model prediksi churn pelanggan pada dua konteks industri yang berbeda, yaitu sektor e-commerce dan sektor telekomunikasi. Paradigma eksperimental dipilih karena penelitian ini melibatkan manipulasi variabel perlakuan (penerapan SMOTE dan RFE) dan pengukuran dampaknya terhadap kinerja model secara terukur dan dapat direplikasi. Seluruh tahapan eksperimen mulai dari pra-pemrosesan, penyeimbangan kelas, seleksi fitur, pemodelan, hingga evaluasi dilakukan secara sistematis dan terdokumentasi agar penelitian ini dapat direproduksi oleh peneliti lain secara independen.

2.2. Populasi, Sampel dan Instrumen Penelitian

Penelitian ini menggunakan dua dataset sekunder publik yang bersumber dari repositori Kaggle guna memfasilitasi studi komparasi lintas domain. Dataset pertama berjudul "E-Commerce Customer Churn Prediction", berdimensi 5.630 baris dengan 20 atribut, yang merepresentasikan rekam jejak transaksi dan metrik kepuasan pelanggan pada platform ritel digital. Dataset kedua adalah "Telco Customer Churn", berdimensi 7.043 baris dengan 21 atribut, yang mencatat profil demografis, layanan berlangganan, dan tagihan bulanan pelanggan di sektor telekomunikasi. Variabel target pada kedua dataset ini adalah atribut Churn, yang direpresentasikan secara biner: nilai 1 untuk pelanggan yang berhenti berlangganan, dan nilai 0 untuk pelanggan yang dipertahankan. Ringkasan karakteristik kedua dataset disajikan pada Tabel 1.

Tabel 1. Karakteristik Dataset Penelitian

Karakteristik	Dataset E-Commerce	Dataset Telekomunikasi
Sumber	Kaggle	Kaggle
Jumlah Baris	5.630	7.043
Jumlah Atribut	20	21
Rasio Kelas Minoritas	±16,8% (churn)	±26,5% (churn)
Tipe Fitur	Numerik & Kategorikal	Numerik & Kategorikal
Variabel Target	Churn (0/1)	Churn (0/1)

Penggunaan dua dataset dengan tingkat ketidakseimbangan kelas dan distribusi fitur yang berbeda ini bertujuan untuk menguji ketangguhan (*robustness*) model secara komprehensif. Seluruh data yang digunakan merupakan data sekunder publik yang tidak mengandung informasi identitas pribadi (*anonymized*), sehingga tidak memerlukan prosedur etik khusus.

Dalam penelitian berbasis komputasi (*data mining*), instrumen penelitian direpresentasikan oleh perangkat lunak (*software*) dan perangkat algoritmik yang digunakan untuk mengekstraksi pola dari data. Lingkungan

eksperimen dalam penelitian ini dibangun menggunakan bahasa pemrograman Python. Beberapa pustaka (*library*) utama yang bertindak sebagai instrumen pengolah data meliputi *Scikit-Learn* untuk pembangunan model klasifikasi dan metrik evaluasi, *Imbalanced-learn* untuk eksekusi sintesis data minoritas, serta *Pandas* untuk manipulasi struktur dataset.

Adapun instrumen algoritma utama yang digunakan sebagai subjek pengujian adalah algoritma *Random Forest*. Untuk mendukung pengujian komparatif, instrumen tambahan yang diaplikasikan mencakup metode *Synthetic Minority Over-sampling Technique* (SMOTE) sebagai instrumen penyeimbangan kelas data latih, dan *Recursive Feature Elimination* (RFE) sebagai instrumen pereduksi dimensi fitur.

2.3. Prosedur Eksperimen

Untuk mengukur kontribusi individual dan gabungan dari RFE dan SMOTE, eksperimen dirancang dalam empat skenario yang dijalankan secara berurutan pada masing-masing dataset, sebagaimana disajikan pada Tabel 2.

Tabel 2. Skenario Eksperimen

Skenario	SMOTE	RFE	Keterangan
S1	Tidak	Tidak	Baseline Random Forest
S2	Tidak	Ya	Random Forest + RFE
S3	Ya	Tidak	Random Forest + SMOTE
S4	Ya	Ya	Random Forest + SMOTE + RFE

Setiap skenario dijalankan menggunakan data latih dan data uji yang identik (pembagian data ditetapkan terlebih dahulu sebelum eksperimen dimulai) agar perbedaan hasil antar skenario semata-mata mencerminkan dampak dari komponen optimasi yang ditambahkan, bukan variasi pembagian data

2.4. Preprocessing Data

Tahapan pra-pemrosesan data bertujuan mentransformasi data mentah dari kedua dataset ke dalam format numerik yang dapat diproses oleh algoritma klasifikasi. Langkah awal meliputi penanganan nilai hilang (*missing values*) pada atribut numerik. Pada dataset E-Commerce, imputasi median diterapkan pada kolom lama berlangganan, sedangkan pada dataset Telekomunikasi, imputasi median diterapkan untuk menangani anomali kekosongan data (*blank strings*) pada atribut total tagihan. Selanjutnya, fitur kategorikal pada kedua dataset ditransformasi menjadi bentuk numerik; fitur biner diproses dengan Label Encoding, sementara fitur multikategori menggunakan One-Hot Encoding. Setelah itu, rentang nilai antar atribut dinormalisasi menggunakan Min-Max Scaler agar komputasi algoritma tidak didominasi oleh atribut dengan rentang nilai besar. Terakhir, dataset dipartisi menjadi 80% data latih (*training set*) untuk membangun model, dan 20% data uji (*testing set*) untuk memvalidasi performa prediksi.

2.5. Penyeimbangan Kelas dengan SMOTE

Distribusi kelas pada dataset *churn* secara inheren tidak seimbang (*imbalanced*). Untuk menghindari bias model terhadap kelas mayoritas, metode *Synthetic Minority Over-sampling Technique* (SMOTE) diaplikasikan secara eksklusif pada data latih. SMOTE mensintesis sampel data minoritas baru dengan menginterpolasi nilai dari sampel minoritas yang ada dan tetangga terdekatnya (*k-nearest neighbors*). Secara matematis, data sintetis x_{new} dirumuskan sebagai:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \quad (1)$$

di mana x_i adalah sampel observasi kelas minoritas x_{zi} merupakan sampel observasi dari tetangga terdekat yang dipilih secara acak, dan λ adalah bobot pengali acak dalam interval $[0, 1]$. Penerapan SMOTE hanya pada data latih bukan pada keseluruhan dataset merupakan prosedur kritis untuk mencegah data leakage yang dapat menyebabkan estimasi performa model yang terlalu optimis (*overly optimistic*). Parameter jumlah tetangga ditetapkan pada $k = 5$ sesuai dengan nilai default yang direkomendasikan dalam literatur [11].

2.6. Seleksi Fitur dengan Recursive Feature Elimination (RFE)

Guna meningkatkan efisiensi komputasi dan mengurangi *noise*, reduksi dimensi dilakukan menggunakan metode RFE berbasis *wrapper*. RFE bekerja secara iteratif: melatih model *Random Forest* awal pada seluruh

himpunan fitur, menghitung tingkat signifikansi tiap atribut (*feature importance*), dan mengeliminasi atribut dengan skor signifikansi terendah. Proses pemangkasan (*pruning*) ini diulang secara rekursif hingga tercapai subset fitur spasial yang optimal untuk mengklasifikasikan status *churn*. Penentuan jumlah fitur optimal dilakukan melalui prosedur cross-validation (CV) dengan $k = 5$, di mana nilai F1-Score pada setiap jumlah fitur dievaluasi secara sistematis. Subset fitur dengan nilai F1-Score CV tertinggi dipilih sebagai input final untuk pelatihan model. Pemisahan tahap seleksi fitur (dilakukan pada data latih) dari evaluasi akhir (dilakukan pada data uji) dijaga ketat untuk menghindari kebocoran informasi.

2.7. Algoritma Random Forest

Pemodelan utama dilakukan menggunakan *Random Forest*, yakni pendekatan *ensemble learning* yang mengintegrasikan sekumpulan besar *Decision Trees* secara independen. Algoritma ini menggunakan teknik *Bootstrap Aggregating* (Bagging), di mana setiap pohon dilatih menggunakan subset data dan subset fitur acak. Untuk tugas klasifikasi komputasional, prediksi akhir (*output*) ditentukan berdasarkan agregasi pemungutan suara terbanyak (*majority voting*) dari seluruh pohon yang terbentuk, sehingga menghasilkan model yang tangguh (*robust*) terhadap varians dan *overfitting*. Konfigurasi hiperparameter awal yang digunakan dalam penelitian ini adalah: jumlah pohon ($n_estimators$) = 100, kriteria pemisahan ($criterion$) = 'gini', kedalaman maksimum pohon (max_depth) = None (tumbuh hingga murni), dan $random_state = 42$ untuk memastikan reproduktibilitas hasil eksperimen. Nilai-nilai ini konsisten diterapkan pada seluruh skenario eksperimen agar perbandingan antar kondisi bersifat adil dan terukur.

2.8. Evaluasi Kinerja Model

Teknik analisis data dan uji validitas model pada penelitian ini dieksekusi pada tahapan evaluasi kinerja. Untuk mengukur tingkat validitas dan reliabilitas algoritma secara empiris, kinerja model prediktif dievaluasi menggunakan *Confusion Matrix* pada data uji. Pendekatan analisis ini menghasilkan matriks klasifikasi berupa *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Keempat nilai metrik ini digunakan untuk mengalkulasi parameter kinerja komputasional dengan rincian sebagai berikut.

Accuracy berguna untuk mengukur seberapa akurat model dalam mengklasifikasikan data secara keseluruhan, baik kelas positif maupun negatif. Persamaannya adalah sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision berguna untuk mengukur seberapa banyak dari prediksi positif yang benar-benar positif. Precision berguna dalam situasi di mana biaya kesalahan prediksi positif (*False Positive*) lebih tinggi, dimana persamaannya adalah sebagai berikut.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall berguna untuk mengukur seberapa banyak dari keseluruhan data aktual positif yang berhasil diprediksi secara tepat oleh model. Recall berguna dalam situasi di mana biaya kegagalan dalam mendeteksi kelas positif (*False Negative*) sangat merugikan, dimana persamaannya adalah sebagai berikut.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score berguna untuk mengukur nilai rata-rata harmonik antara nilai Precision dan Recall. F1-Score berguna dalam situasi di mana dataset memiliki distribusi kelas yang tidak seimbang (*imbalanced data*) guna mencari keseimbangan antara metrik Precision dan Recall, dimana persamaannya adalah sebagai berikut.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Sebagai validasi akhir, metrik AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) digunakan sebagai metrik tambahan untuk mengukur kemampuan diskriminatif model secara keseluruhan terhadap threshold klasifikasi yang bervariasi. Nilai AUC mendekati 1,0 mengindikasikan model dengan kemampuan pemisahan kelas yang sangat baik, sedangkan nilai 0,5 menunjukkan kemampuan yang tidak lebih baik dari tebakan acak. Metrik ini dipilih karena bersifat *threshold-independent* sehingga memberikan evaluasi yang lebih komprehensif pada data tidak seimbang [10].

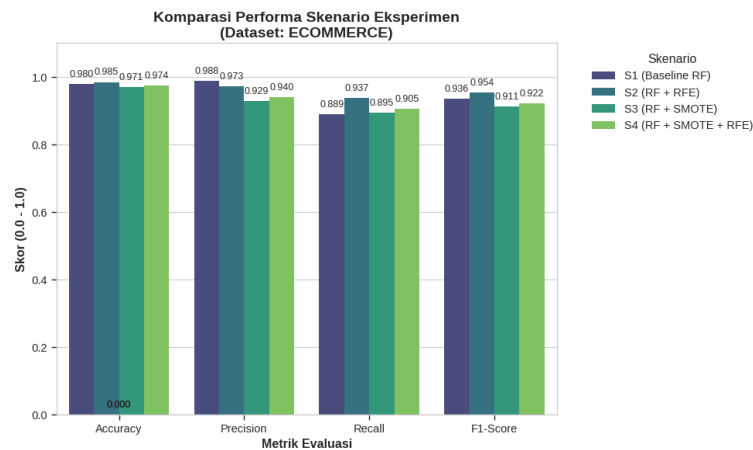
3. HASIL DAN PEMBAHASAN

3.1. Hasil Eksperimen

Eksperimen dijalankan pada empat skenario optimasi untuk masing-masing dataset. Hasil lengkap kinerja model pada *dataset* E-Commerce dan Telekomunikasi disajikan pada Tabel 3 dan Tabel 4 (Visualisasi grafik komparasi secara komprehensif dengan pembulatan tiga angka desimal dapat dilihat pada Gambar 2 dan Gambar 3).

Tabel 3. Hasil Eksperimen Dataset E-Commerce

Skenario	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Fitur
S1 (Baseline RF)	0.9796	0.9883	0.8895	0.9363	0.9988	29
S2 (RF + RFE)	0.9849	0.9727	0.9368	0.9544	0.9991	17
S3 (RF + SMOTE)	0.9707	0.9290	0.8947	0.9115	0.9962	29
S4 (RF+SMOTE+RFE)	0.9742	0.9399	0.9053	0.9223	0.9963	27

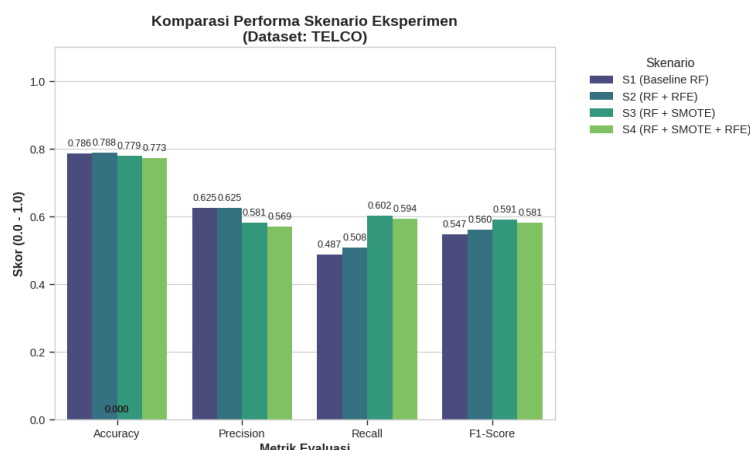


Gambar 2. Komparasi Performa Model pada Sektor E-Commerce

Berdasarkan data pada Tabel 3 dan visualisasi pada Gambar 2, dapat diidentifikasi bahwa pada dataset E-Commerce, skenario S2 (RF+RFE) mencatat kinerja terbaik pada empat dari lima metrik utama: Accuracy (0,9849), Recall (0,9368), F1-Score (0,9544), dan AUC-ROC (0,9991). Dibandingkan *baseline* S1, penerapan RFE meningkatkan Recall sebesar **+4,73 poin persentase** dan F1-Score sebesar **+1,81 poin persentase**, disertai reduksi dimensi fitur dari 29 menjadi 17. Skenario S4 (RF+SMOTE+RFE) memberikan peningkatan Recall sebesar +1,58 poin dibanding *baseline*, namun sedikit di bawah S2 untuk seluruh metrik. Penerapan SMOTE secara terpisah (S3) justru menurunkan Accuracy dan Precision dibanding *baseline*, mengindikasikan bahwa pada dataset ini ketidakseimbangan kelas tidak cukup parah untuk membutuhkan oversampling yang agresif.

Tabel 4. Hasil Eksperimen Dataset Telekomunikasi

Skenario	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Fitur
S1 (Baseline RF)	0.7864	0.6254	0.4866	0.5474	0.8212	30
S2 (RF + RFE)	0.7885	0.6250	0.5080	0.5605	0.8211	20
S3 (RF + SMOTE)	0.7793	0.5814	0.6016	0.5913	0.8247	30
S4 (RF+SMOTE+RFE)	0.7729	0.5692	0.5936	0.5812	0.8273	22



Gambar 3. Komparasi Performa Model pada Sektor Telekomunikasi

Pada dataset Telekomunikasi pada Tabel 4 dan Gambar 3, pola hasil berbeda secara signifikan. Skenario S3 (RF+SMOTE) mencatat Recall tertinggi (0,6016) dan F1-Score tertinggi (0,5913), sedangkan S4 (RF+SMOTE+RFE) unggul pada AUC-ROC (0,8273). Peningkatan Recall dari S1 ke S3 sebesar **+11,50 poin persentase** merupakan peningkatan paling substansial di seluruh eksperimen. Hal ini mengonfirmasi bahwa pada dataset dengan ketidakseimbangan kelas yang lebih besar ($\pm 26,5\%$), SMOTE memberikan kontribusi nyata. Meskipun Accuracy S3 dan S4 sedikit di bawah *baseline*, trade-off ini dapat diterima mengingat Recall metrik paling kritis untuk kasus deteksi *churn* justru meningkat signifikan.

Tabel 5. Efisiensi Reduksi Dimensi Fitur oleh RFE

Dataset / Skenario	Fitur Awal	Fitur Setelah RFE	Reduksi
E-Commerce (S2 & S4)	29	17	41,4%
Telekomunikasi – S2	30	20	33,3%
Telekomunikasi – S4	30	22	26,7%

Tabel 5 merangkum kontribusi RFE dalam hal efisiensi dimensi. Reduksi sebesar 41,4% pada E-Commerce dan 33,3% pada Telekomunikasi (skenario S2) tidak hanya berdampak pada efisiensi komputasi, tetapi juga meningkatkan interpretabilitas model sehingga memudahkan identifikasi fitur-fitur paling berpengaruh terhadap keputusan *churn* pelanggan.

3.2. Pembahasan

Temuan pada dataset E-Commerce menunjukkan bahwa RFE memberikan kontribusi paling dominan dibanding SMOTE. Hal ini dapat dijelaskan oleh karakteristik dataset: rasio ketidakseimbangan kelas yang relatif rendah ($\pm 16,8\%$) membuat model *baseline* sudah mampu belajar dari kelas minoritas dengan cukup baik, sehingga oversampling tidak berdampak signifikan. Sebaliknya, reduksi dimensi dari 29 menjadi 17 fitur melalui RFE secara efektif mengeliminasi *noise* yang menghambat generalisasi model, sejalan dengan temuan Marpaung *et al.* [8] yang membuktikan bahwa integrasi RFE dengan Random Forest mampu meningkatkan performa klasifikasi secara signifikan pada dataset berdimensi tinggi. Peningkatan Recall yang dicapai S2 juga sejalan dengan hasil Putra *et al.* [9] yang menunjukkan bahwa seleksi fitur berbasis RFE meningkatkan sensitivitas model dalam mendeteksi kelas positif pada dataset medis kompleks.

Pada dataset Telekomunikasi, dominasi SMOTE yang terlihat pada skenario S3 konsisten dengan temuan Andespa *et al.* [10] yang menyatakan bahwa SMOTE paling efektif pada data dengan ketidakseimbangan kelas sedang hingga tinggi (di atas 20%). Peningkatan Recall sebesar 11,50 poin persentase yang diperoleh S3 dibanding *baseline* mengonfirmasi argumen Ahsani *et al.* [12] bahwa SMOTE secara kritis meningkatkan sensitivitas model terhadap kelas minoritas. Namun, nilai Accuracy S3 yang sedikit turun mengindikasikan adanya *trade-off* antara sensitivitas dan spesifisitas fenomena yang umum terjadi ketika oversampling diaplikasikan tanpa penyesuaian *threshold* [6]. Rendahnya efek gabungan S4 dibanding S3 pada dataset ini menunjukkan bahwa dalam kondisi tertentu, kombinasi RFE setelah SMOTE dapat mengeliminasi fitur sintesis yang sebetulnya informatif.

Perbandingan lintas sektor menghasilkan *key finding* utama penelitian ini: **efektivitas kombinasi RFE dan SMOTE bersifat kontekstual dan bergantung pada tingkat ketidakseimbangan kelas dataset**. Pada sektor E-Commerce dengan *class imbalance* rendah, RFE menjadi komponen kunci peningkatan performa. Pada sektor

Telekomunikasi dengan *class imbalance* lebih tinggi, SMOTE menjadi komponen determinan terutama untuk metrik Recall dan F1-Score yang relevan secara bisnis. Temuan ini memperluas validasi empiris dari studi-studi sebelumnya [9], [10], [12] yang umumnya hanya menguji kombinasi tersebut pada satu domain, dan memberikan panduan praktis bagi praktisi dalam memilih strategi optimasi yang tepat sesuai karakteristik data.

Dari perspektif praktis, penelitian ini memberikan implikasi langsung bagi tim *data science* di industri. Untuk platform e-commerce dengan data pelanggan yang relatif seimbang, penerapan RFE sudah cukup untuk meningkatkan kualitas model prediksi *churn* secara signifikan, sekaligus menghemat sumber daya komputasi melalui reduksi dimensi [14]. Sebaliknya, untuk perusahaan telekomunikasi yang umumnya menghadapi ketidakseimbangan kelas lebih tinggi, penerapan SMOTE pada data latih merupakan langkah yang diprioritaskan, khususnya untuk meningkatkan kemampuan model dalam mendeteksi pelanggan yang benar-benar berisiko *churn* (Recall) [15].

3.3. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam interpretasi hasil. Pertama, kedua dataset yang digunakan merupakan data publik dari Kaggle yang mungkin tidak sepenuhnya merepresentasikan kondisi data pelanggan di lingkungan produksi nyata, di mana distribusi fitur dan tingkat *churn* dapat bervariasi secara dinamis. Kedua, penelitian ini tidak melakukan optimasi *hyperparameter* secara menyeluruh (misalnya melalui *GridSearchCV*); hasil yang lebih optimal dimungkinkan dengan tuning yang lebih intensif. Ketiga, perbandingan hanya mencakup empat skenario berbasis Random Forest dan tidak menyertakan algoritma lain seperti XGBoost atau LightGBM sebagai *baseline* pembanding, sehingga generalisasi temuan terhadap algoritma lain memerlukan studi lanjutan.

4. KESIMPULAN

Penelitian ini berhasil mengukur secara kuantitatif kontribusi individual dan gabungan dari RFE dan SMOTE terhadap peningkatan performa prediksi *churn* pelanggan berbasis Random Forest melalui studi komparasi lintas sektor E-Commerce dan Telekomunikasi. Berdasarkan hasil eksperimen pada empat skenario optimasi bertahap, tiga kesimpulan utama dapat ditarik.

Pertama, RFE terbukti sebagai komponen optimasi yang paling dominan pada dataset E-Commerce (*imbalance* $\pm 16,8\%$), di mana penerapannya mampu meningkatkan Recall sebesar +4,73 poin persentase dan F1-Score sebesar +1,81 poin persentase dibandingkan baseline, disertai efisiensi dimensi fitur sebesar 41,4% (29 $\rightarrow$ 17 fitur). Kedua, SMOTE terbukti sebagai komponen determinan pada dataset Telekomunikasi (*imbalance* $\pm 26,5\%$), dengan peningkatan Recall yang paling substansial di seluruh eksperimen sebesar +11,50 poin persentase yang secara bisnis berarti kemampuan deteksi pelanggan *churn* yang jauh lebih tinggi. Ketiga, efektivitas kombinasi SMOTE dan RFE bersifat kontekstual: tidak ada konfigurasi yang secara universal optimal. Tingkat ketidakseimbangan kelas dataset menjadi faktor penentu utama dalam memilih strategi optimasi yang tepat.

Implikasi praktis dari penelitian ini adalah panduan berbasis bukti bagi praktisi: gunakan ambang batas *imbalance* $\pm 20\%$ sebagai titik keputusan strategis. Pada kondisi *imbalance* di bawah ambang ini, RFE cukup sebagai intervensi utama; pada kondisi di atasnya, SMOTE menjadi prioritas. Panduan ini mempermudah tim *data science* dalam mengalokasikan sumber daya optimasi secara efisien tanpa perlu melakukan eksperimen *trial-and-error* yang ekstensif.

Penelitian selanjutnya disarankan untuk: (1) memvalidasi panduan ambang batas *imbalance* pada dataset dengan rentang ketidakseimbangan yang lebih luas; (2) mengintegrasikan optimasi *hyperparameter* via *GridSearchCV* atau Bayesian Optimization ke dalam pipeline eksperimen; (3) memperluas komparasi dengan algoritma *gradient boosting* seperti XGBoost dan LightGBM; serta (4) menguji *robustness* kombinasi SMOTE+RFE+Random Forest pada domain industri lain seperti perbankan, asuransi, atau layanan berlangganan digital.

DAFTAR PUSTAKA

- [1] A. Hermawan, A. Saputra, M. D. Rafi, S. Basmallah, Y. T. S. Putra, and W. Nabila, "Implementing XGBoost model for predicting customer churn in e-commerce platforms," *Repeater: Publikasi Teknik Informatika dan Jaringan*, vol. 3, no. 2, pp. 17–31, Mar. 2025, doi: 10.62951/repeater.v3i2.401.
- [2] D. A. Ardhani and K. D. Tania, "Knowledge discovery on e-commerce customer churn using interpretable

- machine learning: A comparative study of SHAP-based classifiers,” *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2695–2702, Oct. 2025, doi: 10.30871/jaic.v9i5.10811.
- [3] H. Y. Tang and S. Ya’acob, “E-commerce customer churn prediction for the marketplace in Malaysia,” *Open International Journal of Informatics*, vol. 11, no. 2, pp. 58–66, Dec. 2023, doi: 10.11113/oiji2023.11n2.273.
- [4] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, “Customer churn prediction system: a machine learning approach,” *Computing*, vol. 104, no. 2, pp. 271–294, 2022, doi: 10.1007/s00607-021-00908-y.
- [5] R. P. Gronloh and H. Achmadi, “Prediction model using machine learning: analysis of determinants of customer churn at PT XYZ,” *Jurnal Info Sains: Informatika dan Sains*, vol. 14, no. 4, pp. 745–758, Dec. 2024. [Online]. Available: <https://ejournal.seaninstitute.or.id/index.php/InfoSains/article/view/5906>
- [6] R. E. Ako, “Effects of data resampling on predicting customer churn via a comparative tree-based random forest and XGBoost,” *Journal of Computing Theories and Applications*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.
- [7] G. Airlangga, “Comparative study of XGBoost, random forest, and logistic regression models for predicting customer interest in vehicle insurance,” *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 4, pp. 2542–2549, Oct. 2024, doi: 10.33395/sinkron.v8i4.14194.
- [8] S. H. Marpaung, F. M. Sinaga, K. H. Rambe, F. P. Simamora, and K. Kelvin, “Random forest optimization using recursive feature elimination for stunting classification,” *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 8, no. 1, p. 281, Mar. 2025, doi: 10.24014/ijaidm.v8i1.35295.
- [9] D. G. C. Putra and A. T. Putra, “Optimizing random forest for predicting thoracic surgery success in lung cancer using recursive feature elimination and GridSearchCV,” *Recursive Journal of Informatics*, vol. 2, no. 2, pp. 97–105, Sep. 2024, doi: 10.15294/rji.v2i2.73154
- [10] R. Andespa, K. Sadik, C. Suhaeni, and A. M. Soleh, “Evaluating random forest and XGBoost for bank customer churn prediction on imbalanced data using SMOTE and SMOTE-ENN,” *Media Statistika*, vol. 18, no. 1, pp. 25–36, Oct. 2025, doi: 10.14710/medstat.18.1.25-36.
- [11] R. Ridwan, E. H. Hermaliani, and M. Ernawati, “Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian,” *Comput. Sci.*, vol. 4, no. 1, pp. 80–88, 2024, [Online]. Available: <https://jurnal.bsi.ac.id/index.php/co-science/article/view/2990>
- [12] M. Ahsani, N. Taqwimi, B. B. Wahono, and H. Mulyo, “Implementation of random forest algorithm with RFE and SMOTE on cardiotocography dataset,” *Journal of Dinda Data Science, Information Technology, and Data Analytics*, vol. 5, no. 2, pp. 139–146, 2025, [Online]. Available: <http://journal.ittelkom-pwt.ac.id/index.php/dinda>
- [13] J. Shobana and C. G. Gangadhar, “E-commerce customer churn prevention using machine learning-based business intelligence strategy,” *Measurement*, vol. 270, Jan. 2023, Art. no. 110998. doi: 10.1016/j.measurement.2023.110998.
- [14] S. A. Qureshi, A. S. Imran, and M. I. Babar, “A comprehensive framework for customer churn prediction in e-commerce sector using dimensionality reduction,” *IEEE Access*, vol. 11, pp. 91321–91335, 2023, doi: 10.1109/ACCESS.2023.3301234.
- [15] T. Ahmad, A. K. Jilani, and W. Hussain, “Evaluating the business value of recall-optimized churn prediction models in the telecom industry,” *Telecommunication Systems*, vol. 85, no. 2, pp. 245–259, Feb. 2024, doi: 10.1007/s11235-023-01045-8.