SINTA Jurnal Sistem Informasi dan Teknologi Komputasi
Volume 3 Nomor 3 Juli 2026 [page: 165-173]
DOI: 10.61124/sinta.v3i3.320

IMPLEMENTASI ALGORITMA RANDOM FOREST UNTUK PREDIKSI CUSTOMER CHURN PADA PERUSAHAAN TELEKOMUNIKASI

Muhammad Junaidi[1], Roberto Kaban[2]
[1]Rekayasa Perangkat Lunak, Institut Teknologi dan Bisnis Indonesia, Indonesia
[2]Teknik Informatika, Institut Teknologi dan Bisnis Indonesia, Indonesia
*e-mail: muhammadjunaidi0208@gmail.com

Abstrak
Customer churn merupakan permasalahan krusial dalam industri telekomunikasi yang berdampak langsung pada penurunan pendapatan perusahaan. Tantangan utama dalam prediksinya adalah ketidakseimbangan kelas, di mana jumlah pelanggan *non*-churn jauh lebih besar dibandingkan pelanggan *churn*. Penelitian ini mengimplementasikan algoritma *Random Forest* dikombinasikan dengan *Synthetic Minority Oversampling Technique* (SMOTE) untuk mengatasi permasalahan tersebut. Dataset yang digunakan adalah Telco Customer Churn dari Kaggle, terdiri dari 7.043 data pelanggan dengan 21 atribut. Pra-pemrosesan meliputi pembersihan data dan label encoding, kemudian dataset dibagi dengan rasio 80:20, dan SMOTE diterapkan eksklusif pada data pelatihan. Hasil pengujian menunjukkan model mencapai *accuracy* 77,15%, *precision* 55,65%, *recall* 68,45%, dan *F1-score* 61,39%. Analisis *feature importance* mengungkapkan bahwa *Contract, OnlineSecurity*, dan *TechSupport* merupakan faktor paling dominan yang memengaruhi *churn*. Integrasi *Random Forest* dan SMOTE terbukti menghasilkan model prediksi yang andal sekaligus memberikan wawasan strategis bagi perusahaan dalam merancang program retensi pelanggan yang lebih efektif.

Kata kunci : Customer Churn; Random Forest; Machine Learning; SMOTE; Telekomunikasi

Abstract
Customer churn is a critical issue in the telecommunications industry, as it directly impacts a company's revenue. One of the main challenges in churn prediction is class imbalance, where the number of non-churn customers significantly exceeds the number of churn customers. This study implements the Random Forest algorithm combined with the Synthetic Minority Oversampling Technique (SMOTE) to address this problem. The dataset used is the Telco Customer Churn dataset from Kaggle, consisting of 7,043 customer records with 21 attributes. The preprocessing stage includes data cleaning and label encoding, after which the dataset is split into training and testing sets with an 80:20 ratio, and SMOTE is applied exclusively to the training data. The experimental results show that the model achieves an accuracy of 77.15%, precision of 55.65%, recall of 68.45%, and an F1-score of 61.39%. Feature importance analysis reveals that Contract**,** OnlineSecurity, and TechSupport are the most influential factors affecting customer churn. The integration of Random Forest and SMOTE is proven to produce a reliable prediction model while providing strategic insights for companies to design more effective customer retention programs.

Keywords: Customer Churn; Random Forest; Machine Learning; SMOTE; Telecommunications

1. PENDAHULUAN

Industri telekomunikasi merupakan salah satu sektor yang mengalami pertumbuhan paling pesat seiring kemajuan teknologi informasi dan komunikasi. Pertumbuhan ini mendorong persaingan yang semakin ketat antar perusahaan penyedia layanan, sehingga kemampuan mempertahankan pelanggan menjadi faktor kritis dalam menjaga keberlangsungan bisnis. Salah satu tantangan terbesar yang dihadapi perusahaan telekomunikasi adalah fenomena *customer churn*, yaitu kondisi ketika pelanggan memutuskan berhenti menggunakan layanan dan beralih kepada kompetitor. Tingginya tingkat *customer churn* berdampak pada penurunan pendapatan, meningkatnya biaya akuisisi pelanggan baru, serta menurunnya stabilitas bisnis perusahaan [1]. Perkembangan *machine learning* membuka peluang signifikan bagi perusahaan untuk menganalisis perilaku pelanggan berdasarkan data historis yang dimiliki, sehingga dapat membangun model prediksi yang mampu mengidentifikasi pelanggan berpotensi *churn* secara lebih dini dan akurat. Hasil prediksi tersebut dapat dijadikan dasar dalam merancang strategi retensi pelanggan yang lebih efektif serta meningkatkan kualitas layanan secara berkelanjutan [2].

Di antara berbagai algoritma *machine learning* untuk klasifikasi, *Random Forest* merupakan salah satu yang paling banyak diterapkan. *Random Forest* adalah algoritma berbasis *ensemble learning* yang membangun sejumlah *decision tree* secara bersamaan dan menggabungkan hasilnya untuk menghasilkan keputusan akhir. Algoritma ini unggul dalam menangani data berukuran besar, meminimalkan risiko *overfitting*, serta menghasilkan

DOI: 10.61124/sinta.v3i3.320

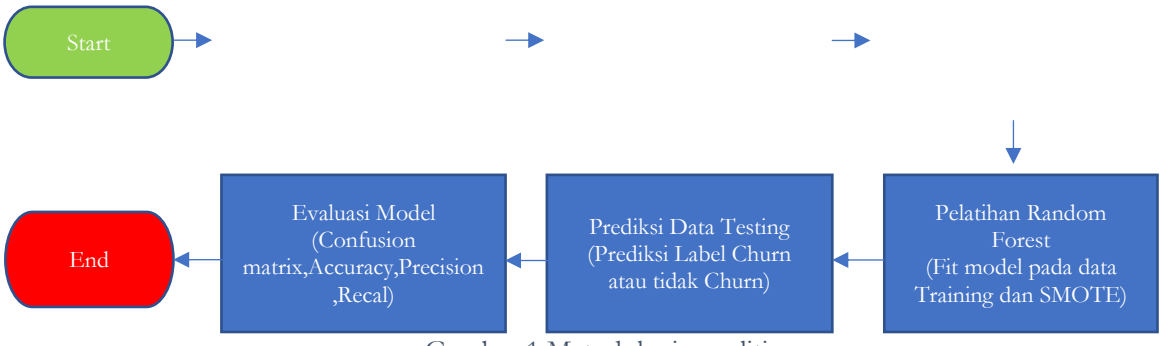
performa klasifikasi yang baik pada berbagai jenis dataset [3]. *Random Forest* mampu menghasilkan performa klasifikasi yang unggul dalam mendeteksi pelanggan berpotensi *churn* sekaligus mengidentifikasi faktor-faktor dominan yang memengaruhi keputusan pelanggan berpindah layanan [4] . Sejalan dengan itu, *machine learning* efektif membantu perusahaan telekomunikasi menganalisis perilaku pelanggan, [1] . Lebih lanjut, *Random Forest* dapat dimanfaatkan sebagai pendekatan *uplift modeling* untuk memprediksi retensi pelanggan secara lebih terukur [5].

Meskipun berbagai penelitian telah menerapkan algoritma *Random Forest* dalam prediksi *customer churn* di industri telekomunikasi, sebagian besar studi terdahulu lebih berfokus pada evaluasi akurasi model tanpa mengintegrasikan teknik penyeimbangan data secara konsisten dan menyeluruh [6]. Selain itu, penelitian yang secara spesifik menganalisis kontribusi fitur-fitur dominan terhadap keputusan *churn* pelanggan pada dataset *Telco Customer Churn* dari Kaggle masih terbatas [7]. Kesenjangan penelitian inilah yang menjadi landasan dan motivasi utama penelitian ini.

Berdasarkan permasalahan yang telah diuraikan, penelitian ini bertujuan untuk mengimplementasikan algoritma *Random Forest* dalam memprediksi customer churn pada perusahaan telekomunikasi dengan menerapkan metode SMOTE untuk menangani ketidakseimbangan data. Dataset yang digunakan adalah Telco Customer Churn yang diperoleh dari Kaggle dengan jumlah 7.043 data dan 21 fitur. Kontribusi penelitian ini terletak pada kombinasi penerapan SMOTE dan *Random Forest* yang dievaluasi secara komprehensif menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*, sehingga diharapkan dapat memberikan solusi prediktif yang lebih andal bagi perusahaan telekomunikasi dalam merancang strategi retensi pelanggan.

2. METODE

Penelitian ini dilaksanakan melalui serangkaian tahapan yang terstruktur, dimulai dari persiapan dataset hingga pengujian performa model klasifikasi secara keseluruhan, tahapan penelitian dirancang untuk menghasilkan mod*rn* y*an a* pada konteks perusahaa*ner* *salah* *lustri* telekomunikasi, se*yang* *bagi* *tensi* pelanggan [8]. Alur penelitian diawali dengan pemuatan dataset Telco Customer Churn,kemudian dilanjutkan dengan tahapan *preprocessing* data yang mencakup pembersihan data dan transformasi data kategorikal menjadi numerik. Selanjutnya, dataset dibagi menjadi data *training* dan *testing*. Untuk mengatasi permasalahan ketidakseimbangan kelas, diterapkan metode *Synthetic Minority Oversampling Technique* (SMOTE) pada data *training*. Setelah itu, model *Random Forest* dilatih menggunakan data yang telah di proses, kemudian digunakan untuk melakukan prediksi pada *testing*. Tahap akhir adalah evaluasi performa model menggunakan *confusion matrix*,*accuracy*,*precision*,*recall* dan *F1-Score*. Tahapan penelitian ini secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1 Metodologi penelitian

2.1 Dataset Penelitian

Dataset yang digunakan dalam penelitian ini adalah dataset Telco Customer Churn yang bersumber dari platform Kaggle https://www.kaggle.com/datasets/blastchar/telco-customer-churn . Dataset ini merupakan dataset publik (data sekunder) yang memuat informasi mengenai pelanggan perusahaan telekomunikasi beserta status apakah pelanggan tersebut melakukan *churn* atau tidak. Dataset terdiri dari 7.043 rekaman data dengan 21 atribut kolom, di mana 20 kolom digunakan sebagai fitur masukan dan 1 kolom digunakan sebagai label target. Salah satu atribut prediktor, yaitu *customerID*, hanya berfungsi sebagai

identitas unik pelanggan sehingga tidak digunakan dalam proses pembentukan model karena tidak memiliki pengaruh terhadap prediksi *churn*. Oleh karena itu, model *Random Forest* dibangun menggunakan 19 atribut prediktor dan 1 atribut target (*Churn*). Ketidakseimbangan kelas pada dataset customer churn merupakan tantangan umum yang berpotensi menurunkan performa model klasifikasi apabila tidak ditangani secara tepat [9]. Oleh karena itu dengan adanya kondisi ketidakseimbangan ini memperkuat alasan diterapkannya metode SMOTE dalam penelitian ini.

2.2 Preprocessing Data

Tahap *preprocessing* data merupakan langkah awal yang bertujuan untuk mempersiapkan dataset sebelum digunakan dalam proses pelatihan model. Pada penelitian ini, *preprocessing* data mencakup dua proses utama, yaitu pembersihan data (*cleaning data*) dan transformasi data kategorikal menjadi data numerik, agar dataset dapat diproses secara optimal oleh algoritma *Random Forest*. Proses transformasi data kategorikal menjadi numerik dilakukan menggunakan teknik *label encoding*. Transformasi data kategorikal ke dalam representasi numerik merupakan tahapan yang krusial dalam *preprocessing* data, mengingat sebagian besar algoritma *machine learning* hanya dapat memproses data dalam bentuk numerik [10]. Teknik ini diterapkan pada sejumlah atribut kategorikal seperti *gender*, *Partner*, *Dependents*, serta berbagai atribut layanan telekomunikasi lainnya. Label target *Churn* juga ditransformasikan menjadi nilai biner, di mana nilai 0 merepresentasikan pelanggan *non-churn* dan nilai 1 merepresentasikan pelanggan *churn*. Setelah transformasi data selesai dilakukan, tahap selanjutnya adalah pemisahan atribut fitur dan label target. Seluruh kolom selain kolom *Churn* digunakan sebagai fitur penelitian, sedangkan kolom *Churn* digunakan sebagai target klasifikasi. Dataset kemudian dibagi menjadi data *training* dan data *testing* menggunakan metode *train-test split* dengan proporsi 80:20. Data pelatihan digunakan dalam proses pembentukan model *Random Forest*, sedangkan data pengujian digunakan untuk mengevaluasi performa model.

2.3 Penerapan SMOTE

Dataset Telco Customer Churn memiliki distribusi kelas yang tidak seimbang (*imbalanced dataset*), di mana jumlah data pelanggan *non-churn* jauh lebih banyak dibandingkan pelanggan *churn*, sehingga model berpotensi lebih condong memprediksi kelas mayoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan metode *Synthetic Minority Oversampling Technique* (SMOTE), yaitu teknik *oversampling* yang membangkitkan data sintetis pada kelas minoritas menggunakan pendekatan k-nearest neighbor sehingga distribusi antar kelas menjadi lebih seimbang dan performa model klasifikasi dapat ditingkatkan [11]. Penerapan SMOTE dilakukan secara eksklusif pada data *training* setelah proses pembagian dataset, dengan tujuan menghindari kebocoran data (*data leakage*).Penerapan SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas pada dataset customer churn, sehingga model yang dihasilkan mampu mempelajari pola pelanggan *churn* secara lebih representatif [12].

2.4 Random Forest

Random Forest adalah algoritma *ensemble learning* yang bekerja dengan membangun sejumlah pohon keputusan secara paralel pada proses pelatihan. Setiap pohon dilatih menggunakan subset data dan subset fitur yang dipilih secara acak melalui teknik *bootstrap sampling*, kemudian hasil prediksi dari seluruh pohon digabungkan melalui mekanisme *majority voting* untuk menghasilkan keputusan akhir [3]. Galat generalisasi pada *Random Forest* akan konvergen menuju suatu batas tertentu seiring bertambahnya jumlah pohon, dan besarnya galat ini bergantung pada kekuatan masing-masing pohon serta korelasi antar pohon dalam hutan[3]. Pada konteks prediksi *customer churn*, *Random Forest* dipilih karena kemampuannya menangani data dengan jumlah fitur yang besar serta menghasilkan prediksi yang lebih stabil dan akurat [4]. Algoritma ini juga mampu mengukur tingkat kepentingan tiap fitur (*feature importance*) secara otomatis, sehingga dapat memberikan wawasan mengenai atribut mana yang paling berpengaruh terhadap keputusan *churn* pelanggan [4]. Dengan kombinasi data pelatihan yang telah diseimbangkan menggunakan SMOTE pada tahap sebelumnya, model *Random Forest* diharapkan dapat mempelajari pola kedua kelas secara lebih merata dan menghasilkan prediksi yang lebih andal.

2.5 Implementasi Random Forest

Evaluasi model dilakukan untuk mengukur performa algoritma *Random Forest* dalam mengklasifikasikan *customer churn*. Metrik evaluasi yang digunakan meliputi *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. *Confusion matrix* menggambarkan distribusi hasil prediksi model secara rinci melalui empat komponen, yaitu *True Positive* (TP) merupakan data pelanggan *churn* yang diprediksi benar, *True Negative* (TN) merupakan data pelanggan *non-churn* yang diprediksi benar, *False Positive* (FP) merupakan data pelanggan *non-churn* yang

DOI: 10.61124/sinta.v3i3.320

diprediksi keliru sebagai *churn*, dan *False Negative* (FN) merupakan data pelanggan *churn* yang diprediksi keliru sebagai *non-churn*. *Accuracy* mengukur proporsi prediksi yang benar secara keseluruhan terhadap seluruh data pengujian, dihitung menggunakan persamaan [13] :

$$(1)\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision mengukur ketepatan model dalam memprediksi kelas positif, yaitu proporsi prediksi *churn* yang benar-benar merupakan pelanggan *churn*, dihitung menggunakan persamaan [13]:

$$(2)\ Precision = \frac{TP}{TP + FP}$$

Recall mengukur kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya, yaitu seberapa besar proporsi pelanggan churn yang berhasil diidentifikasi oleh model. Perhitungan nilai *recall* menggunakan Persamaan sebagai berikut [13] :

$$(3)\ Recall = \frac{TP}{TP + FN}$$

F1-score merupakan rata-rata harmonik dari *precision* dan *recall* yang memberikan penilaian lebih seimbang, terutama pada kondisi data yang tidak seimbang. Perhitungan nilai *F1-score* menggunakan Persamaan sebagai berikut[13] :

$$(4)\ F1\text{-}score = 2 \times \frac{\text{Precision x recall}}{\text{Precision + recall}}$$

Hasil evaluasi menggunakan kelima metrik tersebut digunakan sebagai dasar penilaian kemampuan algoritma *Random Forest* dalam melakukan klasifikasi pelanggan *churn* pada perusahaan telekomunikasi secara menyeluruh dan objektif.

3. HASIL PENELITIAN

Bagian ini memaparkan hasil dari setiap tahapan penelitian yang telah dilakukan, meliputi *preprocessing* data, penerapan SMOTE, evaluasi model *Random Forest*, analisis *confusion matrix*, serta analisis *feature importance*.

3.1 Hasil Preprocessing Data

Dataset Telco Customer Churn yang diperoleh dari Kaggle terdiri dari 7.043 data pelanggan dengan 21 atribut, yang terdiri atas 20 atribut prediktor dan 1 atribut target, yaitu *Churn*. Pada tahap *preprocessing* dilakukan pembersihan data dan transformasi atribut kategorikal ke dalam bentuk numerik menggunakan teknik *label encoding*. Selain itu, atribut *customerID* tidak digunakan dalam proses pemodelan karena hanya berfungsi sebagai identitas unik pelanggan dan tidak memiliki pengaruh terhadap prediksi *churn*, sehingga model dibangun menggunakan 19 atribut sebagai fitur masukan dan 1 atribut sebagai label target. Hasil distribusi kelas setelah *preprocessing* ditunjukkan pada Tabel 1.

Tabel 1. Hasil *Preprocessing* Data

Kelas	Jumlah Data
Non-Churn	5.174
Churn	1.869
Total	7.043

Berdasarkan Tabel 2, proporsi pelanggan *non-churn* mencapai 73,46% sedangkan pelanggan *churn* hanya 26,54%. Ketidakseimbangan ini berpotensi membuat model cenderung memprediksi kelas mayoritas, sehingga diperlukan teknik penyeimbangan data sebelum pelatihan model dilakukan.

3.2 Hasil Pembagian Dataset

Dataset dibagi menggunakan metode *train-test split* dengan rasio 80:20. Data *training* terdiri dari 5.634 data dan data *testing* terdiri dari 1.409 data. Proporsi kelas pada kedua subset tetap dipertahankan sehingga distribusi kelas pada data *training* dan data *testing* tetap merepresentasikan distribusi kelas pada dataset awal. Hasil pembagian dataset ditunjukkan pada tabel 2.

Table 2. Hasil Pembagian Dataset

Kelas	Training	Testing
Non-Churn	4.139	1.035
Churn	1.495	374
Total	5.634	1.409

3.3 Hasil Penerapan SMOTE

Guna mengatasi permasalahan ketidakseimbangan kelas, penelitian ini menerapkan metode *Synthetic Minority Oversampling Technique* (SMOTE) pada data *training*. Metode SMOTE menghasilkan data sintetis pada kelas minoritas sehingga distribusi data antar kelas menjadi lebih proporsional [14]. Perbandingan distribusi data *training* sebelum dan sesudah penerapan SMOTE disajikan pada Tabel 3.

Tabel 3. Hasil Penerapan SMOTE

Kelas	Sebelum SMOTE	Sesudah SMOTE
Non-Churn	4.139	4.139
Churn	1.495	4.139

Berdasarkan Tabel 3, sebelum penerapan SMOTE jumlah data pelanggan *non-churn* pada data pelatihan sebanyak 4.139 data, sedangkan jumlah data pelanggan *churn* hanya sebanyak 1.495 data. Kondisi tersebut menunjukkan adanya ketidakseimbangan kelas yang berpotensi menyebabkan model lebih cenderung mempelajari karakteristik kelas mayoritas. Setelah dilakukan proses *oversampling* menggunakan SMOTE, jumlah data pada kelas minoritas meningkat dari 1.495 menjadi 4.139 data sehingga distribusi kedua kelas menjadi seimbang. Hasil tersebut menunjukkan bahwa metode SMOTE berhasil mengatasi permasalahan ketidakseimbangan kelas dan memberikan kesempatan yang lebih baik bagi model *Random Forest* untuk mempelajari karakteristik pelanggan yang berpotensi melakukan churn.

3.4 Hasil Evaluasi Model Random Forest

Setelah proses penyeimbangan data menggunakan SMOTE, model *Random Forest* dilatih menggunakan data *training* dan diuji menggunakan data *testing* sebanyak 1.409 data. Evaluasi performa model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil evaluasi model disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Model *Random Forest*

Metrik	Nilai
Accuracy	77,15%
Precision	55,65 %
Recall	68,45 %
F1- score	61,39 %

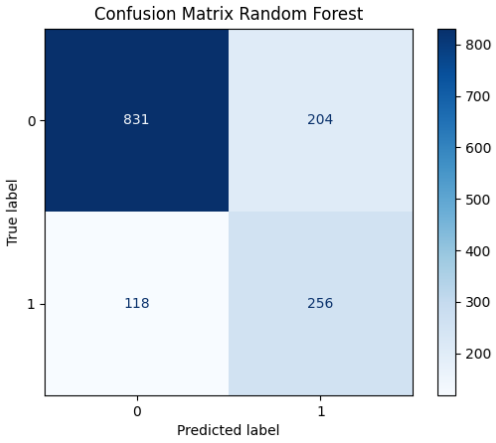
Berdasarkan Tabel 5, model *Random Forest* menghasilkan nilai *accuracy* sebesar 77,15%, yang menunjukkan bahwa model mampu mengklasifikasikan pelanggan dengan tingkat ketepatan yang cukup baik. Nilai *precision* sebesar 55,65% menunjukkan bahwa sebagian besar pelanggan yang diprediksi akan melakukan *churn* benar-benar termasuk ke dalam kelas *churn*. Nilai *recall* sebesar 68,45% mengindikasikan bahwa model mampu mengidentifikasi sebagian besar pelanggan yang sesungguhnya melakukan *churn*.

DOI: 10.61124/sinta.v3i3.320

Adapun nilai *F1-score* sebesar 61,39% menunjukkan keseimbangan yang cukup baik antara *precision* dan *recall* dalam proses klasifikasi pelanggan *churn*.

3.5 Hasil Analisis Confusion Matrix

Guna memperoleh gambaran yang lebih rinci mengenai performa model, evaluasi dilakukan menggunakan *confusion matrix*. *Confusion matrix* digunakan untuk menunjukkan jumlah prediksi yang tepat maupun yang keliru pada masing-masing kelas. Hasil *confusion matrix* model *Random Forest* disajikan pada Gambar 2.



Gambar 2. Hasil Analisis *Confusion Matrix*

Hasil *Confusion Matrix* secara rinci di sajikan dalam tabel 5.

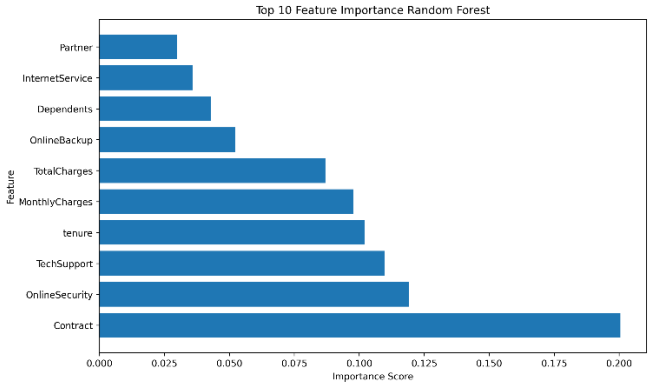
Tabel 5. Hasil *Confusion Matrix*

Aktual/Prediksi	Non-Churn	Churn	Total
Non-Churn	831	204	1.035
Churn	118	256	374
Total	949	460	1.409

Berdasarkan Tabel 5, diperoleh nilai *True Negative* (TN) sebanyak 831 data, *False Positive* (FP) sebanyak 204 data, *False Negative* (FN) sebanyak 118 data, dan *True Positive* (TP) sebanyak 256 data. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengidentifikasi pelanggan yang tidak melakukan *churn*, yang ditunjukkan oleh jumlah prediksi benar pada kelas *non*-churn yang lebih tinggi dibandingkan kesalahan prediksi. Selain itu, model juga mampu mengidentifikasi 256 pelanggan yang benar-benar melakukan *churn*. Meskipun demikian, masih terdapat 118 pelanggan yang sebenarnya melakukan *churn* namun diprediksi sebagai pelanggan *non*-churn. Hal ini menunjukkan bahwa masih terdapat kemungkinan terjadinya kesalahan klasifikasi pada kelas minoritas, sehingga peningkatan performa model masih dapat dilakukan pada penelitian selanjutnya.

3.6 Hasil Analisis Feature Importance

Salah satu keunggulan algoritma *Random Forest* adalah kemampuannya dalam mengukur tingkat kepentingan setiap fitur terhadap hasil prediksi. Analisis *feature importance* dilakukan untuk mengidentifikasi faktor-faktor yang paling memengaruhi kemungkinan pelanggan melakukan *churn*. Hasil analisis *feature importance* disajikan pada Gambar 3.



Gambar 3. Grafik feature importance model *Random Forest*.

Sepuluh fitur dengan nilai *importance* tertinggi disajikan pada Tabel 6.

Table 6. Hasil Analisis *Feature Importance*

Fitur	Importance
Contract	0,200503
OnlineSecurity	0,119158
TechSupport	0,109899
Tenure	0,102178
MonthlyCharges	0,097729
TotalCharges	0,086888
OnlineBackup	0,052142
Dependents	0,042863
InternetService	0,035998
Partner	0,029856

Berdasarkan Tabel 7, fitur *Contract* memiliki nilai *importance* tertinggi sebesar 0,200503, diikuti oleh fitur *OnlineSecurity* sebesar 0,119158 dan *TechSupport* sebesar 0,109899. Hasil tersebut menunjukkan bahwa jenis kontrak pelanggan, layanan keamanan daring, dan layanan dukungan teknis merupakan faktor yang paling berpengaruh terhadap kemungkinan pelanggan melakukan churn. Selain itu, lama pelanggan menggunakan layanan (*tenure*), besarnya biaya bulanan (*MonthlyCharges*), serta total biaya yang telah dikeluarkan pelanggan (*TotalCharges*) juga memberikan kontribusi yang cukup besar terhadap hasil prediksi.

Di samping itu, fitur *OnlineBackup*, *Dependents*, *InternetService*, dan *Partner* turut berkontribusi dalam proses klasifikasi, yang menunjukkan bahwa karakteristik layanan serta kondisi pelanggan juga memiliki keterkaitan terhadap tingkat loyalitas pelanggan. Dengan demikian, hasil analisis *feature importance* menunjukkan bahwa aspek layanan dan jenis kontrak pelanggan merupakan faktor utama yang memengaruhi keputusan pelanggan untuk tetap menggunakan atau menghentikan layanan telekomunikasi.

4. DISKUSI

Penelitian ini berhasil mengimplementasikan algoritma *Random Forest* yang dikombinasikan dengan metode SMOTE untuk memprediksi *customer churn* pada dataset Telco Customer Churn. Model yang dibangun menghasilkan nilai *accuracy* sebesar 77,15%, *precision* 55,65%, *recall* 68,45%, dan *F1-score* 61,39%, mencerminkan kemampuan yang cukup baik dalam mengklasifikasikan pelanggan yang berpotensi melakukan *churn*. Penerapan SMOTE pada data *training* terbukti efektif dalam menyeimbangkan distribusi kelas sehingga model dapat mempelajari karakteristik pelanggan *churn* secara lebih representatif. Berdasarkan *confusion matrix*, model mampu mengklasifikasikan 831 pelanggan *non-churn* secara tepat (*True Negative*) dan 256 pelanggan *churn* secara tepat (*True Positive*), dengan 204 pelanggan *non-churn* yang keliru diprediksi sebagai *churn* (*False Positive*) dan 118 pelanggan *churn* yang keliru diprediksi sebagai *non-churn* (*False Negative*). Hasil tersebut mengindikasikan bahwa model memiliki kemampuan yang memadai dalam mengenali pola perilaku

pelanggan yang berpotensi melakukan *churn*, sehingga dapat dijadikan landasan yang valid dalam menyusun strategi retensi pelanggan yang lebih tepat sasaran.

Temuan penelitian ini memiliki keterkaitan yang kuat dengan sejumlah kajian terdahulu yang membuktikan efektivitas algoritma *Random Forest* dalam prediksi *customer churn*. Ouf et al. mengusulkan kerangka hybrid yang terbukti meningkatkan akurasi prediksi churn pada dataset Telco Kaggle melalui kombinasi beberapa algoritma *machine learning* [15].Selanjutnya Thorat menunjukkan bahwa *Random Forest* mampu menghasilkan prediksi yang akurat sekaligus mengidentifikasi faktor-faktor dominan yang memengaruhi *churn* pelanggan [16]. Sejalan dengan itu, Ahmad et al. memperoleh performa prediksi yang tinggi menggunakan pendekatan *machine learning* pada big data platform di industri telekomunikasi [17]. Sementara itu, Wakhidah et al. membuktikan bahwa *Random Forest* dan *XGBoost* sama-sama memiliki kemampuan klasifikasi yang baik pada dataset Telco Customer Churn [7]. Analisis *feature importance* mengungkapkan bahwa fitur *Contract* memiliki pengaruh terbesar terhadap prediksi *churn* dengan nilai *importance* sebesar 0,200503, diikuti oleh *OnlineSecurity* (0,119158), *TechSupport* (0,109899), *tenure* (0,102178), *MonthlyCharges* (0,097729), dan *TotalCharges* (0,086888). Temuan ini mengindikasikan bahwa jenis kontrak berlangganan, kualitas layanan keamanan dan dukungan teknis, durasi berlangganan, serta besaran biaya layanan merupakan faktor-faktor utama yang memengaruhi kemungkinan pelanggan melakukan *churn*.

Secara keseluruhan, kombinasi algoritma *Random Forest* dan SMOTE terbukti mampu menghasilkan model prediksi *customer churn* dengan performa yang memadai sekaligus mengidentifikasi faktor-faktor dominan yang berkontribusi terhadap perilaku *churn* pelanggan. Meskipun demikian, penelitian ini masih memiliki keterbatasan berupa penggunaan dataset publik dari Kaggle tanpa perbandingan dengan algoritma lain maupun optimasi *hyperparameter*. Penelitian selanjutnya disarankan untuk mengeksplorasi algoritma lain seperti *XGBoost*, *LightGBM*, atau *Artificial Neural Network* guna memperoleh model prediksi yang lebih optimal dan memiliki kemampuan generalisasi yang lebih baik.

5. KESIMPULAN

Berdasarkan hasil penelitian, kombinasi algoritma *Random Forest* dan metode SMOTE terbukti efektif dalam memprediksi *customer churn* pada dataset Telco Customer Churn yang terdiri dari 7.043 data pelanggan. Penerapan SMOTE berhasil mengatasi permasalahan ketidakseimbangan kelas sehingga model dapat mempelajari karakteristik pelanggan *churn* secara lebih representatif. Model yang dibangun menghasilkan nilai *accuracy* sebesar 77,15%, *precision* 55,65%, *recall* 68,45%, dan *F1-score* 61,39%, yang mencerminkan kemampuan klasifikasi yang memadai dalam mendeteksi pelanggan berpotensi *churn*. Analisis *Feature Importance* mengungkapkan bahwa *Contract, OnlineSecurity, TechSupport, tenure, MonthlyCharges,* dan *TotalCharges* merupakan faktor-faktor paling dominan yang memengaruhi kemungkinan pelanggan melakukan *churn.* Dengan demikian, integrasi *Random Forest* dan SMOTE tidak hanya menghasilkan model prediksi *customer churn* yang andal, tetapi juga memberikan wawasan berharga mengenai faktor-faktor utama yang memengaruhi keputusan pelanggan dalam mempertahankan atau menghentikan penggunaan layanan telekomunikasi.

DAFTAR PUSTAKA

[1] Achmad Mauludi Asror and I Kadek Dwi Nuryana, "Prediction and Analysis of Customer Churn at Telkomsel Using Machine Learning Approach," *J. Emerg. Inf. Syst. Bus. Intell. JEISBI*, vol. 6, no. 2, Jul. 2025, doi: 10.26740/jeisbi.v6i2.65825.

[2] Y. Zhou, W. Chen, X. Sun, and D. Yang, "Early warning of telecom enterprise customer churn based on ensemble learning," *PLOS ONE*, vol. 18, no. 10, p. e0292466, Oct. 2023, doi: 10.1371/journal.pone.0292466.

[3] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[4] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[5] R. T. Prasetio, Y. Ramdhani, V. H. C. Putra, and P. Pratiwi, "Algoritma Random Forest Sebagai Uplift Modeling pada Prediksi Retensi Pelanggan Layanan Telekomunikasi," *J. Infortech*, vol. 7, no. 2, pp. 84–88, Dec. 2025, doi: 10.31294/infortech.v7i2.11577.

[6] M. M. S. Nurhidayat and Dyah Anggraini, "Analysis and Classification of Customer Churn Using Machine Learning Models," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 7, no. 6, pp. 1253–1259, Nov. 2023, doi: 10.29207/resti.v7i6.4933.
[7] L. N. Wakhidah, A. K. Zyen, and B. B. Wahono, "Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms," *J. Appl. Inform. Comput.*, vol. 9, no. 1, pp. 89–95, Jan. 2025, doi: 10.30871/jaic.v9i1.8740.
[8] V. Chang, K. Hall, Q. Xu, F. Amao, M. Ganatra, and V. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, vol. 17, no. 6, p. 231, May 2024, doi: 10.3390/a17060231.
[9] M. Z. Alotaibi and M. A. Haq, "Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 3, pp. 14572–14578, Jun. 2024, doi: 10.48084/etasr.7480.
[10] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024, doi: 10.3390/math12162553.
[11] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, p. 87, Jun. 2024, doi: 10.1186/s40537-024-00943-4.
[12] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results Control Optim.*, vol. 14, p. 100342, Mar. 2024, doi: 10.1016/j.rico.2023.100342.
[13] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inform.*, vol. 17, no. 1, pp. 168–192, Jan. 2021, doi: 10.1016/j.aci.2018.08.003.
[14] L. Oriana, A. D. S. Wati, A. P. Ramahdani, N. N. Safira, and E. Ismanto, "Peningkatan Kinerja Model Random Forest untuk Deteksi Kecurangan Kartu Kredit Menggunakan RandomizedSearchCV," vol. 15, no. 2.
[15] S. Ouf, K. T. Mahmoud, and M. A. Abdel-Fattah, "A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry," *J. Big Data*, vol. 11, no. 1, p. 70, May 2024, doi: 10.1186/s40537-024-00922-9.
[16] A. S. Thorat, "A Random Forest Churn Prediction Model: An Investigation of Machine Learning Techniques for Churn Prediction and Factor Identification in the Telecommunications Industry," vol. 71, no. 4, 2022.
[17] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.