

## Prediksi Customer Churn Pada Layanan Telekomunikasi Menggunakan Algoritma Random Forest (Studi Kasus: Telco Customer Churn)

<sup>1\*</sup>Muhammad Valdhano Zainal, <sup>2</sup>Muhammad Roychan Sabila, <sup>3</sup>Ade Surya Darmawan

<sup>123</sup>Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Tangerang Selatan, Indonesia

<sup>1</sup>[valdhano66@gmail.com](mailto:valdhano66@gmail.com), <sup>2</sup>[iamroychan780@gmail.com](mailto:iamroychan780@gmail.com), <sup>3</sup>[adesuryadrmwn@gmail.com](mailto:adesuryadrmwn@gmail.com)

### Abstract

*The high level of competition in the telecommunications industry makes customer churn one of the important issues in retaining customers. This study aims to build a customer churn prediction model using the Random Forest algorithm on the Telco Customer Churn dataset owned by IBM Watson Analytics. The dataset used is 7,043 customer data with 21 attributes, where the class distribution consists of 73.5% non-churn customers and 26.5% churned customers. The research process is carried out based on the CRISP-DM methodology with the help of Orange3 software. The preprocessing stage includes handling missing values, categorical data transformation using the Treat as Ordinal method, and data normalization using Z-score standardization. Model evaluation is carried out using the 10-fold cross-validation method and obtains an AUC value of 0.827, accuracy of 79.2%, F1-Score of 0.557, Precision of 0.640, and Recall of 0.493. Based on the feature importance analysis, the attributes Contract, tenure, Online Security, and Tech Support were the most influential factors in customer churn behavior. This research is expected to help telecommunications companies develop more effective, data-driven customer retention strategies.*

**Keywords:** Customer Churn, Random Forest, Orange3, CRISP-DM, Data Mining, Churn Prediction.

### Abstrak

Tingkat persaingan yang tinggi pada industri telekomunikasi menjadikan customer churn sebagai salah satu permasalahan penting dalam mempertahankan pelanggan. Penelitian ini bertujuan untuk membangun model prediksi customer churn menggunakan algoritma Random Forest pada dataset Telco Customer Churn milik IBM Watson Analytics. Dataset yang digunakan berjumlah 7.043 data pelanggan dengan 21 atribut, di mana distribusi kelas terdiri dari 73,5% pelanggan tidak churn dan 26,5% pelanggan churn. Proses penelitian dilakukan berdasarkan metodologi CRISP-DM dengan bantuan perangkat lunak Orange3. Tahap preprocessing meliputi penanganan missing value, transformasi data kategorikal menggunakan metode Treat as Ordinal, serta normalisasi data menggunakan standarisasi Z-score. Evaluasi model dilakukan menggunakan metode 10-fold cross-validation dan memperoleh nilai AUC sebesar 0,827, akurasi sebesar 79,2%, F1-Score sebesar 0,557, Precision sebesar 0,640, dan Recall sebesar 0,493. Berdasarkan hasil analisis kepentingan fitur, atribut Contract, tenure, OnlineSecurity, dan TechSupport menjadi faktor yang paling memengaruhi perilaku customer churn. Hasil penelitian ini diharapkan dapat membantu perusahaan telekomunikasi dalam menyusun strategi retensi pelanggan yang lebih efektif dan berbasis data.

**Kata Kunci:** Customer Churn, Random Forest, Orange3, CRISP-DM, Data Mining, Prediksi Churn.

### A. PENDAHULUAN

Industri telekomunikasi menjadi salah satu sektor yang berkembang sangat cepat pada era digital saat ini. Semakin luasnya penggunaan layanan seluler dan internet menyebabkan persaingan antar perusahaan telekomunikasi semakin ketat, sehingga pelanggan memiliki banyak pilihan dalam menentukan penyedia layanan yang sesuai dengan kebutuhan mereka. Kondisi tersebut menimbulkan tantangan bagi perusahaan, terutama terkait banyaknya pelanggan yang memutuskan untuk berhenti berlangganan yang dikenal dengan istilah customer churn. Fenomena churn tidak hanya menunjukkan adanya ketidakpuasan

pelanggan terhadap layanan yang digunakan, tetapi juga dapat berdampak langsung pada pendapatan serta keberlangsungan operasional perusahaan.

Permasalahan customer churn menjadi semakin rumit ketika perusahaan tidak dapat mengenali pelanggan yang berpotensi berhenti berlangganan sejak dini. Ketidakmampuan dalam mendeteksi tanda-tanda awal churn menyebabkan perusahaan kehilangan peluang untuk mengambil langkah pencegahan secara tepat waktu. Selain itu, di tengah persaingan bisnis yang semakin ketat, biaya untuk memperoleh pelanggan baru cenderung lebih tinggi dibandingkan mempertahankan pelanggan lama, sehingga

kemampuan dalam memprediksi dan mengurangi churn menjadi hal yang penting bagi keberlangsungan perusahaan.

Dampak customer churn terhadap kondisi finansial perusahaan tergolong sangat besar. Kehilangan pelanggan dapat meningkatkan biaya operasional perusahaan karena perusahaan harus mengeluarkan lebih banyak sumber daya untuk menarik pelanggan baru. Selain menyebabkan hilangnya pendapatan, churn juga dapat memengaruhi reputasi perusahaan, menurunkan efisiensi operasional, serta berdampak pada nilai perusahaan secara keseluruhan. Oleh sebab itu, kemampuan dalam memprediksi churn secara tepat dan proaktif menjadi salah satu faktor penting untuk meningkatkan daya saing perusahaan telekomunikasi di era modern.

Dataset yang digunakan dalam penelitian ini adalah Telco Customer Churn yang dipublikasikan oleh IBM Watson Analytics, yang terdiri dari 7.043 data pelanggan dengan 21 variabel. Dataset ini mencakup berbagai informasi penting, seperti jenis kelamin, lama berlangganan (tenure), jenis kontrak, layanan tambahan yang digunakan, biaya bulanan (MonthlyCharges), metode pembayaran, hingga status churn pelanggan. Variable target dalam dataset ini adalah churn yang bersifat biner, yaitu bernilai Yes apabila pelanggan telah berhenti berlangganan dan No apabila pelanggan masih aktif menggunakan layanan. Berdasarkan data tersebut, dapat dilakukan analisis untuk mengetahui pola perilaku pelanggan serta membangun model prediksi churn untuk mengidentifikasi pelanggan yang berpotensi berhenti berlangganan.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan pendekatan data mining untuk membangun model prediksi customer churn. Dengan memanfaatkan pola dari data historis pelanggan, model yang dibangun dapat mengidentifikasi pelanggan yang berpotensi berhenti berlangganan sejak dini. Hasil prediksi tersebut dapat membantu perusahaan dalam menentukan strategi retensi pelanggan secara lebih efektif dan tepat sasaran sebelum pelanggan benar-benar meninggalkan layanan.

Teknik yang diterapkan dalam penelitian ini adalah teknik klasifikasi, yaitu salah satu metode dalam data mining yang digunakan untuk memprediksi kategori atau kelas suatu data berdasarkan atribut yang dimiliki (Azmi & Voutama, 2024). Teknik ini dipilih karena variabel target churn termasuk ke dalam data kategorikal biner, sehingga permasalahan customer churn dapat diklasifikasikan sebagai masalah klasifikasi biner. Penelitian ini menggunakan algoritma Random Forest, yaitu metode ensemble learning yang membangun sejumlah pohon keputusan secara paralel dan menggabungkan hasil prediksinya untuk memperoleh hasil yang lebih akurat dan konsisten (Nugroho, 2025). Algoritma ini dipilih karena mampu mengolah dataset berdimensi tinggi, mengurangi kemungkinan overfitting, serta dapat menunjukkan tingkat pengaruh setiap fitur terhadap hasil prediksi (Bili et al., 2026). Proses analisis data dilakukan menggunakan

Orange3 dengan mengikuti tahapan metodologi CRISP-DM.

### Rumusan Masalah

Berdasarkan latar belakang diatas, rumusan masalah dalam penelitian ini adalah sebagai berikut:

- Bagaimana penerapan algoritma Random Forest dalam membangun model prediksi customer churn pada dataset Telco Customer Churn?
- Faktor apa saja yang paling berpengaruh terhadap kemungkinan pelanggan untuk melakukan churn?
- Seberapa baik performa model yang dihasilkan berdasarkan metrik evaluasi pada klasifikasi biner yang tidak seimbang?

### Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

- Membuat model klasifikasi menggunakan algoritma Random Forest untuk memprediksi customer churn pada industri telekomunikasi.
- Mengidentifikasi faktor-faktor yang paling berpengaruh terhadap perilaku churn pelanggan.
- Memberikan rekomendasi berbasis data yang dapat digunakan perusahaan dalam Menyusun strategi retensi pelanggan secara efektif.

## B. Tinjauan Pustaka

### Data Mining

Data mining merupakan suatu proses analisis data untuk menemukan pola atau informasi penting dari kumpulan data berukuran besar dengan memanfaatkan teknik statistik, matematika, kecerdasan buatan, serta machine learning (Mutiarachim et al., 2025). Data mining diartikan sebagai proses menggali pengetahuan yang bernilai dan berpotensi dari data dalam jumlah besar, di mana informasi yang diperoleh sebelumnya belum diketahui secara jelas dan tidak bersifat sederhana (Mutiarachim et al., 2025). Dalam bidang bisnis, data mining berperan dalam mengubah data mentah menjadi informasi yang dapat digunakan sebagai dasar dalam pengambilan keputusan strategis (Bili et al., 2026).

### Customer Churn

Customer churn atau attrition merupakan kondisi ketika pelanggan memutuskan untuk berhenti menggunakan layanan dari suatu perusahaan dalam jangka waktu tertentu (Bili et al., 2026). Dalam industri telekomunikasi, churn dapat terjadi secara sukarela, yaitu saat pelanggan memilih berpindah ke layanan kompetitor, maupun secara tidak sukarela akibat faktor tertentu seperti keterlambatan atau kegagalan pembayaran (Mutiarachim et al., 2025). Tingginya tingkat churn menunjukkan rendahnya tingkat kepuasan pelanggan serta menurunnya kemampuan perusahaan dalam bersaing di pasar.

## Klasifikasi

Klasifikasi adalah salah satu teknik dasar dalam data mining yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas tertentu yang telah ditentukan sebelumnya (Azmi & Voutama, 2024). Pada proses ini, algoritma machine learning mempelajari pola dari dataset yang sudah memiliki label, sehingga model dapat digunakan untuk memprediksi kelas pada data baru yang belum pernah diproses sebelumnya (Syawaludin & Hidayat, 2024).

## Random Forest

Random Forest merupakan salah satu algoritma machine learning yang digunakan dalam proses klasifikasi untuk menghasilkan prediksi yang lebih akurat. Algoritma ini bekerja dengan membangun sejumlah pohon keputusan (decision tree) dan menggabungkan hasil prediksi dari setiap pohon untuk menentukan keputusan akhir. Metode Random Forest juga mampu mengurangi risiko overfitting dan memiliki performa yang baik dalam mengolah data dengan jumlah atribut yang cukup banyak (Kencana et al., 2022)

## C. Metodologi Penelitian

### CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) adalah salah satu metode dalam data mining yang digunakan sebagai kerangka kerja penelitian secara terstruktur dan sistematis (Yoraeni & Rakhmah, 2025).

Metode ini memiliki enam tahapan utama, yaitu:

- Business Understanding**  
Tahap untuk memahami permasalahan dan tujuan penelitian yang akan dilakukan.
- Data Understanding**  
Tahap untuk mempelajari karakteristik, struktur, dan kondisi data yang digunakan.
- Data Preparation**  
Tahap persiapan data seperti pembersihan data, pemilihan atribut, dan transformasi data sebelum proses pemodelan.
- Modeling**  
Tahap penerapan algoritma untuk membangun model prediksi atau klasifikasi.
- Evaluation**  
Tahap evaluasi untuk mengukur performa model yang telah dibuat.
- Deployment**  
Tahap penerapan hasil model agar dapat digunakan dalam mendukung pengambilan keputusan.

Melalui tahapan tersebut, proses analisis data dapat dilakukan secara lebih terarah, mulai dari memahami permasalahan hingga menghasilkan model yang dapat mendukung proses pengambilan keputusan (Yoraeni & Rakhmah, 2025).

## Dataset

Dataset yang digunakan pada penelitian ini adalah Telco Customer Churn yang tersedia secara publik melalui platform IBM Watson Analytics Sample Data. Dataset tersebut berisi 7.043 data pelanggan dari perusahaan telekomunikasi fiktif yang berada di California, Amerika Serikat. Data yang tersedia mencakup informasi terkait karakteristik pelanggan, layanan yang digunakan, serta informasi akun pelanggan.

Beberapa atribut yang terdapat pada dataset meliputi data demografis seperti gender, SeniorCitizen, Partner, dan Dependents. Selain itu, dataset juga memuat informasi layanan seperti InternetService, OnlineSecurity, TechSupport, dan StreamingTV, serta informasi akun pelanggan seperti Contract, PaymentMethod, MonthlyCharges, dan TotalCharges. Variabel target dalam penelitian ini adalah Churn, yang terbagi ke dalam dua kategori, yaitu Yes dan No.

Tabel 1 Karakteristik Dataset

Karakteristik	Nilai
Jumlah Data	7.043 Data
Jumlah Atribut	21 Atribut
Atribut Numerik	3 (tenure, MonthlyCharges, TotalCharges)
Atribut Kategorikal	17 (termasuk Churn)
Distribusi kelas Churn (No / Yes)	5.174 (73,5%) / 1.869 (26,5%)
Missing Value	11 nilai kosong pada TotalCharges

## Preprocessing Data

Tahap preprocessing merupakan bagian penting dalam proses data mining karena kualitas data yang digunakan akan memengaruhi hasil prediksi model. Mengingat kualitas data memiliki pengaruh yang signifikan terhadap akurasi prediksi, tahap data preprocessing diperlukan untuk membersihkan data dari nilai yang hilang maupun data pencilan (outlier) agar model yang dihasilkan lebih stabil dan terhindar dari masalah seperti overfitting (Nugroho, 2025).

Proses preprocessing dilakukan menggunakan widget Data Preprocessing pada Orange3 melalui tiga tahapan utama berikut:

- Penanganan Missing Values**  
Tahap pertama dilakukan dengan menangani missing values menggunakan metode Average/Most Frequent. Nilai kosong pada atribut TotalCharges diisi menggunakan nilai rata-rata, sedangkan atribut kategorikal yang memiliki data kosong diisi menggunakan nilai yang paling sering muncul (modus). Metode ini dipilih karena jumlah data kosong yang ditemukan relatif sedikit sehingga tidak memberikan pengaruh yang signifikan terhadap distribusi data.
- Konversi Data Kategorikal**  
Tahap kedua dilakukan dengan mengubah data kategorikal menjadi bentuk numerik menggunakan metode Treat as Ordinal. Atribut seperti Contract, InternetService, PaymentMethod, dan atribut layanan

lainnya dikonversi menjadi nilai integer agar dapat diproses oleh model machine learning. Pendekatan ini dipilih karena mampu menjaga jumlah atribut tetap efisien dibandingkan metode one-hot encoding yang menghasilkan banyak atribut tambahan.

c. Normalisasi Data

Tahap ketiga dilakukan dengan normalisasi data menggunakan standarisasi Z-score dengan nilai rata-rata (mean) 0 dan standar deviasi 1. Proses ini bertujuan untuk menyamakan skala antar fitur numerik yang memiliki rentang nilai berbeda, seperti atribut tenure dan TotalCharges. Normalisasi diterapkan agar data menjadi lebih konsisten dan mempermudah proses analisis model.

**Penerapan Algoritma Random Forest**

Tahap pemodelan dalam penelitian ini menggunakan algoritma Random Forest untuk melakukan klasifikasi pelanggan yang berpotensi mengalami churn. Untuk memperoleh performa model yang optimal, digunakan beberapa parameter dalam proses pembentukan model. Adapun konfigurasi parameter Random Forest yang digunakan adalah sebagai berikut:

Tabel 2 Konfigurasi Parameter Random Forest

Parameter	Nilai	Keterangan
n_estimators	100	Jumlah pohon keputusan
max_depth	Tidak dibatasi	Pohon tumbuh secara penuh
min_samples_split	2	Minimum sampel untuk pemisahan node
random_state	Aktif	Menjaga konsistensi hasil

Model yang telah dikonfigurasi selanjutnya diuji menggunakan metode 10-fold cross validation. Pada metode ini, dataset dibagi menjadi sepuluh bagian, kemudian sembilan bagian digunakan sebagai data pelatihan dan satu bagian lainnya digunakan sebagai data pengujian secara bergantian. Penggunaan metode ini bertujuan agar hasil evaluasi model menjadi lebih konsisten dan tidak bergantung pada satu pembagian data tertentu. Proses evaluasi dilakukan menggunakan metrik Accuracy, AUC, Precision, Recall, dan F1-Score untuk mengukur performa model dalam mengklasifikasikan customer churn.

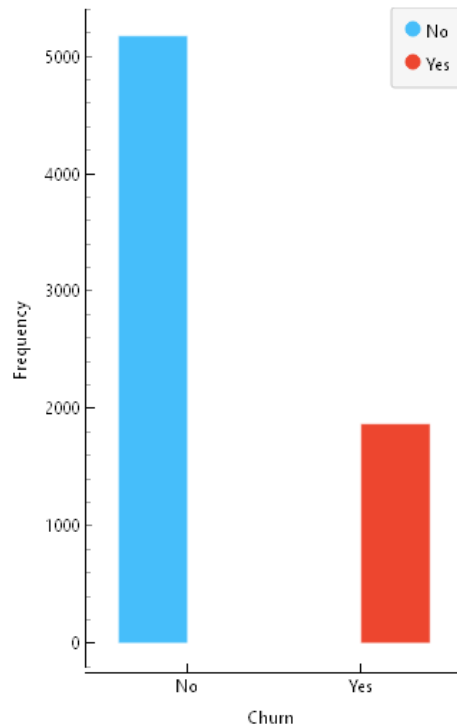
**D. HASIL DAN PEMBAHASAN**

**Eksplorasi Data (EDA)**

1. Analisis Distribusi Variabel Churn

Hasil analisis distribusi menggunakan widget Distributions menunjukkan bahwa variabel target Churn memiliki distribusi kelas yang tidak seimbang. Dari total 7.043 data pelanggan, sebanyak 5.174 pelanggan (73,5%) berada pada kategori tidak churn, sedangkan 1.869 pelanggan (26,5%) termasuk kategori churn. Ketidakeimbangan kelas ini merupakan kondisi yang umum ditemukan pada kasus customer churn dan

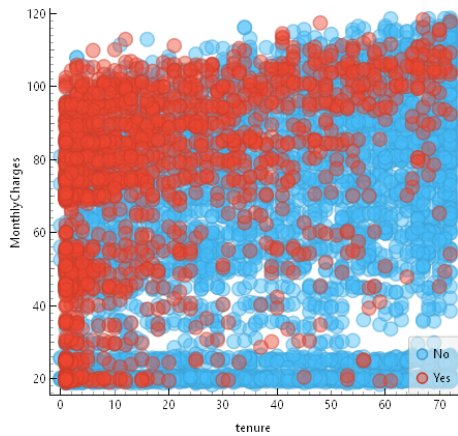
perlu diperhatikan dalam interpretasi performa model, khususnya pada metrik akurasi.



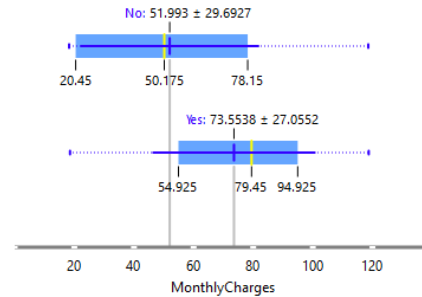
Gambar 1 Distribution Variable Churn

2. Analisis Hubungan tenure dan MonthlyCharges terhadap Churn

Hasil visualisasi Scatter Plot pada fitur tenure dan MonthlyCharges dengan pewarnaan berdasarkan status Churn menunjukkan adanya hubungan yang cukup jelas antara kedua variabel tersebut. Pelanggan yang termasuk kategori churn lebih banyak berada pada area dengan nilai tenure rendah (0–15 bulan) dan MonthlyCharges tinggi, yaitu di atas 60 dolar per bulan. Kondisi ini mengindikasikan bahwa pelanggan baru dengan biaya layanan yang relatif tinggi memiliki kecenderungan lebih besar untuk berhenti berlangganan. Di sisi lain, pelanggan dengan masa berlangganan yang lebih lama didominasi oleh kategori tidak churn, sehingga menunjukkan bahwa loyalitas pelanggan cenderung meningkat seiring lamanya penggunaan layanan. Selain itu, pelanggan dengan biaya bulanan yang lebih rendah, terutama pada kisaran 20–25 dolar per bulan, mayoritas berada pada kategori tidak churn.



Gambar 2 Scatter plot tenure dan MonthlyCharges

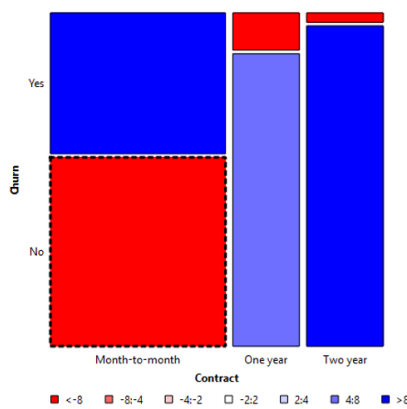


Student's t: 31.040 (p=0.000, N=7043)

Gambar 4 Box Plot fitur numerik terhadap churn

### 3. Analisis Kategorikal terhadap Churn

Hasil analisis menunjukkan bahwa pelanggan dengan jenis kontrak month-to-month memiliki tingkat churn yang lebih tinggi dibandingkan pelanggan dengan kontrak satu tahun maupun dua tahun. Selain itu, pelanggan yang menggunakan layanan internet Fiber Optic juga menunjukkan kecenderungan churn yang lebih besar dibandingkan pengguna layanan DSL atau pelanggan yang tidak menggunakan layanan internet.



Gambar 3 Mosaic Display variabel kategorikal churn

### 4. Analisis Kategorikal terhadap Churn

Hasil visualisasi menunjukkan adanya perbedaan yang cukup jelas pada fitur tenure, di mana pelanggan yang melakukan churn memiliki nilai median tenure yang lebih rendah dibandingkan pelanggan yang tidak churn. Selain itu, fitur MonthlyCharges juga menunjukkan bahwa pelanggan yang melakukan churn cenderung memiliki biaya bulanan yang lebih tinggi dibandingkan pelanggan yang tetap menggunakan layanan.

### Pemodelan Random Forest

Proses Pelatihan Model :

Proses pemodelan dilakukan dengan menggunakan algoritma Random Forest pada data yang sebelumnya telah melalui tahap preprocessing, meliputi penanganan missing value, transformasi data kategorikal, serta normalisasi data. Model dibangun dengan 100 decision tree untuk meningkatkan kestabilan dan akurasi hasil klasifikasi. Fitur Replicable Training diaktifkan agar proses pelatihan menghasilkan output yang konsisten pada setiap percobaan. Parameter max\_depth tidak dibatasi sehingga setiap pohon dapat menangkap pola data secara lebih mendalam, sementara min\_samples\_split ditetapkan bernilai 2 sebagai batas minimum pemisahan node. Model yang telah dilatih kemudian digunakan untuk melakukan prediksi status churn pelanggan, lalu dievaluasi menggunakan metode cross validation guna mengukur kinerja dan generalisasi model klasifikasi.

Berikut adalah hasil evaluasi model:

Tabel 3 Hasil Evaluasi Model

Metrik Evaluasi	Nilai
AUC	0,827
Accuracy (CA)	0,792 (79%)
F1-Score	0,557
Precision	0,640
Recall	0,493
MCC	0,430

### Evaluasi Model

Confusion Matrix

Confusion matrix memberikan gambaran yang lebih jelas tentang kesalahan prediksi model, yang terdiri dari True Positive (pelanggan churn yang diprediksi benar sebagai churn), True Negative (pelanggan tidak churn yang diprediksi benar sebagai tidak churn), False Positive (pelanggan tidak churn tetapi salah diprediksi churn), dan False Negative (pelanggan churn tetapi salah diprediksi tidak churn). Dalam kasus prediksi churn, False Negative menjadi kesalahan paling merugikan bagi bisnis karena pelanggan yang sebenarnya akan berhenti berlangganan tidak terdeteksi oleh model sehingga perusahaan kehilangan kesempatan untuk melakukan tindakan

pengecahan, oleh karena itu peningkatan nilai Recall menjadi hal penting yang perlu diperhatikan pada pengembangan model selanjutnya.

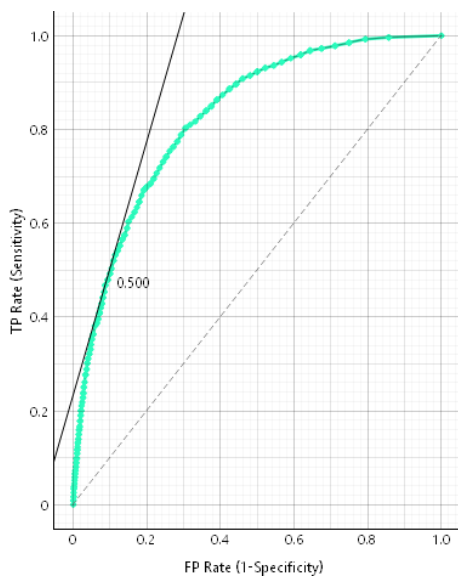
Tabel 4 Confusion Matrix Random Forest

	Predicted No	Predicted Yes
Actual No	83,1%	36,0%
Actual Yes	16,9%	64%

Berdasarkan hasil confusion matrix, model Random Forest menunjukkan kemampuan yang cukup baik dalam mengklasifikasikan pelanggan yang tidak churn, namun masih terdapat sejumlah pelanggan yang berpotensi churn tetapi belum berhasil terdeteksi oleh model. Kondisi ini konsisten dengan nilai Recall sebesar 0,493 dan Precision sebesar 0,640 yang diperoleh pada tahap evaluasi, yang menunjukkan bahwa model masih memerlukan peningkatan dalam mendeteksi kelas minoritas karena ketidakseimbangan distribusi kelas pada dataset.

### Kurva ROC

Analisis ROC (Receiver Operating Characteristic) menunjukkan bahwa model Random Forest mampu menghasilkan performa klasifikasi yang cukup baik. Hal ini dapat dilihat dari bentuk kurva ROC yang berada di atas garis diagonal dan mengarah ke sudut kiri atas grafik. Pola tersebut menandakan bahwa model memiliki kemampuan yang cukup baik dalam membedakan pelanggan yang melakukan churn dan pelanggan yang tidak melakukan churn.



Gambar 5 Kurva ROC

Nilai AUC sebesar 0,827 menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik dalam memprediksi customer churn. Hasil tersebut menunjukkan bahwa model dapat mengidentifikasi pelanggan yang berpotensi melakukan churn dengan tingkat akurasi yang

cukup tinggi. Selain itu, kurva ROC juga dapat dijadikan acuan dalam menentukan nilai threshold klasifikasi sesuai dengan kebutuhan dan strategi bisnis perusahaan.

### Feature Rank

Tabel 5 Feature Rank

Rank	Variable	Info.Gain	Gain Ratio	Gini
1	Contract	0.142	0.099	0.066
2	OnlineSecurity	0.093	0.062	0.047
3	TechSupport	0.091	0.061	0.046
4	InternetService	0.080	0.052	0.041
5	tanure	0.094	0.047	0.049
6	OnlineBackup	0.068	0.044	0.033
7	DeviceProtection	0.063	0.041	0.031
8	PaymentMethod	0.063	0.033	0.036
9	StreamingMovies	0.046	0.030	0.021
10	StreamingTV	0.046	0.030	0.021

Widget Feature Rank menghasilkan peringkat fitur berdasarkan tingkat pengaruhnya terhadap prediksi churn menggunakan metode pengukuran seperti Information Gain dan Gini. Hasil analisis menunjukkan bahwa fitur tenure menjadi fitur yang paling dominan, diikuti oleh MonthlyCharges dan TotalCharges. Temuan tersebut sejalan dengan hasil EDA yang menunjukkan adanya perbedaan distribusi yang cukup jelas antara pelanggan churn dan tidak churn pada ketiga fitur numerik tersebut. Selain itu, pada fitur kategorikal, Contract termasuk fitur dengan tingkat kepentingan yang tinggi, sehingga menunjukkan bahwa jenis kontrak memiliki pengaruh besar terhadap perilaku churn. Layanan internet Fiber Optic dan tidak adanya layanan keamanan online juga memberikan kontribusi yang cukup signifikan terhadap churn, sedangkan fitur seperti gender dan Dependents menunjukkan pengaruh yang relatif kecil terhadap hasil prediksi model.

### Interpretasi

Berdasarkan hasil keseluruhan analisis, dapat diketahui bahwa pelanggan yang memiliki risiko churn tinggi umumnya merupakan pelanggan dengan durasi berlangganan yang singkat (kurang dari 15 bulan), memiliki tagihan bulanan yang tinggi (di atas 60 dolar), menggunakan kontrak month-to-month, memakai layanan internet Fiber Optic, serta tidak menggunakan layanan tambahan seperti OnlineSecurity atau TechSupport. Selain itu, nilai Recall yang relatif rendah (49,3%) menunjukkan bahwa model masih belum mampu mengidentifikasi seluruh pelanggan yang berpotensi melakukan churn. Kondisi tersebut dipengaruhi oleh ketidakseimbangan kelas pada dataset, di mana jumlah pelanggan tidak churn lebih dominan dibandingkan pelanggan churn.

### Rekomendasi Bisnis

Berdasarkan hasil penelitian, terdapat beberapa langkah yang dapat dilakukan perusahaan untuk menurunkan tingkat customer churn. Perusahaan dapat meningkatkan program onboarding bagi pelanggan baru, terutama pada

awal masa berlangganan karena periode tersebut memiliki risiko churn yang cukup tinggi. Selain itu, perusahaan juga dapat memberikan promo atau penawaran khusus agar pelanggan dengan kontrak bulanan tertarik menggunakan kontrak jangka panjang. Perusahaan juga perlu meningkatkan kualitas layanan internet Fiber Optic guna meningkatkan kepuasan pelanggan. Di sisi lain, penyediaan layanan tambahan seperti OnlineSecurity dan TechSupport dengan harga yang lebih terjangkau dapat membantu meningkatkan loyalitas pelanggan dan mengurangi kemungkinan pelanggan berpindah ke layanan kompetitor.

## E. KESIMPULAN DAN SARAN

### Kesimpulan

Berdasarkan hasil penelitian, algoritma Random Forest mampu digunakan untuk memprediksi customer churn pada dataset Telco Customer Churn dengan performa yang cukup baik. Hasil evaluasi menggunakan metode 10-fold cross-validation memperoleh nilai AUC sebesar 0,827 dan akurasi sebesar 79,2%, yang menunjukkan bahwa model mampu membedakan pelanggan churn dan tidak churn dengan cukup optimal. Namun, nilai Recall sebesar 0,493 menunjukkan bahwa model masih belum mampu mendeteksi seluruh pelanggan yang berpotensi churn akibat ketidakseimbangan distribusi kelas pada dataset. Selain itu, fitur Contract, tenure, OnlineSecurity, dan TechSupport menjadi faktor yang paling berpengaruh terhadap perilaku customer churn. Hasil penelitian ini diharapkan dapat membantu perusahaan telekomunikasi dalam menentukan strategi retensi pelanggan yang lebih tepat untuk mengurangi tingkat customer churn.

### Saran

Berdasarkan hasil penelitian yang telah dilakukan, penelitian selanjutnya disarankan untuk menerapkan teknik penanganan data tidak seimbang seperti SMOTE atau oversampling guna meningkatkan kemampuan model dalam mendeteksi pelanggan yang berpotensi churn. Selain itu, penelitian berikutnya juga dapat membandingkan algoritma Random Forest dengan algoritma klasifikasi lainnya seperti Decision Tree, Naive Bayes, atau XGBoost untuk memperoleh performa model yang lebih optimal. Dari sisi implementasi, model prediksi yang telah dibangun

dapat dikembangkan menjadi sistem pendukung keputusan atau diintegrasikan dengan sistem CRM perusahaan agar proses identifikasi pelanggan yang berpotensi churn dapat dilakukan secara lebih cepat dan efektif.

## F. DAFTAR PUSTAKA

- Azmi, A. F., & Voutama, A. (2024). KOMPUTA : Jurnal Ilmiah Komputer dan Informatika PREDIKSI CHURN NASABAH BANK MENGGUNAKAN KLASIFIKASI RANDOM FOREST DAN DECISION TREE DENGAN EVALUASI CONFUSION MATRIX KOMPUTA : Jurnal Ilmiah Komputer dan Informatika. 13(1).
- Bili, I. K. M., Sudiarta, I. W., Yuditia, M., Rosdiana, N. K. A., & Rafiana, P. (2026). Analisis dan Prediksi Customer Churn pada Platform Streaming Berbasis Langganan Menggunakan Metode Random Forest membicarakan suatu layanan setelah jangka waktu tertentu . Dalam dunia bisnis , churn layanan . Data menunjukkan durasi akses , frekuensi peng. November 2025.
- Kencana, A. K., Ananda, F. D., & Hartanto, A. D. (2022). Implementasi Metode Random Forest Klasifikasi untuk Phishing Link. 4(2), 55–59.
- Mutiara, A., Marutho, D., & Yuniarti, N. A. (2025). HYBRID DEEP LEARNING RANDOM FOREST OPTIMASI PEMILIHAN FITUR UNTUK PREDIKSI CHURN INDUSTRI. 6(2), 694–707. <https://doi.org/10.46576/djtechno>
- Nugroho, M. W. (2025). Jurnal JTIC ( Jurnal Teknologi Informasi dan Komunikasi ) Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi. 9(December), 1562–1571.
- Syawaludin, M. A., & Hidayat, R. (2024). Prediksi Churn Pelanggan Multinational Bank Menggunakan Algoritma Machine Learning. 4(2), 89–97.
- Yoraeni, A., & Rakhmah, S. N. (2025). Penerapan Algoritma Naive Bayes untuk Prediksi Kerontokan Rambut Naive Bayes Algorithm Implementation for Hair Loss Prediction. 7(1), 63–70. <https://doi.org/10.30812/bite.v7i1.5201>