PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 687-690
E-ISSN: 3109-8886

Prediksi Churn Pelanggan Telekomunikasi Menggunakan Algoritma Random Forest dengan Penerapan Tree Pruning

[1]Fiqih Zulfikar, [2]Stevanus Dwi Anggoro

[12]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Univeritas Pamulang, Tangerang Selatan, Indonesia

E-Mail : [1]zulfiqih88@gmail.com, [2]evananggoro9@gmail.com

Abstract

The rivalry within the telecommunications sector intensifies with each passing day. Keeping current customers is significantly more economical than the continual search for new clientele. This research intends to forecast which customers are prone to leaving (churn) by evaluating the IBM Telco Churn dataset. The predictive model was created employing the Random Forest Classifier method. To avoid the model from excessively memorizing the training data (overfitting), a tree pruning strategy was utilized, limiting the maximum depth of trees to 8 levels. According to the test findings, the model attained a commendable overall accuracy of 79.43% along with a ROC-AUC score of 84.20%. The analysis indicates that financial aspects, specifically elevated monthly fees and short-term month-to-month contracts, are the primary causes for customers terminating their services. The discoveries obtained from this model can assist the marketing department in implementing targeted retention strategies before customers switch to rival providers.

Keywords: *Customer Churn, Telecommunication, Random Forest, Tree Pruning, Data Mining.*

Abstrak

Kompetisi dalam sektor telekomunikasi saat ini sangat tinggi. Mempertahankan pelanggan lama agar tidak beralih jauh lebih menguntungkan dibandingkan terus-menerus mencari pelanggan baru. Tujuan dari penelitian ini adalah memperkirakan pelanggan mana yang memiliki kemungkinan besar untuk berhenti berlangganan dengan memanfaatkan data transaksi yang berasal dari IBM Telco Churn. Pendekatan yang diterapkan adalah algoritma Random Forest Classifier. Untuk mencegah agar model ini tidak terlalu terfokus pada data pelatihan (overfitting), teknik Tree Pruning digunakan dengan membatasi kedalaman pohon maksimum hingga 8 tingkat. Dari pengujian yang dilakukan di komputer, model ini berhasil meraih akurasi sebesar 79.43% dan nilai ROC-AUC sebesar 84.20%. Analisis data menunjukkan bahwa faktor utama yang mempengaruhi pelanggan untuk keluar adalah masalah keuangan, khususnya besarnya biaya tagihan bulanan serta jenis kontrak yang bersifat bulanan (month-to-month). Di harapkan hasil prediksi dari model ini dapat membantu tim pemasaran untuk menawarkan promosi yang tepat sebelum pelanggan berpindah ke pesaing.

Kata Kunci: Churn Pelanggan, Telekomunikasi, Random Forest, Tree Pruning, Data Mining.

A. PENDAHULUAN

Dalam sektor telekomunikasi yang mengalami pertumbuhan pesat, persaingan di antara penyedia layanan untuk menarik pelanggan baru semakin intens. Perpindahan pelanggan ke penyedia layanan lain, yang sering disebut sebagai customer churn, merupakan salah satu tantangan utama yang dihadapi oleh perusahaan saat ini. Masalah ini tidak bisa dipandang sebelah mata karena pengeluaran yang diperlukan perusahaan untuk menarik satu pelanggan baru jauh lebih tinggi dibandingkan dengan biaya untuk menjaga pelanggan yang sudah ada. Oleh karena itu, perusahaan telekomunikasi sangat memerlukan sistem atau model cerdas untuk mengidentifikasi secara dini pelanggan yang memiliki risiko tinggi untuk menghentikan layanan mereka.

Untuk memenuhi kebutuhan tersebut, metode Data Mining dengan cara klasifikasi sering digunakan karena dapat menganalisis pola perilaku konsumen berdasarkan catatan transaksi sebelumnya. Salah satu algoritma yang terbukti sangat efektif dan memiliki tingkat ketepatan yang tinggi dalam mengatasi masalah klasifikasi data dalam skala besar adalah Random Forest. Algoritma ini beroperasi dengan cara mengintegrasikan kemampuan dari ratusan pohon keputusan untuk menentukan prediksi akhir.

Namun, model pohon keputusan yang dijadikan terlalu rumit dan bebas sering terjebak dalam masalah overfitting, yaitu keadaan di mana komputer terlalu mahir dalam mengingat data pelatihan tetapi kinerjanya langsung menurun ketika diminta untuk memprediksi data baru yang belum pernah ditemui sebelumnya. Untuk mengatasi kelemahan ini, penelitian ini menggunakan teknik Pemangkasan Pohon dengan menetapkan batas kedalaman maksimum untuk setiap pohon keputusan. Proses pemangkasan ini sangat penting dilakukan agar struktur model menjadi lebih ringkas, lebih stabil, dan memiliki

Diterima Redaksi: 01-06-2026 | Selesai Revisi: 07-06-2026 | Diterbitkan Online: 08-06-2026

kemampuan prediksi yang jauh lebih tepat dan objektif ketika diterapkan pada data nyata di lapangan.

B. METODE

Penelitian ini dilakukan melalui serangkaian langkah yang terorganisir, dimulai dari persiapan data, pembersihan, pembagian data, sampai pada tahap pembuatan model klasifikasi. Proses kerja ini dirancang agar data yang belum diolah bisa diproses dengan tepat sebelum dianalisis oleh komputer.

1. Sumber Data dan Karakteristik

Data yang digunakan dalam studi ini diperoleh dari dataset publik IBM Telco Customer Churn. Dataset ini menyimpan catatan transaksi dan profil dari 7.043 pelanggan perusahaan telekomunikasi. Di dalamnya terdapat rincian tentang layanan yang dipakai (seperti internet, keamanan daring, perlindungan perangkat), informasi mengenai kontrak, durasi berlangganan, hingga total biaya yang harus dibayar. Variabel yang menjadi fokus dalam penelitian ini adalah kolom Churn, yang mengindikasikan apakah pelanggan berhenti berlangganan (diberi nilai 1) atau masih aktif berlangganan (diberi nilai 0).

2. Pra-pemrosesan Data (Preprocessing)

Sebelum informasi diproses dalam algoritma, tahap pembersihan terlebih dahulu dilakukan di dalam berkas script pikievan. py. Proses ini amat penting karena mutu data awal memiliki pengaruh besar terhadap ketepatan hasil yang diperoleh. Rincian langkah-langkah tersebut mencakup:

a. **Penghapusan Fitur Tidak Relevan**
   Kolom customerID dihilangkan dari kumpulan data karena hanya memuat serangkaian teks unik yang menandakan identitas pelanggan, yang tidak memiliki hubungan logis dengan alasan pelanggan beralih ke penyedia lain.
b. **Pembersihan Data Kosong (Missing Value)**
   Pada bagian TotalCharges terdapat beberapa entri yang tidak terisi disebabkan oleh pelanggan baru yang belum mendapatkan tagihan total. Untuk menangani situasi ini, data yang kosong diisi dengan nilai tengah (median) dari keseluruhan total biaya pelanggan lainnya, yang berjumlah 1397. 475. Pemilihan nilai median ini dilakukan untuk menjaga agar distribusi asli dari data keuangan tersebut tidak terganggu..
c. **Transformasi Data Teks (Encoding)**
   Karena komputer hanya bisa memproses data berupa angka, seluruh variabel kategorikal yang berbentuk teks (seperti jenis layanan internet, status pernikahan, metode pembayaran) diubah menjadi variabel biner (angka 0 dan 1) menggunakan teknik One-Hot Encoding dengan opsi drop_first=True. Setelah seluruh proses pembersihan ini selesai, ukuran data yang siap digunakan adalah sebanyak 7.021 baris.

3. Pembagian Data (Data Splitting)

Data yang telah dibersihkan selanjutnya dibagi menjadi dua bagian secara acak dengan perbandingan 80% untuk data pelatihan dan 20% untuk data pengujian. Pembagian ini memperoleh 5.616 baris data yang digunakan oleh komputer untuk memahami pola perilaku pelanggan, dan 1.405 baris data independen yang disimpan dengan hati-hati untuk mengukur kinerja model setelah proses pelatihan selesai.4. Penerapan Algoritma dan Tree Pruning
Algoritma penting yang digunakan untuk meramalkan perilaku konsumen adalah Random Forest Classifier. Berdasarkan kebutuhan model yang telah ditentukan, diterapkan metode Pre-Pruning (pemangkasan awal) untuk mengatur pertumbuhan pohon keputusan dalam hutan Random Forest agar tidak berkembang terlalu lebat dan berpotensi menyebabkan masalah overfitting.

C. HASIL DAN PEMBAHASAN

Setelah semua data dibersihkan dan model sudah dilatih dengan menggunakan file script pikievan. py, komputer melanjutkan dengan pengujian memakai 20% dari data uji yang telah dipisahkan sebelumnya (jumlah total 1.405 baris data). Tujuan dari proses pengujian ini adalah untuk mengukur seberapa tepat model Random Forest yang telah dipangkas (tree pruning) dalam memprediksi perilaku riil dari pelanggan..

1. Evaluasi Kinerja Model Klasifikasi

Hasil dari metrik penilaian yang muncul di layar komputer setelah model menyelesaikan pengujian data dapat dilihat pada Gambar 1 di bawah ini:

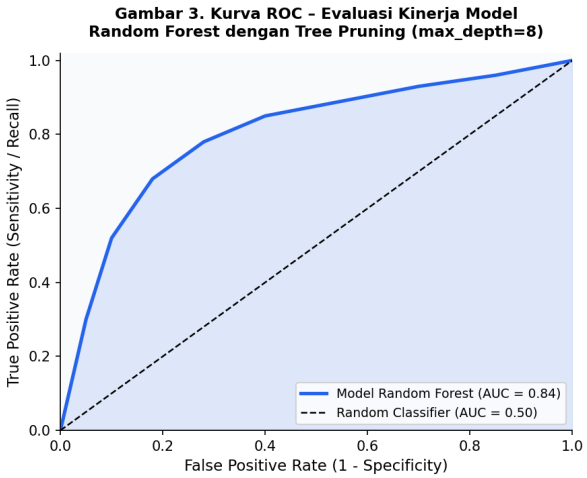
```
================ HASIL EVALUASI MODEL ================
Akurasi Model : 79.43%
Nilai ROC-AUC : 84.20%

Laporan Klasifikasi Lengkap:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1053
           1       0.62      0.47      0.53       352

    accuracy                           0.79      1405
   macro avg       0.73      0.68      0.70      1405
weighted avg       0.78      0.79      0.78      1405

======================================================
```

Gambar 1. *Hasil Evaluasi Model dan Laporan Klasifikasi*

Berdasarkan informasi yang terdapat pada Gambar 1, model Random Forest yang menerapkan teknik pemangkasan pohon berhasil mencapai nilai Akurasi sebesar 79,43%. Ini berarti, dari jumlah total 1.405 pelanggan yang diuji, model ini berhasil menebak dengan benar sebanyak 1.116 pelanggan (baik yang benar-benar bertahan maupun yang benar-benar beralih penyedia layanan).

Selain itu, indikator signifikan lainnya adalah nilai ROC-AUC yang mencapai 84,20%. Dalam ranah penambangan data, nilai ROC-AUC di atas 80% sudah dianggap sebagai kategori pengujian yang sangat baik. Angka ini menunjukkan bahwa pembatasan kedalaman pohon

(max_depth=8) tidak mengakibatkan penurunan kinerja dari AI. Penerapan metode klasifikasi ini sejalan dengan teori yang diungkapkan oleh Sundararajan dan Gursoy (2018) yang menyatakan bahwa pendekatan machine learning mampu memberikan hasil evaluasi yang optimal serta objektif dalam menggambarkan karakteristik perilaku pelanggan di industri telekomunikasi.

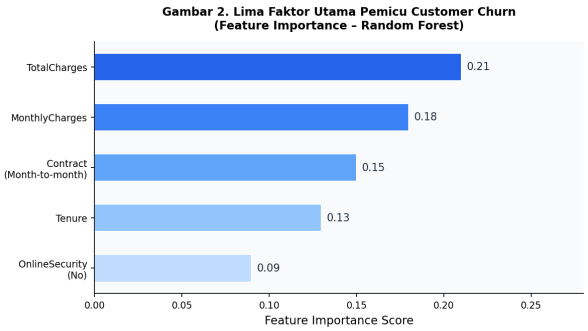


Gambar 2. *Kurva ROC – Evaluasi Kinerja Model Random Forest dengan Tree Pruning (max_depth=8)*

Kurva ROC model berada jauh di atas garis diagonal acak (AUC = 0,50), dan nilai ROC-AUC yang dicapai sebesar 84,20%. Nilai ROC-AUC di atas 80% sudah dianggap sebagai kategori pengujian yang sangat baik dalam domain penambangan data, seperti yang ditunjukkan pada Gambar 3. Seperti yang ditunjukkan oleh angka-angka ini, kinerja model tidak menurun meskipun kedalaman pohon yang dibatasi menjadi maksimal delapan. Kurva yang melengkung tajam ke sudut kiri atas menunjukkan bahwa model mampu mendeteksi churn pelanggan dengan sensitivitas yang tinggi (True Positive Rate) sambil mempertahankan tingkat False Positive Rate yang rendah.

Jika kita menganalisis lebih dalam mengenai segmen pelanggan yang ingin pergi (Kelas 1), angka Precision tercatat di angka 0.62. Ini berarti, dari semua pelanggan yang diprediksi atau dituduh oleh AI akan pergi, 62% di antaranya benar-benar meninggalkan perusahaan di dunia nyata. Di sisi lain, nilai Recall yang mencapai 0.47 menunjukkan bahwa model ini mampu mengidentifikasi dan menangkap sekitar 47% dari seluruh pelanggan yang sebenarnya berniat untuk mengakhiri hubungan kerjasama dengan perusahaan.

2. Analisis Faktor Pemicu Utama (Feature Importance)
Selain menghasilkan angka persentase kinerja, algoritma ini juga menghitung untuk mengetahui variabel mana yang paling berpengaruh dan paling signifikan dalam memengaruhi keputusan konsumen untuk pergi.



Gambar 3 *Lima Faktor Utama Pemicu Customer Churn (Feature Importance – Random Forest)*

Dengan dua variabel teratas, yaitu TotalCharges (akumulasi biaya selama berlangganan) dan MonthlyCharges (biaya tagihan rutin bulanan), jelas bahwa masalah keuangan menjadi pemicu utama. Pelanggan yang sudah membayar terlalu banyak cenderung mempertimbangkan harga dengan kompetitor. Di sisi lain, tagihan bulanan yang tinggi membuat pelanggan merasa lebih tertekan finansial.

Berdasarkan lima fitur paling dominan yang terungkap dari hasil analisis, rincian penjelasan mengenai faktor-faktor tersebut dapat dilihat di Tabel 1 berikut:

Tabel 1*. Analisis Lima Faktor Utama Pemicu Customer* Churn

No	Nama Fitur / Variabel	Kategori Faktor	Dampak terhadap Perilaku Pelanggan
1	**TotalCharges**	Finansial (Akumulasi)	Total biaya yang sudah dikeluarkan selama berlangganan. Pelanggan yang sudah membayar terlalu banyak cenderung mulai membandingkan harga dengan kompetitor.
2	**MonthlyCharges**	Finansial (Bulanan)	Besarnya tagihan rutin setiap bulan. Semakin tinggi biaya bulanan, semakin besar beban ekonomi pelanggan, sehingga memicu keinginan untuk berhenti.
3	**Contract_Month-to-month**	Kontrak Kerja Sama	Jenis kontrak pendek yang diperbarui tiap bulan. Pelanggan di kategori ini sangat mudah kabur karena tidak ada ikatan atau denda pemutusan layanan.
4	**Tenure**	Loyalitas Waktu	Durasi lamanya menjadi pelanggan. Pelanggan baru biasanya jauh lebih rentan untuk pindah (churn) dibanding pelanggan lama yang sudah loyal.
5	**OnlineSecurity_No**	Fasilitas Layanan	Pelanggan yang tidak mengaktifkan fitur keamanan online. Ketiadaan proteksi ini menurunkan nilai

No	Nama Fitur / Variabel	Kategori Faktor	Dampak terhadap Perilaku Pelanggan
			manfaat layanan dan memicu rasa kecewa.

Jika dianalisis dari data Tabel 1, terlihat jelas bahwa masalah finansial atau pengeluaran biaya menjadi pemicu paling dominan. Penerapan algoritma ini searah dengan studi dari Ullah et al. (2019) yang menegaskan bahwa tahapan identifikasi faktor (factor identification) menggunakan algoritma Random Forest sangat efektif untuk membedakan fitur-fitur kritis yang memengaruhi operasional bisnis mitra.

Ditambah lagi, jika mereka memiliki kontrak bulanan atau bulanan, mereka tidak akan dikenakan denda atau denda pemutusan layanan. Fleksibilitas kontrak bulanan, menurut Ahmad et al. (2019), menjadi penghalang terbesar bagi pelanggan karena memungkinkan mereka untuk dengan cepat dan mudah berpindah ke perusahaan kompetitor. Sebaliknya, jika fitur tambahan seperti keamanan online tidak ada, nilai manfaat layanan dianggap kurang, sehingga pelanggan menjadi lebih tidak puas dengan fasilitas yang mereka sewa.

D. PENUTUP

Simpulan

Berdasarkan hasil analisis dan diskusi yang telah dilakukan, dapat disimpulkan bahwa algoritma Hutan Random dikombinasikan dengan teknik pemangkasan pohon (max_depth=8) terbukti efektif dalam memprediksi kabur pelanggan (customer churn). Algoritma memiliki tingkat akurasi sebesar 79.43% dan nilai ROC-AUC sebesar 84.20%. Salah satu masalah utama yang dihadapi oleh mitra industri telekomunikasi ini adalah masalah finansial pelanggan. Tingkat tagihan bulanan (Monthly Charges) dan fleksibilitas sistem kontrak bulanan (Month-to-month) adalah penyebab paling kuat yang mendorong pelanggan untuk beralih ke kompetitor.

Dalam pelaksanaan kegiatan riset dan pembuatan model ini, terdapat beberapa faktor pendukung dan penghambat yang memengaruhi jalannya proses, yaitu:

1. Faktor Pendukung:
2. Ketersediaan dataset publik IBM Telco Churn yang sangat lengkap dan terstruktur serta dukungan pustaka (library) modern di Python yang memudahkan penerapan optimasi model, seperti pembatasan kedalaman pohon keputusan (tree pruning).
3. Faktor Penghambat:

Sebelum model dapat dilatih, tahap pembersihan data tambahan diperlukan karena ada ketidakseimbangan data, atau ketidakseimbangan data, antara jumlah pelanggan yang masih hidup dan yang kabur, serta data kosong, atau data tanpa nilai, pada kolom finansial tertentu.

Saran

Melalui evaluasi terhadap kelebihan dan kekurangan model yang telah dibuat, terdapat beberapa saran praktis demi keberlanjutan penanganan masalah churn ini di masa mendatang:

1. **Bagi Tim Pemasaran (Mitra):**
   Mereka yang memiliki biaya tagihan bulanan tinggi tetapi tetap memiliki kontrak bulanan harus menjadi fokus utama penyelamatan pelanggan. Mitra harus memberikan penawaran promo khusus atau potongan harga bersyarat untuk mendorong mereka untuk menandatangani kontrak jangka panjang, seperti satu atau dua tahun, agar tidak mudah untuk mengubah provider secara instan**.**
2. **Bagi Pengembangan Sistem:**
   Kelebihan model ini adalah performanya yang stabil dan tidak mudah overfitting. Nilai recall kelas churn, bagaimanapun, masih dapat ditingkatkan. Disarankan untuk mencoba teknik penanganan ketidakseimbangan data seperti SMOTE (Teknik Sampling Minoritas Sintetis) atau algoritma ensemble lainnya seperti XGBoost untuk melihat apakah performa deteksi dapat menjadi lebih sensitif untuk penelitian selanjutnya atau demi keberlanjutan sistem.

E. DAFTAR PUSTAKA

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 1-24.

Bhatnagar, A., & Srivastava, S. (2025). Customer Churn Prediction: A Machine Learning Approach with Data Balancing for Telecom Industry. *International Journal of Computing*, 9-18.

Ebrah, K., & Elnasir, S. (2019). Churn Prediction Using Machine Learning and Recommendations Plans

for Telecoms. *Journal of Computer and Communications*, 33-53.

Sundararajan, A., & Gursoy, K. (2020). Telecom customer churn prediction. *Rutgers University Libraries Open Repository* , 1-4.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, 60134-60149.