Perbandingan Performa Algoritma Random Forest Dan Gradient Boosting Dalam Mengklasifikasi Churn Telco

Comparison Of Performance Of Random Forest And Gradient Boosting Algorithm In Classifying Telco Churn

Muhammad Adji Purnama[1], Jilang Ramadhani[2], Yoga Safitra Anugraha[3]
Lusiana Efrizoni [4], Rahmaddeni[5*]
Universitas Sains Dan Teknologi Indonesia[1,2,3,4,5]
E-mail: [1]2110031802125@sar.ac.id, [2]jilangramadhan29@gmail.com,
[3]safitra.yoga101002@gmail.com, [4]lusiana@stmik-amik-riau.ac.id , [5*]rahmaddeni@sar.ac.id
Corresponding author

Abstrak

Customer churn adalah kecenderungan pelanggan berhenti dan berpindah layanan dalam periode tertentu. Ini merupakan masalah utama dalam industri telekomunikasi karena mempengaruhi keuntungan perusahaan. Mempertahankan pelanggan lebih mudah dibandingkan mendapatkan pelanggan baru. Memprediksi churn membantu sektor CRM dalam merancang strategi retensi. Tingkat churn yang tinggi dapat menurunkan pendapatan dan mengganggu stabilitas bisnis. Berdasarkan studi, tingkat churn tahunan di industri telekomunikasi berkisar antara 15% hingga 30%. Data mining, yang memanfaatkan teknik pembelajaran mesin, digunakan untuk menganalisis dan mengekstraksi pengetahuan dari data. Penelitian ini bertujuan untuk membandingkan performa dua algoritma yaitu Random Forest dan Gradient Boosting. Hasil yang didapatkan menggunakan splitting data 80:20 menunjukkan bahwa klasifikasi lebih unggul menggunakan metode Gradient Boosting dibandingkan metode Random Forest dilihat dari tingkat akurasi dan nilai ROC AUC. Metode Gradient Boosting mendapatkan nilai akurasi dan ROC AUC sebesar 83% dan 0.89, Sedangkan metode Random Forest mampu menghasilkan nilai akurasi dan ROC AUC sebesar 81% dan 0.87.

Kata kunci: Churn, Gradient Boosting, Klasifikasi, Random Forest, Telco

Abstract

Customer churn is the tendency of customers to stop and switch services within a certain period. This is a major problem in the telecommunications industry because it affects company profits. Retaining customers is easier than getting new customers. Predicting churn helps the CRM sector in designing retention strategies. A high churn rate can reduce revenue and disrupt business stability. According to studies, the annual churn rate in the telecommunications industry ranges from 15% to 30%. Data mining, which utilizes machine learning techniques, is used to analyze and extract knowledge from data. This research aims to compare the performance of two algorithms, namely Random Forest and Gradient Boosting. The results obtained using 80:20 data splitting show that classification is superior using the Gradient Boosting method compared to the Random Forest method in terms of the level of accuracy and ROC AUC value. The Gradient Boosting method obtained accuracy and ROC AUC values of 83% and 0.89, while the Random Forest method was able to produce accuracy and ROC AUC values of 81% and 0.87.

Keywords: Churn, Gradient Boosting, Classification, Random Forest, Telco

1. PENDAHULUAN

Customer Churn mengacu pada kecenderungan pelanggan untuk beralih dari satu layanan ke layanan lain dalam jangka waktu tertentu. Ini merupakan masalah penting dalam dunia bisnis, terutama dalam industri telekomunikasi, dan berdampak signifikan pada keuntungan

perusahaan karena mendapatkan pelanggan baru jauh lebih mahal daripada mempertahankan pelanggan lama. Sektor manajemen hubungan pelanggan (CRM) perusahaan dapat membantu menentukan strategi untuk mempertahankan pelanggan yang ada dengan memprediksi penurunan pelanggan. Tingkat churn yang tinggi dapat menyebabkan penurunan pendapatan dan mengganggu stabilitas bisnis. Berdasarkan berbagai studi, industri telekomunikasi mengalami tingkat churn tahunan antara 15% hingga 30%, yang berarti perusahaan telekomunikasi kehilangan sebagian besar basis pelanggan mereka setiap tahun. Dampak dari churn ini sangat signifikan, karena dapat mengakibatkan kerugian finansial yang besar serta merusak reputasi perusahaan[1].

Prediksi churn sangat penting bagi perusahaan telekomunikasi karena memungkinkan mereka untuk mengambil langkah proaktif dalam mempertahankan pelanggan. Dengan memprediksi pelanggan yang berpotensi churn, perusahaan dapat merancang strategi retensi yang efektif, seperti menawarkan diskon khusus atau peningkatan layanan. Dampak penting dari telekomunikasi adalah munculnya berbagai bisnis yang melawan tren negatif dalam industri ini. Untuk mendapatkan keunggulan kompetitif, banyak bisnis berinovasi dengan menarik pelanggan baru untuk menggunakan layanan mereka. Persaingan ini mengharuskan beberapa organisasi untuk menangani kerugian akibat kehilangan bisnis (churn) dan merealisasikan keuntungan dari perolehan klien baru. Manfaat dari memprediksi churn meliputi pengurangan biaya pemasaran, karena upaya retensi lebih murah dibandingkan dengan akuisisi pelanggan baru, serta peningkatan loyalitas pelanggan, yang pada gilirannya dapat meningkatkan pendapatan dan keuntungan jangka panjang[2].

Data mining merupakan sebuah proses yang memanfaatkan satu atau beberapa teknik pembelajaran mesin untuk secara otomatis menganalisis dan mengekstraksi pengetahuan. Tujuan utama dari data mining adalah untuk menemukan pola, korelasi, dan wawasan yang dapat membantu dalam pengambilan keputusan bisnis. Dalam berbagai industri, termasuk telekomunikasi, data mining digunakan untuk berbagai aplikasi seperti analisis pelanggan, deteksi penipuan, prediksi tren, dan, yang terpenting, prediksi churn pelanggan. Dengan teknik data mining, perusahaan dapat memahami perilaku pelanggan mereka dengan lebih baik dan membuat keputusan yang lebih tepat berdasarkan data yang tersedia[3].

Berbagai algoritma digunakan dalam data mining untuk prediksi churn. Beberapa algoritma yang populer termasuk decision trees, logistic regression, support vector machines, neural networks, dan ensemble methods seperti Random Forest dan Gradient Boosting.

Random Forest merupakan pengembangan dari metode decision tree, yaitu dengan menerapkan metode bootstrap aggregating dan random feature selection untuk mengumpulkan banyak-nya tree[4]. Dalam metode random forest,sejumlah besar pohon ditumbuhkan untuk membentuk sebuah hutan (forest) dan setelah itu analisis dilakukan pada sekumpulan pohon tersebut[5].

Gradient Boosting merupakan pengembangan lebih lanjut dari algoritma Gradient Tree Boosting berbasis ensemble, yang dapat secara efektif menangani masalah machine learning dengan skala yang lebih besar. teknik ensemble yang membangun model secara bertahap dengan menambahkan model baru yang berfokus pada kesalahan model sebelumnya. Setiap model baru memperbaiki kesalahan yang dibuat oleh model sebelumnya, menghasilkan model akhir yang sangat akurat[6].

Adapun penelitian terdahulu yang dilakukan oleh [7] dengan judul "Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi" algoritma yang digunakan menunjukkan bahwa optimasi algoritma Random Forest menggunakan boosting AdaBoost menghasilkan kinerja yang paling optimal dengan hasil akurasi (0,9913), presisi (0,9831), recall (1,0) dan F1-Score (0,9915) menggunakan data sebanyak 3333 dan 11 variabel. Sedangkan pada penelitian yang di lakukan oleh [8] dengan judul "Prediksi Churn Konsumen Menggunakan Algoritma Random Forest dengan Fuzzy C-Means untuk Meningkatkan Produktivitas Penjualan" Algoritma yang digunakan menunjukkan Hasil akhir pada pemodelan fuzzy c-means berhasil dioptimalkan dengan metode elbow, dan juga validasi cluster sehingga sedikit terjadinya data yang tumpang

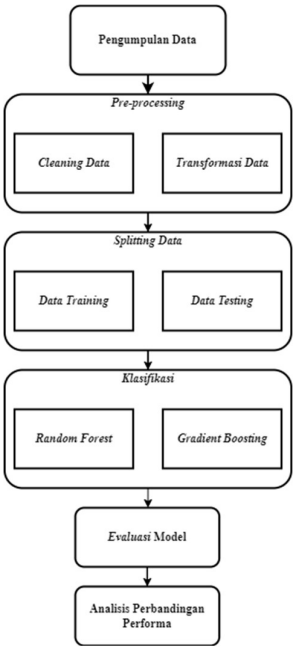
tindih antar cluster, dengan nilai silhouette score sebesar 53% dan random forest dengan nilai akurasi cross-validation sebesar 99.8%.

Berdasarkan penelitian sebelumnya, perbedaan penelitian ini dengan penelitiann sebelumnya penelitian ini melakukan kontribusi signifikan dalam bidang churn pelanggan Telco dengan beberapa kebaruan dengan membandingkan algoritma Random Forest dan Gradient Boosting dalam klasifikasi churn pelanggan Telco. Pemilihan kedua algoritma ini didasarkan pada keunggulan masing-masing dalam menangani data yang kompleks dan memberikan hasil prediksi yang akurat dan dapat diandalkan. Random Forest dipilih karena kestabilannya dan kemampuan memberikan wawasan tentang kepentingan fitur, sementara Gradient Boosting dipilih karena akurasinya yang tinggi dan kemampuannya meningkatkan performa model secara bertahap.

Kebaruan dari penelitian ini tidak hanya membandingkan dari hasil metrik evaluasi seperti akurasi dan ROC AUC saja, namun juga akan menggunakan jumlah data yang lebih besar dan membandingkan Training Time dan Inference Time. Training Time digunakan untuk mengukur durasi yang diperlukan untuk melatih model dengan dataset tertentu, yang penting untuk efisiensi pengembangan model dan pengurangan biaya komputasi. Inference Time digunakan untuk mengukur waktu yang dibutuhkan model untuk menghasilkan prediksi pada data baru, yang sangat penting dalam aplikasi real-time dan efisiensi operasional. Dengan kebaruan yang ada, penelitian ini di harapkan dapat membantu perusahaan telekomunikasi untuk mengoptimalisasi sumber daya, pengurangan biaya, dan pengambilan keputusan lebih cepat.

2. METODE PENELITIAN

Diagram Tahapan Penelitian disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan adalah dataset churn pelanggan Telco California, dataset yang digunakan bersifat publik dan diperoleh dari Kaggle, yang dapat di akses melalui link berikut:

https://www.kaggle.com/datasets/alfathterry/telco-customer-churn-11-1-3 . Dataset berisi 7.043 baris dan 50 variabel.

2.2 Pre-processing

Data preprocessing adalah proses mengubah data mentah menjadi format yang lebih mudah dipahami. Proses ini digunakan untuk memperbaiki kesalahan pada data mentah yang seringkali tidak lengkap dan dalam format yang tidak teratur.

Preprocessing terdiri dari proses validasi dan imputasi data, yang dilakukan secara manual atau otomatis melalui program pengotomatisan proses bisnis (BPA), dengan tujuan menentukan kelengkapan dan akurasi data yang tersaring[9].

2.2.1 *Cleaning* data

Cleaning Data (pembersihan data), adalah proses mengidentifikasi dan mengoreksi atau membuang data yang salah dari kumpulan data. Data yang dibersihkan dapat berupa data yang tidak konsisten, tidak akurat, duplikat, salah format, atau kesalahan lain yang dapat mengganggu proses analisis data selanjutnya. Sebelum memasuki tahap analisis data atau pemodelan machine learning, proses pembersihan data adalah langkah penting dalam persiapan data[10].

2.2.2 *Transformasi* data

Transformasi data merupakan proses mengubah data dari skala pengukuran data asli menjadi skala bentuk lain dengan tujuan untuk memperbaiki karakteristik data dan memudahkan analisis.

Transformasi data dapat dilakukan dengan berbagai cara, seperti mengubah satuan ukuran data, mengubah distribusi data, atau mengubah bentuk data[11].

2.3 Splitting data

Splitting data merupakan proses pembagian data yang digunakan dalam penelitian Penelitian biasanya membagi set data menjadi dua atau lebih bagian untuk menguji model atau algoritma. Data latih digunakan untuk melatih algoritma, sementara data uji digunakan untuk menguji kinerja algoritma[12].

Maka dari itu pada penelitian ini menggunakan *splitting* data 80:20 untuk ke dua algoritma yang di gunakan yaitu *random forest* dan *gradient boosting.*

2.4 Klasifikasi

Data mining dapat mengklasifikasikan. Klasifikasi pertama kali digunakan untuk mengklasifikasikan bidang tanaman berdasarkan spesies tertentu. Ini terjadi pada Carolus von Linne (juga dikenal sebagai Carolus Linnaeus), yang pertama kali mengklasifikasikan spesies berdasarkan karakteristik fisiknya. Dia juga disebut sebagai bapak klasifikas[13].

2.5 Random Forest

Metode Forest Random (RF) meningkatkan nilai akurasi. Tujuan dari teknik ini adalah untuk membangun pohon keputusan yang terdiri dari root node, internal node, dan leaf node dengan menggunakan data dan atribut secara acak. Root node adalah simpul di atas (akar pohon keputusan). Internal node adalah simpul percabangan yang memiliki satu input dan minimal dua output. Leaf node atau terminal node adalah simpul terakhir yang hanya memiliki satu input dan tidak memiliki output. Persamaan persamaan 1 digunakan untuk menghitung nilai entropy dan nilai gain informasi[14].

$$Entropy\ (y) = -\sum_i p(c|Y) log_2 p(c|Y) \quad (1)$$

Y merupakan himpunan kasus dan *p(c|Y)* adalah proporsi nilai *Y* pada kelas *c*.

Information Gain*(Y, a)*

$$= Entropy(Y) - \sum_{v \epsilon Values(a)} \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v) \quad (2)$$

Nilai(a) adalah semua nilai yang mungkin pada himpunan kasus a; Yv adalah subkelas dari Y dengan kelas v yang terkait dengan kelas a; dan semua nilai yang sama dengan a adalah nilai yang sama.

2.6 Gradient Boosting

Algortima Gradient Boosting (GB) di mana metode ini menggunakan mekanisme peningkatan gradient descent dalam rangka evaluasi untuk pembuatan model berikutnya. Friedman (2001) memperkenalkan Gradient Boosting. Algoritma ini kemudian dikembangkan lebih lanjut, dan salah satu aplikasinya adalah eXtreme Gradient Boosting atau XGBoost yang dikembangkan oleh Chen dan Guestrin (2016). Algoritma tersebut merupakan pengembangan dari algortima Gradient Boosting Machine klasik dan hanya digunakan untuk data yang memiliki label dalam proses latihnya[15].

1. Membuat nilai prediksi konstan awal $F_0(x)$.
$$F_0(x) = \log\left[\frac{x}{1-x}\right] (3)$$
dengan $F0(x)$ adalah nilai prediksi kosntan awal, x adalah nilai target klasifikasi, dimana lulus tepat waktu (0) dan lulus tidak tepat waktu.
2. Untuk m = 1 sampai M, hitung.
$$r_{im} = r_i - p \ (4)$$
dengan rim adalah residual pada iterasi m untuk data ke-i, yi adalah nilai variabel Y pada data ke-i, dan p adalah nilai probabilitas prediksi.
3. Kemudian sesuaikan dengan pohon klasifikasi dengan nilai rim dan buat daerah simpul terminal Rjm untuk j = 1, … , . Dengan Rjm adalah himpunan sampel pada simpul terminal untuk iterasi ke-m, j adalah indeks simpul terminal dalam pohon keputusan, dan Jm adalah jumlah simpul terminal yang dihasilkan pohon keputusan pada iterasi ke-m.
4. Untuk j = 1, … ,. Hitung.
$$Y_{jm} = \frac{\sum_{j=1}^{Jm}(y-p)}{\sum_{j=1}^{Jm} p(1-p)} (5)$$
5. Update model.
$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{Jm} Y_{jm} J(x \in R_{jm}) (6)$$

2.7 Evaluation

Evaluation dilakukan setelah proses pengujian model selesai. Tujuan dari evaluasi dan perbandingan algoritma adalah membuktikan bahwa model benarbenar mempresentasikan sesuai pemodelan yang dilakukan dan mengukur performa dari algoritma dengan menentukan tingkat keakurasian dari dua algoritma yang digunakan. Penelitian ini menggunakan pengujian Confusion matrix dan ROC-AUC.

2.7.1 Confusion Matrix

Dalam pengenalan pola dan disiplin lain dari machine learning, jumlah elemen diagonal dari matriks konfusi banyak digunakan untuk mengukur keberhasilan klasifikasi berdasarkan algoritma atau pengamatan manusia[16]

Confusion matrix adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah[17]. Contoh confusion matrix untuk klasifikasi biner di tunjukkan pada Tabel 1.

Tabel 1. Bentuk Confusion Matrix

	Prediksi Positif	**Prediksi Negatif**
Aktual positif	TN	FN
Aktual Negatif	FP	TN

Berdasarkan Tabel 1 dapat di ketahui pada setiap isian dari confusion matrix dijelaskan sebagai berikut :

1) True Positif (TN)= Memberi jumlah banyak data yang diprediksi sebagai positif dan ternyata aktual menyatakan demikian.
2) True Negatif (TN)= Memberi jumlah banyak data yang diprediksi sebagai negative dan ternyata aktual menyatakan demikian.
3) False Positif (FP)= Memberi jumlah banyak data yang salah diprediksi sebagai positif yang seharusnya negatif.
4) False Negatif (FN)= Memberi jumlah banyak data yang salah diprediksi sebagai negative yang seharusnya positif.

Rumus confusion matrix untuk menghitung accuracy, precision, dan recall seperti berikut:

Akurasi, yaitu seberapa baik model benar melakukan klasifikasi, persamaan 7 menunjukan cara memperoleh skor akurasi.

$$\frac{TP+}{Total} \quad (7)$$

Precision, yaitu skor yang mengukur seberapa sering model dengan benar memprediksi positif, untuk mendapatkan skor ini dapat diperoleh pada persamaan 8.

$$\frac{TP+T}{Total} \quad (8)$$

Recall, merupakan skor untuk mengukur seberapa sering model memprediksi positif ketika aktualnya juga positif, skor ini dapat diperoleh melalui persamaan 9.

$\frac{TP}{FN+T}$ (9)F1-Score, merupakan rata-rata harmonic dari Precision dan Recall. F1-Score dapat diperoleh melalui persamaan 10.

$$\frac{precision*recall}{precesion+recall} \quad (10)$$

2.7.2 ROC-AUC

ROC-AUC adalah area karakteristik operasional penerima di bawah kurva. Ini adalah metrik evaluasi yang digunakan untuk mengukur kinerja klasifikasi model, terutama untuk model biner (dua kelas). Metrik ini berfokus pada kemampuan model untuk membedakan kelas positif dan negatif dengan mempertimbangkan trade-off antara tingkat Tingkat Positif Benar (TPR) dan Tingkat Negatif Positif (FPR)[18].

Definisi Istilah Penting dalam Evaluasi Model:

- Positif Sejati (True Positive/TP): Jumlah pengamatan yang benar-benar positif dan diprediksi sebagai positif oleh model.
- Positif Palsu (False Positive/FP): Jumlah pengamatan yang sebenarnya negatif tetapi salah diprediksi sebagai positif oleh model.
- Negatif Sejati (True Negative/TN): Jumlah pengamatan yang benar-benar negatif dan diprediksi sebagai negatif oleh model.

- Negatif Palsu (False Negative/FN): Jumlah pengamatan yang sebenarnya positif tetapi salah diprediksi sebagai negatif oleh model.

Tabel 2. *ROC AUC*

			Ground truth	
		+		**-**
Prediksi	+	TP		FP
	-	FN		TN

True Positive Rate (TPR), juga dikenal sebagai *Sensitivity* atau *Recall*.

$$TPR = \frac{TP}{TP+FN} (11)$$

Tingkat Positif Palsu (FPR) Ini mengukur sejauh mana model memprediksi kasus negatif sebagai positif.

$$FPR = \frac{FP}{FP+TN} (12)$$

Kemudian, Hubungan antara TPR dan FPR pada ambang batas klasifikasi yang berbeda diwakili oleh kurva ROC. Setiap titik pada ROC mewakili ambang batas tertentu. Kinerja model meningkat jika kurva ROC lebih dekat ke sudut kiri atas.

3. HASIL DAN PEMBAHASAN

Bagian ini membahas implementasi algoritma Random Forest dan Gradient Boosting dalam mengklasifikasikan churn pelanggan telco. Model dibangun, diterapkan, dan diuji dengan membagi data menjadi set pelatihan dan set pengujian. Performanya dievaluasi menggunakan Confusin Matrix dan ROC AUC.

3.1 Pengumpulan data

Data yang digunakan dalam penelitian ini berasal dari dataset telco churn yang mencakup informasi demografis, layanan yang digunakan, dan status churn. Data di ambil dari website kaggle Dengan jumlah data 7.043 dan 50 variabel data telco churn california.

3.2 Pre-processing

Selanjutnya sebelum analisis dilakukan preprocessing dilakukan untuk membersihkan dan mengubah data untuk menghasilkan hasil yang akurat dan relevan.

3.2.1 Cleaning data

Data yang digunakan merupakan data telco churn california yang akan dilakukan pengurangan variabel sebanyak 30 dan yang akan digunakan sebanyak 20 variabel. Data ini telah dilakukan proses cleaning.

	Gender	SeniorCitizen	Married	Dependents	TenureinMonths	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtectionPlan	PremiumTechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharge	TotalCharges
0	Male	Yes	No	No	1	No	No	Yes	No	No	Yes	No	No	Yes	Month-to-Month	Yes	Bank Withdrawal	39.65	39.65
1	Female	Yes	Yes	Yes	8	Yes	Yes	Yes	No	Yes	No	No	No	No	Month-to-Month	Yes	Credit Card	80.65	633.30
2	Male	Yes	No	Yes	18	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Month-to-Month	Yes	Bank Withdrawal	95.45	1752.55
3	Female	Yes	Yes	Yes	25	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	Month-to-Month	Yes	Bank Withdrawal	98.50	2514.50
4	Female	Yes	Yes	Yes	37	Yes	Yes	Yes	No	No	No	No	No	No	Month-to-Month	Yes	Bank Withdrawal	76.50	2868.15
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	Female	No	No	No	72	Yes	No	No	No	No	No	No	No	No	Two Year	Yes	Bank Withdrawal	21.15	1419.40
7039	Male	No	Yes	Yes	24	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	One Year	Yes	Mailed Check	84.80	1990.50
7040	Female	No	Yes	Yes	72	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	One Year	Yes	Credit Card	103.20	7362.90
7041	Female	No	Yes	Yes	11	No	No	Yes	Yes	No	No	No	No	No	Month-to-Month	Yes	Bank Withdrawal	29.60	346.45
7042	Male	No	No	No	66	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Two Year	Yes	Bank Withdrawal	105.65	6844.50

7043 rows × 20 columns

Gambar 2. Hasil Cleaning Data

Berdasarkan Gambar 2 Terdapat 30 variabel yang tidak digunakan dari 50 variabel yang ada. variabel tersebut yaitu Customer ID, Age, Under 30, Country, State, City, Zip Code, Latitude, Longitude, Population, Quarter, Referred a Friend, Number of Referrals, Offer, Avg Monthly Long Distance Charges, Internet Type, Avg Monthly GB Download, Device Protection Plan, Streaming Music, Unlimited Data, Total Refunds, Total Extra Data Charges, Total Long Distance Charges, Total Revenue, Satisfaction Score, Customer Status, Churn Score, CLTV, Churn Category, Churn Reason. Kolom tersebut akan dihapus dari tabel dataset yang akan digunakan Handling Missing Value.

3.2.2 Transformasi Data

Selanjutnya transformasi data untuk masing-masing variabel untuk membagi data kategori dan numerik.

	SeniorCitizen	Dependents	TenureinMonths	PhoneService	PaperlessBilling	MonthlyCharge	TotalCharges	ChurnLabel	Gender_Male	Married_Yes	...	OnlineSecurity_Yes	OnlineBackup_Yes	DeviceProtectionPlan_Yes	PremiumTechSupport_Yes	StreamingTV_Yes	StreamingMovies_Yes	Cont
0	1	0	1	0	1	39.65	39.65	1	True	False	...	False	False	True	False	False	True	
1	1	1	8	1	1	80.65	633.30	1	False	True	...	False	True	False	False	False	False	
2	1	1	18	1	1	95.45	1752.55	1	True	False	...	False	False	False	False	True	True	
3	1	1	25	1	1	98.50	2514.50	1	False	True	...	False	True	True	False	True	True	
4	1	1	37	1	1	76.50	2868.15	1	False	True	...	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7038	0	0	72	1	1	21.15	1419.40	0	False	False	...	False	False	False	False	False	False	
7039	0	1	24	1	1	84.80	1990.50	0	True	True	...	True	False	True	True	True	True	
7040	0	1	72	1	1	103.20	7362.90	0	False	True	...	False	True	True	False	True	True	
7041	0	1	11	0	1	29.60	346.45	0	False	True	...	True	False	False	False	False	False	
7042	0	0	66	1	1	105.65	6844.50	0	True	False	...	True	False	True	True	True	True	

7043 rows × 22 columns

Gambar 3. Hasil Transformasi Data

Pada Gambar 3, ialah membagi data kategori, numeric dan one hot encoding. Variabel dengan data categori yaitu Senior Citizen, Dependents, Phone Service, Paperless Billing, Churn Label. Dan variable yang di ubah menjadi numeric adalah Total Charges karena Dikonversi ke tipe data numerik dengan mengatasi nilai yang tidak valid dengan nilai rata-rata (mean). Kemudian variable untuk One-Hot Encoding adalah Gender, Married, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Contract, Payment Method.

Setelah dilakukannya preprocessing data, Selanjutnya data yang di gunakan pada penelitin ini menggunakan variabel sebanyak 20 dan jumlah data 7.043, yang mana di tunjukan oleh tabel 3.

Tabel 3. Dataset Churn Pelanggan Telco California

No	Gender	SeniorCitizen	Married	..	MonthlyCharge	TotalCharges	ChurnLabel
0	Male	Yes	No	..	39.65	39.65	Yes
1	Female	Yes	Yes	..	80.65	633.3	Yes
2	Male	Yes	No	..	95.45	1752.55	Yes
3	Female	Yes	Yes	..	98.50	2514.5	Yes
4	Female	Yes	Yes	..	76.50	2868.15	Yes
…	…	…	…	..	…	…	…
7038	Female	No	No	..	21.15	1419.4	No
7039	Male	No	Yes	..	84.80	1990.5	No

No	Gender	SeniorCitizen	Married	..	MonthlyCharge	TotalCharges	ChurnLabel
7040	Female	No	Yes	..	103.20	7362.9	No
7041	Female	No	Yes	..	29.60	346.45	No
7042	Male	No	No	..	105.65	6844.5	No

Berdasarkan Tabel 3, Merupakan dataset churn pelanggan telco california yang terdiri dari fitur, Tentu, Gender(Jenis kelamin pelanggan: Male/Female), Senior Citizen (Indikator warga senior: 0/1), Married (Status pernikahan: Yes/No), Dependents (Indikator tanggungan: Yes/No), Tenurein Months (Lama berlangganan dalam bulan), PhoneService (Layanan telepon: Yes/No), Multiple Lines (Saluran telepon ganda: Yes/No/No phone service), Internet Service (Layanan internet: DSL/Fiber optic/No), Online Security (Keamanan online: Yes/No/No internet service), Online Backup (Backup online: Yes/No/No internet service), Device Protection Plan (Perlindungan perangkat: Yes/No/No internet service), Premium Tech Support (Dukungan teknis premium: Yes/No/No internet service), Streaming TV (Layanan streaming TV: Yes/No/No internet service), Streaming Movies (Layanan streaming film: Yes/No/No internet service), Contract(Jenis kontrak: Month-to-Month/One Year/Two Year).

3.3 Splitting data

Sebelum melakukan analisis klasifikasi, data dibagi menjadi dua bagian: pelatihan data dan penilaian data. Pelatihan data membantu mengajarkan algoritma untuk membuat model, dan penilaian data digunakan untuk mengukur tingkat keakuratan dan performa yang didapatkan dari data training. Pada penelitian ini Data training dan data testing dibagi dengan proporsi 80% untuk data training dan 20% dari data testing dari total dataset.

Tabel 4. Pembagian Data Training dan Data Testing

Keterangan	Data Training	Data Testing	Total
Proporsi	80%	20%	100%
Jumlah	5.634	1.409	7.043

Berdaskan Tabel 4 Dapat diketahui dari 7.043 dataset yang ada pembagian data untuk training sebanyak 5.634 data dan untuk data testing 1.409 data, pembagian data training dan data testing di lakukan secara *random* dengan bantuan *software* python.

3.4 Klasifikasi

3.4.1 *Random Forest*

```
[ ] # Membuat model Random Forest
    rf_model = RandomForestClassifier(random_state=42)

[ ]
    # Melatih model menggunakan data latih
    rf_model.fit(X_train, y_train)

    RandomForestClassifier
RandomForestClassifier(random_state=42)

    # Melakukan prediksi menggunakan data uji
    rf_predictions = rf_model.predict(X_test)

[ ] # Evaluasi model
    rf_accuracy = accuracy_score(y_test, rf_predictions)
    rf_conf_matrix = confusion_matrix(y_test, rf_predictions)
    rf_classification_rep = classification_report(y_test, rf_predictions)
```

Gambar 4. Tahapan *Random Forest*

Berdasarkan gambar 4 Untuk mengklasifikasikan churn pelanggan menggunakan algoritma Random Forest, menggunakan scikit-learn, sebuah pustaka machine learning di Python. Pertama, model diinisialisasi dengan random_state=42 untuk memastikan reprodusibilitas hasil. Model kemudian dilatih menggunakan data latih (X_train dan y_train). Setelah pelatihan, model digunakan untuk membuat prediksi pada data uji (X_test).

3.4.2 Gradient Boosting

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[ ] gb_model = GradientBoostingClassifier(random_state=42)
    gb_model.fit(X_train, y_train)

    GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Gambar 5. Tahapan *Gradient Boosting*

Berdasarkan gambar 5 Model Gradient Boosting Classifier diinisialisasi dengan random_state=42 untuk memastikan hasil yang dapat direproduksi. Setelah membagi dataset menggunakan train_test_split dengan proporsi data uji sebesar 20% dan inisialisasi random_state=42, proses pelatihan dilakukan menggunakan data latih (X_train dan y_train). Ini melibatkan pembentukan serangkaian estimasi berurutan, terutama pohon keputusan, dengan tujuan mengurangi kesalahan prediksi.

Setelah dilatih, model digunakan untuk melakukan prediksi terhadap data uji (X_test). Hasil prediksi disimpan dalam variabel gb_predictions. Evaluasi performa model dilakukan menggunakan metrik seperti akurasi untuk mengukur persentase prediksi yang benar, matriks kebingungan (confusion matrix) yang menggambarkan jumlah prediksi yang benar dan salah untuk setiap kelas, serta laporan klasifikasi yang menyediakan informasi lebih detail seperti precision, recall, dan F1-score untuk setiap kelas.

3.5 Evaluation Model

Evaluation model yang di hasilkan dari pengklasifikasi menggunakan algoritma *random forest* dan *gradient boosting* adalah confusion matrix dan ROC-AUC.

```
Random Forest Accuracy: 0.8126330731014905
Random Forest Confusion Matrix:
[[919  90]
 [174 226]]
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.91      0.87      1009
           1       0.72      0.56      0.63       400

    accuracy                           0.81      1409
   macro avg       0.78      0.74      0.75      1409
weighted avg       0.81      0.81      0.81      1409
```

Gambar 6. Hasil Confusion Matrix Menggunakan Random Forest*Berdasarkan gambar 6 hasil pengklasifikasikan pada churn pelanggan telco menggukan algoritma random forest mendapatkan hasil 81%.*

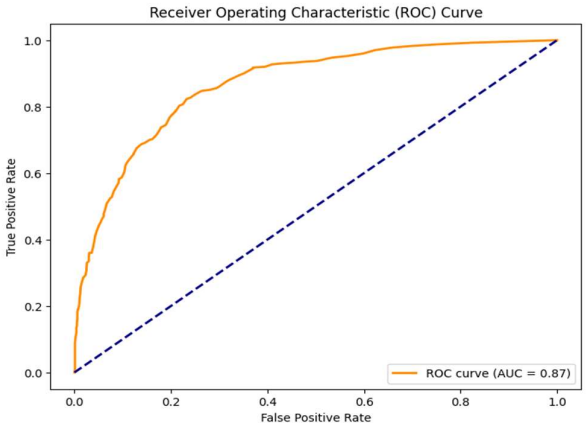
```
Gradient Boosting Accuracy: 0.8261178140525195
Gradient Boosting Confusion Matrix:
[[915  94]
 [151 249]]
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.91      0.88      1009
           1       0.73      0.62      0.67       400

    accuracy                           0.83      1409
   macro avg       0.79      0.76      0.78      1409
weighted avg       0.82      0.83      0.82      1409
```

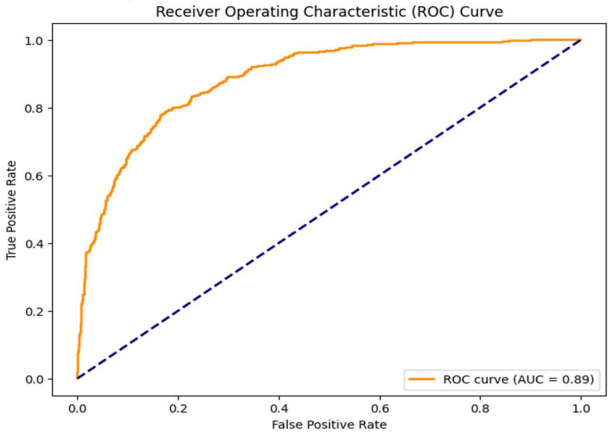
Gambar 7. *Hasil Klasifikasi Menggunakan Gradient Boosting*

Sedangkan pada gambar 7 hasil pengklasifikasian pada churn pelanggan telco menggunakan algoritma *gradient boosting* mendapatkan hasil 82%.



Gambar 8. Hasil ROC-AUC Algoritma Random Forest

Berdasarkan gambar 8, hasil ROC-AUC pada churn pelanggan telco menggukan algoritma *random forest* mendapatkan hasil 87%



Gambar 9. Hasil ROC-AUC Algoritma Gradient Boosting

Berdasarkan gambar 9 hasil ROC-AUC pada churn pelanggan telco menggunakan algoritma *gradient boosting* mendapatkan hasil 89%.

3.6 Analisis Perbandingan Performa

Tabel 5. Hasil Perbandingan

Algoritma	Splitting Data	Evaluasi		Training Time (Seconds)	Inference Time (seconds)
		Confusion Matrix	ROC AUC		
Random Forest	80:20	81%	0.87	1.0329	0.0439
Gradient Boosting	80:20	83%	0.89	1.0142	0.0352

Berdasarkan Tabel 5 setelah melakukan evaluasi terhadap beberapa *spliting data* penelitian ini mengambil nilai *spliting data* terbaik yaitu 80:20. Hasil akurasi dari algoritma random forest dan gradient boosting mendapatkan hasil *acuracy* 0.81 dan *ROC-AUC* 0.87 pada algoritma *random forest*, Sedangkan *gradient boosting* mendapatkan *acuracy* 0.83 *ROC-AUC* 0.89

Dalam hal Training Time, Gradient Boosting memerlukan waktu 1.0142 detik dan Random Forest memerlukan waktu 1.0329 detik. Untuk Inference Time, Gradient Boosting memerlukan waktu 0.0352 detik dan Random Forest yang memerlukan waktu 0.0439 detik.

4. KESIMPULAN DAN SARAN

Penelitian ini membandingkan performa algoritma Random Forest dan Gradient Boosting dalam mengklasifikasi churn pelanggan Telco. Hasil dari penelitian ini menunjukkan bahwa Gradient Boosting unggul dibandingkan Random Forest dalam hal akurasi, ROC AUC, Training Time, dan Inference Time. Gradient Boosting mencapai akurasi 83% dan ROC AUC 0.89%, sementara Random Forest mencapai akurasi 81% dan ROC AUC 0.87%.

Dari sisi efisiensi waktu, Gradient Boosting memerlukan waktu pelatihan (Training Time) sebesar 1.0142 detik dan waktu inferensi (Inference Time) sebesar 0.0352 detik. Sementara itu, Random Forest memerlukan waktu pelatihan sebesar 1.0329 detik dan waktu inferensi sebesar 0.0439 detik, hal ini menunjukkan Gradient Boosting lebih cepat 0.0187 detik dari segi Training time dan 0.0087 detik lebih cepat dari segi Inferece Time dibandingkan dengan random forest karena Gradient Boosting membutuhkan waktu yang lebih singkat untuk kedua proses tersebuts

Berdasarkan Hasil evaluasi menunjukkan bahwa Gradient Boosting lebih efektif dalam mengklasifikasi churn pelanggan telco karena kemampuannya dalam mengoptimalkan kesalahan model secara iteratif dibandingkan Random forest.

Penelitian selanjutnya bisa mengembangkan metode yang lebih kompleks atau mengkombinasikan beberapa algoritma deep learning seperti Neural Networks atau Convolutional Neural Networks (CNN) untuk meningkatkan akurasi prediksi churn pelanggan telco.

DAFTAR PUSTAKA

[1] M. Rizki Kurniawan, P. Nurul Sabrina, and R. Ilyas, "PREDIKSI CUSTOMER CHURN PADA PERUSAHAAN TELEKOMUNIKASI MENGGUNAKAN ALGORITMA C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 5, pp. 3369–3375, 2024, doi: 10.36040/jati.v7i5.7476.

[2] S. D. Damanik and M. I. Jambak, "Klasifikasi Customer Churn pada Telekomunikasi Industri Untuk Retensi Pelanggan Menggunakan Algoritma C4.5," *KLIK Kaji. Ilm.*

Inform. dan Komput., vol. 3, no. 6, pp. 1303–1309, 2023, doi: 10.30865/klik.v3i6.829.

[3] K. Setyoardi and P. F. Ariyani, "IMPLEMENTASI DATA MINING PADA DATA THE JAVANESE CAFE IMPLEMENTATION OF DATA MINING ON THE JAVANESE CAFÉ DATA BASED ON WEB USING APRIORI ALGORITHM," vol. 3, no. 140, pp. 195–203, 2024.

[4] A. F. Ramadhan, S. D. Permai, J. Harefa, and Alexander, "The Comparison of Random Forest and Artificial Neural Network for Customer Churn Prediction in Telecommunication," *Proc. 3rd 2023 Int. Conf. Smart Cities, Autom. Intell. Comput. Syst. ICON-SONICS 2023*, no. December, pp. 224–229, 2023, doi: 10.1109/ICON-SONICS59898.2023.10435087.

[5] I. Slamet and I. Susanto, "Prediksi Nasabah Churn Dengan Algoritma Decision Tree, Random Forest Dan Support Vector Machine," *Escaf*, pp. 1045–1053, 2024.

[6] A. A. Saputra, B. N. Sari, and C. Rozikin, "Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit," *J. Ilm. Wahana Pendidik.*, vol. 10, no. 7, pp. 27–36, 2024.

[7] N. Agian, S. Dinata, G. Abdurrahman, and N. Q. Fitriyah, "Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi," 2023.

[8] P. Studi, T. Informatika, F. Sains, D. A. N. Teknologi, U. Islam, and N. Syarif, "Prediksi Churn Konsumen Menggunakan Algoritma Random Forest dengan Fuzzy C-Means untuk Meningkatkan Produktivitas Penjualan Bisnis Prediksi Churn Konsumen Menggunakan Algoritma Random Forest dengan Fuzzy C-Means untuk Meningkatkan Produktivitas Penjualan," 2023.

[9] Alifian adam, "Data Preprocessing: Pengertian, Manfaat, dan Tahapan Kerjanya," *accurate*, 2022. https://accurate.id/teknologi/data-preprocessing/

[10] S. A. Risyad, "Data Cleaning: Arti, Manfaat, dan Cara Melakukannya," *ibimbing*, 2023. https://dibimbing.id/blog/detail/data-cleaning-arti-manfaat-dan-cara-melakukannya

[11] Anwar Hidayat, "Pengertian Dan Jenis Transformasi Data," *Statistikian*, 2023. https://www.statistikian.com/2013/01/transformasi-data.html

[12] A. Putri, C. S. Hardiana, E. Novfuja, F. Try, and P. Siregar, "Comparison of K-NN , Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction Komparasi Algoritma K-NN , Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," vol. 3, no. April, pp. 20–26, 2023.

[13] P. Bidang, K. Sains, Y. Mardi, J. Gajah, M. No, and S. Barat, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika".

[14] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 578, 2021, doi: 10.32493/informatika.v5i4.7575.

[15] M. Ridwansyah and H. Zakaria, "Implementasi Algortima Gradient Boosting Pada Aplikasi Hutang Piutang Perorangan Secara Berbasis Web Untuk Meningkatan Akurasi Prediksi Pelunasan Hutang (Studi Kasus : PT Naila Kreasi Mandiri)," *Ridwansyah, Muhammad Zakaria, Hadi*, vol. 1, no. 4, pp. 440–451, 2023.

[16] A. Chusyairi, T. Haryanto, and R. N. Hayat, "Prediksi Perubahan Iklim Untuk Pertumbuhan Tanaman Jeruk Keprok Menggunakan Naïve Bayes," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 18, no. 1, p. 23, 2023, doi: 10.30872/jim.v18i1.9352.

[17] J. S. Komputer, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," vol. 5, no. November 2019, pp. 697–711, 2021.

[18] andi, "Memahami dan memodelkan Matriks Evaluasi ROC-AUC dalam Machine Learning," *medium*, 2023.