Jurnal Sistem Informasi Bisnis 01(2025)

http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

Prediksi *Churn* Pelanggan Telekomunikasi dengan Optimalisasi Seleksi Fitur dan *Tuning Hyperparameter* pada Algoritma Klasifikasi C4.5

Soterio Antoh[1], Rudy Herteno[2], Irwan Budiman[3], Dwi Kartini[4], Muhammad Itqan Mazdadi[5]

Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Lambung Mangkurat

Naskah masuk: 23 Oktober 2024; Revisi terakhir: 18 Desember 2024;
Diterima publikasi: 8 Januari 2025; Tersedia daring: 28 Februari 2025
DOI : 10.21456/vol15iss1pp60-67

Abtract

In the telecommunications industry, predicting customer churn is crucial for maintaining business sustainability. High churn rates can negatively impact profitability, necessitating effective retention strategies. This research aims to enhance the accuracy of telecommunications customer churn prediction by optimizing the C4.5 classification algorithm through feature selection and hyperparameter tuning. The methods used include Information Gain for feature selection and hyperparameter tuning with Random Search and Grid Search. This study utilizes the Telco Customer Churn dataset from Kaggle, split into an 80:20 ratio for training and testing data. Six approaches are applied: (1) the basic C4.5 algorithm, (2) C4.5 with Information Gain, (3) C4.5 with Random Search, (4) C4.5 with Grid Search, (5) C4.5 with a combination of Information Gain and Random Search, and (6) C4.5 with a combination of Information Gain and Grid Search. The results indicate that the C4.5 algorithm alone achieves an accuracy of 74.09%, while applying Information Gain increases accuracy to 78.42%. Hyperparameter tuning with Random Search achieves the highest accuracy of 80.05%, whereas Grid Search reaches 77.71%. Combining Information Gain with Random Search results in an accuracy of 78.99%, while combining Information Gain with Grid Search yields an accuracy of 78.85%. These findings suggest that hyperparameter tuning using Random Search significantly improves accuracy compared to other methods, while Information Gain feature selection does not have a significant impact on performance in this context.

Keywords: Classification; C4.5 Algorithm; Customer Churn; Hyperparameters; Feature Selection

Abstrak

Dalam industri telekomunikasi, prediksi churn pelanggan sangat penting untuk menjaga keberlanjutan bisnis. Tingkat *churn* yang tinggi dapat mengurangi profitabilitas, sehingga diperlukan strategi retensi yang efektif. Penelitian ini bertujuan meningkatkan akurasi prediksi *churn* pelanggan dengan mengoptimalkan algoritma klasifikasi C4.5 melalui seleksi fitur dan *tuning hyperparameter*. Metode penelitian ini mencakup seleksi fitur menggunakan *Information Gain* serta *tuning hyperparameter* dengan Random *Search* dan *Grid Search*. Data yang digunakan adalah dataset Telco *Customer Churn* dari Kaggle, yang dibagi dalam rasio 80:20 untuk data latih dan data uji. Penelitian ini melibatkan enam pendekatan: (1) Algoritma C4.5 tanpa optimasi, (2) C4.5 dengan *Information Gain*, (3) C4.5 dengan *Random Search*, (4) C4.5 dengan *Grid Search*, (5) C4.5 dengan kombinasi *Information Gain* dan *Random Search*, serta (6) C4.5 dengan kombinasi *Information Gain* dan *Grid Search*. Hasilnya menunjukkan bahwa Algoritma C4.5 tanpa optimasi memiliki akurasi 74,09%, sedangkan penerapan *Information Gain* meningkatkan akurasi menjadi 78,42%. *Tuning hyperparameter* dengan Random *Search* menghasilkan akurasi tertinggi sebesar 80,05%, sedangkan *Grid Search* mencapai 77,71%. Kombinasi *Information Gain* dengan Random Search memberikan akurasi 78,99%, sementara kombinasi dengan *Grid Search* menghasilkan akurasi 78,85%. Hasil penelitian ini mengindikasikan bahwa *tuning hyperparameter* menggunakan Random *Search* memberikan peningkatan akurasi yang signifikan dibandingkan metode lainnya, sementara seleksi fitur *Information Gain* tidak berdampak signifikan pada peningkatan kinerja dalam konteks prediksi *churn* ini.

Kata Kunci: Klasifikasi; Algoritma C4.5; Churn; Hyperparameter; Seleksi Fitur

*) Corresponding author: rudy.herteno@ulm.ac.id

http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

1. Pendahuluan

Industri telekomunikasi sering menghadapi masalah churn pelanggan, yaitu ketika pelanggan beralih dari satu penyedia layanan ke penyedia lain atau memutuskan berhenti menggunakan layanan telekomunikasi sama sekali. *Churn* dapat menyebabkan kerugian finansial yang signifikan bagi penyedia layanan, serta dapat menurunkan reputasi merek mereka (Damanik & Jambak, 2023). Menjaga pelanggan yang sudah ada menjadi prioritas bagi perusahaan telekomunikasi. Menjaga pelanggan merupakan hal yang krusial karena umumnya biaya untuk menarik pelanggan baru lebih besar dibandingkan dengan mempertahankan pelanggan yang sudah ada. Oleh karena itu, mengidentifikasi pelanggan yang berpotensi untuk beralih dan mengambil tindakan pencegahan adalah strategi penting (Wardani *et al.*, 2018).

Dampak *churn* pelanggan dalam industri telekomunikasi sangat signifikan terhadap daya saing dan keberlanjutan bisnis. Tingginya *churn* mengakibatkan penurunan pendapatan, *berkurangnya* pangsa pasar, dan peningkatan biaya pemasaran serta operasional karena memperoleh pelanggan baru jauh lebih mahal daripada mempertahankan pelanggan yang sudah ada. Selain itu, churn berpotensi menurunkan nilai perusahaan di mata investor dan menyebabkan ketidakstabilan pendapatan, yang berdampak negatif pada citra merek. Ketika pelanggan sering berpindah layanan, kepuasan dan loyalitas pelanggan yang tersisa juga menurun, membuat mereka lebih rentan untuk beralih ke kompetitor. Untuk mengatasi dampak ini, perusahaan perlu berupaya meningkatkan kualitas layanan, menawarkan harga yang kompetitif, dan memberikan insentif untuk memperkuat loyalitas pelanggan. (Puneeth *et al.*, 2022).

Prediksi *churn* pelanggan penting bagi perusahaan telekomunikasi untuk mengurangi dampak negatif pada pendapatan dan reputasi. Namun, prediksi ini menjadi lebih kompleks dalam situasi yang tidak pasti karena banyak faktor yang memengaruhi perilaku pelanggan. (Amin *et al.*, 2019). Memahami faktor-faktor yang memengaruhi perilaku churn pelanggan dan mengembangkan model prediktif yang akurat adalah langkah penting untuk mengurangi churn dan mempertahankan pangsa pasar. Pendekatan berbasis data dan pembelajaran mesin kini populer dalam konteks ini. (Loukili *et al.*, 2022).

Algoritma klasifikasi adalah metode dalam pembelajaran mesin yang digunakan untuk mengkategorikan data ke dalam kelas tertentu. Dalam prediksi churn, algoritma ini menganalisis data pelanggan untuk mengidentifikasi pola yang menunjukkan kemungkinan mereka akan berhenti menggunakan layanan. Dengan teknik seperti C4.5 dapat membantu perusahaan merumuskan strategi retensi yang efektif, mengurangi tingkat *churn*, dan meningkatkan loyalitas pelanggan (Anita *et al.*, 2021)

C4.5 adalah algoritma klasifikasi berbasis pohon keputusan yang populer untuk memprediksi churn pelanggan. Algoritma ini menghasilkan aturan keputusan yang mudah diinterpretasi, membantu perusahaan memahami faktor-faktor yang memengaruhi keputusan pelanggan untuk beralih (Utami *et al.*, 2020). Algoritma C4.5, dikembangkan oleh Ross Quinlan pada tahun 1993 sebagai penyempurnaan dari ID3, menggunakan pendekatan rekursif untuk membangun pohon keputusan untuk klasifikasi data (Saputra *et al.*, 2021).

Seleksi fitur dalam algoritma C4.5 sangat penting untuk meningkatkan kinerja algoritma C4.5 untuk prediksi churn pelanggan. C4.5 membangun pohon keputusan berdasarkan atribut dataset, dan dengan memilih fitur yang relevan, model dapat lebih fokus pada variabel penting dalam memprediksi churn. Menghilangkan fitur yang tidak signifikan membantu mengurangi kompleksitas dan meningkatkan efisiensi komputasi. Dengan mempertimbangkan fitur yang berkorelasi kuat dengan churn, model menjadi lebih akurat dan presisi. Teknik seleksi fitur univariat, seperti Information Gain, membantu mengidentifikasi atribut yang paling berkontribusi pada prediksi, membuat model lebih mudah diinterpretasi dan lebih andal dalam mendukung keputusan bisnis terkait retensi pelanggan (Sana *et al.*, 2022).

Optimasi hyperparameter sangat penting dalam meningkatkan kinerja algoritma C4.5 untuk prediksi churn pelanggan. Dengan menyesuaikan parameter seperti kedalaman maksimum pohon dan jumlah minimum instance per node, model menjadi lebih seimbang antara bias dan varians, mengurangi overfitting, dan meningkatkan generalisasi pada data baru. Teknik seperti random search dan grid search membantu menemukan konfigurasi terbaik untuk memastikan model andal pada data uji. Optimasi ini memberikan akurasi lebih tinggi, presisi lebih baik, dan stabilitas interpretasi yang lebih baik, sehingga mendukung keputusan bisnis yang efektif dalam strategi retensi pelanggan (Loukili *et al.*, 2022).

Penelitian (Utami *et al.*, 2020) mengenai Prediksi *Churn Rate* pada Jasa Telekomunikasi menggunakan algoritma C4.5 menghasilkan model dengan akurasi 87%, presisi 87,5%, dan recall 97%. Namun, penelitian tersebut belum mempertimbangkan optimalisasi seleksi fitur maupun tuning hyperparameter untuk lebih meningkatkan kinerja model.

Penelitian terkait optimalisasi seleksi fitur dapat ditemukan dalam studi yang telah dilakukan oleh (Yoga Siswa *et al.*, 2022) Menggunakan seleksi fitur Information Gain dan teknik pruning dapat meningkatkan kinerja algoritma C4.5 sebesar 3,46% dalam kasus keterlambatan biaya kuliah. Namun, penelitian tersebut belum mempertimbangkan penerapan optimalisasi seleksi fitur dan tuning hyperparameter secara bersamaan dalam konteks prediksi churn pelanggan telekomunikasi.

Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587 On-line:
http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

Penelitian lain juga dilakukan oleh Fajri & Primajaya (2023) Perbandingan teknik *Hyperparameter Grid Search* dan Random *Search* pada algoritma SVM menunjukkan bahwa kedua metode memiliki hasil yang serupa, dengan nilai rata-rata akurasi sebesar 84%. Namun, penelitian tersebut hanya berfokus pada SVM dan tidak mencakup algoritma lain seperti C4.5, khususnya dalam konteks prediksi churn pelanggan.

Penelitian oleh Akbar (2021) membandingkan klasifikasi C4.5 dengan dan tanpa seleksi fitur menggunakan *information gain*. Hasilnya menunjukkan bahwa klasifikasi dengan seleksi fitur lebih efektif, terutama dalam akurasi dan kecepatan komputasi, dengan penurunan atribut sebesar 96%. Namun, penelitian ini terbatas pada klasifikasi bibliografi otomatis dan belum mencakup konteks prediksi churn pelanggan.

Penelitian oleh Sadiq & Ahmed (2019) menggunakan algoritma *Decision Tree* C4.5 dengan *Grid Search* untuk meningkatkan akurasi dalam klasifikasi dan prediksi kinerja siswa. Hasilnya menunjukkan bahwa model C4.5 yang ditingkatkan memiliki kinerja yang lebih baik dibandingkan dengan model C4.5 (J48) di Weka dan algoritma lain yang diuji. Namun, penelitian ini berfokus pada prediksi kinerja siswa dan belum mengaplikasikan pendekatan serupa pada kasus churn pelanggan telekomunikasi.

Pada Penelitian sebelumnya mengenai prediksi churn pelanggan telekomunikasi telah menggunakan berbagai algoritma klasifikasi dan teknik machine learning untuk meningkatkan akurasi, namun seringkali kurang memperhatikan optimalisasi seleksi fitur dan *tuning hyperparameter* secara bersamaan dalam algoritma C4.5. Banyak studi yang membahas seleksi fitur atau *tuning hyperparameter*, tetapi jarang yang mengintegrasikan kedua aspek tersebut dalam satu penelitian komprehensif. Kesenjangan ini menunjukkan potensi peningkatan performa model prediksi churn pelanggan yang belum sepenuhnya dieksplorasi.

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah Bagaimana meningkatkan akurasi Algoritma C4.5 dalam memprediksi churn pelanggan di sektor telekomunikasi, baik sebelum dan setelah penerapan seleksi fitur *Information Gain* serta *Hyperparameter Tuning* menggunakan Random *Search* dan *Grid Search*, termasuk kombinasi keduanya. Penelitian ini bertujuan untuk membandingkan dua pendekatan penerapan seleksi fitur *Information Gain* dan *Hyperparameter Tuning* menggunakan Random *Search* serta *Grid Search*. Penerapan kedua pendekatan ini diharapkan dapat meningkatkan tingkat akurasi Algoritma C4.5 dalam memprediksi churn pelanggan dibandingkan dengan penerapan algoritma C4.5 tanpa optimasi.

2. Kerangka Teori

Customer churn terjadi ketika pelanggan meninggalkan perusahaan atau berhenti menggunakan produk. Ini penting bagi perusahaan, terutama di sektor telekomunikasi, karena berdampak negatif pada pendapatan dan reputasi (Sari *et al.*, 2023). Penelitian terbaru mengenai metode untuk meningkatkan akurasi prediksi churn pelanggan telekomunikasi pada Algoritma C4.5 akan dibahas dalam kerangka teori ini.

2.1. Algoritma C4.5

C4.5 adalah algoritma yang digunakan untuk membangun model pohon keputusan, yang dikembangkan oleh Ross Quinlan pada tahun 1993, adalah versi perbaikan dari ID3. Algoritms C4.5 berfungsi untuk klasifikasi data dengan cara membagi dataset berdasarkan atribut yang paling informatif, dengan tujuan meminimalkan entropi dan meningkatkan ketepatan klasifikasi. Proses kerjanya dimulai dengan pengumpulan data yang terdiri dari atribut dan kelas. Selanjutnya, data dipraproses dengan menghapus entri yang tidak lengkap dan mengatasi nilai yang hilang. Algoritma kemudian menghitung rasio informasi dari setiap atribut untuk menentukan mana yang paling informatif, sebelum membagi data dan membangun pohon keputusan. C4.5 juga menerapkan teknik pemangkasan untuk mengurangi kompleksitas pohon dan menghindari *overfitting*. Keunggulan C4.5 meliputi kemampuannya untuk menangani data kategorikal dan numerik, mengatasi nilai yang hilang, serta menghasilkan model yang mudah dipahami dan diinterpretasikan. (Kubat, 2021).

Menurut Nasrullah (2021) algoritma C4.5 sering kali memberikan akurasi yang lebih tinggi dibandingkan dengan algoritma lain seperti ID3 dan CART, menjadikannya pilihan yang kuat untuk tugas klasifikasi, terutama dalam dataset yang kompleks.

2.2. Information Gain

Information Gain merupakan salah satu teknik penting dalam seleksi fitur yang digunakan untuk meningkatkan kinerja algoritma klasifikasi. Ukuran ini menilai penurunan ketidakpastian data setelah pembagian berdasarkan atribut. Nilai *Information Gain* yang lebih tinggi menunjukkan bahwa atribut tersebut lebih efektif dalam memisahkan kelas. *Information Gain* memberikan manfaat signifikan dalam memilih fitur yang relevan. Dia membantu memilih fitur informatif yang meningkatkan akurasi model dan mengurangi kompleksitas, serta mencegah *overfitting* dengan menghindari fitur yang tidak relevan. Selain itu, dengan mengurangi jumlah fitur, proses pelatihan menjadi lebih cepat dan efisien, serta fitur yang dipilih lebih mudah diinterpretasikan, memungkinkan pemangku kepentingan memahami keputusan model (Saputra *et al.*, 2021).

Menurut Ramanda Hasibuan & MARJI (2019) cara kerja information gain sebagai berikut:

1. Menghitung Entropi: Entropi adalah ukuran ketidakpastian dalam data. Dihitung dengan rumus:

Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587 On-line:
http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

$$Entropi(S) = -\sum_{i=1}^{c} P_i \, log_2(P_i)$$

di mana *Pi* adalah proporsi kelas ke-i dalam dataset *S* dan c adalah jumlah kelas

2. Menghitung Entropi Setelah Pembagian**:** Ketika dataset dibagi berdasarkan fitur tertentu, kita menghitung entropi untuk setiap subset. Jika *A* adalah fitur yang digunakan untuk membagi dataset *S*, maka entropi setelah pembagian dapat dihitung dengan rumus:

$$Entropi(S|A) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropi(S_v)$$

di mana S_v adalah subset dari *S* untuk setiap nilai v dari fitur *A*.

3. Menghitung Information Gain**:** Information Gain dihitung dengan mengurangkan entropi setelah pembagian dari entropi awal:
$IG(S, A) = H(S) - H(S|A)$
Nilai IG yang lebih tinggi menunjukkan bahwa fitur tersebut memberikan lebih banyak informasi tentang kelas target.

2.3. Random Search

Random *search* adalah teknik optimasi hyperparameter yang memilih kombinasi acak dari nilai-nilai hyperparameter untuk menemukan kinerja model terbaik. Berbeda dengan pencarian *grid* yang sistematis, pencarian acak menjelajahi rentang nilai yang lebih luas dan dapat diparalelkan, menjadikannya efisien secara komputasi. Teknik ini mudah diterapkan dengan menentukan rentang *hyperparameter* dan jumlah kombinasi acak yang dievaluasi (Febiani *et al.*, 2022).

Penelitian oleh Rizky *et al.*, (2024) menunjukkan bahwa pendekatan menggunakan penyetelan hiperparameter dengan Metode Random *Search* lebih unggul dalam memprediksi cacat perangkat lunak dibandingkan dengan pendekatan berbasis pohon lainnya seperti DT, RF, dan DF. (Navon & Bronstein, 2022) menyatakan bahwa Random *Search* lebih efisien daripada *Grid Search* untuk optimasi hiperparameter dalam machine learning.

Dengan keunggulan dalam menjelajahi ruang parameter yang lebih luas dan kemudahan penerapan, Random Search menjadi salah satu metode yang disukai dalam proses tuning hyperparameter, khususnya dalam aplikasi yang memerlukan waktu komputasi yang efisien.

2.4. Grid Search

Grid Search adalah metode sistematis untuk menyetel *hyperparameter* dalam algoritma pembelajaran mesin dengan mengevaluasi setiap kombinasi nilai yang mungkin untuk *hyperparameter* yang relevan. Meskipun efektif dalam menemukan nilai terbaik dan memberikan wawasan tentang pengaruhnya terhadap model, metode ini dapat mahal secara komputasi, terutama dalam ruang pencarian yang besar. Meskipun demikian, pencarian grid tetap populer karena kemudahannya dalam penggunaan dan interpretasi hasil (Arden & Safitri, 2022).

Grid Search bekerja dengan mengevaluasi semua kombinasi *hyperparameter* yang mungkin dalam ruang pencarian yang telah ditentukan. Pengguna harus menetapkan satu set nilai untuk setiap *hyperparameter*, dan *Grid Search* akan menguji setiap kombinasi tersebut, melatih dan memvalidasi model untuk masing-masing kombinasi, kemudian mengevaluasi kinerjanya menggunakan metrik yang relevan. Keuntungan dari *Grid Search* adalah kemampuannya untuk menjamin temuan kombinasi hyperparameter yang optimal, serta kemudahan dalam pemahaman proses dan hasilnya. Namun, kelemahan utamanya adalah waktu dan sumber daya yang dibutuhkan, terutama jika jumlah hyperparameter dan variasi nilainya sangat besar, sehingga jumlah kombinasi yang harus diuji bisa sangat banyak. Selain itu, *Grid Search* juga dapat berisiko menyebabkan *overfitting* pada data pelatihan jika tidak hati-hati dalam pemilihan data validasi (Ivan & Prasetyo, 2023).

2.5. Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan merangkum prediksi model terhadap hasil aktual. Setiap baris menunjukkan kelas yang diprediksi, sedangkan setiap kolom menunjukkan kelas yang sebenarnya (Rawal & Agarwal, 2019). Tabel 1 menunjukkan metode uji *confusion matrix* yang digunakan untuk menentukan evaluasi *metrics.*

Table 1. *Confusion Matrix*

Prediksi	Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Tabel 2. menggambarkan bagaimana setiap elemen dalam *confusion matrix* digunakan untuk menghitung metrik evaluasi performa model klasifikasi seperti akurasi, presisi, recall, dan F1-score.

Table 2. Metrik Evaluasi

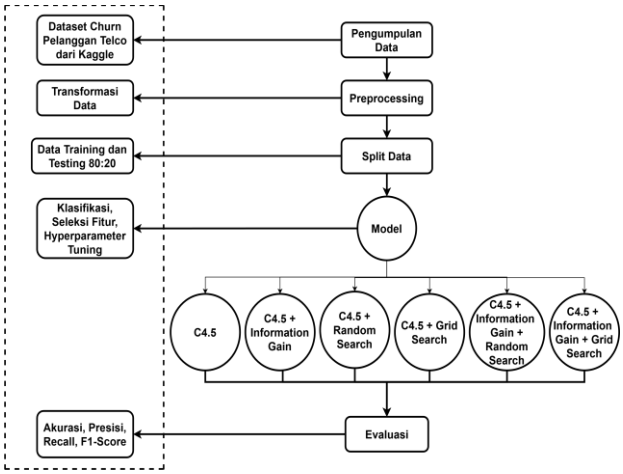
Matrix	Rumus
Akurasi	$\frac{TP + TN}{TP + TN + FP + FN} \; x \; 100\%$
Precision	$\frac{TP}{TP + FP} \; x \; 100\%$
Reccal	$\frac{TP}{TP + FN} \; x \; 100\%$

3. Metode

Penelitian ini bertujuan untuk meningkatkan prediksi *churn* pelanggan telekomunikasi dengan menggunakan seleksi fitur berdasarkan *Information Gain* dan optimasi *hyperparameter* melalui Random *Search* dan *Grid Search* pada Algoritma C4.5.

http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

Gambar 1. di bawah ini menunjukkan diagram alir dari metodologi penelitian yang digunakan. Diagram ini menggambarkan langkah-langkah yang dilakukan mulai dari pengumpulan data, *preprocessing*, pembagian data pemodelan, hingga evaluasi model.



Gambar 1. Diagram Alir Penelitian

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini diambil dari dataset Telco *Customer Churn* yang tersedia di Kaggle, yang dapat diunduh melalui tautan ini https://www.kaggle.com/datasets/blastchar/telco-customer-churn. Dataset ini mencakup informasi mengenai pelanggan telekomunikasi, termasuk berbagai variabel yang berpotensi memengaruhi churn, seperti durasi layanan, penggunaan data, dan kepuasan pelanggan.

3.2. Preprocessing Data

Tahap *preprocessing* melibatkan beberapa langkah penting untuk menyiapkan data agar siap untuk analisis. Langkah-langkah ini mencakup:

1. Pembersihan Data: Menghapus entri yang tidak lengkap atau duplikat.
2. Transformasi Data: Melakukan *encoding* pada variabel kategori dan normalisasi pada variabel numerik untuk memastikan konsistensi.

3.3. Pembagian Data

Data yang telah diproses dibagi menjadi dua set: data pelatihan dan data pengujian. Pembagian dilakukan dengan proporsi 80:20, di mana 80% dari data digunakan untuk melatih model dan 20% untuk menguji kinerja model. Hal ini bertujuan untuk memastikan bahwa model dapat digeneralisasi dengan baik terhadap data yang belum pernah dilihat sebelumnya.

3.4. Pemodelan

Pada tahap pemodelan, algoritma C4.5 diterapkan untuk klasifikasi, diikuti oleh seleksi fitur menggunakan *Information Gain* dan optimasi *hyperparameter* dengan *Random Search dan Grid Search*. Kombinasi teknik ini diterapkan untuk memaksimalkan kinerja model dalam memprediksi churn pelanggan dan dievaluasi menggunakan data pengujian.

3.5. Evaluasi

Setelah model dibangun, tahap evaluasi dilakukan untuk mengukur kinerja model menggunakan *confusion matrix* dan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

4. Hasil dan Pembahasan

4.1 Hasil

4.1.1 Dataset

Data dalam penelitian ini diambil dari dataset Telco *Customer Churn* di Kaggle, terdiri dari 7.043 baris dan 21 atribut, dengan 20 atribut sebagai fitur dan 1 atribut tujuan (*Churn*). Fitur mencakup informasi demografis, layanan, serta rincian kontrak, pembayaran, dan tagihan yang dapat memengaruhi *churn*. Dataset ini umum digunakan untuk analisis perilaku pelanggan dan model prediksi, sebagaimana ditunjukkan pada Gambar 2.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes

5 rows × 21 columns

Gambar 2. Dataset

4.1.2 Tahap Preprocessing Data

Tahap prapemrosesan melibatkan pembersihan data, yaitu menghapus entri yang tidak lengkap atau duplikat, serta transformasi data dengan *encoding* variabel kategori dan normalisasi variabel numerik untuk menjaga konsistensi. Hasil dari tahap preprocessing dapat dilihat pada Gambar 3.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure
0	0	0	0	0	0	1
1	1	1	0	1	0	34
2	2	1	0	1	0	2
3	3	1	0	1	0	45
4	4	0	0	1	0	2
...	...	...	...	...	...	...
7038	7038	1	0	0	1	24
7039	7039	0	0	0	1	72
7040	7040	0	0	0	1	11
7041	7041	1	1	0	0	4
7042	7042	1	0	1	0	66

7043 rows × 21 columns

Gambar 3. Dataset Hasil Preprocessing

Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587 On-line:
http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

4.1.3 Tahap Pembagian Data

Dalam penelitian ini, variabel yang diprediksi adalah Churn, sementara variabel lainnya berfungsi sebagai fitur. Dataset dibagi menjadi data pelatihan (80%) dan data pengujian (20%) menggunakan train_test_split, menghasilkan 5.634 baris untuk pelatihan dan 1.409 baris untuk pengujian, guna memastikan generalisasi model yang baik.

4.1.4 Tahap Pemodelan

Pada tahap pemodelan, beberapa pendekatan diterapkan untuk membangun model klasifikasi menggunakan algoritma C4.5, sebagai berikut:

1. *Implementasi Algoritma C4.5*

Model *Decision Tree* C4.5 dievaluasi tanpa optimisasi hyperparameter menggunakan accuracy_score() untuk akurasi dan classification_report() untuk presisi, recall, dan F1-score. Hasil evaluasi disajikan dalam Tabel 3.

Tabel 3. Hasil Pengujian Menggunakan C4.5

Akurasi: 74,09%			
Kelas	Precision	Recall	F1-Score
0	83%	81%	82%
1	51%	55%	53%

2. *Implementasi Algoritma C4.5 dengan Information Gain*

Data dibagi menjadi fitur dan label, dengan lima fitur teratas dipilih menggunakan SelectKBest berbasis *Information Gain.* Model dilatih tanpa optimasi *hyperparameter* dan dievaluasi berdasarkan akurasi serta laporan klasifikasi, yang ditunjukkan di Tabel 4.

Tabel 4. Hasil Pengujian Menggunakan C4.5 dengan *Information Gain*

Akurasi: 78,41%			
Kelas	Precision	Recall	F1-Score
0	85%	86%	85%
1	60%	56%	58%

3. *Implementasi Algoritma C4.5 dengan Random Search*

Hyperparameter dioptimalkan menggunakan entropi. *RandomizedSearchCV* digunakan untuk mencari kombinasi terbaik dari kedalaman maksimum pohon, jumlah minimum sampel untuk membagi node, dan jumlah minimum sampel per daun. Hasil evaluasi disajikan di Tabel 5.

Tabel 5. Hasil Pengujian Menggunakan C4.5 dengan Random *Search*

Akurasi: 80,05%			
Kelas	Precision	Recall	F1-Score
0	86%	86%	86%
1	62%	62%	62%

4. *Implementasi Algoritma C4.5 dengan Grid Search*

Model diterapkan menggunakan *DecisionTreeClassifier* dengan kriteria entropi. Hyperparameter dioptimasi melalui *GridSearchCV*, dan hasil evaluasi disajikan di Tabel 6.

Tabel 6. Hasil Pengujian Menggunakan C4.5 dengan *Grid Search*

Akurasi: 77,71%			
Kelas	Precision	Recall	F1-Score
0	85%	85%	85%
1	58%	57%	58%

5. *Implementasi Algoritma C4.5 dengan Information Gain dan Random Search*

Algoritma C4.5 menggunakan *Information Gain* untuk pemilihan fitur, dioptimalkan dengan Random Search, dan dievaluasi pada data pengujian. Hasilnya ditampilkan di Tabel 7.

Tabel 7. Hasil Pengujian Menggunakan C4.5 dengan *Information Gain* dan *Random Search*

Akurasi: 78,99%			
Kelas	Precision	Recall	F1-Score
0	81%	93%	87%
1	68%	39%	50%

6. *Implementasi Algoritma C4.5 dengan Information Gain dan Grid Search*

Implementasi mencakup persiapan data, pemilihan fitur dengan *SelectKBest*, pembangunan model *Decision Tree*, optimasi *hyperparameter* melalui *GridSearchCV*, dan penyusunan laporan klasifikasi yang ditampilkan di Tabel 8.

Tabel 8. Hasil Pengujian Menggunakan C4.5 dengan *Information Gain* dan *Grid Search*

Akurasi: 78,85%			
Kelas	Precision	Recall	F1-Score
0	83%	89%	86%
1	63%	50%	56%

4.1.5 Tahap Evaluasi

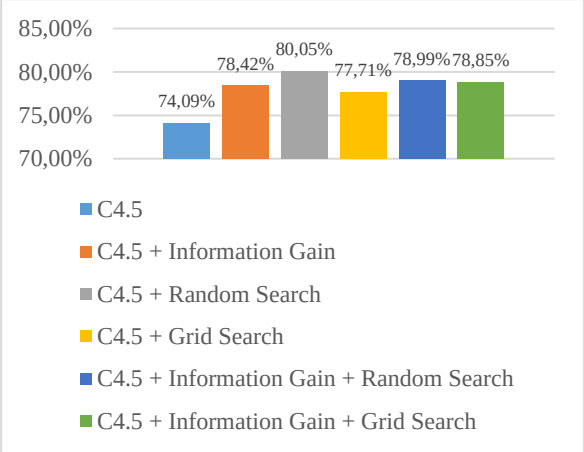
Setelah model dibangun, evaluasi dilakukan menggunakan *confusion matrix* dan metrik seperti akurasi, presisi, recall, dan F1-score, menunjukkan perbandingan kinerja serta peningkatan dari optimasi dan seleksi fitur.

Tabel 9. Hasil Perbandingan

Model	Kelas	Akurasi	Presisi	Recall	F-Score
C4.5	0	74,09%	83%	81%	82%
	1		51%	55%	53%
C4.5 dan Information Gain	0	78,42%	85%	86%	85%

Copyright ©2025, JSINBIS, p-ISSN: 2502-2377, e-ISSN: 2088-3587 On-line:
http://ejournal.undip.ac.id/index.php/jsinbis/article/view/63397

Model	Kelas	Akurasi	Presisi	Recall	F-Score
C4.5 dan Random Search	1	80,05%	60%	56%	58%
	0		86%	86%	86%
C4.5 dan Grid Search	1	77,71%	62%	62%	62%
	0		85%	85%	85%
C4.5 Berbasis Information Gain dan Random Search	1	78,99%	58%	57%	58%
	0		81%	93%	87%
C4.5 Berbasis Information Gain dan Grid Search	1	78,85%	68%	39%	50%
	0		83%	89%	96%
	1		63%	50%	56%



Gambar 4. Hasil Perbandingan

4.2. Pembahasan

Penelitian ini menunjukkan bahwa penerapan metode C4.5 pada data pelatihan menghasilkan akurasi awal sebesar 74,09%. Setelah seleksi fitur dengan *Information Gain*, akurasi meningkat menjadi 78,42%. Optimisasi hyperparameter dengan Random *Search* menghasilkan akurasi tertinggi 80,05%, lebih baik daripada Grid Search yang mencapai 77,71%, menunjukkan efektivitas Random *Search* dalam meningkatkan kinerja model.

Penggunaan *Grid Search* dalam penelitian ini, baik dengan dan tanpa seleksi fitur *Information Gain*, menunjukkan hasil akurasi yang sedikit lebih rendah daripada *Random Search.* Saat digunakan bersama *Information Gain*, akurasi *Grid Search* mencapai 78,85%, sementara Random Search mencapai 78,99%. Hasil ini menekankan bahwa kombinasi *Information Gain* dengan *Random Search* lebih optimal dalam meningkatkan akurasi model prediksi churn pelanggan dibandingkan dengan Grid Search.

Hal ini sejalan dengan penelitian terdahulu Reddy & Chittineni (2021) yang juga menemukan efektivitas Information Gain dalam meningkatkan akurasi, tetapi dengan teknik optimisasi berbeda. Temuan ini menunjukkan bahwa penggabungan *Information Gain* dengan *Random Search* berpotensi menjadi pendekatan yang efisien dan akurat untuk pemodelan churn pelanggan, walaupun penelitian lanjutan tetap diperlukan untuk mencapai hasil yang lebih optimal.

5. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa penerapan algoritma C4.5 pada data pelatihan awal memberikan akurasi sebesar 74,09%. Melalui seleksi fitur menggunakan *Information Gain*, akurasi model meningkat signifikan menjadi 78,42%, menegaskan pentingnya pemilihan fitur yang relevan dalam meningkatkan kinerja prediksi churn pelanggan. Optimisasi *hyperparameter* menggunakan Random *Search* mencapai akurasi tertinggi, yaitu 80,05%, dibandingkan dengan *Grid Search* yang mencapai akurasi 77,71%. Hasil ini menunjukkan bahwa Random *Search* lebih efektif dalam menemukan kombinasi *hyperparameter* yang optimal untuk model C4.5.

Implikasi dari temuan ini adalah bahwa penerapan teknik seleksi fitur dan optimisasi *hyperparameter* secara tepat dapat secara signifikan meningkatkan kinerja model prediksi. Namun, penelitian ini memiliki keterbatasan, termasuk keterbatasan pada dataset yang digunakan dan metode yang diuji. Oleh karena itu, penelitian lebih lanjut diperlukan untuk menguji teknik ini pada berbagai dataset contoh prediksi penyakit jantung dan dengan berbagai algoritma pembelajaran mesin lainnya.

Saran untuk penelitian selanjutnya adalah mengganti dataset *customer churn* dengan dataset dari bidang kesehatan untuk memprediksi perawatan lanjutan pasien. Selain itu, eksplorasi metode lain seperti *ensemble learning*, *deep learning*, dan metode non-parametrik dapat dilakukan untuk meningkatkan performa prediksi churn dan memahami faktor-faktor yang memengaruhi fenomena tersebut.

Daftar Pustaka

Akbar, M. N. (2021). Klasifikasi Bibliografi Otomatis Menggunakan C4.5 Dan Information Gain. *Jurnal Instek (Informatika Sains Dan Teknologi)*, *6*(1). https://doi.org/10.24252/instek.v6i1.18636

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer Churn Prediction In Telecommunication Industry Under Uncertain Situation. *Journal Of Business Research*, *94*.

Anita, Wicaksono, A., & Padilah, T. N. (2021). Pengaruh Jumlah Record Dataset Terhadap Algoritma Klasifikasi Berdasarkan Data Customer Churn. *Jurnal Ilmiah Informatika*, *6*(1). https://doi.org/10.1016/j.jbusres.2018.03.003

Arden, F., & Safitri, C. (2022). Hyperparameter Tuning Algorithm Comparison With Machine Learning Algorithms. *Proceeding - 6th International Conference On Information Technology, Information Systems And Electrical Engineering: Applying Data Sciences And Artificial Intelligence Technologies For Environmental Sustainability,*

Icitisee 2022. http://doi.org/10.1109/ICITISEE57756.2022.10057630

Damanik, S. D., & Jambak, M. I. (2023). Klasifikasi Customer Churn Pada Telekomunikasi Industri Untuk Retensi Pelanggan Menggunakan Algoritma C4.5. *Klik: Kajian Ilmiah Informatika Dan Komputer*, *3*(6). https://doi.org/10.30865/klik.v3i6.829

Fajri, M., & Primajaya, A. (2023). Komparasi Teknik Hyperparameter Optimization Pada Svm Untuk Permasalahan Klasifikasi Dengan Menggunakan Grid Search Dan Random Search. *Journal Of Applied Informatics And Computing*, *7*(1).

Febiani, N., Fauzan, Abd. C., & Huda, M. M. (2022). Implementasi Algoritma Decision Tree C4.5 Dengan Improvisasi Mean Dan Median Pada Dataset Numerik. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *5*(1). https://doi.org/10.37600/tekinkom.v5i1.435

Ivan, J., & Prasetyo, S. Y. (2023). Heart Disease Prediction Using Ensemble Model And Hyperparameter Optimization. *International Journal On Recent And Innovation Trends In Computing And Communication*, *11*, 290–295. http://doi.org/10.17762/ijritcc.v11i8s.7208

Kubat, M. (2021). An Introduction To Machine Learning. In *An Introduction To Machine Learning*. http://doi.org/10.1007/978-3-319-63913-0

Loukili, M., Messaoudi, F., & Ghazi, M. El. (2022). Supervised Learning Algorithms For Predicting Customer Churn With Hyperparameter Optimization. *International Journal Of Advances In Soft Computing And Its Applications*, *14*(3). http://doi.org/10.15849/IJASCA.221128.04

Nasrullah, A. H. (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. *Jurnal Ilmiah Ilmu Komputer*, *7*(2). https://doi.org/10.35329/jiik.v7i2.203

Navon, D., & Bronstein, A. M. (2022). *Random Search Hyper-Parameter Tuning: Expected Improvement Estimation And The Corresponding Lower Bound*. http://doi.org/10.48550/arXiv.2208.08170

Puneeth, B. R., Preethi Salian, K., Prathyakshini, Murthy, A., Salian, S., & Surabhi. (2022). Analysis Of Telecom Churn Using Machine Learning Techniques. *International Conference On Artificial Intelligence And Data Engineering, Aide 2022*. http://doi.org/10.1109/AIDE57180.2022.10060222

Ramanda Hasibuan, M., & Marji. (2019). Pemilihan Fitur Dengan Information Gain Untuk Klasifikasi Penyakit Gagal Ginjal Menggunakan Metode Modified K-Nearest Neighbor (Mknn). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(11).

Rawal, B., & Agarwal, R. (2019). Improving Accuracy Of Classification Based On C4.5 Decision Tree Algorithm Using Big Data Analytics. *Advances In Intelligent Systems And Computing*, *711*. https://doi.org/10.1007/978-981-10-8055-5_19

Reddy, G. S., & Chittineni, S. (2021). Entropy Based C4.5-Sho Algorithm With Information Gain Optimization In Data Mining. *Peerj Computer Science*, *7*. https://doi.org/10.7717/peerj-cs.424

Rizky, M. H., Faisal, M. R., Budiman, I., Kartini, D., & Abadi, F. (2024). Effect Of Hyperparameter Tuning Using Random Search On Tree-Based Classification Algorithm For Software Defect Prediction. *Ijccs (Indonesian Journal Of Computing And Cybernetics Systems)*, *18*(1), 95. https://doi.org/10.22146/ijccs.90437

Sadiq, M. H., & Ahmed, N. S. (2019). Classifying And Predicting Students' Performance Using Improved Decision Tree C4.5 In Higher Education Institutes. *Journal Of Computer Science*, *7*(12). http://doi.org/10.3844/jcssp.2019.1291.1306

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A Novel Customer Churn Prediction Model For The Telecommunication Industry Using Data Transformation Methods And Feature Selection. *Plos One*, *17*(12 December). https://doi.org/10.1371/journal.pone.0278095

Saputra, R. A., Wasiyanti, S., & Pribadi, D. (2021). Information Gain Pada Algoritma C4.5 Untuk Klasifikasi Penerimaan Bantuan Pangan Non Tunai (Bpnt). *Indonesian Journal Of Business Intelligence (Ijubi)*, *4*(1). http://doi.org/10.21927/ijubi.v4i1.1757

Sari, R. P., Febriyanto, F., & Adi, A. C. (2023). Analysis Implementation Of The Ensemble Algorithm In Predicting Customer Churn In Telco Data: A Comparative Study. *Informatica (Slovenia)*, *47*(7). https://doi.org/10.31449/inf.v47i7.4797

Utami, Yohana T., Shofiana, D. A., & Heningtyas, Y. (2020). Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi. *Jurnal Komputasi*, *8*(2). http://doi.org/10.23960/komputasi.v8i2.2647

Wardani, N. W., Dantes, G. R., & Indrawan, G. (2018). Prediksi Customer Churn Dengan Algoritma Decision Tree C4.5 Berdasarkan Segmentasi Pelanggan Untuk Mempertahankan Pelanggan Pada Perusahaan Retail. *Jurnal Resistor (Rekayasa Sistem Komputer)*, *1*(1). http://doi.org/10.31598/jurnalresistor.v1i1.219

Yoga Siswa, T. A., Putra, G. M., & Prafanto, A. (2022). Seleksi Fitur Information Gain Dan Teknik Pruning Untuk Memperbaiki Akurasi Algoritma C4.5 Dalam Kasus Keterlambatan Biaya Kuliah. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, *17*(2). http://doi.org/10.30872/jim.v17i2.11794