Alghoritm : Jurnal Teknologi Informasi (AJTI), Vol. 1, No. 1, Juni 2025
ISSN: xxxx-xxx

Implementasi Model Prediksi *Churn* Pelanggan Menggunakan Algoritma *Random Forest* Pada Website Industri Telekomunikasi

Holilurrahman [1*]; Mohamad Imron [2]
Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Annuqayah
rahmanholilur11@gmail.com[1*;] rohimabd708@gmail.com [2,;]
* Penulis Korespondensi

Abstrak

Industri telekomunikasi di Indonesia berkembang pesat berkat peningkatan jumlah pengguna layanan internet dan telekomunikasi, yang memicu persaingan sengit di antara penyedia layanan. Salah satu tantangan utama yang dihadapi oleh perusahaan telekomunikasi adalah churn pelanggan, yang dapat mengakibatkan penurunan pendapatan yang signifikan. Penelitian ini bertujuan untuk meningkatkan efektivitas strategi retensi pelanggan. Dengan membangun model prediksi churn pelanggan dan mengintegrasikannya ke dalam sebuah website industri telekomunikasi, penelitian ini bertujuan untuk membantu perusahaan mengidentifikasi pelanggan yang berpotensi melakukan churn dan mengambil tindakan pencegahan yang tepat. Metode penelitian ini menggunakan pendekatan AI Project Cycle yang terdiri dari tahap problem scoping, data acquisition, data exploration, modelling, evaluation, dan deployment. Dengan fokus pada elemen-elemen kunci yang memiliki dampak signifikan terhadap tingkat churn pelanggan dalam konteks industri telekomunikasi Indonesia, penelitian ini diharapkan dapat memberikan dampak positif yang signifikan bagi masyarakat, pengembangan ilmu pengetahuan dan teknologi, serta peningkatan nilai ekonomi. Dalam penelitian ini, akurasi prediksi menggunakan algoritma Random Forest mencapai 95%, sedangkan penggunaan Decision Tree menunjukkan akurasi sebesar 92%. Hal ini menunjukkan bahwa Random Forest memberikan peningkatan yang signifikan dalam kinerja prediksi dibandingkan dengan pendekatan Decision Tree. Hasil ini juga lebih tinggi dibandingkan dengan penelitian sebelumnya yang menggunakan beberapa algoritma seperti Decision Tree, SVM, KNN, Logistic Regression, C4.5, AdaBoost, XGBoost.
Kata Kunci: *Churn pelanggan; Model prediksi; Algoritma Random Forest; Strategi retensi pelanggan; AI Project Cycle.*

Abstract

The telecommunications industry in Indonesia is growing rapidly due to the increasing number of internet and telecommunications service users, which has triggered fierce competition among service providers. One of the main challenges faced by telecommunications companies is customer churn, which can result in a significant decrease in revenue. This study aims to improve the effectiveness of customer retention strategies. By building a customer churn prediction model and integrating it into a telecommunications industry website, this study aims to help companies identify customers who are likely to churn and take appropriate preventive measures. This research method uses the AI Project Cycle approach consisting of problem scoping, data acquisition, data exploration, modeling, evaluation, and deployment stages. By focusing on key elements that have a significant impact on customer churn rates in the context of the Indonesian telecommunications industry, this study is expected to provide a significant positive impact on society, the development of science and technology, and increasing economic value. In this study, the prediction accuracy using the Random Forest algorithm reached 95%, while the use of Decision Tree showed an accuracy of 92%. This shows that Random Forest provides a significant improvement in prediction performance compared to the Decision Tree approach. These results are also higher compared to previous studies that used several algorithms such as Decision Tree, SVM, KNN, Logistic Regression, C4.5, AdaBoost, XGBoost.
Keywords: *Customer churn; Prediction models; Random Forest algorithm; Customer retention strategies; AI Project Cycle.*

PENDAHULUAN

Kemajuan industri telekomunikasi di Indonesia berkembang sejalan dengan pertumbuhan jumlah individu yang menggunakan layanan internet dan telekomunikasi. Saat ini, berbagai perusahaan di Indonesia

menawarkan layanan fixed broadband, termasuk Telkom dengan Indihome, XL Axiata dengan XL Home, MNC dengan MNC Play, First Media Tbk dengan First Media, MyRepublic Limited dengan MyRepublic, dan lainnya [1]. Untuk bertahan dalam persaingan yang sengit, perusahaan telekomunikasi harus terus berinovasi. Salah satu upaya inovatif yang mereka lakukan adalah dengan mengembangkan teknologi data mining dan model berbasis *machine learning* untuk analisis, prediksi, dan manajemen penurunan pelanggan [2]. Menurut pengumuman terbaru dari Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia pada tahun 2024 mencapai 221.563.479 jiwa dari total populasi 278.696.200 jiwa penduduk Indonesia tahun 2023. Dari hasil survei penetrasi internet Indonesia 2024 yang dirilis oleh APJII, tingkat penetrasi internet Indonesia mencapai angka 79,5%, yang menandakan peningkatan sebesar 1,4% dibandingkan dengan periode sebelumnya. Hal ini menunjukkan bahwa semakin banyak penduduk Indonesia yang terhubung dengan internet, mencerminkan pertumbuhan yang positif dalam adopsi teknologi di negara ini.

Dalam persaingan tersebut, setiap operator tidak hanya memperhatikan perkembangan produk dan layanan mereka, tetapi juga memberikan fokus yang besar pada kebutuhan pelanggan. Hal ini dikarenakan pelanggan memiliki kebebasan untuk memilih operator yang sesuai dengan preferensi mereka, dan dapat melakukan peralihan (*churn*) kapan saja. *Customer churn* adalah fenomena di mana pelanggan beralih dari satu provider ke provider lain, yang dapat mengakibatkan penurunan pendapatan yang signifikan bagi perusahaan telekomunikasi [3].

Dengan melakukan prediksi terhadap pelanggan yang berpotensi beralih (*churn*), perusahaan memiliki kesempatan untuk mengambil tindakan yang efektif dalam mempertahankan mereka. Untuk mempertahankan pelanggan yang ada, perusahaan harus meningkatkan layanan pelanggan, meningkatkan kualitas produk, dan memiliki pemahaman terhadap pelanggan yang berisiko beralih sebelum hal itu terjadi. Prediksi ini bisa dilakukan dengan menganalisis data pelanggan menggunakan teknik data mining. Dengan mengumpulkan informasi bisnis dari industri telekomunikasi, perusahaan dapat memperoleh wawasan yang berharga dalam memprediksi apakah pelanggan akan tetap setia atau beralih ke penyedia layanan lain [4].

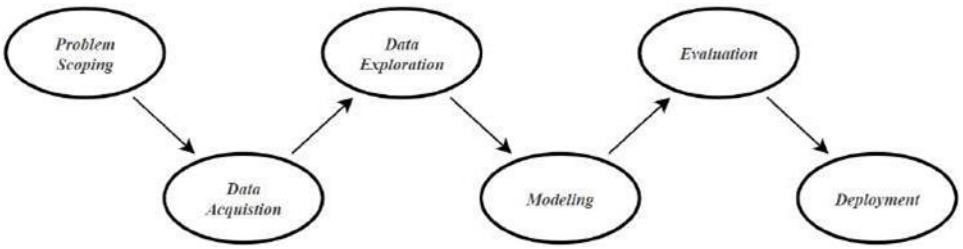
Pada penelitian [5] bertujuan untuk membangun model prediksi churn pelanggan di industri telekomunikasi. Penelitian ini menggunakan metode logistic regression dan correlation-based feature selection, menghasilkan akurasi sebesar 85,75% dengan memanfaatkan correlation-based feature selection dan forward selection. Penelitian selanjutnya yang dilakukan oleh Adhy Mauludin Nur Aziz pada tahun 2023 membangun model klasifikasi yang lebih baik dalam hal akurasi untuk memprediksi churn pelanggan pada perusahaan telekomunikasi. Metode yang digunakan adalah logistic regression dan decision tree, dengan hasil akurasi 80% untuk logistic regression dan 72% untuk decision tree [6]. Pada penelitian Mohammad Amirulhaq Iskandar dan Ulinnuha Latifa dengan judul "Website Prediksi Customer Churn untuk Mempertahankan Pelanggan pada Perusahaan Telekomunikasi." Penelitian ini bertujuan untuk membangun model prediksi churn pelanggan guna mempertahankan pelanggan di perusahaan telekomunikasi. Metode yang digunakan adalah machine learning dengan algoritma K-Nearest Neighbors (KNN), menghasilkan akurasi sebesar 81% [7]. Untuk penelitian yang dilakukan oleh Iqbal Muhammad Latief, Agus Subekti, dan Windu Gata pada tahun 2021 berjudul "Prediksi Tingkat Pelanggan Churn pada Perusahaan Telekomunikasi dengan Algoritma Adaboost." Tujuannya adalah

membangun model prediksi churn pelanggan di perusahaan telekomunikasi. Metode yang digunakan mencakup beberapa teknik klasifikasi data mining seperti random forest dan adaboost, menghasilkan akurasi tertinggi sebesar 80% menggunakan algoritma adaboost [4]. Dan pada penelitian Selfia Hafidatus Sholeha, Mochammad Faid, Moh. Ainol Yaqin Selfia Hafidatus Sholeha, Mochammad Faid, dan Moh. Ainol Yaqin melakukan penelitian berjudul "Prediksi Perpindahan Pelanggan pada Toko Online Menggunakan Metode Tree-Based Gradient Boosted Models." Penelitian ini bertujuan membangun model prediksi churn pelanggan menggunakan metode tree-based gradient boosted models, khususnya model XGBoost, LightGBM, dan CatBoost. Hasil penelitian menunjukkan akurasi sebesar 80,032% dengan model XGBoost [8].

Dalam mengatasi permasalahan yang telah disebutkan, penelitian ini akan mengusulkan prediksi churn pelanggan di industri telekomunikasi menggunakan algoritma *Random Forest* yang dapat meningkatkan hasil akurasi dari penelitian-penelitian sebelumnya. Pemilihan algoritma *Random Forest* didasarkan pada kemampuannya dalam mengklasifikasi data yang tidak lengkap serta mampu menangani jumlah sampel data yang besar dengan baik [9]. *Random forest* adalah sebuah algoritma yang terdiri dari sejumlah besar pohon keputusan. Algoritma ini melakukan prediksi dengan cara melakukan voting terhadap kelas dari setiap pohon, dan kelas dengan jumlah voting terbanyak akan menjadi kelas akhir yang dipilih. Hal ini merupakan salah satu keunggulan dari algoritma *Random forest* dalam melakukan prediksi [10]. Oleh karena itu, diharapkan penelitian ini dapat memberikan sumbangan yang berarti dalam meningkatkan kapasitas perusahaan telekomunikasi dalam meramal perilaku churn pelanggan, serta mengambil langkah-langkah pencegahan yang sesuai.

METODE PENELITIAN

Peneliti ini menggunakan metode *AI project cycle* dalam pengembangan prediksi churn pelanggan di industri telekomunikasi karena metode ini memberikan pendekatan sistematis dan efektif dalam mengelola keseluruhan proses pengembangan solusi berbasis kecerdasan buatan. Penjabaran tentang tahapan proses penelitian yang akan dilakukan oleh penulis dapat dijelaskan dibawah ini :



Gambar.1 *AI Project Cycle*

1. *Problem Scoping*

Problem scoping adalah langkah awal dalam siklus proyek *AI* yang bertujuan untuk mengidentifikasi dan menetapkan batasan masalah. Langkah ini dimaksudkan untuk mengklarifikasi dan mengarahkan objek

penelitian. Dalam proses problem scoping, metode 4W digunakan, yang terdiri dari who, where, what, dan why. Metode ini membantu untuk mengidentifikasi siapa yang terlibat, di mana masalah tersebut terjadi, apa masalahnya beserta faktor pendukungnya, dan mengapa masalah tersebut perlu diatasi serta manfaat apa yang akan diperoleh dari penyelesaiannya [11]. Terdapat metode 4W yang digunakan untuk memfasilitasi proses problem scoping, yaitu:

a. Who: Siapa yang terlibat dalam masalah tersebut?
b. What: Apa masalah dan faktor-faktor pendukung masalah?
c. Where: Dimana kondisi, situasi, atau lokasi masalah diamati?
d. Why: Alasan mengapa masalah tersebut perlu diatasi dan manfaat apa yang bisa diperoleh dari penyelesaiannya?

2. *Data Acquisition*

Data Acquisition adalah langkah awal dalam mengumpulkan data yang diperlukan untuk proyek kecerdasan buatan (AI). Ini merupakan fondasi atau bahan mentah yang kemudian akan diolah dan dianalisis sesuai dengan permasalahan yang diteliti, sehingga dapat menghasilkan solusi terbaik. Ada beberapa metode untuk mendapatkan sumber data tersebut, antara lain [12]

a. Tools/Alat : Laptop.
b. Observasi : Penelitian.
c. Open Data : *Kaggle*

Penelitian ini menggunakan data publik yang diperoleh dari situs web www.kaggle.com. Data tersebut terdiri dari berbagai variabel yang relevan dengan topik penelitian. Variabel yang digunakan dalam analisis meliputi:

a. Status pelanggan televisi (is_tv_subscriber): Menunjukkan apakah pengguna merupakan pelanggan televisi atau bukan.
b. Status pelanggan paket film (is_movie_package_subscriber): Indikator apakah pengguna berlangganan paket film atau tidak.
c. Usia langganan (subscription_age): Durasi dalam waktu yang telah pengguna menjadi pelanggan.
d. Rata-rata tagihan bulanan (bill_avg): Jumlah rata-rata tagihan bulanan pengguna.
e. Sisa kontrak (remaining_contract): Waktu yang tersisa dalam kontrak langganan.
f. Jumlah kegagalan layanan (service_failure_count): Total kegagalan layanan yang dialami oleh pelanggan.
g. Rata-rata kecepatan unduh (download_avg): Rata-rata kecepatan unduh yang diterima oleh pelanggan.
h. Rata-rata kecepatan unggah (upload_avg): Rata-rata kecepatan unggah yang diterima oleh pelanggan.
i. Unduhan melebihi batas (download_over_limit): Indikator apakah pelanggan sering melebihi batas unduhan atau tidak.

3. *Data Exploration*

Data exploration, yang diperkenalkan oleh John Tukey, adalah langkah lanjutan setelah proses data acquisition. Pada tahap ini, data diatur dan dianalisis secara rinci dengan pemahaman karakteristik yang didapat melalui proses preprocessing data yang telah diperoleh sebelumnya [13].

a. Summary Descriptive Statistics

Statistika deskriptif adalah metode analisis statistik yang umum digunakan untuk merangkum dan menyajikan data. Biasanya, statistika deskriptif digunakan sebagai langkah awal untuk memahami dan merapikan data sebelum dilakukan analisis lebih lanjut. Namun, statistika deskriptif juga bisa menjadi analisis independen yang memberikan berbagai informasi tentang distribusi dan karakteristik data [14].

b. *Data Cleaning*

Data cleansing merupakan proses analisis terhadap kualitas data dengan melakukan modifikasi, perubahan, atau penghapusan data yang dianggap tidak diperlukan, tidak lengkap, tidak akurat, atau memiliki format yang salah dalam basis data. Tujuan dari proses tersebut adalah untuk menghasilkan data yang berkualitas tinggi yang dapat dipercaya untuk digunakan dalam analisis dan pengambilan keputusan yang akurat [15]. Pada tahap awal eksplorasi data, langkah pertama adalah mengidentifikasi dan menangani data duplikat. Duplikasi adalah proses menciptakan salinan yang identik dengan aslinya. Menyalin, pada dasarnya, mengacu pada tindakan membuat sesuatu menjadi serupa atau identik dengan yang sudah ada. Setelah itu, penanganan missing value. Missing values merujuk pada situasi di mana terdapat data yang tidak lengkap atau absen dalam suatu dataset. Kehadiran missing values adalah kejadian yang lazim dalam pengolahan data, yang dapat disebabkan oleh berbagai faktor seperti kegagalan perangkat, kesalahan perhitungan, ketidaktepatan pencatatan data, serta berbagai kendala teknis lainnya [16]. Langkah berikutnya adalah penanganan data tidak normal. Data tidak normal atau anomali data adalah kondisi di mana terjadi ketidaksesuaian atau ketidakberaturan dalam dataset [17]. Terakhir, dilakukan proses penghapusan *outlier*. *Outlier* adalah data yang terletak jauh dari pusat distribusi dan memiliki karakteristik yang berbeda dari data lainnya. Mereka bisa memiliki nilai yang sangat rendah atau sangat tinggi [18]. Salah satu cara untuk mengidentifikasi *outlier* adalah dengan menggunakan *Interquartile Range* (IQR), yang memanfaatkan nilai-nilai *quartile* (Q) [19].

c. Visualisasi Data

Visualisasi adalah metode pembelajaran yang memungkinkan konsep materi untuk dipahami secara visual melalui indera penglihatan. Sedangkan visualisasi data adalah istilah umum yang menggambarkan berbagai cara untuk membantu individu memahami makna data dengan cara menggambarkan data dalam bentuk visual yang relevan [20].

4. *Modelling*

Pada tahap pemodelan ini, akan digunakan algoritma *Random Forest* untuk melakukan pembangunan model. *Random Forest* adalah suatu algoritma dalam pembelajaran mesin yang canggih dan mudah diterapkan. Algoritma ini menjadi populer karena kepraktisan dan fleksibilitasnya, karena dapat digunakan untuk berbagai tugas seperti klasifikasi dan regresi [21]. Selain membangun model dengan algoritma *Random Forest*, tahap ini juga akan melibatkan perbandingan performa model dengan menggunakan algoritma Decision Tree merupakan

salah satu metode yang relatif mudah dipahami oleh manusia. Model ini menggunakan struktur pohon atau hierarki untuk melakukan prediksi. Ide dasarnya adalah mengubah data menjadi struktur pohon dan serangkaian aturan keputusan [22].

Dengan melakukan perbandingan performa antara model *Random Forest* dan *Decision Tree*, kita dapat menunjukkan serta memverifikasi keandalan model *Random Forest* dalam melakukan tugas klasifikasi. Hasil dari perbandingan ini akan memberikan wawasan yang berharga tentang keefektifan dan keunggulan dari penggunaan algoritma *Random Forest* dalam konteks spesifik tugas yang sedang dijalankan.

5. *Evaluation*

Evaluastion merupakan langkah selanjutnya dalam proses pembuatan model. Pada tahap ini, model yang telah dibuat diuji dan dianalisis untuk menilai kinerjanya terhadap masalah yang dihadapi serta dampak yang muncul dari hasil dan penggunaan model dan dataset. Berbagai metode evaluasi digunakan, termasuk menilai akurasi data setelah proses pelatihan, untuk menentukan apakah model tersebut sudah memadai atau masih perlu diperbaiki [23],

Dalam tahap ini, confusion matrix akan digunakan untuk mengevaluasi model, termasuk:

a. Akurasi: Persentase prediksi yang benar dari total pengamatan.

Rumus → (TP+TN)/(TP+FP+TN+FN)

b. Presisi: Persentase kasus yang diprediksi AI dan benar-benar terjadi.

Rumus → (TP)/(TP+FP)

c. *Recall:* Mengukur proporsi kasus yang terjadi dan diprediksi secara tepat oleh AI.

Rumus → (TP)/(TP+FN)

Evaluasi ini juga akan melibatkan penggunaan kurva ROC (*Receiver Operating Characteristic*). Kurva ROC adalah metode evaluasi penting untuk menilai kinerja sistem klasifikasi. Ini menggambarkan hubungan antara sensitivitas (*true positive rate*) dan spesifisitas (*false positive rate*), membantu dalam menilai kemampuan sistem dalam membedakan kelas positif dan negatif serta *trade-off* sensitivitas dan spesifisitas [24]. Dengan melibatkan metrik-metrik tersebut, evaluasi akan memberikan gambaran yang lebih lengkap tentang kinerja model, serta membantu menentukan apakah perlu adanya peningkatan atau perbaikan lebih lanjut. Validasi silang (*cross-validation*) juga akan digunakan untuk memastikan bahwa kinerja model konsisten dan tidak dipengaruhi oleh pembagian dataset yang spesifik. *Cross-validation* adalah teknik yang digunakan untuk mengevaluasi performa model yang telah dibuat dengan membagi data secara acak menjadi dua bagian: data training dan data testing [25].

6. *Deployment*

Deployment merupakan proses implementasi model pada sebuah aplikasi atau sistem. Tersedia berbagai platform penyedia untuk deployment atau hosting sebuah website, termasuk Heroku, Vercel, Github, dan Netlify [12].

HASIL PENELITIAN DAN PEMBAHASAN

Hasil penelitian ini diuraikan berdasarkan metode penelitian sebelumnya yang sudah dilakukan seperti berikut :

1. *Data Acquistion*

Untuk memahami perilaku *churn* dan membangun model prediksi yang akurat, diperlukan data yang komprehensif. Penelitian ini memanfaatkan data publik yang tersedia di platform Kaggle. Data yang didapat adalah dataset internet service provider customer churn berjumlah 72.274 baris data dan 11 kolom yang dipublikasikan oleh Mehmet Sabri Kunt tahun 2022, didalamnya terdapat beberapa data numerik seperti berikut:

	id	is_tv_subscriber	is_movie_package_subscriber	subscription_age	bill_avg	reamining_contract
0	15	1	0	11.95	25	0.14
1	18	0	0	8.22	0	NaN
2	23	1	0	8.91	16	0.00
3	27	0	0	6.87	21	NaN
4	34	0	0	6.39	0	NaN
...	...	...	...	...	...	...
72269	1689648	1	1	0.09	0	1.25
72270	1689676	1	0	0.06	1	1.63
72271	1689679	1	0	0.02	0	2.19
72272	1689733	0	0	0.01	0	0.72
72273	1689744	1	1	0.01	0	0.82

72274 rows × 11 columns

Gambar.2 Dataset

a. ID pelanggan: Kolom yang penting untuk identifikasi unik setiap pelanggan dalam dataset.
b. Status pelanggan televisi (*is_tv_subscriber*): Menunjukkan apakah pengguna merupakan pelanggan televisi atau bukan.
c. Status pelanggan paket film (*is_movie_package_subscriber*): Indikator apakah pengguna berlangganan paket film atau tidak.
d. Usia langganan (*subscription_age*): Durasi dalam waktu yang telah pengguna menjadi pelanggan.
e. Rata-rata tagihan bulanan (*bill_avg*): Jumlah rata-rata tagihan bulanan pengguna.
f. Sisa kontrak (*remaining_contract*): Waktu yang tersisa dalam kontrak langganan.
g. Jumlah kegagalan layanan (*service_failure_count*): Total kegagalan layanan yang dialami oleh pelanggan.
h. Rata-rata kecepatan unduh (*download_avg*): Rata-rata kecepatan unduh yang diterima oleh pelanggan.
i. Rata-rata kecepatan unggah (*upload_avg*): Rata-rata kecepatan unggah yang diterima oleh pelanggan.

j. Unduhan melebihi batas (*download_over_limit*): Indikator apakah pelanggan sering melebihi batas unduhan atau tidak.

k. Label *churn*: Variabel kunci yang menunjukkan apakah pelanggan telah melakukan churn (1) atau tidak (0).

2. *Data Exploration*

a) Cleaning Data

a Data Duplikat

Hasil dari dataset.duplicated().sum() menunjukkan bahwa tidak ada data duplikat dalam dataset, yang menunjukkan bahwa setiap entri dalam dataset adalah unik dan tidak ada duplikasi.

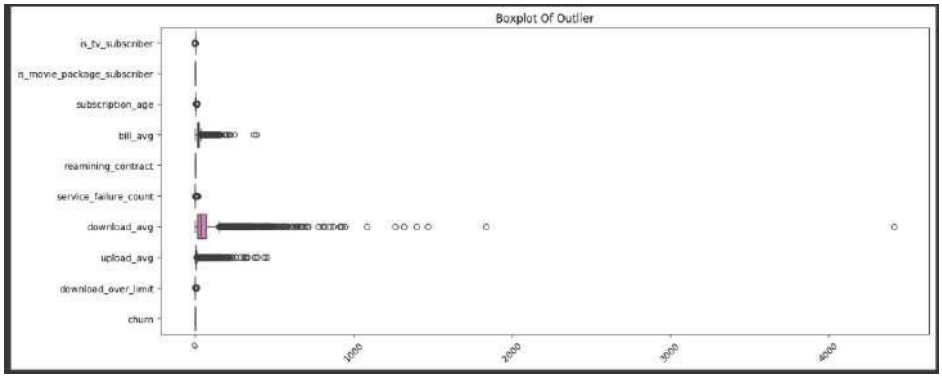
b *Missing Value*

Hasil dari dataset.isnull().sum() menunjukkan jumlah *missing value* pada tiap kolom dalam dataset. Kolom "*id*", "*is_tv_subscriber*", "*is_movie_package_subscriber*", "*subscription_age*", "*bill_avg*", "*service_failure_count*", "*download_over_limit*", dan "*churn*" tidak memiliki *missing value*, karena jumlahnya adalah 0. Namun, terdapat *missing value* pada kolom "*remaining_contract*", "*download_avg*", dan "*upload_avg*". Terdapat 21.572 missing value pada kolom "*remaining_contract*" (lama kontrak yang tersisa), 381 *missing value* pada kolom "*download_avg*" (rata-rata kecepatan unduh), dan juga 381 *missing value* pada kolom "*upload_avg*" (rata-rata kecepatan unggah).

c Data Tidak Normal

Setelah menganalisis dataset, ditemukan nilai yang tidak normal pada kolom subscriber age. Lebih tepatnya, hasil dari describe dataset menunjukkan adanya nilai -0,02, yang secara logika tidak masuk akal karena umur tidak dapat memiliki nilai negatif. Untuk menjaga integritas data, dilakukan penghapusan pada nilai yang tidak normal tersebut dari dataset. Dengan demikian, sisa dataset setelah tahap ini sekitar 50374.

d *Outliers*



Gambar.3 Cek *Outlier*

Dari hasil pengecekan *outlier* yang dilakukan, tampak bahwa kolom-kolom "*bill_avg*", "*download_avg*", dan "*upload_avg*" menjadi paling menonjol. Ini berarti bahwa dalam distribusi nilai- nilai dalam ketiga kolom tersebut, terdapat sejumlah observasi yang secara signifikan berbeda dari mayoritas data. Dengan kata lain, ada nilai-nilai yang jauh dari kisaran nilai yang diharapkan atau yang umum dalam dataset.

Penonjolan kolom-kolom ini dapat memiliki implikasi yang penting dalam analisis lebih lanjut. Adanya *outlier* dalam kolom "*bill_avg*" misalnya, mungkin menandakan adanya pola pengeluaran yang tidak biasa dari sebagian pengguna. Sementara itu, adanya *outlier* dalam kolom "*download_avg*" dan "*upload_avg*" dapat menunjukkan adanya pengguna dengan pola konsumsi bandwidth yang tidak biasa, yang dapat menjadi fokus penelitian lebih lanjut..

Setelah mengidentifikasi kolom "*bill_avg*", "*download_avg*", dan "*upload_avg*" sebagai yang paling menonjol dalam hal *outlier*, langkah selanjutnya adalah menangani *outlier* tersebut menggunakan metode IQR (*Interquartile Range*).

Setelah penanganan *outlier* dilakukan, pengecekan kembali *boxplot* menunjukkan penurunan jumlah *outlier* yang signifikan. Meskipun masih ada beberapa *outlier* yang tersisa, namun penanganan outlier telah berhasil mengurangi dampaknya pada ketiga kolom yang paling menonjol tersebut.

Outlier seringkali dianggap sebagai nilai yang tidak biasa atau mengganggu, namun mereka juga dapat memiliki nilai informatif dalam pemodelan data. *Outlier* dapat mewakili situasi atau kejadian yang tidak biasa tetapi signifikan, yang mungkin memiliki dampak penting terhadap hasil analisis atau prediksi.

Oleh karena itu, penting untuk mempertimbangkan apakah *outlier* harus dihapus sepenuhnya atau dipertahankan dalam analisis. Terkadang, mempertahankan *outlier* dan memperlakukannya secara khusus dalam pemodelan dapat meningkatkan pemahaman dan prediksi atas fenomena yang diamati. Setelah penanganan *outlier* ini menggunakan metode IQR dataset berkurang menjadi 44355.

b) Data Visualization

a. Cetak Jumlah *Churn*

```
Jumlah Churn (1): 18604
Jumlah Tidak Churn (0): 25751
Persentase Pelanggan Churn: 41.94%
```

Gambar.4 Cetak Jumlah *Churn*

3. *Modelling*

Dalam mengeksplorasi model-model prediktif untuk penelitian ini, peneliti telah menetapkan algoritma *Random Forest* sebagai fokus utama. Namun, keputusan ini tidak diambil secara tanpa alasan; sebaliknya, itu merupakan hasil dari pertimbangan yang teliti dan berdasarkan karakteristik unik dari *Random Forest* yang dapat memberikan keunggulan dalam kasus ini. Meskipun demikian, penting untuk mengingat bahwa keputusan ini tidak eksklusif, dan penelitian ini juga akan memperbandingkan kinerja *Random Forest* dengan algoritma klasifikasi lainnya, terutama *Decision Tree*.

Alasan di balik pemilihan *Random Forest* sebagai fokus utama adalah karena kemampuannya untuk mengatasi beberapa kelemahan yang sering terlihat pada *Decision Tree*, yang merupakan salah satu pendahulunya yang lebih sederhana. *Decision Tree* cenderung rentan terhadap *overfitting*, di mana model secara terlalu ketat menyesuaikan diri dengan data pelatihan, sehingga kinerjanya menurun saat diberikan data baru.

Namun, sementara *Random Forest* menawarkan banyak keunggulan, termasuk kekuatan dalam menangani data yang besar dan kompleks serta kemampuan untuk menangani fitur-fitur yang tidak terstruktur atau berkaitan secara kompleks. Oleh karena itu, penelitian ini akan mencakup serangkaian eksperimen yang membandingkan kinerja *Random Forest* dengan *Decision Tree* dalam konteks dataset dan tujuan spesifik penelitian ini. Dengan demikian, memungkinkan peneliti untuk memahami kelebihan dan kekurangan dari masing-masing pendekatan tersebut, serta membuat keputusan yang lebih terinformasi tentang model mana yang paling cocok untuk keperluan analisis dan prediksi dalam penelitian ini.

4. *Evaluation*

a. Confusion Matrix

1. *Random Forest*

Model *Random Forest* menunjukkan hasil yang sangat baik: Akurasi: 95.37%

Classification Report:

Precision untuk kelas 0 (label negatif): 94%

Recall untuk kelas 0: 98%

F1-score untuk kelas 0: 96%

Precision untuk kelas 1 (label positif): 98%

Recall untuk kelas 1: 91%

F1-score untuk kelas 1: 94%

Confusion Matrix:

True Negative (TN): 5114 *False Positive* (FP): 79 *False Negative* (FN): 332 *True Positive* (TP): 3346

Model ini memiliki kemampuan yang sangat baik dalam memprediksi kelas 0 dan 1, dengan tingkat *precision*, *recall*, dan *F1-score* yang tinggi untuk kedua kelas tersebut. Sebagai tambahan, akurasi keseluruhan model juga sangat tinggi, mencapai 95.37%.

2. *Decission Tree*

Hasil dari model *Decision Tree* adalah sebagai berikut: Akurasi: 92.72%

Classification Report:

Precision untuk kelas 0 (label negatif): 95%

Recall untuk kelas 0: 93%

F1-score untuk kelas 0: 94%

Precision untuk kelas 1 (label positif): 90%

Recall untuk kelas 1: 93%

F1-score untuk kelas 1: 91%

Confusion Matrix:

True Negative (TN): 4817 *False Positive* (FP): 376 *False Negative* (FN): 270 *True Positive* (TP): 3408

b. Cross Validation

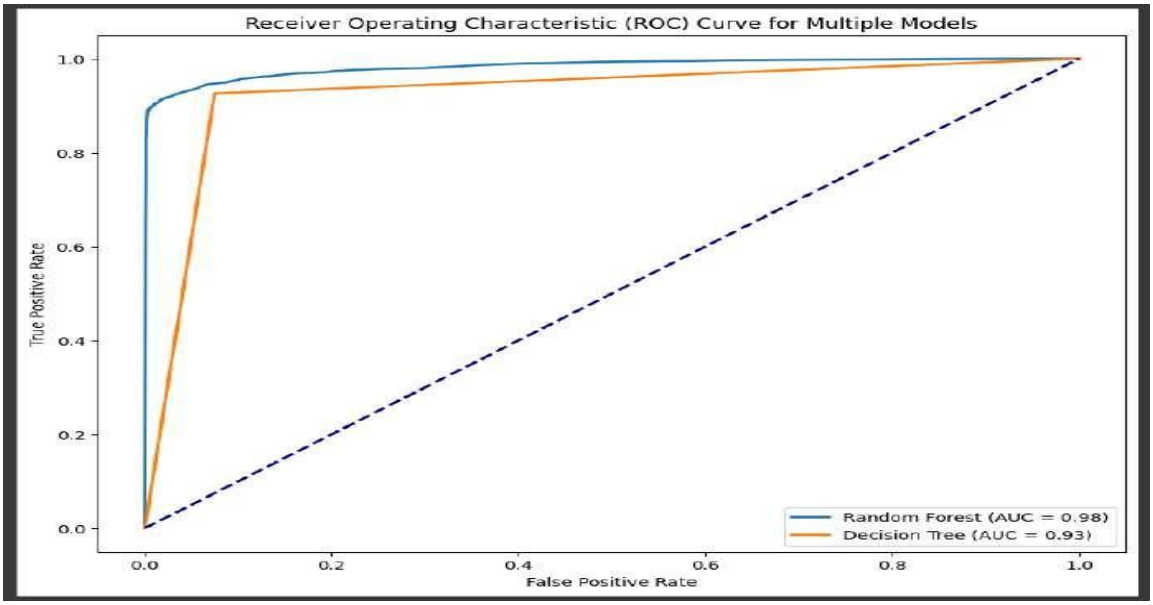
1. *Random Forest*

Hasil dari *cross-validation* menunjukkan bahwa model *Random Forest* memiliki rata-rata akurasi sebesar 95.08% setelah dilakukan validasi silang. Setiap *fold* dari validasi silang juga menunjukkan akurasi yang tinggi, berkisar antara 94.53% hingga 95.48%.

2. *Decission Tree*

Cross-validation menunjukkan bahwa model *Decision Tree* memiliki rata-rata akurasi sebesar 92.81% setelah dilakukan validasi silang. Hasil validasi silang dari setiap *fold* berkisar antara 92.22% hingga 93.34%.

c. Receiver Operating Characteristic (ROC)



Gambar.5 ROC

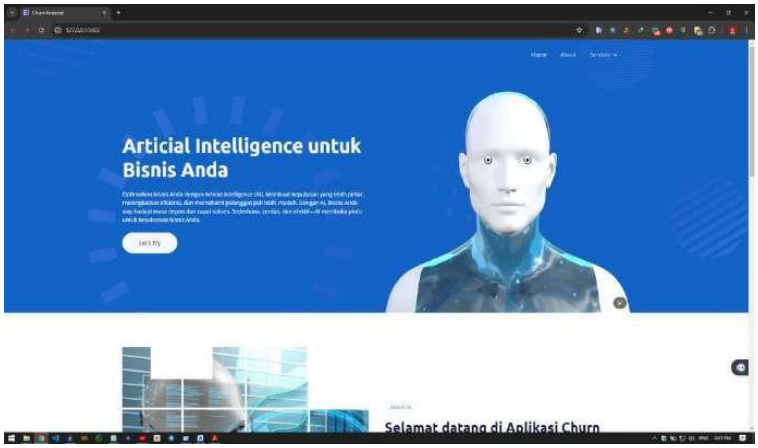
Dalam evaluasi kinerja model klasifikasi, digunakan metrik Area *Under the Receiver Operating Characteristic Curve* (AUC-ROC) sebagai indikator utama. Hasil evaluasi ROC untuk 2 model yang telah dibangun adalah sebagai berikut:

Random Forest menghasilkan AUC sebesar 0.98. Hal ini menunjukkan bahwa model *Random Forest* secara keseluruhan memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif.

Model *Decision Tree* (*Fecission Tree*) menghasilkan AUC sebesar 0.93. Meskipun hasilnya masih baik, namun terdapat perbedaan yang cukup signifikan dibandingkan dengan model *Random Forest*. Hal ini menandakan bahwa model *Decision Tree* memiliki kinerja yang sedikit lebih rendah dalam membedakan antara kelas positif dan negatif dibandingkan dengan model *Random Forest*. Dengan demikian, dari hasil evaluasi ROC ini, dapat disimpulkan bahwa model *Random Forest*, memiliki kinerja yang sangat baik dalam melakukan klasifikasi, sementara model *Decision Tree* memiliki kinerja yang sedikit lebih rendah meskipun tetap memberikan hasil yang dapat diterima.

5. *Deployment*

a. Halaman Utama



Gambar.6 Halaman Utama

Di halaman utama situs web ini, pengguna disuguhi dengan tiga menu utama: "*Home*", "*About*", dan "*Services*" yang mempunyai tiga tiga submenu yaitu "*Form Predict*", "Dataset *Predict*", dan "*Data Visualization*".
Setiap submenu ini memberikan akses ke fitur-fitur khusus yang relevan dengan analisis dan prediksi data. "*Form Predict*" menyediakan fitur untuk prediksi berdasarkan input perdata. Sementara "*Dataset Predict*" menyediakan fitur untuk memprediksi dataset yang dimasukkan pengguna, kemudian dataset hasil prediksinya bisa di download dalam bentuk file CSV. "*Data Visualization*" fitur untuk menampilkan visualisasi data yang menarik untuk memahami informasi yang terkandung dalam dataset.

SIMPULAN

Penelitian ini menggunakan metode *AI Project Cycle* dan algoritma *Random Forest* untuk membangun model prediksi *churn* pelanggan. Faktor-faktor seperti status langganan TV, paket film, usia langganan, dan tagihan rata-rata ditemukan berperan penting dalam memprediksi churn pelanggan. Hasil penelitian menunjukkan bahwa penggunaan algoritma *Random Forest* dapat meningkatkan efisiensi dalam memprediksi churn pelanggan dengan signifikan. Dalam penelitian ini, akurasi prediksi menggunakan algoritma Random Forest mencapai hingga 95%, sementara penggunaan *Decision Tree* menunjukkan akurasi sebesar 92%. Hal ini menandakan bahwa *Random Forest* memberikan peningkatan yang signifikan dalam performa prediksi dibandingkan dengan pendekatan *Decision Tree*.

DAFTAR PUSTAKA

[1] F. Arina and M. Ulfah, "Analisa survival untuk mengurangi customer churn pada perusahaan telekomunikasi," *Journal Industrial Servicess*, vol. 8, no. 1, p. 59, 2022.

[2] M. Zhao, Q. Zeng, M. Chang, Q. Tong, and J. Su, "A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China," *Discrete Dynamics in Nature and Society*, vol. 2021, 2021.

[3] J. Pamina *et al*., "An effective classifier for predicting churn in telecommunication," *JARDCS*, vol. 11, no. 1 Special Issue, pp. 221–229, 2019.

[4] I. M. Latief, A. Subekti, and W. Gata, "Prediksi Tingkat Pelanggan Churn Pada Perusahaan Telekomunikasi Dengan Algoritma Adaboost," *Jurnal Informatika*, vol. 21, no. 1, pp. 34–43, 2021.

[5] D. L. Prianto, I. Ernawati, and N. Chamidah, "Implementasi Churn Prediction Di Industri Telekomunikasi Dengan Metode Logistic Regression Dan Correlation-Based Feature Selection," *n.p.*, pp. 188–196, 2022.

[6] A. Mauludin Nur Aziz *et al*., "Prediksi Customer Churn Menggunakan Logistic Regression dan Decission Tree," *E-Prosiding Teknik Informatika*, vol. 4, no. 1, pp. 11–18, 2023.

[7] M. Amirulhaq Iskandar and U. Latifa, "Website Prediksi Customer Churn Untuk Mempertahankan Pelanggan Pada Perusahaan Telekomunikasi," *JATI*, vol. 7, no. 2, pp. 1308–1316, 2023.

[8] S. H. Sholeha, M. Faid, and M. A. Yaqin, "Prediksi Perpindahan Pelanggan Pada Toko Online Menggunakan Metode Tree-Based Gradient Boosted Models," *Journal of Computer …*, vol. 5, no. 3, pp. 605–614, 2024.

[9] N. A. S. Dinata, G. Abdurrahman, and N. Q. Fitriyah, "Perbandingan Optimasi Algoritma Random Forest Menggunakan Teknik Boosting Terhadap Kasus Klasifikasi Churn Pelanggan Di Industri Telekomunikasi," *Jurnal Aplikasi Sistem Informasi Dan Elektronika*, vol. 5, no. 1, pp. 28–37, 2023.

[10] A. N. Rachmi, "Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn," *n.p.*, pp. 1–101, 2020.

[11] C. Halim, H. D. Purnomo, and T. Wahyono, "Analisis Pengelompokan Wilayah Penyebaran COVID-19 di Indonesia dengan Metode Clustering Menggunakan Algoritma K-Means dan K-Medoids," *INOVTEK Polbeng - Seri Informatika*, vol. 7, no. 2, p. 359, 2022.

[12] F. Azimah *et al*., "Sistem Pendeteksi Gejala Awal Covid-19 Dengan," in *BDCCS*, vol. 4, no. 3, pp. 675–688, 2022.

[13] A. Prasetyo and T. Ridwan, "Analisis Sentimen Terhadap Pemberhentian Tv Analog Pada Twitter Menggunakan Algoritma Naive Bayes," *Jurnal Teknika*, vol. 15, no. 2, pp. 67–74, 2023.

[14] L. D. Martias, "Statistika Deskriptif Sebagai Kumpulan Informasi," *Fihris*, vol. 16, no. 1, p. 40, 2021.

[15] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," *Jurnal Tekno Kompak*, vol. 15, no. 1, p. 131, 2021.

[16] M. R. A. Prasetya, A. M. Priyatno, and Nurhaeni, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," *Jurnal Informasi Dan Teknologi*, vol. 5, no. 2, pp. 52–62, 2023.

[17] D. Karmanita and B. Hendrik, "Anomali Data Mining Menggunakan Metode K-Means Dalam Penilaian Mahasiswa Terhadap Pelayanan Prodi," *Jurnal Media Infotama*, vol. 19, no. 2, pp. 522–527, 2023.

[18] P. R. Fitrayana and D. R. S. Saputro, "Algoritme CLARA untuk Menangani Data Outlier," in *PRISMA*, vol. 5, pp. 721–725, 2022.

[19] R. Siringoringo, R. Perangin Angin, and B. Rumahorbo, "Model Klasifikasi Genetic-XGBoost Dengan T-Distributed Stochastic Neighbor Embedding Pada Peramalan Pasar," *Jurnal Times*, vol. XI, no. 1, pp. 30–36, 2022.

[20] M. Ariandi and S. Rahma Puteri, "Analisis Visualisasi Data Kecamatan Kertapati menggunakan Tableau Public," *JUPITER*, vol. 14, no. 2-b, pp. 366–373, 2022.

[21] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 640, 2021.

[22] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *Jurnal Pilar Nusa Mandiri*, vol. 7, no. 2, p. 217, 2021.

[23] Y. Mufidah, R. Noah, L. Lawalatta, and N. Bragas, "Pengaruh Tingkat Akurasi Dalam Identifikasi Gejala Dan Tanda Penyakit Pada Tanaman," *Jurnal Informatika Progres*, vol. 14, no. 1, pp. 11–15, 2022.

[24] L. Qadrini, A. Sepperwali, and A. Aina, "Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial," *Jurnal Inovasi Penelitian*, vol. 2, no. 7, pp. 1959–1966, 2021.

[25] Fatmawati and N. A. K. Rifai, "Klasifikasi Penyakit Diabetes Retinopati Menggunakan Support Vector Machine," *Jurnal Riset Statistika*, pp. 79–86, 2023.