PROSEMNASPROIT : Prosiding Seminar Nasional Ilmu Teknik
Volume. 2 Nomor. 2, Desember 2025
E-ISSN : 3063-5713; P-ISSN : 3063-4709; Hal. 256-268
DOI: https://doi.org/10.61132/prosemnasproit.v2i2.202
Tersedia: https://prosiding.aritekin.or.id/index.php/PROSEMNASPROIT

Model Machine Learning untuk Klasifikasi Loyalitas Pelanggan Menggunakan Random Forest

Tengku Syahvina Rival Dini[1*], Rani Chantika[2], Pebi Mina Husania[3], Puji Sri Alhirani[4]
[1,2,3,4] Univeristas Islam Negeri Sumatera Utara, Indonesia

Alamat: Jl. Lapangan Golf, Durin Jangak, Kec. Pancur Batu, Kabupaten Deli Serdang, Sumatera Utara
Korespondensi penulis: tengkusyahfina1004@gmail.com

Abstract. *This research develops a machine learning model to classify customer loyalty using the Random Forest algorithm. Customer churn is a critical issue that reduces revenue and increases acquisition costs. A dataset of 50,000 customers from global e-commerce and subscription platforms was processed through data cleaning, imputation, outlier handling, and class balancing with SMOTE. The Random Forest model was built as a baseline and optimized with hyperparameter tuning. Evaluation using accuracy, precision, recall, and F1-score shows that the optimized model achieved 90.81% accuracy and 83.87% F1-score, outperforming previous Naïve Bayes approaches. Feature importance analysis highlights customer service interactions, lifetime value, and demographic factors as key predictors of churn. These findings demonstrate Random Forest's effectiveness in churn prediction and provide practical insights for customer retention strategies*

Keywords: *Customer loyalty, churn prediction, Random Forest, machine learning, SMOTE*

Abstrak. Penelitian ini mengembangkan model machine learning untuk klasifikasi loyalitas pelanggan menggunakan algoritma Random Forest. Churn pelanggan menjadi isu penting karena menurunkan pendapatan dan meningkatkan biaya akuisisi. Dataset berisi 50.000 pelanggan dari platform e-commerce dan layanan berlangganan global diproses melalui pembersihan data, imputasi, penanganan outlier, serta penyeimbangan kelas dengan SMOTE. Model Random Forest dibangun sebagai baseline dan dioptimasi dengan hyperparameter tuning. Evaluasi menggunakan akurasi, presisi, recall, dan F1-score menunjukkan bahwa model teroptimasi mencapai akurasi 90,81% dan F1-score 83,87%, lebih baik dibandingkan pendekatan Naïve Bayes sebelumnya. Analisis feature importance menegaskan bahwa interaksi layanan pelanggan, lifetime value, dan faktor demografi merupakan prediktor utama churn. Hasil ini membuktikan efektivitas Random Forest dalam prediksi churn sekaligus memberikan wawasan strategis bagi retensi pelanggan.

Kata kunci: Loyalitas pelanggan, prediksi churn, Random Forest, machine learning, SMOTE

1. LATAR BELAKANG

Perkembangan bisnis digital dan retail modern menuntut perusahaan untuk mampu mempertahankan loyalitas pelanggan sebagai salah satu faktor utama keberlangsungan usaha. Di tengah pesatnya perkembangan bisnis digital dan retail modern, mempertahankan loyalitas pelanggan menjadi faktor kritis bagi keberlangsungan usaha. Dinamika pasar saat ini menuntut perusahaan untuk berinovasi dan memahami kebutuhan konsumen secara mendalam (Seftiani, 2024). Fenomena pelanggan churn menjadi tantangan serius karena berdampak langsung pada penurunan pendapatan dan meningkatnya biaya akuisisi pelanggan baru. Churn sendiri adalah suatu istilah penting yang merujuk pada fenomena ketika pelanggan berhenti menggunakan produk atau layanan dari suatu perusahaan (Iskandar & Latifa, 2023). Oleh karena itu,

Naskah Masuk: 29 November 2025; Revisi: 30 November 2025; Diterima: 28 Desember 2025;
Terbit: 31 Desember 2025

diperlukan pendekatan berbasis data untuk memprediksi potensi churn sehingga perusahaan dapat melakukan strategi retensi yang lebih efektif.

Penelitian terdahulu telah banyak mengkaji permasalahan ini dengan berbagai algoritma *machine learning*. Misalnya, (Firmansyah & Yulianto, 2021) menggunakan algoritma Naïve Bayes untuk memprediksi customer churn pada bisnis retail. Hasil penelitian tersebut menunjukkan akurasi sebesar 80%, dengan presisi mencapai 100%. Meskipun Naïve Bayes terbukti mampu memberikan prediksi yang cukup baik, algoritma ini memiliki keterbatasan karena asumsi independensi antar variabel yang seringkali tidak sesuai dengan kondisi data riil.

Sebagai pembanding, penelitian ini menggunakan algoritma Random Forest yang dikenal lebih kuat dalam menangani data dengan variabel kompleks dan interaksi antar fitur (Rahma & Putri, 2025). Penelitian ini menekankan pada penggunaan algoritma Random Forest sebagai alternatif yang lebih akurat dibandingkan Naïve Bayes, dengan fokus pada klasifikasi loyalitas pelanggan melalui penampilan hasil akurasi sebagai indikator utama performa model, serta memberikan *insight* tambahan berupa interpretasi hasil akurasi untuk mendukung pengambilan keputusan bisnis sehingga tidak hanya berhenti pada nilai prediksi semata. Dengan demikian, urgensi penelitian ini terletak pada kebutuhan perusahaan untuk memiliki model prediksi yang lebih andal dalam mengidentifikasi pelanggan berisiko churn agar strategi retensi dapat dilakukan lebih tepat sasaran.

2. KAJIAN TEORITIS

Artificial Intelligence (AI) merupakan bidang ilmu komputer yang berfokus pada pengembangan sistem yang mampu meniru kecerdasan manusia dalam melakukan tugas-tugas seperti pengambilan keputusan, prediksi, dan klasifikasi. AI juga merupakan bidang multidisiplin yang bertujuan untuk mengotomatisasi aktivitas yang saat ini membutuhkan kecerdasan manusia (Wahyudi, 2023). Salah satu cabang utama dari AI adalah *machine learning,* yaitu pendekatan yang memungkinkan sistem belajar dari data untuk meningkatkan performa tanpa harus diprogram secara eksplisit. Dengan kata lain, *machine learning* adalah metode yang digunakan dalam AI untuk membangun model prediktif yang dapat mengenali pola dan membuat keputusan berdasarkan data historis (Sidik & Ansawarman, 2022).

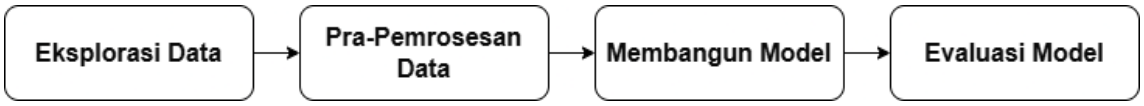
Dalam *machine learning* terdapat tiga kategori utama, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Masing-masing tipe tersebut memiliki tipe data yang berbeda untuk digunakan (Santoso et al., 2021). *Supervised learning* menggunakan data berlabel sehingga model dapat mempelajari hubungan antara masukan atau sebab dan hasilnya atau akibat (Nurhalizah et al., 2024). Contohnya dalam klasifikasi loyalitas pelanggan yang diberi label "*churn*" atau "*loyal*". *Unsupervised learning* digunakan ketika data tidak memiliki label, dengan tujuan menemukan pola atau struktur tersembunyi, misalnya pengelompokan pelanggan berdasarkan perilaku belanja (Harahap et al., 2023). Sementara itu, *reinforcement learning* melibatkan agen yang belajar melalui interaksi dengan lingkungan, menerima *reward* atau *penalty* atas tindakan yang dilakukan, dan banyak digunakan dalam sistem rekomendasi maupun robotika (Baradja & Tjendrowasono, 2024).

Dalam penelitian ini digunakan pendekatan dengan algoritma *Random Forest,* yaitu metode *ensemble learning* yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil (Jananto, 2025). Sama seperti penelitian yang dilakukan oleh Firmansyah sebelumnya, yaitu dengan menggunakan algoritma Naive bayes. Random Forest juga termasuk kedalam bagian dari algoritma *supervised learning.* Random Forest bekerja dengan membangun sejumlah pohon keputusan dari subset data yang berbeda, kemudian menggabungkan hasil prediksi melalui voting mayoritas untuk klasifikasi. Keunggulan algoritma ini adalah kemampuannya menangani data dengan variabel kompleks, interaksi antar fitur, serta mengurangi risiko *overfitting*.

Model *machine learning* dipahami sebagai representasi matematis yang dibangun dari data untuk memprediksi atau mengklasifikasikan suatu fenomena. Dengan kata lain model adalah hasil nyata atau implementasi spesifik dari proses belajar tersebut (Gomede, 2024). Model tersebut merupakan inti dari implementasi AI, karena melalui model inilah sistem dapat melakukan prediksi secara otomatis. Oleh karena itu, hubungan antara AI dan model *machine learning* sangat erat: AI merupakan media kerangka konseptual yang biasanya sudah berupa aplikasi. Lalu *machine learning* adalah cara bagaimana mendapatkan Ai itu sendiri dengan menghasilkan model *machine learning* yang menjadi wujud praktis agar dapat digunakan untuk menyelesaikan masalah nyata, seperti memprediksi loyalitas pelanggan.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk membangun model klasifikasi loyalitas pelanggan. Proses pengembangan model dilakukan secara sistematis melalui tahapan-tahapan yang umum digunakan dalam alur kerja data analis. Setiap tahapan dirancang untuk memastikan kualitas data, ketepatan pemodelan, serta validitas hasil evaluasi. Alur penelitian dapat dilihat sebagaimana pada gambar 1.



Gambar 1. Alur penelitian

A. Eksplorasi Data

Dalam pengembangan model *machine learning*, data merupakan komponen paling fundamental (Widodo & Setiawan, 2025). Algoritma pembelajaran mesin membutuhkan data yang representatif dan berkualitas untuk dapat mengenali pola, membangun model prediktif, serta menghasilkan keputusan yang akurat. Oleh karena itu, data maupun pengolahan datanya merupakan langkah yang krusial.

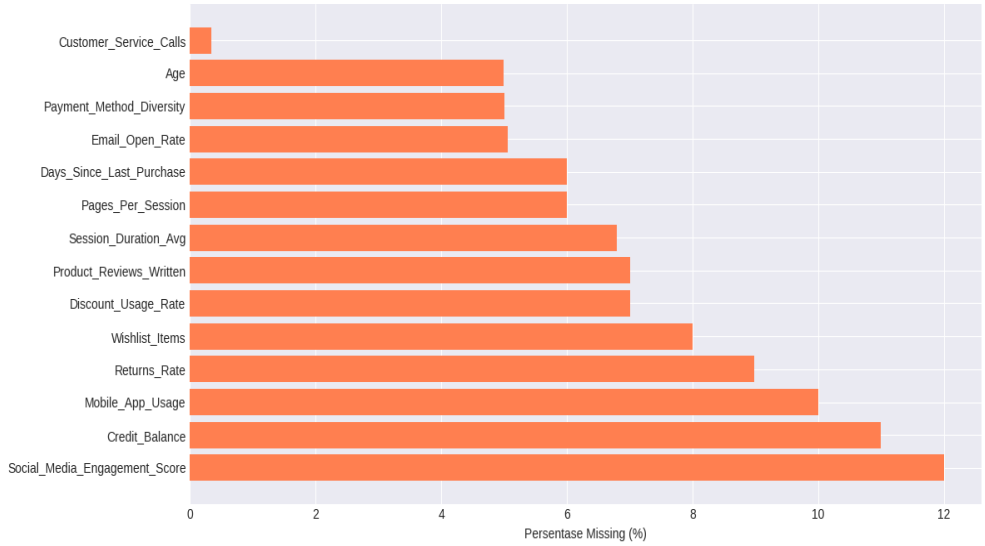
Penelitian ini menggunakan *dataset* yang komprehensif yang berisi data perilaku, demografi, dan transaksi dari 50.000 pelanggan pada *platform e-commerce* dan layanan berlangganan berskala global. *Dataset* ini memiliki 25 atribut atau kolom. *Dataset* ini dirancang untuk memberikan pandangan menyeluruh terhadap interaksi pelanggan, tingkat keterlibatan, dan status loyalitas mereka. Data mencakup pelanggan dari berbagai negara seperti Amerika Serikat, Inggris, Jerman, Kanada, India, Jepang, Prancis, dan Australia, serta merekam perjalanan pelanggan sejak pendaftaran hingga status terkini (Singh, 2025). *Dataset* ini memiliki tipe data campuran, yaitu numerik kontinu (seperti nilai transaksi dan skor keterlibatan), kategorikal (seperti jenis kelamin, negara, metode pembayaran), serta indikator biner (*Churned*). Beberapa kolom juga mengandung nilai *missing values* dan *outliers* yang perlu ditangani pada tahap pra-pemrosesan.

B. Pra-Pemrosesan Data

Tahap pra-pemrosesan data merupakan langkah penting dalam *machine learning* karena kualitas data akan sangat menentukan kualitas model yang dihasilkan. Sebelum data digunakan untuk pelatihan, perlu dilakukan pembersihan terhadap masalah umum yang sering muncul, yaitu *missing value* dan *outlier*.

1) Handling Missing Value

Missing value adalah kondisi ketika suatu variabel dalam *dataset* tidak memiliki nilai atau terisi dengan data kosong (NaN). *Missing value* adalah satu kondisi dimana terdapat nilai yang tidak lengkap atau kosong pada satu atau beberapa kriteria (Sudrajat & Cholid, 2023). Sedangkan *outlier* merupakan nilai yang menyimpang jauh dari pola umum data (Husain & Jamaluddin, 2025). Hal ini dapat terjadi karena berbagai alasan, seperti kesalahan pencatatan, anomali perilaku pelanggan, atau memang mencerminkan kondisi ekstrem tertentu atau juga data yang tidak lengkap saat proses input dan variabel yang memang tidak relevan bagi sebagian pelanggan sehingga tidak diisi. Dalam data yang dilakukan penelitian ini didapatkan *missing value* dan *outlier* sebagai berikut



Gambar 2. Distribusi *missing value* yang terdapat pada beberapa kolom

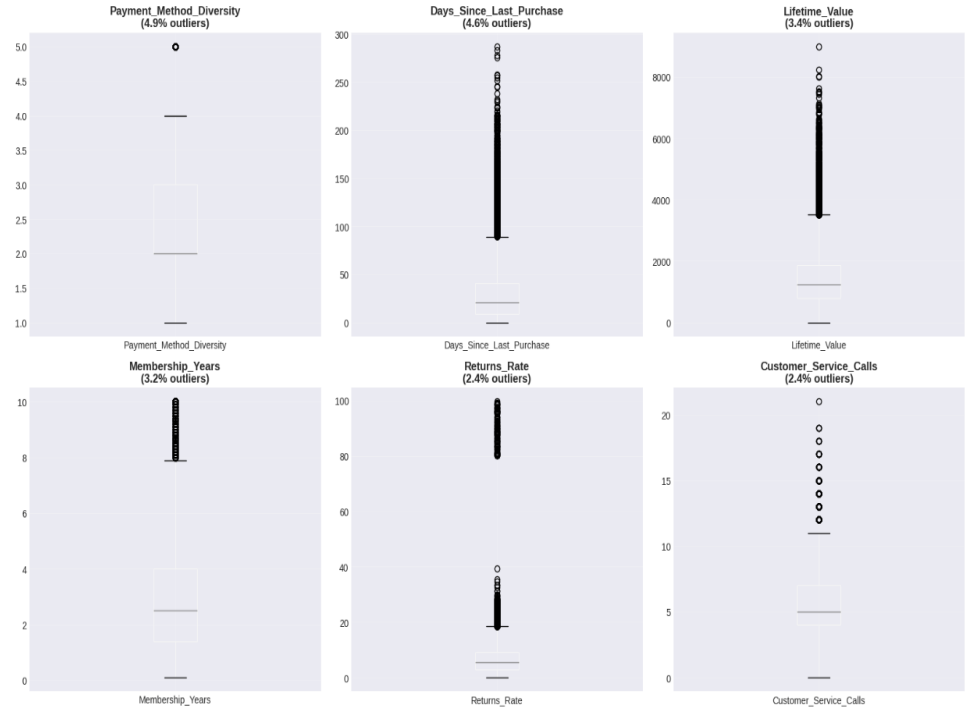
Berdasarkan gambar diatas terlihat bahwa beberapa kolom memiliki jumlah *missing value* diatas 10%. Seperti kolom mobile_app_usage memiliki *missing value* sebesar 10%. Untuk menanganinya dapat menggunakan imputasi seperti *median imputation* bagi atribut atau kolom yang bertipe numerik dan imputasi *modus imputation* bagi atribut yang bertipe kategorikal (Lubis, 2025). Untuk kolom dengan tingkat *missing* lebih dari 10%, dibuat indikator flag tambahan untuk menandai baris yang sebelumnya memiliki nilai kosong. Strategi ini memungkinkan model mengenali pola ketidakhadiran data sebagai sinyal potensial dalam klasifikas (Ramadhani et al., 2024). Hasilnya dapat dilihat sebagai berikut:

```
================================================================================
PROCESSING NUMERICAL FEATURES (20 kolom)
================================================================================
✓ Age                          | Missing: 2,495 ( 4.99%) | Median:    38.00
✓ Session_Duration_Avg         | Missing: 3,399 ( 6.80%) | Median:    26.80
✓ Pages_Per_Session            | Missing: 3,000 ( 6.00%) | Median:     8.40
✓ Wishlist_Items               | Missing: 4,000 ( 8.00%) | Median:     4.00
✓ Days_Since_Last_Purchase     | Missing: 3,000 ( 6.00%) | Median:    21.00
✓ Discount_Usage_Rate          | Missing: 3,500 ( 7.00%) | Median:    40.20
✓ Returns_Rate                 | Missing: 4,491 ( 8.98%) | Median:     5.40
✓ Email_Open_Rate              | Missing: 2,528 ( 5.06%) | Median:    19.70
✓ Customer_Service_Calls       | Missing:   168 ( 0.34%) | Median:     5.00
✓ Product_Reviews_Written      | Missing: 3,500 ( 7.00%) | Median:     2.00
✓ Social_Media_Engagement_Score | Missing: 6,000 (12.00%) | Median:    27.60 + Flag 'Social_Media_Engagement_Score_missing_flag'
✓ Mobile_App_Usage             | Missing: 5,000 (10.00%) | Median:    18.60
✓ Payment_Method_Diversity     | Missing: 2,500 ( 5.00%) | Median:     2.00
✓ Credit_Balance               | Missing: 5,500 (11.00%) | Median:  1896.00 + Flag 'Credit_Balance_missing_flag'
```

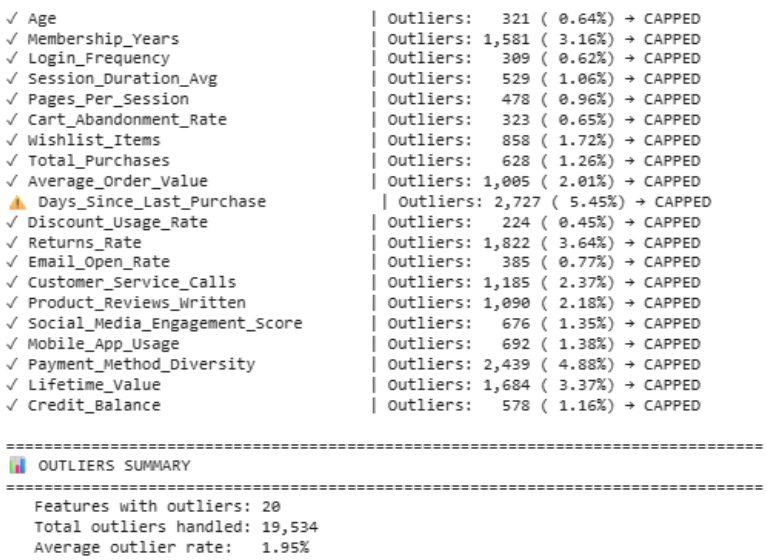
Gambar 3. Imputasi kolom yang bertipe numerik dengan median

2) Handling Outlier

Dalam penelitian ini, *outlier* ditangani menggunakan metode *Interquartile Range* (IQR) *Capping*, yaitu teknik yang mendeteksi nilai ekstrem berdasarkan distribusi kuartil. Nilai yang berada di luar rentang dianggap sebagai *outlier*. Alih-alih menghapus data tersebut, dilakukan proses *capping*, yaitu membatasi nilai *outlier* agar tidak melebihi batas bawah dan atas yang telah ditentukan (Pritalia, 2022). Untuk distribusi *outlier* dapat dilihat pada boxplot di gambar 4. Dan untuk melihat hasil penangannya dengan IQR dapat dilihat pada gambar 5.



Gambar 4. Distribusi outlier pada beberapa atribut



Gambar 5. 20 kolom yang memiliki *outlier* berhasil ditangan dengan IQR

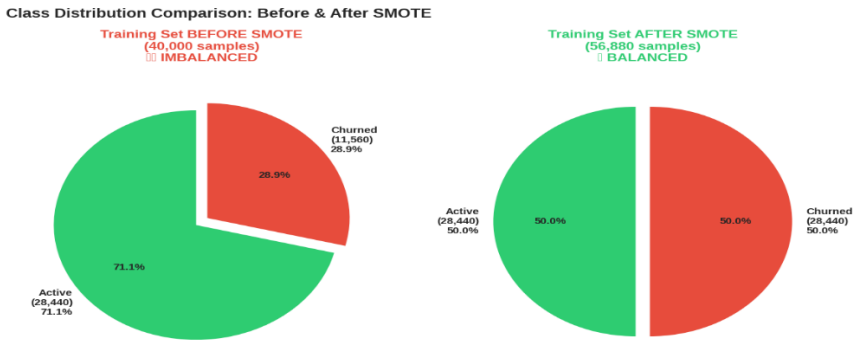
3) Split Data

Setelah data dibersihkan dari *missing values* dan *outlier*, tahap selanjutnya dalam proses pra-pemrosesan adalah pembagian data *(train-test split).* Pembagian data dilakukan dengan rasio 80% untuk pelatihan *(training)* dan 20% untuk pengujian *(testing).* Proses ini menggunakan teknik *stratified sampling*, yaitu pembagian data yang mempertahankan proporsi kelas target agar distribusi antara pelanggan aktif dan churn tetap konsisten di kedua subset. Hasil pembagian menunjukkan bahwa dari total 50.000 data pelanggan, sebanyak 40.000 data digunakan untuk pelatihan dan 10.000 data untuk pengujian.

4) Handling Imbalanced Data

Setelah data dibersihkan dari *missing values* dan *outlier*, tahap selanjutnya atribut dalam proses pra-pemrosesan adalah penanganan ketidakseimbangan label atau target yang didalam *dataset* ini adalah atribut churn. Untuk mengatasi ketidakseimbangan tersebut, digunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE bekerja dengan membuat sampel sintetis dari kelas minoritas (pelanggan churn) berdasarkan interpolasi antara data yang ada. Teknik ini hanya diterapkan pada *training set*, agar model belajar dari data yang seimbang tanpa memengaruhi evaluasi pada *test set* yang tetap mencerminkan distribusi asli (Pratiwi et al., 2025).

Setelah SMOTE diterapkan, jumlah data pelanggan churn dalam *training set* meningkat dari 11.560 menjadi 28.440, sehingga total training set menjadi 56.880 data dengan distribusi kelas yang seimbang (50:50). Sebanyak 16.880 data sintetis berhasil ditambahkan untuk mencapai keseimbangan tersebut. Untuk perbedaanya dapat dilihat pada gambar 6 berikut:



Gambar 6. Perbedaan ketidakseimbangan data setelah Teknik SMOTE

C. Membangun Model

Tahap pembangunan model dilakukan dengan menggunakan algoritma Random Forest Classifier. Tahap pertama, dibangun *baseline model* dengan parameter *default*. Model ini dilatih menggunakan data *training*, kemudian diuji pada data *testing*. *Baseline* model berfungsi sebagai acuan awal untuk melihat performa algoritma Random Forest sebelum dilakukan optimasi lebih lanjut. Selanjutnya, dilakukan *hyperparameter tuning* menggunakan teknik GridSearchCV dengan 3-fold cross-validation. Proses ini bertujuan mencari kombinasi parameter terbaik yang dapat meningkatkan performa model. Parameter yang diuji meliputi jumlah pohon (n_estimators), kedalaman maksimum pohon (max_depth), jumlah minimum sampel untuk split (min_samples_split), jumlah minimum sampel pada daun (min_samples_leaf), serta metode pemilihan fitur (max_features).

D. Evaluasi Model

Model yang telah dilatih kemudian siap untuk dievaluasi menggunakan berbagai metrik seperti *confusion matrix* atau laporan klasifikasi seperti *accuracy*, *precision*, *recall*, *F1-score*, serta dianalisis lebih lanjut melalui *feature importance* untuk mengetahui variabel yang paling berpengaruh terhadap loyalitas pelanggan**.**

Laporan klasifikasi sendiri merupakan ringkasan metrik evaluasi yang dihasilkan dari confusion matrix. Setelah matriks kebingungan dihitung, sistem secara otomatis menghasilkan laporan klasifikasi yang menampilkan nilai metrik untuk setiap kelas, sehingga memudahkan analisis kinerja model secara komprehensif. Laporan ini mencakup beberapa metrik evaluasi sebagai berikut (Azmi & Voutama, 2024).

1. Akurasi (Accuracy): proporsi prediksi yang benar terhadap seluruh data.
2. Presisi (Precision): proporsi prediksi positif yang benar dibandingkan dengan seluruh data yang diprediksi positif.
3. Recall (Sensitivitas): proporsi data positif yang berhasil diprediksi dengan benar dibandingkan dengan seluruh data positif yang sebenarnya.
4. F1-Score: nilai rata-rata harmonik antara presisi dan recall yang digunakan untuk menilai keseimbangan kinerja model.

4. HASIL DAN PEMBAHASAN

Tahap akhir penelitian ini adalah pembangunan dan evaluasi model Random Forest Classifier untuk memprediksi loyalitas pelanggan. Model dilatih menggunakan 56.880 data *training* yang telah diseimbangkan dengan teknik SMOTE, dan diuji pada 10.000 data *testing* dengan distribusi asli.

A. Hasil Model Baseline Random Forest (RF)

Model baseline dengan parameter default menghasilkan akurasi sebesar 90.76% dan F1-score sebesar 83.77%. untuk detailnya dapat dilihat pada tabel 1

Tabel 1. Klasifikasi report model baseline RF

Metrik	Score
Accuracy	90.76%
Precision	84.20%
Recall	82.43%
F1-Score	83.77%

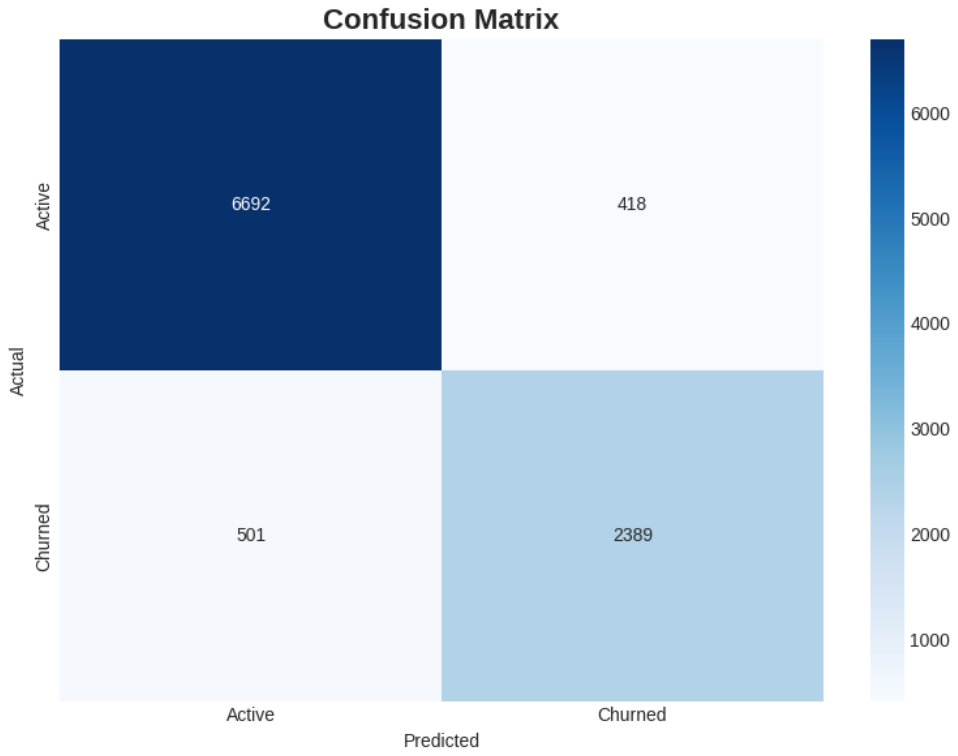
B. Hasil Model RF Dengan Hyperparameter T uning

Setelah dilakukan pemodelan pertama yaitu model *baseline,* selanjutnya dilakukannya model dengan *hyperparameter tuning* menggunakan GridSearchCV dengan 3-fold *cross-validation,* diperoleh kombinasi parameter terbaik yaitu n_estimators=300, max_depth=30, min_samples_split=2, min_samples_leaf=1, dan max_features='sqrt'. Model dengan parameter optimal ini menunjukkan peningkatan performa dengan rata-rata F1-score cross-validation sebesar 91.61%. untuk lebih detailnya dapat dilihat pada tabel 2 dibawah. *Confusion matrix* memperlihatkan bahwa model mampu mengklasifikasikan pelanggan dengan cukup akurat. Hal ini menunjukkan bahwa model berhasil mendeteksi sebagian besar pelanggan churn

dengan tingkat kesalahan yang relatif rendah dan untuk confusion matrix model hyperparameter dapat dilihat pada gambar 7 dibawah:

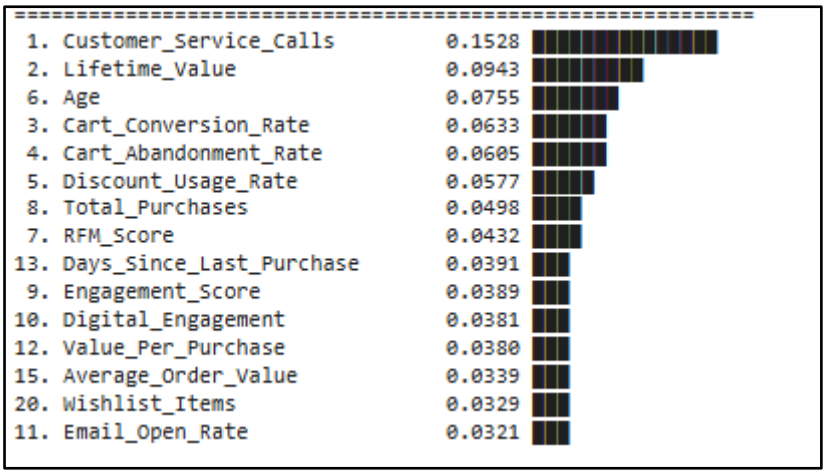
Tabel 2. Klasifikasi report model RF dengan hyperparameter tuning

Metrik	Score
Accuracy	90.81%
Precision	85.11%
Recall	82.66%
F1-Score	83.87%



Gambar 7. Confusion matrix model RF hyperparameter tuning

Hasil analisis fitur menunjukkan bahwa variabel yang paling berpengaruh terhadap prediksi churn adalah Customer_Service_Calls dengan nilai kepentingan sebesar 0.1528, diikuti oleh Lifetime_Value sebesar 0.0943 dan Age sebesar 0.0755. Selain itu, variabel lain seperti Cart Conversion Rate, Cart Abandonment Rate, dan Discount Usage Rate juga memberikan kontribusi yang signifikan terhadap hasil prediksi. Temuan ini memberikan wawasan penting bahwa tingkat interaksi pelanggan dengan layanan pelanggan, nilai seumur hidup pelanggan, serta faktor demografis memiliki peran besar dalam menentukan loyalitas pelanggan dan kecenderungan terjadinya churn.



Gambar 8. Top 15 variabel terpengaruh terhadap model

5. KESIMPULAN DAN SARAN

Secara keseluruhan, model Random Forest yang dibangun menunjukkan performa yang sangat baik dengan akurasi di atas 90% menandakan kemampuan diskriminatif yang tinggi. Model ini siap digunakan untuk implementasi dalam sistem prediksi churn, serta dapat menjadi dasar rekomendasi bisnis dalam menyusun strategi retensi pelanggan. Dengan hasil ini, penelitian berhasil menunjukkan bahwa algoritma Random Forest lebih unggul dibandingkan pendekatan sebelumnya (Naïve Bayes dengan akurasi 80%), sehingga dapat menjadi kontribusi nyata dalam bidang machine learning untuk analisis loyalitas pelanggan.

DAFTAR REFERENSI

Azmi, A. F., & Voutama, A. (2024). Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, 13(1), 111–119. https://doi.org/https://doi.org/10.34010/komputa.v13i1.12639

Baradja, A., & Tjendrowasono, T. I. (2024). Pengaplikasian Deep Reinforcement Q-Learning Untuk Prediksi Perdagangan Valas Otomatis. *Jurnal Rekayasa Sistem Informasi Dan Teknologi*, 1(3), 190–198. https://doi.org/10.59407/jrsit.v1i3.519

Firmansyah, & Yulianto, A. (2021). Prediksi Customer Churn Pada Bisnis Retail Menggunakan Algoritma Naïve Bayes. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, 6(1), 41–47. https://doi.org/http://doi.org/10.33395/remik.v4i1.11196

Gomede, E. (2024). Understanding the Difference Between Algorithms and Models in Machine Learning. Medium.Com. https://medium.com/the-modern-scientist/understanding-the-difference-between-algorithms-and-models-in-machine-learning-71ebacd207fa

Harahap, A. A., Raihan, M., Amani, N., & Andini, P. R. (2023). Comparison of Unsupervised Learning Techniques for Clustering Data on the Number of Villages in Indonesia. Institut Riset Dan Publikasi Indonesia (IRPI). *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 1(1), 163–170.

Husain, A., & Jamaluddin, S. R. W. (2025). Pemodelan Data Angka Kematian Bayi Menggunakan Regresi Robust. *SAINTEK: Jurnal Sains, Teknologi & Komputer*, 1(1), 1–7. https://doi.org/https://doi.org/10.56495/saintek.v1i1.326

Iskandar, M. A., & Latifa, U. (2023). Website Prediksi Customer Churn Untuk Mempertahankan Pelanggan Pada Perusahaan Telekomunikasi. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(2), 1308–1316. https://doi.org/https://doi.org/10.36040/jati.v7i2.6639

Jananto, A. (2025). Studi Komparatif Algoritma Klasifikasi Data Mining pada Prediksi Prestasi Siswa Berbasis Data Sosiodemografis. *Jurnal Teknik Informatika Unika ST. Thomas (JTIUST)*, 10(2), 231–242. http://ejournal.ust.ac.id/index.php/JTIUST/article/view/5840

Lubis, A. H. (2025). Imputasi Data: Konsep, Metode, dan Peran Pentingnya dalam Pelatihan Model dan Pengambilan Keputusan. Andrehasudungan.Blog.Uma.Ac.Id. https://andrehasudungan.blog.uma.ac.id/2025/08/26/imputasi-data-konsep-metode-dan-peran-pentingnya-dalam-pelatihan-model-dan-pengambilan-keputusan/

Nurhalizah, R. S., Ardianto, R., & Purwono. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning : Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika (JIKI)*, 4(1), 61–72. https://doi.org/https://doi.org/10.54082/jiki.168

Pratiwi, F. S., Barata, M. A., & Ardianti, A. D. (2025). Implementasi Metode Smote Dan Random Over- Sampling Pada Algoritma Machine Learning Untuk Prediksi Customer Churn Di Sektor Perbankan. *Jurnal Sistem Informasi Dan Informatika (Simika))*, 8(1), 87–98. https://doi.org/https://doi.org/10.47080/simika.v8i1.3678

Pritalia, G. L. (2022). Analisis Komparatif Algoritme Machine Learning pada Klasifikasi Kualitas Air Layak Minum. *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, 2(1), 43–55. https://doi.org/https://doi.org/10.24002/konstelasi.v2i1.5630

Rahma, S. A., & Putri, T. (2025). Klasifikasi Konsumen Berdasarkan Loyalitas Belanja Online Menggunakan Algoritma Random Forest. *Jurnal ICT : Information Communication & Technology*, 25(1), 81–86. https://doi.org/https://doi.org/10.36054/jict-ikmi.v25i1.304

Ramadhani, D., Soleh, A. M., & Erfiani. (2024). Machine Learning-Based Univariate Time Series Imputation Method for Estimating Missing Values in Non- Stationary Data. *Jurnal Matematika, Statistika Dan Komputasi*, 21(1), 307–320. https://doi.org/10.20956/j.v21i1.36468

Santoso, P., Abijono, H., & Anggreini, N. L. (2021). Algoritma Supervised Learning Dan Unsupervised. *G-Tech Jurnal Teknlogi Terapan*, 4(2), 315–318. https://doi.org/https://doi.org/10.33379/gtech.v4i2.635

Seftiani, A. (2024). Pengaruh Kualitas Pelayanan dan Inovasi Produk Terhadap Keunggulan Bersaing Minuman Boba Pattaya [Universitas Jambi]. https://repository.unja.ac.id/id/eprint/69303

Sidik, A. D., & Ansawarman, A. (2022). Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning. *Formosa Journal of Multidisciplinary Research (FJMR)*, 1(3), 559–568. https://doi.org/10.55927/fjmr.v1i3.745

Singh, D. (2025). Customer Engagement and Churn Analytics Dataset. Kaggle. https://www.kaggle.com/datasets/dhairyajeetsingh/ecommerce-customer-behavior-dataset

Sudrajat, W., & Cholid, I. (2023). K-Nearest Neighbor (K-Nn) Untuk Penanganan Missing Value Pada Data Umkm. *JRSIT: Jurnal Rekayasa Sistem Informasi Dan Teknologi*, 1(2), 54–63. https://doi.org/10.59407/jrsit.v1i2.77

Wahyudi. (2023). studi kasus pengembangan dan penggunaan artificial intelligence (ai) sebagai penunjang kegiatan masyarakat indonesia. *indonesian journal on software engineering*, 9(1), 28–32. https://doi.org/10.31294/ijse.v9i1.15631

Widodo, A. P., & Setiawan, A. (2025). Systematic Literature Review : Peran Machine Learning dalam Manajemen Sumber Daya Manusia. *Optimal: Jurnal Ekonomi Dan Manajemen*, 5(4), 650–665. https://doi.org/https://doi.org/10.55606/optimal.v5i4.8089