

**ANALISIS KOMPARATIF ALGORITMA MACHINE
LEARNING UNTUK KLASIFIKASI PERILAKU BELANJA
KONSUMEN PADA E-COMMERCE**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

AISHA ARYANI

22.11.5118

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS KOMPARATIF ALGORITMA MACHINE
LEARNING UNTUK KLASIFIKASI PERILAKU BELANJA
KONSUMEN PADA E-COMMERCE**

yang disusun dan diajukan oleh

Aisha Aryani

22.11.5118

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 26 Februari 2026

Dosen Pembimbing,



Majid Rahardi, S.Kom., M.Eng.
NIK. 190302393

HALAMAN PENGESAHAN

SKRIPSI

**ANALISIS KOMPARATIF ALGORITMA MACHINE
LEARNING UNTUK KLASIFIKASI PERILAKU BELANJA
KONSUMEN PADA E-COMMERCE**

yang disusun dan diajukan oleh

Aisha Aryani

22.11.5118

Telah dipertahankan di depan Dewan Penguji
pada tanggal 26 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Rizqi Sukma Kharisma, S.Kom., M.Kom.
NIK. 190302215

I Made Artha Agastya, S.T., M.Eng., Ph.D.
NIK. 190302352

Majid Rahardi, S.Kom., M.Eng.
NIK. 190302393

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 26 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Aisha Aryani
NIM : 22.11.5118

Menyatakan bahwa Skripsi dengan judul berikut:

ANALISIS KOMPARATIF ALGORITMA MACHINE LEARNING UNTUK KLASIFIKASI PERILAKU BELANJA KONSUMEN PADA E-COMMERCE

Dosen Pembimbing : Majid Rahardi, S.Kom, M.Eng

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 26 Februari 2026

Yang Menyatakan,



10000
11D10ANX315566987

Aisha Aryani

KATA PENGANTAR

Penulis memanjatkan puji dan syukur ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga skripsi ini dapat diselesaikan dengan baik. Skripsi ini disusun sebagai tugas akhir perkuliahan dan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknologi Informasi, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, serta sebagai bentuk penerapan ilmu yang diperoleh selama masa perkuliahan. Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan, oleh karena itu penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan bimbingan, dukungan, dan bantuan selama proses penyusunan skripsi ini, kepada:

1. Orang tua dan Kakak tercinta atas dukungan dan doa yang tiada henti.
2. Prof. Dr. Mohammad Suyanto, M.M selaku Rektor Universitas Amikom Yogyakarta.
3. Prof. Dr. Kusrini, M.Kom, selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.
4. Eli Pujastuti, M.Kom, selaku Ketua Program Studi Informatika, Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.
5. Majid Rahardi, S.Kom, M.Eng, selaku Dosen Pembimbing yang telah memberikan bimbingan dan arahan selama penyusunan laporan.
6. Tim Dosen Penguji, atas masukan kritis dan saran konstruktif yang memperdalam pemahaman penulis terhadap penelitian ini.
7. Teman-teman saya yang selalu ada buat saya, khususnya ibu peri saudara perempuan, dua sahabat perkuliahan, dan teman rumah yang selalu ada dan memberikan dukungan.
8. Diri sendiri, yang telah berusaha tetap kuat dan bertahan sejauh ini.

Yogyakarta, 26 Februari 2026

Penulis

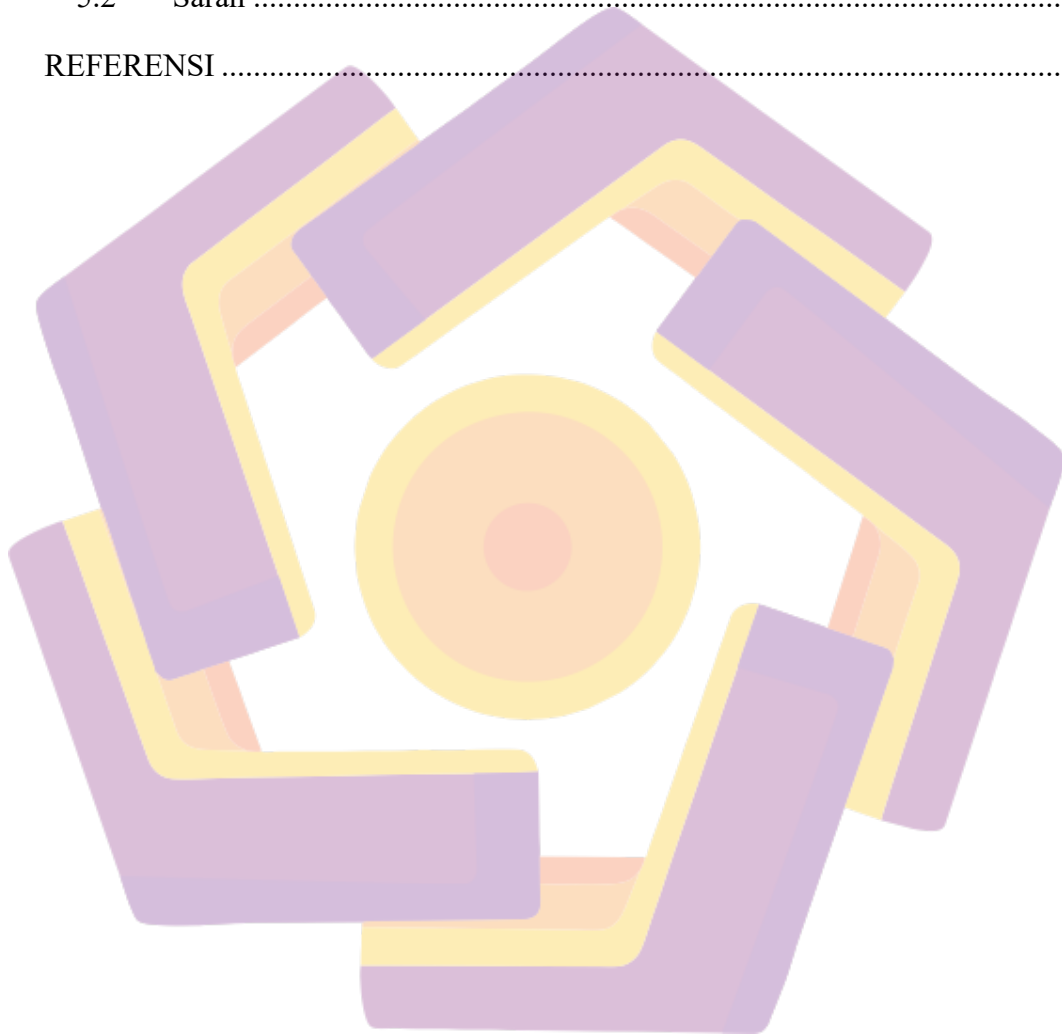
DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
KATA PENGANTAR	v
DAFTAR ISI.....	vi
DAFTAR TABEL.....	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMBANG DAN SINGKATAN	xii
DAFTAR ISTILAH	xiii
INTISARI	xv
<i>ABSTRACT</i>	xvi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.5.1 Manfaat Teoritis.....	5
1.5.2 Manfaat Praktis	5
1.6 Sistematika Penulisan	6
BAB II TINJAUAN PUSTAKA	7
2.1 Studi Literatur	7

2.2	Dasar Teori.....	15
2.2.1	Pengertian E-Commerce	15
2.2.2	Pengertian Perilaku Belanja Konsumen.....	15
2.2.3	Machine learning	16
2.2.4	Klasifikasi	17
2.2.5	Naïve Bayes	17
2.2.6	Logistic regression	18
2.2.7	Random forest.....	19
2.2.8	Preprocessing Data.....	21
2.2.9	Ketidakseimbangan Data (Inbalance Dataset)	21
2.2.10	Evaluasi Model	22
2.2.11	Confusion matrix	22
BAB III METODE PENELITIAN		25
3.1	Objek Penelitian.....	25
3.2	Alur Penelitian	25
3.2.1	Mulai	26
3.2.2	Pengumpulan <i>Dataset</i>	26
3.2.3	Eksplorasi Data (Exploratory Data Analysis).....	27
3.2.4	Pra-Pemrosesan Data	28
3.2.5	Split Data	29
3.2.6	Modeling	29
3.3	Alat dan Bahan.....	31
3.3.1	Data Penelitian	31
3.3.2	Alat dan Instrumen Penelitian.....	31
3.3.2.1	Kebutuhan Perangkat Keras.....	31

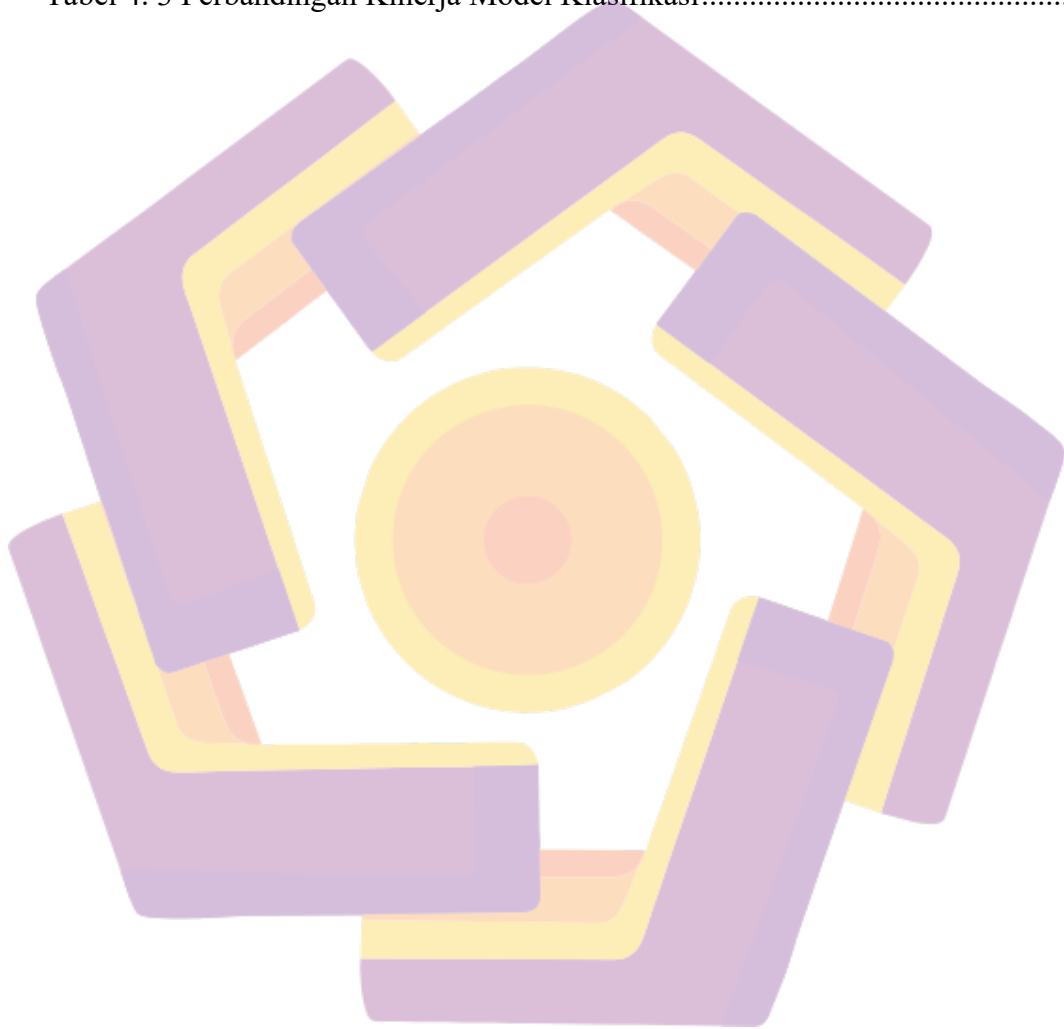
3.3.2.2	Kebutuhan Perangkat Lunak.....	31
3.3.2.3	Library.....	32
3.3.2.4	Platform Pendukung.....	32
BAB IV HASIL DAN PEMBAHASAN		33
4.1	Import Library dan <i>Dataset</i>	33
4.2	EDA	33
4.2.1	Check for Missing Values.....	33
4.2.2	Look for Duplicate Rows.....	34
4.2.3	Outlier Detection.....	34
4.2.4	Analisis Distribusi Variabel Target.....	35
4.2.5	Distribution Of Each Feature	36
4.2.6	Correlation Between Features.....	37
4.3	Preprocessing	38
4.3.1	Claning Data	39
4.3.2	Encode Target	39
4.3.3	Encode Kategorikal.....	40
4.4	Split Data	40
4.4.1	Train Dan Split.....	40
4.4.2	Feature scaling	41
4.4.3	Balancing Data (SMOTE).....	41
4.5	Modeling	42
4.6	Evaluasi.....	42
4.6.1	Evaluasi Random forest	43
4.6.2	Evaluasi Logistic regression	44
4.6.3	Evaluasi Naïve Bayes.....	45

4.7	Perbandingan Semua Model	45
4.8	Data Visualisasi	46
BAB V PENUTUP		52
5.1	Kesimpulan	52
5.2	Saran	52
REFERENSI		54



DAFTAR TABEL

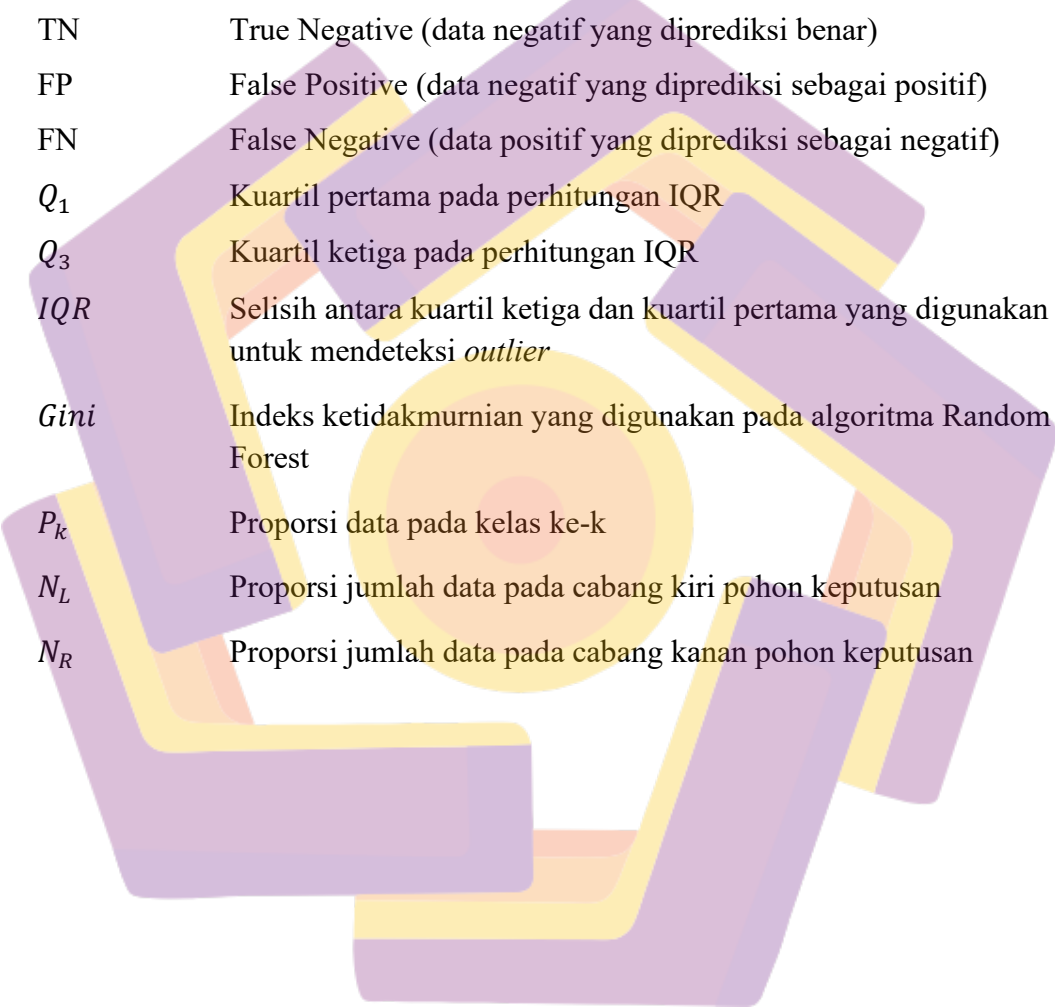
Tabel 2. 1 Keaslian Penelitian	10
Tabel 4. 1 Tabel Hasil Perhitungan Batas Outlier Data Numerik.....	39
Tabel 4. 2 Perbandingan Penerapan SMOTE	42
Tabel 4. 3 Perbandingan Kinerja Model Klasifikasi.....	46



DAFTAR GAMBAR

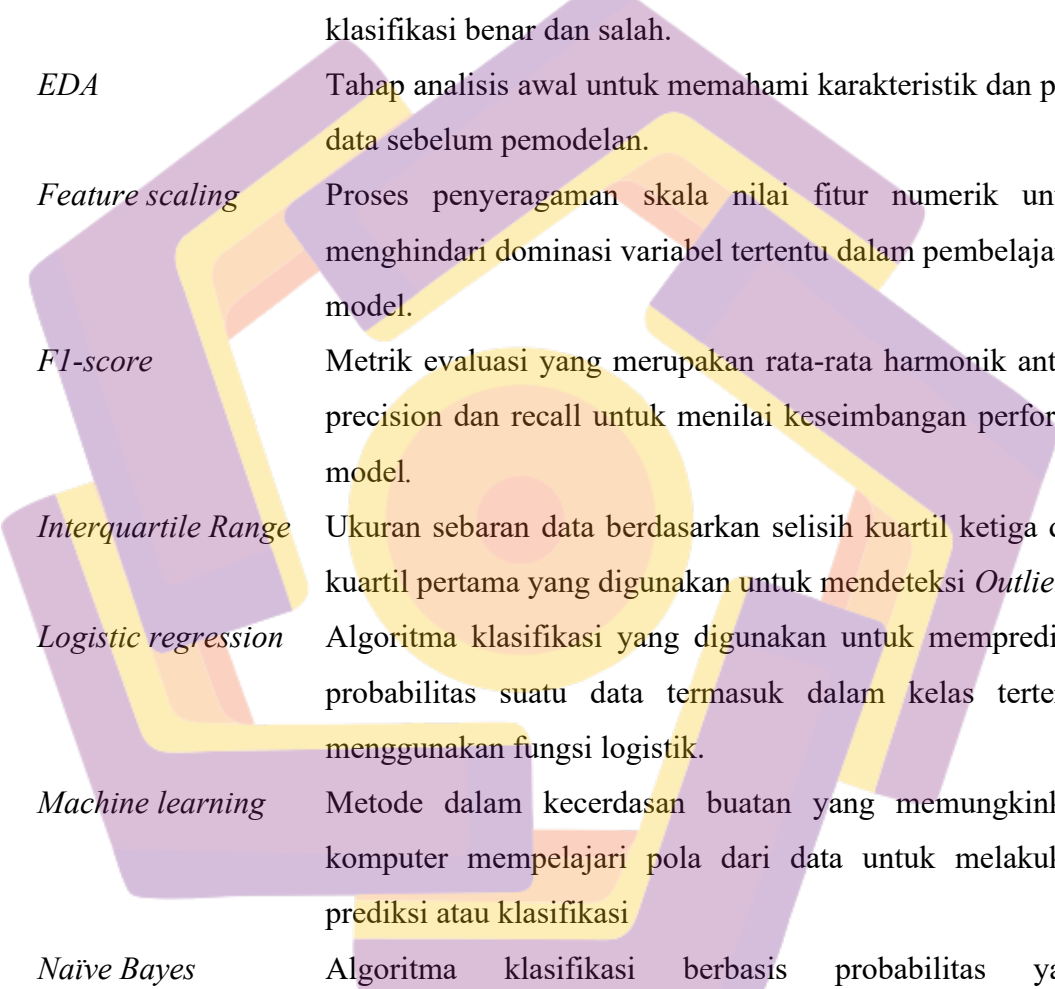
Gambar 2. 1 Contoh Pola Perilaku Belanja	16
Gambar 2. 2 Confusion Matrix	23
Gambar 3. 1 Alur Penelitian Alur Penelitian	26
Gambar 4. 1 Hasil pemeriksaan missing values pada setiap variable.....	34
Gambar 4. 2 Boxplot seluruh variabel numerik untuk deteksi Outlier	35
Gambar 4. 3 Distribusi Variabel Target Subscription Status.....	36
Gambar 4. 4 Distribusi Fitur Numerik, Variabel Target, dan Fitur Kategorikal Utama	37
Gambar 4. 5 Matriks korelasi variabel numerik	38
Gambar 4. 6 Perhitungan Batas Outlier Menggunakan Metode IQR	39
Gambar 4. 7 Confusion matrix Model Random forest.....	43
Gambar 4. 8 Confusion matrix Model Logistic regression.....	44
Gambar 4. 9 Confusion matrix Model Naive Bayes.....	45
Gambar 4. 10 Perbandingan Kinerja Model	47
Gambar 4. 11 Perbandingan Akurasi dan F1-score Model Klasifikasi	48
Gambar 4. 12 analisis terhadap data yang sudah di one hot encoding.....	49
Gambar 4. 13 pengaruh data numerik terhadap label umur vs subscriber dengan menggunakan box plot.....	50

DAFTAR LAMBANG DAN SINGKATAN

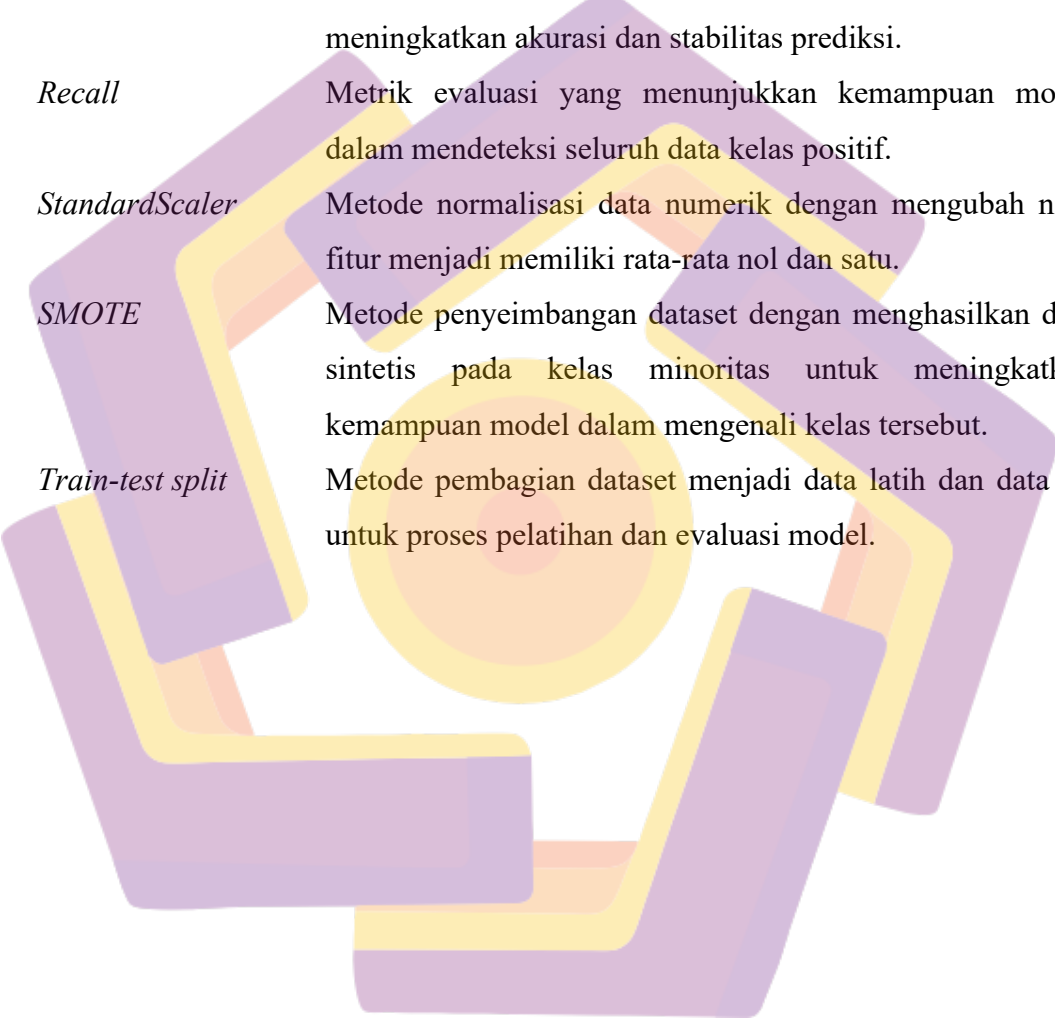


EDA	Exploratory Data Analysis
IQR	Interquartile Range
SMOTE	Synthetic Minority Over-sampling Technique
TP	True Positive (data positif yang diprediksi benar)
TN	True Negative (data negatif yang diprediksi benar)
FP	False Positive (data negatif yang diprediksi sebagai positif)
FN	False Negative (data positif yang diprediksi sebagai negatif)
Q_1	Kuartil pertama pada perhitungan IQR
Q_3	Kuartil ketiga pada perhitungan IQR
IQR	Selisih antara kuartil ketiga dan kuartil pertama yang digunakan untuk mendeteksi <i>outlier</i>
Gini	Indeks ketidakmurnian yang digunakan pada algoritma Random Forest
P_k	Proporsi data pada kelas ke-k
N_L	Proporsi jumlah data pada cabang kiri pohon keputusan
N_R	Proporsi jumlah data pada cabang kanan pohon keputusan

DAFTAR ISTILAH



<i>Accuracy</i>	Metrik evaluasi yang menunjukkan tingkat ketepatan model dalam memprediksi keseluruhan data.
<i>Confusion matrix</i>	Tabel evaluasi yang digunakan untuk membandingkan hasil prediksi model dengan data sebenarnya berdasarkan klasifikasi benar dan salah.
<i>EDA</i>	Tahap analisis awal untuk memahami karakteristik dan pola data sebelum pemodelan.
<i>Feature scaling</i>	Proses penyeragaman skala nilai fitur numerik untuk menghindari dominasi variabel tertentu dalam pembelajaran model.
<i>F1-score</i>	Metrik evaluasi yang merupakan rata-rata harmonik antara precision dan recall untuk menilai keseimbangan performa model.
<i>Interquartile Range</i>	Ukuran sebaran data berdasarkan selisih kuartil ketiga dan kuartil pertama yang digunakan untuk mendeteksi <i>Outlier</i> .
<i>Logistic regression</i>	Algoritma klasifikasi yang digunakan untuk memprediksi probabilitas suatu data termasuk dalam kelas tertentu menggunakan fungsi logistik.
<i>Machine learning</i>	Metode dalam kecerdasan buatan yang memungkinkan komputer mempelajari pola dari data untuk melakukan prediksi atau klasifikasi.
<i>Naïve Bayes</i>	Algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen.
<i>Outlier</i>	Nilai data yang menyimpang jauh dari pola umum <i>Dataset</i> .
<i>Outlier Capping</i>	Teknik penanganan <i>Outlier</i> dengan membatasi nilai ekstrem tanpa menghapus data.
<i>One-hot encoding</i>	Metode pengkodean data kategorikal menjadi variabel biner agar dapat diproses model.



<i>Preprocessing Data</i>	Proses pengolahan dan persiapan data sebelum digunakan dalam pelatihan model.
<i>Precision</i>	Metrik evaluasi yang menunjukkan tingkat ketepatan model dalam memprediksi kelas positif.
<i>Random forest</i>	Algoritma klasifikasi berbasis ensemble learning yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan stabilitas prediksi.
<i>Recall</i>	Metrik evaluasi yang menunjukkan kemampuan model dalam mendeteksi seluruh data kelas positif.
<i>StandardScaler</i>	Metode normalisasi data numerik dengan mengubah nilai fitur menjadi memiliki rata-rata nol dan satu.
<i>SMOTE</i>	Metode penyeimbangan dataset dengan menghasilkan data sintetis pada kelas minoritas untuk meningkatkan kemampuan model dalam mengenali kelas tersebut.
<i>Train-test split</i>	Metode pembagian dataset menjadi data latih dan data uji untuk proses pelatihan dan evaluasi model.

INTISARI

Perkembangan pesat e-commerce menghasilkan data transaksi dalam jumlah besar yang mencerminkan perilaku belanja konsumen, namun kompleksitas data serta ketidakseimbangan kelas antara konsumen yang berlangganan dan tidak berlangganan menjadi permasalahan dalam proses analisis dan pemodelan klasifikasi. Kondisi ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas sehingga menurunkan kemampuan dalam mengenali karakteristik konsumen secara menyeluruh dan berdampak pada kurang optimalnya strategi pemasaran berbasis data. Penelitian ini bertujuan untuk memprediksi status berlangganan konsumen e-commerce serta membandingkan kinerja beberapa algoritma *machine learning* dalam mengklasifikasikan perilaku belanja konsumen. *Dataset* yang digunakan terdiri dari 3.900 data dengan 18 atribut yang merepresentasikan karakteristik transaksi dan perilaku belanja. Metode penelitian meliputi *Exploratory Data Analysis (EDA)*, pembersihan data, penanganan *Outlier* menggunakan metode *Outlier capping* berbasis *Interquartile Range (IQR)*, *encoding* fitur kategorikal menggunakan *One-hot encoding*, pembagian data dengan metode *stratified Train-test split*, *Feature scaling* menggunakan *StandardScaler*, serta penyeimbangan data latih menggunakan metode *SMOTE*. Selanjutnya, proses pemodelan dilakukan menggunakan algoritma *Naive Bayes*, *Logistic regression*, dan *Random forest* dengan evaluasi kinerja berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa ketiga algoritma mampu memprediksi status berlangganan konsumen dengan performa yang baik, di mana *Naive Bayes* menghasilkan kinerja terbaik berdasarkan nilai *F1-score*. Hasil penelitian ini dapat dimanfaatkan oleh pelaku e-commerce sebagai dasar pengambilan keputusan strategis dalam pengelolaan pelanggan dan penyusunan strategi pemasaran berbasis data.

Kata kunci: e-commerce, perilaku belanja konsumen, *machine learning*, klasifikasi, prediksi.

ABSTRACT

The rapid growth of e-commerce has generated large volumes of transaction data that reflect consumer shopping behavior. However, data complexity and class imbalance between subscribed and non-subscribed consumers pose challenges in classification analysis and modeling. This condition may cause classification models to be biased toward the majority class, thereby reducing their ability to accurately capture overall consumer characteristics and resulting in less optimal data-driven marketing strategies. This study aims to predict consumers' subscription status in e-commerce and to compare the performance of several machine learning algorithms in classifying consumer shopping behavior. The dataset used consists of 3,900 records with 18 attributes representing transaction characteristics and shopping behavior. The research methodology includes Exploratory Data Analysis (EDA), data cleaning, Outlier handling using the IQR-based Outlier capping method, categorical feature encoding using One-hot encoding, data splitting using the stratified Train-test split method, Feature scaling using StandardScaler, and balancing the training data using the SMOTE method. Furthermore, the modeling process is conducted using Naive Bayes, Logistic regression, and Random forest algorithms, with performance evaluation based on accuracy, precision, recall, and F1-score metrics. The results indicate that all three algorithms are capable of predicting consumer subscription status with good performance, with Naive Bayes achieving the best performance based on the F1-score metric. The findings of this study can be utilized by e-commerce practitioners as a basis for strategic decision-making in customer management and the development of data-driven marketing strategies.

Keyword: e-commerce, consumer shopping behavior, machine learning, classification, prediction