

**STUDI KOMPARATIF ALGORITMA RANDOM FOREST  
DAN XGBOOST UNTUK ANALISIS PERILAKU  
KONSUMEN E-COMMERCE**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi S1 Informatika



disusun oleh

**AZZILA ISME KAREN**

**20.11.3312**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2024**

**STUDI KOMPARATIF ALGORITMA RANDOM FOREST  
DAN XGBOOST UNTUK ANALISIS PERILAKU  
KONSUMEN E-COMMERCE**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi S1 Informatika



disusun oleh

**AZZILA ISME KAREN**

**20.11.3312**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2024**

**HALAMAN PERSETUJUAN**

**SKRIPSI**

**STUDI KOMPARATIF ALGORITMA RANDOM FOREST  
DAN XGBOOST UNTUK ANALISIS PERILAKU  
KONSUMEN E-COMMERCE**

yang disusun dan diajukan oleh

**AZZILA ISME KAREN**

**20.11.3312**

telah disetujui oleh Dosen Pembimbing Skripsi  
pada tanggal 20 Agustus 2024

Dosen Pembimbing,



**Arif Akbarul Huda, S.Si., M.Eng.**

**NIK. 190302287**

**HALAMAN PENGESAHAN**

**SKRIPSI**

**STUDI KOMPARATIF ALGORITMA RANDOM FOREST  
DAN XGBOOST UNTUK ANALISIS PERILAKU  
KONSUMEN E-COMMERCE**

yang disusun dan diajukan oleh

**AZZILA ISME KAREN**

**20.11.3312**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 20 Agustus 2024

**Susunan Dewan Penguji**

**Nama Penguji**

**Tanda Tangan**

Ike Verawati, M.Kom  
NIK. 190302237

Bayu Setiaji, M.Kom  
NIK. 190302216

Arif Akbarul Huda, S.Si, M.Eng  
NIK. 190302287



Skrripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 20 Agustus 2024

**DEKAN FAKULTAS ILMU KOMPUTER**



Hanif Al Fatta, S.Kom., M.Kom., Ph.D

NIK. 190302096

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

**Nama mahasiswa** : Azzila Isme Karen  
**NIM** : 20.11.3312

Menyatakan bahwa Skripsi dengan judul berikut:

### **STUDI KOMPARATIF ALGORITMA RANDOM FOREST DAN XGBOOST UNTUK ANALISIS PERILAKU KONSUMEN E-COMMERCE**

Dosen Pembimbing : Arif Akbarul Huda, S.Si., 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, 20 Agustus 2024  
Yang Menyatakan,



Azzila Isme Karen

## HALAMAN PERSEMBAHAN

Dengan tulus, penuh penghargaan dan rasa syukur, halaman persembahan ini disusun sebagai ungkapan terima kasih atas dukungan, kontribusi dan dedikasinya atas perjalanan penulisan Skripsi yang berjudul “Studi Komparatif Algoritma *Random Forest* dan *XGBoost* untuk Analisis Perilaku Konsumen *E-commerce*”.

- 1 Tuhan Yang Maha Esa, terima kasih atas rahmat, karunia, dan kekuatan yang telah Engkau anugerahkan selama perjalanan ini.
- 2 Rasulullah *Sallallahu 'Alaihi Wasallam*, motivator dan panutan saya untuk selalu menjadi pengingat untuk memperbaiki diri.
- 3 Ayah dan Ibu tercinta yang tulus memberikan do'a, dukungan, dan dorongan yang tiada henti tercurahkan melalui keringat, usaha dan kekuatan hati yang setia menemani sehingga dapat mencapai titik ini.
- 4 Kepada Paguyuban Perbengkelan yang senantiasa memberikan pesan motivasi dan menerima apapun keadaan saya untuk selalu semangat menjalani hari-hari penuh warna ini.
- 5 Terima kasih kepada adik-adik saya yang selalu meluangkan waktu dengan memberikan berbagai dukungan dan menemani saya melepas penat dikala letih.
- 6 Kepada Bapak Arif Akbarul Huda, S.Si., M.Eng., yang telah memberikan bimbingan, arahan, dan berbagi pengetahuan baru.
- 7 Teman-teman yang telah memberikan semangatnya untuk menyelesaikan skripsi ini.

Dedikasi ini merupakan bentuk penghormatan dan terima kasih yang tulus kepada mereka yang telah berperan penting dalam perjalanan penelitian ini. Semoga setiap kata yang tercurah mampu mencerminkan rasa terima kasih yang mendalam.

## KATA PENGANTAR

Segala puji terhadap *Allah Subhanallahu wa ta'ala* atas rahmat dan karunia-Nya sehingga penulis dapat menuntaskan skripsi bertajuk "STUDI KOMPARATIF ALGORITMA RANDOM FOREST DAN XGBOOST UNTUK ANALISIS PERILAKU KONSUMEN E-COMMERCE" dengan baik. Skripsi ini merupakan tugas akhir serta syarat untuk menyelesaikan studi Program Sarjana (S1) Fakultas Ilmu Komputer Prodi Informatika.

Melalui dukungan dan kontribusi dari berbagai pihak, penulis menuturkan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M., Rektor Universitas Amikom Yogyakarta.
2. Ibu Windha Mega Pradnya D, M.Kom., Ketua Program Studi S1 Informatika.
3. Bapak Arif Akbarul Huda, S.Si., M.Eng., Dosen pembimbing yang telah memberikan arahan, saran dan berbagi pengetahuan baru.

Penulis menyadari bahwa masih terdapat keterbatasan dan kekurangan dalam penyusunan penelitian ini. Penulis mengharapkan kritik dan saran membangun guna membantu penyempurnaan skripsi ini. Semoga penelitian ini memberikan andil yang bermanfaat untuk memahami kinerja algoritma *Random Forest* dan *XGBoost* pada studi perilaku konsumen dalam perkembangan bidang *e-commerce*.

Yogyakarta, 20 Agustus 2024

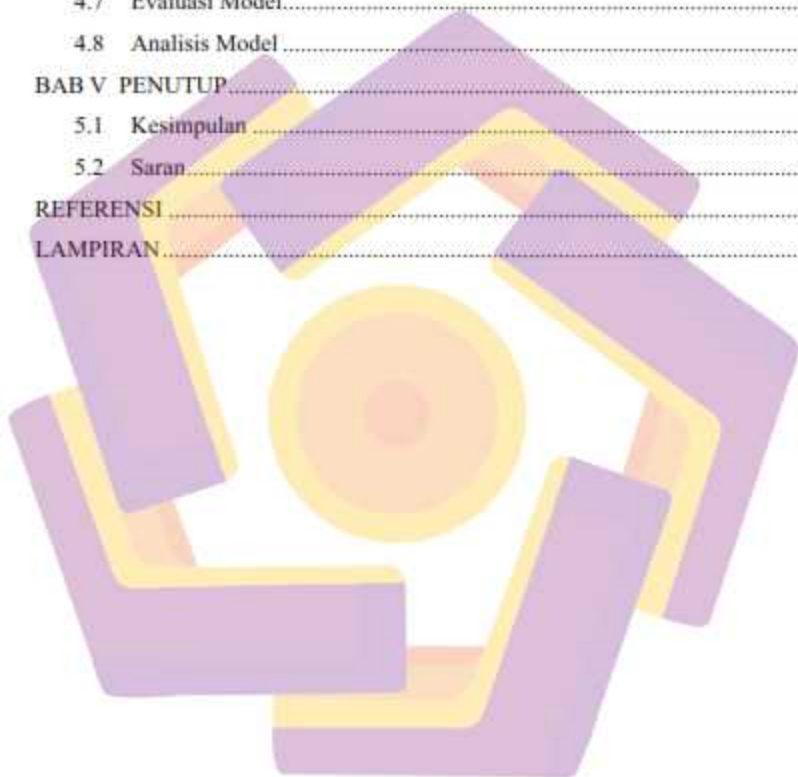
Penulis

## DAFTAR ISI

|  |      |
|--|------|
| HALAMAN JUDUL.....                       | i    |
| HALAMAN PERSETUJUAN.....                 | ii   |
| HALAMAN PENGESAHAN.....                  | iii  |
| HALAMAN PERNYATAAN KEASLIAN SKRIPSI..... | iv   |
| HALAMAN PERSEMBAHAN.....                 | v    |
| KATA PENGANTAR.....                      | vi   |
| DAFTAR ISI.....                          | vii  |
| DAFTAR TABEL.....                        | ix   |
| DAFTAR GAMBAR.....                       | x    |
| DAFTAR LAMBANG DAN SINGKATAN.....        | xii  |
| DAFTAR ISTILAH.....                      | xiii |
| INTISARI.....                            | xiv  |
| <i>ABSTRACT</i> .....                    | xv   |
| <b>BAB I PENDAHULUAN</b> .....           | 1    |
| 1.1 Latar Belakang.....                  | 1    |
| 1.2 Rumusan Masalah.....                 | 2    |
| 1.3 Batasan Masalah.....                 | 2    |
| 1.4 Tujuan Penelitian.....               | 2    |
| 1.5 Manfaat Penelitian.....              | 2    |
| 1.6 Sistematika Penulisan.....           | 3    |
| <b>BAB II TINJAUAN PUSTAKA</b> .....     | 4    |
| 2.1 Studi Literatur.....                 | 4    |
| 2.2 Dasar Teori.....                     | 11   |
| <b>BAB III METODE PENELITIAN</b> .....   | 21   |
| 3.1 Objek Penelitian.....                | 21   |
| 3.2 Alur Penelitian.....                 | 21   |
| 3.3 Alat dan Bahan.....                  | 23   |
| <b>BAB IV HASIL DAN PEMBAHASAN</b> ..... | 25   |
| 4.1 Persiapan Awal.....                  | 25   |

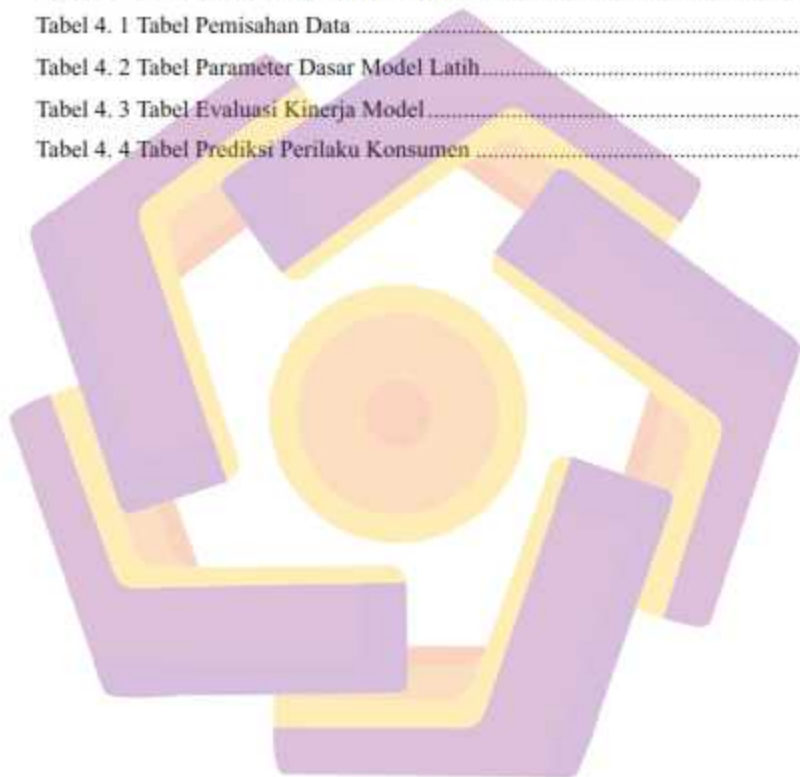


|                    |                                |    |
|--------------------|--------------------------------|----|
| 4.2                | <i>Pre-processing</i> .....    | 27 |
| 4.3                | EDA.....                       | 36 |
| 4.4                | <i>Feature Selection</i> ..... | 37 |
| 4.5                | Pemisahan Data.....            | 37 |
| 4.6                | Pemodelan.....                 | 39 |
| 4.7                | Evaluasi Model.....            | 40 |
| 4.8                | Analisis Model.....            | 45 |
| BAB V PENUTUP..... |                                | 48 |
| 5.1                | Kesimpulan.....                | 48 |
| 5.2                | Saran.....                     | 49 |
| REFERENSI.....     |                                | 50 |
| LAMPIRAN.....      |                                | 54 |



## DAFTAR TABEL

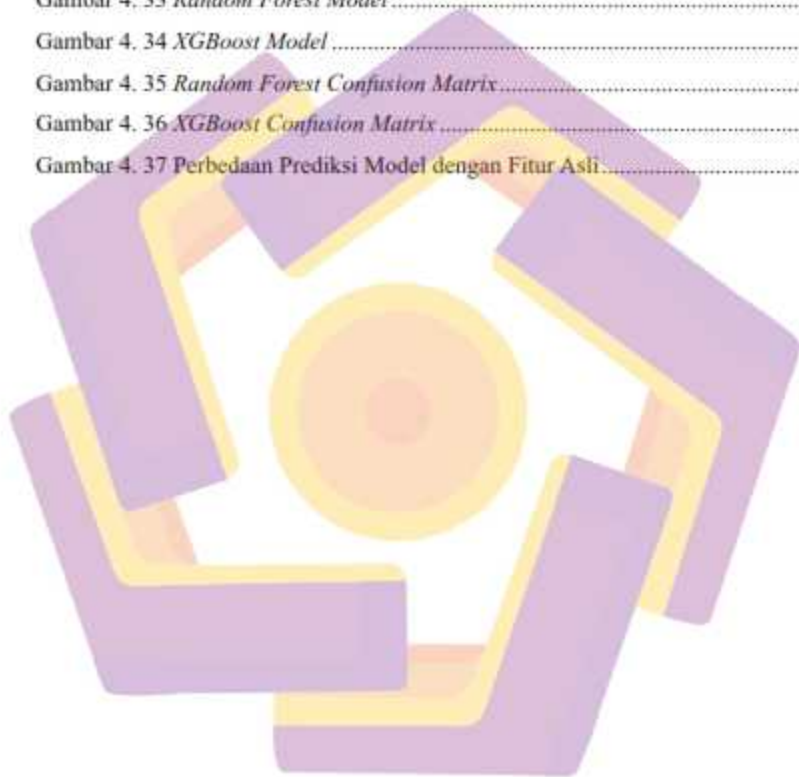
|   |    |
|---|----|
| Tabel 2. 1 Tabel Keaslian Penelitian .....                  | 6  |
| Tabel 2. 2 Tabel Penghitungan <i>Confusion Matrix</i> ..... | 16 |
| Tabel 3. 1 Tabel Atribut <i>Dataset</i> .....               | 21 |
| Tabel 3. 2 Tabel Deskripsi Variabel Penelitian.....         | 24 |
| Tabel 4. 1 Tabel Pemisahan Data .....                       | 39 |
| Tabel 4. 2 Tabel Parameter Dasar Model Latih.....           | 39 |
| Tabel 4. 3 Tabel Evaluasi Kinerja Model.....                | 46 |
| Tabel 4. 4 Tabel Prediksi Perilaku Konsumen .....           | 47 |



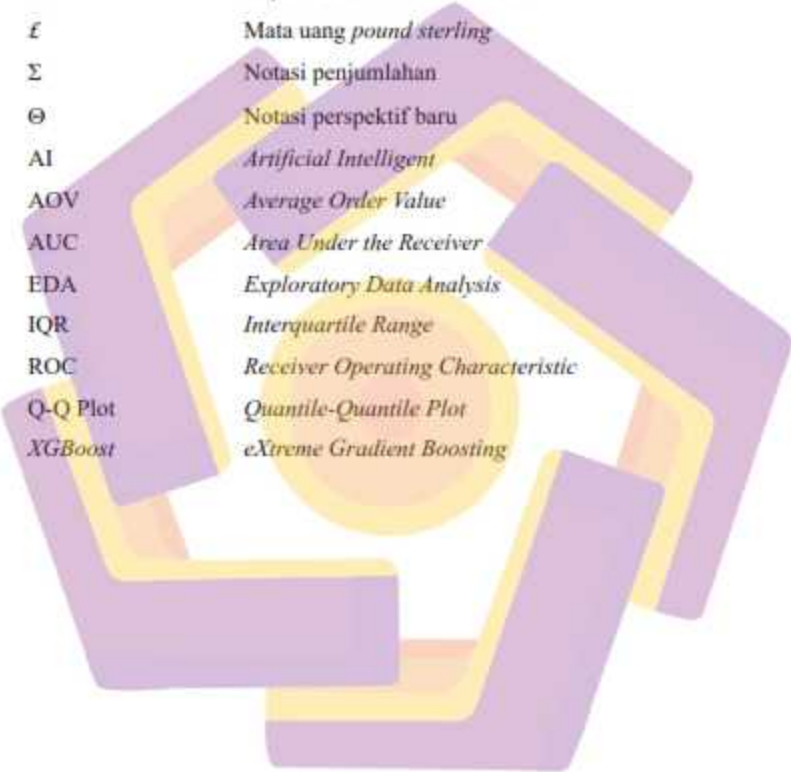
## DAFTAR GAMBAR

|   |    |
|---|----|
| Gambar 3. 1 Alur Penelitian.....                                    | 23 |
| Gambar 3. 2 Bentuk <i>Dataset</i> .....                             | 24 |
| Gambar 4. 1 <i>Mount Dataset</i> .....                              | 26 |
| Gambar 4. 2 Membaca <i>Dataset</i> .....                            | 26 |
| Gambar 4. 3 Tampilan <i>Dataset Ritel Online</i> .....              | 26 |
| Gambar 4. 4 Struktur <i>Dataset Ritel Online</i> .....              | 27 |
| Gambar 4. 5 Riwayat <i>Transaksi Dataset Ritel Online</i> .....     | 27 |
| Gambar 4. 6 Riwayat <i>Pesanan Dibatalkan</i> .....                 | 28 |
| Gambar 4. 7 <i>Persentase Riwayat Pesanan</i> .....                 | 28 |
| Gambar 4. 8 <i>Indeks Karakter Unik</i> .....                       | 29 |
| Gambar 4. 9 <i>Penyaringan Data Entri</i> .....                     | 29 |
| Gambar 4. 10 <i>Pengecekan Missing Values</i> .....                 | 30 |
| Gambar 4. 11 <i>Persentase Missing Values</i> .....                 | 30 |
| Gambar 4. 12 <i>Menangani Missing Values</i> .....                  | 30 |
| Gambar 4. 13 <i>Menangani Duplicated Entries</i> .....              | 31 |
| Gambar 4. 14 <i>Entri Duplikat yang Dihapus</i> .....               | 31 |
| Gambar 4. 15 <i>Visualisasi Outliers menggunakan Q-Q Plot</i> ..... | 31 |
| Gambar 4. 16 <i>Mengangani Outliers pada Fitur Quantity</i> .....   | 32 |
| Gambar 4. 17 <i>Menangani Outliers pada Fitur UnitPrice</i> .....   | 32 |
| Gambar 4. 18 <i>Hasil Menangani Extreme Outliers</i> .....          | 33 |
| Gambar 4. 19 <i>Membuat Fitur Total Pembelian</i> .....             | 33 |
| Gambar 4. 20 <i>Representasi Total Belanja Konsumen</i> .....       | 34 |
| Gambar 4. 21 <i>Membuat DataFrame Baru</i> .....                    | 34 |
| Gambar 4. 22 <i>Tampilan Dataset Ritel Online Baru</i> .....        | 35 |
| Gambar 4. 23 <i>Kriteria Fitur CustomerLoyalty</i> .....            | 35 |
| Gambar 4. 24 <i>Membuat Fitur Loyalitas Pelanggan</i> .....         | 36 |
| Gambar 4. 25 <i>10 Pelanggan Loyal Teratas</i> .....                | 36 |
| Gambar 4. 26 <i>Hubungan Kuantitas dan Harga Satuan</i> .....       | 37 |
| Gambar 4. 27 <i>Struktur Baru Dataset</i> .....                     | 37 |

|   |    |
|---|----|
| Gambar 4. 28 <i>Feature Separation</i> .....                  | 38 |
| Gambar 4. 29 <i>Splitting Data</i> .....                      | 38 |
| Gambar 4. 30 Distribusi Data Asli .....                       | 38 |
| Gambar 4. 31 Menangani Data tidak Seimbang .....              | 39 |
| Gambar 4. 32 Distribusi Data Terkini .....                    | 39 |
| Gambar 4. 33 <i>Random Forest Model</i> .....                 | 39 |
| Gambar 4. 34 <i>XGBoost Model</i> .....                       | 40 |
| Gambar 4. 35 <i>Random Forest Confusion Matrix</i> .....      | 41 |
| Gambar 4. 36 <i>XGBoost Confusion Matrix</i> .....            | 43 |
| Gambar 4. 37 Perbedaan Prediksi Model dengan Fitur Asli ..... | 47 |



## DAFTAR LAMBANG DAN SINGKATAN



|           |  |
|-----------|--|
| $\gamma$  | Mengontrol kompleksitas model                |
| $\lambda$ | Regularisasi L2                              |
| $\Omega$  | Representasi keseluruhan fungsi regularisasi |
| $\omega$  | Representasi bobot individual                |
| $\pounds$ | Mata uang <i>pound sterling</i>              |
| $\Sigma$  | Notasi penjumlahan                           |
| $\Theta$  | Notasi perspektif baru                       |
| AI        | <i>Artificial Intelligent</i>                |
| AOV       | <i>Average Order Value</i>                   |
| AUC       | <i>Area Under the Receiver</i>               |
| EDA       | <i>Exploratory Data Analysis</i>             |
| IQR       | <i>Interquartile Range</i>                   |
| ROC       | <i>Receiver Operating Characteristic</i>     |
| Q-Q Plot  | <i>Quantile-Quantile Plot</i>                |
| XGBoost   | <i>eXtreme Gradient Boosting</i>             |

## DAFTAR ISTILAH

|                            |  |
|----------------------------|--|
| Bias                       | Perbedaan antara nilai prediksi model dan nilai asli.  |
| <i>Credit Card Default</i> | Peminjam yang gagal membayar minimum hutang kartu kredit dalam jangka waktu tertentu.  |
| <i>DataFrame</i>           | Struktur data dua dimensi yang terdiri dari baris dan kolom.   |
| Dataset                    | Kumpulan data yang terstruktur untuk melatih, menguji, dan mengevaluasi model <i>machine learning</i> .  |
| Interpretabilitas          | Kemampuan model untuk memberikan penjelasan yang jelas dan dapat dimengerti atas prediksi atau keputusannya.   |
| <i>Intervening</i>         | Variabel yang mempengaruhi hubungan antara variabel independen dan variabel dependen.  |
| <i>Meta-algorithmic</i>    | Pendekatan tingkat tinggi dengan menggabungkan beberapa algoritma/model untuk mencapai solusi yang lebih kuat menggunakan sistem transparan dan adaptif. |
| <i>Overfitting</i>         | Model terlalu baik untuk memprediksi data latih, namun kurang sensitif untuk memprediksi data baru.  |
| <i>Recency</i>             | Waktu terakhir kali suatu transaksi terjadi.   |
| <i>Recons</i>              | Entri atau baris tunggal berisi informasi atau nilai tertentu yang terdiri dari beberapa bidang atau atribut.  |
| <i>Robust</i>              | Kemampuan model untuk mempertahankan akurasi dan stabilitas dalam menghadapi kondisi tidak ideal.  |
| <i>Skew</i>                | Ukuran asimetri dalam distribusi probabilitas.   |
| Sparsitas                  | Kemampuan model mengidentifikasi dan fokus pada fitur penting dan mengabaikan fitur tidak relevan dalam dataset.   |
| Spesifisitas               | Proporsi contoh negatif yang diidentifikasi dengan benar.  |
| Varians                    | Ukuran seberapa besar variasi prediksi model pada dataset yang sama.   |

## INTISARI

Selama beberapa tahun terakhir, algoritma *Random Forest* dan algoritma *eXtreme Gradient Boosting (XGBoost)* populer untuk analisis perilaku konsumen. Namun, ditengah kepopulerannya, banyak perdebatan mana algoritma yang kompatibel untuk analisis perilaku konsumen di antara keduanya. Hal tersebut dibuktikan oleh beberapa penelitian yang menunjukkan bahwa *Random Forest* kompatibel untuk menangani distribusi data tidak seimbang, sedangkan *XGBoost* kompatibel untuk menangani data besar dan kompleks. Oleh sebab itu, penelitian ini akan membuktikan mana algoritma yang kompatibel dalam menangani *dataset* besar yang tidak seimbang dan kompleks untuk analisis perilaku konsumen *e-commerce* melalui perbandingan kinerja algoritma *Random Forest* dan *XGBoost*. Hasilnya, kedua algoritma tersebut mencapai nilai *accuracy* di atas 98%, dengan algoritma *Random Forest* sedikit lebih unggul (98.43%) dibandingkan algoritma *XGBoost* (98.40%). Begitu pula pada nilai *recall* algoritma *Random Forest* yang dapat mengungguli algoritma *XGBoost* (98.40% vs 98.00%). Sebaliknya, algoritma *XGBoost* lebih unggul daripada algoritma *Random Forest* pada nilai *precision* (97.23% vs 96.48%) dan nilai F1 (97.61% vs 97.43%). Adanya nilai AUC yang sangat tinggi (0.9992 untuk algoritma *Random Forest* dan 0.9993 untuk algoritma *XGBoost*) mengindikasikan bahwa kedua algoritma tersebut mampu membedakan kelas dengan hampir sempurna. Adapun hasil prediksi algoritma *XGBoost* yang sedikit lebih seimbang (253 loyal dan 583 tidak loyal) dibandingkan dengan algoritma *Random Forest* (256 loyal dan 580 tidak loyal) membuktikan bahwa algoritma *XGBoost* berkinerja lebih optimal untuk menangani *dataset* besar yang tidak seimbang dan kompleks. Berdasarkan hasil tersebut, penelitian ini dapat memberikan kontribusi pada pengembangan metode analisis perilaku konsumen pada industri *e-commerce* dan meningkatkan kemampuan algoritma *machine learning* untuk analisis perilaku konsumen *e-commerce*.

**Kata kunci:** Studi Komparatif, *Random Forest*, *XGBoost*, *E-Commerce*, Analisis Perilaku Konsumen

## ABSTRACT

Over the past few years, Random Forest algorithm and eXtreme Gradient Boosting (XGBoost) algorithm are popular for consumer behavior analysis. However, amidst their popularity, there is a lot of debate on which algorithm is compatible for consumer behavior analysis between the two. This is evidenced by several studies that show that Random Forest is compatible for handling unbalanced data distribution, while XGBoost is compatible for handling large and complex data. Therefore, this study will prove which algorithm is compatible in handling large unbalanced and complex datasets for e-commerce consumer behavior analysis through a performance comparison of Random Forest and XGBoost algorithms. As a result, both algorithms achieved accuracy values above 98%, with the Random Forest algorithm slightly superior (98.43%) to the XGBoost algorithm (98.40%). Similarly, the recall value of the Random Forest algorithm can outperform the XGBoost algorithm (98.40% vs 98.00%). In contrast, the XGBoost algorithm is superior to the Random Forest algorithm in the precision value (97.23% vs 96.48%) and F1 value (97.61% vs 97.43%). The existence of very high AUC values (0.9992 for Random Forest algorithm and 0.9993 for XGBoost algorithm) indicates that both algorithms are able to distinguish classes almost perfectly. The slightly more balanced prediction results of the XGBoost algorithm (253 loyal and 583 non-loyal) compared to the Random Forest algorithm (256 loyal and 580 non-loyal) prove that the XGBoost algorithm performs more optimally for handling large unbalanced and complex datasets. Based on these results, this research can contribute to the development of consumer behavior analysis methods in the e-commerce industry and improve the ability of machine learning algorithms for e-commerce consumer behavior analysis.

**Keywords:** Comparative Study, Random Forest, XGBoost, E-Commerce, Consumer Behavior Analysis