Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

Studi Komparasi Kinerja Algoritma AdaBoost dan CatBoost dalam Prediksi Perilaku Pembelian Pelanggan

Princess Iqlima Kafilla*, Fandy Setyo Utomo, Giat Karyono

Fakultas Ilmu Komputer, Magister Ilmu Komputer, Universitas Amikom Purwokerto, Banyumas, Indonesia
Email: [1,*]princess19iqlima@gmail.com, [2]fandy_setyo_utomo@amikompurwokerto.ac.id,
[3]giatkaryono@amikompurwokerto.ac.id
Email Penulis Korespondensi: princess19iqlima@gmail.com
Submitted: **10/07/2025**; Accepted: **31/03/2026**; Published**: 31/03/2026**

Abstrak−Perilaku pembelian pelanggan merupakan faktor krusial dalam pengembangan strategi pemasaran yang efektif. Dengan memanfaatkan analisis prediktif, bisnis dapat mempersonalisasi rekomendasi, mengoptimalkan kampanye pemasaran, serta meningkatkan pengalaman pengguna, yang pada akhirnya berkontribusi pada peningkatan tingkat konversi dan retensi pelanggan. Penelitian ini bertujuan untuk membandingkan kinerja algoritma AdaBoost dan CatBoost dalam memprediksi perilaku pembelian pelanggan. Dataset yang digunakan mencakup atribut demografis dan riwayat perilaku pelanggan, sehingga memungkinkan analisis yang komprehensif. Hasilnya menunjukkan bahwa CatBoost memberikan performa yang lebih baik secara keseluruhan dengan akurasi sebesar 94%, sedangkan AdaBoost menunjukkan nilai recall dan F1-score yang lebih tinggi pada kelas positif. Penelitian ini menyimpulkan bahwa kedua algoritma memiliki keandalan dalam memprediksi perilaku pelanggan, di mana CatBoost lebih unggul dalam menangani fitur kategorikal, sementara AdaBoost menawarkan adaptabilitas yang baik pada dataset yang lebih sederhana. Sebagai langkah lanjut, penelitian mendatang dapat mengeksplorasi implementasi model-model ini dalam skenario real-time.

Kata Kunci: AdaBoost; Analisis Prediktif; CatBoost; Perilaku Pembelian Pelanggan; Personalisasi Pemasaran; Strategi Pemasaran

Abstrak−Customer purchase behavior is a crucial factor in the development of effective marketing strategies. By leveraging predictive analytics, businesses can personalize recommendations, optimize marketing campaigns and improve user experience, ultimately contributing to increased conversion rates and customer retention. This research compares the performance of AdaBoost and CatBoost algorithms in predicting customer purchase behavior. The dataset used includes demographic attributes and customer behavior history, allowing for comprehensive analysis. The results showed that CatBoost performed better overall with an accuracy of 94%, while AdaBoost showed higher recall and F1-score values in the positive class. This study concludes that both algorithms have reliability in predicting customer behavior, where CatBoost is superior in handling categorical features, while AdaBoost offers good adaptability on simpler datasets. As a next step, future research can explore the implementation of these models in real-time scenarios.

Keywords: AdaBoost; CatBoost; Customer Behavior; Marketing Personalization; Marketing Strategy; Predictive Analytics

1. PENDAHULUAN

Sekarang ini, kita bisa dengan gampang mendapatkan berbagai informasi lewat internet. Internet menjadi kebutuhan yang sangat penting bagi sebagian besar orang, selain untuk memenuhi kebutuhan seperti pakaian, makanan, dan tempat tinggal. Dengan perkembangan teknologi informasi yang sangat cepat, telah terjadi perubahan dalam cara hidup setiap orang, terutama dalam berkomunikasi [1]. Pada awal tahun 2023, ada sebanyak 212,9 juta orang di Indonesia yang menggunakan internet, meningkat 3,85 persen dibandingkan tahun sebelumnya. Jumlah penduduk Indonesia pada tahun 2023 mencapai 278,69 juta orang. Dari jumlah tersebut, sekitar 77 persen penduduk Indonesia sudah bisa mengakses internet [2].

Populasi orang Indonesia yang menggunakan internet terus bertambah, sehingga membuat pasar e-commerce semakin menjanjikan. Saat ini, banyak platform seperti marketplace dan aplikasi belanja online muncul, menawarkan berbagai produk dan layanan untuk memenuhi kebutuhan konsumen yang terus berubah [3]. Perkembangan teknologi dan peningkatan akses internet telah mengubah cara orang berbelanja, dengan semakin banyak konsumen beralih menggunakan platform e-commerce untuk memenuhi kebutuhan sehari-hari mereka [4]. Transformasi digital mengubah cara kita membangun ekonomi dengan fokus pada pengetahuan. Saat ini, teknologi digital melahirkan pendekatan baru dalam ekonomi, seperti ekonomi kreatif, ekonomi berbasis jaringan, ekonomi sosial, dan ekonomi platform [5].

Pelanggan atau customer adalah orang atau kelompok yang membayar produk atau jasa dari toko atau bisnis tertentu. Mereka merupakan bagian yang sangat penting dalam membantu pertumbuhan perusahaan, sehingga perlu diberi perhatian khusus saat melakukan analisis tentang pelanggan [6]. Cara pelanggan membeli barang atau jasa merupakan hal yang sangat penting dalam membuat strategi pemasaran yang berhasil. Keputusan pembelian merupakan salah satu aspek dalam perilaku konsumen. Perilaku konsumen sendiri mempelajari bagaimana individu, kelompok, maupun organisasi dalam menentukan pilihan, membeli, menggunakan, hingga mengevaluasi produk berupa barang, jasa, ide, atau pengalaman demi memenuhi kebutuhan dan keinginan mereka. Dalam proses pembelian, konsumen umumnya melewati lima tahapan, yaitu: mengenali masalah, mencari informasi, mengevaluasi berbagai alternatif, mengambil keputusan untuk membeli, dan akhirnya menunjukkan perilaku setelah pembelian [7]. Dalam era digital yang ditandai dengan perkembangan teknologi dan ketersediaan data besar (big data), analisis perilaku pelanggan semakin mendalam dan akurat. Hal ini memungkinkan perusahaan untuk mengidentifikasi tren, preferensi,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

serta kebutuhan pelanggan secara lebih presisi, sehingga dapat menghasilkan keputusan bisnis yang lebih tepat sasaran [8] [9].

Analisis perilaku pelanggan melibatkan studi tentang bagaimana konsumen berinteraksi dengan platform online, mulai dari menjelajahi dan menambahkan produk ke keranjang mereka hingga melakukan pembelian atau meninggalkan proses tersebut. Dengan menganalisis pola-pola ini, bisnis dapat mendapatkan wawasan yang berharga tentang preferensi, motivasi, dan poin-poin penting bagi pelanggan. Salah satu aplikasi analisis perilaku pelanggan yang paling kuat adalah prediksi pembelian, yang memanfaatkan data historis dan teknik pembelajaran mesin untuk meramalkan kemungkinan pelanggan melakukan pembelian. Analisis prediktif membantu bisnis mempersonalisasi rekomendasi, mengoptimalkan kampanye pemasaran, dan meningkatkan pengalaman pengguna, yang pada akhirnya meningkatkan tingkat konversi dan retensi pelanggan [10].

Penggunaan teknik machine learning telah menjadi salah satu pendekatan yang relevan dalam memprediksi perilaku pembelian pelanggan. Machine Learning (ML) adalah cabang ilmu yang berfokus pada pengembangan dan evaluasi algoritma yang memungkinkan komputer untuk memperoleh kemampuan belajar secara mandiri [11]. Algoritma machine learning menggunakan data historis untuk mengenali pola dan memprediksi kejadian di masa depan. Proses pembelajarannya terdiri dari dua tahap utama, yakni tahap pelatihan (training) dan tahap pengujian (testing). Salah satu fungsi utama dalam machine learning adalah klasifikasi multi-kelas (multi-class classification), yaitu proses di mana model mengelompokkan data ke dalam lebih dari dua kategori [8].

Beberapa penelitian telah mengkaji penggunaan algoritma machine learning, khususnya metode ensemble learning dalam memprediksi perilaku pembelian pelanggan. Metode ensemble merupakan pendekatan dalam pembelajaran mesin yang menggabungkan beberapa algoritma pembelajaran untuk memperoleh hasil prediksi yang lebih akurat dibandingkan penggunaan satu algoritma tunggal. Dengan mengkombinasikan berbagai model, metode ini bertujuan meningkatkan kinerja dan keandalan sistem prediksi. Salah satu teknik dalam metode ensemble yang cukup populer adalah AdaBoost (Adaptive Boosting) dan Catboost [11].

Beberapa penelitian terdahulu telah mengeksplorasi penggunaan algoritma machine learning untuk prediksi perilaku pelanggan. Penelitian yang dilakukan oleh Lalwani dkk, tahun 2022 dalam bidang sistem informasi dan analisis data pelanggan mengevaluasi enam model machine learning, yaitu Logistic Regression, Random Forest, Gradient Boosting, SVM, KNN, dan XGBoost untuk prediksi perilaku pembelian. Hasilnya menunjukkan bahwa model ensemble seperti Random Forest dan XGBoost memberikan performa terbaik dengan akurasi di atas 94% [12]. Penelitian lain yang dilakukan oleh Deniz dan Bulbul, tahun 2024 dalam bidang teknologi informasi dan bisnis berbasis data fokus pada prediksi churn pelanggan dan membandingkan empat algoritma boosting yaitu Gradient Boosting, AdaBoost, XGBoost, dan LightGBM. Penelitian ini menyimpulkan bahwa Gradient Boosting mencapai akurasi tertinggi sebesar 92,94%, diikuti oleh XGBoost, AdaBoost, dan LightGBM [9]. Dalam konteks yang lebih spesifik pada algoritma boosting untuk data e-commerce, penelitian yang dilakukan oleh Lalwani dkk membandingkan berbagai model termasuk CatBoost dan XGBoost, dan menemukan bahwa keduanya memberikan performa prediksi terbaik dengan skor F1 masing-masing 0.93 dan 0.92, serta ROC AUC tinggi, dalam tugas prediksi niat pembelian konsumen. Selaras dengan itu, penelitian yang dilakukan oleh Lin pada tahun 2025 yang berfokus pada prediksi churn dalam industri telekomunikasi melaporkan bahwa teknik ensemble seperti AdaBoost dan XGBoost memberikan hasil yang sangat baik, dengan skor AUC tertinggi mencapai 84% [13]. Temuan-temuan ini menunjukkan konsistensi efektivitas algoritma boosting, terutama XGBoost dan varian terkaitnya (seperti CatBoost), dalam berbagai skenario prediksi perilaku pelanggan, baik itu churn maupun niat pembelian, sehingga mendasari relevansi penelitian komparatif terhadap algoritma seperti AdaBoost dan CatBoost.

Meskipun penelitian-penelitian terdahulu memberikan wawasan penting tentang efektivitas berbagai model machine learning, termasuk teknik boosting, dalam memprediksi perilaku pelanggan, perbandingan langsung dan komprehensif antara algoritma AdaBoost dan CatBoost dalam konteks prediksi perilaku pembelian pelanggan e-commerce masih belum sepenuhnya dieksplorasi secara eksplisit. Seperti penelitian yang dilakukan oleh Deniz dan Bulbul maupun Lalwani dkk yang menunjukkan performa terbaik untuk model berbasis boosting (XGBoost, Gradient Boosting, CatBoost), sementara penelitian Deniz dan Bulbul pada 2024 secara khusus menyebutkan kinerja baik AdaBoost dalam prediksi churn, sementara penelitian Lalwani dkk pada tahun 2022 menunjukkan keunggulan CatBoost dalam prediksi perilaku pembelian. Namun, belum ada penelitian yang secara khusus difokuskan untuk membandingkan kedua algoritma ini (AdaBoost dan CatBoost) dalam skenario yang sama untuk prediksi perilaku pembelian, sambil mengevaluasi kekuatan unik masing-masing.

Algoritma Adaptive Boosting (AdaBoost) merupakan algoritma supervised learning yang kuat dalam membangun model klasifikasi dan terbukti efektif dalam menangani masalah ketidakseimbangan kelas dengan memberikan bobot berbeda pada data pelatihan di setiap iterasi untuk meningkatkan akurasi [13]. Sementara itu, CatBoost terkenal dengan kemampuannya menangani fitur kategorikal secara langsung tanpa perlu proses encoding tambahan, yang meningkatkan efisiensi komputasi dan mengurangi potensi bias transformasi data. CatBoost juga dirancang untuk mengurangi overfitting melalui pendekatan gradient boosting berbasis pohon simetris dan menawarkan pemrosesan yang lebih cepat pada dataset besar [14].

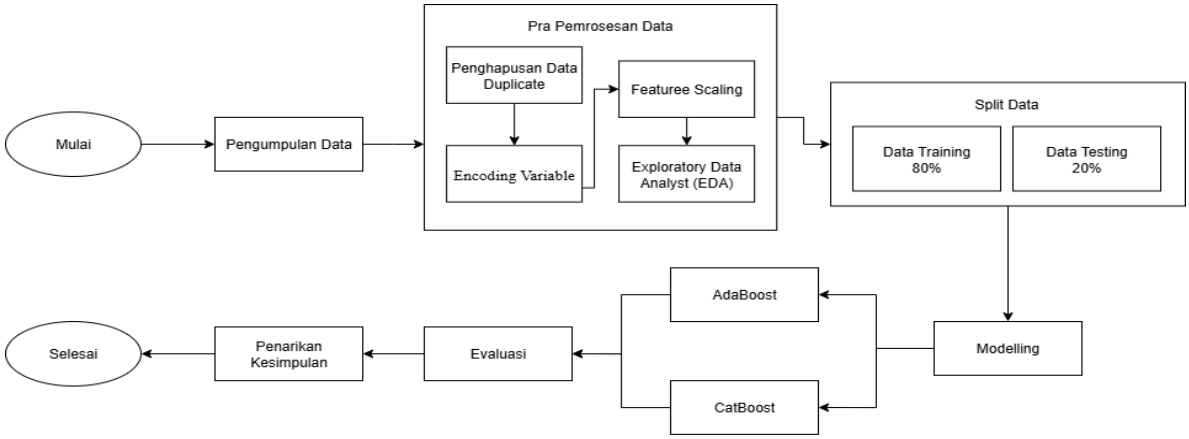
Untuk menjawab celah yang belum terpenuhi dalam studi sebelumnya, penelitian ini membandingkan kinerja dua algoritma boosting yang kurang tereksplorasi secara spesifik dalam prediksi perilaku pembelian pelanggan, yaitu AdaBoost dan CatBoost. Dengan menggunakan dataset yang sama seperti pada penelitian Deniz dan Bulbul maupun Sukrimaa dkk, penelitian ini bertujuan untuk mengevaluasi dan membandingkan kemampuan kedua algoritma dalam

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

memprediksi keputusan pembelian pelanggan berdasarkan matrik evaluasi standar seperti precision, recall, F1-score, dan akurasi. Selain itu, penelitian ini juga mencari model terbaik yang bisa memberikan prediksi lebih baik dibandingkan penelitian sebelumnya. Penelitian ini juga mengidentifikasi fitur-fitur dari pelanggan yang paling berpengaruh dalam proses mereka memutuskan untuk membeli. Hasil penelitian diharapkan bisa menjadi dasar analisis yang kuat bagi perusahaan untuk merancang strategi pemasaran yang lebih efektif, memberikan layanan yang lebih sesuai dengan kebutuhan pelanggan, serta meningkatkan loyalitas dan jumlah pelanggan yang melakukan pembelian.

2. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan melalui beberapa tahapan utama yang sistematis, dimulai dari import data hingga penarikan kesimpulan. Secara garis besar, seluruh alur proses penelitian yang dilakukan dirangkum secara visual pada diagram alir dalam Gambar 1.



Gambar 1. Metode Penelitian

2.1 Pengumpulan Data

Proses pengumpulan data dilakukan dengan cara mengakses dan mengunduh dataset dari situs Kaggle, yang menyediakan akses ke berbagai dataset publik untuk analisis. Pada penelitian ini, dataset yang digunakan adalah dataset asli milik Mr. Rabie El Kharoua yang dipublikasikan di repository tersebut. Dataset ini telah dipilih berdasarkan relevansinya dengan topik yang diteliti dan mencakup informasi yang diperlukan untuk mencapai tujuan penelitian [12].

2.2 *Preprocessing* Data

Preprocessing data ialah melakukan pengolahan data agar data yang salah dapat diperbaiki atau dibersihkan sehingga data dapat digunakan [14]. Tujuan dari langkah ini adalah memastikan kualitas data dalam kondisi terbaik sebelum digunakan dalam analisis atau pembuatan model selanjutnya [12]. Langkah-langkah preprocessing yang diterapkan dalam penelitian ini dijelaskan sebagai berikut:

a. Penghapusan Data Duplikat
   Salah satu langkah awal yang dilakukan dalam pra pemrosesan data adalah menghapus data duplikat, dengan tujuan agar model hanya membaca informasi yang unik dan tidak terpengaruh oleh data yang berulang [15].
b. *Encoding Variable*
   Transformasi variabel kategorikal ke dalam bentuk numerik merupakan langkah penting dalam pra-pemrosesan data untuk pembelajaran mesin. Salah satu metode yang digunakan adalah *label encoding*, yaitu mengubah setiap kategori dalam variabel menjadi label numerik yang unik [12]. Langkah ini penting untuk memungkinkan model memproses variabel kategorikal secara efektif dan akurat [9].
c. *Feature Scaling*
   Tahap ini bertujuan untuk menormalkan variabel numerik, seperti *Age*, *Annual Income*, *Number of Purchases*, dan *Time Spent on Website*, ke dalam skala standar. Normalisasi ini penting dilakukan untuk mencegah fitur dengan rentang nilai yang lebih besar mendominasi proses pembelajaran model secara tidak proporsional [12][9].
d. *Exploratory Data Analyst (EDA)*
   Exploratory Data Analysis (EDA) dilakukan untuk memahami struktur dan distribusi dasar dari data. Tahapan ini mencakup pembuatan statistik ringkasan, visualisasi distribusi data melalui histogram dan box plot, serta identifikasi korelasi antar fitur menggunakan *correlation heatmap*. EDA membantu mengungkap pola dan hubungan antar variabel yang berperan penting dalam proses *feature engineering* dan pemilihan model yang sesuai [9].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

e. *Feature Engineering*
Feature Engineering melibatkan pembuatan fitur baru atau mengubah fitur yang sudah ada untuk meningkatkan daya prediksi model [9].

2.3 *Modelling*

Pada penelitian ini, dilakukan tahap pemodelan dengan memanfaatkan metode ensemble untuk meningkatkan performa prediksi [12]. Metode ensemble yang digunakan yaitu AdaBoost dan CatBoost, karena keduanya memiliki keunggulan dalam meningkatkan akurasi model. AdaBoost bekerja secara iteratif dengan melatih serangkaian model lemah, di mana setiap model difokuskan pada contoh-contoh yang sebelumnya salah diklasifikasikan atau memiliki kesalahan prediksi terbesar. Dengan memberikan bobot lebih besar pada data yang sulit, model-model selanjutnya dapat secara efektif memperbaiki kesalahan klasifikasi dan meningkatkan akurasi keseluruhan [16].

2.4 Evaluasi Hasil

Dalam konteks penelitian ini, istilah True Positive (TP) merujuk pada pelanggan yang benar-benar melakukan pembelian dan berhasil diprediksi sebagai pembeli oleh model. True Negative (TN) menggambarkan pelanggan yang tidak melakukan pembelian dan juga diprediksi dengan tepat sebagai non-pembeli. False Positive (FP) terjadi ketika model memprediksi pelanggan akan melakukan pembelian, padahal sebenarnya tidak. Sedangkan False Negative (FN) menunjukkan kondisi di mana pelanggan seharusnya melakukan pembelian, namun oleh model diklasifikasikan sebagai tidak membeli [17].

a. *Accuracy*
Setelah pengujian, data akan dievaluasi untuk mengukur keakuratan model, salah satunya dengan melihat nilai *Accuracy*. Evaluasi ini bertujuan untuk memahami seberapa baik model dalam membuat prediksi atau mengklasifikasikan data dengan benar. Untuk mengevaluasi sejauh mana model dapat memprediksi perilaku pembelian pelanggan, akurasi menjadi metrik utama yang digunakan Formula 1 [18].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

b. *Precision*
Mengacu pada Formula 2, precision dihitung dengan membandingkan jumlah True Positive terhadap total kombinasi antara True Positive dan False Positive. Hal ini menunjukkan seberapa akurat prediksi positif yang dibuat oleh model dibandingkan dengan seluruh prediksi yang diklasifikasikan sebagai positif [18].

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

c. *Recall*
Nilai recall diperoleh dengan membagi jumlah True Positive dengan total True Positive dan False Negative, sehingga mencerminkan kemampuan model dalam mendeteksi semua kasus positif secara akurat Formula 3 [18].

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

d. *F1-Score*
Langkah berikutnya adalah menghitung F1-Score, yang merepresentasikan rata-rata harmonis dari precision dan recall. Perumusannya disajikan pada Formula 4 [19].

$$F1 - Score = \frac{Precission\ x\ Recall}{Precission+Recall} \quad (4)$$

e. *ROC-AUC*
ROC-AUC mengukur sejauh mana model mampu membedakan antara kelas positif yang benar dan kelas positif yang salah. Skor ROC-AUC berada dalam rentang 0 hingga 1, di mana nilai 0 mencerminkan performa model yang sangat buruk (tidak dapat membedakan antara kelas positif dan negatif), sedangkan nilai 1 menunjukkan performa sempurna (model sepenuhnya mampu membedakan kedua kelas tersebut) [20].

$$AUC = \int TPR(FPR)\, d(FPR) \quad (5)$$

Kelima metrik evaluasi ini akan digunakan sebagai tolak ukur utama untuk membandingkan performa model klasifikasi *AdaBoost* dan *CatBoost* secara komprehensif. Hasil perbandingan ini akan menjadi dasar untuk menentukan model mana yang paling efektif dan andal dalam memprediksi pembelian pelanggan untuk kasus studi ini.

3. HASIL DAN PEMBAHASAN

Hasil penelitian ini berfokus pada evaluasi model prediksi pembelian pelanggan menggunakan algoritma *machine learning* yaitu AdaBoost dan CatBoost. Untuk mengevaluasi performa model, data dibagi menjadi 2 subset yaitu data pelatihan sebesar 80% dan data pengujian sebesar 20%. Model dilatih menggunakan subset data pelatihan dan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

divalidasi menggunakan subset data pengujian. *Feature Scalling* diterapkan untuk memastikan bahwa fitur numerik berada pada skala yang sebanding. Untuk variabel kategorikal menggunakan *one-hot endcoder*. Adapun matriks yang digunakan untuk mengukur kinerja model meliputi *accuracy, precission, recall, f1-score, dan Receiver Operating Characteristic Curve* (ROC-AUC).

3.1 Dataset

Tabel 2 menampilkan beberapa contoh isi dataset yang akan digunakan dan selanjutnya akan dilakukan preprocessing data pada penelitian ini.

Tabel 2. Dataset Penelitian

Age	Gender	Annual Income	Number of Purchases	….	PurchaseStatus
40	1	66120.26793867795	8		1
20	1	23579.773583030514	4		0
27	1	127821.3064316501	11		1
24	1	137798.62311954965	19		1
31	1	99300.96422033315	19		1
66	1	37758.11747465244	14		0
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
39	1	65048.14183385337	13		1
67	1	28775.331068896852	18		1
40	1	57363.247540525364	7		0
63	0	134021.77553236455	16		1
50	0	52625.66597423874	13		1

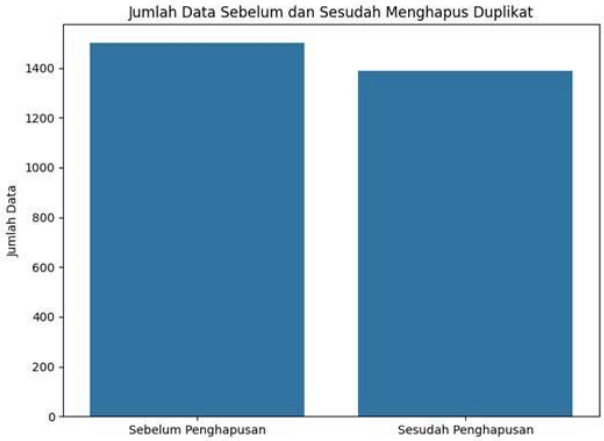
Target variabel yaitu atribut PurchaseStatus yang menunjukkan kemungkinan pelanggan untuk melakukan pembelian, dengan dua kategori utama:

a. 0 (No Purchase): Pelanggan tidak melakukan pembelian, yang mencakup 48% dari total data.
b. 1 (Purchase): Pelanggan melakukan pembelian, yang mencakup 52% dari total data.

Distribusi ini menunjukkan bahwa data relatif seimbang antara pelanggan yang melakukan pembelian dan yang tidak melakukan pembelian [12].

3.2 Preprocessing Data

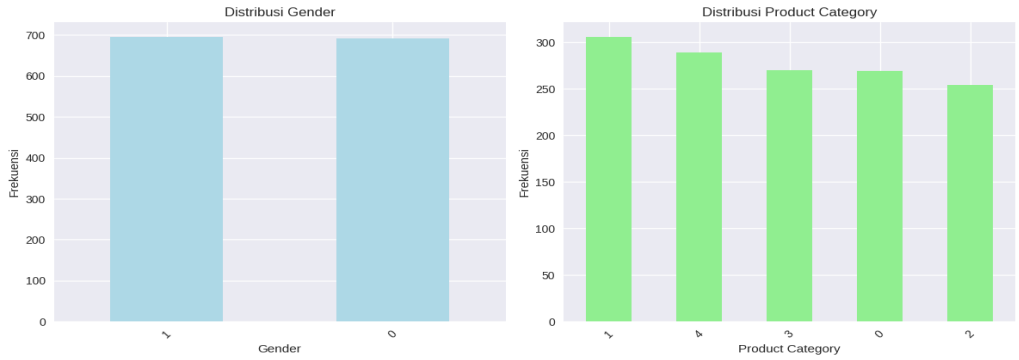
a. *Penghapusan Data Duplikat*
Pada tahap ini, dilakukan penghapusan data duplikat yang terindikasi dalam dataset. Pada Gambar 2 terdapat visualisasi dari total 1500 data, ditemukan sebanyak 112 data duplikat. Oleh karena itu, data duplikat tersebut dihapus sehingga jumlah data yang tersisa menjadi 1388 data unik. Proses ini bertujuan untuk memastikan integritas data dan mencegah pengaruh negatif pada hasil analisis akibat redundansi informasi.



Gambar 2. Jumlah Data Sebelum dan Sesudah Menghapus Data Duplikat

b. *Encoding Variable*
Proses *encoding variable* dilakukan untuk mengubah data non-numerik menjadi bentuk numerik agar dapat diproses oleh algoritma *machine learning*. Pada penelitian ini, teknik *one-hot encoding* digunakan untuk menangani variabel kategorikal, yaitu Gender dan Product Category, sehingga masing-masing kategori pada variabel tersebut direpresentasikan dalam bentuk fitur biner yang terpisah [9].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

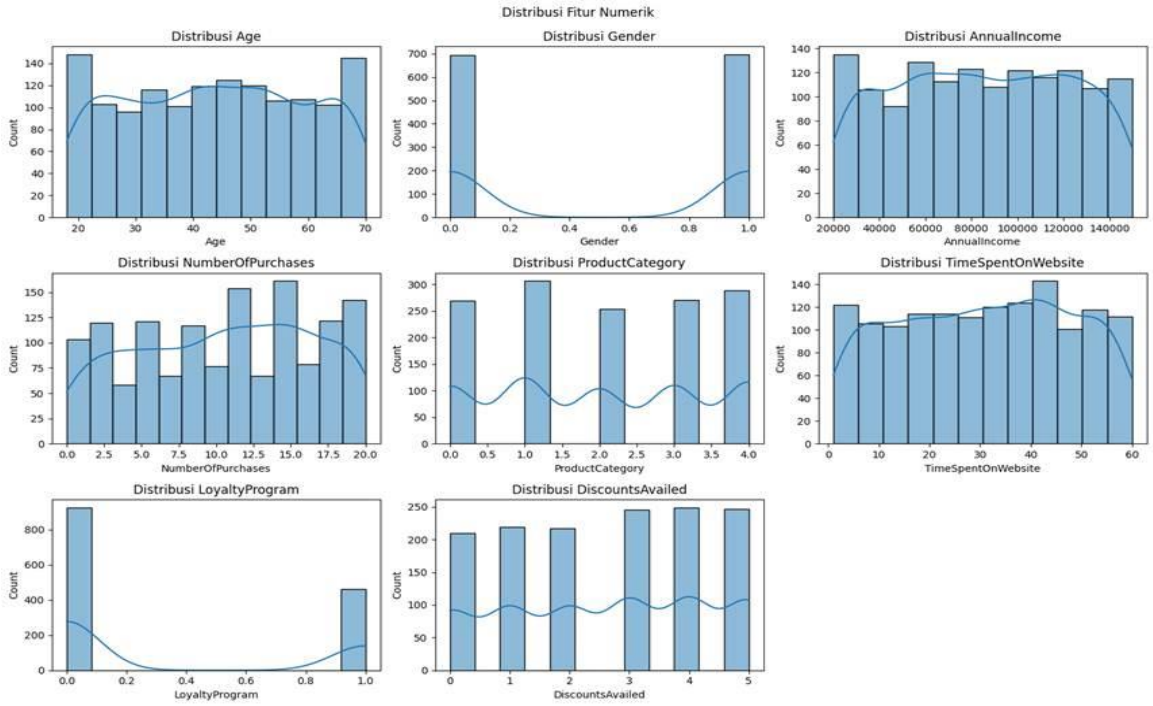


Gambar 3. Distribusi fitur kategorikal yang telah dilakukan *one-hot encoding*

c. *Feature Scaling*
Proses *feature scaling* dilakukan untuk menormalkan variabel numerik seperti Age, Annual Income, Number of Purchases, dan Time Spent on Website ke dalam skala yang seragam. Pada penelitian ini, normalisasi dilakukan menggunakan metode StandardScaler. Langkah ini penting agar variabel dengan rentang nilai yang lebih besar tidak memberikan pengaruh yang berlebihan terhadap kinerja model selama proses pelatihan [9].

d. *Exploratory Data Analyst (EDA)*
Exploratory Data Analysis (EDA) atau analisis data eksploratif merupakan tahap awal yang krusial dalam proses penelitian untuk memahami karakteristik, pola, anomali, dan hubungan yang terdapat dalam dataset. Pada tahap ini, dilakukan investigasi terhadap data sebelum proses pemodelan untuk mendapatkan wawasan mendalam mengenai setiap fitur dan relasinya dengan variabel target, yaitu *PurchaseStatus* (Status Pembelian). Tujuan utama dari EDA dalam penelitian ini adalah untuk mengidentifikasi fitur-fitur yang paling potensial mempengaruhi keputusan pembelian pelanggan. Proses EDA melibatkan dua teknik utama: analisis distribusi untuk memahami sebaran nilai setiap fitur dan analisis korelasi untuk mengukur kekuatan hubungan antar fitur.

1. Analisis Distribusi Fitur
Analisis distribusi dilakukan untuk memahami sebaran data dari setiap fitur numerik. Visualisasi distribusi ini penting untuk mengetahui tendensi sentral, penyebaran, dan potensi adanya data yang tidak seimbang (*imbalanced*). Gambar 4 menyajikan distribusi dari setiap fitur.



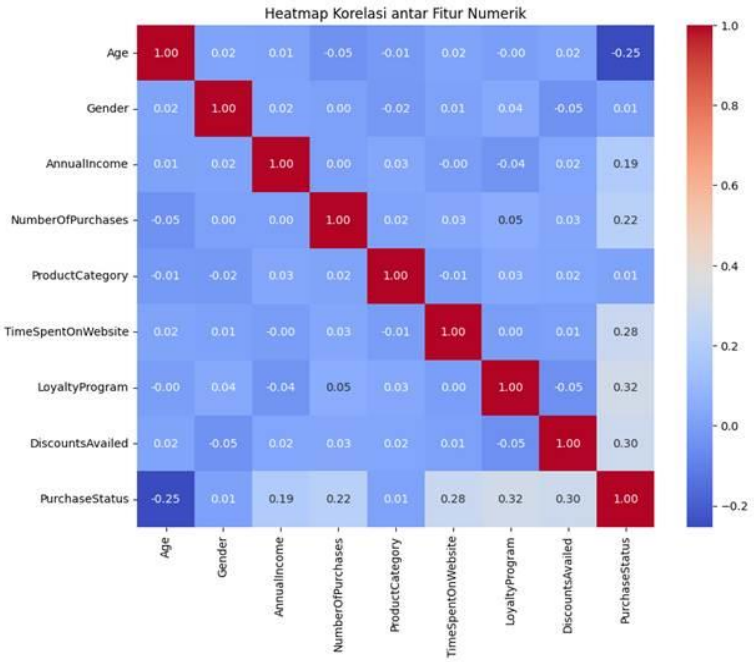
Gambar 4. Distribusi Fitur

Berdasarkan visualisasi pada Gambar 4, distribusi data menunjukkan keragaman sampel yang baik. Fitur seperti *Age*, *NumberOfPurchases*, dan *TimeSpentOnWebsite* memiliki sebaran yang cenderung seragam, sementara AnnualIncome mendekati distribusi normal. Fitur kategorikal seperti *ProductCategory* dan *DiscountsAvailed* juga menunjukkan jumlah sampel yang relatif seimbang di setiap kategorinya. Selain itu, teridentifikasi adanya ketidakseimbangan data (imbalance) pada fitur biner Gender dan *LoyaltyProgram*, di

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

mana jumlah pelanggan yang tidak mengikuti program loyalitas secara signifikan lebih banyak daripada yang mengikutinya namun hal itu tidak memberikan dampak dikarenakan target atributnya adalah PurchaseStatus.

2. Analisis Korelasi Antar Fitur
   Analisis korelasi dilakukan untuk mengukur tingkat dan arah hubungan linear antara pasangan fitur numerik. Heatmap korelasi divisualisasikan pada Gambar 5. Heatmap ini digunakan untuk memvisualisasikan matriks korelasi, di mana warna yang lebih terang atau lebih gelap menunjukkan kekuatan hubungan. Analisis ini sangat penting untuk mengidentifikasi fitur mana yang paling kuat hubungannya dengan variabel target *PurchaseStatus*.



Gambar 5. Korelasi Antar Fitur

Heatmap pada Gambar 5 menyajikan analisis korelasi Pearson untuk mengidentifikasi fitur yang paling berpengaruh terhadap *PurchaseStatus*. Hasilnya menunjukkan bahwa *LoyaltyProgram* (r=0.32), *DiscountsAvailed* (r=0.30), dan TimeSpentOnWebsite (r=0.28) memiliki korelasi positif sedang, yang mengindikasikan ketiganya adalah prediktor potensial yang kuat untuk terjadinya pembelian. Di sisi lain, fitur Age (Usia) menunjukkan korelasi negatif lemah (r=−0.25), sementara fitur lain seperti AnnualIncome, Gender, dan *ProductCategory* hanya memiliki korelasi yang sangat kecil. Analisis ini juga mengonfirmasi tidak adanya multikolinearitas yang kuat antar fitur independen, sebuah kondisi yang baik untuk tahap pemodelan.

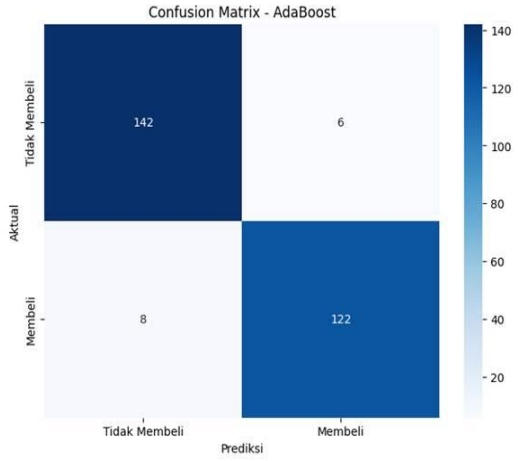
Dari keseluruhan tahap analisis data eksploratif, dapat ditarik kesimpulan bahwa keikutsertaan dalam program loyalitas, pemanfaatan diskon, dan durasi waktu di situs web merupakan fitur-fitur dengan daya prediktif paling signifikan, diikuti oleh usia pelanggan dengan pengaruh negatif. Wawasan ini menjadi fondasi penting yang akan digunakan dalam tahap pemodelan untuk menguji dan membandingkan kinerja algoritma AdaBoost dan CatBoost dalam memprediksi perilaku pembelian pelanggan berdasarkan pola data yang telah teridentifikasi.

e. *Feature Engineering*
   Feature engineering merupakan proses membuat fitur baru atau mentransformasi fitur yang sudah ada guna meningkatkan kemampuan prediktif model [9]. Dalam penelitian ini, dilakukan pembuatan dua fitur tambahan, yaitu:
   1. *Spender Segment*
      Mengelompokkan pelanggan ke dalam kategori *High Spender*, *Medium Spender*, dan *Low Spender* berdasarkan jumlah pembelian yang dilakukan. Segmentasi ini bertujuan untuk memahami perilaku belanja pelanggan dari tingkat pengeluaran yang berbeda.
   2. *Age Group*
      Mengelompokkan pelanggan ke dalam rentang usia seperti 18–30, 31–45, 46–60, dan 61–70. Pengelompokan usia ini memberikan wawasan terkait tren pembelian berdasarkan kelompok umur tertentu.

3.3 Modelling

Model pertama yang dieksekusi adalah *AdaBoost*. Hasil prediksi model terhadap data uji diringkas dalam *confusion matrix* pada Gambar 6. Matriks ini memberikan rincian mengenai jumlah prediksi yang benar dan salah.

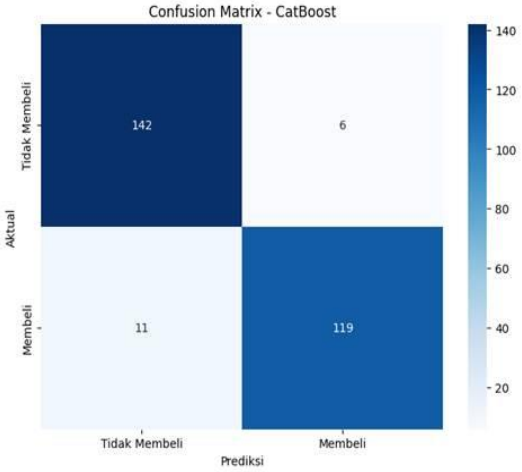
Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947



Gambar 6. *Confusion Matrix* Model AdaBoost

Berdasarkan Gambar 6, terlihat bahwa dari total 278 data uji, model AdaBoost berhasil memprediksi 122 pelanggan yang akan membeli dengan benar dan 142 pelanggan yang tidak akan membeli dengan benar. Namun, model ini membuat kesalahan prediksi sebanyak 6 kasus, di mana pelanggan yang sebenarnya tidak membeli diprediksi akan membeli, dan 8 kasus di mana pelanggan yang seharusnya membeli diprediksi tidak akan membeli.

Model kedua yang dieksekusi adalah *CatBoost*, sebuah algoritma berbasis *gradient boosting* modern. Hasil evaluasinya pada data uji diringkas dalam *confusion matrix* pada Gambar 7.



Gambar 7. *Confusion Matrix* Model CatBoost

Berdasarkan Gambar 7, model *CatBoost* berhasil memprediksi 119 pelanggan yang akan membeli dengan benar (*True Positive*) dan 142 pelanggan yang tidak akan membeli dengan benar (*True Negative*). Kesalahanprediksi yang dibuat model ini terdiri dari 6 kasus *False Positive* dan 11 kasus *False Negative*.

3.4 Evaluasi Hasil

Untuk mengukur performa model dalam melakukan prediksi, digunakan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Hasil evaluasi model AdaBoost terdapat pada Tabel 3.

Tabel 3. Metrik Kuantitatif Model Adaboost

Kelas	*Precision*	*Recall*	*F1-Score*
0 (Tidak Membeli)	0.95	0.96	0.95
1 (Membeli)	0.95	0.95	0.95
Akurasi Keseluruhan	0.95	0.95	0.95
Rata-rata	0.95	0.95	0.95
ROC-AUC			0.9535

Rangkuman metrik kuantitatif dari hasil ini disajikan pada Tabel 3. Model ini mencapai akurasi keseluruhan sebesar 95%. Untuk kelas target 'Membeli', model ini menghasilkan skor presisi 0.95, recall 0.94, dan F1-Score 0.95, yang menunjukkan kinerja yang sangat baik dan seimbang. Adapun untuk rangkuman metrik kuantitatif dari hasil evaluasi model CatBoost ini disajikan pada Tabel 4.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

Tabel 4. Hasil Metrik Kinerja Model CatBoost

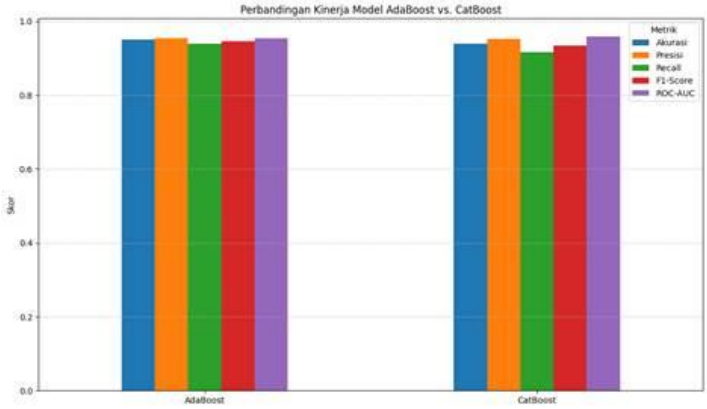
Kelas	*Precision*	*Recall*	*F1-Score*
0 (Tidak Membeli)	0.93	0.96	0.94
1 (Membeli)	0.95	0.92	0.93
Akurasi Keseluruhan	0.94	0.94	0.94
Rata-rata	0.94	0.94	0.94
ROC-AUC			0.9593

Untuk menentukan model terbaik, perbandingan langsung antara *AdaBoost* dan *CatBoost* dilakukan. Tabel 4 menyajikan perbandingan skor untuk setiap metrik, sementara Gambar 6 dan 7 memvisualisasikan perbandingan tersebut.

Tabel 5. Perbandingan Kinerja Model Adaboost dan Catboost

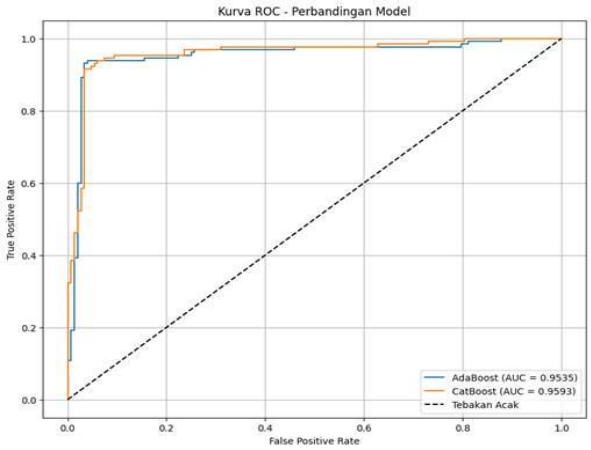
Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*
AdaBoost	0.9496	0.9531	0.9346	0.9457	0.9535
CatBoost	0.9388	0.9520	0.9153	0.9333	0.9592

Hasil perbandingan pada Tabel 5 menunjukkan bahwa kedua model, *AdaBoost* dan *CatBoost*, menunjukkan kinerja yang sangat kompetitif dan hampir setara, dengan masing-masing model memiliki keunggulan pada metrik yang berbeda.



Gambar 8. Grafik Perbandingan Kinerja Metrik Model Adaboost dan CatBoost

Model AdaBoost sedikit lebih unggul dalam metrik F1-Score (0.946 vs 0.933) dan Recall (0.938 vs 0.915). Nilai *recall* yang lebih tinggi mengindikasikan bahwa AdaBoost memiliki kemampuan yang sedikit lebih baik dalam mengidentifikasi semua pelanggan potensial, sehingga meminimalkan risiko kehilangan peluang penjualan (*False Negative*).



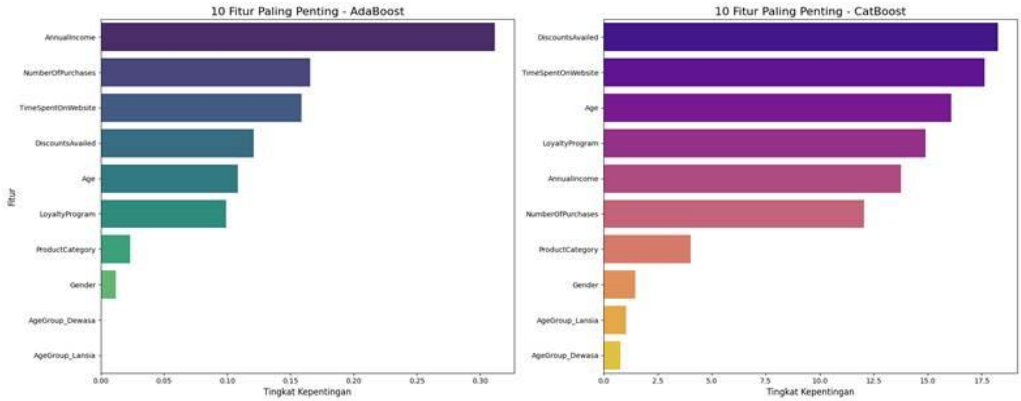
Gambar 9. Kurva ROC-AUC Model Adaboost dan CatBoost

Di sisi lain, model *CatBoost* menunjukkan skor ROC-AUC yang sedikit lebih tinggi 0.959 sementara *AdaBoost* memiliki skor ROC-AUC sebesar 0.953. Hal ini mengindikasikan bahwa secara keseluruhan, *CatBoost* memiliki kemampuan diskriminatif yang sedikit lebih baik dalam memisahkan antara pelanggan yang akan membeli dan yang tidak, pada semua kemungkinan ambang batas klasifikasi. Kedua model menunjukkan tingkat presisi yang hampir

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

identik, artinya kemampuan mereka untuk tidak salah melabeli pelanggan sebagai pembeli (menghindari pemborosan biaya marketing) sangat sebanding.

Untuk memberikan wawasan bisnis, dilakukan analisis fitur penting pada model dengan *F1-Score* tertinggi, yaitu *AdaBoost*. Gambar 10 menyajikan 10 fitur teratas yang paling berpengaruh.



Gambar 10. Perbandingan 10 Fitur Paling Berpengaruh dari Model AdaBoost dan CatBoost

Analisis komparatif pada Gambar 10 menunjukkan beberapa temuan menarik mengenai faktor-faktor kunci yang memengaruhi pembelian. Model *AdaBoost* secara dominan memprioritaskan *AnnualIncome* dan *NumberOfPurchases*, yang mengindikasikan fokus pada kekuatan finansial dan riwayat pelanggan. Sebaliknya, *CatBoost* memberikan perspektif berbeda dengan menempatkan *DiscountsAvailed* dan *TimeSpentOnWebsite* sebagai faktor teratas, yang menekankan pentingnya promosi intensif dan keterlibatan pengguna. Meskipun prioritas utamanya berbeda, kedua model secara konsisten mengakui pentingnya kombinasi antara keterlibatan di situs web, status ekonomi, dan riwayat transaksi sebagai pendorong utama keputusan pembelian, sementara fitur seperti *Gender* dan *ProductCategory* menunjukkan pengaruh yang lebih rendah.

4. KESIMPULAN

Berdasarkan hasil penelitian, algoritma CatBoost dan AdaBoost sangat efektif dalam memprediksi perilaku pembelian pelanggan, di mana CatBoost unggul dengan akurasi keseluruhan sebesar 94%, sementara AdaBoost lebih sensitif dalam mendeteksi pelanggan potensial berkat nilai recall dan F1-score yang lebih tinggi pada kelas positif. Keunggulan CatBoost juga terletak pada kemampuannya menangani fitur kategorikal tanpa pra-pemrosesan tambahan. Wawasan ini sangat penting bagi perusahaan untuk merancang strategi pemasaran yang lebih terarah, seperti personalisasi layanan dan optimalisasi kampanye untuk meningkatkan retensi. Untuk pengembangan selanjutnya, disarankan melakukan optimasi hyperparameter menggunakan metode seperti Optuna atau PSO serta mengeksplorasi algoritma ensemble modern lainnya seperti XGBoost dan LightGBM untuk implementasi model dalam skenario *real-time*.

REFERENCES

[1] M. Kumalasanti, "Pengaruh Konten Pemasaran Tokopedia Terhadap Keputusan Pembelian Pelanggan Di Yogyakarta," *J. Competency Bus.*, vol. 6, no. 01, pp. 77–94, 2022, doi: 10.47200/jcob.v6i01.1306.

[2] D. I. Maharani, "Peluang dan Tantangan Sektor E-Commerce dalam Meningkatkan Perekonomian di Era Transformasi Digital," *J. Simki Econ.*, vol. 7, no. 1, pp. 201–210, 2024, doi: 10.29407/jse.v7i1.493.

[3] A. J. Wahyu *et al.*, "Jejak Inovasi Teknologi E-Commerce Di Indonesia : Perkembangan Dan Tantangan Di Era Digital," *J. Ilmu Sains dan Teknol.*, vol. 1, pp. 12–17, 2025, url: https://artmediapub.id/index.php/JIST/article/view/68

[4] F. Sudirjo, T. Purwati, W. Widyastuti, and Y. U. Budiman, "Analisis Dampak Strategi Pemasaran Digital dalam Meningkatkan Loyalitas Pelanggan: Perspektif Industri E-commerce," *Jurnal Pendidikan Tambusai*, vol. 7, no. 2, pp. 7524–7532, 2023, doi: 10.31004/jptam.v7i2.7422

[5] D. R. Rochmawati, H. Hatimatunnisani, and M. Veranita, "Analisis Strategi Bisnis di Era Transformasi Digital," *MUKASI Jurnal Ilmu Komunikasi.*, vol. 2, no. 3, pp. 223–232, 2023, doi: 10.54259/mukasi.v2i3.2105.

[6] P. A. Wicaksana, I. B. A. Swamardika, and R. S. Hartati, "Literature Review Analisis Perilaku Pelanggan Menggunakan RFM Model," *Maj. Ilm. Teknol. Elektro*, vol. 21, no. 1, p. 21, 2022, doi: 10.24843/mite.2022.v21i01.p04.

[7] M. Sofyan, stiamiacid Ahmad Junaidi, and N. Fitri Rahmawati, "Analisis Kualitas Layanan Dan Persepsi Harga Terhadap Keputusan Pembelian Jasa Ekspedisi Pt. Kadiri Logistik Cargo," *Conf. Ser.*, vol. 1, no. 1, pp. 83–95, 2022, url: https://prosiding.senmabis.nusaputra.ac.id/index.php/prosiding/article/download/31/30

[8] L. Zhang *et al.*, "A Review of Machine Learning in Building Load Prediction," *Applied Energy*, Vol 285, pp. 116452, 2021, doi: 10.1016/j.apenergy.2021.116452

[9] E. Deniz and S. Ç. Bülbül, "Predicting Customer Purchase Behavior Using Machine Learning Models," *Information Technology in Economics and Business*, vol. 1, no. 1, pp. 1–6, 2024, doi: 10.69882/adba.iteb.2024071

[10] S. Iseal and H. Michael, "Customer Behavior Analysis and Purchase Prediction in E-Commerce," no. February, 2025.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2800−2810
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.7947

[11] S. S. Mukrimaa *et al.*, "Machine Learning Teori, Studi Kasus dan Implementasi Menggunakan Python," *J. Penelit. Pendidik. Guru Sekol. Dasar*, vol. 6, no. August, p. 128, 2021, doi: 10.5281/zenodo.5113507.
[12] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer Churn Prediction System: A Machine Learning Approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022, doi: 10.1007/s00607-021-00908-y.
[13] J. Lin, "Application of machine learning in predicting consumer behavior and precision marketing," *PLoS One*, vol. 20, no. 5 May, pp. 1–12, 2025, doi: 10.1371/journal.pone.0321854.
[14] C. Haryanto, N. Rahaningsih, and F. Muhammad Basysyar, "Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 533–539, 2023, doi: 10.36040/jati.v7i1.6343.
[15] E. Hasibuan *et al.*, "Implementasi Machine Learning untuk Prediksi Harga Mobil Bekas dengan Algoritma Regresi Linear berbasis Web," *J. Ilm. Komputasi*, vol. 21, no. 4, pp. 595–602, 2022, doi: 10.32409/jikstik.21.4.3327.
[16] P. Beja-Battais, "Overview of AdaBoost : Reconciling its views to better understand its dynamics," pp. 3–31, 2023, [Online]. Available: http://arxiv.org/abs/2310.18323
[17] A. F. Azmi and A. Voutama, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix," *Komputa J. Ilm. Komput. dan Inform.*, vol. 13, no. 1, pp. 111–119, 2024, doi: 10.34010/komputa.v13i1.12639.
[18] M. F. Alamsyah and A. Wijaya, "Jurnal Informatika : Jurnal pengembangan IT Perbandingan Metode KNN dan Naïve Bayes dalam Deteksi Tingkat Stres Berdasarkan Ekspresi Wajah," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 359–369, 2025, doi: 10.30591/jpit.v10i2.8513.
[19] J. J. A. Limbong, I. Sembiring, and K. D. Hartomo, "Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 2, p. 347, 2022, doi: 10.25126/jtiik.2022924960.
[20] Z. A. Dwiyanti and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," *J. Tekno Insentif*, vol. 17, no. 2, pp. 127–137, 2023, doi: 10.36787/jti.v17i2.1136.