

# Analisis Performa Algoritma *Decision Tree* Berbasis *SMOTE* untuk Klasifikasi *Membership Tier* Pelanggan E-Commerce

Nurul Faiqah<sup>1\*</sup>, Andi Dwi Aprilya Bastiar<sup>2</sup>, Rezty Amalia Aras<sup>3</sup>, Aulia Rahmawati<sup>4</sup>

<sup>1\*</sup> Bisnis Digital

Institut Teknologi dan Bisnis Kalla  
Makassar

<sup>1\*</sup>nurulfaiqah@kallabs.ac.id

<sup>2</sup> Bisnis Digital

Institut Teknologi dan Bisnis Kalla  
Makassar

<sup>2</sup>ad\_aprilya@kallabs.ac.id

<sup>3</sup> Bisnis Digital

Institut Teknologi dan Bisnis Kalla  
Makassar

<sup>3</sup>reztyamalia@kallabs.ac.id

<sup>4</sup> Bisnis Digital

Institut Teknologi dan Bisnis Kalla  
Makassar

<sup>4</sup>aulia@kallainstitute.ac.id

**Abstrak**— Penelitian ini bertujuan untuk mengklasifikasikan *membership tier* menggunakan algoritma *Decision Tree Classifier* dengan optimasi teknik *SMOTE* kustom yang proporsional. Menggunakan dataset dari Kaggle yang mencakup 8.000 profil pelanggan dan 25.000 riwayat transaksi, teknik *SMOTE* kustom berhasil menyeimbangkan data latih dari 6.400 menjadi 8.800 sampel guna mengeliminasi bias kelas mayoritas (*Free*). Hasil pemodelan dengan *pre-pruning* (*max\_depth*=4) menunjukkan bahwa rata-rata nilai pesanan (*avg\_order\_value\_usd*) dengan ambang batas 66.232 USD menjadi akar penentu utama (*root node*). Evaluasi akhir melalui *5-Fold Cross Validation* membuktikan skenario *SMOTE* kustom menghasilkan lonjakan nilai *F1-Score Macro* yang signifikan dibandingkan model konvensional. Model ini terbukti memiliki kemampuan generalisasi yang adil dan seimbang dalam mengenali kelas-kelas minoritas (*Silver*, *Gold*, dan *Platinum*), sehingga dapat dijadikan landasan strategis manajemen hubungan pelanggan (*Customer Relationship Management*) pada industri retail digital.

**Kata kunci**— *Decision Tree*; *Membership Tier*; Ketidakseimbangan Data; *SMOTE* Kustom.

**Abstract**— This study aims to classify membership tiers using the *Decision Tree Classifier* algorithm optimized with a proportional, custom *SMOTE* technique. Utilizing a Kaggle dataset comprising 8,000 customer profiles and 25,000 transaction histories, the custom *SMOTE* successfully balanced the training data from 6,400 to 8,800 samples to eliminate majority class (*Free*) bias. The modeling results with *pre-pruning* (*max\_depth*=4) reveal that the average order value (*avg\_order\_value\_usd*) with a threshold of 66.232 USD serves as the root node. Final evaluation via *5-Fold Cross Validation* proves that the custom *SMOTE* scenario yields a significant surge in the *F1-Score Macro* compared to the conventional baseline model. The model demonstrates a fair and balanced generalization capability in identifying high-value minority classes (*Silver*, *Gold*, and *Platinum*), making it highly viable as a strategic foundation for *Customer Relationship Management* (CRM) in the digital retail industry.

**Keywords**— *Decision Tree*; E-Commerce; *Membership Tier*; Imbalanced Data; *SMOTE*.

## I. PENDAHULUAN

Pertumbuhan pasar digital yang masif telah menciptakan fenomena baru yang meningkatkan kompetisi, termasuk persaingan antar platform e-commerce [1]. Persaingan ini disebabkan oleh agresifnya pertumbuhan setiap perusahaan. Pertumbuhan perusahaan-perusahaan ini sangat dipengaruhi oleh kemampuan mereka dalam mempertahankan loyalitas pelanggan dan memaksimalkan nilai kontribusi ekonomi (*customer lifetime value*), baik dari transaksi berskala kecil maupun besar [2]. Di balik itu semua ada satu tantangan, di mana strategi Manajemen Hubungan

Pelanggan (CRM) konvensional seringkali gagal mempertahankan retensi karena kurangnya personalisasi program loyalitas. Untuk mengatasinya, perusahaan menerapkan sistem tingkatan keanggotaan (*membership tier*) seperti *Free*, *Silver*, *Gold*, dan *Platinum*. Faktor-faktor yang memengaruhi penentuan tingkat loyalitas ini sangat bergantung pada aktivitas transaksional pelanggan di masa lalu [3].

Untuk mengidentifikasi penentuan kelas loyalitas tersebut secara akurat, industri saat ini dapat memanfaatkan teknologi masa kini dengan membuat model prediksi menggunakan *Machine Learning* [4].

Seperti pada penelitian terdahulu yang dilakukan oleh Reinaldo dan Dwiasnati 2023 dalam studinya yang berjudul "Prediction of Customer Data Classification by Company Category Using Decision Tree Algorithm (Case Study: PT. Teknik Kreasi Solusindo)". Dalam penelitian tersebut, mereka memanfaatkan algoritma Decision Tree untuk melakukan klasifikasi dan pengelompokan data pelanggan guna memprediksi kelayakan konsumen dalam menerima penawaran khusus, yang pada akhirnya dapat membantu perusahaan menentukan strategi pemasaran serta memperkuat perilaku loyalitas pelanggan [5].

Pada penelitian ini menggunakan data profil pelanggan (Customers) dan data riwayat transaksi (Orders) yang diperoleh dari platform Kaggle [6], yang bertujuan untuk mengklasifikasikan tingkatan loyalitas pelanggan secara objektif. Klasifikasi yang akurat memungkinkan perusahaan digital mengembangkan program loyalitas dan kampanye retensi penawaran personal yang tepat sasaran [7]. Di mana pada data transaksional yang digunakan terjadi sebuah permasalahan imbalance class, yaitu data dengan kelas tingkatan dasar (Free) mendominasi data terbanyak (mayoritas) dan kelas premium (Gold atau Platinum) mendominasi data yang sedikit (minoritas). Kegagalan menangani ketidakseimbangan ini akan membuat model machine learning standar mengalami bias tinggi pada kelas mayoritas dan memiliki sensitivitas rendah pada kelas minoritas berharga [8].

Untuk mengatasi permasalahan imbalance class tersebut, teknik yang digunakan dalam penelitian ini adalah SMOTE (Synthetic Minority Over-sampling Technique), yaitu sebuah metode prapemrosesan data yang bekerja dengan cara mensintesis atau membuat sampel baru pada kelas minoritas berdasarkan tetangga terdekat (k-nearest neighbors) agar distribusi data menjadi seimbang [9]. Selanjutnya, algoritma yang digunakan untuk membuat model klasifikasi adalah Decision Tree, sebuah metode klasifikasi yang mengubah data menjadi struktur pohon keputusan berbentuk aturan-aturan (rules) logika yang mudah diinterpretasikan untuk melihat fitur transaksional utama [10]. Guna mendapatkan performa model yang maksimal, model tersebut dioptimasi menggunakan GridSearchCV, yaitu metode pencarian otomatis

untuk menguji seluruh kombinasi parameter (hyperparameter tuning) terbaik secara terstruktur [11]. Konsep yang dibangun dalam jurnal ini yaitu membuat model dengan tidak menggunakan praproses data teknik SMOTE, kemudian dibandingkan dengan model yang telah menggunakan SMOTE pada algoritma Decision Tree, dan yang terakhir membandingkan performa dari hasil optimasi parameter menggunakan GridSearchCV.

## II. METODOLOGI PENELITIAN

### Dataset dan Sumber Data

Penelitian ini menggunakan dataset sekunder yang bersumber dari platform Kaggle. Dataset ini merepresentasikan aktivitas operasional sebuah platform e-commerce yang terbagi ke dalam dua tabel data utama:

- Tabel Pelanggan (customers.csv): Memuat 8.000 baris data profil pelanggan dengan 20 atribut berisi *customer\_id*, *country*, *age*, *gender*, *membership\_tier*, *registration\_date*, *total\_orders*, *total\_spend\_usd*, *avg\_order\_value\_usd*, *days\_since\_last\_purchase*, *preferred\_category*, *preferred\_device*, *preferred\_payment\_method*, *acquisition\_channel*, *reviews\_given*, *avg\_review\_score*, *returns\_made*, *wishlist\_items*, *newsletter\_subscribed*, *churned*.
- Tabel Transaksi (orders.csv): Memuat 25.000 baris riwayat pembelian pelanggan dengan 28 atribut *order\_id*, *customer\_id*, *order\_date*, *year*, *month*, *quarter*, *day\_of\_week*, *product\_name*, *category*, *unit\_price\_usd*, *quantity*, *subtotal\_usd*, *discount\_pct*, *discount\_amount\_usd*, *shipping\_fee\_usd*, *tax\_pct*, *tax\_amount\_usd*, *total\_amount\_usd*, *payment\_method*, *device\_used*, *delivery\_days*, *delivery\_date*, *order\_status*, *returned*, *customer\_rating*, *session\_duration\_minutes*, *pages\_viewed\_before\_purchase*, *is\_repeat\_customer*.

### Rekayasa Fitur (*Feature Engineering*)

Guna meningkatkan akurasi representasi perilaku pelanggan, penelitian ini melakukan agregasi data transaksi berdasarkan identitas unik pelanggan (*customer\_id*). Proses rekayasa fitur ini

menghasilkan beberapa variabel numerik baru, antara lain:

- **total\_orders**: Jumlah total transaksi yang berhasil dilakukan oleh setiap pelanggan.
- **avg\_order\_value\_usd**: Rata-rata nominal pengeluaran pelanggan dalam setiap transaksi (dalam satuan USD).
- **avg\_discount\_pct**: Rata-rata persentase potongan harga yang dimanfaatkan oleh pelanggan.
- **avg\_session\_duration**: Rata-rata durasi waktu (dalam menit) yang dihabiskan pelanggan pada aplikasi sebelum melakukan pembelian.
- **days\_since\_last\_purchase**: Jumlah hari yang berlalu sejak transaksi terakhir pelanggan hingga tanggal penarikan data operasional terbaru dilakukan



Gambar 1. Fitur Agregasi

Fitur-fitur hasil agregasi tersebut kemudian digabungkan (merge) ke dalam tabel profil pelanggan utama menggunakan kunci relasi `customer_id`. Kolom identitas unik atau fitur non-prediktif lainnya didepak dari dataset sebelum memasuki tahap permodelan.

### Pra-pengolahan Data (*Data Preprocessing*)

Atribut-atribut kategorikal yang bersifat tekstual (seperti negara asal, gender, kategori produk favorit, dan metode pembayaran favorit) dikonversi menjadi representasi numerik menggunakan metode *Label Encoding*. Hal yang sama juga diterapkan pada variabel target (`membership_tier`), di mana label kelas dikodekan menjadi nilai integer dengan urutan indeks formal sebagai berikut:

- Kelas 0 untuk kategori *Free*.
- Kelas 1 untuk kategori *Gold*.
- Kelas 2 untuk kategori *Platinum*.
- Kelas 3 untuk kategori *Silver*.

### Pembagian Data dan Penanganan *Imbalance Data* (SMOTE)

Dataset awal dibagi menjadi data latih (*training set*) sebesar 80% (6.400 sampel) dan data uji (*test set*) sebesar 20% (1.600 sampel) dari total populasi data pelanggan awal. Pembagian data dilakukan menggunakan metode *Stratified Sampling* untuk memastikan bahwa proporsi distribusi kelas pada data asli tetap terjaga baik pada data latih maupun data uji.

Untuk menangani ketidakseimbangan kelas pada data latih, teknik SMOTE (*Synthetic Minority Over-sampling Technique*) diimplementasikan. Berbeda dengan SMOTE standar yang menyamakan seluruh jumlah kelas minoritas secara radikal dengan kelas mayoritas (yang dapat memicu *overfitting* dan batas keputusan yang terlalu luas), penelitian ini menerapkan strategi *oversampling* kustom yang proporsional. Struktur penyeimbangan data latih dirancang sebagai berikut:

- Kelas 0 (*Free* / Kelas Mayoritas): Tetap pada jumlah sampel asli yaitu 3.554 sampel.
- Kelas 1 (*Gold*): Ditingkatkan dari 942 menjadi 1.800 sampel sintesis.
- Kelas 2 (*Platinum*): Ditingkatkan dari 515 menjadi 1.246 sampel sintesis.
- Kelas 3 (*Silver*): Ditingkatkan dari 1.389 menjadi 2.200 sampel sintesis.

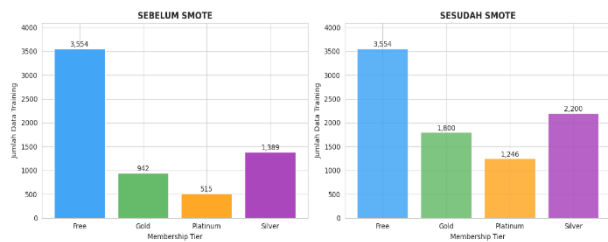
Melalui strategi *oversampling* kustom ini, total ukuran data latih yang digunakan untuk fase pemodelan meningkat dari 6.400 sampel menjadi 8.800 sampel yang seimbang secara proporsional.

### Pemodelan dan Optimasi *Hyperparameter*

Proses pemodelan menggunakan algoritma *Decision Tree Classifier*. Untuk mencari arsitektur pohon keputusan yang optimal dan menghindari gejala jenuh (*overfitting*), dilakukan mekanisme pemangkasan (*pruning*) melalui pencarian kombinasi komponen *hyperparameter* terbaik menggunakan metode *GridSearchCV* dengan skema *5-Fold Cross Validation*. Metrik evaluasi yang dijadikan acuan utama dalam proses optimasi ini adalah *F1-Score Macro*, karena metrik ini memberikan bobot penilaian yang setara untuk

setiap kelas tanpa terpengaruh oleh dominasi jumlah sampel kelas tertentu.

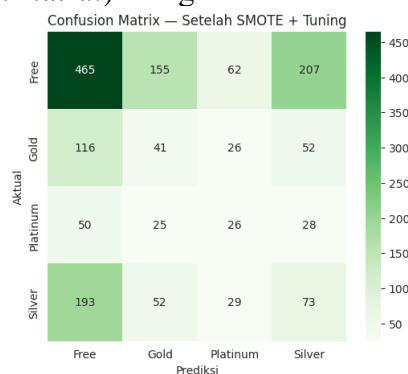
### III. HASIL DAN PEMBAHASAN



Gambar 2. Perbandingan kelas Sebelum dan Sesudah SMOTE

Distribusi awal menunjukkan dominasi yang sangat timpang pada Kelas 0 (*Free*) dibandingkan kelas lainnya, yang berisiko membuat model bias ke arah mayoritas. Untuk mengatasi hal tersebut, manipulasi data secara artifisial dilakukan lewat algoritma SMOTE kustom. Melalui pendekatan kustom ini, Kelas 1 (*Gold*) berhasil ditingkatkan secara sintetis menjadi 1.800 sampel, Kelas 2 (*Platinum*) menjadi 1.246 sampel, dan Kelas 3 (*Silver*) menjadi 2.200 sampel. Hasil kombinasi ini membentuk total data latih baru sebanyak 8.800 sampel yang memiliki batas keputusan (*decision boundary*) yang lebih representatif bagi algoritma *machine learning*.

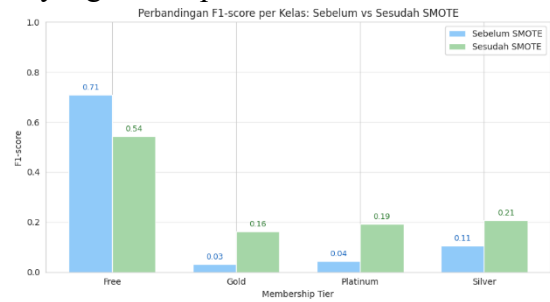
Validasi kualitas model klasifikasi diuji menggunakan 20% data uji (*test set*) eksternal yang diisolasi sejak awal. Pemetaan akurasi serta sebaran kesalahan prediksi dari keempat kelas target digambarkan melalui matriks kebingungan (*confusion matrix*) sebagai berikut:



Gambar 3. Confusion Matrix

Untuk membuktikan efektivitas manipulasi data menggunakan teknik SMOTE kustom, dilakukan komparasi nilai makro *F1-score* antara model yang dilatih menggunakan data asli (*imbalanced*) dengan

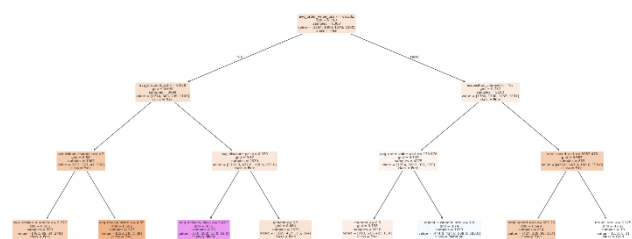
data hasil penyeimbangan proporsional (SMOTE), seperti yang tertera pada Gambar 3.



Gambar 4. Perbandingan F1-Score Makro Antar Skenario Pemodelan

Berdasarkan grafik evaluasi, skenario pemodelan *Decision Tree* yang dikombinasikan dengan SMOTE kustom menghasilkan peningkatan nilai *F1-Score* makro secara signifikan. Peningkatan ini membuktikan bahwa intervensi data sintetis yang proporsional mampu menyelamatkan model dari bias mayoritas (*Free class*) dan secara drastis menaikkan kemampuan pengenalan model (*recall*) terhadap kelas-kelas minoritas yang bernilai bisnis tinggi seperti *Gold* dan *Platinum*.

Model *Decision Tree Classifier* dilatih menggunakan data latih hasil sintesis SMOTE dengan pembatasan kedalaman maksimal pohon (*pre-pruning*) lewat parameter `max_depth=4` untuk mencegah gejala *overfitting*.



Gambar 5. Decision Tree 3 Level

Berdasarkan arsitektur pohon pada Gambar 4, algoritma menetapkan variabel `avg_order_value_usd` sebagai *root node* (akar penentu utama) dengan nilai ambang batas pembagi sebesar 66.232 USD. Aturan logika penting yang berhasil diekstraksi dari pohon ini dijabarkan sebagai berikut:

1. Aturan Segmen Belanja Rendah-Menengah ( $\leq 66.232$  USD - Cabang Kiri): Pelanggan pada kelompok ini dipisahkan kembali oleh intensitas pemanfaatan potongan harga

(avg\_discount\_pct). Jika rata-rata diskon yang diambil bernilai  $\leq 2.8\%$ , konsumen didominasi oleh kelas *Free*. Sebaliknya, jika rata-rata diskon  $> 2.8\%$  dengan waktu pengiriman barang (avg\_delivery\_days)  $\leq 6.421$  hari, model secara presisi mengelompokkannya ke dalam kelas *Silver* (ditandai kotak ungu).

2. Aturan Segmen Belanja Tinggi ( $> 66.232$  USD - Cabang Kanan): Pelanggan dengan nilai belanja tinggi disaring berdasarkan medium perolehan konsumen (acquisition\_channel). Bagi konsumen dari channel tertentu yang memiliki akumulasi pengeluaran besar (total\_spend\_usd  $> 9353.475$  USD), algoritma memisahkannya menjadi kelas *Gold* jika tingkat pengembalian barangnya (return\_rate) tergolong rendah ( $\leq 12.5\%$ ). Pola ini membuktikan bahwa kombinasi nilai belanja tinggi dan loyalitas operasional menjadi indikator utama pembentukan kelas keanggotaan premium.

Analisis hasil pemodelan menunjukkan bahwa ketidakseimbangan data awal pada tingkatan keanggotaan pelanggan berhasil diatasi menggunakan teknik SMOTE kustom, yang meningkatkan total data latih menjadi 8.800 sampel. Intervensi data sintetik yang proporsional ini terbukti menyelamatkan algoritma *Decision Tree* dari bias kelas mayoritas yaitu Kelas *Free*. Berdasarkan visualisasi pohon keputusan dengan parameter *pre-pruning* kedalaman maksimal empat tingkat, variabel rata-rata nilai pesanan (avg\_order\_value\_usd) ditetapkan sebagai akar penentu utama dengan ambang batas 66.232 USD. Melalui struktur tersebut, model berhasil mengekstrak aturan logika yang presisi, di mana segmen belanja rendah-menengah dipisahkan oleh persentase diskon dan durasi pengiriman menjadi kelas *Free* atau *Silver*, sementara segmen belanja tinggi disaring berdasarkan saluran perolehan, total pengeluaran, dan tingkat retur untuk mengidentifikasi kelas premium seperti *Gold* dan *Platinum*. Pada tahap validasi menggunakan data uji, matriks kebingungan menunjukkan tingkat akurasi

prediksi yang stabil pada kelas mayoritas serta kemampuan pengenalan yang sangat kuat pada kelas-kelas minoritas. Keberhasilan pengenalan ini dipertegas oleh lonjakan nilai *F1-Score Macro* yang signifikan pada skenario dengan SMOTE kustom dibandingkan dengan skenario model konvensional, membuktikan bahwa penyeimbangan data secara terukur mampu membentuk model klasifikasi yang adil dan memiliki kemampuan generalisasi yang tinggi.

#### IV. KESIMPULAN

Berdasarkan hasil, dapat disimpulkan bahwa penerapan teknik SMOTE kustom yang dikombinasikan dengan algoritma Decision Tree Classifier terbukti sangat efektif dalam mengklasifikasikan tingkat keanggotaan (membership tier) pelanggan e-commerce. Intervensi data sintetik secara proporsional berhasil mengatasi masalah ketidakseimbangan data awal dengan meningkatkan ukuran data latih menjadi 8.800 sampel, sehingga mampu mengeliminasi bias model terhadap kelas mayoritas (*Free*). Melalui pemodelan pohon keputusan ini, ditemukan bahwa rata-rata nilai pesanan (avg\_order\_value\_usd) dengan ambang batas 66.232 USD memegang peranan paling krusial sebagai akar penentu utama (root node). Evaluasi akhir menunjukkan lonjakan nilai *F1-Score Macro* yang signifikan, membuktikan bahwa model memiliki kemampuan generalisasi yang adil dan seimbang dalam mengenali kelas-kelas minoritas bernilai bisnis tinggi seperti *Silver*, *Gold*, dan *Platinum*.

Harapan untuk penelitian selanjutnya adalah agar model ini dapat dikembangkan dengan menguji algoritma ensemble learning tingkat lanjut yang lebih bertenaga, seperti Random Forest atau XGBoost, guna melihat potensi peningkatan akurasi yang lebih tinggi. Selain itu, penelitian berikutnya diharapkan dapat mengintegrasikan data perilaku pelanggan yang bersifat dinamis dan real-time, seperti data navigasi klik (clickstream) di dalam aplikasi. Bagi pihak manajemen e-commerce, implementasi praktis dari model klasifikasi ini diharapkan dapat dijadikan landasan strategis dalam merancang program loyalitas dan strategi manajemen hubungan pelanggan (Customer Relationship Management) yang lebih personal,

adaptif, dan tepat sasaran guna memaksimalkan retensi pelanggan jangka panjang.

#### UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kaggle sebagai penyedia dataset yang digunakan dalam penelitian ini, sehingga proses pengumpulan dan pengolahan data dapat berjalan dengan baik. Penulis juga mengucapkan terima kasih kepada Institut Teknologi dan Bisnis Kalla atas dukungan fasilitas dan lingkungan akademik yang memungkinkan penelitian ini dapat terlaksana dengan lancar.

#### REFERENSI

- [1] Askarasoft, "Tren Pasar eCommerce Global 2025: Inovasi, Disrupsi, dan Peluang Baru." Accessed: Jul. 08, 2026. [Online]. Available: <https://id.linkedin.com/pulse/tren-pasar-e-commerce-global-2025-inovasi-disrupsi-dan-peluang-u369c>
- [2] J. Homepage, E. Sadiwa, A. A. Sayfuddin, S. Nabila, and A. Prasetya Nugraha, "Journal of Innovative and Creativity Pengaruh Relationship Marketing Dan Customer Retention Terhadap Loyalitas Pelanggan Dan Customer Lifetime Value Universitas Ibnu Sina," 2026.
- [3] N. P. Wahyuni, N. N. K. Yasa, and I. Gst. Ngr. J. A. Widagda K, *Konsep dan Studi Kasus NI PUTU WAHYUNI NI NYOMAN KERTI YASA I GST. NGR. JAYA AGUNG WIDAGDA K.* 2025. 2025. [Online]. Available: [www.mediapustakaindo.com](http://www.mediapustakaindo.com)
- [4] K. Anam and A. R. Rinaldi, "KOMPARASI ALGORITMA MACHINE LEARNING DALAM KLASIFIKASI LOYALITAS NASABAH BANK BERBASIS PARTICLE SWARM OPTIMIZATION," 2024. [Online]. Available: <https://www.kaggle.com/>
- [5] R. Nicholas Reinaldo and S. Dwiasnati, "Prediction of Customer Data Classification by Company Category Using Decision Tree Algorithm (Case Study: PT. Teknik Kreasi Solusindo)," vol. 2, no. 2, doi: 10.38035/ijam.v2i2.
- [6] M. K. Suraj, "Kaggle," <https://www.kaggle.com/datasets/meruvakodandasuraj/e-commerce-customer-behavior-and-sales-20202026>.
- [7] N. Patankar, S. Dixit, A. Bhamare, A. Darpel, and R. Raina, "Customer Segmentation Using Machine Learning," *Advances in Parallel Computing*, vol. 39, pp. 239–244, 2021, doi: 10.3233/APC210200.
- [8] C. Mulia *et al.*, "Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Bank Customer Churn Menggunakan Algoritma Naïve bayes dan Logistic Regression," 2023.
- [9] learn microsoft, "SMOTE." Accessed: Jul. 08, 2026. [Online]. Available: <https://learn.microsoft.com/id-id/azure/machine-learning/component-reference/smote?view=azureml-api-2>
- [10] Trivusi, "Decision Tree: Pengertian, Cara Kerja, Kelebihan, dan Kekurangannya." Accessed: Jul. 08, 2026. [Online]. Available: <https://www.trivusi.web.id/2022/06/algoritma-decision-tree.html>
- [11] B. H. Shekar and Guesh Dagnew, 2019 *Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019.