Jurnal Ilmu Komputer, Sistem Informasi, Teknik Informatika
ISSN: 2964-4763
Vol. 5, No. 1, Maret 2026

Komparasi Kinerja Model *Machine learning* Berbasis Metadata Produk untuk Prediksi Popularitas Produk Elektronik pada Marketplace Lazada Indonesia

Eko Aziz Apriadi[1], Agus Komarudin[2], Ribut Julianto[3]
[1,2,3]Universitas Indonesia Mandiri
Email: [1]ekoazizapriadi72@gmail.com, [2]aguskomarudin689@gmail.com, [3]ributjulianto@uimandiri.ac.id

Abstrack

This study explores the use of product metadata to predict the popularity of electronic products in the Lazada Indonesia marketplace. By using machine learning models, including Logistic Regression, Random Forest, and XGBoost, this study shows that simple metadata such as product category, brand, price, and rating are sufficiently informative to build predictive models. The results indicate that although Logistic Regression delivers the lowest performance due to its linear nature, both Random Forest and XGBoost provide significant improvement. XGBoost achieves the best results with the highest accuracy and F1-score, making it the most effective model for predicting product popularity. These findings highlight the complexity of e-commerce data, which requires more flexible models to capture non-linear patterns and interactions among product features. This study contributes to e-commerce management by providing insights into the use of machine learning for inventory management, promotional strategies, and product placement in digital marketplaces.

Keywords: *E-commerce, Product popularity prediction, Metadata analysis, Machine learning, XGBoost*

1. PENDAHULUAN

Perkembangan ekonomi digital di Asia Tenggara menunjukkan bahwa *e-commerce* telah menjadi salah satu penggerak utama transaksi digital. Laporan *e-Conomy SEA 2025* menunjukkan bahwa ekonomi digital Asia Tenggara telah melampaui US$300 miliar GMV, dengan pertumbuhan yang tetap kuat dari tahun ke tahun. Laporan yang sama juga menegaskan bahwa aktivitas belanja daring telah menjadi bagian dari perilaku digital sehari-hari masyarakat kawasan, sehingga persaingan antarseller dan antarplatform menjadi semakin ketat. Dalam konteks ini, kemampuan untuk membaca pola pasar dari data produk menjadi kebutuhan strategis, bukan lagi sekadar pilihan teknis[1].

Di tengah pertumbuhan tersebut, Lazada menempati posisi penting sebagai salah satu platform *e-commerce* utama di Asia Tenggara dengan cakupan operasional di enam negara, termasuk Indonesia. Bagi penjual dan pengelola platform, jumlah produk yang sangat besar menciptakan tantangan baru, yaitu bagaimana mengidentifikasi produk mana yang berpotensi lebih populer dibandingkan produk lain dalam kategori yang sama. Tantangan ini semakin relevan karena pertumbuhan *e-commerce* regional juga didorong oleh perubahan pola penemuan produk, termasuk penguatan video commerce dan format belanja digital yang makin kompetitif. Situasi ini menuntut pendekatan analitis yang mampu mengolah data produk secara cepat, konsisten, dan terukur[2].

Popularitas produk pada marketplace dapat dipahami sebagai sinyal penting yang merepresentasikan tingkat perhatian, keterlibatan, dan respons pasar terhadap suatu produk. Dalam praktik *e-commerce*, indikator seperti jumlah ulasan dan rating sering kali berperan sebagai bentuk *social proof* yang memengaruhi persepsi konsumen sebelum membeli. Temuan meta-analisis terbaru menunjukkan bahwa ulasan daring memiliki pengaruh signifikan terhadap niat beli, sementara studi meta-analitik lain

menegaskan bahwa kepercayaan, risiko yang dirasakan, keamanan, dan *electronic word-of-mouth* berperan nyata dalam keputusan pembelian digital. Karena itu, prediksi popularitas produk tidak hanya bermanfaat bagi penjual untuk strategi penawaran, tetapi juga relevan bagi platform untuk pengelolaan katalog dan optimasi rekomendasi[3].

Salah satu pendekatan yang menjanjikan untuk menjawab kebutuhan tersebut adalah pemanfaatan metadata produk. Metadata seperti kategori, merek, harga, dan rating memiliki keunggulan karena terstruktur, mudah diakses, dan relatif stabil dibandingkan data teks bebas yang memerlukan pemrosesan bahasa alami yang lebih kompleks. Kajian sistematis tentang AI dalam *e-commerce* menegaskan bahwa data mining dan *machine learning* berperan penting dalam mengekstraksi pola perilaku, tren, dan hubungan tersembunyi pada data platform digital. Dengan demikian, metadata produk dapat diposisikan sebagai sumber informasi yang efisien untuk membangun model prediktif yang lebih praktis dan mudah direplikasi.

Dalam ranah akademik, penerapan *machine learning* di *e-commerce* telah berkembang pada berbagai masalah, seperti prediksi kepuasan pelanggan, prediksi harga, prediksi permintaan, dan identifikasi tren produk. Studi terkini menunjukkan bahwa model *machine learning* mampu memberikan performa yang kuat dalam klasifikasi kepuasan pelanggan e-commerce, sementara penelitian lain membandingkan banyak algoritma untuk prediksi harga komoditas dan pengembangan sistem prediksi tren produk. Selain itu, riset terdahulu juga menunjukkan bahwa prediksi *best-selling products* dapat digunakan untuk meningkatkan pencarian produk pada platform e-commerce. Temuan-temuan ini memperlihatkan bahwa *machine learning* memiliki landasan yang kuat untuk diterapkan pada persoalan popularitas produk[4].

Sebagian besar penelitian *e-commerce* masih banyak berfokus pada analisis sentimen, rekomendasi berbasis perilaku pengguna, atau prediksi variabel tunggal seperti harga dan kepuasan. Kajian yang secara spesifik membandingkan kinerja beberapa model *machine learning* untuk memprediksi popularitas produk berbasis metadata terstruktur pada marketplace Indonesia masih relatif terbatas. Keterbatasan ini terlihat semakin jelas ketika objek penelitian dipersempit pada kategori elektronik, padahal kategori ini dikenal memiliki variasi harga tinggi, banyak merek, serta tingkat kompetisi yang sangat padat. Celah inilah yang membuat penelitian tentang komparasi model *machine learning* berbasis metadata produk menjadi relevan dan memiliki nilai kebaruan[5].

Kebutuhan untuk melakukan komparasi model juga didasarkan pada fakta bahwa tidak ada satu algoritma yang selalu unggul pada semua karakteristik data. Penelitian-penelitian terbaru pada domain *e-commerce* menunjukkan bahwa performa model sangat dipengaruhi oleh jenis fitur, bentuk distribusi data, tujuan prediksi, dan keseimbangan antara akurasi serta kompleksitas model. Dengan kata lain, pemilihan algoritma tidak dapat hanya didasarkan pada popularitas metode, tetapi harus diuji langsung pada data dan konteks masalah yang spesifik. Oleh sebab itu, studi komparatif menjadi langkah metodologis yang penting agar model yang dipilih benar-benar sesuai dengan karakter data marketplace.

Berdasarkan eksplorasi awal terhadap dataset penelitian, data yang digunakan memuat 10.942 produk elektronik dari 5 kategori dan 234 merek pada Lazada Indonesia. Dataset ini juga memperlihatkan rentang harga yang sangat lebar, dari Rp1.000 hingga Rp275.000.000, serta distribusi jumlah ulasan yang sangat timpang, dengan median hanya 2 ulasan tetapi nilai maksimum mencapai 9.631 ulasan. Seluruh data juga memiliki nilai rating, sehingga variabel seperti kategori, merek, harga, dan rating dapat langsung dimanfaatkan sebagai fitur prediktif. Karakter ini menunjukkan bahwa dataset memiliki nilai empiris yang kuat, sekaligus menghadirkan tantangan nyata bagi model *machine learning* dalam menangkap pola popularitas produk[6][7].

Karakter data yang timpang dan heterogen tersebut membuat penelitian ini tidak cukup bila hanya menggunakan satu model. Model linear mungkin lebih sederhana dan mudah dijelaskan, tetapi belum tentu mampu menangkap relasi nonlinier antara harga, rating, merek, dan kategori produk. Sebaliknya, model berbasis ensemble sering kali lebih adaptif terhadap pola kompleks, tetapi perlu diuji secara objektif agar tidak hanya unggul pada akurasi, melainkan juga stabil pada kelas data yang berbeda. Karena itu, komparasi kinerja model menjadi inti penting untuk menentukan pendekatan prediksi yang paling sesuai bagi data metadata produk elektronik di marketplace.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk membandingkan kinerja beberapa model *machine learning* dalam memprediksi popularitas produk elektronik pada marketplace Lazada Indonesia dengan memanfaatkan metadata produk. Penelitian ini diharapkan memberikan kontribusi teoretis berupa penguatan kajian prediksi popularitas produk berbasis data terstruktur, sekaligus kontribusi praktis bagi penjual dan pengelola marketplace dalam menyusun strategi katalog, promosi, dan pengambilan keputusan berbasis data. Dengan fokus pada konteks Indonesia dan kategori elektronik, studi ini juga diharapkan dapat memperkaya literatur informatika terapan yang relevan dengan kebutuhan industri digital saat ini.

2. TINJAUAN PUSTAKA

2.1 *E-commerce* dan Marketplace di Indonesia

Perkembangan *e-commerce* di Indonesia, terutama melalui platform seperti Lazada, telah mengubah cara konsumen berbelanja, mendorong pentingnya analisis data produk untuk memahami pola konsumsi dan tren pasar. Dalam konteks ini, prediksi popularitas produk menjadi krusial bagi penjual dan pengelola marketplace untuk merumuskan strategi pemasaran yang efektif. Beberapa indikator yang umum digunakan untuk mengukur popularitas produk termasuk jumlah ulasan, rating produk, serta fitur-fitur metadata seperti kategori, merek, dan harga. *Machine learning*, khususnya model seperti *Random Forest*, XGBoost, dan Support Vector Machine (SVM), telah terbukti efektif dalam menganalisis data besar dan memprediksi popularitas produk berbasis metadata terstruktur. Penelitian menunjukkan bahwa model-model ini mampu menghasilkan prediksi yang akurat dalam konteks *e-commerce* dengan memanfaatkan data yang relatif mudah diakses dan lebih stabil, seperti yang ditemukan dalam metadata produk[8].

Meskipun sejumlah penelitian telah dilakukan untuk memprediksi popularitas produk e-commerce, sebagian besar berfokus pada penggunaan data ulasan atau teks bebas. Sedangkan, penggunaan metadata produk seperti kategori, harga, dan rating produk untuk prediksi popularitas belum banyak dieksplorasi di Indonesia, khususnya pada marketplace seperti Lazada. Penelitian oleh *Li et al. (2021)* menunjukkan bahwa model berbasis metadata dapat memberikan hasil yang lebih efisien dan terukur dibandingkan dengan model berbasis teks ulasan yang memerlukan pemrosesan bahasa alami yang kompleks. Oleh karena itu, penelitian ini bertujuan untuk membandingkan kinerja beberapa model *machine learning* dalam memprediksi popularitas produk elektronik di Lazada Indonesia dengan memanfaatkan metadata produk, yang diharapkan dapat memperkaya literatur *e-commerce* dan menyediakan solusi praktis untuk pengelolaan katalog produk.

2.2 Popularitas Produk dalam E-Commerce

Popularitas produk dalam *e-commerce* sering kali diukur melalui berbagai indikator, seperti jumlah ulasan, rating produk, dan tingkat penjualan. Produk dengan banyak ulasan dan rating tinggi

cenderung dianggap lebih populer oleh konsumen, karena dinilai lebih terpercaya dan memiliki kualitas yang lebih baik. Ulasan produk yang positif memberikan *social proof* yang mendorong pembelian, sedangkan rating yang tinggi menunjukkan kepuasan pelanggan yang dapat mempengaruhi keputusan calon pembeli. Ulasan pelanggan berperan penting dalam meningkatkan konversi pembelian, di mana produk dengan ulasan banyak dan positif lebih sering dipilih dibandingkan produk dengan ulasan yang sedikit atau buruk[9].

Selain ulasan dan rating, faktor lain seperti harga, merek, dan kategori produk juga berperan besar dalam menentukan popularitas produk. Produk dengan harga yang kompetitif dan merek yang dikenal cenderung lebih menarik bagi konsumen, terutama di pasar yang sangat kompetitif seperti Indonesia. Penelitian menunjukkan bahwa, selain faktor sosial seperti ulasan, harga yang lebih terjangkau dan merek yang kuat juga meningkatkan kemungkinan suatu produk menjadi lebih populer. Oleh karena itu, prediksi popularitas produk dalam *e-commerce* membutuhkan pendekatan yang komprehensif, yang mempertimbangkan berbagai faktor tersebut, untuk memberikan gambaran yang lebih akurat mengenai bagaimana produk akan diterima pasar[10][11].

2.3 Metadata Produk dalam Marketplace

Metadata produk merujuk pada data terstruktur yang mendeskripsikan produk dalam marketplace, seperti kategori, merek, harga, rating, dan jumlah ulasan. Metadata ini memainkan peran penting dalam pengelolaan katalog produk, pencarian produk, serta analisis pasar. Dalam platform e-commerce, metadata memungkinkan sistem untuk menyaring, mengelompokkan, dan menyarankan produk kepada konsumen berdasarkan preferensi mereka. Sebagai contoh, informasi kategori dan merek dapat membantu konsumen menemukan produk yang relevan dengan lebih cepat, sementara harga dan rating memberikan gambaran kualitas dan nilai suatu produk[12].

Penggunaan metadata produk dalam marketplace juga mempermudah penerapan teknik analitik dan *machine learning* untuk memprediksi tren dan popularitas produk. Dengan memanfaatkan metadata, platform *e-commerce* dapat mengembangkan model prediktif yang lebih akurat dan efisien untuk memahami permintaan pasar dan perilaku konsumen. Penelitian [13] menunjukkan bahwa model berbasis metadata dapat mengurangi kompleksitas dalam pemrosesan data, karena metadata sudah terstruktur dan tidak memerlukan pemrosesan tambahan seperti yang diperlukan pada data teks bebas. Metadata produk sangat penting dalam merancang strategi pemasaran yang lebih terarah dan meningkatkan pengalaman berbelanja bagi konsumen di marketplace.

2.4 *Machine learning* dalam *E-Commerce*

Machine learning (ML) telah menjadi alat yang sangat penting dalam *e-commerce*, terutama untuk memproses data besar dan kompleks yang dihasilkan oleh platform digital. Teknologi ini memungkinkan perusahaan untuk membuat prediksi dan mengambil keputusan berbasis data yang lebih akurat, seperti prediksi permintaan, rekomendasi produk, dan analisis perilaku pelanggan. Dalam konteks *e-commerce*, *machine learning* digunakan untuk berbagai aplikasi, mulai dari personalisasi pengalaman berbelanja hingga deteksi penipuan. Menurut *Günther et al. (2020)*, penggunaan ML memungkinkan platform *e-commerce* untuk menyajikan produk yang lebih relevan bagi pengguna dengan menganalisis pola belanja sebelumnya, sehingga meningkatkan konversi dan kepuasan pelanggan[14].

Machine learning juga digunakan untuk memprediksi popularitas produk, yang sangat penting untuk

pengelolaan inventaris dan strategi pemasaran. Model ML, seperti *Random Forest*, XGBoost, dan Support Vector Machine (SVM), dapat menganalisis data terstruktur seperti metadata produk (kategori, harga, rating) untuk memprediksi seberapa populer suatu produk di pasar. Penelitian oleh *Chai et al. (2019)* menunjukkan bahwa model berbasis *machine learning* dapat membantu *e-commerce* dalam memahami tren pasar dan preferensi konsumen secara lebih efisien, sehingga dapat memberikan rekomendasi produk yang lebih tepat dan mempercepat pengambilan keputusan bisnis. Dengan kemampuan untuk menganalisis data dalam jumlah besar, ML membuka peluang untuk optimasi berkelanjutan dalam berbagai aspek operasional *e-commerce* [15].

2.5 *Random Forest* dan XGBoost dalam Prediksi Popularitas Produk

Random Forest adalah algoritma *machine learning* berbasis ensemble yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil. Model ini bekerja dengan cara membangun banyak pohon keputusan yang dilatih dengan subset acak dari data, kemudian menggabungkan hasilnya untuk meningkatkan akurasi prediksi. *Random Forest* sangat efektif dalam menangani data besar dan variabel yang tidak terstruktur, serta mampu mengatasi masalah overfitting yang sering ditemui pada model pohon keputusan tunggal. Penelitian oleh *Breiman (2001)* menunjukkan bahwa *Random Forest* memiliki kemampuan untuk menangani data dengan jumlah fitur yang banyak dan dapat mengidentifikasi fitur yang paling penting dalam prediksi, sehingga menjadikannya pilihan populer untuk berbagai aplikasi prediktif, termasuk dalam prediksi popularitas produk di marketplace[16].

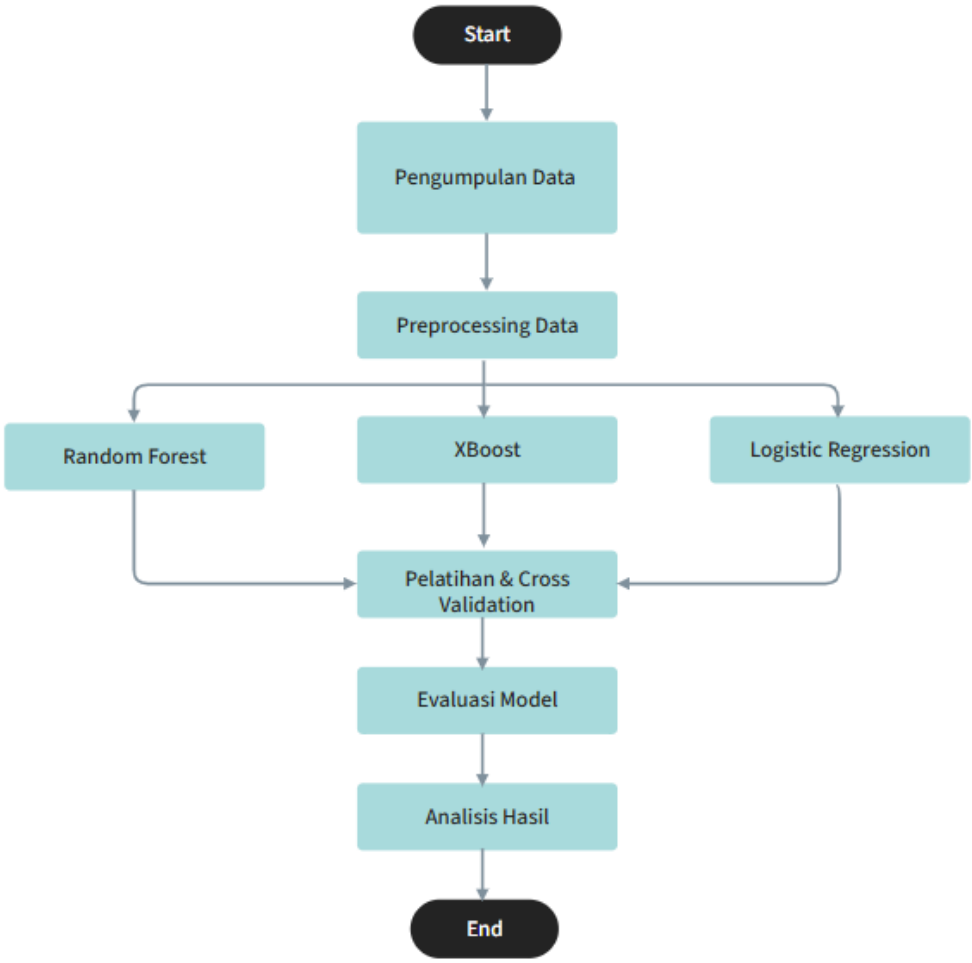
XGBoost, di sisi lain, adalah algoritma boosting yang mengoptimalkan model dengan cara meningkatkan bobot pada data yang sulit diprediksi pada iterasi sebelumnya. Model ini bekerja dengan cara menggabungkan pohon keputusan secara bertahap, dengan tujuan mengurangi kesalahan prediksi secara signifikan. XGBoost terkenal dengan kecepatan pelatihan yang tinggi dan kinerja yang sangat baik dalam kompetisi *Kaggle*, serta mampu mengatasi masalah data yang tidak seimbang dan fitur yang kompleks. Studi oleh [17] menunjukkan bahwa XGBoost memiliki keunggulan dalam menangani masalah klasifikasi dan regresi dengan data yang banyak dan beragam, membuatnya sangat cocok untuk memprediksi popularitas produk berdasarkan metadata, seperti kategori, harga, dan rating. Kedua model ini telah terbukti efektif dalam berbagai studi terkait prediksi permintaan dan tren produk di *e-commerce* [18].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode komparatif untuk membandingkan kinerja beberapa model *machine learning* dalam memprediksi popularitas produk elektronik pada marketplace Lazada Indonesia. Data yang digunakan dalam penelitian ini berasal dari dataset produk yang mencakup informasi metadata seperti kategori, merek, harga, rating, dan jumlah ulasan untuk 10.942 produk elektronik yang terdaftar di Lazada. Data ini kemudian diproses menggunakan teknik preprocessing untuk memastikan bahwa data siap digunakan dalam model *machine learning*. Proses preprocessing mencakup pembersihan data, pengisian nilai yang hilang, serta transformasi variabel kategorikal menggunakan one-hot encoding dan variabel numerikal dengan standardisasi agar dapat digunakan oleh algoritma *machine learning* [19][20].

Model yang akan dibandingkan dalam penelitian ini meliputi *Random Forest*, XGBoost, dan Logistic Regression. Setiap model akan dilatih menggunakan data metadata produk untuk memprediksi

popularitas produk yang diukur berdasarkan jumlah ulasan dan rating yang diterima. Proses pelatihan dan evaluasi dilakukan dengan menggunakan teknik cross-validation untuk menghindari overfitting dan untuk memperoleh evaluasi yang lebih objektif tentang kinerja setiap model. Hasil dari setiap model kemudian dianalisis menggunakan beberapa metrik evaluasi, seperti akurasi, precision, recall, dan F1-score. Analisis komparatif ini bertujuan untuk menentukan model yang paling efektif dalam memprediksi popularitas produk, yang dapat diterapkan dalam strategi pemasaran dan pengelolaan katalog produk di marketplace



Gambar 1. Diagram Alir Penelitian

4. HASIL DAN PEMBAHASAN

4.1 Persiapan Data

Data yang digunakan dalam penelitian ini diperoleh dari kaggle Marketplace Lazada Indonesia, yang mencakup produk elektronik dalam rentang waktu yang tidak terbatas, namun mencerminkan kondisi saat data diambil pada tahun 2023. Jumlah data yang digunakan dalam penelitian ini sebanyak 10.942 *record* dan mencakup 5 kategori produk serta 234 merek. Atribut yang tersedia dalam dataset ini terdiri dari nama produk**,** kategori**,** merek**,** harga**,** rating rata-rata, dan jumlah ulasan. Setiap atribut ini mewakili metadata produk yang relevan untuk analisis prediksi popularitas produk. Data kemudian diproses dan dibersihkan sebelum digunakan dalam tahap modeling untuk memprediksi popularitas produk berdasarkan jumlah ulasan dan rating yang diterima.

Tabel 1. Dataset Produk Elektronik pada Marketplace Lazada Indonesia

	A	B	C	D	E	F	G	H	I
1	itemId	category	name	brandName	url	price	averageRating	totalReviews	retrievedDate
2	1E+08	beli-harddisk-eksternal	TOSHIBA Smart HD LED TV 32" - 32L5650VJ Free Bracket TV - Hitam - Khusus Jabodetabek	Toshiba	https://www.lazada.co.id/products/toshiba-smart-hd-led-tv-32-32l5650vj-free-bracket-tv-hitam-khusus-jabodetabek-i100002528-s100003593.html?search=1	2499000	4	8	02/10/2019
3	1E+08	beli-harddisk-eksternal	TOSHIBA Full HD Smart LED TV 40" - 40L5650VJ - Hitam - Khusus Jabodetabek	Toshiba	https://www.lazada.co.id/products/toshiba-full-hd-smart-led-tv-40-40l5650vj-hitam-khusus-jabodetabek-i100003785-s100006061.html?search=1	3788000	3	3	02/10/2019
4	1E+08	beli-harddisk-eksternal	Samsung 40 Inch Full HD Flat LED Digital TV 40J5000	LG	https://www.lazada.co.id/products/samsung-40-inch-full-hd-flat-led-digital-tv-40j5000-i100004132-s100006644.html?search=1	3850000	3	2	02/10/2019
5	1E+08	beli-harddisk-eksternal	Sharp HD LED TV 24" - LC-24LE175I - Hitam	Sharp	https://www.lazada.co.id/products/sharp-hd-led-tv-24-lc-24le175i-hitam-i100004505-s100007387.html?search=1	1275000	3	11	02/10/2019
6	1E+08	beli-harddisk-eksternal	Lenovo Ideapad 130-15AST LAPTOP MULTIMEDIA I AMD A4-9125 I 8GB DDR4 I 1TB HDD I AMD Radeon R3 I 15,6 HD LED I DVDRW I DOS	Lenovo	https://www.lazada.co.id/products/lenovo-ideapad-130-15ast-laptop-multimedia-i-amd-a4-9125-i-8gb-ddr4-i-1tb-hdd-i-amd-radeon-r3-i-156-hd-led-i-dvdrw-i-dos-i100005037-s100008405.html?search=1	3984100	5	1	02/10/2019
7	1E+08	beli-harddisk-eksternal	Sandisk	Flashdisk	Cruzer	Glide	CZ60	16GB	+
8	1E+08	beli-harddisk-eksternal	Asus	X407UB-BV187T	Laptop	Multimedi	Murah	I	Coreâ„¢
9	1E+08	beli-harddisk-eksternal	Philips 32PHA3052S/70 32" Televisi LED (FREE Bracket LED) - Khusus JABODETABEK	Philips	https://www.lazada.co.id/products/philips-32pha3052s70-32-televisi-led-free-bracket-led-khusus-jabodetabek-i100010722-s100016473.html?search=1	1990000	4	10	02/10/2019
10	1E+08	beli-harddisk-eksternal	Philips	39PHA4251S/70	39'	Televisi	LED	(FREE	Bracket
11	1E+08	beli-harddisk-eksternal	Acer	E5	575G-74E2	-	Core	i7-6500U	-
12	1E+08	beli-harddisk-eksternal	Sharp	LED	32	LE	180	-	32
13	1E+08	beli-harddisk-eksternal	SanDisk	Flashdisk	Cruzer	Glide	CZ60	-	32GB
			Maxtor Hardisk Eksternal M3		https://www.lazada.co.id/products/maxtor-hardisk-eksternal-m3-1tb-25-usb-30-hdd-ext-				

4.2 Statistik Deskriptif Data

Tahap berikutnya adalah melihat gambaran umum data. Statistik deskriptif membantu memahami karakteristik dataset sebelum masuk ke pemodelan. Dari hasil perhitungan, terlihat bahwa data memiliki keragaman yang cukup besar, terutama pada variabel harga dan jumlah ulasan.

```
print("Jumlah data:", len(df)) Untitled-1 9+
1  print("Jumlah data:", len(df))
2  print("Jumlah kategori:", df["category"].nunique())
3  print("Jumlah merek:", df["brandName"].nunique())
4  print("Harga minimum:", df["price"].min())
5  print("Harga maksimum:", df["price"].max())
6  print("Harga rata-rata:", df["price"].mean())
7  print("Rating minimum:", df["averageRating"].min())
8  print("Rating maksimum:", df["averageRating"].max())
9  print("Rating rata-rata:", df["averageRating"].mean())
10 print("Ulasan minimum:", df["totalReviews"].min())
11 print("Ulasan maksimum:", df["totalReviews"].max())
12 print("Ulasan rata-rata:", df["totalReviews"].mean())
13 print("Ulasan median:", df["totalReviews"].median())
```

Gambar 2. Statistik Deskriptif Data

Hasil statistik deskriptif menunjukkan bahwa dataset terdiri dari 10.942 produk dengan 5 kategori dan 235 merek. Harga produk bervariasi antara Rp1.000 hingga Rp275.000.000, dengan harga rata-rata sebesar Rp3.020.218,62. Rating produk berkisar antara 1,0 hingga 5,0, dengan rating rata-rata 4,17. Jumlah ulasan berkisar antara 1 hingga 9.631, dengan rata-rata jumlah ulasan 27,37 dan nilai median 2. Data ini menunjukkan distribusi jumlah ulasan yang sangat timpang, di mana sebagian besar produk hanya memiliki sedikit ulasan, sementara beberapa produk memiliki jumlah ulasan yang sangat tinggi. Kondisi ini menjadikan tugas prediksi popularitas produk lebih menantang dan relevan untuk diuji dengan berbagai model *machine learning*.

4.3 Preprocessing Data

Sebelum pelatihan model, data perlu diproses agar bisa dibaca oleh algoritma *machine learning*. Variabel kategorikal seperti category dan brandName diubah menjadi bentuk numerik menggunakan One-Hot Encoding. Variabel numerik seperti price dan averageRating dinormalisasi dengan

StandardScaler. Setelah itu, data dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%.

C: > Users > Asus > Downloads > stitch_layanan_pengguna_keanggotaan > stitch_layanan_pengguna_keanggotaan > struktur_organisasi_detail > from sklearn.py > ...

```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer

X = df[["category", "brandName", "price", "averageRating"]]
y = df["popularity_class"]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

preprocess = ColumnTransformer([
    ("cat", OneHotEncoder(handle_unknown="ignore"), ["category", "brandName"]),
    ("num", Pipeline([
        ("imputer", SimpleImputer(strategy="median")),
        ("scaler", StandardScaler())
    ]), ["price", "averageRating"])
])

X_train_p = preprocess.fit_transform(X_train)
X_test_p = preprocess.transform(X_test)

print("Bentuk data latih:", X_train_p.shape)
print("Bentuk data uji:", X_test_p.shape)
```

Gambar 3. Preprocessing Data

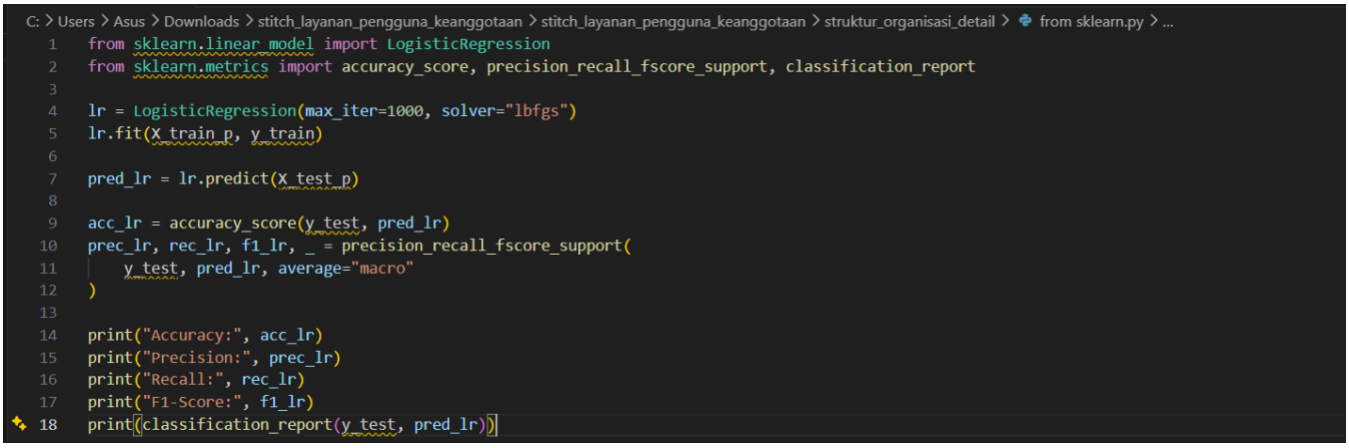
Hasil preprocessing menunjukkan bahwa setelah data diproses dan fitur kategorikal diubah menggunakan *one hot encoding*, bentuk data latih menjadi (8753, 236) dan bentuk data uji menjadi (2189, 236). Proses *one hot encoding* pada variabel kategori dan merek menghasilkan 236 fitur, yang mencerminkan hasil transformasi variabel kategorikal menjadi format numerik yang siap digunakan dalam model klasifikasi. Hal ini menunjukkan bahwa metadata produk telah berhasil diubah menjadi format yang sesuai untuk analisis *machine learning*, memungkinkan model untuk memprediksi popularitas produk berdasarkan data yang telah diproses.

Tabel 2. Hasil Preprocessing Data

Data	Bentuk
Data latih	(8753, 236)
Data uji	(2189, 236)
Jumlah fitur	236

4.4 Pemodelan Menggunakan Logistic Regression

Model pertama yang digunakan adalah Logistic Regression sebagai model dasar. Model ini penting karena sederhana, cepat, dan sering digunakan sebagai pembanding awal. Logistic Regression cenderung bekerja baik pada hubungan yang linier, tetapi kurang kuat dalam menangkap pola yang lebih kompleks.



```
C: > Users > Asus > Downloads > stitch_layanan_pengguna_keanggotaan > stitch_layanan_pengguna_keanggotaan > struktur_organisasi_detail > from sklearn.py > ...
1  from sklearn.linear_model import LogisticRegression
2  from sklearn.metrics import accuracy_score, precision_recall_fscore_support, classification_report
3
4  lr = LogisticRegression(max_iter=1000, solver="lbfgs")
5  lr.fit(X_train_p, y_train)
6
7  pred_lr = lr.predict(X_test_p)
8
9  acc_lr = accuracy_score(y_test, pred_lr)
10 prec_lr, rec_lr, f1_lr, _ = precision_recall_fscore_support(
11     y_test, pred_lr, average="macro"
12 )
13
14 print("Accuracy:", acc_lr)
15 print("Precision:", prec_lr)
16 print("Recall:", rec_lr)
17 print("F1-Score:", f1_lr)
18 print(classification_report(y_test, pred_lr))
```

Gambar 4. Pemodelan Menggunakan Logistic Regression

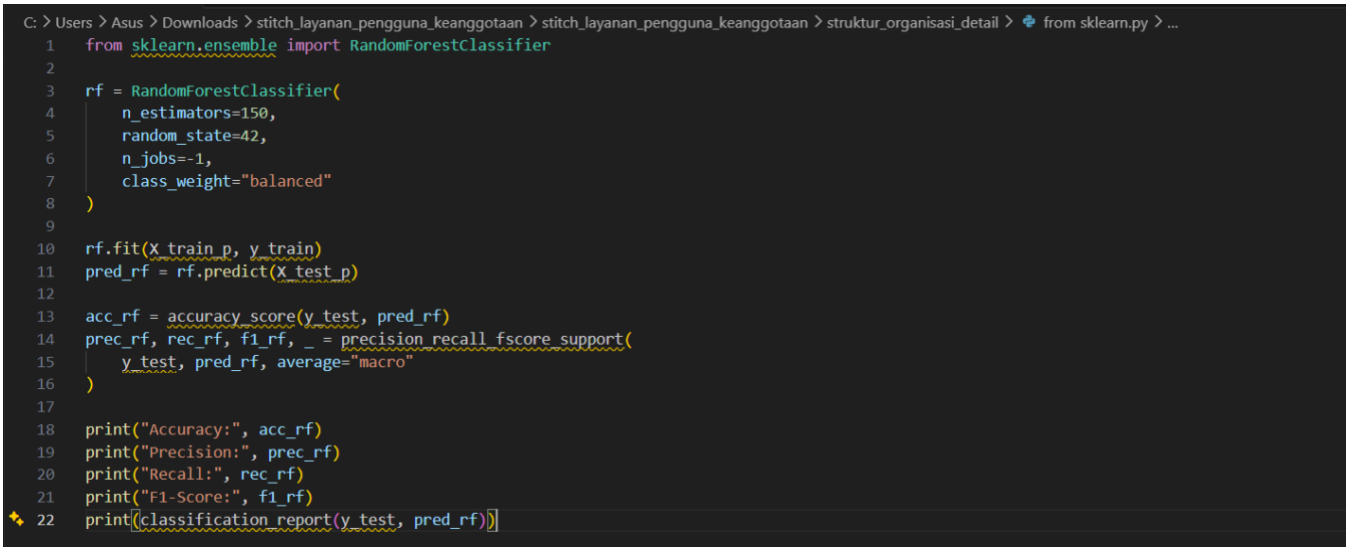
Hasil evaluasi model *Logistic Regression* menunjukkan performa yang relatif rendah dengan accuracy sebesar 0,4550, *precision macro* sebesar 0,4378, recall macro sebesar 0,4550, dan F1-score macro sebesar 0,4201. Model ini cukup baik dalam mengenali kelas Rendah dan Sedang, tetapi sangat lemah pada kelas Tinggi, yang menunjukkan bahwa model ini kurang mampu menangkap kompleksitas hubungan antara metadata produk dan popularitas produk. Hal ini menandakan bahwa hubungan antara variabel tersebut tidak sepenuhnya linier, dan model yang lebih fleksibel perlu diuji untuk meningkatkan performa prediksi.

Tabel 3. Hasil Evaluasi Logistic Regression

Metrik	Nilai
Accuracy	0,4550
Precision macro	0,4378
Recall macro	0,4550
F1-score macro	0,4201

4.5 Pemodelan Menggunakan *Random Forest*

Model kedua adalah *Random Forest*. Algoritma ini bekerja dengan banyak pohon keputusan dan menggabungkan hasilnya menjadi satu keputusan akhir. *Random Forest* umumnya lebih baik dalam menangkap hubungan nonlinier dan interaksi antarvariabel.



```
C: > Users > Asus > Downloads > stitch_layanan_pengguna_keanggotaan > stitch_layanan_pengguna_keanggotaan > struktur_organisasi_detail > from sklearn.py > ...
1  from sklearn.ensemble import RandomForestClassifier
2
3  rf = RandomForestClassifier(
4      n_estimators=150,
5      random_state=42,
6      n_jobs=-1,
7      class_weight="balanced"
8  )
9
10 rf.fit(X_train_p, y_train)
11 pred_rf = rf.predict(X_test_p)
12
13 acc_rf = accuracy_score(y_test, pred_rf)
14 prec_rf, rec_rf, f1_rf, _ = precision_recall_fscore_support(
15     y_test, pred_rf, average="macro"
16 )
17
18 print("Accuracy:", acc_rf)
19 print("Precision:", prec_rf)
20 print("Recall:", rec_rf)
21 print("F1-Score:", f1_rf)
22 print(classification_report(y_test, pred_rf))
```

Gambar 5. Pemodelan Menggunakan *Random Forest*

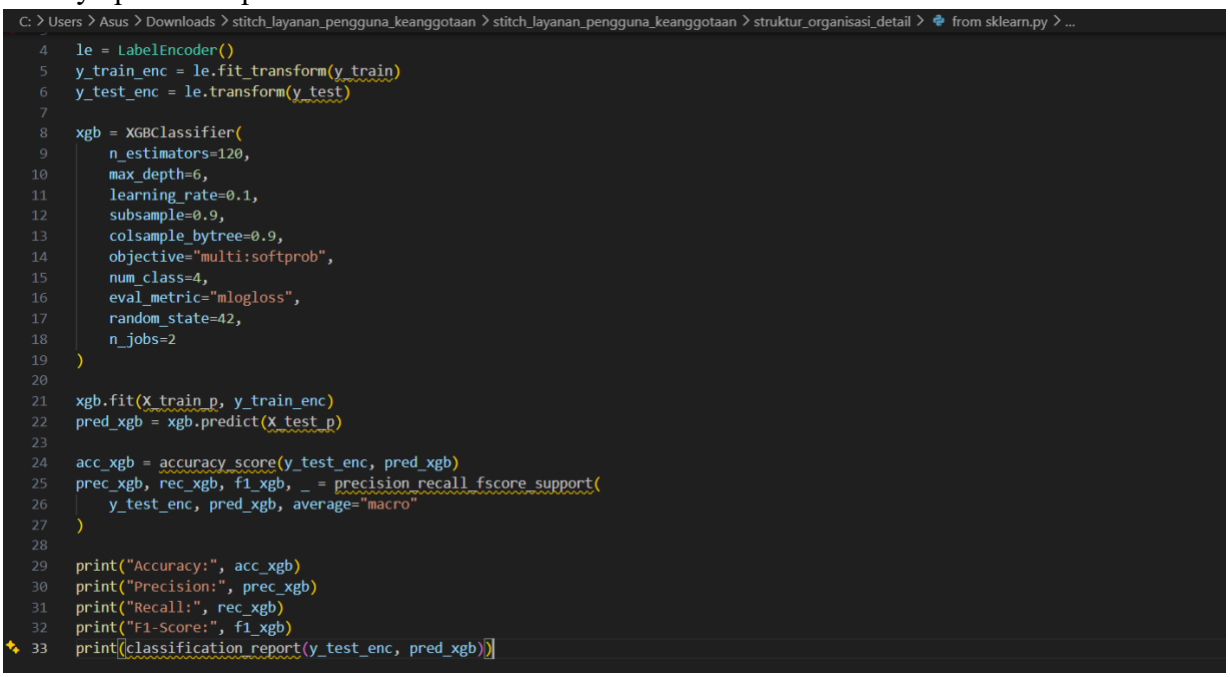
Hasil evaluasi model *Random Forest* menunjukkan performa yang lebih baik dibandingkan dengan Logistic Regression, dengan accuracy sebesar 0,4806, precision macro sebesar 0,4748, recall macro sebesar 0,4805, dan F1-score macro sebesar 0,4771. Model ini lebih stabil dalam mengenali kelas Rendah, Sedang, dan Sangat Tinggi, namun tetap memiliki kesulitan dalam memprediksi kelas Tinggi. Hal ini menunjukkan bahwa pola popularitas produk cukup kompleks dan tidak sepenuhnya dapat dipisahkan dengan jelas menggunakan model sederhana. Meski demikian, *Random Forest* memberikan hasil yang lebih memadai dibandingkan model sebelumnya, sehingga lebih cocok untuk menangkap pola dalam data yang lebih rumit.

Tabel 4. Hasil Pemodelan Menggunakan *Random Forest*

Metrik	Nilai
Accuracy	0,4806
Precision macro	0,4748
Recall macro	0,4805
F1-score macro	0,4771

4.6 Pemodelan Menggunakan XGBoost

Model ketiga adalah XGBoost. Algoritma ini berbasis boosting dan dikenal sangat kuat dalam data tabular. XGBoost membangun pohon keputusan secara bertahap dan memperbaiki kesalahan model sebelumnya pada setiap iterasi.



Gambar 6. Pemodelan Menggunakan XGBoost

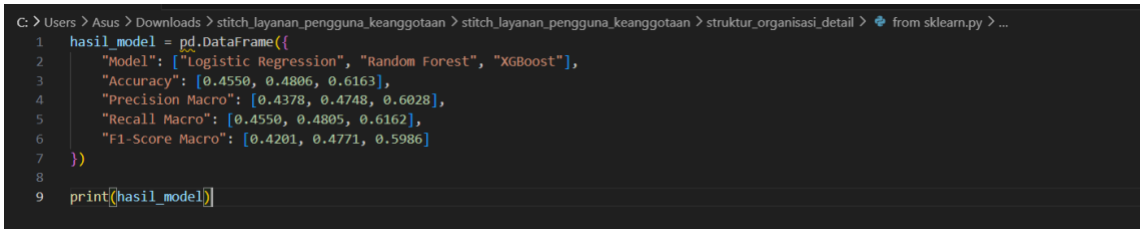
Hasil evaluasi model XGBoost menunjukkan performa terbaik di antara semua model yang diuji, dengan accuracy sebesar 0,6163, precision macro sebesar 0,6028, recall macro sebesar 0,6162, dan F1-score macro sebesar 0,5986. Nilai accuracy dan F1-score macro yang paling tinggi menunjukkan bahwa XGBoost paling efektif dalam memprediksi popularitas produk elektronik berdasarkan metadata. Kinerja XGBoost yang lebih baik menegaskan bahwa hubungan antarfitur dalam data produk cukup kompleks dan membutuhkan model yang dapat menangkap pola-pola bertingkat yang ada.

Tabel 5. Pemodelan Menggunakan XGBoost

Metrik	Nilai
Accuracy	0,6163
Precision macro	0,6028
Recall macro	0,6162
F1-score macro	0,5986

4.7 Perbandingan Kinerja Model

Setelah seluruh model diuji, langkah berikutnya adalah membandingkan hasilnya secara langsung agar terlihat model mana yang paling unggul.



Gambar 7. Perbandingan Kinerja Model

Hasil perbandingan kinerja antara model Logistic Regression, *Random Forest*, dan XGBoost menunjukkan bahwa XGBoost memiliki performa terbaik di antara ketiganya. Dengan accuracy sebesar 0,6163, precision macro sebesar 0,6028, recall macro sebesar 0,6162, dan F1-score macro sebesar 0,5986, XGBoost menunjukkan kemampuan terbaik dalam memprediksi popularitas produk. *Random Forest* berada di posisi kedua dengan performa yang lebih baik dibandingkan Logistic Regression, yang menunjukkan hasil lebih rendah dalam semua metrik evaluasi. Hal ini mengindikasikan bahwa model berbasis ensemble seperti *Random Forest* dan XGBoost lebih efektif dalam menangkap pola yang lebih kompleks pada data metadata produk, sedangkan Logistic Regression lebih terbatas dalam menangani hubungan non-linier.

Tabel 6. Hasil Perbadingan Kinerja Model

Model	Accuracy	Precision Macro	Recall Macro	F1-Score Macro
Logistic Regression	0,4550	0,4378	0,4550	0,4201
Random Forest	0,4806	0,4748	0,4805	0,4771
XGBoost	0,6163	0,6028	0,6162	0,5986

4.8 Pembahasan

Hasil penelitian ini menunjukkan bahwa metadata produk seperti kategori, merek, harga, dan rating sudah cukup informatif untuk digunakan dalam memprediksi popularitas produk elektronik di marketplace Lazada Indonesia. Hal ini membuktikan bahwa meskipun data yang digunakan relatif sederhana, yaitu metadata yang sudah tersedia dalam bentuk terstruktur, informasi ini dapat memberikan gambaran yang kuat mengenai popularitas produk. Dengan menggunakan data ini, model prediksi dapat membantu dalam membuat keputusan yang lebih terarah terkait dengan pemasaran dan strategi katalog.

Meskipun Logistic Regression telah diterapkan, model ini menunjukkan performa yang lebih rendah dibandingkan dengan model lainnya. Hal ini disebabkan oleh keterbatasan Logistic Regression dalam menangkap hubungan yang lebih rumit dan non-linier antar fitur-fitur produk. Sebagai model linear,

Logistic Regression cenderung tidak cukup fleksibel untuk menangani kompleksitas pola dalam data e-commerce, seperti variasi harga dan jumlah ulasan yang sangat berbeda-beda antar produk. Oleh karena itu, meskipun model ini berguna untuk kasus sederhana, dalam konteks data yang lebih kompleks, performanya terbatas.

Sebaliknya, *Random Forest* menunjukkan performa yang lebih baik karena kemampuannya dalam menangkap pola non-linier dan interaksi antar variabel. *Random Forest* bekerja dengan membangun banyak pohon keputusan secara acak dan menggabungkannya untuk menghasilkan hasil yang lebih stabil dan akurat. Hasil ini menunjukkan bahwa model berbasis ensemble lebih efektif dalam menangani data dengan banyak fitur dan variabel yang memiliki hubungan kompleks. Meskipun lebih baik daripada Logistic Regression, *Random Forest* masih memiliki keterbatasan dalam hal prediksi yang sangat kompleks, terutama pada kelas Tinggi.

XGBoost terbukti memberikan hasil terbaik dibandingkan dengan kedua model lainnya. Model ini bekerja dengan mengoptimalkan pohon keputusan secara bertahap melalui proses boosting, yang memungkinkan model untuk memperbaiki kesalahan prediksi pada iterasi berikutnya. XGBoost lebih adaptif dalam menangani data yang besar dan kompleks, serta lebih efisien dalam mengatasi ketidakseimbangan kelas dan interaksi antar fitur. Hal ini menjelaskan mengapa XGBoost menghasilkan accuracy dan F1-score tertinggi, menunjukkan kemampuannya dalam memprediksi popularitas produk lebih akurat dibandingkan model lain.

Temuan ini menegaskan bahwa karakteristik data *e-commerce* memang kompleks dan sulit diprediksi dengan model linear sederhana. Harga yang bervariasi, jumlah ulasan yang timpang, serta keberagaman merek dan kategori produk membuat model berbasis ensemble seperti *Random Forest* dan XGBoost lebih efektif dalam menangkap pola-pola yang ada. Dari sisi praktis, model XGBoost dapat sangat berguna bagi penjual dan pengelola marketplace dalam mengidentifikasi produk yang berpotensi populer lebih awal. Hal ini memungkinkan mereka untuk mengoptimalkan pengaturan promosi, pengelolaan stok, dan penempatan produk dalam katalog digital secara lebih efisien. Model ini juga dapat menjadi alat yang efektif untuk membantu pengelolaan marketplace dalam merumuskan strategi berbasis data yang lebih tepat.

5. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa metadata produk seperti kategori, merek, harga, dan rating dapat digunakan secara efektif untuk memprediksi popularitas produk elektronik di marketplace Lazada Indonesia. Meskipun Logistic Regression memberikan hasil yang terbatas karena sifatnya yang linier, *Random Forest* dan XGBoost menunjukkan peningkatan performa yang signifikan, dengan XGBoost menghasilkan hasil terbaik. Hal ini menunjukkan bahwa model berbasis ensemble lebih efektif dalam menangani kompleksitas data *e-commerce* yang memiliki variabilitas tinggi pada fitur-fitur produk. Temuan ini memberikan kontribusi praktis dengan menunjukkan bahwa XGBoost dapat digunakan untuk membantu pengelolaan produk di marketplace, seperti identifikasi produk yang berpotensi populer lebih awal, pengaturan promosi, dan pengelolaan stok yang lebih efisien.

DAFTAR PUSTAKA

[1] D. Mubarok, K. Adjani, B. D. R. Hutama, M. M. Mutoffar, And R. Indrayani, "Big Data Analytics Dan Machine Learning Untuk Memprediksi Perilaku Konsumen Di E-Commerce," *J. Inform. Dan Rekayasa Elektron.*, Vol. 8, No. 1, Pp. 159–167, 2025.

[2] E. A. Apriadi And M. Bisri, "Optimization Of Bpjs Health Facility Distribution With K-Means Clustering Algorithm," *Int. J. Technol. Comput. Sci.*, Vol. 1, No. 1, Pp. 1–13, 2025.

[3] S. P. Veviarosi *Et Al.*, *Membangun Konektivitas Wireless, Mobile Computing, And Mobile Commerce*. Fahmi Karya, 2025.

[4] A. H. Purwantini, "Informasi Dan E-Commerce," *Sist. Inf. Manaj.*, P. 59.

[5] F. K. Ihtada, "Studi Perbandingan Metode Ekstraksi Fitur Untuk Topic Modeling Berbasis Aspek Dan Sentimen Analisis Pada Ulasan Produk E-Commerce," 2025, *Universitas Islam Negeri Maulana Malik Ibrahim*.

[6] M. I. Akbar, "Implementasi Natural Language Processing Untuk Analisis Sentimen Produk Pada Marketplace," 2025, *Universitas Islam Indragiri*.

[7] I. Budi, *Analisis Sentimen Berdasarkan Data Media Sosial: Studi Kasus Pada Berbagai Domain Di Indonesia*. Universitas Indonesia Publishing.

[8] L. Dwi, "Perbandingan Performa Model Prediksi Customer Churn Berbasis Machine Learning Pada Fashion E-Commerce," 2023.

[9] S. A. H. Bahtiar, "Perbandingan Naïve Bayes Dan Logistic Regression Dalam Sentiment Analysis Pada Review Marketplace Menggunakan Rating-Based Labeling," 2023.

[10] K. B. Sirait And E. A. Apriadi, "Ai-Based Business Management Transformation: Improving Operational Efficiency Through Intelligent Systems," *Int. J. Technol. Comput. Sci.*, Vol. 1, No. 1, Pp. 51–65, 2025.

[11] E. A. Apriadi, S. Lestari, And S. Y. Irianto, "Comparison Of Performance Of K-Nearest Neighbors And Neural Network Algorithm In Bitcoin Price Prediction," *Sink. J. Dan Penelit. Tek. Inform.*, Vol. 8, No. 2, Pp. 617–622, 2024.

[12] E. A. Apriadi *Et Al.*, *Kecerdasan Buatan Teori, Implementasi, Dan Aplikasi Di Era Digital*. Eko Aziz Apriadi, 2025.

[13] E. A. Apriadi, "A Literature Review On The Role Of Ai (Artificial Intelligence) In Industry 4.0 Transformation," *Int. J. Technol. Comput. Sci.*, Vol. 1, No. 1, Pp. 25–37, 2025.

[14] S. Yuliasari And A. Rahmatulloh, "Performance Analysis And Accuracy Of Machine Learning Algorithms For Heart Disease Prediction," *Telemat. J. Telemat. Dan Teknol. Inf.*, Vol. 22, No. 3, Pp. 98–106, 2025.

[15] A. Agustiningsih, Y. Findawati, And I. A. Kautsar, "Classification Of Vocational High School Graduates'ability In Industry Using Extreme Gradient Boosting (Xgboost), Random Forest, And Logistic Regression," *J. Tek. Inform.*, Vol. 4, No. 4, Pp. 977–985, 2023.

[16] S. Sunardi And S. Suyahman, "Analisis Komparasi Prediksi Serangan Ddos Menggunakan Machine Learning," In *Proceeding Of Informatics Collaborations And Dessimenation Meeting*, 2025, Pp. 84–91.

[17] P. K. Akbar, F. R. Noverio, And M. R. Mumtaz, "Klasifikasi Prediksi Risiko Gagal Jantung Menggunakan Metode Machine Learning Studi Komparatif Logistic Regression Random Forest Dan Xgboost," In *Prosiding Seminar Nasional Penelitian Lppm Umj*, 2025.

[18] H. Putra And R. Rumini, "Comparative Study Of Logistic Regression, Random Forest, And Xgboost For Bank Loan Approval Classification," *J. Appl. Informatics Comput.*, Vol. 9, No. 5, Pp. 2822–2835, 2025.

[19] R. A. Wardana, "Klasifikasi Tingkat Nilai Matematika Siswa Menggunakan Metode Machine Learning Berbasis Faktor Sosial Dan Perilaku," In *Prosiding Seminar Nasional Indonesia*, 2026, Pp. 46–63.

[20] A. Astofa, P. Rosyani, R. Rahmawati, And S. Apandi, "Evaluasi Komparatif Algoritma Machine Learning Untuk Prediksi Dini Diabetes," *Bull. Comput. Sci. Res.*, Vol. 6, No. 1, Pp. 558–565, 2025.