JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024, Pp. 123 – 133
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Ensemble Learning pada Kategorisasi Produk *E-Commerce* Menggunakan Teknik *Boosting*

Genta Dwigi Sepbriant [(1)*], Danang Wahyu Utomo [(2)]
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang
e-mail : 111202013101@mhs.dinus.ac.id, danang.wu@dsn.dinus.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 31 Januari 2024, direvisi 16 Maret 2024, diterima 17 Maret 2024, dan dipublikasikan 25 Mei 2024.

Abstract

The development of e-commerce significantly contributes to technological advancement, especially for businesses adopting the concept. The growth of e-commerce has seen a significant increase, reaching 196.47 million users in 2023. In e-commerce, a wide range of product variations is provided to users, which can lead to errors or confusion in product selection. Product categorization is crucial in e-commerce to assist users in navigating efficiently. However, manual categorization is less effective as it can be time-consuming. This study aims to clarify the factors of concern in grouping using the K-Nearest Neighbors (KNN) algorithm in product categorization on the e-commerce platform. This research focuses on whether the novelty lies in the implemented algorithm, the variables used, or the applied grouping parameters. This work applies the XGBoost algorithm to improve the effectiveness of product categorization in e-commerce through ensemble learning approaches. The research findings indicate that boosting algorithms like XGBoost outperform individual algorithms like KNN regarding classification accuracy. This proves that ensemble learning approaches may greatly enhance product classification in e-commerce. The testing process of the implemented e-commerce system in this study also provides confidence in the theoretical and practical benefits of applying this research to enhance efficiency and user experience in product categorization on the e-commerce platform.

Keywords: Product Categorization, E-Commerce, Ensemble Learning, XGBoost, Boosting

Abstrak

Perkembangan *e-commerce* memberikan kontribusi nyata dalam perkembangan tektnologi terutama pada perusahaan yang menjalankan konsep bisnis. Perkembangan *e-commerce* saat ini meningkat secara signifikan dengam mencapai 196,47 juta jiwa pada tahun 2023. Dalam *e-commerce* tentunya banyak memberikan variasi produk pilihan kepada pengguna. Hal ini dapat mengakibatkan kesalahan atau kekeliruan dalam pemilihan produk. Kategorisasi produk menjadi penting dalam sebuah *e-commerce* untuk membantu pengguna menavigasi dengan efisien. Namun, kategorisasi manual kurang efektif karena dapat memakan waktu yang cukup lama. Penelitian ini bertujuan untuk mengklarifikasi faktor-faktor yang menjadi perhatian dalam pengelompokan menggunakan algoritma K-Nearest Neighbors (KNN) pada kategorisasi produk dalam platform *e-commerce*. Fokus penelitian ini adalah pada apakah keunikan (*novelty*) terletak pada algoritma yang diimplementasikan, variabel yang digunakan, atau pada parameter pengelompokkan yang diterapkan. Dengan menggunakan teknik *ensemble learning*, penelitian ini mengimplementasikan algoritma XGBoost untuk meningkatkan efisiensi kategorisasi produk dalam *e-commerce*. Hasil penelitian menunjukkan bahwa algoritma *boosting* seperti XGBoost mampu mengungguli kinerja algoritma individu seperti KNN dalam hal akurasi klasifikasi. Ini menunjukkan bahwa dengan memanfaatkan teknik *ensemble learning*, kategorisasi produk dalam *e-commerce* dapat ditingkatkan secara signifikan. Proses pengujian sistem *e-commerce* yang diimplementasikan dalam penelitian ini turut memberikan keyakinan terkait manfaat penerapan penelitian ini secara teoritis maupun praktis dalam meningkatkan efisiensi dan pengalaman pengguna dalam kategorisasi produk di platform *e-commerce*.

Kata Kunci: Kategorisasi Produk, *E-Commerce*, *Ensemble Learning*, XGBoost, *Boosting*

1. PENDAHULUAN

Saat ini *e-commerce* menunjukkan perkembangan yang signifikan dalam layanan jual beli melalui internet. *E-commerce* memberikan kontribusi nyata dalam perkembangan teknologi terutama bagi perusahaan yang menjalankan konsep bisnis. *E-commerce* menjadi isu dalam kehidupan sehari-hari terutama bagi pelaku bisnisatau Perusahaan (Gomero-Fanny et al., 2021). Para pelaku bisnis, organisasi dan/atau Perusahaan saat ini menjalankan layanan *e-commerce* melalui telepon selular yang dapat memudahkan para pengguna dalam hubungan bisnis dan komunikasi. Selain itu, juga memudahkan dalam melakukan jangkauan pelanggan, perluasan jaringan pemasaran dan transaksi dalam bentuk apapun. Perkembangan *e-commerce* saat ini mampu mengubah aturan bisnis dan segala jenis transaksi (Huang et al., 2019). Pada penelitian lainnya juga menyatakan bahwa saat ini Perusahaan menggunakan perkembangan teknologi dalam platform jual beli *online* (Indasari & Tjahyanto, 2023). *E-commerce* menjadi solusi utama dalam bagi pelaku bisnis dalam menawarkan produknya dalam singkat, sedangkan bagi pengguna (*user*) memudahkan pencarian produk dalam waktu singkat.

Dalam era digital saat ini, jumlah data yang besar dan meningkat memberikan tantangan baru dalam pengolahan informasi. *E-commerce* menjadi salah satu isu yang melibatkan transaksi data dalam jumlah besar. Berdasarkan data pada laman dataindonesia.id menunjukkan bahwa pengguna *e-commerce* tahun 2023 meningkat dari tahun sebelumnya yaitu 10% menjadi 196,47 juta jiwa. Dari data tersebut dapat disimpulkan jika pengguna *e-commerce* melakukan transaksi, maka terjadi transaksi yang melibatkan data dan informasi dalam jumlah besar. Adanya transaksi dalam jumlah besar menjadikan perusahaan melakukan investasi besar terhadap platform *e-commerce*. Menurut Mashalah et al. (2022), terdapat 3 (tiga) faktor yang mendorong perkembangan *e-commerce*: teknologi pendukung, persaingan, dan perilaku pengguna.

Permasalahan umum yang banyak dibahas dalam *e-commerce* adalah banyaknya jumlah produk atau jenis produk yang dipasarkan pada platform *online* (Ansharullah et al., 2023; Donati et al., 2019; Indasari & Tjahyanto, 2023). Banyaknya variasi produk memberikan banyak pilihan produk kepada pengguna yang dapat menyebabkan kesalahan atau kekeliruan dalam pemilihan produk. Bagi perusahaan, hal ini menjadi masalah besar karena ada potensi produk tertentu tidak dipilih oleh konsumen. Bagi pengguna yang masih awam tentang platform *e-commerce*, banyaknya variasi produk dapat menyebabkan kesalahan pemilihan produk karena harus memahami deskripsi produk satu per satu. Penelitian lain menyatakan permasalahan bahwa produk yang sama dijual pada beberapa toko, namun dengan informasi yang berbeda (Ristoski et al., 2018). Perlu adanya kategorisasi produk untuk membantu pengguna awam dalam menemukan produk yang sesuai.

Dalam platform *e-commerce* populer, menyediakan banyak ragam produk yang ditampilkan. Masing-masing produk memberikan deskripsi dan ulasan yang berbeda. Dibutuhkan adanya kategorisasi untuk produk-produk tersebut dengan tujuan: rekomendasi produk, prediksi produk (Jain & Kumar, 2020), dan kategorisasi yang sesuai dengan deskripsi (Tan et al., 2020). Kategorisasi secara manual dengan pemberian label pada masing-masing produk membutuhkan waktu lama dan tidak ada jaminan ketepatan dalam penempatan deskripsi produk dalam suatu kategori. Menurut Jahanshahi et al. (2021), mengusulkan kategori yang sesuai dengan deskripsi membutuhkan waktu yang lama dan kompleksitas semakin meningkat.

Permasalahan lainnya, kategorisasi produk dapat dipengaruhi berdasarkan input nama produk. Adanya kesamaan nama produk dengan deskripsi yang berbeda dapat mempengaruhi ketepatan dalam kategorisasi produk. Pada penelitian yang dilakukan Jahanshahi et al. (2021) menunjukan adanya perbedaan kategori dengan sub kategori berdasarkan input nama produk. Menurut Kim et al. (2021), kategorisasi yang dilakukan manusia sangat sulit dilakukan untuk mendapat hasil yang akurat dan cepat hanya berdasarkan nama. Ketidaktepatan kategorisasi produk juga dapat dilakukan oleh operator (Ozyegen et al., 2022). Rekomendasi kategorisasi oleh operator dapat menghasilkan kategorisasi yang subyektif atau kurang tepat. Adanya penambahan atau *update* produk baru dalam *e-commerce* dapat mempengaruhi tingkat akurasi kategorisasi produk.

Berdasarkan permasalahan di atas, beberapa solusi telah diusulkan untuk kategorisasi produk yang lebih baik yaitu: teknik penyematan kata (Sharma & Sagvekar, 2023); klasifikasi (Pawłowski, 2022; Perdana et al., 2021); dan translasi mesin (Tan et al., 2020). Kategorisasi produk berbasis klasifikasi teks menjadi solusi yang populer dan terbukti mampu menempatkan produk *e-commerce* sesuai dengan kategorinya. Dari penelitian yang dilakukan oleh Patra et al. (2021), hasil menunjukkan bahwa klasifikasi produk menggunakan model pembelajaran mesin (*machine learning* atau ML) mampu memberikan akurasi terbaik dalam mengklasifikan produk berdasarkan kategori yang telah ditentukan. Pada penelitian lainnya, Pothuganti (2019) mengusulkan algoritma *Ordered Weighted Averaging Combination* (OWC) untuk mengenali identitas produk baru yang tidak dikenali. Tujuannya, memudahkan kategorisasi produk yang baru ditambahkan pada platform *e-commerce*. Pengklasifikasi produk otomatis diusulkan oleh Lee & Yoon (2018) dengan tujuan untuk mengklasifikasikan produk berdasarkan deskripsi dan dokumen doc2vec. Teknik, model, dan algoritma klasifikasi yang diusulkan masing-masing memiliki performa baik dalam mengklasifikasikan produk *e-commerce*.

Pada pendekatan *data mining*, model *Machine Learning (*ML) telah diusulkan untuk meningkatkan performa, akurasi, dan kinerja model yang diusulkan. *Ensemble learning* adalah pendekatan yang terdiri dari penggabungan model *Machine Learning* (ML) untuk meningkatkan kinerja kategorisasi atau klasifikasi. Arumnisaa & Wijayanto (2023) mengusulkan *ensemble classifier* untuk meningkatkan akurasi dalam klasifikasi. Hasil akurasi yang dihasilkan lebih tinggi dari algoritma individu. Pada penelitian lainnya diusulkan analisis perbandingan terhadap model *ensemble* dengan *neural network* untuk mengetahui akurasi terbaik dalam klasifikasi produk *e-commerce* (Kalaivani, 2020). Pada perbandingan tersebut, algoritma AdaBoost dengan SVM menghasilkan akurasi terbaik. Pada penelitian lainnya, *ensemble learning* diusulkan dalam klasifikasi *review* produk dengan hasil bahwa algoritma *boosting* mampu mengungguli algoritma individu seperti *K-Nearest Neighbors* (KNN) dalam akurasi klasifikasi (Fayaz et al., 2020). Algoritma *boosting* seperti *extreme gradient boosting* (XGBoost), AdaBoost mampu memberikan akurasi terbaik dibandingkan algoritma individu dalam pembelajaran mesin.

Pada penelitian ini mengusulkan algoritma XGBoost dalam kategorisasi produk *e-commerce*. XGBoost merupakan pendekatan berbasis *boosting* yang terdiri dari beberapa *decision tree* di mana pohon sebelum dan berikutnya akan saling bergantungan. XGBoost menggabungkan berbagai metode pengklasifikasian lemah dengan melatih model baru secara berurutan dengan menggunakan model klasifikasi sebelumnya. Pada eksperimen menggunakan data uji *e-commerce dataset* yang terdiri dari atribut deskripsi dan kategori dengan jumlah *dataset* 50.434 data produk. Uji coba juga menerapkan *hyperparameter tunning* untuk mencari akurasi terbaik.

2. METODE PENELITIAN

2.1 *Ensemble Learning*

Ensemble learning adalah metode dalam *Machine Learning* (ML) dengan menggabungkan dua atau lebih algoritma dengan tujuan meningkatkan akurasi, mengurangi kesalahan, atau bias model yang dihasilkan oleh algoritma individu. Mienye juga menyatakan definisi dari *ensemble learning* yaitu suatu teknik dengan kombinasi algoritma *Machine Learning* (ML) untuk mendapatkan performa superior dibandingkan dengan algoritma yang digunakan secara tunggal (Mienye & Sun, 2022). Peneliti lain menyatakan bahwa *ensemble learning* merupakan metode yang melibatkan beberapa algoritma dengan performa lemah kemudian dikombinasikan untuk mengasilkan performa yang lebih baik (Dong et al., 2020). Secara umum, tujuan utama dari metode *ensemble learning* adalah menggabungkan algoritma ML untuk mencari performa terbaik dari penggabungan yang dilakukan. Menurut Mienye, terdapat 3 (tiga) teknik yaitu: *bagging, boosting,* dan *stacking.* Penelitian ini fokus pada teknik *boosting* yaitu teknik dengan pemrosesan sekuensial berdasarkan proses dari model sebelumnya.

2.2 Pendekatan *Boosting*

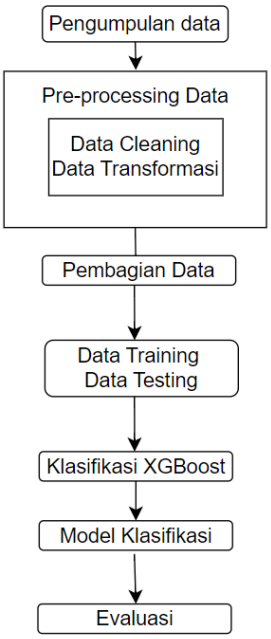
Pendekatan *boosting* adalah metode yang melibatkan penerapan secara *iterative* dari algoritma *boosting* dasar ke versi yang disesuaikan dari data masukan (Mienye & Sun, 2022). Metode *boosting* biasanya menggunakan data masukan untuk melatih model *boosting* lemah, kesalahan pada klasifikasi, dan melatih model tersebut dengan set yang disesuaikan pada klasifikasi sebelumnya. Peneliti lain menyatakan bahwa metode *boosting* ini berfokus pada pengurangan bias daripada variasi dengan meningkatkan *boosting* awal dasar yang memiliki bias tinggi. Secara keseluruhan pendekatan *boosting* adalah metode yang digunakan untuk mengurangi kesalahan dalam analisis data prediktif dan meningkatkan kinerja model pada data kompleks atau sulit diklasifikasikan dan dapat bekerja baik dengan berbagai algoritma.

2.3 Algoritma XGBoost

Algoritma XGBoost merupakan algoritma dengan implementasi *decision-tree* yang menggunakan kerangka dari *Gradient Boosting*. XGBoost merupakan implementasi dari *Gradient Boosting* untuk meningkatkan performa kinerja dan stabilitas. Pada kasus klasifikasi dan regresi, algoritma XGBoost tepat diusulkan karena mengacu pada pohon keputusan terbaik. Menurut Jafarzadeh et al menyatakan bahwa algoritma XGBoost adalah algoritma terbaik jika dibandingkan dengan algoritma ML lainnya (Jafarzadeh et al., 2021). Beberapa keuntungan algoritma XGBoost adalah mampu menangani *dataset* yang besar, data yang hilang, dan mencegah terjadinya *overfitting* (Nobre & Neves, 2019). Adanya penggunaan *tree depth, learning rate* dan *subsampling* menjadikan algoritma XGBoost menghasilkan akurasi lebih baik dibandingkan algoritma *Machine Learning* (ML) lainnya. Adapum formula dan variabel yang digunakan dalam koteks penggunaan algoritma XGBoost untuk klasifikasi dalam *e-commerce* dirumuskan pada Pers. (1). Di mana $F(x)$ adalah fungsi prediksi untuk kelas tertentu, ω_0 adalah bias, M adalah jumlah pohon keputusan (*boosting* rounds), dan $f_m(x)$ adalah fungsi dari pohon keputusan ke-m.

$$F(x) = \omega_0 + \sum_{m=1}^{M} f_m(x) \quad (1)$$

2.4 *Dataset*



Gambar 1 Tahap Eksperimen

Pada eksperimen menggunakan *e-commerce dataset* sebagai uji kategorisasi produk. Penulis melakukan pengambilan data dari *website* https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification. *Dataset* ini merupakan data untuk klasifikasi berdasarkan katagori pada platfrom *e-commerce*, yang terdiri dari 4 kategori utama, yaitu "*Electronics", "Household", "Books"*, dan *"Clothing & Accessories"* yang hampir mencakup 80% dari produk yang biasanya ada disitus *e-commerce*. *Dataset* ini disajikan dalam format ".csv" dengan kolom pertama adalah nama kategori dan kolom kedua data poin (deskripsi produk) dari kategori tersebut. pengaturan atribut adalah *label* dan *desc*.

Pengaturan *dataset* menggunakan alat bantu Google Colaboratory. Pada kolom *label* merupakan kolom kategori dari suatu produk *e-commerce* yaitu: *household, books, clothing & accessories,* dan *electronics*. Kolom *desc* sebagai kolom deskripsi suatu produk yaitu berisi penjelasan informasi produk. *Dataset* menampilkan sampel *dataset* dengan pengaturan atribut adalah *label* dan *desc*. Pengaturan *dataset* menggunakan alat bantu Google Colaboratory. Pada kolom *label* merupakan kolom kategori dari suatu produk *e-commerce* yaitu: *household, books, clothing & accessories,* dan *electronics*. Kolom *desc* sebagai kolom deskripsi suatu produk yaitu berisi penjelasan informasi produk.

2.5 Eksperimen

Tahap eksperimen sesuai dengan tahapan pada Gambar 1. Tahapannya yaitu terdiri dari pengumpulan data, pemrosesan awal data untuk melakukan pembersihan *dataset*, transformasi data, pengaturan kategorisasi produk, dan evaluasi menggunakan *confusion matrix.* Eksperimen menggunakan alat bantu Google Colaboratory dengan bahasa pemrograman Python.

2.5.1 Pemrosesan Awal Data

Pemrosesan awal data digunakan untuk melakukan normalisasi data seperti tahap pembersihan, transformasi data, dan integrasi data untuk disiapkan sebelum digunakan pada tahap analisis. Berikut adalah fungsi yang diterapkan pada pemrosesan awal data:

1) Konversi huruf dalam ukuran kecil atau *lowercase*.
2) Menghilangkan tanda baca seperti titik, koma, tanda tanya, petik satu, petik dua, dan strip.
3) *Stopword removal* menghilangkan kata yang tidak relevan berdasarkan daftar *library* yang ditentukan.
4) Menghilangkan teks numerik.

2.5.2 Tokenisasi

Tokenasi merupakan sebuah proses untuk membagi sejumlah teks baik kalimat maupun paragraf menjadi beberapa bagian tertentu. Dengan kata lain tokenisasi adalah tahapan untuk memotong struktur kalimat menjadi perkata. Fungsi tokenisasi dijalankan pada urutan ke 5 (lima) setelah menghilangkan teks numerik. Sebagai contoh kalimat "*Pitaara Box Offer Exclus collage"* kemudian diubah menjadi *"Pitaara"*, *"Box"*, *"Offer"*, *"Exclus"*, dan *"collage"*.

2.5.3 Lemmatisasi dan *Stemming*

Lemmatisasi adalah proses pengelompokan bentuk infleksi yang berbeda dari sebuah kata sehingga dapat dianalisis sebagai satu *item*. Lemmatisasi mirip dengan *stemming* dengan membawa konteks kata-kata dengan menggabungkan kata-kata yang mirip dengan satu kata. Lemmatisasi contoh kalimat "*textured print which gives*" menjadi "*texture print which give"* sedangkan untuk *stemming* contoh kalimat "*Paper plane design framed wall hanging motivationa*l" menjadi "*Paper plane design frame wall hang motiv*".

2.5.4 Transformasi Data

Berdasarkan jenis *dataset* yang digunakan, *dataset* tidak dapat langsung digunakan pada XGBoost karena *dataset* sendiri berupa data tekstual. Sebaliknya, diperlukan langkah transformasi data dari bentuk teks ke bentuk numerik menggunakan TF-IDF (Jahanshahi et al.,

2021). TF-IDF (*Term Frequency-Inverse Document Frequency*) merupakan gabungan dari dua konsep yaitu menghitung bobot dan nilai dengan menghitung frekuensi yang muncul pada teks berupa kata dokumen (Andriani & Wibowo, 2021). Tujuan dari metode ini yaitu untuk menentukan bobot pada kata-kata dalam teks dengan mengukur tingkat signifikansi suatu kata dalam kumpulan dokumen. Rumus perhitungan TF-IDF dituliskan pada Pers. (2) sampai (4), di mana $Tf_{i\,j}$ merupakan banyaknya kata i pada dokumen ke j, IDf_i yaitu banyaknya dokumen yang mengandung kata ke i, N adalah total dokumen, dan log merupakan logaritma natural.

$$Tf_{ij} = \frac{f_{i,j}}{\sum k\, f_{i,j}} \tag{2}$$

$$IDf_i = log\left(\frac{N}{n_i}\right) + 1 \tag{3}$$

$$TFIDF_{ij} = TF_{ij} \times IDf_i \tag{4}$$

2.5.5 Skenario Uji

Pada kategorisasi produk uji *dataset* dilakukan pengaturan pembagian data latih dan data uji. Besaran persentase data uji adalah 20% atau 0,2 dari ukuran *dataset*. Pengaturan selanjutnya adalah *hyperparameter tuning* yaitu pengaturan parameter pada implementasi fungsi XGBoost berdasarkan parameter *number of tree* atau *n_estimators, learning_rate,* dan *max_depth*. Parameter *n_estimators* digunakan untuk menentukan jumlah pohon keputusan yang dibuat secara paralel, *learning rate* adalah parameter yang digunakan sebagai laju pembelajaran dan *max_depth* adalah parameter yang digunakan untuk penentu kedalaman maksimum. Pada eksperimen ini dilakukan uji coba dengan pengaturan: *n_estimator*s=[100, 200]; *max_depth*=[3, 5, 7]; dan *learning_rate*=[0.1, 0.2]. Teknik *random search* diusulkan untuk mendapatkan hasil optimal pada kombinasi *hyperparameter* antara *n_estimators, max_depth* dan *learning_rate.*

2.5.6 *Confusion Matrix*

Confusion matrix digunakan sebagai alat bantu untuk evaluasi klasifikasi atau kategorisasi produk *e-commerce*. Kesesuaian atau ketetapan deskripsi produk pada suatu kategori diukur dengan *confusion matrix.* Pada Gambar 2, *confusion matrix* berupa tabel yang terdiri dari jumlah data kelas positif dan prediksi benar (TP), kelas negatif prediksi benar (FP), kelas positif prediksi salah (FN) dan kelas negatif prediksi salah (TN).

	Actual positive	***Actual negative***
Predicted positive	TP	FP
Predicted negative	FN	TN

Gambar 2 *Confusion Matrix*

Berdasarkan hasil uji akurasi setelah mengklasifikasikan produk, dilakukan dua kali pengujian seperti yang ditunjukkan pada Tabel 1. Pengujian pertama yakni dengan nilai *n_estimator:* 100 dan *n_estimator:* 200. Hasil dari pengujian pertama menunjukan bahwa *n_estimator:* 100, *learning_rate:* 0,2 dan *max_depth: 7* diperoleh akurasi sebesar 93,41%. Sedangkan pengujian kedua dengan nilai *n_estimator:* 200, pengujian dilakukan *n_estimator:* 200, *learning_rate:* 0,2 dan *max_depth: 7* diperoleh akurasi tertinggi sebesar 94,17%. Berdasarkan hasil tersebut, menunjukan bahwa pengujian yang dilakukan menemukan hasil yang mampu melakukan klasifikasi dengan baik.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 2, MEI, 2024: 123 – 133

Tabel 1 Hasil Uji Akurasi

Hyperparameter			Akurasi
n_estimators	learning_rate	max_depth	
100	0.1	3	86.85%
		5	90.99%
		7	91.83%
	0.2	3	91.35%
		5	92.77%
		7	93.41%
200	0.1	3	91.47%
		5	92.71%
		7	93.41%
	0.2	3	92.86%
		5	93.79%
		7	94.17%

3. HASIL DAN PEMBAHASAN

3.1 *Preprocessing*

Pada tahap pertama eksperimen dimulai dengan persiapan *dataset* kemudian dilakukan pembersihan data. Dari proses normalisasi data menghasilkan data yang berbeda dengan *dataset* sebelumnya. Pada Gambar 3, menunjukkan sampel *dataset* setelah dilakukan normalisasi data. Sampel data menunjukkan deskripsi suatu produk bersih tidak ada tanda baca atau simbol tertentu. Kemudian tulisan dalam bentuk *lowercase* dan kata dalam bentuk kata dasar (setelah melalui proses lemmatisasi dan *stemming*). Hal ini ditujukan untuk memudahkan komputasi pada model XGBoost dalam kategorisasi produk. Intensitas kemunculan masing-masing kata menjadi penentu kategori untuk deskripsi tersebut.

```
'pitaara box romant venic canva paint 6mm thick mdf frame 21 1 14  enclosur materi mdf mount frame 1 14  5
3 6 35 6  size 21 1 14 0  53 6 35 6  enhanc beauti room wall breathtak digit print artwork print technolog
captur everi detail imag print enhanc matt paint canva ensur rich live colour wall art panel mount mdf rea
di hang wall beauti interior home artwork gift live dine room outdoor galleri hotel restaur offic recept k
itchen area balconi bathroom pitaara box offer exclus collect thousand artwork digit paint canva print wal
l poster wall decor product home offic surround provid rang creativ spectacular art product use gift everi
occas everi season tag wall paint canva print modern art abstract design wallart artwork home bedroom dine
live draw room digit print bathroom common area kitchen offic decor stretch stretch frame frame beauti cla
ssi royal special uniqu eleg stylish creativ afford best photo gift fabric balconi interior exterior outdo
or galleri hotel restaur colour color small larg extra larg overs hang giant slim durabl waterproof buy sh
op purchas decor onlin place vintag canva romant venic artwork paint style'
```

Gambar 3 Sampel Data *Preprocessing*

Tahap selanjutnya adalah lemmatisasi dan *stemming*, yaitu mengubah kata atau kalimat menjadi kata dasar, sebagai contoh ditunjukkan pada Gambar 4. Proses *stemming* menjadikan input tiap kata ke bentuk akar atau dasar. Beberapa kata dasar mungkin asing atau dikenali karena ada beberapa yang dipotong. Sebagai contoh kata *romantic* setelah dilakukan stemming menjadi *romant*. Proses selanjutnya adalah transformasi data dengan melakukan TF-IDF yaitu melakukan vektorisasi teks mengubah teks dalam representasi vektor numerik. Hasil dari vektorisasi TF-IDF diterapkan pada model yang diusulkan untuk dihitung nilai akurasinya.

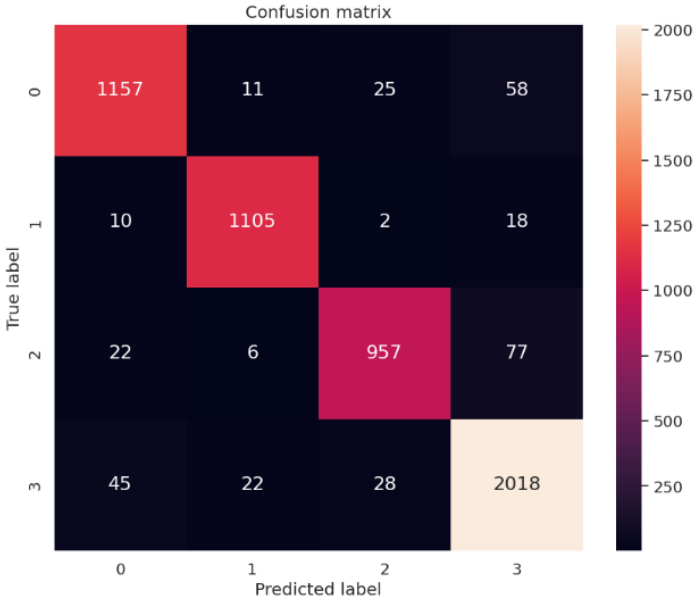
Input: Pitaara Box Romantic Venice Canvas Painting
Output: pitaara box romant venic canva

Gambar 4 Contoh Proses *Stemming*

3.2 *Confusion Matrix*

Proses selanjutnya adalah confusion matrik dengan mengevaluasi kinerja model klasifikasi dengan menyajikan informasi tentang hasil prediksi model terhadap data uji. Dari Gambar 5 menunjukkan *confusion matrix* memiliki empat sel utama, yang masing-masing menyajikan informasi yaitu:

1) *True Positive* (TP): Terletak pada diagonal utama matriks, di mana prediksi dan aktual keduanya positif.
2) *True Negative* (TN): Terletak di luar diagonal utama, di antara sel-sel yang tidak terlibat dalam *True Positive*.
3) *False Positive* (FP): Terletak di kolom yang sesuai dengan prediksi positif tetapi aktualnya negatif. Misalnya, untuk Prediksi 0, FN/FP terletak di baris Aktual 1, 2, dan 3.
4) *False Negative* (FN): Terletak di baris yang sesuai dengan aktual positif tetapi diprediksi sebagai negatif. Misalnya, untuk Aktual 0, FN/FP terletak di kolom Prediksi 1, 2, dan 3.



Gambar 5 Hasil *Confusion Matrix*

Tabel 2 Metrik Evaluasi

	Precision	Recall	F1-Score
0	0.92	0.94	0.93
1	0.97	0.97	0.97
2	0.90	0.95	0.92
3	0.96	0.93	0.94
	Akurasi		**0.94**

3.2.1 *Precision*

Presisi digunakan untuk mengukur seberapa akurat prediksi positif dalam kategorisasi produk. Dalam kasus kategorisasi produk ini, presisi digunakan mengetahui keakuratan produk yang diprediksi masuk dalam kategorisasi yang benar. Pada Tabel 2 menunjukkan bahwa produk yang dikategorikan pada label 0 adalah 94%, label 1 96%, label 2 95%, dan label 3 90%. Hal ini menunjukkan bahwa di atas 90% produk yang dikategorisasikan tepat atau benar, atau dengan kata lain di atas 90% model yang diprediksi positif yaitu benar-benar positif.

3.2.2 *Recall*

Recall digunakan untuk mengetahui nilai prediksi positif sebenarnya dalam kategorisasi produk. Sebagai contoh pada Tabel 2, label 1, *recall* menunjukkan 0,97 atau 97%. Artinya, 97% dari hasil positif sesuai diprediksi dengan tepat menggunakan algoritma XGBoost. Dalam hal ini, semakin tinggi persentase *recall* maka algoritma XGBoost semakin baik dalam menangani kategorisasi produk.

3.2.3 *F1-Score*

F1-Score merupakan matriks evaluasi yang digunakan untuk mencari keseimbangan antara nilai hasil *precision* dan *recall*. Pada Tabel 2, label 1 menunjukkan *F1-Score* adalah 0,97 atau 97% hal ini menunjukkan ada keseimbangan antara *precision* dan *recall*. Hal ini juga menunjukkan bahwa algoritma XGBoost mampu meminimalkan nilai *False Positive* dan *False Negative*.

4. KESIMPULAN

Setelah melakukan analisis dengan mengumpulkan data dan melakukan pengujian, penelitian ini mengusulkan penerapan XGboost dalam kategorisasi produk di platform *e-commerce* sebagai solusi umtuk meningkatkan akurasi klasifikasi. Dalam eksperimen yang dilakukan melalui pengujian dan variasi *hyperparameter*, ditemukan bahwa konfigurasi seperti *n_estimator*: 200, *learning_rate*: 0,2, dan *max_depth*: 7 mampu menghasilkan akurasi tertinggi sebesar 97,17%. Penggunaan matrik evaluasi seperti *precision*, *recall*, dan *F1-score* dalam mengevaluasi kinerja model menunjukan bahwa algoritma XGboost mampu memberikan prediksi yang akurat untuk setiap kategori produk, menawarkan potensi untuk meningkatkan efisiensi bisnis serta pengalaman pengguna dalam memilih produk dengan lebih efisien di lingkungan *e-commerce*. Dengan mengandalkan teknik *ensemble learning*, XGBoost dapat mengatasi tantangan dalam klasifikasi produk, mengoptimalkan proses kategorisasi, dan secara signifikan meningkatkan akurasi, yang pada gilirannya akan memberikan dampak positif pada operasional dan keuntungan bisnis. Kesimpulannya, penelitian ini memberikan wawasan yang berharga tentang penerapan XGBoost dalam konteks *e-commerce*, menyoroti potensi algoritma ini sebagai solusi efektif untuk meningkatkan kualitas layanan dan pengalaman pengguna di platform perdagangan *online*.

DAFTAR PUSTAKA

Andriani, N., & Wibowo, A. (2021). Implementasi Text Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Berbasis Web. *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya*, *2*(2), 130–137. https://conference.upnvj.ac.id/index.php/senamika/article/view/1807

Ansharullah, M. O., Agustin, W., Lusiana, Junadhi, Erlinda, S., & Zoromi, F. (2023). Product Classification Based on Categories and Customer Interests on the Shopee Marketplace Using the Naïve Bayes Method. *JAIA - Journal of Artificial Intelligence and Applications*, *2*(2), 15–22. https://doi.org/10.33372/jaia.v2i2.888

Arumnisaa, R. I., & Wijayanto, A. W. (2023). Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI). *SISTEMASI*, *12*(1), 206. https://doi.org/10.32520/stmsi.v12i1.2501

Donati, L., Iotti, E., Mordonini, G., & Prati, A. (2019). Fashion Product Classification through Deep Learning and Computer Vision. *Applied Sciences*, *9*(7), 1385. https://doi.org/10.3390/app9071385

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/S11704-019-8208-Z/METRICS

Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble Machine Learning Model for Classification of Spam Product Reviews. *Complexity*, *2020*, 1–10. https://doi.org/10.1155/2020/8857570

Gomero-Fanny, V., Ruiz, A., & Andrade-Arenas, L. (2021). Prototype of Web System for Organizations Dedicated to e-Commerce under the SCRUM Methodology. *International Journal of Advanced Computer Science and Applications*, *12*(1), 437–444. https://doi.org/10.14569/IJACSA.2021.0120152

Huang, Y., Chai, Y., Liu, Y., & Shen, J. (2019). Architecture of next-generation e-commerce platform. *Tsinghua Science and Technology*, *24*(1), 18–29. https://doi.org/10.26599/TST.2018.9010067

Indasari, S. S., & Tjahyanto, A. (2023). Automatic Categorization of Multi Marketplace FMCGs Products using TF-IDF and PCA Features. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(2), 198–204. https://doi.org/10.32736/sisfokom.v12i2.1621

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sensing*, *13*(21), 4405. https://doi.org/10.3390/rs13214405

Jahanshahi, H., Ozyegen, O., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2021). *Text Classification for Predicting Multi-level Product Categories*. http://arxiv.org/abs/2109.01084

Jain, S., & Kumar, V. (2020). Garment Categorization Using Data Mining Techniques. *Symmetry*, *12*(6), 984. https://doi.org/10.3390/sym12060984

Kalaivani, P. (2020). Machine Learning Approach to Analyse Ensemble Models and Neural Network Model for E-Commerce Application. *Indian Journal of Science and Technology*, *13*(28), 2849–2857. https://doi.org/10.17485/IJST/v13i28.927

Kim, H., Joo, G., & Im, H. (2021). Product Category Classification using Word Embedding and GRUs. *The Journal of Korean Institute of Information Technology*, *19*(4), 11–18. https://doi.org/10.14801/jkiit.2021.19.4.11

Lee, H., & Yoon, Y. (2018). Engineering doc2vec for automatic classification of product descriptions on O2O applications. *Electronic Commerce Research*, *18*(3), 433–456. https://doi.org/10.1007/S10660-017-9268-5/METRICS

Mashalah, H. Al, Hassini, E., Gunasekaran, A., & Bhatt (Mishra), D. (2022). The impact of digital transformation on supply chains through e-commerce: Literature review and a conceptual framework. *Transportation Research Part E: Logistics and Transportation Review*, *165*, 102837. https://doi.org/10.1016/j.tre.2022.102837

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Nobre, J., & Neves, R. F. (2019). Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. *Expert Systems with Applications*, *125*, 181–194. https://doi.org/10.1016/j.eswa.2019.01.083

Ozyegen, O., Jahanshahi, H., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2022). Classifying multi-level product categories using dynamic masking and transformer models. *Journal of Data, Information and Management*, *4*(1), 71–85. https://doi.org/10.1007/s42488-022-00066-6

Patra, A., Vivek, V., Shambhavi, B. R., Sindhu, K., & Balaji, S. (2021). Product Classification in E-Commerce Sites. In *Advances in Intelligent Systems and Computing: Vol. 1299 AISC* (pp. 485–495). Springer, Singapore. https://doi.org/10.1007/978-981-33-4299-6_40

Pawłowski, M. (2022). Machine Learning Based Product Classification for eCommerce. *Journal of Computer Information Systems*, *62*(4), 730–739. https://doi.org/10.1080/08874417.2021.1910880

Perdana, S. A. P., Aji, T. B., & Ferdiana, R. (2021). Aspect Category Classification dengan Pendekatan Machine Learning Menggunakan Dataset Bahasa Indonesia. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *10*(3), 229–235. https://doi.org/10.22146/jnteti.v10i3.1819

Pothuganti, K. (2019). Open-World Classification Algorithm to Product Identification. *International Journal of Innovative Research in Computer and Communication Engineering*, *7*(12), 4282–4287. https://doi.org/10.2139/ssrn.3719055

Ristoski, P., Petrovski, P., Mika, P., & Paulheim, H. (2018). A machine learning approach for product matching and categorization. *Semantic Web*, *9*(5), 707–728. https://doi.org/10.3233/SW-180300

Sharma, P., & Sagvekar, V. R. (2023). Weighted Ensemble LSTM Model with Word Embedding Attention for E-Commerce Product Recommendation. *Journal of Communications Software and Systems*, *19*(4), 299–307. https://doi.org/10.24138/jcomss-2023-0126

Tan, L., Li, M. Y., & Kok, S. (2020). E-Commerce Product Categorization via Machine Translation. *ACM Transactions on Management Information Systems*, *11*(3), 1–14. https://doi.org/10.1145/3382189