Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Perbandingan Algoritma Naïve Bayes dan Regresi Logistik pada Klasifikasi Produk Olahraga Tokopedia Berdasarkan Gender

Veni Zahara Kartika[1], Jasmine Georgina Sekartaji[2], Evan Aryaputra[3], Taufiqurrahman Syah Effendi[4], Ericson Chandra Sihombing[5], Luluk Muthoharoh[6], Ardika Satria[7], Rizty Maulida Badri[8]

[1,2,3,4,5,6,7] *Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*
[1] veni.121450075@student.itera.ac.id
[2] jasmine.121450159@student.itera.ac.id
[3] evan.121450102@student.itera.ac.id
[4] taufiqurrahman120450051@student.itera.ac.id
[5] ericson.121450026@student.itera.ac.id
[6] luluk.muthoharoh@sd.itera.ac.id
[7] ardika.satria@sd.itera.ac.id

[8] *Jurusan Fisika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan*
[8] riztymaulidabadri@gmail.com

Corresponding author email: luluk.muthoharoh@sd.itera.ac.id

Abstract: *In the digital era, Big Data plays a crucial role in large-scale analysis and strategic decision-making. Tokopedia, one of the largest e-commerce platforms in Indonesia, generates a significant volume of data that can be utilized for analytical purposes. This study applies the Naïve Bayes and Logistic Regression algorithms using PySpark to classify data from the Tokopedia platform. PySpark is chosen for its ability to efficiently process large datasets. The results show that both algorithms perform well, with Logistic Regression achieving a higher prediction accuracy of 0.9967 compared to Naïve Bayes at 0.8684. These findings indicate that Logistic Regression is more effective for large-scale data classification. This study is expected to serve as a reference for the development of automated machine learning–based classification systems to support decision-making in the e-commerce industry.*
Keywords: *Tokopedia, PySpark, Naïve Bayes, Logistic Regression, Big Data*

Abstrak: Di era digital, *Big Data* menjadi kunci dalam analisis skala besar dan pengambilan keputusan strategis. Tokopedia, sebagai salah satu *e-commerce* terbesar di Indonesia, menghasilkan data dalam jumlah besar yang dapat dimanfaatkan untuk analisis. Penelitian ini menerapkan algoritma Naïve Bayes dan Regresi Logistik menggunakan *PySpark* untuk mengklasifikasikan data pada platform Tokopedia. *PySpark* dipilih karena kemampuannya dalam menangani data besar secara efisien. Hasil menunjukkan bahwa kedua algoritma memberikan kinerja yang baik, dengan Regresi Logistik menghasilkan akurasi prediksi lebih tinggi sebesar 0.9967 dibandingkan Naïve Bayes sebesar 0.8684. Temuan ini menunjukkan bahwa Regresi Logistik lebih efektif dalam klasifikasi data pada skala besar. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem klasifikasi otomatis berbasis machine learning untuk mendukung pengambilan keputusan dalam industri *e-commerce*.
Kata kunci: Tokopedia, *PySpark*, *Naïve Bayes*, Regresi Logistik, *Big Data*

I. PENDAHULUAN

Pada era digital yang semakin berkembang dengan pesat, kebutuhan dan penggunaan akan data terus meningkat seiring waktu ditambah dengan perkembangan internet yang semakin cepat, dengan perkembangan ini perubahan yang terjadi pada dunia internet khususnya data semakin cepat juga[1–3]. Dengan perkembangan data yang sangat cepat dan banyak maka data di era digital ini menjadi salah satu bagian penting untuk melakukan analisis dan membantu untuk pengambilan keputusan[4,5]. Pada era digital yang terus berkembang, *Big Data* menjadi salah satu elemen penting khususnya bagi perusahaan global karena perusahaan dapat melakukan analisis data dalam skala yang besar untuk

mendukung pengambilan keputusan yang lebih tepat dan strategis [6,7]. Data dengan volume yang besar dan perkembangan yang cepat disebut dengan istilah *big data* [8,9]. Layanan *big data* di Indonesia menunjukkan pertumbuhan yang positif. Hal ini berbeda dengan pemanfaatannya di kalangan masyarakat yang masih belum maksimal, meskipun pasar dengan layanan *big data* terus mengalami peningkatan sebesar 14,7% pada awal hingga pertengahan tahun 2022. Berdasarkan laporan dari IDC (*International Data Corporation*), menunjukkan bahwa pertumbuhan pasar dalam bidang *big data* dan analitik di Negara Indonesia meningkat hingga 12,5% dibandingkan awal hingga pertengahan tahun 2021 [10].

Berbagai metode pemasaran muncul dalam aliran yang tak berujung yang dipengaruhi oleh teknologi jaringan, pembelian dan penjualan barang sehingga penjualan tidak lagi terbatas pada toko fisik *offline* saja. Banyak orang telah melihat nilai komersial dan potensi pasar *e-commerce* yang telah berkembang dengan cepat [11,12]. Sejauh ini, *e-commerce* terintegrasi menjadi bagian yang tidak terpisahkan dari masyarakat. Berbagai platform *e-commerce* memiliki tingkat yang berbeda-beda, dan persaingan pasar yang sangat ketat dengan memberikan layanan dan pemasaran yang lebih akurat bagi pelanggan dan menciptakan lebih banyak nilai kekayaan. Ini hanya dapat dicapai dengan menganalisis data konsumsi konsumen [13]. Salah satu tempat yang tepat untuk mendapatkan data dalam jumlah besar yaitu dari situs *e-commerce* atau toko *online*. Pada situs toko *online* terdapat variasi data yang bisa diambil dan digunakan seperti data penjualan suatu barang, data rating penjual atau toko, data ulasan, dan masih banyak lainnya [14]. *E-commerce* juga membantu menjangkau khalayak yang lebih luas, meningkatkan jumlah pengunjung ke lembaga budaya, dan memanfaatkan data bisnis untuk pengambilan Keputusan [15].

Salah satu situs toko *online* yang banyak digunakan oleh Masyarakat Indonesia untuk berbelanja adalah Tokopedia. Tokopedia merupakan situs belanja berbasis daring yang berada di Indonesia sejak lama dan didirikan pada tahun 2009. Tokopedia menjadi platform *e-commerce* di Indonesia yang telah berkembang dengan signifikan dan mampu bersaing dengan platform *e-commerce* lainnya [16]. Perkembangan Tokopedia merupakan indikasi perubahan penting dalam lingkungan *e-commerce* di Indonesia, yang didorong oleh kerja sama strategis, inovasi teknologi, dan perluasan akses internet nasional. Data penelitian yang digunakan merupakan data barang yang dijual pada situs Tokopedia dengan kategori produk olahraga dan data pengguna Tokopedia. Dengan menggunakan data tersebut penelitian ini terdiri dari beberapa tujuan penting, yaitu mengetahui berapa jumlah produk olahraga berdasarkan gender laki-laki, perempuan, dan keduanya (*unisex*) serta melakukan prediksi dan klasifikasi produk olahraga berdasarkan gender laki-laki, perempuan, atau keduanya (*unisex*) menggunakan algoritma Naïve Bayes dan Logistic Regression.

Metode Naïve Bayes dan Logistic Regression dapat digunakan untuk melakukan klasifikasi produk. Penelitian yang dilakukan oleh Purnama dan Putra tentang Klasifikasi Penjualan Produk yang menggunakan Algoritma Naive Bayes pada bisnis Konter HP Bayu Cell menunjukkan hasil klasifikasi produk tersebut tergolong kategori "Laris" dibuktikan melalui nilai distribusi Gaussian sebesar 0,005211. Kemudian penelitian lain oleh Juwita et al. tentang prediksi Penjualan Pada Toko VJCakes Pematang Siantar didapatkan Informasi berdasarkan tabel probabilitas setiap variabelnya, disertai dengan tingkat akurasi model sebesar 83,44% terhadap data pengujian, dapat dimanfaatkan oleh pihak VJCakes sebagai dasar dalam pengambilan keputusan yang lebih optimal di masa mendatang. Penelitian oleh Ardelia et al tentang Klasifikasi Harga Ponsel Menggunakan Algoritma Logistic Regression diperoleh Kesimpulan penggunaan logistic regression dapat memprediksi kategori harga ponsel dengan tingkat akurasi yang tinggi yaitu 98% yang didapatkan dengan menggunakan perbandingan data split 90:10.

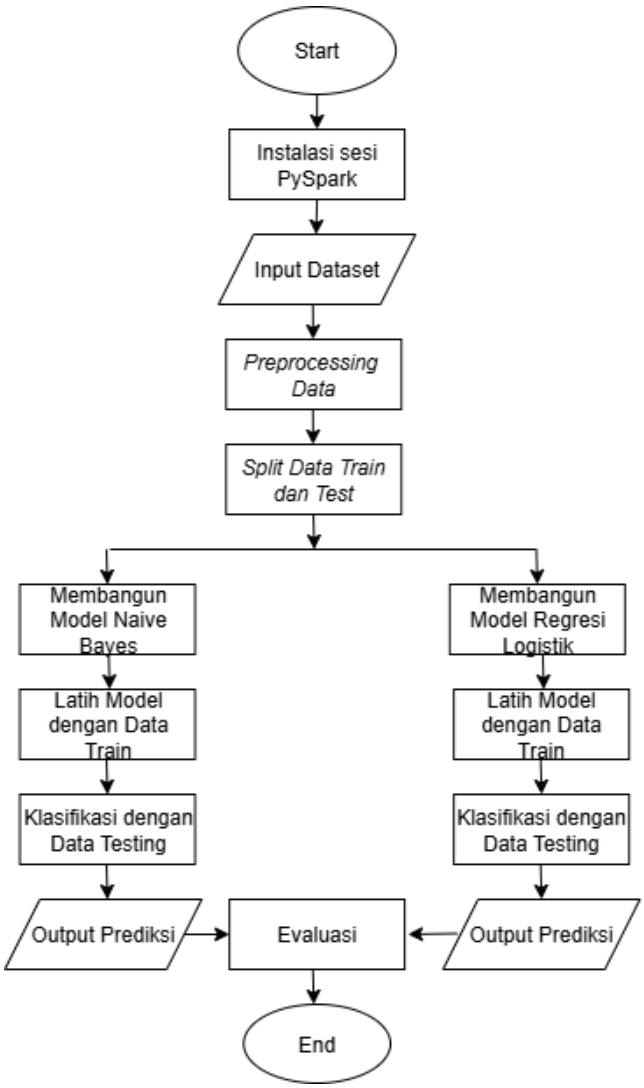
State of the art pada penelitian "Perbandingan Algoritma Naïve Bayes dan Regresi Logistik pada Klasifikasi Produk Olahraga Tokopedia Berdasarkan Gender" mengacu pada penggunaan algoritma klasifikasi seperti **Naïve Bayes** dan **Logistic Regression** yang telah terbukti efektif dalam berbagai

studi sebelumnya untuk memprediksi dan mengelompokkan data berdasarkan kategori tertentu. Dalam konteks *e-commerce*, kedua algoritma ini mampu menangani data berjumlah besar serta memberikan hasil klasifikasi yang cukup akurat. Penelitian-penelitian terdahulu menunjukkan bahwa Naïve Bayes efektif dalam memproses data kategorikal dengan distribusi probabilistik, sementara Logistic Regression unggul dalam prediksi dengan variabel numerik dan menghasilkan akurasi tinggi. Penerapan kedua metode ini pada data Tokopedia, khususnya pada produk olahraga berdasarkan gender, menjadi kontribusi baru yang relevan dengan perkembangan teknologi digital dan tren belanja *online* di Indonesia, sekaligus memperkaya literatur tentang klasifikasi produk dalam *e-commerce* lokal.

II. METODE

Penelitian ini menggunakan beberapa tahapan metode dalam pelaksanaannya sebagai berikut.

II.1. Alur Tahapan Penelitian



Gambar 1. Flowchart Penelitian

II.2. Pengumpulan Data

Penelitian ini menggunakan data terkait nama-nama produk olahraga yang diperoleh dari Tokopedia, dengan cara menggunakan ekstensi Chrome "*Web Scraper*" dan implementasi *Infinite Scroll* untuk menangani pemuatan konten dinamis. Data ini diekstrak dalam format CSV yang berisi informasi seperti nama, jenis, kategori, dan produk.

II.3. Pra-proses Data

Pra-proses data dilakukan sebagai langkah awal dalam melakukan analisis teks untuk menghasilkan model yang akurat dan informatif. Dalam konteks analisis produk olahraga Tokopedia, pra-proses data meliputi beberapa tahapan berikut:

1. Pembersihan teks yang meliputi:
   - Menghilangkan karakter non-ASCII;
   - Menghilangkan *hyperlink* dan *mentions*;
   - Menghilangkan tanda baca dan simbol khusus;
   - Menghilangkan angka;
   - Menghilangkan karakter tunggal;
   - Menghilangkan spasi berlebihan;
   - Mengubah teks kedalam format huruf kecil.
2. Tokenisasi: Proses mengubah teks menjadi kata - kata.
3. *Stemming/Lemmatization*: Proses pengurangan kata menjadi bentuk dasarnya.

II.4. Pemberian Label

Label produk diberikan berdasarkan kata-kata yang muncul pada deskripsi produk, dengan ketentuan sebagai berikut:

- Label 1: Produk ditujukan untuk wanita (mengandung kata "wanita", "perempuan", "girl", "ladies", atau "cewek").
- Label 2: Produk ditujukan untuk pria (mengandung kata "laki-laki", "pria", "boys", "gentlemen", atau "cowok").
- Label 0: Produk ditujukan untuk *unisex* (tidak mengandung kata-kata yang disebutkan pada Label 1 dan Label 2, atau mengandung keduanya).

II.5. Klasifikasi

Tahap pertama dalam melakukan klasifikasi teks adalah membagi data ke dalam dua kelompok, yakni sebanyak 60% dialokasikan sebagai data pelatihan (*training*) untuk melatih model klasifikasi dalam mengenali data dan melakukan klasifikasi, dan 40% sisanya sebagai data pengujian (*testing*) yang digunakan untuk mengevaluasi kinerja model. Penelitian ini mengimplementasikan dua model klasifikasi yang dijabarkan sebagai berikut.

1. Naive Bayes
   Model klasifikasi probabilistik yang efektif untuk teks pendek dengan menggunakan teorema bayes "*Theorem of Probability*".

$$p(A|B) = \frac{p(B|A)\,.p(A)}{p(B)} \quad (1)$$

Keterangan:

$p(A|B)$: Probabilitas bersyarat A terjadi jika B terjadi

$p(B|A)$: Probabilitas bersyarat B terjadi jika A terjadi

$p(A)$: Probabilitas kejadian A

$p(B)$: Probabilitas kejadian B

2. Regresi Logistik
   Model klasifikasi linear yang dapat digunakan untuk memprediksi probabilitas kelas dan mengidentifikasi keterkaitan antara variabel bebas dan variabel terikat yang berupa data kategorik [2].

$$P(Y = 1|X) = \frac{e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p)}}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p)}} \quad (2)$$

Keterangan:

Y : Variabel dependen dengan nilai 0, 1, dan 2

$X_1, X_2, \ldots, X_p$: Variabel independen

$\beta_0, \beta_1, \ldots, \beta_p$: Koefisien regresi

e : Eksponen

II.6. Evaluasi Model:

Performa model dievaluasi menggunakan data *testing*. Matriks evaluasi yang digunakan yakni *confusion Matrix* yang menunjukkan distribusi aktual dan prediksi label. Dalam *confusion matrix* terdapat beberapa metrik performa yang bisa didapatkan. Pertama adalah akurasi, akurasi merupakan perhitungan yang memberikan informasi seberapa akurat model dapat melakukan klasifikasi dengan tepat. Lalu berikutnya ada *precision* yang memberikan informasi tentang tingkat keakuratan data hasil prediksi oleh model dengan data aktualnya. Kemudian ada *recall* merupakan evaluasi yang memberikan informasi model dalam memprediksi kelas positif dengan benar. Lalu yang terakhir ada *f1-score* memberikan informasi tentang keseimbangan antara *precision* dan *recall*. Berdasarkan *confusion matrix* berikut persamaan yang digunakan untuk nilai akurasi, *precision*, *recall*, dan *f1-score*:

$$Akurasi = \frac{Jumlah\ prediksi\ benar}{Total\ jumlah\ data} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

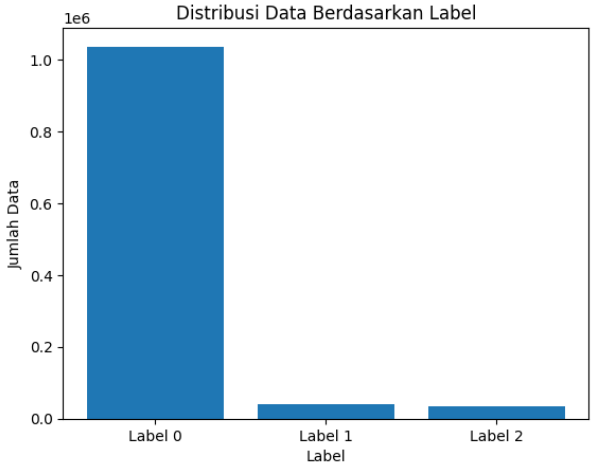
$$F1 = 2 \times \frac{Precision * Recall}{Precision+Recall} \quad (6)$$

Keterangan:

- *True Positive* (TP): kondisi ketika data aktual positif diprediksi dengan benar oleh model.
- *True Negative* (TN): kondisi ketika data aktual negatif diprediksi dengan benar oleh model.
- *False Positive* (FP): kondisi ketika data aktual negatif diprediksi sebagai data positif.
- *False Negative* (FN): kondisi ketika data aktual positif diprediksi sebagai data negatif.

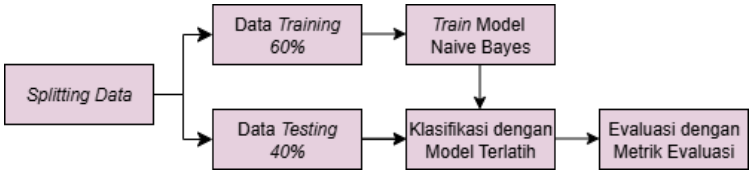
III. HASIL DAN PEMBAHASAN

Dataset "produk_olahraga_tokopedia.csv" terdiri dari 1,048,576 baris dan 4 kolom dengan beberapa atribut yaitu nama, jenis, kategori, dan produk. Dalam pemrosesannya, dilakukan beberapa tahapan kategori seperti text_non_asci, *removed*, dan *clean* terkait erat dengan proses pembersihan dan *preprocessing* teks dalam pemrosesan data. Atribut-atribut tersebut adalah hasil dari proses pembersihan dan *preprocessing* teks yang dilakukan yang keseluruhan bertujuan untuk menghasilkan teks yang lebih konsisten dan mudah diproses dalam analisis berikutnya. Selanjutnya dilakukan pemberian label pada data produk olahraga untuk laki-laki, perempuan, dan keduanya (*unisex).* Produk dikategorikan sebagai untuk laki-laki, perempuan, atau keduanya (unisex) berdasarkan kata kunci yang terdapat dalam deskripsi produk. Kata kunci yang digunakan untuk pelabelan ini meliputi: untuk perempuan adalah 'wanita', 'perempuan', 'girl', 'ladies', dan 'cewek'; untuk laki-laki adalah 'laki-laki', 'pria', 'boys', 'gentlemen', dan 'cowok'. Jika deskripsi produk mengandung kata-kata yang menunjukkan perempuan, dan juga mengandung kata-kata yang menunjukkan laki-laki, maka produk tersebut diberi label *unisex*. Jika hanya mengandung kata-kata untuk perempuan, maka diberi label untuk perempuan, dan jika hanya mengandung kata-kata untuk laki-laki, maka diberi label untuk laki-laki. Hasil yang didapatkan seperti Gambar 2.



Gambar 2. Hasil Distribusi Data Berdasarkan Label

Setelah melalui proses pelabelan, data akan diproses dengan model pertama yakni model Naïve Bayes yang merupakan salah satu algoritma klasifikasi yang menggunakan Teorema Bayes dan memiliki asumsi setiap fitur bersifat saling bebas terhadap satu sama lain. Berikut adalah langkah-langkah dalam penerapan klasifikasi Naïve Bayes seperti pada Gambar 3.



Gambar 3. Alur Pemodelan Naïve Bayes

- Melakukan split data dengan random splitnya [0.6, 0.4].
- Membuat model Naïve Bayes kemudian melatihnya dengan data *train*.
- Lalu melakukan klasifikasi teks pada data *test* dengan model yang sudah dibangun. Sehingga didapat seperti pada Gambar 4 berikut.

clean_text	label	prediction
alat sonar deteksi ikan portable portable fish finder	0	0.0
laser rangefinder the truth model bushnell	0	0.0
tas consina millau bridge	0	0.0
teropong monokular focus and zoom lens	0	0.0
free ongkir berwyn samsak tinju berdiri adjustable tinggi diatur	0	0.0
promo papan skor futsal jual papan score futsal digital murah	0	0.0
beauty tool pcs curved side release buckle paracord bracel	0	0.0
topi atas bolong topi lari tenis golf olahraga outdoor sport cewek	1	1.0
umpan pancing bionik bahan plastik abs ukuran cm	0	0.0
great pcs buatan lembut frog cm swimbait fishing	0	0.0

only showing top 10 rows

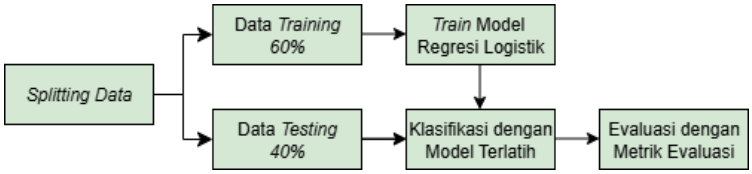
Gambar 4. Hasil Prediksi Naive Bayes

- Menghitung metrik evaluasi dan didapatkan hasil akurasinya 0.8684.
- Menghitung *weighted precision* dari klasifikasi Naïve Bayes dan didapatkan 0.9509.
- Menghitung *weighted recall* dari klasifikasi Naïve Bayes dan didapatkan 0.8684.
- Menghitung *weighted f1-score* dari klasifikasi Naïve Bayes dan didapatkan 0.8962.
- Menampilkan *confusion matrix* dari hasil prediksi model klasifikasi seperti pada Gambar 5 berikut.

label_prediction	0.0	1.0	2.0
0	358307	20686	35663
1	830	15319	175
2	1084	101	12673

Gambar 5. *Confusion Matrix* Model Naïve Bayes

Berdasarkan hasil prediksi pada *confusion matrix* diatas didapat bahwa model Naïve Bayes dapat memprediksi label 0 dengan benar sebanyak 358307 data, label 1 sebanyak 15319 data, dan label 2 sebanyak 12673 data. Setelah melalui pemodelan Naïve Bayes, selanjutnya dilakukan kembali pemodelan dengan model Regresi Logistik yang menjelaskan keterkaitan hubungan variabel dependen yang bersifat kategorik dengan variabel independen yang dapat berupa data kategorik maupun interval. Berikut adalah langkah-langkah dalam penerapan klasifikasi Regresi Logistik seperti pada Gambar 6.



Gambar 6. Alur Pemodelan Naïve Bayes

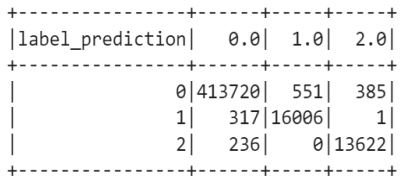
- Melakukan split data dengan random splitnya [0.6, 0.4].
- Membuat model Regresi Logistik kemudian melatihnya dengan data *train*.
- Lalu melakukan klasifikasi teks pada data *test* dengan model yang sudah dibangun. Sehingga didapat seperti pada Gambar 7.

clean_text	label	prediction
alat sonar deteksi ikan portable portable fish finder	0	0.0
laser rangefinder the truth model bushnell	0	0.0
tas consina millau bridge	0	0.0
teropong monokular focus and zoom lens	0	0.0
free ongkir berwyn samsak tinju berdiri adjustable tinggi diatur	0	0.0
promo papan skor futsal jual papan score futsal digital murah	0	0.0
beauty tool pcs curved side release buckle paracord bracel	0	0.0
topi atas bolong topi lari tenis golf olahraga outdoor sport cewek	1	1.0
umpan pancing bionik bahan plastik abs ukuran cm	0	0.0
great pcs buatan lembut frog cm swimbait fishing	0	0.0

only showing top 10 rows

Gambar 7. Hasil Prediksi Regresi Logistik

- Menghitung metrik evaluasi dan didapatkan hasil akurasinya 0.9967.
- Menghitung *weighted precision* dari klasifikasi Regresi Logistik dan didapatkan 0.9967.
- Menghitung *weighted recall* dari klasifikasi Regresi Logistik dan didapatkan 0.9967.
- Menghitung *weighted f1-score* dari klasifikasi Regresi Logistik dan didapatkan 0.9967.
- Menampilkan *confusion matrix* dari hasil prediksi model klasifikasi seperti pada Gambar 8**.**



label_prediction	0.0	1.0	2.0
0	413720	551	385
1	317	16006	1
2	236	0	13622

Gambar 8. Confusion Matrix Regresi Logistik

Berdasarkan hasil prediksi didapat bahwa model regresi logistik dapat memprediksi label 0 dengan benar sebanyak 413720 data, label 1 sebanyak 16006 data, dan label 2 sebanyak 13622 data. Model regresi logistik menunjukkan performa yang cukup baik dalam memprediksi ketiga label (0, 1, dan 2), dengan jumlah prediksi benar terbanyak pada label 0 sebanyak 413.720 data. Hal ini menunjukkan bahwa distribusi data kemungkinan besar tidak seimbang, dengan label 0 (kemungkinan mewakili kategori dominan, seperti produk unisex atau kategori yang paling umum dijual) memiliki jumlah data yang secara signifikan lebih besar dibandingkan dengan label 1 dan label 2. Jumlah prediksi benar untuk label 1 dan label 2 masing-masing adalah 16.006 dan 13.622, yang relatif jauh lebih kecil dibanding label 0. Ini mengindikasikan bahwa model lebih mudah mengenali pola dari label 0 karena datanya lebih banyak (bias data) dan kemungkinan model kurang optimal dalam membedakan fitur khusus antara label 1 dan label 2, terutama jika jumlah data latih untuk kedua label tersebut terbatas. Hal ini memberikan beberapa implikasi seperti kinerja model terlihat tinggi pada kelas mayoritas (label 0), namun perlu dievaluasi ulang untuk minoritas. Diperlukan evaluasi metrik lain seperti *precision*, *recall*, dan *F1-score* per label agar performa model dapat dinilai secara adil, terutama untuk label yang lebih sedikit jumlahnya. Penerapan teknik penanganan data tidak seimbang seperti *oversampling* (SMOTE), *undersampling*, atau penyesuaian bobot kelas dapat membantu meningkatkan prediksi untuk label 1 dan 2. Secara keseluruhan, meskipun akurasi total mungkin terlihat tinggi karena dominasi label 0, namun evaluasi yang lebih mendalam per kelas diperlukan untuk memastikan model bekerja baik di seluruh kategori.

Hasil dari pemodelan, didapatkan perbandingan kinerja antara metode klasifikasi menggunakan algoritma Naïve Bayes dan Regresi Logistik dapat ditampilkan secara kuantitatif dengan mempertimbangkan parameter-parameter evaluasi seperti akurasi, presisi, *recall*, dan *f1-score*. Sehingga memperoleh pemahaman yang komprehensif terkait performa kedua algoritma dalam melakukan prediksi pada kasus klasifikasi nama produk olahraga Tokopedia berdasarkan jenis kelamin. Penggunaan metrik-metrik ini memiliki perbandingan objektif dan memungkinkan penarikan kesimpulan yang valid mengenai keunggulan dari metode yang diterapkan. Hasil perbandingan ini dapat dilihat pada tabel perbandingan hasil klasifikasi Naïve Bayes dan Regresi Logistik.

Berdasarkan tabel perbandingan hasil klasifikasi Naïve Bayes dan Regresi Logistik, dapat dianalisis beberapa hal penting. Metrik akurasi menunjukkan bahwa Regresi Logistik menunjukkan kinerja yang lebih unggul dengan skor 0,9967 dibandingkan Naïve Bayes yang hanya mencapai 0,8684. Ini mengindikasikan bahwa Regresi Logistik dapat mengklasifikasikan data dengan tingkat keakuratan yang lebih tinggi secara keseluruhan. Pada metrik *Weighted Precision*, Regresi Logistik juga unggul dengan skor 0,9967, sedangkan Naïve Bayes memiliki skor yang sedikit lebih rendah, yaitu 0,9509. Hal ini menunjukkan bahwa Regresi Logistik lebih akurat dalam menghasilkan prediksi positif yang relevan, terutama pada kelas mayoritas[17-19]. Untuk metrik *Weighted Recall*, Regresi Logistik kembali mengungguli Naïve Bayes dengan skor 0,9667 dibandingkan 0,8684. Ini menandakan bahwa

Regresi Logistik lebih efektif dalam mengidentifikasi *instance* positif dari semua kelas, terutama pada kelas minoritas. Skor *Weighted F1-Score*, yang merupakan nilai rata-rata dari presisi dan *recall*, juga lebih tinggi pada Regresi Logistik 0,9667 dibandingkan Naïve Bayes 0,8962. Ini mengkonfirmasi kinerja yang lebih baik dari Regresi Logistik secara keseluruhan. *Confusion matrix* menyajikan informasi lebih rinci terhadap jumlah prediksi yang benar serta prediksi yang salah untuk setiap kelas [20-23]. Dari matriks tersebut, terlihat bahwa Regresi Logistik memiliki jumlah prediksi yang benar (pada diagonal utama) lebih tinggi dibandingkan Naïve Bayes. Berdasarkan kombinasi skor akurasi, *weighted precision, recall*, dan *F1-Score* serta analisis *confusion matrix*, Regresi Logistik terbukti lebih andal dalam klasifikasi produk olahraga berdasarkan gender di Tokopedia.

Tabel 1. Tabel Perbandingan Hasil Klasifikasi Naïve Bayes dan Regresi Logistik

Metrik	Klasifikasi Naïve Bayes	Klasifikasi Regresi Logistik
Akurasi	0.8684	0.9967
Weighted Precision	0.9509	0.9967
Weighted Recall	0.8684	0.9967
Weighted F1-Score	0.8962	0.9967

Temuan ini sejalan dengan berbagai studi sebelumnya yang membandingkan kedua algoritma tersebut dalam berbagai domain berbeda. Sebagai contoh, penelitian yang dilakukan oleh Sheikh et al. (2024) dalam hal mendeteksi berita palsu menunjukkan bahwa Regresi Logistik menunjukkan tingkat akurasi yang lebih tinggi, yaitu sebesar 98%, bila dibandingkan dengan algoritma Naïve Bayes yang mencapai sekitar 93% [24]. Demikian pula, dalam studi oleh Atchaya dan Somasundaram (2023) mengenai deteksi penipuan kartu kredit, Regresi Logistik menunjukkan akurasi 93,59%, mengungguli Naïve Bayes yang mencapai 85,88%[25]. Namun, terdapat juga penelitian yang menunjukkan terdapat performa sebaliknya. Misalnya, dalam penelitian yang telah dilakukan oleh Reddy dan Kumar (2023) mengenai prediksi serangan jaringan, Naïve Bayes menunjukkan akurasi 85,90%, lebih tinggi dibandingkan dengan Regresi Logistik yang hanya mencapai 65,30%[26]. Perbedaan performa ini menunjukkan bahwa efektivitas masing-masing algoritma sangat bergantung pada karakteristik data dan konteks aplikasinya. Dalam kasus klasifikasi produk olahraga berdasarkan gender di Tokopedia, Regresi Logistik tampaknya lebih mampu menangani kompleksitas data dan memberikan hasil yang lebih akurat dibandingkan dengan hasil klasifikasi menggunakan Naïve Bayes [27].

IV. KESIMPULAN

Berdasarkan hasil penelitian menunjukkan bahwa Regresi Logistik memiliki kinerja yang lebih baik dibandingkan Naïve Bayes dibuktikan dengan akurasi yang lebih tinggi sebesar 0.9967 dalam mengklasifikasikan nama produk olahraga Tokopedia berdasarkan jenis kelamin, dengan akurasi prediksi yang lebih tinggi. Hal ini menunjukkan bahwa Regresi Logistik lebih baik dalam menangani data dengan kompleksitas moderat pada skala besar. Meskipun demikian, pemilihan algoritma klasifikasi yang optimal tetap harus mempertimbangkan karakteristik data, tujuan analisis, dan kebutuhan spesifik dari penelitian. Temuan ini diharapkan dapat menjadi acuan bagi pengembangan

sistem klasifikasi berbasis *machine learning* dalam konteks *e-commerce*, khususnya dalam pengolahan data besar secara efisien.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Tim SENADA yang telah berkontribusi dalam penyusunan dan penyediaan template ini.

REFERENSI

1. T. W. E. Suryawijaya, "Memperkuat Keamanan Data melalui Teknologi Blockchain: Mengeksplorasi Implementasi Sukses dalam Transformasi Digital di Indonesia," *J. Stud. Kebijak. Publik*, vol. 2, no. 1, pp. 55–68, 2023, doi: 10.21787/jskp.2.2023.55-68.
2. E. A. Wulandari, "Menganalisis Sejarah Perkembangan Perekonomian," vol. 3, no. 1, pp. 1–9, 2024, doi: 10.22437/krinok.v3i1.27465.
3. S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electron. Mark.*, vol. 26, no. 2, pp. 173–194, 2016, doi: 10.1007/s12525-016-0219-0.
4. A. Sobandi, "Pengolahan Data Dalam Sistem Informasi," *Manajerial J. Manaj. dan Sist. Inf.*, vol. 1, no. 1, pp. 89–95, 2002, [Online]. Available: https://ejournal.upi.edu/index.php/manajerial/article/view/16450
5. R. Putri, D. Prasetya, R. N. Azizah, J. Bayu, W. Halwa, and R. H. Nugroho, "Implementasi Penggunaan Data Analytics untuk Mengoptimalkan Pengambilan Keputusan Bisnis di Era Digital," vol. 2, no. 2, pp. 1–12, 2025.
6. A. Kumar, N. Kumar, and K. B. Thapa, "Tunable broadband reflector and narrowband filter of a dielectric and magnetized cold plasma photonic crystal," *Eur. Phys. J. Plus*, vol. 133, no. 7, 2018, doi: 10.1140/epjp/i2018-12073-3.
7. P. A. Narendra, "Perusahaan Big Data di Indonesia Era Digital," *mdmedia*, 2024. https://mdmedia.co.id/whats-up/149/artikel/perusahaan-big-data-di-indonesia-era-digital (accessed May 24, 2025).
8. N. S. Putra, H. Ritchi, and A. Alfian, "Hubungan Big Data Analytics terhadap Kualitas Audit: Penerapan pada Instansi Pemerintah," *J. Ris. Akunt. dan Keuang.*, vol. 11, no. 1, pp. 57–72, 2023, doi: 10.17509/jrak.v11i1.55139.
9. A. B. HARYANTO and E. SETIAWAN, "Impact Of Big Data Analytics On Audit Quality With Audit Delay As Mediator," *Int. J. Environ. Sustain. Soc. Sci.*, vol. 5, no. 4, pp. 814–821, 2024, doi: 10.38142/ijesss.v5i4.1120.
10. Rita Puspita Sari, "Pasar Big Data di Indonesia Masih Terus Berkembang," *Cloudd Computing Indonesia*, 2024. https://www.cloudcomputing.id/berita/big-data-di-indonesia
11. R. Tahir *et al.*, *Transformasi Bisnis di Era Digital (Teknologi Informasi dalam Mendukung Transformasi Bisnis di Era Digital)*, no. August. 2023.
12. P. K. Perdagangan, "Perdagangan Digital (E-Commerce) Indonesia Periode 2023," pp. 1–8, 2024.
13. L. Wang, T. R. A. L. Pertheban, T. Li, and L. Zhao, "Application of business intelligence based on big data in E-commerce data evaluation," *Heliyon*, vol. 10, no. 21, p. e38768, 2024, doi: 10.1016/j.heliyon.2024.e38768.
14. S. A. Rajagukguk, S. Mutmainah, and A. Satria, "Analisis Sentimen Pembelajaran Tatap Muka dengan Apache SPARK," *J. Rekayasa Teknol. Inf.*, vol. 6, no. 2, p. 159, 2022, doi: 10.30872/jurti.v6i2.8162.
15. R. Novendra, Y. Turnandes, G. Nugroho, M. Revnu, and M. R. Manahan, "Systematic Literature Review : Digital Marketing Strategy Tinjauan Pustaka Sistematis : Strategi Digital Marketing," vol. 5, no. 2, pp. 9075–9094, 2024.
16. Y. Tiara Putri, R. Kusumadewi, and E. Saefulloh, "PENGARUH KREDIBILITAS INFLUENCER DAN BRAND AWARENESS TERHADAP MINAT PEMBELIAN DI TOKOPEDIA (Studi Pada Pelanggan Tokopedia yang Bertransaksi Melalui Bank Syariah Indonesia)," *Entrep. J. Bisnis Manaj. dan Kewirausahaan*, vol. 4, no. 2, pp. 205–225, 2023, doi: 10.31949/entrepreneur.v4i2.5651.
17. M. C. Martini, E. Pelle, F. Poggi, and A. Sciandra, "The role of citation networks to explain academic promotions: an empirical analysis of the Italian national scientific qualification," *Scientometrics*, vol. 127, no. 10, pp. 5633–5659, 2022, doi: 10.1007/s11192-022-04485-5.
18. E. da Silva Rocha, F. L. de Morais Melo, M. E. F. de Mello, B. Figueiroa, V. Sampaio, and P. T. Endo, "On

usage of artificial intelligence for predicting mortality during and post-pregnancy: a systematic review of literature," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–17, 2022, doi: 10.1186/s12911-022-02082-3.
19. P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, p. 100032, 2021, doi: 10.1016/j.cmpbup.2021.100032.
20. A. Riansah *et al.*, "PENERAPAN ALGORITMA RANDOM FOREST DAN DECISION TREE," vol. 9, no. 3, pp. 4242–4249, 2025.
21. S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," no. December, 2024, doi: 10.53555/AJBR.v27i4S.4345.
22. A. Abdulkareem, T. Anyim, O. Popoola, J. Abubakar, and A. Ayoade, "Prediction of Induction Motor Faults Using Machine Learning," *Heliyon*, vol. 11, no. 1, p. e41493, 2025, doi: 10.1016/j.heliyon.2024.e41493.
23. M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Comput. Geosci.*, vol. 184, no. June 2023, p. 105531, 2024, doi: 10.1016/j.cageo.2024.105531.
24. N. Sheikh, S. Parveen, A. Rehman, A. Naeem, M. Anjum, and M. Yasin, "Comparative Study of Fake News Detection using Navie Bayes and Logistic Model," vol. 4883, no. July, pp. 1008–1015, 2024, doi: 10.53555/ks.v12i5.3390.
25. P. Atchaya and K. Somasundaram, "Novel Logistic Regression over Naive Bayes Improves Accuracy in Credit Card Fraud Detection," *J. Surv. Fish. Sci.*, vol. 10, no. 1S, pp. 2172–2181, 2023, [Online]. Available: https://sifisheriessciences.com/journal/index.php/journal/article/view/450
26. K. Saikrishna and S. A. Kumar, "Improvement of Naive Bayesian Classification over Logistic Regression for Network Attack Prediction Accuracy," vol. 10, no. Sadineni 2020, pp. 1793–1802, 2023.
27. O. Gumus, E. Yasar, Z. P. Gumus, and H. Ertas, "Comparison of different classification algorithms to identify geographic origins of olive oils," *J. Food Sci. Technol.*, vol. 57, no. 4, pp. 1535–1543, 2020, doi: 10.1007/s13197-019-04189-4.